Proceedings of the 1999 Winter Simulation Conference
P. A. Farrington, H. B. Nembhard, D. T. Sturrock, and G. W. Evans, eds.

EFFICIENT SIMULATION OF A TANDEM JACKSON NETWORK

Dirk P. Kroese

Teletraffic Research Center
University of Adelaide
South Australia 5005, AUSTRALIA

Victor F. Nicola

Department of Electrical Engineering
University of Twente
Enschede 7500 AE, THE NETHERLANDS

ABSTRACT

In this paper we consider a two-node tandem Jackson network. Starting from a given state, we are interested in estimating the probability that the content of the second buffer exceeds some high level L before it becomes empty. The theory of Markov additive processes is used to determine the asymptotic decay rate of this probability, for large L. Moreover, the optimal exponential change of measure to be used in importance sampling is derived and used for efficient estimation of the rare event probability of interest.

Unlike changes of measures proposed and studied in recent literature, the one derived here is a function of the content of the first buffer, and yields asymptotically efficient simulation for any set of arrival and service rates. The relative error is bounded independent of the level L, except when the first server is the bottleneck and its buffer is infinite, in which case the relative error is bounded linearly in L.

1 INTRODUCTION

The tandem Jackson network has received considerable attention as a reference example for the analysis and testing of different methodologies and various techniques to speed up simulations involving rare events. The particular interest in this system stems from the fact that in spite of its (apparent) simplicity, its large deviations behaviour is not yet fully understood. The main difficulty being its multi-dimensional state space and the complicated large deviations behaviour along its boundaries.

Among rare events of interest in the tandem Jackson network, the most studied is the overflow of the total network population (see, e.g., Parekh and Walrand 1989, Anantharam et al. 1990, Frater and Anderson 1989, Frater et al. 1991, Tsoucas 1992, Glasserman and Kou 1995). Exact large deviations analysis leading to an asymptotically optimal change of measure is quite difficult. Instead, a heuristic change of measure is suggested in Parekh and Walrand (1989), which interchanges the arrival rate (to the first queue) and the slowest service rate. The same change of measure is suggested based on time reversal arguments (see, e.g., Anantharam et al. 1990, Frater et al. 1991). However, analysis in Glasserman and Kou (1995) and counter examples in Glasserman and Wang (1997) show that the importance sampling estimator based on this change of measure is not necessarily asymptotically efficient; in fact, it has an infinite variance in some parameter regions. Other rare events of interest are the buffer overflow at the individual network nodes. If the node of interest is the bottleneck (relative to all preceding nodes), then the optimal exponential change of measure is to interchange the arrival rate and the service rate at this (bottleneck) node; the service rates at all other nodes are kept unchanged (see, e.g., Parekh and Walrand 1989, Frater and Anderson 1989). However, this change of measure is not optimal (not even asymptotically efficient) if we are interested in the buffer overflow at a node after the bottleneck. The theory of *effective bandwidth* has been used to derive heuristics for the efficient simulation of a class of feed-forward discrete-time queueing networks, see, e.g., Chang et al. (1994) and De Veciana et al. (1994). (This class essentially resembles a feed-forward fluid-flow network.) Another approach is considered in Kroese and Nicola (1998) to study a fluid-flow line with unreliable nodes. To the best of our knowledge, analogous approaches for application to continuous-time queueing networks has not yet been introduced; not even for a simple tandem Jackson network.

In this paper we consider a two-node tandem Jackson network, and study the buffer overflow event at the second node. We present a new Markov additive process (MAP) representation of the system. (For MAP definitions and properties, see Ney and Nummelin 1987). This MAP is exponentially (and optimally) tilted for use in an importance sampling procedure to estimate the probability of buffer overflow in the second node. Unlike changes of measure considered in the literature, the one we derive here depends on the contents of the first buffer. No complete proof of its optimality is available at this time, but empirical studies in this paper strongly confirm its asymptotic efficiency. The

resulting estimates have relative error which is (asymptotically) bounded independent of the overflow level, except when the first server is the bottleneck and its buffer is infinite. In the latter case, the relative error is (asymptotically) linearly bounded in the overflow level.

In Section 2 we give some preliminaries. A MAP representation of the system and its exponential change of measure are introduced in Section 3. In Sections 3.1 and 3.3, the optimal changes of measure are derived for finite and infinite first buffer, respectively. Empirical results in Section 4 demonstrate the (asymptotic) efficiency of the developed importance sampling estimator. Conclusions and related future research are given in Section 5.

