Traffic-Aware Base Station Sleeping Control and Power Matching for Energy-Delay Tradeoffs in Green Cellular Networks

Jian Wu, *Student Member, IEEE,* Sheng Zhou, *Member, IEEE,* and Zhisheng Niu, *Fellow, IEEE*

***Abstract*—In this paper, traffic-aware sleeping control (SC) and power matching (PM) of a single base station (BS) in cellular networks are studied. The objective is to find the sleeping control and power matching configurations that achieve the Pareto optimal tradeoff between total power consumption and average delay. Two types of sleeping control schemes are considered: The BS goes to sleep whenever there is no active user, and wakes up when N users are assembled or after a period of multiple or single vacation time. We first discuss when to incorporate sleeping control into power matching energy efficiently. The explicit relationship between total power consumption and average delay with varying service rate is analyzed theoretically, indicating that sacrificing delay cannot always be traded for energy saving, and we also provide conditions under which the energy-optimal rate exists. Moreover, the optimal pair of sleeping parameter and service rate to achieve the optimal energy-delay tradeoff, and the energy consumption lower bound are also derived. Both the analytical and simulation results show that tolerable sacrifice of delay performance can be traded for substantial amount of energy saving given that careful designs were made according to our analysis.**

***Index Terms*—Traffic-aware, sleeping control, power matching, energy-delay tradeoff.**

I. Introduction

WIRELESS cellular networks have been growing tremendously during the last decade. Due to the popularity of smartphones, the demand for cellular data traffic has explosively increased, which has triggered a vast expansion of network infrastructures, resulting in dramatically increased energy consumption [2]. To deal with the rising energy cost, a new research area called "green cellular networks" has emerged to enable an energy-efficient cellular networks [3].

For cellular networks, BSs are the dominant components consuming 60-80% of the total energy [4]. The author in [5] shows that we can seize the opportunity to trace the traffic variation in the temporal and spatial domain and adapt the radio resource accordingly, and thus a great amount of energy can be saved. As one of the most popular and efficient traffic-oriented energy saving schemes, BS sleeping has been studied to realize substantial energy saving when the traffic load is low [4]- [7]. Besides, transmit power adaptation to match the traffic load during the working period is also an effective way to avoid the over-provisioning power consumption, because even a small reduction in the BS transmit power enables considerable saving in overall energy consumption due to its influence on the operational power of amplifiers, cooling systems and so on [8]. In this paper we will focus on how to make use of these two schemes, *BS sleeping control* and *power matching*, jointly to achieve a good tradeoff between energy saving and user Quality of Service (QoS).

There has been few work on optimizing these two schemes jointly. In some classical literature [9]- [11], ideas similar to the user number or vacation based sleeping design have been studied, where single server queueing analysis is carried out. Recently in [12], adopting the Markov Decision Process (MDP), the authors prove that the optimal sleeping pattern of serving delay-tolerant jobs for a typical server has a simple hysteretic structure. On the other hand, traditional power adaptation schemes mainly focus on combatting channel fading and controlling interference [13] [14]. However, recently there have been efforts to introduce the traffic-aware power and rate adaptation. The authors in [8] study the power-sharing policies considering the spatial load difference. The similar power-aware speed scaling technique is also studied in computer systems [15] [16]. However, the references above are confined to only one aspect. We combine the sleeping control and power matching in one optimization problem, aiming at providing further insight for the energy-delay tradeoffs in cellular networks.

In our previous work [1], we mainly discuss the optimal queue-aware and load-aware power adaptation schemes, and make a preliminary consideration of the joint load-level power matching and sleeping control design. In this paper, we adopt an analytical approach towards designing optimal sleeping parameter and transmit power (and thus service rate) jointly for energy-delay tradeoffs with non-realtime user requests arrival at a BS. As pointed out in [17], the tradeoff between energy and delay usually deviates from the monotonic curve [18] [19] when practical concerns are considered. Therefore figuring out when and how to trade tolerable delay for energy is important for practical system design.

Specifically, with power matching, both user number N-

Manuscript received December 30, 2012; revised May 14, 2013; accepted June 19, 2013. The associate editor coordinating the review of this paper and approving it for publication was N. Sagias.

The authors are with Tsinghua National Laboratory for Information Science and Technology, Dept. of Electronic Engineering, Tsinghua University, Beijing 100084, P.R. China (e-mail: wujian09@mails.tsinghua.edu.cn, {sheng.zhou, niuzhs}@tsinghua.edu.cn).

Part of this work has been published in IEEE Globecom 2012 [1].

This work is sponsored in part by the National Basic Research Program of China (2012CB316001), the Nature Science Foundation of China (61201191, 60925002, 61021001), and Hitachi Ltd.

Digital Object Identifier 10.1109/TWC.2013.071613.122092

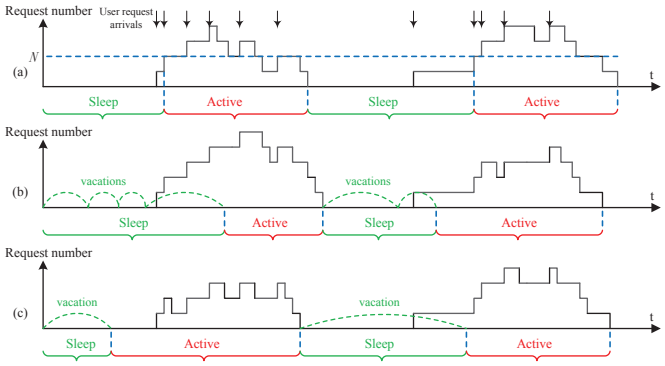


Fig. 1. Examples of different sleeping schemes: (a) *N-based* sleeping control; (b) *V-based* sleeping control (multiple vacations); (c) *V-based* sleeping control (single vacation).

based and vacation time *V-based* sleeping schemes are studied: the BS goes to sleep when the system is empty, and waits for *N users to assemble* or waits *a period of vacation time* before waking up from sleep. The user request arrivals are needed to be observed to wake the BS up once N user requests assemble in the *N-based* sleeping control. However, in practice, sleeping for a period of time is preferred due to its convenience, because there is no need to continuously monitor user request arrivals. Two kinds of vacation time *V-based* sleeping schemes will be analyzed: for the *multiple vacation* case, only discontinuous user requests observation is required to wake the BS up when at least one request is monitored at the end of each vacation period; for the *single vacation* case, no monitoring is needed and the BS just wakes up after a single period of vacation no matter whether there are users in the system or not. Fig. 1 shows examples of different sleeping schemes.

The main contributions of this paper are as follows.

- We optimize the BS sleeping control and power matching jointly for the optimal energy-delay tradeoff issue with the aim of providing references to practical green operation.
- We focus on specifically three joint sleeping control and power matching schemes under the system cost (a weighted combination of total power and average delay) minimization framework, and develop the optimal sleeping parameter and service rate. To this end, we inherit the hysteretic sleeping structure from [12] [32].
- For each "sleeping control + power matching" scheme, we obtain several key results: (i) Under what circumstance is the BS sleeping control beneficial for energy saving, otherwise only the power matching scheme should be used? This condition greatly depends on the traffic load, power consumption and sleeping parameters. (ii) Given the sleeping pattern, what's the explicit relationship between total power consumption and average delay with varying service rate? We find that the relationship curve is not always monotonic, and provide the conditions under which the energy optimal rate exists. (iii) For the optimal energy-delay tradeoff, as delay increases, we obtain the asymptotic values of the total power consumption under different traffic load environments, which are the minimum total power needed to support the offered traffic load.
- Some directions for the practical operation are suggested: (i) The "*N-based* sleeping control + power matching" scheme has the best energy-delay tradeoff relationship; the "*V-based* sleeping control + power matching" scheme, especially the single vacation case, is more practical as it does not need to monitor user requests during sleeping periods. (ii) Joint optimization brings substantial energy saving gain in a light-loaded system; while in a heavy-load system, whether sleeping control should be incorporated depends jointly on the load and delay requirement. Power matching has a wider range of adaptability to the traffic variation while sleeping control is more energy-efficient when the traffic load is relatively low.

In this paper we assume Poisson arrivals for the user requests as most literature did [7] [20] for the simplicity of analysis. For traffic with more bursty nature, Batch Poisson process($M^{[X]}/G/1$), or Markov modulated Poisson process (MMPP) can provide better approximations [35]. Also, the ideal assumption of continuous power/rate adaptation is made, which is needed to make the analytical model tractable explicitly, and can be treated as providing the performance upper bound. The impact of discrete transmission rates on the tradeoff performance is studied in the simulation. Finally, assuming the packet error has been guaranteed by the physical layer technique, the retransmission is not considered.

The rest of this paper is organized as follows: In Section II we describe the system model. Section III gives the power matching scheme without sleeping control as a baseline. In Section IV, with power matching, *N-based*, *multiple vacation* and *single vacation V-based* sleeping schemes are studied. Numerical and simulation results are provided in Section V, and Section VI gives the conclusions.

II. System Model

We consider a single BS scenario where users arrive according to a Poisson process with arrival rate λ. Each user requires a random amount of downlink service with average length l bits, e.g., non-realtime file download with average file size l, and then the user leaves after being served. Assuming the arrival rate can be well estimated [30], with time varying traffic intensity in practice, we just need to operate according to the current arrival rate.

The M/G/1 processor-sharing (PS) model is used here [20] [21]. Assume that the BS service capacity or service rate is x bits per second, which adapts to the system traffic load and is equally shared by all users being served. So the user departure rate is $\mu = x/l$, and the traffic load at the BS is $\rho = \lambda/\mu = \lambda l/x$. The relationship between the service rate x and the transmit power P_t is

$$x = B\log_2(1+\gamma P_t), \quad P_t \in [0, P_t^{max}] \tag{1}$$

where $\gamma = \frac{\eta g}{N_0 B}$, g[1] represents the channel gain, B is the bandwidth, N_0 denotes the noise density, and η is a constant

[1]We first study the basic case: users experience homogeneous channels with gain g. When heterogeneous channel conditions are considered, the multi-class PS model can be used, which will be discussed in our future work.

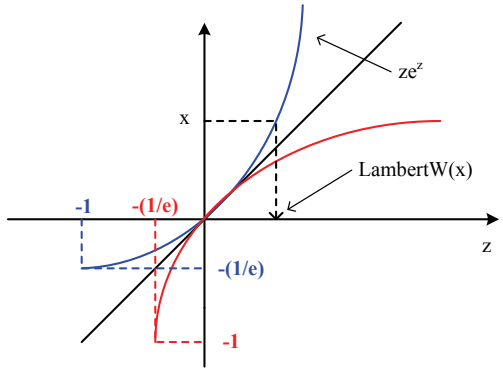


Fig. 2. Lambert function $\mathbf{W}$ is the inverse function of ze^z.

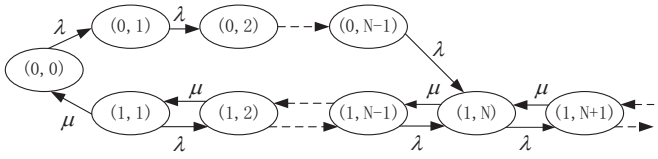


Fig. 3. The 2-D state transition diagram for the *N-based* sleeping scheme.

related to bit error rate (BER) requirement when adaptive modulation and coding is used [22]. Since the service rate x and the transmit power P_t are directly linked in this way, we use service rate matching and power matching interchangeably.

We assume the BS has *active* mode and *sleep* mode, with the power consumption P_{BS} as follows [24] [25]:

$$P_{BS} = \begin{cases} P_o + \Delta_P P_t, & \textit{active} \text{ mode}, \\ P_{sleep}, & \textit{sleep} \text{ mode}. \end{cases} \quad (2)$$

P_o and P_{sleep} are the static power consumption in the active mode and sleep mode respectively, and Δ_P is the slope of the load-dependent power consumption, where the transmit power P_t adapts to the system traffic load. It is also assumed that there is a fixed switching energy cost E_s for each mode transition.

