Energy Efficiency-QoS Tradeoff in Cellular Networks with Base-Station Sleeping

Jingjin Wu*†, Eric W. M. Wong†, Yin-Chi Chan†, and Moshe Zukerman†
*Division of Science and Technology, BNU-HKBU United International College,
Zhuhai, Guangdong, P. R. China
†Department of Electronic Engineering, City University of Hong Kong, Kowloon, Hong Kong
Email: jingjinwu@uic.edu.hk; {eeewong, ycchan26, m.zu}@cityu.edu.hk

***Abstract*—Energy efficiency and Quality of Service (QoS) are two important considerations for the design and planning of cellular networks. One effective approach to handle the tradeoff between power consumption and QoS is to switch some of the Base Stations (BSs) to sleep mode when traffic load is low. In this paper, we model each BS as a processor-sharing queue with vacations, and investigate the performance of three BS sleeping schemes, namely the isolated scheme in which each BS switches mode based on its own load, the cooperative scheme in which traffic is allowed to overflow from sleeping BSs to neighboring active BSs, and a combination of both schemes. We propose a robust, scalable and computationally efficient analytical method to evaluate QoS metrics (including mean delay and blocking probability) and power consumption for each scheme and validate their accuracy by simulations. We also demonstrate the power consumption and QoS trade-off by extensive and statistically reliable experiments, and compare the performance of the three schemes under different network conditions.**

I. Introduction

In recent years, base station (BS) sleeping has become an important technique to reduce energy consumption in cellular mobile networks [1], [2]. Energy saving is achieved by switching BSs (or certain components of them) into a low power-consuming mode called "sleep mode" when traffic load is low. Comparing to other energy-efficient approaches such as upgrading hardware components or adopting renewable energy resources, BS sleeping has the advantage of convenience and cost effectiveness as it can be implemented in existing network infrastructure. As BSs are responsible for a significant proportion of energy consumed in cellular networks (over 80% in certain scenarios), BS sleeping has the potential to save a large amount of energy.

On the other hand, the total capacity of the network is reduced while some BSs are switched to the sleep mode. Therefore, it is important to be able to monitor and evaluate the quality of service (QoS) and understand its implications with BS sleeping. In this paper, we investigate the trade-off between energy saving and QoS metrics, including mean delay and blocking probability, by comparing different BS sleeping schemes.

Specifically, we model each BS as a single server Processor Sharing (PS) queue with Poisson arrivals, exponentially distributed service time, a finite buffer of size k and vacations, namely an M/M/1/k-PS queue with vacations. Henceforth, we will use the notation M/M/1/k-PS queue to mean the general case of this queue with or without vacations. If we mean a specific case we will write specifically if it is with or without vacations. Under this model, data or voice calls from different users are assumed to arrive according to a Poisson process (a common assumption for current cellular networks, see e.g., [3], [4]) and are served simultaneously by the BS. In a PS queue, the service capacity of the server (BS) is shared equally among all the customers being served. Due to the prevalence of mobile multimedia applications today such as web browsing, video streaming, online gaming, peer-to-peer video on demand and video conference, the dominant traffic of cellular networks has been changed from voice calls to packet-switched data. Then the need to avoid situations where large flows of data generated by some users slow down service to users generation small flows justifies the use of a processor sharing queuing discipline that in turn justifies the PS model [3], [5]–[9].

Another notable feature of the multimedia mobile traffic is delay sensitivity, as a minimum data rate needs to be guaranteed for such traffic. To guarantee that admitted connections satisfy delay and data rate requirements, an upper limit is set on the number of admitted connections. In this case, we need to evaluate both the mean delay and the drop rate due to violation of such requirements. In this regard, the drop rate is equivalent to the blocking probability of the M/M/1/k-PS queue. Hereafter, we will refer such "drop rate" as "blocking probability" for consistency.

BS sleeping can be either implemented separately in each single cell (*isolated scheme*) or cooperatively among multiple BSs (*cooperative scheme*). The cooperative scheme is based on dynamic capacity allocation and user association techniques [10], [11]. These techniques enable a user to use the capacity originally assigned to another BS if the first BS it attempts cannot offer the required service due to insufficient capacity or sleeping operation [10], [11].

Simulation is the traditional way of evaluating QoS metrics such as blocking probability and delay in telecommunication networks when an exact analytical solution method is not available. However, simulations are not scalable and cannot be used in realistically sized systems and networks as the running time becomes prohibitive. Therefore, evaluations by analytical approximations with reasonable accuracy and computational efficiency are more desirable for applications such as network design, where computational efficiency is key for searching

978-1-5090-5019-2/17/$31.00 ©2017 IEEE

optimal solutions. In this paper, we will propose analytical approximation methods based on queuing theory and the recently established Information Exchange Surrogate Approximation (IESA) framework [8], [12], [13] to obtain the mean delay and blocking probability, and compare the performance of the *isolated*, *cooperative* and *hybrid* (a combination of isolated and cooperative) schemes under different network conditions.

