Base Station Sleeping Control and Power Matching for Energy-Delay Tradeoffs with Bursty Traffic

Jian Wu, Yanan Bao, Guowang Miao, *Member, IEEE,* Sheng Zhou, *Member, IEEE,* and Zhisheng Niu, *Fellow, IEEE*

***Abstract*—In this paper, we study sleeping control and power matching for a single cell in cellular networks with bursty traffic. The base station (BS) sleeps whenever the system is empty, and wakes up when N users are assembled during the sleep period. The service capacity of the BS in the active mode is controlled through adjusting its transmit power. The total power consumption and average delay are analyzed, based on which the impact of parameter N and transmit power on the energy-delay tradeoff is studied. It is shown that given the average traffic load, the more bursty the traffic is, the less the total power is consumed, while the delay performance of more bursty traffic is better only under certain circumstances. The optimal energy-delay tradeoff is then obtained through joint sleeping control and power matching optimization. The relationship between the optimal control parameters and the asymptotic performance are also provided. Moreover, the influence of the traffic autocorrelation is explored, which shows less impact on the system performance compared with that of the burstiness. Numerical results show the energy saving gain of the joint sleeping control and power matching scheme, as well as the impact of burstiness on the optimal energy-delay tradeoff.**

***Index Terms*—Sleeping control, power matching, energy-delay tradeoff, bursty traffic**

I. Introduction

The exponential growth of mobile data traffic has triggered the vast expansion of network infrastructures, resulting in dramatically increasing network energy consumption [1]. Energy-efficient designs are urgently needed from both environmental and economic aspects. In cellular networks BSs consume nearly 60-80% of the total energy [2]. The total power consumption of a BS consists of both circuit and transmit power consumption. The circuit power is independent of the transmit power, and is consumed because of signal processing, battery backup, site cooling, etc. The transmit power is for reliable data transmission and is mainly consumed by amplifiers, feeder losses and so on [3], [4]. Therefore efforts to reduce both the circuit power and transmit power consumption should be made.

BS sleeping design has been proposed recently to realize substantial reduction of energy consumption, which puts BSs into a sleep mode when the traffic load is low [5]–[10]. Besides, transmit power adaptation to match traffic load requirement in the active mode is also an effective way to save energy. This is because even a small reduction in the transmit power enables considerable saving in overall power consumption due to its impact on the operational power consumption of amplifiers and so on [3], [11]. The delay constrained power/energy minimization problem has been widely studied, where average delay constraints are considered in [12]–[14] and hard delay constraints are studied in [15]–[17]. Here we incorporate average delay as our main design constraint. With a bound on mean delay, the objective is to minimize the average total power consumption. It has been pointed out that the energy-delay tradeoff usually deviates from the monotonic curve [12] when practical factors are considered [18], [19]. As a result, figuring out when and how to trade tolerable delay for energy savings is important for the practical system design. In this paper we will study how to make use of the BS sleeping control and power matching to achieve a good energy-delay tradeoff for energy savings.

The Poisson model has been used a lot when random traffic arrivals are considered in energy-saving design [7]–[10], [20]–[22]. However, in practice the data traffic usually has bursty features. As a result, this paper focuses on the traffic scenario with bursty arrivals, the widely used models of which include the Interrupted Poisson Process (IPP), Markov Modulated Poisson Process (MMPP) and so on. Both the IPP and MMPP have been shown to be powerful in modeling various types of multimedia traffic, while at the same time being analytically tractable [23]–[29]. For example, in 802.16 broadband wireless networks, the superposition of up to four IPPs is used to model the HTTP, TCP and FTP traffic [23], [24]. For the multimedia service over IP network offered by the 3rd Generation Partnership Project (3GPP), MMPP can provide accurate models for the voice traffic, Internet protocol traffic and video traffic respectively [25]. Besides, the superposition of two-state MMPPs can be used to model self-similar traffic [28], [29]. In this manuscript, we focus on the basic IPP and two-phase MMPP models, based on which generalization could be made further. The sleeping scheme with extra active period for LAN switches considering IPP traffic is studied in [30]. Both Poisson and IPP traffic models

J. Wu, S. Zhou and Z. Niu 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@tsinghua.edu.cn; niuzhs@tsinghua.edu.cn).

Y. Bao is with the Computer Science Department, University of California, Davis, CA 95616, USA (e-mail: ynbao@ucdavis.edu).

G. Miao is with KTH Royal Institute of Technology, Stockholm 10044, Sweden (e-mail: guowang@kth.se).

Part of this work has been published in IEEE ICC'14 WS-E2Nets [32].

This work is sponsored in part by the National Basic Research Program of China (973 Program: No. 2012CB316001), the National Science Foundation of China (NSFC) under grant No. 61201191, No. 61321061, No. 61401250, and No. 61461136004, and Hitachi Ltd and ICT-TNG in KTH.

are adopted in the system analysis against a background of optical burst switch networks, where significant differences are found [31]. In [32] we use the IPP traffic model to provide a preliminary analysis on the power consumption and delay performance, based on which substantial extension has been made in this paper.

It is well known that the variance coefficient and the autocorrelation coefficient are two major characteristics of random process [33], [34]. In this paper, first, from the viewpoint of variance coefficient, we model user arrivals using IPP to give the first-step study of the influence of traffic burstiness on the total power consumption, delay performance, as well as the sleeping control and power matching schemes. However, IPP still falls into the category of renewal process, and it cannot capture the autocorrelation feature of the traffic [34]. As a result, we further extend our analysis to the non-renewal two-phase Markov Modulated Poisson Process, which is also known as the Switched Poisson Process (SPP), and explore the influence of the autocorrelation feature.

The N-based BS sleeping control, where the BS goes to sleep when the system is empty and wakes up when N users are accumulated, and power matching schemes are considered jointly in this paper. The main contributions of this paper include:

- Decide when to incorporate the N-based sleeping control, and prove that given the average traffic load, the more bursty the traffic is, the wider the energy-efficient adaptation range of the sleeping threshold N will be.
- Illustrate the impact of the sleeping threshold, transmit power and traffic features on the total power consumption and delay performance. (i) Provide the condition under which the energy-optimal transmit power exists. (ii) Find that given the average traffic load, the more bursty the traffic is, the less the total power is consumed, while the delay performance of the more bursty traffic is better only under certain circumstances. (iii) The total power consumption does not always increase with the average traffic load, which greatly depends on the sleeping threshold N and the transmit power.
- Optimize the sleeping threshold N and transmit power jointly to minimize the total power consumption while guaranteeing the delay requirement. (i) Derive the equations of the optimal control pair. Given the transmit power, the bounds for the optimal sleeping threshold are also obtained, providing approximations of the relationship between the optimal N and the optimal transmit power. (ii) The asymptotic performance of the optimal energy-delay tradeoff is given, which shows that the power consumption lower bound relates to the average traffic load and does not vary with different burstiness. (iii) Find that the traffic region in which the joint sleeping control and power matching scheme performs better than the power matching only scheme is wider for more bursty traffic.
- Explore the impact of traffic autocorrelation feature. In the extension to the SPP traffic model, we find that compared with the variance coefficient, independent of the system utilization, the correlation feature of traffic does not have much effect on the total power consumption. Only when the system is heavily loaded, a larger autocorrelation coefficient will lead to worse delay performance.

The rest of this paper is organized as follows: In Sec. II we describe the system model. Sec. III gives the analysis of the total power consumption and delay performance. The impact of different parameters is investigated in Sec. IV. Sec. V studies the joint sleeping control and power matching optimization. In Sec. VI, we extend the analysis to the non-renewal process model, and explore the influence of the traffic autocorrelation. Numerical results are provided in Sec. VII, and Sec. VIII concludes the paper.

II. System model

A. Traffic model

We consider the downlink of a single BS where users arrive according to an IPP with parameters (λ, r_1, r_2). As shown in Fig. 1, there are on and off periods, which are both exponentially distributed with the average length r_1^{-1} and r_2^{-1} respectively. In on periods, users arrive according to a Poisson process with arrival rate λ, and there is no arrival in off periods. The average user arrival rate is $\hat{\lambda} = \frac{\lambda r_2}{r_1+r_2}$. Each user requests a random amount of best-effort data service with average length l bits, e.g., file download with average file size l, and the user leaves the system after being served.

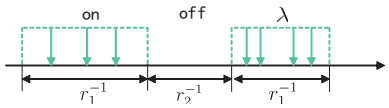


Fig. 1. Arrivals of IPP with parameters (λ, r_1, r_2).

B. BS power consumption model

We assume the BS has the *active* and *sleep* modes, with the power consumption P_{BS} as follows [3]:

$$P_{BS} = \begin{cases} P_o + \Delta_P P_t, & \textit{active} \text{ mode}, \\ P_{sleep}, & \textit{sleep} \text{ mode}. \end{cases} \quad (1)$$

P_o and P_{sleep} are the circuit power consumption in the active and sleep modes 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.

C. BS sleeping control (SC) and power matching (PM)

For the BS sleeping control, the hysteretic sleeping structure is inherited [35], [36]. We focus on the N-based BS sleeping control scheme, where the BS goes to sleep when the system is empty and returns to active mode once N users assemble in the system.

For the power matching, in the active mode, the transmit power P_t of the BS is adapted to match the traffic load. Assume that the BS service capacity is x bits per second, which is equally shared by all users being served. This can be easily achieved using a fair scheduler. The user departure

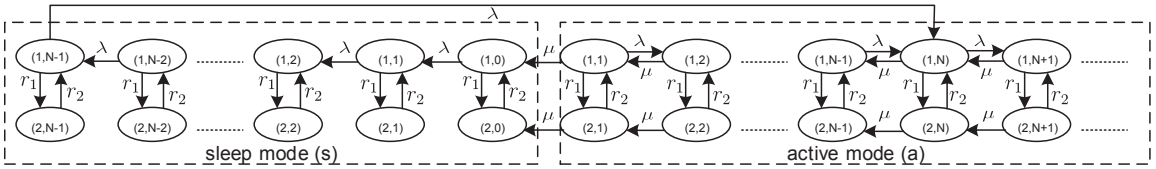


Fig. 2. State transition diagram of the extended IPP/M/1 queueing model for the N-based BS sleeping control and power matching.

rate is $\mu = x/l$. With $x = B\log_2(1+\gamma P_t)$, the relationship between the service rate[1] μ and the transmit power P_t is

$$\mu = \frac{B}{l}\log_2(1+\gamma P_t), \quad P_t \in [0, P_t^{max}] \tag{2}$$

where $\gamma = \frac{\eta g}{N_0 B}$, g represents the channel gain, B is the bandwidth, N_0 denotes the noise density, and η is a constant related to the bit error rate (BER) requirement when adaptive modulation and coding is used [37].

The control variables are the sleeping threshold N and the transmit power P_t. The delay we consider is the response time from the user arriving at the BS and generating its service request until this request is finished and the user leaves the system. In this work, homogeneous channel condition is assumed. When heterogeneous channel models are considered, the power-rate relationship and queueing model will need to be extended. For example, based on references [38]–[40], if we divide the BS coverage into different service zones in ring shapes, the relationship between the transmit power and the average service rate can be represented in a harmonic mean way. Moreover, queueing models with multi-class traffic should be used then in analyzing the user delay performance. In this case, our analyzing method can still be utilized and this will be left to the future work.

III. The IPP/M/1 queueing model with N-based sleeping control and power matching

In this section, given the sleeping threshold $N \geq 1$ and the transmit power P_t in the active mode, we analyze the total power consumption and average delay performance for IPP traffic. Since the user departure rate μ is a function of P_t, our sleeping control and power matching schemes can be modeled using an extended IPP/M/1 queueing model with N-based sleeping and adjustable service rate.

A. The extended IPP/M/1 queueing model

The state transition diagram of the queueing model is shown in Fig. 2. The total state space is divided into two sets, one for the active mode and the other for the sleep mode. In each set, the state space is defined as (i,j), where $i=1$ ($i=2$) represents the on (off) period, and j counts the number of users in the system. $p_{i,j}, (i \in \{1,2\}, j>0)$ is the probability that the BS is in the active mode with state (i,j), and $q_{i,j}, (i \in \{1,2\}, 0 \leq j < N)$ denotes the probability that the BS is in the sleep mode with state (i,j).

[1]Here the terminology "service rate" is used for both the bit service rate x and the request/user service rate μ. The service rate μ will be used in the queueing analysis.

Based on the transition graph, we have the following proposition, which is proved in Appendix A.

Proposition 1: For the extended IPP/M/1 queueing model with N-based sleeping, the probability p_s that the BS is in the sleep mode is

$$p_s = 1 - \frac{\hat{\lambda}}{\mu} = \frac{\mu(r_1+r_2)-\lambda r_2}{\mu(r_1+r_2)}, \tag{3}$$

which is independent of the sleeping threshold N.

