Delay-Constrained Energy-Optimal Base Station Sleeping Control

Xueying Guo, *Student Member, IEEE*, Zhisheng Niu, *Fellow, IEEE*, Sheng Zhou, *Member, IEEE*, and P. R. Kumar, *Fellow, IEEE*

***Abstract*— Base station (BS) sleeping is an effective way to improve the energy-efficiency of cellular networks. However, it may bring extra user-perceived delay. We conduct a theoretical study into the impact of BS sleeping on both energy-efficiency and user-perceived delay. We consider hysteresis sleep and three typical wake-up schemes, namely single sleep, multiple sleep, and N-limited schemes. We model the system as an $M/G/1$ vacation queue, which captures the setup time, the mode-changing cost, as well as the counting or detection cost during the sleep mode. Closed-form expressions for the average power and the Laplace–Stieltjes transform of delay distribution are obtained. The impacts of system parameters on these expressions are analyzed. We then formulate an optimization problem to design delay-constrained energy-optimal BS sleeping policies. We show that the optimal solutions possess a special structure, thereby allowing us to obtain them explicitly or numerically by simple bisection search. In addition, the relationship between the optimal power consumption and the mean delay constraint is analyzed, so as to answer the fundamental question: how much energy can be saved by trading off a certain amount of delay? It is shown that this optimal relationship is linear only when the delay constraint is lower than a threshold. Numerical studies are also conducted, where the impact of detection or counting cost during the sleep mode is explored, and the delay distribution under the optimal policy is obtained.**

***Index Terms*—Base station sleeping, user-perceived delay, vacation queues, energy-delay trade-off.**

I. Introduction

THE ever increasing demand for ubiquitous information access and broadband multimedia service in wireless networks has triggered vast expansion of network infrastructures, resulting in dramatically increased energy consumption. The electricity consumption of the global radio access network is reported to be 77 TWh in 2012, and is estimated to be 109 TWh in 2020 [2]. This rising energy has both ecological and economic impacts, thus triggering the emerging research area called "green cellular networks" [3]. About 60-80% of the energy consumption in cellular networks is consumed by base stations (BSs) [4]. With a large proportion of energy spent on main supply, cooling, idle-mode signaling and processing, a BS with little or no traffic load may nevertheless consume more than 90 percent of its peak energy [5]. Under these circumstances, BS sleeping is drawing increasing attentions recently [4]–[10].

The introduction of BS sleeping is consistent with the concept of traffic-aware network dynamic operation [11], and is facilitated by many newly proposed network architectures such as hyper-cellular networks [12]. In a hyper-cellular network (HCN), the coverage of control signals is decoupled from the coverage of traffic signals. In this way, the data BSs (DBS) for data service functions can be switched-off to adapt to the traffic dynamics, while the signal coverage can be maintained by the control BSs (CBS). The emerging of heterogeneous network (HetNet) also benefits the introduction of BS sleeping.

By introducing BS sleeping, the BS or part of it can be switched off to save energy. However, the user-perceived delay may then deteriorate. Therefore, the impact of BS sleeping on both energy and delay needs to be studied. Also, BS sleeping suffers from several practical constraints such as setup time and BS mode-changing cost, which further complicates the problem.

In this paper, we focus on a single cell setting and carry out a theoretical study into the impact of BS sleeping on both energy consumption and delay performance. This single cell setting is important when we consider the BS sleeping in HetNet or HCN. Since the small cells in HetNet and the data BSs in HCN can be switched off without causing a coverage problem, their sleeping operations can be considered independently, leading effectively to the single cell setting.

Based on such quantitative analyses, our first goal is to design delay-constrained energy-optimal BS sleeping policies. In addition, we investigate the resulting relationship between the optimal energy consumption and delay constraint, so as to answer the following question: how much energy can be saved by trading off a certain amount of delay? This energy-delay trade-off (EDT) is a fundamental problem in wireless communications [13]–[15], and here it is studied in the BS sleeping scenario. The answer to this problem also guides the pricing strategy of the operators.

A hysteresis sleep scheme is employed to avoid frequent BS mode-changing operations, with the hysteresis time as a key

Manuscript received March 29, 2015; revised October 12, 2015; accepted December 6, 2015. Date of publication January 21, 2016; date of current version May 19, 2016. This work was supported in part by the National Basic Research Program of China under Grant 2012CB316001, in part by the National Science Foundation of China under 61571265, Grant 61321061, Grant 61401250, and Grant 61461136004, in part by the NSF under Contract nos. CNS-1302182, in part by the AFOSR Contract FA9550-13-1-0008, in part by the Science Technology Center Grant CCF-0939370, in part by USARO under Contract no. W911NF-15-1-0279, and in part by the Hitachi Ltd. Part of this work was presented at 25th IEEE International Teletraffic Congress [1].

X. Guo, Z. Niu, and S. Zhou are with Tsinghua National Laboratory for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing 100084, China (e-mail: guo-xy11@mails.tsinghua.edu.cn).

P. R. Kumar is with Texas A&M University, College Station, TX 77843 USA.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSAC.2016.2520221

parameter. Three typical wake-up schemes are considered for different scenarios, namely single sleep (SS), multiple sleep (MS), and N-limited schemes. With these schemes, the BS waits for one or several sleep periods, or waits for N tasks to accumulate before it wakes up. Their key parameters are vacation time or accumulation number.

We model the system as an $M/G/1$ queue with vacations, thus obtaining the closed-form expressions for average power consumption and delay distribution for the SS, MS, and N-limited schemes. In many existing cellular systems, the radio resources occupied by an existing task can not be immediately released and re-allocated to newly arrived tasks. Thus, the tasks need to be served in sequence. So we first consider the first-come-first-service (FCFS) service discipline. Subsequently, we generalize our results to the processor-sharing (PS) service discipline, which is a good approximation of CDMA systems.

We then investigate how different sleeping control parameters affect system performance, and further formulate an optimization problem to design delay-constrained energy-optimal BS sleeping policies. The optimal solutions for different wake-up schemes can be derived explicitly or obtained by simple bisection searches. In addition, the explicit relationship between the optimal average power consumption and mean delay constraint is analyzed to answer the energy-delay trade-off problem. Finally, numerical studies are conducted, so that we investigate the impact of the detecting or counting cost during sleep mode, and illustrate the probability density functions (PDFs) of delay under the optimal policy.

The rest of the paper is organized as follows. We review related work in Section II. We describe the hysteresis sleep and three typical wake-up schemes in Section III. We present the system model and formulate the optimization problem in Section IV. We derive the closed-form expressions for average power consumption and delay distribution by queueing theoretic techniques in Section V. We analyze the impacts of sleeping control parameters in Section VI. We proceed to design optimal BS sleeping policies, and analyze the optimal power-delay relationships in Section VII. The results for the processor-sharing service discipline are presented in Section VIII. Numerical results are illustrated in Section IX, followed by the conclusion and future work in Section X.

II. Related Work

There are several existing papers which have jointly considered BS energy-efficiency and user-perceived delay in BS sleeping policy design. Paper [7] proposes a discontinuous transmission scheme in OFDMA cellular systems, which aggregates delay-tolerant traffic to improve system energy-efficiency. Paper [9] encompasses both BS dynamic operations and user association to jointly optimize the flow-level performance and energy consumption. Paper [16] resorts to a game theoretical approach to minimize energy consumption with a flow-level delay constraint in multi-cell scenario. However, all of these studies omit the setup time. Thus, the BS can always sleep whenever it is empty. Consequently, a specific BS sleeping analysis of sleep and wake-up policy design is not taken into consideration. Also, they focus on the design of heuristic policies, and a theoretical study of the impact of BS sleeping on both energy and delay is not their objective.

Paper [17] focuses on a typical vacation server, and employs an MDP model to design optimal BS sleeping policies. The optimal policy was proved to have a hysteretic structure. Note that the optimal policy is stationary, i.e., it decides current control based on current queue length, which motivates us to consider the N-limited scheme. However, the setup time is omitted in this MDP formulation, and the problem is solved numerically.

Energy-delay trade-off (EDT) is a fundamental problem in wireless communications, dating back to Shannon's channel capacity theorem. Paper [13] studies the EDT in fading channels, and shows that it is of a square-root form by asymptotic analysis. Paper [14] further extends this square-root bound to a multi-user scenario. However, these studies focus on the transmit power consumption. Some recent papers consider circuit power consumption as well as transmit power in studying EDT [15], but only the transmission delay is considered. Also, the EDT under sleeping control remains an open problem.

In our previous work [18], [19], we have considered N-limited scheme and analyzed how each single parameter affects the system performance. However, the results for the SS and MS schemes were not demonstrated, and the detecting or counting cost during the sleep mode was not considered. More importantly, a joint optimization of system parameters to design optimal BS sleeping policies was not conducted, and thus the optimal EDT relationship was not studied. Paper [20] jointly considers BS sleeping control and power matching with exponentially distributed service time, where the hysteresis sleep is not included, and the setup time is omitted. This paper is a generalization of our conference paper [1].

III. Hysteresis Sleep and Three Typical Wake-up Schemes

A *hysteresis* sleep scheme is employed: When the system becomes empty, the BS waits for a while. If any new task arrives during this waiting period, the BS immediately starts its service; otherwise, the BS goes to the sleep mode after this period. This hysteresis sleep scheme can efficiently reduce frequent BS mode-changing [21], and thus avoid extra energy and delay penalties resulting from frequent mode-changing.

The choice of *wake-up schemes* depends on both the hardware design of the BS and the network structure of the cellular system [22]. Three typical wake-up schemes are considered:

- *Single Sleep* (SS) scheme: The BS simply wakes up after a specific time. This scheme is characterized by its ease of implementation, and thus is widely applicable. Also, since the BS does not need to be aware of the network status during the sleep, it can go to deep sleep with a relatively low power level. However, the choice of sleep time is crucial since system performance may otherwise deteriorate.
- *Multiple Sleep* (MS) scheme: The BS detects the network status after a specific sleeping period, and only wakes up if there are tasks waiting in the system. Otherwise, it

continues with another sleep period. This scheme is practical when the BS can be partly switched on to detect whether there are waiting tasks during the sleep mode. This scheme may potentially save more energy since it prolongs the sleep time if no tasks arrived during the last sleeping period. However, it consumes extra energy for the BS to detect the system status at the end of each sleeping period.

- N-limited scheme: The BS wakes up when there are N tasks accumulated in the system. To apply this scheme, either there is a low-power counting element in the BS that keeps awake during the sleep mode, or there is network equipment that counts arriving tasks and triggers the wake-up. Since the BS is aware of the network situation during the sleep, this N-limited scheme may potentially lead to better system performance. However, the awake counting element increases the sleep-mode power level of this scheme.

Although this paper applies to both uplink and downlink scenarios, it is hard to implement the N-limited scheme in the uplink scenario within the traditional cellular architecture. Actually, the implementation of the wake-up schemes may depend on the evolution of the cellular network architecture such as the HetNet and the newly proposed hyper-cellular network (HCN) [12], [19].

IV. System Model

Consider a typical cell in a hyper-cellular network where the whole BS is considered as a single server. Here, we consider a higher-level abstraction of the system. Thus, the flow-level sessions, such as web page downloads, are the tasks we deal with. The task arrivals are assumed to follow a Poisson process with rate λ, and each task requires an *i.i.d.* (independently and identically distributed) service time B, which follows a general distribution. Here, the service time is either predetermined by the user requests, or is decided by both the user workload and the wireless transmission condition. Thus, the way to capture the service time distribution depends on the traffic type. For the service discipline, we firstly consider the first-come-first-service (FCFS) discipline. The results for the processor-sharing (PS) discipline are presented in Section VIII.

