Simple and Efficient Importance Sampling Scheme for a Tandem Queue with Server Slow-down

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

Abstract

This paper considers importance sampling as a tool for rare-event simulation. The system at hand is a so-called tandem queue with slow-down, which essentially means that the server of the first queue (or: upstreanm queue) switches to a lower speed when the second queue (downstream queue) exceeds some threshold. The goal is to assess to what extent such a policy succeeds in protecting the first queue, and therefore we focus on estimating the probability of overflow in the downstream queue.

It is known that in this setting importance sampling with traditional state-independent distributions performs poorly. More sophisticated state-dependent schemes can be shown to be asymptotically efficient, but their implementation may be problematic, as for each state the new measure has to be computed. This paper presents an algorithm that is considerably simpler than the fully state-dependent scheme; it requires low computational effort, but still has high efficiency.

1 Introduction

Importance sampling (IS) is a powerful and flexible technique to speed up the Monte Carlo simulations of rare events. The main idea behind IS is to simulate a system under a new probability measure which guarantees more frequent occurrence of the rare event of interest. To obtain an unbiased estimator, the output of the simulations is corrected by so-called likelihood ratios. The challenge is to construct a 'good' new measure. An often-used notion in this respect is that of asymptotic efficiency (or: asymptotic optimality) which essentially means that the variance of the estimator behaves approximately as the square of its first moment. When this is not the case, the estimator may even have infinite variance. We refer to [5] for more background and history of the IS method.

In the current paper, we consider the socalled slow-down network which was introduced in [13], see also [6, 2]. This is a two-node Jackson-like tandem queue, with the additional feature that the rate of the first server slows down when the number of jobs in the second queue is greater or equal than some pre-specified threshold value. When the content of the second buffer drops below the threshold, the first server returns to its normal speed. In this way, the second buffer is offered some protection against frequent overflow. We are interested in estimating the probability of such an overflow before the system becomes empty, starting off from any given state.

One approach to the efficient simulation of this problem is to use a state-independent IS scheme, in which the new measure, which is suggested by large deviations analysis, is static. In this respect we mention the landmark paper by Parekh and Walrand [11], in which amongst others a new measure is constructed for estimating the probability of overflow in a single $M/M/1$ system; this measure prescribes to simulate the system while interchanging the arrival and service rates. Two years later, Sadowsky proved that this type of new measure is asymptotically efficient even for a $GI/GI/m$ queue (with light-tailed service times), see [12]. Application of a similar new measure to a two-node Jackson tandem network (swapping the slowest service rate and arrival rate) was not so encouraging – the method was asymptotically efficient for some parameter values, but has unbounded variance for other values; see [4, 1]. Also for the slow-down model state-independent schemes were established, see [6]. Asymptotic efficiency of those schemes was concluded experimentally, but only for a limited set of parameter settings.

A second approach is to use a state-dependent new measure, i.e., a new measure in which the parameters to be used may depend on the current state of the system during the simulations. In [14] such a scheme for a tandem Jackson network with arbitrary number of nodes was constructed and asymptotic efficiency was shown experimentally, but an analytic proof was lacking. The first important result in that direction (also for a k-node tandem Jackson) was achieved by Dupuis et al. [3], who proposed a change of measure based

on game-theoretic analysis, and proved the scheme to be asymptotically efficient. Also for the slow-down system a state-dependent scheme was proposed and proved to be efficient in [2]. However these results are valid only when the starting state is the origin (i.e. an empty system); moreover they were only shown under some assumptions on the model parameters (i.e. the arrival rate to the first queue, and the service rates of both queues). The latest results on the slow-down model can be found in [10], where a family of state-dependent IS schemes is presented for general starting states and all possible parameter settings. However, the focus of that paper was on the analysis of the decay rate and the proof of asymptotic efficiency, and no numerical experiments were reported. The reason for this is that the step to an actual efficient implementation is not straightforward at all, since it entails amongst others solving a system of two joint cubic equations to determine the state-dependent new transition rates for each step during the simulations. Hence, although the IS scheme itself is asymptotically efficient, the computational effort would still be considerable.

The contribution of the current paper is that we present a simple and efficient IS implementation for simulating the overflow probability in the slow-down model. On the one hand it is as easy to implement as the scheme in [2], or even as that in [6], while also performing comparably in terms of computational demand. On the other hand it allows any given starting state, while inheriting asymptotic efficiency from [10]. No assumptions on the model parameters are needed, but we only present the analysis for the case that was also considered in [2]. We provide a substantial number of numerical results, including a variety of parameter settings, and make the comparison with [2] and [6] where this is possible.

We finish this section by indicating the structure of the paper. We describe the model of interest in detail and provide some importance sampling background in Section 2. We also establish the stability criterion for the slow-down network in this section. In Section 3 we present our IS scheme. The proof of asymptotic efficiency is presented in Section 4. Supporting numerical results are presented in Section 5 and we conclude in Section 6.

2 Model and Preliminaries

2.1 Model

In this section we describe the slow-down network in detail. It consist of two stations with Poisson arrivals at rate λ to the first station. At first any job receives service at the first station. After the first service completion, job is immediately rerouted to the second station. After receiving the second service, job leaves the system. Service times at the second station are exponential with parameter μ_2. We have a more interesting situation at the first station, whose service speed depends on the content of the second queue. Normally, service times at the first station are exponential with parameter μ_1, but if the number of jobs in the second queue exceeds some pre-specified value – the slow-down threshold – then the service times are still exponential, but the parameter of the distribution is μ_1^+, where $\mu_1^+ < \mu_1$. When system 'stabilizes' and the number of jobs in the second queue is again below the slow-down threshold, the rate of the first station returns to its original value μ_1.

For convenience we choose the parameters such that $\lambda + \mu_1 + \mu_2 = 1$, without loss of generality. A clear consequence is that $\lambda + \mu_1^+ + \mu_2 < 1$. Again, as in [10], we assume the waiting rooms at both stations to be infinitely large and we define the discrete-time joint queue-length process $Q_j = (Q_{1,j}, Q_{2,j})$. Here $Q_{i,j}$ is the number of jobs at node i after the j-th transition. We define the possible jump directions of the process Q_j via vectors $v_0 = (1,0)$, $v_1 = (-1,1)$ and $v_2 = (0,-1)$ with corresponding jump rates λ, μ_1 (or μ_1^+) and μ_2. This process is regenerative if we assume stability, as we will do, see Section 2.3. Our main interest is to estimate the probability of reaching some high level B in the second queue before it returns to the origin, starting from any state. Note that in our model the slow-down threshold scales with B, we will determine it as θB in the rest of the work.

We will also consider the scaled process $X_j = Q_j/B$. The advantage of this scaling is that we may use the same state space $[0,\infty) \times [0,1]$ for any value of B. In particular, our target probability is equivalent to the probability that the second component of the scaled process X_j reaches 1 before it returns to the origin.

