Efficient Heuristics for the Simulation of Population Overflow in Series and Parallel Queues

Victor F. Nicola Tatiana S. Zaburnenko
Faculty of Electrical Engineering, Mathematics and Computer Science
University of Twente, P.O. Box 217
7500 AE Enschede, The Netherlands
Email: {v.f.nicola, t.s.zaburnenko@ewi.utwente.nl}

***Abstract*— In this paper we propose state-dependent importance sampling heuristics to estimate the probability of population overflow in Markovian networks of series and parallel queues. These heuristics capture state-dependence along the boundaries (when one or more queues are empty) which is critical for the asymptotic optimality of the change of measure. The approach does not require difficult (and often intractable) mathematical analysis or costly optimization involved in adaptive importance sampling methodologies. Experimental results on tandem and parallel networks with a moderate number of nodes yield asymptotically efficient estimates (often with bounded relative error) where no other state-independent importance sampling techniques are known to be efficient. Insight drawn from simulating basic networks in this paper promises the applicability of the proposed methodology to larger networks with more general topologies.**

I. Introduction

Efficient simulation of queueing networks has long been the focus of much research, owing to its applicability in the modeling, analysis and dimensioning of logistic, production and communication networks. In particular, the analysis of rare yet critical events, such as buffer overflow, has been a challenging problem attracting a huge number of scientists and practitioners over the past few decades. Despite vast theoretical and empirical efforts, progress is markedly slow as evidenced by the lack of generally applicable techniques and tools equipped to deal with the difficulty inherent in simulating rare events in queueing networks.

Among the most effective methodologies researched and applied so far are those based on importance sampling (see, e.g., [9], [18], [2], [19]) and importance splitting (see, e.g., [15], [31], [14]) techniques. (Importance sampling is the methodology adopted in this paper.) However, the success of these techniques has been mostly limited to simulations of single server queues, under restrictive assumptions regarding the underlying arrival and service processes (e.g., having light-tail distributions [26], [29]). Some queueing networks of certian topologies have also been considered (see, e.g., [8], [23], [22]). The overflow event of interest is usually that of an individual buffer or that of the total network population. Most previous work has focused on estimating the probability of the latter event, given some initial network state (typically, starting from an empty network).

Until recently, only state-independent importance sampling heuristics were developed and considered for analysis. In these heuristics, the change of measure is 'static' and independent of the network state (i.e., the number of customers at each node in a Jackson network). A relatively simple (and well known) heuristic change of measure for simulations of population overflow in queueing networks is that proposed in [26] and further investigated in [11] and [12]. However, even for the simplest Jackson queueing network (e.g., 2-node tandem network), the effectiveness of this heuristic is limited to only some region of the (arrival and service) parameters space (see [16], [17], [7]). (We use the term 'effectiveness' interchangeably with 'asymptotic efficiency,' see Section II-B for a precise definition.) Effective bandwidth methods have been used to develop heuristics for simulating overflow of an individual buffer in some specific class of networks, e.g., feed-forward [10] and in-tree [8] fluid flow networks. More recently [20] proposed a heuristic change of measure for simulating the overflow of an individual buffer in queueing networks of general topology and arbitrary routing (i.e., including feedback). Under some regularity conditions, the heuristic is provably effective for any feasible set of network parameters (e.g., arrival and service rates in a Jackson network). This marks a significant advance, since the heuristic is applicable to a broad class of networks under less restrictive (and possibly, non-Markovian) assumptions. State-independence is a common feature of the heuristics mentioned above. Also, none of them is provably effective for simulating population overflow in networks with an arbitrary and feasible set of parameters ([16], [20], [7]).

Based on Markov additive process formulation of a two-node tandem network and large deviations arguments, work in [22] reveals that a state-dependent change of measure is effective where, provably, no effective state-independent change of measure exists ([16], [7]). This finding triggered more research to develop methodologies for obtaining state-dependent importance sampling heuristics that are demonstrably effective (at least empirically). Initially, the work in [6] which uses an adaptive optimization technique based on the method of cross-entropy [28] to approximate the 'optimal' state-dependent change of measure. Later, a similar adaptive approach based on stochastic approximation is introduced in [1]. To date, these approaches appear to be the most promising for application to Markovian (Jackson) networks of general topology. A drawback, however, is the computational and storage demands for large state-space models associated with large networks.

Most recently, tandem networks are considered in [24] and [33] with the aim to develop a (heuristic) state-dependent change of measure that is sufficiently close to the "optimal" without tedious mathematical analyses or costly optimizations involved in adaptive methodologies. The key observation is that the "optimal" change of measure depends on the network state only along and close to the boundaries (when one or more nodes are empty), and tends to become state-independent in the interior of the state-space. Therefore, if we can determine the change of measure along the boundaries and at the interior of the state-space, then we may be able to combine them appropriately to construct a state-dependent change of measure that approximates the "optimal" one in the entire state-space. The proposed methodology is dubbed "state-dependent heuristic" or SDH in short. Experimental results using the so obtained change of measure to simulate tandem networks yield estimates with a bounded relative error for almost any feasible set of network parameters (see [33] and [24]). Only when the lowest service rates are equal (i.e., there is more than one bottleneck), then the relative error increases at most linearly with the overflow level.

In this paper we review and refine some recent work using this heuristic approach to simulate tandem networks, and further extend its utility for the efficient simulation of parallel queues. Experimental results to estimate the probability of population overflow in these (tandem and parallel) networks produce asymptotically efficient estimates, with relative error increasing at most linearly with the overflow level. The proposed heuristics are robust and effective, yet easier-to-implement and could be more efficient than those based on adaptive importance sampling methodologies (e.g., [6]), particularly for large networks.

In Section II we give some preliminaries, introduce the basic model and define the probability of interest. The importance sampling technique is briefly reviewed. In Section III we motivate the proposed SDH and give its formal representation for tandem and parallel networks, respectively. In Section IV we present experimental results and comparisons with other known methods to estimate the probability of population overflow in some example networks. Conclusions highlighting the advantages and challenges associated with the proposed methodology are given in Section V.

II. PRELIMINARIES

The queueing network model and associated notation are introduced in Section II-A. A brief review of importance sampling and some properties of simulation estimators are provided in Section II-B.

A. Model and Notation

Consider a Jackson network consisting of n nodes (queues), each having its own buffer of infinite size. Customers arrive at node i $(i = 1, \ldots, n)$ according to a Poisson process with rate λ_i. The service time of a customer at node i is exponentially distributed with rate μ_i $(i = 1, \ldots, n)$. Customers that leave node i join node j with probability p_{ij} $(i = 1, \ldots, n; j = 1, \ldots, n)$ or leave the network with probability p_{ie} $(i = 1, \ldots, n)$. We also assume that the queueing network is stable, i.e., $\gamma_i < \mu_i$ for all $i = 1, \ldots, n$, where γ_i is the total arrival rate at node i, as determined from the traffic equations

$$\gamma_i = \lambda_i + \sum_{\forall j} \gamma_j \, p_{ji} \; .$$

Let $X_{i,t}$ $(i = 1, \ldots, n)$ denote the number of customers at node i at time $t \geqslant 0$ (including those in service). Then the vector $\mathbf{X_t} = (X_{1,t}, X_{2,t}, ..., X_{n,t})$ is a Markov process representing the state of the network at time t. Denote by S_t the total number of customers in the network (network population) at time t, i.e., $S_t = \sum_{i=1}^{n} X_{i,t}$.

Assuming that the initial network state is $\mathbf{X_0}$ (usually, $\mathbf{X_0} = (0, 0, ..., 0)$ corresponding to an empty network), we are interested in the probability that the network population reaches some high level $L \in \mathbb{N}$ before becoming empty. We denote this probability by $\gamma(L)$ and refer to it as the *population overflow probability*, starting from the initial state $\mathbf{X_0}$. Since the associated event is typically rare, importance sampling may be used to efficiently estimate this probability.