B. The total power consumption and delay performance

To derive the average delay and total power consumption, the generation function is used [41], which is defined as $G(z) = G_1(z)+G_2(z) = (\sum_{m=0}^{N-1} z^m q_{1,m} + \sum_{m=1}^{\infty} z^m p_{1,m}) + (\sum_{m=0}^{N-1} z^m q_{2,m} + \sum_{m=1}^{\infty} z^m p_{2,m})$, $|z| \leq 1$. According to Appendix B, $G(z)$ is derived as

$$G(z) = \frac{1}{g(z)}\Big\{q_{2,0}\big[r_1 z + r_2 z + \mu z - \lambda z^2 - \mu + \lambda z\big] + q_{1,0}\big[\frac{r_1}{r_2}(r_1 z + r_2 z + \mu z - \lambda z^2 - \mu + \lambda z)\sum_{n=1}^{N-1} z^n + (r_1 z + r_2 z + \mu z - \mu)\sum_{n=0}^{N-1} z^n\big]\Big\}, \tag{4}$$

where $g(z) = -\lambda(1+\frac{r_2}{\mu})z^2 + (\lambda+\mu+r_1+r_2)z - \mu$, and

$$q_{1,0} = (1-\frac{\hat{\lambda}}{\mu})\frac{\frac{r_2}{r_1+r_2}}{N - \hat{\lambda}[(\frac{1}{r_2}+\frac{1}{\mu})z_0 - \frac{1}{r_2}]\frac{1-z_0^N}{1-z_0}}, \tag{5}$$

$$q_{2,0} = (1-\frac{\hat{\lambda}}{\mu})\frac{\frac{r_1}{r_1+r_2} - \hat{\lambda}[(\frac{1}{r_2}+\frac{1}{\mu})z_0 - \frac{1}{r_2}]\frac{1-z_0^N}{1-z_0}}{N - \hat{\lambda}[(\frac{1}{r_2}+\frac{1}{\mu})z_0 - \frac{1}{r_2}]\frac{1-z_0^N}{1-z_0}}. \tag{6}$$

z_0 is the unique root the polynomial $g(z)$ possesses in the open interval $(0,1)$, which is

$$z_0 = \frac{(\lambda+\mu+r_1+r_2) - \sqrt{(\lambda+\mu+r_1+r_2)^2 - 4\lambda(\mu+r_2)}}{2\lambda(1+r_2/\mu)}, \tag{7}$$

and its property to be used later is provided in Appendix C.

1) Total Power Consumption: The total power consumption $P_{(N,P_t)}$ is composed of three parts as follows.

$$P_{(N,P_t)} = (1-p_s)(P_o + \Delta_P P_t) + p_s P_{sleep} + E_s F_m. \tag{8}$$

The first two parts are the average power consumption in the active and sleep modes respectively, and the last term is the mode switching cost $E_s F_m$. The mode transition frequency F_m, defined as the number of mode transitions between active and sleep modes per unit time, is

$$F_m = 2\lambda q_{1,N-1} = (1-\frac{\hat{\lambda}}{\mu})\frac{2\hat{\lambda}}{N - \hat{\lambda}[(\frac{1}{r_2}+\frac{1}{\mu})z_0 - \frac{1}{r_2}]\frac{1-z_0^N}{1-z_0}}. \tag{9}$$

This is because the BS will be turned on when there is a new user arrival in state $(1, N-1)$ of the sleep mode, and each turn-on operation will correspond to one turn-off operation. Here $q_{1,N-1} = q_{1,0}$ due to Eq. (A.9) in Appendix A. Then the total power consumption is

$$P_{(N,P_t)} = (1-\frac{\hat{\lambda}}{\mu})[P_{sleep} + \frac{2\hat{\lambda}E_s}{N-\hat{\lambda}[(\frac{1}{r_2}+\frac{1}{\mu})z_0 - \frac{1}{r_2}]\frac{1-z_0^N}{1-z_0}}] + \frac{\hat{\lambda}}{\mu}(P_o + \Delta_P P_t), \quad N \geq 1. \tag{10}$$

2) The average delay: The average number of users in the system $L_{(N,P_t)} = \sum_{m=0}^{N-1} m(q_{1,m}+q_{2,m}) + \sum_{m=1}^{\infty} m(p_{1,m}+p_{2,m})$ can be derived from the generation function as follows.

$$L_{(N,P_t)} = \frac{\mathrm{d}G(z)}{\mathrm{d}z}\bigg|_{z=1} = \frac{\mathrm{d}}{\mathrm{d}z}\frac{g(z)G_1(z)+g(z)G_2(z)}{g(z)}\bigg|_{z=1}. \tag{11}$$

Substituting $g(z)$ and Eqs. (B.5)-(B.6) of Appendix B, and using the Little's Law $L_{(N,P_t)} = \hat{\lambda}D_{(N,P_t)}$, the average delay is derived as

$$D_{(N,P_t)} = \frac{1}{\mu-\hat{\lambda}} + \frac{\lambda-\mu}{(\mu-\hat{\lambda})(r_1+r_2)} + \frac{N}{N-\hat{\lambda}[(\frac{1}{r_2}+\frac{1}{\mu})z_0 - \frac{1}{r_2}]\frac{1-z_0^N}{1-z_0}}(\frac{N-1}{2\hat{\lambda}} + \frac{1}{r_1+r_2}). \tag{12}$$

C. Special case: The IPP/M/1 queueing model with power matching only

In this section we consider the special case that there is no sleeping control, and only the transmit power can be adapted to match the traffic load. This can be modeled using the IPP/M/1 queueing model with an adjustable service rate. Using the similar analysis method, the total power consumption $P_{(P_t)}$ and average delay $D_{(P_t)}$ are as follows. Note that we cannot make $N = 0$ in the previous analysis to get the performance here.

$$P_{(P_t)} = P_o + \frac{\hat{\lambda}}{\mu}\Delta_P P_t, \qquad D_{(P_t)} = \frac{1}{\mu - \lambda(1+\frac{r_2}{\mu})z_0}. \tag{13}$$

Specially, there is $D_{(1,P_t)} = D_{(P_t)}$ for the delay performance.

IV. Effects of System Parameters on the Power and Delay Performance

A. The traffic burstiness

The burstiness of IPP traffic is reflected through the variance coefficient C^2 [42], which is given by

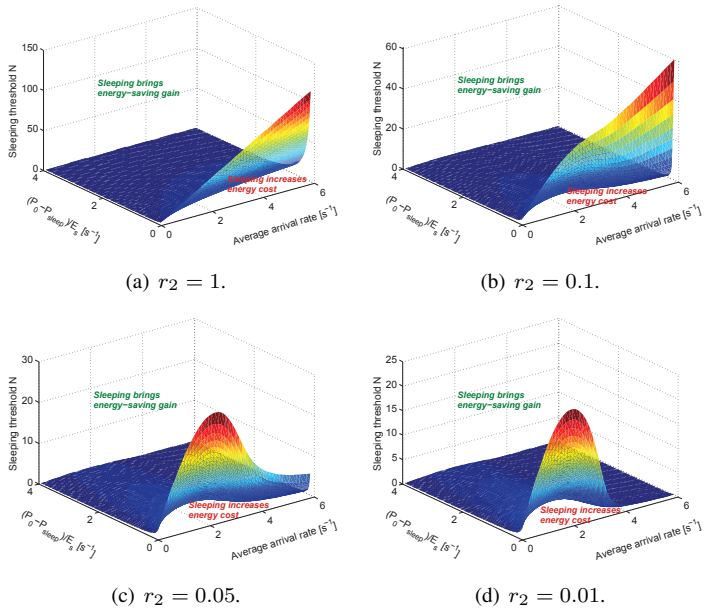
$$C^2 = 1 + \frac{2\lambda r_1}{(r_1+r_2)^2}. \tag{14}$$

With $r_1 = kr_2$, the average arrival rate of IPP traffic is

$$\hat{\lambda} = \frac{\lambda r_2}{r_1+r_2} = \frac{\lambda}{1+k}, \tag{15}$$

and the variance coefficient turns to

$$C^2 = 1 + \frac{2\lambda k}{(1+k)^2 r_2}. \tag{16}$$



(a) $r_2 = 1$. (b) $r_2 = 0.1$. (c) $r_2 = 0.05$. (d) $r_2 = 0.01$.

Fig. 3. The region that sleeping can bring energy saving gain with different r_2 (P_t = 10W, l = 2MB, $k = 1$).

The average arrival rate is independent of r_2 and only relates to λ and k. Given λ and k, the average arrival rate is fixed and the burstiness of IPP traffic is only affected by the parameter r_2, and the smaller r_2 is, the more bursty the traffic will be. In the following, given λ and k, we investigate the impact of the traffic burstiness by varying r_2, so that the average traffic arrival rate is kept the same under different burstiness.

B. Selection of sleeping threshold and transmit power

First we compare the joint sleeping control and power matching scheme with the power matching only case to find when it is energy-efficient to incorporate the BS sleeping control.

Proposition 2: For the IPP traffic with parameters (λ, kr_2, r_2), given the transmit power P_t, it is energy-efficient to incorporate the N-based sleeping control when

$$N + f(N,\mu,\lambda,k,r_2) > \frac{2\lambda E_s}{(1+k)(P_o - P_{sleep})}, \tag{17}$$

where $f(N,\mu,\lambda,k,r_2) = -\frac{\lambda}{1+k}[(\frac{1}{r_2}+\frac{1}{\mu})z_0 - \frac{1}{r_2}]\frac{1-z_0^N}{1-z_0}$ and $z_0 = \frac{(\lambda+\mu+kr_2+r_2)-\sqrt{(\lambda+\mu+kr_2+r_2)^2-4\lambda(\mu+r_2)}}{2\lambda(1+r_2/\mu)}$ with the properties:

$$(a)\ f(N,\mu,\lambda,k,r_2) > 0; \quad (b)\ \frac{\partial f(N,\mu,\lambda,k,r_2)}{\partial r_2} < 0; \quad (c)\ \frac{\partial f(N,\mu,\lambda,k,r_2)}{\partial N} > 0. \tag{18}$$

Proof: The proof is in Appendix D. ■

Remark: The properties of (18) indicate that given the average traffic arrival rate, the more bursty the traffic is, the wider the energy-efficient range of N will be. This can be proved as follows. Given the average traffic load, the burstiness of the traffic is only affected by r_2. Assuming that both (N, r_2) and (N', r_2') make (17) an equality and $r_2 < r_2'$, the objective is to prove

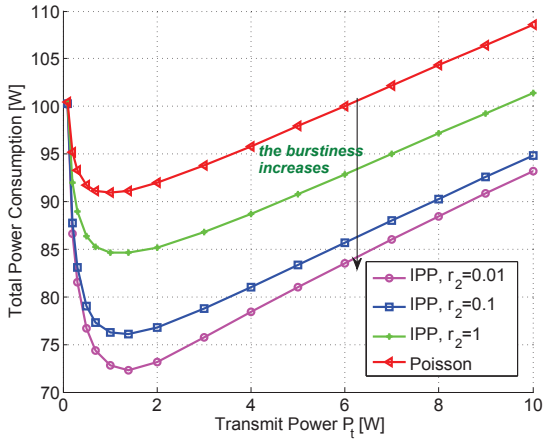


Fig. 4. The total power consumption with varying P_t. ($\lambda = 5, l =$ 2MB, $k = 1$, $N = 3$.)

$N < N'$. First we assume that $N \geq N'$, and there is $N' - N = f(N,\mu,\lambda,k,r_2) - f(N',\mu,\lambda,k,r_2') = [f(N,\mu,\lambda,k,r_2) - f(N,\mu,\lambda,k,r_2')] + [f(N,\mu,\lambda,k,r_2') - f(N',\mu,\lambda,k,r_2')] > 0$, which contradicts with the assumption, and therefore we have $N < N'$. Similarly, making use of $f(N,\mu,\lambda,k,r_2) > 0$ and $\frac{\partial f(N,\mu,\lambda,k,r_2)}{\partial N} > 0$, through comparing this condition with that for the Poisson traffic [10], the IPP traffic always has a wider adaptation range of N.

The condition is depicted in Fig. 3 where the x-axis is average traffic arrival rate and the y-axis is the parameter $\frac{P_o - P_{sleep}}{E_s}$ related to the energy consumption model. The surface is obtained through making (17) an equality. Above the surface, incorporating sleeping control brings energy saving gain. However, below the surface sleeping is harmful due to the extra mode switching energy cost. From Fig. 3(a) to Fig. 3(d), the surfaces are lowered as r_2 decreases from 1 to 0.01. In other words, as the burstiness of the traffic increases from Fig. 3(a) to Fig. 3(d), the region above the surface in which sleeping brings energy saving gain expands.

For the sleeping threshold N, from Eqs. (10) and (12), it is simple to obtain $\frac{\partial P_{(N,P_t)}}{\partial N} < 0$, $\frac{\partial D_{(N,P_t)}}{\partial N} > 0$, which means the total power consumption decreases and the average delay increases with N.

For the transmit power P_t, we have $\frac{\partial D_{(N,P_t)}}{\partial P_t} < 0$, and it is intuitive that the average delay performance gets better as the transmit power P_t increases. However, for the total power consumption and P_t, their relationship is not always monotonic, and it greatly depends on traffic and system parameters as shown in the following proposition which is proved in Appendix E.

