# Fast Simulation of Cellular Networks with Dynamic Channel Assignment

**Felisa J. Vázquez-Abad** *

Department of Computer Science and Operations Research

University of Montreal, Montreal, Canada H3C 3J7

Email: vazquez@iro.umontreal.ca

also *Principal Fellow*, Department of Electrical and Electronic Engineering

The University of Melbourne

**Lachlan L. H. Andrew and David Everitt** †

ARC Special Research Centre for Ultra-Broadband Information Networks

Department of Electrical and Electronic Engineering

The University of Melbourne, Victoria 3010, Australia

Email: {l.andrew,d.everitt}@ee.mu.oz.au

**Abstract**

Blocking probabilities in cellular mobile communication networks using dynamic channel assignment are hard to compute for realistic sized systems. This computational difficulty is due to the structure of the state space, which imposes strong coupling constraints amongst components of the occupancy vector. Approximate tractable models have been proposed, which have product form stationary state distributions. However, for real channel assignment schemes, the product form is a poor approximation and it is necessary to simulate the actual occupancy process in order to estimate the blocking probabilities.

Meaningful estimates of the blocking probability typically require an enormous amount of CPU time for simulation, since blocking events are usually rare. Advanced simulation approaches use importance sampling (IS) to overcome this problem. In this paper we study two regimes under which blocking is a rare event: low load and high cell capacity. Our simulations use the standard clock (SC) method. For low load, we propose a change of measure that we call *static ISSC*, which has bounded relative error. For high capacity, we use a change of measure that depends on the current state of the network occupancy. This is the *dynamic ISSC* method. We prove that this method yields zero variance estimators for single clique models, and we empirically show the advantages of this method over naïve simulation for networks of moderate size and traffic loads.

*Supported in part by NSERC-Canada grant # WFA0184198.

†Supported by the Australian Research Council (ARC).

# 1 Introduction

Efficient design of cellular mobile communications networks requires the ability to determine the quality of service provided by a particular network configuration. A common quality of service measure is the *blocking probability*, which is the probability that a new call will not be admitted to the network due to insufficient network resources. This paper will consider techniques for determining the blocking probability in cellular telephony systems with frequency reuse, including first generation systems such as the Advanced Mobile Phone System, AMPS [Lee, 1995], and second generation systems such as the Global System for Mobile communication, GSM [Mouly and Pautet, 1992, Redl, Weber, and Oliphant, 1995].

In cellular networks, each mobile station communicates with a base station connected to the wireline telephone network. The region in which mobiles connect to a given station is called a *cell*. Each mobile station communicates with its base station using a specific frequency pair or frequency/time-slot pair known as a "channel". To avoid interference, this channel cannot be used in nearby cells; however, it may be reused in cells sufficiently remote that interference caused by the reused channel is below a specified threshold.

In static assignment schemes, each cell is allocated a fixed subset of the available channels, and calls arriving in a cell are connected only when there are free channels available from that subset. While simple to implement, this strategy may result in wasted resources; all the channels for one cell may be in use, but adjacent cells may have free capacity that could be used to connect incoming calls without causing interference. Network capacity can be improved by *dynamic channel assignment* [Cox and Reudink, 1972, Cox and Reudink, 1973], in which channels not currently in use in the nearby cells may be used. It is these systems which are the focus of this paper.

Many techniques have been developed for determining the performance of such networks. For Markov models (Poisson arrivals and exponential holding times), when the system is reversible [Kelly, 1979], the stationary state distribution has a simple product form expression on a state space $\mathcal{S}$ which is a small subset of a hypercube $H$. When there is no mobility of users, there is such a product form solution if the network uses "maximum packing" [Everitt and Macfadyen, 1983], in which calls in progress can be rearranged to use different channels. There are also models of mobility which preserve this property (see, for example, Pallant and Taylor, 1995, Boucherie and Mandjes, 1998). Moreover, the result remains valid even when call holding times have non-exponential distributions [Kelly, 1979]. Product form systems have been studied extensively (see, for example, the survey of Nelson, 1993). The product form expression involves a normalizing constant, from which the blocking probability can be determined directly, without needing to determine specific state probabilities. It can be evaluated exactly by recursive methods [Dziong and Roberts, 1987, Pinsky and Conway, 1992], mean value analysis [Reiser and Lavenberg, 1980], generating function inversion methods [Choudhury, Leung, and Whitt, 1995] or uniform asymptotic approximation [Mitra and Morrison, 1994]. However, these techniques all have exponential complexity in the number of cells.

For systems with a large number of cells, Monte Carlo techniques can be used either to estimate the normalizing constant [Ross and Wang, 1992] or to estimate blocking in a way that avoids the need to calculate it. The simplest approach of the second type is the acceptance/rejection (A/R)

method, in which states are generated in the full hypercube $H$; those lying outside the state space $\mathcal{S}$ are rejected, while for those on the boundary of the feasible region, the proportion of blocked cells is recorded, weighted by the respective arrival rates (see, for example, Everitt and Macfadyen, 1983). As the number of cells grows, generation of a sample point inside the state space $\mathcal{S} \subset H$ may become a rare event, and so importance sampling (IS) has been applied to these methods (see Ross, Tsang, and Wang, 1994, Mandjes, 1997, Lassila and Virtamo, 2000). An alternative approach is to use Markov Chain Monte Carlo (MCMC) techniques such as the Gibbs sampler used by Lassila and Virtamo [1998] and Vázquez-Abad and Andrew [2000]. These generate a Markov chain whose steady state probabilities satisfy the target product form, and they may be simulated more efficiently.

Most dynamic channel assignment implementations do not have such a product form solution. It is common in such cases to use closed form approximations, such as the ubiquitous reduced load approximation, developed for circuit switched networks (see, for example, Kelly, 1991). This approximation works well if there is minimal correlation between blocking due to conflicts with different reuse constraints, but poorly if there is significant correlation. Due to the spatial nature of the reuse constraints in the cellular case, it can be expected that there will be significant correlation. There are many other approximations; see, for example, Zahorjan, Eager, and Sweillam [1988].

A very flexible, straightforward and hence common approach is to directly simulate the arrival and departure process of calls. This allows any performance measure of the system to be estimated. Moreover, it allows arbitrary channel allocation schemes to be compared, including those for which there is no product form solution, or indeed no known closed form solution at all. For these reasons, this is the approach most commonly taken by engineers investigating different dynamic channel assignment systems. However, this approach can be very slow, especially when blocking probabilities are low. In this paper, we present two importance sampling schemes for the efficient simulation of systems with low blocking probabilities, assuming no user mobility.

Section 2 outlines some important background material, starting with the model for the channel occupancy process, and then describes the principles of fast simulation. Section 3 describes the use of quasi-regenerative cycles for the fast simulations in this paper. The two specific rare event regimes are then investigated in Sections 4 (low load) and 5 (high capacity). In both of these regimes, the utilization tends to zero. We conclude in Section 6 that these techniques provide a significant improvement over standard techniques when events are very rare, and indicate scope for further research.

# 2 Motivation and Background

## 2.1 Blocking Probabilities

A cellular network is a collection of $K$ spatially separated base stations, and a collection of users who make calls of limited duration. During a call, a user communicates with the nearest base station by means of one of $C$ channels. The region which is closer to one base station than to any other is called a cell. The principle behind cellular networks is that each of the $C$ channels can be used simultaneously by multiple users across the network, if and only if the so-called “reuse

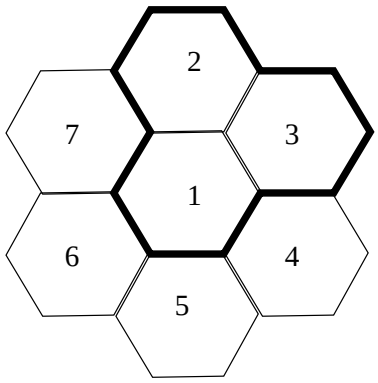



Figure 1: Simple cellular network model

constraints" are satisfied. These constraints ensure that the performance in any given cell is not excessively degraded by the interference caused by other cells using the same channel. The reuse constraints hence depend on the precise layout of the cells. For the examples in this paper, we shall assume that the cells form a hexagonal grid and 3-cell reuse is employed; that is, no channel may be used simultaneously by more than one call in any group of three mutually adjacent cells. In general, a set of cells in which a channel may only be used once is called a "clique". Figure 1 shows a simple seven-cell network, with one clique highlighted. Let $M$ be the number of cliques (in Figure 1, $M = 6$), and let $c_j$ be the $j$th clique, $j = 1, \ldots, M$.

Calls can arrive at the cell in one of two ways. They may be new calls or they may be existing calls being handed off from neighbouring cells due to user mobility. The model used in this paper does not include user mobility. Using dynamic channel assignment, calls arriving to a cell are assigned one of the available channels. If no channel can be allocated without violating a reuse constraint, then the call is *blocked*. Otherwise it is accepted, and uses the selected channel. In practice, the call will generally use the same channel until it departs from the cell. Thus, in general, the state of the system depends on both the number of calls in each cell (the occupancy), and also on which particular channels they use.

There is no useful lower bound on the occupancy of a given cell, $i$, in the states when blocking occurs; it is possible for calls arriving to cell $i$ to be blocked when there are no calls at all in cell $i$, if all the channels are used elsewhere in the cliques to which $i$ belongs. Define the "cluster" associated with cell $i$ to be the union of all cliques containing $i$:

$$\mathcal{C}_i = \bigcup_{c_j \ni i} c_j.$$

It is then possible to say that the occupancy of the *cluster* $\mathcal{C}_i$ must be at least $C$ when calls arriving to cell $i$ are blocked, since each channel must be blocked by at least one of the cliques containing cell $i$. That is,

$$\sum_{j \in \mathcal{C}_i} n_j \geq C \tag{1}$$

is a necessary condition for blocking to occur, where $n_j$ denotes the number of calls (i.e., channels in use) in cell $j$. This is the fundamental property of blocking states that is used in the methods

presented here. Note that when channels are reserved for handovers, as for example in Li and Alfa [2000], then $C$ should be interpreted as the number of channels available to new calls and $n_j$ as the number of calls using these channels.

Most of the techniques described in the introduction rely on having a known closed form for the blocking probability. There is such a closed form for maximum packing channel assignment, proposed by Everitt and Macfadyen [1983], in which channels may be reassigned on the arrival of a new call. However, the operation of reassigning calls is not feasible in practice, and so this closed form is merely a lower bound for the blocking of real channel assignment algorithms. The techniques to be presented in Sections 4 and 5 are applicable to real channel assignment algorithms and are thus of more general applicability than most of the techniques described in the Introduction.