When the queue becomes empty, the server waits for a hysteresis time D, which follows a general distribution. If any new tasks arrive during this period, the server immediately starts its service; otherwise, the server enters the sleep mode.

For the server in the sleep mode, we consider three typical schemes of wake-up:

- Single Sleep (SS) scheme: the server wakes up after a vacation time V, which follows a general distribution.
- Multiple Sleep (MS) scheme: the server takes another vacation of length V if the queue remains empty after a vacation. It is assumed that the vacation times are i.i.d.
- N-limited scheme: the server wakes up when there are N tasks accumulated during the sleep.

After wake-up, the server further needs a setup time S to be warmed up and thereupon starts to provide services. This setup time follows a general distribution.

TABLE I
Variables in System Model

Variable	Notation
Arrival Rate	λ
Service Time	B
Close-down Time	D
Vacation Time (for SS and MS schemes)	V
Accumulating Number (for N-limited scheme)	N
Setup Time	S
Idle Power	P_{ID}
Load-dependent Power Efficiency	η
Transmission Power	P_{TR}
Sleep Power	P_{SL}
Setup Power	P_{ST}
Detecting Energy Consumption (for MS scheme)	E_{DT}
Detecting Power Consumption (for N-limited scheme)	P_{DT}

For the energy consumption model, we adopt the linear model as in [23], i.e.,

$$P = \begin{cases} P_{\mathrm{ID}} + P_{\mathrm{TR}}/\eta, & \text{in active mode;} \\ P_{\mathrm{SL}}, & \text{in sleep mode,} \end{cases}$$

where P_{ID} is the power level when BS is in active mode but idle, P_{TR} is the RF transmit power, η is the efficiency for load dependent power consumption, and P_{SL} is the power level during sleep mode which is smaller than P_{ID}, i.e, $P_{\mathrm{SL}} < P_{\mathrm{ID}}$. Further, it is assumed that during the setup time, the BS has a power consumption P_{ST}, which captures the energy cost resulting from BS mode-changing operation [24], [25], and it is normally larger than P_{ID}, i.e., $P_{\mathrm{ST}} \geq P_{\mathrm{ID}}$.

In addition, for the MS scheme, there is an additional *energy consumption for detection* E_{DT} each time the server finishes a vacation time and detects whether there are waiting tasks. For the N-limited scheme, the sleep mode power level is $P_{\mathrm{SL}} + P_{\mathrm{DT}}$, where P_{DT} is the *additional power consumption for counting tasks in sleep mode*. Without loss of generality, it is assumed that $P_{\mathrm{SL}} + P_{\mathrm{DT}} < P_{\mathrm{ID}}$ when the N-limited scheme is applied.

We summarize these variables in Table I.

Our goal is to design sleeping control parameters (i.e., V and D for the SS and MS schemes, or N and D for the N-limited scheme) so as to minimize average power consumption with a constraint on mean delay. That is,

$$\min_{\{D,V\} \text{ or } \{D,N\}} \quad \mathrm{E}[P] \tag{1}$$

$$\text{s.t.} \quad \mathrm{E}[T] \leq \tau, \tag{2}$$

where $\mathrm{E}[P]$ denotes the average system power consumption, $\mathrm{E}[T]$ denotes the mean delay, τ is the parameter for mean delay constraint, and the minimization over a random variable means the choice of its distribution and the corresponding distribution parameters (recall that D and V are assumed to follow general distributions).

V. Queueing Analysis

In this section, we derive closed-form expressions for delay distribution and average power consumption.

We begin with some notations: For any continuous random variable X, we denote by h_X its expectation, by c_X^2

its squared coefficient of variation, and by $\tilde{X}(s)$ its Laplace-Stieltjes Transform (LST), i.e.,

$$h_X \triangleq \mathrm{E}[X],\ c_X^2 \triangleq \frac{\mathrm{E}[X^2]}{h_X^2}-1,\ \tilde{X}(s) \triangleq \mathrm{E}[\mathrm{e}^{-sX}]. \quad (3)$$

Then, $h_B, h_V, h_S, c_B^2, c_V^2, c_S^2, \tilde{B}(s), \tilde{V}(s), \tilde{S}(s)$ and $\tilde{D}(s)$ are defined accordingly. Denote the *traffic intensity* by ρ, i.e.,

$$\rho \triangleq \lambda h_B. \quad (4)$$

The superscripts (SS), (MS) and (N) are employed to denote the scheme applied. In addition, we denote by,

$$p_{\mathrm{v}} \triangleq \tilde{D}(\lambda) = \mathrm{E}[\mathrm{e}^{-\lambda D}], \quad (5)$$

which is actually the probability that the BS switches to the sleep mode after it becomes empty, i.e., *sleeping probability*.

Further, note that the operation of the server can be divided into different phases with different power levels, as in Fig. 1. In the *busy* phase, the BS serves the tasks with a power level $P_{\mathrm{ID}} + P_{\mathrm{TR}}/\eta$. In the *close-down* phase, the BS waits for the hysteresis time D before it transits to the sleep mode with a power level P_{ID}. In the *sleep* phase, the BS is in the sleep mode, with a power level P_{SL}. In the *setup* phase, the BS transits from the sleep mode to the active mode, with a power level P_{ST}. In the *idle* phase, the BS remains in the active mode with zero load, and has a power level P_{ID}. Note that the idle phase only exists for the SS scheme.[1]

Also, it can be easily seen that, no matter which scheme is applied, the system forms a regenerative process [29] with *regeneration points* at the time epochs when busy phases end. Denote by *regeneration cycle* the time period between two successive regeneration points. Then, the system evolves statistically the same in different regeneration cycles. As a result, the average system power consumption satisfies,

$$\begin{aligned}\mathrm{E}[P] = &\big[\mathrm{E}[L_{\mathrm{BS}}](P_{\mathrm{ID}}+P_{\mathrm{TR}}/\eta)+\mathrm{E}[L_{\mathrm{CD}}]P_{\mathrm{ID}}+\mathrm{E}[L_{\mathrm{SL}}]P_{\mathrm{SL}}\\ &+\ \mathrm{E}[L_{\mathrm{ST}}]P_{\mathrm{ST}}+\mathrm{E}[L_{\mathrm{ID}}]P_{\mathrm{ID}}\big/\left(\mathrm{E}[L_{\mathrm{BS}}]+\mathrm{E}[L_{\mathrm{CD}}]\right.\\ &\left.+\mathrm{E}[L_{\mathrm{SL}}]+\mathrm{E}[L_{\mathrm{ST}}]+\mathrm{E}[L_{\mathrm{ID}}]\right), \end{aligned} \quad (6)$$

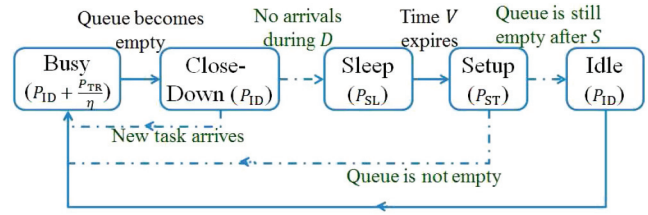
where L_{BS}, L_{CD}, L_{SL}, L_{ST} and L_{ID} denote the lengths of busy phase, close-down phase, sleep phase, setup phase and idle phase, in a single regeneration cycle, respectively. Note that these lengths are statistically the same in different regeneration cycles.

Now we derive the expected lengths of different operational phases in a single regeneration cycle for the SS scheme. First, the proportion of time that is occupied by busy phases in a long-term average is ρ, by recalling (4). That is, $\mathrm{E}[L_{\mathrm{BS}}^{(\mathrm{SS})}] = \rho$. Second, the length of L_{CD} is equivalent to the minimum of D and the length of an exponentially distributed random variable with parameter λ. Thus, $\mathrm{E}[L_{\mathrm{CD}}^{(\mathrm{SS})}] = (1-p_{\mathrm{v}})/\lambda$. Third, by (5), the probability that a regeneration cycle involves a sleep phase is p_{v}. Thus, $\mathrm{E}[L_{\mathrm{SL}}^{(\mathrm{SS})}] = p_{\mathrm{v}}h_V$, and $\mathrm{E}[L_{\mathrm{ST}}^{(\mathrm{SS})}] = p_{\mathrm{v}}h_S$. In addition, in a regeneration cycle which involves a sleep phase, the probability that an idle phase appears is $\tilde{V}(\lambda)\tilde{S}(\lambda)$. Thus, $\mathrm{E}[L_{\mathrm{ID}}^{(\mathrm{SS})}] = p_{\mathrm{v}}\tilde{V}(\lambda)\tilde{S}(\lambda)/\lambda$. Consequently, by combining these results with (6), we obtain the average power consumption for the SS scheme,

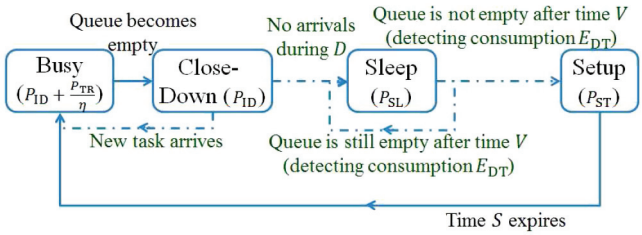
$$\begin{aligned}\mathrm{E}[P^{(\mathrm{SS})}] = &\rho(P_{\mathrm{ID}}+P_{\mathrm{TR}}/\eta)+(1-\rho)\left\{p_{\mathrm{v}}h_V P_{\mathrm{SL}}\right.\\ &\left.+p_{\mathrm{v}}h_S P_{\mathrm{ST}}+\left[\frac{1}{\lambda}(1-p_{\mathrm{v}})+\frac{1}{\lambda}p_{\mathrm{v}}\tilde{V}(\lambda)\tilde{S}(\lambda)\right]P_{\mathrm{ID}}\right\}\\ &\Big/\left[\frac{1}{\lambda}(1-p_{\mathrm{v}})+p_{\mathrm{v}}\left(h_V+h_S+\frac{1}{\lambda}\tilde{V}(\lambda)\tilde{S}(\lambda)\right)\right]. \end{aligned} \quad (7)$$

By a similar argument, we can obtain the average power consumption for the MS and N-limited schemes,

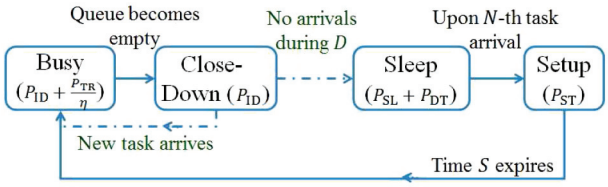
$$\begin{aligned}\mathrm{E}[P^{(\mathrm{MS})}] = &\rho(P_{\mathrm{ID}}+P_{\mathrm{TR}}/\eta)+(1-\rho)\left[(1-p_{\mathrm{v}})P_{\mathrm{ID}}/\lambda\right.\\ &\left.+\frac{p_{\mathrm{v}}h_V}{1-\tilde{V}(\lambda)}P_{\mathrm{SL}}+p_{\mathrm{v}}h_S P_{\mathrm{ST}}+\frac{p_{\mathrm{v}}}{1-\tilde{V}(\lambda)}E_{\mathrm{DT}}\right]\\ &\Big/\left[(1-p_{\mathrm{v}})/\lambda+p_{\mathrm{v}}h_V(1-\tilde{V}(\lambda))^{-1}+p_{\mathrm{v}}h_S\right], \end{aligned} \quad (8)$$



(a) Single Sleep (SS) scheme



(b) Multiple Sleep (MS) scheme



(c) N-limited scheme

Fig. 1. The transition diagram of BS operation phases with different wake-up schemes.

