Characterizing Energy-Delay Tradeoff in Hyper-Cellular Networks with Base Station Sleeping Control

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

***Abstract*—Base station (BS) sleeping operation is one of the effective ways to save energy consumption of cellular networks, but it may lead to longer delay to the customers. The fundamental question then arises: *how much energy can be traded off by a tolerable delay*? In this paper, we characterize the fundamental tradeoffs between total energy consumption and overall delay in a BS with sleep mode operations by queueing models. Here, the BS total energy consumption includes not only the transmitting power but also basic power (for baseband processing, power amplifier, etc) and switch-over power of the BS working mode, and the overall delay includes not only transmission delay but also queueing delay. Specifically, the BS is modeled as an M/G/1 vacation queue with setup and close-down times, where the BS enters sleep mode if no customers arrive during the close-down (hysteretic) time after the queue becomes empty. When asleep, the BS stays in sleep mode until the queue builds up to N customers during the sleep period (N-Policy). Several closed-form formulas are derived to demonstrate the tradeoffs between the energy consumption and the mean delay for different wake-up policies by changing the close-down time, setup time, and the parameter N. It is shown that the relationship between the energy consumption and the mean delay is linear in terms of mean close-down time, but non-linear in terms of N. The explicit relationship between total power consumption and average delay with varying service rate is also analyzed theoretically, indicating that sacrificing delay cannot always be traded off for energy saving. In other words, larger N may lead to lower energy consumption, but there exists an optimal N^* that minimizes the mean delay and energy consumption at the same time. We also investigate the maximum delay (delay bound) for certain percentage of service and find that the delay bound is nearly linear in mean delay in the cases tested. Therefore, similar tradeoffs exist between energy consumption and the delay bound. In summary, the closed-form energy-delay tradeoffs cast light on designing BS sleeping and wake-up control policies which aim to save energy while maintaining acceptable quality of service.**

I. Introduction

Recently, it has been reported that information and communication technology (ICT) industry is becoming a significant part of the world energy consumption and cellular networks are among the main energy consumers in the ICT field. Specifically, base stations (BSs) account for over 80% of the cellular network energy consumption [2]. Therefore, in order to support increasing data transmission rate, energy efficiency is key in future base station operations [3].

One of the key approaches to make the mobile communication networks more energy-efficient is to have the cellular architecture and radio resource allocation more adaptive to network environment and traffic variations, including making some lightly-loaded base stations (BSs) go to sleep. This is the key concept of the so-called TANGO (Traffic-Aware Network planning and Green Operation) [4] and CHORUS (Collaborative and Harmonized Open Radio Ubiquitous Systems) [5] published by the authors earlier. To realize this, a new cellular framework, named *hyper-cellular* networks (HCN)[6], has been proposed, in which the coverage of control signals is decoupled from the coverage of traffic signals so that the traffic coverage can be more adaptive to the dynamics of traffic characteristics and QoS requirements. Due to this elasticity of HCN, some delay-insensitive users may have to experience some delay or other kind of QoS degradation in order to save energy, i.e., energy can be traded off by some delay. The fundamental question then arises: how much energy can be traded off by a tolerable delay?

The so-called energy-delay tradeoff (EDT) problem has a broader meaning rather than the tradeoff only between energy and delay. It means broadly that the tradeoff between the limited network resources (e.g., power level, buffer space, etc) and the tolerable performance (e.g., delay, packet loss, etc) degradation perceived by users. The characterization of EDT exposes the ways in which resources could be saved by compromising on the level of performance. In this paper, we will mainly focus on the tradeoff between total energy consumption and the end-to-end flow level delay caused by BS sleeping in the hyper-cellular networks in which the BSs go to sleep when necessary in order to save energy.

Sleep mode operation is an effective way to save energy while maintaining acceptable quality of service (QoS) [3], [7], [8], [9], [10], [11], [12]. To save energy, a BS can be turned off (or go to sleep mode) when the traffic load is relatively light, but the quality of service will deteriorate accordingly. In [7], the authors investigate energy saving sleep mode operations while maintaining acceptable throughput received by users. Based on a Markovian model, they solve a set of balance

This work is sponsored in part by the National Basic Research Program of China (2012CB316001), the Nature Science Foundation of China (61201191, 61321061, 61401250, 61461136004), NSF under Contract Nos. CNS-1302182 and CCF-0939370, AFOSR under Contract No. FA-9550-13-1-0008, and Hitachi Ltd. Part of this work has been presented in IEEE ICCS2012 [1] as an invited paper.

Zhisheng Niu, Xueying Guo, and Sheng Zhou are with Tsinghua National Laboratory for Information Science and Technology, the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China. Email: niuzhs@tsinghua.edu.cn.

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

equations and obtain the probability that users achieve the target throughput. [8], [9] consider sleep mode operations with blocking probability constraint in a cellular network.

In this paper, we consider the energy-delay tradeoff because delay performance is a key matric in mobile multimedia communications, i.e., *how much energy can be traded off by a certain amount of delay*? In [13], it has been shown that the tradeoffs do exist between average transmission power and average buffer delay by changing the transmission rate, but the BS sleep operation was not considered. We consider energy-delay tradeoffs in BS sleep mode operation and assume fixed transmission rate when the base station is turned on. When a base station is turned off, customers have to wait until the base station wakes up and therefore will experience longer delay. We aim to investigate how much energy can be traded off by the queueing delay in terms of the BS sleep time, setup time, and close-down time. We focus on the cases where the sleep mode operations do not affect customer arrival processes, and do not consider the benefit of coordinated multi-point [14] or cell zooming [15], where a customer can be served by a neighboring base station.

Several closed-form formulas are derived to demonstrate the tradeoffs between the energy consumption and the mean delay for different wake-up policies by changing the close-down time, setup time, and the parameter N. It is shown that the relationship between the energy consumption and the mean delay is linear in terms of mean close-down time, but non-linear in terms of N. The explicit relationship between total energy consumption and average delay with varying service rate is also analyzed theoretically, indicating that sacrificing longer delay does not always help to reduce energy consumption. In other words, larger N may lead to lower energy consumption, but there exists an optimal N^* that minimizes the mean delay and energy consumption at the same time. We also investigate the maximum delay for certain percentage of service, which is closely related to the mean delay. In summary, the closed-form tradeoffs cast light on designing BS sleep control policies which aim to save energy while maintaining acceptable quality of service. We have also investigated the bound on given percentile of overall delay and found that the delay bound is nearly linear in mean delay in the cases tested. Therefore, similar tradeoffs exist between energy consumption and the delay bound. In addition, we propose a two-step optimization method to optimize the sleeping operation parameters and to investigate the optimal energy-delay relationship for all wake-up policies.

