TI 2001-050/4
Tinbergen Institute Discussion Paper

Fast Simulation of a Queue fed by a Superposition of Many (Heavy-Tailed) Sources

Nam Kyoo Boots[1]
Michel Mandjes[2]

[1] Department of Econometrics, Faculty of Economics and Business Administration, Vrije Universiteit Amsterdam, and Tinbergen Institute,
[2] Bell Laboratories/Lucent Technologies, Murray Hill, NJ, USA.

Tinbergen Institute
The Tinbergen Institute is the institute for economic research of the Erasmus Universiteit Rotterdam, Universiteit van Amsterdam and Vrije Universiteit Amsterdam.

Tinbergen Institute Amsterdam
Keizersgracht 482
1017 EG Amsterdam
The Netherlands
Tel.: +31.(0)20.5513500
Fax: +31.(0)20.5513555

Tinbergen Institute Rotterdam
Burg. Oudlaan 50
3062 PA Rotterdam
The Netherlands
Tel.: +31.(0)10.4088900
Fax: +31.(0)10.4089031

Fast simulation of a queue fed by a superposition of many (heavy-tailed) sources

Nam Kyoo Boots* and Michel Mandjes †

Abstract

We consider a queue fed by a large number, say n, of on-off sources with generally distributed on- and off-times. The queueing resources are scaled by n: the buffer is $B \equiv nb$ and link rate is $C \equiv nc$. The model is versatile: it allows us to model both *long range dependent* traffic (by using heavy-tailed distributed on-periods) and *short range dependent* traffic (by using light-tailed on-periods). A crucial performance metric in this model is the steady-state buffer overflow probability.

This overflow probability decays exponentially in the number of sources n. Therefore, if the number of sources grows large, naive simulation is too time-consuming, and we have to use fast simulation techniques instead. Due to the exponential decay (in n), importance sampling with an exponential change of measure essentially goes through, irrespective of the on-times being heavy-tailed or light-tailed. An asymptotically optimal change of measure is found by using large deviations arguments. Notably, the change of measure is not constant during the simulation run, which is essentially different from many other studies (usually relying on large buffer asymptotics).

We provide numerical examples to show that the resulting importance sampling procedure indeed improves considerably over naive simulation. We present some accelerations. Finally, we give short comments on the influence of the shape of the distributions on the loss probability, and we describe the limitations of our technique.

Key words: long-range dependence, importance sampling, queueing theory, large deviations asymptotics, buffer overflow, heavy-tailed random variables

*Department of Econometrics, Vrije Universiteit, De Boelelaan 1105, Room 1A22, 1081 HV Amsterdam, the Netherlands. Email: nboots@econ.vu.nl

†Bell Laboratories/Lucent Technologies, 600 Mountain Ave., Room 2C361, P.O. Box 636, Murray Hill, NJ 07974-0636, USA. Email: michel@research.bell-labs.com. The author is currently also with CWI, Amsterdam, the Netherlands, and Faculty of Applied Mathematics, University of Twente, the Netherlands.

1 Introduction

In communication networks it is important to predict the performance of a network element fed by a given set of traffic sources. It eases the task of doing adequate resource allocation, admission control, and dimensioning of buffers and link rates. A particularly interesting issue is the impact of the traffic characteristics on the performance. This matter attracted renewed attention after the discovery that a wide variety of traffic types show *long range dependence* (LRD), i.e., burstiness on a wide variety of time scales [13]. A LRD traffic stream is characterized by a correlation function of which the decay is slower than exponential in time. This is in stark contrast with short range dependent (SRD) input, where the correlation decays exponentially.
A large body of work on short range dependent models were already available. Particularly, accurate methods for the computation of loss and delay performance of queues with SRD input were developed, see for instance the seminal work [1]. For LRD sources, this queueing analysis could clearly not be used anymore. Assuming that network traffic could be long-range dependent, the logical question is: does this extreme burstiness significantly degrade the performance (usually measured in terms of packet loss and delay)?

Performance evaluation of queues with LRD and SRD traffic. A partial answer is given in the studies of Ryu and Elwalid [21], Heyman and Lakshman [10], and Grossglauser and Bolot [9]. They argue that in realistic scenarios and for stringent delay requirements (i.e., buffers typically not very large), only short term correlations play a role, and hence the better analyzed models based on SRD traffic can be reused. To assess this issue in greater detail, we use the versatile traffic model of on-off sources. These sources alternate between transmitting at a certain peak rate (commonly called a 'burst') and being silent. The activity and silence periods are random variables. The sources feed into a queue with constant capacity. The versatility of the model is reflected by the fact that it covers both LRD and SRD traffic, by using specific choices of the burst and silence distributions. The aggregate of the sources generates LRD traffic if the burst size has a heavy-tailed distribution [13], whereas light-tailed on-periods lead to SRD traffic. In models with heavy-tailed on-times hardly any analytical results exist. The known results describe asymptotics of the loss probability for large values of the buffer size; there are no results that explicitly give the entire buffer content distribution. From a practical point of view, the regime of *large buffers* is probably not the most relevant, as many (real-time) applications require some delay bound. For these applications a more relevant asymptotic regime could be the one with *many sources*, since in practice, many relatively small sources will share the network elements.
Roughly the model is as follows. There are a large number, say n, of on-off sources feeding into the queue. The resources buffer and bandwidth are scaled accordingly: buffer $B \equiv nb$, and link rate $C \equiv nc$. In this regime there are a number of strong *large deviations* results available [4, 14, 17]. Notably the probability of overflow p_n decays exponentially in the number of sources n; the corresponding decay rate is the solution of a variational problem. Here for ease the sources are assumed to be independent and statistically identical.

An obvious drawback of this large deviations approach is that some of the above mentioned many-sources asymptotics [4, 17] are *rough*, in that only the exponential decay rate, say I is derived. The 'subexponential part' $f(n)$ (with $\log f(n) = o(n)$, where $n \to \infty$) of the expansion is not found. Therefore, the resulting naive estimate $p_n = \exp(-nI)$ is not always accurate, even if the number of sources is large. In other words: the asymptotics of the log of the overflow probability are found, rather than the asymptotics of the probability itself. The results in [14] are more precise: there a (subexponential) function $f(\cdot)$ is provided such that $p_n f(n) \exp(nI) \to 1$. However, for given n, still the error made by approximating $p_n \approx \exp(-nI)/f(n)$ is not known.

Simulation. A natural alternative to exact calculations and asymptotic approximations is stochastic simulation. However, the probabilities involved are typically small, which makes them hard to estimate: consequently a considerable amount of simulation effort is required to obtain reliable estimates. This explains the interest in variance reduction techniques, commonly known as 'fast simulation'.
A commonly used fast simulation technique is *importance sampling*, which is often based on an *exponential change of measure* (also called *exponential twisting*). This technique can be explained easily by considering a random walk $(\xi_i)_{i\in\mathbb{N}}$, where the ξ_i are i.i.d. with density g. Assume a negative drift: $\mathbb{E}\xi_i < 0$. We are interested in the probability that this random walk ever exceeds level x, say $\mathbb{P}(x)$. Because of the negative drift $\mathbb{P}(x)$ will be small, particularly for large x, and naive (direct) simulation will typically be slow. The idea of importance sampling based on an exponential change of measure is to replace the density g by an exponentially twisted density $g_\theta(x) = g(x)\exp(\theta x)/M_\xi(\theta)$, where $M_\xi(\theta)$ is the moment generating function $\mathbb{E}\exp(\theta\xi_i)$. The tilting parameter θ has to be chosen positive, and large enough to make sure that the mean under the new density is positive. To compensate for the change of measure (and the increased likelihood of the rare event), the simulation output has to be adapted by using likelihood ratios. Details on this procedure are found in [11].
It is emphasized that the above exponential change of measure does not work for heavy-tailed $(\xi_i)_{i\in\mathbb{N}}$. The reason is that for heavy-tailed ξ_i the normalizing constant $M_\xi(\theta)$ is infinite for all positive θ and thus exponential twisting is infeasible. Similarly for on-off sources with heavy-tailed on-times, it can be argued that we cannot construct an exponential twisting of burst and silence distributions. A general statement is: as long as the loss probability is exponentially decaying in the buffer size B, a variant of the above twisting procedure works, if there is subexponential decay it does not (like in the case of heavy-tailed on-times [16]). This makes the problem of importance sampling with heavy-tailed distributions *hard*, although some partial results are available [2, 3].

Importance sampling in the many-sources domain. However, in the regime of many sources we *do* have an exponential decay, albeit in the number of sources n rather than in the buffer size B. As we show in this paper, this implies that exponential twisting is possible, since it does not involve exponential twisting of the (possibly heavy-tailed) on-times. However, the resulting change of measure is more complicated than in the traditional random walk type of models: it is not constant during the path to overflow. This is the essential difference with exponential twisting in the large buffer domain [11, 12, 18, 20].

The choice of our change of measure results from large deviation theory. We show that the average path under this measure equals the optimal path to overflow identified by Wischik [27]. We are also able to bound the variance of the resulting estimator such that the number of simulation replications (required to get an estimate with predefined accuracy) grows subexponentially in n, whereas p_n decays essentially exponentially.
The main contributions of this paper are twofold. First, we propose an efficient simulation technique to estimate the overflow probability in a queue with n on-off sources. This model is generic in that it captures both LRD and SRD scenarios. Second, our work is among the first papers that describes importance sampling for a model with heavy-tailed on-off sources, cf. [2, 3]. Also fast simulation in the many sources regime is relatively new; in [19] this is considered in a much more restrictive model.

The organization of this paper is as follows. Section 2 presents the model and some preliminaries. Then Section 3 gives our importance sampling procedure, which is evaluated in Section 4. Section 5 gives some considerations on the implementation, simulation results, and discusses the limitation of our recipe. Section 6 contains some remarks and outlook.

2 Model and preliminary results

This section prepares the exposition of our fast simulation procedure (Section 3), and its theoretical assessment (Section 4). In Subsection 2.1 we present the model. Subsection 2.2 provides a number of large deviation asymptotics (both the decay rate of the loss probability and sample path large deviations). These results are needed to construct the importance sampling technique. A scheme for the numerical computation of the decay rate and the optimal path to a buffer overflow are given in Subsection 2.3.

