Optimal Changes of Measure for Buffer Overflows in Tandem Network

Jiyeon Lee
Department of Statistics, Yeungnam University

Abstract

We consider a stable tandem network which consists of two M/M/1 queues. The optimal changes of measure to run the fast simulation for the probability of rare events such as buffer overflows are obtained.

1 Introduction

In a queueing network with finite buffers, a certain proportion of packets are lost due to buffer overflows. While the probability that the buffer overflows occur can be calculated analytically for a single M/M/1 queue(Parekh and Walrand(1989)), the first step equation for a network of queues cannot be solved analytically because the order of the characteristic equation becomes large. Therefore, simulation is often used to find the probability of the buffer overflow or the expected recurrence time of buffer overflows. For a stable system, the events of reaching a large backlog are very infrequent. Hence, direct simulations are very slow and take up a lot of computer time. Besides, there is also the difficulty of implementing a pseudo-random generator function effectively during very long simulation. However, using the idea of importance sampling, the probability of this rare event can be found by fast simulation without incurring the large cost involved in direct simulation.

The idea in importance sampling is as follows. Suppose we are interested in certain (rare) events occurring in a system S that we can simulate on a computer. Then instead of simulating S we simulate a second system $\bar{S}$, which has the property that events in S and $\bar{S}$ correspond in some way. In particular, to the rare events A in S correspond events $\bar{A}$ in $\bar{S}$. The correspondence is such that

1) the events $\bar{A}$ in $\bar{S}$ are more frequent than the events A in S, and

2) the connection between S and $\bar{S}$ allows one to infer $P(A)$ if one knows $\bar{P}(\bar{A})(\bar{P}$ is the probability measure in system $\bar{S}$.)

A major issue in the use of importance sampling is how one should construct $\bar{S}$ from S. To an extent, the problems of obtaining the probability of rare events or the mean time between occurrences of rare events are being replaced by another difficult problem of obtaining the optimal changes of measure. In this paper, the system S will be a network of queues. The system $\bar{S}$ will be a network of queues also, with the same structure as S, but with various parameters such as arrival and service rates that will be different from the corresponding quantities in S.

Asmussen(1982) showed that the asymptotically optimal change of measure for estimating the probability of large build-ups in an M/M/1 queue corresponds to simulating an unstable M/M/1 queue with interchanging arrival rate and service rate.

Based on a heuristic application of large deviations techniques, Parekh and Walrand(1989) proposed importance sampling estimator for overflow probabilities in queueing network. For tandem networks, their estimator interchanges the arrival rate and the slowest service rate, thus generalizing the M/M/1 estimator described above. They evaluated this estimator numerically and found that it generally works well. An optimization step required in Parekh and Walrand(1989) was solved in Frater et al.(1991) for Jackson networks and in Frater and Anderson(1994) for tandem networks. However they considered only the total backlog of the queueing network whose service rates are different, that is, there is only one node which has the largest load. Glasserman and Kou(1995) mentioned the importance sampling based on the above interchange rule tends to be less effective if the service rates are close.

In this paper we find the optimal change of measure for one node's overflow in tandem network by using the h-transform in McDonald(1999) and the time-reversed process in Anantharam et al.(1990). In section 2 we describe the model and the h-transform method of McDonald(1999). The optimal changes of measure due to the h-transform for one node's overflow are obtained in section 3 and 4.

2 Model

We consider two M/M/1 queues in tandem which have respective service rates μ_1 and μ_2. We assume, for stability, that the arrival rate λ satisfies $\lambda < \mu_1$ and $\lambda < \mu_2$. For simplicity, we will refer to such a system by a (λ, μ_1, μ_2)-network. Fig. 1 depicts a (λ, μ_1, μ_2)-network.

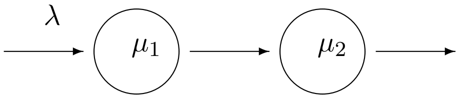

Fig. 1. A (λ, μ_1, μ_2)-network.