II. Related Work and the State-of-the-Art

As one of the fundamental tradeoffs in communications (especially in wireless communications), the relationship between power (energy) consumption and QoS (in particular delay) has been studied for many years. But, most of them are on the *transmit* power and the *tranmission* delay from the perspective of physical layer. For example, it has been clearly shown in Shannon formula that there exists a fundamental tradeoff between them, i.e., using longer time for transmission can save energy or, equivalently speaking, energy consumption per bit in AWGN channels is minimized when bandwidth (W) or transmission time (t) is infinite. This result is extended to fading channels in [13] by specifically stating that the EDT does exist in a squared root form, and then further extended to a multiuser context in [16] showing that the square-root tradeoff is both necessary and achievable. There are also some studies on the EDT in specific systems such as CDMA networks [17] and industrial wireless networks [18].

However, if the circuit and/or basic power in a BS and/or the queueing delay of the packets are taken into account, the relationship between total energy consumption and end-to-end delay will get much more complicated. This is mainly because that the circuit and/or basic power in a BS could be much larger than the transmit power, e.g., the transmit power of a typical GSM base station is about 40W only while the total power of the BS could be as high as 1500W. The queueing delay due to the multiple access from different users could also dominate the end-to-end delay, in particular when the burstiness of the packet arrivals is high. In addition, we should not only consider the average delay but also the delay variation or the delay bound in practical networks, and not only the energy consumption and the delay in a single link but also those in a cell (with multiple links) or in a whole cellular network (with multiple cells) should be considered. As a result, characterizing the EDT in a cellular network is much complicated than before. To do this, we need to combine the information theory (directly linked to the energy consumption) and the queueing theory (directly linked to delay) together, which has been shown unconsummated yet in [19]. As an earlier work, paper [20] has shown that, if the packets can be queued before transmission, the queue itself can be viewed as a memory channel which may increase the capacity of the systems. But, the delay considered here is transmission delay only and the EDT problem was not considered yet.

Recently, if the circuit power and/or basic power is taken into account in addition to the transmit power, it has been shown that the energy and the delay are not always a tradeoff, i.e., energy consumption could be further reduced with shorter delay under some conditions [21]. This is because the circuit power and/or basic power could be linearly reduced when the transmission time decreases. As a result, the total energy consumption may be reduced when shortening the transmission time if the circuit power and/or basic power dominates. However, the authors considered transmission delay only. If the queueing delay is further taken into account, the relationship between energy and delay gets even more complicated, which has not been solved yet. In particular, there have been no any studies so far to discuss the impact of the sleeping control on the EDT.

In order to achieve the optimal energy-delay tradeoff, i.e., reduce the energy consumption as less as possible for a given delay requirement, two typical methods have been proposed. One is to adjust the power level of the BS according to the variation of traffic load rather than keeping the power at a constant level. For example, paper [22] has proposed an adaptive policy in which the service rate (i.e., the transmit power level) is changed while keeping almost flat the delay curve. But, considering the fact that the transmit power only accounts for a small portion of the whole energy consumption of a typical BS, this method is not so effective in terms of the energy efficiency. The other method is to make the BS go to sleep when the traffic load is quite low. This method is much more effective in terms of the energy efficiency because not only the transmit power but also the circuit and/or basic power can be saved by such a sleeping control. For instance, paper [23] has proposed a hysteretic sleeping control policy, in which the server stay in sleep mode when asleep until the queue builds up to the point where the ON threshold is met. After waking up, the server stays awake until all jobs in the queue are processed and then is turned OFF. It concludes that, compared with a baseline policy that never puts the server to sleep, 1) low utilization can result in almost 87% energy saving, and 2) high utilization results in only 7.4% energy saving. But, the EDT problem was not considered. In [24], the authors formulate a total cost minimization problem that allows for a flexible tradeoff between flow-level delay and energy consumption, but the sleep control was not considered. In this paper, we will mainly focus on the impact of sleep control parameters (e.g., close-down time, setup time, and sleeping time) on the EDT in a hyper-cellular network where the BSs go to sleep when the queue gets empty.

III. Hyper-Cellular Networks with BS Sleep Mode Operation

As discussed earlier, we have argued that the future cellular architecture has to be more adaptive to the traffic dynamics and the network environments. However, the deeply coupled structure of the signaling functions and data service functions in the existing cellular networks makes it very difficult to be realized. This is mainly because, even though there is little traffic needed to be transmitted, the base stations have to transmit pilot signals in order to keep the coverage which cannot adapt to the traffic dynamic and network environment easily. In order to make the cell coverage more adaptive to the traffic dynamic and network environment (i.e., soft coverage), the coverage for control functions (i.e., control coverage) and that for data service functions (i.e., data coverage) have to be decoupled so that the data service BSs (DBSs) can be deployed and operated in an on-demand manner while the control BSs (CBSs) will be always on to guarantee the coverage. That is called hyper-cellular networks (HCN) where the cell coverage is more ”soft” and smart than before because the DBSs can be easily switched off (sleep) and on (awake) in accordance with the traffic variations and environment changes.

In fact, the HCN concept can be further extended to the cases where different part of data from different layers of an user or different type of data from different users go through different DBSs, i.e., the DBSs could be further divided into multiple tiers to form different coverage and the different coverage layers can be optimized independently. Under this framework of separation, the data of different QoS requirements (data rate, delay, secrecy, reliability, on-line availability) may be supported by most appropriate radios with the lowest cost (in terms of deployment, energy, spectrum, etc.). As an example, most current applications may require both always-online experience and high data-rate communication on-demand. For these applications, the data to be transmitted over the air can at least be divided into the following two parts:

- The data to ensure the network aware of the state of the user (keep awake), and to ensure allocation of transmission resource for on-demand high rate data
- The high rate service data of the contents

These two parts of data are of quite different QoS requirements, thus in our architecture, they may be served through different coverage or different RATs (Radio Access Technologies). The first part of data is sometimes referred to as control data or signaling data, which is in general of very low rate but should be supported all the time. This is suitable to be served by a large coverage network, i.e., with large cell size. Since large cell size result in smoothing the variation or fluctuation of traffics in each cell, these base stations can be operated in a high efficiency condition.

With the support of the large coverage for the first part of data, the access points (or base stations) for high rate coverage can be muted when the traffic is low, and leave only a small portion of high rate access points to operate in a high efficiency condition.

As for the sleeping time and the wake-up policy, three typical scenario can be considered in general.

