Efficient Simulation of Buffer Overflow Probabilities in Jackson Networks with Feedback

S.K. Juneja
Tata Institute of Fundamental Research
Mumbai, India
juneja@tifr.res.in

V.F. Nicola
University of Twente
Enschede, The Netherlands
nicola@cs.utwente.nl

February 18, 2005

Abstract

Consider a Jackson network that allows feedback and that has a single server at each queue. The queues in this network are classified as a single 'target' queue and the remaining 'feeder' queues. In this setting we develop the large deviations limit and an asymptotically efficient importance sampling estimator for the probability that the target queue overflows during its busy period, under some regularity conditions on the feeder queue-length distribution at the initiation of the target queue busy period. This importance sampling distribution is obtained as a solution to a non-linear program. We especially focus on the case where the feeder queues, at the initiation of the target queue busy period, have the steady state distribution corresponding to these instants. In this setting, we explicitly identify the importance sampling distribution when the feeder queue service rates exceed a specified threshold. We also relate our work to the existing large deviations literature to develop a perspective on successes and limitations of our results.

1 Introduction

In this paper we consider efficient estimation of small probabilities of buffer overflow in a Jackson network setting using the importance sampling simulation technique (see, e.g., Glynn and Iglehart 1989 for an introduction to importance sampling). This problem of efficient simulation of such probabilities is of practical relevance in communication networks design; for example, in determining buffer allocation to keep the overflow probability at an acceptably low level. It may also be of interest in the design of session admission and routing algorithms at a switch (see Heidelberger 1995 and Chang et al. 1994 for a discussion on possible applications).

We consider a Jackson network that allows feedback and has a single server at each queue. Our aim is study queue-length build-up at a pre-specified 'target' queue. Thus, the queues in the network are classified as a target queue and the remaining 'feeder' queues. Specifically, we focus on estimating the probability that the queue-length at the target queue hits a large level N during its busy period (a busy period of the target queue is initiated when an arrival to it finds it empty, and it ends when subsequently the target queue re-empties).

The initial distribution of queue lengths at the feeder queues (i.e., at the instants of busy period initiations at the target queue) plays an important role in our analysis. We focus primarily on two types of initial distributions:

1. *Type-I* initial distribution: Here, with probability 1 each feeder queue is less than or equal to a constant $c(N)$ (i.e., it may depend on N) that is $o(N)$.
2. *Type-II* distribution: Here, the feeder queues have the steady-state distribution corresponding to the instants of target queue busy period initiations. The problem of efficient estimation of this probability is closely related to the problem of estimating the steady-state loss probability when the target queue has a finite buffer of size N , i.e., the fraction of customers lost due to buffer overflow at the target queue in the steady-state (see, e.g., Chang et. al. 1994, L'Ecuyer and Champoux 2001). In many applications this loss probability is an important performance measure of interest.

The main contributions of this paper include:

1. We propose a change of measure as a solution to a set of conditions to estimate the probability of interest. Using this change of measure, under Type-I initial distribution, we develop the large deviations limit for the probability of interest and an asymptotically efficient importance sampling estimator. Additionally, we note that such results hold even when Type-I initial distribution is replaced by the condition that the feeder queues have finite small buffers of $o(N)$ size.
2. We explicitly identify the proposed change of measure when the feeder queue service rates are beyond a specified threshold. We also develop a simple heuristic non-linear program (NLP) to help determine the proposed change of measure in more general settings.
3. We note that under Type-II initial distribution, the problem is more complex and that the proposed change of measure may be asymptotically efficient for estimating the probability of interest under Type-II initial distribution only when the service rates at the feeder queues are beyond another specified threshold. We also develop conditions under which the proposed change of measure, although not asymptotically efficient, may estimate the probability of interest at a provably asymptotically faster rate than the naive simulation (i.e., yields 'asymptotic gain').

The proposed change of measure has a natural generalization to queues with renewal inter-arrival and service streams and batch arrivals and services (see Juneja and Nicola 2003). Extensions that generalize renewal inter-event streams to Markov additive processes and show how super-position of many such independent streams may be handled are also well known (see, e.g., Chang et. al. 1994, Beck et. al. 1999). We restrict our focus to Jackson networks to illustrate the key ideas clearly so as to minimize the notational complexity. Another reason for this restriction is that Juneja and Nicola (2003) make some assumptions that are difficult to verify from model primitives for more general networks but are easily verified for Jackson networks. In particular, in Jackson network settings it is easy to see the successes and limitations of the proposed change of measure under Type-II initial distribution.

Anantharam, Heidelberger and Tsoucas (1990) show that the most likely paths along which queue lengths build up in a Jackson network correspond to the most likely paths (as governed by the law of large numbers) through which the reversed Jackson network empties (In their analysis, and as is standard in the analysis of Jackson networks using large deviations theory, the queue lengths are scaled appropriately so that the paths of the scaled process correspond to the rates of the original network). Note that such paths may be piece-wise linear. Thus, when the target queue empties in the reversed network, some queues may remain empty while others may fill up (at a particular set of rates for the unscaled queue length process). If some queues are filling up with the emptying of the target queue, then once the target queue is empty, the rate dynamics for the remaining network changes. While Anantharam, Heidelberger and Tsoucas (1990) identify the most likely paths along which rare events occur in Jackson network, this may not suffice for conducting importance sampling. As shown in Glasserman and Kou (1995), importance sampling using the measure corresponding to these paths viewed in forward direction, need not be successful (it may lead to large, even infinite variance; see Randhawa and Juneja 2004). Frater, Lennon and Anderson (1990) note that when the feeder queue service rates are beyond a specified threshold, in the reversed network, the target queue empties along a single path. That is, as the target queue empties, their is no build up in any other queue, and the network empties when the target queue empties. One of our contributions is that we prove that the change of measure corresponding to these paths in the forward direction indeed efficiently estimates the probability of large level crossing at target queue under Type-II initial distribution.

As mentioned earlier, the problem of efficiently estimating the probability that the target queue crosses a large level under Type-II initial distribution is closely related to efficiently estimating the steady state probability of the target queue crossing a large level. In the Jackson network setting, the latter probability has a well known product form expression. However, even for Jackson networks, a practically important performance measure such as mean time between large level crossings in the target queue, may not be amenable to closed form analysis. We empirically show that the methodology developed to efficiently estimate the probability of target queue crossing a large level in its busy cycle, also proves useful in efficient estimation of mean time between large level crossings in the target queue (Glynn et al. 1993 show this in reliability settings).

In Section 2, the conditions that the proposed change of measure must satisfy are stated and the main results assuming these conditions are stated and proved. In Section 3, we derive the large deviations limit for the probability of interest under Type-II initial distribution. Here, we also outline the simulation methodology to estimate mean time between large level crossings in the target queue. In Section 4, we explicitly identify the change of measure when the service rates at the feeder queues are beyond a specified threshold. Here, we also briefly discuss the efficiency of the proposed change of measure under Type-II initial distribution. To gain insight into the successes and limitations of the proposed method under Type-II initial distribution and to aid in possible enhancements, we briefly review recent and relevant literature on large deviations behavior in queuing networks in Section 5. Simulation experiments to support our theoretical results and to verify the efficacy of the proposed change of measure are included in Section 6. In this section we also give a heuristic to determine the proposed change of measure. Proofs requiring significant technical details are relegated to the Appendix.

2 Mathematical Framework and Main Results

Consider a queuing network in which all queues are indexed by the set $\mathcal{H}$. These queues are further categorized by one 'target' queue indexed by t and the remaining 'feeder' queues indexed by the set $\mathcal{F}$. Thus, $\mathcal{H} \equiv \{t\} \cup \mathcal{F}$. We assume that each queue has a single server and an infinite buffer. The following notation is associated with the queues in the network under the probability measure $\mathbf{P}$. For each $i \in \mathcal{H}$, let λ_i denote the external arrival rate to queue i and let μ_i denote the service rate at queue i. Let $P = (p_{ik} : i, k \in \mathcal{H})$ denote the matrix of routing probabilities, and let p_{ie} denote the probability that a customer leaving queue i exits the network. We assume that every customer arriving to the network eventually exits the network. Thus, $P^n \to 0$ as $n \to \infty$. We further assume that arrival at any queue has a positive probability of reaching the target queue. Let γ_i denote the total arrival rate at queue $i \in \mathcal{H}$. From the traffic equations it follows that

$$\gamma_i = \lambda_i + \sum_{j \in \mathcal{H}} p_{ji} \min(\gamma_j, \mu_j).$$

We also assume that $0 < \gamma_i < \mu_i$ for all $i \in \mathcal{H}$, and hence the network is stable.

Let $\tilde{\mathbf{Q}} = (\tilde{Q}_s : s \geq 0)$ denote the continuous time Markov chain (CTMC) associated with the Jackson network, where $\tilde{Q}_s = (\tilde{Q}_s(i) : i \in \mathcal{H})$, and each $\tilde{Q}_s(i)$ denotes the queue length at queue i at time s. Our primary focus is on the embedded discrete time Markov chain (DTMC) $\mathbf{Q} = (Q_n : n \geq 0)$ associated with this network, where $Q_n = (Q_n(i) : i \in \mathcal{H})$, and each $Q_n(i)$ denotes the queue length at queue i at the n th jump instant of the network, i.e., just after a service completion and associated routing of customers or external arrivals. Note that $\tilde{Q}_s(i)$ and $Q_n(i)$ include the customer in service at queue i, if any.

Let $\mathcal{Z}_+$ denote the set of non-negative integers. For $N \geq 1$, let $\mathcal{S}_N \subset \mathcal{Z}_+^{|\mathcal{H}|}$ denote the set of states corresponding to N customers in the target queue; $T_N = \inf\{n \geq 1 : Q_n \in \mathcal{S}_N\}$. Similarly, let $\mathcal{O} \subset \mathcal{Z}_+^{|\mathcal{H}|}$ denote the set of states where the target queue is empty; $T_o = \inf\{n \geq 1 : Q_n \in \mathcal{O}\}$. Further, define the event $\mathcal{B}_N = \{T_N < T_o\}$. Let π denote a distribution of Q_0 that corresponds to the network state upon the initiation of busy cycles at the target queue (thus, $Q_0(t) = 1$ under π). In the later sections we consider some specific distributions for π. Wherever useful, we append a subscript π to $\mathbf{P}$ to show its dependence on π. Our primary interest is in efficiently estimating $\mathbf{P}_\pi(\mathcal{B}_N)$ for large N under suitable restrictions on π, using importance sampling.

2.1 Proposed Change of Measure

Let $\mathbf{P}^*$ denote a change of measure under which the original Jackson network remains a Jackson network. Under it, for $i \in \mathcal{H}$, λ_i^* and μ_i^* denote the external arrival and service rates at queue i, γ_i^* denotes the total arrival rate at queue i and the transition probabilities are given by $(p_{ik}^*, p_{ie}^* : i, k \in \mathcal{H})$. Let $\mathcal{D} \subset \mathcal{H}$ denote the set $\{i : \lambda_i = 0\}$.

Assumption 1 *There exists a probability measure $\mathbf{P}^*$ and a set $\mathcal{C} \subseteq \mathcal{F}$ such that $\mu_i^* \geq \mu_i$ for $i \in \mathcal{C}$, $\mu_i^* = \mu_i$ for $i \in \mathcal{F} - \mathcal{C}$, and $\mu_t^* < \mu_t$. Furthermore,*

$$\sum_{i \in \mathcal{H}} (\lambda_i^* + \mu_i^*) = \sum_{i \in \mathcal{H}} (\lambda_i + \mu_i), \tag{1}$$

and there exist constants $c_j > 1$ such that $\lambda_j^* = c_j \lambda_j$ for $j \in \mathcal{H}$ and

$$p_{ij}^* = \frac{c_j}{c_i} \frac{\mu_i}{\mu_i^*} p_{ij},$$

for all $i, j \in \mathcal{H}$ and

$$p_{ie}^* = \frac{1}{c_i} \frac{\mu_i}{\mu_i^*} p_{ie}.$$

In addition, For $i \in \mathcal{C}$,

$$\gamma_i^* = \mu_i^*, \tag{2}$$

and for $i \in \mathcal{F} - \mathcal{C}$,

$$\gamma_i^* < \mu_i. \tag{3}$$

Here, $\mathcal{C}$ denotes the set of feeder queues that are critical under $\mathbf{P}^*$. This set is allowed to be empty. Also, under Assumption 1, the traffic equations reduce to

$$\gamma_i^* = \lambda_i^* + \sum_{j \in \mathcal{F}} p_{ji}^* \gamma_j^* + \min(\gamma_t^*, \mu_t^*) p_{ti}^*, \tag{4}$$

for all $i \in \mathcal{H}$. Note that, under Assumption 1, $\lambda_i^* > \lambda_i$ for $i \in \mathcal{H} - \mathcal{D}$, $\lambda_i^* = \lambda_i = 0$ for $i \in \mathcal{D}$.

Lemma 1 *Under Assumption 1, the target queue is unstable under* $\mathbf{P}^*$*, i.e.,* $\gamma_t^* > \mu_t^*$*.*

Some notation is needed to prove the lemma and for further analysis. Since every arriving customer eventually leaves the system, $(I-P)^{-1} = \sum_{n=0}^{\infty} P^n$ exists and is a non-negative matrix (see, e.g., Chapter 7, Chen and Yao 2001). Let $R = (r_{ij} : i, j \in \mathcal{H})$ equal $(I-P)^{-1}$. Similarly, let P^* denote the routing probability matrix $(p_{ij}^* : i, j \in \mathcal{H})$. Then, $(I-P^*)^{-1}$ exists and is non-negative. Let $R^* = (r_{ij}^* : i, j \in \mathcal{H})$ equal $(I-P^*)^{-1}$. Note that if all the queues in the network are stable or critical under $\mathbf{P}^*$, then r_{ij}^* denotes the expected number of visits to queue j by a customer in queue i before it leaves the network. The proof of Lemma 1 is given in Appendix B.

2.2 Importance Sampling

Note that estimating $\mathbf{P}_\pi(\mathcal{B}_N)$ using naive simulation involves generating a sample of $(Q_n : 0 \le n \le T)$ using $\mathbf{P}_\pi$, where $T = \min(T_o, T_N)$. If the event $\mathcal{B}_N$ occurs, then the indicator function $I(\mathcal{B}_N)$ is set to 1, otherwise it is set to zero. An average of i.i.d. samples of $I(\mathcal{B}_N)$ gives an estimate of $\mathbf{P}_\pi(\mathcal{B}_N)$. It is well known that naive simulation is computationally prohibitive for estimating small probabilities.