2 PRELIMINARIES

Consider a simple Jackson network consisting of two queues in tandem. Customers arrive at the first queue (buffer) according to a Poisson process with rate λ. The service time of a customer at the first queue is exponentially distributed with rate μ_1. Customers that leave the first queue enter the second one. The service time in the second queue has and exponential distribution with rate μ_2. We assume stability of the queueing system, i.e.,

$$\lambda < \min\{\mu_1, \mu_2\}.$$

The size of the first buffer is denoted by b_1 (which may be finite or infinite.) Let X_t and Y_t denote the number of customers in the first and second queue at time t, respectively. We assume that the second buffer is initially non-empty; to simplify notation and without loss of generality, we set $Y_0 = 1$. Let $\mathbb{P}_i$ denote the probability measure under which (X_t) starts from i at time 0 (i.e., $X_0 = i, 0 \le i \le b_1$); and let $\mathbb{E}_i$ denote the corresponding expectation operator. In Section 3 we will consider various changes of measure; we will denote by $\tilde{\mathbb{P}}_i$ any such measure for which (X_t) starts at i. $\tilde{\mathbb{E}}_i$ denotes the corresponding expectation operator. We are interested in the probability that, starting from $(X_0, Y_0) = (i, 1)$, the second queue hits some large level $L \in \mathbb{N}$ before hitting 0. We denote this probability by γ_i and will refer to it as the *overflow probability* of the second buffer, given that the initial number of customers in the first queue is i.

3 EXPONENTIAL CHANGE OF MEASURE

The key to understanding the change of measure that we are going to propose is Asmussen and Rubinstein (1995), where an exponential change of measure for Markov additive processes is discussed in the context of rare event simulation. Basically, a Markov additive process in continuous time is a stochastic process (J_t, S_t), where (J_t) is a finite state Markov chain and (S_t) behaves like a process with stationary and independent increments during the time intervals when (J_t) is in any given state. Moreover, a jump of (J_t) from i to j has a certain probability (depending only on i and j) of triggering a jump of (S_t) at the same time. The size of this jump has a fixed distribution, which depends only on i and j.

To see why the theory of Markov additive processes is relevant for the tandem queue, consider the following process (S_t), defined by

$$S_t = Y_0 + (D_t - E_t), \quad t \ge 0, \tag{1}$$

where (D_t) denotes the departure process from the first queue and (E_t) is a Poisson process with intensity μ_2, independent of (D_t). It is not difficult to see that (X_t, S_t) is a Markov additive process. Namely, during intervals where (X_t) is constant, (S_t) behaves like a pure death process with rate μ_2. Moreover, a downward jump of (X_t) triggers (at the same time) an upward jump of (S_t) of size 1. Now, setting $X_0 = i$ and $S_0 = Y_0 = 1$, observe that the overflow probability γ_i, as defined in the previous section, is exactly the probability that (S_t) hits level L before hitting level 0. We now have a closer look at the process (S_t). We first consider the case where the first buffer has finite capacity b_1; in this case the state space of the driving process (X_t) is finite and the theory of Asmussen and Rubinstein (1995) carries through.

3.1 Finite First Buffer

For each $s \ge 0$, define the matrix $M_t(s)$ whose (i, j)th element is $\mathbb{E}_i\, \mathrm{e}^{sS_t}\, I_{\{X_t = j\}}$. Notice that $M_t(\cdot)$ is a generalization of the *moment generating function* for ordinary random variables. Let $G(s)$ be the tri-diagonal matrix of dimension $b_1 + 1$, given by $G(s) =$

$$\begin{pmatrix} -\lambda - \mu_2 + \mu_2\,\mathrm{e}^{-s} & \lambda & & & \\ \mu_1\,\mathrm{e}^{s} & -\lambda - \mu_1 - \mu_2 + \mu_2\,\mathrm{e}^{-s} & \lambda & & \\ & \ddots & \ddots & \ddots & \\ & & \mu_1\,\mathrm{e}^{s} & -\mu_1 - \mu_2 + \mu_2\,\mathrm{e}^{-s} & \end{pmatrix}$$

It can be shown that

$$M_t(s) = \mathrm{e}^{t\,G(s)},\ t \ge 0, \tag{2}$$

which follows from the observation that

$$M_t'(s) = G(s)\, M_t(s),$$

for all $t > 0$, where the elements of the matrix $G(s)$ are determined from an infinitesimal analysis. For example, for $i = 1, \ldots, b_1$, as $h \downarrow 0$, we have

$$\begin{aligned}
&\mathbb{E}_i\, \mathrm{e}^{sS_h}\, I_{\{X_h = i-1\}} \\
&\quad = \mathbb{E}_i(\, \mathrm{e}^{sS_h} \mid X_h = i-1)\, \mathbb{P}_i(X_h = i-1) \\
&\quad = \mu_1\, h\, \mathrm{e}^{s} + o(h),
\end{aligned}$$

This shows that $(i, i-1)$th element of the matrix $G(s)$ is equal to $\mu_1\, \mathrm{e}^{s}$. Notice that a downward jump of (X_t) leads to an upward jump of (S_t).