[1]Also, note that the SS scheme here is different from the single vacation model with setup and close-down times, as in [26]–[28], since in our case the server goes to the setup phase immediately after a vacation time regardless of whether there are waiting customers.

$$\mathrm{E}[P^{(N)}] = \rho(P_{\mathrm{ID}} + P_{\mathrm{TR}}/\eta) + (1-\rho)\left[(1-p_{\mathrm{v}})P_{\mathrm{ID}}/\lambda + p_{\mathrm{v}}\frac{N}{\lambda}(P_{\mathrm{SL}}+P_{\mathrm{DT}}) + p_{\mathrm{v}}h_S P_{\mathrm{ST}}\right] \Big/ \left[\frac{1}{\lambda}(1-p_{\mathrm{v}}) + p_{\mathrm{v}}\frac{N}{\lambda} + p_{\mathrm{v}}h_S\right]. \quad (9)$$

Now, we consider the overall sojourn time (delay) of tasks. It follows from the system model that the BS sleeping problem is an $M/G/1$ vacation queue with close-down and setup times no matter which scheme is applied. As a result, the Laplace-Stieltjes transforms (LST) of the overall delay T for different schemes are,

$$\tilde{T}^{(\mathrm{SS})}(s) = \frac{\lambda(1-\rho)\tilde{B}(s)}{s-\lambda+\tilde{B}(s)\lambda}\left[\frac{s}{\lambda}(1-p_{\mathrm{v}}) + p_{\mathrm{v}}\tilde{S}(\lambda)\tilde{V}(\lambda)\left(\frac{s}{\lambda}-1\right)+p_{\mathrm{v}}\right] \Big/ \left[1 + p_{\mathrm{v}}\left(\lambda h_S + \lambda h_V + \tilde{S}(\lambda)\tilde{V}(\lambda) - 1\right)\right]; \quad (10)$$

$$\tilde{T}^{(\mathrm{MS})}(s) = \frac{\lambda(1-\rho)\tilde{B}(s)}{s-\lambda+\tilde{B}(s)\lambda}\left[(1-p_{\mathrm{v}})\frac{s}{\lambda} - p_{\mathrm{v}}\tilde{S}(s)\frac{\tilde{V}(s)-\tilde{V}(\lambda)}{1-\tilde{V}(\lambda)} + p_{\mathrm{v}}\right] \Big/ \left\{p_{\mathrm{v}}\lambda\left[h_S + h_V/\left(1-\tilde{V}(\lambda)\right)\right] + 1 - p_{\mathrm{v}}\right\}; \quad (11)$$

$$\tilde{T}^{(N)}(s) = \frac{\lambda(1-\rho)\tilde{B}(s)}{p_{\mathrm{v}}(N+\lambda h_S)+1-p_{\mathrm{v}}}\left\{\frac{p_{\mathrm{v}}\tilde{S}(s)}{\lambda}\frac{\left[\lambda/(s+\lambda)\right]^N - \left[\tilde{B}(s)\right]^N}{\lambda/(s+\lambda) - \tilde{B}(s)} + \frac{1-p_{\mathrm{v}}}{\lambda} + \frac{(1-p_{\mathrm{v}})\left(1-\tilde{B}(s)\right) + p_{\mathrm{v}}\left[1-\tilde{S}(s)\left(\tilde{B}(s)\right)^N\right]}{s-\lambda+\lambda\tilde{B}(s)}\right\}. \quad (12)$$

Here, to obtain (10) and (11), we note that, when the SS or MS scheme is applied, the delay of an existing task in the system is independent of future task arrivals. Thus, the distribution form of Little's Theorem [30] holds, leading to,

$$\tilde{Q}(s) = \tilde{T}(\lambda - \lambda s), \quad (13)$$

where the $\tilde{Q}(s)$ is the probability generating function of the steady-state queue length distribution, i.e., $\tilde{Q}(s) \triangleq \mathrm{E}[s^Q]$. The derivation of $\tilde{Q}(s)$ is shown in Appendix A. For (12), the delay distribution for the N-limited scheme is as in [19].

It directly follows from (10)–(12) that the mean delays are,

$$\mathrm{E}[T^{(\mathrm{SS})}] = \bar{T}_0 + \frac{\lambda p_{\mathrm{v}}\left[h_V^2(1+c_V^2)+h_S^2(1+c_S^2)+2h_Vh_S\right]}{2+2p_{\mathrm{v}}\left[\lambda(h_S+h_V)+\tilde{V}(\lambda)\tilde{S}(\lambda)-1\right]}; \quad (14)$$

$$\mathrm{E}[T^{(\mathrm{MS})}] = \bar{T}_0 + \lambda p_{\mathrm{v}}/\left[2(1-\tilde{V}(\lambda))\right] \cdot \frac{h_V^2(1+c_V^2) + h_S^2(1+c_S^2)(1-\tilde{V}(\lambda)) + 2h_Vh_S}{p_{\mathrm{v}}\left[\lambda h_V/(1-\tilde{V}(\lambda)) + \lambda h_S - 1\right] + 1}; \quad (15)$$

$$\mathrm{E}[T^{(N)}] = \bar{T}_0 + \frac{p_{\mathrm{v}}\left[N(N-1)+2N\lambda h_S+\lambda^2 h_S^2(1+c_S^2)\right]}{2\lambda[p_{\mathrm{v}}(N+\lambda h_S)+1-p_{\mathrm{v}}]}. \quad (16)$$

where,

$$\bar{T}_0 \triangleq h_B + \frac{\rho^2(1+c_B^2)}{2\lambda(1-\rho)}, \quad (17)$$

which is the mean delay of the corresponding $M/G/1$ queue without BS sleeping.

Remark: Each mean delay in (14)–(16) has two terms: one is $\bar{T}_0$, and the other is the additional delay due to BS sleeping. Note that the second term is independent of the service time.

VI. Impacts of Sleeping Control Parameters

In this section, we analyze the impact of sleeping control parameters, and present the results that facilitates optimal sleeping policy design. By sleeping control parameters, we mean the hysteresis time D and the vacation time V in the SS and MS schemes, along with the hysteresis time D and accumulating number N in the N-limited scheme.

Firstly, we note that there is a necessary condition for the introduction of BS sleeping:

$$\mathrm{E}[P] < P_{\mathrm{ID}} + \rho P_{\mathrm{TR}}/\eta. \quad (18)$$

This is because the r.h.s. (right-hand side) of (18) is the average power consumption when BS sleeping is not introduced. As a result, if condition (18) is violated, the introduction of BS sleeping incurs extra delay but cannot save energy. Further, by (7)–(9), this necessary condition is equivalent to:

$$\begin{cases} h_V(P_{\mathrm{ID}} - P_{\mathrm{SL}}) > h_S(P_{\mathrm{ST}} - P_{\mathrm{ID}}), \text{ for the SS scheme;} \\ \frac{h_V(P_{\mathrm{ID}}-P_{\mathrm{SL}})-E_{\mathrm{DT}}}{1-\tilde{V}(\lambda)} > h_S(P_{\mathrm{ST}}-P_{\mathrm{ID}}), \text{ for the MS scheme;} \\ N(P_{\mathrm{ID}}-P_{\mathrm{SL}}-P_{\mathrm{DT}}) > \lambda h_S(P_{\mathrm{ST}}-P_{\mathrm{ID}}), \text{ for the } N\text{-limited,} \end{cases}$$

which provide the basic constraints on the choice of sleeping control parameters.

For the impact of the hysteresis time D, we have the following results.

Lemma 1: For any of the SS, MS and N-limited schemes, the following results hold:

1. The impact of the hysteresis time D on $\mathrm{E}[P]$ and $\mathrm{E}[T]$ is only manifested through p_{v} in (5).
2. Under the condition of (18), for a given λ, the average power consumption $\mathrm{E}[P]$ decreases monotonically with increasing p_{v}.
3. For a given λ, mean delay $\mathrm{E}[T]$ increases monotonically with increasing p_{v}.
4. By changing hysteresis time D (or equivalently changing p_{v} with a given λ), there is a *linear* relationship between $\mathrm{E}[P]$ and $\mathrm{E}[T]$. Specifically, by changing D,

$$\mathrm{E}[P] = P_{\mathrm{ID}} + \rho P_{\mathrm{TR}}/\eta + f(\cdot)(\mathrm{E}[T] - \bar{T}_0), \quad (19)$$

where the $f(\cdot)$ is defined as follows for different schemes,

$$f^{(\mathrm{SS})}(V) = 2(1-\rho)\frac{-h_V(P_{\mathrm{ID}} - P_{\mathrm{SL}}) + h_S(P_{\mathrm{ST}} - P_{\mathrm{ID}})}{h_V^2(1+c_V^2) + h_S^2(1+c_S^2) + 2h_Vh_S}, \quad (20)$$

$$f^{(\mathrm{MS})}(V) = 2(1-\rho) \cdot \frac{-h_V(P_{\mathrm{ID}}-P_{\mathrm{SL}}) + E_{\mathrm{DT}} + h_S(P_{\mathrm{ST}}-P_{\mathrm{ID}})\left(1-\tilde{V}(\lambda)\right)}{h_V^2(1+c_V^2) + \left(1-\tilde{V}(\lambda)\right)h_S^2(1+c_S^2) + 2h_Vh_S}, \quad (21)$$

$$f^{(N)}(N) = 2\lambda(1-\rho) \cdot \frac{-N(P_{\mathrm{ID}} - P_{\mathrm{SL}} - P_{\mathrm{DT}}) + \lambda h_S(P_{\mathrm{ST}} - P_{\mathrm{ID}})}{N^2 - N(1-2\lambda h_S) + \lambda^2 h_S^2(1+c_S^2)}. \quad (22)$$

Proof: Statement 1) follows from (7)–(9) and (14)–(16). Statement 2) and 3) are obtained by analyzing the partial derivative of $E[P]$ in (7)–(9) or $E[T]$ in (14)–(16) with respect to p_{v}.

For statement 4), it follows from 1) and 3) that the relationship between $E[P]$ and $E[T]$ by changing D can be obtained by representing p_{v} as a function of $\mathrm{E}[T]$ following (14)–(16), and then substituting this representation into the expression for $E[P]$ in (7)–(9). Thus, (19) follows. ■

In (19), when $p_{\mathrm{v}} = 0$, i.e., the BS never goes to the sleep mode, the average power consumption is $P_{\mathrm{ID}} + \rho P_{\mathrm{TR}}/\eta$, and the mean delay is $\bar{T}_0$ in (17). Also, it follows from the second and third statements of Lemma 1 that the slope in (19) is always negative under condition (18). In addition, since $p_{\mathrm{v}} \in [0, 1]$, this linear relationship through changing D only exists in a certain range.

Now, for the impact of V or N, we temporarily focus on the deterministic vacation times, i.e., $c_V^2 = 0$, and obtain the following results.

Lemma 2: For deterministic vacation times with length h_V, and integer $N \geq 1$, the following results hold.

1. Under the condition of (18), $\mathrm{E}[P]$ is monotone decreasing in h_V (for the SS and MS schemes), or in N (for the N-limited scheme).
2. When the SS scheme is applied, mean delay is monotone increasing in h_V iff. (if and only if)

$$2(h_V + h_S)\left(1 - p_{\mathrm{v}} + 2p_{\mathrm{v}}\mathrm{e}^{-\lambda h_V}\tilde{S}(\lambda)\right) + \lambda p_{\mathrm{v}}\left\{2(h_V+h_S)^2 - \left(1-\mathrm{e}^{-\lambda h_V}\tilde{S}(\lambda)\right) \cdot \left[h_V^2 + h_S^2(1+c_S^2) + 2h_V h_S\right]\right\} \geq 0, \quad (23)$$

which has a sufficient condition: $c_S^2 \leq 1$.