Estimating $\mathbf{P}_\pi(\mathcal{B}_N)$ using the importance sampling change of measure $\mathbf{P}_\pi^*$ involves generating samples of $(Q_n : 0 \le n \le T)$ using $\mathbf{P}_\pi^*$. The output of each such sample $I(\mathcal{B}_N)$ is unbiased by multiplying it with the likelihood ratio L_T (the ratio of the probability of the generated path under $\mathbf{P}_\pi$ and under $\mathbf{P}_\pi^*$). To evaluate L_T, consider a transition of Markov chain from state Q_n to Q_{n+1} such that $Q_{n+1}(i) = Q_n(i) + 1$ and $Q_{n+1}(j) = Q_n(j)$ for all $j \neq i$. For such a transition, the ratio of the transition probabilities under $\mathbf{P}_\pi$ and $\mathbf{P}_\pi^*$ equals

$$\frac{\lambda_i / (\sum_{j \in \mathcal{H}} [\lambda_j + \mu_j I(Q_n(j) > 0)])}{\lambda_i^* / (\sum_{j \in \mathcal{H}} [\lambda_j^* + \mu_j^* I(Q_n(j) > 0)])}.$$

In particular, in view of (1) and the facts that $\mu_i^* \ge \mu_i$ for $i \in \mathcal{F}$ and $Q_n(t) > 0$ for all $n \le T$, this ratio is upper bounded by $\frac{\lambda_i}{\lambda_i^*} = \frac{1}{c_i}$ for all $n \le T$. Note that it equals $\frac{1}{c_i}$ if $Q_n(i) > 0$ for all $i \in \mathcal{C}$ (since $\mu_i^* = \mu_i$ for $i \in \mathcal{F} - \mathcal{C}$).

Similarly, if transition from Q_n to Q_{n+1} corresponds to a departure from queue i that transitions to queue j (resp., leaves the system) then the ratio of probabilities for such a transition under $\mathbf{P}_\pi$ and $\mathbf{P}_\pi^*$ is upper bounded by $\frac{\mu_i}{\mu_i^*} \frac{p_{ij}}{p_{ij}^*}$ (resp. $\frac{\mu_i}{\mu_i^*} \frac{p_{ie}}{p_{ie}^*}$) for all $n \le T$. Again, note that the upper bound holds as an equality if $Q_n(i) > 0$ for all $i \in \mathcal{C}$.

Thus, we conclude that

$$L_T \le \prod_{i \in \mathcal{H}} \left((\frac{\lambda_i}{\lambda_i^*})^{N_T(i)} (\frac{\mu_i}{\mu_i^*})^{M_T(i)} \prod_{j \in \mathcal{H} \cup \{e\}} (\frac{p_{ij}}{p_{ij}^*})^{M_T(i,j)} \right) \tag{5}$$

a.s., where $N_m(i)$ denotes the number of external arrivals to queue i, $M_m(i)$ denotes the number of service completions from queue i, and $M_m(i,j)$ denotes the number of service completions from queue i to queue j (or to outside the network if $j = e$) up to jump instant m of the Markov chain $(Q_n : n \geq 0)$, and in this definition we use the convention that $0/0 = 1$. Since, for $i \in \mathcal{D}$, $N_T(i) = 0$ a.s., we can rewrite (5) as

$$L_T \leq \prod_{i \in \mathcal{H}} \left((\frac{1}{c_i})^{N_T(i)} (\frac{\mu_i}{\mu_i^*})^{M_T(i)} \prod_{j \in \mathcal{H} \cup \{e\}} (\frac{p_{ij}}{p_{ij}^*})^{M_T(i,j)} \right) \tag{6}$$

a.s. In our analysis it would also be useful to note that

$$L_T = \prod_{i \in \mathcal{H}} \left((\frac{1}{c_i})^{N_T(i)} (\frac{\mu_i}{\mu_i^*})^{M_T(i)} \prod_{j \in \mathcal{H} \cup \{e\}} (\frac{p_{ij}}{p_{ij}^*})^{M_T(i,j)} \right) \tag{7}$$

a.s. along $\tilde{\mathcal{B}}_N$, the set of sample paths in the rare event $\mathcal{B}_N$ where the queues in $\mathcal{C}$ never empty.

2.3 Main Results

The following assumption is essential to Theorem 1:

Assumption 2 *Under Assumption 1,*

$$\log E_{\mathbf{P}_\pi^*} \left((\prod_{i \in \mathcal{F}} c_i^{2(Q_0(i) - Q_T(i))}) I(\mathcal{B}_N) \right)$$

is upper bounded by $o(N)$[*] term.

Note that Assumption 2 implies that

$$\log E_{\mathbf{P}_\pi^*} \left((\prod_{i \in \mathcal{F}} c_i^{(Q_0(i) - Q_T(i))}) I(\mathcal{B}_N) \right) \tag{8}$$

is upper bounded by $o(N)$ term.

Assumption 2 is always satisfied under Type-I initial distribution, i.e., when $Q_0(i)$ is bounded by an $o(N)$ term for each $i \in \mathcal{F}$. We discuss sufficient conditions for this to hold under Type-II initial distribution in Section 4.2.

Theorem 1 *Under Assumption 1 and Equation (8),*

$$\lim_{N \to \infty} \frac{\log(\mathbf{P}_\pi(\mathcal{B}_N))}{N} = -\log c_t. \tag{9}$$

Under additional Assumption 2,

$$\lim_{N \to \infty} \frac{\log(E_{\mathbf{P}_\pi^*}(L_T^2 I(\mathcal{B}_N)))}{N} = -2\log c_t. \tag{10}$$

[*] A function $f(N)$ is said to be $o(N)$ if $\lim_{N \to \infty} f(N)/N = 0$. It is said to be $O(N^x)$ for $x \geq 0$ if there exists a positive constant K such that $f(N) \leq KN^x$ for all N sufficiently large. It is said to be $\Theta(N^x)$ for $x \geq 0$ if there exist positive constants K_1 and K_2, $(K_1 < K_2)$, such that $K_1 N^x \leq f(N) \leq K_2 N^x$ for all N sufficiently large.

We later identify cases where Assumption 2 may not hold, however, the following condition may hold:

$$E_{\mathbf{P}_\pi^*}\left(\prod_{i\in\mathcal{F}} c_i^{kQ_0(i)}\right) < \infty \tag{11}$$

for $k \in [1,2]$.

Theorem 2 *Under Assumption 1 and Equation (11),*

$$\limsup_{N\to\infty} \frac{\log(E_{\mathbf{P}_\pi^*}(L_T^2 I(\mathcal{B}_N)))}{N} \le -k\log c_t. \tag{12}$$

Proof of Upper Bound in Theorem 1: We first show that the right hand side (RHS) upper bound the LHS in (9) and (10). From physical balance considerations the following equality holds a.s.

$$N_T(i) = \sum_{j\in\mathcal{H}\cup\{e\}} M_T(i,j) + Q_T(i) - \sum_{j\in\mathcal{H}} M_T(j,i) - Q_0(i). \tag{13}$$

Plugging this into (6) and reorganizing (6), it follows that

$$L_T \le \left(\prod_{i\in\mathcal{H}} c_i^{Q_0(i)-Q_T(i)}\right)\left(\prod_{i,j\in\mathcal{H}}\left[\frac{\mu_i}{\mu_i^*}\frac{c_j}{c_i}\frac{p_{ij}}{p_{ij}^*}\right]^{M_T(i,j)}\right)\left(\prod_{i\in\mathcal{H}}\left[\frac{\mu_i}{\mu_i^*}\frac{1}{c_i}\frac{p_{ie}}{p_{ie}^*}\right]^{M_T(i,e)}\right) \tag{14}$$

a.s. From Assumption 1, and the fact that on $\mathcal{B}_N$, $Q_T(t) - Q_0(t) = N-1$, it follows that

$$L_T \le (\frac{1}{c_t})^{N-1}\left(\prod_{i\in\mathcal{F}} c_i^{(Q_0(i)-Q_T(i))}\right) \tag{15}$$

a.s. Integrating the above over the set $\mathcal{B}_N$ with respect to $\mathbf{P}_\pi^*$, noting that $\mathbf{P}_\pi(\mathcal{B}_N) = E_{\mathbf{P}_\pi^*}(L_T I(\mathcal{B}_N))$ and (8), we get

$$\lim_{N\to\infty} \frac{\log(\mathbf{P}_\pi(\mathcal{B}_N))}{N} \le -\log c_t. \tag{16}$$

Similarly, the upper bound for (10) holds.

Sketch of Proof of Lower Bound in Theorem 1: We now proceed to establish that $-\log c_t$ is the large deviations lower bound (ldlb) for the probability $\mathbf{P}_\pi(\mathcal{B}_N)$ (since $E_{\mathbf{P}_\pi^*}(L_T^2 I(\mathcal{B}_N)) \ge \mathbf{P}_\pi(\mathcal{B}_N)^2$, the appropriate lower bound for $E_{\mathbf{P}_\pi^*}(L_T^2 I(\mathcal{B}_N))$ then follows). Since this involves a significant degree of technicalities, we first discuss the broad ideas before working out the details in Appendix A. We establish this lower bound by identifying a subset, say $\bar{\mathcal{B}}_N$ of $\mathcal{B}_N$ that has substantial probability under $\mathbf{P}_\pi^*$ in the sense that

$$\liminf_{N\to\infty} \frac{\log \mathbf{P}_\pi^*(\bar{\mathcal{B}}_N)}{N} = 0, \tag{17}$$

and such that $L_T \approx c_t^{-N+o(N)}$ along paths in this subset. More precisely, $L_T \ge c_t^{-N+o(N)}$ uniformly along this set. Once this is achieved, the desired ldlb follows by noting that

$$\mathbf{P}_\pi(\mathcal{B}_N) = E_{\mathbf{P}_\pi^*}(L_T I(\mathcal{B}_N)) \ge E_{\mathbf{P}_\pi^*}(L_T I(\bar{\mathcal{B}}_N)) \ge c_t^{-N+o(N)}\mathbf{P}_\pi^*(\bar{\mathcal{B}}_N).$$

Roughly speaking, such a $\bar{\mathcal{B}}_N$ corresponds to $\tilde{\mathcal{B}}_N$; the set of paths to the rare set in which the critical queues remain busy. Since the target queue is unstable and the queues in $\mathcal{C}$ are critical, this set may be expected to satisfy (17). Along this set (7) holds, and in view of the simplifications shown in the proof of the upper bound

$$L_T = (\frac{1}{c_t})^{N-1} \left(\prod_{i \in \mathcal{F}} c_i^{(Q_0(i) - Q_T(i))} \right)$$

a.s. along $\bar{\mathcal{B}}_N$.

Since the target queue is unstable, T is $\Theta(N)$ with large probability. Also note that the critical queues behave similar to a symmetric random walk. It is well known that a supremum of the absolute value of a symmetric random walk increases at a rate that is proportional to the square root of time (see, e.g., Lemma 7). Thus, it is reasonable to expect that $(Q_T(i) - Q_0(i))$ is $O(\sqrt{N})$ with large probability, for $i \in \mathcal{C}$. For the stable light tailed queues ($i \in \mathcal{F} - \mathcal{C}$), it is well known in great generality that the supremum of the queue length grows at a logarithmic rate with time (see, e.g., Zeevi and Glynn 2001). Hence, we may expect that for a bounded $Q_0(i)$, $(Q_T(i) - Q_0(i))$ is $O(\log N)$ with large probability, for $i \in \mathcal{F} - \mathcal{C}$. Note that $Q_T(t) - Q_0(t) = N - 1$. Thus, it is reasonable to expect that $L_T \approx c_t^{-N+o(N)}$ so that the large deviations lower bound result holds. The proof is made rigorous in Appendix A.

Remark 1 (When feeder queues have finite small buffers) In this setting, a customer arriving to a feeder queue with a full buffer, leaves the system. Note that the proof of the lower bound is unaffected if each feeder queue i has a finite buffer $\Theta(N^{1/2+\epsilon_i})$ for $\epsilon_i > 0$. Also note that Assumption 2 holds for all π when the buffers at feeder queues are $o(N)$.

We now argue that the proof of the upper bound holds even when the feeder queues have finite buffers. Note that along a realization $(X_0, X_1, \ldots, X_T)$ to the rare set, the transitions may be divided into two categories: Those belonging to customers that had to leave the system due to overflow at a feeder queue, and those belonging to others who did not. Thus, L_T may be expressed as the product of the likelihood ratio corresponding to the transitions in Category 1 ($L_T^{(1)}$) and that corresponding to the transitions in Category 2 ($L_T^{(2)}$).

Suppose $N_T(i)$ is redefined to denote the number of external arrivals to queue i up to time T, that belong to Category 2, and $M_T(i,j)$ denotes the number of service completions from queue i to queue j (or to outside the network if $j = e$) up to time T from customers from the second category. Then it is easy to see that RHS in (6) upper bound $L_T^{(2)}$ and that (13) holds under the modified definitions. Thus, the proof of the upper bound holds, if we can show that $L_T^{(1)} \leq 1$.

To see this, consider an arbitrary customer along a path to the rare event that arrives at queue i_0 and is routed to the queues $i_1, \ldots, i_m$ before having to leave the system as the feeder queue i_m ($m \geq 0$) is full. The contributions to the likelihood ratio $L_T^{(1)}$ due to its transitions is upper bounded by

$$\frac{\lambda_{i_0}}{\lambda_{i_0}^*} \frac{\mu_{i_0} p_{i_0 i_1}}{\mu_{i_0}^* p_{i_0 i_1}^*} \cdots \frac{\mu_{i_{m-1}} p_{i_{m-1} i_m}}{\mu_{i_{m-1}}^* p_{i_{m-1} i_m}^*}.$$

Plugging in the values from Assumption 1, this can be seen to equal $\frac{1}{c_m}$ which is upper bounded by 1. Thus, $L_T^{(1)} \leq 1$ and (6) holds in this set-up as well.

In particular, it follows that when the buffers at each feeder queue i is $\Theta(N^{1/2+\epsilon_i})$ for $0 < \epsilon_i < 1/2$, under Assumption 1, (9) and (10) hold for any initial distribution π.

