Efficient simulation of a tandem Jackson network

Dirk P. Kroese* Victor F. Nicola†

Abstract

The two-node tandem Jackson network serves as a convenient reference model for the analysis and testing of different methodologies and techniques in rare event simulation. In this paper we consider a new approach to efficiently estimate the probability that the content of the second buffer exceeds some high level L before it becomes empty, starting from a given state. The approach is based on a Markov additive process representation of the buffer processes, leading to an exponential change of measure to be used in an importance sampling procedure. Unlike changes of measures proposed and studied in recent literature, the one derived here is a function of the content of the first buffer. We prove that when the first buffer is finite this method yields asymptotically efficient simulation for any set of arrival and service rates. In fact, the relative error is bounded independent of the level L; a new result which is not established for any other known method. When the first buffer is infinite, we propose a natural extension of the exponential change of measure for the finite buffer case. In this case, the relative error is shown to be bounded (independent of L) only when the second server is the bottleneck; a result which is known to hold for some other methods derived through large deviations analysis. When the first server is the bottleneck, experimental results using our method seem to suggest that the relative error is bounded linearly in L.

Keywords: Tandem Jackson network, Markov additive processes, rare event simulation, importance sampling, bounded relative error, orthogonal polynomials.

*Department of Mathematics, The University of Queensland, Brisbane 4072, Australia. Email: kroese@maths.uq.edu.au

†Department of Electrical Engineering, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands. Email: nicola@cs.utwente.nl

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 difficulties are 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., [20], [2], [7], [8], [22], [10]). Exact large deviations analysis leading to an asymptotically efficient change of measure is quite difficult. Instead, a heuristic change of measure is suggested in [20], 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., [2], [8]). However, analysis in [10] and counter examples in [11] 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 an asymptotically efficient exponential change of measure is obtained by interchanging the arrival rate and the service rate at this (bottleneck) node; the service rates at all other nodes are kept unchanged (see, e.g., [20], [2] and [7]). However, this change of measure is not 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., [4] and [6]. (This class essentially resembles a feed-forward fluid-flow network.) To the best of our knowledge, an analogous approach 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 formulation of the problem using the theory of Markov additive processes (MAP) (see, e.g., [1], [17], [18]). By exponentially tilting an appropriate MAP representation of the system, we find and implement 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. When the first buffer is finite, we formally prove that the proposed simulation procedure

yields estimates with a relative error that is bounded independently of the overflow level. This result is much stronger than asymptotic efficiency, which is shown not to hold for other known methods (see [10]). When the first buffer is infinite, we propose a change of measure which is a natural extension of the change of measure for the case of finite first buffer. Using the theory of orthogonal polynomials [5] we show that this leads to two types of behaviour for the simulation procedure. When the second buffer is the bottleneck, we again have estimates with bounded relative error. However, when the first buffer is the bottleneck, experimental simulation results suggest that the relative error is (asymptotically) linearly bounded in the overflow level.

A related approach for the simulation of backlogs in fluid flow lines is considered in [14]. More recently, an adaptive importance sampling approach is used to determine (approximately) the 'optimal' state-dependent change of measure for rare event simulation in Jackson queueing networks [3] (optimality here is in the sense of minimizing the variance of the estimator or a related cross-entropy distance measure [21]).

In Section 2 we give some preliminaries and a Markov additive process (MAP) representation of the two-node tandem Jackson network. For this MAP, an exponential change of measure is introduced in Section 3. In Sections 3.1 and 3.2, appropriate changes of measure are derived for finite and infinite first buffer, respectively. A proof of asymptotic efficiency of the corresponding importance sampling procedure in the case of finite first buffer is also presented. Experimental studies using some concrete examples are carried out in Section 4 to examine the (asymptotic) efficiency of the developed importance sampling estimator. Conclusions and related future research are given in Section 5. The Appendix deals with the spectral theory of certain triangular matrices and their corresponding orthogonal polynomials. The lemmas listed there are the main ingredients for the proof of asymptotic efficiency of the proposed simulation procedures.

2 Markov additive process representation

A Markov additive process in continuous time is a stochastic process (J_t, Z_t), where (J_t) is a Markov chain with denumerable state space and (Z_t) has stationary and independent increments during the time intervals when (J_t) is in any given state. That is, given J_t has not changed in the interval (t_1, t_2), then for any $t_1 < s_1 < \ldots < s_n < t_2$, the increments $Z_{s_2} - Z_{s_1}, \ldots, Z_{s_n} - Z_{s_{n-1}}$ are mutually independent, and the total increment during the interval $[t_1, t_2]$ depends on t_1 and t_2 only through the difference $t_2 - t_1$. 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 (Z_t) at the same time. The size of the jump in the process (Z_t) has a fixed distribution, which depends only on i and j. For theoretical properties and limit theorems of MAPs, the reader is referred to [17] and [18].

A Markov additive process (J_t, Z_t) is characterized by the family of matrices $(M_t(s), t \geqslant 0)$, where the (i,j)-th element of $M_t(s)$ is given by

$$[M_t(s)]_{ij} = \mathbb{E}_i \, \mathrm{e}^{s(Z_t - Z_0)} \, I_{\{J_t = j\}} \, ,$$

where $\mathbb{E}_i$ denotes the expectation operator given the initial MAP state $J_0 = i$. Notice that $M_t(\cdot)$ is a generalization of the *moment generating function* for ordinary random variables. Moreover, we have

$$\begin{aligned} \mathbb{E}_i \, \mathrm{e}^{s(Z_{t+h} - Z_0)} \, I_{\{J_{t+h} = j\}} &= \sum_k \mathbb{E}_i \, \mathrm{e}^{s(Z_{t+h} - Z_0)} \, I_{\{J_t = k\}} \, I_{\{J_{t+h} = j\}} \\ &= \sum_k \mathbb{E}_i \, \mathrm{e}^{s(Z_t - Z_0)} \, I_{\{J_t = k\}} \, \mathbb{E}_i \left[\mathrm{e}^{s(Z_{t+h} - Z_t)} \, I_{\{J_{t+h} = j\}} \, | \, J_t = k \right] \\ &= \sum_k [M_t(s)]_{ik} \; \mathbb{E}_k \, \mathrm{e}^{s(Z_h - Z_0)} \, I_{\{J_h = j\}}. \end{aligned}$$

Consequently, if for all k and j

$$[A(s)]_{kj} := \lim_{h \downarrow 0} \frac{1}{h} \, \mathbb{E}_k \left[\mathrm{e}^{s(Z_h - Z_0)} \, I_{\{J_h = j\}} - \delta_{kj} \right]$$

exists (we have used δ in the usual notation of Dirac), then

$$\frac{d}{dt} M_t(s) = M_t(s) \, A(s), \; t \geqslant 0,$$

with $M_0(s) = I$ (the identity matrix). It follows that

$$M_t(s) = \mathrm{e}^{t \, A(s)}, \; t \geqslant 0. \tag{1}$$

The matrix $A(s)$ is identified as the MAP (infinitesimal) generator.

Two-node tandem network

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 an 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 may be finite or infinite; in fact, we will consider both cases. Let X_t and Y_t denote the number of customers in the first and second queue at time t, respectively. Let $\mathbb{P}_i$ denote the probability measure under which (X_t) starts from i at time 0 (i.e., $X_0 = i$, $i \geqslant 0$); and let $\mathbb{E}_i$ denote the corresponding expectation operator. In Section 3 we consider various changes of measure; $\tilde{\mathbb{P}}_i$ denotes any such measure for which (X_t) starts at i, $\tilde{\mathbb{E}}_i$ denotes the corresponding expectation operator. Assuming that the second buffer is initially non-empty, say, $Y_0 = 1$, we are interested in the probability that, starting from $(X_0, Y_0) = (i, 1)$, the content of the second buffer hits some high level $L \in \mathbb{N}$ before hitting 0. We denote this probability by γ_i and refer to it as the *overflow probability* of the second buffer, given that the initial number of customers in the first queue is i.