3.2 Change of Measure

Next, we define a change of measure based on the family of matrices $(G(s))$. For any $s \geq 0$, define $\kappa(s) := \log(\mathrm{sp}(M_t(s)))/t$, where $\mathrm{sp}(M_t(s))$ denotes the spectral radius (or the maximum eigenvalue) of $M_t(s)$. Using (2) we identify $\kappa(s)$ as the largest positive eigenvalue of $G(s)$. Let $\mathbf{w}(s) = \{w_k(s), 0 \leq k \leq b_1\}$ denote the corresponding right-eigenvector. For any $\theta \geq 0$ and any initial state i for the first buffer, we consider the following change of measure, $\tilde{\mathbb{P}}_i$, under which (X_t, S_t) is a Markov additive process (as defined above), but for which (X_t) has a different Q-matrix given by

$$\tilde{Q} = \Delta^{-1}(\mathbf{w}(\theta))\, G(\theta)\, \Delta(\mathbf{w}(\theta)) - \kappa(\theta)\, I, \tag{3}$$

and (S_t) has death rate

$$\tilde{\mu}_2 = \mu_2\, \mathrm{e}^{-\theta}. \tag{4}$$

Here, we have used the notation $\Delta(\mathbf{a})$ to denote a diagonal matrix with entries corresponding to a vector $\mathbf{a}$. Notice that $\tilde{Q}$ is a genuine Q-matrix. Writing out (3), we find that the so-called *conjugate* arrival and service rates of the first queue are given by

$$\tilde{\lambda}(k) = G_{k,k+1}(\theta)\, \frac{w_{k+1}(\theta)}{w_k(\theta)},\ k = 0, 1, \ldots, b_1 - 1, \tag{5}$$

$$\tilde{\mu}_1(k) = G_{k,k-1}(\theta)\, \frac{w_{k-1}(\theta)}{w_k(\theta)},\ k = 1, 2, \ldots, b_1. \tag{6}$$

Note that the conjugate rates depend on k, the content of the first buffer.

It can be shown (for example, by following a particular trajectory) that the likelihood ratio corresponding to $\tilde{\mathbb{P}}_i$ over an interval $[0, T]$ is given by

$$W_T(\theta) = \frac{w_i(\theta)}{w_{X_T}(\theta)}\, \mathrm{e}^{-\theta\, S_T + T\, \kappa(\theta)}. \tag{7}$$

The deterministic time T in (7) may be replaced with a stopping time depending on the history of (X_t, S_t). Let τ be the first time at which (S_t) hits level L or level 0, then (starting with i in the first buffer)

$$\begin{aligned}
\gamma_i &= \mathbb{E}_i\, I_{\{S_\tau = L\}} = \tilde{\mathbb{E}}_i\, W_\tau(\theta)\, I_{\{S_\tau = L\}} \\
&= \tilde{\mathbb{E}}_i\, \frac{w_i(\theta)}{w_{X_\tau}(\theta)}\, \mathrm{e}^{-\theta\, L + \tau\kappa(\theta)}\, I_{\{S_\tau = L\}}.
\end{aligned}$$

The above equation shows how we may estimate γ_i under the new measure. The optimal exponential change of measure for importance sampling is obtained by setting $\theta = \theta^*$, such that $\kappa(\theta^*) = 0$. Under this change of measure,

$$\gamma_i = \tilde{\mathbb{E}}_i\, \frac{w_i(\theta^*)}{w_{X_\tau}(\theta^*)}\, \mathrm{e}^{-\theta^* L}\, I_{\{S_\tau = L\}}. \tag{8}$$

Now, if

$$\inf_L \tilde{\mathbb{P}}_i(S_\tau = L) = \delta, \tag{9}$$

for some strictly positive constant δ, then (8) gives the following lower and upper bounds for γ_i:

$$\delta\, \mathrm{e}^{-\theta^* L}\, \frac{w_i(\theta^*)}{\max_k w_k(\theta^*)} \leq \gamma_i \leq \mathrm{e}^{-\theta^* L}\, \frac{w_i(\theta^*)}{\min_k w_k(\theta^*)}.$$

Since we also have

$$\tilde{\mathbb{E}}_i\, W_\tau^2(\theta^*) \leq\, \mathrm{e}^{-2\theta^* L}\, \frac{w_i(\theta^*)}{\min_k w_k(\theta^*)},$$

it follows that simulation under θ^* yields a *bounded relative error* if (9) holds (or, equivalently, if a lower bound on γ_i decays no faster than $\mathrm{e}^{-\theta^* L}$.) Empirical results in Section 4 support our claim of bounded relative error. Equations (8) and (9) imply that θ^* is the *exponential decay rate* of γ_i; the corresponding *geometric decay rate* is denoted by η.