We introduce the following subsets of the state space, with $x := (x_1, x_2)$:

$$\begin{array}{lll} D := \{x : x_1 > 0, 0 < x_2 < \theta\}, & \partial_1 := \{(0,x_2) : x_2 > 0\}, & \partial_\theta := \{(x_1,\theta) : x_1 \geq 0\}, \\ D^+ := \{x : x_1 > 0, \theta \leq x_2 < 1\}, & \partial_1^+ := \{(0,x_2) : x_2 \in [\theta,1)\}, & \partial_e := \{(x_1,1) : x_1 \geq 0\}, \\ & \partial_2 := \{(x_1,0) : x_1 > 0\}. & \end{array}$$

The full state space is $\bar{D} \cup \bar{D}^+$, where $\bar{D} := D \cup \partial_\theta \cup (\partial_1 \setminus \partial_1^+) \cup \partial_2$ and $\bar{D}^+ := D^+ \cup \partial_e \cup \partial_1^+ \cup \partial_\theta$.

Note that transition v_k is impossible when queue k is empty, i.e., when $X_j \in \partial_k$. We modify the process X_j to deal with this by allowing some self-loop transitions in the following way (see also Figure 1): for $k = 1, 2$,

$$\mathbb{P}(X_{j+1} = X_j | X_j \in \partial_k \setminus \partial_1^+) = \mu_k, \quad \mathbb{P}(X_{j+1} = X_j | X_j \in \partial_1^+) = \mu_1^+ / (\lambda + \mu_1^+ + \mu_2). \tag{1}$$

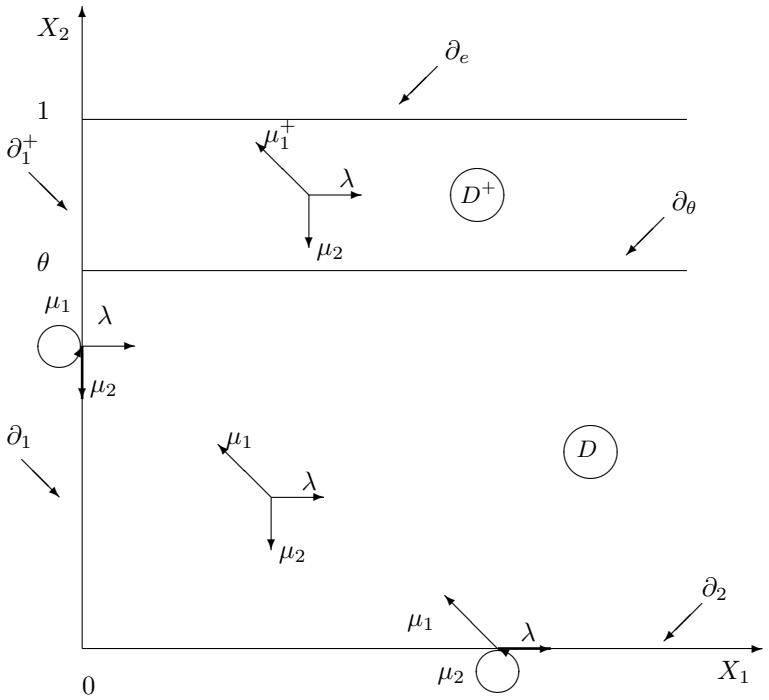

Figure 1: State space and transition structure for the scaled process X_j.

Next, we introduce the stopping time τ_B^s, which is the first time that the process X_j hits level 1, starting from state $x^s = (x_1^s, x_2^s)$, without visits to the origin:

$$\tau_B^s = \inf\{k > 0 : X_k \in \partial_e, X_j \neq 0 \text{ for } j = 1, \ldots, k-1\}, \tag{2}$$

and we define $\tau_B^s = \infty$ if X_j hits the origin before ∂_e. It will also be convenient to let $I(A_B^s)$ be the indicator of the event $\{\tau_B^s < \infty\}$ for the scaled sample path $A_B^s = (X_j, j = 0, \ldots : X_0 = x^s)$. Thus we can write the probability of our interest as

$$p_B^s = \mathbb{E} I(A_B^s) = \mathbb{P}(\tau_B^s < \infty). \tag{3}$$

It is clear that estimating the probability p_B^s through direct, naïve, simulations is not feasible when B grows large. We therefore have to use some alternative techniques to obtain a reliable estimator. In this paper we focus on importance sampling, which we will now describe briefly.

2.2 Background on Importance Sampling

To estimate p_B^s, IS generates samples under a new probability measure $\mathbb{Q}$, with respect to which $\mathbb{P}$ is absolutely continuous. The probability $\mathbb{P}(A_B^s)$ can now alternatively be expressed as

$$\mathbb{P}(A_B^s) = \mathbb{E}^{\mathbb{Q}}[LI], \tag{4}$$

where I is an indicator function and L is the likelihood ratio (also known as Radon-Nikodým derivative) of a realization ('path') ω:

$$L = \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}(\omega). \tag{5}$$

After n iterations we obtain a family of observations $(L_i, I_i), i = 1, \ldots, n$ and are able to construct the unbiased estimator of $\mathbb{P}(A_B)$ by $n^{-1} \cdot \sum_{i=1}^{n} L_i I_i$. We conclude this subsection by introducing the definition of asymptotical efficiency.

Definition 2.1. The IS scheme for $\mathbb{P}(A_B)$ is called *asymptotically efficient* if

$$\liminf_{B\to\infty} \frac{\log \mathbb{E}^{\mathbb{Q}}[L^2 I]}{\log \mathbb{E}^{\mathbb{Q}}[LI]} \geq 2. \tag{6}$$

If the probability of $\{A_B^s\}$ decays exponentially in B, i.e.,

$$-\lim_{B\to\infty} \frac{1}{B} \log \mathbb{P}(A_B^s) \ \in \ (0, \infty),$$

then we can apply the relation $\mathbb{E}^{\mathbb{Q}}[L^2 I] = \mathbb{E}[LI]$ and (4) to simplify expression (6) to obtain

$$\limsup_{B\to\infty} \frac{1}{B} \log \mathbb{E}[LI] \leq 2 \lim_{B\to\infty} \frac{1}{B} \log \mathbb{P}(A_B^s). \tag{7}$$

2.3 Stability

The stability condition for the standard two-node Jackson tandem network is the very simple and well-known criterion:

$$\lambda \ < \ \min(\mu_1, \mu_2).$$

Here we provide the stability condition from [7] for the slow-down network, which was not known before. This criterion is very important in our context. It gives the possibility to identify which parameter settings are interesting and which are not. In other words we know in advance when the system is unstable and hence when the event of our interest is not rare. In the following theorem the integer m is the value of the slow-down threshold (which is chosen to scale with B in the rest of the paper as $m = \theta B$).