In general, the state of the *occupancy process* at time $t$ is given by $\tilde{n}(t) = (\tilde{n}_{1,1}(t), \ldots, \tilde{n}_{K,C}(t))$, where $\tilde{n}_{i,c}(t) = 1$ if channel $c$ is used in cell $i$ at time $t$, and zero otherwise. Sometime we also used a simplified state description, given by $n(t) = (n_1(t), \ldots, n_K(t))$, where $n_i(t)$ represents the number of channels in use in cell $i$ at time $t$. This is the *aggregate occupancy process*. Note that $n(t)$ is completely determined by $\tilde{n}(t)$. Under maximum packing, all calls in progress can be rearranged to different channels on the arrival of a call, and so the behaviour of the system is determined entirely by this aggregate state.

Some of our numerical examples will use the so-called "clique packing" approximation to maximum packing, proposed by Everitt and Macfadyen [1983] and further investigated by Raymond [1991], which considers only constraints local to each clique. A state is feasible under clique packing if *each* of the cliques contains no more calls than there are channels:

$$n^{(c_j)} \leq C \quad \forall j = 1, \ldots, M \tag{2}$$

where $n^{(A)}$ is the number of calls in a set of cells, $A$, in a given network state, $n$.

At each cell $i$, new calls arrive following independent Poisson processes with corresponding intensities $\lambda_i, i = 1, \ldots, K$. A call that arrives at cell $i$ at time $t$ is accepted if there is still at least one channel available. If call rearrangement is not permitted, this requires that there exists a channel, $c$, such that $\tilde{n}_{j,c}(t) = 0$ for all cells $j \in \mathcal{C}_i$. Under clique packing the requirement is simply that

$$\max_{c_j \ni i}(n^{(c_j)}) \leq C - 1.$$

An accepted call on channel $c$ causes $\tilde{n}_{i,c}(t) = \tilde{n}_{i,c}(t^-) + 1$ (whence $n_i(t) = n_i(t^-) + 1$), all other components of the state remaining unchanged. We say that at this time the call is connected. If an incoming call to cell $i$ finds no channels available (under clique packing, the current state satisfies (2) with equality for some $c_j \ni i$) then all channels are used and the call is blocked, with no change to the state. Note that (1) is a necessary condition for blocking whether or not clique packing is used.

Calls stay connected for a random length of time called the holding time, assumed to be independent of the rest of the process history. All holding times are identically distributed with mean $1/\mu$. When a call in cell $i$ departs, the corresponding occupancy component is decreased by one unit. Although the holding times are assumed to be exponential in this paper, the network performance is in fact independent of the holding time distribution for many channel assignment

schemes, including clique packing [Kelly, 1979]. Most models without call rearrangement are not included in these.

This model gives rise to a continuous-time Markov process. Furthermore, because the process, $\tilde{n}(t)$, consists of independent arrivals and departures, it can be expressed as a *quasi birth and death* (QBD) process (see Neuts, 1981). In QBD processes, states can be arranged in layers, such that transitions from layer $s$ can only be to states in layers $s-1$, $s$ or $s+1$. For any $i$, layer $s$ can be defined to consist of states in which cluster $\mathcal{C}_i$ contains $s$ calls. This system is a QBD, since a call arrival within the cluster causes a transition from layer $s$ to layer $s+1$, a departure within the cluster causes a transition from $s$ to $s-1$, and an arrival or departure outside the cluster causes a transition entirely within level $s$. In this representation, all blocking states for cell $i$ are in layers $C$ and higher. The occupancy process, $\tilde{n}(t)$, is a particular case of a QBD where the rates and barriers depend on all components of the process. When the process is in state $\tilde{n}$, the birth rate of component $i$ of the aggregate state, $n$, is $\lambda_i \mathbf{1}_{\{\tilde{n} \notin \tilde{\mathcal{B}}_i\}}$. Here $\mathbf{1}_{\{A\}}$ is the indicator function of event $A$ and $\tilde{\mathcal{B}}_i$ is the set of states that cause blocking in cell $i$, which depends on the channel assignment used. For clique packing, this simplifies to $\lambda_i \mathbf{1}_{\{n \notin \mathcal{B}_i\}}$, where $\mathcal{B}_i$ is the set of aggregated blocking states for cell $i$:

$$\mathcal{B}_i = \left\{ n \in \mathcal{S} : \exists c_j \ni i, n^{(c_j)} = C \right\}, \quad i = 1, \ldots, K. \tag{3}$$

Here $\mathcal{S}$ is the state space consisting of all integer vectors $n = (n_1, \ldots, n_K) \in \mathbb{N}^K$ satisfying (2). When clique packing is not used, define $\mathcal{B}_i$ to be all those $n \in \mathcal{S}$ which are aggregate states corresponding to the blocking states, $\tilde{\mathcal{B}}_i$. Note that for $\tilde{n} \in \tilde{\mathcal{B}}_i$ equation (1) holds. The total death rate at level $s$ is $s\mu$.

The performance measure of interest is the *blocking probability*, defined as the long term probability that an incoming arrival is blocked:

$$B = \lim_{t \to \infty} \frac{\sum_{i=1}^{K} Y_i(t)}{A(t)} = \sum_{i=1}^{K} \left( \frac{\lambda_i}{\lambda_{\text{tot}}} \right) B_i = \sum_{i=1}^{K} \left( \frac{\lambda_i}{\lambda_{\text{tot}}} \right) \pi(\tilde{\mathcal{B}}_i) \tag{4}$$

where $Y_i(t)$ is the total number of calls blocked in cell $i$ up to time $t$, $A(t)$ is the total number of arrivals up to time $t$ and $\lambda_{\text{tot}} = \sum_{i=1}^{K} \lambda_i$ is the total arrival rate. The term $B_i$ is the long term proportion of calls arriving to cell $i$ that are blocked, and $\pi(\tilde{\mathcal{B}}_i)$ is the stationary probability that the state is in the blocking set $\tilde{\mathcal{B}}_i$.

The renewal-reward theorem can be used to re-write (4) in terms of expectations within regenerative cycles. However, as the state space has many components, regeneration cycles are frequently too long to be a feasible basis for simulation. The concept of *quasi-regeneration* was introduced to calculate stationary averages for such systems, as explained by Chang, Heidelberger, and Shahabuddin [1995] and Gaivoronski and Messina [1996]. Consider a random process, $\tilde{n}(t)$, and assume it starts with the stationary distribution $\mathsf{P}(\tilde{n}(0) = m) = \pi(m)$. Consider a set of states, $A$, such that there is an a.s. finite stopping time $\tilde{T}_0$, defined as the first entry time to the set $A$ from its complement $A'$. Let $\tilde{T}_1, \tilde{T}_2, \ldots$ be the subsequent times of entrances to the set $A$ from the set $A'$. Clearly these are also stopping times, and are a.s. finite. Since the process is in steady state, the distribution of the process $\{\tilde{n}(t + \tilde{T}_i) : t > 0\}$, $i \geq 1$, is identical to that of the process

$\{\tilde{n}(t+\tilde{T}_0) : t > 0\}$. The set $A$ is called a quasi-regenerative set. Because the aggregated occupancy process $n(t)$ described above is an irreducible Markovian process on a finite state space, all subsets of the state space are quasi-regenerative sets and a unique stationary measure $\pi$ exists. The times between consecutive entries to the set $A$ are termed "$A$-cycles". Unlike true regenerative cycles, $A$-cycles may not be independent, but they are still identically distributed.

It will be useful to consider different quasi-regeneration sets, $A_i$, for different cells $i$. Following Sadowsky [1991], we will require that $A_i \supset \tilde{\mathcal{B}}_i$. Let $T^{(i)}$ be the random length of an $A_i$-cycle, and $X_i(T^{(i)})$ be the amount of time within an $A_i$-cycle that the process spends in $\tilde{\mathcal{B}}_i$. Then [Breiman, 1992]

$$B_i = \frac{\mathsf{E}[X_i(T^{(i)})]}{\mathsf{E}[T^{(i)}]}. \tag{5}$$

The sets $A_i$ will be chosen in such a way as to minimize the required simulation time. $A$-cycles have been used in a number of papers, such as Sadowsky [1991], Nicola *et al.* [1993] and L'Ecuyer and Champoux [1996], to name but a few.

## 2.2 Fast Simulation Methods

A commonly used figure of merit of an estimator is its relative efficiency (see, for example, Glynn and Whitt, 1992). This quantifies the tradeoff between computational effort and relative mean square error (or equivalently, the relative variance if the estimator is unbiased).

**Definition 1** *Let $\hat{Y}(S)$ denote a consistent estimator of $B$ that uses $S$ samples of a stochastic process. The* relative efficiency *of $\hat{Y}(S)$ is:*

$$\mathcal{E}_r(\hat{Y}(S)) = \left( \frac{B^2}{\mathrm{CPU}[\hat{Y}(S)]\mathsf{Var}[\hat{Y}(S)]} \right), \tag{6}$$

*where $\mathrm{CPU}[\hat{Y}(S)]$ denotes the expected value of the CPU time of the simulation that produces the $S$ samples.*

**Definition 2** *Let $\{n(t)\}$ be a stochastic process defined on a set of outcomes $\Omega$ and let $\epsilon > 0$ be a parameter of the distribution of the process, denoted $\mathsf{P}^\epsilon$. The event $\mathcal{R} \subset \Omega$ is called a* rare event *if $\lim_{\epsilon \to 0} \mathsf{P}^\epsilon(\mathcal{R}) = 0$.*

If $p(\epsilon) = \mathsf{P}^\epsilon(\mathcal{R})$ is estimated via simulation using $\mathbf{1}_{\{\mathcal{R}\}}$ for $S$ consecutive $A$-cycles, then the variance of the estimator is at least $p(\epsilon)(1-p(\epsilon))/S$ if consecutive $A$-cycles introduce positive correlation on the Bernoulli sequence (as is usually the case in our application). The relative error in the estimation, defined as the standard deviation of the estimator divided by the true value, is then bounded below by $\sqrt{(1-p(\epsilon))/(S\,p(\epsilon))} \to \infty$, as $\epsilon \to 0$ for fixed $S$. The efficiency in the estimation of rare events can be improved using a change of measure approach via importance sampling (see Devetsikiotis and Townsend, 1993 and Asmussen and Nielsen, 1995 among others).