Proposition 3: For the IPP traffic with parameters (λ, kr_2, r_2), given the sleeping threshold N,
i) $P_{(N,P_t)}$ monotonically increases with P_t when

$$l \geq \frac{B}{\hat{\lambda}\ln 2}\left\{\mathbf{W}\left[\frac{\gamma}{\Delta_P e}\left(P_o - P_{sleep} - \frac{2\hat{\lambda}E_s}{N+f(N,\hat{\lambda},\lambda,k,r_2)}\right) - \frac{1}{e}\right] + 1\right\}. \quad (19)$$

ii) $P_{(N,P_t)}$ first decreases and then increases with P_t when

$$l < \frac{B}{\hat{\lambda}\ln 2}\left\{\mathbf{W}\left[\frac{\gamma}{\Delta_P e}\left(P_o - P_{sleep} - \frac{2\hat{\lambda}E_s}{N+f(N,\hat{\lambda},\lambda,k,r_2)}\right) - \frac{1}{e}\right] + 1\right\}, \quad (20)$$

and there exists the energy-optimal transmit power P_t^{eo}, which is the unique solution of the following equation

$$\frac{\mu l \ln 2}{B} - 1 = \mathbf{W}\left[\frac{\gamma}{\Delta_P e}\left(P_o - P_{sleep} - \frac{\Delta_P}{\gamma} + 2E_s y(N,\mu,\lambda,k,r_2)\right)\right], \quad (21)$$

with the function $y(N,\mu,\lambda,k,r_2) = -\frac{\hat{\lambda}}{N+f(N,\mu,\lambda,k,r_2)} + \frac{(\mu-\hat{\lambda})f(N,\mu,\lambda,k,r_2)\left[\left(\frac{1}{1-z_0} - \frac{Nz_0^{N-1}}{1-z_0^N}\right)\frac{\mu^2 - \mu(\mu+r_2)z_0 + r_2 z_0\sqrt{(\lambda+\mu+kr_2+r_2)^2 - 4\lambda(\mu+r_2)}}{\mu+r_2} - \mu\right]}{[N+f(N,\mu,\lambda,k,r_2)]^2\sqrt{(\lambda+\mu+kr_2+r_2)^2 - 4\lambda(\mu+r_2)}}$.

Remark: Here $\mathbf{W}$ is the Lambert function, which is defined as $\mathbf{W}(z)e^{\mathbf{W}(z)} = z, z \in \mathbb{C}$ [43]. We just use its real branch $\mathbf{W}_0 : \mathcal{D}_{\mathbf{W}_0} = [-e^{-1}, +\infty) \mapsto [-1, +\infty)$ and denote it as $\mathbf{W}$ for the sake of simplicity. In Eq. (21) μ is a function of the transmit power P_t as $\mu = \frac{B}{l}\log_2(1 + \gamma P_t)$. To explain the physical meaning behind the conditions (19) and (20), we first figure out the structure of the total power consumption in Eq. (8), which can be rearranged into

$$P_{(N,P_t)} = \underbrace{P_o - p_s(P_o - P_{sleep})}_{(1)} + \underbrace{(1-p_s)\Delta_P P_t}_{(2)} + \underbrace{E_s F_m}_{(3)}.$$

The first term decreases with P_t, while the second term and the third term increase with P_t. When the average file size l is large, μ is relatively smaller under condition (19) than that under (20), and thus the sleeping probability is low under condition (19), which makes the working power consumption in the second term and the switching cost in the third term completely outweigh the static power consumption saved from sleeping, so the total power consumption monotonically increases with P_t. Otherwise, a larger μ under condition (20) leads to a higher sleeping probability than that in (19), so the static power consumption saved from sleeping plays the main role at first and the total power decreases with P_t. However, $P_{(N,P_t)}$ will go up as P_t increases further.

Fig. 4 shows the case that the total power consumption first decreases and then increases with P_t, and the energy-optimal transmit power which minimizes the total power consumption exists.

C. Impact of traffic characteristic parameters

First, given the average traffic load, we investigate the impact of burstiness on the total power consumption and delay performance.

Proposition 4: For the IPP traffic with parameters (λ, kr_2, r_2), given the transmit power P_t and sleeping threshold $N(N \geq 1)$, there is always $\frac{\partial P_{(N,P_t)}}{\partial r_2} > 0$.
Remark: The proof below is simple due to the property $\frac{\partial f(N,\mu,\lambda,k,r_2)}{\partial r_2} < 0$ in Proposition 2.

$$\frac{\partial P_{(N,P_t)}}{\partial r_2} = -\frac{\frac{2\lambda E_s}{1+k}\left(1 - \frac{\lambda}{\mu(1+k)}\right)}{[N+f(N,\mu,\lambda,k,r_2)]^2}\frac{\partial f(N,\mu,\lambda,k,r_2)}{\partial r_2} > 0. \quad (22)$$

It indicates that given the control pair (N, P_t) and the average traffic load, less total power will be consumed as the burstiness of the traffic increases, as shown in Fig. 4.

For the average delay, more bursty traffic also has better delay performance only under certain circumstances. Besides the delay comparison between IPP and Poisson traffic, Fig. 5 demonstrates the impact of r_2 on the delay of IPP traffic. There exists a transition area. On its right where N is relatively

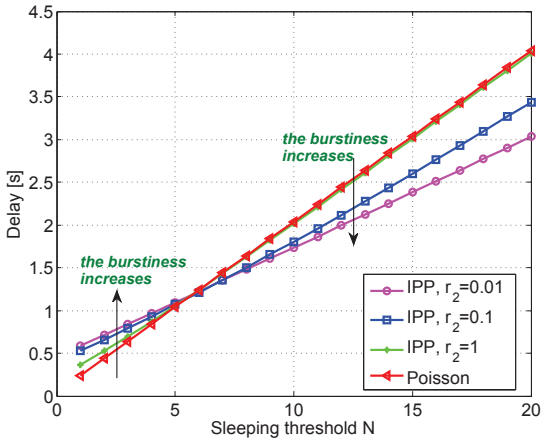
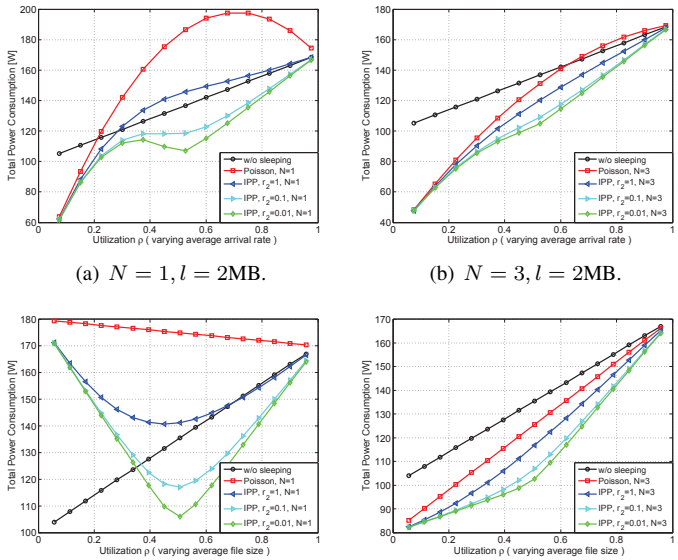


Fig. 5. The delay performance with varying N. ($\lambda = 5, l = 2$MB, $k = 1$, $P_t = 10$W.)



(a) $N = 1, l = 2$MB. (b) $N = 3, l = 2$MB.

(c) $N = 1, \lambda = 6$. (d) $N = 3, \lambda = 6$.

Fig. 6. Power consumption with varying traffic arrival rate (a, b) and file size (c, d). ($k = 1, P_t = 10$W.)

large, the delay performance is better for more bursty traffic. The situation is opposite on its left side. As the traffic load increases, the transition area will move to the right. This can be explained as follow: when the sleeping threshold N is small, which means the number of users needed to wake the BS up is low, the more bursty the traffic is, the more users will be accumulated and wait in the system while the BS is in the active mode, resulting larger average delay; on the contrary, when N is large, it is easier to reach the threshold N and wake the BS up for more bursty traffic, so the BS starts serving the users earlier, and the delay performance is better.

Besides the burstiness of the traffic, the impact of average arrival rate $\hat{\lambda}$ and file size l on the total power consumption is also explored. For the Poisson arrival, explicit conclusions are obtained, which are shown in the following proposition and proved in Appendix F. The corresponding numerical results for the IPP traffic are also provided.

Proposition 5: Given the transmit power P_t and sleeping threshold $N(N \geq 1)$, the total power consumption of Poisson traffic

i) increases with the average arrival rate if

$$N \geq \frac{2\mu E_s}{P_o - P_{sleep} + \Delta_P P_t}, \tag{23}$$

otherwise, there exists $\hat{\lambda} = \frac{\mu}{2} + \frac{N}{4E_s}(P_o - P_{sleep} + \Delta_P P_t)$ which maximizes the total power consumption;

ii) increases linearly with the average file size l if

$$N > \frac{2\hat{\lambda} E_s}{P_o - P_{sleep} + \Delta_P P_t}, \tag{24}$$

otherwise, it is a non-increasing linear function of l.

The impact on the total power consumption for both Poisson and IPP traffic are depicted in Fig. 6, where the x-axis is the system utilization $\rho = \frac{\hat{\lambda}}{\mu}$, and either the average arrival rate or the file size is varying. The results in Fig. 6(b) and Fig. 6(d) with $N = 3$ are intelligible as the total power just increases with the average arrival rate and file size. However, in Fig. 6(a) and Fig. 6(c) with $N = 1$, results turn out to be different. For the IPP traffic in Fig. 6(a), the relationship is not merely monotonic, and it fluctuates more heavily for more bursty traffic. For the impact of the file size of IPP traffic, sometimes there exists an l at which the total power is minimum as shown in Fig. 6(c). Moreover, Fig. 6 also illustrates that BS sleeping cannot always bring energy saving gain as stated in Proposition 2. In Fig. 6(a) and Fig. 6(c) with $N = 1$, frequent mode switching, especially when the average arrival rate is large or the file size is small, may consume more switching cost than the energy saved through sleeping.

V. Optimal Energy-Delay Tradeoffs

After studying the impact of different parameters, we want to find the optimal transmit power P_t and sleeping threshold N that minimize the total power consumption while guaranteeing the average delay requirement D_{th}, i.e.

$$\begin{aligned} \min_{N,P_t} \quad & P_{(N,P_t)} \\ s.t. \quad & D_{(N,P_t)} \leq D_{th}. \end{aligned} \tag{25}$$

First, the traffic region that can satisfy the delay constraint is $T = \{(\lambda, r_1, r_2, l) | D_{(1,P_t^{max})} \leq D_{th}\}$, which can be transformed into

$$T = \{(\lambda, r_1, r_2, l) | \frac{2}{\lambda + r_1 + r_2 - \mu_m + \sqrt{(\lambda + r_1 + r_2 + \mu_m)^2 - 4\lambda(\mu_m + r_2)}} + \frac{\lambda + r_1 + r_2 - \mu_m}{(r_1 + r_2)\mu_m - \lambda r_2} \leq D_{th},\ \mu_m = \frac{B}{l}\log_2(1 + \gamma P_t^{max})\}, \tag{26}$$

and we will solve the optimization problem in this region. It has been mentioned in Sec. IV-B that as N increases, the total power consumption decreases and the average delay increases. As a result, given the transmit power, the optimal N should be the one that makes the delay constraint an equality. Based on this, the following proposition is derived, which is proved in Appendix G.

Proposition 6: For the IPP traffic with parameters (λ, kr_2, r_2), given the transmit power P_t, the optimal sleeping

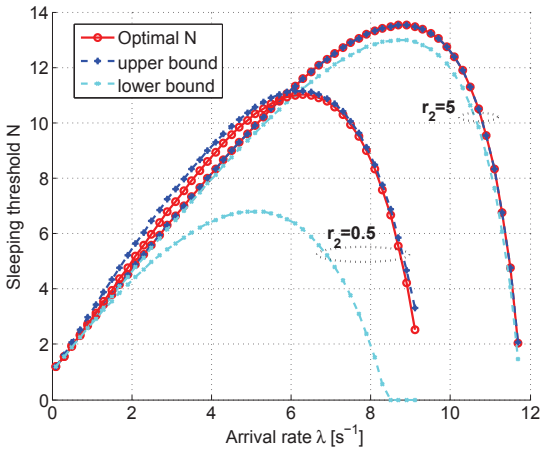


Fig. 7. The optimal sleeping threshold N and its bounds with given transmit power for IPP traffic. ($P_t = 10$W, $D_{th} = 2$s, $k = 1, l = 2$MB.)

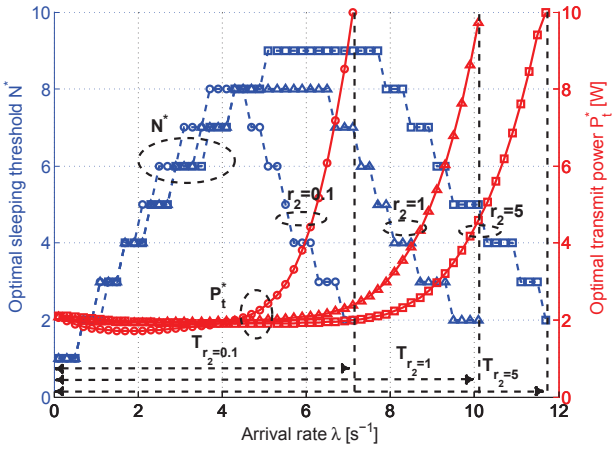


Fig. 8. The optimal sleeping threshold N^* and transmit power P_t^* for IPP traffic with varying arrival rate. ($D_{th} = 2$s, $k = 1, l = 2$MB.)

threshold of the delay constrained total power consumption minimization problem is the unique solution of

$$N(N-E) = A(1-z_0^N), \tag{27}$$

which is bounded by

$$\max\{0, E\} < N < \frac{E+\sqrt{E^2+4A}}{2}, \tag{28}$$

where the parameters $A > 0$ and E are

$$A = \frac{2\lambda^2}{(1+k)^2(1-z_0)}\left[\frac{1}{r_2} - \left(\frac{1}{r_2}+\frac{1}{\mu}\right)z_0\right]\left[D_{th} - \frac{kr_2+r_2+\lambda-\mu}{\mu r_2(1+k)-\lambda r_2}\right], \tag{29}$$

$$E = 1 + \frac{2\lambda}{1+k}\left[D_{th} - \frac{k\lambda+r_2(1+k)^2}{r_2(1+k)(k\mu+\mu-\lambda)}\right]. \tag{30}$$

Remark: Note that μ is used in these equations for simplicity, which is a function of the transmit power, $\mu = \frac{B}{l}\log_2(1+\gamma P_t)$. Moreover, the parameters A and E are functions of μ. As a result, Eq. (27) provides the relationship between the optimal control pair N and P_t of the optimization problem, and the upper and lower bounds of the sleeping threshold in (28) further offer explicit approximations of their relationship. Given transmit power, the optimal N and its bounds are given in Fig. 7. In practice N is integral, and here we just use the initial values solved from Eq. (27) for better illustration. It can be observed that given the average traffic load, the bounds are tighter for the traffic with lower burstiness, and the upper bound provides a better approximation than the lower bound.