Theorem 2.2. *The slow-down network with parameters $(\lambda, \mu_1, \mu_1^+, \mu_2)$ and the slow-down threshold m is stable if and only if*

$$\lambda < \frac{\mu_1 |1-\psi^m||1-\psi^+| + \mu_1^+ \psi^m |1-\psi|}{|1-\psi^m||1-\psi^+| + \psi^m |1-\psi|},$$

with $\psi = \mu_1/\mu_2$ and $\psi^+ = \mu_1^+/\mu_2$.

We refer to [8] for the proof of this theorem. This proof considers two cases separately: $\mu_1^+ \leq \mu_2$ and $\mu_2 < \mu_1^+$. In the first case we use Foster's criterion to design the stability condition and prove it. The proof in the second case is based on some non-trivial stochastic analysis.

It may be interesting to note that the slow-down system can be stable even when $\lambda > \mu_1^+$. The intuition behind this is as follows. Consider the case when both $\lambda > \mu_1^+$ and the condition in Theorem 2.2 hold true. The content of the first queue typically increases when the number of jobs in the second queue is above the slow-down threshold. However, it stays finite because the content of the second queue tends to decrease and the system returns to its normal state in which the number of jobs in the first queue tends to decrease.

3 Importance Sampling

The main purpose of this paper is to design the modification of the IS schemes established in our previous work [10]. We show that the new schemes inherit the property of asymptotic efficiency, but have a much simpler structure. At first, we wish to mention the main difference between the new measure introduced in [10] and the new measure we derive in this paper. In [10], the new measures $(\tilde{\lambda}(x), \tilde{\mu}_1(x), \tilde{\mu}_2(x))$ and $(\tilde{\lambda}^+(x), \tilde{\mu}_1^+(x), \tilde{\mu}_2^+(x))$ are changing (or in other words, have to be recalculated) after every transition in such a way as to follow the most probable path to overflow. In this paper we calculate new measures $(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2)$ and $(\tilde{\lambda}^+, \tilde{\mu}_1^+, \tilde{\mu}_2^+)$ *only once* according to the initial state of the process and continue to use it all the time (but see Remark 3.2). In other words the new measures, provided in this paper, depend on

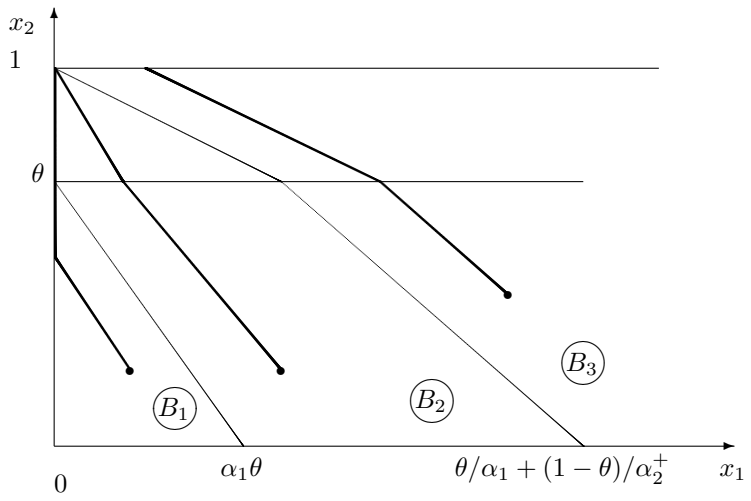

Figure 2: Partition of $\bar{D} \cup \bar{D}$ and some optimal paths to overflow when $\mu_1^+ \leq \mu_2 < \mu_1$.

the starting state x^s in the same manner as the measures from [10] depend on *the current state* x. Here we use some modification along the boundaries, which is similar to the one used in [10].

We decide to restrict the analysis to the case when the bottleneck shifts, i.e., when $\mu_2 < \mu_1^+ < \mu_1$. The rest of the cases (i.e., $\mu_1^+ < \mu_1 \leq \mu_2$ and $\mu_1^+ \leq \mu_2 < \mu_1$) can be dealt with in a similar manner, see also Remark 3.2, and in fact we will present numerical results for these cases as well. Throughout this section we fix the starting state x^s and assume it is situated below the slow-down threshold, i.e. $x^s \in \bar{D}$, as this is the most interesting case.

At first let us recall from [10] the most probable path to overflow and the pair of new measures $(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2)$ and $(\tilde{\lambda}^+, \tilde{\mu}_1^+, \tilde{\mu}_2^+)$ that will ensure that any sample will follow the optimal trajectory with high probability. To this end we assign a 'cost' to any path, minimizing which we obtain the optimal trajectory and corresponding new measure. We refer to [9, 10] for the precise description of this method. To ease the exposition on the new measures we divide the state space as it is shown in Figure 2, which also provides some examples of the most probable overflow trajectories (solid lines). We are particularly interested in the partition of the bottom part of the state space:

$$\begin{aligned} B_1 &:= \{x \in \bar{D} : x_2 \leq -\frac{x_1}{\alpha_1} + \theta\}, \\ B_2 &:= \{x \in \bar{D} : -\frac{x_1}{\alpha_1} + \theta < x_2 < -\alpha_1 x_1 - \frac{\alpha_1}{\alpha_2^+}(1-\theta) + \theta\}, \\ B_3 &:= \{x \in \bar{D} : x_2 \geq -\alpha_1 x_1 - \frac{\alpha_1}{\alpha_2^+}(1-\theta) + \theta\}, \end{aligned} \tag{8}$$

where $\alpha_1 = (\mu_1 - \mu_2)/(\mu_1 - \lambda)$ and $\alpha_2^+ = (\mu_2 - \mu_1^+)/(\mu_2 - \lambda)$.

The new measures for $x^s \in B_1 \cup B_3$ are not difficult. However, in order to find the optimal new measure for $x^s \in B_2$ one first needs to solve the following systems jointly

$$\begin{cases} \tilde{\lambda} = \tilde{\mu}_1 + \frac{\kappa - x_1^s}{\theta - x_2^s}(\tilde{\mu}_1 - \tilde{\mu}_2) \\ \tilde{\lambda} + \tilde{\mu}_1 + \tilde{\mu}_2 = \lambda + \mu_1 + \mu_2 \\ \tilde{\lambda}\tilde{\mu}_1\tilde{\mu}_2 = \lambda\mu_1\mu_2 \\ \tilde{\lambda} \leq \tilde{\mu}_1 \text{ and } \tilde{\mu}_1 > \tilde{\mu}_2 \\ \tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2 > 0 \end{cases} \tag{9}$$

and

$$\begin{cases} \tilde{\lambda}^+ = \tilde{\mu}_1^+ - \frac{\kappa}{1-\theta}(\tilde{\mu}_1^+ - \tilde{\mu}_2^+) \\ \tilde{\lambda}^+ + \tilde{\mu}_1^+ + \tilde{\mu}_2^+ = \lambda + \mu_1^+ + \mu_2 \\ \tilde{\lambda}^+\tilde{\mu}_1^+\tilde{\mu}_2^+ = \lambda\mu_1^+\mu_2 \\ \tilde{\lambda}^+ \leq \tilde{\mu}_1^+ \text{ and } \tilde{\mu}_1^+ > \tilde{\mu}_2^+ \\ \tilde{\lambda}^+, \tilde{\mu}_1^+, \tilde{\mu}_2^+ > 0 \end{cases} \tag{10}$$

with condition

$$\kappa := x_1^s - \frac{\tilde{\mu}_1 - \tilde{\lambda}}{\tilde{\mu}_1 - \tilde{\mu}_2}(\theta - x_2^s) = \frac{\tilde{\mu}_1^+ - \tilde{\lambda}^+}{\tilde{\mu}_1^+ - \tilde{\mu}_2^+}(1-\theta). \tag{11}$$

