SIMULATION

http://sim.sagepub.com

Efficient Simulation of a Tandem Queue with Server Slow-down

D.I. Miretskiy, W.R.W. Scheinhardt and M.R.H. Mandjes

SIMULATION 2007; 83; 751
DOI: 10.1177/0037549707086193

Efficient Simulation of a Tandem Queue with Server Slow-down

D.I. Miretskiy
W.R.W. Scheinhardt
Department of Applied Mathematics
University of Twente
Postbus 217
7500 AE Enschede
The Netherlands
d.miretskiy@math.utwente.nl

M.R.H. Mandjes
University of Amsterdam
Korteweg-de Vries Institute for Mathematics
Plantage Muidergracht 24
1018 TV Amsterdam
The Netherlands

Tandem Jackson networks and more sophisticated variants have found widespread application in various domains. One such variant is the tandem queue with server slow-down, in which the server of the upstream queue reduces its service speed as soon as the downstream queue exceeds some pre-specified threshold, to provide the downstream queue some sort of 'protection'. This paper focuses on the overflow probabilities in the downstream queue. Owing to the Markov structure these can be solved numerically, but the resulting system of linear equations is usually large. An attractive alternative could be to resort to simulation, but this approach is cumbersome due to the rarity of the event under consideration. A powerful remedy is to use importance sampling, i.e. simulation under an alternative measure, where unbiasedness of the estimator is retrieved by weighing the observations by appropriate likelihood ratios. To find a good alternative measure, we first identify the most likely path to overflow. For the standard tandem queue (i.e. no slow-down) this path was known, but we develop an appealing novel heuristic which can also be applied to the model with slow-down. The knowledge of the most likely path is then used to devise importance sampling algorithms, both for the standard tandem system and for the system with slow-down. Our experiments indicate that the corresponding new measure is sometimes asymptotically optimal, and sometimes not. We systematically analyze the cases that may occur.

Keywords: Rare event simulation, importance sampling, server slow-down, tandem Jackson network

1. Introduction

Tandem Jackson networks have found widespread application in various domains as they are simple but powerful and, due to their Markovian structure, amenable to analysis. The standard tandem model, however, is not always realistic. For instance, in many practical situations, the service stations share information about their current buffer content, and use this information to facilitate effective network management. In this paper we study such a model: a tandem queue that consists of two nodes or servers where, in order to protect the second (downstream) queue from overflow, the first (upstream) server keeps track of queue length at the second server and lowers its service rate when the second queue is large. Van Foreest et al. [1] have already introduced this model and studied the consequences for the first queue. Our main interest is to determine the probability of overflow in the *second* queue

SIMULATION, Vol. 83, Issue 11, November 2007 751–768

DOI: 10.1177/0037549707086193

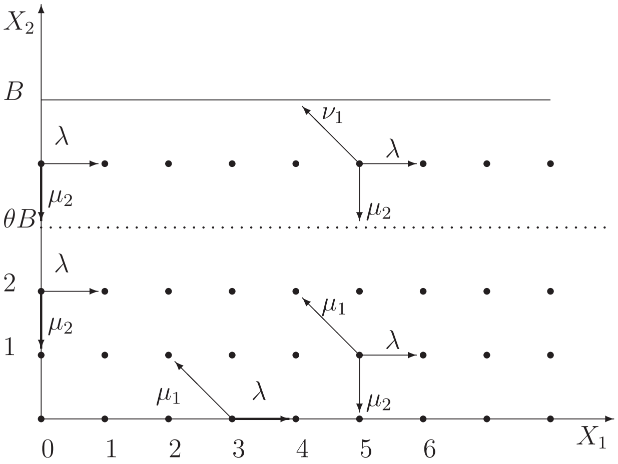

Figure 1. State space and transition structure for $(X_1(t), X_2(t))$

during a busy cycle of the joint system, which is defined as the time between two consecutive arrivals to an empty system.

To be more specific, let us denote the number of jobs in queue i at time t by $X_i(t), i = 1, 2$. Jobs arrive at the first queue according to a Poisson process at rate λ, and the service rate of server i is μ_i. However, the rate of the first server reduces to $\nu_1 < \mu_1$ ('slow-down') at times when $X_2(t)$ is at or above some threshold value. Instead of assuming that the second queue has a finite buffer of capacity B, we prefer to analyze a system in which both buffers are infinitely large and then consider the probability p_B that during a busy cycle of the joint system the second buffer reaches a high level B. The value of the threshold is θB with $\theta \in [0, 1]$, so that it scales with B. For a typical state space representation of the Markov process $\{(X_1(t), X_2(t)), t \geq 0\}$ we refer to Figure 1. Note that when we choose $\theta = 1$, the probability p_B is the same as that for a standard tandem Jackson network.

Even for the standard Jackson network, no explicit expression for p_B is known. For fixed B we can in principle obtain numerical values by truncating the state space in the horizontal direction and then solving a (large) system of linear equations, but this is not very practical when B is large. In that case an attractive alternative would be to use simulations, but due to the rarity of the event of interest it would require an extremely large number of replications to obtain a good estimate of p_B. To avoid this difficulty we employ the Importance Sampling (IS) method, which is one of the most common tools in rare event simulation. The main idea of IS is to make the probability of interest much higher by simulating under an alternative measure, and then weighing the observations with appropriate likelihood ratios.

To obtain a good alternative measure we first identify the most likely path to overflow, i.e. the way in which overflow most probably occurs, conditional on its occurrence. Typically, a good alternative measure is such that the process will closely follow this path with high probability under that measure. For the standard Jackson case the most likely path is already known (Anatharam et al. [2]), as opposed to the model with slow-down. We have developed an appealing method for finding the most likely path, which is heuristic in nature since it is based on a conjecture. When we apply this method to the standard Jackson case, it indeed yields the path of [2]. The shape of this path depends critically on the values of the parameters (arrival and services rates). The path is then translated into an alternative measure (i.e. new arrival and service rates), under which most paths lead to overflow by realizations close to the most likely path. Unfortunately, when performing IS under this measure, it turns out that the measures we find are not asymptotically optimal for all parameter values. We systematically analyze the cases that may occur: first for the standard Jackson network in Section 3, since we believe the results there are interesting on their own, and then for the slow-down model in Section 4. We conclude with some open problems for future research in Section 5. Some detailed explanations about the shape of the most likely path to overflow in the second buffer are provided in the Appendix (for both models).

We finish this section by relating our work to some existing literature. Most results on efficient simulation for tandem Jackson networks deal with the probability that the total network population exceeds some large value during a busy cycle. Parekh and Walrand [3] proposed an alternative measure and Glasserman and Kou [4] found that this measure is not always asymptotically efficient; see also de Boer [5]. A more accurate state-dependent change of measure for the same problem was introduced by Zaburnenko and Nicola [6]. The special case of both servers having equal rates was studied by Sandmann [7]. Kroese and Nicola [8] focus on the second queue reaching a high level as in our study, but during a busy cycle of the second queue which is different from the busy cycle of the system that we consider.

2. Model and Preliminaries

2.1 Model

First we consider a two-node tandem Jackson network with Poisson arrivals at rate λ and two stations with exponentially distributed service times with parameters μ_1 and μ_2. For convenience we choose the parameters such that $\lambda + \mu_1 + \mu_2 = 1$, without loss of generality. Both buffers are assumed to be infinitely large. Let $X(t) = \{(X_1(t), X_2(t)), t \geq 0\}$ be the joint queue-length process. This process is regenerative if we impose the stability condition $\lambda < \min(\mu_1, \mu_2)$, which we will do from now on. The limiting distribution of the process is well known and given by

$$\begin{aligned}\pi(i,j) &= \lim_{t\to\infty}\mathbb{P}(X_1(t)=i, X_2(t)=j)\\ &= (1-\rho_1)(1-\rho_2)\rho_1^i\rho_2^j, \qquad (1)\end{aligned}$$

where $\rho_i = \lambda/\mu_i$. Our main interest is to estimate the probability of reaching some high level B in the second buffer during a busy cycle of the system.

We now describe the slow-down model. Consider again a two-node network as above with infinitely large buffers. In addition, when the number of jobs in the second buffer exceeds some *slow-down threshold*, the first service station decreases its rate such that the (remaining) service times are exponential with parameter $\nu_1 < \mu_1$. We are once more interested in the estimation of the probability of reaching some high level B in the second buffer during a busy cycle, where we assume that the slow-down threshold scales with B as θB for some $\theta \in [0,1]$. It is important that we still choose $\lambda + \mu_1 + \mu_2 = 1$, without loss of generality, but then clearly $\lambda + \nu_1 + \mu_2 < 1$. See Figure 1 for an illustration of the state space and transition structure; taking $\theta = 1$ corresponds to the standard tandem Jackson network since we only consider the process as long as $X_2(t) < B$.

2.2 Objectives

As stated, our main interest is to estimate the probability of reaching some high level B in the second buffer during a busy cycle T, which is defined as the time between two successive epochs at which the process leaves the empty state $(0,0)$. In this sense, the busy cycle consists of a busy period and the subsequent idle period. When we define the random variable T_B as the first entrance time of either level B or state $(0,0)$, i.e.

$$\begin{aligned}T_B &= \min\{t > 0 | X_2(t) = B \quad \text{or}\\ (X_1(t), X_2(t)) &= (0,0)\},\end{aligned}$$

then we can formally define the probability of interest as

$$p_B := \mathbb{P}_{(1,0)}(X_2(T_B) = B), \qquad (2)$$

where $\mathbb{P}_{(1,0)}$ denotes the conditional probability given that $(X_1(0), X_2(0)) = (1,0)$. Note that the starting state is $(1,0)$ here, because every busy cycle starts with an arrival to queue 1. For fixed B, we can obtain this probability analytically by solving a system of equations for $x(i,j) = \mathbb{P}_{(i,j)}(X_2(T_B) = B)$ of the form $x(i,j) = \lambda x(i+1,j) + \mu_1 x(i-1,j+1) + \mu_2 x(i,j-1)$ on the interior; we have similar equations for the boundaries. Unfortunately, it is time consuming to solve (the truncated version of) such a system, which motivated us to choose simulation as a main tool for this paper.

Due to the stability condition the overflow event becomes rare as B grows large, and hence p_B will become small. The following theorem specifies how this happens in the standard tandem Jackson network.