B. Importance Sampling

Importance sampling involves simulating the system under different underlying probability distributions so as to increase the frequency of typical sample paths leading to the rare event. Formally, let w be a sample path over the interval $[0, t]$. Then, the likelihood ratio associated with w is given by $W_t(w) = \frac{P(w)}{\tilde{P}(w)}$, where $P(w)$ and $\tilde{P}(w)$ are the probabilities (or likelihoods) of sample path w under the original and the new measure, respectively. Obviously,

$\tilde{P}(w) > 0$ whenever $P(w) > 0$. Starting from $\mathbf{X_0}$, define τ as the first time S_t hits level L or level 0, then

$$\gamma(L) \;=\; \mathbb{E}\, I_{\{S_\tau = L\}} \;=\; \tilde{\mathbb{E}}\, W_\tau\, I_{\{S_\tau = L\}}\,, \tag{1}$$

where $I_{\cdot}$ is the indicator function taking the value 1 if the event . is true and 0 otherwise, and W_τ is the likelihood ratio over the interval $[0, \tau]$. $\mathbb{E}$ and $\tilde{\mathbb{E}}$ are the expectations under the original and the new changes of measure, respectively. The variance of the estimator $\tilde{\mathbb{E}}\, W_\tau\, I_{\{S_\tau = L\}}$ is given by

$$\tilde{\mathbb{E}}\, {W_\tau}^2\, I_{\{S_\tau = L\}} \;-\; (\gamma(L))^2\,. \tag{2}$$

The relative error is the ratio of the standard deviation of the estimator over its expectation, i.e.,

$$\sqrt{\frac{\tilde{\mathbb{E}}\, {W_\tau}^2\, I_{\{S_\tau = L\}}}{(\gamma(L))^2} \;-\; 1}\,. \tag{3}$$

The estimator $\tilde{\mathbb{E}}\, W_\tau\, I_{\{S_\tau = L\}}$ is said to be *asymptotically efficient* if its relative error grows at sub-exponential (e.g., polynomial) rate as $L \to \infty$ (i.e., as $\gamma(L) \to 0$). Formally, let $\lim_{L\to\infty} \frac{1}{L} \log \gamma(L) = \theta$. That is, θ is the asymptotic decay rate of the overflow probability $\gamma(L)$ as $L \to \infty$. Then, from Equation 3, asymptotic efficiency is obtained if

$$\lim_{L\to\infty} \frac{1}{L} \log \tilde{\mathbb{E}}\, {W_\tau}^2\, I_{\{S_\tau = L\}} \;=\; 2\theta\,. \tag{4}$$

The estimator is said to have *bounded relative error* if its relative error is bounded in L as $\gamma(L) \to 0$. This implies asymptotic efficiency, however, it is a stronger and more desirable property for any importance sampling estimator. It has long been a topic of intensive and mathematically involved research to develop asymptotically efficient changes of measure to estimate overflow probabilities in queueing (and in particular, tandem) networks. Some early work can be found in, e.g., [26], [12], [13], [3], and [30].

It is important to note that a change of measure may, in general, depend on the state of the system, even if the original underlying distributions do not depend on the system state. For instance, the arrival and service rates in a Markovian queueing network are typically fixed and independent of the network state (i.e., the buffer content at each node). However, a change of measure to be used in importance sampling simulation may involve new arrival and service rates that depend on the state of the network. Recent works confirm that state-dependent changes of measure are generally more effective in simulations of rare events in queueing networks (see, e.g., [22], [6]). Therefore, in this paper we aim at developing heuristics to approximate the "optimal" state-dependent change of measure.

III. STATE-DEPENDENT HEURISTICS

Even for the simplest (2-node) tandem network, no state-independent change of measure that is asymptotically efficient over the entire range of feasible network parameters (arrival and service rates) is known to exist ([16], [7]). Only state-dependent change of measures, developed through analysis (e.g., [22]) or determined using adaptive importance sampling (e.g., [6], [1]), have shown to be effective for any feasible set of network parameters. Unfortunately, however, these methodologies have some drawbacks. It is not clear whether the analysis in [22] can be easily extended to larger and more general networks. On the other hand, large state-space limits the effectiveness of adaptive importance sampling to Markovian networks with a small number of nodes ([1], [6]). Evidently, it is of much interest to find more practical and robust approaches to develop effective state-dependent changes of measure.

Roughly speaking, a state-dependent change of measure allows to influence sample paths behaviour more freely than a state-independent one. Therefore, it is more suited to change the average sample path behaviour of the simulated system so as to follow the "optimal" (most likely) path leading to the rare event. In a Jackson network, such a change of measure depends on the state of the network, i.e., the number of customers at each network node. Indeed, theoretical and empirical results in [22] and [6] indicate that an effective (asymptotically optimal) state-dependent change of measure always exists and can often be determined (through analysis or adaptively), also when provably no state-independent change of measure exists. Even when the latter exists, a properly determined state-dependent change of measure is almost always more effective. For the "optimal" change of measure, state dependence is also shown to be strong along the boundaries of the state-space (i.e., when one or more buffers are empty) and diminishes toward the interior of the state-space (i.e., when the contents of all buffers are sufficiently large). Capturing dependencies along the boundaries have shown to be very crucial for the asymptotic efficiency ("optimality") of the change of measure.

The above observation suggests that if we know the "optimal" change of measure along the boundaries and in the interior of the state-space, then we might be able to construct a change of measure that approximates the "optimal" one over the entire state-space. If the approximation is sufficiently good, then the so constructed change of measure may yield asymptotically efficient estimators. To realize the above idea we need to determine the "optimal" change of measure in the interior and along the boundaries of the state-space. In [24] heuristics based on combining known large deviations results and time-reversal arguments are used to construct such a change of measure for the 2-node tandem network (see Section III-A). Empirical results show that it produces asymptotically efficient estimators, with a bounded relative error for almost the entire feasible range of network parameters. In Section III-A we refine and generalize the change of measure in [24] to tandem networks with any number of nodes. In Section III-B we propose a state-dependent heuristic change of measure for the efficient simulation of parallel networks with any number of nodes.

A. SDH for the n-node Tandem Network

Let λ and μ_i $(i = 1, \ldots, n)$ be the arrival rate at the first node and the service rate at the $i-th$ node, respectively. Denote by $\rho_i = \frac{\lambda}{\mu_i}$ the traffic intensity at node i, and assume that $\rho_1 \leqslant \rho_2 \leqslant \ldots \rho_n < 1$. We note, however, that this ordering is not a restriction, since the probability of population overflow is invariant with respect to the placement order of nodes in a Jackson tandem network [32]. Without loss of generality we assume that $\lambda + \sum_{i=1}^{n} \mu_i = 1$.

Let $x_i, i = 1, \ldots, n$, be the number of customers at node i at time t. Then the state of the network, $\mathbf{X_t}$, is given by the vector $\mathbf{x} = (x_1, x_2, ..., x_n)$. The new rates may depend on the network state and, therefore, they are functions of the vector $\mathbf{x}$. Denote by $\tilde{\lambda}(\mathbf{x})$ and $\tilde{\mu}_i(\mathbf{x})$ $(i = 1, \ldots, n)$ the rates under the new change of measure, and by $\mathbf{SDH}_i^{\mathrm{Tn}}(\mathbf{x})$ $(i = 1, \ldots, n)$ the $(n+1) \times (n+1)$ SDH transformation matrix to simulate population overflow in the first i nodes of the n-node tandem network. (The superscript Tn is a reference to the entire n-node tandem network.) Thus, $\mathbf{SDH}_n^{\mathrm{Tn}}(\mathbf{x})$ is a linear operator transforming the original rates into the new rates used to simulate population overflow in the network Tn. (For convenience, we occasionally abuse notation by dropping the vector $\mathbf{x}$).

Consider a 2-node tandem network $(\mathrm{T2})$ with arrival and service rates λ, μ_1 and μ_2, respectively. Denote by $\mathcal{M}_0$ the original change of measure, by $\mathcal{M}_1$, the change of measure: $\tilde{\lambda} = \mu_1$, $\tilde{\mu_1} = \lambda$, $\tilde{\mu_2} = \mu_2$, and by $\mathcal{M}_2$, the change of measure: $\tilde{\lambda} = \mu_2$, $\tilde{\mu_1} = \mu_1$, $\tilde{\mu_2} = \lambda$. In words, our proposed state-dependent heuristic for the 2-node tandem network can be described as follows: Initially (starting from an empty network) apply $\mathcal{M}_0$ (i.e., no change of measure). As the number of customers at node 1 (x_1) increases, gradually go from $\mathcal{M}_0$ to $\mathcal{M}_1$. At the same time, as the number of customers at node 2 (x_2) increases, gradually go from $\mathcal{M}_0$ and/or $\mathcal{M}_1$ to $\mathcal{M}_2$. Thus, for a sufficiently large x_2, $\mathcal{M}_2$ is applied. As x_2 decreases, gradually go from $\mathcal{M}_2$ to $\mathcal{M}_0$ and/or $\mathcal{M}_1$, depending on x_1.

The above state-dependent change of measure for the 2-node tandem network $(\mathrm{T2})$ is formally expressed in the following proposition.

