Optimal Sleeping Mechanism for Multiple Servers With MMPP-Based Bursty Traffic Arrival

Zhiyuan Jiang, Bhaskar Krishnamachari, Sheng Zhou, and Zhisheng Niu, *Fellow, IEEE*

***Abstract*—A fundamental problem in green communications and networking is the operation of servers (routers or base stations) with sleeping mechanism to optimize energy-delay tradeoffs. This problem is very challenging when considering realistic bursty (non-Poisson) traffic. We prove for the first time that the optimal structure of such a sleeping mechanism for multiple servers when the arrival of jobs is modeled by a bursty Markov-modulated Poisson process (MMPP). It is shown that the optimal operation, which determines the number of active (or sleeping) servers dynamically, is hysteretic and monotone, and hence, it is a queue-threshold-based policy. This letter settles a conjecture in the literature that the optimal sleeping mechanism for a single server with interrupted Poisson arrival process, which can be treated as a special case of MMPP, is queue-threshold-based. The exact thresholds are given by numerically solving the Markov decision process.**

***Index Terms*—Green wireless communications, Markov-modulated-Poisson-process, Markov decision process, threshold-based policy.**

I. Introduction

THE REDUCTION of *energy consumption* has attracted more and more attention in several engineering fields, e.g., wireless communication systems and data centers. One of the most effective approaches is to put idle servers into sleeping mode due to the fact that a significant amount of energy is wasted by keeping the idle servers active. Concretely, a base station (BS) consumes 90% of its peak power even when the traffic load is low [1], and a typical idle server consumes 50%-60% of its peak power. On the other hand, the utilization of BSs and servers is usually low, especially with more and more densely deployed infrastructures [2]. Meanwhile, the energy consumption reduction comes with an undesired user *delay* increase, due to the extra job queuing time with possibly sleeping servers. Therefore, the design of sleeping mechanism should consider the tradeoff between energy consumption and queuing delay, and in the meantime avoid frequent server mode switching which costs extra energy.

In practice, the arrival traffic at servers often exhibits a high level of *burstness* [3], whereas existing works usually focus on Poisson-based, non-bursty traffic arrivals. It is important to understand the impact of traffic burstness since, intuitively, it may create more sleeping opportunities. However, the optimization of the energy-delay tradeoff with bursty, non-Poisson traffic becomes very challenging and hence few results are available.

A. Related Work and Main Contributions

The BS and server sleeping mechanisms have attracted wide attention in the literature. It is proved by Kamitsos *et al.* [4] that the optimal structure of the sleeping operations for Poisson arrival and a single server is queue-threshold-based. The proof is built upon the previous work by Lu and Serfozo [5] and Hipp and Holzbaur [6]. In the work by Wu *et al.* [7] and Leng *et al.* [8], the arrival traffic pattern is generalized to interrupted Poisson process (IPP) to capture the traffic burstness. In an IPP process, jobs only arrive during the ON phase and the ON and OFF phases transit to each other based on a Markov process. Specifically, Wu *et al.* [7] calculate the optimum queue threshold with IPP arrival and a single server by fixing the sleeping policy to be N-based, i.e., turning on the server when there are N jobs and turning it off when the queue is empty. However, it is shown in the work by Leng *et al.* [8] that the optimal sleeping policy with IPP arrival has two sets of thresholds, meaning that in each phase of the IPP the thresholds to turn on and off the server are different. Therefore, the N-based policy [7] is in general not optimal with IPP arrival. Towards finding the optimal sleeping policy, Leng *et al.* [8] adopt a partially observable Markov decision process (POMDP) formulation and analyze the optimal sleeping policy with IPP arrivals and a single server numerically. It is proved that the optimal policy is hysteretic but the monotonicity property, which together with the hysteretic property proves the optimal policy structure to be queue-threshold-based, is left as a conjecture.