The main contributions of this paper are as follows:

- We provide a model of a cellular BS with sleeping technique as an M/M/1/k-PS queue with vacations and address the importance of blocking probability due to violation of delay requirement for multimedia mobile traffic. To the best of our knowledge, this is the first work to model a cellular BS as a PS queue with finite buffer, and measure QoS (including mean delay and blocking probability) and energy efficiency at the same time.
- We provide new robust, accurate, scalable and computationally efficient analytical approximation methods to evaluate QoS metrics, including mean delay and blocking probability, in multi-BS cellular networks with BS sleeping.
- We compare the performance of different BS sleeping schemes under different network conditions in terms of the power consumption and QoS tradeoff between energy saving and QoS metrics.

The remainder of this paper is organized as follows: Section II reviews relevant existing work on BS sleeping schemes and related performance evaluation tools. Section III describes network and power consumption models used in this paper. Section IV provides analysis on QoS metrics for different BS sleeping schemes. Numerical results are presented in Section V and the paper is concluded in Section VI.

II. Related work

There has been existing research addressing energy efficiency or QoS by modeling a BS as an M/G/1 queue or its close variants. For example, Song *et al.* [9] studied resource management in a cellular/WLAN integrated network by modeling a WLAN channel as an M/G/1/k-PS queue (G indicates general service times). The authors derived an appropriate size threshold for load sharing between cellular and WLAN components of the network, but their analysis did not involve energy efficiency. Guo *et al.* [14] considered a single server queue with vacations and obtained closed-form results of the tradeoff between energy consumption and delay for different sleeping schemes. The authors also considered a finite-buffer queue with vacations and a PS service discipline, namely an M/G/1/k - PS queue with vacations, but they did not estimate the blocking probability caused by violation of delay requirement. A similar approach was found in [3]. The authors modelled a BS as an M/G/1-PS queue with vacations, and derived the optimal parameters to achieve the optimal delay-power tradeoff for a joint BS sleeping and power matching scheme. Notably, the authors demonstrated that the N-policy (an *isolated scheme*, to be described in details in Section III-A) is better than other policies in terms of achieving optimal delay-power tradeoff.

Both [3] and [14] studied a single BS scenario and did not consider interaction among different BSs in the network. However, the single BS model is rather limited as dynamic user association schemes have been proposed to enable users of sleeping BSs to instead associated with a nearby active BS to continue their services. Tabassum *et al.* [11] proposed to associate users of sleeping BSs to an active BS with maximum mean channel access probability, aiming at improving spectral efficiency and minimizing outage probability of the network. The authors derived analytical approximation results for spectral efficiency and outage probability. However, delay, as another important QoS metric for contemporary cellular networks, is not considered in their work.

Cellular networks with *channel borrowing capabilities* can be considered as *overflow loss systems*, as suggested by Kelly [15]. In such systems, a request can overflow to an alternative BS if the first BS it attempts is busy. Dynamic user association schemes resemble channel borrowing in some way, as users originally associated to sleeping BSs are allowed to be served by other BSs remaining active. The Erlang Fixed-Point Approximation (EFPA) [16] was the classical approximation method for evaluating blocking probability in overflow loss systems. However, it can lead to very inaccurate estimation results in systems where *mutual overflow* effects are present, due to unrealistic assumptions [12], [17]. Wong *et al.* [12] proposed the IESA framework, which has its roots in the EFPA, in order to improve the accuracy of approximation for such systems. The work was further improved in [8] by including other approximation techniques such as moment matching.

Previously, we have verified that IESA is much more accurate than EFPA in terms of blocking probability estimation in a cellular network model with channel borrowing capabilities, and further extended the model to allow BS sleeping with fixed switching patterns (i.e. the *cooperative scheme*) in [13]. We considered blocking probability as the QoS metric by modeling each BS as an M/M/k/k queue (loss system). This paper uses the more realistic M/M/1/k-PS queue with vacations (delay-loss system) and measures both mean delay and blocking probability. Furthermore, this paper considers the *cooperative scheme* as well as two additional BS sleeping schemes (*isolated scheme* and *hybrid scheme*). The schemes will be discussed in detail in the later sections of this paper.

III. System model

A. Network and BS sleeping schemes

To make our analysis more tractable, each BS is modelled as an M/M/1/k-PS queue with vacations. That is, new users arrive at each BS according to a Poisson process with rate λ. Service times of users are exponentially distributed with mean $1/\mu$. This is also based on the properties for the M/G/1/k-PS queuing system [9] where the mean delay and blocking probability are insensitive to the service time distribution.

The M/M/1/k-PS system will not accept further customers when there are k customers in a BS. Therefore, the minimum data rate for each accepted customer is guaranteed. When the value of k increases, more customers may be accepted simultaneously and thus accepted customers would have lower data rate and higher mean delay if offered traffic is high.