The simulation model assumes an underlying discrete time Markovian process $\{U_k\}$, such that the occupancy process $\{\tilde{n}(t)\}$ can be determined solely from the generation of $\{U_k\}$. Use the

notation $\{\tilde{n}(k), k = 1, 2, \ldots\}$ and $\{n(k), k = 1, 2, \ldots\}$ for the embedded processes (with the obvious abuse in notation). Let $f_{\tilde{n}}(u)$ be the conditional density of $U_{k+1}$ given the state $\tilde{n}(k) = \tilde{n}$. Let now $f^*_{\tilde{n}}(\cdot)$ be *another* conditional density, and define for any $T \in \mathbb{N}$

$$L_T = \prod_{k=0}^{T-1} \frac{f_{\tilde{n}(k)}(U_{k+1})}{f^*_{\tilde{n}(k)}(U_{k+1})}, \tag{7}$$

known as the *Radon-Nikodym derivative* of the distribution $\mathsf{P}$ with respect to the distribution $\mathsf{P}^*$, where $\mathsf{P}$ and $\mathsf{P}^*$ are the distributions of the respective Markov processes after $T$ events. If a rare event $\mathcal{R}$ depends only on the history of the process $\{U_k, k \leq T\}$ for some fixed time $T$, then (see, for example, Ross, 1997)

$$\mathsf{E}[\mathbf{1}_{\{\mathcal{R}\}}] = \mathsf{E}^*[L_T \mathbf{1}_{\{\mathcal{R}\}}], \tag{8}$$

where $\mathsf{E}^*$ denotes the expectation w.r.t. the distribution induced by $f^*$. The change of measure (8) is valid also when $T$ is a random stopping time, as explained, for example, in Sadowsky [1991] and Asmussen and Nielsen [1995]. This approach can use arbitrary densities $f^*$ as long as they satisfy an absolute continuity constraint; the restriction of $\mathsf{P}$ to the "important set" $\mathcal{R}$, $\mathsf{P}\rfloor_{\mathcal{R}}$, must be absolutely continuous with respect to the new measure $\mathsf{P}^*$; equivalently $\forall \omega \in \mathcal{R}$, $f_{\tilde{n}}(U(\omega)) > 0 \Rightarrow f^*_{\tilde{n}}(U(\omega)) > 0$. (See, for example, Vázquez-Abad and LeQuoc, 2001.)

**Definition 3** *The unbiased IS estimator for the rare event probability $p(\epsilon)$, $L_T \mathbf{1}_{\{\mathcal{R}\}}$, has* bounded relative error (BRE) *under $\mathsf{P}^*$ if there are constants $b < \infty, \epsilon_0 > 0$ such that*

$$\sup_{\epsilon \leq \epsilon_0} \frac{\sqrt{\mathsf{Var}^*[L_T \mathbf{1}_{\{\mathcal{R}\}}]}}{p(\epsilon)} \leq b. \tag{9}$$

The above definition is widely used in rare event estimation; see, for example, Shahabuddin [1994]. The following lemma is a direct consequence of Definition 3. (See, for example, Chang, Heidelberger, and Shahabuddin, 1995.)

**Lemma 1** *If there are constants $l, u$ and $b$ such that $p(\epsilon) \geq l\epsilon^b$ and $L_T \mathbf{1}_{\{\mathcal{R}\}} \leq u\epsilon^b$ a.s., then the IS estimator for $p(\epsilon)$, $L_T \mathbf{1}_{\{\mathcal{R}\}}$, has BRE.*

By construction, $p(\epsilon) = \mathsf{P}^\epsilon(\mathcal{R}) = \mathsf{E}^*[L_T \mathbf{1}_{\{\mathcal{R}\}}]$ under *any* valid change of measure. Often the new measure depends on $\epsilon$, and hence, in general, $L_T$ will be a function of $\epsilon$, although we will not make this explicit in our notation. Because variances are non negative, it must always be true that

$$\mathsf{E}^*[L_T^2 \mathbf{1}_{\{\mathcal{R}\}}] \geq p^2(\epsilon). \tag{10}$$

Estimators that satisfy (10) with equality are optimal and are called zero variance estimators.

Sadowsky [1991] studies a $GI/GI/s/\infty$ queueing system and estimates the excessive backlog probability: the probability that an arrival finds at least $C$ customers waiting, $C$ being large. By defining the quasi-regenerative set, $A$, to be those states in which all $s$ servers are busy, he finds

the asymptotically optimal change of measure for the probability of excessive backlog occurring within a given $A$-cycle. The approach uses the so-called "conjugate" distributions, which in the Markovian case is known as "rate swapping": simulate an $M/M/s/\infty$ queueing system, with the new rates

$$\lambda^* = s\mu, \quad \mu^* = \lambda/s. \tag{11}$$

This change of measure is applied to the system once its occupancy reaches $s$, and its optimality relies on the fact that the number of servers in use is constant between the start of acceleration and the time when the backlog reaches $C$. This algorithm can also be used for estimating the fraction of customers lost in the long run in a $GI/GI/s/C$ queueing system, $C$ being large compared to $s$.

In Sections 4 and 5 we discuss the implementation of the IS that swaps arrival and service rates for the cellular network problem.

# 3 Model for Fast Simulation

Estimation of blocking probabilities can be especially difficult in the case when blocking is a rare event. The following sections address fast simulation of blocking probabilities for two regimes under which blocking is a rare event: low load and high capacity. For both problems we use the same simulation model, which does not rely on the product form solution.

The occupancy process $\{\tilde{n}(t); t \geq 0\}$, as described in Section 2.1, is a continuous time Markov process. Recall that the arrival rate of calls into cell $i$ is $\lambda_i$, $i = 1, \ldots, K$, and the mean holding time per call is $1/\mu$. Also, recall that the blocking probability can be expressed as in (5). Given a set $A_i$, consider the process started with the stationary distribution, conditional on the event $\{\tilde{n} \in A_i\}$. Define $\tau_i$, as the number of events required to reach a blocking state for cell $i$ or the end of the current quasi-regenerative cycle, and $\mathcal{R}_i$ as the event that the cycle contains a blocking state for cell $i$:

$$\tau_i = \min\{k : S_k \geq T^{(i)} \text{ or } \tilde{n}(S_k) \in \tilde{\mathcal{B}}_i\}, \quad \text{and} \quad \mathcal{R}_i = \{S_{\tau_i} < T^{(i)}\}, \tag{12}$$

where $S_k$ is the epoch of the $k$th event in the system. Then

$$B_i = \frac{\mathsf{E}[X_i(T^{(i)})|\mathcal{R}_i]\,\mathsf{P}(\mathcal{R}_i)}{\mathsf{E}[T^{(i)}]}. \tag{13}$$

For each cell $i$, a change of measure is performed to build an efficient estimator for $\mathsf{P}(\mathcal{R}_i)$. Fix the index $i$ throughout the remaining sections, until the description in Section 5 of the simulations. We now define the set $A_i$.

**Definition 4** *Let the quasi-regenerative set $\tilde{A}_i$ be the set of all states in which the cluster occupancy, $n^{(\mathcal{C}_i)}$, satisfies $n^{(\mathcal{C}_i)} > \theta_i$ for some threshold $0 \leq \theta_i < C$. Let $A_i$ be the corresponding set of aggregate states.*

The optimal choice of $\theta_i$ will be considered in Section 5.2.3.

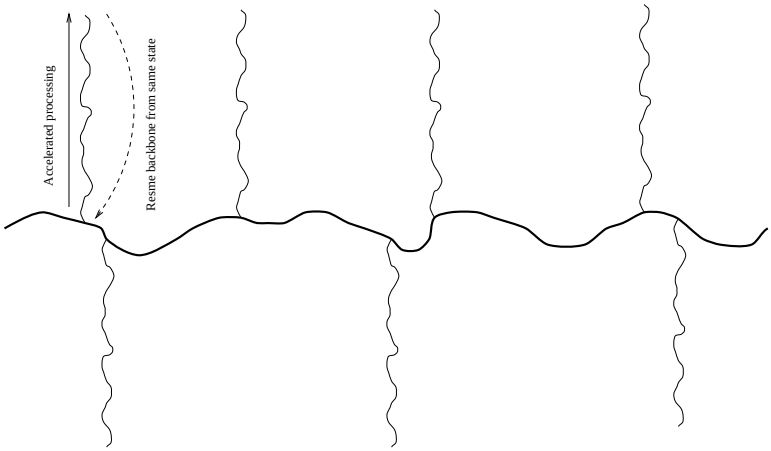



Figure 2: "Backbone and ribs" simulation framework.

The following sections will derive changes of measure with bounded relative error for estimating $\mathsf{P}(\mathcal{R}_i)$, the probability that there is a blocking state within an $A_i$-cycle associated with an arbitrary cell $i$ of the cellular network model. Fast simulation via IS requires simulating the process with cluster acceleration *only until the process enters a blocking state*, i.e., a state in $\tilde{\mathcal{B}}_i$. Once the state enters $\tilde{\mathcal{B}}_i$, the standard measure is used to resume the simulation and to estimate the time spent in blocking states within the current $A_i$-cycle, $\mathsf{E}[X(T^{(i)})|\mathcal{R}_i]$.

However, as mentioned before, an $A_i$-cycle must start with the state distribution being the equilibrium distribution conditional on the process having just entered the set $A_i$ from $A_i'$. In the absence of importance sampling, this will also be the distribution at the end of the $A_i$-cycle. However, importance sampling disturbs the distribution, and the distribution of the state at the end of an accelerated $A_i$-cycle cannot be used as the starting state for the next $A_i$-cycle. Hence, in addition to the simulation with importance sampling mentioned above, we start a second simulation from the same initial state (*i.e.,* the initial state that is used for the $A_i$ cycle with importance sampling). This simulation is done with the original probability measure and its purpose is to recover the initial state for the simulation of the next $A_i$-cycle. This gives rise to the "backbone and ribs" arrangement, seen in Figure 2. A second advantage of this approach is that the length of an $A_i$-cycle, the denominator of (13), may be estimated with lower variance from the non-accelerated $A_i$-cycles. This technique is common in queueing network simulation, and has been implemented by Nicola *et al.* [1993], Chang, Heidelberger, and Shahabuddin [1995] and L'Ecuyer and Champoux [1996], among others.

For each cell $i$, the $A_i$-cycle is used to get estimates only of the blocking probability $B_i$ in a single cell. Thus in order to estimate the blocking probability, it is necessary to run separate simulations for each cell. Fortunately, it is not necessary to simulate the "backbone" separately. Instead, separate "ribs" can be started each time the backbone starts an $A_i$-cycle for *any* cell $i$. Note that a single event may be the start of $A_i$-cycles for more than one $i$, and each of these must be considered.

The ribs may be further simplified by noting that the set $A_i'$ contains no blocking states for cell

$i$. Thus, once an accelerated $A_i$-cycle has left the set $A_i$, the entire time spent in blocking states will already have occurred. This means that the rest of the $A_i$-cycle need not be simulated, even though a considerable amount of time may be spent in the set $A_i'$ before the $A_i$-cycle is truly over.