- *Single Sleep* (SS) Policy: The BS goes to sleep mode only once and then wake up after a specific time. One of the key features of this policy is that it can be easily implemented because the BS does not need to be aware of the network situation during the sleep. In other words, the BS can go to deep sleep with much lower power level. Of course, the longer the sleep time, the more energy can be saved from the sleeping control. But, packets arrived during the sleeping period will suffer from longer delay. Therefore, how to design the optimal sleeping time is a major challenge.
- *Multiple Sleep* (MS) Policy: If there is no any packets arrived during the sleeping period, the BS will go to sleep mode again until it finds some packets to be served when awake. Such a policy can further save energy than the SS policy by prolonging the sleeping time if there is no packets arrived during the last sleep period. But, it may lead to longer delay and, moreover, consume extra energy for the BS to wake up and check the status of the system.
- N-Policy: The sleeping BS wakes up only when N or more packets have been accumulated during the sleeping period. Compared with the SS and MS policies, N-policy could be more energy-efficient because the BS can stay in sleep mode longer if the arrival rate is relatively low without unnecessary wake-ups. The larger the parameter N, the more energy savings but the longer delay. Hence, how to design the optimal parameter N is a key challenge. One of the disadvantages of this policy is that someone has to count the number of arrivals during the sleeping period. If this has to be done by the sleeping BS, the energy saving gain by this policy could be very limited because the BS cannot go to deep sleep in order to count

the packet arrivals. This problem can be easily solved in hyper-cellular networks because the associated CBS can take this responsibility and then send a wake-up signal to the sleeping BS by wireline channel. Overall, N-policy is more suitable for the hyper-cellular networks if the parameter N is well designed.

In [25], the authors has studied the three wake-up policies extensively and proposed an optimal mechanism for the BS sleep control. It also concludes that the fundamental EDT of the three policies is quite similar. Therefore, we consider the N-policy only in the sequel.

IV. BS Sleep Mode Modeling and its Analysis

For a base station which uses code division multiple access (CDMA) technology or orthogonal frequency division multiple access (OFDMA) technology, the whole frequency resource should be turned off if the BS decides to go to sleep. Therefore, a base station can be modeled as a single server queue. The switch-over cost of the base station is considered as the setup time and the power consumption during the setup time. The switch-over cost is also an impediment of frequently turning on and off the base station. In general, longer close-down time will lead to shorter waiting time of packets, because packets which arrive during the close-down time will be served immediately without setup time. However, since the base station is idle in close-down time, it consumes more power compared with the case where the base station enters sleep mode immediately after serving all customers. Another effective way to reduce setup frequency is to turn the base station on when it sees $N > 1$ customers waiting. Larger N reduces the setup frequency and energy cost, but it may result in longer delay. To find the tradeoffs between energy and delay, and to design effective sleep policies, we mainly consider the effects of the close-down time and N on energy and delay in the follows.

A. $M/G/1$ Vacation Queue with Setup and Close-down Times

We model a base station as an $M/G/1$ vacation queue with close-down and setup times. Packets are assumed to arrive to the BS in a Poisson process[1] with parameter λ and request an *i.i.d.* (independently and identically distributed) service time B with mean h_B and squared coefficient of variation C_B^2. During this period, the transmit power of the BS is assumed to be fixed at P_{ON}[2]. When the queue becomes empty, the server keeps waiting for a while D (called *close-down* time) which follows a general distribution with mean h_D and squared coefficient of variation C_D^2. During this period, the transmit power of the BS is assumed to be P_{CD}. If any new packets arrive during the close-down time, the server immediately starts its service without any delay. But if no packets arrive during the close-down time, the server will be switched into sleep mode with a very low power level P_{SL}. The sleeping time (corresponds to the vacation time in queueing models) either follows an *i.i.d.* distribution with mean h_V and squared coefficient of variation C_V^2 in the SS and MS policies or is terminated by the N-th packet arrival in N-policy. Specifically for the N-policy, if N packets have arrived during the sleep period, the server starts to set up (or called warm-up) and then to serve new packets, where the server setup time is also generally distributed with mean h_S and squared coefficient of variation C_S^2 and the power level during the setup phase is P_{ST}. Such a queuing model has been studied in [27] and we will use it in the follows to characterize the EDT in our hyper-cellular networks with sleep mode operation.

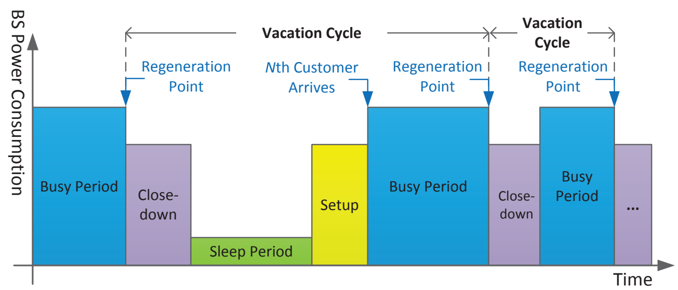


Fig. 1. Power transition and vacation cycle of a single-server vacation queue with setup and close-down times

For the analysis of delay variation and delay bound, we first denote by $\tilde{B(s)}$, $\tilde{D(s)}$, $\tilde{S(s)}$, and $\tilde{T(s)}$ the Laplace-Stieltjes transform of the service time, close-down time, setup time, and the total sojourn time of the packets in the system. Then, the probability that no customers arrive during the close-down time is given by $p_v = \tilde{D}(\lambda)$, which is in fact the probability that the server will get into sleep (vacation) mode anyway (called *sleeping probability* in the sequel). In this case, the close-down time is exactly the given variable D. On the contrary, if there is any packet arrival before the close-down time expires, the close-down phase will be terminated by the packet arrival. Based on the memoryless nature of Poisson arrivals, we can get the *effective* average close-down time as $\tilde{h_D} = (1-p_v)h_D + p_v\lambda^{-1}$. The mean length of a cycle $E[C]$ (time between two successive epochs at which the queue becomes empty) is then given by [28]:

$$
\begin{aligned}
E[C] &= (1-p_v)\left(\frac{1}{\lambda}+\frac{h_B}{1-\rho}\right)+p_v\left(\frac{N}{\lambda}+\frac{h_S+Nh_B}{1-\rho}\right) \\
&= \frac{1-p_v+p_v(N+\rho_S)}{\lambda(1-\rho)}, \qquad (1)
\end{aligned}
$$

where $\rho = \lambda h_B$ and $\rho_S = \lambda h_S$.

We then obtain the Laplace-Stieltjes transform of packet's sojourn time, which includes both the waiting time in the queue and the service time, as follow [28][3].

[1]Although the packet arrivals in real networks are in general much more bursty than Poisson process, we make this simple assumption because this paper aims at providing some closed-form expressions for the energy-delay tradeoff in order to characterize the EDT more explicitly. The extended discussions on the impact of the burstiness of arrival process on sleeping control can be found in [26].

