Optimally Efficient Simulation of Buffer Overflows in Queues with Deterministic Service Times via Importance Sampling

M.R. FRATER*, J. WALRAND** and B.D.O. ANDERSON*

* Department of Systems Engineering, Australian National University
** Department of Electrical Engineering and Computer Science, University of California

Simply because of their rarity, the estimation of the statistics of buffer overflows in queueing systems via direct simulation is often very expensive in computer time. Past work on fast simulation using importance sampling has concentrated on systems with Poisson arrival processes and exponentially distributed service times. However, in practical systems, such as ATM switches, service times are often deterministic and constant. This paper demonstrates how one can generate an asymptotically optimal simulation system (in the sense of variance) for queues with deterministic service times and a variety of arrival processes.

Keywords: Large Deviations, Importance Sampling, Exit Problem, Simulation, Queueing Systems, ATM Switches

1 INTRODUCTION

In a queueing system with finite buffers, some proportion of customers arriving at any queue are lost due to buffer overflows. While this number will be small in a properly dimensioned system, it is of interest because there is often a large cost associated with such a loss. However, the very rarity of the event of losing a customer makes direct simulation very costly in terms of computer time, if not impossible. For some simple systems, such as the M/M/1 queue, it is possible to analytically calculate the mean time between overflows, and simulation is unnecessary. However, for more complex systems, it is not generally possible to calculate the recurrence times of buffer overflows.

Several authors have described methods of using importance sampling to improve the efficiency of simulations of rare events [1,2,3]. These approaches, based on Large Deviations theory, provide asymptotic optimality in the limit as the events of interest become infinitely rare. Such simulations are optimal in the sense that they minimize the variance of a probability estimator, and hence minimize the simulation time required.

A number of other approaches to the problem of finding rare event statistics appear in the literature, including that of Neuts [4], in which the recurrence times of the rare events are evaluated numerically, using linear algebraic techniques for distributions of *Phase Type*. The computational burden of this latter method, while potentially superior to direct simulation, could still be overwhelming for certain problems. Other authors (such as [5]) combine analytic techniques with simulation to obtain rare event statistics efficiently.

The use of importance sampling for estimating the statistics of buffer overflows in queueing networks is addressed by [2,3]. The emphasis in these works is M/M/1 queues and Jackson networks, and it is shown how one can find an asymptotically optimal simulation system for simulating buffer overflows in these systems. However, in practical configurations, it is more usual to have systems whose service time is both deterministic and constant. In this paper, we concentrate on queues with deterministic service times.

Section 2 describes the problem and summarizes a method that can be used for generating an asymptotically optimal simulation system for arbitrary arrival and service distributions, but with specific reference to queues with deterministic service times. Specific results for queues with deterministic service times are presented in Section 3. The arrival processes analysed are Poisson and batch Poisson. A systematic method for generalizing the results for a Poisson arrival stream to perform sub-optimally efficient simulation (but still with a significant speedup factor over direct simulation) of a system with a Markov modulated arrival process is also given. Some simulation results are presented in Section 4.

Work supported by Australian Telecommunications and Electronics Research Board (ATERB), and ANU Centre for Information Science Research (CISR).

Paper received 23 January 1990.
Final revision 21 May 1990.

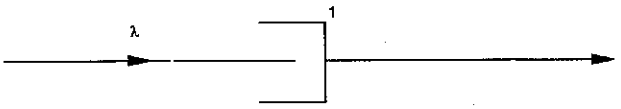


FIG. 1. – Queue With a Deterministic Server and Arrival Process with Average Rate λ.

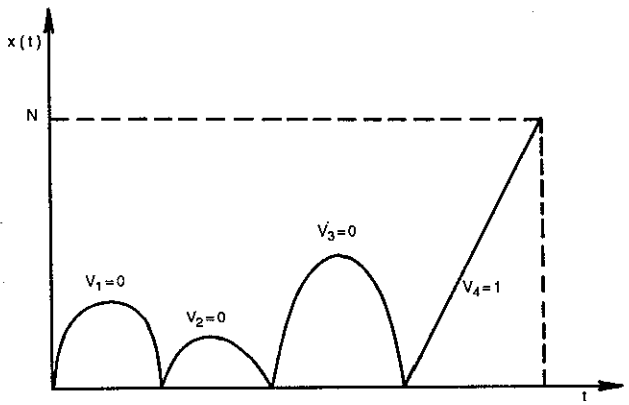


FIG. 2. – Typical Trajectory of a Queue.

