EFFICIENT SIMULATION OF POPULATION OVERFLOW IN 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

ABSTRACT

In this paper we propose a state-dependent importance sampling heuristic to estimate the probability of population overflow in networks of parallel queues. This heuristic approximates the "optimal" state-dependent change of measure without the need for difficult mathematical analysis or costly optimization involved in adaptive methodologies. Comprehensive simulations of networks with an arbitrary number of parallel queues and different traffic intensities yield asymptotically efficient estimates (with relative error increasing sub-linearly in the overflow level) where no other state-independent importance sampling techniques are known to be efficient. The efficiency of the proposed heuristic surpasses those based on adaptive importance sampling algorithms, yet it is easier to determine and implement and scales better for large networks.

1 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. Among the most effective methodologies researched and applied so far are those based on importance sampling (see, e.g., Parekh and Walrand 1989, Asmussen and Rubinstein 1995, Heidelberger 1995, Juneja and Nicola 2005).

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 (e.g., 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 Parekh and Walrand (1989) and further investigated in Frater et al. (1991). However, even for the simplest Jackson queueing network (e.g., 2-nodes in series or in parallel), the effectiveness of this heuristic is limited to only some region of the (arrival and service) parameters space (see Glasserman and Kou 1995, de Boer 2004). (We use the term "effectiveness" interchangeably with "asymptotic efficiency;" see, e.g., Nicola and Zaburnenko 2005b for a precise definition.)

Based on Markov additive process formulation of a two-node tandem network and large deviations arguments, work in Kroese and Nicola (2002) reveals that a state-dependent change of measure is effective where no effective state-independent change of measure exists. Since then, there has been increasingly more research on methodologies to obtain efficient state-dependent importance sampling heuristics, with very encouraging results. In de Boer and Nicola (2002) an adaptive optimization technique based on the method of cross-entropy (Rubinstein 2002) is used to approximate the "optimal" state-dependent change of measure. A drawback of this approach, however, is the excessive computational and storage demands for large state-space models associated with large networks.

In Nicola and Zaburnenko (2005a, 2005b) and Zaburnenko and Nicola (2005), heuristics are proposed to approximate the "optimal" state-dependent change of measure without the need for costly optimizations. 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 with the so obtained heuristic change of measure for tandem networks with multiple nodes yield estimates with a bounded relative error (see Zaburnenko and Nicola 2005, Nicola and Zaburnenko 2005a). In Nicola and Zaburnenko (2005b), changes of measure for feed-forward and feedback networks are pro-

1-4244-0501-7/06/$20.00 ©2006 IEEE

posed following the same heuristic approach. Experimental results reported there are encouraging but not sufficiently robust. This is primarily because an efficient heuristic to simulate parallel networks (which is a key to more complex topologies) was not available then.

In this paper we follow the same heuristic approach to develop a state-dependent change of measure for the efficient simulation of parallel queues with probabilistic routing. Experimental results to estimate the probability of population overflow in networks of up to 4 nodes in parallel produce asymptotically efficient estimates, with relative error increasing (sub-)linearly in the overflow level. The proposed heuristic is effective, robust, easy-to-implement and is shown to be more efficient than those based on adaptive methodologies (e.g., de Boer and Nicola 2002), particularly for large networks. Moreover, the findings provide crucial insights and pave the way to develop more effective and robust heuristic changes of measure (compared to those presented in Nicola and Zaburnenko 2005b) for feed-forward and other complex network topologies.

In Section 2 we introduce the basic model and define the probability of interest. The importance sampling technique is briefly reviewed. In Section 3 we introduce our heuristic approach, then we give a formal representation of the proposed SDH change of measure for parallel networks with probabilistic routing. The heuristic is also motivated using a time-reversal argument. In Section 4 we present experimental results and comparisons with other known methods to estimate the probability of population overflow in some examples of parallel networks. We conclude in Section 5.

2 PRELIMINARIES

The queueing network model and associated notation are introduced in Section 2.1. A brief review of importance sampling and some properties of simulation estimators are provided in Section 2.2.

