Efficient Simulation of a Tandem Queue with Server Slow-down*

D.I. Miretskiy†, W.R.W. Scheinhardt†, M.R.H. Mandjes‡

Abstract

Tandem Jackson networks, and more sophisticated variants, have found widespread application in various domains. One such a 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 prespecified 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 normal 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. Then the knowledge of the most likely path is used to devise importance sampling algorithms, both for the normal 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.

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. In [3] this model was already introduced and the consequences for the first queue were studied, but here our main interest is to determine the probability of overflow in the *second* queue during a busy cycle. Here we define a busy cycle as the time between two consecutive arrivals to an empty system.

To be more specific, let us denote the number of jobs at server i by $X_i(t), i = 1, 2$. Jobs arrive at the first queue according to a Poisson process with 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 the second buffer reaches a high level 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, where the left panel shows the standard Jackson situation, and the right panel shows the slow-down situation. Note that the value of the threshold is θB with $\theta \in (0, 1)$, so that it scales with B.

*Part of this research has been funded by the Dutch BSIK/BRICKS project.

†Postal address: Department of Applied Mathematics, University of Twente, Postbus 217, 7500 AE Enschede, The Netherlands. E-mail address: {d.miretskiy, w.r.w.scheinhardt}@math.utwente.nl

‡Postal address:University of Amsterdam, Korteweg-de Vries Institute for Mathematics, Plantage Muidergracht 24, 1018 TV Amsterdam, The Netherlands. E-mail address: mmandjes@science.uva.nl

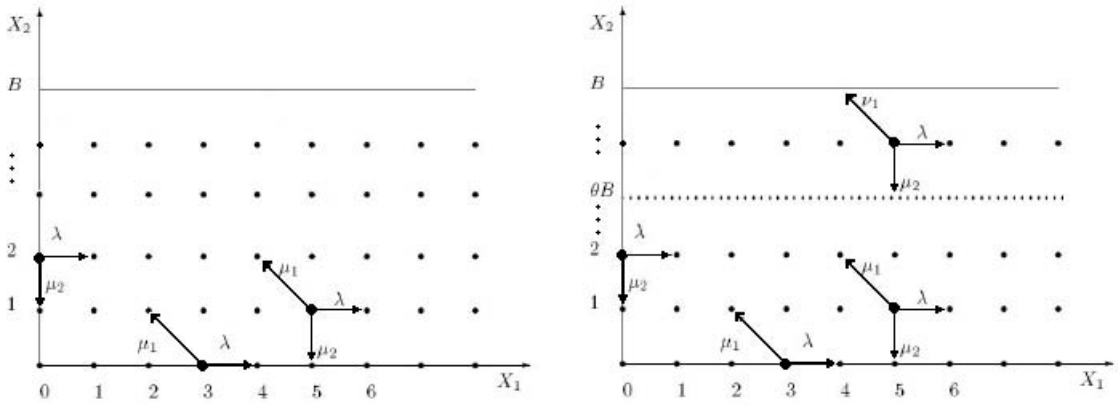

Figure 1: State space and transition structure for $(X_1(t), X_2(t))$

We note that even for the 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 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. For the standard Jackson case this was already known [1], as opposed to the model with slow-down. We have developed an appealing, but heuristic, method for detecting most likely path; when applying the method to the standard Jackson case, it indeed yields the path of [1]. It is remarked that the shape of the 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 Jackson network in Section 2 since we believe the results there are interesting on their own, and then for the slow-down model in Section 3. We conclude with some open problems for future research in Section 4.

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. In [7] an alternative measure was proposed, and in [4] it was found that this measure is not always asymptotically efficient, see also [2]. A more accurate state-dependent change of measure for the same problem was introduced in [10]. The special case of both servers having equal rates was studied in [8]. In [5] the focus is on the second queue reaching a high level as in our study, but the definition of the busy cycle is different from ours.

