# Power Consumption and GoS Tradeoff in Cellular Mobile Networks with Base Station Sleeping and Related Performance Studies

Jingjin Wu, *Member, IEEE*, Eric W. M. Wong, *Senior Member, IEEE*, Yin-Chi Chan, *Member, IEEE* and Moshe Zukerman, *Life Fellow, IEEE*

***Abstract*—Mobile network operators usually consider power consumption and Grade of Service (GoS) as two important aspects in the design and planning of modern cellular networks. Base station (BS) sleeping is an effective approach to reduce the power consumption of the network, by switching some of the BSs to a low-power "sleep mode" during off-peak traffic hours. In this paper, we model each BS with sleeping mechanism as an M/G/1/$K$ queue with vacations, and the entire cellular network as a network of such queues, to incorporate practical factors in BS sleeping, such as close-down and startup periods and additional power consumption for activating a sleeping BS. We investigate the power consumption and GoS under three BS sleeping schemes: (1) the isolated scheme, in which each BS switches between active and sleep modes based on its own real-time traffic load, (2) the cooperative scheme, in which selective BSs are switched to long-term sleep and traffic is allowed to overflow from sleeping BSs to nearby active BSs, and (3) the hybrid scheme, in which some BSs are switched to long-term sleep and other BSs switch modes according to their real-time traffic load. A robust, scalable and computationally efficient analytical method is proposed to evaluate GoS metrics, including mean delay and blocking probability, and power consumption under each scheme. We validate the accuracy of the proposed method, demonstrate the trade-off among power consumption, blocking probability and mean delay, and compare the performance of the three schemes via extensive and statistically reliable numerical experiments.**

***Index Terms*—Base station sleeping, performance analysis, teletraffic model, power-performance tradeoff**

## I. Introduction

Recently, base station (BS) sleeping has emerged as an effective approach to reduce power consumption in cellular mobile networks [1]. Energy saving is achieved by switching BSs (or certain components of them) to a low power-consuming mode called "sleep mode" during non-busy hours when traffic in the network is relatively low. As BSs consume up to 80% of energy in cellular networks, BS sleeping may reduce a considerable amount of power consumption [2].

BS sleeping belongs to a broad family of approaches aiming at improving the energy efficiency of cellular networks by adjusting the transmitting power of BSs [3]. A BS selected to sleep reduces its transmit power to zero while neighboring active BSs increase their transmit power to maintain coverage. Compared to other power-saving approaches for green cellular networks, including applying renewable energy solutions or upgrading hardware components, BS sleeping can be implemented in existing network infrastructure and is thus considered more cost-effective [2]. On the other hand, switching some BSs to sleep mode leads to a reduction in the network capacity. Therefore, network operators must accurately evaluate the Grade of Service (GoS) metrics and investigate the impact of different BS sleeping strategies on the GoS [2], [4]. In this paper, we provide new methods to evaluate GoS measures such as blocking probability and mean delay. Such methods can be used to obtain accurate numerical values for such measures under various BS sleeping schemes that help us assess the trade-off between power consumption and GoS metrics.

In particular, we consider a cellular network, where each BS is modeled as a single-server queue, fed by arrivals that follow a Poisson process, with a finite buffer size of $K$ and generally distributed service times. This queueing model is known as the M/G/1/$K$ queue. The assumption of Poisson arrivals has been applied for modeling the *Busy Hour Traffic*, which refers to network traffic load during the busiest hour, in existing research on teletraffic models [5]. We will demonstrate that our proposed method can still obtain accurate evaluations when this assumption is relaxed in Section V. The generally distributed service time addresses various factors that may affect the service time of a user request in a cellular network, such as the application type, the amount of data to transmit and the channel condition. We refer to the parameter $K$, which represents the maximum number of requests that a BS can serve concurrently, as the capacity of the BS. We further consider vacations with startup and close-down times in the queue to model the operation of BS sleeping [6]. Henceforth, we will use the notation M/G/1/$K$ queue to represent the general case of this queue with or without a range of modeling extensions, including vacations, startup and close-down times. In cases where reference to a specific case is important for clarity, we will specify the particular modeling extension that

J. Wu is with the Department of Statistics, BNU-HKBU United International College, Zhuhai, Guangdong, China, and also with the Department of Electrical Engineering, City University of Hong Kong, Hong Kong SAR, China (Email:jj.wu@ieee.org).

E. W. M. Wong, Y. C. Chan and M. Zukerman are with the Department of Electrical Engineering, City University of Hong Kong, Hong Kong SAR, China (Email: eeewong@cityu.edu.hk, ycchan26@cityu.edu.hk, m.zu@cityu.edu.hk).

A preliminary version of this paper was presented at IEEE GLOBECOM 2017, and was published in its Proceedings.

The work described in this paper was partly supported by College Research Grant from BNU-HKBU United International College [UIC-R201911] and a grant from the Innovation and Technology Fund (ITF) of the Hong Kong Special Administrative Region, China [ITS/191/16].

we consider.

Another key justification of the application of the M/G/1/$K$ queue as the model of each BS is that many popular multimedia mobile services are highly delay-sensitive. Accordingly, a minimum data rate needs to be guaranteed for traffic generated by such services [7]. To ensure that such delay and data rate requirements are satisfied for all admitted user requests, an upper limit should be set on the number of requests allowed to be served by each BS concurrently. For this case, it is important to evaluate accurately both the blocking probability and the mean delay to be able to avoid violation of GoS requirements using the finite-buffer model. In this paper, we define the GoS metric "mean delay" as the average time spent in the network by an admitted user task from the moment when the connection between the user and a BS is established to the moment when the requested service is completed. Therefore, the delay of a task is composed of the service time required by the task, and the waiting time that the task spends in the queue.

There are two broad types of BS sleeping schemes. One is to implement the sleep mode separately and independently in each BS (*isolated scheme*), where a BS is switched to the sleep mode when it experiences an idle period of a certain length [8], [9]. The other is to consider the long-term traffic among multiple nearby BSs (*cooperative scheme*) [10]–[12]. In the latter case, traffic in the service areas of sleeping BSs is distributed to nearby active BSs through cell breathing or beamforming techniques [10], and the network of BSs can be considered as a network of queues. The selection of BSs to sleep is made either according to fixed switching patterns derived from historical traffic analysis and prediction [11], [12], or dynamically based on the real-time impact of each BS to the network [10].

Standard approaches for evaluation of network performance measures such as delay and blocking probability based on simulations or numerical brute force solutions of Markov chains are not scalable to realistically sized networks, especially in cooperative schemes where the traffic loads in different BSs are dependent. Therefore, for applications such as network dimensioning, where computational scalability is key for obtaining high-quality optimal solutions, it is preferable to evaluate GoS metrics by numerical approximations of acceptable accuracy.

In this paper, we propose analytical approximation methods based on teletraffic theory and the recently established Information Exchange Surrogate Approximation (IESA) framework [13]–[16] to obtain the mean delay and blocking probability in a cellular network, and analyse the trade-off between power consumption and GoS under the isolated, cooperative and *hybrid* (a joint application of isolated and cooperative) BS sleeping schemes. This paper is a significant extension of its conference version [16], where simplified assumptions of exponentially distributed service times and deterministic close-down times were made. Also, BS startup times and the additional power consumption for switching on a sleeping BS were not considered in [16]. The main contributions of this paper are summarized as follows:

- We model a cellular network with sleeping mechanism as a network of M/G/1/$K$ queues with vacations, where each BS is considered as a single-server queue. The finite-buffer queuing model addresses the importance of blocking probability due to violation of the data rate requirement for requests generated by multimedia mobile applications. We also consider BS close-down and startup times, which are consistent with real BS operations. To the best of our knowledge, this is the first work to evaluate mean delay, blocking probability, and power consumption of a cellular network at the same time. In addition, our model can be applied to a wide variety of cellular network standards, including 4G and 5G networks, as the M/G/1/$K$ queue has been demonstrated to be suitable for modelling base stations in these networks [8], [17].
- We propose the hybrid BS sleeping scheme which makes decisions based on both real time traffic load at each BS and long-term traffic trends, with the objective of further improving energy saving as compared to existing isolated and cooperative schemes in a multi-BS cellular network.
- We propose new robust, accurate, scalable and computationally efficient analytical approximation methods to evaluate GoS metrics, including blocking probability and mean delay, in cellular networks with isolated, cooperative and hybrid BS sleeping schemes.
- We numerically verify that our approximation results of mean delay and blocking probability are accurate under different network conditions, in Section V. In addition, we numerically demonstrate that our method also provides reasonably accurate estimates for networks with more bursty arrivals, which can be modelled as a Markov Modulated Poisson Process (MMPP).
- We apply the proposed methods to compare the performance of isolated, cooperative and hybrid BS sleeping schemes under different network conditions, in terms of the tradeoff among power consumption, blocking probability, and mean delay.

The remainder of this paper is organized as follows: Section II reviews recent research on different BS sleeping schemes, as well as existing methods to evaluate GoS in cellular networks. Section III introduces three BS sleeping schemes and power consumption models. Section IV describes our proposed methods to evaluate blocking probability, mean delay and power consumption under each BS sleeping scheme in detail. We compare the analytical and simulation results to verify the accuracy and robustness of our proposed methods, and demonstrate the tradeoff among power consumption, blocking probability, and mean delay in Section V. Finally, in Section VI, we present concluding remarks.