The objective is to minimize the system cost, which is a weighted combination of average total power $P^{tot}_{\mathcal{T}_{sleep},x}$ and average delay $D_{\mathcal{T}_{sleep},x}$. Here $\mathcal{T}_{sleep}$ represents the parameter in the BS sleeping control: For the *N-based* sleeping control, $\mathcal{T}_{sleep} = N$; for the *V-based* sleeping control, $\mathcal{T}_{sleep}$ represents the mean time of each vacation; $\mathcal{T}_{sleep} = 0$ denotes that there is no sleeping control. We seek to find the service rate x and sleeping control parameter $\mathcal{T}_{sleep}$ that minimize

$$z(\mathcal{T}_{sleep}, x) = P^{tot}_{\mathcal{T}_{sleep},x} + \beta D_{\mathcal{T}_{sleep},x}. \quad (3)$$

The positive weighting factor β indicates the relative importance of the average delay over the average power which can be thought of as a Lagrange multiplier on an average delay constraint [18] [36]. The "delay" we consider in this paper is the average response time from the user's service request arriving at the BS until this request is finished. From the Little's Law, we know that the mean delay is directly related to the average queue length.

III. Traffic-aware Power Matching Scheme

First, we only focus on the power matching scheme without BS sleeping where the transmit power will be adapted to match the average traffic load.

With service rate x, the total power consumption is

$$P^{tot}_{0,x} = P_o + \frac{\lambda l}{x}\frac{\Delta_P}{\gamma}(2^{\frac{x}{B}} - 1), \quad (4)$$

where $\frac{\lambda l}{x}$ is the busy probability. According to the property of M/G/1 PS queue [33] and the Little's Law, the average delay is

$$D_{0,x} = \frac{l}{x - \lambda l}. \quad (5)$$

The system cost is then given by

$$z(\mathcal{T}_{sleep}, x)|_{\mathcal{T}_{sleep}=0} = P_o + \frac{\lambda l}{x}\frac{\Delta_P}{\gamma}(2^{\frac{x}{B}} - 1) + \frac{\beta l}{x - \lambda l}, \quad (6)$$

where the transmit power is only consumed in the busy period.

By taking $\frac{dz(\mathcal{T}_{sleep}=0,x)}{dx} = 0$, we obtain

$$\mathbf{W}\Big[\frac{\gamma\beta}{e\lambda\Delta_P}\Big(\frac{x}{x-\lambda l}\Big)^2 - \frac{1}{e}\Big] = \frac{\ln 2}{B}x - 1. \quad (7)$$

Note that the convexity of all the objective functions has been proved in Appendix A. Here the Lambert function $\mathbf{W}$ is adopted. It is the inverse of $f(w) = we^w$ and is defined as [26]

$$\mathbf{W}(z)e^{\mathbf{W}(z)} = z, z \in \mathbb{C}. \quad (8)$$

The real branch of the Lambert function is $\mathbf{W}_0 : \mathcal{D}_{\mathbf{W}_0} = [-e^{-1}, +\infty) \mapsto [-1, +\infty)$ as shown in Fig. 2, and we will just use $\mathbf{W}$ in this paper for the sake of simplicity. By solving Eq.(7), we obtain the optimal rate $x^*(\rho)$ as a function of the traffic load. This solution is unique due to the fact that, the left side of Eq.(7) is a decreasing function of x which approaches ∞ as $x \to \lambda l$ and equals to $\mathbf{W}\big(\frac{\gamma\beta}{e\lambda\Delta_P} - \frac{1}{e}\big)$ as $x \to \infty$, while the right side is an increasing function of x, and equals to $\frac{\lambda l \ln 2}{B} - 1$ as $x \to \lambda l$ and approaches ∞ as $x \to \infty$. As $\frac{\partial x^*}{\partial \lambda} > 0$, $\frac{\partial x^*}{\partial l} > 0$ and $\frac{\partial x^*}{\partial \gamma} > 0$, x^* is an increasing function of the traffic load and the channel gain. Note that for the optimized service rate here and in the following discussion, the corresponding transmit power cannot exceed its maximum constraint, otherwise, the maximum power P_t^{max} will be used.

IV. Traffic-aware Sleeping Control with Power Matching

In this section, we consider the BS sleeping control and power matching jointly. "*N-based* sleeping control with power matching (N_SC + PM)", "*multiple vacation based* sleeping control with power matching (MV_SC + PM)" and "*single vacation based* sleeping control with power matching (SV_SC + PM)" schemes are analyzed respectively.

A. N-based sleeping control with power matching (N_SC + PM)

Assume the BS goes to sleep when there is no service request and returns to active mode once N user requests assemble. Here the control variables are N and service rate x. Using an extended-Markov-chain shown in Fig. 3 with departure rate $\mu = x/l$, the static probability distribution is given below, and the proof is in Appendix B. Here we define an extended state space $\{(i, j) : i = 0, j = 0, 1, \ldots, N-1; i =$

$1, j = 1, 2, \ldots\}$ such that if $i = 0$ then j denotes the number of users in the system when the BS is in sleep mode, and if $i = 1$ then j counts the number of users in the system when the BS is in active mode.

$$\Pr(i,j)=\begin{cases}\frac{x-\lambda l}{Nx} & \text{if } i=0;\\ \frac{\lambda l}{Nx}(1-(\frac{\lambda l}{x})^j) & \text{if } i=1, 1\le j\le N;\\ \frac{\lambda l}{Nx}((\frac{\lambda l}{x})^{j-N}-(\frac{\lambda l}{x})^j) & \text{if } i=1, j>N.\end{cases} \tag{9}$$

After some calculation, we obtain the probability in sleep mode as

$$\sum_{j=0}^{N-1}\Pr(0,j)=1-\frac{\lambda l}{x}, \tag{10}$$

and the average queue length is

$$\sum_{j=0}^{N-1} j\Pr(0,j)+\sum_{j=1}^{\infty} j\Pr(1,j)=\frac{\lambda l}{x-\lambda l}+\frac{N-1}{2}. \tag{11}$$

With the sleeping control, the energy cost for mode transitions has to be dealt with. The sleep period starts when the system is empty and lasts until N users assemble with the average assembling time to be N/λ. At the beginning of each active period there are N users in the system, and thus the average working time is $\frac{N}{\mu-\lambda}$. Therefore, the mode transition frequency F_m, which is defined as the mode transition times between active mode and sleep mode per unit time, is

$$F_m=\frac{2}{\frac{N}{\lambda}+\frac{N}{\mu-\lambda}}=\frac{2\lambda}{N}(1-\frac{\lambda}{\mu}), \tag{12}$$

and the mode switching energy cost per unit time is $E_s F_m$.

Then the total power consumption $P_{N,x}^{tot}$ under the *N-based* sleeping control with service rate x is

$$P_{N,x}^{tot}=\frac{\lambda l}{x}[P_o+\frac{\Delta_P}{\gamma}(2^{\frac{x}{B}}-1)]+(1-\frac{\lambda l}{x})[P_{sleep}+\frac{2\lambda E_s}{N}]. \tag{13}$$

Using the Little's Law, the average delay $D_{N,x}$ is given by

$$D_{N,x}=\frac{l}{x-\lambda l}+\frac{N-1}{2\lambda}. \tag{14}$$

By comparing it with the power matching scheme without sleeping control in Section III and analyzing the properties of $P_{N,x}^{tot}[D_{N,x}]$, we have the following two propositions.

Proposition 1: Compared with the power matching only scheme, using the "*N-based* sleeping control + power matching" scheme brings energy saving gain in the total power consumption only when

$$\lambda<\frac{(P_o-P_{sleep})N}{2E_s}. \tag{15}$$

Remark: The situation that energy saving gain exists greatly depends on the relationship between the energy consumption parameter $\frac{P_o-P_{sleep}}{E_s}$, the traffic arrival rate λ and the sleeping parameter N. Fig. 4 shows whether it is energy-efficient to incorporate the *N-based* sleeping control into the power matching scheme. Above the surface $\lambda=\frac{(P_o-P_{sleep})N}{2E_s}$, incorporating sleeping control brings energy saving gain. However, below the surface where sleeping increases energy cost, sleeping is harmful due to the extra mode switching cost. Therefore, under this situation, only adopting the power matching is enough.

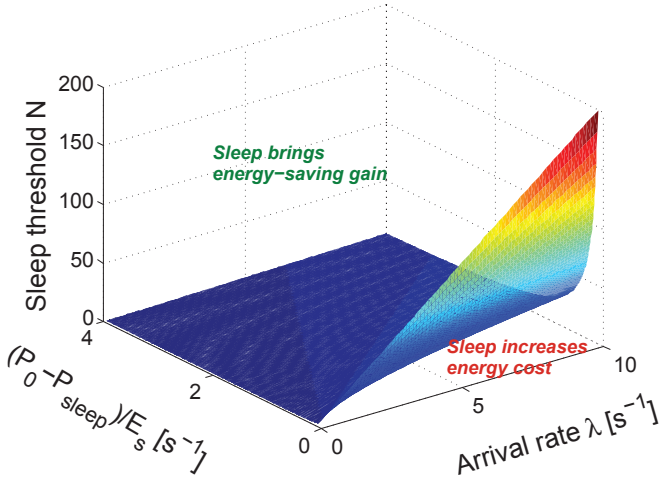


Fig. 4. Whether it is energy efficient to incorporate the *N-based* sleeping control into power matching.

Proposition 2: For the "*N-based* sleeping control + power matching" scheme, with varying service rate x, we have:

1. $P_{N,x}^{tot}$ is a monotonically decreasing function of $D_{N,x}$, if one of the following conditions is satisfied,

1) $\lambda<\frac{(P_o-P_{sleep})N}{2E_s}$,
$l\ge\frac{B}{\lambda\ln 2}\left\{\mathbf{W}[\frac{\gamma}{\Delta_P e}(P_o-P_{sleep}-\frac{2\lambda E_s}{N})-\frac{1}{e}]+1\right\}$.

2) $\lambda\ge\frac{(P_o-P_{sleep})N}{2E_s}$.

2. For $P_{N,x}^{tot}[D_{N,x}]$, the unique energy-optimal rate x_{en}^* exists if the following condition is satisfied,

$\lambda<\frac{(P_o-P_{sleep})N}{2E_s}$,
$l<\frac{B}{\lambda\ln 2}\left\{\mathbf{W}[\frac{\gamma}{\Delta_P e}(P_o-P_{sleep}-\frac{2\lambda E_s}{N})-\frac{1}{e}]+1\right\}$.

The corresponding energy-optimal rate is

$$x_{en}^*=\frac{B}{\ln 2}\left\{\mathbf{W}[\frac{\gamma}{\Delta_P e}(P_o-P_{sleep}-\frac{2\lambda E_s}{N})-\frac{1}{e}]+1\right\}. \tag{16}$$

3. In both of the upper two cases, as the average delay increases, the total power consumption approaches an asymptotic value of $P_o+\frac{\Delta_P}{\gamma}(2^{\frac{\lambda l}{B}}-1)$.

Proof: See Appendix C. ■

Remark: The properties of $P_{N,x}^{tot}[D_{N,x}]$ depend on the traffic parameters λ and l, system bandwidth B, channel condition γ, sleeping parameter N and the power consumption parameters. For the case that the energy-optimal service rate exists, only when $x>x_{en}^*$ delay can be traded off with energy; otherwise, increasing delay will only cause bad energy performance. Interestingly, x_{en}^* is an increasing function of γ, and thus transmitting faster when channels are good indeed saves energy. In addition, transmitting fast is beneficial when the gap between P_o and P_{sleep} is large; otherwise large busy probability will cause too much static energy consumption. As the delay increases, the asymptotic limit of the total power consumption $P_o+\frac{\Delta_P}{\gamma}(2^{\frac{\lambda l}{B}}-1)$ corresponds to the total power consumption when the system is always in active mode with

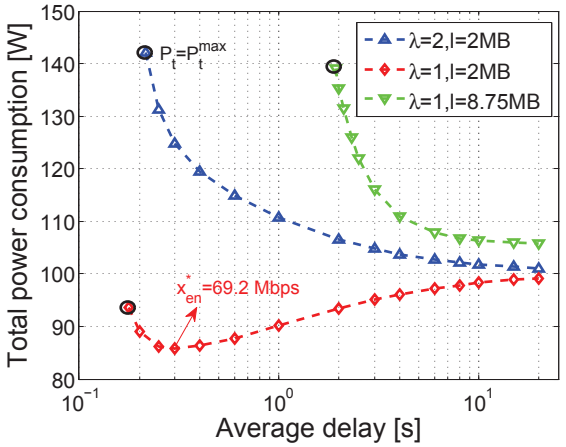


Fig. 5. Example of Proposition 2 with $N = 1$: the relationship of $P_{N,x}^{tot}$ and $D_{N,x}$ with varying service rate x. $P_o = 100$W, $P_{sleep} = 30$W, $E_s = 25$J, $\Delta_P = 7$ [24].