Now we are ready to define the new measures below and above the slow-down threshold, which depend only on the starting state x^s. The new measure below the slow-down threshold, $(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2)$, is as follows

$$(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2) = \begin{cases} (\mu_2, \mu_1, \lambda), & \text{if } x^s \in B_1, \\ \text{solution to (9)}, & \text{if } x^s \in B_2, \\ (\lambda, \mu_1, \mu_2), & \text{if } x^s \in B_3. \end{cases} \tag{12}$$

Above the slow-down threshold the new measure $(\tilde{\lambda}^+, \tilde{\mu}_1^+, \tilde{\mu}_2^+)$ is defined by

$$(\tilde{\lambda}^+, \tilde{\mu}_1^+, \tilde{\mu}_2^+) = \begin{cases} (\sqrt{\frac{\lambda\mu_1^+}{z^+}}, \sqrt{\frac{\lambda\mu_1^+}{z^+}}, \mu_2 z^+), & \text{if } x^s \in B_1, \\ \text{solution to (10)}, & \text{if } x^s \in B_2, \\ (\lambda, \mu_2, \mu_1^+), & \text{if } x^s \in B_3, \end{cases} \tag{13}$$

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

$$\lambda + \mu_1^+ + \mu_2(1 - z^+) = 2\sqrt{\frac{\lambda\mu_1^+}{z^+}}, \tag{14}$$

see also [10].

Now, let us define $\gamma(x)$ to be the residual cost of moving from state x to ∂_e along the path to overflow that started in x^s:

$$\gamma(x) := \begin{cases} \gamma_1(x) + \gamma_2(\kappa, \theta) & \text{if } x \in \bar{D}, \\ \gamma_2(x) & \text{if } x \in \bar{D}^+. \end{cases}$$

with

$$\gamma_1(x) := -(x_1 - \kappa)\log\frac{\tilde{\lambda}}{\lambda} - (\theta - x_2)\log\frac{\tilde{\mu}_2}{\mu_2}, \quad \text{if } x \in \bar{D} \tag{15}$$

being the minimal cost of the bottom part of the path to overflow and

$$\gamma_2(\kappa, \theta) := -\kappa\log\frac{\tilde{\lambda}^+}{\lambda} - (1-\theta)\log\frac{\tilde{\mu}_2^+}{\mu_2}, \quad \text{if } x \in \bar{D}^+ \tag{16}$$

being the minimal cost of the top part of the optimal path to overflow. Note that κ, $(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2)$ and $(\tilde{\lambda}^+, \tilde{\mu}_1^+, \tilde{\mu}_2^+)$ (given by (11), (12) and (13) respectively) are fixed, i.e., they only depend on the fixed initial state x^s, and not on the current state x (as was the case in [10]).

Now we are ready to introduce an important large deviations result. We refer to [10] for the proof of this theorem.

Theorem 3.1. *The exponential decay rate of the p_B^s is equal to the minimal cost of overflow $\gamma(x^s)$, i.e.,*

$$-\lim_{B\to\infty}\frac{1}{B}\log p_B^s = \gamma(x^s).$$

As in [9, 10], the description of the new measure will be based on some function $W(x)$, which is closely related to the decay rate function in Theorem 3.1, and provides an adaptation close to the boundaries to protect the likelihood ratio.

The idea of defining a new measure in terms of such a function is due to [3], where the function was found using a game-theoretic framework. Here we simply define the following:

$$\begin{aligned} W_1(x) &= 2\gamma(x) - \delta, \\ W_2(x) &= 2\gamma(x_1, \delta/2\log\frac{\mu_2}{\lambda}) - \delta, \\ W_3(x) &= 2\gamma(0) - 3\delta, \end{aligned} \tag{17}$$

where δ is some small positive number. We would like to explain the meanings of the functions $W_i(x)$ briefly. The main function $W_1(x)$ provides the pair of state-independent measures (below and above the slow-down threshold) that 'push' any particular sample path to follow the optimal trajectory. Note, that these measures do not change in case a sample path deviates from the optimal trajectory, as was the case in [10]. The functions $W_2(x)$ and $W_3(x)$ provide two different state-independent measures that are used to vanish the affection of the likelihood around the boundaries, where it is not optimal to use the 'main' measure.

Finally, to define the state-dependent new measure we construct the function $W(x)$. For this reason the following mollification procedure is applied, see also [3, 2, 9]

$$W(x) := -\epsilon \log \sum_{i=1}^{3} e^{-W_i(x)/\epsilon}. \tag{18}$$

Here ϵ is a 'smoothness' parameter; a larger value of ϵ corresponds to a smoother function $W(x)$. Moreover, we see that $W(x)$ converges to the non-smooth function $W_1(x) \wedge W_2(x) \wedge W_3(x)$ as $\epsilon \to 0$. We have to mention that function $W_1(x)$ (and consequently $W(x)$) is not continuous around the slow-down threshold, so our use of the word smooth is not entirely correct here.

We briefly discuss some issues which arise due to the different definitions of the function $W_1(x)$ in our paper and in [2]. The function $W_1(x)$ in [2] can be rewritten in our notation as $W_1(x) = 2\min\{\gamma_1(x) + \gamma_2(0,\theta), \gamma_2(x)\} - \delta$. Note that this definition only holds when the starting state is the origin. Such a definition guarantees that $W_1(x)$ is continuous, which simplifies the proof of asymptotic efficiency. However, it also implies that the new measure allows sample paths to deviate significantly from the optimal trajectory. This may happen because the new measure proposed in [2] has a north-east drift in a subspace of D^+, while it should have a strictly north drift in order to follow the optimal trajectory with high probability.