2.1 Model

Traffic. We consider n i.i.d. on-off sources feeding into a buffered resource. This resource is modeled as a queue with infinite buffer size, drained at a constant rate C. The traffic rate of each source alternates between a peak rate, say 1, and 0. The activity periods constitute an i.i.d. sequence of random variables, each of them distributed as a $\mathbb{N}$-valued random variable A. The silence periods are also an i.i.d. sequence, distributed as a $\mathbb{N}$-valued random variables S. Both sequences are mutually independent. Define also

$$A(k) := \text{Traffic generated by a single source in steady state in a time interval of } k \text{ time slots.}$$

Later in our analysis we need the following assumption on the on- and off-times:

Assumption 2.1 *The random variables A and S are such that $\mathbb{E}A^{1+\zeta} < \infty$ (for some positive ζ) and $\mathbb{E}S < \infty$.*

This assumption has several implications – for details we refer to Section 2.1 of [8]. In the first place, the fact that both $\mathbb{E}A$ and $\mathbb{E}S$ are finite ensures that the long-run fraction of time the source spends in the on-state is

$$p := \frac{\mathbb{E}A}{\mathbb{E}A + \mathbb{E}S},$$

and the fraction spent in the off-state is its complement $1 - p$. Also, the *residual* activity period $A^\star$ is well-defined: conditioned on the process being in the on-state, $A^\star$ has distribution

$$F_{A^\star}(k) := \mathbb{P}(A^\star > k) = \frac{1}{\mathbb{E}A} \sum_{\ell=k}^{\infty} \mathbb{P}(A > \ell);$$

the distribution of $S^\star$ is given analogously.

Performance measure. We are interested in the steady-state probability of the buffer content exceeding level B. Hence, we follow a conventional approach in inferring finite-buffer performance from an infinite-buffer model with a threshold at the finite buffer size. As emphasized in the introduction, we focus on the asymptotic regime in which the number of sources grows large and the resources are scaled accordingly [25]. To be more precise, we rescale the resources by the number of sources: $C \equiv nc$ and $B \equiv nb$. This scaling was first introduced by Weiss [25] and has proven to be very powerful, see e.g. [4, 6, 22]. It is assumed that the system is stable and non-trivial:

$$\rho := p < c < 1.$$

In the above defined scaled model we define

$$p_n := \text{steady-state probability that the buffer content exceeds level } nb.$$

Throughout this paper we use the representation

$$p_n = \mathbb{P}\left(\exists k \in \mathbb{N} : A_n(k) - nck > nb\right), \tag{1}$$

where $A_n(k)$ denotes the amount of traffic generated in $\{1, \ldots, k\}$ by the aggregate of the n sources. In this paper, our goal is to estimate this probability by simulation, with some predefined accuracy. Since we use representation (1) for the buffer overflow probability, we simulate the process $\{A_n(k) - nck, k \in \mathbb{N}\}$ which we allow to take any value in the interval $(-\infty, B]$.

Dependence structures. The model presented above offers a high degree of versatility, as it allows us to model a broad variety of dependence structures. Importantly, it covers both short-range dependent and long-range dependent input. To model SRD traffic input streams, we could use light-tailed on-periods. We call a random variable light-tailed if its distribution function has a tail that decays at an exponential or faster rate. We call this class $\mathcal{E}$. Examples are the *Exponential* distribution, or, more generally, the class of *phase-type* distributions.
To model traffic with a dependence structure that ranges over a longer time, we use heavy-tailed on-periods. Examples we consider in this work are the *Pareto* distribution and the *Weibull* distribution.

Notably, in [26] it is shown that the superposition of many on-off sources with Pareto sojourn-times converges to fractional Brownian motion (with an appropriate scaling of the number of sources as well as time), which exhibits the desired LRD features. The heavy-tailed distributions that we use in this paper are in the class of subexponential distributions $\mathcal{S}$:

Definition 2.2 *Suppose X_1 and X_2 are i.i.d. copies of the random variable X. If*

$$\lim_{x\to\infty} \frac{\mathbb{P}(X_1 + X_2 > x)}{\mathbb{P}(X_1 > x)} = 2,$$

the X is said to be subexponential. We write: $X \in \mathcal{S}$.

2.2 Large deviation results for the loss probability

This subsection focuses on the calculation of rough characteristics of the overflow probability p_n. Later in this paper we use these asymptotics to find the change of measure of our importance sampling procedure, and to establish a number of structural properties of the resulting simulation method. We present two theorems: Theorem 2.3 first describes the asymptotics of p_n, Theorem 2.5 describes the system's most likely way to develop from an empty queue towards the rare event of buffer overflow.

For any value of the buffer size b, under fairly general conditions, the probability p_n *decays exponentially in n.* In Theorem 2.3 below it is stated how to compute the corresponding exponential *decay rate*

$$I := -\lim_{n\to\infty} \frac{1}{n} \log p_n,$$

which implies the following rough approximation:

$$p_n \approx e^{-nI}, \ n \text{ large.}$$

Theorem 2.3 has a long history. Botvich and Duffield [4] proved it under very mild conditions on the sources, whereas related results were derived in [6, 22]. An improvement was made by Likhanov and Mazumdar [14]. The version that we use in this paper follows relatively directly from the result in [14].

Theorem 2.3 *Under Assumption 2.1, and for $A^\star \in \{\mathcal{E} \cup \mathcal{S}\}$,*

$$I = \inf_{k\in\mathbb{N}} \sup_{\theta} \left(\theta(b + ck) - \log \mathbb{E}e^{\theta A(k)} \right). \tag{2}$$

Proof. As the proof is given in Mandjes and Borst [16], we limit ourselves to a short sketch. First define

$$I_k := \sup_{\theta} \left(\theta(b + ck) - \log \mathbb{E}e^{\theta A(k)} \right).$$

- Likhanov and Mazumdar [14] show that decay rate (2) applies if

$$\liminf_{k\to\infty} \frac{I_k}{\log k} > 0. \tag{3}$$

  Or, in other words, if there is an $\alpha > 0$ such that $I_k > \alpha \log k$ eventually.

- Proposition 3.3 of [16] proves that $\mathbb{E}A^{1+\zeta} < \infty$ implies, both for $A \in \mathcal{S}$ and $A \in \mathcal{E}$, that for any $\epsilon \in (0, 1-p)$ there is an $\alpha > 0$ such that for k large enough

  $$\mathbb{P}(A(k) > k(p+\epsilon)) < k^{-\alpha}.$$

  In [14] it is shown that this implies (3). □

A corollary that follows from the proof of Theorem 2.3 is the following.

Corollary 2.4 *Under Assumption 2.1, and for $A^\star \in \{\mathcal{E} \cup \mathcal{S}\}$, there is an $\alpha > 0$ and a $k_{\min} \in \mathbb{N}$ such that for $k \geq k_{\min}$,*

$$I_k > \alpha \log k. \tag{4}$$

As is well-known from the theory of importance sampling, an optimal (i.e., zero variance) estimator for the rare event probability is obtained if we would sample from the unknown distribution of the stochastic process *conditioned on the occurrence of the rare event* [11]. In this paper we use importance sampling techniques based on large deviations results to mimic this conditional distribution.
Importantly, decay rate (2) implicitly provides us the time-scale of a typical path to overflow: the optimizing k, say $k^\star$, is the 'most likely' duration of the busy period preceding overflow, *given* overflow occurs. The relevance of this time-scale is clear: To obtain variance reduction, the importance sampling parameters should be chosen such that they 'mimic' the system's 'most likely path to overflow'.
To achieve this, clearly knowledge of time-scale $k^\star$ is not enough; more detailed knowledge of that 'most likely path to overflow' is required. This path, say f, is given by a sample path large deviation result by Wischik [27]. Of course, f reaches overflow at time $k^\star$.
Let us state Wischik's [27] result a little more precisely. Given that, for some k, $A_n(k)/n - ck$ exceeds b, Wischik [27] essentially proves that any deviation (according to some specific metric) of the process $(A_n(k)/n)_{k\in\mathbb{N}}$ from the most likely path f (given below in Theorem 2.5) has an exponentially decreasing probability (in n).

Theorem 2.5 *The most likely path to overflow is given by*

$$f(j) = \frac{\mathbb{E}A(j)\exp(\theta_{k^\star}A(k^\star))}{\mathbb{E}\exp(\theta_{k^\star}A(k^\star))}, \tag{5}$$

$j \in \mathbb{N}$. Specifically, $f(k^\star) = b + ck^\star$.

As said, we may interpret $k^\star$ as the 'most likely epoch of overflow', as it turns out to be the first time $f(k) - ck$ attains level b. In fact, the buffer starts to fill at time 1, in $\{1, \ldots, k^\star\}$ the buffer level increases to level b, whereas after $k^\star$ the net input rate is negative.
The exact statement of Theorem 2.5 is found in [27]. Notably, a number of assumptions on the input traffic have to be fulfilled for this statement to hold. For a discussion on these we refer to Section 2 of [27]. It is noted that they are stronger than our Assumption 2.1.

2.3 Calculation of the decay rate and the optimal path to overflow

As we saw, Theorems 2.3 and 2.5 present analytic expressions of both the decay rate I and the most likely path to overflow f. In our fast simulation procedure we need the numerical value of the decay rate. In this subsection we indicate how this can be found. We also indicate how we can compute the most likely path to overflow numerically.
Abbreviate

$$\begin{array}{rclcrcl} a_k & := & \mathbb{P}(A=k); & & s_k & := & \mathbb{P}(S=k); \\ a_k^\star & := & \mathbb{P}(A^\star=k); & & s_k^\star & := & \mathbb{P}(S^\star=k). \end{array}$$

First we point out how to compute moment generating function $\mathbb{E}\exp(\theta A(k))$. This can be done recursively, as follows. Clearly, in evident notation,

$$\mathbb{E}e^{\theta A(k)} = p \cdot \mathbb{E}_{A^\star} e^{\theta A(k)} + (1-p) \cdot \mathbb{E}_{S^\star} e^{\theta A(k)}.$$

Both terms can be evaluated as follows:

$$\mathbb{E}_{A^\star} e^{\theta A(k)} = \sum_{i=1}^{k-1} a_i^\star e^{\theta i} \mathbb{E}_S e^{\theta A(k-i)} + \sum_{i=k}^{\infty} a_i^\star e^{\theta k}, \;\; \mathbb{E}_{S^\star} e^{\theta A(k)} = \sum_{i=1}^{k-1} s_i^\star \mathbb{E}_A e^{\theta A(k-i)} + \sum_{i=k}^{\infty} s_i^\star,$$

where

$$\mathbb{E}_A e^{\theta A(j)} = \sum_{i=1}^{j-1} a_i e^{\theta i} \mathbb{E}_S e^{\theta A(j-i)} + \sum_{i=j}^{\infty} a_i e^{\theta j}, \;\; \mathbb{E}_S e^{\theta A(j)} = \sum_{i=1}^{j-1} s_i \mathbb{E}_A e^{\theta A(j-i)} + \sum_{i=j}^{\infty} s_i.$$

It follows directly that $\mathbb{E}e^{\theta A(\ell)}$ $(\ell = 1, \ldots, k-1)$ have to be computed to obtain $\mathbb{E}e^{\theta A(k)}$. Now it is not hard to see that the complexity of computing $\mathbb{E}e^{\theta A(k)}$ is $O(\sum_{\ell=1}^{k} O(\ell)) = O(k^2)$. In Section 3 it is explained that we need to compute this moment generating function for $k = 1$ to $k = k_0$, for some fixed positive integer k_0 (larger than $k^\star$).
Having a procedure to find the moment generating function $\mathbb{E}e^{\theta A(k)}$, it is not hard to find I_k, because of the convexity in θ; we call the optimizing argument θ_k. To find I, we compute the infimum over k.