When the MS scheme is applied, mean delay is monotone increasing in h_V iff.

$$\lambda p_{\mathrm{v}}\left\{h_V^2 + h_V h_S\left[2-\mathrm{e}^{-\lambda h_V}(2+\lambda h_V)\right] + h_S^2(1-c_S^2)\left[1-\mathrm{e}^{-\lambda h_V}(1+\lambda h_V)\right]\right\} + (1-p_{\mathrm{v}})\left\{h_V\left[2-\mathrm{e}^{-\lambda h_V}(2+\lambda h_V)\right] - 2h_S\left[1-\mathrm{e}^{-\lambda h_V}(1+\lambda h_V)\right]\right\} \geq 0, \quad (24)$$

with a sufficient condition: $\lambda \leq 1/h_S$ and $c_S^2 \leq 1$.

When the N-limited scheme is applied, mean delay is monotone increasing with N iff.

$$\lambda^2 h_S^2(1-c_S^2)p_{\mathrm{v}} + 2\lambda h_S + 2 \geq 0, \quad (25)$$

which has a sufficient condition: $c_S^2 \leq 1$.

Proof: Statement 1) and (23), (24) are obtained by analyzing the partial derivative of $\mathrm{E}[P]$ in (7)–(9) or E[T] in (14), (15) with respective to h_V or N. For (25), we note that the r.h.s. of $\mathrm{E}[T^{(N)}]$ in (16) is a convex function of N for $N \geq 0$ since its second-order partial derivative with respect to N is positive. Thus, the first order partial derivative of $\mathrm{E}[T^{(N)}]$ with respect to N is a monotonically increasing function of N. As a result, $\mathrm{E}[T^{(N)}]$ is monotonically increasing in N iff.,

$$\mathrm{E}\left[T \,\middle|\, N=1\right] \leq \mathrm{E}\left[T \,\middle|\, N=2\right],$$

which is equivalent to (25). ■

The probability distribution of setup time is important for the monotonicity conditions in Lemma 2. Here, we further show how the distribution of setup time affects system performance.

Lemma 3: With a given h_S, the following results hold:

1. For the SS scheme, under the condition of (18), the deterministic setup time (i.e., setup time with $c_S^2 = 0$) leads to the smallest $\mathrm{E}[P^{(\mathrm{SS})}]$.
2. For the MS and N-limited schemes, $\mathrm{E}[P]$ remains the same for different c_S^2, while $\mathrm{E}[T]$ is an increasing function of c_S^2.

Proof: For statement 1), we note that when (18) holds, $\partial \mathrm{E}[P^{(\mathrm{SS})}]/\partial \tilde{S}(\lambda) \geq 0$. Since $\tilde{S}(\lambda) \geq \mathrm{e}^{-\lambda h_S}$ holds by Jensen's Inequality, statement 1) follows. Statement 2) directly follows from (15)–(16) and (8)–(9). ■

As a result, the system benefits from a setup time with a smaller coefficient of variation. This result sheds light upon the design principle of BSs for equipment providers.

VII. Optimal Policy Design and the Property of Optimal Power-Delay Relationships

Now we proceed to solve the optimization problem (1)–(2). Firstly, by Lemma 1, it is equivalent to solve,

$$\min_{\{p_{\mathrm{v}}, V\} \text{ or } \{p_{\mathrm{v}}, N\}} \quad \mathrm{E}[P] \quad (26)$$

$$\text{s.t.} \quad \mathrm{E}[T] \leq \tau, \quad (27)$$

$$p_{\mathrm{v}} \in [0, 1], \quad (28)$$

where the optimization variable depends on the scheme applied, i.e., $\{p_{\mathrm{v}}, V\}$ for the SS or MS scheme and $\{p_{\mathrm{v}}, N\}$ for the N-limited scheme. This is because, by Lemma 1, when the optimal value of p_{v} is obtained, any probability distribution of hysteresis time D leading to this p_{v} value is optimal.

Note that $\tau \geq \bar{T}_0$ is required, since the introduction of BS sleeping cannot decrease delay, while $\bar{T}_0$ is the mean delay without BS sleeping.

We begin with some notations:

$$V^\star \triangleq \arg\min_V f(V) \text{ and } N^\star \triangleq \arg\min_N f(N), \quad (29)$$

with $f(\cdot)$ as in (20)–(22). That is, $V^\star$ is the vacation time which is of the distribution that minimizes $f(V)$ under the SS or MS scheme, and $N^\star$ is the integer that minimizes $f(N)$ under the N-limited scheme. In addition, we substitute $V = V^\star$ (or $N = N^\star$) and $p_{\mathrm{v}} = 1$ into (14)–(16), and denote the resulting mean delay value by τ_{th}. That is, τ_{th} is the value of mean delay when sleeping control parameters are set as $V = V^\star$ (or $N = N^\star$) and $p_{\mathrm{v}} = 1$.

Before we present the specific expressions for $V^\star$, $N^\star$ and τ_{th} in (35)–(40), we first show the following results.

Theorem 4: For the optimization problem (26)–(28) with $\tau \in [\bar{T}_0, \tau_{\text{th}}]$, the following results hold:

1. The optimal V is $V^\star$(or the optimal N is $N^\star$), which remains the same for different constraints τ. The optimal p_{v} is a monotonically increasing function of τ. To be specific, $p_{\text{v}} = z(V^\star, \tau)$ or $p_{\text{v}} = z(N^\star, \tau)$ with,

$$z^{(\text{SS})}(V,\tau) = 2(\tau-\bar{T}_0)\Big/\Big\{\lambda\Big[h_V^2(1+c_V^2)+2h_Vh_S +h_S^2(1+c_S^2)\Big]-2(\tau-\bar{T}_0)\Big[\lambda(h_V+h_S)+\tilde{V}(\lambda)\tilde{S}(\lambda)-1\Big]\Big\}, \tag{30}$$

$$z^{(\text{MS})}(V,\tau) = 2(\tau-\bar{T}_0)\Big/\left\{\lambda\left[\frac{h_V^2(1+c_V^2)+2h_Vh_S}{1-\tilde{V}(\lambda)} +h_S^2(1+c_S^2)\right]-2(\tau-\bar{T}_0)\left[\frac{\lambda h_V}{1-\tilde{V}(\lambda)}+\lambda h_S-1\right]\right\}, \tag{31}$$

$$z^{(N)}(N,\tau)=2(\tau-\bar{T}_0)/\Big[\lambda h_S^2(1+c_S^2)+2Nh_S +N(N-1)/\lambda-2(\tau-\bar{T}_0)(\lambda h_S+N-1)\Big], \tag{32}$$

for the SS, MS, and N-limited schemes, respectively.

2. The optimal average power consumption, denoted $\bar{P}_{\text{opt}}$, has a linear relationship with constraint τ. Specifically,

$$\bar{P}_{\text{opt}} = P_{\text{ID}} + \rho P_{\text{TR}}/\eta + f(V^\star)(\tau-\bar{T}_0), \tag{33}$$

$$\text{or } \bar{P}_{\text{opt}} = P_{\text{ID}} + \rho P_{\text{TR}}/\eta + f(N^\star)(\tau-\bar{T}_0). \tag{34}$$

with $f(\cdot)$ in (20)–(22).

Proof: We prove the statement 1) in the following sequence. We propose a relaxed problem which has a larger feasible region compared with the original problem. Then we prove that the optimal pairs $\{V, p_{\text{v}}\}$ or $\{N, p_{\text{v}}\}$ in 1) are the solution to this relaxed problem. Finally, we show that these optimal pairs are in the feasible region of the original problem if $\tau \in [\bar{T}_0, \tau_{\text{th}}]$. The details are shown in Appendix B.

For statement 2), since the optimal V or N remains the same for $\tau \in [\bar{T}_0, \tau_{\text{th}}]$, the result follows from the fourth statement of Lemma 1. ■

Now, we provide specific results concerning $V^\star$, $N^\star$ and τ_{th} for different schemes:

1) For the SS scheme: The $V^\star$ satisfies,

$$c_{V^\star}^2 = 0, \tag{35}$$

$$h_V^\star = \frac{h_S}{P_{\text{ID}}-P_{\text{SL}}}\Big[(P_{\text{ST}}-P_{\text{ID}}) +\sqrt{(P_{\text{ST}}-P_{\text{ID}})(P_{\text{ST}}+P_{\text{ID}}-2P_{\text{SL}})+(1+c_S^2)(P_{\text{ID}}-2P_{\text{SL}})^2}\Big], \tag{36}$$

where $h_V^\star \triangleq h_{V^\star} = \text{E}[V^\star]$. That is, $V^\star$ follows a deterministic distribution with length as in (36). This is because, by (29), $V^\star$ is the vacation time distribution that minimizes $f^{(\text{SS})}$ in (20). And in (20), since the numerator is negative when $\text{E}[P] < P_{\text{ID}} + P_{\text{TR}}/\eta$, with a given h_V, a smaller c_V^2 leads to a smaller $f^{(\text{SS})}$. Thus, (35) follows. For (36), it is obtained by analyzing the partial derivative of $f^{(\text{SS})}$ with respect to h_V. The details are omitted here since they are straightforward. Then, by the definition of τ_{th}, we have,

$$\tau_{\text{th}}^{(\text{SS})} = \bar{T}_0 + \frac{\lambda\left[(h_V^\star)^2+2h_V^\star h_S+h_S^2(1+c_S^2)\right]}{2\lambda\left(h_V^\star+h_S\right)+2\text{e}^{-\lambda h_V^\star}\tilde{S}(\lambda)}. \tag{37}$$

2) For the MS scheme: The value $V^\star$ cannot be obtained explicitly. However, we carry out a numerical study as in Section IX-A and find that a smaller c_V^2 always leads to a better power-delay relationship. Hence, we exclusively focus on the deterministic vacation times. Then, it is shown in Appendix C that the optimal length $h_V^\star$ can be simply obtained by a bisection search in the relatively low traffic region (which covers the typical scenarios when BS sleeping is considered). Further, we have,

$$\tau_{\text{th}}^{(\text{MS})} = \bar{T}_0 + \frac{(h_V^\star)^2+2h_V^\star h_S+h_S^2(1+c_S^2)\left(1-\text{e}^{-\lambda h_V^\star}\right)}{2h_V^\star+2h_S\left(1-\text{e}^{-\lambda h_V^\star}\right)}. \tag{38}$$

3) For the N-limited scheme: We have,

$$N^\star = \begin{cases} 1, \text{ when } \left[1-\lambda^2h_S^2(1+c_S^2)\right](P_{\text{ID}}-P_{\text{SL}}-P_{\text{DT}}) \\ \qquad -\lambda h_S(P_{\text{ST}}-P_{\text{ID}})(1+2\lambda h_S)\geq 0; \\ \frac{\lambda h_S(P_{\text{ST}}-P_{\text{ID}})}{P_{\text{ID}}-P_{\text{SL}}-P_{\text{DT}}}\left\{1+\sqrt{\Delta}\right\}, \text{ otherwise, with,} \end{cases} \tag{39}$$

$$\Delta = 1+(2-\frac{1}{\lambda h_S})\frac{P_{\text{ID}}-P_{\text{SL}}-P_{\text{DT}}}{P_{\text{ST}}-P_{\text{ID}}}+(1+c_S^2)\left[\frac{P_{\text{ID}}-P_{\text{SL}}-P_{\text{DT}}}{P_{\text{ST}}-P_{\text{ID}}}\right]^2.$$

Note that in (39), the accumulating number N is relaxed to a real number, and (39) is obtained by analyzing the partial derivative of $f^{(N)}(\cdot)$ with respect to N. To focus on integers $N \geq 1$, $N^\star$ should be replaced by either $\lfloor N^\star \rfloor$ or $\lceil N^\star \rceil$.[2] Also, we have,

$$\tau_{\text{th}}^{(N)} = \bar{T}_0 + \frac{\lambda h_S^2(1+c_S^2)+2N^\star h_S+N^\star(N^\star-1)/\lambda}{2\lambda h_S+2N^\star}. \tag{40}$$

By Theorem 4, we obtain the optimal solutions of the problem (26)–(28) with $\tau \leq \tau_{\text{th}}$ for different schemes. Now, we consider the cases with $\tau > \tau_{\text{th}}$.