Now we are ready to define a new measure, see also (41) in [10].

$$\begin{array}{lll}
\bar{\lambda}(x) = & \lambda e^{-\langle DW(x), v_0\rangle/2} N(x), & \text{if } x \in \bar{D}, \\
\bar{\mu}_i(x) = & \mu_i e^{-\langle DW(x), v_i\rangle/2} N(x), \quad i = 1, 2, & \text{if } x \in \bar{D}, \\
\bar{\lambda}^+(x) = & \lambda/(\lambda + \mu_1^+ + \mu_2) e^{-\langle DW(x), v_0\rangle/2} N^+(x), & \text{if } x \in \bar{D}^+, \\
\bar{\mu}_1^+(x) = & \mu_1^+/(\lambda + \mu_1^+ + \mu_2) e^{-\langle DW(x), v_1\rangle/2} N^+(x), & \text{if } x \in \bar{D}^+, \\
\bar{\mu}_2^+(x) = & \mu_2/(\lambda + \mu_1^+ + \mu_2) e^{-\langle DW(x), v_2\rangle/2} N^+(x), & \text{if } x \in \bar{D}^+.
\end{array} \tag{19}$$

Note that the functions $\tilde{\lambda}(x)$, etc. from the previous section are transition *rates*, while the functions $\bar{\lambda}(x)$, etc. are transition *probabilities* under the new measure (just as λ (resp. $\lambda/(\lambda + \mu_1^+ + \mu_2)$) is a transition probability under the original measure when $x \in \bar{D}$ (resp. $x \in \bar{D}^+$). The functions $N(x)$ and $N^+(x)$ provide the normalization such that the new transition probabilities sum up to 1. More precisely,

$$N(x) := \left[\lambda e^{-\langle DW(x), v_0\rangle/2} + \mu_1 e^{-\langle DW(x), v_1\rangle/2} + \mu_2 e^{-\langle DW(x), v_2\rangle/2}\right]^{-1}$$

and

$$N^+(x) := \left[\frac{\lambda e^{-\langle DW(x), v_0\rangle/2}}{\lambda + \mu_1^+ + \mu_2} + \frac{\mu_1^+ e^{-\langle DW(x), v_1\rangle/2}}{\lambda + \mu_1^+ + \mu_2} + \frac{\mu_2 e^{-\langle DW(x), v_2\rangle/2}}{\lambda + \mu_1^+ + \mu_2}\right]^{-1}.$$

These normalization functions are in fact closely related to the so-called Hamiltonians $\mathbb{H}(DW(x))$ and $\mathbb{H}_s(DW(x))$ in [2, 3]. In fact $\mathbb{H}(DW(x)) = 2\log N(x)$ and $\mathbb{H}_s(DW(x)) = 2\log N^+(x)$.

The new state-dependent measure in every state x is strongly dependent on the gradient of the function $W(x)$. To ease the exposition we express them as follows

$$DW(x) = \sum_{k=1}^{3} \rho_k(x) DW_k(x), \text{ where } \rho_k(x) = \frac{e^{-W_k(x)/\epsilon}}{\sum_{i=1}^{3} e^{-W_i(x)/\epsilon}}. \tag{20}$$

The gradients of auxiliary functions $W_i(x)$ have further representation

$$\begin{aligned} DW_1(x) &= 2\left(\log\frac{\lambda}{\tilde{\lambda}}, \log\frac{\tilde{\mu}_2}{\mu_2}\right), && \text{if } x \in \bar{D} \\ DW_1(x) &= 2\left(\log\frac{\lambda}{\tilde{\lambda}^+}, \log\frac{\tilde{\mu}_2^+}{\mu_2}\right), && \text{if } x \in \bar{D}^+ \\ DW_2(x) &= 2\left(\log\frac{\lambda}{\tilde{\lambda}}, 0\right), \\ DW_3(x) &= (0,0)\,. \end{aligned} \tag{21}$$

To end this section we give an elegant representation of the new measure in (19). Hereto we recall that the first two lines of (21) correspond to state independent measures $(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2)$ and $(\tilde{\lambda}^+, \tilde{\mu}_1^+, \tilde{\mu}_2^+)$, which are solutions to (9) and (10). The third line corresponds to the pair of state-independent measures $(\hat{\lambda}, \hat{\mu}_1, \hat{\mu}_2) := (\tilde{\lambda}, \mu_1\lambda/\tilde{\lambda}, \mu_2)$ and $(\hat{\lambda}^+, \hat{\mu}_1^+, \hat{\mu}_2^+) := (\tilde{\lambda}, \mu_1^+\lambda/\tilde{\lambda}, \mu_2)$, where $\tilde{\lambda}$ and $\tilde{\lambda}^+$ again are solutions to (9) and (10). The last line in (21) corresponds to the 'natural', i.e., unchanged measure. The new measure (19) can now be rewritten as

$$\begin{aligned} \bar{\lambda}(x) = &\quad \tilde{\lambda}^{\rho_1(x)}\hat{\lambda}^{\rho_2(x)}\lambda^{\rho_3(x)}M(x), && \text{if } x \in \bar{D}, \\ \bar{\mu}_1(x) = &\quad \tilde{\mu}_1^{\rho_1(x)}\hat{\mu}_1^{\rho_2(x)}\mu_1^{\rho_3(x)}M(x), && \text{if } x \in \bar{D}, \\ \bar{\mu}_2(x) = &\quad \tilde{\mu}_2^{\rho_1(x)}\hat{\mu}_2^{\rho_2(x)}\mu_2^{\rho_3(x)}M(x), && \text{if } x \in \bar{D}, \\ \bar{\lambda}^+(x) = &\quad (\tilde{\lambda}^+)^{\rho_1(x)}(\hat{\lambda}^+)^{\rho_2(x)}(\lambda)^{\rho_3(x)}M^+(x), && \text{if } x \in \bar{D}^+, \\ \bar{\mu}_1^+(x) = &\quad (\tilde{\mu}_1^+)^{\rho_1(x)}(\hat{\mu}_1^+)^{\rho_2(x)}(\mu_1^+)^{\rho_3(x)}M^+(x), && \text{if } x \in \bar{D}^+, \\ \bar{\mu}_2^+(x) = &\quad (\tilde{\mu}_2^+)^{\rho_1(x)}(\hat{\mu}_2^+)^{\rho_2(x)}(\mu_2)^{\rho_3(x)}M^+(x), && \text{if } x \in \bar{D}^+, \end{aligned} \tag{22}$$

where the weights $\rho_i(x)$ are defined in (20), and $M(x)$, $M^+(x)$ are normalization functions.

Remark 3.2. Here we briefly discuss the IS schemes for the rest of the cases, as will be presented in [8]. If the second buffer is always the bottleneck, the IS scheme is similar to the one presented here. However, when the first buffer is always the bottleneck the situation is more complicated. For most starting states x^s the IS scheme is again similar to the one presented in this section, but when x^s lies inside some subspace C_1, which includes the origin (see [10]), the measure consists of 'two parts'. That is, we have two different functions for $W_1(x)$ inside and outside of C_1, and consequently two new measures. Starting in $x^s \in C_1$, the typical sample path under the new measure moves to the south-east, and then, after hitting the boundary of C_1, to the north-west. This ensures that any sample path will follow the optimal trajectory with high probability.

4 Asymptotic Efficiency

This section is dedicated to analytic proof of the IS scheme presented in the previous section. We first give some lemmas. See [10] for most of the proofs; only Lemma 4.2 is different, but can be proved in a similar way.