We now focus on the eigenvector $\mathbf{w}(\theta^*)$, which we simply denote by $\mathbf{w}$ with entries $\{w_k, 0 \leq k \leq b_1\}$. We normalize $\mathbf{w}$ such that $w_0 = 1$. Given the tri-diagonal form of $G(\theta^*)$ it is easy to see that

$$w_1 = (\lambda + \mu_2 - \mu_2\, \eta)/\lambda\ ,$$

$$w_{k+2} + a_1\, w_{k+1} + a_2\, w_k = 0\ ,\quad k = 0, \ldots, b_1 - 2,$$

where $a_1 = -(\lambda + \mu_1 + \mu_2 - \mu_2\, \eta)/\lambda$ and $a_2 = \mu_1/(\lambda\, \eta)$.

These equations completely specify w_{b_1} in terms of η. However, we also have the boundary condition $w_{b_1}(-\mu_1 - \mu_2 + \mu_2\eta) + w_{b_1-1}\, \mu_1/\eta = 0$. This extra equation enables us to determine η. The following characteristic equation

plays a significant role in the evaluation of the change of measure:

$$z^2 + a_1 z + a_2 = 0, \tag{10}$$

with a_1 and a_2 as defined above.

If the first server is the bottleneck ($\mu_1 < \mu_2$), then Equation (10) has two complex solutions $z\,\mathrm{e}^{\pm i\phi}$, and the eigenvector $\mathbf{w}$ is therefore given by

$$w_k = z^k\,(\cos(k\phi) + c\,\sin(k\phi)), \quad k = 0, \ldots, b_1,$$

with $c = (w_1/z - \cos(\phi))/\sin(\phi)$. It follows from Equations (4), (5) and (6) that the conjugate rates are given by: $\tilde{\mu}_2 = \mu_2\,\eta$,

$$\tilde{\lambda}(k) = \lambda z\,\frac{(\cos((k+1)\phi) + c\,\sin((k+1)\phi))}{(\cos(k\phi) + c\,\sin(k\phi))}, \quad k = 0, \ldots, b_1 - 1,$$

$$\tilde{\mu}_1(k) = \frac{\mu_1}{\eta z}\,\frac{(\cos((k-1)\phi) + c\,\sin((k-1)\phi))}{(\cos(k\phi) + c\,\sin(k\phi))}, \quad k = 1, \ldots, b_1.$$

If the second server is the bottleneck ($\mu_2 < \mu_1$), then Equation (10) has two real solutions, say, z_1 and z_2. The eigenvector $\mathbf{w}$ is therefore given by

$$w_k = c_1\,z_1^k + c_2\,z_2^k, \quad k = 0, \ldots, b_1,$$

with c_1 and c_2 as determined from the two equations: $w_0 = c_1 + c_2 = 1$ and $w_1 = c_1 z_1 + c_2 z_2 = (\lambda + \mu_2 - \mu_2\,\eta)/\lambda$. The corresponding conjugate rates are determined from Equations (4), (5) and (6):

$$\tilde{\mu}_2 = \mu_2\,\eta\,,$$

$$\tilde{\lambda}(k) = \lambda\,\frac{c_1\,z_1^{k+1} + c_2\,z_2^{k+1}}{c_1\,z_1^k + c_2\,z_2^k}, \quad k = 0, \ldots, b_1 - 1,$$

$$\tilde{\mu}_1(k) = \frac{\mu_1}{\eta}\,\frac{c_1\,z_1^{k-1} + c_2\,z_2^{k-1}}{c_1\,z_1^k + c_2\,z_2^k}, \quad k = 1, 2, \ldots, b_1.$$

3.2.1 Remark 1

If the service rates at both nodes are equal (i.e., $\mu_1 = \mu_2$) and the first buffer is finite, then it can be shown that Equation (10) has two complex solutions $z\,\mathrm{e}^{\pm i\phi}$. Therefore, this is a special case of that in which the first node is the bottleneck, and the conjugate rates can be determined similarly.

3.3 Infinite First Buffer

The theory of Markov additive processes as described above only holds for driving Markov process with a finite state space. In particular, we cannot apply the change of measure derived in Section 3.1 to the estimation of γ_i in the case where the first buffer has infinite capacity. However, by reasoning analogously we obtain a similar change of measure; in fact, it has somewhat simpler form than that for the case with finite first buffer.