A (λ, μ_1, μ_2)-network can be described as a Markov jump process $(N(t), t \geq 0)$ on $\mathcal{S} \equiv \mathbf{N}^2$, where $\mathbf{N}$ is the non-negative integers. Let $(x, y) \in \mathcal{S}$ denote the number of customers waiting or being served at each node. The jump rates of this Makov jump process are shown in Fig. 2.

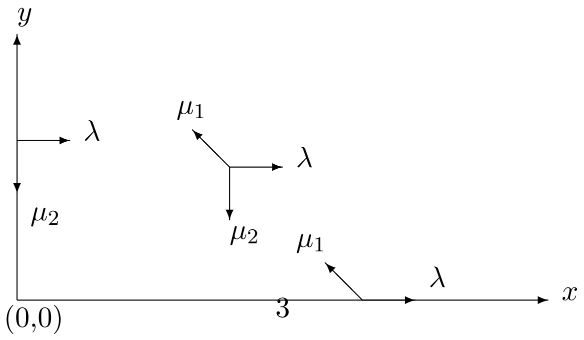

Fig. 2. Jump rates for (λ, μ_1, μ_2)-network.

The generator of L of N is given as an operator on a bounded function g on $\mathcal{S}$:

$$\begin{aligned} Lg(x,y) = \quad & \lambda(g(x+1,y)-g(x,y)) \\ + \quad & \mu_1(x)(g(x-1,y+1)-g(x,y)) \\ + \quad & \mu_2(y)(g(x,y-1)-g(x,y)), \quad (x,y)\in\mathcal{S} \end{aligned}$$

where $\mu_1(x) = \mu_1$ if $x > 0$ and 0 otherwise and where $\mu_2(y)$ is defined analogously. The stationary distribution π of this jump process is given by

$$\pi(x,y) = (1-\frac{\lambda}{\mu_1})(\frac{\lambda}{\mu_1})^x(1-\frac{\lambda}{\mu_2})(\frac{\lambda}{\mu_2})^y \tag{1}$$

under the condition that $\lambda/\mu_i < 1 \; i = 1,2$. The equation (1) implies that, in the steady state at a fixed time, the queue sizes at the different nodes are independent. Furthermore, the queue size at node i has the stationary measure of a birth and death process with birth rate λ and death rate $\mu_i, i = 1,2$.

The event rate of the (λ,μ_1,μ_2)-network is $\lambda+\mu_1+\mu_2$. Without loss of generality we assume $\lambda+\mu_1+\mu_2 = 1$ (otherwise, we can rescale time) so if we regard L as the discrete generator of a Markov chain W on $\mathcal{S}$ then the (λ,μ_1,μ_2)-network is precisely the homogeneization of this chain. Consequently π is also the stationary distribution of W. We assume the kernel K is associated with the Markov chain W with the stationary probability distribution π.

We are interested in the rare event when the one node overloads; that is when the one coordinate of W exceeds a level ℓ. For the moment, we relabel the first coordinate on this coordinate. When the first coordinate overloads, the other node may remain stable even though it is subject to higher load. The coordinate corresponding to this super-stable node is renumbered to $r+1$ through $r+m, \; r+m=2$. Unfortunately when one node overloads it may drive the other node into overload. We assume this node corresponds to coordinates 2 through r. Then for $\vec{x} = (x_1,x_2)$, x_1 denotes the number of customers of the chosen node to overload and x_2 is that of the other node which is super-stable or unstable.

We look for a harmonic function $h(x_1, x_2)$ since in addition to twisting the first component to become transient we must judiciously twist only the other component which remains recurrent after twisting. Furthermore the twist must make nodes 1 through r transient to plus infinity.

