Performance Analysis of Green Cellular Networks with Selective Base-Station Sleeping

Jingjin Wu[a,b], Eric W. M. Wong[b], Jun Guo[c], Moshe Zukerman[b]

[a] *Division of Science and Technology, Beijing Normal University-Hong Kong Baptist University United International College, Zhuhai, Guangdong, P.R. China*
[b] *Department of Electronic Engineering, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong*
[c] *College of Science and Technology, Dongguan University of Technology, Dongguan, Guangdong, P.R. China*

Abstract

Base station (BS) sleeping is one of the emerging solutions for energy saving in cellular networks. It saves energy by selectively switching underutilized BSs to a low power consuming mode ("sleep mode") during low traffic hours while transferring their associated traffic to active BSs nearby. However, while saving energy, BS sleeping causes a reduction in total available capacity of the network, so Grade of Service (GoS) might be degraded, resulting in a trade-off between energy saving and network performance. This paper proposes a robust and computationally efficient analytical approximation technique, which we call Information Exchange Surrogate Approximation for Cellular Networks (IESA-CN), based on the recently established IESA framework for evaluation of GoS, as measured by call blocking probability, in cellular networks with different BS sleeping patterns. By considering the mutual overflow effect between BSs, the newly proposed method is verified by extensive and statistically reliable simulation experiments to significantly improve the accuracy as compared to traditional Erlang Fixed-Point Approximation in a wide range of scenarios.

Keywords: Blocking probability, approximation, base station, cellular network, energy efficiency.

1. Introduction

Energy efficiency is becoming increasingly important for cellular mobile network operators due to environmental and economic concerns [1–3]. Reduction in energy consumption reduces pollution and greenhouse gas, and brings cost-saving benefits to operators and consumers.

Substantial research effort has been directed towards reducing energy consumption in cellular networks. One of the emerging solutions to achieve this goal is base station (BS) sleeping [1–3]. BSs account for 50% – 80% of the total energy consumption by cellular networks. They are usually designed for peak hour traffic. However, traffic demands at individual BSs are highly variable both temporally and spatially [4]. As a result, a significant amount of energy can be wasted in under-utilized BSs, if all BSs operate in active mode all the time. Therefore, it is reasonable to selectively switch some of the BSs to "sleep mode", a low power consuming mode where only minimal structure required for re-activation operation is kept on. Note that a macro BS in sleep mode typically consumes less than 7% of the energy that is required by an active BS operating at full load [5]. Meanwhile, the remaining active BSs cooperatively extend coverage for calls which would have required service from those BSs that entered sleep mode.

While it is desirable to save energy by BS sleeping, it is also important to maintain the grade of service (GoS) when the total network capacity is reduced due to BS sleeping. That is, there exists a trade-off between energy saving and GoS requirements [2, 6–9]. Relaxing GoS constraints normally enables more energy savings, whereas tightening GoS constraints is likely to reduce energy saving by BS sleeping. A decision maker needs quantitative means to assess this trade-off, particularly, to evaluate the GoS for each energy saving scheme.

One GoS measure in cellular network is call blocking probability, defined as the number of calls that are blocked entry or dropped during service divided by the total number of call arrivals [7, 10, 11]. It is an important GoS measure for cellular networks since the GSM era. Even in current cellular network where data traffic has become the dominant traffic, blocking probability is still a useful measure for real-time mobile applications such as video streaming, mobile gaming and video conferencing. In order to satisfy Quality of Service (QoS) requirements such as delay and data rate in addition to the above mentioned blocking probability GoS requirement, there is a need to limit the number of admitted connections served by a given BS. Here,

the blocking probability is equivalent to the probability that a connection is rejected for admission because of the need to meet the application QoS requirements.

Unfortunately, it is often not possible to explicitly obtain an analytical expression for the blocking probability [12]. Computer simulation is therefore used to evaluate the performance of various cellular networks (e.g. [13–15]). However, simulations are time consuming, especially, for large systems. On the other hand, analytical approximations are much more computationally efficient. This is especially important when the evaluation is used as a module in a network design tool for searching optimal solutions where computational efficiency is key for such optimization procedures.

Therefore, analytical approximation methods are needed to evaluate blocking probability in cellular networks. The time required to obtain an approximation is often several orders of magnitude less than the time required for a simulation. The most prominent concern for approximations, however, is the accuracy.

We will base our discussions for evaluating blocking probabilities on the *overflow loss system* model, which forms an important class of teletraffic models for evaluating system performance. The definition of *traffic overflow*, according to [16], is "1. That condition wherein the traffic offered to a portion of a communication system exceeds its capacity and the excess may be blocked or may be provided with alternate routing, or 2. the excess traffic itself". In this paper, we use the shorter term *overflow* to refer to the condition of traffic overflow, and we use the commonly used term of *overflow traffic* (e.g. [17–21]) for the excess traffic itself. We will also use *overflow* as a verb to describe "action" by the access traffic when it is blocked or provided with alternate routing. Overflow loss systems are systems or networks where overflow traffic may exist. In such systems, if all servers in the primary server group are unavailable, overflow traffic is either blocked and cleared from the system, or it overflows to an alternative server group [22]. In cellular networks with BS sleeping, overflow occurs when a call attempts a sleeping or fully-loaded BS. In this case, an active BS nearby with available capacity will serve as the alternative server group for the call rejected by the sleeping or fully-loaded BS [23].

Apart from cellular networks considered in this paper, the applications of overflow loss systems can be found in many other systems ranging from classical telecommunication systems to emerging service sector models including such as circuit switching systems [24], optical burst switching systems [25, 26],

video-on-demand systems [27], call centers [28], and health-care systems [29].

The exact blocking probability in such overflow loss systems can be obtained by solving a set of steady-state equations for a multi-dimensional Markov process, with each dimension represents the state of one server group in the system. However, this approach is not viable in an overflow loss system as such system generally does not have a closed form solution for blocking probability [30, 31]. Also, due to the curse of dimensionality, this approach is not scalable for systems of practical size as the state space increases explosively when the system becomes large [12]. Therefore, it becomes desirable to estimate the blocking probability by analytical approximations.

The classical approximation method for blocking probability in overflow loss systems, known as the Erlang Fixed-Point Approximation (EFPA), was first proposed in 1964 [32]. Kelly [33] suggested that cellular networks with channel borrowing mechanism can be modelled as overflow loss systems and proposed to use EFPA to estimate the blocking probabilities. In this paper, we follow Kelly's suggestions by applying two EFPA-based approximation methods, i.e. the traditional EFPA and a newly-proposed versatile approach named Information Exchange Surrogate Approximation for Cellular Networks (IESA-CN). The IESA-CN approach, based on a recently proposed framework known as Information Exchange Surrogate Approximation (IESA), develops a surrogate model called an information exchange system (IES). It features an information exchange mechanism in which incoming calls may exchange certain congestion information with calls in service [22]. In addition, unlike the original IESA, IESA-CN captures unique features in cellular networks. We will show that, with simulation results as the benchmark, IESA-CN is a significantly more accurate approximation than the conventional approximation method EFPA and can be applied for evaluating blocking probabilities in cellular networks with BS sleeping in a wide range of scenarios. Note that IESA-CN, like EFPA, decouples the system into independent subsystems loaded with Poisson traffic (e.g., Erlang B subsystems [32, 33]), which makes both methods computationally efficient. On the other hand, IESA-CN introduces an information exchange mechanism, which can capture traffic dependence in the system, and hence it can significantly improve the accuracy over EFPA.

This paper is an extension of its conference version [34], that assumes identical traffic in each cell without considering BS sleeping. The contributions of this paper are summarized as follows:

1. We apply the traditional approximation technique EFPA to a cellular network model with BS sleeping technique. To the best of our knowledge, this is the first work that applies EFPA to obtain call blocking probabilities for such models.
2. We develop a suitable surrogate under IESA framework for cellular networks with BS sleeping called Information Exchange System for Cellular Networks (IES-CN). We provide, for the first time, an accurate and computationally feasible analytical approximation of blocking probability for cellular networks with (or without) BS sleeping, asymmetric offered traffic across BSs and handovers. IESA-CN based on the proposed surrogate IES-CN is shown to be accurate in a wide range of scenarios and we demonstrate that IESA-CN, which has a root from the classical EFPA, is a significant improvement over EFPA.

IESA-CN has many potential network design applications including network planning, resource allocation, and admission control. It can apply to homogenous or heterogeneous mobile cellular networks with or without BS sleeping.

The remainder of this paper is structured as follows. Section 2 provides background information and discusses existing work on BS sleeping techniques and approximation methods on cellular networks and overflow loss systems. Section 3 describes the cellular network model to be evaluated. Section 4 defines call attributes used by the IES-CN and describes each approximation method in detail. A case study and its associated numerical results are presented in Section 5. Finally, Section 6 concludes the paper.

2. Related work

2.1. BS sleeping technique

A comprehensive survey of BS sleeping techniques in cellular networks is provided in [2]. Compared to the other green techniques such as upgrading hardware components to more energy-efficient standards, BS sleeping has the advantage of minimal deployment and replacement cost as all the operations could be implemented on existing network infrastructure. Marsan *et al.* [35] compared energy savings achieved by different switching patterns of BSs, but did not take GoS degradation into consideration, and instead assume that the remaining active BSs are able to serve all the calls during low traffic hours. Gong *et al.* [7] proposed a two-stage dynamic algorithm to optimize the

trade-off between energy consumption and blocking probability in renewable energy powered cellular networks without considering overflow traffic. Niu *et al.* [8] provided a delay model for a single cell with BS sleeping based on an M/G/1 queue with vacations. This simple single BS model leads to closed-form results for the trade-off between energy and delay. However, it does not consider handover and overflow traffic present in cellular networks with BS sleeping. Bousia *et al.* [36] proposed a distance-aware BS sleeping algorithm in LTE-Advanced cellular network. The authors demonstrated by simulation that their proposed algorithm outperforms an existing random switching-off algorithm.

BS sleeping techniques have also been applied concurrently with other green strategies, such as heterogeneous cell deployment, to further improve the savings. Cao *et al.* [37] proposed strategies to obtain the optimal BS density (which can be achieved by BS sleeping) in homogeneous and heterogeneous cellular networks with service outage probability constraint based on stochastic geometry theory. Huang *et al.* [13] discussed three energy-efficient control strategies including BS sleeping in heterogeneous cellular networks based on large-scale user behavior by formulating an optimization problem involving BS density, BS power consumption and GoS requirement. The authors used simulations to demonstrate that significant improvement of energy efficiency can be achieved by integrating BS sleeping technique and heterogeneous cell deployment strategy. Chen *et al.* [9] proposed a joint BS and relay stations (RS) sleeping mechanism to maximize energy saving under GoS constraint. The authors also adopted simulations to obtain energy consumption and GoS metric such as throughput.

BS sleeping schemes can also be cooperatively applied by multiple network operators to further reduce energy consumption. Oikonomakou *et al.* [38] proposed a cooperative sleeping scheme in a single macrocell of a heterogenous cellular network owned by multiple network operators. By simulation results, the proposed scheme is shown to achieve notable energy saving while maintaining satisfactory QoS for users. Bousia *et al.* [39, 40] further studied operators cooperation for BS sleeping, they proposed a game-theoretic approach to minimize cost for each operator [39] and a multiobjective framework to solve the problem of resource allocation among operators by comparing different strategies of bidding for resources [40].