[2]More general cases where the transmit power linearly increases as traffic load is considered in [25]

[3]Here we correct an error in the Laplace-Stieltjes transform of waiting time in the queue ($\tilde{W}(s) = \tilde{T}(s)/\tilde{B}(s)$) on page 136 of [28].

$$\begin{aligned}\tilde{T}(s) &= \frac{\tilde{B}(s)}{E[C]}\{\frac{1-p_v}{\lambda}+\frac{(1-p_v)(1-\tilde{B}(s))}{s-\lambda+\lambda\tilde{B}(s)} \\ &+ p_v\frac{\tilde{S}(s)}{\lambda}\frac{[\lambda/(s+\lambda)]^N-[\tilde{B}(s)]^N}{\lambda/(s+\lambda)-\tilde{B}(s)} \\ &+ p_v\frac{1-\tilde{S}(s)(\tilde{B}(s))^N}{s-\lambda+\lambda\tilde{B}(s)}\}. \end{aligned} \tag{2}$$

Differentiating Eq. (2) by s and then setting $s=0$ and substituting Eq.(1) into it, we obtain the mean sojourn time $E(T)$, which is in consistent with the result in [27]:

$$\begin{aligned}E[T] &= h_B+\frac{(1+C_B^2)\rho h_B}{2(1-\rho)} \\ &+ \frac{p_v[N(N-1)+2N\rho_S+(1+C_S^2)\rho_S^2]}{2\lambda[p_v(N+\rho_S)+1-p_v]}. \end{aligned} \tag{3}$$

This is in fact an extended version of the *decomposition theorem* in normal vacation queues [28], i.e., the first two parts are the mean waiting time of the normal M/G/1 queue without vacation, close-down, and setup times and the 3rd part is the additional delay caused by the vacation, the close-down, and the setup operations. It is also noteworthy that not only the first moments of the service and the setup times but also their second moments will have a great impact on the average delay performance, while only the first moment of the close-down time has impact on the average delay performance.

By averaging the power levels of the BS in service (P_{ON}), close-down (P_{CD}), sleeping (P_{SL}), and setup (P_{ST}) phases, we obtain the average power level as follow:

$$\begin{aligned}E[P] &= \rho P_{\mathrm{ON}}+\frac{1}{E[C]}(\frac{1-p_v}{\lambda}P_{\mathrm{CD}}+p_v\frac{N}{\lambda}P_{\mathrm{SL}}+p_vh_SP_{\mathrm{ST}}) \\ &= \frac{1}{E[C]}[\frac{1-p_v}{\lambda}(P_{\mathrm{CD}}-P_{\mathrm{SL}})+p_vh_S(P_{\mathrm{ST}}-P_{\mathrm{SL}})] \\ &+ (1-\rho)P_{\mathrm{SL}}+\rho P_{\mathrm{ON}}. \end{aligned} \tag{4}$$

This directly corresponds to the total energy consumption of the BS because it takes all the cases into account. If we multiply $E[P]$ with the BS operation time, that will lead to the total energy consumption.

B. EDT in terms of Close-down Time

We investigate the *closed-form* relationship between mean power and mean sojourn time by changing the mean close-down time. To do this, we first rewrite Eq.(3) as follows:

$$\begin{aligned}E[T] &= h_B+\frac{(1+C_B^2)\rho h_B}{2(1-\rho)} \\ &+ \frac{[N(N-1)+2N\rho_S+(1+C_S^2)\rho_S^2]}{2\lambda[(N+\rho_S)-1+p_v^{-1}]}, \end{aligned} \tag{5}$$

from which it can be seen that $E[T]$ is monotonically increasing convex function of p_v. Given the distribution of close-down time, as the mean close-down time (h_D) increases, the probability that no customers arrive during the close-down time (p_v) decreases and therefore $E[T]$ decreases too. In other words, $E[T]$ is monotonically decreasing function of h_D.

On the other hand, substituting Eq.(1) into Eq.(4) yields to

$$\begin{aligned}E[P] &= \frac{(1-\rho)[(1-p_v)(P_{\mathrm{CD}}-P_{\mathrm{SL}})+p_v\rho_S(P_{\mathrm{ST}}-P_{\mathrm{SL}})]}{1+p_v(N-1+\rho_S)} \\ &+ (1-\rho)P_{\mathrm{SL}}+\rho P_{\mathrm{ON}} \end{aligned} \tag{6}$$

from which it can been seen that $E[P]$ is monotonically decreasing function of p_v and therefore increasing function of h_D.

To bridging $E[T]$ with $E[P]$, we first represent p_v in terms of $E[T]$ based on Eq. (5) and then substitute p_v into Eq. (6), which yield to

$$\begin{aligned}E[P] &= \rho P_{\mathrm{ON}}+(1-\rho)P_{\mathrm{CD}} \\ &+ \frac{2\lambda(1-\rho)E[T]-2\rho(1-\rho)-(1+C_B^2)\rho^2}{(1+C_S^2)\rho_S^2+2N\rho_S+N(N-1)} \\ &\times [\rho_S(P_{\mathrm{ST}}-P_{\mathrm{SL}})-(N+\rho_S)(P_{\mathrm{CD}}-P_{\mathrm{SL}})]. \end{aligned} \tag{7}$$

Surprisingly, the relationship between $E[P]$ and $E[T]$ is *linear* in terms of mean close-down time for the given system parameters. This is quite different from the previous studies and can be used to guide the design of practical systems. For instance, within the tolerable delay scope, the system should set the mean close-down time as small as possible and the resulting energy saving gain should be linearly enhanced with the reduction of the mean close-down time.

C. EDT in terms of N (Sleeping Time)

In N-policy, the BS is waked up once upon N packets are accumulated in the queue, i.e., N is a control parameter for the sleeping time. Larger N leads to longer sleeping time and therefore less energy consumption. But, in this case, the BS goes to setup phase less frequently and therefore the end-to-end delay may not necessarily increase accordingly. Depending on the parameters of setup time distribution, there should be an optimal value of N, which leads to minimum end-to-end delay and total energy consumption. In order to get this optimal N^*, we generalize N to a real number and then use optimization method to obtain N^* as follows:

$$N^*=\sqrt{\rho_S^2C_S^2+\rho_S+(\frac{1-p_v}{p_v})^2}-\rho_S-\frac{1-p_v}{p_v}. \tag{8}$$

Here, N^* is feasible only when $\rho_S(C_S^2-1)p_v+3p_v-2>0$ holds. If $N>N^*$, the EDT does exist, i.e., larger N will lead to longer end-to-end delay but less energy consumption. In other words, sacrificing delay does not always help to save energy. However, if $N\le N^*$, the EDT does not exist anymore, i.e., larger N will lead to shorter end-to-end delay as well as less energy consumption. This finding is substantially valuable for the practical energy-efficiency design of hyper cellular networks.