In order to compute the optimal path to overflow (5), we need to compute $\mathbb{E}A(\ell)\exp(\theta A(k^\star))$ for $\ell = 1, \ldots, k^\star$. This can also be done recursively as follows:

$$\mathbb{E}A(\ell)e^{\theta A(k)} = p \cdot \mathbb{E}_{A^\star} A(\ell) e^{\theta A(k)} + (1-p) \cdot \mathbb{E}_{S^\star} A(\ell) e^{\theta A(k)}.$$

Both terms can be evaluated as follows:

$$\mathbb{E}_{A^\star} A(\ell) e^{\theta A(k)} = \sum_{i=1}^{\ell-1} a_i^\star e^{\theta i} \left[i\mathbb{E}_S e^{\theta A(k-i)} + \mathbb{E}_S A(\ell-i) e^{\theta A(k-i)} \right] + \ell \sum_{i=\ell}^{k-1} a_i^\star e^{\theta i} \mathbb{E}_S e^{\theta A(k-i)} + \ell \sum_{i=k}^{\infty} a_i^\star e^{\theta k},$$

$$\mathbb{E}_{S^\star} A(\ell) e^{\theta A(k)} = \sum_{i=1}^{\ell-1} s_i^\star \mathbb{E}_A A(l-i) e^{\theta A(k-i)},$$

where

$$\mathbb{E}_A A(\ell)e^{\theta A(j)} = \sum_{i=1}^{\ell-1} a_i e^{\theta i}\left[\mathbb{E}_S A(l-i)e^{\theta A(j-i)} + i\mathbb{E}_S e^{\theta A(j-i)}\right] + \ell\sum_{i=\ell}^{j-1} a_i e^{\theta i}\mathbb{E}_S e^{\theta A(j-i)} + \ell\sum_{i=j}^{\infty} a_i e^{\theta j},$$

$$\mathbb{E}_S A(\ell)e^{\theta A(k)} = \sum_{i=1}^{\ell-1} s_i \mathbb{E}_A A(j-i)e^{\theta A(j-i)}.$$

3 Fast simulation procedure – importance sampling

This section describes the importance sampling procedure. In Section 3.1 we review the general framework of rare event simulation and importance sampling. Then we formalize our algorithm in Section 3.2. Section 3.3 presents the required change of measure.

3.1 Rare event simulation and importance sampling

Let U_n be the event of a buffer overflow, i.e., $p_n = \mathbb{P}(U_n)$ with

$$U_n = \{\exists k \in \mathbb{N} : A_n(k) - nck > nb\}.$$

Since we assume many sources n and because $p_n \downarrow 0$ $(n \to \infty)$ (cf. Theorem 2), we are in setting of *rare event simulation.* Rare event simulation has an intrinsic problem, as will be explained below.

Infeasibility of naive methods. Let $\hat{p}_n$ be an estimator of p_n. In order to guarantee its accuracy, one aims for a small *relative error* (RE), defined as the ratio of the standard deviation of $\hat{p}_n$ and the estimated quantity p_n.

Requirement 3.1 *The relative error RE of the simulation experiment should be below δ.*

Naive simulation, i.e., just simulating sample paths and estimating p_n by the fraction of sample paths that lie in U_n, is *not* efficient: with N_n defined as the number of simulation replications, then [24, page 335-336]

$$N_n \sim \frac{1}{\delta^2 \cdot p_n}.$$

In other words, the number of samples needed is inversely proportional to the probability to be estimated. Consequently, since the buffer overflow probability decays exponentially (in n), N_n blows up at an exponential rate (keeping the relative error RE fixed). This explains why naive simulation is not a feasible method for estimating rare events. Clearly, variance reduction is needed. To assess the quality of variance reduction techniques, a number of optimality criteria have been developed.

Optimality notions. If the number of needed simulation replications stays bounded for a fixed relative error as n goes to infinity, then one says that the simulation estimator has a *bounded relative error.*

Usually it is not easy to develop simulation algorithms with a bounded relative error, and hence one settles for some weaker optimality notion. A commonly used benchmark is *asymptotic optimality* (also known as *asymptotic efficiency*), see e.g. Heidelberger [11]. In the setting of probabilities which decay at an exponential or faster rate, we have the following definition:

Definition 3.2 *We call an estimator* $\hat{p}_n$ *of* p_n asymptotically optimal *if*

$$\lim_{n\to\infty} \frac{1}{n} \log\left(\mathbb{E}\,\hat{p}_n^2\right) = 2 \lim_{n\to\infty} \frac{1}{n} \log p_n. \tag{6}$$

In Section 4 we show that our proposed method is asymptotically optimal.
From $\mathbb{V}\mathrm{ar}\,\hat{p}_n^2 = \mathbb{E}\,(\hat{p}_n^2) - (p_n)^2 \geq 0$ it is easy to verify that the left hand side in (6) is not smaller than the right hand side. Hence, the best possible estimator achieves equality. Informally, asymptotic optimality entails that the number of simulation replications that are needed to obtain a fixed relative error may grow as n grows, but this growth is at a smaller than exponential rate.

Variance reduction. The variance reduction technique we use to improve over ordinary Monte-Carlo simulation is *importance sampling*, see the survey paper [11] for an extensive treatment. The idea of importance sampling can be explained as follows. Let in our original stochastic model all random variables be defined on a probability space, corresponding to measure $\mathbb{P}$. Then, in the simulations the system is simulated under measure $\mathbb{Q}$ (with $\mathbb{P}$ absolutely continuous relative to $\mathbb{Q}$). The new measure $\mathbb{Q}$ should be chosen such that the rare event under consideration occurs more frequently. To get an unbiased estimate, the observations are weighed by a *likelihood ratio*, measuring the difference in likelihood of the simulation output in both models.
More formally, the procedure can be described as follows. Denote in the sequel expectation with respect to $\mathbb{P}$ by $\mathbb{E}(\cdot)$, and expectation with respect to $\mathbb{Q}$ by $\mathbb{E}^{(\mathbb{Q})}(\cdot)$. Simulate the queue until it is decided whether event U_n occurs or not; in the former case $I(U_n) := 1$, in the latter case $I(U_n) := 0$. Then it is a standard result that unbiasedness is recovered if the observation $I(U_n)$ is weighed by likelihood ratio $\mathrm{d}\mathbb{P}/\mathrm{d}\mathbb{Q}(\omega) =: L(\omega)$:

$$p_n = \mathbb{P}(U_n) = \mathbb{E}^{(\mathbb{Q})}\left[I(U_n)\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\right].$$

This L is determined by the sample paths ω generated in the individual simulation experiment: $L(\omega)$ is defined as the ratio of the probability density of ω under the original measure $\mathbb{P}$, and the density under the importance sampling measure $\mathbb{Q}$. Details on the calculation of these likelihood ratios are given in Section 3.2 and 3.3.

Large deviations. A convenient choice of $\mathbb{Q}$ can be obtained by using large deviation theory. The theory of sample path large deviations, cf. Theorem 2.5, provides us the most likely path f to a buffer overflow. The idea is to construct the change of measure $\mathbb{Q}$ such that typical sample paths drawn under $\mathbb{Q}$ resemble this f. In Section 3.3 we give our new measure $\mathbb{Q}$, in Section 4.3 we show that it follows on average the path given in Theorem 2.5.

If we use this change of measure, it turns out that we can bound the likelihood ratio of overflow at time $k^\star$ with e^{-nI} – such bounds are typically required to prove asymptotic optimality, see also Section 4. However, the likelihood ratio of overflow at *another* time $k \neq k^\star$ is not bounded so tightly. We solve this problem by partitioning U_n into several disjoint subsets $(U_n(k))_{k\in\mathbb{N}}$, and to estimate the probabilities $\mathbb{P}(U_n(k))$ separately by suitable changes of measure $\mathbb{Q}_k$.

Partitioning of the overflow event. Truncation. Let $K := \inf\{k \in \mathbb{N} : A_n(k) - nck > nb\}$ be the epoch of (the first) buffer overflow. Then the event of overflow for the first time at time k is given by $U_n(k) := \{K = k\}$. Defining $p_n(k) := \mathbb{P}(U_n(k))$, and noticing that the events $(U_n(k))_{k\in\mathbb{N}}$ are disjoint, it is clear that

$$U_n = \bigcup_{k=1}^{\infty} U_n(k) \quad \text{and} \quad p_n = \sum_{k=1}^{\infty} p_n(k).$$

Notice that overflow is only possible for k larger than $b/(1-c)$ (all sources send at peak rate all the time). Hence, the above summation does not necessarily start at $k = 1$. However, for notational ease we neglect this issue.