2 Tandem Jackson Network

2.1 Model and Preliminaries

In this section 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

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

where $\rho_i = \lambda/\mu_i$. For a typical state space representation we refer to the left panel of Figure 1.

Our main interest is to estimate the probability of reaching some large level B in the second buffer, as B grows large, during a busy cycle, which is defined as the time between two successive epochs at which the process leaves the empty state $(0,0)$. When we define the random variable T_B as the first entrance time of either level B or state (0,0), i.e.,

$$T_B = \min\{t > 0 | X_2(t) = B \text{ or } (X_1(t), X_2(t)) = (0,0)\},$$

then we can formally define the probability of interest in the following way

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

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 $x(i,j) = \lambda x(i+1,j) + \mu_1 x(i-1,j+1) + \mu_2(i,j-1)$ on the interior; for the boundaries we have similar equations. Unfortunately it is time consuming to solve such a system, which motivated us to choose simulations 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. Although it is not entirely trivial, we omit the proof, which is based on regenerative theory and the relation between p_B and the stationary probabilities $\pi(\cdot, B)$.

Theorem 1. *The overflow probability p_B is asymptotically geometric in B with parameter ρ_2. More precisely,*

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

Theorem 1 is important in itself, as it gives us already a rough estimate for the probability of interest (2) for large B. In fact it says 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. We will explain this in more detail in the sequel.

2.2 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 probably occurs, if it occurs. This question has already been answered in [1]. Here, time-reversibility of the tandem Jackson network is used 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 up to same value that is proportional to B (in the sequel this value is denoted by $-\alpha^{-1}B$, where α can be expressed in the arrival- and service rates). 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 (i.e., different from [1]), to find the optimal path. This method is heuristic by nature, but has some 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 from [1], since we do not know the explicit form for the stationary distribution in that case, and therefore we cannot use time-reversibility there. However our heuristic method *can* be applied here.

To proceed, let us first formulate a general conjecture, upon which our further research will be based.

Conjecture 2. *Consider a Markov chain for which the state space can be written as a finite union of disjoint sets on which the transition parameters are constant. It is our conjecture that in such models the typical path leading to the rare event consists of a finite concatenation of subpaths on the various subsets that are straight lines.*

If this conjecture is true, the great benefit is that the solution boils down to optimizing over a finite number of possible path-types, i.e., we reduced the problem to a combinatorial problem.

In our tandem model we have a simple situation with only four sets on which the transition parameters are constant, viz. the sets $\{(0,0)\}$, $\{(n,0)\}$, $\{(0,m)\}$ and $\{(n,m)\}$ where $n,m>0$. We check all possible paths consisting of a concatenation of straight lines. For each shape of the path we minimize some 'cost function' over all tilted parameters, which enables us to find the overall optimum in the end. The family of cost functions I we consider is defined by (also see [9], p. 14, 20)

$$I(\tilde{\lambda} \mid \lambda) = \lambda - \tilde{\lambda} + \tilde{\lambda}\log(\frac{\tilde{\lambda}}{\lambda}), \tag{4}$$

and $I(\tilde{\mu}_1 \mid \mu_1)$ and $I(\tilde{\mu}_2 \mid \mu_2)$ in the same way. We can think of the value $I(\tilde{\lambda} \mid \lambda)$ as the cost we need to pay to let a Poisson process with parameter λ behave like a Poisson process with parameter $\tilde{\lambda}$, per time unit. Note that the function (4) is convex and equals 0 at $\tilde{\lambda}=\lambda$.