## II. Related work

The single-server queuing model has been widely used to model traffic variations through a single access point, such as a cellular BS [18]. A particularly popular queuing model for BS sleeping is the M/G/1 queue with vacations, which has been applied to obtain closed-form expressions of power consumption and delay under different variants of isolated BS sleeping schemes [8], calculate the system parameters to attain the optimal tradeoff between power consumption and delay under a joint BS sleeping and power matching scheme [9], or identify the optimal sleeping policy by formulating the problem as a partially observable Markov Chain [19]. Notably, in [8] and [9], the authors stated that the $N$-policy sleeping scheme has better performance in terms of the power-delay tradeoff than other isolated schemes.

A major limitation of the single BS model is that it does not address the impact of cell breathing and beamforming techniques, which enable active BSs to increase their transmission power and extend coverage to serve users whose closest BS has switched to sleep. Tabassum *et al.* [12] proposed a dynamic user association scheme where users arriving at a sleeping BS are reassigned to the active BS with the greatest mean channel access probability. The authors demonstrated that the proposed scheme can improve spectral efficiency and minimize outage probability. Multi-cell cooperative sleeping algorithms in heterogeneous networks have also been studied to improve grid energy savings where hybrid energy sources are available [20], or reduce the power consumption while guaranteeing the minimum BS coverage requirement [21]. Kong *et al.* [22] modelled a heterogeneous network as multiple interacting queues to demonstrate a potential tradeoff between delay and signal-to-interference ratio. Renga *et al.* [23] investigated the integration of demand-based cooperative BS sleeping with harvesting of renewable energy through a Markovian model, and demonstrated that large potential savings can be achieved by the integrated techniques.

One important issue that has been largely ignored in existing research is the dependency of the traffic loads at nearby BSs, which may be caused by the movement of mobile users or resource sharing among nearby BSs. Kelly [24] suggested that cellular networks with *channel borrowing capabilities*, a classical resource sharing mode in first-generation mobile networks, can be considered as *overflow loss systems*. In such systems, a user request can *overflow* to an alternative server (BS) if the first server it attempts is unavailable, and will be blocked only if all alternative servers are unavailable. Dynamic user association schemes in modern cellular networks resemble channel borrowing in some way, as the reassociation of users from a sleeping BS to an active BS is analogous to the sleeping BS "borrowing" a channel from the active BS.

The Erlang Fixed-Point Approximation (EFPA) [25] has been the classical approximation method of choice for evaluating blocking probability in overflow loss systems. The key idea of EFPA is to decompose the system into independent Erlang B subsystems to reduce the computational complexity. However, EFPA is known to be very inaccurate for systems with *mutual overflow* effects [13], [26]. The IESA framework, which has its roots in the EFPA but applies the decomposition-based approach on a surrogate system rather than the original system, was proposed in order to improve the accuracy of the approximation [13]. The framework was further improved by integrating with other approximation techniques such as moment matching [14].

Previously, we have considered a cellular network model with dynamic user association capabilities and cooperative BS sleeping with fixed switching patterns [15]. By decomposing the network into independent BSs loaded with Poisson traffic, we verified that IESA had significant improvement over EFPA on the accuracy of estimating the blocking probability. In the conference version of this paper [16], we have demonstrated the accuracy of IESA in estimating blocking probability and mean delay in a delay-loss system with exponential service time distribution. This paper is based on the more realistic and general model for each BS of a mobile cellular network, namely, an M/G/1/$K$ queue with vacations, close-down and startup times, and measures GoS and power consumption under isolated, cooperative and hybrid BS sleeping schemes in cellular networks.

## III. Network model

### A. BS sleeping schemes

For tractability, we consider each BS as a finite-buffer single-server queue, with vacations. New user requests arrive at each BS according to a Poisson process, with rate $\lambda$. Each user request has a service time requirement of $R$ which is i.i.d. (independently and identically distributed) and follows a general distribution. The service time is determined by the amount of data required to be transmitted by the user request, and the transmission rate offered by the wireless connection. Note that the assumption of i.i.d. service time distribution of various tasks does not rule out long range dependence (LRD) in the aggregate traffic, which is commonly observed in traffic generated by multimedia mobile applications. LRD may result if the variance of these time durations is very high (e.g. if they are Pareto distributed with certain parameter values) [27]. Unless otherwise specified, we assume that each BS is identical in operations with the First-Come First-Served (FCFS) service discipline. We will derive the blocking probability, mean delay and power consumption under this assumption.

When a BS has $K$ requests in service, it will not admit any more incoming requests until one of the in-service requests completes its service and leaves the BS. The value of $K$, referred as the capacity of the BS, can be adjusted for the requirements of specific applications. When $K$ increases, more requests may be admitted simultaneously and thus admitted requests perceive a longer mean waiting time under high traffic scenarios. The finite capacity of $K$ addresses the latency (delay) requirement of many tasks in modern and future cellular networks. If an incoming user request sees too many (i.e. $K$) other requests queuing for service at a BS upon its arrival, which indicates a long waiting time that will violate its latency requirement based on the FCFS discipline, it will leave the congested BS to attempt other BSs. If all BSs covering the user's location are not available, the request will not be served, which can be regarded as a cancellation of the request due to violation of the latency requirement.

We describe in detail the three BS sleeping schemes considered in this paper as follows.

- *Isolated ($N$-policy) scheme*: as described in [8], [9], a BS will enter a close-down period of $C$ at the departure of its last serving request, and switch to sleep if no new requests arrive during this period. We will refer this kind of sleep mode as *short-term sleep* in the rest of the paper.

This scheme is appropriate for femto BSs which allow frequent mode changes. A sleeping BS will reactivate when $N$ or more users have arrived since the beginning of the last sleep period. A BS reactivating from sleep needs a startup period of $S$ to warm up before starting to serve users. Under this scheme, each BS can be modelled as an independent M/G/1/$K$ queue with vacations, close-down times, and startup times.

- *Cooperative scheme*: as described in [12], [15], selective BSs in the network switch to sleep during low-traffic hours. We will refer this kind of sleep mode as *long-term sleep* in the rest of the paper. The selection of sleeping BSs and the duration of the sleep periods can be determined by traffic analysis and prediction techniques [28]. This scheme is appropriate for traditional macro BSs which only allow one or two switches per day. The startup time and close-down time can be ignored in the cooperative scheme, as switching of modes is infrequent. Under this scheme, active BSs can extend coverage to serve users originally covered by BSs switched to sleep or with no idle capacity, resulting in dependency among states of different BSs. We will apply the decomposition-based IESA approach to evaluate blocking probability, mean delay and power consumption as exact analytical evaluations are computationally prohibitive.
- *Hybrid scheme*: as described in [16], after selected BSs are switched to long-term sleep as in the cooperative scheme, the remaining BSs are allowed to enter short-term sleep based on the $N$-policy. This scheme is appropriate for heterogeneous cellular networks, where marco BSs (following the cooperative scheme for long-term sleep) and femto BSs (following the $N$-policy for short-term sleep) co-exist. As in the cooperative scheme, dependency among BSs exists as the traffic load has to be redistributed to active BSs. Meanwhile, the startup times and close-down times of BSs following the $N$-policy need to be taken into account.

To clarify, in this paper we use "BS sleeping" to mean that some of the radio transceivers in BSs are placed in a halt state in which the power consumption of these transceivers are reduced. This is consistent with our experiments on a real BS site in Hong Kong, which we will describe in more detail in Section V. In addition, there is minimal re-association or migration cost for users in service under the three schemes considered in this paper. Under the isolated scheme (or the isolated component of the hybrid scheme), the BS enters sleep mode only after a close-down period in which no user is served by the BS. For the cooperative scheme (or the cooperative component of the hybrid scheme), switching of modes is infrequent and thus the re-association or migration effect is negligible.

### B. Power consumption

We consider three major sources of power consumption in a BS. The first one is transmission power, mostly consumed by power amplifiers when the BS is transmitting to/from mobile users. The amount of transmission power consumed in a BS is dependent on its carried traffic. The second source is static power for air conditioning and signal processing, which does not change with carried traffic. An idle active BS still consumes a considerable amount of static power, while a sleeping BS consumes much less [29], [30]. The final source of power consumption is the extra power needed each time when the BS is reactivated from sleep (referred to as "switching cost" in some literature).

We denote $P_{\text{SL}}, P_{\text{ST}}$, and $P_{\text{ID}}$ as the power consumption of a BS in the sleeping, startup, and idle (active but serving no users) phases, respectively. For simplicity, we assume that a sleeping BS has a power consumption of $P_{\text{SL}}$ regardless of whether it is in short-term or long-term sleep. The value of $P_{\text{SL}}$ accounts for the extra power consumption required for operations supporting BS sleeping, such as monitoring the number of requests accumulated for BSs in short-term sleep and coverage extension by neighboring active BSs for BSs in long-term sleep to resolve the hidden node problem. Beamforming can be applied to mitigate the interference and thus maintain a certain level of signal-to-noise-plus-interference-ratio (SINR) for reliable transmissions. As discussed in the previous paragraph, we have $P_{\text{SL}} < P_{\text{ID}}$. Also, we assume that $P_{\text{ID}} < P_{\text{ST}}$ to account for the switching cost during a startup. For an active BS, we consider a linear power consumption model where the instantaneous power consumption is $P_{\text{A}} = P_{\text{ID}} + \frac{A}{K} P_{\text{TR}}$, where $A$ is the instantaneous carried traffic of the BS, $K$ is the maximum number of requests that a BS can serve at the same time, and $P_{\text{TR}}$ is the maximum transmission power. Note that for a BS during the close-down period, the power consumption is still $P_{\text{I}}$ as the BS remains active with no users in service.