2 PROBLEM FORMULATION

2.1 The Model

We consider a queue with a finite buffer of size N, a deterministic server, and some arrival process with average rate λ, as shown in Figure 1. We will assume without loss of generality that the virtual service rate is 1. By sampling the number of customers in the queue immediately after each virtual service, we can form a discrete-time Markov chain whose state is the number of customers in the queue.

If we run the system starting from the empty state, we will observe a trajectory something like that shown in Figure 2. We will use the term *cycle* to denote each piece of trajectory starting with the queue empty and ending with the first time that either the buffer is empty again or overflows. Let τ be the time for an overflow to occur, starting with the buffer empty, and α be the probability that a cycle ends in an overflow. Then we have:

$$\mathrm{E}[\tau] = \frac{\mathrm{E}[J_k]}{\alpha} \tag{1}$$

where J_k is the length of cycle k [2]. (We note that this assumes that the cycles are independent, as is common.)

The expected length of a cycle $\mathrm{E}[J_k]$ is of moderate size. Hence, this quantity is estimated easily via direct simulation. However, the probability that a cycle ends in an overflow α will be very small when overflows are rare. We will use importance sampling to create a system from which α can be estimated much more quickly than is possible via direct simulation.

2.2 Importance Sampling

The idea in importance sampling is as follows. Suppose that we are interested in certain (rare) events in a system S that we can simulate on a digital computer. Instead of simulating S, we simulate a second system $\overline{S}$, which has the property that the events in S and $\overline{S}$ correspond in some way. In particular, to the rare events A in S correspond events $\overline{A}$ in $\overline{S}$ (which may be the same as the events A). The correspondence is such that

1. the events $\overline{A}$ in $\overline{S}$ are more frequent that the events A in S, and
2. the connection between S and $\overline{S}$ allows one to infer $\mathrm{P}(A)$ if one knows $\overline{\mathrm{P}}(\overline{A})$. ($\overline{\mathrm{P}}(\overline{A})$ is the probability of the event $\overline{A}$ in $\overline{S}$.)

Let $V_k = \mathbf{1}_{\{\text{the buffer overflows in cycle } k\}}$.

Then in our original system S we have:

$$\mathrm{E}[V_k] = \alpha \tag{2}$$

Let L_k denote the likelihood ratio $\frac{dP}{d\overline{P}}$ during cycle k, i.e. the ratio of the probabilities of the trajectories under the measures P and $\overline{P}$ in S and $\overline{S}$. We observe that the L_k are i.i.d. and

$$\overline{\mathrm{E}}[L_k V_k] = \mathrm{E}[V_k] = \alpha \tag{3}$$

Hence, if we simulate the system $\overline{S}$ for p cycles, we can estimate the probability that a cycle ends in an overflow α from:

$$\widehat{\alpha} = \frac{L_1 V_1 + L_2 V_2 + \ldots + L_p V_p}{p} \tag{4}$$

Now we have not yet suggested how the system $\overline{S}$ might be chosen in order to ensure that a good speedup is obtained, or better still, to maximize the speedup obtained. Nor have we defined precisely what we mean by speedup. In many ways, we have replaced one difficult problem (finding the probability of overflow) with another.

2.3 Optimal Simulation – Large Deviations

The problem of finding the best system to use in importance sampling can be posed as an optimization problem as follows. Let A be a rare event for a system S, with $\alpha = \mathrm{P}(A) \ll 1$. For a direct Monte Carlo simulation involving n independent experiments, we could estimate α via:

$$\widehat{\alpha}_n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_A(\omega_i) \tag{5}$$

where the ω_i are the i.i.d. outcomes of the experiments, and $\mathbf{1}_A$ takes value 1 when the event A has occurred, and zero otherwise. The variance of $\hat{\alpha}_n$ is easily computed as

$$\mathrm{E}[\alpha - \hat{\alpha}_n]^2 = \frac{1}{n}(\alpha - \alpha^2) \tag{6}$$