Proposition 1 (SDH for the 2-node Tandem Network)
Define $[a]^+ = \max(a, 0)$ and $[a]^1 = \min(a, 1)$, and let $1 \leq b_i \leq \infty$, $i = 1, 2$, be a fixed integer. The following equation expresses the proposed state-dependent change of measure for the 2-node tandem network [24]:

$$\begin{bmatrix} \tilde{\lambda} \\ \tilde{\mu}_1 \\ \tilde{\mu}_2 \end{bmatrix} = \mathbf{SDH}_2^{\mathrm{T2}} \begin{bmatrix} \lambda \\ \mu_1 \\ \mu_2 \end{bmatrix}, \tag{5}$$

$$\mathbf{SDH}_2^{\mathrm{T2}} = \left[\frac{x_2}{b_2}\right]^1 \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} + \left[\frac{b_2 - x_2}{b_2}\right]^+ \mathbf{SDH}_1^{\mathrm{T2}},$$

$$\mathbf{SDH}_1^{\mathrm{T2}} = \left[\frac{x_1}{b_1}\right]^1 \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} + \left[\frac{b_1 - x_1}{b_1}\right]^+ \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

The first matrix is the identity matrix with the first and the third rows interchanged; this corresponds to interchanging the arrival rate λ with the service rate μ_2. The second matrix is the identity matrix with the first and the second rows interchanged; this corresponds to interchanging the arrival rate λ with the service rate μ_1. The third matrix is the identity matrix, corresponding to no change of measure. Note that the new arrival and service rates (essentially) do not correspond to a stable network, and they depend on the state of the network (x_1, x_2), the number of customers at each node.

Consistent with an earlier variant of the same heuristic in [24], and as set in our experiments with 2-node tandem networks, let $b_1 = 1$. Then the above change of measure depends only on x_2. In this case, the heuristic implies a gradual "transition" between two changes of measure $(\mathcal{M}_1)$ and $(\mathcal{M}_2)$ (as indicated schematically in Figure 4): Along the boundary, $x_2 = 0$, the change of measure is $(\mathcal{M}_1)$. In the interior, $x_2 \geq b_2$, the change of measure is $(\mathcal{M}_2)$. In the interim, $1 < x_2 < b_2$, the new rates are simply linear interpolation of their values at $x_2 = 0$ and $x_2 = b_2$.

Let us follow a sample path starting from an arrival to an empty network. The proposed change of measure (with $b_1 = 1$) implies the following: Initially, and while $x_2 = 0$, exchange the arrival rate (λ) with the service rate at node 1 (μ_1), i.e., start with overloading the first node (making it unstable) while the second node is stable. As the number of customers at the second node increases in the range $(1 < x_2 < b_2)$, gradually and simultaneously reduce the load on the first node while increasing the load on the second node. When the number of customers at the second node reaches level b_2 and while $x_2 \geq b_2$, exchange the arrival rate (λ) with the service rate at node 2 (μ_2); i.e., overload the second node (making it unstable) while the first node is stable (resp. "critical") if $\mu_2 < \mu_1$ (resp. $\mu_2 = \mu_1$). In the interior $(x_2 \geq b_2)$, the new rates do not depend on the network state (i.e., neither x_1 nor x_2).

Time Reversal Argument

The effectiveness of the change of measure in Proposition 1 may be explained using time-reversal argument [21].

However, by no means this should be interpreted as a formal validation of its asymptotic efficiency. The reverse time process is also a 2-node tandem network (see Figure 2); however, arrivals (rate λ) enter the network at Node 2 (service rate μ_2) and exit from Node 1 (service rate μ_1).

Roughly speaking, according to [3], in the limit as $L \to \infty$, the most likely path to the rare set (i.e., population overflow) in the forward time process is the same path by which the reverse time process evolves, given that the latter starts from the rare set. Since both Node 1 and Node 2 may be non-empty upon entry into the rare set, the hitting state (x_1, x_2) is somewhere along the line $x_1 + x_2 = L$. Let $\mu_2 \leq \mu_1$, and the reverse time process starts at (L_1, L_2) such that $L_1 + L_2 = L$. Node 2 has arrival rate λ and initially its departure rate is μ_2, thus it empties at rate $(\mu_2 - \lambda)$. In the meantime, Node 1 has input rate μ_2 and a departure rate μ_1, thus it also empties at rate $(\mu_1 - \mu_2)$. If $\mu_1 = \mu_2$, then node 1 is "critical" and does not empty; this corresponds to Path III in Figure 3. If and when Node 2 empties first, its arrival and departure rates are equal to λ. At that time, Node 1 has arrival rate λ and departure rate μ_1, thus it empties at rate $(\mu_1 - \lambda)$. This corresponds to Path II in Figure 3. If and when Node 1 empties first, its arrival and departure rates are equal to μ_2. At that time, Node 2 has arrival rate λ and departure rate μ_2, thus it empties at rate $(\mu_2 - \lambda)$. This corresponds to Path I in Figure 3.

Note that departures (resp. arrivals) in reverse time correspond to arrivals (resp. departures) in forward time. It follows that along the most likely path from an empty network to population overflow (in forward-time), there are two possible scenarios depending on the entry state (L_1, L_2) into the rare set, which in turn depends on the arrival and service rates [19]: One scenario corresponds to Path I, in which Node 2 builds up first while Node 1 is stable (i.e., $\tilde{\lambda} = \mu_2, \tilde{\mu_1} = \mu_1, \tilde{\mu_2} = \lambda$). At some point, also Node 1 starts to build up until the rare set is hit (i.e., $\tilde{\lambda} = \mu_1, \tilde{\mu_1} = \mu_2, \tilde{\mu_2} = \lambda$). This scenario is more likely when $\mu_2 \ll \mu_1$. A second scenario corresponds to Path II (or Path III), in which Node 1 builds up first while Node 2 is stable (i.e., $\tilde{\lambda} = \mu_1, \tilde{\mu_1} = \lambda, \tilde{\mu_2} = \mu_2$). At some point, also Node 2 starts to build up until the rare set is hit (i.e., $\tilde{\lambda} = \mu_1, \tilde{\mu_1} = \mu_2, \tilde{\mu_2} = \lambda$). This scenario is more likely when μ_2 is less than, but sufficiently close to, μ_1. Paths III is simply the limit of Path II when $\mu_1 = \mu_2$.

Now, if $\mu_2 \leqslant \mu_1$ (as we assume above), then the heuristic in [26] exchanges λ and μ_2 leaving μ_1 unchanged; i.e., Node 1 is stable, and Node 2 builds up all the way until the rare set is hit. This corresponds to the Path PW in Figure 3. It is interesting note that for $\mu_2 \ll \mu_1$ Path I is the most likely and it gets closer to Path PW, which explains the effectiveness of the heuristic in [26] for sufficiently small μ_2. For larger μ_2 (closer to μ_1) the most likely path deviates further form Path PW and gets closer to Path II, which clarifies the ineffectiveness of the heuristic in [26].

On the other hand, by appropriately setting b_1 and b_2, the state-dependent heuristic in Proposition 1 can (roughly) capture the most likely path to overflow (i.e., Path I, Path II or Path III, depending on the network parameters). Indeed, this clarifies the robustness and effectiveness of this heuristic over the entire feasible parameter range (as evidenced from experimental results in Section IV-A).

The heuristic in Proposition 1 can be generalized to n nodes in tandem as follows.

Proposition 2 (SDH for the n-node Tandem Network) Let $\mathbf{\Theta}$ be a vector with the original network parameters, i.e., $\mathbf{\Theta}^T = [\lambda, \mu_1, \ldots, \mu_n]$. Similarly, $\tilde{\mathbf{\Theta}}(\mathbf{x})$ is a vector with the new parameters (depending on the network state $\mathbf{x}$) for simulating the network under importance sampling. The SDH for an n-node tandem network is given by

$$\tilde{\mathbf{\Theta}} = \mathbf{SDH}_n^{\mathrm{Tn}} \ \mathbf{\Theta},$$

with the transformation matrix $\mathbf{SDH}_n^{\mathrm{Tn}}$ expressed recursively as follows:

$$\mathbf{SDH}_k^{\mathrm{Tn}} = \left[\frac{x_k}{b_k}\right]^1 \mathbf{I}_k^{\mathrm{Tn}} + \left[\frac{b_k - x_k}{b_k}\right]^+ \mathbf{SDH}_{k-1}^{\mathrm{Tn}}, k = 1, \ldots, n, \tag{6}$$

where $\mathbf{SDH}_0^{\mathrm{Tn}} = \mathbf{I}_0^{\mathrm{Tn}}$ and $\mathbf{I}_k^{\mathrm{Tn}}$ $(k = 1, \ldots, n)$ is the identity matrix of dimension $(n+1)$ with the first and the $(k+1)$-st rows interchanged.

Note that, except for $n = 1$ (single server), by setting $b_i = \infty$ for $i = 1, \ldots, n-1$ and $b_n = 1$, the above heuristic does not reduce to the well known heuristic of interchanging the arrival rate (λ) and the slowest service rate (μ_n) [26]. The latter exchanges λ and μ_n upon arrival to an empty network and does not depend on the network state; i.e., importance sampling is applied upon the first arrival to node 1 and continues until the network empties or the rare event is reached. The heuristic in Proposition 2 exchanges λ and μ_n upon the first arrival to node n and only as long as it is non-empty, i.e., importance sampling is applied only during the busy periods of node n and continues until it empties or the rare event is reached.