The simulation model is the standard clock technique of Vakili [1991], which corresponds to the dynamical description of a multidimensional birth and death process. Consider the embedded occupancy processes $\{\tilde{n}(k)\}$ and $\{n(k)\}$. Let $N(k) = \sum_{j=1}^{K} n_j(k)$ be the total occupancy at step $k$. An exponential random variable with intensity $\Lambda_{N(k)}$ is used to determine the *inter-event time*, $T_{k+1}$, or the time until the next event. Here

$$\Lambda_{N(k)} = \sum_{j=1}^{K} \lambda_j + N(k)\mu.$$

Next, the event *type* $D_{k+1}$ is determined as a discrete random variable with distribution

$$D_{k+1} = \begin{cases} a_j : \text{arrival at cell } j & \text{w.p. } \dfrac{\lambda_j}{\Lambda_{N(k)}} \\ d_j : \text{departure of call at cell } j & \text{w.p. } \dfrac{n_j(k)\mu}{\Lambda_{N(k)}} \end{cases}.$$

The embedded occupancy process is updated by first setting

$$n_j(k+1) = \begin{cases} n_j(k) + 1 & \text{if } D_{k+1} = a_j \text{ and } \tilde{n}(k) \notin \tilde{\mathcal{B}}_i, \\ n_j(k) & \text{if } D_{k+1} = a_j \text{ and } \tilde{n}(k) \in \tilde{\mathcal{B}}_i, \\ n_j(k) - 1 & \text{if } D_{k+1} = d_j \end{cases}$$

and then determining the state, $\tilde{n}(k+1)$, using the channel assignment rule. Recall that $S_k$ is the epoch of the $k$th event in the system.

When call rearrangement is permitted, $\tilde{n}(k) \notin \tilde{\mathcal{B}}_i$ if and only if $n(k) \notin \mathcal{B}_i$, and so the evolution of the aggregate process is determined only by its state. When calls cannot be rearranged, the random variables $D_k$ and the inter-event times are as described above, but an additional decision is made concerning which channel an arrival is to, or a departure is from. These decisions affect $\tilde{n}(k+1)$ but not $n(k+1)$. Without rearrangement, the decision whether or not to block an arrival cannot be made on the basis of $n(k)$ alone, but $n(k)$ must still satisfy (1) for blocking to occur.

To denote use of IS for the standard clock simulation model we use the acronym ISSC. Similar work has been done in the context of reliability by Heidelberger, Shahabuddin, and Nicola [1994].

# 4 Static ISSC Estimation for Light Traffic

In the light traffic regime, assume that $\lambda_i = k_i\epsilon, i = 1, \ldots K$, with $\epsilon \to 0$. (This is analogous to the regime used by, for example, Shahabuddin [1994] in the context of reliability.) In the $GI/G/s/\infty$ case, the servers are busy from the start of the $A$-cycle containing a blocking state, until a blocking state is reached. In our model, however, channels are not continuously busy within $A_i$-cycles until a blocking state is reached. Thus swapping rates as in (11) will not be optimal. Instead, consider

the change of measure that swaps aggregate arrival rates per cluster and inverse expected holding times.

**Proposition 1** *Consider the ISSC simulation model with initial state $\tilde{n}(0)$ such that $n^{(\mathcal{C}_i)}(0) = \theta_i = 0$, and $A_i$-cycles as defined in Definition 4. Arrivals at the cluster $\mathcal{C}_i$ have rate $\lambda^* = \mu$ and service rate for the calls in the cluster is $\mu^* = \lambda$. Other inter-arrival and holding times (outside the cluster) have the original exponential distribution. Call the underlying measure $\mathsf{P}^*$. Then*

$$\mathsf{P}(\mathcal{R}_i) = \mathsf{E}^* \left[ \exp\left( -(\mu - \lambda) \sum_{k=1}^{\tau_i - 1} (n^{(\mathcal{C}_i)}(k) - 1) T_{k+1} \right) \left( \frac{\lambda}{\mu} \right)^{a-d} \mathbf{1}_{\{\mathcal{R}_i\}} \right],$$

*where $n^{(\mathcal{C}_i)}(k) = \sum_{j \in \mathcal{C}_i} n_j(k)$ is the total occupancy of the cluster, $a(j)$ is the total number of arrivals to cell $j$ prior to event number $\tau_i$ (including blocked calls), $a = \sum_{j \in \mathcal{C}_i} a(j)$ and $d$ is the corresponding number of call completions (excluding blocked calls).*

**Proof :** Given the state $n(k)$ of the process at the time of the $k$-th event, $T_{k+1} \sim \exp(\Lambda^*_{N(k)})$. The new event rate is the random variable

$$\Lambda^*_{N(k)} = \mu + \sum_{j \notin \mathcal{C}_i} \lambda_j + \lambda \sum_{j \in \mathcal{C}_i} n_j(k) + \mu \sum_{j \notin \mathcal{C}_i} n_j(k),$$

and the event types are now

$$D_{k+1} = \begin{cases} \text{arrival at cell } j \in \mathcal{C}_i & \text{w.p. } \dfrac{\lambda_i}{\lambda} \dfrac{\mu}{\Lambda^*_{N(k)}} \\ \text{arrival at cell } j \notin \mathcal{C}_i & \text{w.p. } \dfrac{\lambda_j}{\Lambda^*_{N(k)}} \\ \text{departure of call at cell } j \in \mathcal{C}_i & \text{w.p. } \dfrac{n_j(k)\lambda}{\Lambda^*_{N(k)}} \\ \text{departure of call at cell } j \notin \mathcal{C}_i & \text{w.p. } \dfrac{n_j(k)\mu}{\Lambda^*_{N(k)}} \end{cases}.$$

On $\mathcal{R}_i$, we perform this change of measure within an $A_i$-cycle until the $\tau_i$-th event occurs, which is the first time that the state is in $\tilde{\mathcal{B}}_i$. Using the framework of Section 2.2, the corresponding Radon-Nikodym derivative is given by (7) as

$$\begin{aligned} L_{\tau_i} &= \prod_{k=1}^{\tau_i - 1} \left( \frac{\Lambda_{N(k)}}{\Lambda^*_{N(k)}} \right) e^{(\Lambda^*_{N(k)} - \Lambda_{N(k)}) T_{k+1}} \\ &\quad \times \prod_{k=1}^{\tau_i - 1} \left( \frac{\Lambda^*_{N(k)}}{\Lambda_{N(k)}} \right) \left( \sum_{j \in \mathcal{C}_i} \left( \frac{\lambda_j}{\mu(\lambda_j/\lambda)} \right) \mathbf{1}_{\{D_{k+1} = a_j\}} + \left( \frac{\mu}{\lambda} \right) \mathbf{1}_{\{D_{k+1} \in \mathcal{D}_i\}} + \mathbf{1}_{\{D_{k+1} \notin \mathcal{D}_i \cup \mathcal{A}_i\}} \right) \end{aligned} \tag{14}$$

where $\mathcal{A}_i = \{a_j : j \in \mathcal{C}_i\}$ is the set of event types which are arrivals to the cluster and similarly $\mathcal{D}_i = \{d_j : j \in \mathcal{C}_i\}$ is the set of event types which are departures of calls within the cluster. From their definitions, it follows that

$$\Lambda^*_{N(k)} - \Lambda_{N(k)} = (\mu - \lambda) + (\lambda - \mu) \sum_{j \in \mathcal{C}_i} n_j(k) = -(\mu - \lambda)(n^{(\mathcal{C}_i)}(k) - 1).$$

Simplifying the expression above,

$$\begin{aligned} L_{\tau_i} &= \exp\left(-(\mu-\lambda)\sum_{k=1}^{\tau_i - 1}(n^{(\mathcal{C}_i)}(k)-1)T_{k+1}\right)\left[\prod_{j\in\mathcal{C}_i}\left(\frac{\lambda}{\mu}\right)^{a(j)}\right]\left(\frac{\mu}{\lambda}\right)^{d} \\ &= \exp\left(-(\mu-\lambda)\sum_{k=1}^{\tau_i - 1}(n^{(\mathcal{C}_i)}(k)-1)T_{k+1}\right)\left(\frac{\lambda}{\mu}\right)^{a-d}. \end{aligned} \tag{15}$$

Application of (8) proves the claim. ◁

**Lemma 2** *When $\lambda < \mu$,*

$$L_{\tau_i}\mathbf{1}_{\{\mathcal{R}_i\}} < (\lambda/\mu)^{C-\theta_i}, \qquad \mathsf{P}^* - \textit{w.p.1.} \tag{16}$$

*In particular, $L_{\tau_i}\mathbf{1}_{\{\mathcal{R}_i\}} < 1, \mathsf{P}^* -$ w.p.1., guaranteeing variance reduction when using the ISSC estimator.*

**Proof :** On $\mathcal{R}_i$, there are at least $C$ on-going calls within the cluster at the hitting time $\tau_i$. At the time the $A_i$-cycle begins, there are (by Definition 4) $\theta_i + 1$ calls within the cluster. Because the stopping time within an $A_i$-cycle counts only the transitions from the start of the $A_i$-cycle up until a blocking state is reached, and $\tilde{\mathcal{B}}_i \subset \tilde{A}_i$, it follows that on the set $\mathcal{R}_i$ we have $n^{(\mathcal{C}_i)}(k) \geq \theta_i + 1, k = 1, \ldots, \tau_i$, whence

$$(\mu - \lambda)\sum_{k=1}^{\tau_i - 1}(n^{(\mathcal{C}_i)}(k) - 1)T_{k+1} \geq 0. \tag{17}$$

Moreover, $a = \sum_{j\in\mathcal{C}_i} a(j) \geq d + C - \theta_i$. Using (15) gives the result. ◁

**Theorem 1** *The ISSC estimator for $p(\epsilon)$, suggested by Proposition 1, has BRE as $\epsilon \to 0$ when $\theta_i = 0$ and $\lambda_j = k_j\epsilon$, for all cells $j$.*

**Proof :** The proof is an application of Lemma 1. The upper bound is obtained with the result of Lemma 2:

$$L_{\tau_i}\mathbf{1}_{\{\mathcal{R}_i\}} \leq \left(\frac{\lambda}{\mu}\right)^C = u\epsilon^C, \tag{18}$$

where $u = \left(\sum_{j \in \mathcal{C}_i} k_j/\mu\right)^C$.

It remains to show $p(\epsilon) \geq l\epsilon^C$. In order for blocking of arrivals to cell $i$ to occur, it is sufficient for the occupancy of a single clique, $c_j \ni i$, to reach $C$. Let a "minimal path" be a trajectory in which the first $C$ events in an $A_i$-cycle are arrivals to the same clique within the accelerated cluster. All minimal paths will lead to blocking states, and thus their probability is a lower bound for $p(\epsilon)$. The probability of such minimal paths is the probability that each of the first $C$ events be an arrival to the same clique, and so

$$p(\epsilon) \geq \left(\frac{\bar{\lambda}_i}{\Lambda_{N(C)}}\right)^C \geq \left(\frac{\bar{k}_i \epsilon}{C\mu + \lambda_{\text{tot}}}\right)^C = l\epsilon^C, \tag{19}$$

where $\bar{\lambda}_i$ is the smallest aggregate clique rate within cluster $\mathcal{C}_i$ and $l = (\bar{k}_i/(C\mu + \lambda_{\text{tot}}))^C$, with $\bar{k}_i$ the smallest of $\sum_{s \in c_j} k_s$ over the cliques $c_j \subset \mathcal{C}_i$.