Formally, when the first buffer has infinite capacity, the matrix $G(\theta)$ of Section 3.1 becomes an infinite dimensional matrix of the same tri-diagonal form. Putting, $w_0 = 1$, $w_1 = (\lambda + \mu_2 - \mu_2\,\eta)/\lambda$, and

$$w_{k+2} + a_1\,w_{k+1} + a_2\,w_k = 0, \quad k = 0, 1, \ldots,$$

where $a_1 = -(\lambda + \mu_1 + \mu_2 - \mu_2\,\eta)/\lambda$ and $a_2 = \mu_1/(\lambda\,\eta)$, we see that $\mathbf{w}$ is completely specified by the geometric decay rate $\eta = \mathrm{e}^{-\theta^*}$. However, in this case we have no boundary condition to obtain η. The determination of η depends on which server is the bottleneck.

If the first server is the bottleneck ($\mu_1 < \mu_2$), then η is such that the characteristic equation (10) has only one solution, say, z. (The reader may verify this by considering the approximate model where the capacity b_1 of the first buffer is large but finite.) Consequently, the eigenvector $\mathbf{w}$ is of the form

$$w_k = z^k(1 + c\,k), \quad k = 0, 1, 2, \ldots,$$

with $c = (\lambda + \mu_2 - \mu_2\,\eta)/(\lambda z) - 1$. The conjugate rates (as defined in Section 3.1) also follow from Equations (4), (5) and (6): $\tilde{\mu}_2 = \mu_2\,\eta$,

$$\tilde{\lambda}(k) = \lambda\,z\,\frac{1 + c\,(k+1)}{1 + c\,k}, \quad k = 0, 1, \ldots,$$

$$\tilde{\mu}_1(k) = \frac{\mu_1}{\eta\,z}\,\frac{1 + c\,(k-1)}{1 + c\,k}, \quad k = 1, 2, \ldots.$$

Empirical results in Section 4 indicate that when the first buffer is infinite and is the bottleneck, the change of measure proposed above yields estimates with relative error which is (asymptotically) linear in the overflow level L; a formal proof is not yet available.

3.3.1 Remark 2

By using some algebra it is not difficult to show that if the first server is the bottleneck then η is a root of the cubic equation

$$-4\lambda\mu_1 + (\lambda^2 + 2\lambda\mu_1 + \mu_1^2 + 2\lambda\mu_2 + 2\mu_1\mu_2 + \mu_2^2)\,\eta \\ -2\,\mu_2\,(\lambda + \mu_1 + \mu_2)\,\eta^2 + \mu_2^2\,\eta^3 = 0. \qquad (11)$$

Similarly, z can be shown to be the root of

$$\mu_1\mu_2 + (-\lambda^2 - \lambda\mu_1 - \lambda\mu_2)\,z^2 + 2\lambda^2\,z^3 = 0. \qquad (12)$$

If the second server is the bottleneck ($\mu_2 < \mu_1$), then the conjugate rates are even simpler in form. In this case, $w_k = z^k$, $k \geq 0$, with $z = 1/\eta = \mu_2/\lambda$. (This can again be verified by considering the approximate model with large b_1.) The corresponding conjugate rates are: $\tilde{\lambda} = \mu_2$, $\tilde{\mu}_1 = \mu_1$ and $\tilde{\mu}_2 = \lambda$, i.e., we interchange the arrival rate and the smallest service rate. Empirical results in Section 4 indicate that this change of measure yields estimates with bounded relative error. Note that this is consistent with the optimal change of measure obtained from large deviations analysis of other, but related, overflow probabilities in queueing networks (see, e.g., Parekh and Walrand 1989, Frater and Anderson 1989).

3.3.2 Remark 3

For an infinite first buffer, as μ_1 approaches μ_2 from below, it can be shown that η and z from Equations (11) and (12) approach λ/μ_2 and μ_2/λ, respectively. Therefore, when the service rates are equal, the conjugate rates are obtained by interchanging the arrival rate and the service rates (i.e., $\tilde{\lambda} = \mu_1 = \mu_2$ and $\tilde{\mu_1} = \tilde{\mu_2} = \lambda$.) Empirical results in Section 4 indicate that this change of measure (which is the commonly used heuristic) yields estimates with relative error that is (asymptotically) bounded linearly in L. This agrees with observations made in the literature (see, e.g., Glasserman and Kou 1995, Heidelberger 1995) that the above change of measure is less effective when the service rates are equal.

4 EXAMPLES

We give four concrete examples of the tandem Jackson network with two servers. In the first example, we consider a system in which the second server is the bottleneck. In the second and third examples, the first server is the bottleneck. The interesting case of equal service rates is considered in the fourth example. We are interested in the estimation of the overflow probability in the second buffer $\gamma = \gamma_1$ (i.e., starting from $X_0 = 1$ and $Y_0 = 1$), for both cases: finite and infinite first buffer.