For instance, consider for any i a 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 tilted parameters, such that $\tilde{\mu}_1 > \tilde{\mu}_2$ and $\tilde{\lambda} > \tilde{\mu}_1$, in order to have a drift upward and to the right. The total cost of such a path, per unit length in vertical direction is

$$\frac{I(\tilde{\lambda} \mid \lambda) + I(\tilde{\mu}_1 \mid \mu_1) + I(\tilde{\mu}_2 \mid \mu_2)}{\tilde{\mu}_1 - \tilde{\mu}_2}. \tag{5}$$

Here, the numerator represents the total cost per unit time, and the denominator is the average speed by which the process moves up. If we would replace the denominator by $\tilde{\lambda} - \tilde{\mu}_1$, we would find the cost per unit length in horizontal direction. Finally we mention that the slope of this path is given by

$$\alpha = \frac{\tilde{\mu}_1 - \tilde{\mu}_2}{\tilde{\lambda} - \tilde{\mu}_1}. \tag{6}$$

Minimizing (5) over the three tilted parameters, such that also $\tilde{\mu}_1 > \tilde{\mu}_2$ and $\tilde{\lambda} > \tilde{\mu}_1$ hold will then give the optimal values for the tilted parameters and the slope of the path for this particular shape. By considering all possible shapes that satisfy our conjecture, we then obtain the globally optimal change of measure and corresponding path. The cost function itself will yield $-\log d$ as its optimum value, where d is the geometric decay rate of the probability of interest; therefore we should have $d=\rho_2$ for the tandem model, see Theorem 1.

We will now split our problem 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.

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 tilted parameters for such a path we need to solve the following minimization problem

$$I_2 = \inf\left\{ \frac{I(\tilde{\lambda} \mid \lambda) + I(\tilde{\mu}_1 \mid \mu_1) + I(\tilde{\mu}_2 \mid \mu_2)}{\tilde{\lambda} - \tilde{\mu}_2} \right\}, \tag{7}$$

where the infimum is taken over all tilted variables $\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2$, such that $\tilde{\mu}_1 > \tilde{\lambda}$ and $\tilde{\lambda} > \tilde{\mu}_2$, ensuring a drift to the left and upward. Note that the denominator is again the average speed at which the process moves up. After taking partial derivatives with respect to all tilted 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 and conclude that for this case I_2 is in fact the minimal cost per unit movement in the vertical direction, and indeed ρ_2 is the decay rate.

Proposition 3. *Assume that Conjecture 2 holds and that $\lambda < \mu_2 < \mu_1$ (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*

$$(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2) = (\mu_2, \mu_1, \lambda). \tag{8}$$

Now let us see what happens if the first queue is the bottleneck.

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

We present the minimization problem for the path to overflow as described by [1]. Thus, assume we have tilted parameters that satisfy $\tilde{\mu}_1 < \tilde{\mu}_2$ and $\tilde{\mu}_1 < \tilde{\lambda}$, to ensure a path along the horizontal axis, with drift down and to the right. 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

$$I_1 = \inf \left\{ -\alpha^{-1} \frac{I(\tilde{\lambda} \mid \lambda) + I(\tilde{\mu}_1 \mid \mu_1) + I(\tilde{\mu}_2 \mid \mu_2)}{\tilde{\lambda} - \tilde{\mu}_1} + \frac{I(\bar{\lambda} \mid \lambda) + I(\bar{\mu}_1 \mid \mu_1) + I(\bar{\mu}_2 \mid \mu_2)}{\bar{\mu}_1 - \bar{\mu}_2} \right\},$$

where the infimum is taken over all tilted and barred variables that satisfy the given boundary conditions, and α is the slope of the second part of the path, i.e., $\alpha = (\bar{\mu}_1 - \bar{\mu}_2)/(\bar{\lambda} - \bar{\mu}_1)$, cf. (6). The solution for this problem can be found in two steps, first minimizing the first term over the tilted 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 been checked as well, and indeed none of them has lower cost.

Proposition 4. *Assume that Conjecture 2 holds and that $\lambda < \mu_1 < \mu_2$ (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 $\alpha = \frac{\bar{\mu}_1 - \bar{\mu}_2}{\bar{\mu}_1 - \lambda}$, and the decay rate is ρ_2. The corresponding change of measure is given by*

$$(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2) = (\mu_1, \lambda, \mu_2) \quad \text{until } X_1(t) = -\alpha^{-1}B, \tag{9}$$

$$(\bar{\lambda}, \bar{\mu}_1, \bar{\mu}_2) = (\mu_1, \mu_2, \lambda) \quad \text{afterwards.} \tag{10}$$

We conclude that the results which we found using our heuristic perfectly coincide with the results from [1]. Although a formal proof of the method is still lacking, this evidence shows 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.