Theorem 1. *For the standard tandem Jackson network, the overflow probability p_B is asymptotically geometric in B with parameter ρ_2. More precisely,*

$$\lim_{B\to\infty}\frac{1}{B}\log p_B = \log \rho_2. \qquad (3)$$

Proof. The proof uses arguments that are somewhat similar to those used in the proof of theorem 1 in Anantharam [9]. Denoting the indicator function of an event A by $\mathbf{1}(A)$, we can define the time spent by the process at level B during a busy cycle T as

$$I_{(\cdot,B)} = \int_0^T \mathbf{1}(X_2(t) = B)dt. \qquad (4)$$

A simple renewal argument tells us that $\pi(0,0) = \lambda^{-1}/\mathbb{E}_{(1,0)}T$, and also that

$$\begin{aligned}\pi(\cdot,B) &= \frac{\mathbb{E}_{(1,0)}I_{(\cdot,B)}}{\mathbb{E}_{(1,0)}T}\\ &= p_B\mathbb{E}_{(1,0)}(I_{(\cdot,B)} \mid I_{(\cdot,B)} > 0)\lambda\pi(0,0).\end{aligned}$$

Thus, using equation (1), we can write

$$p_B = \frac{1}{\lambda(1-\rho_1)}\frac{{\rho_2}^B}{\mathbb{E}_{(1,0)}(I_{(\cdot,B)} \mid I_{(\cdot,B)} > 0)}. \qquad (5)$$

Using equation (1), we can also show that $\mathbb{E}_{(1,0)}(I_{(i,B)} \mid I_{(\cdot,B)} > 0) = \rho_1^i\mathbb{E}_{(1,0)}(I_{(0,B)} \mid I_{(\cdot,B)} > 0)$ which leads us to

$$\begin{aligned}&\mathbb{E}_{(1,0)}(I_{(\cdot,B)} \mid I_{(\cdot,B)} > 0)\\ &= \frac{\mathbb{E}_{(1,0)}(I_{(0,B)} \mid I_{(\cdot,B)} > 0)}{1-\rho_1}. \qquad (6)\end{aligned}$$

Now we condition on the number of jobs in the first queue when the second queue reaches level B, and note that for all $i > 0$,

$$(I_{(0,B)} \mid X_1(T_B) = i) \overset{d}{=} \begin{cases} 0, \text{ if } (0,B) \text{ is not visited during the cycle,}\\ (I_{(0,B)} \mid X_1(T_B) = 0),\\ \text{if } (0,B) \text{ is visited during the cycle.}\end{cases}$$

We can therefore write

$$\mathbb{E}_{(1,0)}(I_{(0,B)} \mid I_{(\cdot,B)} > 0)$$

$$= \sum_{i=0}^{\infty} \mathbb{E}_{(1,0)}(I_{(0,B)} \mid X_1(T_B) = i, I_{(\cdot,B)} > 0)\mathbb{P}(X_1(T_B)$$

$$= i \mid I_{(\cdot,B)} > 0) \leq \sum_{i=0}^{\infty} \mathbb{E}_{(1,0)}(I_{(0,B)} \mid X_1(T_B)$$

$$= 0, I_{(\cdot,B)} > 0)\mathbb{P}(X_1(T_B) = i \mid I_{(\cdot,B)} > 0)$$

$$= \mathbb{E}_{(1,0)}(I_{(0,B)} \mid X_1(T_B) = 0, I_{(\cdot,B)} > 0)$$

$$= \frac{1}{(\lambda + \mu_2)q_B}, \tag{7}$$

where $q_B = \mathbb{P}_{(0,B)}((0,0)$ reached before $(0,B))$. It is clear that the sequence q_B is strictly decreasing in B and that $\lim_{B\to\infty} q_B = q_\infty \in (0,1)$ exists, which implies that for any positive B,

$$q_B > q_\infty. \tag{8}$$

Combining equations (5)–(8), we derive a lower bound on p_B. Using the simple fact that $\mathbb{E}_{(1,0)}(I(\cdot,B) \mid I(\cdot,B) > 0) \geq 1/\mu_2$ in equation (5), we also find an upper bound. This leads us to

$$\frac{\lambda + \mu_2}{\lambda} q_\infty \rho_2^B \leq p_B \leq \frac{\rho_2^{B-1}}{1-\rho_1},$$

from which we have

$$\log \rho_2 \leq \lim_{B\to\infty} \frac{1}{B} \log p_B \leq \log \rho_2.$$

□

Theorem 1 is important in itself, as it provides already a rough estimate for the probability of interest (equation (2)) for large B. In fact, Theorem 1 states that p_B is of the form $f(B)\rho_2^B$ where $\log f(B)/B \to 0$ as B grows large. To obtain p_B more precisely, we will use estimates based on simulations. Secondly, the theorem is important as it will help us to verify the asymptotic optimality of the estimators involved in these simulations.

2.3 The Optimal Path

In order to find a good change of measure for IS simulations, the first step is usually to find the 'optimal path to overflow', i.e. the way in which overflow most probably occurs, conditional on its occurrence. To this end one usually considers the 'B-scaled network', which corresponds to $\frac{X(t)}{B} = \left(\frac{X_1(t)}{B}, \frac{X_2(t)}{B}\right)$ in our tandem system. Our target probability is equivalent to the probability that the second component of this scaled process reaches 1 before the process returns to the origin. The optimal path is often used to state a 'law of large numbers for rare events', in the spirit of identifying a path $x^*(t) = \left(x_1^*(t), x_2^*(t)\right)$ such that

$$\mathbb{P}\left(\left\|\frac{X(t)}{B} - x^*(t)\right\| > \varepsilon \,\middle|\, \frac{X_2(T_B)}{B} = 1\right) \to 0$$

as $B \to \infty$ for all $\varepsilon > 0$, where $\|\cdot\|$ is some metric.

Such a path has already been identified for general Jackson networks by Anantharam et al. [2] (and hence also for our tandem system). Anantharam et al. [2] use the time-reversed process to find the shape of the most probable path to overflow. In fact, it is shown that this path can have two different forms, depending on the relation between μ_1 and μ_2. If the second server is the bottleneck ($\mu_2 < \mu_1$), the optimal path to overflow has a very simple shape: the second buffer fills up gradually, while the first queue remains virtually empty. On the other hand, when the first queue is the bottleneck we have a more complicated situation, in which the path consists of two parts. During the first part the second queue stays virtually empty while the number of jobs in the first buffer grows to some value that is proportional to B. During the second part, the number of jobs in the first buffer decreases (virtually to 0) while the second buffer fills up to B.

In the remainder of this section we present another method to find the optimal path. This method is heuristic by nature, but has important advantages. First, it not only yields the shape of the optimal path, but also gives a 'good' change of measure, which will ensure that most simulation runs under this new measure will be close to the optimal path. Secondly, we note that in the slow-down model, which is our ultimate interest in this paper, we cannot use the method described by Anantharam et al. [2] since we do not know the explicit form of the stationary distribution in that case. We therefore cannot use an analysis based on the time-reversed process. Our heuristic method, however, *can* be applied here.

The method is based on the use of a family of cost functions I defined by

$$I(\tilde{\lambda} \mid \lambda) = \lambda - \tilde{\lambda} + \tilde{\lambda} \log\left(\frac{\tilde{\lambda}}{\lambda}\right) \tag{9}$$

(see also Shwartz and Weiss [10]). Note that the function (9) is convex and equals 0 at $\tilde{\lambda} = \lambda$. Intuitively, we can think of the value $I(\tilde{\lambda} \mid \lambda)$ as the cost we need to pay to allow a Poisson process with parameter λ behave like a Poisson process with parameter $\tilde{\lambda}$, per time unit. The idea behind this heuristic is the following. If $N(t)$ is the number of arrivals generated by a Poisson process of rate λ, we may be interested in $\mathbb{P}(N(t) \approx \tilde{\lambda}t)$. Assume for ease that t is an integer; then $N(t)$ is distributed as the sum of

t independent and identically distributed (i.i.d.) Poisson random variables, each with mean λ. Cramér's theorem then says

$$\mathbb{P}(N(t) \approx \tilde{\lambda} t)$$

$$\approx \exp\left\{-t \sup_{\theta}\left[\theta\tilde{\lambda} - \log \mathbb{E}\exp(\theta N(1))\right]\right\}, \quad (10)$$

and it is easy to verify that the latter expression reduces to

$$\exp(-tI(\tilde{\lambda} \mid \lambda)). \quad (11)$$

For instance, consider for any i a straight path from (0, 0) to (i, B) through the interior of the state space, staying away from the boundaries. We then need to replace the parameters by 'tilded' parameters $\tilde{\lambda}$, $\tilde{\mu}_1$ and $\tilde{\mu}_2$, such that $\tilde{\mu}_1 > \tilde{\mu}_2$ and $\tilde{\lambda} \geq \tilde{\mu}_1$, in order to have northeast drift . The total cost of such a path per unit length in the vertical direction is

$$\frac{\mathbb{I}(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2)}{\tilde{\mu}_1 - \tilde{\mu}_2}, \quad (12)$$

where $\mathbb{I}(\tilde{\lambda},\ \tilde{\mu}_1,\ \tilde{\mu}_2) = I(\tilde{\lambda} \mid \lambda) + I(\tilde{\mu}_1 \mid \mu_1) + I(\tilde{\mu}_2 \mid \mu_2)$ represents the total cost per unit time, and the denominator is the average speed by which the process moves up. If we replaced the denominator by $\tilde{\lambda} - \tilde{\mu}_1$, we would find the cost per unit length in the horizontal direction. Finally, we mention that the (negative) slope of this path is given by

$$\alpha = \frac{\tilde{\mu}_1 - \tilde{\mu}_2}{\tilde{\lambda} - \tilde{\mu}_1}. \quad (13)$$

Minimizing expression (12) over the three tilded parameters, such that $\tilde{\mu}_1 > \tilde{\mu}_2$ and $\tilde{\lambda} \geq \tilde{\mu}_1$ also hold, will then give the optimal values for the tilded parameters and the slope of the path for this particular shape. Equations (10) and (11) indicate that the exponent of the negative of expression (12) can be used as an approximation of the probability of interest, but conditional on a path of the given type (with $\tilde{\lambda} \geq \tilde{\mu}_1$). By considering all possible path types we will obtain an approximation of the probability of overflow in the second buffer during a busy cycle, as well as the most probable path (i.e. the minimizing values of $\tilde{\lambda}$, $\tilde{\mu}_1$ and $\tilde{\mu}_2$). The same ideas can be applied to the slow-down model.

Note that we can also associate cost to a non-straight path $r(t) = (x_1(t), x_2(t))$ with $t \in [0, T]$ for some T, namely as $\int_0^T \ell(r(t), r'(t))dt$, where ℓ is the so-called local rate function [10, equation (5.5)]. If $r'(t)$ is locally equal to $(\tilde{\lambda} - \tilde{\mu}_1, \tilde{\mu}_1 - \tilde{\mu}_2)$ with $\tilde{\lambda}\ \tilde{\mu}_1\ \tilde{\mu}_2 = \lambda\ \mu_1\ \mu_2$, the relation with our cost functions is given by $\ell(r(t), r'(t)) = \mathbb{I}(\tilde{\lambda},\ \tilde{\mu}_1,\ \tilde{\mu}_2)$, which can be seen by noting that the solution to equation (5.2) in Shwartz and Weiss [10] is given by $\theta_1 = \log(\tilde{\lambda}/\lambda)$ and $\theta_2 = -\log(\tilde{\mu}_2/\mu_2)$.

To proceed, let us formulate a conjecture and a theorem upon which our research will be based. Consider the slow-down system, with $\theta \in [0, 1]$. The state space consists of four subsets on which the transition parameters are constant, namely: $\{(0, 0)\}$, $\{(n, 0)\}$, $\{(0, m)\}$, $\{(n, m)\}$ and $\{(n, m')\}$ where $n, m > 0$ and $m < \theta B \leq m'$. See Figure 1.

Conjecture 2. *The path with minimal cost in terms of cost-function (9), is the most probable path between any two points.*

Theorem 3. *Consider the slow-down system, with $\theta \in$ [0, 1] and assume that Conjecture 2 is true, then the typical path which starts in the empty state and leads to overflow in a single node or in the total queue consists of a concatenation of subpaths on the various subsets that are straight lines; each subset is traversed at most once.*

The great benefit of Theorem 3 is that the solution now boils down to optimizing over a finite number of possible path-types, i.e. we reduced the problem to a combinatorial problem. The proof of the theorem is based on the following two lemmas.