It follows by Lemma 1 that in this case, ISSC has BRE. ◁

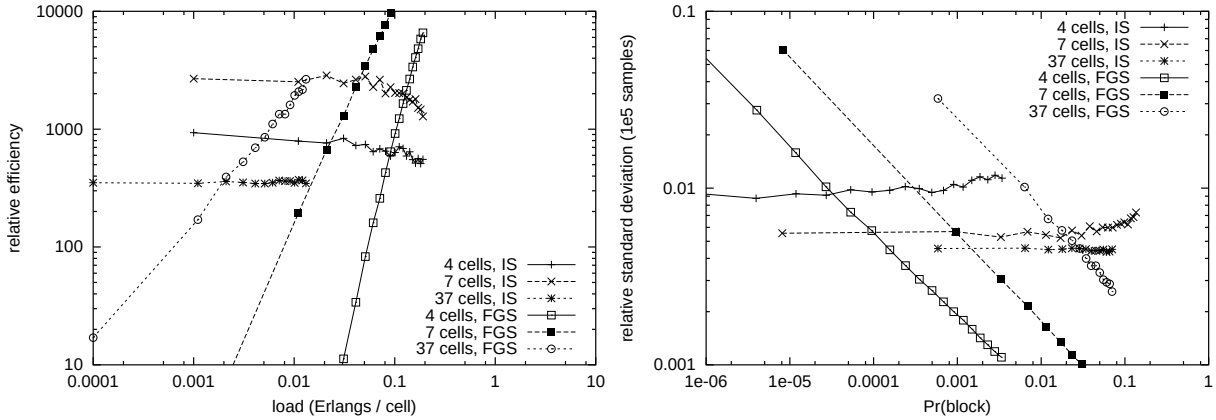
Note that the condition $\theta_i = 0$ is not a major limitation, since the modal cluster occupancy will be zero, and so $\theta_i = 0$ gives the shortest $A_i$-cycles, which is desirable for simulations.

The ISSC with $\theta_i = 0$ and $C = 2$ was tested for $\epsilon \to 0$ on three sizes of network: 4, 7 and 37 cells. The load in Erlangs per cell is given by $\lambda_i/\mu$, which was equal for all cells. A single backbone was used to estimate all $B_i$. The first $10^7$ ribs were simulated for each $i$, and used to estimate the numerator of (13). The variance corresponding to each $B_i$ was estimated using batch means, with 100 batches of $10^5$ $A_i$-cycles. The estimation of this variance, as well as that of the variance of the estimator of $B$, is described in more detail in Section 5.2.2.

Figure 3 compares the results for clique packing with those of the filtered Gibbs sampler of Vázquez-Abad and Andrew [2000]. The results clearly show that the relative error of the estimated network blocking probability is bounded as $\epsilon \to 0$.

Figure 4 shows the performance of ISSC for a seven cell network when existing calls cannot be rearranged. In this case, the network has no product form solution. Channels were assigned using the first fit algorithm, which starts searching from channel 1 and selects the first available channel. This was shown by Yates [1997] to produce significantly less blocking than random selection.

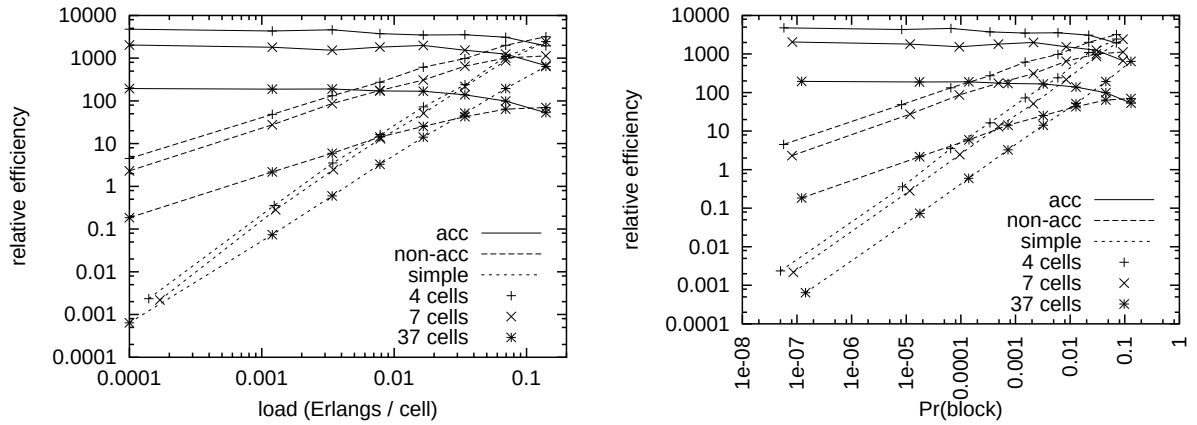
Along with the "non-accelerated $A_i$-cycles", Figure 4 includes results for a "simple" estimation scheme. This simulates only the backbone and simply counts the proportions of calls which are blocked. For high blocking, this has higher relative efficiency than the simulation based on quasi-regeneration. Its CPU time is lower partly due to its ability to estimate blocking over the entire network at once, rather than focusing on a single cell, $i$, in each $A_i$-cycle, and partly due to the elimination of "bookkeeping" associated with tracking the $A_i$-cycles. However, for lower blocking it performs even worse than the non-accelerated $A_i$-cycles, because it does not keep track of the proportion of time that the network is in a blocking state, but simply the number of calls blocked.




(a) relative efficiency against load

(b) relative error against blocking

Figure 3: Relative efficiency and relative error for importance sampling (IS) and filtered Gibbs sampler (FGS) for light loads




(a) relative efficiency against load

(b) relative efficiency against blocking

Figure 4: Relative efficiency for light loads when ribs are accelerated, when ribs not accelerated, and when the simple algorithm is used.

# 5 Dynamic ISSC Estimation for High Capacity

## 5.1 Change of Measure

It is unlikely that future networks will be operated at extremely low load, as was assumed in the previous section. Engineers are more interested in the behaviour as the capacity increases. This is particularly true of wavelength-continuous wavelength division multiplexing (WDM) trunk networks, which are mathematically analogous to the cellular networks described so far. It is possible to overcome the restriction that $\lambda < \mu$ used in Lemma 2, and to investigate limiting regimes other than $\lambda/\mu \to 0$, by allowing the arrival and service rates to be state dependent in the new probability measure.

This section addresses the simulation of the regime of $C \to \infty$ with $\lambda_i/\mu$ independent of $C$. Swapping $\lambda$ and $\mu$ as in the previous section will not yield BRE for high capacity regimes (even when $\lambda < \mu$) because $\mathsf{P}^*(\mathcal{R}_i) \to 0$ as $C \to \infty$. A change of measure, which is optimal in the case of a single clique cellular system with $C$ channels, is presented and then applied to the general cellular case. Notice that the former is equivalent to a queueing system with $C$ parallel servers and no waiting, and can be solved trivially using the Erlang loss formula. However, the change of measure and corresponding simulation analysis are presented to give the intuition for more general cellular models. In the general cellular context it is suboptimal, but provides a dramatic improvement over simulation using the original measure.

### 5.1.1 The single-clique case

In the case of a single-clique network, whether $\tilde{n}(t) \in \tilde{\mathcal{B}}_i$ or not, is completely determined by the aggregate state $n(t)$, so we may just use the process $\{n(t)\}$. Note that $\{n(t)\}$ is a birth and death process on $\mathcal{S} = \{0, \ldots, C\}$ with birth rate $\lambda$ and death rate $s\mu$, for $s \in \mathcal{S}$, and call $\{n(k)\}$ the embedded random walk (with the obvious abuse of notation). Let the quasi-regenerative set be $A = \{\theta + 1, \ldots, C\}$, and let $M \in \mathbb{R}$ be the random length of an $A$-cycle. Let $\tau \in \mathbb{N}$ be the first hitting time (in the embedded random walk) of $C$ within the $A$-cycle; that is, $\tau = \min\{k : S_k \geq M$ or $n(k) = C\}$, where $S_k$ is the epoch of the $k$-th event. Finally let $\mathcal{R} = \{S_\tau < M\}$. Consider a dynamic change of measure, where the birth and death process $n^*(t)$ has rates $\lambda^*(s)$ and $s\mu^*(s)$.

**Theorem 2** *Consider the ISSC estimator for $\mathsf{P}(\mathcal{R})$ using the dynamic rates*

$$\lambda^*(s) = \lambda + s(\mu - \mu^*(s)), \tag{20a}$$

$$\mu^*(s+1) = \frac{\lambda\mu}{\lambda^*(s)} \tag{20b}$$

*for $s \geq \theta + 1$, starting with $\mu^*(\theta+1) = 0$, and $\mu^*(s) = \mu$, $\lambda^*(s) = \lambda$ for $s < \theta + 1$. This has BRE in the limit of $C \to \infty$, with the likelihood ratio*

$$L_\tau = \prod_{s=\theta+1}^{C-1} \frac{\lambda}{\lambda + s(\mu - \mu^*(s))} \tag{21}$$

*when a blocking state is reached. Moreover, the variance of the estimate is zero even for finite $C$.*

**Proof :** With $\mu^*(\theta+1) = 0$, $\mathsf{P}^*(\mathcal{R}) = 1$ since the $A$-cycle is not allowed to end until a blocking state is reached. This violates the usual absolute continuity condition that for every $\omega \in \mathcal{S}$, $\mathsf{P}(\omega) > 0 \Rightarrow \mathsf{P}^*(\omega) > 0$. However, as mentioned in Section 4, $\mathsf{P}\rfloor_{\mathcal{R}}$ is absolutely continuous with respect to $\mathsf{P}^*$. This follows because on the event $\mathcal{R}$, a blocking state is reached before the $A$-cycle is over, thus no trajectory on $\mathcal{R}$ can have a transition from $\theta+1$ back to $\theta$ before $\tau$, as that would start a new $A$-cycle. Thus for every $\omega \in \mathcal{R}$, $\mathsf{P}(\omega) > 0 \Rightarrow \mathsf{P}^*(\omega) > 0$, and the change of measure is valid for the estimation of $\mathsf{P}(\mathcal{R})$.

On any path leading to the blocking boundary, $C$, any transition due to a call departure from state $s$ ($s \geq \theta+2$) to $s-1$ must necessarily be followed in some future stage by a matching transition from $s-1$ to $s$; otherwise it is impossible to achieve full occupancy. The corresponding factors contributing to $L_\tau$ are then

$$\left(\frac{s\mu}{s\mu^*(s)}\right)\left(\frac{\lambda}{\lambda^*(s-1)}\right) = \left(\frac{\lambda^*(s-1)}{\lambda}\right)\left(\frac{\lambda}{\lambda^*(s-1)}\right) = 1,$$

therefore all such loops cancel out their contributions. The only remaining contributions to $L_\tau$ are the factors for the "minimal" blocking path $\theta+1 \to \theta+2 \to \theta+3 \to \ldots \to C$, which yields (21). This is a deterministic function of $C$. Since $\mathsf{P}^*(\mathcal{R}) = 1$, $L_\tau \mathbf{1}_{\{\mathcal{R}\}} = \mathsf{E}[L_\tau \mathbf{1}_{\{\mathcal{R}\}}]$ w.p.1, and is thus optimal. ◁

Note that since this change of measure is exactly optimal for any combination of $\lambda$, $\mu$ and $C$, it is optimal for any scaling regime. This includes the two scaling regimes considered in this paper, and also the important regime where $C \to \infty$ with $\lambda/C\mu$ held fixed, which is not explicitly addressed here.