2.1 Model and Notation

Consider a queueing network consisting of n nodes in parallel, each having its own (infinite) buffer. Customers arrive according to a Poisson process with rate λ. Upon arrival a customer is routed to one of the n parallel node according to some routing (scheduling) policy. An example of a “static” policy is probabilistic routing (considered in this paper), by which an arrival is assigned to node i with a fixed probability p_i. For this policy, the arrival process at node i is also Poisson with rate $\lambda_i = \lambda p_i$ $(i = 1,\ldots,n)$. An example of a “dynamic” scheduling policy is the JSQ (Join Shortest Queue) which, if applicable, may be preferred because of its load balancing feature and some optimality properties (see, e.g., Ephremides et al. 1980, Winston 1997). The service time at node i is exponentially distributed with rate μ_i $(i = 1,\ldots,n)$, after which the customer leaves the network. 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 $\mathbf{X}_t = (X_{1,t},\ldots,X_{n,t})$ 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, $\mathbf{X}_0 = (0,\ldots,0)$ 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.

2.2 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 (for a more comprehensive review see, e.g., Heidelberger 1995). Formally, let w be a sample path over the interval $[0,t]$. Then, the likelihood ratio associated with w is given by $W_t(w) = P(w)/\tilde{P}(w)$, 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 $\cdot$ 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 the asymptotic decay rate of $\tilde{\mathbb{E}} W_\tau{}^2 I_{\{S_\tau=L\}}$ is equal to that of $(\gamma(L))^2$, i.e.,

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

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 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. State-dependent changes of measure are generally more effective in simulations of rare events in queueing networks (see, e.g., Kroese and Nicola 2002, de Boer and Nicola 2002). Therefore, in this paper (as in Zaburnenko and Nicola 2005 and Nicola and Zaburnenko 2005a, 2005b) we aim at developing heuristics to approximate the "optimal" state-dependent change of measure.

3 A STATE-DEPENDENT HEURISTIC

Recent theoretical and empirical studies in Kroese and Nicola (2002) and de Boer and Nicola (2002) indicate that the "optimal" (or asymptotically efficient) change of measure depends on the network state, i.e., the number of customers at the network nodes. Furthermore, this crucial dependence is strong along the boundaries of the state-space (i.e., when one or more buffers are empty) and diminishes in the interior of the state-space (i.e., when contents of all buffers are sufficiently large).

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. In Nicola and Zaburnenko (2005a), heuristics based on combining known large deviations results and time-reversal arguments are used to construct such a change of measure for tandem networks. Empirical results show that it produces asymptotically efficient estimates for all feasible network parameters (the relative error is bounded for tandem networks having a single bottleneck). In the following we propose a heuristic state-dependent change of measure to efficiently simulate networks of parallel queues with probabilistic routing.

3.1 SDH for Parallel Networks with Probabilistic Routing

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

Let $x_i, i = 1,\dots,n$, 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 $\tilde{\lambda}_i(\mathbf{x})$ and $\tilde{\mu}_i(\mathbf{x})$ $(i = 1,\dots,n)$ the arrival and service rates at node i under the new change of measure, and by $\mathbf{SDH}_i(\mathbf{x})$ $(i = 1,\dots,n)$ the 2×2 linear operator (matrix) transforming the original rates into the new rates at node i $(i = 1,\dots,n)$. (For convenience, we occasionally abuse notation by dropping the vector $\mathbf{x}$). Define $[a]^+ = \max(a,0)$ and $[a]^1 = \min(a,1)$, then the change of measure at node i $(i = 1,\dots,n)$ is given by:

$$\begin{bmatrix} \tilde{\lambda}_i(\mathbf{x}) \\ \tilde{\mu}_i(\mathbf{x}) \end{bmatrix} = \mathbf{SDH}_i(\mathbf{x}) \begin{bmatrix} \lambda_i \\ \mu_i \end{bmatrix}, \tag{5}$$

$$\begin{aligned} \mathbf{SDH}_i(\mathbf{x}) &= \left[\frac{b_i - x_i}{b_i}\right]^+ \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \\ &\quad + \left[\frac{x_i}{b_i}\right]^1 \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \end{aligned} \tag{6}$$

for some integer $b_i \geq 1$, and $i = 1,\dots,n$. The first matrix is the identity matrix, corresponding to no change of measure. The second 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. Note that the change of measure at node i depends on the network state only through x_i. In a scalar form, the new rates are given by:

$$\tilde{\lambda}_i(x_i) = \left[\frac{b_i - x_i}{b_i}\right]^+ \lambda_i + \left[\frac{x_i}{b_i}\right]^1 \mu_i$$

and

$$\tilde{\mu}_i(x_i) = \left[\frac{b_i - x_i}{b_i}\right]^+ \mu_i + \left[\frac{x_i}{b_i}\right]^1 \lambda_i\,.$$

Note also that the equality $\sum_{i=1}^{n}(\tilde{\lambda}_i(\mathbf{x}) + \tilde{\mu}_i(\mathbf{x})) = 1$ holds under the above change of measure.