2.3 Importance sampling

When we simulate our system, we use a new (changed) measure in order to increase the probability of overflow, until this happens; for the remainder of the busy cycle we will use the old measure. To compensate 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 in the following way:

$$p_B = \mathbb{E}^* L(X) \mathbf{1}(X), \tag{11}$$

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. The idea of IS is that we simulate the system N times under the new measure, and then estimate the probability by the sample mean:

$$\hat{p}_B = \frac{1}{N} \sum_{i=1}^{N} L(X_i) \mathbf{1}(X_i). \tag{12}$$

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 (12) the simulation effort grows subexponentially in B if it is asymptotically optimal. Let us explain what this means. Since the variance is

always nonnegative 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 call it asymptotically efficient or asymptotically optimal. A formal definition is as follows

Definition 5. *An IS estimator is called asymptotically optimal, if*

$$\lim_{B\to\infty} \frac{\log \mathbb{E}^* L^2(X)\mathbf{1}(X)}{\log p_B} = 2. \tag{13}$$

Corollary 6. *In the tandem case the IS estimator is asymptotically optimal if*

$$\lim_{B\to\infty} \frac{\log \mathbb{E}^* L^2(X)\mathbf{1}(X)}{B \log \rho_2} = 2. \tag{14}$$

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

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

Again we start our analysis with the simplest problem: 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 3.

To construct the probability estimator of a path, we need to know the likelihood of a sample path, which is just the product of the likelihoods of all individual transitions made during the path until either level B is reached, 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 in 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

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

Similary 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

$$L_2(X) = \left(\frac{\lambda}{\mu_2}\right)^{B-1+R} \left(\frac{\mu_1+\mu_2}{\lambda+\mu_1}\right)^{H}, \tag{15}$$

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 (12) is asymptotically optimal. Corollary 6 and (15) together give that we need

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

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

$$\gamma \left(\frac{\mu_1+\mu_2}{\lambda+\mu_1}\right)^2 < 1. \tag{16}$$

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 2 we present a contour plot of the left-hand side of (16) 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 (16) holds, so that the estimator

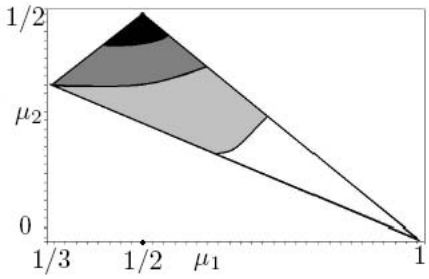

Figure 2: Contour plot of the left-hand side of (16) for the two-node tandem Jackson network, case 2, under the new measure (8). Values less than 0.5 are in white, less than 1 in light grey, less than 1.5 in dark grey and greater than 1.5 in black.

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 (14), which we also did empirically. The graphs in the left panel of Figure 3 represent for two different parameter settings the estimate of (14), given by

$$\hat{\psi}_B = \frac{\log \frac{1}{N} \sum_{i=1}^{N} L^2(X_i)\mathbf{1}(X_i)}{B \log \rho_2}, \tag{17}$$

for $B = 50$ as N, the number of replications, increases until 10^6. The values of the parameters (λ, μ_1, μ_2) are respectively $(0.1, 0.7, 0.2)$ (top graph) and $(0.3, 0.36, 0.34)$ (bottom graph). This is empirical evidence that for the first parameter setting we have an asymptotically efficient estimator, while for the second setting we do not.