We consider the following three BS sleeping schemes in this paper.

- *Isolated (N-policy) scheme*: as described in [3] and [14], the BS will switch to sleep mode when it has been idle (serving no users) for a close-down period of t^*. A sleeping BS will be reactivated when N or more users have been accumulated during the sleeping period. The Markov Chain representation of this policy is presented in Fig. 1. In the 2-D Markov Chain, the letter "A" or "S" indicates the state of the BS (active or sleep) and the number represents the number of customers in the BS.
- *Cooperative scheme*: BSs are selectively switched to sleep based on fixed or dynamic patterns according to traffic load. Under this scheme, users associated with a BS that has gone to sleep will be re-associated and served by one of active BSs nearby [11].
- *Hybrid scheme*: some BSs are switched to sleep based on fixed patterns as in the cooperative scheme. Other BSs follow the N-policy to sleep and reactivate.

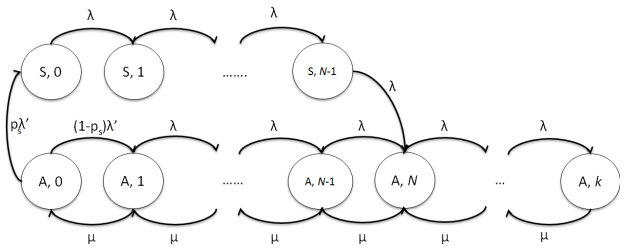

Fig. 1: A Markov Chain representation of N-policy sleeping scheme.

Note that we do not consider the start-up time of BSs when switching from sleep to active mode in this paper. For macro BSs in conventional cellular networks, the setup time is too long for the isolated schemes to be feasible [4], [18], while for femto BSs in 4G LTE networks, such start-up time is negligible as compared to the service time and delay requirement of users [4], [18].

B. Power consumption model

Power consumption of a BS can be generally divided into two parts, namely variable power consumption such as power amplifiers which depends on the traffic load carried by the BS, and static power consumption such as air conditioning and signal processing which is a constant amount as long as the BS is active [19], [20]. A sleeping BS consumes much less power as compared to an active one.

Following the discussions above, the power consumption of a BS is given by

$$P_{\mathrm{BS}} = \begin{cases} P_{\mathrm{static}} + \dfrac{\mathcal{A}}{C} P_v^{\max} & \text{when active,} \\ P_{\mathrm{sleep}} & \text{when sleeping,} \end{cases} \tag{1}$$

where P_{static} is static power consumption, $\mathcal{A}$ is traffic load of the BS, C is the capacity of the BS, and $P_v^{\max}$ is the variable power consumption if the BS is fully loaded.

IV. Analysis of QoS metrics

In this section, we derive the analytical expressions of QoS metrics including mean delay and blocking probability for each scheme.

A. Isolated scheme

We first derive the state probability equations of the isolated scheme. The probability that a BS will go to sleep is equal to the probability that no customer arrives during the close-down period t^* after the BS becomes idle. As the arrivals follow a Poisson process, the probability p_s is given by:

$$p_s = e^{-\lambda t^*}. \tag{2}$$

Due to the close-down period, the effective arrival rate λ' at state $(A,0)$ is different from λ. By definition, the system will leave state $(A,0)$ under either of the following two conditions: 1) the BS enters the sleep mode after the close-down period expires (transit to state $(S,0)$ with a probability of p_s); 2) a customer arrives before the close-down period expires (transit to state $(A,1)$ with a probability of $1-p_s$). Therefore, the mean time that the system spends at state $(A,0)$ is equal to

$$t' = (1-p_s)\frac{1}{\lambda} + p_s t^*. \tag{3}$$

The value of λ' can be then calculated by

$$\lambda' = \frac{1}{t'} = \frac{1}{(1-p_s)\frac{1}{\lambda} + p_s t^*}. \tag{4}$$

Based on (2) and (4), the transition probabilities from state $(A,0)$ to states $(S,0)$ and $(A,1)$ are $p_s\lambda'$ and $(1-p_s)\lambda'$, respectively. Transition probabilities between other states are intuitive based on the Markov Chain in Fig. 1. We denote the steady-state probability of state (M,n) as $\pi_{M,n}$, and set $\mathcal{A} = \lambda/\mu$ and $\mathcal{A}' = \lambda'/\mu$. All steady-state probabilities can be expressed in terms of $\pi_{A,0}$ by:

$$\pi_{S,0} = \frac{\lambda'}{\lambda} p_s \pi_{A,0}, \tag{5}$$

$$\pi_{S,i} = \pi_{S,0}, (1 \le i \le N-1), \tag{6}$$

$$\pi_{A,i} = \begin{cases} A'\pi_{A,0} & \text{for } i = 1, \\ A\pi_{A,N-1} + A' p_s \pi_{A,0} & \text{for } 1 < i \le N, \\ A\pi_{A,N-1} & \text{for } N < i \le k. \end{cases} \tag{7}$$

Combining (5), (6), (7) along with the normalization equation:

$$\sum_{i=0}^{k} \pi_{A,i} + \sum_{j=0}^{N-1} \pi_{S,j} = 1, \tag{8}$$

we can obtain all the steady-state probabilities $\pi_{M,n}$.