In all the experimental results presented here, the same number of replications, namely, 10^6, is used to obtain each estimate (using importance sampling). The actual simulation effort, however, increases slightly for higher overflow levels. For each estimate in Tables 1,2,3 and 4, we also include its relative error RE (standard deviation divided by the mean) and its invariance constant IC (assuming that $\gamma_1 \propto e^{-\theta^* L}$, IC is the constant of proportionality.)

For the tandem Jackson network being considered, numerical values of the overflow probabilities can be obtained using the algorithm outlined in Garvels and Kroese (1999). For the purpose of validation, these values are also listed in the tables.

4.1 Example 1 ($\lambda = 1, \mu_1 = 4$ and $\mu_2 = 2$)

The second server is the bottleneck.

For a finite first buffer, $b_1 = 9$, the geometric decay rate η of the second buffer is approximately[1] 0.49967. Equation (10) has two real solutions $z_1 = 2.00198$ and $z_2 = 3.99868$, and the eigenvector $\mathbf{w}$ is given by

$$w_k = c_1\,z_1^k + c_2\,z_2^k, \quad k = 0, \ldots, b_1,$$

with $c_1 = 1.00066$ and $c_2 = -0.00066$. This leads to a change of measure which is very close to interchanging the arrival rate and the slowest (second) service rate, i.e., $\tilde{\lambda} \approx 2$, $\tilde{\mu}_1 \approx 4$ and $\tilde{\mu}_2 \approx 1$.

For an infinite first buffer, we find that $\eta = 1/2$, and the eigenvector $\mathbf{w}$ is given by

$$w_k = 2^k, \quad k = 0, 1, 2, \ldots .$$

This leads to $\tilde{\lambda} = 2$, $\tilde{\mu}_1 = 4$ and $\tilde{\mu}_2 = 1$, i.e., interchanging the arrival and the slowest service rates.

The resulting estimates and their relative errors are displayed in Table 1. For both cases, finite and infinite first buffer, the estimates (for an increasing overflow level, L) exhibit bounded relative error. This is consistent with well established theoretical and empirical results (see, e.g., Parekh and Walrand 1989, Frater and Anderson 1989).

4.2 Example 2 ($\lambda = 1, \mu_1 = 2$ and $\mu_2 = 3$)

The first server is the bottleneck.

For a finite first buffer, $b_1 = 9$, we find that $\eta = 0.28898$. Equation (10) has two complex solutions $z\,e^{\pm i\phi}$, with $z = 2.63077$ and $\phi = -0.22144$. The eigenvector $\mathbf{w}$ is therefore given by

$$w_k = z^k\,(\cos(k\phi) + c\,\sin(k\phi)), \quad k = 0, \ldots, b_1,$$

[1] All numerical values are rounded to 5 significant digits.

with $c = (w_1/z - \cos(\phi))/\sin(\phi) = -0.98048$. The conjugate rates are determined as in Section 3.1.

For an infinite first buffer, η is a solution of (11):

$$-8 + 36\,\eta - 36\,\eta^2 + 9\,\eta^3 = 0,$$

and z is a solution of (12):

$$3 - 3\,z^2 + z^3 = 0.$$

The numerical values are $\eta = 0.31194$ and $z = 2.53209$, and the eigenvector $\mathbf{w}$ satisfies

$$w_k = z^k(1 + c\,k), \quad k = 0, 1, \ldots,$$

with $c = w_1/z - 1 = 0.21014$. The change of measure is obtained accordingly (as in Section 3.3).

The resulting estimates and their relative errors are displayed in Table 2. For a finite first buffer, the estimates (for an increasing overflow level, L) exhibit bounded relative error. When the first buffer is infinite, the estimates are accurate but their relative error increases linearly with L.

4.3 Example 3 ($\lambda = 1, \mu_1 = 4/3$ and $\mu_2 = 2$)

The first server is the bottleneck.

Using the same procedure as in the above example, for a finite first buffer, $b_1 = 9$, we find that $\eta = 0.41467$. Equation (10) has two complex solutions $z\,\mathrm{e}^{\pm i\phi}$, with $z = 1.79315$ and $\phi = -0.21466$, and the eigenvector $\mathbf{w}$ is given by

$$w_k = z^k\,(\cos(k\phi) + c\,\sin(k\phi)), \quad k = 0, \ldots, b_1,$$

with $c = (w_1/z - \cos(\phi))/\sin(\phi) = -1.09603$.