As said above, in our simulation procedure we use a sequence of measures $\mathbb{Q}_k$ to estimate the probabilities $p_n(k)$ by estimators $\hat{p}_n(k)$, with $k \in \mathbb{N}$. Since the buffer overflow probability is the sum of *infinitely* many of such probabilities we truncate at k_0: p_n is estimated by $\hat{p}_n := \sum_{k=1}^{k_0} \hat{p}_n(k)$ for some large k_0. Obviously, epoch k_0 should be chosen such that the error made is small, where the error is defined as the *relative bias* (RB):

$$\text{RB} = \frac{p_n - \mathbb{E}\hat{p}_n}{p_n}.$$

Obviously RB is larger than zero, since $\hat{p}_n$ underestimates $p_n(b, c)$. In this paper we impose the requirement that the RB is smaller than some small predefined ϵ:

Requirement 3.3 *For any fixed $\epsilon > 0$, k_0 is chosen such that the relative bias RB is below ϵ. Equivalently: $\mathbb{E}\hat{p}_n \leq p_n \leq \mathbb{E}\hat{p}_n/(1-\epsilon)$.*

Notice that our estimator is *biased*: $\mathbb{E}\hat{p}_n \neq p_n$. However, we are not loosing much if we choose ϵ small. From a practical point of view there is not much difference between an unbiased estimator with 10% RE on the one hand, and a biased estimator ($\epsilon = 0.05$) with 5% RB.

3.2 The algorithm

In this section we give a description of our algorithm in pseudo code. Here δ is the relative error and ϵ is the relative bias.

```
Find decay rate I                                  [See Section 2.3].
Determine k0 such that RB < epsilon                [See Section 4.1].
```

```
M := 0
FOR k in {1,...,k0} DO
     Calculate change of measure Q(k)              [see Section 3.3].
END
REPEAT
     FOR k in {1,...,k0} DO
          Simulate realization w under Q(k)
          Determine if I = 1 or 0
          Determine likelihood ratio L(w)          [see Section 3.3].
          Update mean M(k) and variance V(k) of kth estimator
     END
     Update mean M and variance V of estimator
UNTIL RE = sqrt(V)/M < delta
```

For the sample means and sample variances we use the standard formulas. In the above algorithm, we need for all $k \in \{1, \ldots, k_0\}$ the change of measure $\mathbb{Q}_k$. The calculation of this importance sampling distribution is the subject of the next subsection.

3.3 The exponential change of measure

As explained in Section 3.1, we estimate p_n by estimating the individual $p_n(k)$, all of them with a specific change of measure. As $p_n(k)$ decays exponentially, it is a natural choice to use an *exponential twist* of $A(k)$:

$$\mathbb{Q}_k\left(A(k) = x\right) = \frac{e^{\theta_k x}\mathbb{P}(A(k) = x)}{\mathbb{E}\exp\left(\theta_k A(k)\right)}, \tag{7}$$

where θ_k is the optimizing θ in

$$\sup_{\theta}\left(\theta(b + ck) - \log \mathbb{E}e^{\theta A(k)}\right).$$

We will use the abbreviation $\mathbb{Q}$ for $\mathbb{Q}_{k^\star}$. We say that we twist the distribution of $A(k)$ by an exponential amount of θ_k. Unfortunately, the new measure $\mathbb{Q}_k$ does not provide us immediately the change of measure of the on-times and off-times during the time interval $\{1, \ldots, k\}$. Below we will propose a change of measure of these random variables; later we will show that this change of measure coincides with the desired distribution (7).

Change of measure. For any of the n sources, we propose the following change of measure. Like under the original measure $\mathbb{P}$, the source alternates between on and off, but the on- and off-times are *time-dependent*:

- First we draw the 'initial state', i.e., active or silent. The source is on with probability

$$\rho_k := \frac{\pi_A \mathbb{E}_{A^\star} e^{\theta_k A(k)}}{\mathbb{E} e^{\theta(k) A(k)}},$$

and off with probability $1 - \rho_k$.

- The durations of the initial on or off-state are twisted as follows:

$$\mathbb{Q}_k(A^\star = i) = \frac{a_i^\star e^{\theta_k i} \mathbb{E}_S e^{\theta_k A(k-i)}}{\mathbb{E}_{A^\star} e^{\theta_k A(k)}} \quad , \quad \mathbb{Q}_k(S^\star = i) = \frac{b_i^\star \mathbb{E}_A e^{\theta_k A(k-i)}}{\mathbb{E}_{S^\star} e^{\theta_k A(k)}}.$$

for $i < k$ and

$$\mathbb{Q}_k(A^\star = k) = \frac{\sum_{i=k}^\infty a_i^\star e^{\theta_k k}}{\mathbb{E}_{A^\star} e^{\theta_k A(k)}} \quad , \quad \mathbb{Q}_k(S^\star = k) = \frac{\sum_{i=k}^\infty b_i^\star}{\mathbb{E}_{S^\star} e^{\theta_k A(k)}}.$$

- Similarly, a burst or silence starting at time ℓ is twisted as follows:

$$\mathbb{Q}_k(A = i \mid \ell) = \frac{a_i e^{\theta_k i} \mathbb{E}_S e^{\theta_k A(k-\ell-i)}}{\mathbb{E}_A e^{\theta_k A(k-\ell)}} \quad , \quad \mathbb{Q}_k(S = i \mid \ell) = \frac{b_i \mathbb{E}_A e^{\theta_k A(k-\ell-i)}}{\mathbb{E}_S e^{\theta_k A(k-\ell)}}.$$

for $i < k - \ell$ and

$$\mathbb{Q}_k(A = k - \ell \mid \ell) = \frac{\sum_{i=k-\ell}^\infty a_i e^{\theta_k (k-\ell)}}{\mathbb{E}_A e^{\theta_k A(k-\ell)}} \quad , \quad \mathbb{Q}_k(S = k - \ell \mid \ell) = \frac{\sum_{i=k-\ell}^\infty b_i}{\mathbb{E}_S e^{\theta_k A(k-\ell)}}.$$

Let $X(j) = 1$ (0) represent the event that the source is in the on (off) state at time j, and introduce the short notation $\mathbb{P}(i_1, \ldots, i_k) := \mathbb{P}(X(1) = i_1, \ldots, X(k) = i_k)$; define $\mathbb{Q}_k(i_1, \ldots, i_k)$ analogously (replace $\mathbb{P}$ by $\mathbb{Q}_k$). It is not hard to verify that

$$\mathbb{Q}_k(i_1, \ldots, i_k) = \frac{\mathbb{P}(i_1, \ldots, i_k) e^{\theta(k) \sum_{j=1}^k i_j}}{\mathbb{E} e^{\theta_k A(k)}},$$

as required. Thus we arrive at the following Proposition:

Proposition 3.4 *The above change of measure coincides with the desired new distribution* (7).

We now point out how to calculate the likelihood ratios, to be used in the algorithm of Section 3.2. Suppose the n i.i.d. values of $A(k)$ are sampled, and have values $\omega_1, \ldots, \omega_n$. Then it can be checked that the likelihood ratio of the experiment is

$$L(\omega_1, \ldots, \omega_n) := \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}_k}(\omega_1, \ldots, \omega_n) = e^{-\theta_k \sum_{i=1}^n \omega_i} \left(\mathbb{E} e^{\theta_k A(k)} \right)^n. \tag{8}$$

It is important to observe that, using the above change of measure, the likelihood ratio is small in the regions of interest, which is a desirable property of importance sampling distributions. This is because $A_n(k) > nb + nck$ implies that $LI(U_n(k))$ is bounded from above by e^{-nI_k} :

$$LI(U_n(k)) \leq e^{-n\theta_k(b+ck)} \left(\mathbb{E} e^{\theta_k A(k)} \right)^n = e^{-nI_k}. \tag{9}$$

Notice that the exponential change of measure *changes* during the simulation run. This is essentially different from many earlier studies [12, 18, 20]. In those studies a constant exponential change of measure is derived. The main difference with our work is that we look at the many-sources regime, whereas there it is focused on large-buffers asymptotics. Importantly, the techniques of [12, 18, 20] do not allow for heavy tails, whereas our many-sources-based approach *does*.

4 Optimality properties of the importance sampling procedure

In this section we prove that the proposed change of measure has a number of desirable properties. First we analytically derive an expression for the 'simulation horizon', k_0, given Requirement 3.3. In Section 4.2 we show that this choice of k_0 implies that the proposed procedure is asymptotically optimal. We conclude this section by proving that our change of measure follows the optimal path identified by Wischik [27].

4.1 Derivation of simulation horizon k_0

As explained in Section 3, the simulation is truncated at epoch k_0. In this section we describe how to choose this k_0. Recall that k_0 has to be chosen such that the relative bias of $\hat{p}_n$ is smaller than some small preselected number ϵ, i.e., k_0 has to be chosen such that

$$\text{RB} = \frac{p_n - \sum_{k=1}^{k_0} p_n(k)}{p_n} = \frac{\sum_{k_0+1}^{\infty} p_n(k)}{p_n} < \epsilon.$$

We find an upper bound on RB by deriving an upper bound on $\sum_{k=k_0+1}^{\infty} p_n(k)$ and a lower bound on p_n. This gives us a procedure to find a k_0 that guarantees that the relative bias RB does not exceed ϵ.

- First we find a *lower bound* on p_n. Obviously,

$$p_n = \mathbb{P}(\exists k < \infty : A_n(k) - nck > nb) \geq \mathbb{P}(A_n(k^\star) > nb + nck^\star) \geq \mathbb{P}(A_n(k^\star) = \lceil nb + nck^\star \rceil).$$

  Notice that the $A(k^\star)$ are distributed on $\{0, \ldots, k^\star\}$. Because of this finite state space, we may invoke Inequality (2.1.13) of Dembo and Zeitouni [7]. It implies that the latter probability is not smaller than

$$(n+1)^{-(k^\star+1)} \exp\left(-nJ\left(\frac{1}{n}\lceil nb + nck^\star \rceil\right)\right), \quad \text{with} \quad J(x) := \sup_\theta \left(\theta x - \log \mathbb{E}e^{\theta A(k)}\right).$$

  We could use this lower bound in our calculation of k_0, but we might wish to replace it by a cleaner expression. This is done as follows. Clearly, for large n,

$$nJ\left(\frac{1}{n}\lceil nb + nck^\star \rceil\right) \leq nJ\left(b + ck^\star + \frac{1}{n}\right) \approx nJ(b + ck^\star) + J'(b + ck^\star).$$

  In the last expression $J(b + ck^\star)$ equals $I_{k^\star} = I$. Also $J'(b + ck^\star)$ reduces to $\theta_{k^\star}$, due to Exercise 5 of [5, pag. 74].