Consider the estimation of the steady-state probability that the content of the second buffer is at or exceeds a given level, L. One way to estimate this probability is by observing a large number of 'true' regeneration cycles (e.g., with an empty system as a regeneration point). Then, an estimate of the steady-state overflow probability is the ratio of two (possibly independent) estimators [13]; one for the expected time at or above level L in a cycle (the numerator), and the other for the expected cycle time (the denominator). The numerator could be estimated independently and more efficiently using importance sampling. A disadvantage is that true regenerations may not be sufficiently frequent (i.e., the number of events in a regeneration cycle is excessive), thus causing the simulation to be less efficient.

An alternative way to estimate the steady-state overflow probability at the second buffer is to observe a large number of 'pseudo' regenerations. Here, a pseudo regeneration is defined to be a sample path between two successive entries to the set of states $(X_t, Y_t) = (\cdot, 1)$ due to a departure from node 1. These pseudo regenerations are not i.i.d.; however, a ratio estimator of the overflow probability (similar to that for true regenerations) still holds [13], provided that the method of batch means is used to form a valid confidence interval (each batch consists of a sufficiently large number of pseudo cycles). Such an estimator is sometimes referred to as the A-cycle approach (see, for example, [19]). An advantage is that pseudo regenerations are more frequent than true regenerations, thus leading to less costly simulation. Here too, importance sampling can be used to efficiently estimate the expected time at or above level L in a pseudo regeneration cycle. The change of measure is the same as that used to efficiently estimate γ_i (defined above) for any $i = 0, 1, 2, \ldots$, and which is considered in this paper.

As will be explained in Remark 1, the choice of the probability γ_i (to be estimated) is also made to facilitate the exposition in this paper; other probabilities of interest may be estimated using the same approach.

The content of the second buffer (Y_t) can be expressed as

$$Y_t = Y_0 + (D_t - O_t), \;\; t \geqslant 0,$$

where (D_t) denotes the departure process from the first queue and (O_t) denotes the departure process from the second queue. To see why the theory of Markov additive processes is relevant for the tandem queue, consider the process (S_t), defined by

$$S_t = Y_0 + (D_t - E_t), \;\; t \geqslant 0, \tag{2}$$

where (E_t) is a Poisson process with intensity μ_2, independent of (D_t). Given that $Y_0 > 0$, denote by τ_0 the first time Y_t hits 0, i.e.,

$$\tau_0 := \inf_t \{t > 0 \, : \, Y_t = 0\} \, .$$

Clearly, the process (S_t) is identical to (Y_t) in the interval $[0, \tau_0]$. Hence, starting at $t = 0$, the process of interest (X_t, Y_t) is identical to the process (X_t, S_t) *only* until (S_t) hits 0 for the first time.

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 as 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 $S_0 = Y_0 = 1$, note 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.

Remark 1 It is important to note that the appropriate MAP representation depends on the probability being considered. For example, one way to remove the restriction $Y_0 > 0$ (in the definition of the probability γ_i) is by considering the MAP (J_t, Y_t) in which the driving process (J_t) is the two-dimensional Markov chain (X_t, Y_t). Note that the additive process (Y_t) is a jump process; the jump size is 0 if (X_t) makes an upward jump, +1 if (X_t) makes a downward jump, and is -1 if (Y_t) makes a downward jump. The same MAP can be used to study the more conventional probability that from some arbitrary starting state (e.g., an empty system), the second buffer hits some high level L before the system empties.
If we are interested in the overflow probability of the overall network population, starting from an empty system and before the system empties again, then an appropriate MAP representation would be $(J_t, X_t + Y_t)$. Here too, (J_t) is the two-dimensional Markov chain (X_t, Y_t). The additive process $(X_t + Y_t)$ is again a jump process; the jump size is +1 if (X_t) makes an upward jump, 0 if (X_t) makes a downward jump, and is -1 if (Y_t) makes a downward jump.
In the above MAPs, the driving process $(J_t) = (X_t, Y_t)$ has the two-dimensional state space in $\mathbb{N} \times \mathbb{N}$ when the first and second buffers are infinite. The state space is finite in one dimension (resp. both dimensions) if there is a limit on the capacity

of the first buffer (resp. the total network population). Equation (1) still holds, however, $M_t(s)$ and $A(s)$ are now obtained by unfolding the two-dimensional state space into a one-dimensional state space. Further investigation is underway to use these MAP representations for efficient simulation. □

Other Jackson networks

Other tandem Jackson networks can be treated similarly. For example, consider a 3-node tandem network; we are interested in the overflow probability at the third node, starting from a given state $(i, j, 1)$, where i and j may assume any values. An appropriate MAP representation is one in which the driving process (J_t) is the two-dimensional Markov chain (X_t, Y_t), representing the buffer content at the first and the second node, respectively. More general Jackson networks may be represented similarly; however, depending on the measure of interest, the dimension (and hence the state-space) of the MAP process may increase with the number of nodes, leading to more tedious yet essentially similar mathematical and algorithmic treatment. Needless to say, state-space explosion is an inherent problem that we do not claim to have overcome using our approach.

3 Exponential change of measure

The key to understanding the change of measure that we are going to propose is in [1], where an exponential change of measure for Markov additive processes is discussed in the context of rare event simulation.

3.1 Finite first buffer

We first consider the case where the first buffer has a finite capacity N. In this case the state space of the driving process (X_t) is finite in $\{0, \ldots, N\}$ and the theory in [1] carries through. Consider the MAP (X_t, S_t) defined in Section 2. To construct the corresponding MAP generator (i.e., the matrix $A(s)$ in (1)), we need to determine the infinitesimal expectations $\mathbb{E}_i \left[\mathrm{e}^{s(S_h - S_0)} I_{\{X_h = j\}} - \delta_{ij} \right]$ as $h \downarrow 0$, for all i, j in $\{0, \ldots, N\}$ (note that $S_0 = 1$ and $\delta_{ij} = 0$ for $j \neq i$, as defined in Section 2). For instance, since a downward jump of (X_t) leads to an upward jump of (S_t), it

follows that, for $i = 1, \ldots, N$, as $h \downarrow 0$, we have

$$\begin{aligned}\mathbb{E}_i\left[\mathrm{e}^{s(S_h - S_0)}\, I_{\{X_h = i-1\}} - \delta_{i,i-1}\right] &= \mathbb{E}_i\left[\mathrm{e}^{s(S_h - S_0)} \mid X_h = i-1\right] \mathbb{P}_i(X_h = i-1) \\ &= \mathrm{e}^s\,(\mu_1\, h + o(h)) \\ &= \mu_1\, h\, \mathrm{e}^s + o(h).\end{aligned}$$

Hence, the $(i, i-1)$-th element of the matrix $A(s)$ exists and is equal to $\mu_1\, \mathrm{e}^s$. Other elements of the matrix $A(s)$ can be determined similarly. Therefore, for the MAP (X_t, S_t), (1) holds with $A(s)$ given by the $(N+1, N+1)$-tri-diagonal matrix

$$G_N(s) = \begin{pmatrix} -\lambda - \mu_2 + \mu_2\, \mathrm{e}^{-s} & \lambda & & & \\ \mu_1\, \mathrm{e}^s & -\lambda - \mu_1 - \mu_2 + \mu_2\, \mathrm{e}^{-s} & \lambda & & \\ & \ddots & \ddots & \ddots & \\ & & \mu_1\, \mathrm{e}^s & -\mu_1 - \mu_2 + \mu_2\, \mathrm{e}^{-s} & \end{pmatrix}.$$

Note that the MAP generator $G_N(s)$ is equal to the matrix $\hat{Q}^{(N+1)}(\mathrm{e}^{-s})$, defined in (27) of Appendix A.