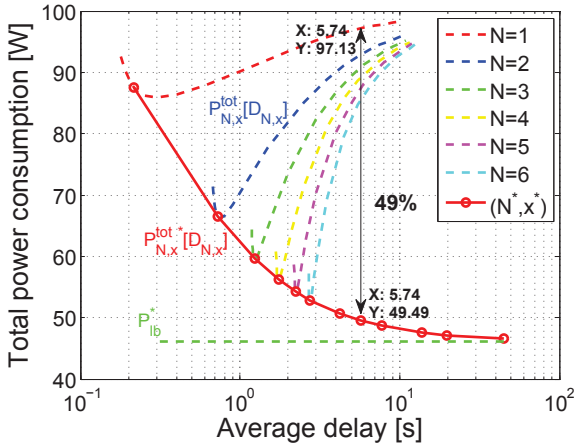


Fig. 6. The energy-delay tradeoff of the "*N-based* sleeping control + power matching" scheme. $\lambda = 1, l = 2$MB.

system utilization $\frac{\lambda l}{x}$ approaching 1. Fig. 5 shows one example of Proposition 2 with $N = 1$ where for each curve the service rate x is changing. The upper two curves correspond to the two monotonic cases respectively, and for the curve the energy optimal rate exists, we have $x_{en}^* = 69.2$Mbps.

After studying the properties of $P_{N,x}^{tot}[D_{N,x}]$, we are interested in minimizing the system cost

$$z(\mathcal{T}_{sleep}, x)|_{\mathcal{T}_{sleep}=N} = P_{N,x}^{tot} + \beta D_{N,x}. \tag{17}$$

Assume (N^*, x^*) is the optimal control pair for a given β, and P_{N^*,x^*}^{tot} and D_{N^*,x^*} are the corresponding total power and delay. Then P_{N^*,x^*}^{tot} must be the minimum power that the average delay is less than D_{N^*,x^*}. Define $P_{N,x}^{tot\,*}[D_{N,x}]$ to be the minimum total power so that the average delay is less than $D_{N,x}$, so we have $P_{N,x}^{tot\,*}[D_{N^*,x^*}] = P_{N^*,x^*}^{tot}$. We refer to $P_{N,x}^{tot\,*}[D_{N,x}]$ as the optimal power-delay curve. Varying β and finding the optimal control pair for each value can provide different points on the power-delay curve, which are also known as Pareto-optimal points [18] [36].

Proposition 3: For the optimal power-delay curve $P_{N,x}^{tot\,*}[D_{N,x}]$, we have the following properties:
1. Given the tradeoff parameter β, the optimal control pair is denoted by (N^*, x^*), where x^* is the unique solution of the following equation

$$\mathbf{W}\Big[\frac{\gamma}{\Delta_P e}\Big[\frac{\beta x^2}{\lambda(x-\lambda l)^2} + P_o - P_{sleep} - \frac{2\lambda E_s}{N} - \frac{\Delta_P}{\gamma}\Big]\Big] = \frac{\ln 2}{B}x - 1.$$

The optimal sleeping parameter N^* is

$$N^* = 2\lambda\Big[\frac{E_s}{\beta}\Big(1 - \frac{\lambda l}{x}\Big)\Big]^{1/2}. \tag{18}$$

2. As the average delay increases, $P_{N,x}^{tot\,*}[D_{N,x}]$ approaches an asymptotic value of P_{lb}^* where

1) for $\rho \geq \frac{B}{\ln 2}\Big\{\mathbf{W}\big[\frac{\gamma(P_o - P_{sleep})}{\Delta_P e} - \frac{1}{e}\big] + 1\Big\}$,

$$P_{lb}^* = P_{N\to\infty, x\to\lambda l}^{tot} = P_o + \frac{\Delta_P}{\gamma}(2^{\frac{\lambda l}{B}} - 1). \tag{19}$$

2) for $\rho < \frac{B}{\ln 2}\Big\{\mathbf{W}\big[\frac{\gamma(P_o - P_{sleep})}{\Delta_P e} - \frac{1}{e}\big] + 1\Big\}$,

$$P_{lb}^* = P_{N,x}^{tot}\Big\{N\to\infty, x = \frac{B}{\ln 2}\Big\{\mathbf{W}\big[\frac{\gamma(P_o - P_{sleep})}{\Delta_P e} - \frac{1}{e}\big] + 1\Big\}\Big\}$$
$$= P_{sleep} + \frac{\lambda l(P_o - P_{sleep} - \frac{\Delta_P}{\gamma})\ln 2}{B\mathbf{W}\big[\frac{\gamma(P_o - P_{sleep})}{\Delta_P e} - \frac{1}{e}\big]}. \tag{20}$$

Proof: See Appendix D. ■

Remark: Here x^* and N^* are obtained by solving $\frac{\partial z(\mathcal{T}_{sleep}=N,x)}{\partial x} = 0$ and $\frac{\partial z(\mathcal{T}_{sleep}=N,x)}{\partial N} = 0$ jointly. First, solve this equation $\mathbf{W}\Big[\frac{\gamma}{\Delta_P e}\Big[\frac{\beta x^2}{\lambda(x-\lambda l)^2} + P_o - P_{sleep} - \sqrt{\frac{\beta E_s}{1-\frac{\lambda l}{x}}} - \frac{\Delta_P}{\gamma}\Big]\Big] = \frac{\ln 2}{B}x - 1$ for the optimal service rate x^*; and then get N^* with Eq.(18). It can be observed that N^* is related to the switching cost in a square root form, which is similar to the result derived by Heyman [10] because it turns out that N only affects the average delay and the switching cost in the objective. N^* should be an integer and is chosen from $\{\lfloor N^* \rfloor, \lceil N^* \rceil\}$ which minimizes $z(\mathcal{T}_{sleep} = N, x)$ due to the convexity of the objective function. For the optimal tradeoff curve, the "endpoint" at the right depicts the smallest possible value of $P_{N,x}^{tot}$, without any consideration of $D_{N,x}$. The asymptotic limit P_{lb}^* is the minimum total power needed to support the offered traffic load.

In Fig. 6, the dashed lines show the $P_{N,x}^{tot}[D_{N,x}]$ curves with varying service rate x. It can be observed that with a larger N, more energy is saved through the sacrifice of delay performance. For the solid line, different points correspond to different values of β. At each point, $P_{N,x}^{tot\,*}$ and $D_{N,x}{}^*$ are the corresponding total power consumption and delay under the optimal control pair (N^*, x^*). Comparing the optimal tradeoff curve with the dashed lines shows how the joint optimization significantly improves the energy-delay performance: It not only removes undesirable energy-delay pairs which make the tradeoff line go up but also achieves significant energy savings (e.g., nearly 50% with average delay to be 5.74 seconds). Note that the energy-optimal points on the dashed lines are not on the solid optimal trade-off curve.

B. Multiple vacation based sleeping control with power matching (MV_SC + PM)

A vacation queue is used to model this situation [27] [34]. Assume that the BS goes to sleep once there is no user request, and it will be asleep for a period of time, which is treated as taking a "vacation". For the multiple vacation case, if the BS returns from a vacation and finds no user request waiting, it begins another vacation immediately in this manner until it finds at least one user request waiting upon returning from a vacation.

The control variables of this scheme are the average length of each vacation v and service rate x. The random variable V is used to represent the length of each vacation, which is assumed to be independently and identically distributed. We have its mean, second moment and Laplace-Stieltjes transform (LST) [34] denoted as $\mathbb{E}\{V\} = v$, $\mathbb{E}\{V^2\}$, and $\Gamma_V(z) = \mathbb{E}\{e^{-zV}\}$, respectively.

With the concept of the "vacation cycle" [34] comprised of the consecutive sleep period and active period, F_m is given by

$$F_m = \frac{2}{\frac{v}{1-\Gamma_V(\lambda)} + \frac{\lambda v}{1-\Gamma_V(\lambda)}\frac{1}{\mu-\lambda}} = \frac{2(1-\frac{\lambda}{\mu})(1-\Gamma_V(\lambda))}{v}. \quad (21)$$

$\Gamma_V(\lambda)$ is the probability that the BS returns from a vacation and finds no user waiting, and thus the average number of vacations in a multiple vacation period is $\frac{1}{1-\Gamma_V(\lambda)}$, and the sleep period length is $\frac{v}{1-\Gamma_V(\lambda)}$. With average $\frac{\lambda v}{1-\Gamma_V(\lambda)}$ users waiting to be served at the beginning of each active period, the average active period length is $\frac{\lambda v}{1-\Gamma_V(\lambda)}\frac{1}{\mu-\lambda}$. Then the mode switching energy cost per unit time is $E_s F_m$.

The fraction of time the BS spends on vacation is

$$\eta_v = \frac{F_m v}{2(1-\Gamma_V(\lambda))} = 1 - \frac{\lambda}{\mu}. \quad (22)$$

Therefore the total power consumption is obtained as $P_{v,x}^{tot} = (1-\eta_v)(P_o + \Delta_P P_t) + \eta_v P_{sleep} + E_s F_m$, which equals to

$$\begin{aligned} P_{v,x}^{tot} &= \frac{\lambda l}{x}[P_o + \frac{\Delta_P}{\gamma}(2^{\frac{x}{B}} - 1)] \\ &+ (1 - \frac{\lambda l}{x})[P_{sleep} + \frac{2E_s(1-\Gamma_V(\lambda))}{v}]. \end{aligned} \quad (23)$$

The average number of users in the system is given in Eq.(24). Here we assume that the users have an exponentially distributed service requirement. This is in agreement with the mean queue length in an M/M/1-FCFS (First Come First Service) queue where the first two terms come from the well-known Pollaczek-Khintchine formula [33]. Indeed, in the case of exponential service requirements, the queue length distribution for PS and FCFS are the same [34].

$$Q_v = \frac{\lambda}{\mu} + \frac{(\frac{\lambda}{\mu})^2}{1-\frac{\lambda}{\mu}} + \frac{\lambda \mathbb{E}\{V^2\}}{2v}. \quad (24)$$

Using the Little's Law, the average delay $D_{v,x}$ is

$$D_{v,x} = \frac{l}{x - \lambda l} + \frac{\mathbb{E}\{V^2\}}{2v}. \quad (25)$$

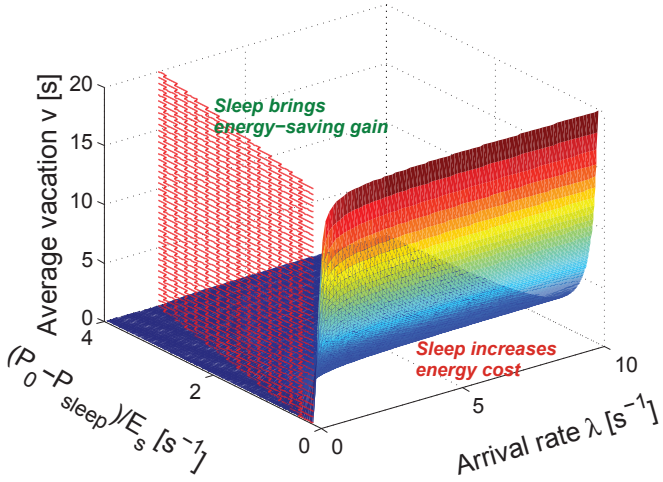


Fig. 7. Whether it is energy efficient to incorporate the *multiple vacation* sleeping control into power matching.