Lemma 4. *The optimal path between two states in the same subset is a straight line.*

Proof. This follows from Lemma 5.16 of Shwartz and Weiss [10], but to provide more insight we present an alternative proof for the following special case. Let us consider two states in the interior $x = (x_1, x_2)$ and $y = (y_1, y_2)$, both below or both above θB. We need to show that the optimal path from x to y is a straight line. To this end we consider a path from x to y via an additional point $z = (z_1, z_2)$ and find the values of z_1 and z_2 which minimize the total cost of such a path. First consider states x, y and z such that $x_1 \leq z_1 \leq y_1$ and $x_2 \leq z_2 \leq y_2$. The optimal cost per unit length in the vertical direction of such a path is

$$\inf\left\{(z_2 - x_2)\frac{\mathbb{I}(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2)}{\tilde{\mu}_1 - \tilde{\mu}_2}\right.$$

$$\left. + \quad (y_2 - z_2)\frac{\mathbb{I}(\bar{\lambda}, \bar{\mu}_1, \bar{\mu}_2)}{\bar{\mu}_1 - \bar{\mu}_2}\right\}. \quad (14)$$

The infimum is taken over variables $\tilde{\lambda}$, $\tilde{\mu}_1$, $\tilde{\mu}_2$, $\bar{\lambda}$, $\bar{\mu}_1$, $\bar{\mu}_2$, z_1 and z_2 that satisfy $\tilde{\lambda} \geq \tilde{\mu}_1$, $\tilde{\mu}_1 \geq \tilde{\mu}_2$, $\bar{\lambda} \geq \bar{\mu}_1$, $\bar{\mu}_1 \geq \bar{\mu}_2$. Note that the knowledge of starting and ending points for each subpath gives us a possibility to express $\tilde{\lambda}$ and $\bar{\lambda}$ from expression (14) in terms of the other variables, i.e.

$$\tilde{\lambda} = \frac{(z_1 - x_1)(\tilde{\mu}_1 - \tilde{\mu}_2) + (z_2 - x_2)\tilde{\mu}_1}{z_2 - x_2}$$

and

$$\bar{\lambda} = \frac{(y_1 - z_1)(\bar{\mu}_1 - \bar{\mu}_2) + (y_2 - z_2)\bar{\mu}_1}{y_2 - z_2}.$$

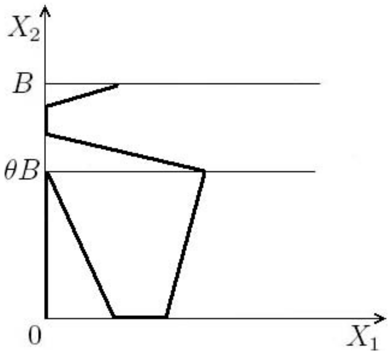

Figure 2. Example of a path with two vertical and two horizontal subpaths

As a result we obtain

$$z_1 = \frac{x_1 - y_1}{x_2 - y_2} z_2 + \frac{x_2 y_1 - x_1 y_2}{x_2 - y_2} \tag{15}$$

for any z_2. Equality (15) guarantees that z lies on the line which connects x and y. The same statement can be proved for arbitrary choices of x, y and z in an equal manner. This completes the proof. □

Lemma 5. *The optimal path does not have more then one subpath in each subset.*

Proof. At first we will focus on a path that has two subpaths in the interior. It is clear that the path could not be optimal if it includes two consecutive subpaths in the same subset, by Lemma 4, so we concentrate on paths that have two subpaths in the interior, with a connecting subpath on one of the boundaries in between; see Figure 2. We will show that it is not optimal to have two subpaths in the interior below θB, using the path from Figure 2 as a typical example.

The minimal cost of the first four subpaths is

$$\inf \left\{ \theta B \frac{\mathbb{I}(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2)}{\tilde{\lambda} - \tilde{\mu}_2} + \theta B \frac{\mathbb{I}(\bar{\lambda}, \bar{\mu}_1, \bar{\mu}_2)}{\bar{\mu}_2 - \bar{\mu}_1} \right.$$

$$\left. + \quad \beta \frac{\mathbb{I}(\hat{\lambda}, \hat{\mu}_1, \hat{\mu}_2)}{\hat{\lambda} - \hat{\mu}_1} + \theta B \frac{\mathbb{I}(\breve{\lambda}, \breve{\mu}_1, \breve{\mu}_2)}{\breve{\mu}_1 - \breve{\mu}_2} \right\}, \tag{16}$$

where

$$\beta = \theta B \left(\frac{\bar{\lambda} - \bar{\mu}_1}{\bar{\mu}_2 - \bar{\mu}_1} + \frac{\breve{\lambda} - \breve{\mu}_1}{\breve{\mu}_1 - \breve{\mu}_2} \right) > 0,$$

and the infimum is taken over all $\tilde{\lambda}$, $\tilde{\mu}_1$, $\tilde{\mu}_2$, $\bar{\lambda}$, $\bar{\mu}_1$, $\bar{\mu}_2$, $\hat{\lambda}$, $\hat{\mu}_1$, $\hat{\mu}_2$, $\breve{\lambda}$, $\breve{\mu}_1$ and $\breve{\mu}_2$, such that $\tilde{\lambda} > \tilde{\mu}_2$, $\tilde{\lambda} < \tilde{\mu}_1$, $\bar{\mu}_1 < \bar{\mu}_2$, $\hat{\lambda} > \hat{\mu}_1$, $\hat{\mu}_1 < \hat{\mu}_2$ and $\breve{\mu}_1 > \breve{\mu}_2$.

We split the problem into two cases. When $\mu_1 > \mu_2$ we obtain after optimization $(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2) = (\mu_2, \mu_1, \lambda)$, $(\bar{\lambda}, \bar{\mu}_1, \bar{\mu}_2) = (\hat{\lambda}, \hat{\mu}_1, \hat{\mu}_2) = (\mu_1, \lambda, \mu_2)$ and $(\breve{\lambda}, \breve{\mu}_1, \breve{\mu}_2) = (\mu_1, \mu_2, \lambda)$. Since the second and third parts have the same cost of moving in the horizontal direction, it is obvious that a path that does not contain the first part (following the vertical boundary) will always have a lower cost than the path in Figure 2. When $\mu_1 < \mu_2$, it can be shown in a similar way that the cheapest path from $(0, 0)$ to $(x, 0)$ is again a straight line which follows the horizontal boundary; see also Section 3, Case 1. This means that the cost of the path in Figure 2 is always bounded from below by the cost of a path which satisfies the lemma.

Note that a path which consists of two subpaths in the interior and a connecting subpath on the vertical axis also cannot be optimal. Using the same arguments, one can prove that the optimal path does not contain two (or more) subpaths in the interior above θB. This completes the proof. □

3. Tandem Jackson Network

For the standard tandem Jackson network, we consider the minimal costs of all possible path types that satisfy Theorem 3 with $\theta = 1$. As a result, we obtain the most probable path to overflow as the path with globally minimal cost, and the associated change of measure. The cost function itself will yield $-\log \rho_2$ as its optimum value, since we already know that ρ_2 is the geometric decay rate of the probability of interest for the tandem model; see Theorem 1.

We split the problem for the standard tandem Jackson network into two cases: (1) $\lambda < \mu_1 < \mu_2$, i.e. the first server is the bottleneck; and (2) $\lambda < \mu_2 < \mu_1$, i.e. the second server is the bottleneck. At the end of this subsection we will focus on the special case in which $\lambda < \mu_1 = \mu_2$ and argue that Case 1 can be extended to $\lambda < \mu_1 \leq \mu_2$.

Case 2, i.e. $\lambda < \mu_2 < \mu_1$

We prefer to start our analysis with Case 2, since this is the simplest problem. We consider a path that follows the vertical axis. To find the optimally tilded parameters for such a path we need to solve the minimization problem

$$I_2 = \inf \left\{ \frac{\mathbb{I}(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2)}{\tilde{\lambda} - \tilde{\mu}_2} \right\}, \tag{17}$$

where the infimum is taken over all tilded variables $\tilde{\lambda}$, $\tilde{\mu}_1$, $\tilde{\mu}_2$, such that $\tilde{\mu}_1 > \tilde{\lambda}$ and $\tilde{\lambda} > \tilde{\mu}_2$, ensuring a northwest drift (i.e. to the left and up). Note that the denominator is again the average speed at which the process moves up; it is $\tilde{\lambda} - \tilde{\mu}_2$ instead of $\tilde{\mu}_1 - \tilde{\mu}_2$ since the first queue is stable ($\tilde{\lambda} < \tilde{\mu}_1$), so the second queue fills up at rate $\tilde{\lambda}$ rather than $\tilde{\mu}_1$. After taking partial derivatives with respect to all tilded variables and setting them equal to zero, some algebra leads us to the solutions $(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2) = (\lambda, \mu_1, \mu_2)$ and $(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2) = (\mu_2, \mu_1, \lambda)$. However, only the second solution satisfies both boundary conditions $\tilde{\mu}_1 > \tilde{\lambda}$ and $\tilde{\lambda} > \tilde{\mu}_2$, so the minimal cost of this type of path is $I_2 = -\log(\rho_2)$ per unit vertical length.

We checked all other possible shapes of the path to overflow (for a detailed account, see Appendix A1) and conclude that for this case I_2 is in fact the minimal cost per unit movement in the vertical direction, and ρ_2 is indeed the decay rate.

Proposition 6. *If $\lambda < \mu_2 < \mu_1$ (Case 2) then the optimal path to overflow of the second buffer has the following shape:* $(0, 0) \to (0, B)$ *and the decay rate is ρ_2. The corresponding change of measure is given by*

$$(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2) = (\mu_2, \mu_1, \lambda). \tag{18}$$

We note that the notation $(x_1, x_2) \to (y_1, y_2)$ represents a straight line from state x to state y (for the scaled process); in this case the path follows the vertical boundary due to the northwest drift under the change of measure.

We now investigate the result of the first queue being the bottleneck.

Case 1, i.e. $\lambda < \mu_1 < \mu_2$

We present the minimization problem for the path to overflow as described by Anantharam et al. [2]. We therefore assume we have tilded parameters that satisfy $\tilde{\mu}_1 < \tilde{\mu}_2$ and $\tilde{\mu}_1 < \tilde{\lambda}$ to ensure a path along the horizontal axis, southeast drift. For the second part of the path we have parameters $\bar{\lambda}$, $\bar{\mu}_1$ and $\bar{\mu}_2$ such that $\bar{\mu}_1 > \bar{\mu}_2$ and $\bar{\lambda} \leq \bar{\mu}_1$. The minimization problem is then given by

$$I_1 = \inf \left\{ -\alpha^{-1} \frac{\mathbb{I}(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2)}{\tilde{\lambda} - \tilde{\mu}_1} + \frac{\mathbb{I}(\bar{\lambda}, \bar{\mu}_1, \bar{\mu}_2)}{\bar{\mu}_1 - \bar{\mu}_2} \right\},$$

where the infimum is taken over variables $\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2, \bar{\lambda}, \bar{\mu}_1$ and $\bar{\mu}_2$ that satisfy the given boundary conditions and α is the (negative) slope of the second part of the path, i.e. $\alpha = (\bar{\mu}_1 - \bar{\mu}_2)/(\bar{\lambda} - \bar{\mu}_1)$ (cf. equation (13)). The solution for this problem can be found in two steps: first minimizing the first term over the tilded variables, which yields $(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2) = (\mu_1, \lambda, \mu_2)$, and then solving the remaining problem, yielding $(\bar{\lambda}, \bar{\mu}_1, \bar{\mu}_2) = (\mu_1, \mu_2, \lambda)$. The total path of this shape will cost us $I_1 = -\log(\rho_2)$ per vertical unit. Paths with other shapes have also been checked, and indeed none of them has lower cost.