## IV. Analysis of Power Consumption and GoS Metrics

We now introduce analytical methods to evaluate power consumption, mean delay, and blocking probability for each sleeping scheme. For a random variable $Y$, we let $Y(x)$ be its cumulative distribution function (CDF), and $\mathbb{E}[Y]$ be its expectation.

### A. Power consumption

As the power consumption of a BS depends on its mode of operation, it is reasonable to consider the long-term average power consumption. The phases of a BS under the isolated scheme is a *regenerative process* that has a *regeneration cycle* which consists of the active, close-down, sleep, and startup phases in sequence, as shown in Fig. 1.

The lengths of these periods have been shown to be statistically the same in different regeneration cycles [31]. Therefore, if we denote by $L_{\text{A}}$, $L_{\text{CD}}$, $L_{\text{SL}}$ and $L_{\text{ST}}$ as the lengths of active, close-down, sleep, and startup phases in a single regeneration cycle respectively, then the average power

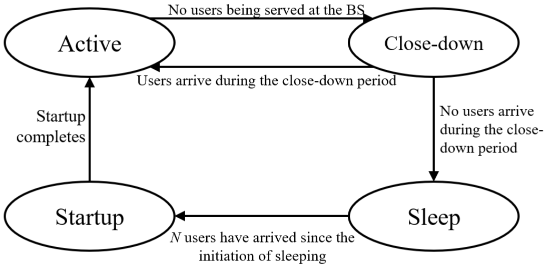


Fig. 1. The regeneration cycle of BS phases under the isolated scheme.

consumption of a BS under the isolated scheme is the weighted average of power consumption in the different phases, namely,

$$\begin{aligned}&\mathbb{E}[P^{\mathrm{I}}] = \\ &(\mathbb{E}[L_{\mathrm{A}}]\mathbb{E}[P_{\mathrm{A}}] + \mathbb{E}[L_{CD}]P_{\mathrm{ID}} + \mathbb{E}[L_{\mathrm{SL}}]P_{\mathrm{SL}} + \mathbb{E}[L_{\mathrm{ST}}]P_{\mathrm{ST}})\, / \\ &(\mathbb{E}[L_{\mathrm{A}}] + \mathbb{E}[L_{\mathrm{CD}}] + \mathbb{E}[L_{\mathrm{SL}}] + \mathbb{E}[L_{\mathrm{ST}}])\,,\end{aligned} \tag{1}$$

where $\mathbb{E}[P_{\mathrm{A}}] = P_{\mathrm{ID}} + \frac{\mathbb{E}[A]}{K}P_{\mathrm{TR}}$ and $\mathbb{E}[A]$ is the average carried traffic of the BS. Note that in (1), $\mathbb{E}[L_{\mathrm{A}}]\mathbb{E}[P_{\mathrm{A}}] = \mathbb{E}[L_{\mathrm{A}}P_{\mathrm{A}}]$ as $L_{\mathrm{A}}$ and $P_{\mathrm{A}}$ are uncorrelated, as $L_A$ depends on the specific BS sleeping scheme adopted and corresponding parameters (such as $N$ and $\mathbb{E}[C]$) under the selected scheme, while $P_A$ only depends on the specifications of hardware components in BSs, e.g., power efficiency of power amplifiers.

Under the cooperative scheme, switches of BS states are relatively infrequent. If we focus on the power consumption (and saving) during low-traffic periods when some of the BSs are in sleep, the power consumption of a BS depends only on the binary states of active or long-term sleep, that is,

$$\mathbb{E}[P^{\mathrm{C}}] = \begin{cases} P_{\mathrm{SL}} & \text{if the BS is selected to sleep,} \\ \mathbb{E}[P_{\mathrm{A}}] & \text{if the BS remains active.} \end{cases} \tag{2}$$

Under the hybrid scheme, the BSs that are selected to long-term sleep will always have a power consumption of $P_{SL}$, while those following the $N$-policy will have the same power consumption as in (1). That is,

$$\mathbb{E}[P^{\mathrm{H}}] = \begin{cases} P_{\mathrm{SL}} & \text{if the BS is selected to} \\ & \text{long-term sleep,} \\ \mathbb{E}[P^{\mathrm{I}}] & \text{if the BS is selected to} \\ & \text{follow the } N\text{-policy.} \end{cases} \tag{3}$$

For all three schemes, the power consumption of the network is the sum of power consumptions of all BSs.

### B. Isolated scheme

We now derive the state probability equations for a BS under the isolated scheme. We begin by denoting $R$ as the service time of a user, and $C$ and $S$ as the close-down time and startup time of a BS, respectively.

An active BS will enter short-term sleep if no user arrives for a period of $C$ after the BS becomes idle. Under the assumption of Poisson arrivals, the probability $p_s$ of no user arrivals for a period of $C$ is given by:

$$p_s = \mathbb{E}[e^{-\lambda C}]. \tag{4}$$

We then derive the expectation of $L_{\mathrm{A}}$, $L_{\mathrm{CD}}$, $L_{\mathrm{SL}}$ and $L_{\mathrm{ST}}$ for a BS under the isolated scheme as in (1). The average length of a close-down period is $\mathbb{E}[C]$ if no users arrive during the period (with a probability of $p_s$), after which the BS enters short-term sleep, and $1/\lambda$ if there is any user arrival before the period ends (with a probability of $1 - p_s$), after which the BS immediately returns to the active phase. Therefore,

$$\mathbb{E}[L_{CD}] = \frac{1 - p_s}{\lambda} + p_s\mathbb{E}[C]. \tag{5}$$

The purpose of the close-down phase is to avoid frequent switching of BS states.

The long-term average proportion of time that a BS, modelled as an M/G/1/$K$ queue following the $N$-policy, is active (busy serving users) in a regeneration cycle as in Fig. 1 is

$$\mathbb{E}[L_{\mathrm{A}}] = \frac{\mathbb{E}[A]}{1 - \mathbb{E}[A]}(\mathbb{E}[L_{\mathrm{CD}}] + \mathbb{E}[L_{\mathrm{SL}}] + \mathbb{E}[L_{\mathrm{ST}}]). \tag{6}$$

Note that as the M/G/1/$K$ queue has a finite state-space, its stability is always guaranteed. The carried traffic $\mathbb{E}[A]$ is less than $1$ even when the offered traffic $\lambda\mathbb{E}[R]$ exceeds $1$.

Finally, since the probability that a regeneration cycle involves sleeping and startup phases is equal to $p_s$,

$$\mathbb{E}[L_{\mathrm{SL}}] = p_s\frac{N}{\lambda}, \tag{7}$$

and

$$\mathbb{E}[L_{\mathrm{ST}}] = p_s\mathbb{E}[S]. \tag{8}$$

We next consider the set of embedded Markov points at which either a startup or a service is completed. We define the state of a BS at the end of the $n$-th embedded Markov point by

$$\xi_n = \begin{cases} 0 & \text{startup completion;} \\ 1 & \text{service completion.} \end{cases} \tag{9}$$

Let $Q_n$ denote the number of users immediately after the $n$-th embedded Markov point. Let $k$ be a non-negative integer, we define $\{q_k\}$ as the steady state probability for a BS to have $k$ users waiting after a startup completion, and $\{\pi_k\}$ as the steady state probability for a BS to have $k$ users in service after a service completion, that is

$$q_k = \lim_{n\to\infty} P[\xi_n = 0, Q_n = k], 0 \le k \le K, \tag{10}$$

$$\pi_k = \lim_{n\to\infty} P[\xi_n = 1, Q_n = k], 0 \le k \le K - 1. \tag{11}$$

Denote

$$s_k = \int_0^\infty \frac{(\lambda x)^k}{k!}e^{-\lambda x}\mathrm{d}S(x) \tag{12}$$

as the probability that $k$ users arrive during a startup time, and

$$r_k = \int_0^\infty \frac{(\lambda x)^k}{k!}e^{-\lambda x}\mathrm{d}R(x) \tag{13}$$

as the probability that $k$ users arrive during a service time, where the integrand is the probability that $k$ users arrive in $x$ units of time, times the probability of a startup/service lasting $x$ units of time. We define $r_j^c = \sum_{k=j}^{\infty} r_k$ and $s_j^c = \sum_{k=j}^{\infty} s_k$. The $\{q_k\}$ and $\{\pi_k\}$ terms satisfy the following equations:

$$q_k = 0, k \leq N-1, \tag{14a}$$

$$q_k = \pi_0 p_s s_{k-N}, N \leq k \leq K-1, \tag{14b}$$

$$q_K = \pi_0 p_s s_{K-N}^c, \tag{14c}$$

$$\pi_k = \sum_{j=1}^{k+1} (q_j + \pi_j) r_{k-j+1} + \pi_0 (1-p_s) r_k, k \leq K-2, \tag{14d}$$

$$\pi_{K-1} = \sum_{j=1}^{K-1} (q_j + \pi_j) r_{K-j}^c + q_K + \pi_0 (1-p_s) r_{K-1}^c, \tag{14e}$$

$$\sum_{k=N}^{K} q_k + \sum_{k=0}^{K-1} \pi_k = 1. \tag{14f}$$

From (14a), (14b) and (14c), we can further derive that the probability that an arbitrary embedded Markov point is the completion of a startup is

$$p_s \pi_0 = \sum_{k=N}^{K} q_k. \tag{15}$$

Therefore, the probability that an arbitrary embedded Markov point is the completion of a service is $1 - p_s \pi_0$.