In order to simplify the expression, the mean close-down time has been set to zero and therefore $p_v=1$ in the following analysis. In this case, mean power of N-policy is equivalent to a 1-policy system which has $E[S]/N$ mean setup time, given by

$$E[P]=\rho P_{\mathrm{ON}}+(1-\rho)P_{\mathrm{SL}}+\frac{(1-\rho)\rho_S}{N+\rho_S}(P_{\mathrm{ST}}-P_{\mathrm{SL}}). \tag{9}$$

However, the mean sojourn time is not necessarily monotonically increasing in N, given by

$$E[T]=E[B]+\frac{\rho h_B(1+C_B^2)}{2(1-\rho)}+\frac{N(N-1)+2N\rho_S+\rho_S^2(1+C_S^2)}{2\lambda(N+\rho_S)}. \tag{10}$$

By changing N, the explicit relationship between $E[P]$ and $E[T]$ is then given by

$$E[T]=\frac{\rho_S C_S^2+1}{2\lambda}A+\frac{h_S}{2A}+h_B+\frac{\rho h_B(1+C_B^2)}{2(1-\rho)}-\frac{1}{2\lambda}, \tag{11}$$

where

$$A=\frac{E[P]-(\rho P_{\mathrm{ON}}+(1-\rho)P_{\mathrm{SL}})}{(P_{\mathrm{ST}}-P_{\mathrm{SL}})(1-\rho)}. \tag{12}$$

D. Delay bound and delay variation

In practical systems, end users usually care more about the maximum delay or the violation probability of a predefined delay bound. That is to say, rather than the mean delay, the delay bound and energy tradeoff is more important in real system design. This brings even serious challenges to us because, in order to characterize the delay bound, we need to have the analytical results of the tail distribution of the waiting time. Although we have obtained the LST of the sojourn time as shown in Eq.(2), it is not trivial to obtain the tail distribution in an analytical form. In replace, we obtain the probability density function of the overall delay numerically by inverse Laplace transform of Eq. (2) and then asymptotically calculate the tail probability of the delay bound by Chebyshev's inequality or large deviation theorem. The numerical results will be shown in the following section.

Another metrics for delay variation is the variance of the time delay, which can be obtained from the LST shown in Eq.(2). But, unfortunately, the result is in general quite complicated. Just as an example, we show the result for $N=1$ as follow.

$$\begin{aligned}\mathrm{Var}[T] &= \left[12\lambda^2(1+p_v\rho_S)^2\right]^{-1}\Big\{4\rho_S^3 m_S^3 p_v\,(1+p_v\rho_S)\\ &+ 12\rho^2 c_B^2(1+p_v\rho_S)^2+12\rho_S^2 p_v(1+c_S^2-p_v)\\ &+ \frac{\rho^3(1+p_v\rho_S)^2}{(1-\rho)^2}\left[4m_B^3(1-\rho)+3\rho(1+c_B^2)^2\right]\\ &- 3p_v^2\rho_S^4(1+c_S^2)^2\Big\},\end{aligned} \tag{13}$$

where $m_B^3=E[B^3]/h_B^3$ and $m_S^3=E[S^3]/h_S^3$.

V. Numerical Results

In this section we provide numerical results to demonstrate the impacts of close-down time and N on base station performance and energy saving, and the tradeoffs between mean power and mean sojourn time. We also investigate the relationship between mean sojourn time and $T_{\max}^{0.01}$. In all these cases, customers arrive as a Poisson process with rate λ. When the base station is transmitting data in power level P_{ON}, the corresponding average service time is assumed to be h_B, which can be obtained via Shannon formula. When the base station is in setup or close-down phases, $P_{\mathrm{ST}}=P_{\mathrm{CD}}=0.9P_{\mathrm{ON}}$, which is justified in [3] as the power consumption in idle state. We assume that base station power consumption during sleep is $0.2P_{\mathrm{ON}}$.

A. Impact of Close-down Time

Figure 2 depicts the effects of the close-down time on the mean sojourn time and mean power. The setup time is deterministic and equal to h_B. The close-down and service times follow exponential distribution. We consider different load conditions, and assume $N=1$. We observe that as the close-down time increases, the mean power increases and mean sojourn time decreases. In light load conditions, sleep mode brings more benefits on energy saving.

Figure 3 depicts the linear relationship between the mean power and mean sojourn time for different N. Other parameters, such as λ, h_B, h_S, C_S^2, C_B^2, have effects on the slope of the linear function, but the linear relationship always exists.

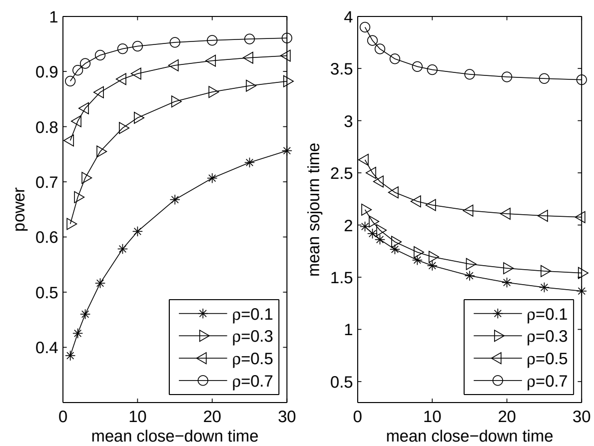


Fig. 2. Mean sojourn time (normalized by h_B) and mean power (normalized by P_{ON}) vs. mean close-down time (normalized by h_B) in an 1-policy $M/M/1$ queue with exponentially distributed close-down time and deterministic setup time. $h_S=h_B$.

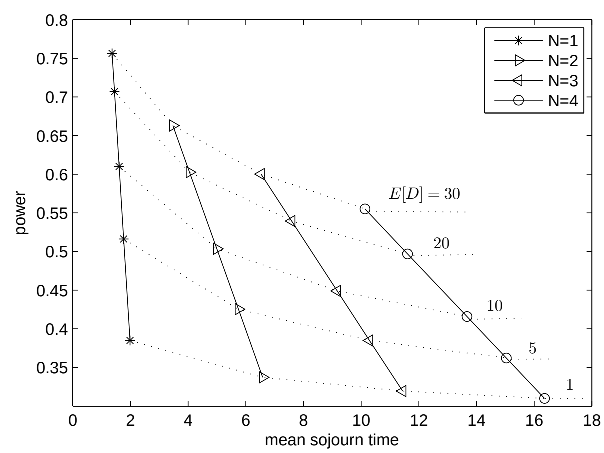


Fig. 3. Mean sojourn time (normalized by h_B) vs. mean power (normalized by P_{ON}) in an N-policy $M/M/1$ queue with exponentially distributed close-down time and deterministic setup time (changing the close-down time). $\rho=0.1$, $h_S=h_B$.

B. Impact of N

Figure 4 depicts the effects of N on the mean sojourn time and mean power. Mean setup time equals to h_B. We consider both light load and heavy load conditions, i.e, $\rho = \rho_S = 0.1$ and 0.8, respectively. We also consider the effects of the deviation of setup time, and let $C_S^2 = 0$ and 25, respectively. The deviation of setup time does not affect the mean power, which is given by Eq. (9). Let T_{wait} denote the waiting time of a customer that arrives during the sleep time until the server starts to setup, i.e., the time interval between one customer arrival and the epoch when the N^{th} customer arrives during the sleep phase. In light load conditions, the inter-arrival time between customers is long, and T_{wait} dominates the mean sojourn time for large N. We observe that in light load conditions, mean sojourn time is increasing in N. In heavy load conditions, T_{wait} is comparable with the setup time. By increasing N, the server goes to setup less often, and the benefit may outweigh the cost of longer T_{wait}, especially when the deviation of setup time is large. Therefore, there may exist $N > 1$ that minimizes the mean sojourn time given by Eq. (10).