Note also that for a fixed $\mu^*(\theta+1)$, the rates are independent of $C$. The optimal adaptivity to $C$ comes from the fact that the rates change as the actual current occupancy changes.

Asymptotically, the change of measure under the new rates is analogous to the static change of measure for an $M/M/K/\infty$ queue that swaps arrival and service rate — as in (11), as the next Lemma shows. Note that acceleration of an $M/M/K/\infty$ queue can be static, since the number of active servers remains constant as the occupancy tends to infinity, leading to a constant service rate during the acceleration. Rate swapping thus gives a static change of measure. In contrast, the number of active servers in the system considered here is equal to the instantaneous occupancy, and the total service rate changes throughout the acceleration. The change of measure of (20) is dynamic, reflecting this change.

**Lemma 3** *Under the update rule of Theorem 2, for any initial* $0 \leq \mu^*(\theta+1) < \mu$,

$$\begin{aligned} \lim_{s\to\infty} \lambda^*(s)/s &= \mu \\ \lim_{s\to\infty} \mu^*(s)s &= \lambda. \end{aligned}$$

**Proof :** We will first show that $\mu^*(s) \to 0$ as $s \to \infty$. By induction, $0 \leq \mu^*(s) < \mu$ for all $s \geq \theta+1$, and hence $\{\mu^*(s)\}$ has a convergent subsequence. To see that $\mu$ is not an accumulation point, note that this would imply $\mu^*(s) = \mu - \phi(s)$ for some $\phi(s) = o(1)$, $\phi(s) \not\equiv 0$. By (20b), that

would in turn imply $\lambda^*(s) - \lambda \sim \phi(s)$, but by (20a), $\lambda^*(s) - \lambda \sim s\phi(s)$, which is a contradiction. Thus there is a strictly increasing sequence $s(m) \in \mathbb{N}$ and a $\mu^* \in [0, \mu)$ such that $\mu^*(s(m)) \to \mu^*$ as $m \to \infty$. For any such sequence $s(m)$, there is a $\delta = \mu - \mu^* > 0$ such that

$$\begin{aligned} \lambda^*(s(m+1)) - \lambda^*(s(m)) &= (s(m+1) - s(m))(\mu - \mu^*) + o(1) \\ &\geq \delta + o(1). \end{aligned}$$

Thus $\lambda^*(s(m)) \to \infty$, and by (20b), $\mu^*(s(m)) \to \mu^* = 0$. Since the sequence $s(m)$ was arbitrary, 0 is the unique accumulation point, and $\mu^*(s) \to 0$ as $s \to \infty$. From (20a),

$$\frac{\lambda^*(s)}{s} = (\mu - \mu^*(s)) + \frac{\lambda}{s} \to \mu,$$

and the result follows from (20b).

◁

### 5.1.2 The general cellular network case

Consider again the standard clock simulation model of the cellular network, and let the change of measure for estimating $\mathsf{P}(\mathcal{R}_i)$ be such that the total event rate is the same as for the original measure: $\Lambda_N^* = \Lambda_N$, when the total occupancy is $N$. When an event occurs, it is an arrival to (or a departure from) cell $j \notin \mathcal{C}_i$ with probability $\lambda_j/\Lambda_N$ (respectively, $n_j\mu/\Lambda_N$) just as for the original measure. Arrivals to (departures from) the cluster $\mathcal{C}_i$ will now occur with probability $\lambda^*(s)/\Lambda_N$ (respectively, $s\mu^*(s)/\Lambda_N$). The proportion of arrivals to the cluster that go to cell $j \in \mathcal{C}_i$ remains fixed at $\lambda_j/\lambda$. Let $s(k) = n^{(\mathcal{C}_i)}(k)$ be the embedded random walk, under the new distribution of the process, starting with $k = 1$ as the start of the $A_i$-cycle.

In the standard clock simulation, $\mu^*(s(k))$ and $\lambda^*(s(k))$ determined the event type, where $\mu^*(\cdot)$ and $\lambda^*(\cdot)$ satisfy the recurrence relation (20), starting from $\mu^*(\theta_i + 1) = 0$ (or more generally from $0 \leq \mu^*(\theta_i + 1) < \mu$). Under this change of measure, the rates are no longer constant, but depend on the state (more specifically, on the cluster occupancy), hence the name "dynamic ISSC".

It is straightforward to calculate the Radon-Nikodym derivative

$$L_{\tau_i} = \prod_{k=1}^{\tau_i - 1} \left( \left( \frac{\lambda}{\lambda^*(s(k))} \right) \mathbf{1}_{\{D_{k+1} \in \mathcal{A}_i\}} + \left( \frac{\mu}{\mu^*(s(k))} \right) \mathbf{1}_{\{D_{k+1} \in \mathcal{D}_i\}} + \mathbf{1}_{\{D_{k+1} \notin \mathcal{D}_i \cup \mathcal{A}_i\}} \right),$$

independent of the inter-arrival times, where $\mathcal{A}_i$ and $\mathcal{D}_i$ are as in (14).

In the cellular network case, it remains true that to reach a state $\tilde{n}(\tau_i) \in \tilde{\mathcal{B}}_i$ all backward transitions in the cluster $\mathcal{C}_i$ from $s$ to $s-1$ will cancel out forward transitions from $s-1$ to $s$. This follows from the observation that the cluster occupancy itself can only increase or decrease by 1 at each event that changes its occupancy. However, it is possible that an arrival to a cell $j \in \mathcal{C}_i$, $j \neq i$, will be blocked even when $\tilde{n}(k) \notin \tilde{\mathcal{B}}_i$, if $\tilde{n}(k) \in \tilde{\mathcal{B}}_j$. Since these events do not cause a change in the occupancy, $s$, their effect is not cancelled out by departure events. However, the contribution of these events to $L_{\tau_i}$ is $\lambda/\lambda^*(s) < 1$, and they cannot cause an increase in variance with respect to

the original measure. Moreover, these events become less frequent in the rare event scenario, since $\mathsf{P}^*$ does not focus on $\tilde{\mathcal{B}}_j$.

Hence the only transitions that will contribute to the final expression for $L_{\tau_i}$ are blocked arrivals, and the forward transitions from $\theta_i + 1$ to the full occupancy $n^{(\mathcal{C}_i)}(\tau_i)$. Now the ISSC estimator is

$$L_{\tau_i} = \prod_{s=\theta_i+1}^{n^{(\mathcal{C}_i)}(\tau_i)-1} \left( \frac{\lambda}{\lambda + s(\mu - \mu^*(s))} \right)^{1+b(s)},$$

where $b(s)$ is the number of blocked arrivals to cluster $\mathcal{C}_i$ while it is in state $s$. The final cluster occupancy $n^{(\mathcal{C}_i)}(\tau_i)$ satisfies $C \leq n^{(\mathcal{C}_i)}(\tau_i) \leq mC$, where $m$ is the number of cliques in $\mathcal{C}_i$, which depends on the interconnectivity of the network. The variance of $L_{\tau_i}$ is thus dependent on the variation of the distribution of the cluster occupancy when a blocking state is first reached.

## 5.2 Implementation Considerations

### 5.2.1 Subsampling the ribs

The correlation between consecutive $A_i$-cycles in the backbone can be very significant. In order to reduce this, it is possible to subsample the ribs, i.e., start a rib for every $k$th $A_i$-cycle in the backbone. This greatly increases the amount of work required to simulate the backbone. However, an $A_i$-cycle which is blocked may be very much longer than a "typical" $A_i$-cycle, especially when the load is a small fraction of the number of channels, and so the backbone is often a small proportion of the simulation time. Moreover, the backbone is shared between many cells, making subsampling very worth while. This approach is valid for any $k$, but the following heuristic arguments can be used to select a value to increase the efficiency.

Estimating the variance of a ratio can be performed following Alexopoulos and Seila [1998], even when both numerator and denominator are sample averages of Markov processes with exponentially decaying covariances, instead of iid random variables. Let

$$Z_l = kX_l \mathbf{1}_{\{l \bmod k = 0\}} - B_i T_l$$

be a random variable obtained during the $l$-th $A_i$-cycle. Here $X_l$ represents a sample of $X(T^{(i)})$ and $T_l$ represents a sample of $T^{(i)}$. The scaling of $X_l$ by $k$ cancels the subsampling by $k$, so that $E[\sum_{l=1}^{k} Z_l] = 0$. Then the estimator obtained with $S$ consecutive $A_i$-cycles satisfies

$$\mathsf{Var}[\hat{B}_i(S)] = \mathsf{Var}\left[ \frac{\frac{k}{S}\sum_{l=1}^{S} X_l \mathbf{1}_{\{l \bmod k = 0\}}}{\frac{1}{S}\sum_{l=1}^{S} T_l} \right] \approx K_1 \, \mathsf{Var}\left[ \frac{1}{S}\sum_{l=1}^{S} Z_l \right],$$

where $K_1$ is a constant depending on $E[(T^{(i)})^2]$, but independent of the sample size $S$. Recall that $T_l$ is estimated from the backbone, while $X_l$ is estimated from an independent simulation of a rib. Neglecting the resulting cross-correlation gives

$$\mathsf{Var}\left[ \frac{1}{S}\sum_{l=1}^{S} Z_l \right] \approx K_X(k/S) + K_T/S, \tag{22}$$

where constants $K_X$ and $K_T$ depend on the autocovariances of sequences $\{X_l\}_{l=1}^S$ and $\{B_i T_l\}_{l=1}^S$ respectively, but are independent of $S$. In particular, if the sequences $\{X_l\}$ and $\{T_l\}$ have no autocorrelation, then $K_X = \mathsf{Var}[X_l]$ and $K_T = \mathsf{Var}[B_i T_l]$. These constants can be estimated from a single simulation by estimating the variance (22) using different rates of subsampling, $k_1$ and $k_2$; the difference between these estimates is approximately $K_X(k_2 - k_1)/S$. Note that $K_X$ depends on $k$ unless $\{X_l\}$ is uncorrelated. Thus (22) should be seen as a linearisation, and the values of $k_1$ and $k_2$ chosen accordingly.