Alternatively, consider a probability measure $\overline{P}$ associated with a system $\overline{S}$, with P absolutely continuous with respect to $\overline{P}$, such that the same event spaces apply for S and $\overline{S}$. Using $\overline{S}$ we can obtain a second estimate

$$\hat{\overline{\alpha}}_n = \frac{1}{n}\sum_{i=1}^{n} 1_A(\overline{\omega}_i)L(\overline{\omega}_i) \tag{7}$$

where $L = \frac{dP}{d\overline{P}}$ and the $\overline{\omega}_i$ are the i.i.d. outcomes of n experiments using $\overline{S}$. The variance of $\hat{\overline{\alpha}}$ is different to (6), and is obtainable as

$$\frac{1}{n}\left(\int_A L^2(\omega)d\overline{P}(\omega) - \alpha^2\right) \tag{8}$$

We want this to be as accurate as possible. So we want to adjust all the probabilities in S to new ones in $\overline{S}$ so that

$$(\sigma^*)^2 = \int_A L^2(\omega)d\overline{P}(\omega) \tag{9}$$

is minimized. This corresponds to minimizing the time necessary for simulation. In fact, the system $\overline{S}$ that we will find will be asymptotically optimal in the limit as the buffer size tends to infinity.

Given a system $\overline{S}$ minimizing $(\sigma^*)^2$, we can use (4) to find the value of α for the original system S from (much faster) simulation performed on $\overline{S}$.

Let $x(k)$ be the state of a Markov chain formed by sampling the system S, which is now assumed to be a queue. Recall that in Section 2.1, we indicated that with a deterministic server, sampling occurs just after each virtual service; in general, the precise details of the sampling will vary with the arrival and service distributions, and will be outlined separately later in the text. We assume that the state-transition equation for $x(\cdot)$ can be written in the form

$$x(k+1) = x(k) + w(k) \tag{10}$$

where $w(\cdot)$ is a random process, defined such that $\mathrm{E}[w(k) \mid x(k)] < 0$ for $x(k) > 0$, and with an appropriate boundary condition at $x(k) = 0$ to prevent $x(k+1)$ becoming negative. Hence the Markov chain is asymptotically stable in the sense that, on average, its state will tend towards zero. In solving the optimization problem, we will ignore the boundary condition at $x(\cdot) = 0$ that stops the state going negative.

Let $F(\cdot)$ be the jump distribution of the Markov chain $x(\cdot)$ associated with the system S we wish to simulate. Its Cramer transform[1] $h(y)$ is given by:

$$h(y) = \inf_{s\in\Re}\left[sy - \log\int_{-\infty}^{\infty} e^{sz}\mathrm{d}F(z)\right]. \tag{11}$$

Let

$$V(T, y_0, \ldots, y_{T-1}) = \sum_{k=0}^{T-1} h(y_k) \tag{12}$$

where y_k is the value of y at time k. We wish to minimize $V(\cdot,\cdot)$ with respect to T and the y_k, subject to the constraint

$$\sum_{k=0}^{T-1} y(k) = N \tag{13}$$

It turns out that, so long as the distribution function $F(\cdot)$ is time and state invariant, the optimal value of y_k is a constant, and does not vary with k. Let y^* be the optimal value of the y_k. Then y^* is the unique positive solution of:

$$h(y^*) = y^*\frac{d}{dy}h(y^*) \tag{14}$$

This calculation is shown in the appendix. Also, the constraint (13) implies $y^* > 0$. As is shown in the appendix, because $h(\cdot)$ is convex, there is always exactly one positive solution of (14).

It has been shown (see e.g. [1,2] for a continuous-time form) that this value of y^* is the average rate of increase of the asymptotically optimal simulation system of S. That is, if we denote the state of the optimal simulation system by $\overline{x}$ and its state-transition equation is:

$$\overline{x}(k+1) = \overline{x}(k) + \overline{w}(k) \tag{15}$$

then

$$\mathrm{E}[\overline{w}(k)] = y^* \tag{16}$$

(In examples below, we shall indicate how to find the distribution of $\overline{w}$.) Note that because $y^* > 0$, the simulation system is unstable in the sense that its state will, on average, increase with time.

It is (14) that will be used in the following sections to find the parameters of the optimal simulation system for a number of types of queue with deterministic service times in Section 3. The simple example of the M/M/1 queue is given in the next section.

[1] More information on the Cramer transform can be found in, for example, [2,3,6].

2.4 Simple Example – M/M/1 Queue

Before presenting the original results of the paper, dealing with queues with deterministic service times, we will first summarize how the results of the previous section, and in particular (14), can be applied to find an optimal simulation system for buffer overflows in an M/M/1 queue.