Lemma 5: For the problem (26)–(28) with $\tau > \tau_{\text{th}}$, the optimal solution satisfies at least one of the following,

$$p_{\text{v}} = 1; \quad \text{or} \quad \text{E}[T] = \tau. \tag{41}$$

Proof: The statement is proved by contradiction. Assume the optimal solution satisfies both $p_{\text{v}} < 1$ and $\text{E}[T] < \tau$. Then by Lemma 1, there exists a larger p_{v} satisfying $\text{E}[T] \leq \tau$ that leads to smaller $\text{E}[P]$. ■

It follows from Lemma 5 that the optimization problem (26)–(28) can be reduced to two single-variable optimization problems by applying (41). To be specific, by (41), either $p_{\text{v}} = 1$ or $p_{\text{v}} = z(V, \tau)$ (or $p_{\text{v}} = z(N, \tau)$ for N-limited scheme) is satisfied. Thus, only the V (or N) need to be selected in solving the problem. In addition, the problem has the following special property under mild conditions.

[2]It can be shown that $f^{(N)}(\cdot)$ monotonically decreases with N when $N < N^\star$, while it monotonically increases with N when $N > N^\star$. Thus, $\lfloor N^\star \rfloor$ or $\lceil N^\star \rceil$ is optimal.

Theorem 6: For the problem (26)–(28) with $\tau > \tau_{\text{th}}$, we exclusively consider deterministic vacation times. Then, when the following two conditions hold:

$$\text{(a) } c_S^2 \le 1, \qquad \text{(b) for the MS scheme, } \lambda \le 1/h_S,$$

the optimal solution satisfies:

1. The optimal p_{v} remains at 1,
2. The optimal h_V or N is an increasing function of τ satisfying $\text{E}[T] = \tau$.

Proof: We only prove the results for the MS scheme since the proofs for the SS or N-limited scheme are similar and simpler. The superscript (MS) is omitted here to simplify the notation.

To prove statement 1), by Lemma 5, we only need to show that with an additional constraint $\text{E}[T] = \tau$, the optimal $p_{\text{v}} = 1$. Firstly, solving the problem (26)–(28) with additional constraint $\text{E}[T] = \tau$ is equivalent to solving,

$$\min_{h_V,\, p_{\text{v}}} \quad f(h_V) \tag{42}$$

$$\text{s.t.} \quad p_{\text{v}} = z(h_V, \tau) \in [0, 1], \tag{43}$$

with $z(\cdot, \tau)$ in (31) (here, for a deterministic V, we denote $z(h_V, \tau) = z(V, \tau)$). This is obtained following a similar argument as in the proof of Theorem 4 (recall how problem the (57)–(58) is derived). Also, when the condition of (a) and (b) hold, we have the following results,

i) A larger p_{v} leads to a larger $\text{E}[T]$; and a larger h_V also leads to a larger $\text{E}[T]$.
ii) Given τ, a smaller h_V leads to a larger $z(h_V, \tau)$.
iii) Since $z(h_V^\star, \tau_{\text{th}}) = 1$, $h_V > h_V^\star$ is required to satisfy $z(h_V, \tau) \le 1$ with $\tau > \tau_{\text{th}}$.
iv) For $h_V > h_V^\star$, a smaller h_V leads to a smaller $f(\cdot)$.

Here, result i) follows from Lemma 1 and Lemma 2. Result ii) follows from i) and the fact that $\text{E}[T] = \tau$ when $p_{\text{v}} = z(h_V, \tau)$, and hence result iii) holds. Result iv) follows from Lemma C.1. As a result, by combining iii), iv), the smallest h_V such that $p_{\text{v}}^\star(h_V, \tau) \le 1$ is the optimal h_V when solving (42), (43). Further, by ii), this smallest possible h_V satisfies $z(h_V, \tau) = 1$. Thus, the optimal sleeping probability is $p_{\text{v}} = 1$.

For statement 2), when (a)–(c) hold, by Lemma 2, a larger h_V leads to a larger $\text{E}[T]$, and a smaller $\text{E}[P]$. Thus, with a given $p_{\text{v}} = 1$, the optimal h_V is the one satisfying $\text{E}[T] = \tau$. ■

Note that the conditions in Theorem 6 are not restrictive: We focus on deterministic vacation times since a numerical study as in Section A finds that a smaller c_V^2 always leads to a better power-delay relationship. For (a), it follows from Lemma 3 that a smaller c_S^2 leads to better system performance with respect to both power and delay. As a result, the equipment provider should make it as small as possible, and it is not harsh to require $c_S^2 \le 1$. Condition (b) requires a relatively low traffic load for the MS scheme, which is the typical case when BS sleeping is considered.

Now we obtain specific results concerning the optimal h_V and N in Theorem 6 with $\tau > \tau_{\text{th}}$.

1) For the SS or MS scheme: It follows from Theorem 6 and Lemma 2 that the optimal h_V can be simply obtained by a bisection search for solving $\text{E}[T] = \tau$.

2) For the N-limited scheme: The optimal N is,

$$\max\left\{1, \left\lfloor \lambda(\tau - \bar{T}_0 - h_S) + \frac{1}{2} + \lambda\sqrt{(\tau - \bar{T}_0 + \frac{1}{2\lambda})^2 - h_S^2 c_S^2 - \frac{h_S}{\lambda}} \right\rfloor \right\}. \tag{44}$$

Further, the optimal average power consumption satisfies,

$$\begin{aligned} \bar{P}_{\text{opt}}^{(N)} &= \rho(P_{\text{ID}} + P_{\text{TR}}/\eta) + (1-\rho)(P_{\text{SL}} + P_{\text{DT}}) \\ &\quad + (1-\rho)(P_{\text{ST}} - P_{\text{SL}} - P_{\text{DT}})h_S \Big/ \left[(\tau - \bar{T}_0) + \frac{1}{2\lambda} \right. \\ &\quad \left. + \sqrt{(\tau - \bar{T}_0 + \frac{1}{2\lambda})^2 - h_S^2 c_S^2 - \frac{h_S}{\lambda}} \right], \text{ with } \tau > \tau_{\text{th}}^{(N)}. \end{aligned} \tag{45}$$

Note that in (45) the accumulating number N is relaxed to positive real numbers to get the optimal power consumption. (That is, the optimal N is set as the second term in (44) without rounding to obtain (45).)

Based on these results, we analyze the relationship between the optimal power consumption and mean delay.

Corollary 7: For the N-limited scheme, when $c_S^2 \le 1$ holds, the relationship between the optimal power consumption, denoted $\bar{P}_{\text{opt}}$, and the delay constraint τ is shown in (34) for $\tau \in [\bar{T}_0, \tau_{\text{th}}]$, and in (45) for $\tau > \tau_{\text{th}}$. It satisfies,

1. $\bar{P}_{\text{opt}}$ is a monotone decreasing function of τ.
2. For $\tau \in [\bar{T}_0, \tau_{\text{th}}]$,

$$\left.\frac{\partial \bar{P}_{\text{opt}}}{\partial \tau}\right|_{\tau < \tau_{\text{th}}} \equiv f(N^\star) < 0.$$

3. For $\tau > \tau_{\text{th}}$, the partial derivative $\partial \bar{P}_{\text{opt}}/\partial \tau$ is monotonically increasing in τ. Further,

$$f(N^\star) < \left.\frac{\partial \bar{P}_{\text{opt}}}{\partial \tau}\right|_{\tau > \tau_{\text{th}}} < 0, \text{ and } \lim_{\tau \to \infty} \frac{\partial \bar{P}_{\text{opt}}}{\partial \tau} = 0.$$

Proof: The results are obtained by analyzing the partial derivatives of the r.h.s. of (39) and (45). ■

As a result, with the optimal BS sleeping policy, relaxing the delay constraint always reduces average power consumption. However, the amount of power saved by trading off a certain amount of delay keeps decreasing after $\tau > \tau_{\text{th}}$.

VIII. Results for Processor-Sharing Service Discipline

So far, the FCFS discipline has been considered to analyze the case when tasks are served in sequence. However, our results can be applied to more general scenarios.

We begin with some notations. When the server is in busy phase with queue length q, let $\phi(l, q)$ denote the service effort directed to the task in position l, for $l = 1, \cdots, q$. That is, with the total service rate h_B^{-1}, when the queue length is q, the service rate allocated to the task in position l is $\phi(l, q)h_B^{-1}$. We assume $\sum_{l=1}^{q} \phi(l, q) = 1$.

This notation can represent several service disciplines. We illustrate three of them:

1) *First-come-first-service (FCFS) discipline*: $\phi(1,q)=1$, and $\phi(l,q)=0, \forall l \neq 1$. It applies to the scenario when tasks are served in sequence.

2) *Processor-sharing (PS) discipline*: $\phi(l,q)=1/q, \forall l$. That is, all the tasks in the system are served simultaneously, and receive an equal fraction of service rate. This is applicable to the round-robin case.

3) *Processor-sharing with finite servers*:

$$\phi(l,q)=\begin{cases}1/q, & l=1,\cdots,q;\ q=1,\cdots,q_{\text{th}};\\ 1/q_{\text{th}}, & l=1,\ldots,q_{\text{th}};\ q=q_{\text{th}}+1,\cdots;\\ 0, & l=q_{\text{th}}+1,\cdots,q;\ q=q_{\text{th}}+1,\cdots.\end{cases}$$

Here, with the positive integer q_{th}, the system forms a q_{th}-server queue. Although the service ability is shared among tasks, the total number of tasks served is limited. This is applicable to CDMA systems with a finite number of code sequences.

Lemma 8: For any work-conserving service discipline with $\sum_{l=1}^{q}\phi(l,q)=1,\ \forall q$, the average power consumption is as in (7)–(9) for the SS, MS, and N-limited schemes, respectively.

Proof: Recalling Section V, the key to analyzing average power consumption is to analyze the expected lengths of operation phases in a single regeneration cycle. For the close-down, sleep, setup, and idle phases, their lengths are not affected by the service discipline. Thus, $\mathrm{E}[L_{\text{CD}}]$, $\mathrm{E}[L_{\text{SL}}]$, $\mathrm{E}[L_{\text{ST}}]$, and $\mathrm{E}[L_{\text{ID}}]$ remain the same for different service disciplines, and equal to that for the FCFS service discipline.

Now we consider the busy phase. Recall that a busy phase only ends when the unfinished workload in the system becomes zero. In addition, the total unfinished workload in the system evolves stochastically the same for different work-conserving service disciplines, since the system service ability is always fully utilized given $\sum_{l=1}^{q}\phi(l,q)=1,\ \forall q$. As a result, $\mathrm{E}[L_{\text{BS}}]$ remains the same for different service disciplines, leading to the results. ■

Corollary 9: When service time B is exponentially distributed, for any work-conserving service discipline with $\sum_{l=1}^{q}\phi(l,q)=1,\ \forall q$, the expressions for average power consumption in (7)–(9) and mean delay in (14)–(16), and all the results in Section VI-VII hold.

Proof: Recalling Lemma 8 and Section VI–VII, we only need to show that the mean delay remains the same as in (14)–(16) for different service disciplines.