Lemma 4.1. *The likelihood $L(A)$ of a path $A = (X_j, j = 0, \ldots, \sigma)$ satisfies*

$$\begin{aligned} \log L(A) &= \frac{B}{2}\sum_{j=0}^{\sigma-1}\langle DW(X_j), X_{j+1} - X_j\rangle \\ &+ \sum_{k=1}^{2}\frac{1}{2}\sum_{j=0}^{\sigma-1}\langle DW(X_j), v_k\rangle I_{\{X_j = X_{j+1} \in \partial_k\}} \\ &- \sum_{j=0}^{\sigma-1}\left(\log N(x) I_{\{X_j \in D\}} + \log N^+(x) I_{\{X_j \in D^+\}}\right). \end{aligned} \tag{23}$$

Lemma 4.2. *For any path $A = (X_j, j = 0, ..., \sigma)$ the first term in (23) satisfies*

$$\left| \frac{B}{2} \sum_{j=0}^{\sigma-1} \langle DW(X_j), X_{j+1} - X_j \rangle - \frac{B}{2}(W(X_\sigma) - W(X_0)) \right| \leq \frac{C}{B\epsilon}\sigma + R,$$

for sufficiently large $B\epsilon$, where C is some positive constant and R is the random error due to discontinuity of the function $W(x)$:

$$R = \frac{B}{2} \log \left(\frac{\tilde{\lambda}}{\tilde{\lambda}^+} \right) \left| \sum_{i=1}^{\sigma^+} (-1)^i (\kappa - \eta_i) \right|, \tag{24}$$

where $B\eta_i$ is the number of jobs in the first buffer prior to the i-th crossing of the slow-down threshold and σ^+ is the number of the slow-down threshold crossings up to time σ.

Lemma 4.3. *For any path $A = (X_j, j = 0, ..., \sigma)$ the second term in (23) has the following upper bound*

$$\sum_{k=1}^{2} \frac{1}{2} \sum_{j=0}^{\sigma-1} \langle DW(X_j), v_k \rangle I_{\{X_j = X_{j+1} \in \partial_k\}} \leq 2 \log \frac{\mu_2}{\lambda} e^{-\delta/\epsilon} \sigma.$$

Lemma 4.4. *For any $x \in D$ we have $\log N(x) \geq 0$, and for any $x \in D^+$ we have*

$$\log N^+(x) \geq -C^\star e^{-h/\epsilon}$$

for some positive, finite constants $C^\star$ and h.

Lemma 4.5. *For any sequence υ_B such that $\lim_{B\to\infty} \upsilon_B = 0$ and τ_B^s defined by (2), the following limit holds:*

$$\lim_{B\to\infty} \frac{1}{B} \log \mathbb{E}(e^{\upsilon_B \tau_B^s} | I(A_B^s) = 1) = 0.$$

We also need the following conjecture, which will be proved in [8] using large deviations methods. The intuition here is that as B grows large in (24), the random variable σ^+ will essentially not grow with B, while due the scaled position(s) where the threshold is crossed by the sample path will become close to the point where the most probable path crosses the threshold, i.e., the η_i will converge to κ.

Conjecture 4.6. *For any scaled sample path A_B^s we believe the following holds true*

$$\lim_{B\to\infty} \frac{1}{B} \log \mathbb{E}(e^R | I(A_B^s) = 1) = 0,$$

where the discontinuity error R is defined in Lemma 4.2.

Finally we make the same assumption as in [3, 9], which tells us how we should choose the values of ϵ and δ.

Assumption 4.7. *The parameters $\delta \equiv \delta_B$ and $\epsilon \equiv \epsilon_B$ are strictly positive and satisfy the following limit conditions as $B \to \infty$: (i) $\epsilon_B \to 0$, (ii) $\delta_B \to 0$, (iii) $B\epsilon_B \to \infty$, (iv) $\epsilon_B/\delta_B \to 0$.*

We are now ready to present the main result of the paper.

Theorem 4.8. *Under Assumption 4.7 and if Conjecture 4.6 holds, the new measure (22) based on (12) and (13) is asymptotically efficient.*

Proof. Here we provide the proof of the efficiency of the new measure defined by (19). At first we provide an upper bound for the log-likelihood expression in Lemma 4.1. An upper bound for the first term of this log-likelihood follows from Lemma 4.2 and the following inequalities

$$W(X_0) = W(x^s) \geq 2\gamma(x^s) - \epsilon \log(3) - 3\delta \quad \text{and} \quad W(X_{\tau_B^s}) \leq -\log \left(\frac{\tilde{\lambda}^+}{\lambda} \right) X_{1,\tau_B^s} - \delta \leq -\delta,$$

see [10] for more details. The upper bound for the first term itself is

$$\frac{B}{2}\sum_{j=0}^{\tau_B^s-1}\langle DW(X_j), X_{j+1}-X_j\rangle \le \frac{B}{2}(-2\gamma(x^s)+\eta(B))+\frac{C}{B\epsilon}\tau_B^s+R, \tag{25}$$

where C is some positive constant, $\eta(B)$ is such that $\lim_{B\to\infty}\eta(B)=0$, and R is the discontinuity error defined in Lemma 4.6.

Lemma 4.3 provides the upper bound for the second term of the expression in Lemma 4.1

$$\sum_{k=1}^{2}\frac{1}{2}\sum_{j=0}^{\tau_B^s-1}\langle DW(X_j), v_k\rangle I_{\{X_j=X_{j+1}\in\partial_k\}} \le 2\log\frac{\mu_2}{\lambda}e^{-\delta/\epsilon}\tau_B^s. \tag{26}$$

The last term of the log-likelihood expression can be bounded using Lemma 4.4

$$-\frac{1}{2}\sum_{j=0}^{\tau_B^s-1}\left(\log N(X_j)I_{\{X_j\in D\}}+\log N^+(X_j))I_{\{X_j\in D^+\}}\right) \le C^\star e^{-h/\epsilon}\tau_B^s, \tag{27}$$

where h and $C^\star$ are some positive constants. Combining (25), (26) and (27) we obtain a bound for the likelihood ratio (23) as in [10],

$$\log(L(A_B^s)) \le -B\gamma(x^s)+B\eta(B)+\chi(B)\tau_B^s+R,$$

where

$$\chi(B)=2\log\frac{\mu_2}{\lambda}e^{-\delta/\epsilon}+\frac{C}{\epsilon B}+C^\star e^{-h/\epsilon}\to 0 \text{ as } B\to\infty,$$

see Assumption 4.7. Now for any path A_B^s we have:

$$\begin{aligned}
&\frac{1}{B}\log\mathbb{E}\left[L(A_B^s)I(A_B^s)\right]=\frac{1}{B}\log(\mathbb{E}\left[L(A_B^s)|I(A_B^s)=1\right]\mathbb{P}\left[I(A_B^s)=1\right])\\
&\le\frac{1}{B}\log\left(\mathbb{E}\left[e^{-B\gamma(x^s)+B\eta(B)+\chi(B)\tau_B^s+R}|I(A_B^s)=1\right]p_B^s\right)\\
&=-\gamma(x^s)+\eta(B)+\frac{1}{B}\log\mathbb{E}\left[e^{\chi(B)\tau_B^s}|I(A_B^s)=1\right]+\frac{1}{B}\log\mathbb{E}\left[e^{R}|I(A_B^s)=1\right]+\frac{1}{B}\log p_B^s.
\end{aligned}$$

