Efficient Simulation of Fluid Queues with Many Sources

Ad Ridder, Vrije Universiteit Amsterdam[*]

February 10, 1999

Abstract

In this paper we study the rare event of overflow in a Markov fluid queue with finite buffer and many input sources. The probability of this rare event will be estimated by simulations. We present a highly efficient importance sampling procedure to speed up the simulations. The implemented change of meausure is suggested after a large deviations analysis of the overflow probability. This analysis yields the optimal path to overflow. A number of experiments support the procedure.

1 Introduction

In this paper we study the rare event of overflow in a finite buffer queue with many input sources. The probability of this rare event will be estimated by simulations. As we shall verify by experiments, the probability decays exponentially when the number of input sources increases, due to the effect of the law of large numbers. This phenomenon is justified by large deviations theory. That means that the sample size of the simulation grows exponentially when we require reliable estimates. Hence, to implement an efficient simulation procedure for a large number of sources we need a variance reduction technique. In this paper we suggest importance sampling. Importance sampling means that the underlying probability measure of the system is changed. In simulations with importance sampling we generate realizations of the random variables according to alternative distributions which are induced by the new measure. Unbiased estimators are obtained by weighing each realization with its likelihood ratio.

In our particular problem, we deal with simulating a continuous-time Markov chain representing the states of the traffic sources. In the ordinary simulations the transition rate matrix is homogeneous, in the importance sampling simulations the matrix is inhomogeneous, i.e., time dependent. The time dependent transition rates follow after an analysis of the most likely way how an overflow occurs, the so-called optimal path concept. Loosely speaking, all realizations of the chain in the simulations which have led to an overflow, lie 'close' to the optimal path.

Experimentally we verify the efficiency of our proposed procedure. These experiments show that in the importance sampling the sample size grows at a very low rate when the number of traffic sources increases. In fact, realizations mimic the optimal path, and therefore we conjecture that our procedure is asymptotically optimal.

[*] Full address: Dept. of Econometrics, Vrije Universiteit, De Boelelaan 1105, 1081 HV Amsterdam, the Netherlands; e-mail aridder@econ.vu.nl

In the next two sections we describe the queueing model and the overflow problem in more detail. Also we give simulation results and the methodology of the simulation experiments. In Section 4 we present our importance sampling procedure together with a number of simulation experiments. Section 5 deals with the concept of optimal path.

2 The Model

The queueing model is decribed as a Markov fluid queue with many Markovian on/off sources, finite buffer, and constant output rate.

An *on/off source* generates a flow of traffic cells at constant rate during a random on-time and is silent during a random off-time. On- and off-times repeat. In a *Markovian on/off source* the on- and off-times have phase-type distributions. Hence, a Markovian on/off source is modelled as a continuous-time Markov chain $X = \{X(s) : s \geq 0\}$ on a finite state space $\{1, 2, \ldots, d\}$. The matrix of transition rates is denoted by $\Lambda = \big(\lambda_{ij}\big)_{i,j=1}^{d}$. The states are partitioned in on-states and off-states. When the chain resides in an on-state i, the source transmits at a constant $r_i = \sigma > 0$. In an off-state i there is no input, $r_i = 0$. In this way we say that $r := \big(r_1, r_2, \ldots, r_d\big)$ is the vector of transmission rates. The Markov chain X is called a *source-chain*.

The *buffer* is fed by a large number n of independent identical Markovian on/off sources. The buffer size is $B < \infty$ and the buffer capacity (output rate) is C. We assume that the buffer can handle average inputs. That is, let π be the stationary distribution of a source-chain X, then the average input per source is $\langle \pi, r \rangle := \sum_{i=1}^{d} \pi_i r_i$. Hence, we assume $\langle \pi, r \rangle < C/n$.

Queueing systems of this type have been studied abundantly for many purposes. We mention only the early work of Kosten [5], and the main reference Anick, Mitra & Sondhi [1]. Their major importance lies in their applications to analyse design problems in ATM (Asynchronous Transfer Mode) networks, see for instance the COST project [7].

3 The Problem

Suppose that we observe at time 0 a buffer content B_0 ($0 \leq B_0 < B$), and that the total input rate I_0 at time 0 exceeds C. Then we wish to estimate the *probability of overflow* at or before some fixed small time t. We denote this probability by H_n. Refering to the application to ATM, knowledge of the order of magnitude of H_n is of importance for Call Acceptance Control [7, Chapter 5]. There is no explicit expression available for this transient probability. Most of the analysis of fluid queues is devoted to stationary behaviour. In case of reversible Markovian sources, Tanaka, Hasida & Takahashi [9] give a Laplace transform representation of a solution of the partial differential equations which describe the system.