Given an M/M/1 queue, with Poisson arrival stream at average rate λ and exponentially distributed service times with parameter μ, we wish to find a new system that we can use to find the probability of overflow α with the least cost in simulation time. We assume without loss of generality that $\lambda + \mu = 1$. We form a Markov chain by sampling the state of the buffer immediately after each arrival or service takes place. Ignoring the boundary condition at $x = 0$, the transition function can be written:

$$x(k+1) = x(k) + \begin{cases} 1 & \text{probability } \lambda \\ -1 & \text{probability } \mu \end{cases} \tag{17}$$

The Cramer transform of the Bernoulli jump distribution associated with this queue is [2]:

$$h(y) = \frac{1}{2}\left[(1+y)\log\frac{1+y}{2\lambda} + (1-y)\log\frac{1-y}{2\mu}\right] \tag{18}$$

Substituting for $h(\cdot)$ in (14), and rejecting the solution with $y < 0$, it turns out that:

$$y^* = \mu - \lambda \tag{19}$$

Now, if our optimal simulation queue has arrival rate λ^* and service rate μ^*, the average rate of increase of this system is $y = \lambda^* - \mu^*$. If we assume, without loss of generality, that $\lambda^* + \mu^* = 1$, then we have, on account of the claim immediately following (12), as well as (19):

$$\lambda^* = \mu \tag{20}$$
$$\mu^* = \lambda \tag{21}$$

which corresponds to swapping the arrival and service rates in passing from the original system to the optimal simulation system. This is well known as the optimal simulation system for simulating buffer overflows in an M/M/1 queue, see [2].

For an M/M/1 queue, the likelihood ratio for the k'th cycle is given by

$$L_k = \left(\frac{\lambda}{\mu}\right)^N \tag{22}$$

This follows by an easy calculation set out in [2].

3 OPTIMAL SIMULATION SYSTEMS

3.1 M/D/1

If we sample the output of an M/D/1 queue, having Poisson arrival stream with rate λ and deterministic service rate 1, immediately after each service, the probability that the state of the queue has increased by z is (for $z \geq -1$):

$$\frac{1}{(z+1)!}\lambda^{z+1}e^{-\lambda} \tag{23}$$

The associated jump distribution is evidently:

$$\mathrm{d}F(z) = \sum_{i=-1}^{\infty} \frac{1}{(z+1)!}\lambda^{z+1}e^{-\lambda}\delta(z-i)\mathrm{d}z \tag{24}$$

We can now use (18) to find the Cramer transform of the distribution:

$$h(y) = (y+1)\log\frac{y+1}{\lambda} + \lambda - (y+1) \tag{25}$$

and hence the derivative of $h(\cdot)$ is:

$$h'(y) = \log\frac{y+1}{\lambda} \tag{26}$$

Hence, the average rate of increase of the state of the optimal simulation system (i.e. y^*, the optimal value of y) is the unique positive solution of:

$$\log\frac{y^*+1}{\lambda} + \lambda - (y^*+1) = 0 \tag{27}$$

Now, we can choose the deterministic service rate in the simulation system as we please, provided that we do it in such a way that (27) has a solution. For convenience, we will make the service rate in the simulation system the same as that in the original system, i.e. 1. Hence, the optimal value of the arrival rate in the simulation system is $\lambda^* = y^* + 1$, and $y^* > 0$ requires $\lambda^* > 1$. Replacing (y^*+1) by λ^* in (27), it can be seen from the plots in Figure 3 that there is exactly one solution for λ^* for which the arrival rate is greater than the service rate.

If we take a second order Taylor series expansion of the logarithm terms in (27), we can find an approximate explicit equation for λ^*:

$$\lambda^* = 2 - \lambda \tag{28}$$

for λ close to 1.

For an M/D/1 queue, the likelihood ratio for the k'th cycle is readily found as

$$L_k = \left(\frac{\lambda}{\lambda^*}\right)^N \tag{29}$$

where ℓ_k is the number of time steps in cycle k.

3.2 Batch-Poisson Arrival Process

If we take an M/D/1 queue, as described in the previous section, with Poisson arrival stream with rate λ, but allow each arrival to be a batch, whose length (i.e. the number of customers in the batch) is distributed according to some distribution, say $F_a(\cdot)$, we have what is called a "batch-Poisson" arrival process. We will proceed with the analysis as in the previous section, firstly for a completely general batch-length distribution $F_a(\cdot)$, and then for the case where $F_a(\cdot)$ is an exponential distribution. For convenience, we will assume that the whole of a batch arrives at the one time, rather than spread out in time as would be the normal case. This assumption avoids the need to analyse overlapping batches, and is justified provided that the batch lengths are small compared to the maximum length of the queue (N), and that the inter-arrival times of customers within a batch are not too long.