Change of measure

Next, we define a change of measure based on the family of matrices $(G_N(s))$. For any $s \geqslant 0$, define $\kappa_N(s) := \log(\mathrm{sp}(M_t(s)))/t$, where $\mathrm{sp}(M_t(s))$ denotes the spectral radius (or the maximum eigenvalue) of $M_t(s)$. Using (1) we identify $\kappa_N(s)$ as the largest positive eigenvalue of $G_N(s)$. Let $\boldsymbol{w}(s) = \{w_k(s), 0 \leqslant k \leqslant N\}$ denote the corresponding right-eigenvector. By the Perron-Frobenius theory of positive matrices, we see that $\boldsymbol{w}(s)$ is a strictly positive vector (if we take one of the elements, e.g., $w_0(s)$, strictly positive).

For any $s \geqslant 0$ 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 also a MAP, but for which the Markov chain (X_t) has a different Q-matrix (infinitesimal generator) given by [1]

$$\tilde{Q}_N(s) \;=\; \Delta^{-1}\left(\boldsymbol{w}(s)\right) G_N(s)\, \Delta\left(\boldsymbol{w}(s)\right) \;-\; \kappa_N(s)\, I, \tag{3}$$

and (S_t) has death rate

$$\tilde{\mu}_2(s) \;=\; \mu_2\, \mathrm{e}^{-s}\,. \tag{4}$$

Here, we have used the notation $\Delta(\mathbf{a})$ to denote a diagonal matrix with entries equal to the elements of a vector $\mathbf{a}$.

Remark 2 Note that the original Q-matrix of (X_t) is the tri-diagonal matrix

$$\begin{pmatrix} -\lambda & \lambda & & \\ \mu_1 & -\lambda-\mu_1 & \lambda & \\ & \ddots & \ddots & \ddots \\ & & \mu_1 & -\mu_1 \end{pmatrix}.$$

To see that $\tilde{Q}_N(s)$ is a genuine Q-matrix, observe that the off-diagonal entries of $\tilde{Q}_N(s)$ are strictly positive, because $\boldsymbol{w}(s)$ is a positive vector. It remains to be proved that the rows of $\tilde{Q}_N(s)$ sum up to 0. Denoting by $\mathbf{1}$ the $(N+1)$-dimensional vector of 1's, and by $\mathbf{0}$ the $(N+1)$-dimensional vector of 0's, we have

$$\tilde{Q}_N(s)\mathbf{1} \;=\; \Delta^{-1}\left(\boldsymbol{w}(s)\right)\, \kappa_N(s)\, \boldsymbol{w}(s) \;-\; \kappa_N(s)\, \mathbf{1} \;=\; \mathbf{0}\,.$$

□

Writing out (3), we find that the so-called *conjugate* arrival and service rates of the first queue are given by

$$\tilde{\lambda}(k;s) \;=\; \lambda\, \frac{w_{k+1}(s)}{w_k(s)}\,, \;\; k = 0, 1, \ldots, N-1, \tag{5}$$

$$\tilde{\mu}_1(k;s) \;=\; \mu_1\, \mathrm{e}^s\, \frac{w_{k-1}(s)}{w_k(s)}\,, \;\; k = 1, 2, \ldots, N. \tag{6}$$

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

Remark 3 Since the rows of $\tilde{Q}_N(s)$ sum up to 0, from (3) we find that

$$\begin{aligned} \kappa_N(s) &= -\lambda - \mu_2 + \tilde{\lambda}(0;s) + \tilde{\mu}_2(s)\,, \\ \kappa_N(s) &= -\lambda - \mu_1 - \mu_2 + \tilde{\lambda}(k;s) + \tilde{\mu_1}(k;s) + \tilde{\mu}_2(s), \;\; k \in \{1, \ldots, N-1\}, \\ \kappa_N(s) &= -\mu_1 - \mu_2 + \tilde{\mu_1}(N;s) - \tilde{\mu}_2(s)\,. \end{aligned}$$

□

Importance sampling

We may estimate γ_i by importance sampling using any of the new measures defined above (each $s > 0$ corresponds to a different measure). Indeed, if we define τ as the first time at which (S_t) hits level L or level 0, then

$$\gamma_i \;=\; \mathbb{E}_i\, I_{\{S_\tau = L\}} \;=\; \tilde{\mathbb{E}}_i\, W_\tau(s)\, I_{\{S_\tau = L\}}\,, \tag{7}$$

where $W_\tau(s)$ is the likelihood ratio corresponding to $\tilde{\mathbb{P}}_i$ over the interval $[0,\tau]$. Two question now arise. First, what is this likelihood ratio? Second, what is the best choice of s under which to perform an importance sampling simulation?

To answer the first question, consider a "trajectory" of (X_t, S_t) in the interval $[0,\tau]$, during which n_τ state transitions have occurred, say, at $t_1, t_2, \ldots, t_{n_\tau}$ (note that $t_{n_\tau} = \tau$, since τ is a stopping time of (X_t, S_t)). These transitions are due to either an arrival to the first queue, a departure from the first queue, or a departure from the second queue. Given that $X_0 = i$, $i \in \{0, \ldots, N\}$, and making use of Remark 3, it can easily be shown that the likelihood ratio due to a transition at time t_1 is given by:

- if arrival to the first node, then

$$\frac{w_i(s)}{w_{i+1}(s)}\, e^{-t_1\,\kappa_N(s)}\,, \quad i = 0, 1, \ldots, N-1$$

- if departure from the first node, then

$$\frac{w_i(s)}{w_{i-1}(s)}\, \mathrm{e}^{-s}\, \mathrm{e}^{-t_1\,\kappa_N(s)}\,, \quad i = 1, 2, \ldots, N$$

- if departure from the second node, then

$$\mathrm{e}^{s}\, \mathrm{e}^{-t_1\,\kappa_N(s)}\,, \quad i = 0, 1, \ldots, N\,.$$

Repeating the above argument at the transition times $t_2, \ldots, t_{n_\tau}$, yields the following likelihood ratio, corresponding to $\tilde{\mathbb{P}}_i$ over the interval $[0,\tau]$,

$$W_\tau(s) \;=\; \frac{w_i(s)}{w_{X_\tau}(s)}\, \mathrm{e}^{-s\,S_\tau\,+\,\tau\,\kappa_N(s)}\,. \tag{8}$$

Having established the likelihood ratio corresponding to trajectories in the interval $[0,\tau]$, we now need to determine an appropriate s under which to carry out the importance sampling procedure. The above likelihood ratio suggests that we should take $s = s_N$ such that $\kappa_N(s_N) = 0$; in fact there is only one such s_N.

Lemma 1 The equation

$$\kappa_N(s) = 0 \tag{9}$$

has exactly one solution s_N in the interval $(0,\infty)$.

Proof. This follows directly from Lemma 4(b), because $\kappa_N(s)$ is the largest eigenvalue of $\hat{Q}^{(N+1)}(\mathrm{e}^{-s})$. □

To simplify the notation in what follows, we abbreviate the vector $\boldsymbol{w}(s_N)$ to $\boldsymbol{w} = (w_0, \ldots, w_N)$. We also abbreviate $\tilde{\lambda}(k; s_N)$ to $\tilde{\lambda}(k)$, $\tilde{\mu_1}(k; s_N)$ to $\tilde{\mu_1}(k)$, and $\tilde{\mu_2}(s_N)$ to $\tilde{\mu_2}$.

We now show that performing importance sampling with $s = s_N$ yields an estimator which is asymptotically efficient.

Theorem 1 Let s_N be the unique solution to (9), then under the measure $\tilde{\mathbb{P}}_i$ defined in (4), (5) and (6), the random variable

$$\Gamma \;:=\; \frac{w_i}{w_{X_\tau}}\,\mathrm{e}^{-s_N\,L}\,I_{\{S_\tau = L\}} \tag{10}$$

has an expectation γ_i and a relative error bounded in L.