Here, we consider the option of executing *simulation studies* in order to find estimates of H_n. When n (number of sources) is large, the event of reaching overflow is a *rare event*, even for small buffer sizes B. This is due to the law of large numbers, which says that almost all sources behave according to their mean. Table 1 below illustrates this phenomenon. The model which is simulated, has the following characteristics.

3.1 Example

The on-times have an Erlang distribution with mean 0.5, and squared coefficient of variation 0.25. The off-times are Hyperexponentially distributed with mean 1.0, and squared coefficient of variation 1.5. The input rate in the on-states is $\sigma = 4.5$. This yields an average input per source $\langle \pi, r \rangle = 1.5$. The other buffer parameters vary and are listed in the table.

n	B	B_0	C	I_0	t	$\hat{H}_n$	#runs	time
10	1	0	25	25	0.25	1.74e-2	9681	37s
20	2	0	50	50	0.25	8.82e-3	19269	151s
30	3	0	75	75	0.25	1.36e-4	1260234	4h 2m
40	4	0	100	100	0.25	7.65e-5	2236711	9h 37m
20	2	1	50	55	0.15	3.70e-2	4454	21s
40	4	2	100	110	0.15	6.14e-3	27715	263s
60	6	3	150	165	0.15	1.17e-3	146129	34m 30s
80	8	4	200	220	0.15	2.35e-4	726271	3h 48m

Table 1. Standard simulations of Example 3.1.
Simulation times on a NeXT workstation (Motorola 68040, 25MHZ)

Notice that the buffer characteristics B, B_0, C, I_0 increase proportionally to the numvber of sources n, in such a way that the load of the buffer equals $n\langle \pi, r \rangle / C = 0.6$.

3.2 Simulation Methodology

A simulation run starts at time 0 and ends either at time t without an overflow, or at some earlier time epoch when buffer overflow occurs. The estimates $\hat{H}_n$ in Table 1 are produced by executing so many runs until the relative width of the 95%-confidence interval is 15% to both sides of the estimate. Equivalently saying: the relative error (RE) of the estimate is approximately 0.0765.

In each simulation run the consecutive states of the sources are realized as follows. The state of the simulation model at time s is represented by $Y(s) = \big(Y_1(s), Y_2(s), \ldots, Y_d(s)\big)$, where

$$Y_i(s) := \#\{\text{number of sources in state } i \text{ at time } s\}.$$

In state $Y(s)$ the total input rate into the buffer equals $\sum_{i=1}^d Y_i(s) r_i$. Clearly, $Y = \{Y(s) : s \geq 0\}$ is a continuous-time Markov chain with multi-dimensional states $y = (y_1, y_2, \ldots, y_d)$ such that $y_i \in \{0, 1, \ldots, n\}$ and $\sum_{i=1}^d y_i = n$. The rate of the transition $y \rightsquigarrow y - e_i + e_j$ $(i \neq j)$ is $y_i \lambda_{ij}$. In our terminology we call Y the *simulation-chain.*

A constant γ is determined such that

$$\gamma \geq \max_y \sum_{\substack{i,j \\ i \neq j}} y_i \lambda_{ij}.$$

In the simulation we realize *jump epochs* according to a Poisson(γ) process. And, only at these jump epochs the simulation-chain Y can change state. The transition probability of changing state y into $y - e_i + e_j$ is $y_i\lambda_{ij}/\gamma$ for $i \neq j$. Formally, we applied the method of *uniformization.*

To start a simulation run we have to specify the initial state $Y(0)$ of the simulation-chain. Recall that we have specified at time 0 only the initial buffer content B_0 and the initial input rate I_0. It suffices to specify the initial distribution α of the source-chain X, i.e., $\alpha_i = P(X(0) = i)$. Since all sources are identical, we may choose an initial distribution with mean input $\langle \alpha, r\rangle = I_0/n$. We choose the *most likely* distribution α with mean input I_0/n. That is, α minimizes the entropy of deviating from the equilibrium distribution π, because $I_0/n > C/n > \langle \pi, r\rangle$. From Sanov's theorem, e.g. in Shwartz & Weiss [8, Section 2.4] it follows that α solves

$$\inf \sum_{i=1}^{d} \alpha_i \log(\alpha_i/\pi_i) \quad \text{s.t.} \quad \langle \alpha, r\rangle = I_0/n.$$

Now we have specified all the ingredients of simulating the model of Example 3.1, which resulted in Table 1.