A number of factors could affect the effectiveness of different BS sleeping strategies in terms of energy saving and GoS impact. Tabassum *et al.* [14] analyzed the impact of different scheduling and user association schemes by

active BSs on the spectral efficiency of users originally associated with BSs that entered sleep mode. Wu *et al.* [6] analyzed the energy saving and delay trade-off, and developed a scheme to choose the optimal time to trigger BS sleep mode based on minimal total power needed to support a certain offered traffic load. Han *et al.* [23] considered four progressive BS switching patterns according to traffic patterns in order to maximize energy saving. The authors considered the outage probability of users based on path-loss and fading effects. Notably, all the GoS, or QoS measures, used in [9, 13, 14, 23, 36, 38–40], i.e. delay, outage probability or spectral efficiency, were obtained by Monte-Carlo simulations.

2.2. Blocking probability approximations for cellular networks without BS sleeping

While most existing publications (e.g., [13, 14, 23]) evaluate blocking probability in cellular networks by computer simulation, there is literature on blocking probability evaluation for cellular network (without BS sleeping) via analytical means. One example is the above discussed work by Cao et al. [37]. In addition, Raymond [41] proposed a method to estimate minimal blocking probability for simple cellular networks with dynamic channel allocation and flow control. Lagrange and Godlewski [42] approximated blocking probability in a two-tier cellular network consisting of micro-cells and umbrella-cells, where calls first attempt the micro-cells and overflow to umbrella-cells if no channels in micro-cells are available. Huang *et al.* [20] proposed another approximation method called multiservice overflow approximation (MOA) to evaluate blocking probability in a similar two-tier cellular network structure.

None of [37], [41], [42] or [20] considered possible mutual traffic overflow [43] among BSs, due to *channel borrowing* or *user association* techniques, by which a user is able to use a channel, or capacity originally assigned to another BS, if the first BS it attempts cannot offer the required service due to insufficient capacity or sleep mode operation [2, 14, 34, 44]. Such *mutual overflow effect* does not exist in models with hierarchical structure such as [38, 42], where the overflows are only assumed to occur unidirectionally from the micro-cells to the umbrella-cells, but not the other way around, or bidirectionally between two micro-cells. By contrast, in non-hierarchical cellular network structure where overflow traffic flows bilaterally, congestion in a particular BS would cause increasing overflow traffic to other BSs, which may in turn yield more overflow traffic to the original BS. Mutual overflow adversely affects the accuracy of approximations, such as those proposed

in [20, 37, 41, 42] when they are used for cellular networks with mutual overflow.

2.3. Blocking probability approximations for overflow loss systems

With channel borrowing, calls that arrive at fully-loaded or sleeping BSs can be served by neighboring BSs with idle capacity. Therefore, cellular networks with BS sleeping can be modelled as an overflow loss system. A BS here is regarded as a *server group*, and each channel (serving a single call) in the BS is a *server*. In the case without channel borrowing, Everitt [44] further mentioned that EFPA can be used to approximate blocking probability in cellular networks with the consideration of call mobility among cells with high accuracy. Mitchell and Sohraby [45] showed that EFPA is very accurate in assessing blocking probabilities for new and handover calls with different control strategies in a multi-cell multi-class cellular network model with symmetric traffic loading.

References [44] and [45] do not discuss channel borrowing mechanism or user association process, which are important parts of the technical basis of BS sleeping. Particularly, the user association process allows users originally associated with BSs that entered the sleep mode to be reassociated with nearby active BSs [2].

Overflow traffic in cellular networks with BS sleeping, especially because of the mutual overflow effect caused by the user association process, could lead to very inaccurate estimation of blocking probability by EFPA [43]. Here we consider EFPA in the sense of [33] without moment matching as was done in [32]. In this sense, EFPA is a classical approximation method based on Poisson and independence assumptions. The Poisson assumption assumes that if the arrival process is Poisson, the overflow traffic also follows a Poisson process. The independence assumption assumes that all server groups are mutually independent. However, it is known that the mean rate of overflow traffic is higher than its variance implying that modelling it by a Poisson process introduces errors, and that the server groups are not statistically independent because a busy server group is likely to imply that other server groups are also heavily loaded at the same time.

Due to these two assumptions, EFPA dramatically reduces the computing time as compared to the original multi-dimensional Markov process. However, they also lead to inaccurate estimates in various scenarios [22]. Several publications have proposed ways to combat the errors of EFPA, e.g., using

moment matching techniques to reduce errors due to the Poisson assumption [46] or derive conditional probabilities to reduce errors due to the independence assumption [47]. However, the improvement of EFPA using moment matching techniques is marginal in systems involving mutual overflow (where the independence error is dominant) while the conditional probabilities derivation method is not scalable [48].

Another approach is to apply the technique used in the traditional EFPA, i.e., decoupling the system into independent Erlang B subsystems, to a certain surrogate of the original system. For example, in the surrogate model used in Overflow Priority Classification Approximation (OPCA) [48], calls are classified based on the number of overflows they experience. A preemptive priority regime, where a junior call with a lower number of overflows is entitled to preemptive priority over a senior call with a greater number of overflows, is incorporated into the original model. That is, a senior call in service must give up its own channel during its service period if a junior call requests it. Alternatively, this preemptive process can be viewed as if the arriving junior call and the senior call in service exchanged their identity upon the arrival. In this way, the congestion information carried by the senior call can be used by the junior call.

By decoupling the system of the surrogate model into independent Erlang B subsystems as in EFPA, OPCA is able to reduce the errors due to the independence assumption in EFPA by capturing state dependencies among overflow traffic. It has been shown to be quite accurate in systems where all calls have full access to all server groups. However, cellular networks can be viewed as partially accessible networks as it is unlikely for a call to visit all the BSs in the system during its lifetime. The rudimentary approach of OPCA is shown to be inaccurate in such systems [22].

The IESA framework, with its roots in EFPA and OPCA, was proposed to estimate blocking probability in partially-accessible networks [22]. Instead of literally swapping the calls as in OPCA, the calls in IESA only exchange certain congestion information while retaining their own identities and overflow records in the IES (i.e. the surrogate model associated with IESA). IESA could be quite accurate if a suitable surrogate is chosen for the system concerned.

The IESA framework initially proposed in [22] was further modified in [49, 50] to estimate the blocking probabilities in Video on Demand (VoD) systems. However, it will not yield an accurate approximation for blocking probability in cellular networks due to the inherent differences between these

two systems. In VoD systems, a request is possibly able to overflow to any disks where a copy of the requested movie is available. However, in cellular networks, when a call is rejected by a BS due to sleeping or insufficient capacity, it can only overflow to nearby BSs due to limited signal strength (the *locality feature*). In this sense, the statistical dependencies among states of local BSs in cellular networks are stronger as they are more highly correlated. Furthermore, in cellular networks, the set of BSs that a given call is allowed to overflow to is changed when the call performs a handover from one BS to another (the *mobility feature*). This situation does not exist in VoD systems where the set of accessible disks is determined upon the initiation of a request and remains fixed throughout its lifetime. Therefore, the surrogate used in [22], [49] and [50] is inappropriate, and its corresponding blocking probability approximation is inaccurate for the current problem.

Following the discussions above, the challenge and contribution of the paper is to design a suitable surrogate for cellular networks with BS sleeping, and derive an accurate approximation of blocking probability of the surrogate under the IESA framework. We will demonstrate later in the paper that the locality and mobility features uniquely present in cellular networks can be addressed by means of a single parameter already available in the original IESA framework. The fact that adjusting a single parameter can adapt the original approximation (as demonstrated in [22], [49] and [50]) under the IESA framework to a different network model illustrates the flexibility and versatility of the IESA framework for estimating the blocking probability of various overflow loss systems.

3. Network Model

Consider a cellular network with multiple interconnected BSs. We define U as the set of all BSs in the network, and let $\Gamma_i \subset U$ denote the set of BSs that a call originated from BS i is allowed to overflow. The number of BSs to which a call originated from i has access is denoted by $n_i = |\Gamma_i|$. The set of traffic source cells that have access to i is denoted by Φ_i. Note that here we do not prioritize handover calls over new calls (e.g., [45, 51, 52]). Priority schemes involving granting preemptive priority to handover calls, allowing handover calls to wait in the buffer, or reserving a proportion of channels exclusively for handover calls can be incorporated into our model by treating new calls and handover calls as two different classes, and applying approximation techniques to each of them. Therefore, the *new call blocking*

probability and *handover forced termination probability* are equal and we will refer to both as *blocking probability* to avoid ambiguity.

We mainly focus, in this paper on a homogeneous cellular network without inter-layer overflow traffic such as that from micro cells to underlying macro cells as in [13, 38]. In Section 5 we also consider a case with irregular cell layouts. We note that moment matching [46, 50], which is the main technique to approximate blocking probabilities in multi-layer overflow loss systems, can be incorporated with the IESA framework to evaluate the blocking probabilities in heterogeneous cellular networks.

We also note that in current and future cellular networks, packet-switched data becomes the dominant traffic due to the popularity of multimedia mobile applications. A multi-service multi-rate loss model is required for such systems [53]. There has been work on applying EFPA or EFPA-based approximations to multi-service multi-rate systems (e.g. [54]). In this paper we assume single rate for simplicity and will leave the extension to multi-service multi-rate model for future research.

We assume that new calls arrive at BS i following a Poisson process with rate λ_i. As we discussed previously, the assumption of Poisson arrivals is fundamental in EFPA. It is also a common assumption in existing research on cellular networks (e.g. [8, 36]). We will show later that because the system is not very sensitive to the burstiness in the arrival process, our proposed method is also fairly accurate for systems with more bursty arrivals, that may be modeled by a Markov modulated Poisson process (MMPP).

Call service times are independent and exponentially distributed with mean $1/\mu_i$. Call sojourn times in each BS are also assumed to be independent and exponentially distributed with mean $1/\delta_i$. For simplicity, we assume the values of μ_i and δ_i are equal across all BSs, thus they can be denoted as μ and δ. In this sense, each BS can be modeled as an $M/M/c/c$ queue (e.g., [41, 52]). The state of BS i is denoted by S_i, where $S_i = 0$ if i is sleeping, while $S_i = 1$ if i is active. We will also demonstrate later that the blocking probabilities are nearly insensitive [55] to service and sojourn time distributions. The classical *symmetric random walk model* is adopted to characterize call mobility. In such a model, a call leaving a cell will move to any one of the neighboring cells with equal probability [56]. Also, for simplicity and without loss of generality, we assume all BSs to have the same number of channels, denoted as $c = c_i$ for all $i \in U$.

The traffic offered (in Erlangs, similarly hereinafter) to a particular BS i is denoted as $A_i = \lambda_i/\mu$. The total traffic offered to the system is thus

$A = \sum_{i \in U} A_i$.

A call will leave its current serving BS under either one of two following conditions: 1) the call completes its service and leaves the system, and 2) the call performs a handover to a neighboring cell served by another BS (due to the mobility of calls).

Assuming that service and sojourn times are independent and exponentially distributed, the probability θ that a call in the network performs a handover given that it has not been completed, is given by

$$\theta = \frac{\delta}{\mu + \delta}. \tag{1}$$

We define $\hat{B}_i$ as the blocking probability at BS i. Given $\hat{B}_i$, the combined arrival rate of new and handover calls λ'_i for i can be obtained by

$$\lambda'_i = \lambda_i + \sum_{j \in \Upsilon_i} \lambda'_j (1 - \hat{B}_j)\theta \tag{2}$$

in which Υ_i is the set of direct neighbor BSs of i.