Actually only the last one of the multiple consecutive vacations contributes to the additional delay caused by the vacation, so the additional delay is just the expected residual life time of a vacation, which is $\frac{\mathbb{E}\{V^2\}}{2v}$.

Similar to Proposition 1, we can get Proposition 4 as follows.

Proposition 4: Compared with the power matching only scheme, using the "multiple vacation sleeping control + power matching" scheme brings energy saving gain in the total power consumption only when

$$\frac{P_o - P_{sleep}}{2E_s} > \frac{1-\Gamma_V(\lambda)}{v}. \quad (26)$$

Especially, for the exponentially distributed vacation (EXP), the condition becomes

$$v > \frac{2E_s}{P_o - P_{sleep}} - \frac{1}{\lambda}; \quad (27)$$

while for the deterministically distributed vacation (DET) where each vacation length is equal to v, the condition is

$$v > \frac{2E_s}{P_o - P_{sleep}} + \frac{\mathbf{W}\big[-\frac{2\lambda E_s}{P_o - P_{sleep}} e^{-\frac{2\lambda E_s}{P_o - P_{sleep}}}\big]}{\lambda}. \quad (28)$$

In both of the upper two cases, the energy-saving region that has nothing to do with v is

$$\frac{P_o - P_{sleep}}{2E_s} \geq \lambda. \quad (29)$$

Proof: See Appendix E. ■

Remark: Similar to Fig. 4, whether it is energy efficient to incorporate the multiple vacation sleeping control with exponential vacation into the power matching scheme is shown in Fig. 7. Especially, the region on the left of the red plane in dashed lines corresponds to the situation that $\lambda \leq \frac{P_o - P_{sleep}}{2E_s}$. For the deterministic vacation, a similar figure can be obtained and it is omitted here.

The properties of $P_{v,x}^{tot}[D_{v,x}]$ are analyzed in Proposition 5. The proof is similar to Proposition 2 and is omitted.

Proposition 5: For the "multiple vacation sleeping control + power matching" scheme with varying service rate x, taking

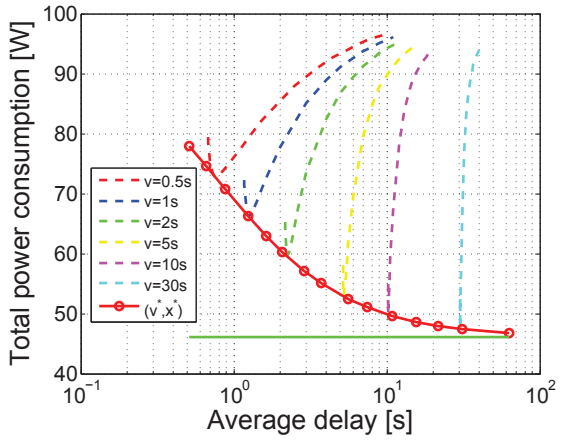


Fig. 8. The energy-delay tradeoff of the "*multiple vacation* sleeping control + power matching" scheme. $\lambda = 1$,$l = 2$MB.

the exponential vacation as an example, we have:
1. $P_{v,x}^{tot}$ is a monotonically decreasing function of $D_{v,x}$, if one of the following conditions is satisfied,

$$1)\quad \frac{1}{\lambda} > \frac{2E_s}{P_o - P_{sleep}} - v,$$
$$l \geq \frac{B}{\lambda \ln 2}\Big\{\mathbf{W}\big[\frac{\gamma}{\Delta_P e}(P_o - P_{sleep} - \frac{2\lambda E_s}{1+\lambda v}) - \frac{1}{e}\big] + 1\Big\}.$$
$$2)\quad \frac{1}{\lambda} \leq \frac{2E_s}{P_o - P_{sleep}} - v.$$

2. For $P_{v,x}^{tot}[D_{v,x}]$, the unique energy-optimal rate x_{ev}^* exists if the following condition is satisfied,

$$\frac{1}{\lambda} > \frac{2E_s}{P_o - P_{sleep}} - v,$$
$$l < \frac{B}{\lambda \ln 2}\Big\{\mathbf{W}\big[\frac{\gamma}{\Delta_P e}(P_o - P_{sleep} - \frac{2\lambda E_s}{1+\lambda v}) - \frac{1}{e}\big] + 1\Big\}.$$

The corresponding energy-optimal rate is

$$x_{ev}^* = \frac{B}{\ln 2}\Big\{\mathbf{W}\big[\frac{\gamma}{\Delta_P e}(P_o - P_{sleep} - \frac{2\lambda E_s}{1+\lambda v}) - \frac{1}{e}\big] + 1\Big\}. \quad (30)$$

3. In both of the upper two cases, as the average delay increases, the total power consumption approaches an asymptotic value of $P_o + \frac{\Delta_P}{\gamma}(2^{\frac{\lambda l}{B}} - 1)$.

Then we turn to minimize the system cost

$$z(\mathcal{T}_{sleep}, x)|_{\mathcal{T}_{sleep}=v} = P_{v,x}^{tot} + \beta D_{v,x}. \quad (31)$$

Assume (v^*, x^*) is the optimal control pair for a given β, and the Pareto-optimal points on the optimal power-delay curve $P_{v,x}^{tot*}[D_{v,x}]$ can be obtained by varying β similar to the *N-based* scheme.

Proposition 6: For the optimal power-delay curve $P_{v,x}^{tot*}[D_{v,x}]$, given the tradeoff parameter β, in the optimal control pair (v^*, x^*),
1. For the exponentially distributed vacation (EXP), x^* is the unique solution of the following equation

$$\mathbf{W}\Big[\frac{\gamma}{\Delta_P e}\Big[\frac{\beta x^2}{\lambda(x-\lambda l)^2} + P_o - P_{sleep} - \frac{2\lambda E_s}{1+\lambda v} - \frac{\Delta_P}{\gamma}\Big]\Big] = \frac{\ln 2}{B}x - 1,$$

and the optimal sleeping parameter v^* is

$$v^* = \Big[\frac{2E_s}{\beta}(1 - \frac{\lambda l}{x})\Big]^{1/2} - \frac{1}{\lambda}. \quad (32)$$

2. For the deterministically distributed vacation (DET), v^* is the unique solution of the following equation

$$\mathbf{W}\Big[\frac{\gamma}{\Delta_P e}\Big[\frac{\beta x^2}{\lambda(x-\lambda l)^2} + P_o - P_{sleep} - \frac{2E_s(1-e^{-\lambda v})}{v} - \frac{\Delta_P}{\gamma}\Big]\Big] = \frac{\ln 2}{B}x - 1,$$

and the optimal service rate x^* is

$$x^* = \frac{\lambda l}{1 - \frac{\beta v^2}{4E_s[1-e^{-\lambda v}(1+\lambda v)]}}. \quad (33)$$

The proof is similar to Proposition 3 and is omitted here. Fig. 8 provides the curves of $P_{v,x}^{tot}[D_{v,x}]$ versus varying service rates under different vacation parameters in the dashed lines. The optimal tradeoff curve $P_{v,x}^{tot*}[D_{v,x}]$ in the solid line removes the undesirable energy-delay pairs which make the $P_{v,x}^{tot}[D_{v,x}]$ relationships go up through the joint optimization of sleeping parameter v and service rate x. Note that the asymptotic lower bound is the same with the asymptotic limit of the *N-based* sleeping control obtained in Proposition 3.

C. Single vacation based sleeping control with power matching (SV_SC + PM)

Besides the multiple vacation situation, we study the single vacation based sleeping control with power matching in this section. Assume that the BS goes into sleep once there is no user request, and it will be asleep to take a single vacation. Then the BS will wake up no matter whether there are user requests in the system or not at the end of this vacation [34].

The random variable V is used to represent the vacation duration with its mean to be $\mathbb{E}\{V\} = v$ as in the previous section. Similar to the multiple vacation case from the "vacation cycle", the mode transition frequency is given by

$$F_m = \frac{2}{v + \Gamma_V(\lambda)(\frac{1}{\lambda} + \frac{1}{\mu-\lambda}) + \frac{\lambda v}{\mu-\lambda}} = \frac{2(1-\frac{\lambda}{\mu})}{v + \frac{\Gamma_V(\lambda)}{\lambda}}. \quad (34)$$

Here the average sleep period length is v. If the BS returns from the vacation finding no user waiting, it becomes idle to wait for an arrival. Together with the average λv users assembled in the single vacation period, the average active period is $\Gamma_V(\lambda)(\frac{1}{\lambda} + \frac{1}{\mu-\lambda}) + \frac{\lambda v}{\mu-\lambda}$. Then the mode switching energy cost per unit time to be $E_s F_m$.

The fraction of time the BS spends on vacation η_v is

$$\eta_v = \frac{vF_m}{2} = \frac{\lambda v(1-\frac{\lambda}{\mu})}{\lambda v + \Gamma_V(\lambda)}. \quad (35)$$

Different from the *N_based* and *multiple vacation* sleeping controls where the BS will always be busy in an active period, in the single vacation case, if the BS wakes up finding no users, it will go through an idle period first before entering into the busy period. Using $\widetilde{P}_t$ to denote the average transmit power consumption in the active period, it is obtained that

$$\widetilde{P}_t = \big[\Gamma_V(\lambda)(1 - \frac{1/\lambda}{2/F_m - v}) + (1 - \Gamma_V(\lambda))\big]P_t, \quad (36)$$

where the first part of the summation corresponds to the case that there is no user request in the system when the BS wakes up from the vacation, and the second part $(1 - \Gamma_V(\lambda))P_t$ represents the situation that there is at least one user request waiting when the BS wakes up from the vacation. Therefore

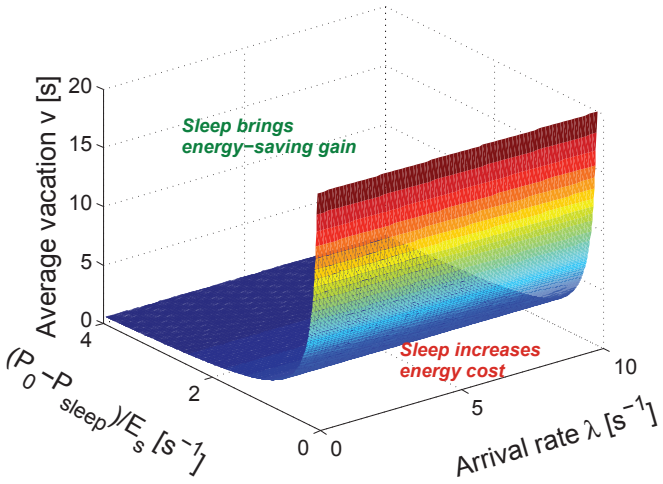


Fig. 9. Whether it is energy efficient to incorporate the *single vacation* sleeping control into power matching.

the total power consumption is given as $P_{v,x}^{tot} = (1-\eta_v)\big[P_o + \Delta_P \widetilde{P}_t\big] + \eta_v P_{sleep} + E_s F_m$. After some calculation, it is as follows.

$$P_{v,x}^{tot} = \frac{\frac{\lambda^2 vl}{x} + \Gamma_V(\lambda)}{\lambda v + \Gamma_V(\lambda)} P_o + \frac{\lambda l}{x}\Big[\frac{\Delta_P}{\gamma}(2^{\frac{x}{B}} - 1)\Big] + \frac{\lambda v(1-\frac{\lambda l}{x})}{\lambda v + \Gamma_V(\lambda)}\big[P_{sleep} + \frac{2E_s}{v}\big]. \quad (37)$$

The mean number of users in the system is given in Eq.(41). Similar to Eq.(24), the third term is the result of the vacation.

$$Q_v = \frac{\lambda}{\mu} + \frac{(\frac{\lambda}{\mu})^2}{1-\frac{\lambda}{\mu}} + \frac{\lambda^2 \mathbb{E}\{V^2\}}{2(\lambda v + \Gamma_V(\lambda))}. \quad (41)$$

Using the Little's Law, the average delay $D_{v,x}$ is

$$D_{v,x} = \frac{l}{x - \lambda l} + \frac{\lambda \mathbb{E}\{V^2\}}{2(\lambda v + \Gamma_V(\lambda))}. \quad (42)$$

Proposition 7: Compared with the power matching only scheme, using the "single vacation sleeping control + power matching" scheme brings energy saving gain in the total power consumption only when

$$v > \frac{2E_s}{P_o - P_{sleep}}. \quad (43)$$

Remark: The proof is similar to Proposition 4 and is omitted here. Whether it is energy efficient to incorporate the single vacation sleeping control into the power matching scheme is shown in Fig. 9. It is shown that the plane is parallel with the coordinate axis of the traffic load, which means that this has nothing to do with the traffic load situation. The reason is that, different from the multiple vacation case where the vacation terminates based on the monitoring of user requests, in the single vacation case, the termination is unrelated to the traffic.