- Now we look for an *upper bound* on $\sum_{k_0+1}^{\infty} p_n(k)$. In Corollary 2.4 we showed that $I_k > \alpha \log k$ for some positive constant α and all $k \geq k_{\min}$. Noticing that $p_n(k)$ is smaller than $\mathbb{P}(A_n(k) - nck > nb)$, a Chernoff bound argument implies that

  $$p_n(k) \leq e^{-nI_k}.$$

  Suppose k_0 is larger than $k_{\min}$. Then, with n larger than $1/\alpha$,

  $$\begin{aligned} \sum_{k=k_0+1}^{\infty} p_n(k) &\leq \sum_{k=k_0+1}^{\infty} e^{-nI_k} \leq \sum_{k=k_0+1}^{\infty} e^{-n(\alpha \log k)} \\ &\leq \int_{k_0}^{\infty} x^{-n\alpha} \mathrm{d}x = \frac{k_0^{-n\alpha+1}}{n\alpha - 1}. \end{aligned}$$

We are left with the task of finding the smallest k_0 such that

$$\frac{k_0^{-n\alpha+1}}{n\alpha - 1} \cdot (n+1)^{k^\star+1} \cdot e^{nI} \cdot e^{\theta_{k^\star}} < \epsilon.$$

A straightforward calculation gives that k_0 could be chosen as the smallest integer larger than

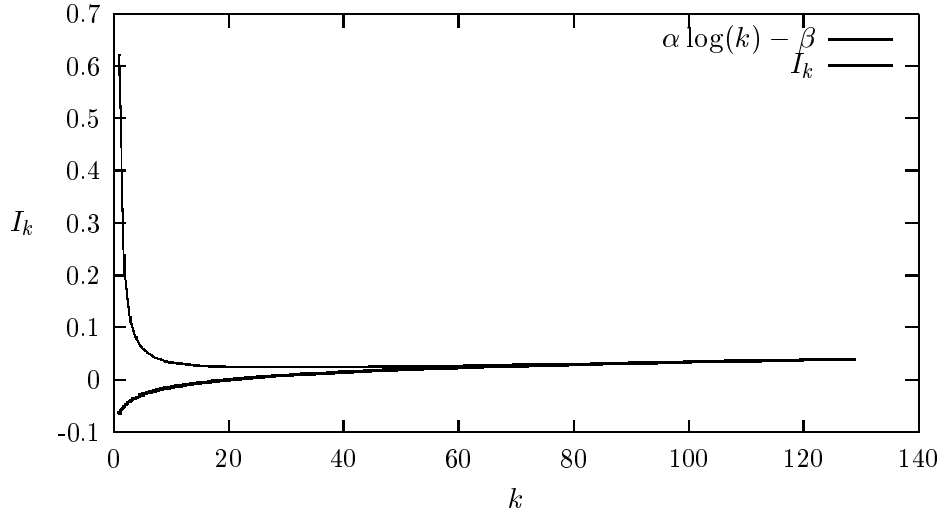
$$\exp\left(\frac{nI + \theta_{k^\star}}{n\alpha - 1}\right) \cdot \left(\frac{(n+1)^{k^\star+1}}{(n\alpha-1)\epsilon}\right)^{\frac{1}{n\alpha-1}}. \tag{10}$$

Call this 'simulation horizon' $k_0(n)$. It is not hard to see that the first factor tends to a constant as $n \to \infty$, whereas the second factor tends to 1. It is not hard to see that $k_0(n)$ is bounded. A fortiori, $\log k_0(n) = o(n)$, a property that we need in Section 4.2.

Our numerical experiments showed that, to reduce k_0, it is often beneficial to use bounds of the form $I_k > \alpha \log k - \beta$ (with $\alpha, \beta > 0$), instead of bounds like $I_k > \alpha \log k$. Then the $k_0(n)$ looks as in (10), but with I replaced by $I + \beta$. In order to find the best α and β (i.e., the ones that minimize the value of k_0), the following heuristic procedure can be followed: (1) choose a k, and solve α and β from $I_k = \alpha \log k - \beta$ and $I_k - I_{k-1} = (\mathrm{d}/\mathrm{d}k)\alpha \log k$. (2) Compute the resulting value of k_0 with the procedure described above and check whether α and β are feasible, i.e. $n > \alpha^{-1}$ and $I_\ell \geq \alpha \log \ell - \beta$ for $\ell = k_0, \ldots, k_{\max}$ for some large $k_{\max}$. (3) Repeat this for a sequence of values of k, and use the one that minimizes k_0 (provided that the corresponding α and β are feasible).

In Figure 1 we applied the algorithm above for a typical example. We present the graph of the functions I_k and $\alpha \log k - \beta$ for the optimal α and β. Note that the latter function lies just above I_k, especially for larger values of k. This indicates that we have chosen α and β and thus k_0 economically.

Figure 1: Computation of α, β and k_0 for $A \sim$Pareto



4.2 Asymptotic optimality

We now prove that our simulation procedure is asymptotically optimal, given the simulation horizon $k_0(n)$ derived in the previous subsection.

Proposition 4.1 *The proposed procedure is asymptotically optimal if* $\log k_0(n) = o(n)$. *In particular, choosing* k_0 *according to* (10) *is sufficient for asymptotic optimality.*

Proof. From (9), for all $j \in \mathbb{N}$, it holds that $\mathbb{E}^{(\mathbb{Q})}\left(L^j I(U_n(k))\right) \leq e^{-jnI_k}$. This immediately gives

$$\mathbb{E}^{(\mathbb{Q})}\left(\left(\sum_{k=1}^{k_0} LI(U_n(k))\right)^2\right) \leq \sum_{k=1}^{k_0} e^{-2nI_k} + 2\sum_{k=1}^{k_0}\sum_{\ell=1}^{k-1} e^{-nI_k - nI_\ell} \leq k_0^2 e^{-2nI},$$

using $I \leq I_k$ for $k \in \mathbb{N}$. This immediately gives that $\hat{p}_n$ is an asymptotically optimal estimator of p_n if $\log k_0 \equiv \log k_0(n)$ is $o(n)$, cf. Condition (6). □

In Definition 3.2 we focused on estimators with a subexponentially growing number of 'experiments' that is required to get a certain RE (in the scaling parameter n). Here, an experiment is defined as the effort that is done to get a single observation, so in fact $k_0(n)$ 'runs' (where the ith run has a length of i epochs). This aspect is not taken care of by our 'asymptotic optimality' notion. This problem can be solved by using more sophisticated versions of the asymptotic optimality criterion. We could consider estimators for which the *amount of 'work'* (expressed for instance in CPU time) grows subexponentially in n. Clearly, from a practical point of view, this seems a fairer notion. However, because $k_0(n)$ is bounded, it is straightforward that our procedure will also be optimal in that sense.
Although it is not reflected in the above optimality notions, our importance sampling algorithm still consumes considerable simulation time if $k_0(n)$ turns out to be large, because of the $k_0(n)$ runs per

experiment. Clearly, this plays an important role if b is large. In Subsection 5.1 we describe a heuristic to reduce the number of these runs as a method to speed up the simulation algorithm.

4.3 Relation to the optimal path

In Proposition 4.1 we established the asymptotic optimality property of our importance sampling procedure. We now present our second proposition supporting the choice of our change of measure. We prove that the average path under the importance sampling measure $\mathbb{Q}$ corresponding to $k^\star$ coincides with the optimal path to overflow that was identified by Wischik [27].

Proposition 4.2 *The average path of the process under the importance sampling measure corresponding to $k = k^\star$ coincides with the most likely path identified by Wischik* [27].

Proof. The probability that, under $\mathbb{Q}$, a source is in the on-state at time $j \in \{1, \ldots, k^\star\}$ is given by

$$\begin{aligned}
\sum_{i_k, k\neq j} \mathbb{Q}(i_1, \ldots, i_{j-1}, 1, i_{j+1}, \ldots, i_{k^\star}) &= \sum_{i_k, k\neq j} \mathbb{P}(i_1, \ldots, i_{j-1}, 1, i_{j+1}, \ldots, i_{k^\star}) \frac{e^{\theta_{k^\star}\left(\sum_{\ell=1,\ell\neq j}^{k^\star} i_\ell + 1\right)}}{\mathbb{E}e^{\theta_{k^\star} A(k^\star)}} \\
&= \sum_{i_1,\ldots,i_{k^\star}} \mathbb{P}(i_1, \ldots, i_{j-1}, 1, i_{j+1}, \ldots, i_{k^\star}) i_j \frac{e^{\theta_{k^\star}\left(\sum_{\ell=1}^{k^\star} i_\ell\right)}}{\mathbb{E}e^{\theta_{k^\star} A(k^\star)}} \\
&= \frac{\mathbb{E}X(j) e^{\theta_{k^\star} A(k^\star)}}{\mathbb{E}e^{\theta_{k^\star} A(k^\star)}}.
\end{aligned}$$

So the mean amount of traffic sent by a single source in $\{1, \ldots, j\}$ is

$$\sum_{i=1}^{j} \frac{\mathbb{E}X(i) e^{\theta_{k^\star} A(k^\star)}}{\mathbb{E}e^{\theta_{k^\star} A(k^\star)}} = \frac{\mathbb{E}A(j) e^{\theta_{k^\star} A(k^\star)}}{\mathbb{E}e^{\theta_{k^\star} A(k^\star)}} = f(j),$$

where the last equation is due to (5). □

The path to overflow depends on the distributions of the on- and off-times. These are treated in detail in [15]. We will reflect on some of them here. As demonstrated in [15], the shape of the off-times does not really affect the qualitative behavior of the queue (i.e., $I(b)$ as a function of b), whereas the shape of the on-times does. For that reason, in the experiments below, we leave the distribution of the off-times constant (Geometric). The on-times are chosen respectively Geometric (light tail), Weibull ('moderately' heavy tail), and Pareto (heavy tail). The exact definitions of these distributions are given in Section 5.2.

I. Distribution of activities and silences during path to overflow. We here focus on the distributions of the residual bursts (silences), given that the source is on (off) at time 0, under the new measure.
As follows implicitly from [17], for small b there is hardly any difference between the new distributions. However, there are significant differences for larger b as can be seen in Figure 2, 3 and 4 where we plotted the distributions of $A^\star$ and $S^\star$ under both the original and the importance sampling measure. We use $\mathbb{E}A = 5$, $\mathbb{E}S = 10$ and $c = 0.37$ in all the figures in this subsection.