In this letter, we generalize the existing work by considering the Markov-modulated-Poisson-process (MMPP) traffic arrival and multiple servers. The optimal sleeping policy structure is proved to be queue-threshold-based, and hence the conjecture by Leng *et al.* [8, Conjecture 1] is settled since IPP and a single server can be considered as a special case. Numerical results are also given to shed light upon the optimum queue thresholds.

Manuscript received October 3, 2017; revised November 21, 2017; accepted December 1, 2017. Date of publication December 12, 2017; date of current version June 19, 2018. This work was supported in part by the Nature Science Foundation of China under Grant 61701275, Grant 91638204, Grant 61571265, and Grant 61621091, in part by the China Post-Doctoral Science Foundation, and in part by the Hitachi Research and Development Headquarter. The associate editor coordinating the review of this paper and approving it for publication was M. Dong. *(Corresponding author: Sheng Zhou.)*

Z. Jiang, S. Zhou, and Z. Niu are with the Tsinghua National Laboratory for Information Science and Technology, Tsinghua University, Beijing 100084, China (e-mail: zhiyuan@tsinghua.edu.cn; sheng.zhou@tsinghua.edu.cn; niuzhs@tsinghua.edu.cn).

B. Krishnamachari is with the Ming Hsieh Department of Electrical Engineering, University of Southern California, Los Angeles, CA 90089 USA (e-mail: bkrishna@usc.edu).

Digital Object Identifier 10.1109/LWC.2017.2782252

II. System Model and Problem Formulation

We consider M servers, each has two operation modes, active and sleeping. The M servers serve jobs in a single queue with a buffer size of B. Denote the number of active servers as W, and $W \in \{0, \ldots, M\}$. We assume that jobs arrive at the queue according to an MMPP to capture the burstiness of the traffic. Jobs arrive during the S-th phase of MMPP based on the Poisson process with rate λ_S, where the MMPP arrival phase is denoted as k_S and $S \in \{1, \ldots, N\}$. The MMPP is parameterized by the N-state continuous time Markov chain with phase transition matrix as $\boldsymbol{R} = \begin{bmatrix} -\sigma_1 & \cdots & \sigma_{1N} \\ \vdots & \ddots & \vdots \\ \sigma_{N1} & \cdots & -\sigma_N \end{bmatrix}$, where $\sigma_{S_1S_2}$ denotes the transition rate from phase k_{S_1} to k_{S_2} of MMPP, and $\sigma_i = \sum_{j=1, j\neq i}^{N} \sigma_{ij}$. The service time is assumed to be independently and identically distributed according to an exponential distribution over jobs with mean service time of μ^{-1} for each active server. Based on the queuing theory, the service rate for W active servers is $W\mu$. The memoryless property of the arrival[1] and departure processes enables us to formulate the problem as a continuous-time MDP. The system state is denoted as (S, Q, W), where $S \in \{1, \ldots, N\}$, $W \in \{0, \ldots, M\}$ and $Q \in \{0, \ldots, B\}$. The state (S, Q, W) denotes that there are Q jobs in the queue, the number of active servers is W and the arrival MMPP is in the S-th phase. The control action space is $\{0, \ldots, M\}$, wherein an action u_a turns a servers to the active mode.[2] In the case of a is smaller than the number of current active servers (W), the action means to turn $W - a$ servers to the sleeping mode.

We adopt the discrete-time approximation of the continuous-time MDP, whereby the time is divided into time slots and time duration of each time slot, i.e., denoted by Δ, is sufficiently small such that there is at most one event (job arrival, departure, or arrival phase shift) occurrence in one time slot [9, Chapter 5.5]. The decision is made at each time slot, and the time index is conveyed in the brackets. The system states evolve as follows

$$W(t+1) = a(t), \tag{1}$$

$$S(t+1) = \begin{cases} \bar{S}, & \text{if arrival phase transits to } k_{\bar{S}} \text{ phase;} \\ S(t), & \text{no phase transition happens,} \end{cases} \tag{2}$$

$$Q(t+1) = \begin{cases} Q(t)+1, & \text{if } Q(t) < B \text{ and a job arrives;} \\ Q(t)-1, & \text{if } Q(t) > 0 \text{ and a job is served;} \\ Q(t), & \text{otherwise.} \end{cases} \tag{3}$$