Remark 1 Note that $b_i \geq 1$ is the number of boundary levels along x_i for which the change of measure depends on x_i (we also refer to it as the "dependence range"). These are the only variable parameters in the above heuristic, and their proper selection is crucial for achieving asymptotic efficiency, particularly for larger networks. In general, the "best" values of b_i, $i = 1, \ldots, n$ (yielding estimates with the lowest variance) may depend on the set of network parameters as well as the overflow level L. However, empirical results suggest robustness. For example, for a given

parameter point, by setting $b_i = b$ at all nodes $i = 1, \ldots, n$ (as we do in most experiments with tandem networks in Section IV-A), the change of measure remains effective for a range of b values around the optimal. However, for some regions in the parameter space, the effectiveness may be more sensitive to b, whose "best" value may vary from one parameter point to another and may also depend on L.

B. SDH for the n-node Parallel Network

Let λ_i and μ_i be, respectively, the arrival rate and the service rate at node i, and denote its traffic intensity by $\rho_i = \frac{\lambda_i}{\mu_i} < 1$ $(i = 1, \ldots, n)$. Without loss of generality we assume that $\sum_{i=1}^{n} (\lambda_i + \mu_i) = 1$.

The new rates may depend on the network state and, therefore, are they are functions of the vector $\mathbf{x} = (x_1, x_2, ..., x_n)$. Denote by $\tilde{\lambda}_i(\mathbf{x})$ and $\tilde{\mu}_i(\mathbf{x})$ the corresponding rates at node i under the new change of measure, and by $\mathbf{SDH}_i(\mathbf{x})$ the 2×2 linear operator (matrix) transforming the original rates into the new rates at node i, i.e.,

$$\begin{bmatrix} \tilde{\lambda}_i \\ \tilde{\mu}_i \end{bmatrix} = \mathbf{SDH}_i \begin{bmatrix} \lambda_i \\ \mu_i \end{bmatrix}, \; i = 1, \ldots, n. \tag{7}$$

As before, define $[a]^+ = \max(a, 0)$ and $[a]^1 = \min(a, 1)$. The following proposition gives the state-dependent heuristic change of measure for n parallel nodes.

Proposition 3 (SDH for the n-node Parallel Network)
Let $\mathbf{\Theta}$ be a vector with the original network parameters, i.e., $\mathbf{\Theta}^T = [\lambda_1, \mu_1, \ldots, \lambda_n, \mu_n]$. Similarly, $\tilde{\mathbf{\Theta}}(\mathbf{x})$ is a vector with the new network parameters, and define the $2n \times 2n$ transformation matrix $\mathbf{SDH}^{\text{Pn}}(\mathbf{x})$ as follows (occasionally, we abuse notation by dropping the vector $\mathbf{x}$).

$$\mathbf{SDH}^{\text{Pn}} = \begin{bmatrix} \mathbf{SDH}_1 & 0 & \ldots & 0 \\ 0 & \mathbf{SDH}_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \mathbf{SDH}_n \end{bmatrix}, \tag{8}$$

with

$$\begin{aligned} \mathbf{SDH}_i = & \begin{bmatrix} \frac{x_i}{b_i} \end{bmatrix}^1 \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \\ & + \begin{bmatrix} \frac{b_i - x_i}{b_i} \end{bmatrix}^+ \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \end{aligned} \tag{9}$$

for some integer $b_i \geq 1$, and $i = 1, \ldots, n$. Then the SDH for an n-node parallel network is given by

$$\tilde{\mathbf{\Theta}} = \mathbf{SDH}^{\text{Pn}} \; \mathbf{\Theta} \, .$$

The superscript Pn refers to a network of n parallel nodes. In Equation 9 for $\mathbf{SDH}_i$, the first matrix is the identity matrix with the first and the second rows interchanged, which corresponds to interchanging the arrival and service rates at node i. The second matrix is the identity matrix, corresponding to no change of measure. Note that the equality $\sum_{i=1}^{n} (\tilde{\lambda}_i + \tilde{\mu}_i) = 1$ holds under the above change of measure.

Remark 2 Note that b_i is the number of boundary levels for which the change of measure at node i depends on the content x_i (we also refer to it as the dependence range at node i). Proper selection of the b_i's is crucial for achieving asymptotic efficiency. Empirical results suggest that the "optimal" b_i's (yielding estimates with the lowest variance) depend on the traffic intensities ρ_i's at all network nodes as well as the overflow level L.

According to the above change of measure, all nodes may be "pushed" (overloaded) simultaneously, however, to different extents depending on their respective ratios of content x_i relative to b_i. This is a state-dependent change of measure, by which busy nodes $(x_i \geq 1)$ are "pushed" harder for higher x_i/b_i.

The well-known heuristic in [26] suggests interchanging the arrival and service rates at the bottleneck node (with the highest ρ_i). This is a state-independent change of measure, which is shown to work well only in a limited region of the network parameters space (namely, when the utilization at the bottleneck node is sufficiently higher than those at all other nodes). For a single node, say, node i, our change of measure, with $b_i = 1$, is identical to that in [26]; both are asymptotically efficient.

Time Reversal Argument

As for the tandem network, the effectiveness of the change of measure in Proposition 3 for the simulation of parallel networks may be explained using time-reversal argument [21]. The reverse time process is also an n-node parallel network. At node i $(i = 1, \ldots, n)$, the arrival and service rates are λ_i and μ_i, respectively (i.e., same as in the forward time process). However, the reverse time process starts from the hitting state into the rare set, say, $(L_1, L_2, \ldots, L_n)$ with $\sum_{i=1}^{n} L_i = L$. In the reverse time, the number of customers at node i $(i = 1, \ldots, n)$ is initially L_i and it empties at rate $(\delta_i = \mu_i - \lambda_i)$. The (reverse) time needed to clear the backlog at node i is therefore given by $\frac{L_i}{\delta_i}$. Clearly, the order in which the backlogs at different nodes disappear depends on the initial (hitting) state as well as the arrival and service rates at each node. Intuitively, the bottleneck node (with the highest ρ_i) is likely to have the largest backlog upon hitting the rare set, and because it empties at a slower rate, its backlog is likely to be the last to disappear. (In forward time, this implies that the bottleneck node is likely to start its build up sooner than other nodes.) Note that it may take some time for the network to empty after all backlogs disappear; this also depends on the traffic intensities and the overflow level L.

Note that departures (resp. arrivals) in reverse time correspond to arrivals (resp. departures) in forward time. It follows that along the most likely path from an empty network to population overflow, each node starts building up a backlog after some (own) initial period. The build up at node i continues at rate $\delta_i = \mu_i - \lambda_i$ until the population overflow level L is reached. Highly loaded nodes are likely to start their backlog build up sooner than lightly loaded nodes. If the traffic intensity at the bottleneck node is sufficiently higher than at other nodes, then the most likely path to overflow involves a build up only at the bottleneck node. This is consistent with the heuristic in [26] which exchanges the arrival and service rates only at the bottleneck node, and therefore clarifies its effectiveness in this case.

By appropriately setting b_i, for $i = 1, \ldots, n$, the state-dependent heuristic in Proposition 3 can (roughly) capture the most likely path to overflow in a network of n parallel nodes. The above time reversal argument along with some experimentation may provide helpful insights into how to properly set the b_is at the different nodes. Empirical results in Section IV-B show that the heuristic is very effective and robust over the entire feasible parameter range .

IV. EXPERIMENTAL RESULTS

Importance sampling to estimate the probability of population overflow $(\gamma(L))$ involves generating, say, N, independent and identically distributed (i.i.d.) busy cycles (i.e., starting with an empty network). Starting a cycle at time 0, define τ_L as the instant when the network population reaches level L for the first time. Similarly, define τ_0 as the instant when the network population returns to 0 for the first time. The indicator function $I_i(\tau_L < \tau_0)$ takes the value 1 if the population overflow (level L) is reached in cycle i, otherwise it takes the value 0.