Proof. First, from (7) and (8) and the fact that $\kappa_N(s_N) = 0$, we have

$$\gamma_i \;=\; \tilde{\mathbb{E}}_i\,\frac{w_i}{w_{X_\tau}}\,\mathrm{e}^{-s_N\,L}\,I_{\{S_\tau = L\}} \;=\; \tilde{\mathbb{E}}_i\Gamma\,. \tag{11}$$

Next, we claim that under the new measure,

$$\tilde{\lambda}(k) \;>\; \lambda\,,\;\; k = 0, \ldots, N-1, \tag{12}$$

$$\tilde{\mu_1}(k) \;>\; \mu_1\,,\;\; k = 1, \ldots, N, \tag{13}$$

$$\tilde{\mu_2} \;<\; \lambda\,. \tag{14}$$

This follows directly from Lemma 5 of Appendix A, with $\boldsymbol{v}$ having dimension $n = N+1$, $\hat{u}_n = \mathrm{e}^{-s_N}$, and $(v_1, \ldots, v_n)^T = (w_0, \ldots, w_N)^T$. Specifically, (12) follows from Lemma 5 (b) and (5), (13) follows from Lemma 5 (c) and (6), and (14) follows from (31) and (4).

Since $\mu_1 > \lambda$, it follows that, under the new measure, the output rate from the second queue is strictly less than its input rate. Hence, under the new measure, the second queue is unstable and a high level L can be reached before hitting 0 with a probability which is bounded from below by some constant $c > 0$, i.e.,

$$\inf_L\,\tilde{\mathbb{P}}_i(S_\tau = L) \;=\; c\,, \tag{15}$$

where c does not depend on L. Using (15), (11) gives the following lower and upper bounds for γ_i :

$$c\,\mathrm{e}^{-s_N L}\,\frac{w_i}{\max_k w_k} \;\leqslant\; \gamma_i \;\leqslant\; \mathrm{e}^{-s_N L}\,\frac{w_i}{\min_k w_k}\,. \tag{16}$$

Similarly,

$$\tilde{\mathbb{E}}_i\,\Gamma^2 \;=\; \tilde{\mathbb{E}}_i\,W_\tau^2(s_N)\,I_{\{S_\tau=L\}} \;\leqslant\; \tilde{\mathbb{E}}_i\left[W_\tau^2(s_N)\,|\,S_\tau=L\right] \;\leqslant\; \mathrm{e}^{-2\,s_N L}\,\frac{w_i^2}{\min_k w_k^2}\,. \qquad (17)$$

Hence, denoting the variance under the new measure by $\tilde{\mathbb{V}}_i$, we have

$$\frac{\tilde{\mathbb{V}}_i\,\Gamma}{\left(\tilde{\mathbb{E}}_i\,\Gamma\right)^2} \;\leqslant\; \frac{\left(\max_k w_k\right)^2}{c^2\,\min_k w_k^2} - 1\,.$$

In other words, Γ has a bounded relative error, independent of L. □

Remark 4 Equation (16) shows that, as a function of L, γ_i decays exponentially with rate s_N, i.e.,

$$\gamma_i \;\propto\; \mathrm{e}^{-L\,s_N} \;=\; \left(\mathrm{e}^{-s_N}\right)^L\,.$$

Alternatively, we may say that, as a function of L, γ_i decays geometrically with rate $\eta_N = \mathrm{e}^{-s_N}$. □

The above theorem shows (formally) that we can estimate γ_i efficiently by generating samples $\Gamma_1,\ldots,\Gamma_n$ of Γ from n independent importance sampling simulation runs (using the conjugate rates), with a stopping time when the content of the second buffer hits either level 0 or level L. An estimator for γ_i is

$$\hat{\Gamma} \;=\; \frac{1}{n}\sum_{j=1}^{n}\Gamma_j\,.$$

An estimator for the variance of Γ is the usual sample variance, which can be used to construct a confidence interval for the estimator $\hat{\Gamma}$.

To obtain the conjugate rates, we need to calculate η_N and $\boldsymbol{w}$. There exist various efficient techniques to find the largest eigenvalue of a matrix; any of these can be used to solve (9) for s_N and hence η_N. To determine the eigenvector $\boldsymbol{w}$, we normalize $\boldsymbol{w}$ such that $w_0 = 1$. Using the tri-diagonal form of $G(s_N)$, it is easy to see that

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

and

$$w_{k+2} \;+\; a_1\,w_{k+1} \;+\; a_2\,w_k \;=\; 0,\;\; k = 0,\ldots,N-2\,, \qquad (18)$$

where $a_1 = -(\lambda+\mu_1+\mu_2-\mu_2\,\eta_N)/\lambda$ and $a_2 = \mu_1/(\lambda\,\eta_N)$. These equations completely determine $\boldsymbol{w}$ in terms of η_N. Alternatively, in view of the difference equation (18), a solution for w_k may be expressed in the following form

$$w_k \;=\; c_1\,z_1^k \;+\; c_2\,z_2^k\,,\;\; k = 0,\ldots,N\,, \qquad (19)$$

where c_1, c_2 are constants determined from w_0 and w_1; z_1 and z_2 are the (possibly complex) roots of the following characteristic polynomial (which is obtained by substituting the above solution for w_k in (18)):

$$\zeta^2 \,+\, a_1\,\zeta \,+\, a_2 \;=\; 0\,, \tag{20}$$

with a_1 and a_2 as defined above.

It turns out that if the first buffer is the bottleneck (i.e., $\mu_1 < \mu_2$) or if $\mu_1 = \mu_2$, then (20) has two complex solutions, say, $z\,\mathrm{e}^{\pm i\phi}$ (here $i := \sqrt{-1}$), and therefore

$$w_k \;=\; z^k\;((c_1 + c_2)\,\cos(k\phi) \,+\, (c_1 - c_2)\,\sin(k\phi))\;,\;\; k = 0, \ldots, N,$$

where c_1 and c_2 are determined from the two equations:

$$w_0 \;=\; 1 \;=\; c_1 \,+\, c_2\,,$$

$$w_1 \;=\; (\lambda + \mu_2 - \mu_2\,\eta_N)\,/\,\lambda \;=\; z\;((c_1 + c_2)\,\cos(\phi) \,+\, (c_1 - c_2)\,\sin(\phi))\;.$$

It follows that

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

with $c = (c_1 - c_2) = (w_1/z - \cos(\phi))/\sin(\phi)$. Finally, the conjugate rates in (4), (5) and (6), are given by:

$$\tilde{\mu}_2 \;=\; \mu_2\,\eta_N\,,$$

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

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

If the second buffer is the bottleneck (i.e., $\mu_2 < \mu_1$), then (20) has two real solutions, say, z_1 and z_2. Therefore, w_k is given by (19) with c_1 and c_2 as determined from the two equations:

$$w_0 \;=\; 1 \;=\; c_1 \,+\, c_2\,,$$

$$w_1 \;=\; (\lambda + \mu_2 - \mu_2\,\eta_N)\,/\,\lambda \;=\; c_1 z_1 \,+\, c_2 z_2\,.$$

The corresponding conjugate rates are determined from (4), (5) and (6):

$$\tilde{\mu}_2 \;=\; \mu_2\,\eta_N\,,$$

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

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

3.2 Infinite first buffer

When the first buffer has infinite capacity, the process (X_t, S_t) defined in (2) is still a Markov additive process, but now the state space of the Markov process (X_t) is infinite. Equation (1) still holds, but now $A(s)$ is given by the infinite-dimensional tri-diagonal matrix

$$G(s) \;=\; \begin{pmatrix} -\lambda - \mu_2 + \mu_2\,\mathrm{e}^{-s} & \lambda & & \\ \mu_1\,\mathrm{e}^{s} & -\lambda - \mu_1 - \mu_2 + \mu_2\,\mathrm{e}^{-s} & \lambda & \\ & \ddots & \ddots & \ddots \end{pmatrix}.$$

Note that the MAP generator $G(s)$ is equal to $Q(\mathrm{e}^{-s})$, defined in (26) of Appendix A.