We show that the queue length distribution remains the same for different service disciplines then the results for mean delay follows from Little's theorem.

Now, we consider how the queue length evolves. With exponentially distributed service time, for a certain service discipline, the departure rate in busy phase for a task of location l is $h_B^{-1}\phi(l,q)$. Then the total departure rate is $\sum_{l=1}^{q}h_B^{-1}\phi(l,q)=h_B^{-1}\sum_{l=1}^{q}\phi(l,q)=h_B^{-1}$, which remains the same for different service disciplines. In addition, the arrival process and the operation of the server during the close-down, sleep, setup, and idle phases remains the same for different service disciplines. As a result, queue length evolves stochastically the same for different service disciplines. ■

Remark: Actually, the results in Lemma 8 and Corollary 9 hold even for *priority processor sharing* service disciplines (or called discriminatory processor sharing [31]). When priority PS is applied, the tasks are classified into different classes with different weighting factors, and the service ability is split among the tasks according to their weights. The results for the priority PS can be obtained by a similar argument as in Lemma 8 and Corollary 9.

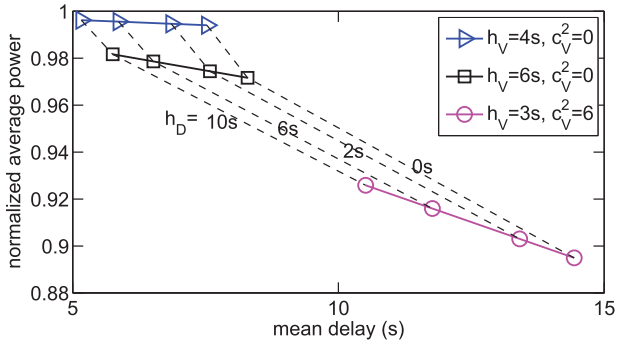


Fig. 2. For the MS scheme, the average power consumption (normalized by the case without sleep mode) vs. mean delay obtained by changing close-down time for different vacation times with mean value h_V and coefficient of variation c_V^2 are shown. (The results are for the case of exponentially distributed close-down time D with different expected values, exponentially distributed setup time S with $h_S=5$s and $E_{\text{DT}}=0.5$J.)

IX. Numerical Results

In this section, we provide the results of numerical studies. We refer to [32] and [33] for system parameters, and set $P_{\text{ID}}=130$ W, $1/\eta=4.7$, $P_{\text{SL}}=75$ W, $P_{\text{TR}}=20$ W. It is further assumed that $P_{\text{ST}}=2P_{\text{ID}}$, $\lambda=0.1\text{s}^{-1}$, and $\rho=0.1$. (So $h_B=1$s.)

A. Impacts of Sleeping Control Parameters and Setup Time

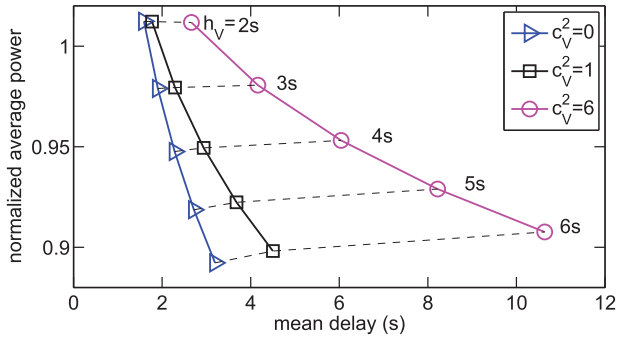
Fig. 2 shows the relationship between average power consumption and mean delay resulting from changing the hysteresis time. The relationship turns out to be linear, which is consistent with Lemma 1.

Fig. 3 shows the relationship between power consumption and mean delay obtained by changing h_V for different vacation time distributions. It can be seen that a smaller c_V^2 leads to a better power-delay relationship, which is consistent with (35).

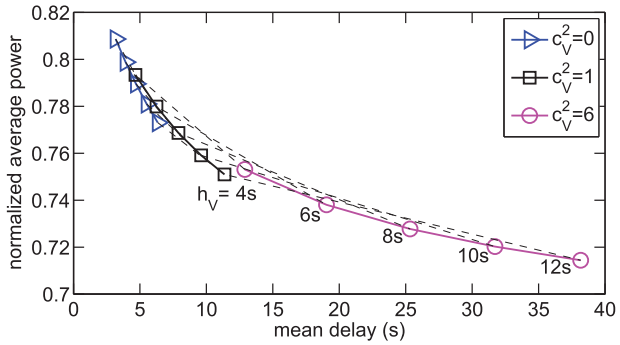
Fig. 4 shows the impact of setup time with respect to both power consumption and mean delay. It can be seen that for a given h_S, a smaller c_S^2 leads to better system performance, which is consistent with Lemma 3.

B. Optimal Power-Delay Relationships

Fig. 5 shows the relationships between the optimal power consumption and mean delay constraint with the optimal sleeping control parameters. It can be seen that when $E_{\text{DT}}=0$ and $P_{\text{DT}}=0$, i.e., when the additional cost to detect or count incoming tasks in the sleep mode is omitted, both the MS and N-limited schemes outperform the simple SS scheme.



(a) SS scheme, with deterministic hysteresis time $D = 0$s.



(b) MS scheme, with exponentially distributed hysteresis time and $h_D = 5$s.

Fig. 3. The average power consumption (normalized for the case without sleep mode) vs. mean delay obtained by changing mean vacation time (h_V) for different vacation time distributions are shown. (The results are for the case of exponentially distributed setup time with $h_S = 1$s and $E_{\mathrm{DT}} = 0$.)

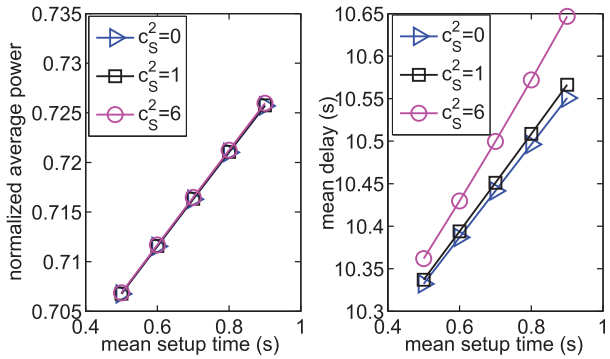


Fig. 4. For the SS scheme, the average power consumption (normalized by the case without sleep mode) and mean delay vs. mean setup time for different coefficients of variation of setup time (c_S^2) are shown. (The results are for the case of constant vacation time with $h_V = 20$s, and exponentially distributed hysteresis time with $h_D = 1$s.)

However, these advantages diminish when E_{DT} and P_{DT} increase.

C. Delay Distribution

Fig. 6 shows the probability density functions (PDFs) of delay under the optimal sleeping control parameters for different schemes. From (10)–(12), with given parameters, the PDF of delay can be obtained numerically by inverting the Laplace-Stieltjes transform. Also, in this case, the coefficient of variation of delay for SS, MS, and N-limited scheme are 0.33, 0.32, and

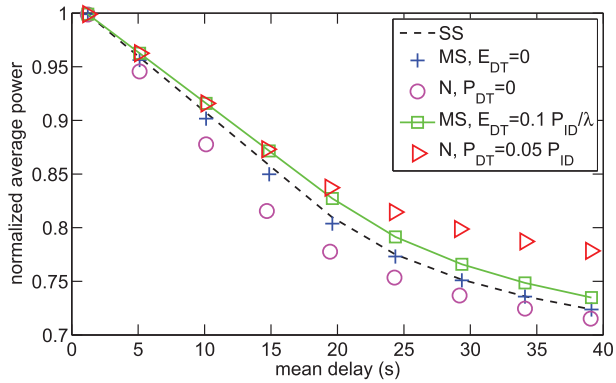


Fig. 5. Optimal power consumption (normalized by the case without sleep mode) vs. mean delay constraint for different schemes are shown. (The results are for the case of an exponentially distributed setup time with $h_S = 5$s.)

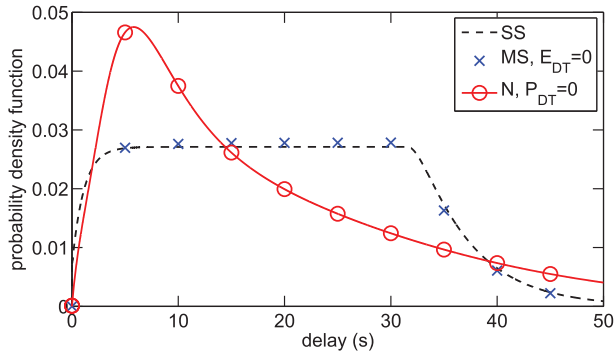


Fig. 6. Probability density functions of delay under optimal BS sleeping control for different schemes under mean delay constraint $\tau = 19.61$s are shown. (The results are for the case of an exponentially distributed setup time S with $h_S = 5$s.)

0.68, respectively. It can be seen that the N-limited scheme has a larger coefficient of variation of delay compared with the other two schemes. This is because, when N-limited scheme is applied, for a task arriving during the sleep mode, its delay depends greatly on the next task's arrival time, which increases the randomness.

X. Conclusions and Future Work

We have carried out a theoretical study into the impact of BS sleeping on both BS energy-efficiency and user-perceived delay. In a single cell scenario, a hysteresis sleep and three typical wake-up schemes, namely the single sleep (SS), multiple sleep (MS), and N-limited schemes, have been considered. Based on a queueing theoretic model, we have derived the closed-form expressions for the average system power and the LST of the user-perceived delay for each wake-up scheme. Both the first-come-first-service (FCFS) service discipline and the processor-sharing (PS) service discipline are considered in this work.

We have further formulated an optimization problem to design delay-constrained energy-optimal BS sleeping policies. We have found that the optimal sleeping control parameter pairs are of a special structure, and thereby determined them either explicitly or numerically by simple bisection searches. We have

also obtained the explicit relationship between the optimal average power and the mean delay constraint. In addition, no matter which scheme is applied, this optimal relationship is shown to be linear if the mean delay constraint is smaller than a threshold. This means that, for a certain amount of delay constraint relaxation, the amount of power saved remains the same when the delay constraint is smaller than the threshold. However, we have further found that this power saved by trading-off a certain amount of delay keeps decreasing when the delay constraint is larger than the threshold, and asymptotically approaches zero.

We have also conducted numerical studies and found that, although the MS or N-limited scheme outperforms the SS scheme when zero detection or counting cost are incurred during the sleep mode, the advantage diminishes when these costs increase. We have explored the PDFs of delay numerically for different schemes with the optimal parameters, and have found that the variance of delay for the N-limited scheme is the largest.

Future work includes the extension of the results to other wake-up schemes such as a randomized N-limited scheme, and the extension to bursty arrivals such as a Batch Poisson model. In addition, an extension to a multi-cell scenario situation can be considered based on these single cell results.