Also consider the case where the feeder queue buffer sizes are finite and $\Theta(N)$ or larger. Again, the proof of the lower bound is unaffected by this restriction. The proof of the upper bound follows from the above discussion (Assumption 2 has to be shown in this case, it is not automatically satisfied as when the buffers at the feeder queues are of size $o(N)$).

Proof of Theorem 2 From (15) it follows that on $\mathcal{B}_N$,

$$L_T \leq (\frac{1}{c_t})^N \left(\prod_{i \in \mathcal{F}} c_i^{Q_0(i)} \right) \tag{18}$$

a.s. Since L_T is non-negative,

$$L_T^{2-k} \leq \max(1, L_T) \leq 1 + L_T \quad a.s.$$

Therefore,

$$E_{\mathbf{P}_\pi^*}(L_T^2 I(\mathcal{B}_N)) \leq E_{\mathbf{P}_\pi^*}(L_T^k I(\mathcal{B}_N)) + E_{\mathbf{P}_\pi^*}(L_T^{k+1} I(\mathcal{B}_N)). \tag{19}$$

However, since $E_{\mathbf{P}_\pi^*}(L_T^{k+1} I(\mathcal{B}_N)) = E_{\mathbf{P}_\pi}(L_T^k I(\mathcal{B}_N))$, (19) and (18) imply that

$$E_{\mathbf{P}_\pi^*}(L_T^2 I(\mathcal{B}_N)) \leq (\frac{1}{c_t})^{k(N-1)} \left(E_{\mathbf{P}_\pi^*}(\prod_{i \in \mathcal{F}} c_i^{kQ_0(i)}) + E_{\mathbf{P}_\pi}(\prod_{i \in \mathcal{F}} c_i^{kQ_0(i)}) \right).$$

The result follows by noting (11) and that each $Q_0(i)$ has same distribution under $\mathbf{P}_\pi^*$ and $\mathbf{P}_\pi$. □

If $\mathbf{P}_\pi^*$ is set equal to $\mathbf{P}_\pi$ (i.e., under naive simulation), then $L_T = 1$ a.s. and

$$\lim_{N \to \infty} \frac{1}{N} \log E_{\mathbf{P}_\pi}(L_T^2 I(\mathcal{B}_N)) = -\log c_t.$$

Thus, if Theorem 2 holds for some $1 < k \leq 2$, then asymptotically one may expect exponential improvement (i.e., asymptotic gain) over naive simulation (asymptotic efficiency holds if $k = 2$).

2.4 Motivation for the Proposed Change of Measure

We briefly motivate $\mathbf{P}_\pi^*$ as this may be useful in future extensions. The fact that the feeder queues are either stable or critical in $\mathbf{P}_\pi^*$ is motivated by observations in Chang et al. (1994) and Kroese and Nicola (2002). Chang et al. (1994) consider a special class of queuing networks referred to as the intree networks. These intree networks have deterministic routing and no feedback. As in our model, they divide the network into a target queue that is located at the root and other feeder queues which are like leafs of a tree feeding the target queue. They also focus on estimating the probability that the buffer at the target queue overflows during its busy period. They propose a change of measure under which the feeder queues are either stable or critical and the target queue is unstable and they report a large amount of variance reduction compared to naive simulation for a few examples (also see, L'Ecuyer and Champoux 1996, 2001 where they empirically study the measure proposed by Chang et al. 1994 on larger and more general networks and propose many refinements).

Kroese and Nicola (2002) consider a two-node tandem Jackson network where the second node is the target queue of interest. They propose a state-dependent change of measure (depending on the queue length at the first buffer) to estimate the probability that the target queue buffer overflows during its busy period and prove its effectiveness under a Type-I distribution. Again, under their proposed change of measure, either the first queue remains stable and its service rate unchanged or the first queue is critical (i.e., its service rate is altered to become equal to its new arrival rate).

The specific form of the proposed change of measure can be derived by imposing the 'sample path independence' (SPI) property proposed in Juneja (2001). A change of measure satisfies the SPI property if between any two states along the most likely paths to the rare event the likelihood ratio is sample path independent (likelihood ratio here is simply the ratio of the original and the new probability of the path between the two states). In our setting, this is equivalent to the likelihood ratio being equal to one on 'cycles' along most likely paths to the rare event (cycle is a path with identical first and last state). It can be seen that SPI property holds for the proposed change of measure along the set $\tilde{\mathcal{B}}_N$, i.e., the set of paths to the rare set in which the critical queues remain busy.

3 Continuous Time Analysis

In this section we use the well known product form steady state distributions to aid in developing the large deviations limit for $P_\pi(\mathcal{B}_N)$ when π corresponds to Type-II initial distribution. Whenever the queueing process moves from the set $\mathcal{S}_N^c$ to $\mathcal{S}_N$ (for any set A, A^c denotes its complement) we say that a level has been crossed. In this section, we also develop a ratio representation for estimating the mean time between level crossings in the target queue (MTBL). This proves useful in developing an effective importance sampling based simulation methodology to estimate MTBL.

3.1 Large Deviations Limit under Type-II Initial Distribution

Recall that $\tilde{\mathbf{Q}} = (\tilde{Q}_s : s \geq 0)$ denotes the CTMC associated with the Jackson network. For each k, let α_k denote the k th time an arrival to the target queue finds it empty. Note that the Type-II initial distribution corresponds to the steady state distribution associated with the discrete time Markov chain $(\tilde{Q}_{\alpha_k} : k \geq 0)$. Let ψ denote this distribution. To emphasize the fact that the CTMC may not regenerate at the stopping times $(\alpha_k : k \geq 0)$ we refer to them as *pseudo-regeneration times* and the process $\tilde{\mathbf{Q}}$ between two successive pseudo-regeneration times as a *pseudo-regeneration cycle*. Let $\phi = (\phi(x) : x \in \mathcal{Z}_+^{|\mathcal{H}|})$ denote the stationary distribution associated with the CTMC $\tilde{\mathbf{Q}}$. Then, for $x = (x_i \in \mathcal{Z}_+ : i \in \mathcal{H})$, the following product form is well-known:

$$\phi(x) = \prod_{i \in \mathcal{H}} (\frac{\gamma_i}{\mu_i})^{x_i}(1 - \frac{\gamma_i}{\mu_i}).$$

Proposition 1 *For a stable Jackson network,*

$$\lim_{N\to\infty} \frac{\log \mathbf{P}_\psi(\mathcal{B}_N)}{N} = -\log \frac{\mu_t}{\gamma_t}. \qquad (20)$$

Let $\phi(A) = \sum_{x \in A} \phi(x)$ for any $A \subset \mathcal{S}$.

Let τ denote the length of a pseudo-regeneration cycle. Note that the following ratio representation holds for the steady state measure ϕ when the network is stable (see, e.g., Cogburn 1975):

$$\phi(A) = \frac{E_{\mathbf{P}_\psi}(\int_{s \leq \tau} I(\tilde{Q}_s \in A) ds)}{E_{\mathbf{P}_\psi}(\tau)}.$$

Let $D = \int_{s \leq \tau} I(\tilde{Q}_s \in \mathcal{S}_N) ds$. Also note that $\phi(\mathcal{S}_N) = (\frac{\gamma_t}{\mu_t})^N$, so that

$$\frac{\mathbf{P}_\psi(\mathcal{B}_N) E_{\mathbf{P}_\psi}(D|\mathcal{B}_N)}{E_{\mathbf{P}_\psi}(\tau)} = (\frac{\gamma_t}{\mu_t})^N.$$

Note that $E_{\mathbf{P}_\psi}(D|\mathcal{B}_N)$ is bounded from below by at least one service time in the target queue. Thus, the large deviations upper bound

$$\lim_{N \to \infty} \frac{\log(\mathbf{P}_\psi(\mathcal{B}_N))}{N} \leq -\log \frac{\mu_t}{\gamma_t},$$

holds. The proof of the lower bound requires more steps and is given in the Appendix B.

3.2 Mean Time between Target Queue Level Crossings

We now develop a ratio representation for estimating MTBL along the lines suggested in Glynn et al. (1993) in the reliability settings. Let $\tilde{N}_l(s)$ denote the number of level crossings in the target queue up to time s. Then, in a stable Jackson network there exists a constant $\zeta > 0$ such that

$$\lim_{s \to \infty} \frac{s}{\tilde{N}_l(s)} = \zeta, \quad a.s.$$

The quantity ζ is defined to be MTBL. Let $\tilde{N}_\alpha(s)$ denote the number of pseudo-regeneration cycles completed till time s. It is easy to see that

$$\zeta = \lim_{s \to \infty} \frac{s}{\tilde{N}_l(s)} = \lim_{s \to \infty} \frac{s/\tilde{N}_\alpha(s)}{\tilde{N}_l(s)/\tilde{N}_\alpha(s)} = \frac{E_{\mathbf{P}_\psi}(\tau)}{E_{\mathbf{P}_\psi}(\tilde{M})}, \quad a.s.,$$

where, (recall that) τ denotes the length of a pseudo-regeneration cycle, $\tilde{M}$ denotes the number of target queue level crossings in such a cycle (see Glynn et. al. 1993).

In practice, the numerator $E_{\mathbf{P}_\psi}(\tau)$ can be estimated cheaply via naive simulation as no rare events are involved. However, the denominator $E_{\mathbf{P}_\psi}(\tilde{M})$ is difficult to accurately estimate via naive simulation as the target queue has to cross the level N in a cycle for $\tilde{M}$ to have a positive value. Also, because of its small value, its accurate estimation is critical to accurate estimation of MTBL. Note that $E_{\mathbf{P}_\psi}(\tilde{M}) = \mathbf{P}_\psi(\mathcal{B}_N) E_{\mathbf{P}_\psi}(\tilde{M}|\mathcal{B}_N)$, so that if we can sample from the conditional distribution, $E_{\mathbf{P}_\psi}(\tilde{M}|\mathcal{B}_N)$ may be easy to estimate via naive simulation and the problem of efficient estimation of $E_{\mathbf{P}_\psi}(\tilde{M})$ reduces to that of efficient estimation of $\mathbf{P}_\psi(\mathcal{B}_N)$. This suggests that the change of measure that asymptotically efficiently estimates $\mathbf{P}_\psi(\mathcal{B}_N)$ may be used to efficiently estimate $E_{\mathbf{P}_\psi}(\tilde{M})$ as well. In practice, we estimate $E_{\mathbf{P}_\psi}(\tilde{M})$ by simulating the embedded DTMC $(Q_n : n \geq 0)$. Then, we may re-express

$$E_{\mathbf{P}_\psi}(\tilde{M}) = E_{\tilde{\mathbf{P}}_\psi}(L_T * \tilde{M}),$$

where $\tilde{\mathbf{P}}_\psi$ denotes a measure under which the transition probabilities correspond to the transition probabilities P^* until time $T = \min(T_N, T_o)$, and if $T_N < T_o$, then between time T_N and T_o they correspond to the transition probabilities P. Glynn et. al. (1993) observed in an analogous setting that if $\mathbf{P}^*_\psi$ estimates $\mathbf{P}_\psi(\mathcal{B}_N)$ efficiently, then $\tilde{\mathbf{P}}_\psi$ described above, estimates $E_{\tilde{\mathbf{P}}_\psi}(L_T * \tilde{M})$ efficiently. We verify this on a queueing network example in Section 6. The details of the simulation methodology used to estimate MTBL are along the lines proposed by Glynn et. al. (1993) and are outlined in Section 6.4.

4 An Explicit and Asymptotically Efficient Measure

In this section we give sufficient conditions for the existence of the proposed change of measure in a simple explicit form. The proposed change of measure may exist under much less restrictive conditions, however, not in such an explicit form. In general settings it may be determined using the heuristic outlined in Section 6.1.

4.1 Feeder Service Rates are Sufficiently Large

Recall that $R = (r_{ij} : i, j \in \mathcal{H})$ equals $(I - P)^{-1}$. Since the network is stable, each r_{ij} equals the expected number of visits to queue j by a customer starting from queue i, before it leaves the system. Note that $r_{it} \leq r_{tt}$.

Proposition 2 *Suppose, for each $i \in \mathcal{F}$ the service rates at the feeder queues satisfy the inequality*

$$\mu_i > \gamma_i(1 + \frac{r_{it}}{r_{tt}}(\frac{\mu_t}{\gamma_t} - 1)). \tag{21}$$

Then, there exists a $\mathbf{P}^*$ that satisfies Assumption 1 with $\mathcal{C}$ empty and has the following parameters:

$$\mu_t^* = \frac{(r_{tt} - 1)\mu_t + \gamma_t}{r_{tt}}. \tag{22}$$

For each $i \in \mathcal{H}$,

$$c_i = (1 + \frac{r_{it}}{r_{tt}}(\frac{\mu_t}{\gamma_t} - 1)). \tag{23}$$

For $i \in \mathcal{F}$ and $j \in \mathcal{H}$,

$$p_{ij}^* = \frac{c_j}{c_i} p_{ij}, \tag{24}$$

and

$$p_{ie}^* = \frac{1}{c_i} p_{ie}. \tag{25}$$

Also, for $j \in \mathcal{H}$,

$$p_{tj}^* = \frac{c_j}{c_t} \frac{\mu_t}{\mu_t^*} p_{tj} = \frac{c_j \gamma_t}{\mu_t^*} p_{tj}, \tag{26}$$

and

$$p_{te}^* = \frac{1}{c_t} \frac{\mu_t}{\mu_t^*} p_{te} = \frac{\gamma_t}{\mu_t^*} p_{te}. \tag{27}$$

For each $i \in \mathcal{H}$, $\lambda_i^* = c_i \lambda_i$ and

$$\gamma_i^* = c_i \gamma_i. \tag{28}$$

Example 1 To see such a $\mathbf{P}^*$ in a simple setting, consider a stable network of queues in tandem and suppose that the target queue is the last queue in this network (if there were queues behind the target queue, they may be ignored as they play no role in the target queue build-up). Arrivals occur at the first queue at rate $\lambda_1 > 0$. After getting served at queue 1 they go to queue 2 and so on. There are no external arrivals at any queue except the first. Suppose that the target queue is the bottleneck, i.e., $\mu_t < \mu_i$ for $i \in \mathcal{F}$. Since, $r_{it} = 1$, and $\gamma_i = \lambda_1$ for all $i \in \mathcal{H}$, the inequality (21) holds. Then, $\mathbf{P}^*$ exists with $\mathcal{C}$ empty and, $\mu_t^* = \lambda_1$, $c_i = \frac{\mu_t}{\lambda_1}$, and $\gamma_i^* = \mu_t$ for $i \in \mathcal{H}$, since $\lambda_1^* = \mu_t$. The routing probabilities remain unaffected.