Importance sampling

Importance sampling may be applied by extending the change of measure proposed for the finite first buffer case (defined by (3) and (4), with $s = s_N$) to the case where the first buffer is infinite. By Lemma 6, as $N \to \infty$, $\eta_N = \mathrm{e}^{-s_N}$ increases to a constant η, with either $\eta = \lambda/\mu_2$ (when $\mu_2 < \mu_1$) or else η is the unique $u \in (0,1)$ satisfying

$$2\sqrt{\frac{\lambda\,\mu_1}{u}} \;=\; \lambda + \mu_1 + \mu_2(1-u)\,. \tag{21}$$

Having determined η from the above, we naturally set $s = -\log\eta$ in (3) and (4) to obtain a change of measure in which (X_t) has the Q-matrix

$$\tilde{Q} \;=\; \Delta^{-1}(\boldsymbol{w})\,Q(\eta)\,\Delta(\boldsymbol{w})\,, \tag{22}$$

and (S_t) has death rate

$$\tilde{\mu}_2 \;=\; \mu_2\,\eta\,, \tag{23}$$

where $\boldsymbol{w}$ is the right eigenvector of $Q(\eta)$ corresponding to the eigenvalue 0. (Note that, by Lemmas 4 and 6, $Q(\eta)$ indeed has an eigenvalue 0.) Moreover, setting $w_0 = 1$, the eigenvector $\boldsymbol{w}$ satisfies,

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

and

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

where $a_1 = -(\lambda + \mu_1 + \mu_2 - \mu_2\,\eta)/\lambda$ and $a_2 = \mu_1/(\lambda\,\eta)$. The vector $\boldsymbol{w}$ is now completely determined in terms of η. Similar to the finite first buffer case, the solution for w_k can also be expressed as follows:

$$w_k \;=\; c_1\,z_1^k \;+\; c_2\,z_2^k\,,$$

where c_1 and c_2 are constants, and z_1 and z_2 are the (possibly complex) roots of the the characteristic equation:

$$\zeta^2 \;+\; a_1\,\zeta \;+\; a_2 \;=\; 0\,, \tag{24}$$

with a_1 and a_2 as defined above. We now distinguish two cases.

If the second buffer is the bottleneck (i.e., $\mu_2 < \mu_1$), then we have $\eta = \lambda/\mu_2$. Equation (24) has two real solutions, $z_1 = \mu_1/\lambda$ and $z_2 = \mu_2/\lambda$, of which only the second satisfies the boundaries: w_0 and w_1. It follows that

$$w_k \;=\; (\mu_2/\lambda)^k \;=\; (1/\eta)^k\,,\;\; k \geqslant 0\,.$$

The corresponding conjugate rates are:

$$\begin{aligned}\tilde{\lambda} &= \mu_2\,,\\ \tilde{\mu}_1 &= \mu_1\,,\\ \tilde{\mu}_2 &= \lambda\,.\end{aligned}$$

That is, we interchange the arrival rate and the smallest service rate. Experimental results in Section 4 suggest that this change of measure yields estimates with bounded relative error. Note that this is consistent with the state-independent (and asymptotically efficient) change of measure obtained from large deviations analysis of related overflow probabilities in queueing networks (see, e.g., [7], [20]).

If the first buffer is the bottleneck (i.e., $\mu_1 < \mu_2$) or if $\mu_1 = \mu_2$, then η is the unique solution of (21) in (0,1). With some algebra it can be shown that, for this η, the characteristic equation (24) has only one real solution, say, z. Consequently, w_k has the form:

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

with $c = (\lambda + \mu_2 - \mu_2\,\eta)/(\lambda z) - 1$. It follows that the conjugate rates are given by:

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

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

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

Experimental results in Section 4 seem to suggest that this change of measure yields estimates with a relative error that (asymptotically) grows linearly with the overflow level L.

Remark 5 With some algebra, it is not difficult to show that if $\mu_1 \leqslant \mu_2$, then (24) has a double root, say, z, which satisfies the following equation

$$\lambda z \;=\; \frac{\mu_1}{\eta z}\,. \tag{25}$$

It follows that

$$\lim_{k\to\infty} \tilde{\lambda}(k) \;=\; \lambda z \;=\; \frac{\mu_1}{\eta z} \;=\; \lim_{k\to\infty} \tilde{\mu_1}(k)\,.$$

In other words, under the new measure, the first queue is unstable and becomes "critical," i.e., $\tilde{\lambda}(k) = \tilde{\mu_1}(k)$, as $k \to \infty$. In fact, when $\mu_1 = \mu_2 = \mu$, it can be shown from (21) and (24) that $\eta = \lambda/\mu$ and $z = \mu/\lambda$. Therefore, when the service rates are equal, the conjugate rates are state-independent, obtained by interchanging the arrival rate and the service rate at the second server (i.e., $\tilde{\lambda} = \tilde{\mu}_1 = \mu$ and $\tilde{\mu_2} = \lambda$). Experimental results in Section 4 seem to indicate that this change of measure yields estimates with a relative error that is (asymptotically) bounded linearly in L. This agrees with observations made in the literature (see, e.g., [10], [13]) that the commonly used heuristic in [20] is less effective when the service rates are equal. □

4 Experimental results

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 experimental results presented here, the same number of replications, namely, 10^6, is used to obtain each estimate using importance sampling. It is important to note that the properties of the estimator (established or claimed) in this paper are with respect to the number of replications rather than the actual simulation effort (e.g., CPU time). The latter, being proportional to the expected number of simulated events per replication, increases (roughly) linearly with the overflow level,

L. For each estimate in Tables 1,2,3 and 4, we also include its relative error RE (standard deviation divided by the mean) and the actual CPU time. For the tandem Jackson network being considered, numerical values of the overflow probabilities can be obtained using the algorithm outlined in [9]. For the purpose of validation, these values are also listed in the tables, along with the corresponding estimates.

Example 1 ($\lambda = 1, \mu_1 = 4$ and $\mu_2 = 2$. The second server is the bottleneck.)

For a finite first buffer, $N = 9$, we find that the geometric decay rate η of the second buffer is approximately[1] 0.49967. Equation (20) has two real solutions $z_1 = 2.00198$ and $z_2 = 3.99868$, and the eigenvector $\boldsymbol{w}$ is given by $w_k = c_1 z_1^k + c_2 z_2^k$, $k = 0, 1, \ldots, N$, 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 $\boldsymbol{w}$ is given by $w_k = 2^k$, $k = 0, 1, 2, \ldots$. This leads to $\tilde{\lambda} = 2$, $\tilde{\mu}_1 = 4$ and $\tilde{\mu}_2 = 1$, i.e., interchanging the arrival rate and the service rate of the slowest (second) server.

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., [20], [7], [10]).

(λ, μ_1, μ_2)	L	γ (Numerical)	$\hat{\gamma}$ (IS)	RE (IS)	CPU (sec)
(1, 4, 2)	20	1.428e-6	1.43e-6	0.11%	90
$N = 9$	25	4.446e-8	4.44e-8	0.11%	114
	50	1.303e-15	1.30e-15	0.11%	238
	60	1.264e-18	1.27e-18	0.11%	287
	100	1.120e-30	1.12e-30	0.11%	484
(1, 4, 2)	20	1.432e-6	1.43e-6	0.11%	90
$N = \infty$	25	4.472e-8	4.47e-8	0.11%	114
	50	1.332e-15	1.33e-15	0.11%	238
	60	1.301e-18	1.30e-18	0.11%	287
	100	1.183e-30	1.18e-30	0.11%	484

Table 1: Estimates of the overflow probability in Example 1.

[1] We have rounded all numerical values to 5 significant digits.

Example 2 ($\lambda = 1, \mu_1 = 2$ and $\mu_2 = 3$. The first server is the bottleneck.)

For a finite first buffer, $N = 9$, we find that $\eta = 0.28898$. Equation (20) has two complex solutions $z\,\mathrm{e}^{\pm i\phi}$, with $z = 2.63077$ and $\phi = -0.22144$. The eigenvector $\boldsymbol{w}$ is therefore given by $w_k = z^k(\cos(k\phi) + c\sin(k\phi))$, $k = 0, 1, \ldots, N$, with $c = (w_1/z - \cos(\phi))/\sin(\phi) = -0.98048$. The change of measure is obtained accordingly (as in Section 3.1).