In each cycle, the change of measure is applied until either the population overflow event is reached or the network population returns to 0. Let W_i be the likelihood ratio associated with cycle i (as defined in Section II-B), then an unbiased estimator $\tilde{\gamma}$ of $\gamma(L)$ is given by

$$\tilde{\gamma} \;=\; \frac{1}{N} \sum_{i=1}^{i=N} I_i \, W_i \; . \tag{10}$$

The second moment of $I\,W$ is estimated by

$$\tilde{\gamma^2} \;=\; \frac{1}{N} \sum_{i=1}^{i=N} I_i \, {W_i}^2 \; . \tag{11}$$

The variance and the relative error of the importance sampling estimator $\tilde{\gamma}$ are given by $\mathrm{VAR}(\tilde{\gamma}) = (\tilde{\gamma^2} - (\tilde{\gamma})^2)/(N-1)$ and $\mathrm{RE}(\tilde{\gamma}) = \sqrt{\mathrm{VAR}(\tilde{\gamma})}\,/\,\tilde{\gamma}$, respectively. Another useful measure for comparing the efficiency of different estimators is the "relative time variance" (RTV) product, which is defined as the simulation time (in seconds) multiplied by the squared relative error of the estimator. As the estimate becomes more stable, its RTV tends to a constant value, which is smaller for a more efficient estimator. For example, if RTV_2 (for Estimator 2) is larger than RTV_1 (for Estimator 1), then it will take Estimator 2 a longer simulation time to reach the same accuracy. For efficiency comparisons we use the variance reduction ratio, $\mathrm{VRR} = \mathrm{RTV}_2\,/\,\mathrm{RTV}_1$, which represents the efficiency gain when using Estimator 1 relative to that when using Estimator 2.

In the following sections, two sets of experiments are presented; one for tandem networks and the second for parallel networks. Each set consists of six experiments; three for a small network with 2 nodes and the other three for a larger network with 4 nodes. In order to illustrate the utility of our approach, all parameter points are chosen in regions where the well-known heuristic in [26] is shown (formally or empirically) to be ineffective. In all simulation experiments, the same number of replications, namely, 10^6, is used to obtain estimates of the population overflow probability $\gamma(L)$. For each estimate in these tables, we include the relative error RE$\%$ (in percentage). For the purpose of comparing the heuristics in this paper (termed SDH) and the adaptive importance sampling methodology (termed SDA) in de [6], we also include the ratio VRR (relative to SDA). Hence, VRR > 1 implies efficiency gain of SDH over SDA. Estimates obtained using the well known heuristic in [26] (termed PW) are also presented, although these are not necessarily accurate or stable. In general, numerical results are difficult to obtain for larger and/or higher overflow levels (i.e., for larger sate-space). Whenever feasible, numerical results (for example, using the algorithm outlined in [4]) are included to verify the correctness of the simulation estimates. Otherwise, the corresponding table entry is marked with a "$*$". In the absence of numerical results, agreement of different estimators (e.g., using SDH and SDA) may be an indication of correctness.

A. Simulation of Tandem Networks

The experiments in this section are designed to demonstrate that the state-dependent change of measure proposed in Section III-A always yield asymptotically efficient estimates (mostly with bounded relative error), also in those regions where no state-independent change of measure is known to be asymptotically efficient. In the experiments with 2-node tandem networks, we set $b_1 = 1$ and $b_2 = b$, with b yielding stable estimates having the lowest relative error. In experiments with 4-node tandem networks, we set $b_i = b$ at all nodes $i = 1, \ldots, n$. Again, b is set to yield stable estimates with the lowest (or close to the lowest) relative error. Similar to SDH, adaptive methodologies (such as SDA) assume state-dependence only over a (small) number of boundary layers (say, b) which must be properly determined to ensure the effectiveness and efficiency of these methods. Too small b may not capture

crucial dependencies close to the boundaries. Too large b may render SDH ineffective, but it will only reduce the efficiency of SDA. In either SDH or SDA, the "optimal" b which maximizes the efficiency (minimizes the RTV) may be determined by repeating the simulation for successively increasing b. Experimental results with SDH and SDA are obtained using their respective "optimal" b.

For the 2-node tandem network, it is proven or shown empirically (see [16] and [7]) that the state-independent heuristic (PW) in [26] yields estimates with bounded relative error only in some (non-contiguous) regions of the feasible parameter space. (The feasible parameter space is that corresponding to stable networks.) Thus, as depicted in Figure 1 for the 2-node tandem network [7], the feasible parameter space may be divided into two regions, depending on the asymptotic properties of the PW estimator:
BRE Region - PW is asymptotically efficient (with bounded relative error); corresponds to Region I (BRE) in Figure 1.
NAE Region - PW is not asymptotically efficient (with exponentially growing or infinite relative error); corresponds to Regions II (ERE) and III (IRE) in Figure 1.

Empirical studies seem to confirm that the above division of the feasible parameter space holds also for tandem networks with any number of nodes (i.e., for any feasible set of network parameters, PW is either BRE or NAE). For the n-node tandem network, sufficient conditions for the asymptotic (and non-asymptotic) efficiency of the PW heuristic are given in [16]. These conditions are strong and do not cover the entire parameter space, i.e., not all feasible parameter points may be determined as BRE or NAE.

We experiment with tandem networks having 2 and 4 nodes, respectively, with the parameters chosen in the NAE Region (i.e., where the heuristic in [26] is not effective). For each network three experiments with different parameter sets are executed; two with symmetric loads (low and high) and the third with asymmetric loads. Typically, it is most difficult to efficiently estimate the probability of overflow when some service rates are equal (or almost equal).

Table I displays numerical and simulation results for the symmetric 2-node tandem network (with low loads): $\lambda = 0.04$ and $\mu_1 = \mu_2 = 0.48$ (i.e., $\rho_1 = \rho_2 = 0.083$). Table II displays numerical and simulation results for the symmetric 2-node tandem network (with high loads): $\lambda = 0.2$ and $\mu_1 = \mu_2 = 0.4$ (i.e., $\rho_1 = \rho_2 = 0.5$). Table III displays numerical and simulation results for the asymmetric 2-node tandem network: $\lambda = 0.18$, $\mu_1 = 0.42$ and $\mu_2 = 0.4$ (i.e., $\rho_1 = 0.43$ and $\rho_2 = 0.45$). Experimental results in Tables I, II and III show that unlike PW, SDH (as described in Section III-A) yields correct (compare with numerical results) and asymptotically efficient estimates with a (seemingly) bounded relative error.

Table IV displays numerical (if feasible; otherwise the table entry is marked by $*$) and simulation results for the symmetric 4-node tandem network (with low loads): $\lambda = 0.04$ and $\mu_i = 0.24$ (i.e., $\rho_i = 0.167$) for $i = 1, 2, 3, 4$. Table V displays simulation results for the symmetric 4-node tandem network (with high loads): $\lambda = 0.1$ and $\mu_i = 0.225$ (i.e., $\rho_i = 0.044$) for $i = 1, 2, 3, 4$. Table VI displays simulation results for the asymmetric 4-node tandem network: $\lambda = 0.1$, $\mu_1 = 0.28$, $\mu_2 = 0.24, \mu_3 = 0.21$ and $\mu_4 = 0.17$ (i.e., $\rho_1 = 0.357$, $\rho_2 = 0.417, \rho_3 = 0.476$ and $\rho_4 = 0.588$). Experimental results in Tables IV, V and VI show that unlike PW, SDH (as described in Section III-A) yields correct (compare with SDA results) and asymptotically efficient estimates with a (seemingly) bounded relative error.

To converge properly, our basic (non-optimized) implementation of SDA may require many iterations, each with a large number of cycles (i.e., long simulation time). On the other hand, if and when it converges, it gives very small relative error. (For more on SDA and its implementation details see [6].) For the examples presented here, SDH typically requires only a few minutes to achieve relative errors less than 1%, and could be more efficient than SDA (VRR > 1) even though its displayed relative error may be higher. In fact, experiments not presented here show that a finer "tuning" of the b_is in SDH (e.g., by allowing different, rather than equal, b_is at different network nodes) may yield further reduction of the relative error.

B. Simulation of Parallel Networks

In this section we experiment with 2- and 4-node (symmetric and asymmetric) parallel networks. Network parameters are chosen in regions where the heuristic in [26] is not effective. This is typically the case in symmetric parallel networks (i.e., all nodes have the same utilization) or when the higher utilizations are sufficiently close.

Table VII displays numerical and simulation results for the symmetric 2-node parallel network (with low loads): $\lambda_1 = \lambda_2 = 0.1$ and $\mu_1 = \mu_2 = 0.4$ (i.e., $\rho_1 = \rho_2 = 0.25$). Table VIII displays numerical and simulation results for the symmetric 2-node parallel network (with high loads): $\lambda_1 = \lambda_2 = 0.15$ and $\mu_1 = \mu_2 = 0.35$ (i.e., $\rho_1 = \rho_2 = 0.43$). Table IX displays numerical and simulation results for the asymmetric 2-node parallel network: $\lambda_1 = 0.12, \lambda_2 = 0.08$ and $\mu_1 = \mu_2 = 0.4$ (i.e., $\rho_1 = 0.3, \rho_2 = 0.2$). Experimental results in Tables VII, VIII and IX show that unlike PW, SDH (as described in Section III-B) yields correct (compare with numerical results), stable and asymptotically efficient estimates with relative error increasing (sub-)linearly in the overflow level L. Note that the "best" b_1 and b_2 are equal only in the symmetric networks. In the asymmetric network, $b_1 = 2$ and $b_2 (> 2)$ increases with the overflow level L. Also SDA produces correct and stable results; however, it is mostly less efficient than SDH (as indicated by VRR > 1).