As before, we sample the state of the queue immediately after each service, and use the resulting Markov chain for our analysis. In any one second, the state of the queue will change by:

$$x_1 + \ldots + x_J - 1 \tag{30}$$

where the x_i are the batch lengths, distributed according to $F_a(\cdot)$, and J, which has a Poisson distribution with average rate λ, is the number of batches arriving in this second. (It should be noted that the x_i are i.i.d.) The Cramer transform of this jump distribution can be written:

$$\begin{aligned} h(y) &= \sup_{s\in\Re}\left(sy - \log \mathrm{E}\left[e^{s(x_1+\ldots+x_J-1)}\right]\right) \quad (31)\\ &= \sup_{s\in\Re}\left(sy - \log e^{-s}\mathrm{E}_F\left[\left(\mathrm{E}_{F_a}\left[e^{sx}\right]\right)^J\right]\right)\\ &= \sup_{s\in\Re}\left(sy - \log e^{-s}\int_{-\infty}^{\infty}[B(s)]^z \bullet \right.\\ &\qquad \left.\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}e^{-\lambda}\delta(z-k)dz\right)\\ &= \sup_{s\in\Re}\left(sy - \log e^{-s-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k B^k(s)}{k!}\right)\\ &= \sup_{s\in\Re}\left(sy + s + \lambda - \lambda B(s)\right) \quad (32) \end{aligned}$$

where

$$B(s) = \int_{-\infty}^{\infty} e^{sx}\mathrm{d}F_a(x) \tag{33}$$

is the expectation of e^{sx} with respect to $F_a(\cdot)$. The subscript on the expectation denotes the distribution on which the expectation is based. The Cramer transform of the jump distribution associated with the M/D/1 queue corresponds to the degenerate case where $B(s) = e^s$.

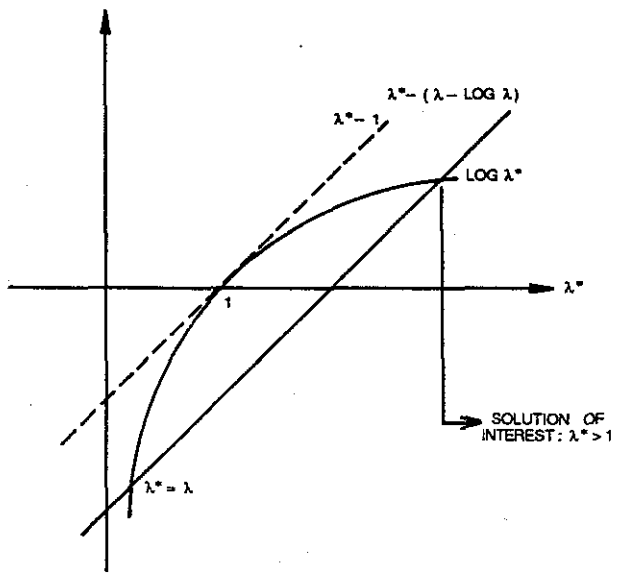


FIG. 3. — Plots of Log λ^* and $\lambda^* - (\lambda - \text{Log } \lambda)$.

Given the distribution $F_a(\cdot)$ of the batch lengths, in principle we can use (32) to find the Cramer transform of the jump distribution of the Markov chain, and hence to solve (14) for the optimal rate of increase of the simulation system. However, in our experience, even for simple distributions the algebra is complicated, and in practice it is likely that we will only be able to find direct solutions in a few special cases.

The special case presented here is where the batch lengths are exponentially distributed. No justification is offered for this choice except that it appears to be the easiest to analyse. In practical terms, it has several disadvantages, including the fact that it allows batch lengths of zero, and also non-integer batch lengths. However, the use of continuous batch length distributions is not completely unknown in the literature, (see e.g. [7].)