We now consider the mean length of the interval between two successive Markov points in order to calculate the blocking probability and the mean delay. For a Markov point with $\xi_n = 1$ and $Q_n = 0$ (comprising $\pi_0$ of all Markov points), if the BS goes to sleep after the Markov point (with a probability of $p_s$), then the mean length of the interval right after the Markov point is the sum of the mean lengths of a vacation period that consists of $N$ inter-arrival times and a startup time, namely $N/\lambda + \mathbb{E}[S]$; conversely if the BS does not go to sleep (with a probability of $1 - p_s$), then the length of the interval right after is one inter-arrival time plus one service time, with a mean of $1/\lambda + \mathbb{E}[R]$. For any other Markov points, the length of the interval right after is always one service time, with a mean of $\mathbb{E}[R]$. The mean interval between two successive Markov points, denoted by $\eta$, is thus

$$\eta = p_s \pi_0 \left( \frac{N}{\lambda} + \mathbb{E}[S] \right) + \frac{(1-p_s)\pi_0}{\lambda} + (1 - p_s \pi_0) \mathbb{E}[R]. \tag{16}$$

The average carried traffic $\mathbb{E}[A]$ is given by the fraction of the time that the BS is busy serving users, which is the ratio of the last term in (16) to $\eta$ [32], namely,

$$\mathbb{E}[A] = \frac{(1 - p_s \pi_0) \mathbb{E}[R]}{\eta}. \tag{17}$$

The blocking probability under the isolated scheme, $P(B^I)$, is then

$$P(B^{\mathrm{I}}) = 1 - \frac{\mathbb{E}[A]}{\rho}, \tag{18}$$

and thus the throughput $\gamma^{\mathrm{I}}$ is

$$\gamma^{\mathrm{I}} = \lambda (1 - P(B^{\mathrm{I}})) = \frac{\mathbb{E}[A]}{\mathbb{E}[R]} = \frac{1 - (1-p_s)\pi_0}{\eta}. \tag{19}$$

The mean queue size is the expectation of the number of requests in the queue at an arbitrary moment, which is

$$\mathbb{E}[Q^{\mathrm{I}}] = \frac{1}{\lambda \eta} \sum_{k=1}^{K-1} k \pi_k + K \left( 1 - \frac{\mathbb{E}[A]}{\rho} \right). \tag{20}$$

Finally, by (20) and Little's Theorem, the mean delay of users under the isolated scheme is

$$\mathbb{E}[W^{\mathrm{I}}] = \frac{\mathbb{E}[Q^{\mathrm{I}}]}{\gamma^{\mathrm{I}}} = \frac{\mathbb{E}[R]}{\lambda \mathbb{E}[A] \eta} \sum_{k=1}^{K-1} k \pi_k + \frac{K}{\lambda} \left( \frac{\mathbb{E}[A]}{\rho} - 1 \right). \tag{21}$$

Note that although our analysis is based on the FCFS discipline, it can be extended to more general situations. Particularly, the power consumption is the same for both the FCFS discipline and the Processor Sharing (PS) discipline, where the bandwidth of a BS is divided equally to all admitted customers [8], [9]. For the blocking probability and mean delay, the results are applicable to the PS discipline with exponential service time distribution [33][1]. Also, the results for blocking probability and mean delay are insensitive to the distribution of the close-down time beyond its mean [6], [32].

### C. Cooperative scheme

In the cooperative scheme, as we need to analyze a multi-BS system where the states of BSs are interdependent, obtaining exact solutions for GoS metrics is computationally prohibitive (mostly due to the curse of dimensionality) [34]. Instead, we will apply the IESA framework to obtain approximations of the GoS metrics and power consumption with satisfactory levels of both accuracy and computational efficiency.

Due to the limited space, we only describe IESA with the necessary information needed for this paper and refer the readers to [13]–[15] for more detailed explanations on IESA. The key idea of IESA is to apply decomposition-based approximation methods to a surrogate system instead of the real system. Traditional decomposition-based approaches, such as EFPA, significantly underestimate blocking probability in systems with a mutual overflow effect due to two inherent simplifying assumptions during the decomposition process, namely the Poisson assumption that assumes that all traffic streams follow the Poisson process, and the independence assumption under which all subsystems are independent. In fact, overflow traffic is known to be non-Poisson even if the new traffic is Poisson, and states of the BSs are statistically dependent on each other due to mutual overflow. In cellular networks, overflow traffic is generated when user requests are rejected by the BS that they attempt, due to either the capacity limit or sleeping operation.

The surrogate system for IESA is specially designed with an information exchange mechanism to increase the proportion of

[1] The blocking probability and mean delay of M/G/1/$K$-PS queue without vacations are insensitive to the service time distribution beyond the mean, but this insensitivity property does not hold for the blocking probability and mean delay of M/G/1/$K$-PS queue with vacations [32], [33].

new traffic as compared to the real system. This increases the validity of both the Poisson and the independence assumptions. Therefore, when the decomposition-based approximation is applied to the surrogate model, the errors caused by the two simplifying assumptions are reduced.

To facilitate the information exchange mechanism, each user request is assigned two attributes in the surrogate system. The first attribute is overflow record of the request, denoted by $\Delta$, which contains the set of BSs that have rejected the request for admission due to capacity limit or sleep. The second attribute, denoted by $\Omega$, is a counter on the number of overflows experienced by the request itself or other requests being served and is an estimate of the number of unavailable BSs due to either sleeping or capacity limit in the network.

In cellular networks, a request can only be served by nearby BSs that provide coverage to the mobile user initiating the request. Therefore, for simplicity, we consider the first BS (which could be the closest BS to the user) that a request attempts, determines the set of BSs that the request can overflow to. We say that a request *originates from* the first BS that it attempts, and let $\Gamma_i$ denote the set of BSs that a request originated from BS $i$ is allowed to overflow.

A new request starts with $\Delta = \emptyset$ and $\Omega = 0$ at its initiation. When a request $x$, originated from BS $i_0$, arrives at BS $i \in \Gamma_{i_0}$ with attributes $(\Delta_x, \Omega_x)$, it will be admitted if BS $i$ is active and has not reached its capacity. Otherwise, it will be rejected and will overflow to one of the BSs in $\Gamma_{i_0} \setminus (\Delta_x \cup i)$. In this latter case, we consider the request with the highest $\Omega$ being served by $i$ at the moment of $x$'s rejection. We denote such a request by $y$ with attributes $(\Delta_y, \Omega_y)$. If $\Omega_x \geq \Omega_y$, $x$ will update its attributes to $\{\Delta_x \cup i, \Omega_x + 1\}$. However, if $\Omega_y > \Omega_x$, $x$ will exchange its second attribute with $y$ and then overflow with attributes $\{\Delta_x \cup i, \Omega_y + 1\}$, while $y$ will update its attributes to $\{\Delta_y, \Omega_x\}$.

To avoid confusion, we emphasize that the surrogate model described here is solely for the purpose of performance evaluation in this paper. We do not assign the above attributes or implement the above information exchange process for requests in the real network. Instead, the information exchange process helps to estimate whether all unattempted BSs are available or not for an overflow request, based on its attributes $\Delta$ and $\Omega$. If all unattempted BSs are presumed to be unavailable, the request will leave the system immediately and be counted as a blocked request, without trying to access any of the unattempted BSs. As in [13]–[15], we denote $p(\Omega^*, |\Delta|, \Omega_x)$ as the probability that a request with attributes $\{\Delta, \Omega_x\}$ is blocked immediately, where $\Omega^*$ is a parameter of the surrogate model representing the maximum limit on the value of $\Omega$ for any request.

Specifically, $p(\Omega^*, |\Delta|, \Omega)$ for a request originated from BS $i$ is evaluated as:

$$p(\Omega^*, |\Delta|, \Omega) = \begin{cases} 0 & \text{if } \Omega < n_i, \\ \\ \dfrac{\binom{\Omega-|\Delta|}{n_i-|\Delta|}}{\binom{\Omega^*-|\Delta|}{n_i-|\Delta|}} & \text{if } \Omega \geq n_i, \end{cases} \tag{22}$$

where $|\Delta| \leq n_i \leq \Omega^*$, and $n_i = |\Gamma_i|$.