Table X displays simulation results (numerical results were not feasible for the 4-node parallel network; the

corresponding table entries are marked by $*$) for the symmetric 4-node parallel network (with low loads): $\lambda_i = 0.05$ and $\mu_i = 0.2$ for $i = 1, 2, 3, 4$ (i.e., $\rho_i = 0.25$, for $i = 1, 2, 3, 4$). Table XI displays simulation results for the symmetric 4-node parallel network (with high loads): $\lambda_i = 0.08$ and $\mu_i = 0.17$, for $i = 1, 2, 3, 4$ (i.e., $\rho_i = 0.47$, for $i = 1, 2, 3, 4$). Table XII displays simulation results for the asymmetric 4-node parallel network: $\lambda_1 = 0.06, \lambda_2 = \lambda_3 = 0.04, \lambda_4 = 0.02$ and $\mu_i = 0.2$, for $i = 1, 2, 3, 4$ (i.e., $\rho_1 = 0.3, \rho_2 = \rho_3 = 0.2, \rho_4 = 0.1$). Experimental results in Tables X, XI and XII show that unlike PW, SDH (as described in Section III-B) yields correct (numerical results are not feasible, but agreement with SDA estimates suggest correctness), stable and asymptotically efficient estimates with relative error increasing (sub-)linearly in the overflow level L. Note that the "best" b_is are equal when the loads are symmetric. For asymmetric loads, $b_1 = 2$ and $b_i > 2$, for $i = 2, 3, 4$, and increases with the overflow level L. Finer "tuning" of the b_is (by allowing them to be different) has shown to yield further reduction in the relative error.

In the experiments presented here, SDH typically requires only a few minutes to achieve relative errors less than 1% and is evidently much more efficient than SDA (VRR $\gg 1$) even though its displayed relative error may be higher. Also SDA produces correct and stable results; however, it is not clear why it performs much worse (relative to SDH) for 4-node parallel networks than it does for 4-node tandem networks (compare with VRRs in Tables IV, V and VI).

V. CONCLUSIONS AND FURTHER WORK

In this paper we have proposed and experimented with a heuristic approach to approximate the "optimal" state-dependent change of measure for the efficient simulation of networks with nodes in series or in parallel. The developed changes of measure (which we refer to as SDH) are used to estimate (using importance sampling) the probability of population overflow in tandem and parallel queueing networks. Experimental results indicate that the heuristics yield asymptotically efficient estimates, with relative error growing at most (sub-)linearly with the overflow level L. The efficiency of the obtained changes of measure compares well with those determined using adaptive importance sampling methodologies. Yet, our approach does not require costly pre-computation and avoids complicated (or intractable) mathematical analyses. Moreover, its effectiveness is not diminished for large networks with huge state-space.

Needless to say, the utility of the approach needs to be tested on larger networks and more complex topologies, including feed-forward and feedback networks. Also, simple and robust guidelines for selecting the number of boundary layers (dependence range) is an important challenge.

REFERENCES

[1] Ahamed, T.P.I., V.S. Borkar, and S.K. Juneja (2004). Adaptive importance sampling technique for Markov chains using stochastic approximation. *Operations Research*. Accepted.

[2] Asmussen, S., and R.Y. Rubinstein (1995). Steady state rare events simulation in queueing models and its complexity properties. In *Advances in Queueing: Theory, Methods and Open problems*, ed. J.H. Dshalalow, 429–461. CRC Press, New York.

[3] Anantharam, V., P. Heidelberger, and P. Tsoucas (1990). Analysis of rare events in continuous time Markov chains via time reversal and fluid approximation. IBM Research Report RC 16280, Yorktown Heights, New York.

[4] de Boer, P.T. (2000). Analysis and efficient simulation of queueing models of telecommunication systems. PhD Thesis, University of Twente.

[5] de Boer, P.T., V.F. Nicola, and R.Y. Rubinstein (2000). Dynamic importance sampling simulation of queueing networks: An adaptive approach based on cross-entropy. In *Proceedings of the 2000 Winter Simulation Conference*, IEEE Computer Society Press, 646–655.

[6] de Boer, P.T., and V.F. Nicola (2002). Adaptive state-dependent importance sampling simulation of Markovian queueing networks. *European Transactions on Telecommunications* 13 (4): 303–315.

[7] de Boer, P.T. (2004). Analysis of sate-independent IS measures for the two-node tandem queue. *International Workshop on Rare Event Simulation (RESIM'04)*, Budapest, Hungary.

[8] Chang, C.S., P. Heidelberger, S. Juneja, and P. Shahabuddin (1994). Effective bandwidth and fast simulation of ATM in-tree networks. *Performance Evaluation* **20** 45–65.

[9] Cottrell, M., J.-C. Fort, and G. Malgouyres (1983). Large deviations and rare events in the study of stochastic algorithms. *IEEE Trans. on Automatic Control* **28** (9) 907–920.

[10] De Veciana, G., C. Courcoubetis, and J. Walrand (1994). Decoupling bandwidths for networks: A decomposition approach to resource management for networks. In *Proceedings of INFOCOM'94*, IEEE Press, 466–473.

[11] Frater, M.R., and B.D.O. Anderson (1989). Fast estimation of the statistics of excessive backlogs in tandem networks of queues. *Australian Telecommun. Res.* **23** (1) 49–55.

[12] Frater, M.R., T.M. Lenon, and B.D.O. Anderson (1991). Optimally efficient estimation of the statistics of rare events in queueing networks. *IEEE Trans. Autom. Control* 36: 1395–1405.

[13] Frater, M.R., and B.D.O. Anderson (1994). Fast simulation of buffer overflows in tandem networks of GI/GI/1 queues. *Ann. Oper. Res.* **49** 207–220.

[14] Garvels, M.J.J. (2000). The splitting method in rare event simulation. PhD Thesis, University of Twente.

[15] Glasserman, P., P. Heidelberger, P. Shahabuddin, and T. Zajic (1996). Multilevel splitting for estimating rare event probabilities. *Operations Research* **47** (4) 585–600.

[16] Glasserman, P., and S-G. Kou (1995). Analysis of an importance sampling estimator for tandem queues. *ACM Transactions of Modeling and Computer Simulation* 5 (1): 22–42.

[17] Glasserman, P., and Y. Wang (1997). Counterexamples in importance sampling for large deviations probabilities. *Ann. Appl. Probab.* **7** (3) 731–746.

[18] Glynn, P., and D.L. Iglehart (1989). Importance sampling for stochastic simulations. *Management Science* **35** 1367–1392.

[19] Heidelberger, P. (1995). Fast simulation of rare events in queueing and reliability models. *ACM Transactions of Modeling and Computer Simulation* 5 (1): 43–85.

[20] Juneja, S.K., and V.F. Nicola (2005). Efficient simulation of buffer overflow probabilities in Jackson networks with feedback. *ACM Transactions of Modeling and Computer Simulation.* To appear.

[21] Kelly, F.P. (1979). *Reversibility and Stochastic Networks.* Wiley, New York.

[22] Kroese, D.P., and V.F. Nicola (2002). Efficient simulation of a tandem Jackson network. *ACM Transactions of Modeling and Computer Simulation* 12 (2): 119–141.

[23] L'Ecuyer, P., and Y. Champoux (2001). Estimating small cell loss ratios in ATM switches via importance sampling. *ACM Transactions of Modeling and Computer Simulation* 11 (1): 76–105.
[24] Nicola, V.F., and T.S. Zaburnenko (2005). Importance sampling simulation of population overflow in two-node tandem networks. In *Proceedings of the 2nd International Conference on the Quantitative Evaluation of Systems (QEST'05)*, Torino, Italy.
[25] Nicola, V.F., and T.S. Zaburnenko (2005). Efficient importance sampling heuristics for the simulation of population overflow in Jackson networks. In *Proceedings of the 2005 Winter Simulation Conference*, IEEE Computer Society Press, 538–546.
[26] Parekh, S., and J. Walrand (1989). A quick simulation method for excessive backlogs in networks of queues. *IEEE Transactions on Automatic Control* 34: 54–66.
[27] Randhawa, R.S., and S.K. Juneja (2004). Combining importance sampling and temporal difference control variates to simulate Markov chains. *ACM Transactions of Modeling and Computer Simulation* 14 (1): 1–30.
[28] Rubinstein, R.Y. (2002). The cross-entropy method and rare events for maximal cut and bipartition problems. *ACM Transactions of Modeling and Computer Simulation* 12 (1): 27–53.
[29] Sadowsky, J.S. (1991). Large deviations theory and efficient simulation of excessive backlogs in a $GI/GI/m$ queue. *IEEE Trans. on Automatic Control* **36** 1383–1394.
[30] Tsoucas, P. (1992). Rare events in series of queues. *J. Appl. Probab.* **29** 168–175.
[31] Villen-Altamirano, M., and J. Villen-Altamirano (2002). Analysis of RESTART simulation: theoretical basis and Sensitivity study. *European Transactions on Telecommunications* **13** (4) 373–386.
[32] Weber, R.R. (1979). The interchangeability of $M/M/1$ queues in series. *Journal of Applied Probability* 16: 690–695.
[33] Zaburnenko, T.S., and V.F. Nicola (2005). Efficient heuristics for simulating population overflow in tandem networks. In *Proceedings of the 5th St. Petersburg Workshop on Simulation (SPWS'05)*, ed. S.M. Ermakov, V.B. Melas, and A.N. Pepelyshev, 755–764. St. Petersburg University Publishers.