Remark 1 Note that b_i is the number of boundary levels for which the change of measure at node i depends on its 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. In general, the "optimal" b_i's (yielding estimates with lowest variance) depend on the set of network parameters as well as the overflow level L. Empirical results suggest dependence (in a yet formally non-apparent way) 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 empty nodes (at which $x_i = 0$) are not "pushed" at all, and busy nodes (at which $x_i \geqslant 1$) are "pushed" harder for higher x_i/b_i. The well-known heuristic in Parekh and Walrand (1989) 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 Parekh and Walrand (1989); both are asymptotically efficient.

3.2 Time Reversal Argument

The effectiveness of the change of measure in Section 3.1 for the simulation of parallel networks may be explained using time-reversal argument (see Kelly 1979 and Anantharam et al. 1990). The reverse time process is also an n-node parallel network. At node i $(i = 1,\dots,n)$, 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, $(L_1, L_2,\dots,L_n)$ with $\sum_{i=1}^{n} L_i = L$. In the reverse time, the number of customers at node i $(i = 1,\dots,n)$ 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 (respectively, arrivals) in reverse time correspond to arrivals (respectively, 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 Parekh and Walrand (1989) 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 $i = 1,\dots,n$, the state-dependent heuristic in Section 3.1 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 4.1 show that the heuristic is very effective and robust over the entire feasible parameter range.

4 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 2.2), then an unbiased estimator $\tilde{\gamma}$ of $\gamma(L)$ is given by

$$\tilde{\gamma} = \frac{1}{N} \sum_{i=1}^{i=N} I_i W_i \, . \tag{7}$$

The second moment of IW is estimated by

$$\tilde{\gamma^2} = \frac{1}{N} \sum_{i=1}^{i=N} I_i W_i^{\,2} \, . \tag{8}$$

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 comparisons we use the variance reduction ratio, $\mathrm{VRR} = \mathrm{RTV}_2 / \mathrm{RTV}_1$, which represents the efficiency gain of Estimator 1 relative to that of Estimator 2.

The experiments in this section are designed to demonstrate that the state-dependent change of measure proposed in Section 3.1 always yields asymptotically efficient estimates (mostly with bounded relative error), also for parameter settings where the state-independent change of measure in Parekh and Walrand (1989) is shown to be ineffective. Comparisons with the more recent and effective adaptive importance sampling methodology in de Boer and Nicola (2002) are also of interest and will be included. Similar to SDH, this adaptive methodology (here referred to as SDA) assumes state-dependence only over a (small) number of boundary layers (say, b_i at node i) which must be properly determined to ensure the effectiveness and efficiency of these methods. Too small b_i may not capture crucial dependencies close to the boundaries. Too large b_i may render SDH ineffective, but it will only reduce the efficiency of SDA. In either SDH or SDA, the "optimal" b_i's maximizing the efficiency (minimizing the RTV) may be determined by repeating the simulation for some "reasonable" (e.g., from experience) combinations of b_i's. Experimental results with SDH and SDA presented in this section are obtained using the corresponding "best" setting of b_i's.

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). To compare the heuristic in this paper (termed SDH) with the adaptive methodology (termed SDA) in de Boer and Nicola (2002), we also include VRR (relative to SDA). Hence, $\mathrm{VRR} > 1$ implies efficiency gain of SDH over SDA. Estimates obtained using the heuristic in Parekh and Walrand (1989) (termed PW) are also presented, however, these are not necessarily accurate or stable. Whenever feasible, numerical results (e.g., using the algorithm outlined in de Boer 2000) are included to verify the correctness of the simulation estimates. Otherwise (e.g., for larger networks and/or higher overflow levels), the corresponding table entry is marked with a "$*$". In these cases, agreement of the SDH and SDA estimates may be an indication of correctness.

4.1 Simulation of Parallel Networks with Probabilistic Routing

In this section we experiment with symmetric and asymmetric parallel networks of 2, 3, and 4 nodes. Network parameters are chosen in regions where the heuristic in Parekh and Walrand (1989) 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.

For the symmetric 2-node parallel network: $\lambda_1 = \lambda_2 = 0.15$ and $\mu_1 = \mu_2 = 0.35$ (i.e., $\rho_1 = \rho_2 = 0.43$). 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 1 and 2 show that unlike PW, SDH (as described in Section 3.1) 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 case. In the asymmetric case, $b_1 = 2$ and $b_2 (> b_1)$ increases with the overflow level L. Also SDA produces correct and stable estimates; however, it appears to be less efficient than SDH (as indicated by VRR ratios significantly higher than one).