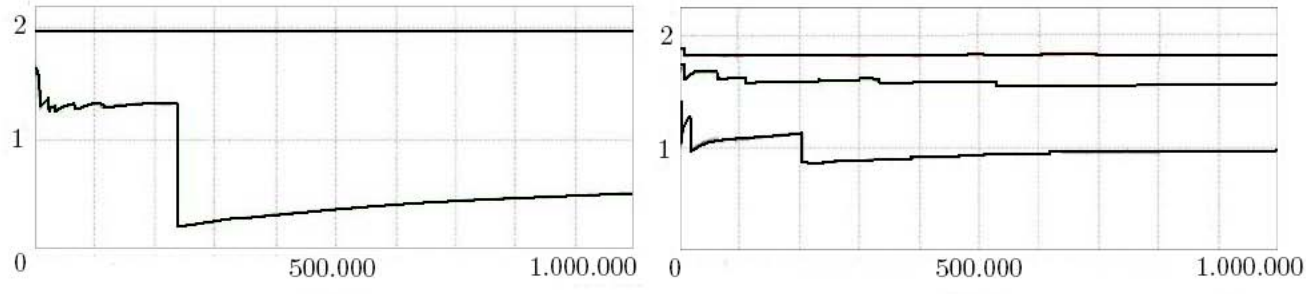

Figure 3: $\hat{\psi}_{50}$ against N. Left (right) panel correponds to case 2 (case 1).

Finally, for the same two parameter settings but 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 (14). Simulations for this table (and upcoming 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 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 do the $N = 10^6$ replications to estimate the overflow probability with confidence interval of width $4.3 \cdot 10^{-9}$ for the first parameter setting in Table 1. Compare this to straightforward simulations where, for a larger confidence interval of width $4 \cdot 10^{-8}$ 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 for 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 uncomparably smaller then in the case of an asymptotically efficient change of measure.

Remark 7. *When we compare our region of asymptotic efficiency with that in [2], they seem to coincide, although [2] considers the probability that the* total *network population, i.e.* $X_1(t) + X_2(t)$*, reaches some high*

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

Table 1: Simulation results for the two-node tandem Jackson network, case 2

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. $\lambda < \mu_1 < \mu_2$

Now let us focus on the case where the first queue is the bottleneck of the system. In Proposition 4 we showed that a good change of measure for this problem is given by (9)–(10).

The likelihood ratio of an arbitrary path to overflow now has a more complicated structure than in case 2:

$$L_1 = \left(\frac{\lambda}{\mu_2}\right)^{B-1-U} \left(\frac{\lambda}{\mu_1}\right)^{R-1} \left(\frac{\mu_1+\mu_2}{\lambda+\mu_2}\right)^{V_1} \left(\frac{\mu_1+\lambda}{\mu_2+\lambda}\right)^{V_2} \left(\frac{\mu_1+\mu_2}{\mu_1+\lambda}\right)^{H_2}, \tag{18}$$

where V_1 is the number of visits to the vertical axis under measure (9); V_2 and H_2 are the numbers of visits to vertical and horizontal axes under measure (10) respectively; U is the number of jobs in the second buffer 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 propose that V_1, V_2 and H_2 are geometrical random variables with parameters γ_1, γ_2 and γ_3 respectively. This proposition received confirmation from simulation experiments. Also assuming independence as B grows large, the inequality that should hold for asymptotic efficiency is now given by

$$\gamma_1 \left(\frac{\mu_1+\mu_2}{\lambda+\mu_2}\right)^2 \gamma_2 \left(\frac{\mu_1+\lambda}{\mu_2+\lambda}\right)^2 \gamma_3 \left(\frac{\mu_1+\mu_2}{\mu_1+\lambda}\right)^2 < 1. \tag{19}$$

Unfortunately, simulations show that (19) never holds under the change of measure (9)–(10). On the other hand, the right panel of Figure 3 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 various values of B we also 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.