- We see that for Geometric on-times, the residual silences (bursts) are relatively short (long) under $\mathbb{Q}$, compared to $\mathbb{P}$. The probability that a sources stays in the on-state (or off-state) during the entire path to overflow is extremely small. The intuition is that under $\mathbb{Q}$ the sources alternate between on and off, but with a longer on-time and shorter off-time than under $\mathbb{P}$.

- For Weibull and Pareto on-times, the off-times under the importance measure show almost no deviant behavior from their normal statistical law, but the bursts are relatively large: There is a relatively large fraction of sources that transmits during the entire path to overflow. Here the intuition is that there are essentially two types of sources: a number of them has one single huge on-time during the entire path to overflow, whereas the remaining sources alternate like they would do under $\mathbb{P}$.

Figure 2: Distributions of the residual on- and off-times for $A \sim$Geometric

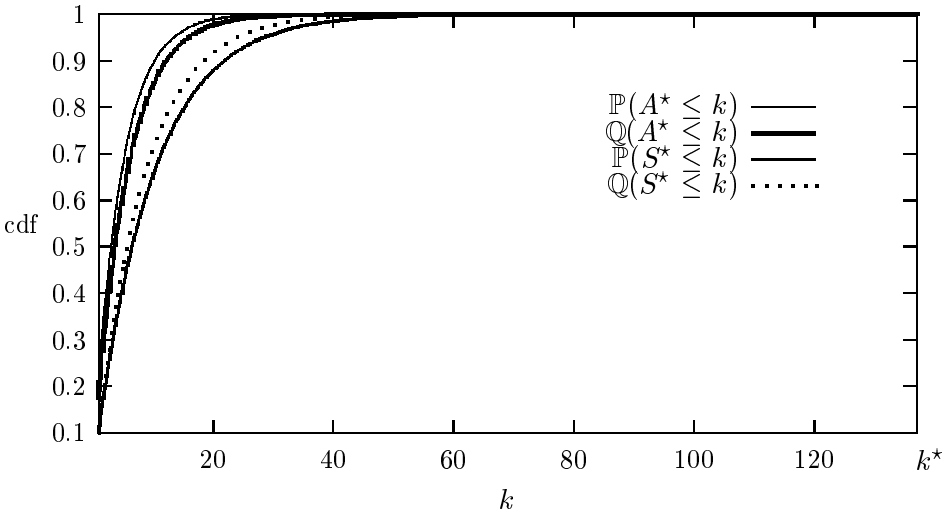


Figure 3: Distributions of the residual on- and off-times for $A \sim$Pareto

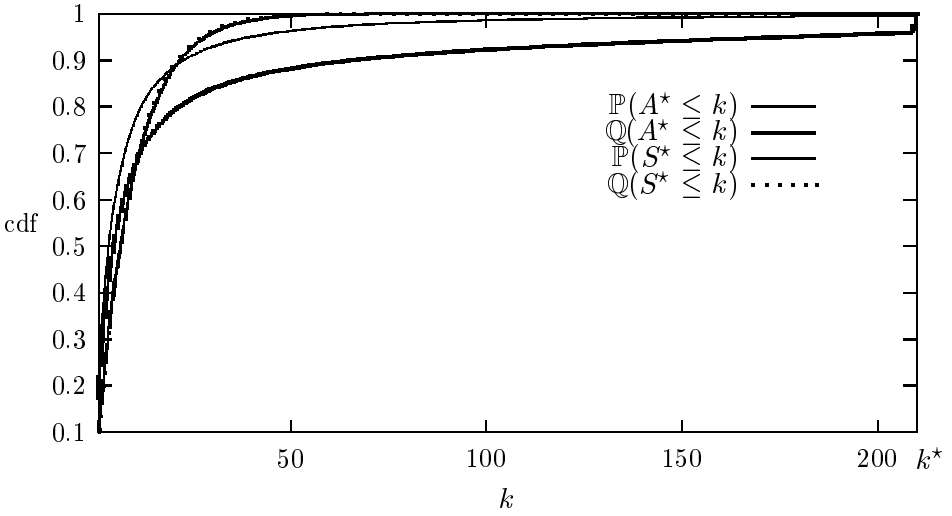
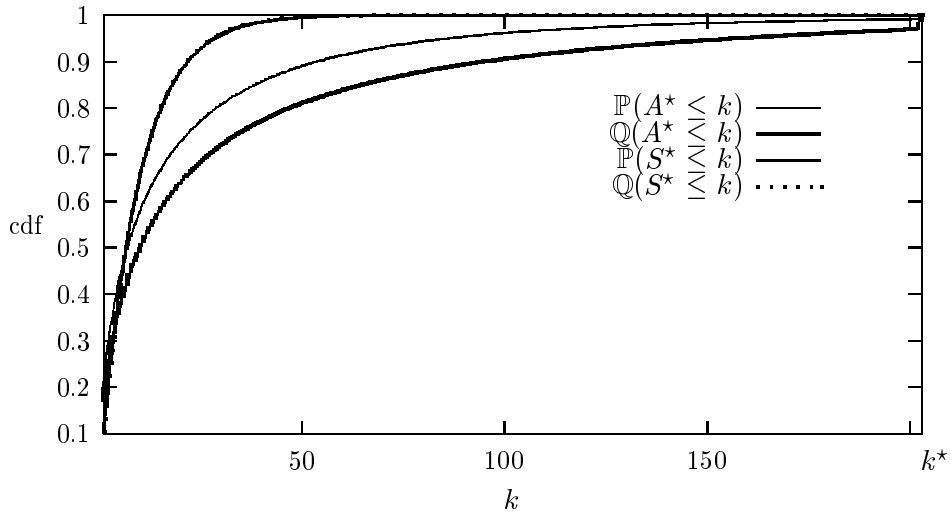


Figure 4: Distributions of the residual on- and off-times for $A \sim$Weibull



An alternative technique for rare event simulation is *ReSTART*. This variance reduction technique can roughly be explained as follows. Suppose the chance on the buffer overflow over level B must be estimated. In this setting *ReSTART* (in its most simple form) is implemented by introducing a threshold at, say, $B/2$. Each time a sample path reaches level $B/2$ for the first time it is split into several subpaths which evolve independently from then on. For *ReSTART* to be successful as a variance reduction technique it is necessary that the rare event is split into two parts: One part involving unlikely realizations of random variables that are drawn before the threshold $B/2$ has been reached, and the other part involving unlikely realizations of random variables that are drawn after level $B/2$ has been reached.

As we saw above, for heavy-tailed on times, a buffer overflow is likely to be caused by a fraction of sources which transmit during the entire path to overflow. In other words: a buffer overflow is likely to be caused by the fact that a fraction of the sources have to transmit during the entire path to overflow, in particular during the part of the path to overflow where the threshold $B/2$ has not yet been reached. This explains why *ReSTART* does not work so well here.

II. Path to overflow: number of transmitting sources, and time to overflow. We review some of the results from [15, 16, 17]. Consider the optimal epoch of overflow $k^\star(b)$ as a function of the buffer size. For small b, $k^\star(b)$ is more or less invariant in the distribution, for given means $\mathbb{E}A$ and $\mathbb{E}S$. For larger b, the value of $k^\star(b)$ increases linearly for Exponential and Weibull on-times, and in a superlinear way for Pareto on-times (like $b \log b$). This implies that for Pareto bursts the net input rate during the path to overflow is small if b is large: it looks like $(\log b)^{-1}$. The off-time distribution does not play an essential role other than via its first moment.

In Figure 5 and 6 we plotted the evolution of the fraction of the sources which are in the on-state during the optimal trajectory to overflow for a typical example. These graphs can be obtained easily from the optimal paths (to be calculated numerically as described in Section 2.3). For very small b there is hardly any difference between the fraction of sources in the on-state during the optimal trajectories for the different on-time distributions. In Figure 5 we plotted these fractions for $b = 0.5$ (which is in the intermediate buffer range). The net rate of sources is positive if the fraction of the sources in the on-state is larger than 0.37. We see that during the optimal trajectory to overflow the buffer starts to fill immediately, first very slowly, later the sources begin to conspire and at the end of the trajectory the net input rate of the buffer process drops down to almost zero.
In Figure 6 we raised the buffer capacity to $b = 5$ (large b). Here we see a clear difference between Geometric (light-tailed) on-times on one hand and Weibull and Pareto (heavy-tailed) on-times on the other hand. For Geometric on-times the fraction of sources in the on-state is constant during the largest part of the trajectory to overflow. This is because all the sources conspire to fill the buffer; during the path to overflow they alternate between on and off. On the other hand, for Weibull and Pareto on-times the buffer fills because of the deviant behavior of some of the sources: they have very long bursts during the optimal trajectory to overflow, as we saw in Figure 3 and 4.

Figure 5: Fraction of the sources in the on-state during the optimal trajectory to overflow for $b = 0.5$

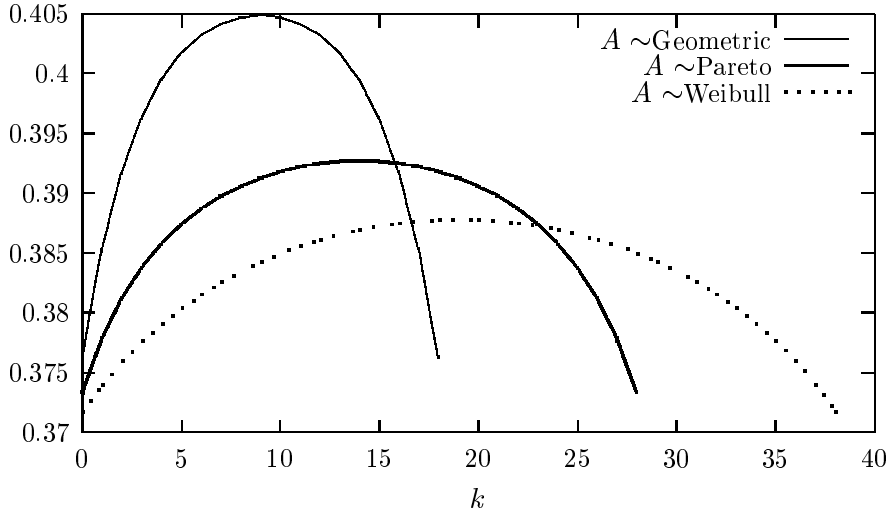
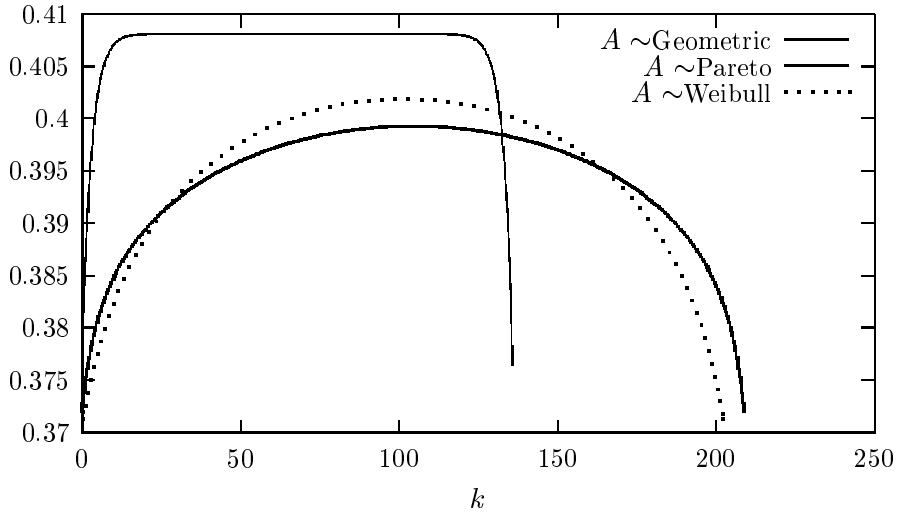