The mean queue size $E(Q)$ is given by:

$$E(Q) = \sum_{i=0}^{k} i\pi_{A,i} + \sum_{j=0}^{N-1} j\pi_{S,j}. \tag{9}$$

The blocking probability $E(B)$ is given by

$$E(B) = \pi_{A,k}. \tag{10}$$

By Little's law, the mean delay $E(D)$ is given by

$$E(D) = \frac{E(Q)}{\lambda(1 - E(B))}. \tag{11}$$

The proportion of time that the BS spends in sleep mode is given by

$$p_{\text{sleep}} = \frac{\sum_{j=0}^{N-1} \pi_{S,j}}{\sum_{i=0}^{k} \pi_{A,i} + \sum_{j=0}^{N-1} \pi_{S,j}}. \tag{12}$$

By (1) and (12), the average power consumption for a BS in the isolated scheme is

$$P^{\text{iso}} = p_{\text{sleep}} P_{\text{sleep}} + p_{\text{active}} P_{\text{active}}, \tag{13}$$

in which $p_{\text{active}} = 1 - p_{\text{sleep}}$, and $P_{\text{active}} = P_{\text{static}} + \frac{\mathcal{A} P_v^{max}}{C}$.

B. Cooperative scheme

For the cooperative and hybrid schemes, as we need to analyze a multi-BS system, obtaining exact analytical solutions for QoS metrics is computationally prohibitive due to the curse of dimensionality [21]. Therefore, we will make use of the IESA framework to obtain a reasonably accurate estimation of the QoS metrics in a computationally efficient manner.

We provide a brief description of the IESA framework here. More detailed explanation can be found in [8], [12], [13]. As mentioned before, IESA was proposed to evaluate blocking probability in systems where mutual overflow exists. The key idea of IESA is to apply traditional EFPA-based approximation to a surrogate model. This is because EFPA underestimates blocking probability in such system due to certain inherent assumptions. The surrogate system for IESA is specially designed to increase the validity of these assumptions. Therefore, when the EFPA-based approximation is applied to the surrogate model, these approximation errors due to these assumptions are reduced.

The surrogate model is formally described as follows. Each request (customer) has three attributes, namely the identity I, overflow record Δ, and estimation of congestion level Ω. I contains the "identity" information on the request which does not change during its service period, such as its origin and expected service time. Δ represents the set of BSs that has rejected admission of the request due to sleep or violation of delay requirement. Ω contains information on the number of overflows ever experienced by the request itself or other existing requests in the network and serves as an estimate of the level of congestion in the network.

Let Γ_i denote the set of BSs that a request originated from BS i is allowed to overflow. A new request has $\Delta = \emptyset$ and $\Omega = 0$. When request ζ originated from BS m with attributes $I_\zeta, \Delta_\zeta, \Omega_\zeta$ arrives at BS i (i and m can be the same), it will be admitted if its admission does not cause violation of the delay requirement for ongoing requests. Otherwise, if the most senior (highest Ω) request κ in service has $\Omega_\kappa < \Omega_\zeta$, the incoming request ζ will overflow to one of the BSs in $\Gamma_m - i$ and its attributes become $\{I_\zeta, \Delta_\zeta \cup i, \Omega_\zeta + 1\}$. However, if $\Omega_\kappa \geq \Omega_\zeta$, requests κ and ζ will exchange their third attribute, Ω, before request ζ's overflow. In this way, the overflow request will have attributes $\{I_\zeta, \Delta_\zeta \cup i, \Omega_\kappa + 1\}$ and the request in service will have $\{I_\kappa, \Delta_\kappa, \Omega_\zeta\}$.

By the information exchange mechanism, an overflow request retains its identity (I) and actual overflow record (Δ) while gathering network congestion information (Ω) from other customers. By definition, $|\Delta| \leq \Omega$ for any arriving request in the network.

The attributes and the information exchange process described previously are designed for a special mechanism that estimates the probability that all of the unattempted BSs are not available. The estimation is based on the values of Δ and Ω of an overflow request. If all of the unattempted BSs are presumed unavailable, the request will be blocked and cleared immediately without attempting the remaining BSs. Similar as in [8], [12], [13], we define $p_{k^*,|\Delta_\zeta|,\Omega_\zeta}$ as the probability that a request ζ with the attributes $\{I_\zeta, \Delta_\zeta, \Omega_\zeta\}$, gives up attempting in a surrogate model with parameter k^*. The parameter k^* is defined as the maximum allowable value of the attribute Ω of any request in the surrogate model and is a measure of the level of dependency in the real system. The value of k^* depends on the specific system. Denote $n_i = |\Gamma_i|$, then $p_{k^*,|\Delta_\zeta|,\Omega_\zeta}$ is evaluated as:

$$p_{k^*,|\Delta|,\Omega} = \begin{cases} 0 & \text{if } \Omega < n_i, \\ \dfrac{\binom{\Omega-|\Delta|}{n_i-|\Delta|}}{\binom{k^*-|\Delta|}{n_i-|\Delta|}} & \text{if } \Omega \geq n_i, \end{cases} \tag{14}$$

where $|\Delta| \leq n_i \leq k^*$.