The state transition probability given action u_a is (the time index is omitted for simplicity)

$$\Pr\{(S, Q, W) \to (S, Q+1, a)\} = \lambda_S \Delta \mathbb{1}(Q < B). \tag{4}$$

$$\Pr\{(S, Q, W) \to (S, Q-1, a)\} = a\mu \Delta \mathbb{1}(Q > 0). \tag{5}$$

$$\Pr\{(S, Q, W) \to (\bar{S}, Q, a)\} = \sigma_{S\bar{S}} \Delta,\ \bar{S} \in \mathcal{N}_S. \tag{6}$$

$$\begin{aligned} &\Pr\{(S, Q, W) \to (S, Q, a)\} \\ &= 1 - a\mu\Delta\mathbb{1}(Q > 0) - \lambda_S \Delta \mathbb{1}(Q < B) - \sum_{\bar{S} \in \mathcal{N}_S} \sigma_{S\bar{S}} \Delta, \end{aligned} \tag{7}$$

where $\mathcal{N}_S = \{1, \ldots, N\} \backslash \{S\}$. All other transition probabilities are zeros.

We consider the active energy consumption cost, switching energy cost, and delay cost of the system. The objective is to minimize the total discounted cost [5], [7], [8], i.e.,

$$\begin{aligned} \min_{a(t) \in \{0, \ldots, M\}} \mathbb{E}\Bigg[\sum_{t=1}^{\infty} r^{t-1} &(\max(a(t) - W(t), 0) E_{\text{sw}} \\ &+ \omega Q(t) + a(t) E_{\text{on}} \Bigg], \end{aligned} \tag{8}$$

where the switching energy consumption is denoted as E_{sw}, E_{on} denotes the energy consumption of the server being active for one time slot, $r \in [0, 1)$ is the discount factor which reflects how important the immediate cost is, and the tradeoff between delay and energy cost is represented by ω. Although only the server start-up energy consumption is considered for switching energy cost based on real systems, the inclusion of shut-down cost would not affect the results since a start-up is always followed by a shut-down to complete a busy cycle.

III. Optimal Policy Structure

In Theorem 1 of the work by Leng *et al.* [8], it is proved that the optimal policy with IPP arrival is a hysteretic policy, i.e., if the policy chooses to switch to a better mode, then it would stay in that mode if it is already in the mode. The extension of the hysteretic property to the MMPP case is straightforward given the work by Hipp and Holzbaur [6, Th. 1] and by examining the switching cost function which is defined as

$$s(W(t), a(t)) = \max(a(t) - W(t), 0) E_{\text{sw}}. \tag{9}$$

It is conjectured by Leng *et al.* [8, Conjecture 1] that the optimal policy for IPP arrival is also a monotone policy, i.e., given $S(t)$ the optimal action $a^*(t) \triangleq f(S(t), Q(t), W(t))$ is non-decreasing with $Q(t)$. Consequently, assuming the conjecture is upheld, it is shown that the optimal policy for IPP arrival is a threshold-based policy, which is described by the *active* and *sleeping* thresholds at ON and OFF phases of IPP, respectively. In what follows, we not only settle the monotone conjecture, but also extend to MMPP arrival case, and thus prove the optimal policy with MMPP arrival and multiple servers is queue-threshold-based.