For an infinite first buffer, we find (as in the above example) that $\eta = 0.45520$, $z = 1.71147$, and the eigenvector $\mathbf{w}$ is given by

$$w_k = z^k(1 + c\,k), \quad k = 0, 1, \ldots,$$

with $c = w_1/z - 1 = 0.22094$.

The resulting estimates and their relative errors are displayed in Table 3. As in the above example, for a finite first buffer, the estimates exhibit bounded relative error. When the first buffer is infinite, the relative error increases linearly with L.

4.4 Example 4 ($\lambda = 1, \mu_1 = 2$ and $\mu_2 = 2$)

Equal service rates at both nodes.

As noted in Remark 1 (Section 3.1), for a finite first buffer, $b_1 = 9$, we follow the same procedure as if the first server is the bottleneck. We find that $\eta = 0.47847$. Equation (10) has two complex solutions $z\,\mathrm{e}^{\pm i\phi}$, with $z = 2.0445$ and $\phi = -0.15$, and the eigenvector $\mathbf{w}$ is given by

$$w_k = z^k\,(\cos(k\phi) + c\,\sin(k\phi)), \quad k = 0, \ldots, b_1,$$

with $c = (w_1/z - \cos(\phi))/\sin(\phi) = -0.0704$.

For an infinite first buffer, the conjugate rates are obtained by exchanging the arrival and service rates (see Remark 3 in Section 3.3).

The resulting estimates and their relative errors are displayed in Table 4. Here too, for a finite first buffer, the estimates (for an increasing overflow level, L) exhibit bounded relative error. When the first buffer is infinite, the relative error increases linearly with L.

4.5 Remark

According to the theory in Section 3, the derived change of measure holds for any starting state, provided that $Y_0 \geq 1$.

Table 1: Estimates of the overflow probability in Example 1. (The second server is the bottleneck.)

(λ, μ_1, μ_2)	L	$\hat{\gamma}$ (IS)	RE (IS)	IC	γ (Numerical)
(1, 4, 2)	20	1.43e-6	0.11%	1.516	1.428e-6
$b_1 = 9$	25	4.44e-8	0.11%	1.514	4.446e-8
	50	1.30e-15	0.11%	1.515	1.303e-15
	60	1.27e-18	0.11%	1.519	1.264e-18
	100	1.12e-30	0.11%	1.516	1.120e-30
(1, 4, 2)	20	1.43e-6	0.11%	1.500	1.432e-6
$b_1 = \infty$	25	4.47e-8	0.11%	1.500	4.472e-8
	50	1.33e-15	0.11%	1.500	1.332e-15
	60	1.30e-18	0.11%	1.500	1.301e-18
	100	1.18e-30	0.11%	1.500	1.183e-30

Table 2: Estimates of the overflow probability in Example 2. (The first server is the bottleneck.)

(λ, μ_1, μ_2)	L	$\hat{\gamma}$ (IS)	RE (IS)	IC	γ (Numerical)
(1, 2, 3)	20	1.88e-11	0.24%	1.138	1.878e-11
$b_1 = 9$	25	3.75e-14	0.24%	1.127	3.759e-14
	50	1.24e-27	0.24%	1.123	1.247e-27
	60	5.06e-33	0.24%	1.128	5.063e-33
	100	1.38e-54	0.24%	1.128	1.377e-54
(1, 2, 3)	20	2.05e-11	0.49%	0.270	2.048e-11
$b_1 = \infty$	25	4.63e-14	0.56%	0.206	4.610e-14
	50	4.32e-27	0.87%	0.086	4.305e-27
	60	2.94e-32	0.98%	0.067	2.956e-32
	100	8.59e-53	1.38%	0.034	8.595e-53

Table 3: Estimates of the overflow probability in Example 3. (The first server is the bottleneck.)

(λ, μ_1, μ_2)	L	$\hat{\gamma}$ (IS)	RE (IS)	IC	γ (Numerical)
(1, 4/3, 2)	20	1.15e-8	0.23%	0.508	1.150e-8
$b_1 = 9$	25	1.40e-10	0.23%	0.506	1.405e-10
	50	3.89e-20	0.23%	0.505	3.887e-20
	60	5.83e-24	0.23%	0.503	5.843e-24
	100	2.99e-39	0.23%	0.503	2.982e-39
(1, 4/3, 2)	20	1.35e-8	0.52%	0.092	1.348e-8
$b_1 = \infty$	25	1.96e-10	0.60%	0.069	1.966e-10
	50	2.19e-19	0.95%	0.027	2.203e-19
	60	6.50e-23	1.07%	0.021	6.541e-23
	100	6.78e-37	1.52%	0.010	6.790e-37

Table 4: Estimates of the overflow probability in Example 4. (Equal service rates at both nodes.)