We define A'_i as the effective offered traffic to i after taking the mobility of calls into consideration. It is given by

$$A'_i = \frac{\lambda'_i}{\mu + \delta}. \tag{3}$$

Taking the mobility feature into account, let B_i^{call} denote the probability that a call originated from BS i cannot complete its service due to blocking or dropping. Given $\hat{B}_i$ and θ, we can derive B_i^{call} by another set of fixed-point equations [57]:

$$B_i^{call} = \hat{B}_i + \theta(1 - \hat{B}_i) \frac{1}{|\Upsilon_i|} \sum_{j \in \Upsilon_i} B_j^{call}. \tag{4}$$

The first term ($\hat{B}_i$) in (4) represents the probability that the call is blocked upon its initiation at i, and the second term represents the probability that the call is successfully admitted to i but dropped upon a handover attempt before its completion.

Henceforth, we assume for any $i \in U$ that Υ_i and Γ_i coincide and are the set of its direct neighbors. This means that a call arriving at i is able to use (borrow) channels in all its direct neighboring BSs if BS i itself is sleeping or

has no vacant channels. Meanwhile, a call can perform multiple handovers across neighboring BSs during its service period.

4. Approximations

In this section, we begin by defining the set of call attributes used for IES-CN. Then we show how to apply EFPA and IESA-CN to our network model.

4.1. Call attributes and notations

In order to perform the information exchange mechanism and estimate blocking probability, we assign several attributes to calls in the model. The first attribute is the call identity I, which contains information including the call's origin, the call's expected service time and sojourn time in each server group, and the elapsed time since the call's inception. The second attribute, denoted by Δ, is defined in [22] as the set of server groups that the call has already attempted and overflowed, or overflow record of the call. Correspondingly in our model, Δ represents the set of BSs that has rejected admission of the call due to no available channels. The third attribute, denoted by Ω, represents an estimate of network congestion, which can be used to capture the statistical dependencies in the entire network. We will discuss this in further detail later when we describe IESA-CN.

We denote call ζ, with its first, second and third attributes being I_ζ, Δ_ζ, Ω_ζ, respectively, as an (I_ζ, Δ_ζ)-call or $(I_\zeta, \Delta_\zeta, \Omega_\zeta)$-call. Attributes of a call may be updated or exchanged during the call's sojourn time in the network. The specific rules of updating or exchanging depend on the approximation method used. We further assume that call ζ generated in BS i determines its next overflow destination by random hunting in $\Gamma_i - \Delta_\zeta$, which is the set of BSs that the call has access to and not yet attempted. As the nature of random hunting requires keeping track of each random sequence of BSs that a call attempts, we define $\Psi(X, x), x = 0, 1, ..., |X|$ as the set of choices of x elements from X. By definition, $\Psi(X, 0) = \emptyset$.

The attributes are created for the information exchange mechanism in the surrogate model under the IESA framework. Although such mechanism does not exist in the true model, the value of the attributes are helpful in calculating blocking probabilities by EFPA. Here, the "true model" is the original cellular network model for which we aim to evaluate estimate the blocking probability by simulation or approximations. Accordingly, notations

for attributes and their values are only relevant to EFPA, IESA-CN, and IES-CN (the surrogate model of IESA-CN) and they appear in the following sections.

4.2. EFPA

EFPA decouples a system of k server groups (in our case k BSs) into k independent Erlang B subsystems [32, 33]. This is consistent with our previous assumption to model BSs as $M/M/c/c$ queues. The load offered to each BS includes the original traffic, the handover traffic, plus all the traffic that overflows to it from other BSs either due to BS sleeping, or unavailability of free channels. We introduce the following notations in order to describe the model more systematically.

For each BS $i \in U$ in EFPA, we define (using the superscript E to represent EFPA):

- $a^{\mathrm{E}}_{i,m,n,\boldsymbol{s}}$ – Traffic offered to i with n overflows from source m and have overflowed sequentially along the path $\boldsymbol{s} = \{s_1, s_2, ...s_n\}$ ($m \in \Phi_i$; $\boldsymbol{s} \subset \Gamma_m$, $n < n_m$).

- $a^{\mathrm{E}}_{i,n}$ – Traffic offered to i with n overflows, namely summing all eligible $a^{\mathrm{E}}_{i,m,n,\boldsymbol{s}}$:

$$a^{\mathrm{E}}_{i,n} = \sum_{m \in \Phi_i, n < n_m} \sum_{\boldsymbol{s} \subset \Psi(\Gamma_m - \{i\}, n)} a^{\mathrm{E}}_{i,m,n,\boldsymbol{s}}. \tag{5}$$

- A^{E}_i – Total combined traffic offered to i, namely:

$$A^{\mathrm{E}}_i = \sum_{n=0}^{\hat{n_i}-1} a^{\mathrm{E}}_{i,n}, \tag{6}$$

  where $\hat{n_i} = \max_{m \in \ \Phi_i} n_m$.

- $v^{\mathrm{E}}_{s_n,n,m,\boldsymbol{s}}$ – Overflow traffic from i with n overflows originated from m that have overflowed sequentially along the path $\boldsymbol{s} = \{s_1, s_2, ..., s_{n-1}, s_n\}$ ($m \in \Phi_{s_n}, \boldsymbol{s} \subset \Psi(\Gamma_m, n)$, $n \leq n_m$, if $n = n_m$, the traffic will be cleared out).

- B^{E}_i – Probability that all channels in i are busy.

By the Poisson assumption and Erlang B formula [33, 58], we obtain the relationship between B_i^{E} and A_i^{E} as:

$$B_i^{\mathrm{E}} = \begin{cases} \mathrm{E}(A_i^{\mathrm{F}}, c) & \text{for all } i \text{ with } S_i = 1; \\ 1 & \text{for all } i \text{ with } S_i = 0, \end{cases} \tag{7}$$

where $\mathrm{E}(A, c)$ is the Erlang B formula where A is the total offered traffic in Erlang and c is the number of channels available.

With the independence assumption of EFPA, the offered traffic $a^{\mathrm{E}}_{s_n,m,n-1,\boldsymbol{s}-\{s_n\}}$, which has overflowed from $n-1$ BSs sequentially along the path $\boldsymbol{s}$, will again overflow from the s_n with probability $B^{\mathrm{E}}_{s_n}$ becoming overflow traffic $v^{\mathrm{E}}_{s_n,n,m,\boldsymbol{s}}$, so that

$$v^{\mathrm{E}}_{s_n,n,m,\boldsymbol{s}} = a^{\mathrm{E}}_{s_n,m,n-1,\boldsymbol{s}-s_n} B^{\mathrm{E}}_{s_n}. \tag{8}$$

The overflow traffic will subsequently be offered to another randomly chosen i in $\Gamma_m - \boldsymbol{s}$, namely a BS accessible by calls originated from m and not yet attempted by the overflow call (not in the path $\boldsymbol{s}$). As there are $n_m - n$ BSs in the set $\Gamma_m - \boldsymbol{s}$, the overflow traffic will be offered to each BS with probability $\frac{1}{n_m - n}$. Accordingly, we have

$$a^{\mathrm{E}}_{i,n,m,\boldsymbol{s}} = \frac{v^{\mathrm{E}}_{s_n,n,m,\boldsymbol{s}}}{n_m - n} \text{ for } i \in \Gamma_m - \boldsymbol{s}. \tag{9}$$

Combining (8) and (9), we can derive $a^{\mathrm{E}}_{i,n,m,\boldsymbol{s}}$ and A_i^{E} as

$$a^{\mathrm{E}}_{i,n,m,\boldsymbol{s}} = \frac{A_m}{n_m} \prod_{j=0}^{n} \frac{B^{\mathrm{E}}_{s_j}}{n_m - j}, \tag{10}$$

and

$$A_i^{\mathrm{E}} = \sum_{m \in \Phi_i} \frac{A_m}{n_m} \left[1 + \sum_{n=0}^{n_m - 1} \sum_{\boldsymbol{s} \subset \Psi(\Gamma_m - \{d\}, n)} \prod_{j=1}^{n} \frac{B^{\mathrm{E}}_{s_j}}{n_m - j} \right], \tag{11}$$

respectively.

Together, (7) and (11) constitute a set of fixed-point equations, which can be solved by the successive substitution method [59]. The iteration is continued until the differences between two consecutive results of B_i^{E} for all $i \in U$ are less than a preset threshold. It follows that the overall blocking probability of the true model estimated by EFPA, denoted by $\hat{B}^{\mathrm{E}}$, is given by

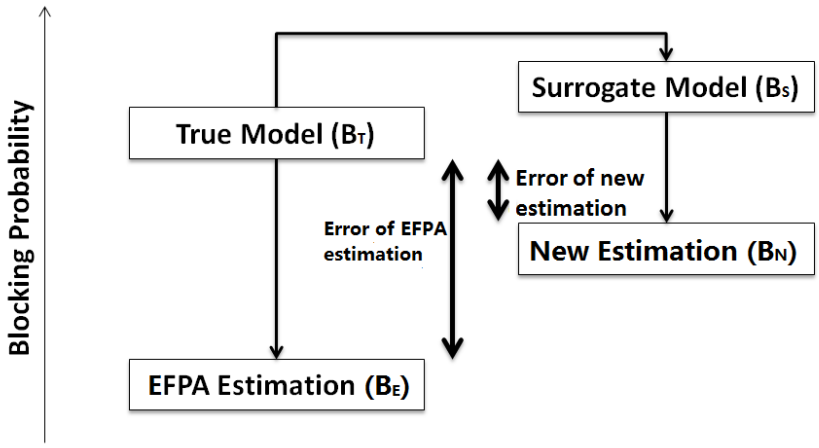


Figure 1: Conceptual illustration of the blocking probability approximation based on IESA framework.

$$\hat{B}^{\mathrm{E}} = 1 - \frac{\sum_{i \in U} A_i^{\mathrm{E}}(1 - B_i)}{A}. \tag{12}$$

We can also obtain the blocking probability of calls from a specific BS. Based on our definitions, if the traffic $a_{i,m,n_m-1,\boldsymbol{s}}$ is blocked once again (with the probability B_i^{E}), it will become $v_{i,m,n_m,\boldsymbol{s}+i}$ and be cleared out of the system as its overflow count reaches n_m. Therefore, the blocking probability for calls originated from m can be expressed as:

$$\hat{B}_m^{\mathrm{E}} = \frac{\sum_{\boldsymbol{s} \subset \Psi(\Gamma_m, n_m)} v^{\mathrm{E}}_{s_{nm},m,n_m,\boldsymbol{s}}}{A_m}. \tag{13}$$

4.3. IESA-CN: conceptual description

The key of IESA framework is applying EFPA, i.e. decoupling the system into independent Erlang B subsystems, to a surrogate model that has similar blocking probability with that of the true model. Doing so preserves the advantages of EFPA such as computational simplicity. Another advantage is that the IESA framework can apply to applications for which EFPA has been used. An illustration of the concept is depicted in Fig. 1.

More specifically, a surrogate model is designed to replace the non-hierarchical traffic structure in the true model with a hierarchical traffic structure based on congestion information received by a call when it overflows from one server

group to the other. In the case that a call is rejected admission at one BS (server group), due to sleeping or insufficient capacity, the overflow call is proactively made to leave the system with a certain probability, which depends on the system congestion level provided by the information exchange mechanism developed in the surrogate model (to be described in detail later). Such "quitter calls" are proactively made to leave the system without attempting all the remaining accessible BSs. These quitter calls have the attribute that they are very likely to be blocked if they were allowed to overflow and attempt other accessible BSs. As a result, the surrogate model has similar blocking probability to that of the true model, and importantly, it has far less mutual overflow (which is known to adversely affect the accuracy of blocking probability evaluation in EFPA). On the other hand, as there is a positive probability that one of the skipped-over BSs could have served those quitter calls if they would be allowed to attempt the remaining accessible BSs, the surrogate system will have a higher blocking probability than the true system, namely $B_S > B_T$.