Theorem 1: The optimal policy to the formulated MDP is a monotone policy, i.e., $\forall$ S, W, and $Q_1 \geq Q_2$,

$$f(S, Q_1, W) \geq f(S, Q_2, W). \tag{10}$$

Proof: The main technique to prove the theorem is inspired by the proof of Theorem 1 in the work by Lu and Serfozo [5]. However, the arrival process is Poisson-based and the cost-to-go function is required to be submodular [5]. In fact, it can be shown through numerical simulations that the cost-to-go function with MMPP arrival is, in general, not a submodular function. To address this issue, we present a stronger result

[1] Although the arrival MMPP is not a renewal process, the arrival phase transition is still memory-less based on the MMPP definition.

[2] Obviously, considering the switching cost, it is better to turn an additional $a - W$ servers to active mode when there are W ($W \leq a$) active servers, rather than to close some servers and turn on more.

in Lemma 1 (Appendix) which indicates that only a partial submodular condition is sufficient, i.e., it suffices that the cost-to-go function is submodular with respect to Q and a. Define the cost-to-go function as

$$V_t(S,Q,W) = \min_{a\in\{0,\ldots,M\}} \{s(W,a) + w_{t-1}(S,Q,a)\}, \tag{11}$$

$$\begin{aligned} w_t(S,Q,a) = {} & \omega Q + aE_{\text{on}} + r\Bigg[\sum_{\bar{S}\in\mathcal{N}_S} \sigma_{S\bar{S}}\Delta V_t(\bar{S},Q,a) \\ & + \lambda_S\Delta\mathbb{1}(Q<B)V_t(S,Q+1,a) \\ & + a\mu\Delta\mathbb{1}(Q>0)V_t(S,Q-1,a) \\ & + \Bigg(1 - a\mu\Delta\mathbb{1}(Q>0) - \lambda_S\Delta\mathbb{1}(Q<B) \\ & - \sum_{\bar{S}\in\mathcal{N}_S}\sigma_{S\bar{S}}\Delta\Bigg)V_t(S,Q,a)\Bigg], \end{aligned} \tag{12}$$

and define

$$u_t(S,Q,W,a) \triangleq s(W,a) + w_t(S,Q,a). \tag{13}$$

To prove Theorem 1, we will first show that Theorem 1 is true in a finite horizon of length T by induction. The generalization to infinite horizon follows standard methods as shown by Lu and Serfozo [5, Th. 2]. In particular, we will show that the following statements are valid.

(i) The optimal policy is non-decreasing in Q.

(ii) $\forall S, t$, and $Q_1 \le Q_2$, $W_1 \le W_2$, $V_t(S,Q_2,W_1) - V_t(S,Q_1,W_1) \ge V_t(S,Q_2,W_2) - V_t(S,Q_1,W_2)$.

(iii) Define $V'_t(S,Q,W) = V_t(S,Q+1,W) - V_t(S,Q,W)$.[3] Then $\forall t, S$ and Q, $0 \le V'_t(S,Q,0) \le V'_t(S,Q+1,M)$.

Induction basis: For $t = 1$, the one-step cost-to-go function is

$$\begin{aligned} V_1(S,Q,W) = \min_{a_1\in\{0,\ldots,M\}} \{ & \max(a_1 - W, 0)E_{\text{sw}} \\ & + \omega Q + a_1 E_{\text{on}}\}. \end{aligned} \tag{14}$$

It is obvious that the optimal control action a_1^* to minimize V_1 does not depend on Q. Therefore, (i)-(iii) are satisfied with equality.

Induction steps: Suppose (i)-(iii) are valid for $k \le t$. Then, for $\forall a \le b$, and S,