Proof of Proposition 2 : It is easily seen that $\mu_t^* < \mu_t$ and that $c_i > 1$ for $i \in \mathcal{H}$. Also, it follows from (21) that $\mu_i > c_i\gamma_i = \gamma_i^*$, and hence, $\mathcal{C}$ is empty. Thus, in Assumption 1, all that we need to show is that the above specified parameters satisfy (1), the equalities

$$\sum_{j\in\mathcal{H}} p_{ij}^* + p_{ie}^* = 1, \tag{29}$$

and,

$$\gamma_i^* = \lambda_i^* + \sum_{j\in\mathcal{F}} p_{ji}^*\gamma_j^* + \mu_t^* p_{ti}^*, \tag{30}$$

for $i \in \mathcal{H}$. To see that (1) holds, in view of (23) and the fact that $\lambda_i^* = \lambda_i c_i$, all we need to show is that

$$\sum_{i\in\mathcal{H}} \lambda_i \frac{r_{it}}{r_{tt}}(\frac{\mu_t}{\gamma_t} - 1) + (\mu_t^* - \mu_t) = 0.$$

This is easily seen by noting that $\sum_{i\in\mathcal{H}} \lambda_i r_{it} = \gamma_t$ and by substituting for μ_t^* from (22) in the above equation.

Consider (29) for $i \in \mathcal{F}$. In view of (24) and (25) it suffices to show that

$$\sum_{j\in\mathcal{H}} p_{ij}c_j + p_{ie} = c_i.$$

Again, substituting for c_j from (23), this equation reduces to

$$1 + \frac{1}{r_{tt}}(\frac{\mu_t}{\gamma_t} - 1)\sum_{j\in\mathcal{H}} p_{ij}r_{jt} = 1 + \frac{r_{it}}{r_{tt}}(\frac{\mu_t}{\gamma_t} - 1),$$

which is true as $\sum_{j\in\mathcal{H}} p_{ij}r_{jt} = r_{it}$.

To see (29) for $i = t$, note that $c_t = \frac{\mu_t}{\gamma_t}$. Thus, it suffices to show that

$$\sum_{j\in\mathcal{H}} p_{tj}c_j + p_{te} = \frac{\mu_t^*}{\gamma_t} = (\frac{r_{tt} - 1}{r_{tt}})\frac{\mu_t}{\gamma_t} + \frac{1}{r_{tt}}, \tag{31}$$

where the RHS follows from (22). In the LHS plug in the expression for c_j from (23) to get

$$1 + \frac{1}{r_{tt}}(\frac{\mu_t}{\gamma_t} - 1)\sum_{j\in\mathcal{H}} p_{tj}r_{jt}.$$

Noting that $r_{tt} = 1 + \sum_{j\in\mathcal{H}} p_{tj}r_{jt}$ in the above equation, its equivalence to RHS in (31) is easily seen.

To see that (30) is true, substitute values to re-express it as:

$$c_i\gamma_i = c_i\lambda_i + \sum_{j\in\mathcal{F}} c_i p_{ji}\gamma_j + c_i\gamma_t p_{ti}.$$

The truth of this equation follows from the original traffic equations. □

Remark 2 Suppose that the service rate μ_i at each feeder queue i is large enough so that the target queue is the bottleneck, i.e., for each $i \in \mathcal{F}$

$$\gamma_t/\mu_t > \gamma_i/\mu_i.$$

This network satisfies the condition of Proposition 2, i.e., γ_i^* is less than μ_i for each $i \in \mathcal{F}$. To see this, note that for $i \in \mathcal{F}$, $r_{it} \leq r_{tt}$. Then,

$$\mu_i > \gamma_i\frac{\mu_t}{\gamma_t} \geq \gamma_i(1 + \frac{r_{it}}{r_{tt}}(\frac{\mu_t}{\gamma_t} - 1)) = \gamma_i^*.$$

It can be checked that when the target queue is the bottleneck, $\mathbf{P}^*$ specified in Proposition 2 agrees with the change of measure proposed by Parekh and Walrand (1989), Frater et al. (1991) in the Jackson network setting. They focus on estimating the probability that the total network population exceeds a specified buffer in the network's busy cycle (i.e., between consecutive times that the complete network is empty). It has been shown that when $\mathbf{P}^*$ is applied to estimate this probability, the resulting estimator may have large (even infinite variance) for certain parameters (see Glasserman and Kou 1995, Randhawa and Juneja 2004).

However, in our analysis the same change of measure is applied to estimate the probability that the bottleneck queue overflows in its busy cycle, the resulting estimator is asymptotically efficient under Type-I initial distribution. It always has a finite variance and is asymptotically efficient for a large set of parameters under Type-II initial distribution (discussed in Section 4.2). The reason for this is that to estimate the latter probability (of overflow at the bottleneck queue), importance sampling is used to generate sample paths where the bottleneck queue always remains busy. While in the estimation of the former probability, importance sampling is applied along the overall regenerative cycle that includes paths where the bottleneck queue may be empty numerous times. For certain parameters, $\mathbf{P}^*$ may assign little probability to such paths compared to original measure, leading to variance increase.

4.2 Asymptotic Efficiency under Type-II Initial Distribution

Recall that ψ denotes the Type-II initial distribution.

Theorem 3 *Suppose that for each $i \in \mathcal{F}$*

$$\mu_i > \gamma_i \left(1 + \frac{r_{it}}{r_{tt}}(\frac{\mu_t}{\gamma_t} - 1)\right)^2. \tag{32}$$

Then, Proposition 2 holds and $\mathbf{P}^*_\psi$ specified in the Proposition satisfies Assumption 2. Thus, Theorem 1 holds and $\mathbf{P}^*_\psi$ is asymptotically efficient for estimating $\mathbf{P}_\psi(\mathcal{B}_N)$.

Proof: Clearly, Proposition 2 holds. From arrival theorems (see, e.g., Walrand 1988) it is easy to infer that in steady state, arrivals to the target queue that see it empty see the remaining queues in their steady state product form distribution. That is, under ψ, the feeder queue lengths have their steady state product form distribution and thus are mutually independent. This leads to simplified analysis of Assumption 2. Since $Q_T(i) \geq 0$ and $(Q_0(i) : i \in \mathcal{F})$ have the same distribution under $\mathbf{P}^*_\psi$ and $\mathbf{P}_\psi$ and are mutually independent under these measures, it follows that:

$$\begin{aligned} E_{\mathbf{P}^*_\psi}\left(\prod_{i\in\mathcal{F}} c_i^{2(Q_0(i)-Q_T(i))}\right) &\leq E_{\mathbf{P}^*_\psi}\left(\prod_{i\in\mathcal{F}} c_i^{2Q_0(i)}\right) \\ &= \Pi_{i\in\mathcal{F}}\, E_{\mathbf{P}_\psi}(c_i^{2Q_0(i)}). \end{aligned}$$

Now, $\mathbf{P}_\psi(Q_0(i) = n) = (\frac{\gamma_i}{\mu_i})^n\ (1 - \frac{\gamma_i}{\mu_i})$. Since Proposition 2 holds, $c_i = \frac{\gamma_i^*}{\gamma_i}$. Then,

$$E_{\mathbf{P}_\psi}(c_i^{2Q_0(i)}) = \sum_{n=0}^{\infty}(\frac{\gamma_i^*}{\gamma_i})^{2n}\ (\frac{\gamma_i}{\mu_i})^n\ (1 - \frac{\gamma_i}{\mu_i}).$$

This is finite iff

$$\frac{(\gamma_i^*)^2}{\gamma_i\mu_i} < 1 \tag{33}$$

for each $i \in \mathcal{F}$. The result follows from (32), (28) and (23). □

Suppose that there exists a $1 < k_i \leq 2$ for $i \in \mathcal{F}$ such that

$$\mu_i > \gamma_i \left(1 + \frac{r_{it}}{r_{tt}}(\frac{\mu_t}{\gamma_t} - 1)\right)^{k_i}. \tag{34}$$

Let $k = \min_{i \in \mathcal{F}} k_i$. Clearly, in this case Proposition 2 holds. From analysis in proof of Theorem 3, it follows that for all $\tilde{k} < k$,

$$\log E_{\mathbf{P}^*_\psi} \left(\prod_{i \in \mathcal{F}} c_i^{\tilde{k} Q_0(i)} \right) < \infty. \tag{35}$$

Then, from Theorems 1 and 2 it is easily seen that

$$\lim_{N \to \infty} \frac{\log(\mathbf{P}_\psi(\mathcal{B}_N))}{N} = -\log c_t \tag{36}$$

and

$$\limsup_{N \to \infty} \frac{\log(E_{\mathbf{P}^*_\psi}(L_T^2 I(\mathcal{B}_N)))}{N} \leq -k \log c_t. \tag{37}$$

Thus, in such cases we may expect our estimator to perform much better than naive simulation for large values of N. Also note that when Proposition 2 holds, $c_t = \frac{\mu_t}{\gamma_t}$ so that (36) agrees with (20). Of course, Proposition 1 shows that (36) holds for stable Jackson networks, even if the service rates at feeder queues do not satisfy (34) or even (21).

Example 2 Again consider Example 1 with the tandem network having only two queues, where the first queue (indexed by 1) is the feeder queue and the second queue is the target queue (indexed by t). First, assume that the second queue is the bottleneck; i.e., $\lambda_1 < \mu_t < \mu_1$. As discussed in Example 1, the conditions of Proposition 2 hold and the proposed change of measure has $\mathcal{C}$ empty and $\lambda_1^* = \gamma_1^* = \gamma_2^* = \mu_t$, $\mu_1^* = \mu_1$, $\mu_t^* = \lambda_1$ and $c_1 = c_t = \frac{\mu_t}{\lambda_1}$. From Theorem 3 it follows that under Type-II initial distribution, the proposed estimator is asymptotically efficient when $\mu_1 > \frac{\mu_t^2}{\lambda_1}$. Furthermore, from (37), it follows that for $\mu_t < \mu_1 \leq \frac{\mu_t^2}{\lambda_1}$, a large variance reduction over naive simulation may be expected as N becomes large. Note that the decay rate of the probability $\mathbf{P}_\psi(\mathcal{B}_N)$ equals $\log \frac{\lambda_1}{\mu_t}$. Of course, in this case the exact value of $\mathbf{P}_\psi(\mathcal{B}_N)$ can be determined as when the queue length process in the first queue is at steady-state, the arrival process to the second queue is Poisson with rate λ_1. Thus, in steady state the second queue is an $M/M/1$ queue with arrival rate λ_1 and service rate μ_t. It is well known that the probability of hitting level N in this queue's busy cycle equals

$$\frac{\mu_t/\lambda_1 - 1}{(\mu_t/\lambda_1)^N - 1}. \tag{38}$$

Now consider the case where the first queue is the bottleneck; i.e., $\lambda_1 < \mu_1 < \mu_t$. In this case, Assumption 1 is satisfied with $\mathcal{C}$ consisting of the first queue, $\lambda_1^* = \gamma_1^* = \gamma_t^* = \mu_1^*$ and λ_1^* and μ_t^* are the unique solution to the two equations

$$\frac{\lambda_1 \mu_1 \mu_t}{\lambda_1^{*2} \mu_t^*} = 1 \tag{39}$$

and

$$\lambda_1 + \mu_1 + \mu_t = 2\lambda_1^* + \mu_t^*$$

such that $\lambda_1^* > \lambda_1$ (equivalently, $\mu_t^* < \mu_t$). Here $c_1 = \frac{\lambda_1^*}{\lambda_1}$ and $c_t = \frac{\lambda_1^{*2}}{\lambda_1 \mu_1}$. It is easy to check that under this measure, the likelihood ratio of any cycle along *any* path where both the queues do not empty, equals 1 (e.g., consider cycle $(5,5),(6,5),(5,6),(5,5)$. Here the likelihood of the three transitions equals LHS of (39)), i.e., the SPI property (discussed in Section 2.4) holds.

To see that (8) (and hence Assumption 2) does not hold in this case under Type-II distribution, note that if (8) were true, the large deviation limit would equal $-\log c_t$ and this has to coincide with the large deviations limit for the probability (38), i.e., $\log \frac{\lambda_1}{\mu_t}$. This, however is untrue as $c_t = \frac{\lambda_1^{*2}}{\lambda_1 \mu_1}$ can be seen to differ from $\frac{\mu_t}{\lambda_1}$. We discuss this case further in the next section.

5 Related Developments in Large Deviations Theory

The large deviations approach to analyze rare events in networks such as the Jackson networks, typically considers the large deviations behavior of the sequence of scaled processes $(\hat{\mathbf{Q}}^{\mathbf{n}} : n \geq 1)$ associated with the network, where $\hat{\mathbf{Q}}^{\mathbf{n}} = (\hat{Q}_s^n : s \geq 0)$ and

$$\hat{Q}_s^n = \frac{1}{n} \tilde{Q}_{ns}.$$

Note that in a stable Jackson network with $\tilde{Q}_0 = 0$, $\hat{Q}_s^n$ converges to zero a.s. for any $s \geq 0$ as $n \to \infty$. Significant large deviations literature focusses on identifying the most likely paths along which the process $(\hat{Q}_s^n : s \geq 0)$ hits a set not containing the origin in the non-negative orthant $\Re_+^{|\mathcal{H}|}$ asymptotically as $n \to \infty$ (see, e.g., Ignatyuk et al. 1994, Dupuis and Ellis 1995, Atar and Dupuis 1999, Dupuis and Ramanan 2002). Ignatyuk et al. (1994) identify the most likely paths along which queues build up in scaled two queue networks modelled as discrete-time Markov chains. In particular, they show that these paths may be piece-wise linear. Avram et al. (2000) show this in the setting of semi-martingale reflected Brownian motion, which typically provide a good approximation for heavily loaded queueing networks. Roughly speaking, the 'most likely paths' in the scaled network may be interpreted as the most likely arrival and service rates, and the routing probabilities observed by the unscaled process conditioned on the occurrence of the rare event. Ignatiouk-Robert (2000) conducts a comprehensive analysis of the large deviations behavior of Jackson networks. In particular, she proposes a probabilistic method to analyze these networks, and using it develops an explicit form of the large deviations rate function associated with the process $(\hat{Q}^n : n \geq 1)$. Further, she also designs an algorithm to get its closed form expression.