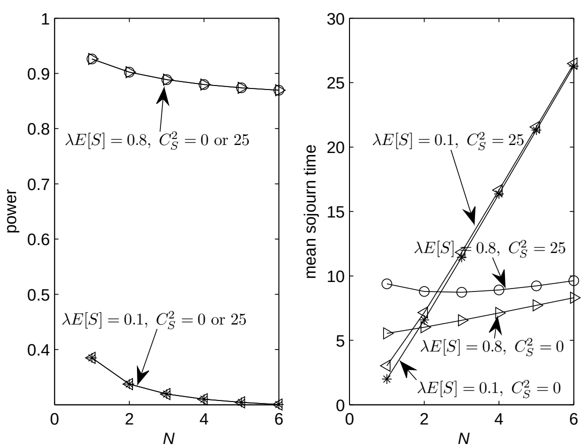


Fig. 4. Mean sojourn time (normalized by h_B) and mean power (normalized by P_{ON}) vs. N in an N-policy $M/M/1$ queue with close-down and setup times. $h_S = h_B$.

Figures 5 and 6 depict the relationships between the mean sojourn time and mean power. Since larger N always reduces mean power, but not necessarily increases the mean sojourn time, mean power is not a monotonically decreasing function in the mean sojourn time as depicted in Fig. 6.

C. Mean Delay vs. Delay Bound

We consider the relationship between the mean sojourn time and $T_{\max}^{0.01}$. From the cases we studied, the relationship between $E[T]$ and $T_{\max}^{0.01}$ is almost linear and depicted in Fig. 7. We obtain these cases by changing the close-down time. For simplicity, we assume that the setup times follow exponential distribution, and thus simplify the calculations of inverse Laplace transforms and tail probabilities. We consider the cases where the service times follow exponential and hyper-exponential distribution, where $p_1 = 0.8,\ p_2 = 0.2,\ \mu_1 = 8,\ \mu_2 = 2/9,\ C_B^2 = 7.125$. We observe that although larger deviation of setup and service times leads to significantly larger $T_{\max}^{0.01}$, the nearly linear relationship still exists.

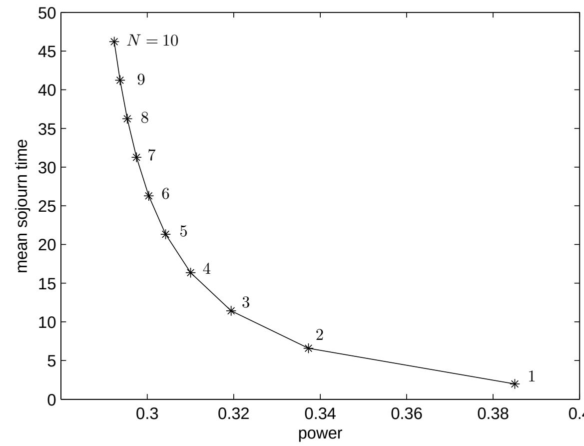


Fig. 5. Mean sojourn time (normalized by h_B) vs. mean power (normalized by P_{ON}) in an N-policy $M/M/1$ queue with close-down and setup times (changing N). $h_S = h_B$, $\rho_S = 0.1$, $C_S^2 = 0$.

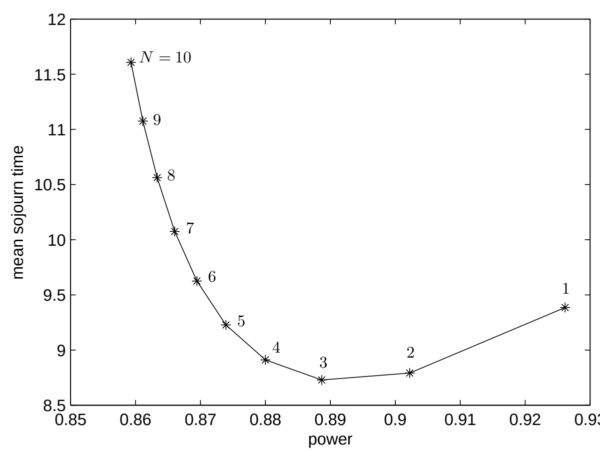


Fig. 6. Mean sojourn time (normalized by h_B) vs. mean power (normalized by P_{ON}) in an N-policy $M/M/1$ queue with close-down and setup times (changing N). $h_S = h_B$, $\rho_S = 0.8$, $C_S^2 = 25$.

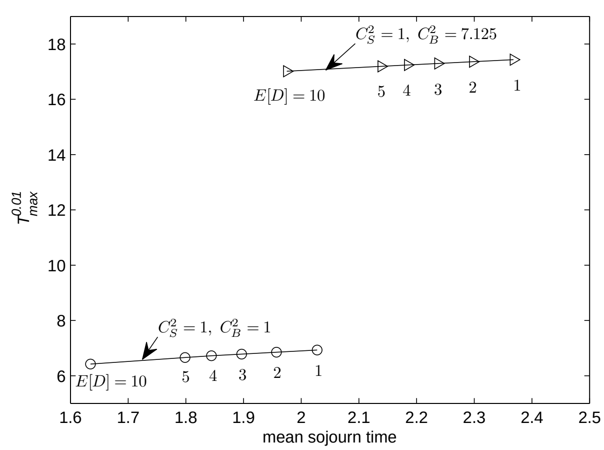


Fig. 7. $T_{\max}^{0.01}$ vs. mean sojourn time (both normalized by h_B) in a 1-policy $M/M/1$ queue with close-down and setup times (changing the close-down time).

The cases in Fig. 8 have the same distributions of service, setup, close-down times as in Fig. 7. The only difference is that we aim to investigate the effect of N, and therefore calculate the sojourn time by changing N rather than the close-down time. We observe that by changing N from 1 to 5, the mean sojourn time is also nearly linear with $T_{\max}^{0.01}$. Moreover, $T_{\max}^{0.01}$ is not very sensitive to the deviation of the service time. One reason is that the effect of setup time on service delay diminishes as N increases. Another reason is that T_{wait} dominates the delay as N increases. Note that we are considering lightly loaded condition for the base station sleeping operation.