$$\begin{aligned} & w'_t(S,Q,b) - w'_t(S,Q,a) \\ & = r\Bigg[\sum_{\bar{S}\in\mathcal{N}_S}\sigma_{S\bar{S}}\Delta(V'_t(\bar{S},Q,b) - V'_t(\bar{S},Q,a)) \\ & \quad + \lambda_S\Delta(V'_t(S,Q+1,b) - V'_t(S,Q+1,a)) \\ & \quad + a\mu\Delta(V'_t(S,Q-1,b) - V'_t(S,Q-1,a)) \\ & \quad + \mu(b-a)\Delta(V'_t(S,Q-1,b) - V'_t(S,Q,b)) \\ & \quad + \Bigg(1 - a\mu\Delta - \lambda_S\Delta - \sum_{\bar{S}\in\mathcal{N}_S}\sigma_{S\bar{S}}\Delta\Bigg) \\ & \quad \cdot (V'_t(S,Q,b) - V'_t(S,Q,a))\Bigg] \\ & \le -r\mu(b-a)\Delta(V'_t(S,Q,b) - V'_t(S,Q-1,b)), \end{aligned} \tag{15}$$

[3]Increment of other functions over Q is denoted identically.

where the inequality is based on the induction hypothesis (ii). Combining the induction hypotheses (ii) and (iii), it follows that

$$\begin{aligned} rClV'_t(S,Q-1,b) \le V'_t(S,Q-1,0) \le V'_t(S,Q,M) \\ \le V'_t(S,Q,b). \end{aligned} \tag{16}$$

Therefore, we obtain $w'_t(S,Q,b) - w'_t(S,Q,a) \le 0$, and it follows that $u_t(S,Q,W,a)$ satisfies the conditions in Lemma 1. Hence, (i) and (ii) are proved by noticing that the minimization operation preserves partial submodularity.

To prove (iii), we obtain

$$\begin{aligned} & w'_t(S,Q,0) - w'_t(S,Q+1,M) \\ & = r\Bigg[\sum_{\bar{S}\in\mathcal{N}_S}\sigma_{S\bar{S}}\Delta(V'_t(S,Q,0) - V'_t(S,Q+1,M)) \\ & \quad + \lambda_S\Delta(V'_t(S,Q+1,0) - V'_t(S,Q+2,M)) \\ & \quad + \mu M\Delta(V'_t(S,Q+1,M) - V'_t(S,Q,M)) \\ & \quad + \Bigg(1 - a\mu\Delta - \lambda_S\Delta - \sum_{\bar{S}\in\mathcal{N}_S}\sigma_{S\bar{S}}\Delta\Bigg)(V'_t(S,Q,0) \\ & \quad - V'_t(S,Q+1,M))\Bigg] \\ & \le -r\mu M\Delta(V'_t(S,Q,M) - V'_t(S,Q-1,M)) \le 0, \end{aligned} \tag{17}$$

where the inequality stems from combining the induction hypothesis (ii) and (iii). It follows that V'_{t+1} is non-negative since V'_t is. Denote

$$x \triangleq f(S,Q,W),\ y \triangleq f(S,Q+2,W), \tag{18}$$

we obtain

$$\begin{aligned} & V'_{t+1}(S,Q,0) \\ & \quad = \min_a u_{t+1}(S,Q+1,W,a) - \min_a u_{t+1}(S,Q,W,a) \\ & \quad \le u_{t+1}(S,Q+1,W,x) - u_{t+1}(S,Q+1,W,x) \\ & \quad = w'_t(S,Q,x) \le w'_t(S,Q,0) \le w'_t(S,Q+1,M) \\ & \quad \le w'_t(S,Q+1,y) = u'_{t+1}(S,Q+1,M,y) \\ & \quad \le V'_{t+1}(S,Q+1,M). \end{aligned} \tag{19}$$

Therefore, the hypothesis (iii) is proved. Note that the corner cases wherein $Q = 0$ or $Q = B$ can be dealt with appropriately, and for brevity the details are not shown. With this, the induction proof is completed. ■

IV. Numerical Results

In this section, the MDP is solved numerically to obtain the optimum queue thresholds. Each time slot is 10 milliseconds. Two arrival phases are considered in the MMPP, where the arrival rates are 5 (ON phase) and 0 (OFF phase) jobs per second, respectively. The definition of ON and OFF phases is identical with that in the IPP; jobs only arrive during the ON phase based on the Poisson model; the duration of both phases obeys i.i.d. exponential distributions. The phase transition rates are 0.5 s^{-1} and 0.25 s^{-1} in ON phase and