Proposition 7. *If $\lambda < \mu_1 < \mu_2$ (Case 1), then the optimal path to overflow of the second buffer has the following shape:* $(0, 0) \to (-\alpha^{-1}B, 0) \to (0, B)$, *where $\alpha = (\bar{\mu}_1 - \bar{\mu}_2)/(\bar{\lambda} - \bar{\mu}_1)$, and the decay rate is ρ_2. The corresponding change of measure is given by*

$$(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2) = (\mu_1, \lambda, \mu_2) \quad \textit{until } X_1(t) = -\alpha^{-1}B, \tag{19}$$

$$(\bar{\lambda}, \bar{\mu}_1, \bar{\mu}_2) = (\mu_1, \mu_2, \lambda) \quad \textit{after}. \tag{20}$$

Case with equal service rates, i.e. $\lambda < \mu_1 = \mu_2$

The special case where the service rates are equal can, in principle, be added to Case 1 or to Case 2. The constant $-\alpha^{-1}$ in equation (19) now equals zero and consecutive utilization of measures (19) and (20) is in fact equivalent to using the new measure in equation (18). Both methods seem to provide the same path $(0, 0) \to (0, B)$, so it might appear at first sight that it does not matter which case we extend. However, in the case of equal service rates, the optimal path is a vertical line upwards *in the interior, without horizontal drift*, which is different from a path on the vertical axis as in Case 2. This is why we prefer Case 1 over Case 2 to include the possibility that $\mu_1 = \mu_2$.

We conclude that the results which we found using our heuristic perfectly coincide with the results of Anantharam et al. [2]. Although a formal proof of the method is still lacking, this suggests that it should also yield good results for the slow-down model, and indeed it does. But let us first see what happens if we use the changes of measure we found in an IS simulation for the tandem model itself.

3.1 Importance Sampling

The principle of Importance Sampling (IS) is that when we simulate our system, we use a new (changed) measure in order to increase the probability of overflow; for the remainder of the busy cycle we will use the old measure. To compensate for the use of the new measure, we need to calculate the *likelihood ratio* $L(X)$ for each random sample path X that is generated in this way. This likelihood ratio of a path X equals the probability that X occurs under the original measure, divided by the probability that X occurs under the new measure. Using this, the overflow probability can be represented as

$$p_B = \mathbb{E}^* L(X)\mathbf{1}(X), \tag{21}$$

where $\mathbb{E}^*$ denotes expectation under the new measure and $\mathbf{1}(X)$ is an indicator function which equals 1 if the rare event of our interest occurs in the sample path X, and 0 otherwise. Thus, we may simulate the system N times under the new measure and then estimate the probability by the sample mean:

$$\hat{p}_B = \frac{1}{N} \sum_{i=1}^{N} L(X_i)\mathbf{1}(X_i). \tag{22}$$

It is obvious that the number of replications to obtain confidence intervals of a given accuracy via direct simulation grows to infinity exponentially fast in B. For an IS estimator as in equation (22), the simulation effort grows subexponentially in B if it is asymptotically optimal. Since the variance is always non-negative, it is clear we have $\log \mathbb{E}^* L^2(X)\mathbf{1}(X) \geq 2 \log \mathbb{E}^* L(X)\mathbf{1}(X)$ for any IS estimator. If for some estimator we have equality as $B \to \infty$, this is a 'good' estimator, and we refer to it as asymptotically efficient or optimal. A formal definition is as follows.

Definition 8. *An IS estimator is referred to as asymptotically optimal if*

$$\lim_{B\to\infty} \frac{\log \mathbb{E}^* L^2(X)\mathbf{1}(X)}{\log p_B} = 2. \tag{23}$$

Corollary 9. *In the tandem case, the IS estimator is asymptotically optimal if*

$$\lim_{B\to\infty} \frac{\log \mathbb{E}^* L^2(X)\mathbf{1}(X)}{B \log \rho_2} = 2. \tag{24}$$

Proof. This is a direct consequence of Theorem 1. □

In the sequel we will use

$$\hat{\psi}_B = \frac{\log \frac{1}{N} \sum_{i=1}^{N} L^2(X_i)\mathbf{1}(X_i)}{B \log \rho_2} \tag{25}$$

as an estimator for the left-hand side of equation (24) to test asymptotic optimality, where N is the number of simulation runs performed under the new measure.

Case 2, i.e. $\lambda < \mu_2 < \mu_1$

We begin our analysis with the simplest problem once more: the tandem Jackson network where the second node is the bottleneck. The optimal path to overflow is known, and under the new measure we will simply interchange λ and μ_2 as follows from Proposition 6.

To construct the probability estimator of a path, we need to know the likelihood of a sample path. This is simply the product of the likelihoods of all individual transitions made during the path until either level B or state $(0, 0)$ is reached, whichever happens first. As an example, let us introduce the likelihood ratio for a transition corresponding to an arrival at the first buffer (i.e. a jump to the right). It is important to note that the likelihood ratios in the interior and on the boundaries may be different. Let us first provide the likelihood ratio for a 'horizontal' jump from some state in the interior. This is given by the ratio of probabilities to make such a jump under the old and new measures, i.e. the ratio of $\lambda/(\lambda + \mu_1 + \mu_2)$ and $\tilde{\lambda}/(\tilde{\lambda} + \tilde{\mu}_1 + \tilde{\mu}_2)$, which gives $L = \lambda/\mu_2$. On the vertical boundary the ratio turns out to be the same, but on the horizontal boundary the likelihood ratio is different, i.e.

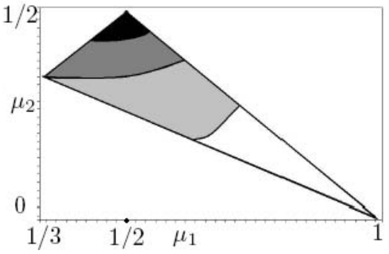

Figure 3. Contour plot of the left-hand side of equation (27) for the two-node tandem Jackson network, Case 2, under the new measure (18). Values less than 0.5 are in white, less than 1 in light gray, less than 1.5 in dark gray and greater than 1.5 in black

$$L' = \frac{\frac{\lambda}{\lambda+\mu_1}}{\frac{\tilde{\lambda}}{\tilde{\lambda}+\tilde{\mu}_1}} = \frac{\lambda}{\mu_2}\frac{\mu_1+\mu_2}{\lambda+\mu_1}.$$

Similarly, we can calculate the likelihood ratios for other types of jumps. Taking these into account, we can find the likelihood ratio of an entire path to overflow as

$$L_2(X) = \left(\frac{\lambda}{\mu_2}\right)^{B-1+R} \left(\frac{\mu_1+\mu_2}{\lambda+\mu_1}\right)^{H}, \tag{26}$$

where R is the number of jobs in the first buffer when the second reaches level B for the first time, and H is the total number of visits to the horizontal axis under the new measure, both belonging to path X.

Now let us see when equation (22) is asymptotically optimal. Corollary 9 and equation (26) together give

$$\mathbb{E}^* \left(\frac{\mu_1+\mu_2}{\lambda+\mu_1}\right)^{2H} = \sum_{i=1}^{\infty} \left(\frac{\mu_1+\mu_2}{\lambda+\mu_1}\right)^{2i} \mathbb{P}^*(H = i) < \infty$$

(note that R can be safely ignored, since $\lambda < \mu_2$).

If H is asymptotically geometric, i.e. if for some constants c and γ we have $\mathbb{P}^*(H = i) \approx c\gamma^i$ as $i \to \infty$, then this holds when γ satisfies

$$\gamma \left(\frac{\mu_1+\mu_2}{\lambda+\mu_1}\right)^2 < 1. \tag{27}$$

Although we have no formal proof, our simulation results confirm that the number of visits to the horizontal axis during a busy cycle indeed has an almost geometrical distribution. In Figure 3, we present a contour plot of the left-hand side of equation (27) as a function of μ_1 and μ_2; note that $\lambda = 1 - \mu_1 - \mu_2$ so that the domain is given by the triangular region $0 < 1 - \mu_1 - \mu_2 < \mu_2 < \mu_1$. The figure illustrates in which parameter region equation (27)

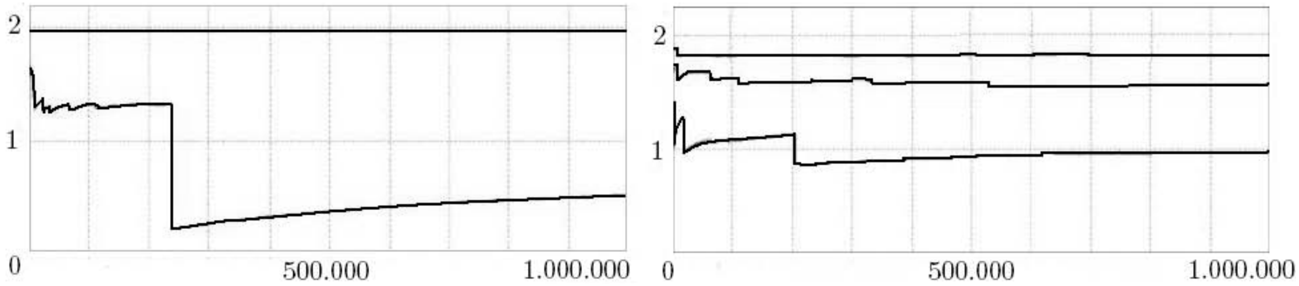

Figure 4. $\hat{\psi}_{50}$ against N. Left (right) panel corresponds to Case 2 (Case 1)

Table 1. Simulation results for the two-node tandem Jackson network, Case 2

$(\lambda, \mu_1, \mu_2) = (0.1, 0.7, 0.2)$			$(\lambda, \mu_1, \mu_2) = (0.3, 0.36, 0.34)$		
B	$\hat{\psi}_B$	p_B	B	$\hat{\psi}_B$	p_B
20	1.93	$1.11 \times 10^{-6} \pm 2.15 \times 10^{-9}$	20	0.67	$6.0 \times 10^{-2} \pm 6.25 \times 10^{-4}$
50	1.97	$1.03 \times 10^{-15} \pm 2.00 \times 10^{-18}$	50	1.3	$1.5 \times 10^{-3} \pm 6.35 \times 10^{-5}$
100	1.99	$9.21 \times 10^{-31} \pm 1.78 \times 10^{-33}$	100	1.6	$2.91 \times 10^{-6} \pm 6.95 \times 10^{-8}$

holds, so that the estimator is asymptotically efficient. Note, however, that we cannot be sure that it is not asymptotically efficient in the remaining part of the domain.