The value of $\Omega^*$ depends on the level of dependency in the real system, which in turn is related to parameters such as the total number of servers (BSs) and the number of servers (BSs) that a certain user has access to [13], [15]. In [15], we proposed a regression method to obtain a quasi-optimal value of $\Omega^*$ that can give reasonably accurate approximations for a certain system. Specifically, we first consider a small set of independent cases with different system parameters as the training set, and evaluate the approximation accuracy in each case with different values of $\Omega^*$. Then, the $\Omega^*$ values with the smallest approximation error are recorded and a regression function is constructed with relevant system parameters as explanatory variables and $\Omega^*$ as the response variable. The regression function is then used to predict the value of $\Omega^*$ in different system settings.

According to (22), there is a positive probability for a request to be blocked before it attempts all BSs in $\Gamma_i$. The probability is higher for requests with higher values of $\Omega$, which indicates that the system is more congested. Also, $p(\Omega^*, |\Delta|, \Omega) = 1$ if $|\Delta| = n_i$ (which is consistent with the behavior in the real system) or $\Omega = \Omega^*$.

We further define the following:

- $\lambda_{i,o,n,\mathbf{s},j}$: Arrival rate for traffic to BS $i$ originated from BS $o$ with $\Delta = \mathbf{s} = \{s_1, ..., s_n\}$ ($s_1 = o, i \notin \mathbf{s}$), $|\Delta| = n$ and $\Omega = j$;
- $\lambda_{i,n,j} = \sum_{\mathbf{s}} \lambda_{i,o,n,\mathbf{s},j}$: Total effective arrival rate to BS $i$ with $|\Delta| = n$ and $\Omega = j$, summing over all possible origins and overflow sequences;
- $\lambda_{i,j} = \sum\limits_{l=0}^{m} \sum\limits_{m=0}^{j} \lambda_{i,l,m}$: Total combined arrival rate to $i$ with $\Omega \leq j$;
- $v_{i,o,n,\mathbf{s},j}$: Overflow traffic from BS $i$, originated from BS $o$, with $\Delta = \mathbf{s} = \{s_1, ..., s_n\}$ with $s_1 = o$, $s_n = i$, $|\Delta| = n$, and $\Omega = j$;
- $B_{i,j}$: Blocking probability of for requests with $\Omega = j$ at BS $i$.

The surrogate is in fact a hierarchical system based on the value of $\Omega$, where traffic with higher values of $\Omega$ is considered to be on "higher" levels. This traffic hierarchy, along with the information exchange and immediate blocking mechanisms in (22), helps to capture the state dependencies among different BSs. The blocking probability of traffic with a lower $\Omega$ is not affected by traffic with higher values of $\Omega$. Therefore, denoting $P_b(\lambda, R, K)$ as the blocking probability of an M/G/1/$K$ queue (without vacations as we do not consider short-term sleep in the cooperative scheme) with arrival rate $\lambda$ and service time requirement $R$, the relationship between the blocking probability $B_{i,j}$ and $\lambda_{i,j}$ at level $j$ for BS $i$ is

$$B_{i,j} = \begin{cases} P_b(\lambda_{i,j}, R, K) & \text{if BS } i \text{ is active;} \\ 1 & \text{if BS } i \text{ is in long-term sleeping,} \end{cases} \tag{23}$$

where $0 \leq j \leq \Omega^*$. One way to calculate the value of $P_b(\lambda_{i,j}, R, K)$ is by referring to (4) to (18), with $\mathbb{E}[C] \to \infty$ as the BS will not enter short-term sleep in the cooperative scheme.

For each BS $i \in U$, we start with the initial value $\lambda_{i,0} = \lambda_{i,0,0} = \lambda_i$, which represents the arrival rate of new requests to BS $i$.

Overflow traffic $v_{i,o,n,\mathbf{s},j}$ can be evaluated as

$$v_{i,o,n,\mathbf{s},j} = \mathbb{E}[R] \left( \sum_{\ell=n-1}^{j-2} \lambda_{i,o,n-1,\mathbf{s},\ell}(B_{i,j-1} - B_{i,j-2}) + \lambda_{i,o,n-1,\mathbf{s},j-1} B_{i,j-1} \right), \tag{24}$$

where the first term in the bracket represents overflow when an incoming request with $|\Delta| = n-1$ and $\Omega \leq j-2$ finds BS $i$ fully occupied with the highest $\Omega$ of the requests in service equal to $j-1$ (with probability $B_{i,j-1} - B_{i,j-2}$). The second term represents overflow when an incoming request with $|\Delta| = n-1$ and $\Omega = j-1$ finds BS $i$ fully occupied with $\Omega \leq j-1$ for all requests in service (with probability $B_{i,j-1}$) [15].

The overflow traffic $v_{i,o,n,\mathbf{s},j}$ will then be randomly offered to one of the $|\Gamma_{s_1} - n|$ unattempted BSs in $\Gamma_{s_1}$ with equal probability, namely $(1 - p(\Omega^*, n, j)) / (|\Gamma_{s_1}| - n)$, or give up attempting and exit the network with probability $p(\Omega^*, n, j)$. Thus, for any BS $i$,

$$\lambda_{i,o,n,\mathbf{s},j} = \frac{1 - p(\Omega^*, n, j)}{\mathbb{E}[R](|\Gamma_{s_1}| - n)} v_{s_n,n,\mathbf{s},j}. \tag{25}$$

We can then iteratively repeat (24) and (25) to calculate the overflow traffic and arrival rate for each level $j \leq \Omega$. Traffic with $\Omega = \Omega^* - 1$, namely those offered to the highest level, includes all lower levels and is equal to the total offered traffic to a BS. Subsequently, traffic carried by BS $i$ is $\lambda_{i,\Omega^*-1}\mathbb{E}[R](1 - B_{i,\Omega^*-1})$. The network blocking probability under the cooperative scheme is

$$P(B^{\mathrm{C}}) = 1 - \frac{\sum_{i \in U} \lambda_{i,\Omega^*-1}(1 - B_{i,\Omega^*-1})}{\sum_{i \in U} \lambda_i}, \tag{26}$$

where $U$ is the set of all BSs in the network and $\sum_{i \in U} \lambda_{i,\Omega^*-1}(1 - B_{i,\Omega^*-1})$ is the sum of carried traffic by all BSs.

As IESA decomposes the network into independent subsystems (individual BSs), we can refer to our previous analysis for the isolated scheme to calculate the mean delay for each BS. By replacing $\lambda$ with $\lambda_{i,\Omega^*-1}$ for each BS $i \in U$ and setting $\mathbb{E}[C] \to \infty$ (so that the BSs do not enter short-term sleep), the mean delay $\mathbb{E}[W_i^{\mathrm{C}}]$ for each BS can be obtained following the same analysis as in (4) to (17), (20) and (21). For BSs selected for long-term sleeping, we set $K = 0$ such that no requests would be admitted in such BSs. The mean delay of all requests is the weighted average of mean delay of each BS, namely,

$$\mathbb{E}[W^{\mathrm{C}}] = \frac{\sum\limits_{i \in U} \lambda_i \mathbb{E}[W_i^{\mathrm{C}}]}{\sum\limits_{i \in U} \lambda_i}. \tag{27}$$

With the assumption of random routing of overflow traffic, the IESA algorithm described above has a polynomial time complexity of $O(|U|n^*\Omega^*)$, where $|U|$ is the total number of BSs in the network and $n^* = \max_{i \in U} n_i$ is the maximum number of neighbors that any BS in the network has. As we have $|U| \geq \Omega^* \geq n^*$ by nature of the IESA algorithm [13], [15], the worst-case of time complexity is $O(|U|^3)$. We will verify the time complexity numerically later in Section V.

### *D. Hybrid scheme*

As mentioned in Section III, the hybrid scheme is a joint application of the isolated and cooperative schemes, where some BSs are selected to enter long-term sleep, while other BSs follows the $N$-policy for potential short-term sleep. To obtain the blocking probability $P(B^{\mathrm{H}})$ and mean delay $\mathbb{E}[W^{\mathrm{H}}]$ for the hybrid scheme, we can mostly follow the analysis of the cooperative scheme by IESA in Section IV-C, as overflow traffic is also present in the hybrid scheme.

However, when calculating the blocking probability $P_b(\lambda_{i,j}, R, K)$ of traffic at level $j$ in BS $i$ as in Equation (23), we should consider appropriate distributions for $C$ and $S$ instead of simply setting $\mathbb{E}[C] \to \infty$ as in the cooperative scheme, as the state probabilities for a BS selected not to enter long-term sleep should be calculated based on an M/G/1/$K$ queue with vacations, close-down and startup times.

## V. Numerical results

We now verify the accuracy of analytical methods described in the previous section by comparing analytical results with numerical simulations.

We collected and analysed real data for power consumption and mobile traffic from 27 January to 24 February 2017 from a BS site in Hong Kong operated by SmarTone Mobile Communications Limited [35], as shown in Fig. 2 (this figure was also reported in the conference paper [16]). It shows that the power consumption of the site when there is no mobile traffic is about 1867.6 W and the maximum power consumption during the period is about 2150 W. As we are limited by the information that we can publish, we assume that the site is composed of 7 identical BSs and the BSs are operating at their full capacity at the time of maximum power consumption[2], such that the static power consumption for an active BS is $P_{\mathrm{ID}} \approx 266.8$ W and the maximum transmission power of a BS is $P_{\mathrm{TR}} \approx 40.43$ W. We further assume $P_{\mathrm{SL}} \approx 10$ W, based on a separate experiment we carried out on energy saving by implementing BS sleeping in real networks. As discussed in Section III-B, the value of $P_{\mathrm{SL}}$ has accounted for the increasing transmission power by neighboring active BSs for coverage extension.