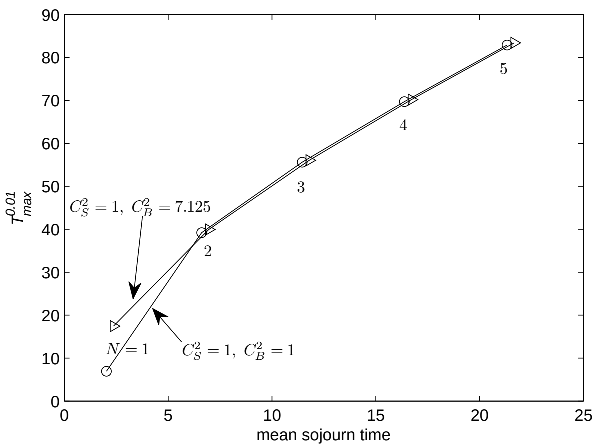


Fig. 8. $T_{\max}^{0.01}$ vs. mean sojourn time (both normalized by h_B) in an N-policy $M/M/1$ queue with close-down and setup times (changing N).

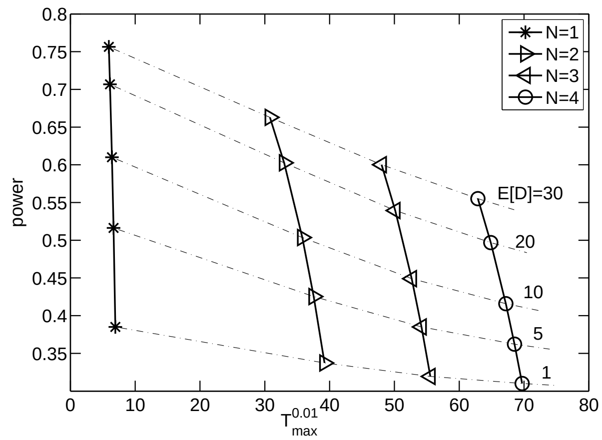


Fig. 9. Mean power (normalized by P_{ON}) vs. $T_{\max}^{0.01}$ (normalized by h_B) in an N-policy M/M/1 queue with exponential distributed close-down time and deterministic setup time (changing the close-down time). $\rho = 0.1$, $h_S = h_B$.

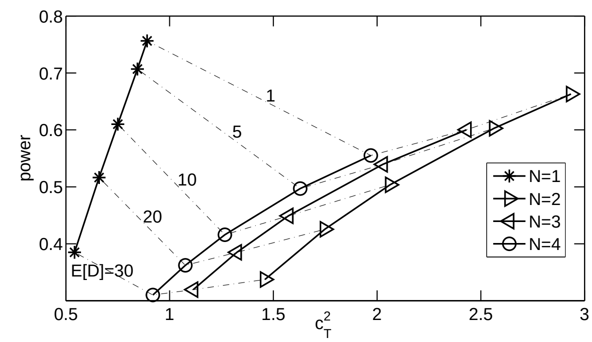


Fig. 10. Mean power (normalized by P_{ON}) vs. squared coefficient of variation c_T^2 in an N-policy M/M/1 queue with exponential distributed close-down time and deterministic setup time (changing the close-down time). $\rho = 0.1$, $h_S = h_B$

Fig.(9) shows the tradeoff between mean power and delay bound while changing parameters $E[D]$ and N. It can be clearly seen that the tradeoff does exist and almost in linear form, i.e., great energy savings can be achieved if the delay bound can be extended a little bit by reducing $E[D]$ or increasing N. However, as shown in Fig.(10), this may make the delay more bursty, especially when reducing $E[D]$. When increasing N but keeping $E[D]$ fixed, the situation is a little bit complicated. Specifically, $N = 2$ leads to the highest burstiness of the delay, while $N = 1$ leads to the lowest, which is indeed a surprising result. These insights provide a clear guideline for the sleeping control design.

VI. Conclusion

In this paper, we have shown that energy can be traded off by delay in hyper cellular networks with base station sleeping control. The general conclusion is that the tradeoffs between energy consumption and delay depend on base station sleeping/weakup policies as well as many other system parameters. Specifically, the closed-form relationships between mean power and mean overall delay have been derived for three wake-up policies based on an $M/G/1$ vacation queueing model with setup and close-down times. By changing the close-down time, mean power is a monotonically decreasing linear function of the mean delay, while the deviation of setup time does not have any impact on energy consumption. By increasing N, mean power decreases, but there may exist $N > 1$ that minimizes the mean delay, in which case energy may not be monotonically decreasing in delay. A nearly linear relationship is also observed between the mean delay and the bound on given percentile customer delay from the cases we tested, which is not very sensitive to the distributions of service time. Therefore, similar tradeoffs exist between mean power and the delay bound. For the control policies discussed in this paper, by limiting the mean delay to a corresponding level, they guarantee with given high probability that customers be served within their tolerable delay.

All in all, we conclude that, with the BS sleeping control, the energy and the delay are not always a tradeoff, i.e., it is possible to make the system energy consumption and the end-to-end delay decrease simultaneously by carefully designing the system parameters. This casts light on designing BS sleeping and wake-up control policies in order to save energy while maintaining acceptable quality of service.

Acknowledgment

The authors would like to thank Mr. Jianan Zhang at MIT for his initial contribution to this paper.

References

[1] Z. Niu, J. Zhang, X. Guo, , and S. Zhou, "On Energy-Delay tradeoff in Base Station Sleep Mode Operation," in *2012 IEEE International Conference on Communications and Systems (ICCS2012)*, 2012.

[2] F. Richter, A. J. Fehske, and G. Fettweis, "Energy efficiency aspects of base station deployment strategies for cellular networks," in *IEEE Vehicular Technology Conference*, 2009.

[3] M. A. Imran and *et, al*, "Energy efficiency analysis of the reference systems, areas of improvements and target breakdown," EARTH, Tech. Rep., 2011. [Online]. Available: https://bscw.ict-earth.eu/pub/bscw.cgi/d71252/EARTH_WP2_D2.3_v2.pdf

[4] Z. Niu, "TANGO: Traffic-Aware Network Planning and Green Operation," *IEEE Wireless Commun. Mag.*, vol. 10, no. 10, Oct. 2011.

[5] S. Zhou, Z. Niu, S. Tanabe, and P. Yang, "CHORUS: Framework for Scalable Collaboration in Heterogeneous Networks with Cognitive Synergetic," *IEEE Wireless Commun. Mag.*, vol. 10, no. 4, pp. 133–139, Aug. 2013.

[6] Z. Niu, S. Zhou, S. Zhou, X. Zhong, and J. Wang, "Energy efficiency and resource optimized hyper-cellular mobile communication system architecture and its technical challenges," *Science China: Information Science*, vol. 42, no. 10, pp. 1191–1203, Oct. 2012.

[7] S.-E. Elayoubi, L. Saker, and T. Chahed, "Optimal control for base station sleep mode in energy efficient radio access networks," in *IEEE INFOCOM*, 2011, pp. 106–110.

[8] J. Gong, S. Zhou, and Z. Niu, "A dynamic programming approach for base station sleeping in cellular networks," *IEICE Trans. Commun.*, vol. E95.B, pp. 551–562, Feb. 2012.