Another way to assess asymptotic efficiency is to directly evaluate equation (24), which we have also done empirically. The lines in the left-hand graph of Figure 4 represent the estimate of equation (24), for $B = 50$, for two different parameter settings as the number of replications grows from zero to 10^6. Here (λ, μ_1, μ_2) have the values $(0.1, 0.7, 0.2)$ and $(0.3, 0.36, 0.34)$ for the upper and lower lines, respectively. This is empirical evidence that for the first parameter setting we have an asymptotically efficient estimator, while for the second setting we do not.

Finally, for the same two parameter settings but for various values of B we present in Table 1 some estimates for the overflow probabilities with 95% confidence intervals and estimates for the left-hand side of equation (24). Simulations for this table (and following tables) are based on $N = 10^6$ independent replications of the busy cycle.

Using the IS method we can decrease simulation time considerably. The time effort per run grows roughly linearly (not exponentially) in B, which implies that the total time effort also grows linearly in B. For $B = 20$ it takes 9 seconds to carry out the $N = 10^6$ replications to estimate the overflow probability with a confidence interval of width 4.3×10^{-9} for the first parameter setting in Table 1. Compare this to straightforward simulations where, for a larger confidence interval of width 4×10^{-8} we need $N \gg 10^6$, taking more than 2 hours. We do not have such a situation in the second column. For $B = 20$ it takes 37 seconds to obtain the estimate for the overflow probability and confidence intervals using IS, and 40 seconds for a similar result using direct simulations (again these values correspond to the first parameter settings). In this case, IS simulations yield somewhat smaller simulation times compared to direct simulations, but the speedup is incomparably smaller than in the case of an asymptotically efficient change of measure.

Remark 10. *When we compare our region of asymptotic efficiency with that of de Boer [5] they seem to coincide, although de Boer considers the probability that the* total *network population, i.e. $X_1(t) + X_2(t)$, reaches some high level B. However, since the optimal paths for both problems coincide for the current Case 2, the similarity need not surprise us.*

Case 1, i.e. $\lambda < \mu_1 < \mu_2$

Let us focus on the case where the first queue is the bottleneck of the system. In Proposition 7 we showed that a good change of measure for this problem is given by Equations (19) and (20).

The likelihood ratio of an arbitrary path to overflow now has a more complicated structure than in Case 2, i.e.

$$L_1(X) = \left(\frac{\lambda}{\mu_2}\right)^{B-1-U} \left(\frac{\lambda}{\mu_1}\right)^{R-1} \left(\frac{\mu_1+\mu_2}{\lambda+\mu_2}\right)^{V_1} \times \left(\frac{\mu_1+\lambda}{\mu_2+\lambda}\right)^{V_2} \left(\frac{\mu_1+\mu_2}{\mu_1+\lambda}\right)^{H_2}, \qquad (28)$$

where V_1 is the number of visits to the vertical axis under measure (19); V_2 and H_2 are the numbers of visits to the vertical and horizontal axes, respectively, under measure (20); U is the number of jobs in the second buffer

Table 2. Simulation results for the two-node tandem Jackson network, Case 1

$(\lambda, \mu_1, \mu_2) = (0.13, 0.17, 0.7)$			$(\lambda, \mu_1, \mu_2) = (0.3, 0.33, 0.37)$		
B	$\hat{\psi}_B$	p_B	B	$\hat{\psi}_B$	p_B
20	1.58	$7.5 \times 10^{-15} \pm 1.2 \times 10^{-15}$	20	0.3	$2.6 \times 10^{-2} \pm 2.39 \times 10^{-3}$
50	1.88	$5.64 \times 10^{-37} \pm 1.21 \times 10^{-37}$	50	1.09	$3.81 \times 10^{-5} \pm 2.8 \times 10^{-5}$
100	1.93	$1.73 \times 10^{-73} \pm 1.73 \times 10^{-74}$	100	1.34	$8.68 \times 10^{-10} \pm 4.05 \times 10^{-10}$

when the first buffer reaches level $\alpha^{-1}B$ for the first time; and R is the number of jobs in the first buffer when the number of jobs in the second buffer reaches level B for the first time. We assume that V_1, V_2 and H_2 are geometrical random variables with parameters γ_1, γ_2 and γ_3 respectively, which is indeed confirmed by simulation experiments. Assuming independence as B grows large, the inequality that should hold for asymptotic efficiency is now given by

$$\gamma_1 \left(\frac{\mu_1+\mu_2}{\lambda+\mu_2}\right)^2 \gamma_2 \left(\frac{\mu_1+\lambda}{\mu_2+\lambda}\right)^2 \times \gamma_3 \left(\frac{\mu_1+\mu_2}{\mu_1+\lambda}\right)^2 < 1. \qquad (29)$$

Unfortunately, simulations show that equation (29) never holds under the change of measure (19) and (20). On the other hand, the right-hand panel of Figure 4 suggests that in Case 1 we may have asymptotical efficiency for some parameters. The variance of the estimator strongly depends on the parameter settings. From top to bottom we have $(\lambda, \mu_1, \mu_2) = (0.13, 0.17, 0.7)$, $(0.25, 0.35, 0.4)$ and $(0.3, 0.33, 0.37)$.

For two of these parameter settings and for various values of B we present in Table 2 some simulation results. It is clear that IS gives a considerable variance reduction and speedup compared to normal simulation, also when the estimator is (arguably) not asymptotically efficient.

Remark 11. *It is possible to consider various changes of measure that will result in the same optimal path. For instance, instead of switching from measure (19) to (20) once, we can also switch back and forth between these measures depending on the current value of $X_2(t)$. Analysis of equation (28) shows that in particular, visits to the horizontal axis during the second part of the cycle are harmful (i.e. they may result in a large value of the likelihood). We tried to exclude these by using the following more complicated change of measure. Begin with measure (19); always switch to measure (20) if $X_1(t) = -\alpha^{-1}B$ and $X_2(t) > 0$; always switch to the natural measure ($\tilde{\lambda} = \lambda$, $\tilde{\mu}_1 = \mu_1$, $\tilde{\mu}_2 = \mu_2$) if $X_1(t) > -\alpha^{-1}B$ and $X_2(t) = 0$; switch back to measure (19) if $X_1(t) \leq -\alpha^{-1}B$ and $X_2(t) = 0$. Empirically, it turns out that this change of measure (and other variants) is not asymptotically efficient either, although the variance of the estimator is a little less.*

4. Slow-down System

In this section we focus on the slow-down system in which the rate of the first server depends on the content of the second buffer. We can identify different cases as in the tandem model, depending on the values of the parameters, and distinguish three cases: (3) $\mu_2 < \nu_1 < \mu_1$; (4) $\nu_1 < \mu_2 < \mu_1$; and (5) $\nu_1 < \mu_1 < \mu_2$. The cases in which $\mu_1 = \mu_2$ or $\nu_1 = \mu_2$ can be dealt with in the same manner as for the standard tandem Jackson network. Cases 3 and 4 are comparable to Case 2 in the tandem model, where the second server is the bottleneck. The difference is in the situation when the number of jobs in the second buffer exceeds the slow-down threshold θB. In Case 3 the second server remains the bottleneck, i.e. $\nu_1 > \mu_2$, while in Case 4 the first server becomes the bottleneck, i.e. $\nu_1 < \mu_2$. When the first server is the bottleneck there is only one possibility, in which the first server remains the bottleneck.

As mentioned earlier, we cannot use a reversibility argument as in Anantharam et al. [2] in the analysis of the slow-down system. However, we can employ our cost function approach, based on Theorem 3.

Case 3, i.e. $\mu_2 < \nu_1 < \mu_1$

Let us start from the situation in which the second server remains the bottleneck, analyzing the path that follows the vertical axis as in Case 2. This path now consists of two parts: below the slow-down threshold and above it. Using the same arguments as in equation (17) we need to find

$$I_3 = \inf \left\{ \theta \frac{\mathbb{I}(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2)}{\tilde{\lambda} - \tilde{\mu}_2} + (1-\theta) \frac{\mathbb{I}(\bar{\lambda}, \bar{\nu}_1, \bar{\mu}_2)}{\bar{\lambda} - \bar{\mu}_2} \right\}, \qquad (30)$$

where the infimum is taken over variables $\tilde{\lambda}$, $\tilde{\mu}_1$, $\tilde{\mu}_2$, $\bar{\lambda}$, $\bar{\mu}_1$ and $\bar{\mu}_2$, such that we have northwest drift below the threshold (i.e. $\tilde{\mu}_1 > \tilde{\lambda}$ and $\tilde{\lambda} > \tilde{\mu}_2$) and above it (i.e. $\bar{\nu}_1 > \bar{\lambda}$ and $\bar{\lambda} > \bar{\mu}_2$). This can easily be solved by splitting it into two separate minimization problems that are completely analogous to equation (17), so the outcome will be to interchange the values of λ and μ_2. We have checked all other possible shapes of the path to overflow (see Appendix A2) and conclude that indeed $I_3 = -\log \rho_2$ is the minimal cost for unit movement in the vertical direction.

Proposition 12. *If $\mu_2 < \nu_1 < \mu_1$ (Case 3) then the optimal path to overflow of the second buffer has the following shape:* $(0,0) \to (0, \theta B) \to (0, B)$. *The corresponding change of measure is given by*

$$\begin{aligned}(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2) &= (\mu_2, \mu_1, \lambda) \quad \text{and}\\ (\bar{\lambda}, \bar{\nu}_1, \bar{\mu}_2) &= (\mu_2, \nu_1, \lambda), \qquad (31)\end{aligned}$$

and the decay rate is ρ_2.

Case 4, i.e. $\nu_1 < \mu_2 < \mu_1$

Now let us concentrate on the network where the bottleneck shifts from the second server to the first server when the slow-down threshold is reached. We focus on a path that follows the vertical axis until the slow-down threshold, after which the process moves with northeast drift. The following minimization problem corresponds to this type of path:

$$I_4 = \inf \left\{ \theta \frac{\mathbb{I}(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2)}{\tilde{\lambda} - \tilde{\mu}_2} + (1-\theta) \frac{\mathbb{I}(\bar{\lambda}, \bar{\nu}_1, \bar{\mu}_2)}{\bar{\nu}_1 - \bar{\mu}_2} \right\}, \quad (32)$$

where we take the infimum over variables $\tilde{\lambda}$, $\tilde{\mu}_1$, $\tilde{\mu}_2$, $\bar{\lambda}$, $\bar{\mu}_1$ and $\bar{\mu}_2$, such that $\tilde{\mu}_1 > \tilde{\lambda}$, $\tilde{\lambda} > \tilde{\mu}_2$, $\bar{\nu}_1 > \bar{\mu}_2$ and $\bar{\lambda} \geq \bar{\nu}_1$. Again we can decompose the optimization problem into two parts. The first part of equation (32) has the same solution as the first part of equation (30), and hence as equation (17). The second part of problem (32) has a more complicated solution, that in fact corresponds to the boundary case in which the path has no horizontal drift, i.e. $\bar{\lambda} = \bar{\nu}_1$. It is given by

$$\bar{\lambda} = \bar{\nu}_1 = \sqrt{\frac{\lambda \nu_1}{z}}, \quad \bar{\mu}_2 = \mu_2 z, \qquad (33)$$

where z is the unique solution in (0, 1) of the equation

$$\lambda + \nu_1 + \mu_2(1-z) = 2\sqrt{\frac{\lambda \nu_1}{z}}. \qquad (34)$$

As an aside, we note that this is the same equation as the equation [11] of Kroese et al. 30. Indeed, the decay rate behavior deduced in that article can also be obtained using our heuristic. Since all other path types turn out to have higher cost (see Appendix A2), this is the optimal path with corresponding cost $I_4 = -\log\left(\rho_2^{\theta} z^{1-\theta}\right)$ per vertical unit.