APPENDIX

A. Queue Length Distributions

Denote by Q the queue length, and by $\tilde{Q}(s)$ the probability generating function (GF) of this discrete random variable, i.e., $\tilde{Q}(s) \triangleq \mathrm{E}[s^Q]$. Then, the GFs of steady-state queue length distribution are,

$$\tilde{Q}^{(\mathrm{SS})}(s) = \tilde{B}(\lambda-\lambda s)(1-\rho)/[\tilde{B}(\lambda-\lambda s)-s] \cdot \frac{1-s-p_{\mathrm{v}}\left[\tilde{V}(\lambda-\lambda s)\tilde{S}(\lambda-\lambda s)+\tilde{V}(\lambda)\tilde{S}(\lambda)(s-1)-s\right]}{p_{\mathrm{v}}\left[\lambda(h_V+h_S)+\tilde{V}(\lambda)\tilde{S}(\lambda)\right]+(1-p_{\mathrm{v}})}, \tag{46}$$

$$\tilde{Q}^{(\mathrm{MS})}(s) = \tilde{B}(\lambda-\lambda s)(1-\rho)/[\tilde{B}(\lambda-\lambda s)-s] \cdot \frac{1-s-p_{\mathrm{v}}\left[\tilde{S}(\lambda-\lambda s)(\tilde{V}(\lambda-\lambda s)-\tilde{V}(\lambda))/(1-\tilde{V}(\lambda))-s\right]}{p_{\mathrm{v}}\left[\lambda h_V/\left(1-\tilde{V}(\lambda)\right)+\lambda h_S\right]+1-p_{\mathrm{v}}}, \tag{47}$$

$$\tilde{Q}^{(N)}(s) = \frac{\tilde{B}(\lambda-\lambda s)(1-\rho)}{\tilde{B}(\lambda-\lambda s)-s} \cdot \frac{1-s-p_{\mathrm{v}}[s^N\tilde{S}(\lambda-\lambda s)-s]}{p_{\mathrm{v}}(N+\lambda h_S)+(1-p_{\mathrm{v}})}, \tag{48}$$

for the SS, MS and N-limited schemes, respectively.

Proof: Define the *non-busy period* as the time period from the epoch when a busy phase ends to the epoch when the next busy phase starts (recall the busy phase in Fig. 1). Then let K denote the *number of tasks that arrive during a non-busy period*. It is clear that the random number K is i.i.d. in different non-busy periods no matter which scheme is applied. Further, we denote by h_K the expected value of K, and by $\tilde{K}(s)$ the probability-generating function of K, i.e., $h_K \triangleq \mathrm{E}[K]$ and $\tilde{K}(s) \triangleq \mathrm{E}[s^K]$.

Then, it follows that,

$$\tilde{Q}(s) = \frac{\tilde{B}(\lambda-\lambda s)(1-\rho)[1-\tilde{K}(s)]}{h_K[\tilde{B}(\lambda-\lambda s)-s]}. \tag{49}$$

This is obtained as follows: First, by analyzing the vacation queues at the instants when tasks finish service and leave the system, we obtain the departing queue length distribution (or, more precisely, its generating function). Then, by the equality of queue length distributions at arrival and departure epochs [34], and by PASTA property, this is also the steady-state queue length distribution, leading to (49). The reader is referred to [26] for the details.

Now, we derive $\tilde{K}(s)$ for different schemes to complete the proof. First, we note that, considering a Poisson process with arrival rate λ, for any continuous random variable X, the number of arrivals during a time period of length X, denoted K_X, has the probability generating function,

$$\tilde{K}_X(s) = \sum_{k=0}^{+\infty} \mathrm{P}(K_X=k)s^k = \sum_{k=0}^{+\infty} \mathrm{E}\left[\frac{(\lambda X)^k}{k!}e^{-\lambda X}\right]s^k = \mathrm{E}\left[\mathrm{e}^{-(\lambda-\lambda s)X}\right] = \tilde{X}(\lambda-\lambda s), \tag{50}$$

where the first equality follows from the definition of probability generating function, the second equality results from the property of the Poisson process, and the last equality follows the definition of LST. (Recall from (3) that $\tilde{X}(\cdot)$ is the LST of X.) As a result, by applying (50) and based on discussing whether the server goes to sleep mode during a non-busy period, the LSTs of random variable K are,

$$\tilde{K}^{(\mathrm{SS})}(s) = p_{\mathrm{v}}\left[\tilde{V}(\lambda-\lambda s)\tilde{S}(\lambda-\lambda s)+\tilde{V}(\lambda)\tilde{S}(\lambda)(s-1)\right] + (1-p_{\mathrm{v}})s, \tag{51}$$

$$\tilde{K}^{(\mathrm{MS})}(s) = p_{\mathrm{v}}\frac{\tilde{V}(\lambda-\lambda s)-\tilde{V}(\lambda)}{1-\tilde{V}(\lambda)}\tilde{S}(\lambda-\lambda s)+(1-p_{\mathrm{v}})s, \tag{52}$$

$$\tilde{K}^{(N)}(s) = p_{\mathrm{v}}s^N\tilde{S}(\lambda-\lambda s)+(1-p_{\mathrm{v}})s, \tag{53}$$

for SS, MS and N-limited schemes, respectively.

Combining (50) with (51)–(53), the proof is completed. ■

B. Proof of the first Statement in Theorem 4

Let us remove the constraint (28) and consider the relaxed optimization problem (26), (27) for the time being. Then, the optimal solution of the relaxed problem must satisfy,

$$\mathrm{E}[T] = \tau. \tag{54}$$

This is because, by Lemma 1, a larger p_{v} leads to smaller $\mathrm{E}[P]$ but larger $\mathrm{E}[T]$. Thus, if (54) is violated by the optimal solution, we can increase p_{v} to further decrease the power consumption, leading to a contradiction. As a result, the relaxed problem (26), (27) is equivalent to,

$$\min_{\{p_{\rm v},V\} \text{ or } \{p_{\rm v},N\}} \quad \mathrm{E}[P] \tag{55}$$

$$\text{s.t.} \quad p_{\rm v} = z(V,\tau) \text{ or } p_{\rm v} = z(N,\tau). \tag{56}$$

Here, constraint (56) is equivalent to (54). We further substitute (56) into (55), and get an equivalent problem,

$$\min_{\{p_{\rm v},V\}} \quad P_{\rm ID} + \rho P_{\rm TR}/\eta + f(V)(\tau - \bar{T}_0) \tag{57}$$

$$\text{s.t.} \quad p_{\rm v} = z(V,\tau), \tag{58}$$

for the SS or MS scheme. Replace V by N in (57), (58) for the N-limited scheme. Thus, the pairs in statement 1) are the optimal solution for the relaxed problem (26)–(27).

Now, since $p_{\rm v} = z(\cdot,\tau)$ results from $\mathrm{E}[T] = \tau$, and noting that $\mathrm{E}[T]$ is an increasing function of $p_{\rm v}$ by Lemma 1, we see that $z(\cdot,\tau)$ is an increasing function of τ. By further noting that $z(V^\star, \bar{T}_0) = 0$ and $z(V^\star, \tau_{\rm th}) = 1$ (or $z(N^\star, \bar{T}_0) = 0$ and $z(N^\star, \tau_{\rm th}) = 1$) by the definitions of $\bar{T}_0$ and $\tau_{\rm th}$, statement 1) holds for the optimization problem (26)–(28) with $\tau \in [\bar{T}_0, \tau_{\rm th}]$.

C. Derivation of $h_V^\star$ for the MS Scheme

In this section, the superscript $^{(\rm MS)}$ is omitted to simplify notation. For deterministic V with length h_V, we denote $f(V)$ by $f(h_V)$. Then, $h_V^\star \triangleq \arg\min_{h_V} f(h_V)$. Next, we show that under a mild condition, $h_V^\star$ satisfies $g(h_V) = 0$ with $g(\cdot)$ in (59), and can be simply obtained by a bisection search on $(0, +\infty)$.

Condition C.1: $\lambda^2 h_S^2(1 + c_S^2) \le 2$.

Note that when conditions (a) and (b) in Theorem 6 are satisfied, Condition C.1 follows. It is shown in Section VII (below Theorem 6) that conditions (a) and (b) are not restrictive and cover the typical low-traffic load scenarios when BS sleeping is considered. We denote,

$$\begin{aligned} g(h_V) \triangleq & \left[\lambda h_S^2(1+c_S^2)\mathrm{e}^{-\lambda h_V} + 2h_S + 2h_V\right] \\ & \cdot \left[h_V(P_{\rm ID} - P_{\rm SL}) - E_{\rm DT} - h_S(P_{\rm ST} - P_{\rm ID})(1 - \mathrm{e}^{-\lambda h_V})\right] \\ & - \left[(P_{\rm ID} - P_{\rm SL}) - \lambda h_S(P_{\rm ST} - P_{\rm ID})\mathrm{e}^{-\lambda h_V}\right] \\ & \cdot \left[2h_V h_S + h_V^2 + h_S^2(1+c_S^2)(1 - \mathrm{e}^{-\lambda h_V})\right]. \end{aligned} \tag{59}$$

Then, the partial derivative of $f(h_V)$ with respect to h_V is,

$$\frac{\partial f(\cdot)}{\partial h_V} = \frac{2(1-\rho)\, g(h_V)}{\left[h_V^2 + 2h_V h_S + h_S^2(1+c_S^2)(1 - \mathrm{e}^{-\lambda h_V})\right]^2}. \tag{60}$$

Lemma C.1: When Condition C.1 holds, there is a unique positive $h_V^\star$ such that $g(h_V^\star) = 0$. Further, for any $h_V \in (0, h_V^\star]$, $g(h_V) < 0$; while for any $h_V > h_V^\star$, $g(h_V) > 0$.

Proof: We prove the result by analyzing the higher-order partial derivatives of $g(h_V)$ with respect to h_V. Firstly,

$$\begin{aligned} \partial^3 g(h_V)/\partial h_V^3 = {} & \lambda^3 h_S \mathrm{e}^{-\lambda h_V} \Big\{ -\lambda(P_{\rm ST} - P_{\rm ID})h_V^2 \\ & - h_V \left[\lambda h_S(1+c_S^2)(P_{\rm ID} - P_{\rm SL}) + 2(\lambda h_S - 2)(P_{\rm ST} - P_{\rm ID}) \right] \\ & + 2h_S(1+c_S^2)(P_{\rm ID} - P_{\rm SL} + \frac{\lambda E_{\rm DT}}{2}) + 4h_S(P_{\rm ST} - P_{\rm ID}) \Big\}. \end{aligned} \tag{61}$$

By noting that the terms in $\{\cdot\}$ in (61) are quadratic functions of h_V, and that $\partial^3 g(h_V)/\partial h_V^3|_{h_V=0} > 0$, we have, $\partial^3 g(h_V)/\partial h_V^3$ is first positive, then zero, and then negative with increasing $h_V > 0$, which means that there is a break point $\hat{h}_V > 0$ such that,

$$\begin{aligned} & \partial^3 g(h_V)/\partial h_V^3|_{h_V=\hat{h}_V} = 0; \quad \partial^3 g(h_V)/\partial h_V^3 > 0, \forall h_V \in (0, \hat{h}_V); \\ & \text{and } \partial^3 g(h_V)/\partial h_V^3 < 0, \forall h_V > \hat{h}_V. \end{aligned}$$

It follows that $\partial^2 g(h_V)/\partial h_V^2$ is monotonically increasing in h_V for $h_V \in (0, \hat{h}_V)$, and is monotonically decreasing in h_V for $h_V > \hat{h}_V$. We can continue this argument to determine the sign of $\partial^2 g(h_V)/\partial h_V^2$ for different h_V, and the monotonicity region for $\partial g(h_V)/\partial h_V$, and finally show that $g(h_V)$ is negative, then zero, and then positive with increasing $h_V > 0$, leading to the result. ■

REFERENCES

[1] X. Guo, S. Zhou, Z. Niu, and P. Kumar, "Optimal wake-up mechanism for single base station with sleep mode," in *Proc. 25th Int. Teletraffic Congr., (ITC)*, Sept. 2013, pp. 1–8.

[2] A. Fehske, G. Fettweis, J. Malmodin, and G. Biczok, "The global footprint of mobile communications: The ecological and economic perspective," *IEEE Commun. Mag.*, vol. 49, no. 8, pp. 55–62, 2011.