Proposition 8: For the "single vacation sleeping control + power matching" scheme with varying service rate x,

1. $P_{v,x}^{tot}$ is a monotonically decreasing function of $D_{v,x}$, if one of the following conditions is satisfied,

1) $v > \dfrac{2E_s}{P_o - P_{sleep}}$,

$$l \geq \frac{B}{\lambda \ln 2}\Big\{\mathbf{W}\big[\frac{\gamma\lambda v(P_o - P_{sleep} - \frac{2E_s}{v})}{\Delta_P e(\lambda v + \Gamma_V(\lambda))} - \frac{1}{e}\big] + 1\Big\}.$$

2) $v \leq \dfrac{2E_s}{P_o - P_{sleep}}$.

2. For $P_{v,x}^{tot}[D_{v,x}]$, the unique energy-optimal rate x_{ev}^* exists if the following condition is satisfied,

$$v > \frac{2E_s}{P_o - P_{sleep}},$$

$$l < \frac{B}{\lambda \ln 2}\Big\{\mathbf{W}\big[\frac{\gamma\lambda v(P_o - P_{sleep} - \frac{2E_s}{v})}{\Delta_P e(\lambda v + \Gamma_V(\lambda))} - \frac{1}{e}\big] + 1\Big\}.$$

The corresponding energy-optimal rate is

$$x_{ev}^* = \frac{B}{\ln 2}\Big\{\mathbf{W}\big[\frac{\gamma\lambda v(P_o - P_{sleep} - \frac{2E_s}{v})}{\Delta_P e(\lambda v + \Gamma_V(\lambda))} - \frac{1}{e}\big] + 1\Big\}. \quad (44)$$

3. In both of the upper two cases, as the average delay increases, the total power consumption approaches an asymptotic value of $P_o + \frac{\Delta_P}{\gamma}(2^{\frac{\lambda l}{B}} - 1)$.

The proof is similar to Proposition 2 and is omitted here.

Proposition 9: For the optimal power-delay curve $P_{v,x}^{tot\,*}[D_{v,x}]$, given the tradeoff parameter β, in the optimal control pair (v^*, x^*), v^* is the unique solution of the following equation

$$\mathbf{W}\Big[\frac{\gamma}{\Delta_P e}\Big[\frac{\beta x^2}{\lambda(x-\lambda l)^2} + \frac{\lambda v(P_o - P_{sleep} - \frac{2E_s}{v})}{\lambda v + \Gamma_V(\lambda)} - \frac{\Delta_P}{\gamma}\Big]\Big] = \frac{\ln 2}{B}x - 1.$$

1. For the exponentially distributed vacation (EXP), the optimal service rate x^* is

$$x^* = \lambda l\big[1 - \frac{\beta(\lambda^3 v^4 + 2\lambda^2 v^3 + 4\lambda v^2 + 2v)}{(P_o - P_{sleep})(1+2\lambda v) + 2E_s\lambda^2 v(2+\lambda v)}\big]^{-1}. \quad (45)$$

2. For the deterministically distributed vacation (DET), the optimal service rate x^* is

$$x^* = \lambda l\big[1 - \frac{\frac{\beta}{2}(\lambda v^2 + 2v + \lambda v^2 e^{\lambda v})}{(P_o - P_{sleep})(1+\lambda v) + 2E_s\lambda(e^{\lambda v} - 1)}\big]^{-1}. \quad (46)$$

Remark: For all the three "sleeping control + power matching" schemes, explicit relationship between the optimal service rate and the sleeping parameter exists, e.g., Eq.(18) in the "N_SC + PM" scheme, Eq.(32) or Eq.(33) in the "MV_SC + PM" scheme, and Eq.(45) or Eq.(46) in the "SV_SC + PM" scheme. The optimal control pair for each of them can be found readily by simple searching algorithms such as the Newton method. For $P_{v,x}^{tot\,*}[D_{v,x}]$, as the average delay increases, the analysis result of the asymptotic limit for both the *multiple* and *single vacation based* sleeping controls is the same as that of the *N-based* scheme in Proposition 3. As a comparison, the main analysis results of different schemes are listed in Table I. The "additional delay" represents the increase in the average delay compared with the power matching only scheme. Moreover, for a better comparison, the explicit relationships between the total power consumption and average delay with varying service rate of different schemes are also provided: Eq.(38) for the "PM only" scheme, Eq.(39) for the "N_SC + PM" scheme, and Eq.(40) for the "MV_SC + PM" scheme.

$$P_{0,x}^{tot}[D_{0,x}]=\frac{\Delta_P/\gamma}{1+\frac{1}{\lambda D_{0,x}}}\left[2^{\frac{l}{B}\left(\frac{1}{D_{0,x}}+\lambda\right)}-1\right]+P_o, \tag{38}$$

$$P_{N,x}^{tot}[D_{N,x}]=\frac{\Delta_P/\gamma}{1+\frac{1}{\lambda(D_{N,x}-\frac{N-1}{2\lambda})}}\left[2^{\frac{l}{B}\left(\frac{1}{D_{N,x}-\frac{N-1}{2\lambda}}+\lambda\right)}-1+\left(P_o-P_{sleep}-\frac{2\lambda E_s}{N}\right)\frac{\gamma}{\Delta_P}\right]+P_{sleep}+\frac{2\lambda E_s}{N}, \tag{39}$$

$$P_{v,x}^{tot}[D_{v,x}]=\frac{\Delta_P/\gamma}{1+\frac{1}{\lambda(D_{v,x}-\frac{\mathbb{E}\{V^2\}}{2v})}}\left[2^{\frac{l}{B}\left(\frac{1}{D_{v,x}-\frac{\mathbb{E}\{V^2\}}{2v}}+\lambda\right)}-1+\left(P_o-P_{sleep}-\frac{2E_s(1-\Gamma_V(\lambda))}{v}\right)\frac{\gamma}{\Delta_P}\right]+P_{sleep}+\frac{2E_s(1-\Gamma_V(\lambda))}{v}. \tag{40}$$

TABLE I
COMPARISON OF DIFFERENT SCHEMES.

Scheme	N_SC+PM	MV_SC(exp)+PM	MV_SC(det)+PM	SV_SC(exp)+PM	SV_SC(det)+PM
Additional delay	$\frac{N-1}{2\lambda}$	v	$\frac{v}{2}$	$\frac{v}{1+1/\lambda v(1+\lambda v)}$	$\frac{v}{2(1+e^{-\lambda v}/\lambda v)}$
Energy-saving Region	$\frac{P_o-P_{sleep}}{2E_s}>\frac{\lambda}{N}$	$\frac{P_o-P_{sleep}}{2E_s}>\frac{1-\Gamma_V(\lambda)}{v}$		$\frac{P_o-P_{sleep}}{2E_s}>\frac{1}{v}$	
Energy-optimal Service rate	$\frac{B}{\ln 2}\left\{\mathbf{W}\left[\frac{\gamma(P_o-P_{sleep}-\frac{2\lambda E_s}{N})}{\Delta_P e}-\frac{1}{e}\right]+1\right\}$	$\frac{B}{\ln 2}\left\{\mathbf{W}\left[\frac{\gamma(P_o-P_{sleep}-\frac{2E_s(1-\Gamma_V(\lambda))}{v})}{\Delta_P e}-\frac{1}{e}\right]+1\right\}$		$\frac{B}{\ln 2}\left\{\mathbf{W}\left[\frac{\gamma(P_o-P_{sleep}-\frac{2E_s}{v})}{\Delta_P e(1+\Gamma_V(\lambda)/\lambda v)}-\frac{1}{e}\right]+1\right\}$	
Optimal tradeoff Asymptotic limit	If $\rho<\frac{B}{\ln 2}\left\{\mathbf{W}\left[\frac{\gamma(P_o-P_{sleep})}{\Delta_P e}-\frac{1}{e}\right]+1\right\}$, $P_{lb}^*=P_{sleep}+\frac{\lambda l\ln 2}{B}\frac{(P_o-P_{sleep}-\frac{\Delta_P}{\gamma})}{\mathbf{W}\left[\frac{\gamma(P_o-P_{sleep})}{\Delta_P e}-\frac{1}{e}\right]}$; else $P_{lb}^*=P_o+\frac{\Delta_P}{\gamma}(2^{\frac{\lambda l}{B}}-1)$.				

V. NUMERICAL AND SIMULATION RESULTS

In this section, first simulations are made to validate our theoretical analysis, and then comparisons between energy-delay tradeoffs of different schemes are provided. A single urban micro-cell scenario is assumed. According to the ITU test environments [29], the system bandwidth $B=10$MHz, the maximum transmit power $P_t^{max}=10$W, and the path loss model $g=36.7\lg d+33.05$ (dB), where we set $d=100$m in the simulation. The noise power density $N_0=-174$dBm/Hz, and $\eta=-1.5/\ln(5\varepsilon)=0.283$ corresponds to the BER requirement of $\varepsilon=10^{-3}$ [22]. We take the micro BS energy consumption parameters $P_o=100$W, $\Delta_P=7$, $P_{sleep}=30$W and set $E_s=25$J [24].

Users arrive according to a Poisson process, and each user requests exactly one file whose size is exponentially distributed with mean $l=2$MB from the BS and leaves the system after being served. The system operates in a time-slotted fashion, and the BS schedules users in a round robin way, serving one user in each time slot with its duration set to be 1 ms. In this paper we assume that the average user arrival rate can be well estimated [30]. At the beginning of each simulation, given the arrival rate λ and the tradeoff parameter β, the optimization is carried out to obtain the corresponding BS sleeping and transmit power settings according to which the system will be operated. For each configuration, the system was simulated for 100 million time slots.

For the energy-delay tradeoff comparisons, the following four types of schemes will be demonstrated.

1) P_t^{max} only: The BS transmits with the maximum power P_t^{max} and there is no sleeping, which is the most common setting in traditional systems;

2) SC + P_t^{max}: Only sleeping control is adopted and there is no power matching, where the BS transmits with P_t^{max};

3) PM only: The scheme discussed in Section III;

4) SC + PM: The schemes discussed in Section IV.

The first two schemes are evaluated to provide a performance baseline. To the best of our knowledge, these are the qualified benchmarks because there is no scheme from current literature focuses on the energy and user-level delay (response time) tradeoffs and jointly optimizes the power matching and sleeping control.