Proposition 13. *If $\nu_1 < \mu_2 < \mu_1$ (Case 4), then the optimal path to overflow of the second buffer has the following shape:* $(0,0) \to (0, \theta B) \to (0^+, B)$. *The corresponding change of measure is given by equations (18) and (33), and the decay rate is $\rho_2^{\theta} z^{1-\theta}$.*

The optimal path in this case looks very similar to the optimal path in Case 3. Indeed, they coincide below θB, where the drift is to the north and east. Above θB the path is also vertical, but there is an essential difference since there is no horizontal drift here. The notation 0^+ in Proposition 13 is meant to express this difference.

Case 5, i.e. $\nu_1 < \mu_1 < \mu_2$

This case has the least interest from a practical point of view, but we include it for the sake of completeness. In this section we will provide the shape of the most likely path to overflow in the second buffer.

Proposition 14. *If $\nu_1 < \mu_1 < \mu_2$ (Case 5), then the optimal path to overflow of the second buffer has the following shape:* $(0,0) \to (\beta_1 B, 0) \to (\beta_2 B, \theta B) \to (0, B)$, *where $\beta_2 = (1-\theta)(\hat{\nu}_1 - \hat{\lambda})/(\hat{\nu}_1 - \hat{\mu}_2)$ and $\beta_1 = \beta_2 + \theta(\bar{\mu}_1 - \bar{\lambda})/(\bar{\mu}_1 - \bar{\mu}_2)$. The corresponding change of measure is given by*

$$\begin{aligned}(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2) &= (\mu_1, \lambda, \mu_2) \textit{ until } X_1 = \beta_1 B,\\ (\bar{\lambda}, \bar{\mu}_1, \bar{\mu}_2) &= (\mu_1, \mu_2, \lambda) \textit{ until } X_2 = \theta B,\\ (\hat{\lambda}, \hat{\nu}_1, \hat{\mu}_2) &= (\mu_1, \frac{\lambda \nu_1}{x \mu_1}, x\mu_2) \textit{ after,}\end{aligned}$$

where x is the unique solution of $\mu_1 \mu_2 x^2 + \mu_1(\mu_1 - \lambda - \nu_1 - \mu_2)x + \lambda \nu_1 = 0$ which guarantees $\hat{\lambda} < \hat{\nu}_1$ and $\hat{\nu}_1 > \hat{\mu}_2$, and the constants β_1 and β_2 are given by $\beta_2 = (1-\theta)(\lambda \nu_1 - \mu_1^2 x)/(\lambda \nu_1 - \mu_1 \mu_2 x^2)$ and $\beta_1 = \beta_2 + \theta(\mu_2 - \mu_1)/(\mu_2 - \lambda)$. The decay rate is $\rho_2^{\theta} x^{1-\theta}$.

We omit all calculations due to the similarity to Case 1.

4.1 Importance Sampling

In this section we present our results for the IS simulations for the system with slow-down threshold. The estimator for the overflow probability has the same form as equation (22), and again we are interested in asymptotic efficiency. In particular, we will compare the asymptotically efficient parameter region with that of the two-node Jackson network in Case 2. Beforehand, it is clear that the first should always be contained in the latter. Let us first focus on the case where the second buffer remains the bottleneck, for which we have a much stronger result.

Case 3, i.e. $\mu_2 < \nu_1 < \mu_1$

In this case we use the change of measure given by equation (31). The rate of the first server is always left unchanged, being equal to either μ_1 or ν_1 depending on the current state of the second buffer.

Proposition 15. *Assume that* $\lambda < \mu_2 < \nu_1 < \mu_1$ *(Case 3), then the overflow probability estimators under measure (18) for the tandem Jackson network and equation (31) for the slow-down network are asymptotically efficient in the same parameter regions.*

Proof. The likelihood ratio of an arbitrary path X that reaches level B is very similar to equation (26), namely

$$L_3(X) = \left(\frac{\lambda}{\mu_2}\right)^{B-1+R'} \left(\frac{\mu_1+\mu_2}{\lambda+\mu_1}\right)^{H'},$$

where R' is the number of jobs in the first buffer when the second one reaches level B for the first time and H' is the total number of visits to the horizontal axis under the new measure. It is enough to show that the second moments of L_2 and L_3 are asymptotically identical to prove the proposition. It is clear that the distribution of R' is not important since $\lambda/\mu_2 < 1$. On the other hand, the distribution of H' does play a role and in fact determines whether or not the estimator is asymptotically efficient for a certain parameter setting. Fortunately, we have that H' converges in distribution to H as $\theta B \to \infty$, so that a comparison with equation (26) gives the statement of the proposition. □

As an illustration, we simulate the system for two different parameter settings. In the first we take $(\lambda, \mu_1, \mu_2) = (0.1, 0.7, 0.2)$ and $\nu_1 = 0.3$, and for the second we take $(\lambda, \mu_1, \mu_2) = (0.3, 0.36, 0.34)$ and $\nu_1 = 0.35$. In these (and all further) simulations we will use $\theta = 0.8$ to define the slow-down threshold. Note the correspondence to the examples in Section 3.1, and that in both cases we have $\nu_1 > \mu_2$. Indeed, in the first case the estimator is asymptotically optimal and in the second case it is not. This can be illustrated by a diagram similar to Figure 4, and also by the values of the estimator of the left-hand side of equation (23) in Table 3, which is now given by

$$\hat{\psi}'_B = \frac{\log \frac{1}{N}\sum_{i=1}^{N} L^2(X_i)\mathbf{1}(X_i)}{\log \hat{p}_B}. \qquad (35)$$

The reason for using equation (35) instead of equation (25) is that we do not have an analogue to Theorem 1 for the slow-down case and hence no analogue to Corollary 9. The speedups obtained in this table are comparable to those in Table 1.

Case 4, i.e. $\nu_1 < \mu_2 < \mu_1$

In this case we use the change of measure given in Proposition 13. Specifically, we always use equation (18) when the number of jobs in the second buffer is below θB and equation (33) otherwise, until level B is reached for the first time. It is now more difficult to obtain an analogue to Proposition 15, since the process may have some cycles around level θB that influence the total likelihood of the path. Notice that this was not true in Case 3, due to the fact that the change of measure used below and above the threshold was essentially the same. Consider then a typical path to overflow of the form

$$\begin{aligned}(1,0) &\to (X_1(t_1), \theta B) \to (X_1(t_2), \theta B)\\ &\to \ldots \to (X_1(t_n), \theta B) \to (R, B),\end{aligned}$$

where the $t_i, i = 1, 2, \ldots, n$ are the $n \geq 1$ subsequent time epochs at which the process visits level θB before moving to level B. It can be shown that the total likelihood of such a path is given by

$$\begin{aligned} L_4(X) &= \left(\frac{\lambda}{\mu_2}\right)^{\theta B} z^{(1-\theta)B} \left(\frac{\mu_1+\mu_2}{\lambda+\mu_1}\right)^{H}\\ &\times \left(\frac{\sqrt{\frac{\lambda\nu_1}{z}}+\mu_2 z}{\lambda+\mu_1}\right)^{V} \left(\frac{\lambda}{\mu_2 z}\right)^{C} \left(\frac{\lambda}{\mu_2}\right)^{D}\\ &\times \left(\sqrt{\frac{\lambda z}{\nu_1}}\right)^{U}. \end{aligned}$$

Here H is the number of visits to the horizontal axis and V is the number of visits to the vertical axis above the threshold. Furthermore, C is the number of subpaths $(X_1(t_i), \theta B) \to (X_1(t_{i+1}), \theta B)$ below θB (starting with a downward jump), D is the 'total horizontal distance covered during these subpaths', i.e. $D = \sum(X_1(t_{i+1}) - X_1(t_i))$ where the sum is taken over subpaths that start with a jump downward, and similarly U is the total horizontal distance covered during subpaths above θB. Since it is difficult to see how the likelihood behaves (e.g. the random variables D and U may take positive or negative values), we content ourselves with some simulation results for the same scenarios as in Cases 2 and 3, taking $\nu_1 < \mu_2$. These can be found in Table 4.

Again it seems clear that the change of measure (33) is asymptotically efficient for the first parameter setting, but not for the second in which the loads of both queues are close to 1. Based on these and other simulation results we found that in the current Case 4, the region of asymptotical efficiency is somewhat smaller than we found in Case 2.

Table 3. Simulation results for the slow-down system, Case 3

(0.1, 0.7, 0.2) → (0.1, 0.3, 0.2)			(0.3, 0.36, 0.34) → (0.3, 0.35, 0.34)		
B	$\hat{\psi}'_B$	p_B	B	$\hat{\psi}'_B$	p_B
20	1.95	$7.94 \times 10^{-7} \pm 1.89 \times 10^{-9}$	20	0.7	$5.8 \times 10^{-2} \pm 4.91 \times 10^{-4}$
50	1.98	$6.5 \times 10^{-16} \pm 1.68 \times 10^{-18}$	50	1.37	$1.46 \times 10^{-3} \pm 3.97 \times 10^{-5}$
100	1.99	$5.59 \times 10^{-31} \pm 1.48 \times 10^{-33}$	100	1.66	$2.64 \times 10^{-6} \pm 9.51 \times 10^{-8}$

Table 4. Simulation results for the slow-down system, Case 4

(0.1, 0.7, 0.2) → (0.1, 0.15, 0.2)			(0.3, 0.36, 0.34) → (0.3, 0.32, 0.34)		
B	$\hat{\psi}'_B$	p_B	B	$\hat{\psi}'_B$	p_B
20	1.92	$3.79 \times 10^{-7} \pm 1.09 \times 10^{-9}$	20	0.45	$5.6 \times 10^{-2} \pm 1.01 \times 10^{-4}$
50	1.95	$1.26 \times 10^{-16} \pm 5.08 \times 10^{-19}$	50	1.34	$1.17 \times 10^{-4} \pm 2.85 \times 10^{-5}$
100	1.98	$3.54 \times 10^{-32} \pm 1.89 \times 10^{-34}$	100	1.45	$1.69 \times 10^{-6} \pm 1.23 \times 10^{-7}$

Case 5, i.e. $\nu_1 < \mu_1 < \mu_2$

In this section we briefly provide simulation results for Case 5, where the first server is always the bottleneck. The new measure is given by Proposition 14. See Table 5 for results, in which the left part corresponds to the left part of Table 2; note that the overflow probabilities are much smaller, due to the slow-down property of the system.

5. Future Work

We conclude by mentioning a number of potential areas of future research. From a mathematical point of view, the most substantial gap lies in Conjecture 2; a rigorous proof would provide more solid support for the heuristic motivation of our change of measure. We also hope to extend it, as well as Theorem 3, to a larger class of systems. We believe it should be possible to show that, for any two-dimensional Markov chain for which the state space can be written as a finite union of convex disjoint sets on which the transition parameters are constant, the typical path leading to the rare event consists of a concatenation of subpaths on the various subsets that are straight lines.

