EFFICIENT HEURISTICS FOR SIMULATING POPULATION OVERFLOW IN PARALLEL NETWORKS

Tatiana S. Zaburnenko
Victor F. Nicola
Faculty of Electrical Engineering, Mathematics and Computer Science
University of Twente, P.O. Box 217
7500 AE Enschede, THE NETHERLANDS

(Extended 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 costly optimization involved in other recently proposed adaptive algorithms. Preliminary results from simulations of networks with up to 4 parallel queues and different traffic intensities yield asymptotically efficient estimates (with relative error increasing sublinearly in the overflow level) where state-independent importance sampling is ineffective.

1 INTRODUCTION

Importance sampling is one of the most effective methodologies for the efficient simulation of queueing networks involving rare events (see, e.g., Parekh and Walrand 1989, 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 (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 Parekh and Walrand (1989). 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 Nicola and Zaburnenko (2005) for a precise definition.)

Recent theoretical and empirical studies (see, e.g., Kroese and Nicola 2002 and de Boer and Nicola 2002) reveal that state-dependent change of measures are generally more effective, also where no effective state-independent change of measure exists. 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 Zaburnenko and Nicola (2005) and Nicola and Zaburnenko (2005), heuristics are proposed to approximate the "optimal" state-dependent change of measure without the need for a costly optimization. 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. The proposed heuristics are effective, easy to implement and could be more efficient than those based on adaptive methodologies (e.g., de Boer and Nicola 2002), particularly for large networks. Experimental results for tandem networks with multiple nodes yield asymptotically efficient estimates, mostly with a bounded relative error (see Zaburnenko and Nicola 2005, Nicola and Zaburnenko 2005).

In this paper we follow a similar heuristic approach to develop a state-dependent change of measure for the efficient simulation of rare events in parallel queues. In Section **??** we introduce the model and notation. In Section **??** we motivate and outline the SDH for parallel networks. In Section **??** we present experimental results and comparisons with the well-known heuristic in Parekh and Walrand (1989) for the estimation of the probability of network population overflow.

2 MODEL AND NOTATION

Consider a queueing network consisting of n nodes in parallel, each having its own (infinite) buffer. At node i ($1 \leqslant i \leqslant n$) customers arrive according to a Poisson process with rate λ_i. The service time is exponentially distributed with rate μ_i, after which customers exit the network. Let $X_{i,t}$ ($1 \leqslant i \leqslant n$) 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 (for a review see, e.g., Heidelberger 1995).

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

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

The estimator $\tilde{\mathbb{E}}\, W_\tau\, I_{\{S_\tau = L\}}$ is said to be *asymptotically efficient* if its relative error grows at subexponential (e.g., polynomial) rate as $L \to \infty$ (i.e., as $\gamma(L) \to 0$). The estimator is said to have *bounded relative error* if its relative error is bounded in L as $\gamma(L) \to 0$. 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.

3 STATE-DEPENDENT HEURISTICS

Recent theoretical and empirical studies in Kroese and Nicola (2002) and de Boer and Nicola (2002) indicate that the "optimal" 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). This 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 (2005), 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, mostly with a bounded relative error. Here we propose a heuristic state-dependent change of measure to efficiently simulate networks of parallel queues.

SDH for the n-node Parallel Network:

Let λ_i and μ_i, respectively, be the arrival and service rates 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 $\sum_{i=1}^{n} (\lambda_i + \mu_i) = 1$. Denote by $\tilde{\lambda}_i$ and $\tilde{\mu}_i$ the corresponding rates at node i under the new change of measure, and by $\mathbf{SDH}_i$ the 2×2 linear operator (matrix) transforming the original rates into the new rates at node i $(i = 1, \ldots, n)$. Define $[a]^+ = \max(a, 0)$ and $[a]^1 = \min(a, 1)$, then the change of measure at node i $(i = 1, \ldots, n)$ is given by:

$$\begin{bmatrix} \tilde{\lambda}_i \\ \tilde{\mu}_i \end{bmatrix} = \mathbf{SDH}_i \begin{bmatrix} \lambda_i \\ \mu_i \end{bmatrix}, \tag{3}$$

$$\begin{aligned} \mathbf{SDH}_i = {} & \left[\frac{b_i - x_i}{b_i}\right]^+ \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \\ & + \left[\frac{x_i}{b_i}\right]^1 \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \end{aligned} \tag{4}$$

for some integer $b_i \geq 1$. 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; this corresponds to interchanging the arrival and service rates at node i. Note that, under the above change of measure, the equality $\sum_{i=1}^{n} (\tilde{\lambda}_i + \tilde{\mu}_i) = 1$ holds.