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

Table 2: Simulation results for the two-node tandem Jackson network, case 1

Remark 8. *It is possible to consider various changes of measure that will result in the same optimal path. For instance, instead of switching from measure (9) to (10) once, we can also switch back and forth between these measures, depending on the current value of $X_2(t)$. Analysis of (18) 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. We start*

with measure (9); we always switch to measure (10) if $X_1 = \alpha^{-1}B$ and $X_2 > 0$; we always switch to the natural measure ($\tilde{\lambda} = \lambda, \tilde{\mu}_1 = \mu_1, \tilde{\mu}_2 = \mu_2$) if $X_1 > \alpha^{-1}B$ and $X_2 = 0$; we switch back to measure (9) if $X_1 \leq \alpha^{-1}B$ and $X_2 = 0$. Empirically it turns out that this change of measure (and other variants) is also not asymptotically efficient, although the variance of the estimator is a little less.

3 The slow-down system

In this section we will focus on the slow-down system which is very similar to the two-node tandem Jackson network studied in section 2, but now the rate of the first server depends on the content of the second buffer.

3.1 Model and preliminaries

Let us consider a two-node network with Poisson arrivals at rate λ and two stations with exponentially distributed service times with parameters μ_1 and μ_2. Again both queues are assumed to be infinitely large. In addition, when the number of jobs in the second buffer exceeds some *slow-down threshold* θB, that scales with the large overflow level B, with $\theta \in (0, 1)$, the first service station decreases its rate such that the (remaining) service times are exponential with parameter $\nu_1 < \mu_1$. Again we are interested in the estimation of the probability of reaching some large level B in the second buffer during a busy cycle as B grows large. Now we have a different stability condition $\lambda < \min(\nu_1, \mu_2)$, which guarantees rarity of the overflow event as B grows large. 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 the right panel of Figure 1 for the state space and transition structure.

As in the tandem model we can identify different cases, depending on the values of the parameters, but now we have three cases: **(3)** $\lambda < \mu_2 < \nu_1 < \mu_1$, **(4)** $\lambda < \nu_1 < \mu_2 < \mu_1$ and **(5)** $\lambda < \nu_1 < \mu_1 < \mu_2$

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. In the remainder of the paper we will focus only on cases 3 and 4, where the system initially has its bottlenek at the second server. One reason is that in the previous section we were not sure whether an asymptotically efficient estimator for the corresponding case 1 exists, so that there is less hope we will find one in our current case 5. Another reason is that from an applications point of view, case 4 is the most interesting and effective, shifting the bottleneck from the second server to the first to protect the second buffer from overflow; case 3 can be added without difficulty because it is essentially the same as case 2 in our earlier model.

3.2 The optimal path

As mentioned earlier, we cannot use a reversibility argument as in [1] in the analysis of the slow-down system. However, we can employ our cost function approach, based on Conjecture 2. Notice in particular that for the current model, the state space can be written as the union of the sets $\{(0,0)\}$, $\{(n,0)\}$, $\{(0,m)\}$, $\{(n,m)|0 < m < \theta B\}$ and $\{(n,m)|m \geq \theta B\}$ with $n, m > 0$, where on each of these sets the transition structure is constant.

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

Let us start from the situation in which the second server stays the bottleneck, analysing 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 (7) we need to find

$$I_3 = \inf \left\{ \theta \frac{I(\tilde{\lambda} \mid \lambda) + I(\tilde{\mu}_1 \mid \mu_1) + I(\tilde{\mu}_2 \mid \mu_2)}{\tilde{\lambda} - \tilde{\mu}_2} + (1-\theta) \frac{I(\bar{\lambda} \mid \lambda) + I(\bar{\nu}_1 \mid \nu_1) + I(\bar{\mu}_2 \mid \mu_2)}{\bar{\lambda} - \bar{\mu}_2} \right\}, \tag{20}$$

where the infimum is taken over all tilted and barred variables, such that the horizontal and vertical drifts are to the left and upward, both below the threshold (i.e., $\tilde{\mu}_1 > \tilde{\lambda}$ and $\tilde{\lambda} > \tilde{\mu}_2$) and above the threshold (i.e., $\bar{\nu}_1 > \bar{\lambda}$ and $\bar{\lambda} > \bar{\mu}_2$). This can easily be solved by splitting it in two separate minimization problems that are

completely analoguous to (7), 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 and conclude that indeed $I_3 = -\log \rho_2$ is the minimal cost for unit movement in the vertical direction.