The optimal sleeping threshold and transmit power can be obtained according to Algorithm 1. In the first step, with $\mu_m = \frac{B}{l}\log_2(1+\gamma P_t^{max})$, N_m is the largest integer satisfying the delay constraint. This is because the average delay increases with N and decreases with P_t, as stated in Sec. IV-B. As a result, for each integer $N \in [1, N_m]$, this reduces to a constrained one-dimensional optimization problem, which can be solved using the property in Proposition 3. At last, in all the N_m pairs of solutions, pick the pair (N^*, P_t^*) that minimizes the total power consumption.

Fig. 8 shows the optimal sleeping threshold and transmit power of the optimization problem with different arrival rates. First, it can be observed that the traffic region in which the delay constraint can be satisfied is different with varying traffic burstiness, and the region is wider for less bursty traffic. As the traffic arrival rate increases, the optimal sleeping threshold N^* first increases and then decreases. For the optimal transmit power, it can be seen that the power matching mainly plays its role when the traffic load is relatively high.

Algorithm 1 Solve the optimal sleeping threshold and transmit power

Input:
$\lambda,\ r_1,\ r_2,\ l,\ \mu_m,\ B,\ \gamma,\ P_o,\ P_{sleep},\ \Delta_P,\ D_{th}$;
Output:
$N^*,\ P_t^*$;
1: Solve the unique non-zero sleeping threshold N_r satisfying Eq. (27) with $\mu = \mu_m$, and set $N_m = \lfloor N_r \rfloor$;
2: **for** each integer $N \in [1, N_m]$ **do**
3: **if** Condition (19) holds **then**
4: Solve the service rate $\mu^*(N)$ satisfying Eq. (27);
5: **else**
6: Solve the service rates $\mu_1(N)$ and $\mu_2(N)$ satisfying Eq. (21) and Eq. (27) respectively, and set $\mu^*(N) = \max\{\mu_1(N), \mu_2(N)\}$;
7: **end if**
8: **end for**
9: $P_t^*(N) = \frac{1}{\gamma}(2^{\frac{\mu^*(N)l}{B}} - 1)$; $\quad (N^*, P_t^*) = \underset{N, P_t^*(N)}{\operatorname{argmin}} P_{(N, P_t^*(N))}$.

Next, we focus on the asymptotic limit of the optimal energy-delay tradeoff, which serves as the total power consumption lower bound.

Proposition 7: For the optimal energy-delay tradeoff of IPP traffic with parameters (λ, kr_2, r_2), as the delay increases, the total power consumption approaches an asymptotic value of P_{lb}, which is a function of the average arrival rate $\hat{\lambda} = \frac{\lambda}{1+k}$, and it has the following two cases:
i) if $\hat{\lambda}l \geq \frac{B}{\ln 2}\left\{\mathbf{W}\left[\frac{\gamma(P_o - P_{sleep})}{\Delta_P e} - \frac{1}{e}\right] + 1\right\}$,

$$P_{lb} = P_{(N,P_t)}\Big|_{N\to\infty, P_t \to \frac{1}{\gamma}(2^{\frac{\hat{\lambda}l}{B}}-1)} = P_o + \frac{\Delta_P}{\gamma}(2^{\frac{\hat{\lambda}l}{B}} - 1), \tag{31}$$

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

$$P_{lb} = P_{(N,P_t)}\Big|_{N\to\infty, P_t = \frac{\frac{1}{\Delta_P}(P_o - P_{sleep}) - \frac{1}{\gamma}}{\mathbf{W}\left[\frac{\gamma}{\Delta_P e}(P_o - P_{sleep}) - \frac{1}{e}\right]} - \frac{1}{\gamma}}$$

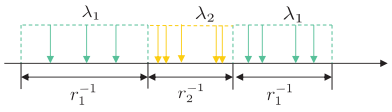


Fig. 9. Arrivals of SPP with parameters $(\lambda_1, \lambda_2, r_1, r_2)$.

$$= P_{sleep} + \frac{\hat{\lambda} l (P_o - P_{sleep} - \frac{\Delta_P}{\gamma}) \ln 2}{B \mathbf{W}[\frac{\gamma (P_o - P_{sleep})}{\Delta_P e} - \frac{1}{e}]}. \quad (32)$$

Proof: See Appendix H. ■

Remark: The two different cases of the total power consumption lower bound is partitioned based on the traffic load value $\hat{\lambda} l$ of the system. Both of the lower bounds are obtained as $N \to \infty$.

- For the first case, when the traffic load is relatively heavier, the lower bound is derived as $P_t \to \frac{1}{\gamma}(2^{\frac{\hat{\lambda} l}{B}} - 1)$, which corresponds to the system utilization $\frac{\hat{\lambda}}{\mu} \to 1$. In this case, P_{lb} has an exponential relationship with the traffic load $\hat{\lambda} l$.
- For the second case, when the traffic load is relatively lower, the lower bound is obtained at $P_t = \frac{\frac{1}{\Delta_P}(P_o - P_{sleep}) - \frac{1}{\gamma}}{\mathbf{W}[\frac{\gamma}{\Delta_P e}(P_o - P_{sleep}) - \frac{1}{e}]} - \frac{1}{\gamma}$, which is also the energy optimal transmit power $P_t^{eo}|_{N \to \infty}$ of Eq. (21). In this case, P_{lb} has a linear relationship with the traffic load $\hat{\lambda} l$.

The asymptotic limit is related to the average arrival rate $\hat{\lambda} = \frac{\lambda}{1+k}$ and the average file size l. As a result, once they are given, P_{lb} does not vary with different burstiness.

VI. Extension to Non-Renewal Process Traffic Model

In this section we extend to the non-renewal process, and investigate the impact of the autocorrelation of traffic.

A. Markov Modulated Poisson Process

We consider the two-phase Markov-Modulated Poisson Process, which is also known as the Switched Poisson Process (SPP). The traffic arrival switches between two Poisson processes with arrival rates λ_1 and λ_2, and the time it stays in each process is exponentially distributed with the average length to be ${r_1}^{-1}$ and ${r_2}^{-1}$ respectively, as shown in Fig. 9. For the SPP with parameters $(\lambda_1, \lambda_2, r_1, r_2)$, the average arrival rate $\hat{\lambda}_s$, variance coefficient C_s^2 and autocorrelation coefficient θ are provided as follows [33].

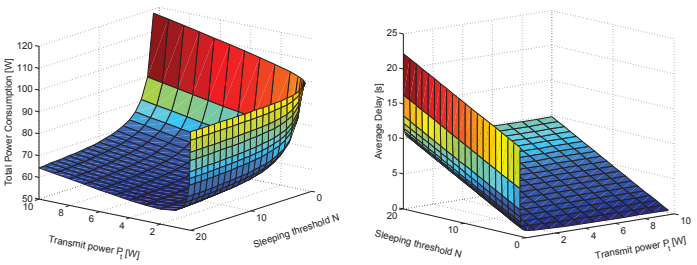
$$\hat{\lambda}_s = \frac{\lambda_1 r_2 + \lambda_2 r_1}{r_1 + r_2}, \quad (33)$$

$$C_s^2 = 1 + \frac{2 r_1 r_2 (\lambda_1 - \lambda_2)^2}{(\lambda_1 \lambda_2 + \lambda_1 r_2 + \lambda_2 r_1)(r_1 + r_2)^2}, \quad (34)$$

$$\theta = \frac{\lambda_1 \lambda_2}{\lambda_1 \lambda_2 + \lambda_1 r_2 + \lambda_2 r_1} \frac{C_s^2 - 1}{2 C_s^2}. \quad (35)$$

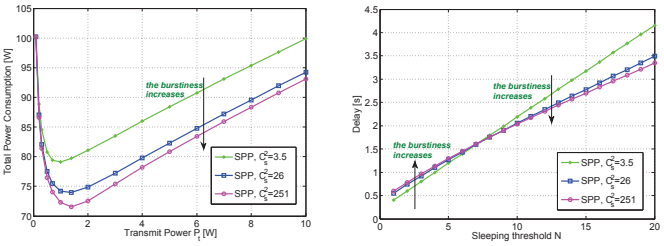
According to Fig. 10, using $q_{i,j}^s$ $(i \in \{1,2\}, 0 \le j < N)$ to denote the probability that the BS is in the sleep mode with state (i, j) for the SPP traffic, the total power consumption and average delay are

$$P_{(N,P_t)}^s = \frac{\hat{\lambda}_s}{\mu}(P_o + \Delta_P P_t) + (1 - \frac{\hat{\lambda}_s}{\mu}) P_{sleep} + 2E_s(\lambda_1 q_{1,N-1}^s + \lambda_2 q_{2,N-1}^s), \quad (36)$$



(a) Total power consumption. (b) Average delay.

Fig. 11. Total power consumption and average delay of SPP v.s. P_t and N. $(\lambda_1 = 1, \lambda_2 = 2, r_1 = r_2 = 1, l = 2\text{MB}.)$



(a) $N = 3$. (b) $P_t = 10$W.

Fig. 12. (a) Total power consumption of SPP v.s. P_t; (b) Average delay of SPP v.s. N. $(\hat{\lambda}_s = 2.5, \theta = 0.35, r_1 = r_2, l = 2\text{MB}.)$

$$D_{(N,P_t)}^s = \frac{1}{\mu - \hat{\lambda}_s} + \frac{\mu}{\hat{\lambda}_s(\mu - \hat{\lambda}_s)(r_1 + r_2)} \Big[\lambda_1 + \lambda_2 - \hat{\lambda}_s - \frac{\lambda_1 \lambda_2}{\mu} + (r_1 + r_2) \sum_{m=0}^{N-1} m(q_{1,m}^s + q_{2,m}^s) - \lambda_1 \sum_{m=0}^{N-1} q_{2,m}^s - \lambda_2 \sum_{m=0}^{N-1} q_{1,m}^s\Big]. \quad (37)$$

However, different from that in the IPP model where explicit expressions can be found, as stated in Appendix I, $q_{i,j}^s$ $(i \in \{1,2\}, 0 \le j < N)$ can only be derived by iteration. Fig. 11 shows how the total power consumption and delay vary with the transmit power P_t and the sleeping threshold N.

B. Impact of the variance and autocorrelation coefficients

In Fig. 12, the average arrival rate and variance coefficients are set the same to those in Fig. 4 and Fig. 5. When $\hat{\lambda}_s$ and θ are given, less total power is consumed with a larger C_s^2 as shown in Fig. 12(a), while the delay performance is better with a larger C_s^2 only under certain conditions in Fig. 12(b). This result is consistent with the impact of C^2 on the IPP traffic. Fig. 13(a) depicts how the total power consumption of SPP varies with the system utilization $\rho = \frac{\hat{\lambda}_s}{\mu}$ and C_s^2 given θ. It is observed when ρ is small, the total power consumption almost does not vary with C_s^2, while when ρ is relatively large, less total power is consumed as C_s^2 increases. In Fig. 13(b), C_s^2 is given, and it shows the impact of θ on the total power consumption. We can see that no matter how large the value of ρ is, the total power consumption almost keeps unchanged for different θ, which means that the correlation feature of traffic does not have much effect on the total power consumption. Fig. 13(c) shows that when the system is heavily loaded, a large θ will make the delay increase.

At last, comparisons are made for the results of Poisson [10], IPP and SPP traffic models to show the connections

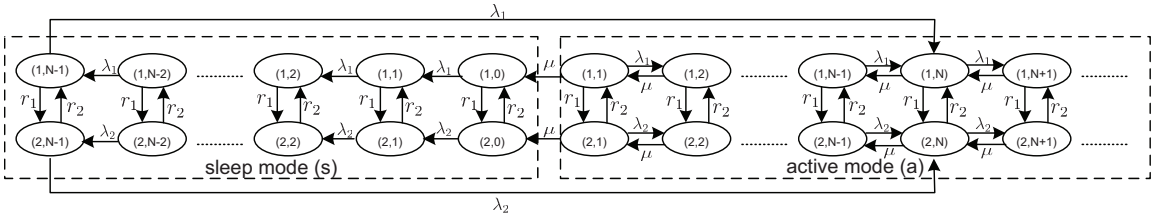
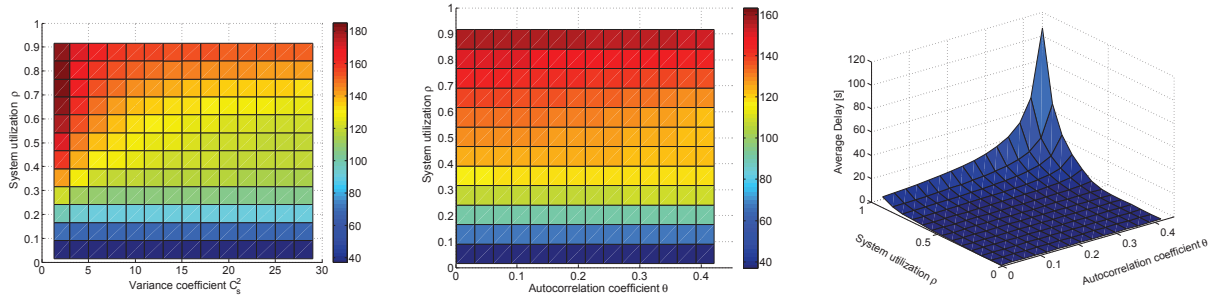


Fig. 10. State transition diagram of the extended SPP/M/1 queueing model for the N-based BS sleeping control and power matching.