As mentioned in the Introduction, the fact that paths to rare events in Jackson networks are piece-wise linear also follows from the work in Anantharam, Heidelberger and Tsoucas (1990). They show that starting from an empty network, the most likely paths to the rare event associated with queue lengths in Jackson networks, corresponds to the most likely path (in reverse direction) followed by the reversed Jackson network starting from the rare set till it empties (also see Frater, Lennon and Anderson 1991).

It is well known that a good change of measure assigns large probability to the most likely paths to the rare event (see, e.g., Heidelberger 1995). The analysis in Ignatyuk et al. (1994) and Anantharam, Heidelberger and Tsoucas (1990) is useful in explaining why the proposed change of measure may succeed or fail in estimating the buffer overflow probability under Type-II initial distribution. Again consider a two-queue tandem Jackson network discussed in Example 2, and suppose that the second queue is the bottleneck, i.e., $\lambda_1 < \mu_t < \mu_1$. From the analysis in these papers, (see, e.g., Theorem 3.6.3, Ignatyuk et al. 1994), it can be inferred that for the scaled network to reach the state $(0, x_t)$, $x_t > 0$, from the origin, involves a single path corresponding to an arrival rate of μ_t, service rate of μ_1 at queue 1 and service rate of λ_1 at queue t, in the unscaled network. Our proposed change of measure duly emphasizes these paths and successfully simulates this network.

When the first queue is the bottleneck, i.e., $\lambda_1 < \mu_1 < \mu_t$, from their analysis it can be inferred that to reach the state $(0, x_t)$, $x_t > 0$, from the origin, involves two linear component paths. Along the first component, the arrival rate to queue 1 equals μ_1, the service rate at queue 1 equals λ_1 and the service rate at queue t remains the same at μ_t. Queue 1 builds up along this path till it reaches the level

$$\frac{\mu_t - \mu_1}{\mu_t - \lambda_1} x_t.$$

Thereafter, along the second component, the arrival rate to queue 1 equals μ_1, the service rate at queue 1 equals μ_t and the service rate at queue t equals λ_1, so that now queue 1 empties as queue t builds up till the state $(0, x_t)$ is reached. This suggests that a good importance sampling change of measure should have one set of rates till the first queue builds up to an appropriate level and another set of rates with which the first queue empties and the second queue builds up till the large level threshold is hit. More generally, this suggests that in general Jackson and other networks, a change of measure that is appropriately piecewise constant ought to be used. Obviously, the proposed change of measure (where the first queue is critical and the second queue is unstable) assigns a constant set of rates and does not sufficiently emphasize the most likely paths to the rare event. This explains why in our simulations with this tandem queue example, under Type-II initial distribution, the proposed change of measure is not very efficient.

6 Experimental Results

In this section, the theoretical results obtained in this paper are supported by means of simulation experiments on simple Jackson networks. We first propose a heuristic non-linear program (NLP) to determine the change of measure described in Assumption 1. We offer no proof of its validity. However, we make a positive note that in our experiments the proposed heuristic always gave a solution that satisfies Assumption 1, almost instantaneously using a standard excel solver software. This was the case even when the service rates are sufficiently large at feeder queues (so that an explicit representation of the associated parameters is known). We experimented with networks having up to five queues (see Example 4; here, the excel solver took less than 0.2 seconds to solve the NLP). For the two-queue tandem network example, the change of measure for different parameter values also follows from the discussion in Example 2.

All experiments are performed on a Sun Blade 1000 running Solaris-8 on a Sparcv9 750 MHz processor with 1,024 MB RAM. In our experiments we also compare the gain in efficiency from using importance sampling over using brute force naive simulation. For this purpose, we define the computation reduction factor (CRF) as the ratio of the computational effort required under naive simulation and the computation effort required under IS to achieve a specified relative error. Suppose that σ^2 ($\hat{\sigma}^2$) denotes the variance of the naive (IS) estimator and τ ($\hat{\tau}$) denotes the expected computation effort required to generate one sample under naive (IS) simulation. Then, it is well known that when independent samples are generated in both the cases, and their number is large, CRF approximately equals

$$\frac{\sigma^2 \tau}{\hat{\sigma}^2 \hat{\tau}}$$

for any specified level of relative error (see, e.g., Glynn and Whitt 1992). In our simulations, we estimate σ^2, $\hat{\sigma}^2$, τ and $\hat{\tau}$ and report the estimated CRF. In our settings, $\mathbf{P}_\pi(\mathcal{B}_N)$, is small, it is computationally costly and even prohibitive to estimate the variance of the indicator $I(\mathcal{B}_N)$ (i.e., $\mathbf{P}_\pi(\mathcal{B}_N)(1-\mathbf{P}_\pi(\mathcal{B}_N))$), using naive simulation. To overcome this drawback, we use the IS estimator of the probability, call it, $\hat{p}$, and set the estimator of σ^2 as $\hat{p}(1-\hat{p})$. Note that in this case, the parameter τ is easily estimated via simulation.

6.1 Determining the Proposed Change of Measure

The notation with an over-line denote the variables of the NLP. The objective is simply to maximize $\bar{c}_t$ subject to the following constraints:

$$\sum_{i\in\mathcal{H}-\mathcal{D}} \bar{\lambda}_i + \sum_{i\in\mathcal{H}} \bar{\mu}_i = \sum_{i\in\mathcal{H}-\mathcal{D}} \lambda_i + \sum_{i\in\mathcal{H}} \mu_i.$$

Each $\bar{c}_i = \frac{\bar{\lambda}_i}{\lambda_i}$ for $i \in \mathcal{H}-\mathcal{D}$, and $\bar{c}_i \geq 1$ for $i \in \mathcal{H}$. For all $i,j \in \mathcal{H}$,

$$\bar{p}_{ij} = \frac{\bar{c}_j}{\bar{c}_i}\frac{\mu_i}{\bar{\mu}_i}p_{ij},$$

$$\bar{p}_{ie} = \frac{1}{\bar{c}_i}\frac{\mu_i}{\bar{\mu}_i}p_{ie},$$

and for all $i \in \mathcal{H}$,

$$\sum_{j\in\mathcal{H}} \bar{p}_{ij} + \bar{p}_{ie} = 1.$$

The variables $(\bar{\gamma}_i : i \in \mathcal{H})$ satisfy the relations

$$\bar{\gamma}_i = \bar{\lambda}_i + \sum_{j\in\mathcal{F}} \bar{p}_{ji}\bar{\gamma}_j + \bar{\mu}_t\bar{p}_{ti}$$

and for all $i \in \mathcal{F}$

$$\bar{\gamma}_i \leq \bar{\mu}_i.$$

In addition, $\bar{\mu}_i \geq \mu_i$ for all $i \in \mathcal{F}$ and $\bar{\mu}_t \leq \mu_t$. The set $\mathcal{C}$, then corresponds to the queues that are critical under the solution to the NLP.

Note that a feasible solution to the above set of constraints may not satisfy Assumption 1 as we may have $\bar{\gamma}_i < \bar{\mu}_i$ (so that the queue is not critical), while the service rate is changed, i.e., $\bar{\mu}_i > \mu_i$. To eliminate this possibility, we could further impose constraints $(\bar{\mu}_i - \bar{\gamma}_i) * (\bar{\mu}_i - \mu_i) = 0$ for $i \in \mathcal{F}$. This ensures that the service processes may be modified only in the critical feeder queues. However, in our experiments these constraints were automatically satisfied by the solution of the NLP that did not incorporate them.

6.2 Type-I Initial Distribution

Example 3 Consider a two-queue tandem Jackson network with $(\lambda_1, \mu_1, \mu_t) = (1, 2, 3)$. The second queue is the target queue. The initial distribution corresponds to having a single customer in each of the two queues and our interest is in estimating the probability that the second queue reaches level N before it empties. The proposed change of measure $(\lambda_1^*, \mu_1^*, \mu_t^*)$ is determined to be $(2.532, 2.532, 0.936)$. As in Kroese and Nicola (2002), where the same example is considered, we generated 10^6 runs for $N = 10, 25$ and 50. The results illustrating the efficiency gain through importance sampling are shown in Table 1. Note that a 95% relative confidence interval is simply the estimated probability ± twice the estimated relative error (RE).

N	Estimate	RE	$\hat{\tau}$ $(\times 10^{-6})$	τ $(\times 10^{-6})$	Est. CRF
10	5.31×10^{-6}	0.23%	72.78	5.64	2.76×10^{3}
25	4.61×10^{-14}	0.35%	187.35	5.65	5.34×10^{10}
50	4.33×10^{-27}	0.51%	556.09	5.65	9.11×10^{22}

Table 1: Two-queue tandem network with Type-I initial distribution described in Example 3. Here $\hat{\tau}$ and τ denote the average CPU time required to generate one sample under IS and naive simulation, respectively.

Example 4 Consider a larger Jackson network comprising 5 queues with a single server at each queue. The target queue is indexed by t and the remaining queues are indexed by 1, 2, 3 and 4. The external arrival rates $(\lambda_t, \ldots, \lambda_4)$ equal $(0.2, 0.8, 0.5, 0.3, 0.6)$. The service rates $(\mu_t, \ldots, \mu_4)$ equal $(1.3, 1.4, 2.4, 1.7, 0.9)$. The routing probability matrix P is given by

$$\begin{pmatrix} 0.20 & 0.06 & 0.03 & 0.08 & 0.05 \\ 0.05 & 0.20 & 0.05 & 0.15 & 0.05 \\ 0.05 & 0.08 & 0.05 & 0.03 & 0.10 \\ 0.07 & 0.10 & 0.04 & 0.03 & 0.10 \\ 0.02 & 0.04 & 0.05 & 0.02 & 0.08 \end{pmatrix}.$$

From the traffic equations, it follows that the total arrival rates to each queue $(\gamma_t, \ldots, \gamma_4)$ equal $(0.44, 1.22, 0.67, 0.57, 0.88)$ and the respective traffic intensities $(\rho_t, \ldots, \rho_4)$ equal

$$(0.34, 0.87, 0.28, 0.34, 0.975).$$

Under the new change of measure (determined by solving the heuristic NLP) the external arrival rates $(\lambda_t^*, \ldots, \lambda_4^*)$ equal $(0.5909, 0.9265, 0.5630, 0.3501, 0.6255)$, the service rates $(\mu_t^*, \ldots, \mu_4^*)$ equal $(0.6247, 1.4056, 2.4000, 1.7000, 0.9138)$, and the new routing probability matrix P^* is given by

$$\begin{pmatrix} 0.4162 & 0.0489 & 0.0238 & 0.0658 & 0.0367 \\ 0.1270 & 0.1992 & 0.0484 & 0.1505 & 0.0448 \\ 0.1312 & 0.0823 & 0.0500 & 0.0311 & 0.0926 \\ 0.1772 & 0.0993 & 0.0386 & 0.0300 & 0.0893 \\ 0.0558 & 0.0438 & 0.0532 & 0.0220 & 0.0788 \end{pmatrix}.$$

The new total arrival rates to each queue $(\gamma_t^*, \ldots, \gamma_4^*)$ equal $(1.30, 1.41, 0.76, 0.67, 0.91)$ and the respective traffic intensities $(\rho_t^*, \ldots, \rho_4^*)$ equal $(2.08, 1, 0.32, 0.39, 1)$. Again, Type-I initial distribution corresponds to having a single customer in each of the five queues. 10^5 independent replications are simulated. The results for N equal to 10, 25, and 50 are given in Table 2.

N	Estimate	RE	$\hat{\tau}$ $(\times 10^{-5})$	τ $(\times 10^{-5})$	Est. CRF
10	2.90×10^{-5}	0.29%	71.6	7.1	4.15×10^{3}
25	1.67×10^{-12}	0.34%	198.5	7.1	1.85×10^{10}
50	1.94×10^{-24}	0.40%	412.3	7.1	5.66×10^{21}

Table 2: 5-queue network with Type-I initial distribution described in Example 4. Here $\hat{\tau}$ and τ denote the average CPU time required to generate one sample under IS and naive simulation, respectively.

6.3 Type-II Initial Distribution

In this section we consider examples where the initial distribution corresponds to ψ, i.e., the instants when an arrival to the target queue finds it empty while the feeder queue lengths have a stationary probability distribution corresponding to such instants.

These initial states are generated by following the procedure outlined in Chang et al. (1994). The network is run under the original measure for a large period of time so that at the instants of arrivals to an empty target queue the feeder queues have approximately their associated stationary distribution (specifically, we start the network by having one customer in each queue and run it till the thousandth time an arrival to the target queue finds it empty). Thereafter, every time an arrival to the target queue finds it empty, we run two simulations from that state; one with the change of measure to obtain samples of $L_T * I(\mathcal{B}_N)$, and the other with the original measure to generate the next sample of the initial state (i.e., the next instant when an arrival to the target queue finds it empty).

Note that since the initial states for successive simulation runs are no longer independent, the outputs from the successive runs are no longer independent. To get around this problem we use the method of batch means. We divide the entire simulation output into batches, each having 100 successive samples. The batch means obtained from successive batches are (approximately) i.i.d. and can be used to determine the relative error of the estimate (as, e.g., in Chang et al. 1994).

Recall that for estimating CRF, we need to estimate the ratio of variances under IS and under naive simulation (along with the ratio of computational effort per sample under IS and naive simulation). In our experiments, the estimate of the ratio of variances is taken to equal the estimated variance of each batch generated using IS with the variance of average of hundred independent samples generated under naive simulation estimated as $\hat{p}(1-\hat{p})/100$ (recall that $\hat{p}$ denotes the IS estimator of the probability). Here, we make a simplifying assumption that it is feasible to generate independent samples from ψ under naive simulation (given the enormous values of estimated CRF, it is reasonable to expect similar order of magnitude results even under more accurate comparison methodologies).