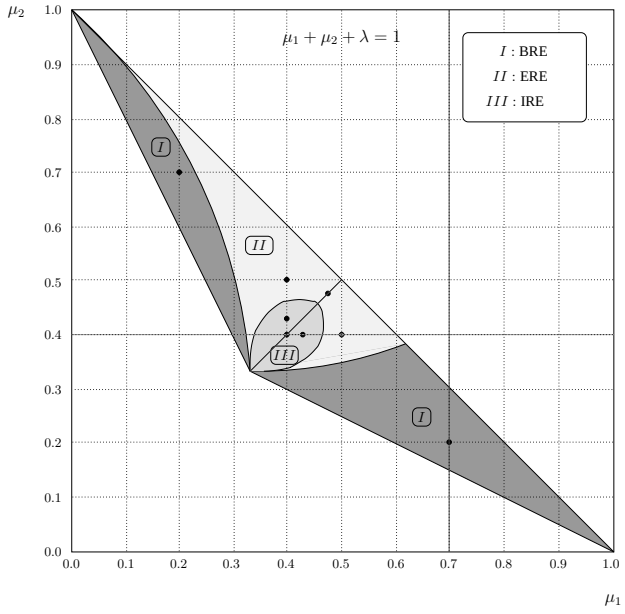

Fig. 1. Asymptotic efficiency of PW in the feasible parameter space (as empirically determined in [7])

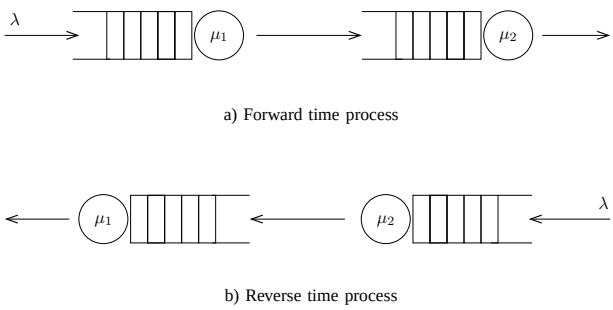

Fig. 2. Time reversal of the 2-node tandem network

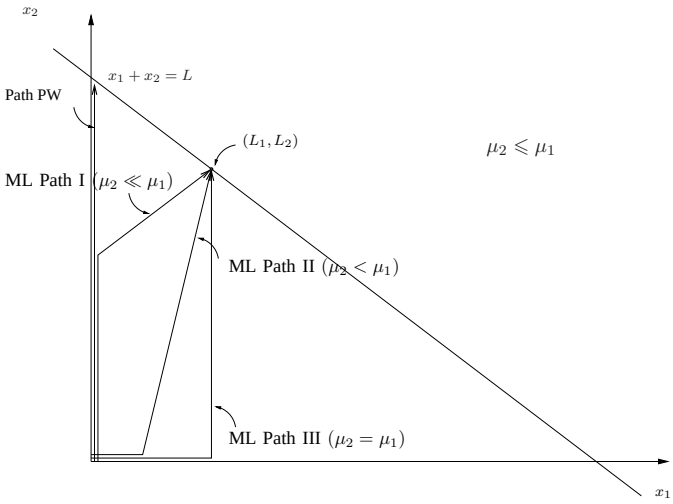

Fig. 3. Most likely path to population overflow in the 2-node tandem network with $\mu_2 \leqslant \mu_1$

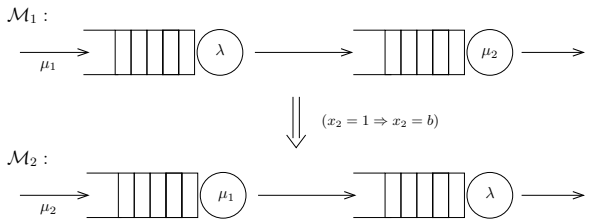

Fig. 4. SDH change of measure for the 2-node tandem network

TABLE I

2-NODE TANDEM NETWORK - SYMMETRIC ($\lambda = 0.04, \mu_1 = \mu_2 = 0.48$) ($\rho_1 = \rho_2 = 0.083$)

L	**Numerical**	**PW**	**SDA**		**SDH**		
	$\gamma(L)$	$\tilde{\gamma}(L) \pm$ RE%	b	$\tilde{\gamma}(L) \pm$ RE%	b	$\tilde{\gamma}(L) \pm$ RE%	VRR
25	2.8722e-025	2.6050e-025 $\pm$ 8.64	3	2.8729e-025 $\pm$ 0.06	3	2.8767e-025 $\pm$ 0.13	4.11
50	6.0327e-052	2.3672e-052 $\pm$ 4.67	3	6.0340e-052 $\pm$ 0.07	3	6.0367e-052 $\pm$ 0.12	3.89
100	1.3270e-105	3.5984e-106 $\pm$ 19.5	3	1.3255e-105 $\pm$ 0.07	3	1.3301e-105 $\pm$ 0.17	1.63

TABLE II

2-NODE TANDEM NETWORK - SYMMETRIC ($\lambda = 0.2, \mu_1 = \mu_2 = 0.4$) ($\rho_1 = \rho_2 = 0.5$)

L	**Numerical**	**PW**	**SDA**		**SDH**		
	$\gamma(L)$	$\tilde{\gamma}(L) \pm$ RE%	b	$\tilde{\gamma}(L) \pm$ RE%	b	$\tilde{\gamma}(L) \pm$ RE%	VRR
25	7.1526e-07	6.8876e-07 $\pm$ 3.71	7	7.1532e-07 $\pm$ 0.04	9	7.1637e-07 $\pm$ 0.14	1.72
50	4.3521e-14	2.7323e-14 $\pm$ 5.99	7	4.3509e-14 $\pm$ 0.06	9	4.3556e-14 $\pm$ 0.11	5.37
100	7.8097e-29	2.5780e-29 $\pm$ 11.8	7	7.8150e-29 $\pm$ 0.10	9	7.7953e-29 $\pm$ 0.13	3.48

TABLE III

2-NODE TANDEM NETWORK - ASYMMETRIC ($\lambda = 0.18, \mu_1 = 0.42, \mu_2 = 0.4$) ($\rho_1 = 0.43, \rho_2 = 0.45$)

L	**Numerical**	**PW**	**SDA**		**SDH**		
	$\gamma(L)$	$\tilde{\gamma}(L) \pm$ RE%	b	$\tilde{\gamma}(L) \pm$ RE%	b	$\tilde{\gamma}(L) \pm$ RE%	VRR
25	3.8066e-08	4.2325e-08 $\pm$ 8.53	5	3.8046e-08 $\pm$ 0.04	6	3.8089e-08 $\pm$ 0.10	1.91
50	1.0684e-16	1.1551e-16 $\pm$ 13.8	7	1.0681e-16 $\pm$ 0.02	6	1.0687e-16 $\pm$ 0.08	1.79
100	5.3355e-34	3.8945e-34 $\pm$ 3.00	6	5.3357e-34 $\pm$ 0.03	6	5.3390e-34 $\pm$ 0.07	2.51

TABLE IV

4-NODE TANDEM NETWORK - SYMMETRIC ($\lambda = 0.04, \mu_1 = \mu_2 = \mu_3 = \mu_4 = 0.24$) ($\rho_1 = \rho_2 = \rho_3 = \rho_4 = 0.167$)

L	**Numerical**	**PW**	**SDA**		**SDH**		
	$\gamma(L)$	$\tilde{\gamma}(L) \pm$ RE%	b	$\tilde{\gamma}(L) \pm$ RE%	b	$\tilde{\gamma}(L) \pm$ RE%	VRR
25	5.0207e-016	4.1762e-016 $\pm$ 11.6	3	5.0277e-016 $\pm$ 0.14	4	5.1093e-016 $\pm$ 0.94	4.07
50	*	3.1024e-035 $\pm$ 37.0	3	1.3532e-034 $\pm$ 0.26	4	1.3560e-034 $\pm$ 1.04	3.58
100	*	6.7039e-074 $\pm$ 62.0	3	1.2775e-072 $\pm$ 0.86	5	1.2809e-072 $\pm$ 1.52	16.4