We define $a_{i,j,n}$ as traffic offered to BS i with n overflows and $\Omega = j$, and $\mathcal{A}_{i,j}$ as total combined traffic offered to i with $\Omega \leq j$, the relationship between these two parameters is

$$\mathcal{A}_{i,j} = \sum_{l=0}^{j} \sum_{m=0}^{l} a_{i,l,m}. \tag{15}$$

By definition, we have $\mathcal{A}_{i,j} = \mathcal{A}_{i,j-1} + \sum_{n=0}^{min(j,n_m)} a_{i,j}$ for $j = 1, 2, ..., k^* - 1$ with initial values $\mathcal{A}_{i,0} = a_{i,0,0} = \mathcal{A}_i =$

λ_i/μ_i.

The surrogate is actually a hierarchical system based on the value of Ω. Blocking probability at a certain level is not affected by the traffic on higher levels (traffic with higher value of Ω). Therefore, if we denote $p_b^{PS}(\mathcal{A}, k)$ as the blocking probability of an M/M/1/k-PS queue (without vacations) with offered traffic $\mathcal{A}$, we can obtain the relationship between the blocking probability $B_{i,j}$ and $\mathcal{A}_{i,j}$ at each level j as

$$B_{i,j} = \begin{cases} p_b^{PS}(\mathcal{A}_{i,j}, k) & \text{if BS } i \text{ is active;} \\ 1 & \text{if BS } i \text{ is sleeping,} \end{cases} \tag{16}$$

where $0 \leq j \leq k^*$.

In normal circumstances, traffic originated from m and blocked at a BS is allowed to overflow to unattempted BSs in Γ_m. However, due to the giving up mechanism described by (14), a proportion of the overflow traffic will be dropped prematurely. The dropped traffic will not be included when calculating the blocking probability of the next level.

The traffic offered to the highest level of the system, namely level $k^* - 1$, is the total offered traffic as it includes all the levels below. Therefore, $\mathcal{A}_{i,k^*-1}(1 - B^I_{i,k^*-1})$ is the total carried traffic by BS i. The system blocking probability can thus be measured by 1 minus the ratio of carried traffic to the offered traffic. Thus we can derive the system blocking probability by IESA as:

$$\hat{B} = 1 - \frac{\sum_{i \in U} \mathcal{A}_{i,k^*-1}(1 - B_{i,k^*-1})}{\sum_{i \in U} \mathcal{A}_i}, \tag{17}$$

where U is the set of all BSs in the system.

Denote $\lambda_i = \mathcal{A}_{i,k^*}/\mu$. Referring back to our previous analysis for the isolated scheme, by replacing λ with λ_i and setting $t^* = \infty$ (as the BSs will not enter sleep mode due to N-policy), we can obtain the approximated mean delay following the same analysis as in (2) to (11).

The power consumption of an active BS i in the cooperative scheme is given by

$$P_i^{\text{coop}} = P_{\text{static}} + \frac{\mathcal{A}_i P_v^{\max}}{C_i}. \tag{18}$$

C. Hybrid scheme

The hybrid scheme is the joint application of the isolated and cooperative schemes. To obtain the QoS metrics for the hybrid scheme, we can mostly follow the analysis of the cooperative scheme in the Section IV-B. However, when calculating the blocking probability of traffic at each level in a single BS as in Equation (16), we should replace the term $p_b^{PS}(\mathcal{A}_{i,j}, k)$ in (16) by $E(B)$ in (10) following the analysis in Section IV-A ((2) to (10)), as the state probabilities for a BS selected to be active based on the switching patterns in the hybrid scheme follow an M/M/1/k-PS queue with vacations and N-policy. The power consumption of such a BS is given by

$$P_i^{\text{hyb}} = p_{\text{active}} \left(P_{\text{static}} + \frac{\mathcal{A}_i P_v^{\max}}{C_i} \right) + p_{\text{sleep}} P_{\text{sleep}}. \tag{19}$$

V. Numerical results

In this section, we first demonstrate the accuracy of our proposed approximation for cooperative and hybrid schemes by comparing simulation and approximation results. Then, we demonstrate the power consumption and QoS tradeoffs of each scheme in a wide range of scenarios.