For the symmetric 3-node parallel network: $\lambda_i = 0.1$ and $\mu_i = 0.2$, for $i = 1,2,3$ (i.e., $\rho_i = 0.5$, for $i = 1,2,3$). For the asymmetric 3-node parallel network: $\lambda_1 = 0.1, \lambda_2 = 0.075, \lambda_3 = 0.025$ and $\mu_i = 0.25$, for $i = 1,2,3$ (i.e., $\rho_1 = 0.4, \rho_2 = 0.3, \rho_3 = 0.1$). Experimental results in Tables 3 and 4 show that unlike PW, SDH (as described in Section 3.1) yields correct (numerical results are not feasible for higher overflow levels, 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" $b_i (i = 1,2,3)$ are equal only in the symmetric case. In the asymmetric case, $b_1 = 2$ and $b_2 = b_3 > b_1$; $b_i (i = 2,3)$ increases with the overflow level L. Also SDA produces correct and stable estimates; however, it appears to be much less efficient than SDH (as indicated by VRR ratios much higher than one).

For the symmetric 4-node parallel network: $\lambda_i = 0.05$ and $\mu_i = 0.2$, for $i = 1,2,3,4$ (i.e., $\rho_i = 0.25$, for $i = 1,2,3,4$). For the asymmetric 4-node parallel network: $\lambda_1 = 0.06, \lambda_2 = 0.04, \lambda_3 = 0.04, \lambda_4 = 0.02$ 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 5 and 6 show that unlike PW, SDH (as described in Section 3.1) 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" $b_i (i = 1,2,3,4)$ are equal only in the symmetric case. In the asymmetric case, $b_1 = 2$ and $b_2 = b_3 = b_4 > b_1$; $b_i (i = 2,3,4)$ increases with the overflow level L. Also SDA produces correct and stable estimates; however, it appears to be much less efficient than SDH (as indicated by VRR ratios much higher than one).

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 de Boer and Nicola 2002.) For the examples presented here, SDH typically requires only a few minutes to achieve relative errors less than 1% and is evidently more efficient than SDA (VRR > 1) even though its relative error (shown in the tables) may be higher.

5 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 to estimate (using importance sampling) the probability of population overflow in networks of parallel queues with probabilistic routing. Experimental results suggest asymptotically efficient estimates, mostly with bounded relative error. The efficiency of the obtained change of measure compares well with those determined using adaptive methodologies. However, our approach does not require costly pre-computation and is easy to implement. Moreover, its effectiveness is not diminished for larger networks, i.e., it is scalable.

Simple and robust guidelines for selecting the number of boundary layers (dependence range) need to be developed. Application of our approach to parallel networks with other (e.g., dynamic) routing policies would be of interest. The findings and supporting empirical results reported in this paper provide better insight and will help develop more effective and robust heuristics for other and more complex topologies.

REFERENCES

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.

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.

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

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.

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.

Ephremides, A., P. Varaiya, and J. Walrand. 1980. A simple dynamic routing problem. *IEEE Transactions on Automatic Control* 25 (8): 690–693.

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 Transactions on Automatic Control* 36: 1395–1405.

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

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

Juneja, S.K., and V.F. Nicola. 2005. Fast simulation of buffer overflow in queueing networks with probabilistic routing. *ACM Transactions on Modeling and Computer Simulation* 15 (4): 281–315.

Kelly, F.P. 1979. *Reversibility and Stochastic Networks.* Wiley, New York.

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

Nicola, V.F., and T.S. Zaburnenko. 2005a. 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),* 220–229.

Nicola, V.F., and T.S. Zaburnenko. 2005b. Efficient importance sampling heuristics for the simulation of population overflow in Jackson networks. In *Proceedings of the 2005 Winter Simulation Conference (WSC’05),* ed. M.E. Kuhl, N.M. Steiger, F.B. Armstrong, and J.A. Joines, 538–546.

Parekh, S., and J. Walrand. 1989. A quick simulation method for excessive backlogs in networks of queues. *IEEE Transactions on Automatic Control* 34 (1): 54–66.

Rubinstein, R.Y. 2002. The cross-entropy method and rare events for maximal cut and bipartition problems. *ACM Transactions on Modeling and Computer Simulation* 12 (1): 27–53.

Winston, J. 1997. Optimality of the shortest line discipline. *Journal of Applied Probability* 14: 181–198.

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.