For simplicity without loss of generality, we assume that the mean arrival rate $\lambda$ arrivals/s is the same for all seven BSs, and mean service time $\mathbb{E}[R] = 1$ s. We consider a wide range of arrival rates, including low arrival rates during idle hours and high arrival rates during busy hours. Based on our analysis in Section IV, the results are insensitive to the distribution of the close-down time $C$ beyond its mean. Therefore, we further assume that the distribution of $C$ is exponential. In

[2] It is common in related studies to assume that BSs of the same type have the same power consumption profile (e.g., [11], [12], [22]).

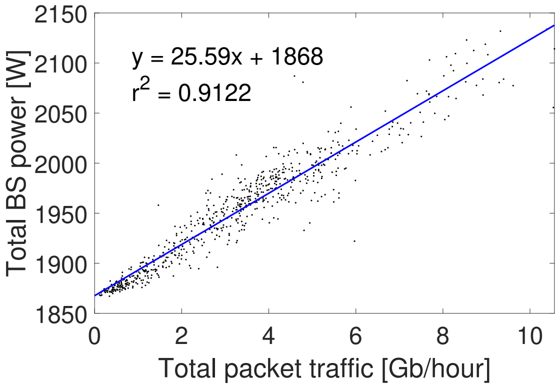


Fig. 2. Power consumption of a real BS site.

the cooperative and hybrid schemes, we assume that one of seven BSs is switched to long-term sleep, while all active BSs are able to serve traffic intended for other BSs that are full or sleeping as the BSs in the same site are geographically close to each other. As we consider a single system setting, the parameter $\Omega^*$ for IESA throughout this section is set to 7 (equal to $|U|$), which gives the most accurate approximation under this particular setting [15].

The 95% confidence interval for all simulation results presented in this section, based on Student's $t$-distribution, are within 3% of the observed mean.

### *A. Accuracy and computational efficiency of proposed analytical methods*

In Figs. 3 and 4, we demonstrate the accuracy of our proposed analytical methods by comparing the derived analytical results with simulation results. We set $K = 10$ for all three schemes in Fig. 3 and $K = 100$ in Fig. 4. The parameters related to short-term sleep for the isolated and hybrid schemes are $N = 3$, $\mathbb{E}[C] = 2$ s, and $\mathbb{E}[S] = 2$ s. The distributions for both $R$ and $S$ are exponential. The analytical and simulation results are quite close to each other in all cases. As expected, under the same arrival rate, the blocking probability is lower and the mean delay is higher for a larger $K$. In general, when the arrival rate is low, BSs are more likely to benefit from the short-term sleep.

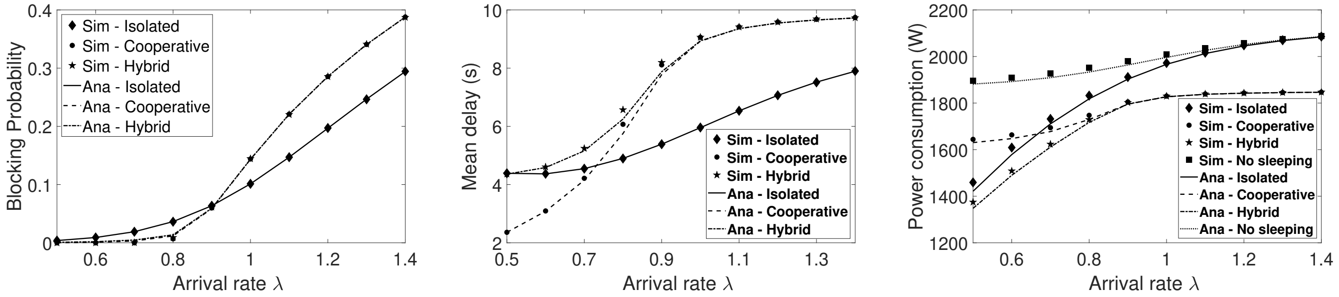
We then change the distributions of $R$ and $S$. As isolated and cooperative schemes can be regarded as special cases of the hybrid scheme, we present the results only for the hybrid scheme in Fig. 5. We consider four different situations with three types of distributions, namely exponential distribution, deterministic (degenerate) distribution, Pareto distribution with shape parameter 2.001 (Pareto-1, with finite mean and variance), and Pareto distribution with shape parameter 1.98 (Pareto-2, with infinite variance, but finite mean). Apart from the distributions, the other parameters are the same as in Fig. 3. The results demonstrate that power consumption and GoS are nearly insensitive to the distributions of service and startup times (beyond their means), so our analytical methods are sufficiently robust to variations in these distributions.

To demonstrate that our methods are also applicable for non-Poisson arrivals, we consider the MMPP arrival process under the hybrid scheme with exponential service and startup times. We demonstrate the simulation results for MMPP arrivals with two different *arrival state duration parameters* (which are the rate of transition between two states with different mean arrival rates [36], referred to as *mode duration parameters* in some publications [37]) in Fig. 6, where "MMPP-1" represents an arrival process with the arrival state duration parameters for both states equal to 1 and "MMPP-2" represents an arrival process with the arrival state duration parameters for both states equal to 100. The other parameters are the same as in Fig. 3. To the best of our knowledge, no effective analytical methods are available yet for evaluating the blocking probability and the mean delay for networks of multiple queues with MMPP arrivals. The results show that, even for MMPP arrivals with different arrival state duration parameters, our analytical methods based on Poisson arrivals still give a fairly accurate evaluation on the blocking probability, mean delay and power consumption for all three BS sleeping schemes, with less than 10% difference between the analytical and simulation results in most cases. Therefore, our method can be a desirable estimation tool for networks with MMPP arrivals.

In terms of computational efficiency, our analytical method based on IESA for cooperative and hybrid schemes can obtain a unique solution after a bounded number of iterations [13], while the brute-force Markov chain solution has exponential complexity in terms of the number of BSs [34]. In Fig. 7, we present numerical results for the running time by simulation (with the 95% confidence intervals for all results kept within 3%) and our analytical method as a function of the number of BSs in the network (similar to the empirical method used in [38]) under the hybrid scheme on a laptop with a 2 GHz Intel i7 processor and 8 GB RAM, by the Curve Fitting Toolbox of MATLAB. For the ease of demonstration, we assume that in a network with $|U|$ BSs, $n_i = |U| - 1$ for all $i \in U$, and set $\Omega^* = |U|$ in all cases, so that the complexity is $O(|U|^3)$ as discussed in Section IV-C. The numerical results in Fig. 7 demonstrate that both simulation and our analytical method have a polynomial complexity $O(|U|^3)$, while the running time of the simulation is at least five times longer than that of our analytical method. In certain cases, e.g., the one demonstrated in our conference paper [16] with exponential service times, deterministic close-down times, no setup times and 7 BSs in the network, the running time of our analytical method could be five orders of magnitude less than that of simulation. Given its low runtime and acceptable level of accuracy, our approximation method can search and find optimal tradeoff solutions in mobile cellular networks that tradeoff GoS and power consumption.

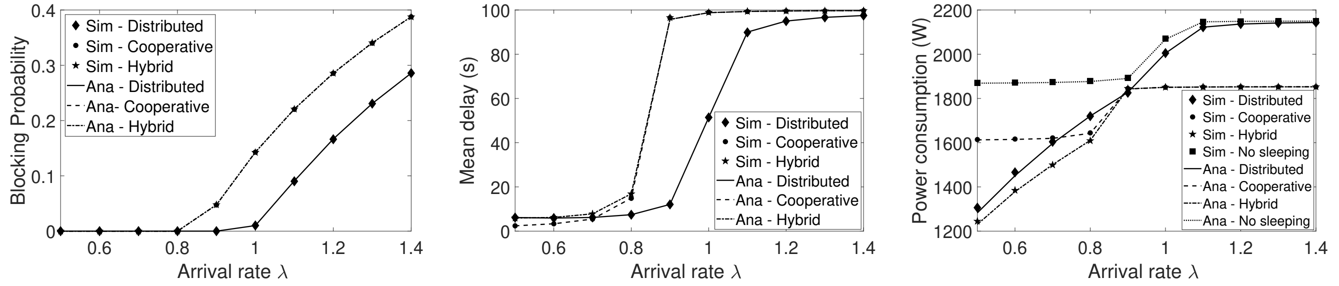
### *B. Trade-off between GoS metrics and power consumption*

We now apply our analytical methods to investigate the tradeoff between power consumption and GoS metrics including mean delay and blocking probability by adjusting the parameters including $K$, $N$, and $\mathbb{E}[C]$ under different BS sleeping schemes. In this subsection, we set $\mathbb{E}[S] = 2$ s, $\lambda = 0.6$, and assume that both $R$ and $S$ are exponentially distributed.