Moreover, proactively giving up overflow traffic leads to a larger proportion of the total traffic offered to a BS formed by new traffic in the surrogate model as compared to the true model, while the proportion formed by overflowed traffic is decreased accordingly (a proof of this was provided in [48] for a special case). As a result, when an EFPA-based approximation (the "new estimation" in Fig. 1) is applied to the surrogate model, the approximation errors resulted from the Poisson and independence assumptions can be reduced. In this sense, the gap between "new estimation" and "surrogate model" in Fig. 1 is narrower than that between "true model" and "estimation by EFPA" (i.e. $B_S - B_N < B_T - B_E$). As we have $B_S > B_T$ and $B_S - B_N < B_T - B_E$, we can deduce that the new estimation will always obtain a higher blocking probability than the EFPA estimation (i.e. $B_N > B_E$). If we can choose an appropriate surrogate that has similar blocking probability with that of the true model so that the positive difference of B_S over B_T is not significant, the new approximation results will be closer to the real blocking probabilities than those by direct application of EFPA (i.e. $|B_T - B_N| < |B_T - B_E|$).

The "new estimation" in Fig. 1 was formally proposed as IESA in [22]. It has been proven in a simple overflow loss system that the blocking of the new estimation obtained by IESA is always between those of exact solution and EFPA, meaning that IESA is at least as good as EFPA in terms of accuracy [22, 60]. Moreover, under critical loading condition, it has been

proven that IESA is much more superior to EFPA [60].

The information exchange mechanism in the surrogate model for IESA, which entails the third attribute while retaining the first two attributes, was originally proposed in [22] in order to improve the accuracy of approximations in partially accessible overflow loss systems. The surrogate for IESA-CN is an adaptation to mobile cellular networks of the original IESA. In particular, IESA-CN includes modeling of locality and mobility that are unique features of cellular networks.

We now formally describe the surrogate for IESA-CN. A new call just initiated has $\Delta = \emptyset$ and $\Omega = 0$. When call ζ originated from BS m with attributes $I_\zeta, \Delta_\zeta, \Omega_\zeta$ arrives at i, it will be admitted if the BS still has vacant channels available. Otherwise, if the most senior call κ in service has $\Omega_\kappa < \Omega_\zeta$, the incoming call 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 $\Omega_\kappa \geq \Omega_\zeta$, call κ and call ζ will exchange their third attribute, Ω, before call ζ's overflow. In this way, the overflow call will have attributes $\{I_\zeta, \Delta_\zeta \cup i, \Omega_\kappa + 1\}$ and the call in service will have $\{I_\kappa, \Delta_\kappa, \Omega_\zeta\}$.

For a handover call ζ, the attributes Δ_ζ and Ω_ζ are reset to $\emptyset$ and 0, respectively upon a handover. This is because the original congestion information becomes irrelevant as the set of BSs that it can access also changes upon a handover. Note that the reset mechanism does not exist in the original IESA.

The additional attribute Ω represents an estimate of the number of busy BSs in the network. For every call, we have $|\Delta| \leq \Omega$ because the number of BSs that the call has already attempted (and overflowed from) is a lower bound for the estimate of the number of BSs that are busy in the network. In this way, an overflow call retains its identity (I) and actual overflow record (Δ) while gathering network congestion information (Ω) from other calls.

We introduce a special mechanism in IESA-CN to approximate the probability that all of the unattempted accessible BSs are not available. The mechanism uses the values of Δ and Ω of an overflow call. In the event that all of the unattempted BSs are presumed unavailable, the call will give up attempting the remaining BSs and will immediately be cleared out of the system. As in [22], we define $P_{k^*,|\Delta_\zeta|,\Omega_\zeta}$ as the probability of a call ζ with the attributes $\{I_\zeta, \Delta_\zeta, \Omega_\zeta\}$ gives up attempting in a system with parameter k^*. The parameter k^* is by definition the maximum allowable value of the attribute Ω of any call in the surrogate model and is a measure of the level of dependency in the real system ($k^* \leq n_i$ as $\Omega \leq |\Delta|$ at all times). $P_{k^*,|\Delta_\zeta|,\Omega_\zeta}$

is evaluated as:

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

where $|\Delta| \leq n_i \leq k^*$. From (14), one can infer that a call with a given value of attribute Ω is more unlikely to be blocked if the value of k^* is higher.

As the approximation results are affected by the design of the surrogate model, choosing an appropriate value of parameter k^* which can correctly reflect the level of dependency in the network and the ability to spread out congestion information is therefore crucial for the accuracy of the approximation under IESA-CN. A handover in cellular networks is generally considered an independent event as the sojourn time of a call in each cell/BS is often assumed to be exponentially distributed [34, 61, 62]. This is also one of the reasons that we reset the call attributes Ω and Δ upon a handover. Overflows, however, cause state dependencies among adjacent BSs and the ability to spread out congestion will affect those dependencies. On the other hand, this ability to spread out congestion depends on the degree of traffic overflow, which in turn depends on both traffic offered to each BS and on the mobility of the calls. More specifically, heavy traffic leads to more overflows and handovers, hence making congestion (as well as congestion information) easier to spread out around the network. As a result, a larger k^* value is required. Similarly, higher mobility of calls (higher handover rate) indicates more handovers during a call's lifetime, and as a result requires a larger k^* as well. This expectation is confirmed by numerical experiments presented later in Fig. 7.

Note that in previous work on approximations in VoD systems under the original IESA such as [22], [49] and [50], k^* is a constant equal to the total number of server groups. Therefore, the level of statistical dependency in such systems is rather fixed, and can be represented by a constant value of k^*. However, due to the locality and mobility features in cellular networks, we need to choose an appropriate value of k^* in IESA-CN for specific network conditions.

According to our tests, of the two features described above, mobility dominates over locality. If we consider a cellular network without any handover, the behavior of the system is similar to a VoD system where the value of k^*

is a constant for IESA. In such scenarios, the optimal k^* value is around the number of BSs within two hops distance as overflow alone (without handover) is unlikely to spread the congestion information beyond that scope.

However, if handovers exist, choosing an appropriate k^* value is crucial for designing a surrogate that can lead to accurate estimations of blocking probabilities. We use regression analysis to forecast the quasi-optimal k^*. The dependent variable is k^*, while the independent variables include the handover rate δ and the average offered traffic per active channel a_{avg} [63].

In practice, we can obtain the quasi-optimal value of k^* for different values of a_{avg} and δ as follows. For a particular cellular network model, a small set of independent cases with symmetric distribution of traffic offered to every BS can be used as the training set for prediction. The training and predicting processes can be done by, for example, the curve fitting toolbox of MATLAB. Then, we use IESA-CN algorithm described later and together with those predicted values of k^* in order to estimate the blocking probability for the general cases with arbitrary distribution of offered traffic. We acknowledge that using machine learning technique requires running simulations to obtain blocking probabilities for the training set. However, considering the difference in computational efficiency between approximation and simulation, this approach is still much more computationally efficient than obtaining blocking probabilities by simulation for every possible set of system parameters [64].

4.4. IESA-CN: detailed description

For IESA-CN, we define (using the superscript I to represent IESA-CN):

- $a^{\mathrm{I}}_{i,m,j,n,\boldsymbol{s}}$ – Traffic offered to BS i with n overflows ($|\Delta| = n$) and $\Omega = j$ from source m and have overflowed sequentially along the path $\boldsymbol{s} = s_1, s_2, ...s_n$ ($m \in \Phi_i$; $\boldsymbol{s} \subset \Gamma_m$; $n < n_m$; $j = 0, 1, ..., k^* - 1$).

- $a^{\mathrm{I}}_{i,j,n}$ – Traffic offered to i with n overflows and $\Omega = j$, namely summing all eligible $a^{\mathrm{I}}_{i,m,j,n,\boldsymbol{s}}$:

$$a^{\mathrm{I}}_{i,j,n} = \sum_{m \in \Phi_i; n < n_m} \sum_{\boldsymbol{s} \subset \Psi(\Gamma_m - \{i\}, n)} a^{\mathrm{I}}_{i,n,j,m,\boldsymbol{s}}. \tag{15}$$

- $\hat{a}^{\mathrm{I}}_{i,j,n}$ – Traffic offered to i with n overflows ($|\Delta| = n$) and Ω up to j, namely

$$\hat{a}^{\mathrm{I}}_{i,n,j} = \sum_{l=n}^{j} a^{\mathrm{I}}_{i,n,l}. \tag{16}$$

- $A^{\mathrm{I}}_{i,j}$ – Total combined traffic offered to i up to level j, namely

$$A^{\mathrm{I}}_{i,j} = \sum_{l=0}^{j} a^{\mathrm{I}}_{i,l}. \tag{17}$$

- $v^{\mathrm{I}}_{i,n,j,m,\boldsymbol{s}}$ – Overflow traffic from i with n overflows and $\Omega = j$ originated from m that have overflowed sequentially along the path $\boldsymbol{s} = s_1, s_2, ..., s_{n-1}, i$ ($m \in \Phi_i, \boldsymbol{s} \subset \Psi(\Gamma_m, n)$, $n \leq n_m$; $n \leq j \leq k^*$).

- $z^{\mathrm{I}}_{s_n,n,j,m,\boldsymbol{s}}$ – Blocked traffic (due to the special giving up mechanism in IESA-CN) from i with n overflows and $\Omega = j$ originated from m that have overflowed sequentially along the path $\boldsymbol{s} = s_1, s_2, ..., s_{n-1}, s_n$ ($m \in \Phi_i, \boldsymbol{s} \subset \Psi(\Gamma_m, n)$, $n \leq n_m$; $j \leq k^*$).

- $B^{\mathrm{I}}_{i,j}$ – Probability that all channels in i are busy at level j serving calls with $|\Delta| \leq min(j, \hat{n}_i - 1)$ and $|\Delta| \leq \Omega < j$.