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 (particularly the traffic intensities ρ_i's) as well as the overflow level L.

According to the above change of measure, empty nodes are not "pushed" (overloaded) at all, and busy nodes are "pushed" simultaneously, however, to different extents depending on their respective ratios of 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 works only in a limited region of the network parameters space (namely, when the utilization at the bottleneck node is sufficiently high relative to 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.

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, then unbiased estimators of the first and second moments of $I\,W$ are given by

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

The variance and the relative error of the importance sampling estimator $\tilde{\gamma}$ (of $\gamma(L)$) are given by $VAR(\tilde{\gamma}) = (\tilde{\gamma^2} - (\tilde{\gamma})^2)\,/\,(N-1)$ and $RE(\tilde{\gamma}) = \sqrt{VAR(\tilde{\gamma})}\,/\,\tilde{\gamma}$, respectively.

In the following we experiment with 2- and 4-node parallel networks. The intent is to demonstrate the effectiveness of our proposed heuristic (termed SDH) compared to that in Parekh and Walrand (1989) (termed PW). For the 2-node parallel network: $\lambda_1 = \lambda_2 = 0.15$ and $\mu_1 = \mu_2 = 0.35$ (i.e., a symmetric network with $\rho_1 = \rho_2 = 0.43$). For the 4-node parallel network: $\lambda_i = 0.05$ and $\mu_i = 0.2$, for $i = 1, 2, 3, 4$ (i.e., a symmetric network with $\rho_i = 0.25$, for $i = 1, 2, 3, 4$). (Typically, PW fails to efficiently estimate the probability of population overflow when the node utilizations are equal or sufficiently close.)

In all simulation experiments, the same number of replications, namely, 10^6, is used to obtain estimates of the population overflow probability $\gamma(L)$ using both, our state-dependent heuristic in Section **??** (SDH) and the heuristic in Parekh and Walrand (1989) (PW). For each estimate we include the relative error (in percentage). Whenever feasible, numerical results (for example, using the algorithm outlined in de Boer 2000) are included to verify the correctness of the simulation estimates. Otherwise, the corresponding table entry is marked with a "$*$".

Experimental results in Tables **??** and **??** show that PW yields incorrect and unstable estimates. On the other hand, our proposed SDH yields correct and asymptotically efficient estimates with a relative error increasing sub-linearly in the overflow level L. Due to symmetry in the above examples, the "best" value for the dependence range (b_i) is the same at all nodes. Typically, SDH requires only a few minutes to achieve relative errors less than 1%. Moreover, the SDH approach does not require difficult analysis or costly precomputation, and its effectiveness is not diminished for networks with larger state-space. However, simple and robust guidelines for selecting the number of boundary layers (dependence range) need to be developed and tested through analysis and/or extensive experimentation.

REFERENCES

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 Trans. 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.

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

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

Juneja, S.K., and V.F. Nicola. 2004. Efficient simulation of buffer overflow probabilities in Jackson networks with feedback. *ACM Trans. on Modeling and Computer Simulation.* Under final revision.

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

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.

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

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

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.

Table 1: 2-Node Parallel Network ($\rho_1 = \rho_2 = .43$)

L	**Num.**	**PW**	**SDH**	
	$\gamma(L)$	$\tilde{\gamma}(L) \pm RE\%$	**b**	$\tilde{\gamma}(L) \pm RE\%$
25	1.980e-08	1.193e-08 ± 12	4	1.976e-08 ± 0.14
50	2.581e-17	8.517e-18 ± 13	6	2.583e-17 ± 0.17
100	2.093e-35	2.303e-35 ± 86	7	2.092e-35 ± 0.27

Table 2: 4-Node Parallel Network ($\rho_i = .25, 1 \leq i \leq 4$)

L	**Num.**	**PW**	**SDH**	
	$\gamma(L)$	$\tilde{\gamma}(L) \pm RE\%$	**b**	$\tilde{\gamma}(L) \pm RE\%$
25	*	8.510e-13 ± 12	4	7.364e-12 ± 0.30
50	*	1.829e-27 ± 48	5	5.043e-26 ± 0.40
100	*	4.624e-58 ± 08	6	3.171e-55 ± 0.87