Also,

$$\text{CPU}[\hat{B}_i(S)] = l_X S/k + l_T S,$$

where $l_X$ and $l_T$ are the mean lengths of $A_i$-cycles corresponding to the ribs and the backbone, respectively. The relative efficiency (6) is then maximized by setting

$$k^* \approx \sqrt{\frac{l_X}{K_X}\frac{K_T}{l_T}}. \tag{23}$$

These values can be estimated coarsely from a pilot simulation, which can also serve as the "warm-up" to achieve steady state. When the accelerated simulation algorithm is used (i.e., the backbone is not accelerated but the ribs are), $k^*$ is of the order of 10, to within one order of magnitude. When the non-accelerated simulation algorithm is used (i.e., when neither the backbone nor ribs are accelerated), $k^*$ is actually often less than 1, indicating that there may be value in running multiple ribs from the same point in the backbone. For the numerical results in this paper, $k$ was not selected from (23). For cases where the ribs were expected to be short, including all cases in Section 4, $k = 1$ was used. For more difficult cases, $k = 10$ was used. These values were selected from prior simulations.

### 5.2.2 Variance estimation

The above expressions for variance are very approximate, and are only appropriate for determining suitable subsampling rates. The variance of the estimator, $\hat{B}_i$, was determined using the methods described in Alexopoulos and Seila [1998]. This uses batch means to determine the variances and covariance of the estimators for the numerator and denominator of (13) which can be used to derive the variance of $\hat{B}_i$. This in turn can be used to construct a confidence interval for $B_i$. Since the estimates for the individual $B_i$s are derived from the same backbone, the variance of the overall estimator, $\hat{B}$, also contains some covariance terms. These terms are generally expected to be small because the main source of variance is estimating the numerator of (13), and the ribs for different cells are simulated independently, albeit with dependent initial states. Hence these covariance terms may be neglected, as they are in the numerical results presented in this paper. If precise confidence intervals are required, then one option is to estimate the $B_i$'s using independent backbones, with the obvious substantial reduction in efficiency.

### 5.2.3 Choice of quasi-regenerative cycles

When the load, $\lambda/\mu$, is not negligible, the probability that a cluster will be completely empty is small. Thus if $\theta$ is too small, like $\theta = 0$ as is used for single server queues, then the $A_i$-cycles

become unmanageably long. This has several implications. The most obvious result is that the simulation time increases in proportion. The seriousness of this is to some extent alleviated by the fact that longer $A_i$-cycles produce better estimates of the proportion of time spent in blocking states within an $A_i$-cycle.

The more serious problem with long $A_i$-cycles is that the blocking states become a small proportion of the $A_i$-cycles, even given that blocking occurs. The assumption behind the IS scheme proposed here is that $A_i$-cycles which contain blocking states are rare events, but if an $A_i$-cycle does contain blocking states, they form a significant proportion of it. Thus the system is accelerated until the first blocking state is reached, and is then allowed to relax back to finish its $A_i$-cycle. If $\lambda/\mu$ and $C$ are both large, then empty clusters become rarer than blocking, and most $A_i$-cycles contain a period in blocking states, reducing the effectiveness of the acceleration.

The length of $A_i$-cycles is minimized by maximizing the rate of crossing the boundary between sets $A_i$ and $A_i'$ of the embedded Markov chain. Note that the stationary rate at which the process crosses from any state with $n^{(\mathcal{C}_i)} = \theta_i$ to $n^{(\mathcal{C}_i)} = \theta_i + 1$ equals the stationary rate at which it crosses from $n^{(\mathcal{C}_i)} = \theta_i + 1$ to $n^{(\mathcal{C}_i)} = \theta_i$. This rate is state dependent:

$$\lambda\pi(\tilde{n}), \quad \tilde{n} \in \tilde{\mathcal{S}}(\theta_i),$$

where $\tilde{\mathcal{S}}(\theta_i)$ is the subset of the state space where $n^{(\mathcal{C}_i)} = \theta_i$. This is maximized at the mode of the stationary distribution $\pi(\cdot)$, which is $\theta_i \approx \sum_{j\in\mathcal{C}_i} \lambda_j/\mu$, assuming that blocking does not significantly distort the state distribution, and that the mean and mode of $\pi$ are close.

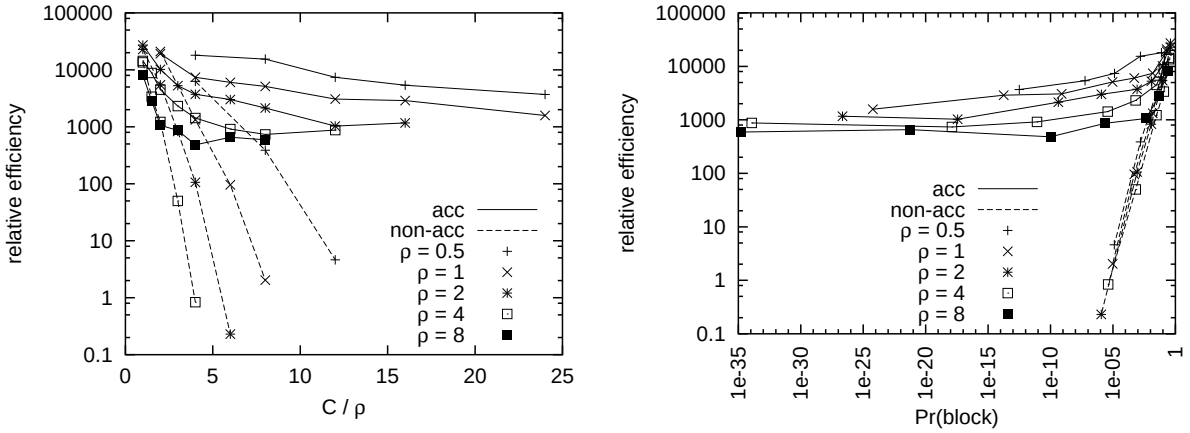
Next, in order to guarantee that $A_i \supset \tilde{\mathcal{B}}_i$, and hence that $L_{\tau_i} \leq 1$ whenever blocking occurs, it was required that $\theta_i < C$. Thus we use the value

$$\theta_i = \max\left(C - 1, \sum_{j\in\mathcal{C}_i} \lambda_j/\mu\right). \tag{24}$$

There may also be benefit in using values of $\theta_i$ larger than $\sum_{j\in\mathcal{C}_i} \lambda_j/\mu$ (but less than $C$). It means less simulation for the ribs, but longer $A_i$-cycles in the backbone. However, the $A_i$-cycles in the backbone are shared between all cells.

## 5.3 Simulation Results

The dynamic change of measure was shown to have BRE for the probability of blocking states occurring within an $A_i$-cycle, $\mathsf{P}(\mathcal{R}_i)$, in the case of a single cell (or more generally a single clique). However, (13) shows that the efficiency of estimating the blocking probability also depends on the efficiency of estimating $\mathsf{E}[X_i(T^{(i)})|\mathcal{R}_i]$ and $\mathsf{E}[T^{(i)}]$. Figure 5 shows the relative efficiency for the actual blocking probability in the single cell case, for the accelerated and non-accelerated methods. Here $\rho = \lambda_i/\mu$ is the load in Erlangs. Again 100 batches of $10^5$ $A_i$-cycles were used. Subsampling was by a factor of $k = 1$ (no subsampling). As $C$ increases, the proportion of time in each $A_i$-cycle spent in a blocking state decays, even on those $A_i$-cycles which contain blocking. This accounts for the slight reduction in efficiency as the blocking rate decreases. However, this reduction is very much smaller than that which occurs without acceleration.




(a) relative efficiency against normalized capacity

(b) relative efficiency against blocking

Figure 5: Relative efficiency for importance sampling (ISSC) and $A$-cycle framework without IS, both in a single cell network.

Figure 6 shows the relative efficiency of the accelerated and non-accelerated methods for a seven-cell clique packing system, for a range of loads, $\rho = \lambda_i/\mu$, and a range of normalized capacities, $C/\rho$. The load on each cell was the same.

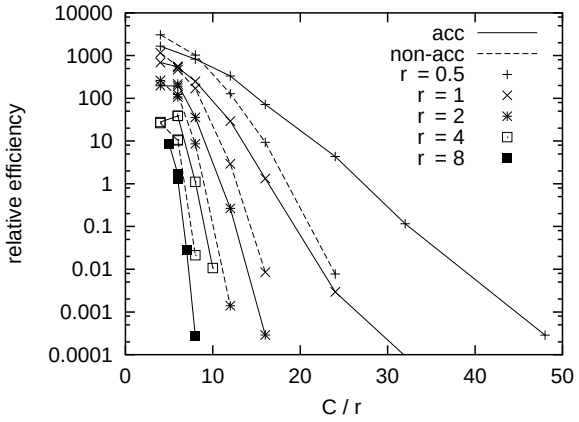
Again 100 batches of $10^5$ $A_i$-cycles were simulated from a backbone shared by all cells, $i$. The ISSC subsamples the $A_i$-cycles in the backbone by a factor of $k = 10$. The simulations without acceleration use $k = 1$, as the rib $A_i$-cycles are shorter and the variance of the corresponding estimator is higher. These results do not suggest that ISSC has BRE for network blocking as $C \to \infty$. However, IS substantially reduces the rate at which the performance degrades for large $C$.

The reason for the reduced efficiency is that the acceleration is applied to all cells in a cluster. For a constant load, as $C$ increases the (true) expected cluster occupancy on blocking satisfies $\mathsf{E}[n^{(c_i)}|\tilde{\mathcal{B}}_i]/C \to 1$, since arrivals at each cell are independent, and only one clique need be full. However, because the acceleration is applied to all cells in the cluster, the cells outside the clique which caused blocking are also filled up. Thus the expected cluster occupancy at blocking under the new measure is significantly larger than under the original measure. That is, outcomes with a high cluster occupancy are accelerated too much, thus increasing the variance.

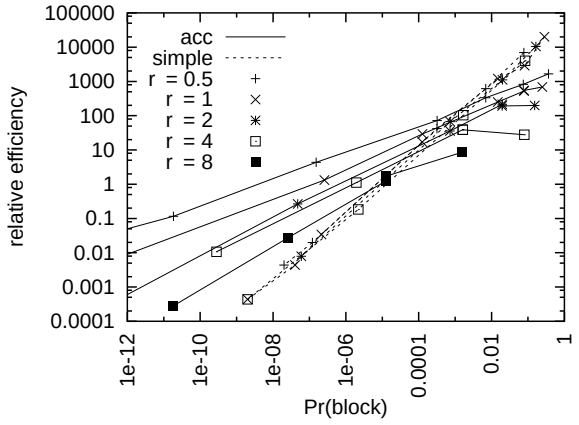
It seems from Figure 6(a) that the improvement decreases as the load increases. However, this is largely due to the fact that the rate at which the blocking decreases for increasing $C$ is different. Figure 6(b) shows the relative efficiency against the blocking probability. This shows that the change in slope of the curves is similar over a range of loads.

Figure 6(b) also uses the "simple" estimator used in Figure 4 which merely counts blocked calls. In this case, this shows an approximately constant improvement by a factor of around 10 compared with the non-accelerated $A_i$-cycles.