Applying Lemma 4.5 and Conjecture 4.6 to the third and fourth terms of this expression, and using Theorem 3.1 we conclude that:

$$\lim_{B\to\infty}\frac{1}{B}\log\mathbb{E}\left[L(A_B^s)I(A_B^s)\right]\le -2\gamma(x^s)=2\lim_{B\to\infty}\frac{1}{B}\log p_B^s,$$

which completes the proof. □

5 Numerical Results

In Tables 1 and 2 we present simulation results for two different parameter settings using the new measure defined in (22). Instead of performing a fixed number of simulation runs such as in much of the IS literature, we simulated until the relative error of the estimator reached the value of 10^{-2}. In the tables we present 95% confidence intervals for p_B^s, the number of needed replications (# runs), the used machine time in seconds, and the number of 'succesful' replications (# succ.), i.e. the number of runs that resulted in buffer overflow.

We compare several starting states x^s, three values of the overflow level B, and two values of ϵ; the value of δ was taken to be $\delta=-\frac{1}{3}\epsilon\log\epsilon$. Note that the starting states in Table 1 belong to B_1, B_2 and B_3 respectively; we only include results for starting states on the horizontal boundary, as these are more difficult to obtain than results for starting states in the interior. In Table 2 we only considered $x^s=(0,0)$ since for the other states the event of interest was not rare, and hence the results are not interesting. Also,

		$\epsilon = 0.01$				$\epsilon = 0.001$			
x^s	B	p_B^s	# succ.	# runs	time	p_B^s	# succ.	# runs	time
$(0,0)$	20	$3.79 \cdot 10^{-7} \pm 7.44 \cdot 10^{-9}$	18,565	41,985	10	$3.79 \cdot 10^{-7} \pm 7.44 \cdot 10^{-9}$	15,576	28,332	8
	50	$1.28 \cdot 10^{-16} \pm 2.50 \cdot 10^{-18}$	36,999	193,128	55	$1.28 \cdot 10^{-16} \pm 2.52 \cdot 10^{-18}$	33,542	58,332	45
	100	$3.48 \cdot 10^{-32} \pm 6.82 \cdot 10^{-34}$	66,473	1,097,097	230	$3.54 \cdot 10^{-32} \pm 6.95 \cdot 10^{-34}$	56,982	109,992	163
$(0.7B,0)$	20	$6.11 \cdot 10^{-3} \pm 1.19 \cdot 10^{-4}$	8,946	8,946	1	$6.12 \cdot 10^{-3} \pm 1.20 \cdot 10^{-4}$	8,946	8,946	1
	50	$3.79 \cdot 10^{-6} \pm 7.44 \cdot 10^{-8}$	24,969	24,969	10	$3.73 \cdot 10^{-6} \pm 7.32 \cdot 10^{-8}$	24,665	24,665	9
	100	$3.25 \cdot 10^{-11} \pm 6.37 \cdot 10^{-13}$	49,528	49,528	37	$3.28 \cdot 10^{-11} \pm 6.43 \cdot 10^{-13}$	51,365	51,365	36
$(1.5B,0)$	20	$5.18 \cdot 10^{-1} \pm 1.01 \cdot 10^{-2}$	11,287	12,888	< 1				
	50	$1.35 \cdot 10^{-1} \pm 2.65 \cdot 10^{-3}$	66,997	77,942	2				
	100	$1.05 \cdot 10^{-2} \pm 2.05 \cdot 10^{-4}$	316,351	367,327	21				

Table 1: Simulation results for $\theta = 0.8$ and $(\lambda, \mu_1, \mu_1^+, \mu_2) = (0.1, 0.7, 0.15, 0.2)$

		$\epsilon = 0.01$				$\epsilon = 0.001$			
x^s	B	p_B^s	# succ.	# runs	time	p_B^s	# succ.	# runs	time
$(0,0)$	20	$5.62 \cdot 10^{-2} \pm 1.11 \cdot 10^{-4}$	33,371	98,230	2	$5.63 \cdot 10^{-2} \pm 1.11 \cdot 10^{-4}$	39,496	91,596	2
	50	$1.18 \cdot 10^{-3} \pm 2.31 \cdot 10^{-5}$	99,116	295,633	19	$1.19 \cdot 10^{-3} \pm 2.33 \cdot 10^{-5}$	99,567	241,332	18
	100	$1.63 \cdot 10^{-6} \pm 3.19 \cdot 10^{-8}$	143,194	382,120	55	$1.63 \cdot 10^{-6} \pm 3.21 \cdot 10^{-8}$	128,864	320,120	49

Table 2: Simulation results for $\theta = 0.8$ and $(\lambda, \mu_1, \mu_1^+, \mu_2) = (0.3, 0.36, 0.32, 0.34)$.

for $x^s = (1.5B, 0)$ in Table 1 we omitted results for $\epsilon = 0.001$ as these were indistinguishable from those with $\epsilon = 0.01$. Clearly, the IS scheme provides fast and reliable estimates. In some cases, especially when B grows large, the running times may be sensitive to the choice of ϵ and δ.

We also performed a few straightforward simulations (i.e., without IS) for comparison, using the same relative error of 10^{-2}. For the parameter settings of Table 1 with $B = 20$, this took 4521 seconds ($\pm 5 \cdot 10^9$ runs) for $x^s = (0,0)$, and 16 seconds ($\pm 2 \cdot 10^6$ runs) for $x^s = (0.7B, 0)$. In the settings of Table 2 with $B = 50$ it took 118 seconds ($\pm 10^7$ runs).

To enable comparison with the state-independent scheme in [6] and the state-dependent scheme in [2], we also fixed the number of runs to be 10^6 and compared the relative errors, see Table 3. Here, $x^s = (0,0)$, $\theta = 0.8$, and in the state-dependent schemes $\epsilon = 0.03/\sqrt{B}$ and $\delta = -\epsilon \log \epsilon$. As can be expected, both state-dependent schemes provide good estimates, but the performance of the state-independent scheme strongly depends on the parameters.

Finally, we present some results for the cases in which either the first queue is always the bottleneck (see left part of Table 4) or the second queue is always the bottleneck (right part of Table 4). In both cases we fixed the relative error, but note that we took it to be 0.05 instead of 0.01 when the first queue is the bottleneck. The choice of $x^s = (0.35B, 0)$ in the case where $(\lambda, \mu_1, \mu_1^+, \mu_2) = (0.25, 0.35, 0.28, 0.4)$ corresponds to the point where the optimal path from $(0,0)$ to ∂_e leaves the horizontal axis. For the case $(\lambda, \mu_1, \mu_1^+, \mu_2) = (0.3, 0.36, 0.35, 0.34)$ we did not include results for $x^s = (3B, 0)$, since the 'new' measure here coincides with the old measure, i.e. it is optimal to use straightforward simulations here.