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

Also by the Erlang B formula, we can obtain the relationship between $B^I_{i,j}$ and $A^I_{i,j}$ at each level j as

$$B^{\mathrm{I}}_{i,j} = \begin{cases} \mathbf{E}(A^{\mathrm{I}}_{i,j}, c_i) & \text{for all } i \text{ with } S_i = 1; \\ 1 & \text{for all } i \text{ with } S_i = 0, \end{cases} \tag{18}$$

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

We analyze the origin of overflow traffic $v^{\mathrm{I}}_{s_n,n,j,m,\boldsymbol{s}}$ for two scenarios. Firstly, with probability $B^{\mathrm{I}}_{s_n,j-1} - B^{\mathrm{I}}_{s_n,j-2}$, all channels of s_n at level $j-1$ are not available. Equivalently, all channels are serving calls with seniority up to $\Omega = j - 1$. In this scenario, the traffic $\hat{a}^{\mathrm{I}}_{s_n,n-1,j-2,m,\boldsymbol{s}-\{s_n\}}$ with $\Omega \leq j-2$ offered to s_n will overflow with information exchange (with the most senior call with $\Omega = j - 1$) and thus forms the overflow traffic $v^{\mathrm{I}}_{s_n,n,j,m,\boldsymbol{s}}$. On the other hand, with probability $B^{\mathrm{I}}_{s_n,j-1}$, all channels of s_n at level $j-1$ are busy serving calls with $\Omega \leq j-1$. In this scenario, the offered traffic $a^{\mathrm{I}}_{s_n,n-1,j-1,m,\boldsymbol{s}-\{s_n\}}$ to s_n simply overflow without information exchange and also contributes to the overflow traffic $v^{\mathrm{I}}_{s_n,n,j,m,\boldsymbol{s}}$. Thus, for $j = 1, 2, ..., k^*$, we derive $v^{\mathrm{I}}_{s_n,n,j,m,\boldsymbol{s}}$ as

$$\begin{aligned} v^{\mathrm{I}}_{s_n,n,j,m,\boldsymbol{s}} &= \hat{a}^{\mathrm{I}}_{s_n,n-1,j-2,m,\boldsymbol{s}-\{s_n\}}(B^{\mathrm{I}}_{s_n,j-1} - B^{\mathrm{I}}_{s_n,j-2}) + a^{\mathrm{I}}_{s_n,n-1,j-1,m,\boldsymbol{s}-\{s_n\}} B^{\mathrm{I}}_{s_n,j-1} \\ &= \hat{a}^{\mathrm{I}}_{s_n,n-1,j-1,m,\boldsymbol{s}-\{s_n\}} B^{\mathrm{I}}_{s_n,j-1} - \hat{a}^{\mathrm{I}}_{s_n,n-1,j-2,m,\boldsymbol{s}-\{s_n\}} B^{\mathrm{I}}_{s_n,j-2}. \end{aligned} \tag{19}$$

Referring back to (14), with a probability of $P_{k^*,n,j}$, the overflow traffic $v^{\mathrm{I}}_{s_n,n,j,m,\boldsymbol{s}}$ is prevented from further hunting for available BSs even if it has not yet attempted all BSs in Γ_m. On the other hand, if the overflow traffic has $|\Delta| = n_m$, i.e., has already attempted all accessible BSs in Γ_m, or the exchanged information indicates that no BSs is possibly available ($\Omega = k^*$), the probability $P_{k^*,n_m,j}$ will be equal to 1. This ensures that calls with $|\Delta| = n_m$ or $\Omega = k^*$ are always immediately cleared out. As defined previously, traffic blocked in this manner will become $z^{\mathrm{I}}_{s_n,n,j,m,\boldsymbol{s}}$, namely

$$z^{\mathrm{I}}_{s_n,n,j,m,\boldsymbol{s}} = v^{\mathrm{I}}_{s_n,n,j,m,\boldsymbol{s}} P_{k^*,n,j}. \tag{20}$$

On the other hand, with probability $1-P_{k^*,n,j}$, the overflow traffic $v^{\mathrm{I}}_{s_n,n,j,m,\boldsymbol{s}}$ will continue to attempt another BS in $\Gamma_m - \boldsymbol{s}$ as in EFPA. Every $i \in \Gamma_m - \boldsymbol{s}$ will be chosen with probability $\frac{1}{n_m-1}$. Accordingly, we have

$$a^{\mathrm{I}}_{i,n,j,m,\boldsymbol{s}} = \frac{v^{\mathrm{I}}_{s_n,n,j,m,\boldsymbol{s}}(1 - P_{k^*,n,j})}{n_m - n}. \tag{21}$$

We can then compute $A^{\mathrm{I}}_{i,j}$ and $B^{\mathrm{I}}_{i,j}$ at each level iteratively based on (16), (17) and (21).

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, $A^{\mathrm{I}}_{i,k^*-1}(1 - B^{I}_{i,k^*-1})$ is the total carried traffic by 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-CN as:

$$\hat{B}^{\mathrm{I}} = 1 - \frac{\sum_{i\in U} A^{\mathrm{I}}_{i,k^*-1}(1 - B^{\mathrm{I}}_{i,k^*-1})}{A}. \tag{22}$$

The blocking probability for calls originated from BS m can be calculated by summing all $z^{\mathrm{I}}_{i,n,j,m,\boldsymbol{s}}$ together, namely

$$\hat{B}^{\mathrm{I}}_m = \sum_{n=0}^{n_m} \sum_{\boldsymbol{s}\subset\Psi(\Gamma_m,n)} \sum_{j=n_m}^{k^*} z^{\mathrm{I}}_{s_{n_m},m,j,n_m,\boldsymbol{s}}. \tag{23}$$

Note that both (12) and (22) does not take the mobility effect into account. We can constitute a set of fixed-point equations by combining (2), (3) and an equation to calculate $\hat{B}$ in terms of A_i for each method, namely, (12) for EFPA and (22) for IESA-CN, to calculate the blocking probability with consideration of the mobility effect.

5. Numerical Results

In this section, we present numerical results for the model described in Section 3 to demonstrate the accuracy, versatility, and computational efficiency of IESA framework.

We consider a cellular network model with 49 interconnected and wrapped-around hexagonal cells as shown in Fig. 2, and each cell is served by a single BS. The wrapped-around design avoids boundary effect and has been popular in cellular network research (see e.g., [65–67]). Assume that each BS has 10 channels, a Markov Chain for the model would have a state space of 10^{49}, which is computationally prohibitive.

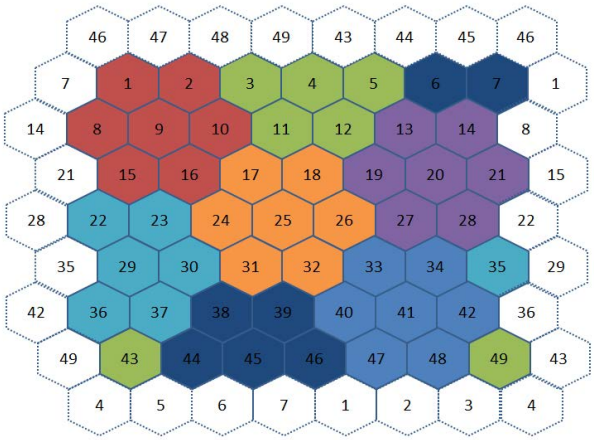
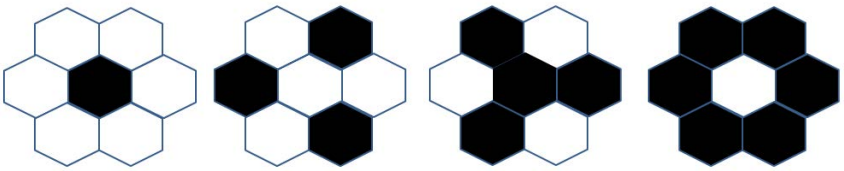


Figure 2: 49-cell hexagonal configuration network model with wrapped-around design.

We have conducted extensive numerical experiments under a wide range of system parameters. The approximation results by IESA-CN and EFPA are compared with simulation results serving as a benchmark, which are obtained by MATLAB in the form of an observed mean from multiple independent runs. We use simulation results as the benchmark as no exact analytical results are available for our model. The confidence intervals are at the 95% level based on the Student's t-distribution. Markov chain simulation is used for the cases where both service time and sojourn time are exponentially distributed, while discrete event simulation is used for the other cases.

We compute the error between the approximation and the simulation in terms of the relative error. Given an approximation result r and a simulation result s, the relative error is $(r - s)/s$. Note that our choice here of a linear



(a) Pattern A (b) Pattern B (c) Pattern C (d) Pattern D

Figure 3: BS sleeping patterns

scale for the relative difference is made for convenience of illustration. Other relevant alternatives such as log scale are also acceptable for assessing the errors.

For simplicity without loss of generality, in this paper we consider four different BS sleeping patterns based on the 7-cell cluster. The 49-cell network model can be decomposed into 7 identical 7-cell clusters as shown in Fig. 2. Each pattern switches different number of BSs in a cluster to sleep mode [23]. As shown in Fig. 3, a dark cell indicates that the BS serving the cell is in sleep mode, while a light cell denotes that the BS is active. All patterns have at least one active BS next to a BS in sleep mode, which ensures that traffic arriving at any sleeping BS could be served by a neighboring active BS.

In this paper, we consider fixed sleeping schemes to compare the performance of EFPA and IESA-CN in approximating the blocking probability. This case can be extended to a dynamic case if the time spent under each sleeping scheme is sufficiently long. The long time duration of being in a given sleeping scheme can be justified by practical considerations associated with transition time requirements for BSs to switch between active and sleep modes. Then, an approximation for the overall blocking probability can be obtained by a weighted average of the individual cases. See equivalent discussion in Case 2A in [68].

To demonstrate that our approach could be applied to cellular network models with general (i.e., asymmetric or unbalanced) distribution of offered traffic, we designate one of seven clusters in the 49-cell model as the "hot" cluster [4, 51, 69]. The BSs in the hot cluster are offered heavier traffic than the rest of the network. Such spatial traffic distributions can be found in both research work and practical scenarios [4, 51, 69]. We denote A_n as the traffic offered to each BS not in the hot cluster, and α as the ratio of traffic offered to each BS in the hot cluster to traffic offered to each BS outside the hot cluster. The traffic offered to each BS in the hot cluster is thus αA_n.

Note that Pattern D (6 out of 7 BSs are in sleep mode) will simplify the 49-cell system to seven isolated clusters, because a call arriving at a BS in sleep mode has only one active BS to overflow to. Therefore, in this case, the approximation results of EFPA and IESA-CN are identical as the information exchange mechanism cannot be activated for the IES surrogate.

5.1. Power savings of switching patterns

Power consumption of a base station comprises of two parts, namely traffic load dependent power consumption such as power amplifiers, and static power consumption such as air conditioning which is consumed as long as the BS is active [36].

Moreover, if a BS extends its coverage to serve customers originally associated with another BS that has been switched to sleep mode, it will consume more power to serve the users that are relatively far away due to the path-loss effect.

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

$$P_{BS} = \begin{cases} P_{static} + \tau P_v^{max} + \hat{\tau}\hat{P}_v^{max} & \text{when active,} \\ \\ P_{sleep} & \text{when sleeping,} \end{cases} \tag{24}$$

where τ and $\hat{\tau}$ are the loading of local traffic and traffic transferred from neighboring sleeping BSs, respectively. P_{static} represents static power consumption, τP_v^{max} represents traffic load dependent power consumption attributed to local traffic, and $\hat{\tau}\hat{P}_v^{max}$ represents variable power consumption attributed to transferred traffic from sleeping BSs. The exact difference between $\hat{P}_v^{max}$ and P_v^{max} depends on various factors such as the path-loss exponent, inter-distance of BSs and distribution of user locations [36, 70].

Assume that $P_{sleep} = 1W$, $P_{static} = 100W$, $P_v^{max} = 160W$ and $\hat{P}_v^{max} = 190W$, average power consumptions of Patterns A, B and C and the case where all BSs are kept active are depicted in Fig. 4. When offered traffic per BS is 2 to 3 Erlangs as shown in the figure, up to 50% power consumption can be saved if Pattern C is chosen. Meanwhile, as we will show later in this section, all three patterns (A, B and C) can maintain the blocking probability below 10^{-2}, which is an acceptable level for cellular networks [23, 42, 52, 66].