If $\beta = \{i_1, i_2, \ldots, i_d\}$, we say $\vec{x}$ is on the boundary $\mathcal{S}_\beta$ if $x_i = 0$ for $i \in \beta$ but $x_i > 0$ for $i \notin \beta$. Denote the interior of the orthant by $\text{int}(\mathcal{S})$ if $x_i > 0$ for all i. We decompose $\vec{x} \in \mathcal{S}$ as

$$\vec{x} = (\tilde{x}, \hat{x}) \text{ where } \tilde{x} \in \mathbf{N}^r, \hat{x} \in \mathbf{N}^m, r + m = 2.$$

Similarly we decompose W into components $(\tilde{W}, \hat{W})$. To fine h we now remove the boundaries $\Delta := \cup_{k=1}^r \mathcal{S}_{\{k\}}$ for the first r coordinates. Define $\mathcal{S}^\infty = \mathcal{I}^r \times \mathbf{N}^m$, where $\mathcal{I}$ denotes the integers. Define $\text{int}(\mathcal{S}^\infty) = \{\vec{x} \in \mathcal{S}^\infty : x_i > 0, i = r+1, \ldots, r+m\}$. If $\beta = \{i_1, i_2, \ldots, i_d\} \subseteq \{r+1, \ldots, r+m\}$, we say $\vec{x} \in \mathcal{S}_\beta^\infty$ if $x_i = 0$ for $i \in \beta$.

On $\mathcal{S}^\infty$ we assume transitions for a chain W^∞ are given by a probability transition kernel K^∞ of the form,

$$K^\infty(\vec{x}, \vec{y}) = \hat{K}^\infty(\hat{x}, \hat{y}) f(\tilde{y} - \tilde{x} | \hat{x}, \hat{y})$$

where $\hat{K}^\infty(\hat{x}, \hat{y})$ is a transition kernel from $\mathbf{N}^m$ to $\mathbf{N}^m$ and $f(\cdot|\hat{x}, \hat{y})$ is a probability mass function given each pair $(\hat{x}, \hat{y})$. We assume the probability transition kernel $K(\vec{x}, \vec{y})$ of W agrees with $K^\infty(\vec{x}, \vec{y})$ when $\vec{x}, \vec{y} \in \mathcal{S}^\infty \setminus \Delta$. Consequently the chain W behaves like W^∞ outside the boundary Δ.

In McDonald(1999) it is shown that in great generality one can construct the harmonic function for the kernel K^∞ of the form $h(\vec{x}) := a_1^{x_1} a_2^{x_2} \cdots a_{r+m}^{x_{r+m}}$ with $a_2 = \cdots = a_r = 1$. We assume the existence of h such that $\mathcal{L}^\infty$ defined below is a discrete generator for all $\vec{x} \in \mathcal{S}^\infty$. For a bounded function g in $\mathcal{S}^\infty$

$$\mathcal{L}^\infty g(\vec{x}) \quad := \quad \frac{1}{h(\vec{x})} L^\infty (h \cdot g)(\vec{x})$$

$$= \sum_{\vec{y}\in\mathcal{S}^\infty}[g(\vec{y})-g(\vec{x})]\frac{h(\vec{y})}{h(\vec{x})}K^\infty(\vec{x},\vec{y}).$$

We call this the generator of the twisted network $\mathcal{W}^\infty = (\tilde{\mathcal{W}}^\infty, \hat{\mathcal{W}}^\infty)$ and we denote the kernel by $\mathcal{K}^\infty$. Hence

$$\mathcal{K}^\infty(\vec{x},\vec{y}) = \frac{h(\vec{y})}{h(\vec{x})}K^\infty(\vec{x},\vec{y}). \tag{2}$$

Of course, the solution h must produce a twisted process such that $\tilde{\mathcal{W}}^\infty$ drifts to plus infinity while $\hat{\mathcal{W}}^\infty$ must be a stable Markov chain. If this fails then we must try again by twisting another set of coordinates; that is we must redefine the super stable nodes.

We perform the h-transform or twist of the nodes which causes the workload of the twisted network to overload. Thus $\mathcal{K}^\infty(\vec{x},\vec{y})$ is the kernel corresponding to the change of measure which induces the importance sampling estimator.