We have collected power consumption and traffic data from a real BS site in Hong Kong, as shown in Fig. 2. Static power consumption is about 1867.6W and the maximum power consumption is about 2150W. As we are limited by the information that we can publish, we assume that the site is composed of 7 identical BSs (such that $P_{\text{static}} \approx 266.8$W and $P_{\text{v}}^{\max} \approx 40.43$W for each BS) and one BS is switched to sleep for the cooperative and hybrid schemes (assume $P_{\text{sleep}} \approx 10$W). Traffic offered to each BS is the same with mean arrival rate $\lambda = 0.8$ arrivals/s (if not specified otherwise) and mean service time $1/\mu = 1$s. The parameter k^* for IESA is set to 12.

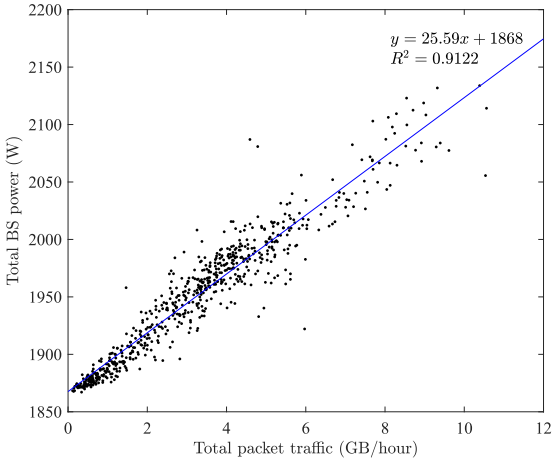

Fig. 2: Power consumption of a real BS site.

A. Accuracy and computational efficiency of approximations

Figs. 3 and 4 show the analytical and simulation results for blocking probability and mean delay for all three schemes. Parameters related to the N-policy are set as $N = 3, t^* = 3$ for the isolated and hybrid scheme. From the results we can validate that our analytical results for blocking probability and mean delay in the isolated scheme are very accurate. Meanwhile, the approximations for mean delay and blocking probability in both cooperative and hybrid schemes are also quite close to simulation results. Specifically, for cooperative and hybrid schemes, we choose the traffic load corresponding to the blocking probability range $10^{-3} - 10^{-2}$, which is considered practical for cellular networks and of particular interest of existing research (e.g. [10]). The results show that the estimation errors of blocking probability and mean delay in both schemes are less than 20%.

In terms of computational efficiency, the running time of analytical approximation (about 0.2 second) is about five orders

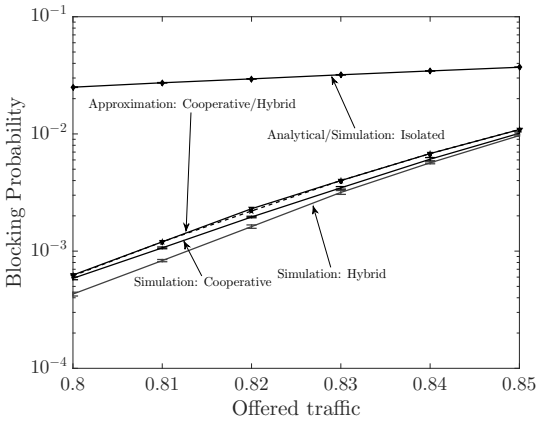

Fig. 3: Accuracy of approximations for blocking probability.

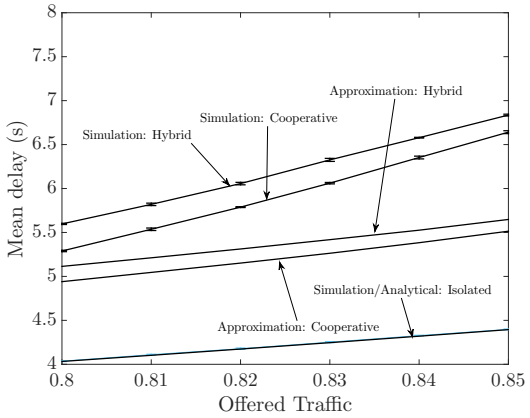

Fig. 4: Accuracy of approximations for mean delay.

of magnitude lower than that of simulation (about 3 hours). Given its reasonable accuracy, our proposed approximation can be used for searching optimal solutions to tradeoff between power consumption and QoS in cellular networks.

B. Power consumption and QoS tradeoff

We now investigate the tradeoff between power consumption and QoS metrics including mean delay and blocking probability.

In Fig. 5, we set $k = 10$ to investigate the power-blocking tradeoff under the same maximum allowable mean delay constraint. We change the parameter N for the isolated and hybrid schemes to obtain different values for power consumption and blocking probability. Note that power consumption for the cooperative and hybrid schemes is defined as the average power consumption per BS. As we discussed previously, N is the number of arrivals that must be accumulated for a sleeping BS to re-activate. Therefore, when N increases, power consumption decreases and blocking probability increases, as each BS spends more time in sleep mode. The results also show that the isolated scheme is the most flexible scheme in trading off blocking probability for power consumption among all three schemes. However, the cooperative scheme can achieve a lower blocking probability for the same amount of power consumption than the isolated scheme, and the hybrid scheme has the potential to further reduce power consumption while attaining similar blocking probability as the cooperative scheme.