Figure 7 shows the results for the same simulation parameters as Figure 6, but for a seven cell




(a) relative efficiency against normalized capacity




(b) relative efficiency against blocking

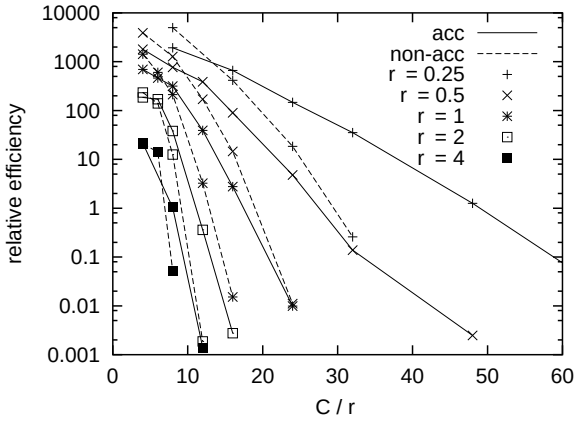
Figure 6: Relative efficiency for importance sampling (ISSC), $A_i$-cycle framework without IS, and the simple simulation, all in a 7 cell network with clique packing.

network when existing calls are not rearranged and first-fit channel assignment is used. Note that in the range which is of most interest to engineers, with blocking between $10^{-6}$ and $10^{-2}$, the acceleration consistently outperforms the non-accelerated simulation.

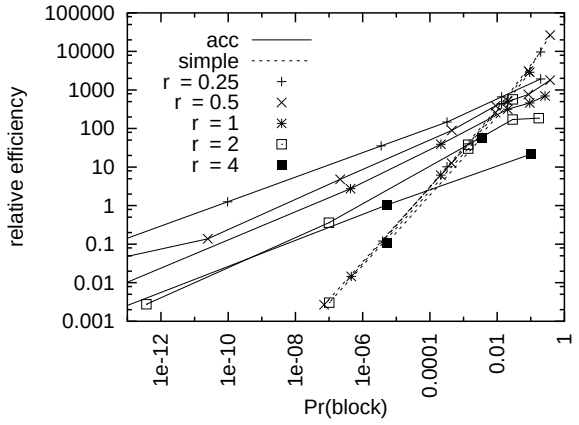
# 6 Concluding Remarks

This paper has addressed fast simulation for estimating blocking probabilities in cellular networks. Blocking is a rare event when the load is low, or the number of available channels is high. We implemented the main ideas of fast simulation using the standard clock framework for simulation. In the case of low load, the proposed change of measure yields an estimator that has bounded relative error. For high capacity systems we propose a change of measure that yields a zero variance estimator for the single clique case; we were unable to prove any efficiency results with this change of measure (in the rare event setting) for the general network case. Nonetheless, this change of measure provides significant improvements over standard simulation in more general networks when events are rare. The reason for the suboptimality is that trajectories with large numbers of calls in a cluster get accelerated disproportionately.

There is much scope for improvement of this technique. The performance for relatively high blocking probability ($> 10^{-3}$) is poor because of the variable number of calls in a cluster when blocking first occurs. This may be improved by reducing the acceleration applied to cells in the cluster which are not in the fullest clique. Also, the current need to use separate $A_i$-cycles to estimate the blocking probability of each cell limits the scalability of the technique. It will also be important to expand the technique to other performance measures and more general system models, such as determining the probability of dropping due to blocked handovers in a system




(a) relative efficiency against normalized capacity




(b) relative efficiency against blocking

Figure 7: Relative efficiency for importance sampling (ISSC), $A_i$-cycle framework without IS, and the simple simulation, all in a 7 cell network *without* call rearrangement.

incorporating user mobility.

# Acknowledgement

The authors would like to thank the guest editors and referees, especially Perwez Shahabuddin, who thoroughly reviewed early versions of this paper. His insightful suggestions helped us to create a much improved contribution.

# References:

Alexopoulos, C., and Seila, A. 1998. Output data analysis. In *Handbook of Simulation*, ed. J. Banks, chapter 7, 225–272. New York, NY: John Wiley and Sons.

Asmussen, S., and Nielsen, H. M. 1995. Ruin probabilities via local adjustment coefficients. *J. Appl. Probab.*, 32:736–755.

Boucherie, R. J., and Mandjes, M. 1998. Estimation of performance measures for product form cellular mobile communications networks. *Telecommunication Systems*, 10:321–354.

Breiman, L. 1992. *Probability*. Classics in Applied Mathematics, Philadelphia, PA: SIAM.

Chang, C.-S., Heidelberger, P, and Shahabuddin, P. 1995. Fast simulation of packet loss rates in a shared buffer communications switch. *ACM Trans. Model. Comput. Simul.*, 5(4):306–325.

Choudhury, G. L., Leung, K. K, and Whitt, W. 1995. An algorithm to compute blocking probabilities in multi-rate multi-class multi-resource loss models. *Adv. Appl. Prob.*, 27:1104–1143.

Cox, D. C., and Reudink, D. O. 1972. Dynamic channel assignment scheme in large cellular-structured mobile communication systems. *IEEE Trans. Commun.*, COM-26:432–438.

Cox, D. C., and Reudink, D. O. 1973. Increasing channel occupancy in large scale mobile radio systems: dynamic channel reassignment. *IEEE Trans. Vehic. Technol.*, VT-22:218–222.

Devetsikiotis, M., and Townsend, K. 1993. Statistical optimization of dynamic importance sampling parameters in efficient simulation of communication networks. *IEEE/ACM Trans. Networking*, 1(3):293–305.

Dziong, Z., and Roberts, J. W. 1987. Congestion probabilities in a circuit-switched integrated services network. *Perf. Eval.*, 7:267–284.

Everitt, D., and Macfadyen, N. W. 1983. Analysis of multicellular mobile radiotelephone systems with loss. *Br. Telecom Technol. J.*, 1(2):37–45.

Gaivoronski, A., and Messina, E. 1996. Optimization of stationary behavior of general stochastic discrete event dynamic systems. In *Proceedings of International Workshop on Discrete Event Systems, WODES '96*, ed. R. Smedinga, M. P. Spathopoulos, and P. Kozák, 238–243. London, UK: Institute of Electrical Engineers.

Glynn, P. W., and Whitt, W. 1992. The asymptotic efficiency of simulation estimators. *Operations Research*, 40(3):505–520.

Heidelberger, P., Shahabuddin, P, and Nicola, V. 1994. Bounded relative error in estimating transient measures of highly dependable non-Markovian systems. *ACM Trans. Model. Comput. Simul.*, 4(2):137–164.

Kelly, F. P. 1979. *Reversibility and Stochastic Networks*. New York, NY: John Wiley and Sons.

Kelly, F. P. 1991. Loss networks. *The Annals of Probability*, 1:319–378.

Lassila, P., and Virtamo, J. 2000. Nearly optimal importance sampling for Monte Carlo simulation of loss systems. COST report COST257TD(00), Helsinki University of Technology.

Lassila, P. E., and Virtamo, J. T. 1998. Efficient Monte Carlo simulation of product form systems. In *Proc. Nordic Teletraffic Seminar (NTS) 14*, 355–366. Copenhagen, Denmark: Available from `http://keskus.hut.fi/tutkimus/cost257/publ/efmcsim.pdf`.

L'Ecuyer, P., and Champoux, Y. 1996. Importance sampling for large ATM-type queueing networks. In *Proceedings of the 1996 Winter Simulation Conference*, 309–316. Piscataway, NJ: IEEE Press.

Lee, W. C. Y. 1995. *Mobile Cellular Telecommunications*. 2nd ed. New York, NY: McGraw Hill.

Li, W., and Alfa, A. S. 2000. Channel reservation for handoff calls in a PCS network. *IEEE Trans. Vehic. Technol.*, 49(1):95–104.

Mandjes, M. 1997. Fast simulation of blocking probabilities in loss networks. *European Journal of Operations Research*, 101:393–405.

Mitra, D., and Morrison, J. A. 1994. Erlang capacity and uniform approximations for shared unbuffered resources. *IEEE/ACM Trans. Networking*, 2(6):558–570.

Mouly, M., and Pautet, M.-B. 1992. *The GSM System for Mobile Communications*. Olympia, WA: Telecom Publishing.

Nelson, R. D. 1993. The mathematics of product form queuing networks. *Computing Surveys*, 25(3):339–369.

Neuts, M. F. 1981. *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach*. Baltimore, MD: Johns Hopkins University Press.

Nicola, V. F., Shahabuddin, P, Heidelberger, P, and Glynn, P. 1993. Fast simulation of steady-state availability in non-Markovian highly dependable systems. In *Proc. Twenty-Third International Symposium on Fault-Tolerant Computing*, 38–47. Piscataway, NJ: IEEE Press.

Pallant, D. L., and Taylor, P. G. 1995. Modeling handovers in cellular mobile networks with dynamic channel allocation. *Operations Research*, 43(1):33–42.

Pinsky, E., and Conway, A. E. 1992. Computational algorithms for blocking probabilities in circuit-switched networks. *Ann. Operat. Res.*, 35:31–41.

Raymond, P.-A. 1991. Performance analysis of cellular networks. *IEEE Trans. Commun.*, 39(12):1787–1793.

Redl, S. M., Weber, M. K, and Oliphant, M. W. 1995. *An Introduction to GSM*. Norwood, MA: Artech House.

Reiser, M., and Lavenberg, S. S. 1980. Mean-value analysis of closed multichain queuing networks. *J. ACM*, 27(2):313–322.

Ross, K. W., Tsang, D. H. K, and Wang, J. 1994. Monte Carlo summation and integration applied to multiclass queuing networks. *J. ACM*, 41(6):1110–1135.

Ross, K. W., and Wang, J. 1992. Monte-Carlo summation applied to product-form loss networks. *Probability in the Engineering and Information Sciences*, 6:323–348.

Ross, S. M. 1997. *Simulation*. 2nd ed. Boston: Academic Press.

Sadowsky, J. S. 1991. Large deviations theory and efficient simulation of excessive backlogs in a $GI/GI/m$ queue.

*IEEE Trans. Autom. Control*, 36(12):1383–1394.

Shahabuddin, P. 1994. Importance sampling for the simulation of highly reliable Markovian systems. *Management Science*, 40(3):333–352.

Vakili, P. 1991. Using a standard clock technique for efficient simulation. *Operations Research Letters*, 10:445–452.

Vázquez-Abad, F. J., and Andrew, L. May, 2000. Filtered Gibbs sampler for estimating blocking probabilities in WDM optical networks. In *Proc. 14th European Simulation Multiconference*, ed. D. Landeghem, 548–555. Ghent, Belgium: Society for Computer Simulation.

Vázquez-Abad, F. J., and LeQuoc, P. 2001. Sensitivity analysis for ruin probabilities. *Journal of the Operational Research Society*, 52(1):71–81.

Yates, J. 1997. Performance analysis of dynamically-reconfigurable wavelength division multiplexed networks. PhD thesis, University of Melbourne, Australia.

Zahorjan, J., Eager, D. L, and Sweillam, H. 1988. Accuracy, speed and convergence of approximate mean value analysis. *Perf. Eval.*, 8:255–270.