3 Overloading the first node

Since the first node is overloaded we take $\Delta = \{(x,y); x = 0, y \in \mathbf{N}\}$ and $\mathcal{S}^\infty = \mathcal{I}\times\mathbf{N}$. To calculate the twist constants α and β for the harmonic function $h(x,y) = \alpha^x\beta^y$, remark that the constraint in the interior, int$(\mathcal{S})$, is

$$\lambda\alpha + \mu_1\alpha^{-1}\beta + \mu_2\beta^{-1} = 1.$$

The constraint on the x-axis, $\mathcal{S}_{\{2\}}$, is

$$\lambda\alpha + \mu_1\alpha^{-1}\beta = \lambda + \mu_1.$$

Subtracting the later constraint from the first yields $\mu_2\beta^{-1} = \mu_2$. Consequently $\beta = 1$. Substituting into the first constraint gives $\alpha = \mu_1/\lambda$. (Of course the other solution is $\alpha = 1$.) Therefore the harmonic function is

$h(x,y) = (\frac{\mu_1}{\lambda})^x$ and from the equation (2) the kernel $\mathcal{K}^\infty$ is given by

$$\begin{aligned}
\mathcal{K}^\infty((x,y),(x+1,y)) &= \frac{h(x+1,y)}{h(x,y)} K^\infty((x,y),(x+1,y)) \\
&= \frac{\mu_1}{\lambda} \cdot \lambda = \mu_1 \\
\mathcal{K}^\infty((x,y),(x-1,y+1)) &= \frac{h(x-1,y+1)}{h(x,y)} K^\infty((x,y),(x-1,y+1)) \\
&= \frac{\lambda}{\mu_1} \cdot \mu_1 = \lambda \\
\mathcal{K}^\infty((x,y),(x,y-1)) &= \frac{h(x,y-1)}{h(x,y)} K^\infty((x,y),(x,y-1)) \\
&= 1 \cdot \mu_2 = \mu_2
\end{aligned}$$

Hence the twisted process is the (μ_1, λ, μ_2)-network which is obtained by interchanging the arrival rate λ and the service rate μ_1 of the first node. Notice that the first node of the (μ_1, λ, μ_2)-network has the load μ_1/λ, larger than 1, which implies that this node overloads. On the other hand the second node remains stable since its load λ/μ_2 is smaller than 1.

4 Overloading the second node

4.1 For the case of $\mu_2 < \mu_1$

In this case we consider $\Delta = \{(x,y); x \in \mathbf{N}, y = 0\}$ and $\mathcal{S}^\infty = \mathbf{N} \times \mathcal{I}$. As before the constraint in the interior is

$$\lambda\alpha + \mu_1\alpha^{-1}\beta + \mu_2\beta^{-1} = 1.$$

The constraint on the y-axis, $\mathcal{S}_{\{1\}}$, is

$$\lambda\alpha + \mu_2\beta^{-1} = \lambda + \mu_2.$$

The solutions of the above equations are $\alpha = \beta = \mu_2/\lambda$, so that the harmonic function $h(x,y)$ is given by $h(x,y) = (\mu_2/\lambda)^{x+y}$. Therefore we have the kernel $\mathcal{K}^\infty$ as followings;

$$\mathcal{K}^\infty((x,y),(x+1,y)) = \frac{h(x+1,y)}{h(x,y)} K^\infty((x,y),(x+1,y))$$

$$\begin{aligned}
&= \frac{\mu_2}{\lambda}\cdot\lambda = \mu_2 \\
\mathcal{K}^{\infty}((x,y),(x-1,y+1)) &= \frac{h(x-1,y+1)}{h(x,y)}K^{\infty}((x,y),(x-1,y+1)) \\
&= (\frac{\mu_2}{\lambda})^{-1}\frac{\mu_2}{\lambda}\cdot\mu_1 = \mu_1 \\
\mathcal{K}^{\infty}((x,y),(x,y-1)) &= \frac{h(x,y-1)}{h(x,y)}K^{\infty}((x,y),(x,y-1)) \\
&= (\frac{\mu_2}{\lambda})^{-1}\cdot\mu_2 = \lambda \qquad (3)
\end{aligned}$$