5.2. Insensitivity of service and sojourn time distributions

Here we aim to examine the sensitivity of the blocking probability to the service or sojourn time distributions. To this end, we consider three dis-

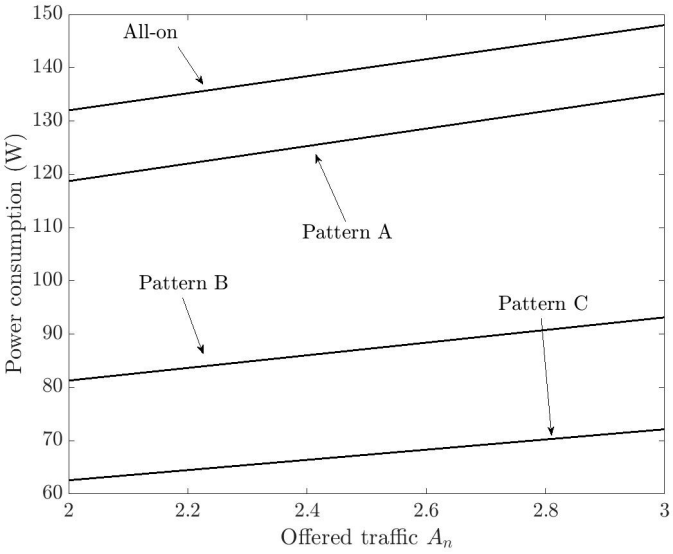


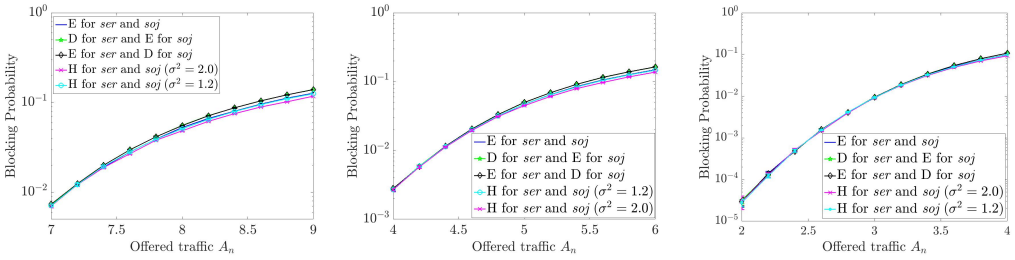
Figure 4: Average power consumptions achievable for different switching patterns.

tributions, namely, exponential (the most common assumption for cellular networks, with a variance of 1.0), deterministic and hyperexponential (with variances of 1.2 and 2.0) distributions. In Fig. 5 we demonstrate that the blocking probabilities are nearly insensitive [55] to the shape of the distributions of either service or sojourn time. This suggests that our proposed approximation method can be applied to systems with non-exponential distribution of service and sojourn time.

5.3. Numerical evidence of Fig. 1

In Fig. 6, we present blocking probabilities of the true and surrogate models obtained by simulation as well as approximation results obtained by EFPA and IESA-CN, respectively. The results confirm our conceptual illustration depicted by Fig. 1. As discussed previously, the surrogate model has relatively higher blocking probability than the true model while the approximations underestimate the blocking probabilities. These two effects appear to compensate each other. Therefore, as shown in the figure, IESA-CN reduces the approximation error as compared to the EFPA.

Please note that only in Fig. 6 we distinguish between "True model Simulation" and "Surrogate model simulation". This is the only figure where we provide numerical support for the conceptual illustration presented in Fig. 1,



(a) Pattern A (b) Pattern B (c) Pattern C

Figure 5: Simulation results of blocking probabilities with different distributions of service time (*ser*) and sojourn time (*soj*); D, E, and H represent deterministic, exponential and hyperexponential distributions, respectively; $\mu = 1, \delta = 1, \alpha = 1.2$.

so the simulation results for the surrogate model are only provided in Fig. 6. In all other figures, the term "simulation" refers to the simulation results of the true model.

5.4. Relationship between the value of k^* and approximation result

As mentioned previously, the optimal value of the parameter k^* in IESA-CN, which is an estimate of the maximum number of BSs that a call has access to, is influenced by the handover rate in the network. Therefore, we firstly present the relationship between k^* value and the approximation results in Fig. 7a for $A_n = 8, \delta = 0$, Fig. 7b for $A_n = 8, \delta = 1$ and Fig. 7c for $A_n = 9, \delta = 1$ in the 49-cell model with all BSs active. The values of the other input parameters are as shown in the captions of the figures.

In all three figures, the inverse relationship between k^* and approximated blocking probability is consistent with Eqn. (14). As the handover rate or offered traffic increases, the number of BSs that a call is expected to visit is likely to increase. In line with our expectation that the network congestion is easier to occur and to spread out to the entire network if a typical call visits more cells during its lifetime, the k^* value that gives the most accurate estimation result also increases from 9 in Fig. 7a to 11 in Fig. 7b and 25 in Fig. 7c. Meanwhile, as the value of k^* increases, the approximation result by IESA-CN approaches that by EFPA. The intuitive explanation is that when k^* is large, the giving up probability obtained by Eqn. (14) approaches zero. Without the giving up mechanism, the surrogate model is the same as the true model and thus the approximation results by applying IESA-CN and EFPA are identical.

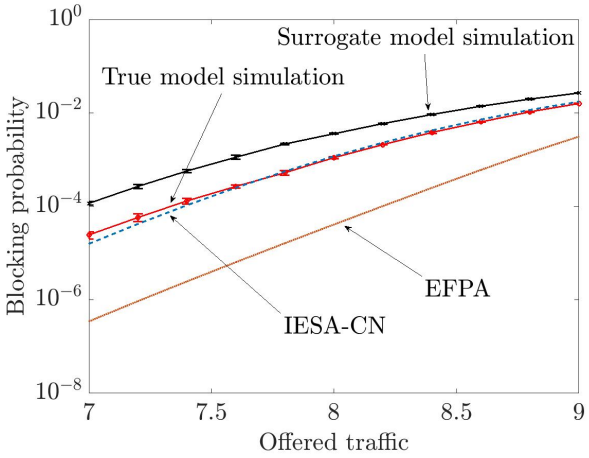
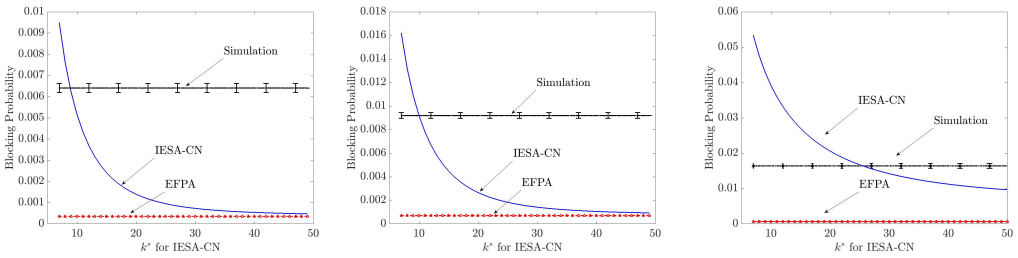


Figure 6: Simulation and approximation results for true and surrogate models ($\mu = 1, \delta = 0, \alpha = 1$).



(a) $A_n = 8, \delta = 0$ (b) $A_n = 8, \delta = 1$ (c) $A_n = 9, \delta = 1$

Figure 7: The choice of the parameter k^* in IESA-CN ($\mu = 1, \alpha = 1$).

5.5. Accuracy of approximations in different network setups

In Fig. 8 – Fig. 11, we demonstrate the sensitivity of accuracy of EFPA and IESA-CN to input parameters including arrival rate λ, handover rate δ and level of asymmetrical traffic distribution α. The curve "simulation" in Fig. 8 represents the simulation results of the true model (corresponding to the "true model simulation" curve in Fig. 6) and is the benchmark for assessing the relative errors in Fig. 9 – Fig. 11. From the results, we observe consistency with our discussion earlier in the chapter. EFPA significantly underestimates blocking probabilities in most cases due to the Poisson and independent assumptions, while IESA-CN significantly improves the accuracy of approximations for Patterns A, B and C. In addition, IESA-CN

provides a conservative estimation as it gives blocking probabilities that are higher than actual values, which is preferable and often adopted for the purpose of network design [64]. On the other hand, both IESA-CN and EFPA are quite accurate for Pattern D as shown in Figs. 8c and 9c, where there is no mutual traffic overflow so that the independence and Poisson errors due to overflow traffic do not exist. We also show the results for networks with all BS turned on in Figs. 10a and 11a. We see that IESA-CN is also accurate in a general cellular network with no BS in sleep mode. For the parameter sets considered in Figs. 8 – 11, we also performed similar runs for Pattern B and the results were very similar to those for Pattern C presented in Figs 8b, 9b, 10c and 11c, respectively.

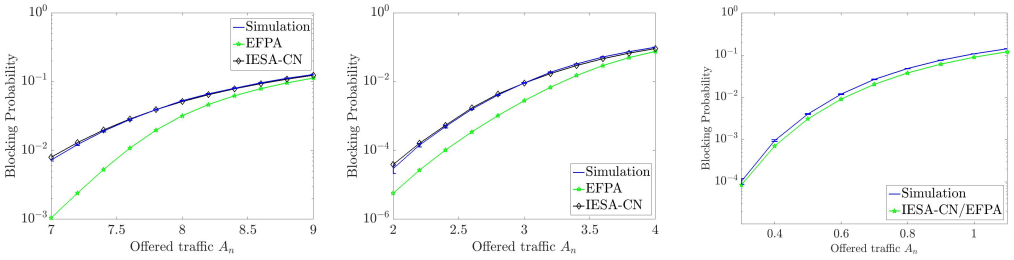
As demonstrated in this section, for all cases studied the proposed IESA-CN is consistently more accurate than EFPA. While, despite that the computational efficiency of IESA-CN is not as well as EFPA, it is much more efficient than the simulation method. Note that for almost all the cases studied the errors of the system and hot-cluster blocking probabilities estimated by IESA-CN are within 20% of the midpoint of the 95% confidence interval of the simulation results based on the Student's-t distribution.

Moreover, blocking probabilities in the range 10^{-3} – 10^{-2} is considered practical for cellular networks and of particular interest of existing research (e.g., [23, 42, 52, 66]). As shown in the figures in this section, the accuracy of IESA is particularly high for the cases where the blocking probability is in this range.

In addition, IESA-CN is a conservative estimation which gives blocking probabilities that are higher than actual values in most cases. In contrast, EFPA is an aggressive estimation which gives lower-than-actual estimations. In many engineering applications such as network planning, conservative estimations are normally more desirable.

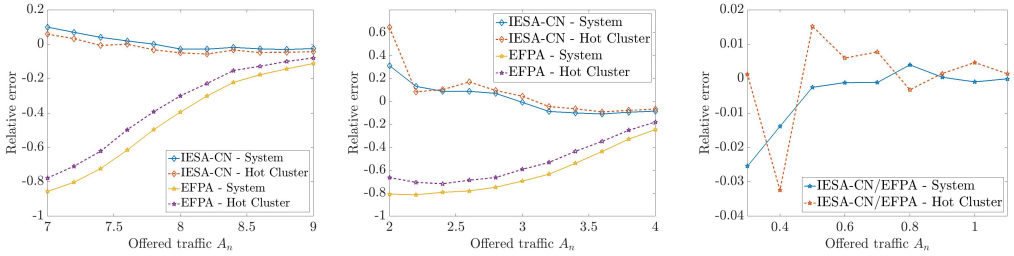
5.6. Non-Poisson arrivals

In Fig. 12, we demonstrated by simulation that if the arrival process is MMPP, the blocking probability will be slightly higher than Poisson arrival with the same offered traffic. Furthermore, as shown in Fig. 12, we note that IESA-CN is still a fairly accurate approximation even if the arrival process is MMPP. To the best of our knowledge, no analytical results are available for MMPP arrivals so far. Therefore, it is desirable to use IESA-CN as an estimation tool for systems with MMPP arrivals when simulation results are not available.



(a) Pattern A (b) Pattern C (c) Pattern D

Figure 8: Simulation and approximation results of system blocking probabilities with different offered traffic A_n ($\mu = 1, \delta = 1, \alpha = 1.2$).