4 Importance Sampling

When we need to find estimates of the overflow probability of orders smaller than those in Table 1, ordinary simulations are not appropriate anymore. Importance sampling is a technique of obtaining variance reduction and therefore speeds up the simulation. However, it not guaranteed to work. When wrongly implemented, the variances go up. Therefore, an analysis of the queueing model is necessary. An approach that seems to be successful, is to apply the *large deviations theory* to come up with an asymptotic expression of the rare event probability. This expression is a rough approximation of the probability, but, moreover, it may suggest the new probability measure under which to simulate (*change of measure*). In many occasions this method yields highly efficient or even *asymptotically optimal* estimates. For the technical details and references to applications of this approach we refer to the surveys of Asmussen & Rubinstein [2] and Heidelberger [4].

Also our paper applies importance sampling after large deviations analysis. However, the change of measure is suggested by the *most likely or optimal path to overflow* rather than by the asymptotic expressions of the overflow probability. The optimal path is an abstract concept which gives insight in the statistical behaviour of the traffic sources during the time to overflow. Hence, it gives understanding of the way how the rare event of overflow has occurred. In this sense it relates to the RESTART method introduced by Villén-Altamirano & Villén-Altamirano [10]. In Glasserman, Heidelberger, Shahabuddin & Zajic [3] it is argued that the RESTART method is highly efficient in case one can find the optimal path.

4.1 Change of Measure

In the next section we give more details on the optimal path. Here we concentrate on the change of measure. First we introduce some notation:

$$\begin{aligned}
A(s) &:= \int_0^s r_{X(u)}\,du && \text{total input of source during } [0,s], \\
M(\theta;s) &:= E\exp[\theta A(s)] && \text{mgf of input } A(s), \\
M_i(\theta;s) &:= E(\exp[\theta A(s)]|X(0)=i) && \text{conditional mgf of input} \qquad (1) \\
L_n(s) & && \text{empirical measure of } n \text{ sources at time } s. \qquad (2)
\end{aligned}$$

Notice that $L_n(s)_i = Y_i(s)/n$, the fraction of traffic sources in state i in the simulation procedure of Section 3.2. When we wish to discriminate between sources, we write $A_\ell(s)$ for the input of the ℓ-th source.

Next we scale buffer size and capacity: $b = (B - B_0)/n$ and $c = C/n$. We assume that each source-chain X is distributed initially at time 0 according to α, for instance the most likely distribution given in Section 3.2. Then one can show by applying Cramér's theorem ([8, Section 1.2]) that

$$\begin{aligned}
&\lim_{n\to\infty} \frac{1}{n} \log H_n \qquad (3) \\
&\quad = \lim_{n\to\infty} \frac{1}{n} \log P\Big(\frac{1}{n}\sum_{\ell=1}^{n} A_\ell(t) \geq b + ct \Big| L_n(0) = \alpha\Big) \\
&\quad = -\sup_\theta \Big[\theta(b+ct) - \sum_{i=1}^{d} \alpha_i \log M_i(\theta;t)\Big]. \qquad (4)
\end{aligned}$$

Let θ^* solve the last optimization program. It is well known that the function in (4) as a function of θ is concave, and thus is $\theta^* > 0$ unique. We let the new measure be such that the source-chain X becomes inhomogeneous. Its matrix of transition rates at time s, $\Lambda(s)$, has entries

$$\lambda_{ij}(s) := \lambda_{ij}\frac{M_j(\theta^*;t-s)}{M_i(\theta^*;t-s)}, \quad \text{for} \quad i \neq j \text{ and } s \in [0,t]. \qquad (5)$$

With these transition rates we executed simulations in the same model as Example 3.1. Again so many simulation runs are done until RE ≈ 0.0765. The required number is compared to the corresponding in the ordinary simulations. The table clearly indicates the huge improvement in simulations runs. In the next subsection we will make remarks on the simulation times.

n	B	B_0	C	I_0	t	standard	ImpSam
10	1	0	25	25	0.25	9681	767
20	2	0	50	50	0.25	19269	624
30	3	0	75	75	0.25	1260234	1339
40	4	0	100	100	0.25	2236711	1267
20	2	1	50	55	0.15	4454	496
40	4	2	100	110	0.15	27715	755
60	6	3	150	165	0.15	146129	1096
80	8	4	200	220	0.15	726271	1324

Table 2. Required number of runs in Example 3.1.