Where $F_a(\cdot)$ is an exponential distribution with parameter a, the mean of $F_a(\cdot)$ will be $\frac{1}{a}$ and the mean arrival rate will be $\frac{\lambda}{a}$, which must be less than 1 for the queue to be asymptotically stable and for overflows to be rare events. The expectation of e^{sx} for this distribution, B(s), is given by:

$$B(s) = \frac{a}{a-s} \qquad \text{for } s < a \tag{34}$$

Therefore,

$$h(y) = \sup_s\left[sy + s + \lambda - \frac{\lambda a}{a-s}\right], \tag{35}$$

and substituting for $h(\cdot)$ in (14) yields two possible solutions:

$$y^* = \begin{cases} \frac{a}{\lambda} - 1 & \text{(which is greater than zero)} \\ \frac{\lambda}{a} - 1 & \text{(which is less than zero)} \end{cases} \tag{36}$$

If λ^* and a^* are the parameters of the arrival process in the simulation system corresponding to λ and a in the original system, then the mean rate of customers arriving in the simulation system will be $\frac{\lambda^*}{a^*}$. Taking the service rate in the optimal simulation system to be 1, as before, we have:

$$y^* = \frac{\lambda^*}{a^*} - 1 \tag{37}$$

and hence, noting that we require $y^* > 0$, we have:

$$\frac{\lambda^*}{a^*} = \frac{a}{\lambda} \tag{38}$$

Even though we know the value of $\frac{\lambda^*}{a^*}$, we have no way of finding either λ^* or a^* separately from the Large Deviations analysis. In other words, we can find the total average arrival rate for the simulation system, but cannot separate this into the optimal average rate of arrival of batches and the optimal average batch length. This may seem to be a problem, in that the batch length will certainly affect tail probabilities, and hence one would expect that the speedup would depend on the relative choice of λ^* and a^*. However, we have assumed that the mean batch length $\frac{1}{a}$ is much less than the buffer size, i.e. that this mean is small. Under these circumstances, the tail probabilities for long sample lengths should be very similar. Hence, we can choose λ^* and a^* as we like within the constraint of (38), without affecting the asymptotic speedup. However, smaller additional speedup factors may be obtained by adjusting the relative values of λ^* and a^*. By this, we mean that the additional speedup will not be exponential in the buffer size, and hence does not show up in the Large Deviations analysis presented above.

We have criticised above the assumption of exponentially distributed batch lengths. In an attempt to partially counter the criticism, we examined batches whose length was distributed according to a shifted exponential distribution (i.e. the probability density is $[\exp -a(x-1)]$ for $x \geq 1$,) but were unable to compute the solution analytically.

3.3 MM/D/1

The Markov-modulated Poisson arrival process is a generalization of the Poisson, in which at any given time the arrival stream is Poisson, but where the average rate of the stream varies with time according to the state of some Markov process. While it has been possible to directly obtain solutions to the optimization problem for Poisson and batch-Poisson arrival streams, the case of a Markov-modulated Poisson arrival process presents significantly greater difficulties.

However, while global optimality appears to be difficult to achieve, since at any given time the input stream is Poisson, we can easily achieve a form of local optimality by taking the solution of (27) for the simulation system, i.e. at each time step, we behave as if the average arrival rate is constant, and hence pretend that we are simulating an M/D/1 queue. For example, suppose that the arrival rate in the original system is switched between λ_1 and λ_2 in a Markov manner. Let y_1^* and y_2^* be solutions of (27) when $\lambda = \lambda_1$ and λ_2, and let $\lambda_1^* = y_1^* + 1$ and $\lambda_2^* = y_2^* + 1$. Then the simulation system will switch its rates just as for the original system, save that it uses the rates λ_1^* and λ_2^* instead of λ_1 and λ_2.

The following assumptions appear to be implicit in claiming that this is a "good" thing to do globally:

- that the modulating process does not depend directly on time;
- that overflows occur due to "malicious noise sequences" in the Poisson arrival stream, rather than due to the behaviour of the modulating Markov process.

There is no reason to suspect that this approach will provide asymptotic optimality in general, but it is to be expected that significant speedup over direct simulation would be obtained.

4 SIMULATION RESULTS

A number of simulations were performed on M/D/1 queues in order to ascertain both the amount of speedup that is obtained by estimating the overflow probability α via simulation of the optimal systems described above, rather than direct simulation, and also the amount of computer time required for the fast simulation. All simulations were run until the relative standard deviation was less than 0.1, corresponding to a 95% confidence of the estimation error being less than 20%.