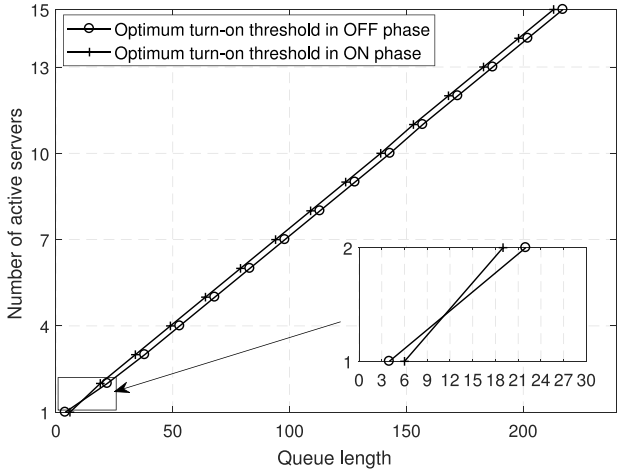

Fig. 1. Optimum queue thresholds for turning on servers.

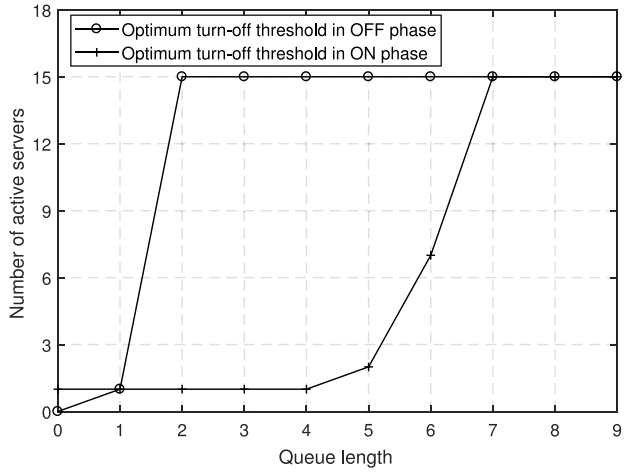

Fig. 2. Optimum queue thresholds for turning off servers.

OFF phase, respectively. The mean service time for a single server is 0.12 *s*. The number of available servers is 15. The buffer size is 250. The turn-on energy consumption of a server is 200 joules. The energy consumption of an active server in a time slot is 2.5 joules. The tradeoff parameter $\omega = 0.2$. These parameters are obtained from realistic cellular systems [8], [10]. The discount factor $r = 0.999$. The solution to the MDP is obtained by standard policy iterations over infinite horizon. In Fig. 1, it is shown that the optimum turn-on queue thresholds, both in ON phase and OFF phase, are almost linear with the number of active servers. The threshold to turn on one server in OFF phase is smaller than that in ON phase, indicating that the optimal action in OFF phase with no active server is to turn on service sooner to reduce the delay cost. The gap between other thresholds in ON phase and OFF phase, which correspond to turning on more than one servers, is constant. Moreover, servers are turned on more aggressively in ON phase with at least one active servers. The turn-off thresholds are shown in Fig. 2. It is observed that the optimal action is not to turn off all the servers until the queue is emptied in OFF phase. Compared with the turn-on thresholds, the optimal action is to turn off servers only when the queue length is relatively quite small, and the servers are turned off very quickly when the queue length decreases beyond a certain point (about 10 jobs in Fig. 2).

V. Conclusion

In this letter, we prove that the optimal sleeping mechanism with MMPP arrival and multiple servers is queue-threshold-based. This result settles a conjecture in the literature and extends to MMPP and multiple-server scenario. Through numerical results, it is shown that the optimal sleeping mechanism with multiple servers exhibits a slow activation, rapid and late (only when the queue length is quite small) shutdown feature.