TABLE V

4-NODE TANDEM NETWORK - SYMMETRIC ($\lambda = 0.1, \mu_1 = \mu_2 = \mu_3 = \mu_4 = 0.225$) ($\rho_1 = \rho_2 = \rho_3 = \rho_4 = 0.44$)

L	**Numerical**	**PW**	**SDA**		**SDH**		
	$\gamma(L)$	$\tilde{\gamma}(L) \pm$ RE%	b	$\tilde{\gamma}(L) \pm$ RE%	b	$\tilde{\gamma}(L) \pm$ RE%	VRR
25	*	3.2961e-06 $\pm$ 4.76	3	5.1771e-06 $\pm$ 0.16	6	5.1053e-06 $\pm$ 1.04	2.42
50	*	4.1536e-14 $\pm$ 59.6	5	6.4774e-14 $\pm$ 0.18	7	6.3828e-14 $\pm$ 0.89	9.28
100	*	4.5737e-32 $\pm$ 13.4	5	1.2525e-30 $\pm$ 0.43	10	1.2808e-30 $\pm$ 1.18	13.2

TABLE VI

4-NODE TANDEM NETWORK - ASYMMETRIC ($\lambda = 0.1, \mu_1 = 0.28, \mu_2 = 0.24, \mu_3 = 0.21, \mu_4 = 0.17$)
($\rho_1 = 0.36, \rho_2 = 0.42, \rho_3 = 0.48, \rho_4 = 0.59$)

L	**Numerical**	**PW**	**SDA**		**SDH**		
	$\gamma(L)$	$\tilde{\gamma}(L) \pm$ RE%	b	$\tilde{\gamma}(L) \pm$ RE%	b	$\tilde{\gamma}(L) \pm$ RE%	VRR
25	*	5.1899e-05 $\pm$ 1.17	3	5.43812e-005 $\pm$ 0.10	11	5.4232e-05 $\pm$ 0.28	5.40
50	*	9.4430e-11 $\pm$ 2.85	3	9.62647e-011 $\pm$ 0.17	11	9.6495e-11 $\pm$ 0.25	10.6
100	*	2.7979e-22 $\pm$ 1.59	4	2.90145e-022 $\pm$ 0.27	12	2.8865e-22 $\pm$ 0.25	37.8

TABLE VII

2-NODE PARALLEL NETWORK - SYMMETRIC ($\lambda_i = 0.1, \mu_i = 0.4$) ($\rho_1 = \rho_2 = 0.25$)

L	**Numerical**	**PW**	**SDA**		**SDH**		
	$\gamma(L)$	$\tilde{\gamma}(L) \pm$ RE%	b	$\tilde{\gamma}(L) \pm$ RE%	b_1, b_2	$\tilde{\gamma}(L) \pm$ RE%	VRR
25	6.4837e-14	2.4899e-14 $\pm$ 7.45	3	6.4826e-14 $\pm$ 0.06	3,3	6.4818e-14 $\pm$ 0.12	2.76
50	1.1675e-28	3.7971e-29 $\pm$ 36.2	4	1.1684e-28 $\pm$ 0.06	4,4	1.1650e-28 $\pm$ 0.15	1.67
100	1.8445e-58	1.9774e-59 $\pm$ 14.5	5	1.8527e-58 $\pm$ 0.08	5,5	1.8511e-58 $\pm$ 0.25	0.77

TABLE VIII

2-NODE PARALLEL NETWORK - SYMMETRIC ($\lambda_i = 0.15, \mu_i = 0.35$) ($\rho_1 = \rho_2 = 0.43$)

L	**Numerical**	**PW**	**SDA**		**SDH**		
	$\gamma(L)$	$\tilde{\gamma}(L) \pm$ RE%	b	$\tilde{\gamma}(L) \pm$ RE%	b_1, b_2	$\tilde{\gamma}(L) \pm$ RE%	VRR
25	1.9796e-08	1.1928e-08 $\pm$ 11.7	4	1.9800e-08 $\pm$ 0.06	4,4	1.9814e-08 $\pm$ 0.14	1.58
50	2.5813e-17	8.5168e-18 $\pm$ 12.7	5	2.5834e-17 $\pm$ 0.06	6,6	2.5904e-17 $\pm$ 0.17	0.98
100	2.0926e-35	2.3032e-35 $\pm$ 86.2	6	2.0923e-35 $\pm$ 0.07	7,7	2.0895e-35 $\pm$ 0.26	0.66

TABLE IX

2-NODE PARALLEL NETWORK - ASYMMETRIC ($\lambda_1 = 0.12, \mu_1 = 0.4, \lambda_2 = 0.08, \mu_2 = 0.4$) ($\rho_1 = 0.3, \rho_2 = 0.2$)

L	**Numerical**	**PW**	**SDA**		**SDH**		
	$\gamma(L)$	$\tilde{\gamma}(L) \pm$ RE%	b	$\tilde{\gamma}(L) \pm$ RE%	b_1, b_2	$\tilde{\gamma}(L) \pm$ RE%	VRR
25	5.6704e-13	7.2661e-13 $\pm$ 22.1	3	5.6480e-13 $\pm$ 0.12	2,5	5.6600e-13 $\pm$ 0.15	5.87
50	4.8047e-26	4.7674e-26 $\pm$ 3.88	3	4.7993e-26 $\pm$ 0.16	2,7	4.8188e-26 $\pm$ 0.20	3.30
100	3.4493e-52	3.3333e-52 $\pm$ 3.01	3	3.4434e-52 $\pm$ 0.21	2,10	3.4563e-52 $\pm$ 0.28	3.23

TABLE X

4-NODE PARALLEL NETWORK - SYMMETRIC ($\lambda_i = 0.05$, $\mu_i = 0.2$) ($\rho_i = 0.25$)

L	**Numerical**	**PW**	**SDA**		**SDH**		
	$\gamma(L)$	$\tilde{\gamma}(L) \pm$ RE%	b	$\tilde{\gamma}(L) \pm$ RE%	b_1, b_i	$\tilde{\gamma}(L) \pm$ RE%	VRR
25	*	8.5099e-13 $\pm$ 12.0	4	7.3197e-12 $\pm$ 0.08	4,4	7.3465e-12 $\pm$ 0.30	33.8
50	*	1.8289e-27 $\pm$ 48.1	4	5.0880e-26 $\pm$ 0.14	5,5	5.1083e-26 $\pm$ 0.41	43.0
100	*	4.6236e-58 $\pm$ 7.58	5	3.1658e-55 $\pm$ 0.14	5,5	3.1384e-55 $\pm$ 0.78	19.2

TABLE XI

4-NODE PARALLEL NETWORK - SYMMETRIC ($\lambda_i = 0.08, \mu_i = 0.17$) ($\rho_i = 0.47$)

L	**Numerical**	**PW**	**SDA**		**SDH**		
	$\gamma(L)$	$\tilde{\gamma}(L) \pm$ RE%	b	$\tilde{\gamma}(L) \pm$ RE%	b_1, b_i	$\tilde{\gamma}(L) \pm$ RE%	VRR
25	*	7.2898e-06 $\pm$ 9.09	3	1.7915e-05 $\pm$ 0.13	8,8	1.7924e-05 $\pm$ 0.24	23.5
50	*	1.1733e-13 $\pm$ 18.0	4	9.8414e-13 $\pm$ 0.26	8,8	9.8027e-13 $\pm$ 0.31	120.
100	*	4.1708e-30 $\pm$ 16.5	5	3.4284e-28 $\pm$ 0.59	9,9	3.4008e-28 $\pm$ 0.49	386.

TABLE XII

4-NODE PARALLEL NETWORK - ASYMMETRIC ($\lambda_1 = 0.06, \lambda_2 = 0.04, \lambda_3 = 0.04, \lambda_4 = 0.02; \mu_1 = \mu_2 = \mu_3 = \mu_4 = 0.2$)
($\rho_1 = 0.3, \rho_2 = \rho_3 = 0.2, \rho_4 = 0.1$)

L	**Numerical**	**PW**	**SDA**		**SDH**		
	$\gamma(L)$	$\tilde{\gamma}(L) \pm$ RE%	b	$\tilde{\gamma}(L) \pm$ RE%	b_1, b_i	$\tilde{\gamma}(L) \pm$ RE%	VRR
25	*	2.8583e-12 $\pm$ 18.9	4	2.4917e-12 $\pm$ 0.15	2,6	2.5012e-12 $\pm$ 0.35	135.
50	*	1.8266e-25 $\pm$ 2.59	4	2.1002e-25 $\pm$ 0.22	2,8	2.1268e-25 $\pm$ 0.64	56.7
100	*	1.4262e-51 $\pm$ 7.11	4	1.3031e-51 $\pm$ 0.37	2,10	1.5248e-51 $\pm$ 1.37	22.0