In Fig. 6, we adjust the value of k to guarantee that the blocking probability is less than 1% and obtain corresponding values of mean delay and power consumption. The results show that the cooperative and hybrid schemes can attain a shorter mean delay as compared to the isolated scheme, while maintaining the same level of power consumption. Meanwhile, similar to the previous figure, by adjusting the parameter N, the isolated and hybrid schemes can reduce power consumption at the expense of relatively higher delay.

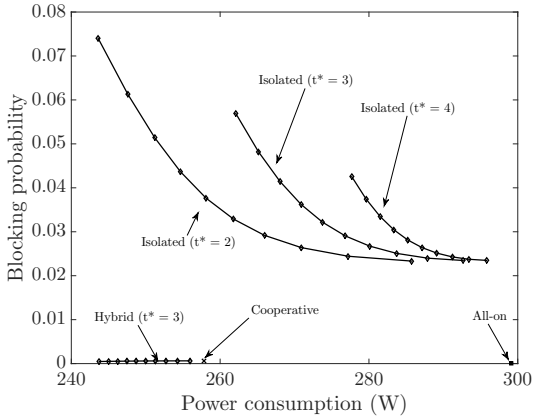

Fig. 5: Tradeoff between power consumption and blocking probability subject to $k = 10$ (constraint on maximum allowable delay).

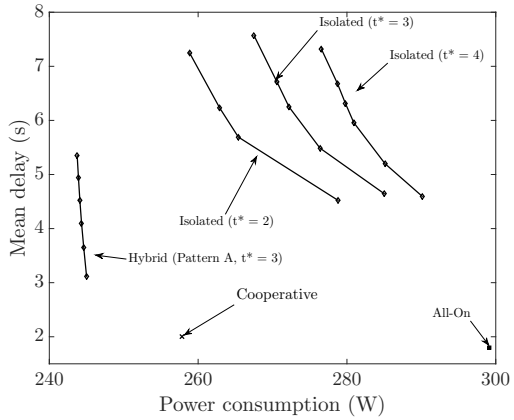

Fig. 6: Tradeoff between power consumption and mean delay subject to blocking probability $< 1\%$.

Fig. 7 shows the tradeoff between mean delay and blocking probability for isolated and cooperative schemes in a more intuitive way. By adjusting the parameter k while keeping all other parameters constant, we set different values for

maximum allowable delay of a customer. As we expected, when the maximum delay requirement is more strict, requests are more likely to be dropped as they fail to satisfy the requirement for every scheme. Also, as power consumptions for all schemes are in a narrow range $(240 - 260\text{W})$, we can conclude that the cooperative and hybrid schemes can achieve a better delay-blocking tradeoff as compared to the isolated scheme for a similar amount of power consumption.

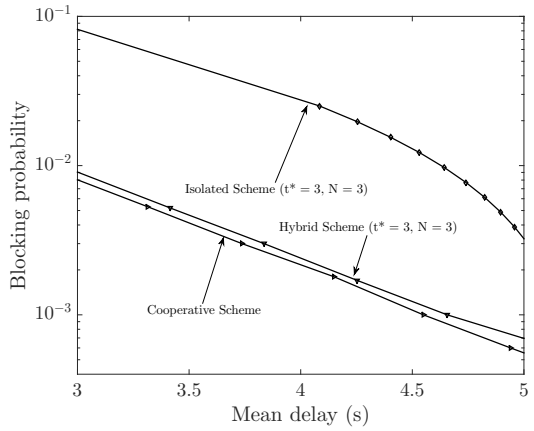

Fig. 7: Tradeoff between blocking probability and mean delay.

VI. Conclusions

We have evaluated the performance of three BS sleeping schemes in terms of the tradeoff by modelling each BS as an M/M/1/k-PS queue with vacations. We have also provided accurate, robust, scalable and computationally efficient analytical means to evaluate QoS and power consumption in cellular networks with BS sleeping. Then, numerical results have shown that the cooperative and hybrid schemes can achieve a better tradeoff as compared to the isolated scheme.

Our new analytical results are useful for network design and optimization applications when there is a need to search for optimal solutions. Such a search involves a prohibitively large number of calculations of mean delay, blocking probability and power consumption under a wide range of conditions and scenarios, and our analytical methods can significantly reduce the computation time.

In the future, we plan to extend the work to more general scenarios, including but not limited to networks with multi-layer heterogeneous cells, and networks with asymmetrical offered traffic to each BS.

Acknowledgement

The work described in this paper was partly supported by College Research Grant from BNU-HKBU United International College [UIC-R201703] and a grant from the Innovation and Technology Funding (ITF) of the Hong Kong Special Administrative Region, China [ITS/191/16].