Appendix

Lemma 1 (Partial Submodular Condition): If $\forall\ S,\ W,\ t$, and $Q_1 \le Q_2$, $a_1 \ge a_2$,

$$\begin{aligned} &u_t(S, Q_2, W, a_1) - u_t(S, Q_1, W, a_1) \\ &\quad \le u_t(S, Q_2, W, a_2) - u_t(S, Q_1, W, a_2), \end{aligned} \tag{20}$$

the optimal policy is a monotone policy.

Proof: Given $\forall\ S$, W, and t, define

$$g(Q) \triangleq \underset{a\in\{0,\ldots,M\}}{\arg\min}\ u_t(S, Q, W, a). \tag{21}$$

Then $\forall\ Q_1 \le Q_2$,

$$\begin{aligned} &u_t(S, Q_1, W, \min[g(Q_1), g(Q_2)]) - u_t(S, Q_1, W, g(Q_1)) \\ &= u_t(S, Q_1, W, g(Q_2)) - u_t(S, Q_1, W, \max[g(Q_1), g(Q_2)]) \\ &\le u_t(S, Q_2, W, g(Q_2)) - u_t(S, Q_2, W, \max[g(Q_1), g(Q_2)]) \\ &\le 0. \end{aligned} \tag{22}$$

This implies that $\min[g(Q_1), g(Q_2)] = g(Q_1)$, and thus $g(Q_1) \le g(Q_2)$. ■

References

[1] E. Oh, B. Krishnamachari, X. Liu, and Z. Niu, "Toward dynamic energy-efficient operation of cellular network infrastructure," *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 56–61, Jun. 2011.

[2] H. Kim and G. de Veciana, "Leveraging dynamic spare capacity in wireless systems to conserve mobile terminals' energy," *IEEE/ACM Trans. Netw.*, vol. 18, no. 3, pp. 802–815, Jun. 2010.

[3] Z. Niu, "TANGO: Traffic-aware network planning and green operation," *IEEE Wireless Commun.*, vol. 18, no. 5, pp. 25–29, Oct. 2011.

[4] I. Kamitsos, L. Andrew, H. Kim, and M. Chiang, "Optimal sleep patterns for serving delay-tolerant jobs," in *Proc. Int. Conf. Energy-Efficient Comput. Netw.*, Passau, Germany, 2010, pp. 31–40.

[5] F. V. Lu and R. F. Serfozo, "M/M/1 queueing decision processes with monotone hysteretic optimal policies," *Oper. Res.*, vol. 32, no. 5, pp. 1116–1132, 1984.

[6] S. K. Hipp and U. D. Holzbaur, "Decision processes with monotone hysteretic policies," *Oper. Res.*, vol. 36, no. 4, pp. 585–588, 1988.

[7] J. Wu, Y. Bao, G. Miao, S. Zhou, and Z. Niu, "Base-station sleeping control and power matching for energy-delay tradeoffs with bursty traffic," *IEEE Trans. Veh. Technol.*, vol. 65, no. 5, pp. 3657–3675, May 2016.

[8] B. Leng, X. Guo, X. Zheng, B. Krishnamachari, and Z. Niu, "A wait-and-see two-threshold optimal sleeping policy for a single server with bursty traffic," *IEEE Trans. Green Commun. Netw.*, vol. 1, no. 4, pp. 528–540, Dec. 2017.

[9] R. G. Gallager, *Discrete Stochastic Processes*, vol. 321. New York, NY, USA: Springer, 2012.

[10] Z. Niu, X. Guo, S. Zhou, and P. R. Kumar, "Characterizing energy-delay tradeoff in hyper-cellular networks with base station sleeping control," *IEEE J. Select. Areas Commun.*, vol. 33, no. 4, pp. 641–650, Apr. 2015.