Table 1 shows the increases in speed obtained, as well as the number of simulation steps required for the various simulations. On the Sparcstation 1 on which these simulations were run, each simulation step took approximately 2.5 microseconds. Hence, it can be seen that if the buffer size grows large, then direct simulation becomes impossible very quickly. Even for the small buffer size (10) used in these simulations, direct simulation of the two most lightly loaded cases would have required over one hour of CPU time and over one billion simulation steps.

The table illustrates that the more "stable" the original system is, i.e. the smaller λ is and the larger the recurrence time of buffer overflows, the more "unstable" is the importance sampling system (and thus the shorter is the simulation time, and of course the greater the speedup factor). This phenomenon appears quite general in its occurrence in importance sampling applied to queues.

Table 1 – Simulation results. The buffer size is 10 in all cases.

λ	λ^*	$\widehat{\alpha}$	Simulation Times		Speedup
			Direct*	Fast	
0.1	3.71495	3.7×10^{-17}	$\gg 10^9$	37	$\gg 10^7$
0.3	2.36456	3.2×10^{-10}	$\gg 10^9$	80	$\gg 10^7$
0.5	1.75643	2.5×10^{-6}	130630807	142	9.2×10^5
0.7	1.37547	4.4×10^{-4}	760922	1466	520

*The simulations were run for a maximum of 10^9 steps.

5 CONCLUSION

This paper has presented methods using importance sampling and Large Deviations theory for performing asymptotically optimal simulation of queues with deterministic servers and a variety of different arrival processes.

These ideas have yet to be extended to networks of queues, and the relationship between the arrival and service processes and the optimal control problem used to generate the optimal simulation system needs to be more fully explored.

A APPENDIX: SOLUTION OF MINIMIZATION PROBLEM

In this section, we present the calculation behind the result used in the main text to find the rate of increase of the state of the optimal simulation system. The following properties of the Cramer transform will be used [1,2,3]:

1. $h''(x) > 0$, i.e. $h(\cdot)$ is strictly convex;
2. $h(\cdot) \geq 0$;
3. $h(m) = 0$;
4. $h'(m) = 0$.

where m is the mean of the distribution of which $h(\cdot)$ is the Cramer transform.

Let

$$V(T, y_0, \ldots, y_{T-1}) = \sum_{k=0}^{T-1} h(y_k) \tag{A.1}$$

We wish to minimize $V(\cdot,\cdot)$ with respect to T and the y_k, subject to the constraint:

$$\sum_{k=0}^{T-1} y_k = N \tag{A.2}$$

The lagrangian is given by:

$$\mathcal{L} = \sum_{k=0}^{T-1} h(y_k) - g\left[\sum_{k=0}^{T-1} y_k - N\right] \tag{A.3}$$

where g is the Lagrange multiplier.

We temporarily fix T. Differentiating $\mathcal{L}$ with respect to y_k, and equating this to zero, we obtain:

$$\frac{\partial \mathcal{L}}{\partial y_k} = h'(y_k) - g \tag{A.4}$$

$$= 0 \tag{A.5}$$

Let y_k^* be the optimal value of y_k. So long as $h'' > 0$, (A.4) implies:

$$y_k^* = y^* \qquad \forall k \in [0, T-1] \tag{A.6}$$

for some unique y^*, yet to be identified. Hence, the equation for the lagrangian (A.3) can be rewritten:

$$\mathcal{L} = Th(y^*) - gTy^* + gN \tag{A.7}$$

and its derivative with respect to the escape time T:

$$\frac{\partial \mathcal{L}}{\partial T} = h(y^*) - gy^* \tag{A.8}$$

$$= 0 \tag{A.9}$$

Substituting for g using (A.4), at the optimal value of y the cost function $h(\cdot)$ and its derivative are related by:

$$h(y^*) = y^* \frac{\mathrm{d}}{\mathrm{d}y} h(y^*) \tag{A.10}$$

i.e., the line tangent to $h(\cdot)$ at y^* passes through the origin.

As was stated above, the Cramer transform $h(\cdot)$ is convex (property 1). Hence, there are exactly two solutions to (A.10). This can be seen from the plot in Figure 4. The first is negative, corresponding to a stable system, in the sense that the state tends towards zero. The average rate of descent of this system is the same as that of the original system, by property 3 of the Cramer transform. This solution does not satisfy the constraint (A.2), and therefore will be discarded. The other solution, for which y^* is positive, is the solution that we seek, and is the average rate of increase of our optimal simulation system.