For the standard tandem model, we found an expression for the logarithmic decay rate in Theorem 1; this value could be used when checking asymptotic optimality. For the tandem model with server slow-down, however, we lack knowledge of such logarithmic asymptotics. A goal would therefore be to prove the analogue of Theorem 1 for the model with server slow-down. For this model, we are also interested in the estimation of the overflow probability and a deeper understanding of the nature of the behavior of estimator (22).

Finally, for both models we wish to construct *state-dependent* IS schemes that are asymptotically efficient for all parameter settings, as in e.g. Dupuis et al. [12].

Appendix

We show how the optimal path to overflow in the second buffer can be found for the standard tandem model and the slow-down model. We begin with the standard model.

A.1 Cost and Shape of the Optimal Path for the Standard Tandem Model

We only need to consider four possible types of paths due to Theorem 3, illustrated in Figure 5. Note, however, that for paths of type (4) which follow the horizontal axis and the interior but not the vertical axis, the slope in the interior need not be positive. To obtain the optimal path we should calculate the minimal cost for each of these types of path, and then take the minimum over these four outcomes. Note that the answer will depend on the case we consider. As an example, we will consider the minimal cost of paths of type (4) for Case 2, i.e. the case in which $\lambda < \mu_2 < \mu_1$.

A.1.1 Paths of Type (4)

It is clear that the cost of such paths consists of two parts: the cost of the subpath on the horizontal boundary and the cost of the subpath in the interior. Formally, for the cost of the entire path we have

$$\inf\left\{\omega\frac{\mathbb{I}(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2)}{\tilde{\lambda} - \tilde{\mu}_1} + B\frac{\mathbb{I}(\bar{\lambda}, \bar{\mu}_1, \bar{\mu}_2)}{\bar{\mu}_1 - \bar{\mu}_2}\right\}, \qquad (36)$$

where ω is the length of the horizontal part. The infimum is taken over variables $\tilde{\lambda}$, $\tilde{\mu}_1$, $\tilde{\mu}_2$, $\bar{\lambda}$, $\bar{\mu}_1$ and $\bar{\mu}_2$ which satisfy $\tilde{\lambda} > \tilde{\mu}_1$, $\tilde{\mu}_1 < \tilde{\mu}_2$ and $\bar{\mu}_1 > \bar{\mu}_2$. Below we will

Table 5. Simulation results for the slow-down system, Case 5

$(0.13, 0.17, 0.7) \to (0.13, 0.14, 0.7)$			$(0.25, 0.35, 0.4) \to (0.25, 0.28, 0.4)$		
B	$\hat{\psi}'_B$	p_B	B	$\hat{\psi}'_B$	p_B
20	1.69	$4.44 \times 10^{-15} \pm 1.43 \times 10^{-15}$	20	0.93	$1.28 \times 10^{-4} \pm 2.79 \times 10^{-5}$
50	1.86	$1.31 \times 10^{-37} \pm 8.23 \times 10^{-38}$	50	1.69	$1.07 \times 10^{-12} \pm 1.15 \times 10^{-13}$
100	1.93	$3.21 \times 10^{-75} \pm 5.02 \times 10^{-76}$	100	1.83	$5.34 \times 10^{-25} \pm 7.61 \times 10^{-26}$

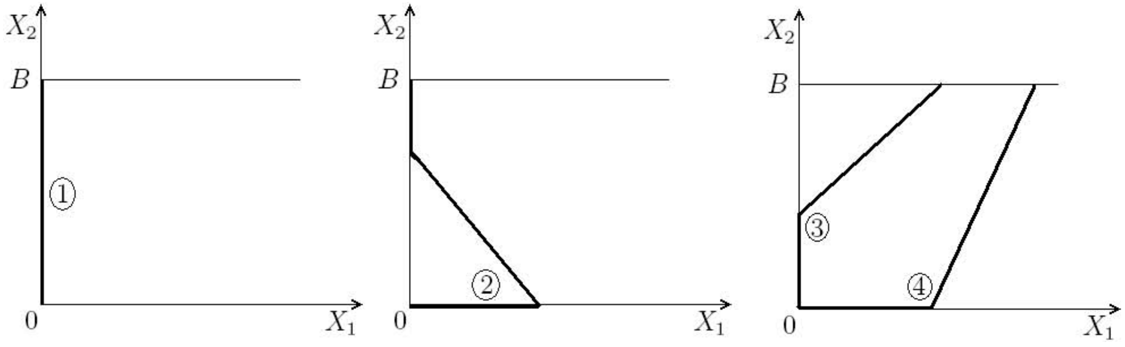

Figure 5. General forms of paths to overflow in second buffer

treat separately the cases in which the path in the interior has northeast drift, i.e. $\bar{\lambda} \geq \bar{\mu}_1$, or northwest drift, i.e. $\bar{\lambda} < \bar{\mu}_1$.

In the first case ($\bar{\lambda} \geq \bar{\mu}_1 > \bar{\mu}_2$) the horizontal subpath does not help us to reach high values in the second buffer, but it increases the cost of the overall path. The first step to optimize equation (36) is therefore to set $\omega = 0$. Now let us optimize the path in the interior under the condition $\bar{\lambda} \geq \bar{\mu}_1 > \bar{\mu}_2$; formally, we need to find a set of parameters which minimizes

$$\frac{\mathbb{I}(\bar{\lambda}, \bar{\mu}_1, \bar{\mu}_2)}{\bar{\mu}_1 - \bar{\mu}_2}.$$

Calculating the partial derivatives with respect to $\bar{\lambda}, \bar{\mu}_1$ and $\bar{\mu}_2$, setting them equal to zero and solving the resulting system we find the solutions:

$$(\bar{\lambda}, \bar{\mu}_1, \bar{\mu}_2) \quad = \quad (\lambda, \mu_1, \mu_2) \quad \text{and}$$

$$(\bar{\lambda}, \bar{\mu}_1, \bar{\mu}_2) \quad = \quad (\lambda, \mu_2, \mu_1).$$

Neither of these parameter settings satisfy our condition $\bar{\lambda} \geq \bar{\mu}_1 > \bar{\mu}_2$. The minimum is therefore attained at the boundary where $\bar{\lambda} = \bar{\mu}_1$, which corresponds to a vertical path in the interior. Minimizing over the two remaining variables, we find the minimizers of equation (36) to be the same as in Case 4 for the slow-down model (see Section 4), namely $\bar{\lambda} = \bar{\mu}_1 = \sqrt{\lambda\mu_1/z}$ and $\bar{\mu}_2 = \mu_2 z$ where z is the unique solution in $(0, 1)$ of equation (34). The corresponding minimal cost is given by $-\log(z)$. We emphasize that the vertical path we found lies in the interior along the vertical boundary, which is different from the (optimal) path of type (1) in which there is a horizontal drift towards the vertical axis.

Let us now consider the other subtype of paths of type (4), for which we have $\bar{\lambda} < \bar{\mu}_1$ and $\bar{\mu}_2 < \bar{\mu}_1$. This time we should set $\omega = -\alpha^{-1}B$, where α is the slope of the second subpath (see equation (13), with tildes replaced by bars), for if we choose $\omega < -\alpha^{-1}B$ we will have a path of type (2) instead of (4). On the other hand, if we choose $\omega > -\alpha^{-1}B$, say $\omega = -\alpha^{-1}B + \delta$ with $\delta > 0$, then the cost of the path $(0, 0) \to (-\alpha^{-1}B + \delta, 0) \to (\delta, B)$ is equal to the cost of the path $(0, 0) \to (-\alpha^{-1}B, 0) \to (0, B)$ plus the cost of the subpath $(0, 0) \to (\delta, 0)$. In other words, it is optimal to set $\delta = 0$. The minimization of equation (36) is now similar to that of Section 3 (Case 1), the only difference being $\mu_1 < \mu_2$. As a result, the infimum is not attained at $(\bar{\lambda},\ \bar{\mu}_1,\ \bar{\mu}_2) = (\mu_1,\ \mu_2,\ \lambda)$, since then $\bar{\lambda} < \bar{\mu}_1$ is not satisfied, but rather at the boundary where $\bar{\lambda} = \bar{\mu}_1$. In other words, we find the same vertical path as above, with $\omega = 0$. This is therefore the optimum over all paths of type (4).

A.1.2 Paths of Other Types

In Section 3 (Case 2) it was found that the minimal cost for paths of type (1) is $-\log(\rho_2)$. Since $z < \rho_2$, this means that it is cheaper to follow the vertical axis, than to follow a vertical path through the interior.

The cost of a path of type (2) consists of two parts: the cost of the subpath following the vertical axis and the cost of the remainder of the path. The optimal cost for the

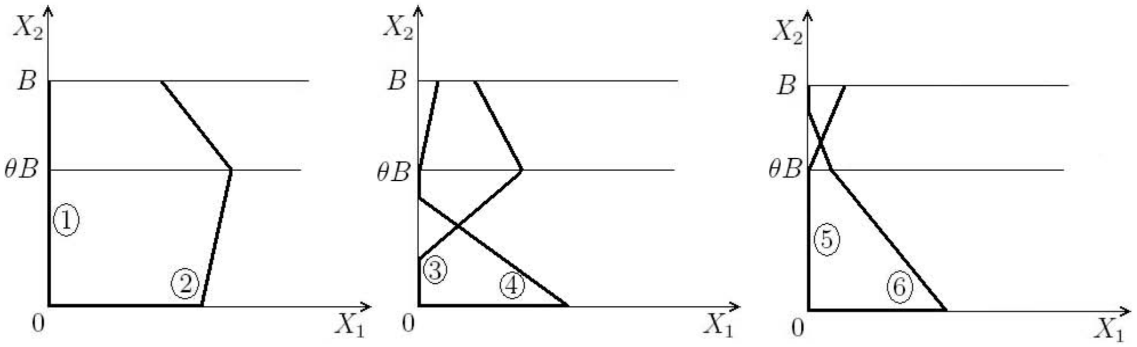

Figure 6. General forms of paths to overflow in second buffer

vertical boundary part is $-\log(\rho_2)$; for the remainder of the path the optimal shape is a vertical line in the interior (see the case for path (4) where $\bar{\lambda} < \bar{\mu}_1$ and $\bar{\mu}_2 < \bar{\mu}_1$). However, since the cost of this is higher than following the vertical axis, we see that the optimal path of type (2) is actually the limiting case where the starting point and end point of the interior subpath are both equal to the origin. In other words, the optimal path of type (2) coincides with the path of type (1).

For paths of type (3) we can use similar arguments to show that here the optimal path is also the same as the path of type (1).

A.2 Cost and Shape of the Optimal Path for the Slow-down Model

We now show how the most likely path to overflow is found in the second buffer for the slow-down model. Theorem 3 provides all possible paths types, which are illustrated in Figure 6. Note that for paths of type (2) the slopes in the interior may have any sign, which also holds for the upper part of path type (3). Paths of type (2), (3) and (6) include paths in which the first part (following the horizontal or vertical boundary) is absent.

In the following we restrict ourselves to the most interesting case, Case 4, where the bottleneck shifts from the second to the first queue after hitting the slow-down threshold, i.e. $\nu_1 < \mu_2 < \mu_1$. For Cases 3 and 5, similar calculations hold. In Proposition 13 we claim that the optimal path is the special case of path type (5), in which the second part of the path is a vertical line in the interior. In the remainder of the appendix when we mention the path of type (5), we are referring to this special case and we will compare other path types to this.