(λ, μ_1, μ_2)	L	$\hat{\gamma}$ (IS)	RE (IS)	IC	γ (Numerical)
(1, 2, 2)	20	2.55e-7	0.19%	0.646	2.557e-7
$b_1 = 9$	25	6.40e-9	0.19%	0.645	6.397e-9
	50	6.32e-17	0.19%	0.643	6.340e-17
	60	3.99e-20	0.19%	0.645	3.987e-20
	100	6.23e-33	0.19%	0.645	6.235e-33
(1, 2, 2)	20	2.77e-7	0.29%	0.290	2.787e-7
$b_1 = \infty$	25	7.68e-9	0.31%	0.258	7.661e-9
	50	1.56e-16	0.38%	0.176	1.559e-16
	60	1.39e-19	0.40%	0.160	1.382e-19
	100	9.63e-32	0.46%	0.122	9.618e-32

When $Y_0 = 0$ (i.e., starting with an empty second buffer), the process (Y_t) stays at level 0 for a while before taking off to higher levels. For any such starting state (for example, an empty system), empirical results (not included here) show that the same change of measure yields estimates with a bounded relative error, except when the first server is the bottleneck and its buffer is infinite. In this case, the relative error increases sharply with L, suggesting that a different (exponential) change of measure to be used along the boundary (while $(Y_t) = 0$) should perhaps be sought.

Indeed, when we use the conditional transition probabilities (given an overflow of the second buffer) as a change of measure on $(Y_t) = 0$, the relative error of the resulting estimates increases linearly (but slowly) with L. Unfortunately, determining the conditional transition probabilities along the boundary $(Y_t) = 0$ is of the same order of complexity as determining the probability we are trying to estimate.

5 CONCLUSIONS

This paper represents an introduction and a preliminary study of a new approach for the analysis and efficient simulation of rare events in queueing networks. We have introduced a MAP (Markov additive process) representation of a two-node tandem Jackson network. An exponential change of measure is used in an importance sampling procedure to estimate the probability of overflow in the second buffer. The optimal tilting parameter and the corresponding *conjugate* rates are determined by solving an appropriate eigenvalue problem. Unlike heuristics proposed and studied in the literature, our approach yields conjugate rates which, in general, depend on the content of the first buffer. Importance sampling simulations with this change of measure yield asymptotically efficient estimators, with a bounded relative error, except when the first node is the bottleneck and its buffer is infinite; in this case the relative error is bounded linearly in the overflow level. Further research is now being conducted to examine the feasibility and effectiveness of this new approach for other rare events of interest in Jackson networks.

REFERENCES

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

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.

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.

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

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

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

Garvels, M.J.J., and D.P. Kroese. 1999. On the entrance distribution in RESTART simulation. In *Proceedings of the Second Workshop on Rare Event Simulation (RESIM'99)*, Enschede, The Netherlands, 65–88.

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

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

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

Kroese, D.P., and V.F. Nicola. 1998. Efficient simulation of backlogs in fluid flow lines. *Int. J. Electron. Commun. AEÜ* 52 (3): 165–171.

Ney, P., and E. Nummelin. 1987. Markov additive processes I. Eigenvalue properties and limit theorems. *The Annals of Probability* 15 (2): 561–592.

Ney, P., and E. Nummelin. 1987. Markov additive processes II. Large deviations. *The Annals of Probability* 15 (2): 593–609.

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

Tsoucas, P. 1992. Rare events in series of queues. *J. Appl. Probab.* 29: 168–175.

AUTHOR BIOGRAPHIES

DIRK P. KROESE holds a Ph.D degree in Mathematical Sciences from the University of Twente, The Netherlands, where he has been a Lecturer from 1990 to 1998. He has held visiting staff positions at Princeton University, U.S.A. and the University of Melbourne, Australia. Currently he is a Research Fellow at the Teletraffic Research Centre in Adelaide, Australia. His interests include queueing theory,

performance analysis, efficient simulation, point processes and fluid queues.

VICTOR F. NICOLA holds the Ph.D. degree in computer science from Duke University, North Carolina, the B.S. and the M.S. degrees in electrical engineering from Cairo University, Egypt, and Eindhoven University of Technology, The Netherlands, respectively. From 1979, he held faculty and research staff positions at Eindhoven University and at Duke University. In 1987, he joined IBM Thomas J. Watson Research Center, Yorktown Heights, New York, as a Research Staff Member. Since 1993, he has been an Associate Professor at the Department of Electrical Engineering, University of Twente, The Netherlands. His research interests include performance and reliability modeling, fault-tolerance, queueing theory, analysis and simulation methodologies, with applications to computer and telecommunication systems.