Proposition 9. *Assume that Conjecture 2 holds and that $\lambda < \mu_2 < \nu_1 < \mu_1$ (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*

$$(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2) = (\mu_2, \mu_1, \lambda) \quad \textit{and} \quad (\bar{\lambda}, \bar{\nu}_1, \bar{\mu}_2) = (\mu_2, \nu_1, \lambda), \tag{21}$$

and the decay rate is ρ_2.

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

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 upward with drift to the right. The following minimization problem corresponds to this type of path:

$$I_4 = \inf \left\{ \theta \frac{I(\tilde{\lambda} \mid \lambda) + I(\tilde{\mu}_1 \mid \mu_1) + I(\tilde{\mu}_2 \mid \mu_2)}{\tilde{\lambda} - \tilde{\mu}_2} + (1-\theta) \frac{I(\bar{\lambda} \mid \lambda) + I(\bar{\nu}_1 \mid \nu_1) + I(\bar{\mu}_2 \mid \mu_2)}{\bar{\nu}_1 - \bar{\mu}_2} \right\}, \tag{22}$$

where we take the infimum over all tilted and barred variables, such that $\tilde{\mu}_1 > \tilde{\lambda}$ and $\tilde{\lambda} > \tilde{\mu}_2$ and $\bar{\nu}_1 > \bar{\mu}_2$ and $\bar{\lambda} \geq \bar{\nu}_1$. Again we can decompose the optimization problem into two parts. The first part of (22) has the same solution as the first part of (20), (and hence as (7)). The second part of problem (22) 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

$$\bar{\lambda} = \bar{\nu}_1 = \sqrt{\frac{\lambda \nu_1}{z}}, \ \bar{\mu}_2 = \mu_2 z, \tag{23}$$

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

$$\lambda + \nu_1 + \mu_2(1-z) = 2\sqrt{\frac{\lambda \nu_1}{z}}. \tag{24}$$

As an aside we note that this is the same equation as (30) in [6], and indeed the decay rate behavior as found in that paper can also be obtained using our heuristic. Since all other paths have higher cost, this is the optimal path, and the corresponding cost is $I_4 = -\log \rho_2$ per vertical unit.

Proposition 10. *Assume that Conjecture 2 holds and that $\lambda < \nu_1 < \mu_2 < \mu_1$ (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 (8) and (23), 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 horizontal drift is to the left, and the path is vertical. 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 10 is meant to express this difference.

3.3 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 (12), 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. $\lambda < \mu_2 < \nu_1 < \mu_1$

In this case we use the change of measure given in (21). 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 11. *Assume that $\lambda < \mu_2 < \nu_1 < \mu_1$ (case 3), then the overflow probability estimators under the measure (8) for the tandem Jackson network and (21) 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 (15), namely

$$L_3(X) = \left(\frac{\lambda}{\mu_2}\right)^{B-1+R'} \left(\frac{\mu_1+\mu_2}{\lambda+\mu_1}\right)^{H'}, \tag{25}$$

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 $(\lambda/\mu_2) < 1$. The distribution of H' on the other hand 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 $H' \xrightarrow{d} H$, as $\theta B \to \infty$. Looking back to (15) for the definition of the random variable H we obtain 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 $\nu_1 = 0.35$. Note the correspondence to the examples in Section 2.3, and that for both cases $\nu_1 > \mu_2$. Indeed in the first case the estimator is asymptotically optimal, and in the second case it is not, which can be illustrated by a similar picture as Figure 3, and also by the values of the estimator of the left-hand side of (13) in Table 3, which is now given by

$$\hat{\psi}'_B = \frac{\log \frac{1}{N}\sum_{i=1}^{N} L^2(X_i)\mathbf{1}(X_i)}{\hat{p}_B}, \tag{26}$$

The reason we use (26) instead of (17) is that we do not have an analogue to Theorem 1 for the slow-down case and hence no analogue to Corollary 6. The speedups obtained in this table are comparable to those in Table 1.