For an infinite first buffer, we find η such that (20) has only one solution z. Algebraically, η is a solution of $-8 + 36\eta - 36\eta^2 + 9\eta^3 = 0$, and $z = \sqrt{\mu_1/(\eta\lambda)}$. The numerical values are $\eta = 0.31194$ and $z = 2.53209$, and the eigenvector $\boldsymbol{w}$ satisfies $w_k = z^k(1 + c\,k)$, $k = 0, 1, 2, \ldots$, with $c = w_1/z - 1 = 0.21014$. The change of measure is obtained accordingly (as in Section 3.2).

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 seems to increase linearly with L.

(λ, μ_1, μ_2)	L	γ (Numerical)	$\hat{\gamma}$ (IS)	RE (IS)	CPU (sec)
(1, 2, 3)	20	1.878e-11	1.88e-11	0.24%	76
$N = 9$	25	3.759e-14	3.75e-14	0.24%	94
	50	1.247e-27	1.24e-27	0.24%	189
	60	5.063e-33	5.06e-33	0.24%	234
	100	1.377e-54	1.38e-54	0.24%	379
(1, 2, 3)	20	2.048e-11	2.05e-11	0.49%	82
$N = \infty$	25	4.610e-14	4.63e-14	0.56%	102
	50	4.305e-27	4.32e-27	0.87%	201
	60	2.956e-32	2.94e-32	0.98%	236
	100	8.595e-53	8.59e-53	1.38%	384

Table 2: Estimates of the overflow probability in Example 2.

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, $N = 9$, we find that $\eta = 0.41467$. Equation (20) has two complex solutions $z\,\mathrm{e}^{\pm i\phi}$, with $z = 1.79315$ and $\phi = -0.21466$, and the eigenvector $\boldsymbol{w}$ is given by $w_k = z^k(\cos(k\phi) + c\sin(k\phi))$, $k = 0, 1, \ldots, N$, 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 $\boldsymbol{w}$ is given by $w_k = z^k(1 + c\,k)$, $k = 0, 1, 2 \ldots$,

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 (for an increasing overflow level, L) exhibit bounded relative error. When the first buffer is infinite, the relative error seems to increase linearly with L.

(λ, μ_1, μ_2)	L	γ (Numerical)	$\hat{\gamma}$ (IS)	RE (IS)	CPU (sec)
(1, 4/3, 2)	20	1.150e-8	1.15e-8	0.23%	73
$N = 9$	25	1.405e-10	1.40e-10	0.23%	92
	50	3.887e-20	3.89e-20	0.23%	185
	60	5.843e-24	5.83e-24	0.23%	222
	100	2.982e-39	2.98e-39	0.23%	370
(1, 4/3, 2)	20	1.348e-8	1.35e-8	0.52%	73
$N = \infty$	25	1.966e-10	1.96e-10	0.60%	91
	50	2.203e-19	2.19e-19	0.95%	177
	60	6.541e-23	6.50e-23	1.07%	211
	100	6.790e-37	6.78e-37	1.52%	342

Table 3: Estimates of the overflow probability in Example 3.

Example 4 ($\lambda = 1, \mu_1 = 2$ and $\mu_2 = 2$. Equal service rates at both nodes.)

For a finite first buffer, $N = 9$, we follow the same procedure as when the first server is the bottleneck. We find that $\eta = 0.47847$. Equation (20) has two complex solutions $z\,\mathrm{e}^{\pm i\phi}$, with $z = 2.0445$ and $\phi = -0.15$, and the eigenvector $\boldsymbol{w}$ is given by $w_k = z^k(\cos(k\phi) + c\sin(k\phi))$, $k = 0, 1, \ldots, N$, 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 rate and the service rate at the second server, i.e., $\tilde{\lambda} = 2$, $\tilde{\mu}_1 = 2$ and $\tilde{\mu}_2 = 1$ (see Remark 5).

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 seems to increase linearly with L.

Remark 6 According to the theory in Section 3, the derived change of measure holds for any starting state, provided that $Y_0 \geq 1$. 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

(λ, μ_1, μ_2)	L	γ (Numerical)	$\hat{\gamma}$ (IS)	RE (IS)	CPU (sec)
(1, 2, 2)	20	2.557e-7	2.55e-7	0.19%	76
$N = 9$	25	6.397e-9	6.40e-9	0.19%	96
	50	6.340e-17	6.32e-17	0.19%	194
	60	3.987e-20	3.99e-20	0.19%	234
	100	6.235e-33	6.23e-33	0.19%	393
(1, 2, 2)	20	2.787e-7	2.77e-7	0.29%	77
$N = \infty$	25	7.661e-9	7.68e-9	0.31%	96
	50	1.559e-16	1.56e-16	0.38%	188
	60	1.382e-19	1.39e-19	0.40%	224
	100	9.618e-32	9.63e-32	0.46%	366

Table 4: Estimates of the overflow probability in Example 4.

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 (and slowly) with L. This, however, is not practical, since 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. Instead, by considering an appropriately modified MAP representation of the tandem network (as noted in Remark 1), our methodology (in Section 3) may be adapted to derive another exponential change of measure which holds also along the boundary $(Y_t) = 0$. □

5 Conclusions

In this paper, 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 buffer overflow in the second node. The 'optimal' twisting 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. It is shown (formally and

empirically) that importance sampling simulations with this change of measure yield asymptotically efficient estimators, with a bounded relative error. Only when the first node is the bottleneck and its buffer size is infinite, experimental results seem to indicate that the relative error is bounded linearly in the buffer overflow level.

This paper represents only an introduction and a preliminary study of a new mathematical approach for the analysis and efficient simulation of rare events in queueing networks. Further research is now being conducted to examine its feasibility and effectiveness for other rare events of interest in tandem Jackson networks. A related approach based on a MAP representation of a two-node fluid line is used in [14] to devise an asymptotically efficient simulation of rare overflow events. A different approach, based on adaptive importance sampling, to approximate the 'optimal' state-dependent change of measure has also been used recently to estimate rare event probabilities in Jackson queueing networks [3]. Further investigation of this approach is worthwhile, as it may also hold some promise. Generalizations to feed-forward and possibly non-Markov queueing networks would constitute an important step towards the applicability of the approach, e.g., for the development and evaluation of resource allocation and routing algorithms in communication networks.

Acknowledgement

The authors wish to thank the anonymous associate editor and the referees for their critical yet constructive comments, which undoubtedly have benefited the paper.

References

[1] S. Asmussen 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.

[2] V. Anantharam, 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.

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

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

[5] T.S. Chihara (1978). *An Introduction to Orthogonal Polynomials*, Gordon and Breach, New York.

[6] G. De Veciana, 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.

[7] M.R. Frater 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.

[8] M.R. Frater, 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.

[9] M.J.J. Garvels 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, March 1999.

[10] P. Glasserman 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.

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

[12] J.F.C. Kingman (1961). A convexity property of positive matrices. *Quart. J. Math.* **12** 283–284.

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

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

[15] D.P. Kroese, W.R.W. Scheinhardt and P.G. Taylor. Spectral Properties of the Tandem Jackson Network, Seen as a Quasi-Birth-and-Death Process. In preparation.

[16] M.F. Neuts (1981). *Matrix-Geometric Solutions in Stochastic Models*, John Hopkins University Press, Baltimore.

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

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

[19] V.F. Nicola, P. Shahabuddin, P. Heidelberger and P.W. Glynn (1993). Fast simulation of steady-state availability in non-Markovian highly dependable systems. *Proceedings of the Twenty-Third International Symposium on Fault-Tolerant Computing,* IEEE Computer Society Press, 38–47.

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

[21] R.Y. Rubinstein (1999). The cross-entropy method for combinatorial and continuous optimization. *Methodology and Computing in Applied Probability* **1** 127–190.

[22] P. Tsoucas (1992). Rare events in series of queues. *J. Appl. Probab.* **29** 168–175.