(a) Total power consumption. (b) Total power consumption. (c) Average delay.

Fig. 13. (a) Total power consumption of SPP v.s. system utilization ρ and variance coefficient C_s^2, $\theta = 0.1$. (b-c) Total power consumption and delay performance of SPP v.s. system utilization ρ and autocorrelation coefficient θ, $C_s^2 = 10$. ($r_1 = r_2, l =$ 2MB, $N = 1, P_t =$ 10W.)

among them. First, table I lists some of our analytical results, making a comparison between IPP and Poisson models given the same average arrival rate $\hat{\lambda}$. First of all, the results of IPP will degenerate to those of Poisson with $r_1 = 0$ $(k = 0)$. Since this condition leads to $z_0|_{r_1=0} = \frac{\mu}{\mu+r_2}$, the function $f(N,\mu,\lambda,k,r_2) = -\frac{\lambda}{1+k}[(\frac{1}{r_2}+\frac{1}{\mu})z_0 - \frac{1}{r_2}]\frac{1-z_0^N}{1-z_0}$ given in Proposition 2 turns to zero, and the function $y(N,\mu,\lambda,k,r_2)$ given in Proposition 3 turns to $-\frac{\hat{\lambda}}{N}$.

1) With the same sleep probability $1-\frac{\hat{\lambda}}{\mu}$, the total power consumption only differs in the mode switching cost through the term $f(N,\mu,\lambda,k,r_2)$. Both the total power consumption and average delay of IPP turn into those of Poisson with $r_1 = 0$ $(k = 0)$. 2) The energy-saving region has definite physical meanings: In a sleep-active operation cycle, the energy saved from sleeping $\frac{N+f(N,\mu,\lambda,k,r_2)}{\hat{\lambda}}(P_o - P_{sleep})$ (or $\frac{N}{\hat{\lambda}}(P_o - P_{sleep})$) should be larger than the switching energy cost $2E_s$. Here $\frac{N+f(N,\mu,\lambda,k,r_2)}{\hat{\lambda}}$ (or $\frac{N}{\hat{\lambda}}$) is the average length of sleep period in a cycle, which is derived through $(1-\frac{\hat{\lambda}}{\mu})\frac{2}{F_m}$. 3) The energy-optimal service rate of IPP is given in an implicit equation. With $y(N,\mu,\lambda,k,r_2) = -\frac{\hat{\lambda}}{N}$, it is the same with that of Poisson. 4) The IPP and Poisson models share the same asymptotic performance of the optimal energy-delay tradeoff, as long as the traffic parameters $\hat{\lambda}$ and l are given.

Although explicit expressions cannot be provided for SPP, it also shares similarities with the other two models. Readers can refer Eqs. (36) and (37) for the power and delay performance. Actually, with $\lambda_1 = \lambda$ and $\lambda_2 = 0$, the results of SPP will degenerate to those of IPP in the table. Because in this case, the iterations in Eqs. (I.3) and (I.4) of SPP in Appendix I turn into Eq. (A.9) of IPP in Appendix A, which makes the probabilities in Eqs. (36) and (37) have the same explicit form as those in Eqs. (5) and (6). With $\hat{\lambda}_s = \hat{\lambda}$, besides the same sleeping probability, the implicit energy-saving region $\frac{1-\frac{\hat{\lambda}_s}{\mu}}{\lambda_1 q^s_{1,N-1}+\lambda_2 q^s_{2,N-1}}(P_o - P_{sleep}) > 2E_s$ is consistent with the physical explanation above. Moreover, SPP should share the same asymptotic limit, since the mode switching cost goes to zero as the sleeping threshold N approaches infinity.

VII. Numerical Results

In this section, we evaluate the system performance. 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. The noise power density $N_0 = -174$dBm/Hz, and $\eta = -1.5/\ln(5\varepsilon) = 0.283$ corresponds to the BER requirement of $\varepsilon = 10^{-3}$ [37]. We take the micro BS energy consumption parameters $P_o = 100$W, $\Delta_P = 7$, $P_{sleep} =$ 30W and set $E_s =$ 25J [3]. For the traffic model, $k = 1$, $r_1 = r_2$ and $l =$ 1MB or 2MB.

A. The optimal tradeoff performance

For the IPP traffic, in Fig. 14, the solid line gives the optimal energy-delay tradeoff obtained through the joint optimization. For each point, the x-axis is the delay requirement and the y-axis corresponds to the minimum total power consumption satisfying this requirement. For a better comparison, the dashed lines representing the energy-delay relationship before the joint optimization are also provided. For each of them, N is fixed and the transmit power is varying. 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. The optimal energy-delay tradeoff of SPP traffic is also given in Fig. 14. With $\hat{\lambda} = \hat{\lambda}_s = 1.5$, the difference between the optimal tradeoff curves of IPP and

TABLE I
COMPARISON UNDER DIFFERENT TRAFFIC MODELS.

Traffic model	Interrupted Poisson Process (λ, kr_2, r_2), $\hat{\lambda}=\frac{\lambda}{1+k}$	Poisson Process $(\hat{\lambda})$ [10]
Mode Transition Frequency F_m	$(1-\frac{\hat{\lambda}}{\mu})\frac{2\hat{\lambda}}{N+f(N,\mu,\lambda,k,r_2)}$	$(1-\frac{\hat{\lambda}}{\mu})\frac{2\hat{\lambda}}{N}$
Total Power Consumption	$\frac{\hat{\lambda}}{\mu}(P_o+\Delta_P P_t)+(1-\frac{\hat{\lambda}}{\mu})(P_{sleep}+\frac{2\hat{\lambda}E_s}{N+f(N,\mu,\lambda,k,r_2)})$	$\frac{\hat{\lambda}}{\mu}(P_o+\Delta_P P_t)+(1-\frac{\hat{\lambda}}{\mu})(P_{sleep}+\frac{2\hat{\lambda}E_s}{N})$
Average Delay	$\frac{1}{\mu-\hat{\lambda}}+\frac{\lambda-\mu}{(\mu-\hat{\lambda})(r_1+r_2)}+\frac{N}{N+f(N,\mu,\lambda,k,r_2)}(\frac{N-1}{2\hat{\lambda}}+\frac{1}{r_1+r_2})$	$\frac{1}{\mu-\hat{\lambda}}+\frac{N-1}{2\hat{\lambda}}$
Energy-saving Region	$\frac{N+f(N,\mu,\lambda,k,r_2)}{\hat{\lambda}}(P_o-P_{sleep})>2E_s$	$\frac{N}{\hat{\lambda}}(P_o-P_{sleep})>2E_s$
Energy-optimal Service rate μ	$\frac{\mu l\ln 2}{B}-1=\mathbf{W}[\frac{\gamma}{\Delta_P e}(P_o-P_{sleep}-\frac{\Delta_P}{\gamma}+2E_s y(N,\mu,\lambda,k,r_2))]$	$\frac{B}{l\ln 2}\{\mathbf{W}[\frac{\gamma}{\Delta_P e}(P_o-P_{sleep}-\frac{\Delta_P}{\gamma}-\frac{2E_s\hat{\lambda}}{N})]+1\}$
Optimal tradeoff Asymptotic limit	If $\hat{\lambda}l<\frac{B}{\ln 2}\{\mathbf{W}[\frac{\gamma(P_o-P_{sleep})}{\Delta_P e}-\frac{1}{e}]+1\}$, $P_{lb}=P_{sleep}+\frac{\hat{\lambda}l\ln 2}{B}\frac{P_o-P_{sleep}-\frac{\Delta_P}{\gamma}}{\mathbf{W}[\frac{\gamma(P_o-P_{sleep})}{\Delta_P e}-\frac{1}{e}]}$; else $P_{lb}=P_o+\frac{\Delta_P}{\gamma}(2^{\frac{\hat{\lambda}l}{B}}-1)$.	

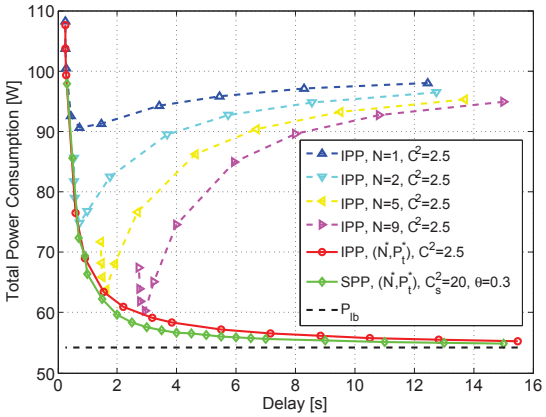


Fig. 14. The energy-delay relationships with different N for IPP traffic, and the optimal energy-delay tradeoffs for both IPP and SPP traffic. ($\hat{\lambda}=\hat{\lambda}_s=1.5, k=1, r_1=r_2, l=$ 2MB.)

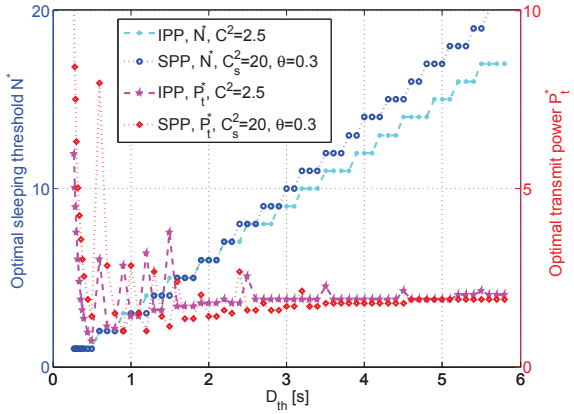


Fig. 15. The optimal control variables of IPP and SPP traffic. ($\hat{\lambda}=\hat{\lambda}_s=1.5, k=1, r_1=r_2, l=$ 2MB.)

SPP is not large, even though the SPP traffic has a much larger variance coefficient. Moreover, the total power consumptions of both IPP and SPP approach the asymptotic value obtained in Proposition 7.

Corresponding to the optimal energy-delay tradeoffs in Fig. 14, the optimal control variables: the sleeping threshold and the transmit power, are depicted in Fig. 15. As the delay requirement gets loose, the optimal sleeping threshold increases. The optimal transmit power decreases under the same sleeping threshold as D_{th} increases, and there exists oscillations for different N^*.

B. Impact of the traffic burstiness

Since the impact of the traffic autocorrelation is limited, and the impact of the variance coefficient on the system performance of SPP is similar to that of IPP, in the following, we will focus on the IPP to explore the influence of traffic burstiness.

Given the delay requirement, the minimum total power consumptions with different traffic arrival rates are depicted in Fig.16. For the case with power matching only, according to the analysis in Sec. III-C, only the transmit power is optimized in the delay constrained total power minimization problem. It elucidates that when the traffic load is low, the joint optimization scheme always consumes less power compared with the power matching only scheme. On the other hand, with high traffic load the power matching only scheme is more energy-efficient. The reason is that when the traffic load is high, the mode transition energy cost may exceed the energy saved from the less sleeping opportunity. Moreover, observing the traffic region in which the joint optimization scheme is better in Fig. 16, we can see that this region is wider for IPP traffic with a larger variance coefficient, which means that the joint sleeping control and power matching scheme has a wider adaptability to more bursty traffic.

In Fig. 17, the optimal energy-delay tradeoffs for the IPP traffic with different burstiness are demonstrated. In Fig. 17(a) with low traffic load, the more bursty the traffic is, the better the energy-delay tradeoff is. Nevertheless in Fig. 17(b) with relatively heavy traffic load, the tradeoff performance is worse for more bursty traffic. This indicates the impact of burstiness on the tradeoff greatly depends on the traffic load. Note that the minimum average delay that can be achieved is bounded by the maximum transmit power.

VIII. CONCLUSION

In this paper, we have analyzed the N-based BS sleeping control and power matching schemes for both the IPP and SPP traffic models. Theoretical analyses are provided on the impact of the sleeping threshold, transmit power and traffic features on the total power consumption and delay performance. Given

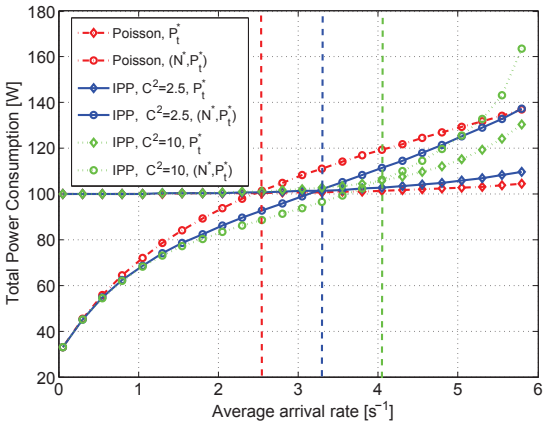


Fig. 16. The total power consumption comparison between the joint control with (N^*, P_t^*) and power matching only with P_t^*. ($l = 1$MB $D_{th} = 0.3$s and $k = 1$.)