Figure 6: Fraction of the sources in the on-state during the optimal trajectory to overflow for $b = 5$



5 Implementation issues and numerical results

This section focuses on the practical implementation and numerical results. In Section 5.1 we point out how to reduce the number of simulation runs per experiment from k_0 to a considerably lower value. We also point out how we can obtain a smaller value of k_0 heuristically. Section 5.2 assesses the speed up, compared to naive simulation. We conclude this section by discussing the limitations of our method.

5.1 Accelerations

Reducing the number of runs per simulation experiment. In realistic scenarios, simulation horizon k_0 can be pretty large, particularly for large b. Since each simulation replication consists of k_0 sample paths, importance sampling can be rather time consuming. We discuss a heuristic to accelerate the simulation algorithm described in Section 3.2 by reducing the simulation effort per simulation replication. A disadvantage of this heuristic is that the variance of the simulation estimate is bounded less tightly. In this case we cannot prove asymptotic optimality anymore. In the simulation procedure as described in Section 3.2, each probability $\mathbb{P}(U_n(k))$ is estimated separately using its own simulation runs with its own change of measure. The change of measure corresponding to a buffer overflow that occurs for the first time at time k can also be used to estimate $\mathbb{P}(U_n(\ell))$ for $\ell < k$. We use this fact in the following way to reduce the number of runs per simulation experiment:
Define $i_0 = 1$, $i_1 = k^\star$, $i_2 = k^\star + \Delta$, $i_3 = k^\star + 2\Delta, \ldots, i_{j-1} = k^\star + \max\{l \in \mathbb{N} : k^\star + l\Delta < k_0\}\Delta$ and $i_j = k_0$ for some positive integer Δ. A way to reduce the simulation time is to simulate for

$$\sum_{\ell=i_j}^{i_{j+1}} \mathbb{P}(U_n(\ell)) = \mathbb{P}\left(\bigcup_{\ell=i_j}^{i_{j+1}} U_n(\ell)\right)$$

in one simulation experiment using the change of measure corresponding to i_{j+1}. Of course, more sophisticated versions of the procedure described above are possible.

One run per simulation experiment. In order to reduce the number of runs per simulation experiment to one, we can simulate for p_n by using the change of measure corresponding to $k^\star$. Since this change of measure is only defined for $A(k)$ for $k \leq k^\star$, we have to extend this change of measure for residual bursts and silences that end after $k^\star$ and for bursts and silences that start after $k^\star$. We do this as follows for the residual bursts and silences:

$$\mathbb{Q}_k\left(A^\star = i\right) = \frac{a_i^\star e^{\theta_k i}\mathbb{E}_S e^{\theta_k A(k-i)}}{\mathbb{E}_{A^\star} e^{\theta_k A(k)}} \quad , \quad \mathbb{Q}_k\left(S^\star = i\right) = \frac{b_i^\star \mathbb{E}_A e^{\theta_k A(k-i)}}{\mathbb{E}_{S^\star} e^{\theta_k A(k)}}.$$

for $i < k$ and

$$\mathbb{Q}(A^\star = i) = \frac{a_i^\star e^{\theta_k k}}{\mathbb{E}_{A^\star} e^{\theta_k A(k)}} \quad , \quad \mathbb{Q}(S^\star = i) = \frac{b_i^\star}{\mathbb{E}_{S^\star} e^{\theta_k A(k)}}$$

for $i \geq k^\star$. Similarly, a burst or silence starting at time ℓ is twisted as follows:

$$\mathbb{Q}_k\left(A = i \mid \ell\right) = \frac{a_i e^{\theta_k i}\mathbb{E}_S e^{\theta_k A(k-\ell-i)}}{\mathbb{E}_A e^{\theta_k A(k-\ell)}} \quad , \quad \mathbb{Q}_k\left(S = i \mid \ell\right) = \frac{b_i \mathbb{E}_A e^{\theta_k A(k-\ell-i)}}{\mathbb{E}_S e^{\theta_k A(k-\ell)}},$$

for $i < k - \ell$ and

$$\mathbb{Q}_k\left(A = i \mid \ell\right) = \frac{a_i e^{\theta_k (k-\ell)}}{\mathbb{E}_A e^{\theta_k A(k-\ell)}} \quad , \quad \mathbb{Q}_k\left(S = i \mid \ell\right) = \frac{b_i}{\mathbb{E}_S e^{\theta_k A(k-\ell)}},$$

for $i \geq k - \ell$. The intuition behind the above change of measure is that till $k^\star$ it gives on average the optimal path to overflow and after $k^\star$ we 'stop' using importance sampling. We have not been able to prove asymptotic optimality of this procedure.

Cutting down the simulation horizon. The simulation horizon k_0 can be very large in many practical scenarios. Therefore, it makes sense to use heuristic methods to cut down k_0 without violating the maximum relative bias condition of the estimator for p_n.
We propose a heuristic to derive a higher lower bound on p_n than derived in Section 4.1. According to the Bahadur-Rao theorem (see, e.g., Theorem 3.7.4 of [7]), $\gamma p_n(k^\star) \sim \sqrt{n}^{-1}\exp(-nI)$ $(n \to \infty)$ for a constant γ. The inequality $p_n > p_n(k^\star)$ suggests to use the heuristic bound $p_n > \sqrt{n}^{-1}\exp(-nI)$. We can compare $\sqrt{n}^{-1}\exp(-nI)$ with the on simulation based estimator of p_n to check whether this inequality is justified. Similarly to (10), we can choose

$$k_0 = \left\lceil \left(\frac{\exp(n(I+\beta))\sqrt{n}}{(n\alpha - 1)\epsilon} \right)^{\frac{1}{n\alpha - 1}} \right\rceil . \tag{11}$$

5.2 Results

In this subsection we present numerical results. We compare the importance sampling algorithm (with and without accelerations) with naive simulation and with two asymptotic approximations. We use the asymptotic approximation $p_n \approx \exp(-nI(b))$ which is induced by the large deviations results from Section 2.2 and the asymptotic approximation $p_n \approx \sqrt{n}^{-1}\exp(-nI)$ which is induced by the Bahadur-Rao theorem (see also Section 5.1).

Comparison between the estimates of p_n. The standard effort of any simulation algorithm is defined as the the variance per simulation replication times the CPU time per simulation replication. For standard simulation the variance per simulation replication is $p_n(1-p_n)$ and this variance is estimated by using the accurate estimate for p_n obtained by importance sampling (without the acceleration described in Section 5.1). The *efficiency ratio* of a simulation technique is defined as the ratio of the standard effort of naive simulation upon the standard effort of the simulation algorithm. We use the efficiency ratio to compare the efficiency of the different simulation algorithms with each other.
To compare the asymptotic approximations with the simulation algorithms, we compute the relative deviation of the asymptotic approximations from the on simulation based estimates.

The on- and off-time distributions. The on- and off times are $\mathbb{N}$-valued random variables. Like in Section 4.3, we choose Geometrically distributed off-periods. For the on-periods we choose the Geometric(q_1) distribution (light tail) with

$$\mathbb{P}(A = k) = (1-q_1)^{k-1}q_1 \ (0 < q_1 < 1),$$

the Weibull(κ, τ) distribution ('moderately' heavy tail) with

$$\mathbb{P}(A = k) = e^{-[\tau(k-1)]^\kappa} - e^{-[\tau k]^\kappa} \ (0 < \kappa < 1, \tau > 0),$$

and the Pareto(α, β) distribution ('very' heavy tail) with

$$\mathbb{P}(A = k) = [\beta/(\beta + k - 1)]^\alpha - [\beta/(\beta + k)]^\alpha \ (\alpha, \beta > 0).$$

It is not hard to develop procedures that give, for a given value of $\mathbb{E}A$, q_1 (Geometric), τ (Weibull, for given κ), and β (Pareto, for given α).

Values of the parameters. We choose $n = 200$, $\mathbb{E}A = 5$, $\mathbb{E}B = 10$, $c = 0.4$, $\alpha = 2.5$ and $\kappa = 0.4$. We choose the maximum relative bias ϵ equal to 0.05. This results in the Pareto(2.5,6.707), Weibull(0.4,0.7688) and the Geometric(0.2) distribution. We compute k_0 from the formula (11).

Results. The results are presented in Table 1, 2 and 3. First we give the simulation results using three different algorithms. The algorithm based on one simulation run per simulation replication is denoted with '1 run', the simulation algorithm that simulates for each $\mathbb{P}(U_k)$ separately is denoted with 'many runs', and the simulation that reduces the number of runs per simulation replication is denoted with 'some runs' (we use $\Delta = 10$). The percentages denote the relative half-width of their 99% confidence intervals (based on the Normal distribution). The numbers between parentheses denote the efficiency ratio (we use the estimate of p_n from algorithm 'some runs' as an approximation for the true value of p_n). We compute the variance per simulation replication for naive simulation via the well-known formula $p_n(1 - p_n)$.
We also give two approximations. Here the number between the brackets denotes the ratio of the approximation and the (estimated) true value of p_n. For each scenario we use 10,000 simulation replications for the algorithms 'many runs' and 'some runs', and we use 1,000 simulation replications for algorithm '1 run'. We choose a fixed number of simulation replications rather than simulating till the relative error has decreased beneath some prefixed level δ. In this way the computer program does not need to memorize all the changes of measure.