Example 5 Consider a collection of two-queue tandem Jackson networks with $(\lambda_1, \mu_t) = (2, 3)$ and μ_1 set to 2.5, 3.5, 4.5 and 5.5. Again, the second queue is the target queue. For $\mu_1 = 2.5$, the proposed change of measure $(\lambda_1^*, \mu_1^*, \mu_t^*)$ is determined to be $(2.7785, 2.7785, 1.943)$. For $\mu_1 \geq 3$, the change of measure is given by $(3, \mu_1, 2)$. From the discussion in Example 2, it can be seen that the resulting estimator is asymptotically efficient when $\mu_1 > 4.5$. The results of the simulation are given in Table 3. For each set of parameters, 10^4 batches were generated, where as mentioned earlier, each batch comprised 100 target queue busy cycles. When $\mu_1 > \mu_t$, the estimator appears to be stable but not its relative error, which though small, is variable (as suggested by a smaller estimated relative error for a larger buffer size). The theoretical value of the probability is calculated using (38).

Example 6 Consider the 5-queue Jackson network of Example 4. The performance of the proposed change of measure under Type-II initial distribution for different target values of N is shown in Table 4. For each set of parameters, 10^3 batches were generated, where again, each batch comprised 100 target queue busy cycles. Note that the estimator appears to be quite stable under the new change of measure, even though the feeder Queues 1 and 4 are critical and it is not clear if Assumption 2 holds (clearly, Equation 33 does not hold for Queues 1 and 4).

μ_1	N	Estimate	RE	$\hat{\tau}$	τ	Est. CRF
2.5	25	1.93×10^{-5}	3.39%	0.019	0.0031	7.26×10^{0}
	50	1.05×10^{-9}	29.84%	0.036	0.0031	1.21×10^{3}
	100	2.28×10^{-19}	12.7%	0.070	0.0031	2.22×10^{12}
3.5	25	1.94×10^{-5}	1.19%	0.017	0.0031	6.40×10^{1}
	50	8.02×10^{-10}	2.63%	0.033	0.0031	1.73×10^{5}
	100	1.24×10^{-18}	3.50%	0.063	0.0031	3.22×10^{13}
4.5	25	1.98×10^{-5}	0.63%	0.018	0.0031	2.17×10^{2}
	50	7.83×10^{-10}	0.55%	0.035	0.0031	3.68×10^{6}
	100	1.27×10^{-18}	2.03%	0.070	0.0031	8.66×10^{13}
5.5	25	1.98×10^{-5}	0.33%	0.018	0.0031	7.58×10^{2}
	50	7.86×10^{-10}	0.32%	0.037	0.0031	1.03×10^{7}
	100	1.23×10^{-18}	0.37%	0.073	0.0031	2.47×10^{15}

Table 3: Two-queue tandem network with Type-II initial distribution described in Example 5. The theoretical values of the probabilities for $N = 25, 50$ and 100 equal 1.98×10^{-5}, 7.84×10^{-10} and 1.23×10^{-18}, respectively. Here $\hat{\tau}$ and τ denote the average CPU time required to generate one batch of samples under IS and naive simulation, respectively.

N	Estimate	RE	$\hat{\tau}$	τ	Est. CRF
10	3.84×10^{-5}	0.49%	0.098	0.024	2.66×10^{3}
25	3.45×10^{-12}	1.02%	0.22	0.024	2.85×10^{9}
50	6.20×10^{-24}	2.99%	0.43	0.024	1.01×10^{20}

Table 4: 5-queue network with Type-II initial distribution described in Example 6. Here $\hat{\tau}$ and τ denote the average CPU time required to generate one batch of samples under IS and naive simulation, respectively.

Example 7 Again consider the 5-queue network of Example 4 with the difference that (μ_1, μ_4) are both increased to 2 and μ_t is reduced to 1. The total arrival rates to each queue $(\gamma_t, \ldots, \gamma_4)$ are the same as before and thus equal $(0.44, 1.22, 0.67, 0.57, 0.88)$ and the respective traffic intensities $(\rho_t, \ldots, \rho_4)$ equal $(0.44, 0.61, 0.28, 0.34, 0.44)$.

The resulting change of measure has the following parameters: the external arrival rates $(\lambda_t^*, \ldots, \lambda_4^*)$ equal $(0.4543, 0.8863, 0.5418, 0.3331, 0.6236)$, the service rates $(\mu_t^*, \ldots, \mu_4^*)$ equal $(0.5609, 2, 2.4, 1.7, 2)$, and the routing probability matrix P^* is given by

$$\begin{pmatrix} 0.3566 & 0.0522 & 0.0255 & 0.0697 & 0.0408 \\ 0.1025 & 0.2000 & 0.0489 & 0.1503 & 0.0469 \\ 0.1048 & 0.0818 & 0.0500 & 0.0307 & 0.0959 \\ 0.1432 & 0.0998 & 0.0390 & 0.0300 & 0.0936 \\ 0.0437 & 0.0426 & 0.0521 & 0.0214 & 0.0800 \end{pmatrix}.$$

The new total arrival rates to each queue $(\gamma_t^*, \ldots, \gamma_4^*)$ equal $(1.00, 1.35, 0.73, 0.64, 0.91)$ and the respective traffic intensities $(\rho_t^*, \ldots, \rho_4^*)$ equal $(1.78, 0.67, 0.30, 0.37, 0.46)$. It can be easily seen that (33) holds for each feeder queue, and hence Assumption 2 also holds. It follows from Theorem 3 that the estimator is provably asymptotically efficient; this is also suggested by the simulation results displayed in Table 6. For each set of parameters, 10^3 batches were generated, where again, each batch comprised 100 target queue busy cycles.

N	Estimate	RE	$\hat{\tau}$	τ	Est. CRF
10	3.46×10^{-4}	0.36%	0.11	0.030	5.81×10^{2}
25	1.57×10^{-9}	0.39%	0.32	0.030	3.87×10^{7}
50	1.98×10^{-18}	0.41%	0.52	0.030	1.75×10^{16}

Table 5: 5-queue network with Type-II initial distribution described in Example 7. Here $\hat{\tau}$ and τ denote the average CPU time required to generate one batch of samples under IS and naive simulation, respectively.

6.4 Estimating MTBL

We now estimate MTBL for Example 7 by modifying the methodology described in the beginning of Section 6.3 along the lines suggested in Glynn et. al. (1993). Rather than generating sample paths of the CTMC associated with the Jackson network (to generate samples of τ, the length of the pseudo-regeneration cycle), we generate sample paths of the embedded DTMC, and use the expected holding times at any state as surrogates for the holding times at that state. The simulation methodology proceeds as described earlier: The network is run for a long period of time (as in Section 6.3, a thousand target queue busy cycles are simulated) using the original probability measure. Thereafter, every time an arrival to the target queue finds it empty, i.e., a pseudo-regeneration cycle is initiated, two simulations are run as before. One simulated path is generated using the change of measure until, either the rare event $\mathcal{B}_N$ occurs or the pseudo-regeneration cycle ends, whichever occurs first. In case $\mathcal{B}_N$ occurs, the remaining path till the pseudo-regeneration cycle ends, is generated using the original measure, and the sample of $L_T * \tilde{M}$ is obtained (recall that $\tilde{M}$ denotes the number of target queue level crossing events in a pseudo-regeneration cycle). Note that $\tilde{M}$ equals 0 if level crossing does not occur. As before, the other simulation path is generated using the original measure to generate the sample of the initial state of the next pseudo-regeneration cycle. This path also provides us with a sample of τ.

We now describe the procedure to construct confidence intervals in this setting. Again, divide the entire simulation output into batches, each having k samples of $L_T * \tilde{M}$ and τ (in our experiments, $k = 100$). Let n denote the total number of batches and let X_i and Y_i denote the average of samples of $L_T * \tilde{M}$ and τ, respectively, from the batch i. Then, for sufficiently large batch size k, the sequence $((X_i, Y_i) : i \leq n)$ may be considered to be a collection of n approximately i.i.d. vectors. Our point estimator of ζ equals

$$\frac{\sum_{i \leq n} Y_i}{\sum_{i \leq n} X_i}.$$

Then, for large n,

$$\sqrt{n}\left(\frac{\sum_{i \leq n} Y_i}{\sum_{i \leq n} X_i} - \zeta\right)$$

is approximately normally distributed with mean 0 and variance

$$\sigma^2 = \frac{\sigma_Y^2 - 2\zeta\sigma_{XY} + \zeta^2\sigma_X^2}{\mu_X^2}$$

where σ_Y^2 denotes the variance of Y_i, σ_X^2 denotes the variance of X_i, σ_{XY} denotes the covariance of X_i and Y_i and μ_X denotes the mean of X_i, under the measures used to generate them in the simulation (see, e.g., Glynn et al. 1993). This variance is estimated from the generated samples and the normal approximation is then used to construct confidence intervals (alternatively, one may use the methodology proposed by Calvin, Glynn and Nakayama 2001 to directly construct confidence intervals without resorting to batch means).

For comparison purposes we also estimate MTBL using naive simulation. Again, the network is run for a long period of time using the original measure. Thereafter, every time an arrival to the target queue finds it empty, a simulation path is generated using the original measure to generate the sample of the initial state of the next pseudo-regeneration cycle. However, in this case the generated path provides us a sample of τ as well as $\tilde{M}$. The methodology for constructing the confidence interval is same as in the previous case. The results of the simulation experiments for Example 7 are given in Table 6.4. They illustrate that IS is very effective compared to naive simulation even for estimating MTBL.

	Importance Sampling				Naive Simulation				
N	# Batches	Estimate	RE	$\hat{\tau}$	# Batches	Estimate	RE	τ	Est. CRF
5	10^4	134.28	0.16%	0.124	10^5	134.67	0.25%	0.022	4.6
10	10^4	8,159.6	0.17%	0.242	4×10^5	8,182.7	1.03%	0.023	134.8
20	10^4	2.96×10^7	0.18%	0.481	1.6×10^6	3.11×10^7	29%	0.023	2.3×10^5

Table 6: Estimation of MTBL for the 5-queue network with Type-II initial distribution described in Example 7. Here $\hat{\tau}$ and τ denote the average CPU time required to generate one batch of samples under IS and naive simulation, respectively.

7 Conclusions

Efficient simulation of queueing networks is amongst the more difficult open problems in simulation. In this paper, we made some initial progress in this setting by devising successful techniques to simulate the probability of a single queue build-up in the Jackson network setting when the service rates at the other queues in the network are sufficiently large. When the buffers at the feeder queues are finite and small (or more generally, under Type-I conditions) then the methodology developed may work for all stable Jackson networks. In future work, these ideas may be generalized to efficiently simulate generalized Jackson networks (see Juneja and Nicola 2003 for initial work in this direction).

A key limitation of our approach is that we rely on a change of measure that emphasizes a single linear path to the rare event (of the scaled network), while it is well known that in many cases, the most likely way a rare event happens is through a piece-wise linear path (with multiple linear components). It is noteworthy that in the existing literature, no successful implementation involving a change of measure that is appropriately piecewise constant (so that a piece-wise linear path is emphasized) has been proposed. The key problem is that in importance sampling, it is not sufficient that the new measure emphasizes the most likely paths to the rare event, but also that it does not assign much lesser than the original probability to any path to the rare event (to avoid build-up of square of the likelihood ratio that can potentially blow-up the second moment of the estimate).

Appendix A: Proof of Large Deviations Lower Bound

As suggested earlier, the proof becomes simpler if the paths involving the boundaries of queues in $\mathcal{C}$ are not considered. We do this by determining a large deviations lower bound (ldlb) for $\mathbf{P}_{\overline{\pi}}(\mathcal{B}_N)$, where under the initial distribution $\bar{\pi}$, X_0 corresponds to the state where length of queues in $\mathcal{C}$ equals $N^{3/4}$, $Q_0(t) = 1$ and the remaining queues in $\mathcal{F} - \mathcal{C}$ are of length 1 (The value $3/4$ is chosen for notational convenience, the analysis is unaffected if it is replaced by any value $\in (1/2, 1)$).

Lemma 2

$$\liminf_{N\to\infty} \frac{\log \mathbf{P}_{\pi}(\mathcal{B}_N)}{N} \geq \liminf_{N\to\infty} \frac{\log \mathbf{P}_{\bar{\pi}}(\mathcal{B}_N)}{N}.$$

Proof: Note that starting from any state s such that $\pi(s) > 0$, one can restrict analysis to a path to the state corresponding to $\bar{\pi}$ along which the target queue length remains ≥ 1 and $< N$ and that has a probability $O(k^{N^{3/4}})$ for a positive constant $k < 1$. Then, $\mathbf{P}_{\pi}(\mathcal{B}_N) \geq \pi(s) O(k^{N^{3/4}}) \mathbf{P}_{\bar{\pi}}(\mathcal{B}_N)$ and the result follows. □

We complete the proof by showing that:

$$\liminf_{N\to\infty} \frac{\log \mathbf{P}_{\bar{\pi}}(\mathcal{B}_N)}{N} \geq -\log c_t.$$

Let d denote the total number of queues in the network. Let $\bar{\mathcal{B}}_N$ denote the subset of $\mathcal{B}_N$ where the queue length in queues in $\mathcal{C}$ remains between $1/2N^{3/4} - d\sqrt{N}$ (N is assumed to be sufficiently large so that this is positive) and $3/2N^{3/4}$ and every queue in $\mathcal{F} - \mathcal{C}$ remains less than $\sqrt{N}$ up to time T (these boundaries on the lengths of queues are arbitrary; what we really need is that critical queues do not vary by more than $N^{1/2+\epsilon}$ for any small $\epsilon > 0$ in either direction from their initial value and the stable queues remain less than $(\log N)^{1+\epsilon}$ up to time T).