AUTHOR BIOGRAPHIES

VICTOR F. NICOLA is an Associate Professor at the Faculty of Electrical Engineering, Mathematics and Computer

Science, University of Twente, The Netherlands. Before that he held positions at IBM Thomas J. Watson Research Center, New York, at Duke University, North Carolina, and at Eindhoven University, The Netherlands. He was also a Visiting Professor at the Norwegian University of Science and Technology and at Simula Research Laboratory, Norway. He is on the Editorial Board of the *International Journal of Simulation Modelling,* and served as a Guest Editor for the *ACM Transactions on Modeling and Computer Simulation.* His research interests include performance and reliability modeling and analysis; (rare event) simulation and optimization methodologies; with applications to high performance networked computing systems and broadband wireless/mobile communication. His e-mail address is: <v.f.nicola@ewi.utwente.nl>.

TATIANA S. ZABURNENKO is a PhD candidate at the Faculty of Electrical Engineering, Mathematics and Computer Science, University of Twente, The Netherlands. Her research interests are in the area of (rare event) simulation with applications in computer and communication networks. Her e-mail address is: <t.s.zaburnenko@ewi.utwente.nl>.

Table 1: 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 ± 11.7	4	1.9800e-08 ± 0.06	4,4	1.9814e-08 ± 0.14	1.58
50	2.5813e-17	8.5168e-18 ± 12.7	5	2.5834e-17 ± 0.06	6,6	2.5904e-17 ± 0.17	0.98
100	2.0926e-35	2.3032e-35 ± 86.2	6	2.0923e-35 ± 0.07	7,7	2.0895e-35 ± 0.26	0.66

Table 2: 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 ± 22.1	3	5.6480e-13 ± 0.12	2,5	5.6600e-13 ± 0.15	5.87
50	4.8047e-26	4.7674e-26 ± 3.88	3	4.7993e-26 ± 0.16	2,7	4.8188e-26 ± 0.20	3.30
100	3.4493e-52	3.3333e-52 ± 3.01	3	3.4434e-52 ± 0.21	2,10	3.4563e-52 ± 0.28	3.23

Table 3: 3-Node Parallel Network – Symmetric $(\lambda_i = 0.1,\ \mu_i = 0.2)$ $(\rho_i = 0.5)$

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	8.3550e-06	4.9906e-06 ± 20.8	5	8.3574e-06 ± 0.07	7,7	8.3550e-06 ± 0.19	7.26
50	*	2.7409e-13 ± 17.0	4	1.0608e-12 ± 0.38	8,8	1.0566e-12 ± 0.22	64.0
100	*	1.5623e-28 ± 8.69	5	3.7658e-27 ± 0.93	9,9	3.8483e-27 ± 0.32	114.

Table 4: 3-Node Parallel Network – Asymmetric $(\lambda_1 = 0.1, \lambda_2 = 0.075, \lambda_3 = 0.025; \mu_i = 0.25)$ $(\rho_1 = 0.4, \rho_2 = 0.3, \rho_3 = 0.1)$

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	8.2980e-10	7.8038e-10 ± 2.17	3	8.2976e-10 ± 0.16	2,6	8.3082e-10 ± 0.19	27.4
50	*	9.1128e-20 ± 2.64	3	9.3142e-20 ± 0.16	2,10	9.3459e-20 ± 0.25	7.39
100	*	1.1746e-39 ± 2.97	3	1.1589e-39 ± 0.39	2,14	1.1798e-39 ± 0.38	7.11

Table 5: 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 ± 12.0	4	7.3197e-12 ± 0.08	4,4	7.3465e-12 ± 0.30	33.8
50	*	1.8289e-27 ± 48.1	4	5.0880e-26 ± 0.14	5,5	5.1083e-26 ± 0.41	43.0
100	*	4.6236e-58 ± 7.58	5	3.1658e-55 ± 0.14	5,5	3.1384e-55 ± 0.78	19.2

Table 6: 4-Node Parallel Network – Asymmetric $(\lambda_1 = 0.06, \lambda_2 = 0.04, \lambda_3 = 0.04, \lambda_4 = 0.02; \mu_i = 0.2)$ $(\rho_1 = 0.3, \rho_2 = \rho_3 = 0.2, \rho_4 = 0.1)$

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 ± 18.9	4	2.4917e-12 ± 0.15	2,6	2.5012e-12 ± 0.35	135.
50	*	1.8266e-25 ± 2.59	4	2.1002e-25 ± 0.22	2,8	2.1268e-25 ± 0.64	56.7
100	*	1.4262e-51 ± 7.11	4	1.3031e-51 ± 0.37	2,10	1.5248e-51 ± 1.37	22.0