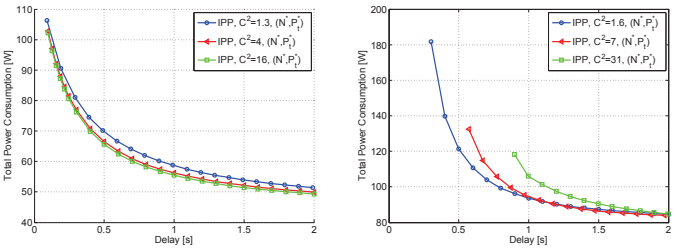
the average traffic load, more energy can be saved with larger traffic burstiness. Besides, the influence of the autocorrelation coefficient on the system performance is relatively weak, compared with that of the variance coefficient. The optimal energy-delay tradeoff is also obtained by solving a delay constrained total power minimization problem, where the relationship between the optimal control parameters is provided. Moreover, the asymptotic limit of the optimal tradeoff is explored, which gives a guideline for the best energy saving gain we can approach.

In conclusion, we mention directions in which this work can be extended. Besides the N-based sleeping control, the analysis could be extended to other sleep patterns under bursty traffic model, e.g., single/multiple-vacation based sleeping control. Moreover, when the multi-cell scenario is considered, it could still be simplified to single-cell model by incorporating the transferred traffic from/to adjacent cells if static inter-cell interference is assumed. Otherwise, the dynamic interference relating to the transmit power and sleeping threshold will make the service of users among different cells coupled. In this case, a more complicated multi-server coupled queueing model is needed.

Appendix A
Proof of Proposition 1

The global balance equations are given as follows:

$$(\lambda+r_1)q_{1,m}=\lambda q_{1,m-1}+r_2q_{2,m},\ (1\le m\le N-1) \quad (A.1)$$
$$r_2q_{2,m}=r_1q_{1,m},\ (1\le m\le N-1) \quad (A.2)$$
$$(\lambda+r_1)q_{1,0}=\mu p_{1,1}+r_2q_{2,0}, \quad (A.3)$$
$$r_2q_{2,0}=\mu p_{2,1}+r_1q_{1,0}, \quad (A.4)$$
$$(\lambda+\mu+r_1)p_{1,1}=\mu p_{1,2}+r_2p_{2,1}, \quad (A.5)$$
$$(\lambda+\mu+r_1)p_{1,m}=\mu p_{1,m+1}+\lambda p_{1,m-1}+r_2p_{2,m},\ (m\ge 2, m\ne N) \quad (A.6)$$
$$(\lambda+\mu+r_1)p_{1,N}=\mu p_{1,N+1}+\lambda(p_{1,N-1}+q_{1,N-1})+r_2p_{2,N}, \quad (A.7)$$
$$(\mu+r_2)p_{2,m}=\mu p_{2,m+1}+r_1p_{1,m},\ (m\ge 1). \quad (A.8)$$



(a) $\lambda = 3, l = 1$MB. (b) $\lambda = 6, l = 2$MB.

Fig. 17. The optimal energy-delay tradeoffs with different burstiness of IPP traffic. ($k = 1$.)

After some algebraic operations, we obtain the following local balance equations.

$$q_{1,m}=q_{1,m-1}=\tfrac{r_2}{r_1}q_{2,m},\ (1\le m\le N-1) \quad (A.9)$$
$$\lambda q_{1,0}=\mu(p_{1,1}+p_{2,1}), \quad (A.10)$$
$$\lambda(q_{1,m}+p_{1,m})=\mu(p_{1,m+1}+p_{2,m+1}),\ (1\le m\le N-1) \quad (A.11)$$
$$\lambda p_{1,m}=\mu(p_{1,m+1}+p_{2,m+1}),\ (m\ge N). \quad (A.12)$$

Summing Eqs. (A.10)-(A.12) over all m and plugging in Eq. (A.9), we obtain

$$\lambda p_1=\mu[(p_1-Nq_{1,0})+(p_2-q_{2,0}-\frac{r_1}{r_2}(N-1)q_{1,0})], \quad (A.13)$$

where $p_1=\frac{r_2}{r_1+r_2}$ and $p_2=\frac{r_1}{r_1+r_2}$ are the probabilities that the system is in on and off periods respectively. Then the sleeping probability p_s is

$$p_s=\textstyle\sum_{m=0}^{N-1}q_{1,m}+\sum_{m=0}^{N-1}q_{2,m}=Nq_{1,0}+q_{2,0}+\frac{r_1}{r_2}(N-1)q_{1,0}$$
$$=p_1(1-\tfrac{\lambda}{\mu})+p_2=1-\tfrac{\lambda r_2}{\mu(r_1+r_2)}. \quad (A.14)$$

Appendix B
The Generation Function

We rewrite Eqs. (A.1) and (A.3) as follows.

$$(\lambda+\mu+r_1)q_{1,m}=\lambda q_{1,m-1}+r_2q_{2,m}+\mu q_{1,m}, \quad (B.1)$$
$$(\lambda+\mu+r_1)q_{1,0}=\mu p_{1,1}+r_2q_{2,0}+\mu q_{1,0}. \quad (B.2)$$

For Eqs. (B.1)-(B.2) and Eqs. (A.5)-(A.7), we multiply each of them by $z^m(m=0,1,\cdots)$ appropriately and sum over all m. This process results in

$$(\lambda+\mu+r_1)G_1(z)=r_2G_2(z)+\lambda zG_1(z)+\mu\sum_{n=0}^{N-1}z^nq_{1,0}$$
$$+\frac{\mu}{z}[G_1(z)-\sum_{n=0}^{N-1}z^nq_{1,0}]. \quad (B.3)$$

Similarly, using Eqs. (A.2)(A.4)(A.8), we have

$$(\mu+r_2)G_2(z)=r_1G_1(z)+\mu[q_{2,0}+\frac{r_1}{r_2}\sum_{n=1}^{N-1}z^nq_{1,0}]$$
$$+\frac{\mu}{z}[G_2(z)-q_{2,0}-\frac{r_1}{r_2}\sum_{n=1}^{N-1}z^nq_{1,0}]. \quad (B.4)$$

With a polynomial defined as $g(z) = -\lambda(1+\frac{r_2}{\mu})z^2 + (\lambda+\mu+r_1+r_2)z - \mu$, utilizing Eqs. (B.3)-(B.4), we have the following equations

$$g(z)G_1(z) = q_{1,0}[r_1\sum_{n=1}^{N-1} z^{n+1} + (r_2 z + \mu z - \mu)\sum_{n=0}^{N-1} z^n] + r_2 q_{2,0} z, \tag{B.5}$$

$$g(z)G_2(z) = (r_1 z + \mu z - \lambda z^2 - \mu + \lambda z)(q_{2,0} + q_{1,0}\frac{r_1}{r_2}\sum_{n=1}^{N-1} z^n) + r_1 q_{1,0}\sum_{n=0}^{N-1} z^{n+1}. \tag{B.6}$$

Based on these, the generation function in Eq. (4) is derived. Since $g(z_0)G(z_0) = 0$ in Eq. (4), we obtain the following equation of $q_{1,0}$ and $q_{2,0}$.

$$q_{2,0}r_2 + q_{1,0}[\frac{z_0 - z_0^N}{1-z_0}r_1 + \frac{1-z_0^N}{1-z_0}(\mu + r_2 - \frac{\mu}{z_0})] = 0. \tag{B.7}$$

Combining Eq. (B.7) and Eq. (A.13), $q_{1,0}$ and $q_{2,0}$ are obtained.

Appendix C
Proof of the bounds of z_0

With $\Delta = \sqrt{(\lambda+\mu+r_1+r_2)^2 - 4\lambda(\mu+r_2)}$, z_0 has the following property:

$$\frac{\mu}{\mu+r_2+\Delta} < z_0 < \frac{\mu(r_2+\Delta)}{(\mu+r_2)(r_1+r_2+\Delta)}. \tag{C.1}$$

In order to prove the left side, substituting Eq. (7) into it, it equals the following inequality.

$$\sqrt{(\lambda+\mu+r_1+r_2)^2 - 4\lambda\mu(1+r_2/\mu)} > r_1 + \lambda + \frac{(r_2+\mu)(r_1-\lambda)}{r_1+\lambda}. \tag{C.2}$$

This inequality holds if the right side is negative, otherwise we have

$$\begin{aligned}\Delta &= \sqrt{(r_1+\lambda)^2 + 2(r_2+\mu)(r_1-\lambda) + (r_2+\mu)^2} \\ &> \sqrt{(r_1+\lambda)^2 + 2(r_2+\mu)(r_1-\lambda) + \frac{(r_2+\mu)^2(r_1-\lambda)^2}{(r_1+\lambda)^2}} \\ &= r_1 + \lambda + \frac{(r_2+\mu)(r_1-\lambda)}{r_1+\lambda}.\end{aligned}$$

Substituting $\Delta = \lambda+\mu+r_1+r_2-2\lambda z_0(1+\frac{r_2}{\mu})$ into the right side of Eq. (C.1), it equals the following inequality after some manipulation.

$$z_0(1+\frac{r_2}{\mu})(\lambda-\mu) < \lambda - \mu + r_1. \tag{C.3}$$

First, based on $\Delta = \sqrt{(\mu+r_1+r_2-\lambda)^2 + 4\lambda r_1} > \mu + r_1 + r_2 - \lambda$, it is easy to prove that $z_0 < \frac{\mu}{\mu+r_2}$.

1) If $\lambda \geq \mu > \frac{\lambda r_2}{r_1+r_2}$, with $0 < z_0 < \frac{\mu}{\mu+r_2}$ we have $z_0(1+\frac{r_2}{\mu})(\lambda-\mu) \leq \lambda - \mu < \lambda - \mu + r_1$.

2) If $\lambda < \mu$, with $z_0(1+\frac{r_2}{\mu}) = \frac{\lambda+\mu+r_1+r_2-\Delta}{2\lambda}$ it equals $\Delta < r_1 + r_2 + \mu - \lambda + \frac{2\lambda r_1}{\mu-\lambda}$, which can be proved as $\Delta = \sqrt{(r_1+r_2+\mu-\lambda)^2+4\lambda r_1} < \sqrt{(r_1+r_2+\mu-\lambda)^2+4\lambda r_1 + \frac{4\lambda r_1(r_1+r_2)}{\mu-\lambda} + \frac{4\lambda^2 r_1^2}{(\mu-\lambda)^2}} = \sqrt{(r_1+r_2+\mu-\lambda+\frac{2\lambda r_1}{\mu-\lambda})^2}$.

Appendix D
Proof of Proposition 2

Using Eq. (10) and Eq. (13), it is energy-efficient to incorporate the sleeping control when

$$P_{(N,P_t)} - P_{(P_t)} = (1-\frac{\hat{\lambda}}{\mu})[P_{sleep} - P_o + \frac{2\hat{\lambda}E_s}{N - \hat{\lambda}[(\frac{1}{r_2}+\frac{1}{\mu})z_0 - \frac{1}{r_2}]\frac{1-z_0^N}{1-z_0}}] < 0.$$

With $r_1 = kr_2$ and $f(N,\mu,\lambda,k,r_2) = -\frac{\lambda}{1+k}[(\frac{1}{r_2}+\frac{1}{\mu})z_0 - \frac{1}{r_2}]\frac{1-z_0^N}{1-z_0}$, it can be transformed into

$$N + f(N,\mu,\lambda,k,r_2) > \frac{2\lambda E_s}{(1+k)(P_o - P_{sleep})}. \tag{D.1}$$

Divide the $f(N,\mu,\lambda,k,r_2)$ function into $f_1(N,\mu,\lambda,k,r_2) = -\frac{\lambda}{1+k}[(\frac{1}{r_2}+\frac{1}{\mu})z_0 - \frac{1}{r_2}]$ and $f_2(N,\mu,\lambda,k,r_2) = \frac{1-z_0^N}{1-z_0}$. Making use of Eq. (7) and the bounds of z_0 in Appendix C, we have

$$\begin{aligned}\frac{\partial z_0}{\partial r_2} &= \frac{1}{2\lambda(1+\frac{r_2}{\mu})}[1 + k - \frac{(kr_2+r_2+\mu+\lambda)(k+1) - 2\lambda}{\sqrt{(\lambda+\mu+kr_2+r_2)^2 - 4\lambda(\mu+r_2)}} - \frac{2\lambda}{\mu}z_0] \\ &= \frac{\mu - [(1+k)(\mu+r_2) + \sqrt{(\lambda+\mu+kr_2+r_2)^2 - 4\lambda(\mu+r_2)}]z_0}{(\mu+r_2)\sqrt{(\lambda+\mu+kr_2+r_2)^2 - 4\lambda(\mu+r_2)}} \\ &< \frac{\mu - [(\mu+r_2) + \sqrt{(\lambda+\mu+kr_2+r_2)^2 - 4\lambda(\mu+r_2)}]z_0}{(\mu+r_2)\sqrt{(\lambda+\mu+kr_2+r_2)^2 - 4\lambda(\mu+r_2)}} < 0,\end{aligned} \tag{D.2}$$

$$\begin{aligned}\frac{\partial f_1(N,\mu,\lambda,k,r_2)}{\partial r_2} = \frac{\lambda(\mu+r_2)}{r_2^2\mu(1+k)}\Big\{ & z_0[1 + \frac{r_2(1+k)}{\sqrt{(\lambda+\mu+kr_2+r_2)^2 - 4\lambda(\mu+r_2)}}] \\ & - \frac{\mu}{\mu+r_2}[1 + \frac{r_2}{\sqrt{(\lambda+\mu+kr_2+r_2)^2 - 4\lambda(\mu+r_2)}}]\Big\} < 0.\end{aligned} \tag{D.3}$$

With $z_0 < \frac{\mu}{\mu+r_2}$ in Appendix C, we have $f_1(N,\mu,\lambda,k,r_2) > -[(\frac{\hat{\lambda}}{r_2}+\frac{\hat{\lambda}}{\mu})\frac{\mu}{\mu+r_2} - \frac{\hat{\lambda}}{r_2}] = 0$.