Lemma 3 *Under Assumption 1,*

$$\liminf_{N\to\infty} \frac{\log \mathbf{P}_{\bar{\pi}}(\bar{\mathcal{B}}_N)}{N} \geq -\log c_t + \liminf_{N\to\infty} \frac{\log \mathbf{P}^*_{\bar{\pi}}(\bar{\mathcal{B}}_N)}{N}. \tag{40}$$

Proof: Note that along a sample path in the set $\bar{\mathcal{B}}_N$, L_T equals

$$\prod_{i\in\mathcal{H}} c_i^{-(Q_T(i)-Q_0(i))} \geq \prod_{i\in\mathcal{H}} c_i^{-Q_T(i)} \quad a.s.$$

Note that $Q_T(t) = N$. Also, in $\bar{\mathcal{B}}_N$, $Q_T(i) \leq 3/2N^{3/4}$ for $i \in \mathcal{C}$ and $Q_T(i) \leq \sqrt{N}$ for $i \in \mathcal{F} - \mathcal{C}$. It follows that

$$\mathbf{P}_{\bar{\pi}}(\bar{\mathcal{B}}_N) = E_{\mathbf{P}_{\bar{\pi}}}(L_T I(\bar{\mathcal{B}}_N)) \geq c_t^{-N} \left(\prod_{i\in\mathcal{C}} c_i^{-3/2\, N^{3/4}} \right) \left(\prod_{i\in\mathcal{F}-\mathcal{C}} c_i^{-\sqrt{N}} \right) \mathbf{P}^*_{\bar{\pi}}(\bar{\mathcal{B}}_N),$$

from which the lemma follows. □

In view of the above lemma, to establish Theorem 1 it suffices to show that

$$\liminf_{N\to\infty} \mathbf{P}^*_{\bar{\pi}}(\bar{\mathcal{B}}_N) > 0. \tag{41}$$

Note that the target queue under $\mathbf{P}^*_{\bar{\pi}}$ is unstable. Thus, there is a positive probability that starting from any state it never empties. Hence, $\liminf_{N\to\infty} \mathbf{P}^*_{\bar{\pi}}(\mathcal{B}_N) > 0$. However, to show that (41) holds further analysis is needed.

Consider another stochastic process $(\bar{X}_n : n \geq 0)$ where $\bar{X}_n = (\bar{Q}_n(i) : i \in \mathcal{C} \cup \{t\})$, that differs from $(X_n : n \geq 0)$ under $\mathbf{P}^*_{\bar{\pi}}$ in distribution in that the service completions at queues in $\mathcal{F} - \mathcal{C}$ are instantaneous. Thus, an arrival to these queues (external arrival or an arrival routed after service completion from queues in $\mathcal{C} \cup \{t\}$) at time n instantaneously reaches either a queue in $\mathcal{C}$, queue t or it leaves the system at time n itself, following the routing governed by the transition matrix under $\mathbf{P}^*_{\bar{\pi}}$.

This process has the desirable property that whenever all the queues are non-empty, under $\mathbf{P}^*_{\bar{\pi}}$ each queue $i \in \mathcal{C}$ has a positive probability $p_i \leq 1/2$ of increasing by one or decreasing by one, and probability $1 - 2p_i$ of remaining unchanged. Thus, the queue length behaves like a symmetric simple random walk.

We briefly sketch how $(\bar{X}_n : n \geq 0)$ and $(X_n : n \geq 0)$ may be constructed on the same probability space as path by path arguments are useful in our analysis. Our interest is in analyzing these processes with identical initial distributions corresponding to $N^{3/4}$ customers in queues in $\mathcal{C}$, 1 customer in queue t and zero customers in queues in $\mathcal{F} - \mathcal{C}$ for both the Markov chains. We observe these systems till a queue in $\mathcal{C} \cup \{t\}$ becomes empty in either of these systems. We refer to $(X_n : n \geq 0)$ as the primary chain and $(\bar{X}_n : n \geq 0)$ as the modified chain.

Suppose the stream of uniform numbers $(U_0, U_1, \ldots)$ governs the arrival and service completion decisions for the modified chain, i.e., for each i, U_i decides whether transition from $\bar{X}_i$ to $\bar{X}_{i+1}$ corresponds to an arrival or a service completion at a specified queue. Suppose that the stream of uniform numbers $(V_0, V_1, \ldots)$ governs the probabilistic routing of the modified chain. The same streams are used for generating the primary chain, however now additional decisions need to be taken. Suppose at a state X_n, $\sum_{i \in \mathcal{F}-\mathcal{C}} Q_n(i) > 0$. Then a variable from another uniform stream $(W_0, W_1, \ldots)$ is used to decide whether a departure from any of the queues in $\mathcal{F} - \mathcal{C}$ is the next event. If yes, then a number from this stream decides the queue in $\mathcal{F} - \mathcal{C}$ that has a departure. If not, or if $\sum_{i \in \mathcal{F}-\mathcal{C}} Q_n(i) = 0$, then for i th such event, for each i, U_i is used to decide whether an arrival to a specific queue in $\mathcal{H}$ has occurred or whether a service completion in a specific queue in $\mathcal{C} \cup \{t\}$ has occurred in a synchronous manner as used for deciding the transition $\bar{X}_i$ to $\bar{X}_{i+1}$ (since our analysis is restricted to the time that the two chains have queues in $\mathcal{C} \cup \{t\}$ non-empty, such synchronization is easily achieved). All probabilistic routing are made using $(V_1, V_2, \ldots)$ for both the chains in complete synchronization, i.e., arrival j to queue i follows identical routing in both the systems for all j and i.

Let $\bar{\tau}$ denote the time at which a queue in $\mathcal{C} \cup \{t\}$ becomes empty in the modified chain and let τ be similarly defined for the primary chain. For the primary chain, for each i, let τ_i denote the time at which U_i is used to decide the next state. Note that the arrival or service completion event in the transition $\bar{X}_i \rightarrow \bar{X}_{i+1}$ is identical to the transition $X_{\tau_i} \rightarrow X_{\tau_i + 1}$ for $i < \bar{\tau}$ and $\tau_i < \tau$, from our construction.

Define $\xi(j) = \max\{i : \tau_i \leq j\}$. Thus, $j - \xi(j)$ denotes the number of transitions involving service completions in queues $\mathcal{F} - \mathcal{C}$ along the chain $(X_n : n \leq j)$. The following lemma relates the queue length in the two Markov chains.

Lemma 4 *Under the above construction, for $j \leq \tau$, for $k \in \mathcal{C} \cup \{t\}$,*

$$Q_j(k) \leq \bar{Q}_{\xi(j)}(k) \leq Q_j(k) + \sum_{l \in \mathcal{F}-\mathcal{C}} Q_j(l)$$

a.s. It thus follows that $\xi(j) \leq \bar{\tau}$ a.s.

Proof: Note that the relative sequence of arrivals to the various queues in the network and service completions from queues in $\mathcal{C} \cup \{t\}$ are identical in $(X_0, X_1, \ldots, X_j)$ and $(\bar{X}_0, \bar{X}_1, \ldots, \bar{X}_{\xi(j)})$ by construction. They differ in that the transitions from service completions in queues in $\mathcal{F} - \mathcal{C}$ may occur in $(X_0, X_1, \ldots, X_j)$ and that all the customers in queues $\mathcal{F} - \mathcal{C}$ in state X_j have already reached queues in $\mathcal{C} \cup \{t\}$ in $\bar{X}_{\xi(j)}$ or have left the system, based on assigned probabilistic routing. Thus, $\bar{Q}_{\xi(j)}(k)$ contains $Q_j(k)$ and all the customers in queues $\mathcal{F} - \mathcal{C}$ in state X_j that will end up in queue k before hitting any other queue in $\mathcal{C}$ or leaving the system. The results therefore follow. $\Box$

Define $\bar{T}_N = \inf\{n : \bar{Q}_n(t) \geq N + d\sqrt{N}\}$. Now consider the paths of $(\bar{X}_i : i \leq \bar{T}_N)$ where the lengths of queues in $\mathcal{C}$ remain between $1/2\ N^{3/4}$ and $3/2\ N^{3/4}$. We call this event $\bar{\mathcal{B}}_{1,N}$. Note that the target queue is unstable and hence $E_{\mathbf{P}^*_{\bar{\pi}}} T_N = O(N)$. In particular, given $\epsilon > 0$, due to Markov inequality, there exists a constant $K_\epsilon > 0$ such that $\mathbf{P}^*_{\bar{\pi}}(T_N \leq K_\epsilon N) \geq 1 - \epsilon$. Let $\bar{\mathcal{B}}_{2,N}$ denote $\{T_N \leq K_\epsilon N\}$.

In addition, consider the event that the maximum in each queue in $\mathcal{F} - \mathcal{C}$ in the primary chain $(X_i : i \geq 0)$ up to time $K_\epsilon N$ remains bounded by $\sqrt{N}$. We refer to this event as $\bar{\mathcal{B}}_{3,N}$. Recall that $\mathcal{B}_N$ is simply the event that the target queue exceeds level N before emptying in the primary Markov chain. Recall the definition of $\bar{\mathcal{B}}_N \subset \mathcal{B}_N$ given earlier in this section. The following lemma is useful in proving (41).

Lemma 5 *For N sufficiently large,*

$$\bar{\mathcal{B}}_{1,N} \cap \bar{\mathcal{B}}_{2,N} \cap \bar{\mathcal{B}}_{3,N} \cap \mathcal{B}_N \subset \bar{\mathcal{B}}_N. \tag{42}$$

Proof:

Suppose that in the LHS of (42), there exists a sample path of positive probability along which there exists a queue $k \in \mathcal{C}$, and time $\tilde{T}$ that $Q_{\tilde{T}}(k)$ is either less than $\frac{1}{2}N^{3/4} - d\sqrt{N}$ or greater than $\frac{3}{2}N^{3/4}$, while till time $\tilde{T} - 1$ all queues in $\mathcal{C}$ are within these limits. Suppose that $\tilde{T} \leq T_N$. It suffices to show that such a sample path is not possible.

Now, up till $\tilde{T}$ all queues in $\mathcal{C} \cup \{t\}$ are busy in the primary chain (recall that $T_N < T_o$ in $\mathcal{B}_N$, where T_o denotes the first time the target queue empties). Thus, from Lemma 4, it follows that for $n \leq \tilde{T}$,

$$\bar{Q}_{\xi(n)}(t) \leq Q_n(t) + \sum_{l \in \mathcal{F} - \mathcal{C}} Q_n(l) \leq N + d\sqrt{N}.$$

In particular, $\xi(\tilde{T}) \leq \bar{T}_N$. Also, from Lemma 4 it follows that $\bar{Q}_{\xi(n)}(i) \geq Q_n(i)$ for $n \leq \tilde{T}$, for $i \in \mathcal{C} \cup \{t\}$. Thus, none of the queues become empty in the modified system till $\xi(\tilde{T})$. Since, for $k \in \mathcal{C}$, $\frac{1}{2}N^{3/4} \leq \bar{Q}_{\xi(\tilde{T})}(k) \leq \frac{3}{2}N^{3/4}$ from Lemma 4 we get the desired contradiction. $\Box$

Now, to see that (41) holds, it suffices to show that $\liminf_{N \to \infty} \mathbf{P}^*_{\bar{\pi}}(\bar{\mathcal{B}}_{1,N} \cap \bar{\mathcal{B}}_{2,N} \cap \bar{\mathcal{B}}_{3,N} \cap \mathcal{B}_N) > 0$. Since the target queue is unstable under $\mathbf{P}^*_{\bar{\pi}}$, we have $\liminf_{N \to \infty} \mathbf{P}^*_{\bar{\pi}}(\mathcal{B}_N) > 0$. Since, $\mathbf{P}^*_{\bar{\pi}}(\bar{\mathcal{B}}_{3,N}) \geq 1 - \epsilon$ for an arbitrary ϵ, it suffices to establish that both $\mathbf{P}^*_{\bar{\pi}}(\bar{\mathcal{B}}_{1,N})$ and $\mathbf{P}^*_{\bar{\pi}}(\bar{\mathcal{B}}_{2,N})$ are arbitrarily close to 1 for sufficiently large N.

The fact that the probability $\mathbf{P}(\bar{\mathcal{B}}_{2,N})$ is arbitrarily close to 1 for large N follows from Zeevi and Glynn (2001), where they show that the maximum of queue length of a stable queue in a Jackson network grows as a logarithmic function of time. Thus, the probability that the queue length exceeds $\sqrt{N}$ by time kN for any constant $k > 0$ tends to zero as $N \to \infty$.

To see that $\mathbf{P}^*_{\bar{\pi}}(\bar{\mathcal{B}}_{1,N})$ tends to 1 as N increases to infinity we need some notation and analysis. Let

$$\bar{T}^i_N = \inf\{n \geq 1 : |\bar{Q}_n(i) - N^{3/4}| \geq 1/2 N^{3/4}\}$$

for $i \in \mathcal{C}$. Then it is easily seen that

$$\bar{\mathcal{B}}_{1,N} = \cap_{i\in\mathcal{C}}\{\bar{T}_N^i > \bar{T}_N\}.$$

Note that along $\bar{\mathcal{B}}_{1,N}$ the queues in $\mathcal{C}$ do not empty up to time $\bar{T}_N$. Thus, for the purpose of ascertaining $\mathbf{P}^*_{\bar{\pi}}(\bar{\mathcal{B}}_{1,N})$, we can view each of these queues as a symmetric random walk without boundaries.

Note that the proposition that $\mathbf{P}^*_{\bar{\pi}}(\bar{\mathcal{B}}_{1,N}) \to 1$ as $N \to \infty$ follows if we can establish that

$$\mathbf{P}^*(\bar{T}_N^i > \bar{T}_N) \to 1$$

as $N \to \infty$ for each $i \in \mathcal{C}$.

Lemma 6 *The probability* $\mathbf{P}^*(\bar{T}_N^i > \bar{T}_N) \to 1$ *as* $N \to \infty$.

Proof: Note that

$$\begin{aligned} E(\bar{T}_N^i) &= E(\min(\bar{T}_N, \bar{T}_N^i)) + & (43)\\ &\quad \mathbf{P}^*(\bar{T}_N^i > \bar{T}_N)E(\bar{T}_N^i - \bar{T}_N|\bar{T}_N^i > \bar{T}_N). & (44)\end{aligned}$$

Note that $E(\min(\bar{T}_N, \bar{T}_N^i)) \le E(\bar{T}_N)$ and hence is $O(N)$ since the target queue is unstable and thus $E(\bar{T}_N)$ is $O(N)$.