A.2.1 Paths of Type (2)

We divide this type in three subtypes: paths without a first horizontal part; paths with a horizontal part and northwest drift in the interior below the slow-down threshold; and paths with horizontal part and northeast drift in the interior below the slow-down threshold.

In the latter two cases we need not consider paths with northeast drift above θB. For paths of the third subtype it is not optimal to follow the horizontal boundary in the eastern direction, so the optimal version of such a path belongs to the first subtype. It follows from Section 3 that any path of the second subtype with northeast drift above θB is also not optimal.

We first study paths without the first horizontal part. The optimal cost of such a path is

$$\inf\left\{\theta B\frac{\mathbb{I}(\bar{\lambda},\bar{\mu}_1,\bar{\mu}_2)}{\bar{\mu}_1-\bar{\mu}_2}+(1-\theta)B\frac{\mathbb{I}(\hat{\lambda},\hat{\nu}_1,\hat{\mu}_2)}{\hat{\nu}_1-\hat{\mu}_2}\right\}, \quad (37)$$

where the infimum is taken over variables $\bar{\lambda}$, $\bar{\mu}_1$, $\bar{\mu}_2$, $\hat{\lambda}$, $\hat{\mu}_1$ and $\hat{\mu}_2$ that satisfy $\bar{\lambda} \geq \bar{\mu}_1$, $\bar{\mu}_1 > \bar{\mu}_2$ and $\hat{\nu}_1 > \hat{\mu}_2$ as well as the additional condition

$$\theta\frac{\bar{\lambda}-\bar{\mu}_1}{\bar{\mu}_1-\bar{\mu}_2} \geq (1-\theta)\frac{\hat{\nu}_1-\hat{\lambda}}{\hat{\nu}_1-\hat{\mu}_2},$$

which ensures that the path will not hit the vertical boundary below level B if $\hat{\nu}_1 > \hat{\lambda}$.

Paths of the second subtype consist of three parts: one part following the horizontal boundary and two parts traversing the interior with northwest drift (below and above θB). The optimal cost of such a path is

$$\inf\left\{\omega_1\frac{\mathbb{I}(\tilde{\lambda},\tilde{\mu}_1,\tilde{\mu}_2)}{\tilde{\lambda}-\tilde{\mu}_1}+\theta B\frac{\mathbb{I}(\bar{\lambda},\bar{\mu}_1,\bar{\mu}_2)}{\bar{\mu}_1-\bar{\mu}_2}\right.$$
$$\left.+\quad(1-\theta)B\frac{\mathbb{I}(\hat{\lambda},\hat{\nu}_1,\hat{\mu}_2)}{\hat{\nu}_1-\hat{\mu}_2}\right\}, \quad (38)$$

where

$$\omega_1=(1-\theta)B\frac{\hat{\nu}_1-\hat{\lambda}}{\hat{\nu}_1-\hat{\mu}_2}+\theta\frac{\bar{\mu}_1-\bar{\lambda}}{\bar{\mu}_1-\bar{\mu}_2},$$

which guarantees that $(0, B)$ is the end point of the path (note that any path which hits level B at some point (x, B) with $x > 0$ is not optimal, using the same arguments as for paths of type (4) in the standard tandem network). The infimum is taken over all variables $\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2, \bar{\lambda}, \bar{\mu}_1, \bar{\mu}_2, \hat{\lambda}, \hat{\nu}_1$ and $\hat{\mu}_2$ which satisfy $\tilde{\lambda} > \tilde{\mu}_1, \tilde{\mu}_1 < \tilde{\mu}_2, \bar{\lambda} \leq \bar{\mu}_1, \bar{\mu}_1 > \bar{\mu}_2, \hat{\lambda} \leq \hat{\nu}_1$ and $\hat{\nu}_1 > \hat{\mu}_2$.

It remains to consider paths of the third subtype. These first follow the horizontal boundary, traverse the interior below θB with northeast drift and finally traverse the interior above θB which has northwest drift. The minimal cost of such a path is

$$\inf \left\{ \omega_2 \frac{\mathbb{I}(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2)}{\tilde{\lambda} - \tilde{\mu}_1} + \theta B \frac{\mathbb{I}(\bar{\lambda}, \bar{\mu}_1, \bar{\mu}_2)}{\bar{\lambda} - \bar{\mu}_1} \right.$$

$$\left. + \quad (1-\theta) B \frac{\mathbb{I}(\hat{\lambda}, \hat{\nu}_1, \hat{\mu}_2)}{\hat{\nu}_1 - \hat{\mu}_2} \right\}, \tag{39}$$

where the choice of

$$\omega_2 = (1-\theta) B \frac{\hat{\nu}_1 - \hat{\lambda}}{\hat{\nu}_1 - \hat{\mu}_2} - \theta B \frac{\bar{\lambda} - \bar{\mu}_1}{\bar{\mu}_1 - \bar{\mu}_2}$$

guarantees that $(0, B)$ is the end point of the path. The infimum is taken over all variables $\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2, \bar{\lambda}, \bar{\mu}_1, \bar{\mu}_2, \hat{\lambda}, \hat{\nu}_1$ and $\hat{\mu}_2$ that satisfy $\tilde{\lambda} > \tilde{\mu}_1, \tilde{\mu}_1 < \tilde{\mu}_2, \bar{\lambda} \geq \bar{\mu}_1, \bar{\mu}_1 > \bar{\mu}_2, \hat{\lambda} \leq \hat{\nu}_1$ and $\hat{\nu}_1 > \hat{\mu}_2$.

Optimization of the costs in equations (37)–(39) provides the optimal shape for paths of type (2), which turns out to be a vertical line in the interior. More precisely, the optimal values of the parameters with bars and hats are given by equations as in equations (33)–(34); as a result, the horizontal part always vanishes (i.e. $\omega_1 = \omega_2 = 0$). It is therefore clear that the cost of any path of type (2) is higher then the optimal cost of path (5).

A.2.2 Paths of Type (3)

Let us continue with paths of type (3) which consist of three parts in general: a first part following the vertical boundary; a second part in the interior below θB with northeast drift; and a third part in the interior above θB with northwest drift (northeast drift here is excluded since the optimal path can only end in $(0, B)$, as before). The optimal cost of such a path is

$$\inf \left\{ \omega_3 \frac{\mathbb{I}(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2)}{\tilde{\lambda} - \tilde{\mu}_2} + (\theta B - \omega_3) \frac{\mathbb{I}(\bar{\lambda}, \bar{\mu}_1, \bar{\mu}_2)}{\bar{\lambda} - \bar{\mu}_1} \right.$$

$$\left. + \quad (1-\theta) B \frac{\mathbb{I}(\hat{\lambda}, \hat{\nu}_1, \hat{\mu}_2)}{\hat{\nu}_1 - \hat{\mu}_2} \right\},$$

where

$$\omega_3 = \theta B - (1-\theta) B \frac{\hat{\nu}_1 - \hat{\lambda}}{\hat{\nu}_1 - \hat{\mu}_2} \frac{\bar{\mu}_1 - \bar{\mu}_2}{\bar{\lambda} - \bar{\mu}_1}$$

is the length of the first part, and the infimum is taken over variables $\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2, \bar{\lambda}, \bar{\mu}_1, \bar{\mu}_2, \hat{\lambda}, \hat{\nu}_1$ and $\hat{\mu}_2$ which satisfy $\tilde{\lambda} > \tilde{\mu}_2, \tilde{\lambda} < \tilde{\mu}_1, \bar{\lambda} \geq \bar{\mu}_1, \bar{\mu}_1 > \bar{\mu}_2$ and $\hat{\nu}_1 > \hat{\mu}_2$. Indeed, a path which hits the vertical axis below level B and follows it afterwards is not optimal (Lemma 5); any path hitting level B at any point with positive first coordinate is not optimal either (see explanations for paths of type (2)). After optimization we obtain that $(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2) = (\mu_2, \mu_1, \lambda)$ and $\hat{\lambda} = \hat{\nu}_1$. As a result, $\omega_3 = \theta B$ and we find the optimal path of type (5) is the optimal outcome for paths of type (3).

A.2.3 Paths of Other Types

The calculations provided in Section 4 and our knowledge about the shape of the most probable path for the standard Jackson network ensures that paths of type (1) have higher cost than the optimal path of type (5). These paths coincide below θB, but (1) is more expensive above θB. After optimization, path (4) as well as path (3) converge to path (5). Finally, we point out that the cost of paths of type (6) is higher than the cost of type (2), because following the vertical axis above θB is more expensive than traversing the interior along it, following a vertical path. This completes our discussion about the optimal shape for Case 3 for the slow-down model.

6. Acknowledgements

This research was partly funded by the Dutch BSIK/BRICKS project.

7. References

[1] van Foreest, N. D., M. R. H. Mandjes, J. C. W. van Ommeren and W. R. W. Scheinhardt. 2005. A tandem queue with server slow-down and blocking. *Stochastic Models*, 21, 695–724.

[2] 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.

[3] 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.

[4] Glasserman, P. and S.-G. Kou. 1995. Analysis of an importance sampling estimator for tandem queues. *ACM Transactions on Modeling and Computer Simulation*, 5, 22–42.

[5] de Boer, P.-T. 2006. Analysis of state-independent importance-sampling measures for the two-node tandem queue. *ACM Transactions on Modeling and Computer Simulation*, 16, 225–250.

[6] Zaburnenko, T. S. and V. F. Nicola. 2005. Efficient heuristics for simulating population overflow in tandem networks. In *Proceedings of the Fifth Workshop on Simulation*, pp. 755–764.

[7] Sandmann, W. 2004. Fast simulation of excessive population size in tandem Jackson networks. In *Proceedings of the 12th Annual IEEE International Symposium MASCOTS'04*, pp. 347–354.

[8] Kroese, D. P. and V. F. Nicola. 2002. Efficient simulation of a tandem Jackson network. *ACM Transactions on Modeling and Computer Simulation*, 12, 119–141.

[9] Anantharam, V. 1989. The optimal buffer allocation problem. *IEEE Transactions on Information Theory*, 35, 721–725.

[10] Shwartz, A. and A. Weiss. 1995. *Large deviations for performance analysis. Queues, communications and computing*. London: Chapman & Hall.

[11] Kroese, D. P., W. R. W. Scheinhardt and P. G. Taylor. 2004. Spectral properties of the tandem Jackson network, seen as quasi-birth-and-death process. *The Annals of Applied Probability*, 14, 2057–2089.

[12] Dupuis, P., A. D. Sezer and H. Wang. 2007. Dynamic Importance Sampling for Queueing Networks. *The Annals of Applied Probability*, 17, 1306–1346.

***Denis Miretskiy** is a graduate student at the University of Twente, Enschede, The Netherlands. He received his M.Sc. degree in 2005 from Utrecht University, The Netherlands, and his specialist degree in 2004 from Volgograd State University, Russia.*

***Werner Scheinhardt** is an assistant professor at the University of Twente, Enschede, The Netherlands, where he received both his doctoral degree (1998) and M.Sc. degree (1994). He is also affiliated to CWI, Amsterdam, The Netherlands.*

***Michel Mandjes** is a professor at the University of Amsterdam, The Netherlands. He received both his doctoral degree (1996) and M.Sc. degree (1993) from the Free University of Amsterdam, The Netherlands. He also worked as a member of the technical staff at Bell Laboratories/Lucent Technologies, Murray Hill NJ, United States and as department head at CWI, Amsterdam.*