Due to the fact that $f_2(N,\mu,\lambda,k,r_2) > 0$ is a non-decreasing function of z_0 and $\frac{\partial z_0}{\partial r_2} < 0$, we have $\frac{\partial f_2(N,\mu,\lambda,k,r_2)}{\partial r_2} \leq 0$. Combining with $f_1(N,\mu,\lambda,k,r_2) > 0$ and $\frac{\partial f_1(N,\mu,\lambda,k,r_2)}{\partial r_2} < 0$, there is $f(N,\mu,\lambda,k,r_2) > 0$, and its a decreasing function of r_2.

Appendix E
Proof of Proposition 3

With Eq. (10) and $P_t = \frac{1}{\gamma}(2^{\frac{\mu l}{B}} - 1)$, there is

$$\begin{aligned}\frac{\partial P_{(N,P_t)}}{\partial \mu} = \frac{\hat{\lambda}e\Delta_P}{\gamma\mu^2}\Big\{ & e^{\frac{\mu l \ln 2}{B} - 1}(\frac{\mu l \ln 2}{B} - 1) - \frac{\gamma}{\Delta_P e}(P_o - P_{sleep} - \frac{\Delta_P}{\gamma} \\ & + \frac{2E_s[\mu(\mu-\hat{\lambda})\frac{\partial f(N,\mu,\lambda,k,r_2)}{\partial \mu} - \hat{\lambda}(N + f(N,\mu,\lambda,k,r_2))]}{[N + f(N,\mu,\lambda,k,r_2)]^2})\Big\}\end{aligned} \tag{E.1}$$

in which

$$\begin{aligned}\frac{\partial f(N,\mu,\lambda,k,r_2)}{\partial \mu} = \frac{\hat{\lambda}(1-z_0^N)}{1-z_0}\Big\{ & \frac{z_0}{\mu^2} - \frac{\partial z_0}{\partial \mu}[\frac{1}{r_2} + \frac{1}{\mu} + \frac{z_0(r_2+\mu)-\mu}{r_2\mu}(\frac{1}{1-z_0} \\ & - \frac{N z_0^{N-1}}{1-z_0^N})]\Big\},\end{aligned} \tag{E.2}$$

$$\frac{\partial z_0}{\partial \mu} = \frac{\mu}{\mu+r_2}\Big\{\frac{r_2 z_0}{\mu^2} + \frac{1-(1+\frac{r_2}{\mu})z_0}{\sqrt{(\lambda+\mu+kr_2+r_2)^2 - 4\lambda(\mu+r_2)}}\Big\}. \tag{E.3}$$

Combining these equations, we have

$$\begin{aligned}\frac{\partial P_{(N,P_t)}}{\partial \mu} = \frac{\hat{\lambda}e\Delta_P}{\gamma\mu^2}\Big\{ & e^{\frac{\mu l \ln 2}{B} - 1}(\frac{\mu l \ln 2}{B} - 1) - \frac{\gamma}{\Delta_P e}(P_o - P_{sleep} \\ & - \frac{\Delta_P}{\gamma} + 2E_s y(N,\mu,\lambda,k,r_2)\Big\},\end{aligned} \tag{E.4}$$

with $y(N,\mu,\lambda,k,r_2)$ given in Proposition 3. Making use of the fact that $\frac{\partial y(N,\mu,\lambda,k,r_2)}{\partial \mu} < 0$, the proof of which is omitted due

to space limitation, we have $\frac{\partial^2 P_{(N,P_t)}}{\partial \mu^2} > 0$. With the stability requirement $\mu > \hat{\lambda}$, only when

$$\frac{\partial P_{(N,P_t)}}{\partial \mu}|_{\mu\to\hat{\lambda}} = \frac{e\Delta_P}{\gamma\hat{\lambda}}\Big\{e^{\frac{\hat{\lambda}l\ln 2}{B}-1}\big(\frac{\hat{\lambda}l\ln 2}{B}-1\big) - \frac{\gamma}{\Delta_P e}\big(P_o - P_{sleep} - \frac{\Delta_P}{\gamma} - \frac{2\hat{\lambda}E_s}{N+f(N,\hat{\lambda},\lambda,k,r_2)}\big)\Big\} \geq 0, \quad \text{(E.5)}$$

which can be transformed into $l \geq \frac{B}{\hat{\lambda}\ln 2}\big\{\mathbf{W}\big[\frac{\gamma}{\Delta_P e}(P_o - P_{sleep} - \frac{2\hat{\lambda}E_s}{N+f(N,\hat{\lambda},\lambda,k,r_2)}) - \frac{1}{e}\big] + 1\big\}$, we always have $\frac{\partial P_{(N,P_t)}}{\partial \mu} > \frac{\partial P_{(N,P_t)}}{\partial \mu}|_{\mu\to\hat{\lambda}} \geq 0$, and $P_{(N,P_t)}$ monotonically increases with μ and P_t. Otherwise, $P_{(N,P_t)}$ first decreases and then increases with μ, and there exists the energy-optimal P_t^{eo} that minimizes $P_{(N,P_t)}$.

Appendix F
Proof of Proposition 5

According to Table I, the total power consumption of Poisson traffic is $P^p_{(N,P_t)} = \frac{\hat{\lambda}}{\mu}(P_o + \Delta_P P_t) + (1 - \frac{\hat{\lambda}}{\mu})(P_{sleep} + \frac{2\hat{\lambda}E_s}{N})$. Taking $\frac{\partial P^p_{(N,P_t)}}{\partial \hat{\lambda}} = 0$, we have

$$\hat{\lambda}^* = \frac{\mu}{2} + \frac{N}{4E_s}(P_o - P_{sleep} + \Delta_P P_t) > 0. \quad \text{(F.1)}$$

If $\hat{\lambda}^* < \mu$, which can be transformed into

$$N < \frac{2\mu E_s}{P_o - P_{sleep} + \Delta_P P_t}, \quad \text{(F.2)}$$

$P^p_{(N,P_t)}$ first increases and then decreases with $\hat{\lambda}$, and achieves its maximum at $\hat{\lambda}^*$. However, if $\hat{\lambda}^* \geq \mu$, $P^p_{(N,P_t)}$ is a monotonically increasing function of $\hat{\lambda}$ in the stability region $\hat{\lambda} < \mu$.

Similarly, for the average file size l, there is

$$\frac{\partial P^p_{(N,P_t)}}{\partial l} = \frac{\hat{\lambda}}{x}(P_o + \Delta_p P_t - P_{sleep} - \frac{2\hat{\lambda}E_s}{N}). \quad \text{(F.3)}$$

Therefore $P^p_{(N,P_t)}$ is a linear increasing function of l when $N > \frac{2\hat{\lambda}E_s}{P_o + \Delta_p P_t - P_{sleep}}$.

Appendix G
Proof of Proposition 6

Since there is $D_{(N,P_t)} = D_{th}$ at the optimal point given P_t, it can be transformed into

$$N(N - E) = A(1 - z_0^N) \quad \text{(G.1)}$$

with $A = -\frac{2\hat{\lambda}}{1-z_0}[(\frac{\hat{\lambda}}{r_2} + \frac{\hat{\lambda}}{\mu})z_0 - \frac{\hat{\lambda}}{r_2}][D_{th} - \frac{kr_2 + r_2 + \lambda - \mu}{r_2(\mu - \hat{\lambda})(1+k)}] > 0, E = 1 + 2\hat{\lambda}\big[D_{th} - \frac{\lambda - \hat{\lambda} + kr_2 + r_2}{(\mu - \hat{\lambda})r_2(1+k)}\big]$. According to Eq. (G.1), the optimal N can be treated as the unique non-zero intersection of two functions, as shown in Fig. 18. When $E \leq 0$, the lower and upper bounds are denoted as (N_l^-, N_u^-); the bounds are (N_l^+, N_u^+) with $E > 0$. Based on Fig. 18, with $N_l^- = 0$ and $N_l^+ = E$, the lower bound of the optimal N is $N_l = \max\{0, E\}$; with $N_u^- = N_u^+ = N_u$, the upper bound $N_u = \frac{E + \sqrt{E^2 + 4A}}{2}$ is the positive solution of

$$N(N - E) = A. \quad \text{(G.2)}$$

As a result, the optimal N given P_t satisfies $N_l < N < N_u$, i.e. $\max\{0, E\} < N < \frac{E + \sqrt{E^2 + 4A}}{2}$.

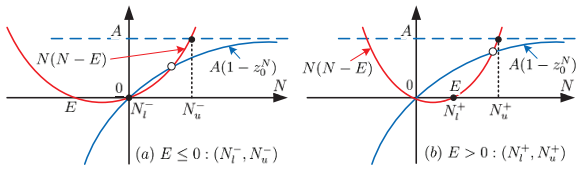


Fig. 18. The bounds for the optimal N.

Appendix H
Proof of Proposition 7

To get the lower bound of the total power consumption $P_{(N,P_t)}$, we make $N \to \infty$ in Eq. (10). With $\mu = \frac{B}{l}\log_2(1 + \gamma P_t)$, we can substitute $P_t = \frac{1}{\gamma}(2^{\frac{\mu l}{B}} - 1)$ into Eq. (10) and obtain

$$\begin{aligned} P_{(N,P_t)}|_{N\to\infty} &= P_{sleep} + \frac{\hat{\lambda}}{\mu}(P_o - P_{sleep} + \Delta_P P_t) \\ &= P_{sleep} + \frac{\hat{\lambda}}{\mu}[P_o - P_{sleep} + \frac{\Delta_P}{\gamma}(2^{\frac{\mu l}{B}} - 1)]. \end{aligned} \quad \text{(H.1)}$$

Taking $\frac{\partial P_{(N,P_t)}|_{N\to\infty}}{\partial \mu} = \frac{e\Delta_P\hat{\lambda}}{\gamma\mu^2}\big[e^{\frac{\mu l\ln 2}{B}-1}(\frac{\mu l\ln 2}{B} - 1) - [\frac{\gamma(P_o - P_{sleep})}{\Delta_P e} - \frac{1}{e}]\big] = 0$, we get

$$\mu^* = \frac{B}{l\ln 2}\big\{\mathbf{W}\big[\frac{\gamma(P_o - P_{sleep})}{\Delta_P e} - \frac{1}{e}\big] + 1\big\}. \quad \text{(H.2)}$$

Making use of the fact that for $t > 0$, $\delta \leq 1$, the inequality $\frac{t^2}{2} - t + 1 - \delta e^{-t} > 0$ always holds, we have

$$\begin{aligned} \frac{\partial^2 P_{(N,P_t)}|_{N\to\infty}}{\partial \mu^2} = \frac{2e\Delta_P\hat{\lambda}}{\gamma\mu^3}e^{\frac{\mu l\ln 2}{B}-1}\big[\frac{1}{2}(\frac{\mu l\ln 2}{B})^2 - \frac{\mu l\ln 2}{B} + 1 \\ -(1 - \frac{\gamma(P_o - P_{sleep})}{\Delta_P})e^{-\frac{\mu l\ln 2}{B}}\big] > 0. \end{aligned} \quad \text{(H.3)}$$

Therefore, $P_{(N,P_t)}|_{N\to\infty}$ has its minimum at the extreme point μ^*.

If $\mu^* \leq \hat{\lambda}$, i.e. $\hat{\lambda}l \geq \frac{B}{\ln 2}\big\{\mathbf{W}\big[\frac{\gamma(P_o - P_{sleep})}{\Delta_P e} - \frac{1}{e}\big] + 1\big\}$, it is not in the stability region $(\hat{\lambda}, \infty)$. Since $P_{(N,P_t)}|_{N\to\infty}$ increases with μ in this case, the asymptotic value given below is obtained when $\mu \to \hat{\lambda}$ with $P_t = \frac{1}{\gamma}(2^{\frac{\mu l}{B}} - 1) \to \frac{1}{\gamma}(2^{\frac{\hat{\lambda}l}{B}} - 1)$.

$$P_{lb} = P_{(N,P_t)}\big|_{N\to\infty, P_t \to \frac{1}{\gamma}(2^{\frac{\hat{\lambda}l}{B}} - 1)} = P_o + \frac{\Delta_P}{\gamma}(2^{\frac{\hat{\lambda}l}{B}} - 1). \quad \text{(H.4)}$$

If $\mu^* > \hat{\lambda}$, i.e. $\hat{\lambda}l < \frac{B}{\ln 2}\big\{\mathbf{W}\big[\frac{\gamma(P_o - P_{sleep})}{\Delta_P e} - \frac{1}{e}\big] + 1\big\}$, μ^* falls into the stability region. So P_{lb} is derived at μ^* in Eq. (H.2), with the transmit power to be

$$\begin{aligned} P_t^* &= \frac{1}{\gamma}(2^{\frac{\mu^* l}{B}} - 1) = \frac{1}{\gamma}(2^{\frac{1}{\ln 2}\{\mathbf{W}[\frac{\gamma}{\Delta_P e}(P_o - P_{sleep}) - \frac{1}{e}] + 1\}} - 1) \\ &= \frac{1}{\gamma}\Big(\frac{\frac{\gamma}{\Delta_P e}(P_o - P_{sleep}) - \frac{1}{e}}{\mathbf{W}[\frac{\gamma}{\Delta_P e}(P_o - P_{sleep}) - \frac{1}{e}]}e - 1\Big) \\ &= \frac{\frac{1}{\Delta_P}(P_o - P_{sleep}) - \frac{1}{\gamma}}{\mathbf{W}[\frac{\gamma}{\Delta_P e}(P_o - P_{sleep}) - \frac{1}{e}]} - \frac{1}{\gamma}. \end{aligned} \quad \text{(H.5)}$$