[3] Z. Hasan, H. Boostanimehr, and V. Bhargava, "Green cellular networks: A survey, some research issues and challenges," *IEEE Commun. Surv. Tuts.*, vol. 13, no. 4, pp. 524–540, 4th Quart. 2011.

[4] B. Rengarajan, G. Rizzo, and M. Marsan, "Bounds on QoS-constrained energy savings in cellular access networks with sleep modes," in *Proc. 23th Int. Teletraffic Congr.*, ((ITC)), Sept. 2011, pp. 47–54.

[5] 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.

[6] G. Micallef, L. Saker, S. E. Elayoubi, and H.-O. Sceck, "Realistic energy saving potential of sleep mode for existing and future mobile networks," *J. Commun.*, vol. 7, no. 10, pp. 740–748, 2012.

[7] R. Gupta and E. Strinati, "Base-station duty-cycling and traffic buffering as a means to achieve green communications," in *Proc. IEEE Veh. Technol. Conf. (VTC)*, Sep. 2012, pp. 1–6.

[8] C. Turyagyenda, K. Al-Begain, and N. Albeiruti, "A novel sleep mode operation for energy efficient LTE cellular networks: A sum product algorithm implementation," in *Proc. Int. Conf. Next Gener. Mobile Apps Serv. Technol.*, Sep. 2013, pp. 159–164.

[9] K. Son, H. Kim, Y. Yi, and B. Krishnamachari, "Base station operation and user association mechanisms for energy-delay tradeoffs in green cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 8, pp. 1525–1536, Sep. 2011.

[10] Z. Niu, Y. Wu, J. Gong, and Z. Yang, "Cell zooming for cost-efficient green cellular networks," *IEEE Commun. Mag.*, vol. 48, no. 11, pp. 74–79, Nov. 2010.

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

[12] Z. Niu, S. Zhou, S. Zhou, X. Zhong, and J. Wang, "Energy efficiency and resource optimized hyper-cellular mobile communication system architecture and its technical challenges," *Sci. China Inf. Sci.*, vol. 42, no. 10, pp. 1191–1202, 2012.

[13] R. Berry and R. Gallager, "Communication over fading channels with delay constraints," *IEEE Inf. Theory*, vol. 48, no. 5, pp. 1135–1149, May 2002.

[14] M. Neely, "Optimal energy and delay tradeoffs for multiuser wireless downlinks," *IEEE Trans. Inf. Theory*, vol. 53, no. 9, pp. 3095–3113, Sep. 2007.

[15] G. Miao, N. Himayat, Y. G. Li, and A. Swami, "Cross-layer optimization for energy-efficient wireless communications: A survey," *Wireless Commun. Mobile Comput.*, vol. 9, no. 4, pp. 529–542, Apr. 2009.

[16] M. Hossain, K. Koufos, and R. Jantti, "Minimum-energy power and rate control for fair scheduling in the cellular downlink under flow level delay constraint," *IEEE Trans. Wireless Commun.*, vol. 12, no. 7, pp. 3253–3263, Jul. 2013.

[17] I. Kamitsos, L. Andrew, H. Kim, and M. Chiang, "Optimal sleep patterns for serving delay-tolerant jobs," in *Proc. 1st Int. Conf. Energy-Efficient Comput. Netw.*, 2010, pp. 31–40.
[18] Z. Niu, J. Zhang, X. Guo, and S. Zhou, "On energy-delay tradeoff in base station sleep mode operation," in *Proc. IEEE 13th Int. Conf. Commun. Syst. (ICCS)*, Nov. 2012, pp. 235–239.
[19] 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. Sel. Areas Commun.*, vol. 33, no. 4, pp. 641–650, Apr. 2015.
[20] J. Wu, S. Zhou, and Z. Niu, "Traffic-aware base station sleeping control and power matching for energy-delay tradeoffs in green cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 8, pp. 4196–4209, 2013.
[21] S.-E. Elayoubi, L. Saker, and T. Chahed, "Optimal control for base station sleep mode in energy efficient radio access networks," in *Proc. IEEE INFOCOM*, Apr. 2011, pp.106–110.
[22] I. Ashraf, F. Boccardi, and L. Ho, "Power savings in small cell deployments via sleep mode techniques," in *Proc. IEEE Int. Symp. Pers. Indoor Mobile Radio Commun. Workshops (PIMRC)*, sept. 2010, pp. 307–311.
[23] G. Auer *et. al.*, "How much energy is needed to run a wireless network?," *IEEE Wireless Commun.*, vol. 18, no. 5, pp. 40–49, Oct. 2011.
[24] B. Stark, G. Szarka, and E. Rooke, "Start-up circuit with low minimum operating power for microwatt energy harvesters," *IET Circuits Devices Syst.*, vol. 5, no. 4, pp. 267–274, Jul. 2011.
[25] L. Wang, X. Feng, X. Gan, J. Liu, H. Yu, and D. Zhang, "Small cell switch policy: A consideration of start-up energy cost," in *Proc. IEEE Int. Conf. Commun. China*, Oct. 2014, pp. 231–235.
[26] H. Takagi, *Queueing Analysis: A Foundation of Performance Evaluation. Volume 1: Vacation and Priority Systems*. Amsterdam, The Netherlands: Elsevier, 1991.
[27] Y. Sakai, Y. Takahashi, Y. Takahashi, and T. Hasegawa, "A composite queue with vacation/set-up/close-down times for SVCC in IP over ATM networks," *J. Oper. Res. Soc. Jpn.*, vol. 41, no. 1, pp. 68–80, Mar. 1998.
[28] Z. Niu and Y. Takahashi, "A finite-capacity queue with exhaustive vacation/close-down/setup times and Markovian arrival processes," *Queueing Syst.*, vol. 31, no. 1, pp. 1–23, Mar. 1999.
[29] S. M. Ross, *Introduction to Probability Models*. New York, NY, USA: Academic, 1997.
[30] J. D. C. Little, "Or forum–Little's law as viewed on its 50th anniversary," *Oper. Res.*, vol. 59, no. 3, pp. 536–549, 2011.
[31] S. Aalto, U. Ayesta, S. Borst, V. Misra, and R. Núñez Queija, "Beyond processor sharing," *SIGMETRICS Perform. Eval. Rev.*, vol. 34, no. 4, pp. 36–43, 2007.
[32] M. A. Imran *et al.*, "D2.3: Energy efficiency analysis of the reference systems, areas of improvements and target breakdown," EARTH, Tech. Rep. INFSO-ICT-247733, 2011 [Online]. Available: https://www.ict-earth.eu/
[33] H. Holtkamp *et al.*, "D2.2: Definition and parameterization of reference systems and scenarios," EARTH, Tech. Rep. INFSO-ICT-247733, 2011 [Online]. Available: https://www.ict-earth.eu/
[34] R. B. Cooper, *Introduction to Queueing Theory*. Amsterdam, The Netherlands: Elsevier 1981.

Xueying Guo (S'14) received the B.E. degree in electronic engineering from Tsinghua University, Beijing, China, in 2011. She is currently pursuing the Ph.D. degree at Tsinghua University. From October 2013 to October 2014, she was a Visiting Scholar at Computer Engineering and System Group, Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX, USA. Her research interests include radio resource management of wireless networks, cyber-physical systems, and green wireless communications. She was the recipient of the Best Student Paper Award from the 25th International Teletraffic Congress (ITC) in 2013.

Zhisheng Niu (F'12) received the degree from Beijing Jiaotong University, Beijing, China, in 1985, and the M.E. and D.E. degrees from Toyohashi University of Technology, Toyohashi, Japan, in 1989 and 1992, respectively. From 1992 to 1994, he worked for Fujitsu Laboratories Ltd., Japan, and in 1994, joined with Tsinghua University, Beijing, China, where he is now a Professor with the Department of Electronic Engineering. He is also a Guest Chair Professor with Shandong University, Shandong University, China. His research interests include queueing theory, traffic engineering, mobile Internet, radio resource management of wireless networks, and green communication and networks. He has been an active volunteer for various academic societies, including the Director for Conference Publications (2010–2011) and the Director for Asia-Pacific Board (2008–2009) of the IEEE Communication Society, Membership Development Coordinator (2009–2010) of the IEEE Region 10, Councilor of IEICE-Japan (2009-11), and Council Member of Chinese Institute of Electronics (2006–2011). He is now a Distinguished Lecturer (2012–2015) and the Chair of Emerging Technology Committee (2014–2015) of the IEEE Communication Society, a Distinguished Lecturer (2014–2016) of the IEEE Vehicular Technologies Society, a member of the Fellow Nomination Committee of IEICE Communication Society (2013–2014), standing committee member of Chinese Institute of Communications (CIC, 2012–2016), and the Associate Editor-in-Chief of the IEEE/CIC joint publication *China Communications*. He is now the Chief Scientist of the National Basic Research Program (so called "973 Project") of China on Fundamental Research on the Energy and Resource Optimized Hyper-Cellular Mobile Communication System (2012-2016), which is the first national project on green communications in China. He is a fellow of IEICE. He was the recipient of the Outstanding Young Researcher Award from the Natural Science Foundation of China in 2009 and the Best Paper Award from the IEEE Communication Society Asia-Pacific Board in 2013. He was also the corecipient of the Best Paper Awards from the 13th, 15th, and 19th Asia-Pacific Conference on Communication (APCC) in 2007, 2009, and 2013, respectively, the International Conference on Wireless Communications and Signal Processing (WCSP'13), and the Best Student Paper Award from the 25th International Teletraffic Congress (ITC25).

Sheng Zhou (M'12) received the B.E. and Ph.D. degrees in electronic engineering from Tsinghua University, Beijing, China, in 2005 and 2011, respectively. From January to June 2010, he was a visiting student at the Wireless System Laboratory, Department of Electrical Engineering, Stanford University, Stanford, CA, USA. He is currently an Assistant Professor with the Department of Electronic Engineering, Tsinghua University. His research interests include cross-layer design for multiple antenna systems, cooperative transmission in cellular systems, and green wireless communications. He was the corecipient of the Best Paper Award at the Asia-Pacific Conference on Communication in 2009 and 2013, the 23th IEEE International Conference on Communication Technology in 2011, and the 25th International Tele-traffic Congress in 2013.

P. R. Kumar (F'88) received the B.Tech. degree in electrical engineering (electronics) from the I.I.T. Madras, Chennai, India, in 1973, and the M.S. and D.Sc. degrees in systems science and mathematics from Washington University, St. Louis, MO, USA, in 1975 and 1977, respectively. From 1977–1984, He was a Faculty Member with the Department of Mathematics, University of Maryland Baltimore County, Baltimore, MD, USA. From 1985 to 2011, he was a Faculty Member with the Department of Electrical and Computer Engineering and the Coordinated Science Laboratory, University of Illinois, Champaign, IL, USA. Currently, he is with Texas A&M University, College Station, TX, USA, where he holds the College of Engineering Chair in Computer Engineering. He has worked on problems in game theory, adaptive control, stochastic systems, simulated annealing, neural networks, machine learning, queuing networks, manufacturing systems, scheduling, wafer fabrication plants and information theory. His research interests include wireless networks, sensor networks, cyber-physical systems, and the convergence of control, and communication and computation. He is a member of the National Academy of Engineering of the USA, and the Academy of Sciences of the Developing World. He received an honorary doctorate from ETH, Zurich, and is a Guest Chair Professor at Tsinghua University, Beijing, China. He is an Honorary Professor at IIT Hyderabad. He was the recipient of the IEEE Field Award for Control Systems, the Donald P. Eckman Award of the American Automatic Control Council, the Fred W. Ellersick Prize of the IEEE Communications Society, and the Outstanding Contribution Award of ACM SIGMOBILE. He was also the recipient of the Daniel C. Drucker Eminent Faculty Award from the College of Engineering at the University of Illinois.