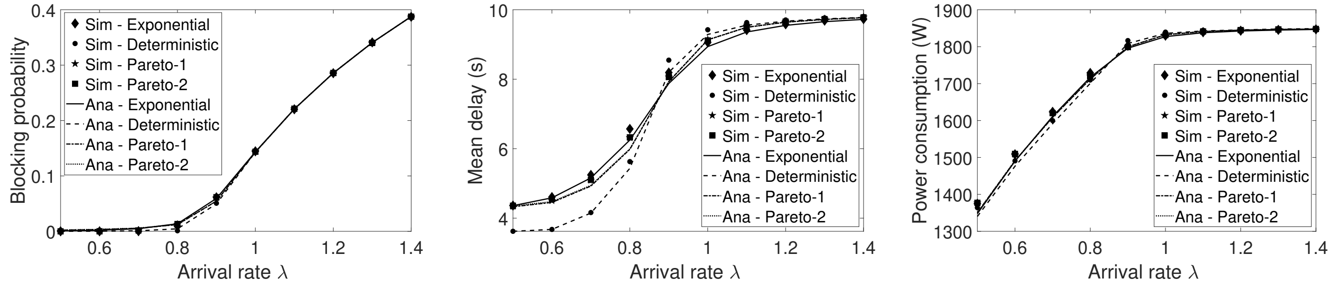
(a) (b) (c)

Fig. 3. Accuracy of proposed analytical methods for three BS sleeping schemes with $K = 10$ in terms of (a) Blocking probability; (b) Mean delay; (c) Power consumption. (Sim = simulation, Ana = analytical)



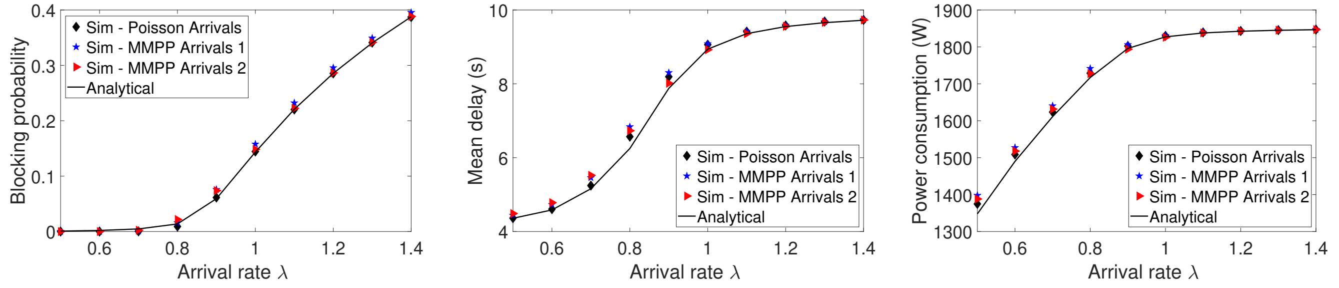
(a) (b) (c)

Fig. 4. Accuracy of proposed analytical methods for three BS sleeping schemes with $K = 100$ in terms of (a) Blocking probability; (b) Mean delay; (c) Power consumption. (Sim = simulation, Ana = analytical)



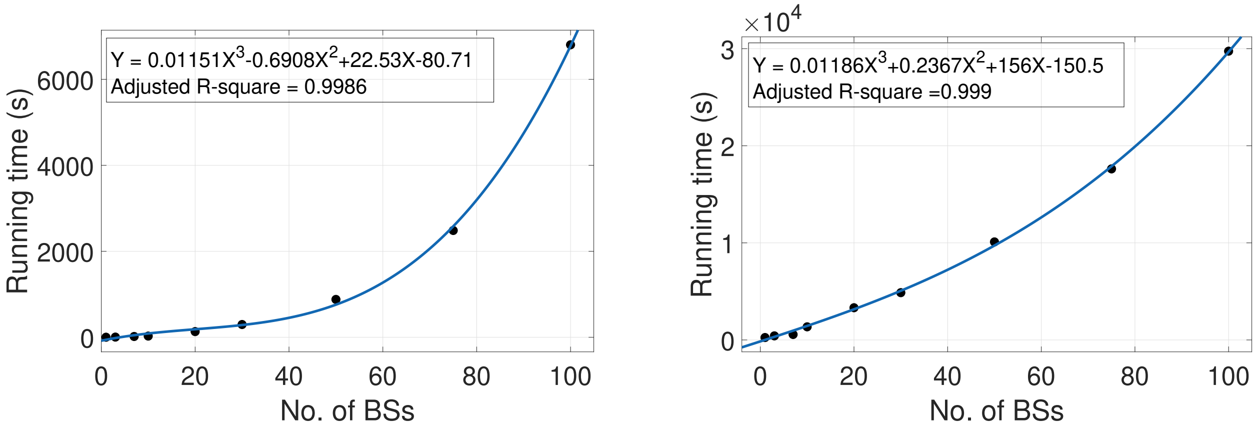
(a) (b) (c)

Fig. 5. Accuracy of proposed analytical methods for different distributions of service and startup times under the hybrid scheme in terms of (a) Blocking probability; (b) Mean delay; (c) Power consumption. (Sim = simulation, Ana = analytical)



(a) (b) (c)

Fig. 6. Comparison of analytical results and simulation results with Poisson and MMPP arrivals (arrival state duration parameters for both states = 1 for "MMPP-1" and = 100 for "MMPP-2") in terms of (a) Blocking probability; (b) Mean delay; (c) Power consumption. (Sim = simulation, Ana = analytical)



(a)

(b)

Fig. 7. Running time of (a) analytical methods; (b) simulation on networks with different number of BSs.

In Fig. 8, we keep $K = 10$ to focus on the tradeoff between the blocking probability and power consumption by changing the parameter $N$ under the isolated and hybrid schemes. The results verify that, when $N$ increases, power consumption decreases and blocking probability increases, as more arrivals must be accumulated at a BS in short-term sleep before it begins the startup. The results also show that, for the same amount of power consumption, the cooperative and hybrid schemes can achieve lower blocking probabilities compared to the isolated scheme. This is because cooperation among BSs in handling incoming traffic when some of BSs are not available leads to higher utilization of capacity in the whole network. Furthermore, the isolated and hybrid schemes are flexible in attaining a desirable level of trade-off between blocking probability and power consumption according to the requirement of the network operator by adjusting $N$ or $\mathbb{E}[C]$.

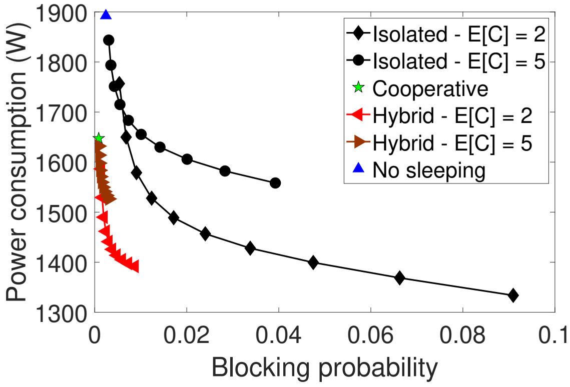


Fig. 8. Tradeoff between power consumption and blocking probability by adjusting $N$ when $K = 10$.

We then adjust the values of $K$ and $N$ to keep the blocking probability just under $1\%$, and demonstrate corresponding values of mean delay and power consumption in Fig. 9. The results show that, similar to Fig. 8, given the same level of power consumption, a shorter mean delay can be attained under the hybrid and cooperative schemes than the isolated scheme. Meanwhile, by increasing the value of $N$ while holding all other things equal, the power consumption can be reduced at the cost of a relatively higher delay under the isolated and hybrid schemes, as a larger $N$ leads to longer periods of short-term sleep under these two schemes.

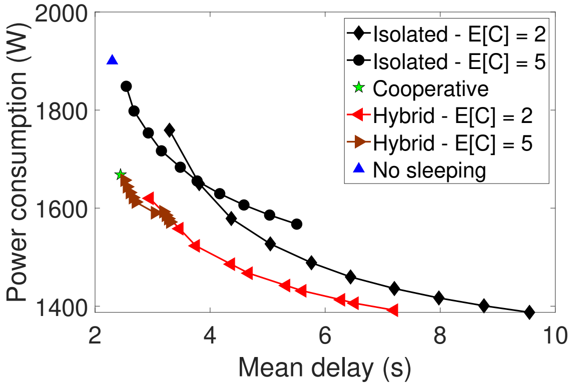


Fig. 9. Tradeoff between power consumption and mean delay by adjusting the values of $N$ and $K$.

Finally, we change the value of $K$ while keeping all other parameters constant, to demonstrate the tradeoff between mean delay and blocking probability under the three schemes in Fig. 10. By doing so, we are in fact adjusting the maximum allowable delay for admitted requests. A smaller $K$ leads to more strict requirement for the maximum delay, as each admitted request is expected to wait for fewer existing requests to finish their services. This also increases the blocking probability for every scheme as it is more likely that a request fails to satisfy the requirement and has to be blocked. In terms of the delay-blocking tradeoff, the cooperative and hybrid schemes are again superior to the isolated scheme.