4.2 Implementation Issues

The simulation-chain Y is time-inhomogeneous —under the new measure— with transition rates $y_i\lambda_{ij}(s)$. The consecutive states in the simulations are realized by applying uniformization. First, let us indicate how to find the minimal uniformization factor

$$\gamma_{\min} := \max_{s\in[0,t]} \max_{y} \sum_{\substack{i,j \\ i\neq j}} y_i\lambda_{ij}(s).$$

Define

$$\lambda_i(s) := \sum_{j:j\neq i} \lambda_{ij}(s), \quad i = 1, 2, \ldots, d. \tag{6}$$

Conjecture 1 *On* $[0,t]$ *is*
(i) $\lambda_i(s)$ *increasing if* i *is an on-state,*
(ii) $\lambda_i(s)$ *decreasing if* i *is an off-state.*

From the expressions (1), (5) and (6) we see that $\lambda_i(t) = \lambda_i := \sum_{j:j\neq i}\lambda_{ij}$. Hence,

$$\gamma_{\min} = n\Big[\max_{i:\text{on-state}} \lambda_i, \max_{i:\text{off-state}} \lambda_i(0)\Big]$$

Then set $\gamma \geq \gamma_{\min}$, and let the jumps of Y occur at the epochs of a Poisson(γ) process.

When a jump occurs at time s, we could use the transition probabilities $y_i\lambda_{ij}(s)/\gamma$ for the transition $y \rightsquigarrow y - e_i + e_j$ $(i \neq j)$. But that would mean a large number of calculations. At any jump epoch s we would have to calculate the conditional mgf's $M_i(\theta^*; t-s)$. And these are numerically calculated via the exponential matrix

$$B(\theta^*; t-s) := \exp\big[(\Lambda + \theta^* R)(t-s)\big], \tag{7}$$

where $R = \operatorname{diag}\{r\}$, see Mandjes & Ridder [6, Section 6]. Therefore we choose for the following alternative. We divide interval $[0,t]$ in K subintervals. Say that $[t_k, t_{k+1}]$ is the k-th subinterval. Then all transition probabilities of jumps occurring in this subinterval obey the same rule. We implemented for the k-th subinterval the averages of its end points:

$$y_i\big(\lambda_{ij}(t_k) + \lambda_{ij}(t_{k+1})\big)/2\gamma.$$

We have tested several combinations of uniformization constants γ and number of subintervals K in Example 3.1, in case the ordinary simulations require 31686 runs in 378sec. (The buffer data are $n = 50, B = 5, B_0 = 2.5, C = 125, I_0 = 137.5, t = 0.15$.) Table 3 lists the results. The case of applying rate matrix $\Lambda(s)$ at any jump epoch s is represented by the two rows of $K = \infty$. Each cell of the table shows both the number of required runs and the simulation time (again until RE ≈ 0.0765). From the results of these experimants we conclude that a small number of subintervals K and the smallest uniformization constant γ suffice for obtaining efficient estimates. Either increasing K of γ does not give much improvement in runs, but the simulation time degrade.

The idea of the importance sampling is to generate realizations which 'mimic' the optimal path. Hence, we conjectured that a large number of jump epochs and applying each time an update of the rate matrix $\Lambda(s)$ would give the best results in terms of required runs. Table 3 shows that this is true, but the improvements are marginally.

K	$\gamma_{\min}$	$1.5\gamma_{\min}$	$2\gamma_{\min}$	$2.5\gamma_{\min}$
5	767	639	744	634
	102	89	107	95
10	711	707	557	628
	180	183	147	170
15	669	691	609	656
	235	251	224	247
20	669	569	572	602
	301	267	276	295
∞	674	682	603	628
	906	1389	1626	2128

Table 3. Required number of runs and simulation time in seconds.

5 Optimal Path

The optimal or most likely path to overflow is conceptually the realization of the simulation-chain Y on time period $[0, t]$ such that, when there would be a very large ('infinite') number of traffic sources, and when it is given that an overflow occurs at time t, it happens ('most likely') along this particular realization.