(a) Pattern A (b) Pattern C (c) Pattern D

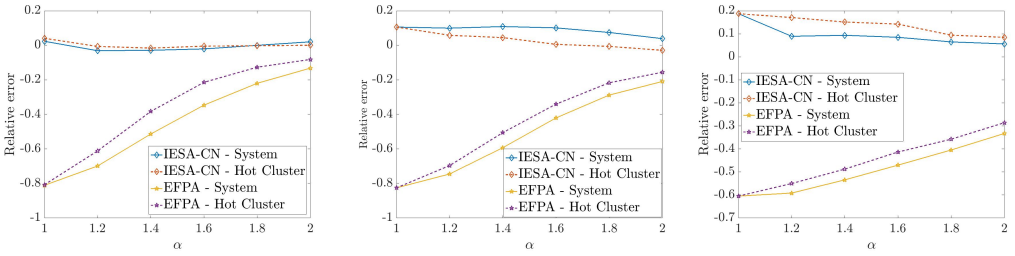
Figure 9: Relative errors of blocking probabilities with different offered traffic ($\mu = 1, \delta = 1, \alpha = 1.2$).

5.7. Irregular network topology

In addition to networks with homogeneous BS layout as in Fig. 2, we demonstrate the approximation results of IESA-CN in a network with irregular topology based on Poisson distributed BSs (e.g. [71]) as shown in Fig. 13. Here we assume that the offered traffic to each BS is the same and equal to A_n. We consider two cases, one with all BSs active and the other with one BS sleeping (the BS shown with a red "X" in Fig. 13). The results are shown in Figs. 14a and 14b, respectively. For both cases, IESA-CN is demonstrated to provide reasonably accurate and relatively conservative estimations of blocking probabilities.

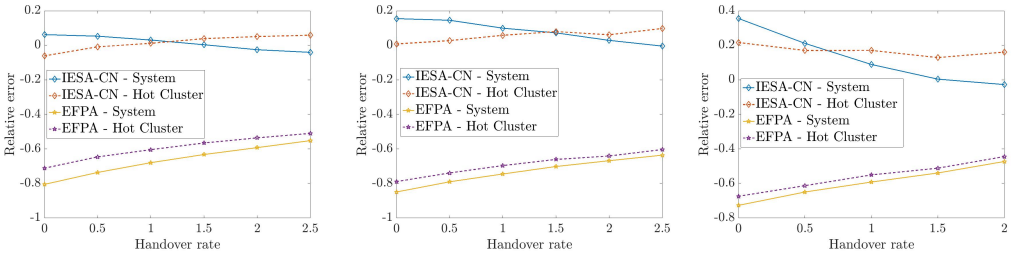
5.8. CPU running time

Table 1 shows the CPU running time of each evaluation method for selected switching patterns. The running time of IESA-CN algorithm is approximately two orders of magnitude higher than that of EFPA due to the



(a) All-on case ($A_n = 8.6$) (b) Pattern A ($A_n = 7$) (c) Pattern C ($A_n = 2.6$)

Figure 10: Relative errors of blocking probability approximations with different levels of asymmetric traffic distribution ($\mu = 1, \delta = 1$).



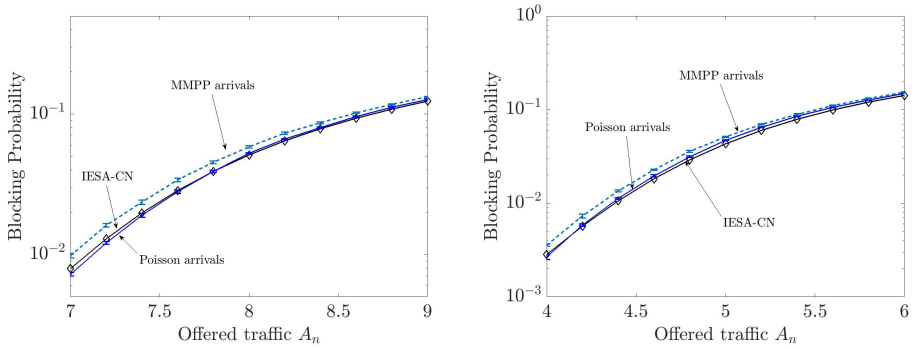
(a) All-on case ($A_n = 8.6$) (b) Pattern A ($A_n = 7$) (c) Pattern C ($A_n = 2.6$)

Figure 11: Relative errors of blocking probability approximations with different handover rates ($\mu = 1, \alpha = 1.2$).

additional computations required for hierarchical application of EFPA to the surrogate model. However, considering the improvement in accuracy and the fact that EFPA is extremely fast, this increase in running time is acceptable. On the other hand, IESA-CN is much faster than the Markov chain simulation (which is faster than the discrete event simulation).

6. Conclusions

IESA is a versatile and promising framework proposed to improve the accuracy of the conventional EFPA in order to address the challenging problem of blocking probability estimation in partially accessible overflow loss systems such as cellular networks. Due to its root from EFPA, IESA framework can apply to a wide range of applications including those for which EFPA has been used. In this work, we have proposed a suitable surrogate and the corresponding IESA-CN method under this framework for a cellular network



(a) Pattern A (b) Pattern B

Figure 12: Blocking probability comparison of Poisson and MMPP arrivals ($\mu = 1, \delta = 1, \alpha = 1.2$)

model with (or without) BS sleeping. Unlike the original IESA, IESA-CN captures the *locality* and *mobility* features uniquely present in cellular networks. Numerical results have confirmed that our approximation under this surrogate is accurate, robust and computationally efficient considering the available alternatives: EFPA, exact Markov chain solution or computer simulation. We have also demonstrated that our approximation is particularly accurate in the blocking probability range normally used in practice.

To the best of our knowledge, our proposed approximation framework is the first workable approach in terms of accuracy and computational efficiency for cellular networks with (or without) BS sleeping operation, asymmetric offered traffic across BSs and call mobility. The framework can then be used for a number of applications including network design, admission control and resource allocation, which we leave for future research. We also acknowledge the modern and future technological developments always give rise to further model extensions. In our case, there is a scope for extending the model to include other considerations, such as, scenarios with different kinds of traffic and heterogeneous cellular networks with multi-layer topology. These extensions may be addressed by incorporate the IESA framework with other approximation techniques such as moment matching, and other teletraffic models such as multi-service multi-rate queues. Incorporation of such issues in the model is planned for future work.

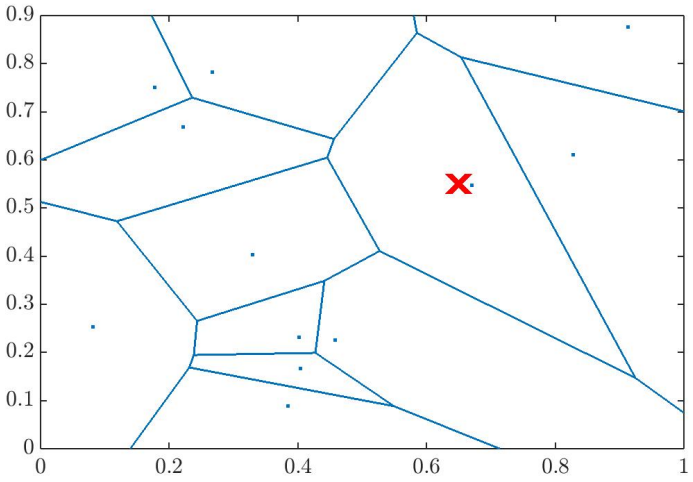


Figure 13: Poisson distributed BSs. The cell boundaries are shown and form a Voronoi tessellation.

Acknowledgment

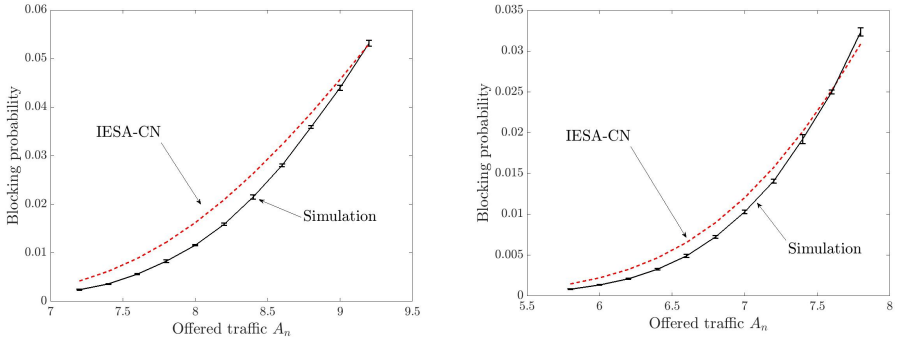
The work described in this paper was partly supported by College Research Grant of Beijing Normal University-Hong Kong Baptist University United International College [UIC-R201703] and grants from the Hong Kong Innovation and Technology Funding (ITF) [ITS/191/16].

References

[1] M. Ismail, W. Zhuang, E. Serpedin, K. Qaraqe, A survey on green mobile networking: From the perspectives of network operators and mobile users, IEEE Communications Surveys and Tutorials 17 (3) (2015) 1535–1556.

[2] J. Wu, Y. Zhang, M. Zukerman, E. K.-N. Yung, Energy-efficient base-stations sleep-mode techniques in green cellular networks: A survey, IEEE Communications Surveys and Tutorials 17 (2) (2015) 803–826.

[3] S. Tombaz, K. W. Sung, S.-W. Han, J. Zander, An economic viability analysis on energy-saving solutions for wireless access networks, Computer Communications 75 (2016) 50 – 61.



(a) All BSs active. (b) One BS in sleep.

Figure 14: Blocking probability approximation for the network depicted in Figure 13 with Poisson distributed BSs ($\mu = 1, \delta = 1$).

[4] D. Lee, S. Zhou, X. Zhong, Z. Niu, X. Zhou, H. Zhang, Spatial modeling of the traffic density in cellular networks, IEEE Wireless Communications 21 (1) (2014) 80–88.

[5] O. Arnold, F. Richter, G. Fettweis, O. Blume, Power consumption modelling of different base station types in heterogeneous cellular networks, in: Future Network and Mobile Summit, Dresden, Germany, 2010, pp. 1 – 8.

[6] J. Wu, S. Zhou, Z. Niu, Traffic-aware base station sleeping control and power matching for energy-delay tradeoffs in green cellular networks, IEEE Transactions on Wireless Communications 12 (8) (2013) 4196–4209.

[7] J. Gong, J. Thompson, S. Zhou, Z. Niu, Base station sleeping and resource allocation in renewable energy powered cellular networks, IEEE Transactions on Communications 62 (11) (2014) 3801–3813.

[8] Z. Niu, X. Guo, S. Zhou, P. Kumar, Characterizing energy-delay tradeoff in hyper-cellular networks with base station sleeping control, IEEE Journal on Selected Areas in Communications 33 (4) (2015) 641–650.

[9] H. Chen, Q. Zhang, F. Zhao, Energy-efficient joint BS and RS sleep scheduling in relay-assisted cellular networks, Computer Networks 100 (2016) 45 – 54.

Method	**Pattern**	**Running time**	**No. of iterations**
EFPA	All-on	1.29s	48
IESA-CN	All-on	118.96s	35
Markov chain simulation	All-on	2.95h	N/A
EFPA	A	0.97s	38
IESA-CN	A	107.20s	36
Markov chain simulation	A	2.96h	N/A
EFPA	B	0.69s	30
IESA-CN	B	71.63s	22
Markov chain simulation	B	2.95h	N/A
EFPA	C	0.31s	30
IESA-CN	C	36.29s	20
Markov chain simulation	C	2.77h	N/A

Table 1: CPU running time of evaluation methods.