A Appendix

Preliminaries

For each $u \in (0,1]$, let $Q(u)$ be the infinite dimensional tri-diagonal matrix

$$Q(u) = \begin{pmatrix} -\lambda - \mu_2 + \mu_2\, u & \lambda & & \\ \mu_1/u & -\lambda - \mu_1 - \mu_2 + \mu_2\, u & \lambda & \\ & \ddots & \ddots & \ddots \end{pmatrix}. \tag{26}$$

Let $Q^{(n)}(u)$ denote the n-th upper-left corner truncation of $Q(u)$. That is, the $(n \times n)$-matrix obtained from $Q(u)$ by deleting the rows and columns $n+1, n+2, \ldots$.

Define the $(n \times n)$-matrix $\hat{Q}^{(n)}$ as

$$\hat{Q}^{(n)}(u) = \begin{pmatrix} -\lambda - \mu_2 + \mu_2\, u & \lambda & & \\ \mu_1/u & -\lambda - \mu_1 - \mu_2 + \mu_2\, u & \lambda & \\ & \ddots & \ddots & \ddots \\ & & \mu_1/u & -\mu_1 - \mu_2 + \mu_2\, u \end{pmatrix}. \tag{27}$$

For a fixed u, the sequence of polynomials $P_1(x), P_2(x), \ldots$, is defined by the following recursion:

$$\begin{aligned} P_0(x) &= 1, \\ \lambda\, P_1(x) &= x + \lambda + \mu_2(1-u), \\ \lambda\, P_n(x) &= (x + \lambda + \mu_1 + \mu_2(1-u))\; P_{n-1}(x) \;-\; \frac{\mu_1}{u}\, P_{n-2}(x),\; n \geqslant 2. \end{aligned} \tag{28}$$

Also, for $n \geqslant 1$, let

$$\hat{P}_n(x) \;=\; P_n(x) - P_{n-1}(x)\,,$$

and

$$K_n(x) \;=\; u\, P_n(x) - P_{n-1}(x)\,.$$

We will write $P_n(x;u)$ when we wish to emphasize the dependence of P_n on u (and similarly for $\hat{P}_n$ and K_n).

We now derive some properties for the polynomials and matrices defined above.

Lemma 2 (a) The zeros of P_n are the eigenvalues of $Q^{(n)}$.

(b) The zeros of $\hat{P}_n$ are the eigenvalues of $\hat{Q}^{(n)}$.

Proof. Let I_n denote the identity matrix of dimension n. The characteristic polynomial of $Q^{(1)}$ is

$$\det\left(x\, I_1 - Q^{(1)}\right) \;=\; x + \lambda + \mu_2(1-u)\,,$$

Because the matrices $Q^{(n)}$ are tri-diagonal, we have

$$\det\left(x\, I_2 - Q^{(2)}\right) \;=\; (x + \lambda + \mu_1 + \mu_2(1-u))\, \det\left(x\, I_1 - Q^{(1)}\right) \;-\; \frac{\mu_1}{u}\,\lambda\,,$$

and for $n \geqslant 3$

$$\begin{aligned} \det\left(x\, I_n - Q^{(n)}\right) = {} & (x + \lambda + \mu_1 + \mu_2(1-u))\, \det\left(x\, I_{n-1} - Q^{(n-1)}\right) \\ & - \frac{\mu_1}{u}\,\lambda\, \det\left(x\, I_{n-2} - Q^{(n-2)}\right)\,. \end{aligned}$$

We see that $\lambda^n\, P_n(x)$ is the characteristic polynomial of $Q^{(n)}$, and thus, for each $n = 1, 2, \ldots$, the zeros of $P_n(x)$ are the eigenvalues of $Q^{(n)}$. This proves (a).

To show (b), observe that the characteristic polynomial of $\hat{Q}^{(n)}$ satisfies

$$\begin{aligned}\det\left(x\,I_n - \hat{Q}^{(n)}\right) &= (x+\mu_1+\mu_2(1-u))\,\det\left(x\,I_{n-1} - Q^{(n-1)}\right)\\ &\quad -\frac{\mu_1}{u}\lambda\,\det\left(x\,I_{n-2} - Q^{(n-2)}\right)\\ &= \det\left(x\,I_n - Q^{(n)}\right) \;-\; \lambda\,\det\left(x\,I_{n-1} - Q^{(n-1)}\right)\\ &= \lambda^n\left(P_n(x) - P_{n-1}(x)\right).\end{aligned}$$

Hence, the zeros of $\hat{P}_n(x)$ are the eigenvalues of $\hat{Q}^{(n)}$. □

Lemma 3 (a) P_n has n distinct real zeros $x_{n,1} < \ldots < x_{n,n}$, and these zeros *interlace*. That is, for all $n \geqslant 2$,

$$x_{n,i} \;<\; x_{n-1,i} \;<\; x_{n,i+1}\,,\;\; i = 1,\ldots,n-1.$$

(b) $\hat{P}_n$ has n distinct real zeros $\hat{x}_{n,1} < \ldots < \hat{x}_{n,n}$ which interlace. Moreover, $\hat{x}_{n,n} > x_{n,n}$ and, for all $n \geqslant 2$,

$$x_{n,i} \;<\; \hat{x}_{n,i} \;<\; x_{n,i+1}\,,\;\; i = 1,\ldots,n-1. \tag{29}$$

(c) K_n has n distinct real zeros $\bar{x}_{n,1} < \ldots < \bar{x}_{n,n}$ which interlace. Moreover, $\bar{x}_{n,n} > \hat{x}_{n,n}$ and, for all $n \geqslant 2$,

$$\hat{x}_{n,i} \;<\; \bar{x}_{n,i} \;<\; \hat{x}_{n,i+1}\,,\;\; i = 1,\ldots,n-1. \tag{30}$$

Proof. This can be shown by induction on n: Obviously, P_1 has one real zero. Assume that $P_1,\ldots,P_{n-1}$ have interlacing real zeros. Because the coefficient of x^i in polynomial P_i is *positive*, by the interlacing property we have $P_{n-2}(x_{n-1,n-1}) > 0$, $P_{n-2}(x_{n-1,n-2}) < 0$, $P_{n-2}(x_{n-1,n-3}) > 0$, etc. The sign of $P_{n-2}(x_{n-1,1})$ is positive if n is even, and negative otherwise.

Consequently, by (28) we have $P_n(x_{n-1,n-1}) < 0$, $P_n(x_{n-1,n-2}) > 0$, etc. This shows that between each two subsequent zeroes of P_{n-1} must lie exactly one zero of P_n. Moreover, since $P_n(x_{n-1,n-1}) < 0$, exactly one zero of P_n must be greater than $x_{n-1,n-1}$. Finally, if n is even, P_n must become positive before $x_{n-1,1}$, and if n is odd, it must become negative before $x_{n-1,1}$, showing that there must lie one zero before $x_{n-1,1}$. This proves (a).

To prove (b), first observe that $\hat{P}_n$ satisfies the same recursion as P_n (i.e., (28) with the P's replaced by $\hat{P}$'s). The interlacing property immediately follows. Moreover, since

$$\hat{P}_n(x_{n,n}) \;=\; 0 \;-\; P_{n-1}(x_{n,n}) \;<\; 0$$

and

$$\hat{P}_n(x_{n,n-1}) \;=\; 0 \;-\; P_{n-1}(x_{n,n-1}) \;>\; 0\,,$$

$\hat{P}_n$ must have a zero greater than $x_{n,n}$ and a zero in the interval $(x_{n,n-1}, x_{n,n})$. Continuing this argument for the other zeros of P_n we conclude that (29) is true.

Finally, (c) is proved in the same way as (b), after observing that

$$K_n(x) \;=\; \hat{P}_n(x) \;-\; (1-u)P_n(x)\,.$$

□

Next we consider the largest eigenvalues of $Q^{(n)}$ and $\hat{Q}^{(n)}$ as functions of u.

Lemma 4 (a) The largest eigenvalues $x_{n,n}$ and $\hat{x}_{n,n}$ of $Q^{(n)}$ and $\hat{Q}^{(n)}$, respectively, are convex functions of $u \in (0,1]$.