## VI. Conclusions

In this paper, we proposed accurate, robust, and scalable analytical means to evaluate blocking probability, mean delay and power consumption in cellular networks with three BS sleeping schemes, by modeling the network as a network

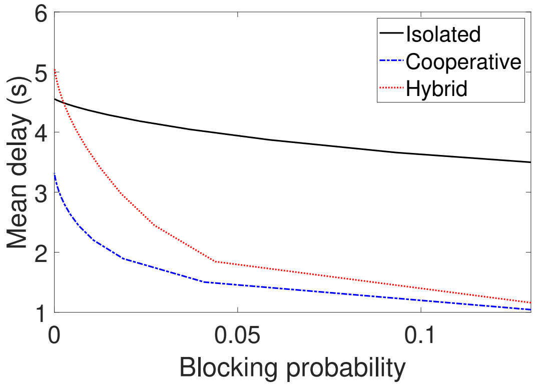


Fig. 10. Tradeoff between blocking probability and mean delay by adjusting the value of $K$.

of M/G/1/$K$ queues with vacations and applying the IESA framework. The accuracy and computational efficiency of our proposed method were verified by numerical simulations. Furthermore, we demonstrated that the cooperative and hybrid schemes can achieve better tradeoffs in terms of blocking-power, delay-power and blocking-delay than the isolated scheme. Compared to our conference version of this paper [16], we considered additional practical issues including the startup times and additional power consumption when activating a sleeping BS, and different distributions of service and startup times. We also demonstrated that our proposed methods are robust for MMPP arrivals over a large range of arrival state duration parameters.

Our new analytical methods, especially those for the cooperative and hybrid schemes where no existing feasible methods are available, are useful for applications such as network design, dimensioning and optimization, where numerous evaluations of GoS metrics and power consumption are needed to identify the optimal solutions for a board range of scenarios.

## 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] X. Ge, H. Jia, Y. Zhong, Y. Xiao, Y. Li, and B. Vucetic, "Energy efficient optimization of wireless-powered 5G full duplex cellular networks: A mean field game approach," *IEEE Transactions on Green Communications and Networking*, vol. 3, no. 2, pp. 455–467, June 2019.

[4] E. Mugume and D. K. C. So, "Deployment optimization of small cell networks with sleep mode," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 10, pp. 10 174–10 186, Oct 2019.

[5] Cisco, "Traffic analysis," Cisco Technology White Paper, Jul 2001.

[6] H. Takagi, "M/G/1/K queues with N-policy and setup times," *Queueing Systems*, vol. 14, no. 1, pp. 79–98, Mar 1993. [Online]. Available: https://doi.org/10.1007/BF01153527

[7] "End-to-end quality of service (QoS) concept and architecture," 3GPP TS 23.107 ver. 11.0.0 Rel. 11, 2012.

[8] 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.

[9] 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.

[10] E. Oh, K. Son, and B. Krishnamachari, "Dynamic base station switching-on/off strategies for green cellular networks," *IEEE Transactions on Wireless Communications*, vol. 12, no. 5, pp. 2126–2136, May 2013.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] J. Wu, E. W. M. Wong, Y. Chan, and M. Zukerman, "Energy efficiency-QoS tradeoff in cellular networks with base-station sleeping," in *Proc. IEEE GLOBECOM 2017*, Dec 2017, pp. 1–7.

[17] J. Li, N. Zhang, Q. Ye, W. Shi, W. Zhuang, and X. Shen, "Joint resource allocation and online virtual network embedding for 5G networks," in *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, Dec 2017, pp. 1–6.

[18] 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.

[19] B. Leng, X. Guo, X. Zheng, B. Krishnamachari, and Z. Niu, "A wait-and-see two-threshold optimal sleeping policy for a single server with bursty traffic," *IEEE Transactions on Green Communications and Networking*, vol. 1, no. 4, pp. 528–540, Dec 2017.

[20] Y. Chiang and W. Liao, "Green multicell cooperation in heterogeneous networks with hybrid energy sources," *IEEE Transactions on Wireless Communications*, vol. 15, no. 12, pp. 7911–7925, Dec 2016.

[21] F. H. Panahi, F. H. Panahi, G. Hattab, T. Ohtsuki, and D. Cabric, "Green heterogeneous networks via an intelligent sleep/wake-up mechanism and D2D communications," *IEEE Transactions on Green Communications and Networking*, vol. 2, no. 4, pp. 915–931, Dec 2018.

[22] F. Kong, X. Sun, V. C. M. Leung, and H. Zhu, "Delay-optimal biased user association in heterogeneous networks," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 8, pp. 7360–7371, Aug 2017.

[23] D. Renga, H. Al Haj Hassan, M. Meo, and L. Nuaymi, "Energy management and base station on/off switching in green mobile networks for offering ancillary services," *IEEE Transactions on Green Communications and Networking*, vol. 2, no. 3, pp. 868–880, Sep. 2018.

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

[25] 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.

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

[27] R. G. Addie, T. D. Neame, and M. Zukerman, "Performance analysis of a Poisson-Pareto queue over the full range of system parameters," *Computer Networks*, vol. 53, no. 7, pp. 1099–1113, May 2009.

[28] J. Hu, W. Heng, G. Zhang, and C. Meng, "Base station sleeping mechanism based on traffic prediction in heterogeneous networks," in *2015 International Telecommunication Networks and Applications Conference (ITNAC)*, Nov 2015, pp. 83–87.

[29] 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.

[30] 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.

[31] H. C. Tijms, *Stochastic models: an algorithmic approach*. John Wiley & Sons Chichester, 1994.

[32] T. T. Lee, "M/G/1/N queue with vacation time and exhaustive service discipline," *Operations Research*, vol. 32, no. 4, pp. 774–784, 1984.

[33] S. Borst, R. Núñez-Queija, and B. Zwart, “Sojourn time asymptotics in processor-sharing queues,” *Queueing Systems*, vol. 53, no. 1, pp. 31–51, Jun 2006.

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

[35] SmarTone Mobile Communications. [Online]. Available: https://www.smartone.com/en/?s=0

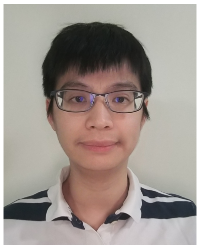
[36] K. Kang and C. Kim, “Performance analysis of statistical multiplexing of heterogeneous discrete-time Markovian arrival processes in an ATM network,” *Computer Communications*, vol. 20, no. 11, pp. 970 – 978, 1997.

[37] M. Zukerman, “Introduction to queueing theory and stochastic teletraffic models.” [Online]. Available: http://www.ee.cityu.edu.hk/~zukerman/classnotes.pdf

[38] C. Xing, R. G. Addie, Y. Peng, R. Lin, F. Li, W. Hu, V. Abramov, and M. Zukerman, “Resource provisioning for a multi-layered network,” *IEEE Access*, vol. 7, no. 1, pp. 16 226–16 245, 2019.

**Jingjin Wu** (S’15–M’16) received the B.Eng. degree (with first class honors) in information engineering in 2011 and the Ph.D. degree in electronic engineering in 2016 from City University of Hong Kong, Kowloon, Hong Kong. Since 2016, he has been an Assistant Professor with the Department of Statistics, BNU-HKBU United International College, Zhuhai, Guangdong, China. His current research focuses on performance evaluation, statistical analysis, design and optimization of wireless networks.

**Eric W. M. Wong** (S’87–M’90–SM’00) received the B.Sc. and M.Phil. degrees in electronic engineering from the Chinese University of Hong Kong, Hong Kong, in 1988 and 1990, respectively, and the Ph.D. degree in electrical and computer engineering from the University of Massachusetts, Amherst, MA, USA, in 1994. He is an Associate Professor with the Department of Electrical Engineering, City University of Hong Kong, Hong Kong. His research interests include analysis and design of telecommunications and computer networks, energy-efficient data center design, green cellular networks and optical networking.



**Yin-Chi Chan** (S’15–M’17) received the B.Math. degree from the University of Waterloo, Waterloo, Canada, in 2010, and the M.Sc. and PhD degrees from City University of Hong Kong, Hong Kong in 2011 and 2017, respectively. He is currently a Research Associate with the Department of Electrical Engineering, City University of Hong Kong, Hong Kong. His research interest is currently focused on approximative methods for the performance evaluation, optimization, and design of communications and service systems.

**Moshe Zukerman** (M’87–SM’91–F’07) received the B.Sc. degree in industrial engineering and management and the M.Sc. degree in operations research from the Technion – Israel Institute of Technology, Haifa, Israel, and the Ph.D. degree in engineering from University of California, Los Angeles, in 1985. He was an independent Consultant with the IRI Corporation and a Postdoctoral Fellow with the University of California, Los Angeles, in 1985–1986. In 1986–1997, he was with the Telstra Research Laboratories (TRL), first as a Research Engineer and, in 1988–1997, as a Project Leader. He also taught and supervised graduate students at Monash University in 1990–2001. During 1997-2008, he was with The University of Melbourne, Victoria, Australia. In 2008 he joined City University of Hong Kong as a Chair Professor of Information Engineering, and a team leader. He has served on various editorial boards such as Computer Networks, IEEE Communications Magazine, IEEE Journal of Selected Ares in Communications, IEEE/ACM Transactions on Networking and the International Journal of Communication Systems.