If $\mu_2 < \mu_1$, then the twisted process (μ_2, μ_1, λ)-network has the overloaded second node and the stable first node. Hence the harmonic function $h(x,y) = (\mu_2/\lambda)^{x+y}$ is what we look for and the transition probabilities in equation (3) is the optimal changes of measure for fast simulation.

However, if $\mu_2 > \mu_1$, then the first node of the twisted process also overloads. It follows that that we have to seek another harmonic function. Since the first node as well as the second node overload, at this time we remove both the x-axis and the y-axis, i.e. we consider $\Delta = \mathcal{S}_{\{1\}} \cup \mathcal{S}_{\{2\}}$ and $\mathcal{S}^{\infty} = \mathcal{I}^2$. Then the harmonic function is of the form $h(x,y) = 1^x\beta^y$. In this case we have only one constraint in the interior such that

$$\lambda + \mu_1\beta + \mu_2\beta^{-1} = 1$$

which implies $\beta = \mu_2/\mu_1$. Thus the harmonic function is given by $h(x,y) = (\mu_2/\mu_1)^y$ so that the twisted process is the (λ, μ_2, μ_1)-network. But at this time the first node of the (λ, μ_2, μ_1)-network is super-stable. It means that we can't find the harmonic function by interchanging the parameters. To solve this problem we use the reversed process in the next subsection 4.2.

4.2 For the case of $\mu_2 > \mu_1$

Anantharam et al.(1990) showed that given a rare queue length vector has occurred, the process got there by following in reverse direction the path by which the reversed network empties from the rare state. Shwartz and Weiss(1993) basically show that the large-deviation paths for which the rare event occurs are identical to the time-reversal paths, provided the jump rates are constant in the interior of the state space. By using these results we can get the harmonic functions which imply the optimal changes of measure.

In the reversed network of the (λ, μ_1, μ_2)-network, arrivals occur at rate λ to the second node, and jobs flow from the second node to the first node and then exit the network. Therefore the reversed network of (λ, μ_1, μ_2)-network is the (λ, μ_2, μ_1)-network in the reversed direction. Its jump rates are depicted in Fig. 3. Since $\mu_2 > \mu_1 > \lambda$ the time-reversal path starting at the rate state $y = \ell$ goes with the slope $\frac{\lambda - \mu_2}{\mu_2 - \mu_1}$ until the x-axis is reached. On the x-axis the path directs to the empty state because $\mu_1 > \lambda$. This time-reversal path is pictured in Fig. 3.

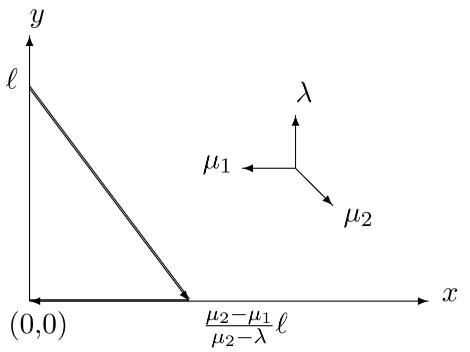

Fig. 3. The path and jump rates of the time-reversed network

Interpreting this path for the forward-time process implied that buildup occurs in two phases. The first phase corresponds to starting at the origin (0,0) and building up the first node until the level $\frac{\mu_2-\mu_1}{\mu_2-\lambda}\ell$. Then, in phase 2 when both nodes are non-empty, the system is run to get $y = \ell$. We can now find two harmonic functions and twisted processes to produce the phase 1 and 2, respectively.