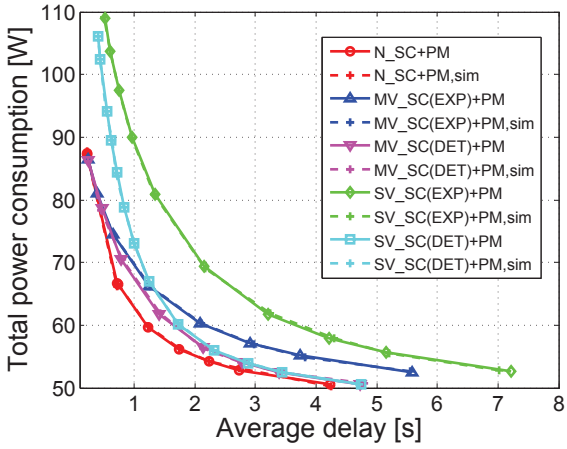
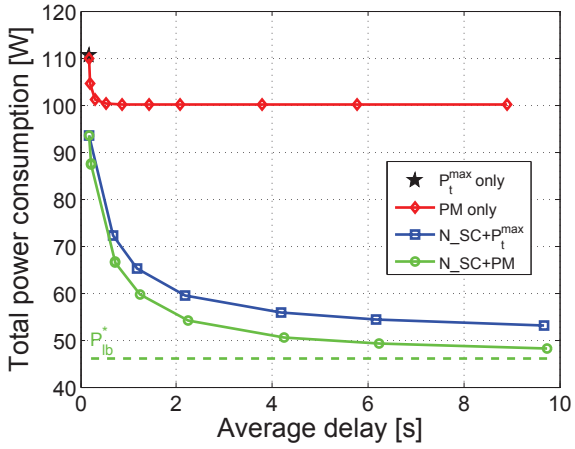


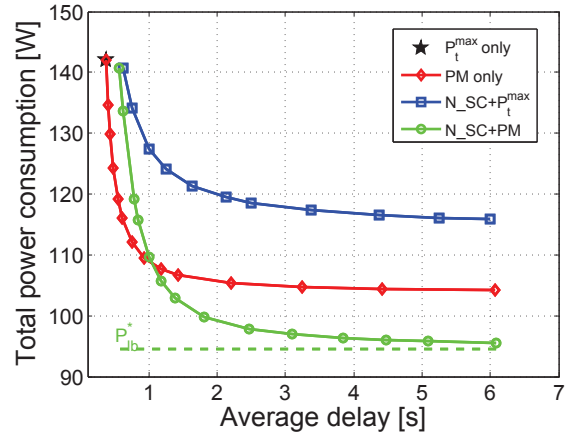
Fig. 10. Comparison of the optimal tradeoff relationships with both analytical and simulation results, $\lambda=1$, $l=2$MB.

As shown in Fig. 10, the solid curves are the results from our analysis, while the dashed curves are obtained from the simulation described above. It can be observed that under different schemes, the analytical results match the simulation results well. For clarification of scheme comparisons in the following results, we just plot the simulation results .

Fig. 10 gives the optimal energy-delay tradeoffs of different "sleeping control + power matching " schemes. Making a general comparison, we can observe that the "*N-based* sleeping control + power matching" scheme has the best energy-delay tradeoff, while the exponentially distributed "single vacation sleeping control + power matching" shows the worst performance among these five schemes. For the *vacation based* sleeping control, the deterministic vacation provides

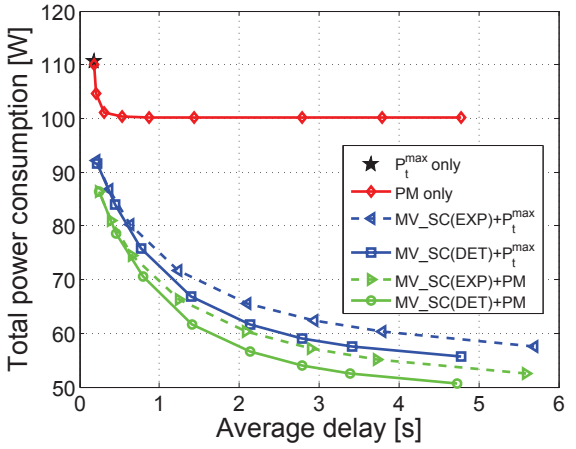


(a) $\lambda = 1$.

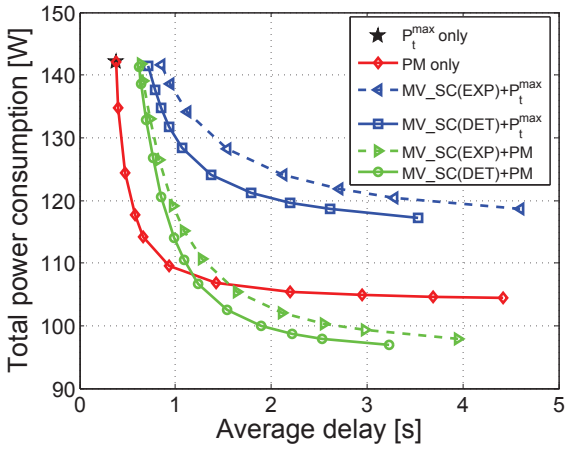


(b) $\lambda = 4$.

Fig. 11. Energy-delay tradeoff comparison: P_t^{max} only, PM only, N_SC + P_t^{max} and N_SC + PM.



(a) $\lambda = 1$.



(b) $\lambda = 4$.

Fig. 12. Energy-delay tradeoff comparison: P_t^{max} only, PM only, MV_SC + P_t^{max} and MV_SC + PM.

better energy-delay tradeoff than the exponential vacation. According to the analysis, with the same service rate, the multiple vacation case has larger average delay than the single vacation case, however, after the joint optimization of service rate and sleeping parameter, the "MV_SC + PM" scheme has better energy-delay performance than the "SV_SC + PM" scheme.

Fig. 11 provides the comparison of the four types of schemes demonstrated above, in which the *N-based* sleeping control is used in the joint optimization. First, the point of the "P_t^{max} only" scheme guarantees the minimum delay, and it has the maximum total power consumption without sleeping control. Second, compare the "PM only" and "N_SC + P_t^{max}" schemes. In *active* mode, the static power consumption $P_o = 100$W, and the load-dependent power consumption $\Delta_P P_t$ varying in $[0, 70]$W; in *sleep* mode, the static power consumption $P_{sleep} = 30$W. As shown in Fig. 11(a), when the traffic load is low, "N_SC + P_t^{max}" provides much larger energy saving gain than the "PM only" scheme; when the traffic load is increased as in Fig. 11(b), the "PM only" scheme outperforms "N_SC + P_t^{max}" in the energy saving gain due to the reduced sleeping opportunity and the corresponding increased influence of the load-dependent power consumption. Third, for the joint optimized "N_SC + PM" scheme, in Fig. 11(a), the arrival rate $\lambda = 1$, and this falls into the region that it is energy efficient to incorporate BS sleeping into the PM. It is shown that the jointly optimized scheme has the best performance and as much as 50% energy saving can be achieved; in Fig. 11(b) where the traffic load is relatively heavy, it no longer outperforms the "PM only" scheme all the time in the energy-delay tradeoff. For the asymptotic limit as delay increases with "N_SC + PM", taking $\lambda = 1$ as an example, the power consumption limit is $P_{sleep} + \frac{\lambda l \ln 2}{B}(P_o - P_{sleep} - \frac{\Delta_P}{\gamma})/\mathbf{W}\left[\frac{\gamma(P_o - P_{sleep})}{\Delta_P e} - \frac{1}{e}\right] = 46.1$W, which matches well with the simulation result.

In the comparison of the four types of schemes in Fig. 12, the multiple vacation based sleeping control is adopted, and results similar to that in Fig. 11 can be obtained.

According to the comparison results in Fig. 11 and Fig. 12, decision should be made carefully to choose one of them according to the load and delay requirements: when the traffic load is low as $\lambda = 1$, the joint "SC + PM" scheme should be used, where as much as 50% energy saving can be achieved compared with the "PM only" scheme under the same average delay requirement. When the traffic load is high

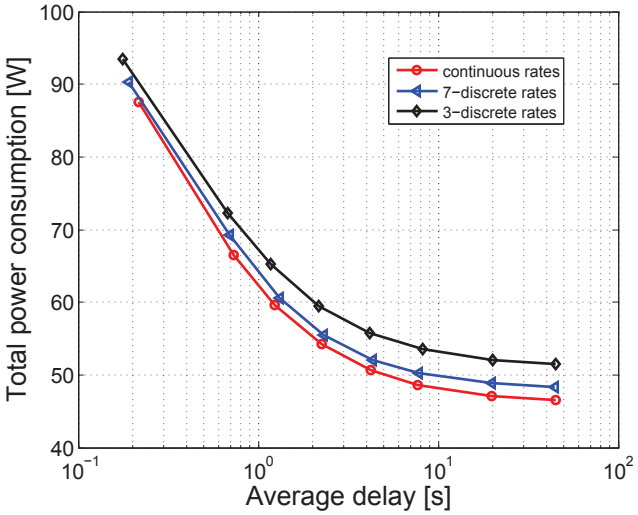
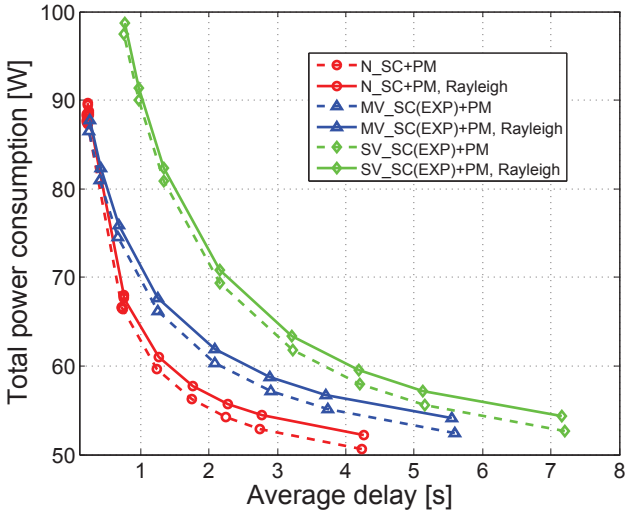
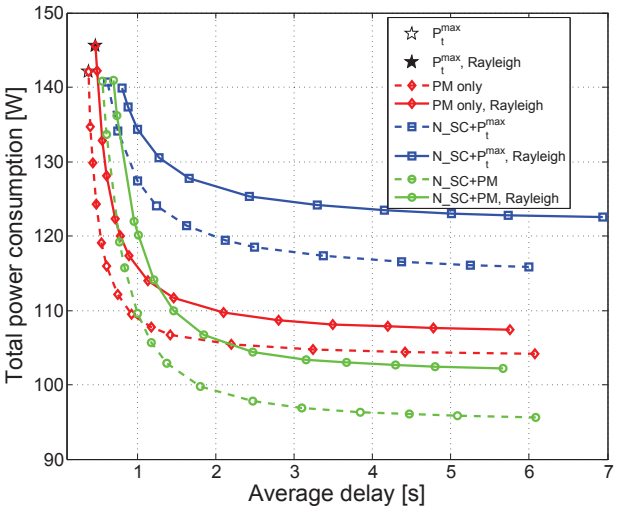


Fig. 13. The optimal energy-delay tradeoffs with "*N-based* sleeping control+power matching" with different rate sets.



(a) $\lambda = 1$.



(b) $\lambda = 4$.

Fig. 14. Energy-delay tradeoffs of different schemes with and without Rayleigh fading.

as $\lambda = 4$, with a small average delay requirement, the "PM only" scheme should be adopted; when relatively longer delay can be tolerated, the joint "SC + PM" scheme should be used to save more energy. In brief, the PM has a wider range of adaptability to the traffic variation while the SC is more energy-efficient when the traffic load is relatively low.

The impact of discrete transmission rates in practical implementation is shown in Fig. 13. Because of the convexity of the objective function, after solving the optimal rate under the continuous assumption, we can choose the one, which has a less objective function value, in the two discrete rates that are closet to the optimal value. According to the modulation and coding set [23], we observe the energy-delay tradeoffs when 3 and 7 discrete transmission rates are adopted respectively. It can be seen that performance loss exists due to the discrete rates, and the loss diminishes when more discrete rates are available.