6 Conclusions

In this paper we constructed an asymptotically efficient IS scheme for estimating the probability of overflow in the second buffer of a slow-down network. In previous work [10] we also proposed an asymptotically efficient scheme, but there we refrained from including numerical experiments, as we felt that those form a topic of research in their own right. This is due to the fact that it is still a rather nontrivial step from an asymptotically efficient procedure, as the ones presented in [10], to an actual, efficient implementation of the algorithm. It is noted that several aspects, which are not captured by the notion of asymptotic efficiency,

	$(\lambda, \mu_1, \mu_1^+, \mu_2) = (0.1, 0.7, 0.15, 0.2)$			$(\lambda, \mu_1, \mu_1^+, \mu_2) = (0.3, 0.36, 0.32, 0.34)$		
B	st.-ind., [6]	st.-dep., [2]	current	st.-ind., [6]	st.-dep., [2]	current
20	$1.49 \cdot 10^{-3}$	$2.63 \cdot 10^{-3}$	$3.54 \cdot 10^{-3}$	$0.92 \cdot 10^{-3}$	$5.30 \cdot 10^{-3}$	$6.00 \cdot 10^{-3}$
50	$2.06 \cdot 10^{-3}$	$7.87 \cdot 10^{-3}$	$8.00 \cdot 10^{-3}$	$12.50 \cdot 10^{-3}$	$8.40 \cdot 10^{-3}$	$11.00 \cdot 10^{-3}$
100	$2.75 \cdot 10^{-3}$	$19.71 \cdot 10^{-3}$	$17.01 \cdot 10^{-3}$	$39.69 \cdot 10^{-3}$	$12.20 \cdot 10^{-3}$	$11.00 \cdot 10^{-3}$

Table 3: Comparison of relative errors for three IS schemes

		$(\lambda, \mu_1, \mu_1^+, \mu_2) = (0.25, 0.35, 0.28, 0.4)$, $RE = 0.05$				$(\lambda, \mu_1, \mu_1^+, \mu_2) = (0.3, 0.36, 0.35, 0.34)$, $RE = 0.01$			
x^s	B	p_B^s	# succ.	# runs	time	p_B^s	# succ.	# runs	time
$(0,0)$	20	$1.11 \cdot 10^{-4} \pm 1.09 \cdot 10^{-5}$	45,685	83,436	2	$5.86 \cdot 10^{-2} \pm 1.44 \cdot 10^{-3}$	32,283	76,169	2
	50	$3.43 \cdot 10^{-11} \pm 3.36 \cdot 10^{-12}$	79,901	148,256	7	$1.42 \cdot 10^{-3} \pm 2.79 \cdot 10^{-5}$	112,128	269,968	21
	100	$5.72 \cdot 10^{-22} \pm 5.60 \cdot 10^{-23}$	235,502	439,006	42	$2.64 \cdot 10^{-6} \pm 5.18 \cdot 10^{-8}$	275,112	661,247	121
$(0.35B, 0)$	20	$6.18 \cdot 10^{-4} \pm 6.06 \cdot 10^{-5}$	38,190	40,333	1	$2.11 \cdot 10^{-1} \pm 4.15 \cdot 10^{-3}$	37,178	42,163	2
	50	$2.56 \cdot 10^{-9} \pm 2.50 \cdot 10^{-10}$	92,005	92,234	5	$1.50 \cdot 10^{-2} \pm 2.95 \cdot 10^{-4}$	82,133	92,301	15
	100	$4.64 \cdot 10^{-18} \pm 4.55 \cdot 10^{-19}$	206,100	206,182	25	$2.15 \cdot 10^{-4} \pm 4.21 \cdot 10^{-6}$	114,694	124,994	35
$(3B, 0)$	20	$1.62 \cdot 10^{-1} \pm 1.58 \cdot 10^{-2}$	21,106	23,496	1				
	50	$1.15 \cdot 10^{-3} \pm 1.13 \cdot 10^{-4}$	43,378	52,840	7				
	100	$1.90 \cdot 10^{-7} \pm 1.86 \cdot 10^{-8}$	78,229	91,231	25				

Table 4: Simulation results for non-shifting bottleneck cases ($\theta = 0.8$)

play a crucial role: it matters for instance very much whether a new measure requires computation of new transition rates 'on the fly', or whether these can be precomputed. These issues have been taken into account in the present paper.

The major advantage of the scheme presented here over the earlier algorithm in [10] lies in the low complexity of the resulting procedure. The change-of-measure is virtually state-independent, and hence the computation of the new transition rates hardly contributes to the time needed to obtain a reliable estimate. In [2] another asymptotically efficient IS scheme was proposed. That scheme has a similar complexity as the one described in this paper. A substantial advantage of our IS scheme is that it can be applied to any starting state (i.e., not just the origin).

References

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

[2] P. Dupuis, K. Leder, and H. Wang. Large deviations and importance sampling for a tandem network with slow-down. *Queueing Systems: Theory and Applications*, 57(2-3):71–83, 2007.

[3] P. Dupuis, A.D. Sezer, and H. Wang. Dynamic importance sampling for queueing networks. *Annals of Applied Probability*, 17(4):1306–1346, 2007.

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

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

[6] D.I. Miretskiy, W.R.W. Scheinhardt, and M.R.H. Mandjes. Efficient simulation of a tandem queue with server slow-down. *Simulation*, 83(11):751–767, 2007.

[7] D.I. Miretskiy, W.R.W. Scheinhardt, and M.R.H. Mandjes. Tandem queue with server slow-down. *ACM Sigmetrics Performance Evaluation Review*, 35(3):51–52, 2007.

[8] D.I. Miretskiy, W.R.W. Scheinhardt, and M.R.H. Mandjes. Stability and importance sampling for a Slow-down tandem queue. *Unpublished manuscript*, 2008.

[9] D.I. Miretskiy, W.R.W. Scheinhardt, and M.R.H. Mandjes. State-dependent importance sampling for a Jackson tandem network. *Submitted*, 2008. See also Memorandum 1867, Dept. of Applied Mathematics, University of Twente, *URL: http://eprints.eemcs.utwente.nl/12734/*.

[10] D.I. Miretskiy, W.R.W. Scheinhardt, and M.R.H. Mandjes. State-dependent importance sampling for a Slow-down tandem queue. *Submitted*, 2008. See also REPORT PNA-R0811, CWI, *URL: http://ftp.cwi.nl/CWIreports/PNA/PNA-R0811.pdf/*.

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

[12] J. S. Sadowsky. Large deviations theory and efficient simulation of excessive backlogs in a $GI/GI/m$ queue. *IEEE Transactions on Automatic Control*, 36(12):1383–1394, 1991.

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

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