(i) the phase 1

To reach $x = \frac{\mu_2-\mu_1}{\mu_2-\lambda}\ell$ firstly, we remove the y-axis. The constraint in the interior is given by

$$\lambda\alpha + \mu_1\alpha^{-1}\beta + \mu_2\beta^{-1} = 1.$$

The constraint on the x-axis, $\mathcal{S}_{\{2\}}$, is

$$\lambda\alpha + \mu_1\alpha^{-1}\beta = \lambda + \mu_1.$$

It follows that the harmonic function is of the form $h(x,y) = (\frac{\mu_1}{\lambda})^x$ which is the same as that in the section 3. Hence the twisted process is the (μ_1, λ, μ_2)-network.

(ii) the phase 2

In order to move along the path with the slope $\frac{\lambda-\mu_2}{\mu_2-\mu_1}$ from $x = \frac{\mu_2-\mu_1}{\mu_2-\lambda}\ell$ to $y = \ell$ the harmonic function $h(x,y)$ should be

$$h(x,y) = (\frac{\mu_1}{\lambda})^x(\frac{\mu_2}{\lambda})^y.$$

and the kernel $\mathcal{K}^\infty$ is given by

$$\begin{aligned}
\mathcal{K}^\infty((x,y),(x+1,y)) &= \frac{h(x+1,y)}{h(x,y)}K^\infty((x,y),(x+1,y)) \\
&= \frac{\mu_1}{\lambda}\cdot\lambda = \mu_1 \\
\mathcal{K}^\infty((x,y),(x-1,y+1)) &= \frac{h(x-1,y+1)}{h(x,y)}K^\infty((x,y),(x-1,y+1)) \\
&= (\frac{\mu_1}{\lambda})^{-1}\frac{\mu_2}{\lambda}\cdot\mu_1 = \mu_2 \\
\mathcal{K}^\infty((x,y),(x,y-1)) &= \frac{h(x,y-1)}{h(x,y)}K^\infty((x,y),(x,y-1)) \\
&= (\frac{\mu_2}{\lambda})^{-1}\cdot\mu_2 = \lambda
\end{aligned}$$

so that the twisted process is the (μ_1, μ_2, λ)-network.

References

[1] V. Anantharam, P. Heidelberger, and P. Tsoucas, *Analysis of rare events in continuous time Markov chains via time reversal and fluid approximation*, Rep. RC 16280, IBM, Yorktown Height, New York.

[2] S. Asmussen, *Conditioned Limit Theorems Relaing a Random Walk to Its Associate, with Applications to Risk Processes and the GI/G/1 queue*, Adv. Appl. Prob. **14** (1982), 143-170.

[3] M. R. Frater, T. M. Lennon and B. D. O. Anderson, *Optimally Efficient Estimation of the Statistics of Rare Events in Queueing Networks*, IEEE Trans. Auto. Control **36** (1991), 1395-1405.

[4] M. R. Frater and B. D. O. Anderson, *Fast Simulation of Buffer Overflows in Tandem Networks of GI/GI/1 Queues*, Ann. Oper. Res. **49** (1994), 207-220.

[5] P. Glasserman and S-G Kou, *Analysis of an Importance Sampling Estimator for Tandem Queues*, ACM Trans. Model. Comput. Simul. **5** (1995), 22-42.

[6] D. McDonald, *Asymptotics of First Passage Times for Random Walk in an Orthant*, Ann. Appl. Prob. **9** (1999), 110-145.

[7] S. Parekh and J. Walrand, *A Quick Simulation Method for Excessive Backlogs in Networks of Queues*, IEEE Trans. Auto. Control **34** (1989), 54-66.

[8] A. Shwartz and A. Weiss, *Induced rare events: analysis via time reversal and large deviations*, Adv. Appl. Prob. **25** (1993), 667-689.

[9] J. Walrand, *An Introduction to Queueing Network*, Prentice-Hall International, Inc., 1988.

[10] R. R. Weber, *The Interchangeability of* $\cdot/M/1$ *Queues in Series*, J. Appl. Prob. **16** (1979), 690-695.