[9] E. Oh and B. Krishnamachari, "Energy savings through dynamic base station switching in cellular wireless access networks," in *IEEE GLOBECOM*, Dec. 2010, pp. 1 –5.

[10] F. B. I. Ashraf and L. Ho, "Power savings in small cell deployments via sleep mode techniques," in *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications Workshops*, 2010, pp. 307–311.

[11] S. W. Fuhrmann and R. B. Cooper, "Sleep mode techniques for small cell deployments," *IEEE Commun. Mag.*, vol. 49, no. 8, pp. 72–79, Aug. 2011.

[12] S. Mclaughlin, P. Grant, J. Thompson, H. Haas, D. Laurenson, C. Khirallah, Y. Hou, and R. Wang, "Techniques for improving cellular radio base station energy efficiency," *IEEE Wireless Commun.*, vol. 18, no. 5, pp. 10 –17, Oct. 2011.

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

[14] R. Irmer, H. Droste, P. Marsch, M. Grieger, G. Fettweis, S. Brueck, H.-P. Mayer, L. Thiele, and V. Jungnickel, "Coordinated multipoint: Concepts, performance, and field trial results." *IEEE Commun. Mag.*, vol. 49, no. 2, pp. 102–111, 2011.

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

[16] M. J. Neely, "Optimal energy and delay tradeoffs for multiuser wireless downlinks," *IEEE J. Selected Area in Commun.*, vol. 53, no. 9, 2007.

[17] F. Meshkati, H. V. Poor, and S. C. Schwartz, "Energy efficiency-delay tradeoffs in cdma networks: A game-theoretic approach," *IEEE Trans. Inform. Theory*, vol. 55, no. 7, 2009.

[18] N. Petreska, H. Al-Zubaidy, and J. Gross, "Power minimization for industrial wireless networks under statistical delay constraints," in *26th International Teletraffic Congress (ITC26)*, 2014.

[19] A. Ephremides and B. Hajek, "Information theory and communication networks: an unconsummated union," *IEEE Trans. Inform. Theory*, vol. 44, no. 10, 1998.

[20] V. Anantharam and S. Verdu, "Bit through queues," *IEEE Trans. Inform. Theory*, vol. 42, no. 1, 1996.

[21] G. Miao and G. Y. Li, "Cross-layer energy-efficient wireless communications: A survey," *Wireless Commun. Mobile Comp.*, 2009.

[22] A. Bianco, M. R. Casu, P. Giaccone, and M. Ricca, "Joint Delay and Power Control in Single-Server Queueing Systems," in *2013 IEEE Online Green Communications (IEEE OnlineGreenComm'13)*, 2013.

[23] I. Kamitsos, L. Andrew, H. Kim, and M. Chiang, "Optimal Sleep Patterns for Serving Delay-Tolerant Jobs," in *1st International Conference on Energy-Efficient Computing and Networking*, 2010.

[24] K. Son, H. Kim, Y. Yi, and B. Krishnamachari, "Base Station Operation and User Association Mechanism for Energy-Delay tradeoffs in Green Cellular Networks," *IEEE J. Selected Area Commun.*, vol. 29, no. 8, pp. 1525–1536, Aug. 2011.

[25] X. Guo, S. Zhou, Z. Niu, and P. R. Kumar, "Optimal wake-up mechanism for single base station with sleep mode," in *2013 International Teletraffic Congress (ITC2013)*, 2013.

[26] J. Wu, Y. Bao, G. Miao, and Z. Niu, "Base Station Sleeping and Power Control for Bursty Traffic in Cellular Networks," in *2014 IEEE International Conference on Communications (ICC2014) Workshop on Energy-Efficient Networks*, 2014.

[27] M. Yadin and P. Naor, "Queuing systems with a removable service station," *Operations Research Quarterly*, vol. 14, no. 4, pp. 393–405, Dec. 1963.

[28] H. Takagi, *Queueing analysis: a foundation of performance evaluation. Volume 1: Vacation and Priority Systems*. Elsevier Science, 1991.

Zhisheng Niu (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.

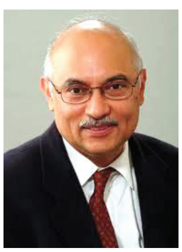
Dr. Niu has been an active volunteer for various academic societies, including Director for Conference Publications (2010-11) and Director for Asia-Pacific Board (2008-09) of IEEE Communication Society, Membership Development Coordinator (2009-10) of IEEE Region 10, Councilor of IEICE-Japan (2009-11), 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.

Xueying Guo (S'14) received her B.E. in Electronic Engineering from Tsinghua University, Beijing, China in 2011, where she is currently working toward her Ph.D. degree in the same department. From Oct. 2013 to Oct. 2014, she was a visiting student at Computer Engineering and System Group, Electrical and Computer Engineering Department of Texas A&M University. She received the Best Student Paper Award from the 25th International Teletraffic Congress (ITC) in 2013. Her research interests include resource allocation in cellular networks, cyber-physical systems, and green wireless communications.

Sheng Zhou (M'12) received the B.E. and Ph.D. degrees in electronic engineering from Tsinghua University, Beijing, China, in 2005 and 2011, respectively. From January to June 2010, he was a visiting student at the Wireless System 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.



P. R. Kumar (F'88) received the B.Tech. degree in electrical engineering (electronics) from I.I.T. Madras, India, in 1973, and the M.S. and D.Sc. degrees in systems science and mathematics from Washington University, St. Louis, MO, USA, in 1975 and 1977, respectively.

From 1977-1984, Dr. Kumar was a faculty member in the Department of Mathematics at the University of Maryland Baltimore County. From 1985-2011, he was a faculty member in the Department of Electrical and Computer Engineering and the Coordinated Science Laboratory at the University of Illinois. Currently he is at Texas A & M University, College Station, TX, USA, where he holds the College of Engineering Chair in Computer Engineering. He has worked on problems in game theory, adaptive control, stochastic systems, simulated annealing, neural networks, machine learning, queuing networks, manufacturing systems, scheduling, wafer fabrication plants and information theory. His current research interests are wireless networks, sensor networks, cyber-physical systems, and the convergence of control, communication and computation.

Dr. Kumar received the IEEE Field Award for Control Systems, the Donald P. Eckman Award of the American Automatic Control Council, the Fred W. Ellersick Prize of the IEEE Communications Society, and the Outstanding Contribution Award of ACM SIGMOBILE.He is a member of the National Academy of Engineering of the USA, and the Academy of Sciences of the Developing World. He was awarded an honorary doctorate by ETH, Zurich, and a Guest Chair Professor at Tsinghua University, Beijing, China. He is an Honorary Professor at IIT Hyderabad. He was awarded the Daniel C. Drucker Eminent Faculty Award from the College of Engineering at the University of Illinois.