Finally, the influence of fasting fading is shown in Fig. 14. We consider the round robin scheduling, and assume that users experience independent and identically distributed Rayleigh fading. Different from Eq.(1), the BS service capacity is provided by averaging over the fast fading variations, and then the objective function is minimized to find the optimal control pair of the sleeping parameter and transmit power. It can be seen that the tradeoff with fading is worse than that in the case without fading. Because the average received power is same in two cases, while in the fading case the service capacity user receives is averaged over a concave power-rate function, and thus it is less than the case without fading. The gap is larger in Fig. 14(b) because power/rate variations have larger influence when the traffic load is high. Note that our conclusions about different schemes still hold with the impact of fast fading under these assumptions as shown in Fig. 14(a) and Fig. 14(b). In fact, when fast fading is considered, the impact of fast fading is related to the scheduling schemes. Not only will the transmit power-service capacity relationship be different, but also the related queueing model may be varied [20] [28]. What's more, compared with the energy-saving gain brought by sleeping control and power matching, how large the gain from various scheduling schemes, like channel-aware opportunistic scheduling, will be is still an open question. However, scheduling is not the focus of this paper and this will be considered in the future work.

VI. Conclusion

In this article, we jointly design the traffic-aware BS sleeping control and power matching schemes for a single BS to achieve flexible energy-delay tradeoffs. Both *N-based* and *V-based* sleeping control schemes are analyzed. The conditions for incorporating these different types of sleeping control into the power matching energy efficiently are obtained. By exploring the relationship between total power consumption and average delay of various schemes, we find that sacrificing delay cannot always be traded for energy saving, and energy-optimal service rate exists under certain conditions which greatly depend on the energy consumption and traffic load parameters. Then the optimal energy-delay tradeoffs are provided by jointly optimizing the service rate and the sleeping

parameter. Both the optimal control pair and the asymptotic limit of total power consumption are derived, which provide the basis for practical operation and give the power consumption lower bound of the joint schemes respectively. In brief, power matching has a wider range of adaptability to the traffic variation while sleeping control is more energy-efficient when the traffic load is relatively low.

In conclusion, we mention several directions in which this work can be extended. When users of multiple classes, e.g. users with different bandwidth requirements, are considered, the egalitarian processor sharing model can be changed to a discriminatory one by introducing weights to various user classes [31]. Moreover, different from the single cell scenario, in multi-cell scenario, issues of user association or user handover among different cells become important. Because these will influence the sleeping probability of different cells and thus have impact on network level energy-delay tradeoffs. Also, the inter-cell interferences make the service of users coupled, which means that the decision of one cell about its sleep parameter and transmit power will impact decisions of other cells, so coordination among BSs is needed in the design to achieve flexible tradeoffs.

APPENDIX A
PROOF OF THE CONVEXITY OF OBJECTIVE FUNCTIONS

Making use of the fact that for $t>0$, $\theta\le 1$, $f(t)=t^2/2-t+1-\theta e^{-t}>0$, we have

$$\frac{\partial^2 z(\mathcal{T}_{sleep},x)}{\partial x^2}=\frac{2\lambda l\Delta_P}{\gamma x^3}e^{\frac{x\ln 2}{B}}\Big[\frac{1}{2}\big(\frac{x\ln 2}{B}\big)^2-\frac{x\ln 2}{B}+1-\theta e^{-\frac{x\ln 2}{B}}\Big]+\frac{2\beta l}{(x-\lambda l)^3}>0,$$

where $\theta=1$ for Eq.(6); $\theta=1-\frac{\gamma}{\Delta_P}(P_o-P_{sleep}-\frac{2\lambda E_s}{N})$ for Eq.(17); $\theta=1-\frac{\gamma}{\Delta_P}(P_o-P_{sleep}-\frac{2E_s(1-\Gamma_v(\lambda))}{v})$ for Eq.(31); $\theta=1-\frac{\gamma}{\Delta_P}(P_o-P_{sleep}-\frac{2E_s}{v})\frac{\lambda v}{\lambda v+\Gamma_v(\lambda)}$ for single vacation case. Satisfying the conditions in Proposition 1, 4 and 7 respectively, we have $\theta\le 1$. Ignoring the integer requirement of N first, we can also obtain similarly that $\frac{\partial^2 z(\mathcal{T}_{sleep},x)}{\partial \mathcal{T}_{sleep}{}^2}>0$ for all schemes except $\mathcal{T}_{sleep}=0$, which are omitted here.

APPENDIX B
PROOF OF EQ.(9) AND EQ.(11)

Combining the following global balance equations with $\sum_{j=0}^{N-1}\Pr(0,j)+\sum_{j=1}^{\infty}\Pr(1,j)=1$, the distribution in Eq.(9) will be obtained by solving them jointly.

$$\begin{cases}\lambda\Pr(0,j)=\mu\Pr(1,1), & j=0,\cdots,N-1\\ (\lambda+\mu)\Pr(1,1)=\mu\Pr(1,2),\\ (\lambda+\mu)\Pr(1,j)=\lambda\Pr(1,j-1)+\mu\Pr(1,j+1), & j\ne 0,1,N\\ (\lambda+\mu)\Pr(1,N)=\lambda[\Pr(1,N-1)+\Pr(0,N-1)]+\mu\Pr(1,N+1).\end{cases}$$

For the average queue length in Eq.(11),

$$\begin{aligned}&\sum_{j=1}^{N-1}j\Pr(0,j)+\sum_{j=1}^{\infty}j\Pr(1,j)\\ &=\sum_{j=1}^{N-1}j\frac{x-\lambda l}{Nx}+\sum_{j=1}^{N}j\frac{\lambda l}{Nx}[1-(\frac{\lambda l}{x})^j]+\sum_{j=N+1}^{\infty}j\frac{\lambda l}{Nx}[(\frac{\lambda l}{x})^{j-N}-(\frac{\lambda l}{x})^j]\\ &=\frac{x-\lambda l}{Nx}\frac{(N-1)N}{2}+\frac{\lambda l}{Nx}\Big[\frac{N(N+1)}{2}+\sum_{j=1}^{\infty}(j+N)(\frac{\lambda l}{x})^j-\sum_{j=1}^{\infty}j(\frac{\lambda l}{x})^j\Big]\\ &=\frac{N-1}{2}+\frac{\lambda l}{Nx}\Big[N+N\sum_{j=1}^{\infty}(\frac{\lambda l}{x})^j\Big]\\ &=\frac{N-1}{2}+\frac{\lambda l}{x-\lambda l}.\end{aligned}$$

APPENDIX C
PROOF OF PROPOSITION 2

The derivative of $P_{N,x}^{tot}$ over the service rate x is

$$\frac{dP_{N,x}^{tot}}{dx}=\frac{\lambda l\Delta_P e}{\gamma x^2}\Big[e^{\frac{x\ln 2}{B}-1}\big(\frac{x\ln 2}{B}-1\big)+\frac{1}{e}-\frac{\gamma}{\Delta_P e}\big(P_o-P_{sleep}-\frac{2\lambda E_s}{N}\big)\Big]. \quad (48)$$

1. If $\lambda\ge\frac{(P_o-P_{sleep})N}{2E_s}$, we have $\frac{dP_{N,x}^{tot}}{dx}>0$. Combined with $\frac{dD_{N,x}}{dx}<0$, $P_{N,x}^{tot}$ is a monotonically decreasing function of $D_{N,x}$.
2. If $\lambda<\frac{(P_o-P_{sleep})N}{2E_s}$, the following equation should be solved for the extreme points.

$$e^{\frac{x\ln 2}{B}-1}\big(\frac{x\ln 2}{B}-1\big)=\frac{\gamma(P_o-P_{sleep}-\frac{2\lambda E_s}{N})}{\Delta_P e}-\frac{1}{e}. \quad (49)$$

Since the left side of Eq.(49) is equal to $-\frac{1}{e}$ at $x=0$ and monotonically increases for $x>0$, it will eventually cross the value on the right side of this equation just once which is larger than $-\frac{1}{e}$. Let x_{en} denote the unique point that satisfies Eq.(49) with

$$x_{en}=\frac{B}{\ln 2}\Big\{\mathbf{W}\big[\frac{\gamma(P_o-P_{sleep}-\frac{2\lambda E_s}{N})}{\Delta_P e}-\frac{1}{e}\big]+1\Big\}. \quad (50)$$

$P_{N,x}^{tot}$ monotonically increases for $x>x_{en}$ and monotonically decreases for $x<x_{en}$.

To keep stability of the system, $x>\lambda l$ must be satisfied.

1) If $l\ge\frac{B}{\lambda\ln 2}\Big\{\mathbf{W}\big(\frac{\gamma(P_o-P_{sleep}-\frac{2\lambda E_s}{N})}{\Delta_P e}-\frac{1}{e}\big)+1\Big\}$, we have $x_{en}\le\lambda l$, and $\frac{dP_{N,x}^{tot}}{dx}>0$ in $(\lambda l,\infty)$, thus $P_{N,x}^{tot}$ is a monotonically decreasing function of $D_{N,x}$.

2) If $l<\frac{B}{\lambda\ln 2}\Big\{\mathbf{W}\big(\frac{\gamma(P_o-P_{sleep}-\frac{2\lambda E_s}{N})}{\Delta_P e}-\frac{1}{e}\big)+1\Big\}$, we have $x_{en}>\lambda l$. $P_{N,x}^{tot}$ monotonically decreases for $x\in(\lambda l,x_{en})$ and monotonically increases for $x>x_{en}$, thus the energy-optimal service rate $x_{en}^*=x_{en}$ exists. Therefore, $P_{N,x}^{tot}$ is a monotonically increasing function of $D_{N,x}$ when $x\in(\lambda l,x_{en})$ and is a monotonically decreasing function of $D_{N,x}$ for $x>x_{en}$.

APPENDIX D
PROOF OF PROPOSITION 3

For the power consumption lower bound of the optimal energy-delay tradeoff curve when $\rho<\frac{B}{\ln 2}\Big\{\mathbf{W}\big[\frac{\gamma(P_o-P_{sleep})}{\Delta_P e}-\frac{1}{e}\big]+1\Big\}$, assume that $t=\mathbf{W}\big[\frac{\gamma(P_o-P_{sleep})}{\Delta_P e}-\frac{1}{e}\big]$.

$$\begin{aligned}P_{lb}^*&=P_{N,x}^{tot}\Big\{N\to\infty,x=\frac{B}{\ln 2}\Big\{\mathbf{W}\big[\frac{\gamma(P_o-P_{sleep})}{\Delta_P e}-\frac{1}{e}\big]+1\Big\}\Big\}\\ &=\frac{\lambda l}{x}[P_o+\frac{\Delta_P}{\gamma}(2^{\frac{x}{B}}-1)]+(1-\frac{\lambda l}{x})P_{sleep}|_{x=\frac{B(t+1)}{\ln 2}}\end{aligned}$$

$$= P_{sleep} + \frac{\lambda l \ln 2}{B(t+1)}[P_o - P_{sleep} + \frac{\Delta_P}{\gamma}(e^{t+1} - 1)]$$
$$= P_{sleep} + \frac{\lambda l \ln 2}{B(t+1)}[P_o - P_{sleep} + \frac{\Delta_P}{\gamma}(\frac{\frac{\gamma(P_o - P_{sleep})}{\Delta_P}}{t} - 1)] \quad (51a)$$
$$= P_{sleep} + \frac{\lambda l \ln 2}{B(t+1)}[P_o - P_{sleep} - \frac{\Delta_P}{\gamma} + \frac{P_o - P_{sleep} - \frac{\Delta_P}{\gamma}}{t}]$$
$$= P_{sleep} + \frac{\lambda l \ln 2}{B(t+1)}(P_o - P_{sleep} - \frac{\Delta_P}{\gamma})\frac{t+1}{t}$$
$$= P_{sleep} + \frac{\lambda l (P_o - P_{sleep} - \frac{\Delta_P}{\gamma}) \ln 2}{B\mathbf{W}[\frac{\gamma(P_o - P_{sleep})}{\Delta_P e} - \frac{1}{e}]}. \quad (51b)$$

For the equation (51a), the property of the Lambert $\mathbf{W}$ function is adopted. Due to the definition of the Lambert $\mathbf{W}$ function, $\mathbf{W}(z)e^{\mathbf{W}(z)} = z, z \in \mathbb{C}$, we have $e^{\mathbf{W}(z)} = \frac{z}{\mathbf{W}(z)}, z \in \mathbb{C}$.