References

[1] M. Ismail, W. Zhuang, E. Serpedin, and K. Qaraqe, "A survey on green mobile networking: From the perspectives of network operators and mobile users," *IEEE Communications Surveys Tutorials*, vol. 17, no. 3, pp. 1535–1556, Third quarter 2015.

[2] J. Wu, Y. Zhang, M. Zukerman, and E. K. N. Yung, "Energy-efficient base-stations sleep-mode techniques in green cellular networks: A survey," *IEEE Communications Surveys Tutorials*, vol. 17, no. 2, pp. 803–826, Second quarter 2015.

[3] 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 Transactions on Wireless Communications*, vol. 12, no. 8, pp. 4196–4209, Aug. 2013.

[4] Y. H. Chiang and W. Liao, "Genie: An optimal green policy for energy saving and traffic offloading in heterogeneous cellular networks," in *Proc. 2013 IEEE International Conference on Communications (ICC)*, Jun. 2013, pp. 6230–6234.

[5] J. Huang, L. Xu, M. Zeng, H. Yan, Q. Duan, and C.-C. Xing, "Hybrid scheduling for quality of service guarantee of multimedia data flows in software defined networks," in *Proceedings of the 8th International Conference on Mobile Multimedia Communications*, ICST, Brussels, Belgium, 2015, pp. 110–116.

[6] Y. Xu, S. E. Elayoubi, E. Altman, and R. El-Azouzi, "Impact of flow-level dynamics on QoE of video streaming in wireless networks," in *Proc. IEEE INFOCOM 2013*, Apr. 2013, pp. 2715–2723.

[7] L. Saker, S. E. Elayoubi, R. Combes, and T. Chahed, "Optimal control of wake up mechanisms of femtocells in heterogeneous networks," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 3, pp. 664–672, Apr. 2012.

[8] Y.-C. Chan, J. Guo, E. W. M. Wong, and M. Zukerman, "Surrogate models for performance evaluation of multi-skill multi-layer overflow loss systems," *Performance Evaluation*, vol. 104, pp. 1–22, Oct. 2016.

[9] W. Song and W. Zhuang, "Multi-service load sharing for resource management in the cellular/WLAN integrated network," *IEEE Transactions on Wireless Communications*, vol. 8, no. 2, pp. 725–735, Feb. 2009.

[10] F. Han, Z. Safar, and K. J. R. Liu, "Energy-efficient base-station cooperative operation with guaranteed QoS," *IEEE Transactions on Communications*, vol. 61, no. 8, pp. 3505–3517, Aug. 2013.

[11] H. Tabassum, U. Siddique, E. Hossain, and M. J. Hossain, "Downlink performance of cellular systems with base station sleeping, user association, and scheduling," *IEEE Transactions on Wireless Communications*, vol. 13, no. 10, pp. 5752–5767, Oct. 2014.

[12] E. W. M. Wong, J. Guo, B. Moran, and M. Zukerman, "Information exchange surrogates for approximation of blocking probabilities in overflow loss systems," in *Proc. The 25th International Teletraffic Congress (ITC)*, Sep. 2013.

[13] J. Wu, E. W. M. Wong, J. Guo, and M. Zukerman, "Performance analysis of green cellular networks with selective base-station sleeping," *Performance Evaluaiton*, vol. 111, pp. 17–36, May 2017.

[14] X. Guo, Z. Niu, S. Zhou, and P. R. Kumar, "Delay-constrained energy-optimal base station sleeping control," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 5, pp. 1073–1085, May 2016.

[15] F. Kelly, "Blocking probabilities in large circuit-switched networks," *Advances in Applied Probability*, vol. 18, pp. 473–505, 1986.

[16] R. B. Cooper and S. Katz, "Analysis of alternate routing networks with account taken of nonrandomness of overflow traffic," Technical Report, Bell Telephone Lab. Memo, 1964.

[17] A. Girard, *Routing and Dimensioning in Circuit-Switched Networks*. Addison-Wesley, 1980.

[18] E. Ternon, P. Agyapong, L. Hu, and A. Dekorsy, "Database-aided energy savings in next generation dual connectivity heterogeneous networks," in *Proc. 2014 IEEE Wireless Communications and Networking Conference (WCNC)*, Apr. 2014, pp. 2811–2816.

[19] H. Holtkamp, G. Auer, S. Bazzi, and H. Haas, "Minimizing base station power consumption," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 2, pp. 297–306, Feb. 2014.

[20] A. Bousia, A. Antonopoulos, L. Alonso, and C. Verikoukis, "Green distance-aware base station sleeping algorithm in LTE-advanced," in *Proc. 2012 IEEE International Conference on Communications (ICC)*, Jun. 2012, pp. 1347–1351.

[21] R. C. McNamara, "Applications of spanning trees to continuous-time Markov processes, with emphasis on loss systems," Ph.D. dissertation, University of Colorado, 2004.