Based on Eq. (H.1), the lower bound in this case is

$$\begin{aligned} P_{lb} &= P_{sleep} + \frac{\hat{\lambda}}{\mu^*}(P_o - P_{sleep} + \Delta_P P_t^*) \\ &= P_{sleep} + \frac{\hat{\lambda}l(P_o - P_{sleep} - \frac{\Delta_P}{\gamma})\ln 2}{B\mathbf{W}[\frac{\gamma(P_o - P_{sleep})}{\Delta_P e} - \frac{1}{e}]}. \end{aligned} \quad \text{(H.6)}$$

APPENDIX I
GENERATION FUNCTION FOR THE SPP MODEL

Using similar method to those in Appendix B, the generation function $G^s(z) = G_1^s(z) + G_2^s(z)$ satisfies

$$g^s(z)G_1^s(z)=\mu z[(\lambda_2+\mu+r_2-\lambda_2 z-\frac{\mu}{z})\sum_{m=0}^{N-1}z^m q_{1,m}^s+r_2\sum_{m=0}^{N-1}z^m q_{2,m}^s], \tag{I.1}$$

$$g^s(z)G_2^s(z)=\mu z[(\lambda_1+\mu+r_1-\lambda_1 z-\frac{\mu}{z})\sum_{m=0}^{N-1}z^m q_{2,m}^s+r_1\sum_{m=0}^{N-1}z^m q_{1,m}^s], \tag{I.2}$$

where $g^s(z) = \lambda_1\lambda_2 z^3-(\lambda_1\lambda_2+\lambda_1 r_2+\lambda_1\mu+\lambda_2 r_1+\lambda_2\mu)z^2+(\lambda_1\mu+r_1\mu+\lambda_2\mu+r_2\mu+\mu^2)z-\mu^2$, and unique solution z_0^s exists for $g^s(z_0^s) = 0$ in $(0, 1)$. So there is $g^s(z_0^s)G^s(z_0^s) = 0$. Besides, the sleeping probability for the SPP also satisfies $\sum_{m=0}^{N-1}(q_{1,m}^s + q_{2,m}^s) = 1 - \frac{\lambda_1 r_2+\lambda_2 r_1}{\mu(r_1+r_2)}$.

Till now we have two equations for $2N$ variables. Different from the case of IPP where explicit expressions exist for $q_{1,0}$ and $q_{2,0}$, we need to use the following iterations, which make the $2N$ variables into two variables, $q_{1,0}^s$ and $q_{2,0}^s$. For $0 < m \leq N-1$,

$$q_{1,m}^s=\frac{1}{\lambda_1\lambda_2+\lambda_1 r_2+\lambda_2 r_1}\left[(\lambda_1\lambda_2+\lambda_1 r_2)q_{1,m-1}^s+\lambda_2 r_2 q_{2,m-1}^s\right], \tag{I.3}$$

$$q_{2,m}^s=\frac{1}{\lambda_1\lambda_2+\lambda_1 r_2+\lambda_2 r_1}\left[(\lambda_1\lambda_2+\lambda_2 r_1)q_{2,m-1}^s+\lambda_1 r_1 q_{1,m-1}^s\right]. \tag{I.4}$$

As a result, $q_{1,0}^s$ and $q_{2,0}^s$ can be solved numerically, and the generation function $G^s(z)$ can also be derived. Based on these, the total power consumption and delay performance can be obtained.

REFERENCES

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

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

[3] 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.

[4] A. J Fehske, F. Richter, and G. P Fettweis, "Energy Efficiency Improvements through Micro Sites in Cellular Mobile Radio Networks," *2nd International Workshop on Green Communications, GLOBECOM 2009*, pp. 1-5, Dec. 2009.

[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," *Mobicom*, Sep. 2009.

[7] 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," *IEEE Globecom*, Anaheim, Dec. 2012.

[8] 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. Select. Areas Commun.*, vol. 29, no. 8, Sep. 2011.

[9] J. Wu, Z. Yang, S. Zhou and Z. Niu, "A traffic-aware dynamic energy-saving scheme for cellular networks with heterogeneous traffic," *IEEE ICCT*, Jinan, Sep. 2011.

[10] 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, Aug. 2013.

[11] J. Kwak, K. Son, Y. Yi, and S. Chong, "Greening Effect of Spatio-Temporal Power Sharing Policies in Cellular Networks with Energy Constraints," *IEEE Transactions on Wireless Communications*, vol. 11, no. 12, pp. 4405-4415, Dec. 2012.

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

[13] D. Rajan, A. Sabharwal, and B. Aazhang, "Delay-bounded packet scheduling of bursty traffic over wireless channels," *IEEE Trans. Inform. Theory*, vol. 50, no. 1, pp. 125-144, Jan. 2004.

[14] B. E. Collins and R. L. Cruz, "Transmission policies for time varying channels with average delay constraints," in *Proc. 1999 Allerton Conf. on Commun., Control, & Comp.*, Monticello, IL, 1999.

[15] E. Uysal-Biyikoglu, B. Prabhakar, and A. El Gamal, "Energy-efficient packet transmission over a wireless link," *IEEE/ACM Trans. Netw.*, vol. 10, pp. 487-499, Aug. 2002.

[16] W. Chen, M. J. Neely, and U. Mitra, "Energy-efficient transmissions with individual packet delay constraints," *IEEE Trans. Inf. Theory.*, vol. 54, no. 5, pp. 2090-2109, May 2008.

[17] Y. Jin, J. Xu, and L. Qiu, "Energy-efficient scheduling with individual packet delay constraints and non-ideal circuit power," *Journal of Communications and Networks*, vol.16, no.1, pp. 36-44, Feb. 2014.

[18] 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.

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

[20] E. Hyytiä, R. Righter, S. Aalto, "Energy-aware Job Assignment in Server Farms with Setup Delays under LCFS and PS," *26th International Teletraffic Congress (ITC)* , pp. 1-9, 2014.

[21] Z. Niu, J. Zhang, X. Guo, and S. Zhou, "On Energy-Delay Tradeoff in Base Station Sleep Mode Operation," *12th IEEE Int. Conf. Communication Systems (ICCS)*, pp. 235-239, 2012.

[22] T. Zhao, J. Wu, S. Zhou, and Z. Niu, "Energy-Delay Tradeoffs of Virtual Base Stations With a Computational-Resource-Aware Energy Consumption Model," *14th IEEE Int. Conf. Communication Systems (ICCS)*, Macau, China, Nov. 2014.

[23] C. R. Baugh, J. Huang, R. Schwartz, and D. Trinkwon, "Traffic model for 802.16 TG3 MAC/PHY simulations," *IEEE 802.16 Broadband Wireless Access Working Group*, Technical Report, Mar. 2001.

[24] S. Andreev, A. Anisimov, Y. Koucheryavy, and A. Turlikov, "Practical Traffic Generation Model for Wireless Networks," in *4th ERCIM eMobility Workshop*, Luleå, Sweden, May 2010.

[25] S. Ghandali, S. M. Safavi, "Modeling multimedia traffic in IMS network using MMPP," *Proc. International Conf. on Electronic Computer Technology (ICECT)*, vol. 6, pp. 281-286, April, Kanyakumari, 2011.

[26] S. Shah-Heydari, T. Le-Ngoc, "MMPP models for multimedia traffic," *Telecommun. Syst.*, vol. 15, pp. 273-293, 2000.

[27] D. Niyato, E. Hossain, A.S. Alfa, "Performance analysis of multi-service wireless cellular networks with MMPP call arrival patterns," *IEEE Global Telecommunications Conference (GLOBECOM)*, vol. 5, pp. 3078-3082, Dallas, Nov. 2004.

[28] A. T. Anderson and B. F. Nielsen, "A Markovian approach for modeling packet traffic with long-range dependence," *IEEE Journal on Selected Areas in Communications*, vol. 16, pp. 719-732, Jun. 1998.

[29] T. Yoshihara, S. Kasahara, and Y. Takahashi, "Practical time-scale fitting of self-similar traffic with Markov-modulated Poisson process," *Kluwer Telecommunication Systems*, vol. 17, pp. 185-211, Jun. 2001.

[30] H. Tamura, Y. Yahiro, Y. Fukuda, K. Kawahara and Y. Oie, "Performance analysis of energy saving scheme with extra active period for LAN switches," in *Proc. IEEE Global Telecommunications Conf.*, pp. 198-203, Nov. 2007.

[31] T. Venkatesh and C. S. R. Murthy, *An Analytical Approach to Optical Burst Switched Networks*. New York: Springer-Verlag, 2010.

[32] J. Wu, Y. Bao, G. Miao and Z. Niu, "Base Station Sleeping and Power Control for Bursty Traffic in Cellular Networks," *IEEE ICC'14 WS-E2Nets*, Sydney, Jun. 2014.

[33] Z. Niu, *Fundamental theory of communication networks*, Teaching materials of Tsinghua Univ. (in Chinese).

[34] R. Gallager, *Discrete stochastic processes*, vol. 101, Boston: Kluwer Academic Publishers, 1996.

[35] I. Kamitsos, L. Andrew, H. Kim and M. Chiang, "Optimal sleep patterns for serving delay-tolerant jobs," *International Conference on Energy-Efficient Computing and Networking*, Apr. 2010.

[36] L. I. Sennott, *Stochastic Dynamic Programming and the Control of Queueing Systems*. New York: Wiley, 1999.
[37] X. Qiu, and K. Chawla, "On the performance of adaptive modulation in cellular systems," *IEEE Transacations on Communications*, vol. 47. no. 6, pp. 884-895, Jun. 1999.
[38] S. Borst, "User-level performance of channel-aware scheduling algorithms in wireless data networks," in *Proc. IEEE INFOCOM*, San Francisco, USA, pp. 321-331, Mar. 2003.
[39] T. Bonald, A. Proutière, "Wireless downlink channels: user performance and cell dimensioning," in *Proc. ACM Mobicom*, 2003.
[40] B. Rengarajan and G. de Veciana, "Architecture and abstractions for environment and traffic aware system-level coordination of wireless networks: The downlink case," in *Conference on Computer Communications (INFOCOM)*, pp. 502-510, 2008.
[41] U. Yechiali and P. Naor, "Queueing problems with heterogeneous arrivals and service," *Operations Res.*, vol. 19, pp. 722-734, 1971.
[42] T. L. Saaty, *Elements of queueing theory*. New York: McGraw-Hill, 1961.
[43] 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.

Jian Wu received her B.S. and Ph.D. degrees in Electronic Engineering from Beijing Jiaotong University in 2009 and Tsinghua University in 2015, respectively. She is currently a research scholar in Computer Science Department at University of California, Davis. She received the Best Paper Award from the 23th IEEE International Conference on Communication Technology (ICCT) in 2011. Her research interests include green wireless communications, cloud computing and multimedia wireless networks.

Yanan Bao received his B.S. and M.S. degrees in Electronic Engineering from Tsinghua University, China, in 2010 and 2013, respectively. He is currently a Ph.D. student at University of California, Davis. His research interests include wireless green communications, machine learning and data mining.

Guowang Miao received a B.S. and M.S. degree from Tsinghua University and a M.S. degree and Ph.D. degree from Georgia Institute of Technology, Atlanta, GA, USA. He once worked in Intel Labs as a research engineer and also in Samsung Research America as a Senior Standards engineer. In 2011, he won an Individual Gold Award from Samsung Telecom America for his contributions in LTE-A standardization. He joined KTH Royal Institute of Technology in Feb 2012 as an assistant professor and then starting from Feb 2015, he is a tenured associate professor in the same institution. His research interest is in the design and optimization of wireless communications and networking. He is the lead author of "Energy and Spectrum Efficient Wireless Network Design", a book published by Cambridge University Press. He has published more than fifty research papers on premier journals or conferences, had several patents granted, and many more patents filed. He has been a technical program committee member of many international conferences and is on the editorial board of several international journals. He was an exemplary reviewer for IEEE Communications Letters in 2011.

Sheng Zhou (S'06, 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 Lab, 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.

Dr. Zhou coreceived 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.

Zhisheng Niu (M'98-SM'99-F'12) graduated from Beijing Jiaotong University, China, in 1985, and got his M.E. and D.E. degrees from Toyohashi University of Technology, Japan, in 1989 and 1992, respectively. During 1992-94, he worked for Fujitsu Laboratories Ltd., Japan, and in 1994 joined with Tsinghua University, Beijing, China, where he is now a professor at the Department of Electronic Engineering. He is also a guest chair professor of Shandong University, China. 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 (2012-15) and Chair of Emerging Technology Committee (2014-15) of IEEE Communication Society, a distinguished lecturer (2014-16) of IEEE Vehicular Technologies Society, a member of the Fellow Nomination Committee of IEICE Communication Society (2013-14), standing committee member of Chinese Institute of Communications (CIC, 2012-16), 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 Award from IEEE Communication Society Asia-Pacific Board in 2013. He also co-received the Best Paper Awards from the 13th, 15th and 19th Asia-Pacific Conference on Communication (APCC) in 2007, 2009, and 2013, respectively, International Conference on Wireless Communications and Signal Processing (WCSP'13), and the Best Student Paper Award from the 25th International Teletraffic Congress (ITC25). 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 both IEEE and IEICE.