[10] M. K. Karray, Analytical evaluation of QoS in the downlink of OFDMA wireless cellular networks serving streaming and elastic traffic, IEEE Transactions on Wireless Communications 9 (5) (2010) 1799–1807.

[11] A. Antonopoulos, C. Verikoukis, Traffic-aware connection admission control scheme for broadband mobile systems, IEEE Communications Letters 14 (8) (2010) 719–721.

[12] R. C. McNamara, Applications of spanning trees to continuous-time Markov processes, with emphasis on loss systems, Ph.D. thesis, University of Colorado (2004).

[13] Y. Huang, X. Zhang, J. Zhang, J. Tang, Z. Su, W. Wang, Energy-efficient design in heterogeneous cellular networks based on large-scale user behavior constraints, IEEE Transactions on Wireless Communications 13 (9) (2014) 4746–4757.

[14] H. Tabassum, U. Siddique, E. Hossain, M. Hossain, Downlink performance of cellular systems with base station sleeping, user association, and scheduling, IEEE Transactions on Wireless Communications 13 (10) (2014) 5752–5767.

[15] I. Katzela, M. Naghshineh, Channel assignment schemes for cellular mobile telecommunication systems: a comprehensive survey, IEEE Personal Communications 3 (3) (1996) 10–31.

[16] National Telecommunication Information Administration, Telecommunications: Glossary of Telecommunications Terms, Government Institutes, 1997.

[17] R. I. Wilkinson, Theories for toll traffic engineering in the U. S. A., Bell System Technical Journal 35 (2) (1956) 421–514.

[18] J. Matsumoto, Y. Watanabe, Individual traffic characteristics queueing systems with multiple Poisson and overflow inputs, IEEE Transactions on Communications 33 (1) (1985) 1–9.

[19] A. Kuczura, The interrupted poisson process as an overflow process, Bell System Technical Journal 52 (3) (1973) 437–448.

[20] Q. Huang, K.-T. Ko, V. Iversen, Approximation of loss calculation for hierarchical networks with multiservice overflows, IEEE Transactions on Communications 56 (3) (2008) 466–473.

[21] E. Karasan, E. Ayanoglu, Effects of wavelength routing and selection algorithms on wavelength conversion gain in WDM optical networks, IEEE/ACM Transactions on Networking 6 (2) (1998) 186–196.

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

[23] F. Han, Z. Safar, K. Liu, Energy-efficient base-station cooperative operation with guaranteed QoS, IEEE Transactions on Communications 61 (8) (2013) 3505–3517.

[24] G. R. Ash, Dynamic Routing in Telecommunications Networks, McGraw-Hill, 1997.

[25] S. Li, M. Wang, E. W. M. Wong, V. Abramov, M. Zukerman, Bounds of the overflow priority classification for blocking probability approximation in OBS networks, IEEE/OSA Journal of Optical Communications and Networking 5 (4) (2013) 378–393.

[26] Z. Rosberg, A. Zalesky, H. Vu, M. Zukerman, Analysis of OBS networks with limited wavelength conversion, IEEE/ACM Transactions on Networking 14 (5) (2006) 1118–1127.

[27] J. Guo, E. W. M. Wong, S. Chan, P. Taylor, M. Zukerman, K. S. Tang, Performance analysis of resource selection schemes for a large scale video-on-demand system, IEEE Transactions on Multimedia 10 (1) (2008) 153–159.

[28] G. Koole, J. Talim, Exponential approximation of multi-skill call centers architecture, in: Proc. QNETs, 2000, pp. 23–32.

[29] N. Litvak, M. van Rijsbergen, R. J. Boucherie, M. van Houdenhoven, Managing the overflow of intensive care patients, European Journal of Operational Research 185 (3) (2008) 998–1010.

[30] N. M. van Dijk, E. van der Sluis, Call packing bound for overflow loss systems, Performance Evaluation 66 (1) (2009) 1 – 20.
URL http://www.sciencedirect.com/science/article/pii/S0166531608000564

[31] A. Farbod, B. Liang, Structured admission control policy in heterogeneous wireless networks with mesh underlay, in: Proc. IEEE INFOCOM 2009, 2009, pp. 495–503.

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

[33] F. Kelly, Blocking probabilities in large circuit-switched networks, Advances in Applied Probability 18 (1986) 473–505.

[34] J. Wu, J. Guo, E. W. M. Wong, M. Zukerman, Approximation of blocking probabilities in mobile cellular networks with channel borrowing, in: Proc. IEEE HPSR 2015, Budapest, Hungary, 2015.

[35] M. Marsan, L. Chiaraviglio, D. Ciullo, M. Meo, Multiple daily base station switch-offs in cellular networks, in: Fourth International Conference on Communications and Electronics (ICCE), Hue, Vietnam, 2012, pp. 245–250.

[36] A. Bousia, A. Antonopoulos, L. Alonso, C. Verikoukis, Green distance-aware base station sleeping algorithm in LTE-advanced, in: Proc. IEEE ICC 2012, 2012, pp. 1347–1351.

[37] D. Cao, S. Zhou, Z. Niu, Optimal combination of base station densities for energy-efficient two-tier heterogeneous cellular networks, IEEE Transactions on Wireless Communications 12 (9) (2013) 4350–4362.

[38] M. Oikonomakou, A. Antonopoulos, L. Alonso, C. Verikoukis, Cooperative base station switching off in multi-operator shared heterogeneous network, in: Proc. IEEE GLOBECOM 2015, 2015, pp. 1–6.

[39] A. Bousia, E. Kartsakli, A. Antonopoulos, L. Alonso, C. Verikoukis, Game-theoretic infrastructure sharing in multioperator cellular networks, IEEE Transactions on Vehicular Technology 65 (5) (2016) 3326–3341.

[40] A. Bousia, E. Kartsakli, A. Antonopoulos, L. Alonso, C. Verikoukis, Multiobjective auction-based switching off scheme in heterogeneous networks to bid or not to bid?, IEEE Transactions on Vehicular Technology 65 (11) (2016) 9168–9180.

[41] P. Raymond, Performance analysis of cellular networks, IEEE Transactions on Communications 39 (12) (1991) 1787 – 1793.

[42] X. Lagrange, P. Godlewski, Teletraffic analysis of a hierarchical cellular network, in: IEEE 45th Vehicular Technology Conference, Vol. 2, 1995, pp. 882–886.

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

[44] D. Everitt, Traffic engineering of the radio interface for cellular mobile networks, Proceedings of the IEEE 82 (9) (1994) 1371–1382.

[45] K. Mitchell, K. Sohraby, An analysis of the effects of mobility on bandwidth allocation strategies in multi-class cellular wireless networks, in: Proc. IEEE INFOCOM 2001, 2001.

[46] A. A. Fredericks, Congestion in blocking systems a simple approximation technique, Bell Syst. Tech. J. 59 (1980) 805–827.

[47] J. M. Holtzman, Analysis of dependence effects in telephone trunking networks, Bell Syst. Tech. J. 50 (1971) 2647–2662.

[48] E. W. M. Wong, A. Zalesky, Z. Rosberg, M. Zukerman, A new method for approximating blocking probability in overflow loss networks, Computer Networks 51 (2007) 2958–2975.

[49] Y. C. Chan, J. Guo, E. W. M. Wong, M. Zukerman, Performance analysis for overflow loss systems of processor-sharing queues, in: Proc. IEEE INFOCOM 2015, Hong Kong, 2015.

[50] Y. C. Chan, J. Guo, E. W. M. Wong, M. Zukerman, Surrogate models for performance evaluation of multi-skill multi-layer overflow loss systems, Performance Evaluation 104 (2016) 1–22.

[51] D. W. McMillan, Traffic modelling and analysis for cellular mobile networks, Ph.D. thesis, University of Melbourne (1993).

[52] M. Sidi, D. Starobinski, New call blocking versus handoff blocking in cellular networks, Wireless networks 3 (1997) 15–27.

[53] V. G. Vassilakis, M. D. Logothetis, The wireless Engset multi-rate loss model for the handoff traffic analysis in W-CDMA networks, in: 2008 IEEE 19th International Symposium on Personal, Indoor and Mobile Radio Communications, 2008, pp. 1–6.

[54] M. Wang, S. Li, E. W. M. Wong, M. Zukerman, Performance analysis of circuit switched multi-service multi-rate networks with alternative routing, Journal of Lightwave Technology 32 (2) (2014) 179–200.

[55] V. Gupta, M. Harchol Balter, K. Sigman, W. Whitt, Analysis of join-the-shortest-queue routing for web server farms, Performance Evaluation 64 (9-12) (2007) 1062–1081.

[56] R. Ramjee, R. Nagarajan, D. Towsley, On optimal call admission control in cellular networks, in: Proc. IEEE INFOCOM '96, Vol. 1, 1996, pp. 43–50 vol.1.

[57] G. Foschini, B. Gopinath, Z. Miljanic, Channel cost of mobility, IEEE Transactions on Vehicular Technology 42 (1993) 4.

[58] M. Zukerman, Introduction to queueing theory and stochastic teletraffic models.
URL http://www.ee.cityu.edu.hk/ zukerman/classnotes.pdf

[59] A. G. Hart, S. Martinez, Sequential iteration of the Erlang fixed-point equations, Information Processing Letters 81 (6) (2002) 319 – 325.

[60] E. W. M. Wong, B. Moran, A. Zalesky, Z. Rosberg, M. Zukerman, On the accuracy of the OPC approximation for a symmetric overflow loss model, Stochastic Models 29 (2013) 149–189.

[61] E. Del Re, R. Fantacci, G. Giambene, Handover and dynamic channel allocation techniques in mobile cellular networks, IEEE Transactions on Vehicular Technology 44 (1995) 229–237.

[62] M. M. Zonoozi, P. Dassanayake, User mobility modeling and characterization of mobility patterns, IEEE Journal on Selected Areas in Communications 15 (7) (1997) 1239–1252.

[63] Y. Kodratoff, R. S. Michalski, Machine learning: an artificial intelligence approach, Morgan Kaufmann, 2014.

[64] S. Chatziperis, P. Koutsakis, M. Paterakis, A new call admission control mechanism for multimedia traffic over next-generation wireless cellular networks, IEEE Transactions on Mobile Computing 7 (1) (2008) 95–112.

[65] D. Everitt, D. Manfield, Performance analysis of cellular mobile communication systems with dynamic channel assignment, IEEE Journal on Selected Areas in Communications 7 (8) (1989) 1172–1180.

[66] Z. Niu, Y. Wu, J. Gong, Z. Yang, Cell zooming for cost-efficient green cellular networks, IEEE Communications Magazine 48 (11) (2010) 74–79.

[67] S. Ni, L. Hanzo, Genetically enhanced performance of a UTRA-like time-division duplex CDMA network, in: 2005 IEEE VTC 2005-Spring, Vol. 4, 2005, pp. 2279–2283.

[68] M. Zukerman, I. Rubin, On multi channel queueing systems with fluctuating parameters, in: Proc. IEEE INFOCOM '86, Miami, Florida, 1986, pp. 600–608.

[69] S. Das, H. Viswanathan, G. Rittenhouse, Dynamic load balancing through coordinated scheduling in packet data systems, in: Proc. IEEE INFOCOM 2003, Vol. 1, 2003, pp. 786–796.

[70] H. Holtkamp, G. Auer, S. Bazzi, H. Haas, Minimizing base station power consumption, IEEE Journal on Selected Areas in Communications 32 (2) (2014) 297–306.

[71] J. G. Andrews, F. Baccelli, R. K. Ganti, A tractable approach to coverage and rate in cellular networks, IEEE Transactions on Communications 59 (11) (2011) 3122–3134.