(b) There exist unique numbers u_n and $\hat{u}_n$ in the interval $(0,1)$ such that

$$x_{n,n}(u_n) \;=\; 0 \quad \text{and} \quad \hat{x}_{n,n}(\hat{u}_n) \;=\; 0\,.$$

Proof. Consider the non-negative matrix $H(z) := (\lambda + \mu_2 + \mu_1)I_n + Q^{(n)}(z)$. Let $\alpha(z)$ denote its largest eigenvalue (which is equal to the largest eigenvalue of $Q^{(n)}(z)$ plus $\lambda + \mu_1 + \mu_2$). Note that the logarithm of each element of $H(z)$ is a convex function in $z \in (0,1]$. From a well-known result for non-negative matrices [12], it follows that $\log\alpha(z)$ (and therefore $\alpha(z)$) is also a convex function on $(0,1]$. This shows that the largest eigenvalue of $Q^{(n)}(z)$ is a convex function on $(0,1]$. The same applies to the largest eigenvalue of $\hat{Q}^{(n)}(z)$. This proves (a).

To show (b), first note that for sufficiently small z, $\hat{x}_{n,n}(z) > 0$ and $x_{n,n}(z) > 0$. Also, $\hat{x}_{n,n}(1) = 0$, because $Q_0^{(n)} + \hat{Q}_1^{(n)} + Q_2^{(n)}$ is the generator (Q-matrix) of a Markov process. A well-known condition for stability (see, e.g., pages 16-19 of [16]) is that the derivative $\hat{x}'_{n,n}(1) > 0$. These facts, combined with (a) show that there is a unique $\hat{z}_n$ such that $\hat{x}_{n,n}(\hat{z}_n) = 0$. To complete the proof for $x_{n,n}$, observe that $x_{n,n}(z) < \hat{x}_{n,n}(z)$. □

Lemma 5 Let $\boldsymbol{v} = (v_1, \ldots, v_n)^T$ be the right eigenvector of $\hat{Q}^{(n)}(\hat{u}_n)$ corresponding to the eigenvalue 0. Then,

(a) $v_i > 0\,, \quad i = 1, \ldots, n,$

(b) $\dfrac{v_{i+1}}{v_i} > 1\,, \quad i = 1, \ldots, n-1,$

(c) $\dfrac{v_i}{v_{i+1}} > \hat{u}_n\,, \quad i = 1, \ldots, n-1.$

Proof. By definition of $\hat{u}_n$, 0 is the largest eigenvalue of $\hat{Q}^{(n)}(\hat{u}_n)$. The corresponding eigenvector is given by

$$(P_0(0;\hat{u}_n),\ \ldots\ ,\ P_{n-1}(0;\hat{u}_n))^T\ .$$

The largest zeros of $P_i(x;\hat{u}_n)$, $i = 1, \ldots, n-1$, are all smaller than $\hat{x}_{n,n} = 0$. Moreover, the leading coefficient of P_i is positive. Hence, P_i is increasing from $x_{i,i}$ onwards. Consequently, $P_i(0;\hat{u}_n) > 0$, $i = 1, \ldots, n-1$. Obviously, $P_0(0;\hat{u}_n) > 0$. This proves (a).

To prove (b), we need to show that $P_i(0;\hat{u}_n) - P_{i-1}(0;\hat{u}_n) > 0$. Or, equivalently,

$$\hat{P}_i(0;\hat{u}_n)\ >\ 0\,,\ \ i = 1, \ldots, n-1.$$

The argument is similar to that of (a). By definition of $\hat{u}_n$, the largest zero of $\hat{P}_i(x;\hat{u}_n)$ is 0. Because of the interlacing property of the polynomials $\{P_i\}$, all the largest zeros of $\hat{P}_1, \ldots, \hat{P}_{n-1}$ are less than 0 and the polynomials are increasing from those zeros onwards. Hence (b) follows.

Finally, to prove (c) it suffices to show that

$$K_i(0;\hat{u}_n)\ <\ 0\,,\ \ i = 1, \ldots, n-1.$$

We know that $\bar{x}_{i,j} < 0$, for $i = 1, \ldots, n-1$ and $j = 1, \ldots, i-1$. Hence, all the second largest zeros of $K_1, \ldots, K_{n-1}$ are less than 0. By the interlacing property we also know that $\bar{x}_{i,i} > \bar{x}_{1,1}$, for $i = 2, 3, \ldots$. Since for $i = 1, \ldots, n-1$, K_i is positive after its last 0, it remains to show that $\bar{x}_{1,1} > 0$.

The zero of $K_1(x)$ is $(u-1)(\mu_2 u - \lambda)/u$, which, for $u = \hat{u}_n$, is positive if

$$\hat{u}_n\ <\ \lambda/\mu_2\,. \tag{31}$$

This latter inequality holds, because if it did not, then $P_1(x)$ and (by the interlacing property) all other $P_n(x)$ would have a zero greater than 0. This concludes the proof. □

Lemma 6 If $\mu_1 \leqslant \mu_2$, then as $n \to \infty$, $\hat{u}_n$ increases to the unique $u \in (0,1)$ satisfying

$$2\sqrt{\frac{\lambda\,\mu_1}{u}}\ =\ \lambda + \mu_1 + \mu_2(1-u)\,.$$

On the other hand, if $\mu_1 > \mu_2$, then as $n \to \infty$, $\hat{u}_n$ increases to λ/μ_2.

Proof. The result basically follows from the theory of orthogonal polynomials [5]. We give here only the main ideas. For a full proof we refer to [15]. The crucial step is that the $\{P_n\}$ form a so-called *orthogonal polynomial sequence*, and satisfy

$$\int_S P_n(x)\, P_m(x)\, \psi(dx) \;=\; \Big(\frac{\mu_1}{u\lambda}\Big)^n\, \delta_{n,m}\,,$$

where ψ is a measure with support S given by

$$S \;=\; \begin{cases} [\sigma(u), \xi(u)] & \text{if } u \leqslant \lambda/\mu_1\,, \\ [\sigma(u), \xi(u)] \cup \{\chi(u)\} & \text{if } u > \lambda/\mu_1\,, \end{cases}$$

with

$$\sigma(u) \;=\; -\lambda - \mu_1 - \mu_2(1-u) - 2\sqrt{\frac{\lambda\,\mu_1}{u}} \tag{32}$$

$$\xi(u) \;=\; -\lambda - \mu_1 - \mu_2(1-u) + 2\sqrt{\frac{\lambda\,\mu_1}{u}} \tag{33}$$

$$\chi(u) \;=\; (\frac{\lambda}{u} - \mu_2)(1-u)\,. \tag{34}$$

The limiting behaviour of the zeroes of P_n is closely related to the measure ψ. In particular,

$$\begin{array}{ll} \{x_{n,1}\}_{n=1}^{\infty} & \text{is a strictly decreasing sequence with limit } \sigma(u); \\ \{x_{n,n-1}\}_{n=1}^{\infty} & \text{is a strictly increasing sequence with limit } \xi(u); \\ \{x_{n,n}\}_{n=1}^{\infty} & \text{is a strictly increasing sequence with limit } \chi_1(u)\,, \end{array}$$

where (see [5])

$$\chi_1(u) \;=\; \sup S \;=\; \begin{cases} \xi(u) & \text{if } u \leqslant \lambda/\mu_1\,, \\ \chi(u) & \text{if } u > \lambda/\mu_1\,. \end{cases}$$

The lemma now follows easily from the observations above. Namely, if $\mu_1 \leqslant \mu_2$, then $x_{n,n}(u)$ and hence also $\hat{x}_{n,n}(u)$ tend to $\xi(u)$. Consequently, the corresponding u_n and $\hat{u}_n$ tend to the unique $u \in (0,1)$ satisfying

$$2\sqrt{\frac{\lambda\,\mu_1}{u}} = \lambda + \mu_1 + \mu_2(1-u)\,.$$

On the other hand, if $\mu_1 > \mu_2$, then $x_{n,n}(u)$ and $\hat{x}_{n,n}(u)$ tend to $\chi(u)$. Consequently, the corresponding u_n and $\hat{u}_n$ tend to λ/μ_2. □