Since $(\bar{Q}_n(i) : n \ge 0)$ is a symmetric random walk for each $i \in \mathcal{C}$, from Lemma 7 it can be seen that

$$E(\bar{T}_N^i) = \frac{N^{1.5}}{2p_i}.$$

(Recall that p_i is the probability with which $\bar{Q}_n(i)$, $i \in \mathcal{C}$ increases or decreases by one step in the next transition when all queues in $\mathcal{C} \cup \{t\}$ are busy). Similarly from Lemma 7 it also follows that,

$$E(\bar{T}_N^i - \bar{T}_N|(\bar{T}_N^i > \bar{T}_N)) \le \frac{N^{1.5}}{2p_i}.$$

Plugging these in (43), we get

$$\frac{N^{1.5}}{2p_i}(1 - \mathbf{P}^*(\bar{T}_N^i > \bar{T}_N)) \le O(N).$$

Thus, the result follows. □

Hitting Times for a Symmetric Simple Random Walk

Let $(Z_n : n \ge 1)$ be a sequence of i.i.d. random variables such that Z_n takes value 1 with probability (w.p.) p, -1 w.p. p and 0 w.p. $1 - 2p$. Consider the associated symmetric simple random walk $(S_n : n \ge 0)$, where $S_n = \sum_{i=1}^n Z_i + S_0$ and $S_0 = x$ (x is an integer). Suppose that a is any positive integer so that $|x| < a$. Let $T = \inf\{n \ge 1 : |S_n| = a\}$. Let E_x denote the expectation operator when $S_0 = x$.

Lemma 7

$$E_xT = (a^2 - x^2)/(2p), \quad (45)$$

In particular, from the above lemma it follows that

$$\sup_{x\in Z^+:|x|<a} E_xT \leq a^2/(2p). \tag{46}$$

Proof: The proof relies on the observation that the random sequence $(S_n^2 - 2pn : n \geq 0)$ is a martingale and T is a stopping time (see any standard text on Martingales, e.g., Williams 1991). Therefore by Martingale Stopping Time Theorem:

$$E_xS_T^2 - 2pE_xT = x^2. \tag{47}$$

The result follows since $E_xS_T^2 = a^2$. □

Appendix B

Proof of Lemma 1

We need some notation before proceeding with the proof. Let $\alpha_i = \frac{\mu_i^*}{\mu_i}$ for $i \in \mathcal{H}$ and for each such i, define the function $g_i(\cdot)$ so that

$$g_i(\theta) = -(\log c_i + \log \alpha_i)\theta + \log(\sum_{j\in\mathcal{H}} c_j^\theta p_{ij} + p_{ie}).$$

For $i \in \mathcal{H}$ define the convex functions $f_i^a(\cdot)$ and $f_i^s(\cdot)$ as follows:

$$f_i^a(\theta) = -\log(1+\theta(c_i-1)) + \theta\log c_i$$

and

$$f_i^s(\theta) = -\log(1+\theta(\alpha_i-1)) + \theta\log\alpha_i.$$

Note that these functions are defined for values that include $[0,1]$ as a proper subset.

Proof of Lemma 1: Note that $\log(\sum_{j\in\mathcal{H}} c_j^\theta p_{ij} + p_{ie})$ is the log-moment generating function of a random variable (rv) that takes value $\log c_j$ with probability p_{ij} and 0 with probability p_{ie}. It is well known that log-moment generating function of a rv is convex (see, e.g., Dembo and Zeitouni 1991). Thus, $g_i(\cdot)$ is convex. Note that $g_i(0) = 0$ and due to Assumption 1, $g_i(1) = 0$. Thus,

$$(\log c_i + \log \alpha_i) = \log(\sum_{j\in\mathcal{H}} c_j p_{ij} + p_{ie}). \tag{48}$$

In addition, it follows from convexity that the first derivative $g_i'(1) \geq 0$.

This, along with the fact that $p_{ij}^* = \frac{c_j}{c_i\alpha_i}p_{ij}$ and (48) implies that for $i \in \mathcal{H}$

$$-(\log c_i + \log\alpha_i) + \sum_{i\in\mathcal{H}} p_{ij}^* \log c_j \geq 0.$$

Let $\overset{\rightarrow}{\log} c = (\log c_i : i \in \mathcal{H})$ and $\overset{\rightarrow}{\log} \alpha = (\log \alpha_i : i \in \mathcal{H})$ be column vectors. Then, in matrix notation, we may re-write the above inequalities as:

$$-\overset{\rightarrow}{\log} \alpha \geq (I - P^*) \overset{\rightarrow}{\log} c\ .$$

Since $R^* = (I-P^*)^{-1}$ is non-negative, it follows that for $i \in \mathcal{H}$

$$-R^* \overset{\rightarrow}{\log} \alpha \geq \overset{\rightarrow}{\log} c$$

or

$$-\sum_{j\in\mathcal{H}} r^*_{ij}\log\alpha_j \geq \log c_i. \tag{49}$$

Also note that $f^a_i(0) = f^s_i(0) = 0$ and $f^a_i(1) = f^s_i(1) = 0$. Again, this implies that $f^{a\prime}_i(1) > 0$ (since $c_i > 1$, each $f^a_i(\cdot)$ is strictly convex) and $f^{s\prime}_i(1) \geq 0$. Therefore, for $i \in \mathcal{H}$:

$$-\frac{c_i - 1}{c_i} + \log c_i > 0 \tag{50}$$

and

$$-\frac{\alpha_i - 1}{\alpha_i} + \log\alpha_i \geq 0. \tag{51}$$

Simplifying and multiplying (50) on both sides by λ_i and similarly simplifying and multiplying (51) on both sides by μ_i, we get for each $i \in \mathcal{H}$:

$$\lambda^*_i \log c_i \geq \lambda^*_i - \lambda_i \tag{52}$$

and

$$\mu^*_i \log\alpha_i \geq \mu^*_i - \mu_i$$

Summing up both sides for all $i \in \mathcal{H}$, and noting (1) we get

$$\sum_{i\in\mathcal{H}} \lambda^*_i \log c_i + \sum_{i\in\mathcal{H}} \mu^*_i \log\alpha_i > 0.$$

The inequality above is strict as there exists at least one i such that the inequality (52) is strict (i.e., when $\lambda_i > 0$).

Combining this with (49), it follows that

$$-\sum_{i\in\mathcal{H}}\sum_{j\in\mathcal{H}} \lambda^*_i r^*_{ij}\log\alpha_j + \sum_{i\in\mathcal{H}} \mu^*_i \log\alpha_i > 0. \tag{53}$$

We complete the proof by first assuming that the target queue is not unstable (i.e., it is stable or critical) and then show a contradiction. Note that if all queues are either stable or critical, then we have for $i \in \mathcal{H}$,

$$\gamma^*_j = \sum_{i\in\mathcal{H}} r^*_{ij}\lambda^*_i. \tag{54}$$

Also note that $\mu^*_i = \gamma^*_i$ for $i \in \mathcal{C}$ and $\alpha_i = 1$ and hence $\log\alpha_i = 0$ for $i \in \mathcal{F} - \mathcal{C}$. Thus, $\sum_{i\in\mathcal{F}} \mu^*_i \log\alpha_i = \sum_{i\in\mathcal{F}} \gamma^*_i \log\alpha_i$.

Using these relations in (53), we see that

$$-\sum_{j\in\mathcal{H}} \gamma^*_j \log\alpha_j + \sum_{i\in\mathcal{F}} \gamma^*_i \log\alpha_i + \mu^*_t \log\alpha_t > 0$$

or

$$(\mu^*_t - \gamma^*_t)\log\alpha_t > 0.$$

Since, $\alpha_t < 1$ and hence $\log\alpha_t < 0$, it follows that $\gamma^*_t > \mu^*_t$. This gives the desired contradiction. □

Proof of Lower Bound in Proposition 1:

Let $\mathcal{A}_N$ denote the event that there are between N and $2N$ customers in the target queue and less than N customers in each of the feeder queues.

Recall that $\phi(\mathcal{A}_N)$ denotes the steady state probability of being in $\mathcal{A}_N$. It is easy to see that

$$\lim_{N\to\infty} \frac{\log \mathbf{P}_\psi(\mathcal{A}_N)}{N} = -\log \frac{\mu_t}{\gamma_t}.$$

Using the ratio representation, we have

$$\phi(\mathcal{A}_N) = \frac{E_{\mathbf{P}_\psi}(\bar{D})}{E_{\mathbf{P}_\psi}(\tau)},$$

where $\bar{D}$ denotes the time spent in $\mathcal{A}_N$ is a pseudo-regenerative cycle. Observe that for $\bar{D}$ to take a positive value, $\mathcal{B}_N$ must occur. Thus,

$$E_{\mathbf{P}_\psi}(\bar{D}) = P_\psi(\mathcal{B}_N) E_{\mathbf{P}_\psi}(\bar{D}|\mathcal{B}_N).$$

To complete, the proof it suffices to upper bound $E_{\mathbf{P}_\psi}(\bar{D}|\mathcal{B}_N)$ by a polynomial in N. Note that this term is upper bounded by the expected time taken by the complete network to empty given that there are $2N$ customers in the target queue and N customers in the feeder queue. We upper bound this using the arguments in Glasserman and Kou (1995). For $x \in \mathcal{Z}_+^{|\mathcal{H}|}$, let E_x denote the expectation operator under the original probability $\mathbf{P}$ when $\tilde{Q}_0 = x$. Let $\tilde{T}$ denote the time taken for the complete network to empty. Set

$$t^* = \max\{E_x(\tilde{T}) : \sum_{i\in\mathcal{H}} x_i = 1\}.$$

Then, as argued in Glasserman and Kou (1995), we have

$$E_{\mathbf{P}_\psi}(\bar{D}|\mathcal{B}_N) \leq t^*(|\mathcal{H}|+1)N,$$

and thus the large deviations lower bound follows.

REFERENCES

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.

AVRAM, F., J. G. DAI AND J. J. HASENBEIN. 2001. Explicit Solutions for Variational Problems in the Quadrant, *Queueing Systems*, **37**, 261-291.

ATAR, R. AND P. DUPUIS. 1999. Large Deviations and Queuing Networks: Methods for Rate Function Identification. *Stochastic Process. Appl.* **84**, 255-296.

BECK, B., A.R. DABROWSKI AND D.R. MCDONALD. 1999. A Unified Approach to Fast Teller Queues and ATM. *Adv. Appl. Prob.* **31**, 758-787.

CALVIN, J.M., P. W. GLYNN AND M. K. NAKAYAMA. 2001. Steady state simulation analysis: importance sampling using the semi-regenerative method. *Proceedings of the 2001 Winter Simulation Conference.* IEEE Press, 441-450.

CHANG, C.S., P. HEIDELBERGER, S. JUNEJA and P. SHAHABUDDIN. 1994. Effective Bandwidth and Fast Simulation of ATM Intree Networks. *Performance Evaluation* **20**, 45-65.

CHEN, H. and D. D. YAO. 2001. *Fundamentals of Queueing Networks: Performance, Asymptotics and Optimization*, Springer-Verlag, New York.

COGBURN, R. 1975. A uniform theory for sums of Markov chain transition probabilities.*The Annals of Probability*, **3**, 191-214.

DEMBO, A. AND O. ZEITOUNI. 1992. *Large Deviations Techniques and Applications.* Jones and Bartlett, Boston, MA.

DUPUIS, P. AND R.S. ELLIS. 1995. The Large Deviations Principle for a General Class of Queuing Systems. *Trans. Amer. Math. Soc.* **347**, 2689-2751.

DUPUIS, P. AND K. RAMANAN. 2002. A Time-reversed Representation for the Tail Probabilities of Stationary Reflected Brownian Motion. *Stochastic Process. Appl.* **98**, 253-287.

FRATER, M.R., T.M. LENON AND B.D.O. ANDERSON. 1991. Optimally Efficient Estimation of the Statistics of Rare Events in Queuing Networks. *IEEE Trans. on Automatic Control* **36**, 12, 1395-1404.

GLASSERMAN, P. AND S. KOU. 1995. Analysis of an Importance Sampling Estimator for Tandem Queues. *ACM Transactions on Modeling and Computer Simulation* **5**, 1, 22-42.

GLYNN, P.W., P. HEIDELBERGER, V. F. NICOLA AND P. SHAHABUDDIN. 1993. Efficient Estimation of the Mean Time Between Failures in Non-Regenerative Dependability Models. Proceedings of the 1993 Winter Simulation Conference. G.W. Evans, M. Mollaghasemi, E.C. Russell and W.E. Biles (eds.). IEEE Press, 311-316.

GLYNN, P. AND D. IGLEHART.1989. Importance Sampling for Stochastic Simulations. *Management Science* **35**, 11, 1367-1392.

GLYNN, P.W. AND W. WHITT. 1992. The asymptotic efficiency of simulation estimators. *Operations Research* **40**, 505-520.

HEIDELBERGER, P. 1995. Fast Simulation of Rare Events in Queuing and Reliability Models. *ACM Transactions on Modeling and Computer Simulation* **5**, 1, 43-85.

IGNATIOUK-ROBERT, I. 2000. Large Deviations of Jackson Networks. *The Annals of Applied Probability* **10**, 3, 962-1001.

IGNATYUK, I.A., V.A. MALYSHEV AND V.V. SCHERBAKOV. 1994. Boundary Effects in Large Deviations Problems. *Russian Math. Surveys* **49**, 2, 41-99.

JUNEJA, S. 2001. Importance Sampling and the Cyclic Approach. *Operations Research* **49**, 6, 900-912.

JUNEJA, S. AND V.F. NICOLA. 2003. Efficient Simulation of Buffer Overflow Probabilities in Queuing Networks with Probabilistic Routing. Technical Report, Tata Institute of Fundamental Research STCS-03/01. Available at authors web pages.

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.

L'ECUYER, P. AND Y. CHAMPOUX. 2001. Estimating small cell-loss ratios in ATM switches via importance sampling. *ACM Trans. Model. Comput. Simul.* **11(1)**, 76-105.

L'ECUYER, P. AND Y. CHAMPOUX. 1996. Importance Sampling for Large ATM-Type Queueing Networks. *Winter Simulation Conference,* IEEE Press.

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.

RANDHAWA, R.S. AND S. JUNEJA. 2003. Combining Importance Sampling and Temporal Difference Control Variates to Simulate Markov Chains. Undergoing revision.

WILLIAMS, D. 1991. *Probability with Martingales.* Cambridge University Press, Cambridge.

WALRAND, J. 1988. *An Introduction to Queuing Networks.* Prentice Hall, Englewood, New

Jersey.

ZEEVI, A. AND P. GLYNN. 2001. Estimating Tail Decays for Stationary Sequences via Extreme Values. Preprint.