To formalize this idea, we consider the space $\mathcal{M}$ of probability measures on $\{1, 2, \ldots, d\}$, and paths $f : [0, t] \to \mathcal{M}$. $f_i(s)$ represents a realization of the fraction of sources in state i at time s. Recall the large deviations *decay rate* of the overflow probability (expression (3)), and the empirical distributions (expression (2)). Then one can show that there is a particular path f^* such that

$$\lim_{n\to\infty} \frac{1}{n} \log H_n = \lim_{n\to\infty} \frac{1}{n} \log P\big(L_n(s) \approx f^*(s);\ s \in [0, t]\big).$$

Therefore, f^* is the most likely path. A formal treatment lies in the theory of large deviations for sample paths and is described in detail in Shwartz & Weiss [8]. However, their method gives an implicit representation of the optimal path, viz. the solution of a variational problem in function space. In [6] we determined the optimal path explicitly in the Markov fluid model with a problem which is slightly different from the one we consider here. Namely, we answered the questions, (i) what is the most likely overflow time t^* given that the system started in equilibrium (at time $-\infty$) and the buffer started to fill at time 0, and (ii) what is the optimal path of this event? We conjecture that the current problem, described in Section 3, can be tackled by the same approach as in [6], and that we obtain the following expression for the optimal path.

Conjecture 2 *Assume that $\alpha \in \mathcal{M}$ is the initial distribution of any source-chain X, and assume overflow time t is small. Then for $s \in [0,t]$ and $j \in \{1,2,\ldots,d\}$:*

$$f_j^*(s) = \sum_{i=1}^{d} \frac{\alpha_i}{M_i(\theta^*;t)} B_{ij}(\theta^*;s) M_j(\theta^*;t-s), \tag{8}$$

where θ^* solves the sup program (4), $M_i(\cdot)$ is the conditional mgf (1), and $B_{ij}(\cdot)$ the entries of the exponential matrix (7).

Using (8) it is easy to show that the optimal path satisfies the (inhomogeneous) Kolmogorov's forward differential equations:

$$f_i'(s) = \sum_{j:j\neq i} f_j(s)\lambda_{ji}(s) - \sum_{j:j\neq i} f_i(s)\lambda_{ij}(s),$$

with transition rates $\lambda_{ij}(s)$ given in (5). Hence, the realizations in importance sampling simulations mimic the optimal path. And therefore we state the following.

Conjecture 3 *The change of measure induced by the transition rates (5) gives asymptotically optimal estimates of the overflow probability.*

6 Conclusions and Further Research

We have considered an importance sampling procedure in estimating small overflow probabilities by simulations. Large deviations analysis of the model gives us the optimal path, i.e., the most likely way how overflow has occurred. The optimal path suggests the change of measure which induces the nonhomogeneous transition rates $\Lambda(s)$ of the simulation-chain. Further investigation is needed to prove formally that the importance sampling estimates are asymptotically optimal. Furthermore, the optimal path expression (8) in Section 5 is valid for small overflow times t only. Current work in progress is concerned about the optimal path for larger t. That means also that the importance sampling procedure should be adapted for larger t-values.

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 R.Y. Rubinstein. Steady state rare events simulation in queueing models and its complexity properties. In J.H. Dshalalow (ed.), *Advances in Queueing, Theory, Methods, and Open Problems*, 429–461. CRC press, Boca Raton, Florida , 1995.

[3] P. Glasserman, P. Heidelberger, P. Shahabuddin & T. Zajic. A large deviations perspective on the efficiency of multilevel splitting. *IEEE Transactions on Automatic Control*, 43:1666–1679, 1998.

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

[5] L. Kosten. Stochastic theory of a multi-entry buffer, part 1. *Delft Progress Report, Series F*, 1:44–50, 1974.

[6] M. Mandjes and A. Ridder. Optimal trajectory to overflow in a queue fed by large number of sources. To appear in: *Queueing Systems*, 1999.

[7] J. Roberts, U. Mocci, J. Virtamo (Eds.). *Broadband Network Teletraffic.* Performance Evaluation and Design of Broadband Multiservice networks. Final Report of Action COST 242. Lecture Notes in Computer Science 1155. Springer, Heidelberg, 1996.

[8] A. Shwartz and A. Weiss. *Large deviations for performance analysis, queues, communication, and computing.* Chapman and Hall, New York, 1995.

[9] T. Tanaka, O. Hasida and Y. Takahashi. Transient analysis of fluid models for ATM statistical multiplexer. *Performance Evaluation*, 23:145–162, 1995.

[10] M. Villén-Altamirano and J. Villén-Altamirano. RESTART: a method for accelerating rare event simulations. In J.W. Cohen and C.D. Pack (eds.), *Queueing, performance and Control in ATM*, 71–76. Elsevier, Amsterdam, 1991.