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

Table 3: Simulation results for the slow-down system, case 3

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

In this case we use the change of measure given in Proposition 10. Specifically, we always use (8) when the number of jobs in the second buffer is below θB and (23) else, until level B is reached for the first time. It is more difficult now to obtain an analogue of Proposition 11, since the process may have some cycles around level θB that influence the total likelihood of the path. Since it is difficult to see how the likelihood behaves, we content ourselves with some simulation results for the same scenarios as in cases 2 and 3, but now taking $\nu_1 < \mu_2$. These can be found in Table 3.3.

Again it seems clear that the change of measure (23) 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 then we found in case 2.

$(0.1, 0.7, 0.2) \to (0.1, 0.15, 0.2)$			$(0.3, 0.36, 0.34) \to (0.3, 0.32, 0.34)$		
B	$\hat{\psi}'_B$	p_B	B	$\hat{\psi}'_B$	p_B
20	1.92	$3.79 \cdot 10^{-7} \pm 1.09 \cdot 10^{-9}$	20	0.45	$5.6 \cdot 10^{-2} \pm 1.01 \cdot 10^{-4}$
50	1.95	$1.26 \cdot 10^{-16} \pm 5.08 \cdot 10^{-19}$	50	1.34	$1.17 \cdot 10^{-3} \pm 2.85 \cdot 10^{-5}$
100	1.98	$3.54 \cdot 10^{-32} \pm 1.89 \cdot 10^{-34}$	100	1.45	$1.69 \cdot 10^{-6} \pm 1.23 \cdot 10^{-7}$

Table 4: Simulation results for the slow-down system, case 4

4 Future work

We conclude our paper by mentioning a number of potential lines 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 plan to continue studying the behavior of estimator (12) under the state-independent change of measures (8), (9), (10) and to find analytical expressions for asymptotically efficient parameter regions. In the case these state-independent change of measures yield unsatisfactory results, we could resort to state-dependent schemes.

For the normal tandem model we found in Theorem 1 an expression for the logarithmic decay rate; 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. Also for this model we are interested in the estimation of the overflow probability, the precise form of asymptotically efficient regions, and a deeper understanding of the nature of the behavior of estimator (12).

References

[1] V. Anantharam, P. Heidelberger and P. Tsoucas (1990). Analysis of rare events in continuous time Markov chains via time reversal and fluid approximation. *IBM Research Report RC 16280.*

[2] P.-T. de Boer (2006). Analysis of state-independent importance-sampling measures for the two-node tandem queue. *ACM Transactions on Modeling and Computer Simulation* **16** (3), 225–250.

[3] N.D. van Foreest, M.R.H. Mandjes, J.C.W. van Ommeren, W.R.W. Scheinhardt (2005). A tandem queue with server slow-down and blocking. *Stochastic Models* **21** (2-3), 695–724.

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

[5] D.P. Kroese, V.F. Nicola (2002). Efficient simulation of a tandem Jackson network. *ACM Transactions on Modeling and Computer Simulation* **12** (2), 119–141.

[6] D.P. Kroese, W.R.W. Scheinhardt, P.G. Taylor (2004). Spectral properties of the tandem Jackson network, seen as quasi-birth-and-death process. *The Annals of Applied Probability* **14** (4), 2057–2089.

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

[8] W. Sandmann (2004). Fast simulation of excessive population size in tandem Jackson networks. Proceedings of the IEEE Computer Society's 12th Annual International Symposium MASCOTS'04.

[9] A. Shwartz, A. Weiss (1995). *Large deviations for performance analysis. Queues, communications and computing.* Chapmann & Hall, London, UK.

[10] T.S. Zaburnenko, V.F. Nicola (2005). Efficient heuristics for simulating population overflow in tandem networks. Proceedings of the Fifth Workshop on Simulation, 755–764.