APPENDIX E
PROOF OF PROPOSITION 4

For the difference of the total power consumption under the two schemes, we have

$$P_x^{tot} - P_{v,x}^{tot} = (1 - \frac{\lambda l}{x})(P_o - P_{sleep} - \frac{2E_s(1 - \Gamma_V(\lambda))}{v}). \quad (52)$$

To obtain the energy-saving gain, the difference has to be positive, which leads to Eq.(26).

For the exponentially distributed vacation, substituting $\Gamma_V(\lambda) = \frac{1}{1+\lambda v}$ into Eq.(26), Eq.(27) can be obtained readily. Then taking the right side to be non-positive gives Eq.(29).

For the deterministically distributed vacation, $\Gamma_V(\lambda) = e^{-\lambda v}$. Substituting it into Eq.(26), after some manipulation we have

$$(\lambda v - \frac{2\lambda E_s}{P_o - P_{sleep}})e^{\lambda v - \frac{2\lambda E_s}{P_o - P_{sleep}}}$$
$$\geq -\frac{2\lambda E_s}{P_o - P_{sleep}} e^{-\frac{2\lambda E_s}{P_o - P_{sleep}}} \quad (53a)$$
$$\geq -e^{-1}. \quad (53b)$$

The inequality (53b) holds due to the fact that $x \leq e^{x-1}$, which can be transformed into $xe^{-x} \leq e^{-1}$. Then we have

$$\lambda v - \frac{2\lambda E_s}{P_o - P_{sleep}} \geq \mathbf{W}\big[-\frac{2\lambda E_s}{P_o - P_{sleep}} e^{-\frac{2\lambda E_s}{P_o - P_{sleep}}}\big], \quad (54)$$

which gives the result of Eq.(28).

$$v > \frac{2E_s}{P_o - P_{sleep}} + \frac{\mathbf{W}\big[-\frac{2\lambda E_s}{P_o - P_{sleep}} e^{-\frac{2\lambda E_s}{P_o - P_{sleep}}}\big]}{\lambda}$$
$$= \frac{2E_s}{P_o - P_{sleep}} + \frac{\mathbf{W}\big[\mathbf{W}^{-1}\big[-\frac{2\lambda E_s}{P_o - P_{sleep}}\big]\big]}{\lambda} \quad (55a)$$
$$= \frac{2E_s}{P_o - P_{sleep}} - \frac{\frac{2\lambda E_s}{P_o - P_{sleep}}}{\lambda} = 0. \quad (55b)$$

For the equality in (55a), the definition of the Lambert $\mathbf{W}$ function, $\mathbf{W}(z)e^{\mathbf{W}(z)} = z$, is adopted, and the condition that equation (55a) holds is

$$\frac{2\lambda E_s}{P_o - P_{sleep}} \leq 1, \quad (56)$$

which proves Eq.(29) for the deterministic vacation.

REFERENCES

[1] J. Wu, Y. Wu, S. Zhou, and Z. Niu, "Traffic-aware power adaptation and base station sleep control for energy-delay tradeoffs in green cellular networks," in *2012 IEEE Globecom*.

[2] G. Fettweis and E. Zimmermann, "ICT energy consumption-trends and challenges," in *Proc. 2008 International Symp. Wireless Personal Multimedia Commun.*, pp. 1–4.

[3] Z. Hasan, H. Boostanimehr, and V. K. Bhargava, "Green cellular networks: a survey, some research issues and challenges," *IEEE Commun. Surveys Tuts.*, vol. 13, pp. 524–540, 2011.

[4] M. A. Marsan, L. Chiaraviglio, D. Ciullo, and M. Meo, "Optimal energy savings in cellular access networks," *2009 IEEE ICC GreenComm Workshop*.

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

[6] S. Zhou, J. Gong, Z. Yang, Z. Niu, and P. Yang, "Green mobile access network with dynamic base station energy saving," in *2009 Mobicom*.

[7] 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, Sept. 2011.

[8] J. Kwak, K. Son, Y. Yi, and S. Chong, "Greening effect of spatio-temporal power sharing policies in cellular networks with energy constraints," *IEEE Trans. Wireless Commun.*, vol. 11, no. 12, pp. 4405–4415, Dec. 2012.

[9] M. Yadin and P. Naor, "Queueing systems with a removable service station," *Operations Research*, 1963.

[10] D. P. Heyman, "Optimal operating policies for M/G/1 queueing systems," *Operations Research.*, vol. 16, pp. 362–382, Mar. 1968.

[11] D. P. Heyman, "The T-policy for the M/G/1 Queue," *Management Science.*, vol. 23, no. 7, pp. 775–778, Mar. 1977.

[12] I. Kamitsos, L. Andrew, H. Kim, and M. Chiang, "Optimal sleep patterns for serving delay-tolerant jobs," *2010 International Conf. Energy-Efficient Comput. Netw.*

[13] G. J. Foschini and Z. Miljanic, "A simple distributed autonomous power control algorithm and its convergence," *IEEE Trans. Veh. Technol.*, vol. 42, no. 4, pp. 641–646, Nov. 1993.

[14] R. Yates, "A framework for uplink power control in cellular radio systems," *IEEE J. Sel. Areas Commun.*, vol. 13, no. 7, pp. 1341–1347, Sept. 1995.

[15] A. Wierman, L. L. H. Andrew, and A. Tang, "Power-aware speed scaling in processor sharing systems," in *Proc. 2009 IEEE INFOCOM*, pp. 2007–2015.

[16] Y. Yao, L. Huang, A. Sharma, L. Golubchik, and M. J. Neely, "Data centers power reduction: a two time scale approach for delay tolerant workloads," in *2012 IEEE INFOCOM*.

[17] Y. Chen, S. Zhang, S. Xu, and G. Y. Li, "Fundamental tradeoffs on green wireless networks," *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 30–37, June 2011.

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

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

[20] S. Borst, "User-level performance of channel-aware scheduling algorithms in wireless data networks," in *Proc. 2003 IEEE INFOCOM*, pp. 321–331.

[21] I. E. Telatar and R. G. Gallager, "Combining queueing theory with information theory for multi-access," *IEEE J. Sel. Areas Commun.*, vol. 13, pp. 963–969, Aug. 1995.

[22] X. Qiu and K. Chawla, "On the performance of adaptive modulation in cellular systems," *IEEE Trans. Commun.*, vol. 47. no. 6, pp. 884–895, June. 1999.

[23] P. Mogensen *et al.*, "LTE capacity compared to the Shannon bound," in *Proc. 2007 IEEE VTC – Spring*, pp. 1234–1238.

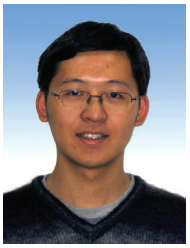
[24] G. Auer *et al.*, "D2.3: energy efficiency analysis of the reference systems, areas of improvements and target breakdown," INFSO-ICT-247733 EARTH, Tech. Rep., Nov. 2010. Available: https://www.ict-earth.eu/publications/deliverables/deliverables.html

[25] G. Auer, V. Giannini, C. Desset, I. Gódor, P. Skillermark, M. Olsson, M. A. Imran, D. Sabella, M. J. Gonzalez, O. Blume, and A. Fehske, "How much energy is needed to run a wireless network?" *IEEE Trans. Wireless Commun.*, vol. 18, pp. 40–49, Oct. 2011.

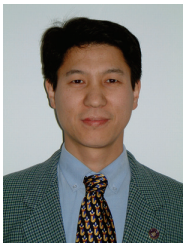
[26] R. M. Corless, G. H. Gonnet, D. E. G. Hare, D. E. Knuth, and D. J. Jeffrey, "On the Lambert W function," *Adv. Computat. Math.*, vol. 5, pp. 329–359, 1996.

[27] Y. Levy and U. Yechiali, "Utilization of idle time in an M/G/1 queueing system," *Informs*, vol. 22, no. 2, pp. 202–201, Oct. 1975.
[28] T. Bonald and A. Proutière, "Wireless downlink data channels: user performance and cell dimensioning," in *Proc. 2003 ACM Mobicom*, pp. 339–352.
[29] Ericsson, "Radio characteristics of the ITU test environments and deployment scenarios R1-091320," 3GPP TSG-RAN1#56bis, Seoul, Korea, Mar. 2009.
[30] Y. Shu, M. Yu, J. Liu, and O. W. W. Yang, "Wireless traffic modeling and prediction using seasonal ARIMA models," in *Proc. 2003 IEEE ICC*, pp. 1675–1679.
[31] G. Fayolle, I. Mitrani, and R. Iasnogorodski, "Sharing a processor among many job classes," *J. ACM*, vol. 27, pp. 519–532, 1980.
[32] L. I. Sennott, *Stochastic Dynamic Programming and the Control of Queueing Systems*. Wiley, 1999.
[33] J. Walrand, *An Introduction to Queueing Networks*. Prentice Hall, 1998.
[34] H. Takagi, *Queueing Analysis: A Foundation of Performance Evaluation, Vol. I, Vacation and Priority Systems, PartI*. North-Holland, 1991.
[35] T. L. Saaty, *Elements of queueing theory*. McGraw-Hill, 1961.
[36] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2003.

Jian Wu received her B.S. degree in Communication Engineering from Beijing Jiaotong University, Beijing, China, in 2009. She is currently a Ph.D student in Electronic Engineering Department at Tsinghua University, Beijing, China. She received the Best Paper Award from the 23th IEEE International Conference on Communication Technology (ICCT) in 2011. Her research interests include green wireless cellular communications and wireless multimedia networking.



Sheng Zhou (S06, M12) received his B.S. and Ph.D. degrees in Electronic Engineering from Tsinghua University, China, in 2005 and 2011, respectively. He is now a postdoctoral scholar in Electronic Engineering Department at Tsinghua University, Beijing, China. From January to June 2010, he was a visiting student at Wireless System Lab, Electrical Engineering Department, Stanford University, CA, USA. He is a co-recipient of the Best Paper Award from the 15th Asia-Pacific Conference on Communication (APCC) in 2009, and the 23th IEEE International Conference on Communication Technology (ICCT) in 2011. His research interests include cross-layer design for multiple antenna systems, cooperative transmission in cellular systems, and green wireless cellular communications.



Zhisheng Niu graduated from Northern Jiaotong University (currently Beijing Jiaotong University), Beijing, China, in 1985, and got his M.E. and D.E. degrees from Toyohashi University of Technology, Toyohashi, Japan, in 1989 and 1992, respectively. After spending two years at Fujitsu Laboratories Ltd., Kawasaki, Japan, he joined with Tsinghua University, Beijing, China, in 1994, where he is now a professor at the Department of Electronic Engineering, deputy dean of the School of Information Science and Technology, and director of Tsinghua-Hitachi Joint Lab on Environmental Harmonious ICT. He is also a guest chair professor of Shandong University. His major research interests include queueing theory, traffic engineering, mobile Internet, radio resource management of wireless networks, and green communication and networks.

Dr. Niu has been an active volunteer for various academic societies, including Director for Conference Publications (2010-11) and Director for Asia-Pacific Board (2008-09) of IEEE Communication Society, Membership Development Coordinator (2009-10) of IEEE Region 10, Councilor of IEICE-Japan (2009-11), and council member of Chinese Institute of Electronics (2006-11). He is now a distinguished lecturer (2012-13) of IEEE Communication Society, standing committee member of both Communication Science and Technology Committee under the Ministry of Industry and Information Technology of China and Chinese Institute of Communications (CIC), vice chair of the Information and Communication Network Committee of CIC, editor of *IEEE Wireless Communication Magazine*, and associate editor-in-chief of IEEE/CIC joint publication *China Communications*.

Dr. Niu received the Outstanding Young Researcher Award from Natural Science Foundation of China in 2009 and the Best Paper Awards (with his students) from the 13th and 15th Asia-Pacific Conference on Communication (APCC) in 2007 and 2009, respectively. He is now a fellow of both IEEE and IEICE.