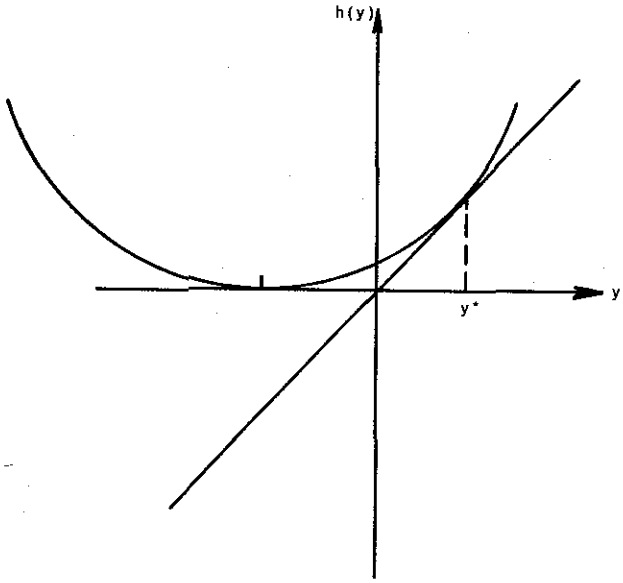


FIG. 4. — The Behaviour of the Cost Function at the Point Minimizing the Performance Index Subject to the Constraints.

REFERENCES

[1] Cottrell, M., Fort, J.C. and Malgouyres, G., "Large Deviations and Rare Events in the Study of Stochastic Algorithms," *I.E.E.E. Trans. Automatic Control*, Vol. AC-28, September 1983, pp. 907–918.

[2] Parekh, S. and Walrand, J., "A Quick Simulation of Excessive Backlogs in Networks of Queues," *I.E.E.E. Trans. Automatic Control*, Vol. 34, January 1989, pp. 54–66.

[3] Frater, M.R. and Anderson, B.D.O., "Fast Estimation of the Statistics of Excessive Backlogs in Tandem Networks of Queues," *Australian Telecommunication Research*, Vol. 23, No. 1, May 1989, pp. 49–55.

[4] Neuts, M.F., *Matrix-geometric Solutions in Stochastic Models*, Baltimore, Maryland: Johns Hopkins University Press, 1981.

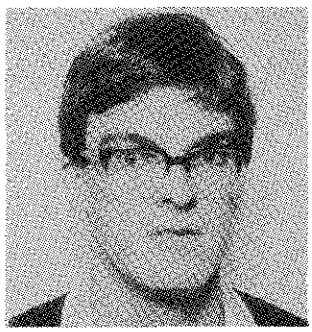
[5] Frost, V.S., Larue, W.W. and Shanmugan, K.S., "Efficient Techniques for Simulation of Computer Communications Networks," *I.E.E.E. J. Select. Areas Commun.*, Vol. 6, January 1988, pp. 146–157.

[6] Freidlin, M.I. and Wentzell, A.D., *Random Perturbations of Dynamical Systems*, New York: Springer-Verlag, 1984.

[7] Wolff, R.W., *Stochastic Modeling and the Theory of Queues*, Englewood Cliffs, New Jersey: Prentice Hall, 1989.

[8] Bertsekas, D.P. and Gallager, R.G., *Data Networks*, Englewood Cliffs, New Jersey: Prentice-Hall, 1987.

BIOGRAPHIES



Michael R.FRATER received the B.Sc. degree in mathematics and physics (1986) and the B.E. degree in electrical engineering (1988) from the University of Sydney.

He is currently a graduate student in the Department of Systems Engineering at the Australian National University, working towards the Ph.D. degree. His research interests include aspects of communications, signal processing and control.

Prof. J.WALRAND. (Biography not available.)

Brian D.O. ANDERSON received the B.Sc. degree in mathematics and the B.E. degree in electrical engineering from the University of Sydney, and the Ph.D. degree from Stanford University in 1966.

He is currently Professor and Head of the Department of Systems Engineering at the Australian National University; from 1967 through to 1981 he was Professor of Electrical Engineering at the University of Newcastle. He is co-author of several books and his research interests are in control, telecommunications and signal processing.

Dr. Anderson is a Fellow of the Royal Society, the Australian Academy of Technological Sciences, the Australian Academy of Science and an Honorary Fellow of the Institution of Engineers, Australia. He is currently President-Elect of IFAC.