Table 1: Estimates of p_{200} for Geometric(0.2) on-times

	$b = 0.1$	$b = 0.5$	$b = 1$
	$k_0 = 30$, $k^\star = 5$	$k_0 = 41$, $k^\star = 13$	$k_0 = 52$, $k^\star = 20$
1 run	$1.16E{-}3 \pm 12.3\%$ (55)	$2.04E{-}7 \pm 13.7\%$ (2.6E5)	$1.05E{-}11 \pm 21.1\%$ (1.9E9)
many runs	$1.06E{-}3 \pm 12.3\%$ (1.9E2)	$2.30E{-}7 \pm 13.3\%$ (6.2E5)	$1.21E{-}11 \pm 12.8\%$ (1.1E10)
some runs	$1.19E{-}3 \pm 15.3\%$ (10)	$2.39E{-}7 \pm 19.4\%$ (2.8E4)	$1.26E{-}11 \pm 26.4\%$ (1.8E8)
$\exp(-nI)$	$5.23E{-}3$ (440%)	$1.23E{-}6$ (516%)	$6.47E{-}11$ (514%)
$\sqrt{n}^{-1}\exp(-nI)$	$3.70E{-}4$ (31%)	$8.73E{-}8$ (36%)	$4.58E{-}12$ (36%)

Table 2: Estimates of p_{200} for Pareto(2.5,6.707) on-times

	$b = 0.1$	$b = 0.5$	$b = 1$
	$k_0 = 65$, $k^\star = 6$	$k_0 = 98$, $k^\star = 19$	$k_0 = 131$, $k^\star = 32$
1 run	$1.58E{-}3 \pm 9.5\%$ (80)	$4.01E{-}6 \pm 10.5\%$ (2.4E4)	$1.67E{-}8 \pm 15.3\%$ (2.8E6)
many runs	$1.68E{-}3 \pm 11.4\%$ (99)	$4.21E{-}6 \pm 12.8\%$ (2.0E4)	$1.82E{-}8 \pm 13.3\%$ (3.4E6)
some runs	$1.61E{-}3 \pm 7.6\%$ (23)	$4.12E{-}6 \pm 6.8\%$ (7.6E3)	$1.84E{-}8 \pm 7.3\%$ (1.2E6)
$\exp(-nI)$	$6.69E{-}3$ (478%)	$1.96E{-}5$ (477%)	$8.86E{-}5$ (481%)
$\sqrt{n}^{-1}\exp(-nI)$	$4.73E{-}4$ (30%)	$1.93E{-}6$ (34%)	$6.26E{-}9$ (34%)

Table 3: Estimates of p_{200} for Weibull(0.4,0.7688) on-times

	$b = 0.1$	$b = 0.5$	$b = 1$
	$k_0 = 177$, $k^\star = 9$	$k_0 = 243$, $k^\star = 26$	$k_0 = 292$, $k^\star = 43$
1 run	$3.44E{-}3 \pm 7.7\%$ (60)	$1.02E{-}4 \pm 7.8\%$ (2.0E3)	$5.52E{-}6 \pm 9.0\%$ (2.8E4)
many runs	$3.49E{-}3 \pm 11.7\%$ (16)	$1.05E{-}4 \pm 13.5\%$ (2.7E3)	$6.45E{-}6 \pm 13.3\%$ (3.9E3)
some runs	$3.45E{-}3 \pm 4.9\%$ (10)	$1.14E{-}4 \pm 5.2\%$ (1.6E2)	$6.02E{-}6 \pm 7.1\%$ (1.4E3)
$\exp(-nI)$	$1.31E{-}2$ (486%)	$4.91E{-}3$ (491%)	$2.75E{-}5$ (456%)
$\sqrt{n}^{-1}\exp(-nI)$	$9.26E{-}3$ (34%)	$3.47E{-}5$ (35%)	$1.94E{-}6$ (32%)

All three importance sampling algorithms produce accurate estimates for p_n. The time needed is considerably smaller than under naive simulation – of course, the smaller the probability to be estimated, the larger the efficiency ratio. The efficiency ratio is typically in the order $10^4 - 10^5$ if p_n is about 10^{-6}, and in the order of 10^7 if p_n is about 10^{-8}. There is no clear-cut answer to the question which method works best, since this seems to depend on the specific scenario.
We see that the asymptotic approximations are not very accurate, but they seem to be off by almost a constant factor. This can be helpful to find (relatively) accurate approximations for p_n for scenarios with parameter values for which even importance sampling is time consuming.

5.3 Discussion

Although our importance sampling procedure clearly outperforms naive simulation, the method has some limitations. Some of these are 'general' limitations that arise when estimating the buffer overflow probability via equation (1).

- For some scenarios, given some prefixed relative bias, the simulation horizon k_0 is way too large to guarantee that a simulation replication will end in a reasonable amount of time. In some cases deriving a smaller k_0 using tighter (heuristic) bounds will help, but in other cases not. Particularly for heavy-tailed on-times, k_0 tends to be large.

- The value of k_0 can also be large for large b or highly loaded queues (the latter means that the drift of the process $\{A_n(k) - ck\}_k$ is, even under the new measure, hardly positive).

- When the number of sources grows large, the simulation effort per replication grows proportionally. Obviously, relying on equation (1), this is hard to prevent.

6 Remarks and outlook

For the model with a large number of on-off sources, we found the change of measure that 'mimics' the most likely path to overflow. However, this most likely path is given in Wischik [27] for many other input processes (for instance Gaussian inputs). For these input processes it would be interesting to find the change of measure that goes with the optimal path.

Also the extension to networks (for instance tandems, or feedforward networks) in the many-sources regime is not explored yet. Finally, we could consider other service disciplines: in the present study we focused on FIFO service, whereas in real networks also priority disciplines and generalized processor sharing may be implemented.

References

[1] D. Anick, D. Mitra, and M. Sondhi. Stochastic theory of a data-handling system with multiple sources. *The Bell System Technical Journal*, 61: 1871 – 1894, 1982.

[2] S. Asmussen and K. Binswanger. Simulation of ruin probabilities for subexponential claims. *Astin Bulletin*, 27: 297 – 318, 1996.

[3] N.K. Boots and P. Shahabuddin. Simulating GI/G/1 queues and insurance risk processes with subexponential distributions. *Preprint.*

[4] D. Botvich and N. Duffield. Large deviations, the shape of the loss curve, and economies of scale in large multiplexers. *Queueing Systems*, 20: 293 – 320, 1995.

[5] J. Bucklew. Large deviation techniques in decision, simulation, and estimation. Wiley, New York, 1990.

[6] C. Courcoubetis and R. Weber. Buffer overflow asymptotics for a buffer handling many traffic sources. *Journal of Applied Probability*, 33: 886 – 903, 1996.

[7] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications.* Jones and Bartlett, Boston, 1993.

[8] V. Dumas and A. Simonian. Asymptotic bounds for the fluid queue fed by subexponential on-off sources. *Preprint.*

[9] M. Grossglauser and J.-C. Bolot. On the relevance of long-range dependence in network traffic. *IEEE/ACM Transactions on Networking,* 7: 629 – 640, 1999.

[10] D. Heyman and T. Lakshman. What are the implications of long-range dependence for VBR traffic engineering? *IEEE/ACM Transactions on Networking,* 4: 301 – 317, 1996.

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

[12] G. Kesidis and J. Walrand. Quick simulation of ATM buffers with on-off multiclass Markov fluid sources. *ACM Transactions on Modeling and Computer Simulation*, 3: 269 – 276, 1993.

[13] W. Leland, M. Taqqu, W. Willinger, and D. Wilson. On the self-similar nature of Ethernet traffic. *IEEE/ACM Transactions on Networking*, 2: 1 – 15, 1994.

[14] N. Likhanov and R. Mazumdar. Cell loss asymptotics in buffers fed with a large number of independent stationary sources. *Proceedings IEEE Infocom,* 339 – 346, 1998.

[15] M. Mandjes and N.K. Boots. The shape of the loss curve, and the impact of long-range dependence on network performance. *In preparation.*

[16] M. Mandjes and S. Borst. Overflow behavior in queues with many long-tailed inputs. To appear in: *Advances in Applied Probability*, 2000. Also: CWI report PNA-R9911, available at

`http://www.cwi.nl/static/publications/reports/PNA-1999.html`

[17] M. Mandjes and J.H. Kim. Large deviations for small buffers: an insensitivity result. To appear in: *Queueing Systems,* 2000.

[18] M. Mandjes and A. Ridder. Finding the conjugate of Markov fluid processes. *Probability in the Engineering and Informational Sciences*, 9: 297 – 315, 1995.

[19] M. Mandjes and A. Ridder. A large deviations analysis of the transient of a queue with many Markov fluid inputs: approximations and fast simulation. To appear in: *ACM Transactions on Modeling and Computer Simulation.*

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

[21] B. Ryu and A. Elwalid. The importance of long-range dependence of VBR video traffic in ATM traffic engineering: myths and realities. *Computer Communication Review,* 26: 3 – 14, 1996.

[22] A. Simonian and J. Guibert. Large deviations approximation for fluid queues fed by a large number of on/off sources. *IEEE Journal of Selected Areas in Communications*, 13: 1017 – 1027, 1995.

[23] M. Villen-Altamirano and J. Villen-Altamirano. RESTART: a method for accelerating rare events simulations. *Proceedings ITC 13,* Copenhagen, 1991.

[24] J. Walrand. *An introduction to queueing networks.* Prentice-Hall, New Jersey, 1988.

[25] A. Weiss. A new technique of analyzing large traffic systems. *Advances of Applied Probability*, 18: 506 – 532, 1986.

[26] W. Willinger, M. Taqqu, R. Sherman, and D. Wilson. Self-similarity through high-variability: statistical analysis of Ethernet LAN traffic at the source level. *Computer Communication Review,* 25: 100 – 113, 1995.

[27] D. Wischik. Sample path large deviations for queues with many inputs. Submitted to: *Annals of Applied Probability*, 1998.