Proceedings of the 1992 Winter Simulation Conference
ed. J. J. Swain, D. Goldsman, R. C. Crain, and J. R. Wilson

APPLICATION OF FAST SIMULATION TECHNIQUES TO SYSTEMS WITH CORRELATED NOISE*

Michael R. Frater

Department of Electrical Engineering
University College
Australian Defence Force Academy
Campbell ACT 2600
Australia

ABSTRACT

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. In addition, arrivals from services, such as variable bit rate video are often correlated. These may be modeled by an auto-regressive process. This paper demonstrates how one can generate an asymptotically optimal simulation system (in the sense of variance) for queues with deterministic service times and auto-regressive arrival processes.

1 INTRODUCTION

In a queueing system with finite buffers, some proportion of customers arriving at any queue is 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.

In broadband ISDN, all services, whether voice, video or data, will be transferred on a packet network. The data will be broken up into cells, each containing 48 bytes of information. Because the cell size is deterministic and constant, a queueing model with deterministic virtual service times is required to model the switching elements in such a network. For such queueing models, there are no simple closed form solutions for quantities such as the invariant probability, and calculation of cell loss rates is typically not feasible.

In such a network, it is not possible to completely prevent cell loss. Because of constraints imposed by the different service types, it is expected that the cell loss probability will be of the order of 10^{-9}. In a real network, these losses might be expected to occur on timescales of the order of minutes or hours. However, in estimating the loss rate by simulation on a digital computer, it is possible that years of CPU time may be required, even on a fast computer, such as a Sun SPARCstation (see e.g. Frater (1990)).

With these required simulation times, it is fair to say that simulation is not just difficult, but impossible. In digital computers, the effects of pseudo-random number generators can make simulation unreliable where a large number of calls to the pseudo-random number generator are involved, *even for simulations that may be feasible in terms of computer time,* (see e.g. Heath and Sanchez (1986)). In fact, the main conclusion of Heath and Sanchez (1986) is that the period of the pseudo-random number generator must be at least some multiple of the *square* of the number of samples required. It is not difficult to construct examples of, for example, queueing networks where this criterion requires periods in excess of 10^{30}.

Several authors have described methods of using importance sampling to improve the efficiency of simulations of rare events, (see, e.g. Cottrell *et al* (1983), Parekh and Walrand (1986) or Frater and Anderson (1991)). These approaches, based on large de-

*Work supported by the Australian and Overseas Telecommunications Corporation, Research and Development Contract No. 7315.

viations 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.

The use of importance sampling for estimating the statistics of buffer overflows in queueing networks is addressed by Parekh and Walrand (1989), Frater and Anderson (1989) and Frater *et al* (1991). 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.

As described above, these assumptions are unsatisfactory in many practical applications, such as broadband ISDN. In such systems, it is more usual to have systems whose service time is both deterministic and constant. Some progress towards developing efficient techniques for queues with deterministic servers were reported in Frater *et al* (1990). In this paper, the application of large deviations theory and importance sampling to the simulation of buffer overflows in M/D/1 queues was described.

However, the arrival traffic generated by the various services anticipated for the broadband ISDN is not suitable for modelling by a Poisson process. Many alternative have been postulated. For example, it has been proposed by Maglaris *et al* (1988) that an autoregressive process be used for modeling the traffic generated by a variable bit rate video source. In this paper, we will demonstrate how the similar techniques to those applied in Frater *et al* 1990 can be used to provide efficient simulation of buffer overflows in a queue with such an arrival process and with a deterministic server.

2 PROBLEM FORMULATION

We consider a queue with a finite buffer of size N and a deterministic server with virtual rate μ. Let $\lambda(k)$ be the number of arrivals that occur during the kth sampling interval. We will assume that the transitions of $\lambda(\cdot)$ are controlled by a Markov process. By sampling, we can form a discrete-time Markov chain whose state is the number of customers in the queue. Let $x(\cdot)$ be a Markov chain formed by sampling the number of customers resident in the queue in an appropriate manner. (For example, we might sample the state immediately after each service.) Then $x(\cdot)$ evolves as:

$$x(k+1) = x(k) + \lambda(k) - \mu \tag{1}$$

We will assume that there is an appropriate boundary condition that prevents $x(\cdot)$ from becoming negative.

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[1]. Then we have:

$$\mathrm{E}[\tau] = \frac{\mathrm{E}[J_k]}{\alpha} \tag{2}$$

where J_k is the length of cycle k. (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. In this paper, we will describe an with efficient method for estimating α by simulation using importance sampling.

3 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:

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

Let L_k denote the likelihood ratio $\frac{\mathrm{d}P}{\mathrm{d}\overline{P}}$ 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

$$\overline{\mathrm{E}}[L_k V_k] = \mathrm{E}[V_k] = \alpha \tag{4}$$

[1] In many situations, other quantities, such as the cell loss probability, may be of more interest. In such cases, a slightly different technique will be required.

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:

$$\widehat{\alpha} = \frac{L_1 V_1 + L_2 V_2 + \ldots + L_p V_p}{p} \tag{5}$$

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.

4 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:

$$\widehat{\alpha}_n = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_A(\omega_i) \tag{6}$$

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 $\widehat{\alpha}_n$ is easily computed as

$$\mathrm{E}[\alpha - \widehat{\alpha}_n]^2 = \frac{1}{n}(\alpha - \alpha^2) \tag{7}$$

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

$$\widehat{\overline{\alpha}}_n = \frac{1}{n} \sum_{i=1}^{n} 1_A(\overline{\omega}_i) L(\overline{\omega}_i) \tag{8}$$

where $L = \frac{\mathrm{d}P}{\mathrm{d}\overline{P}}$ and the $\overline{\omega}_i$ are the i.i.d. outcomes of n experiments using $\overline{S}$. The variance of $\widehat{\overline{\alpha}}$ is different to (7), and is obtainable as

$$\frac{1}{n}\left(\int_A L^2(\omega)\mathrm{d}\overline{P}(\omega) - \alpha^2\right) \tag{9}$$

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

$$(\sigma^*)^2 = \int_A L^2(\omega)\mathrm{d}\overline{P}(\omega) \tag{10}$$

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 (5) to find the value of α for the original system S from (much faster) simulation performed on $\overline{S}$.

5 FAST SIMULATION SYSTEM FOR AUTOREGRESSIVE PROCESS

Let $x(k)$ be the state of a Markov chain formed by sampling the system S, which is now assumed to be a queue with a virtual service rate that is deterministic and constant. We assume that the state-transition equation for $x(\cdot)$ can be written in the form

$$\lambda(k+1) - \Lambda = a\,(\lambda(k) - \Lambda) + bw(k) \tag{11}$$

$$x(k+1) = x(k) + \lambda(k) - \mu \tag{12}$$

with $\lambda(0) = x(0) = 0$, $0 < a < 1$. $w(\cdot)$ is a white noise process. Λ is a constant that determines the equilibrium point of the arrival rate $\lambda(\cdot)$. We assume that there is an appropriate boundary condition that prevents both $\lambda(\cdot)$ and $x(\cdot)$ from becoming negative. We also assume that μ is large enough that 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 conditions that prevent the state from becoming negative.

Let $F(\cdot)$ be the jump distribution of the random process $w(\cdot)$ associated with the system S we wish to simulate. Its Cramer transform $h(y)$ is given by:

$$h(y) = \inf_{s \in \Re}\left[sy - \log \int_{-\infty}^{\infty} \mathrm{e}^{sz}\mathrm{d}F(z)\right]. \tag{13}$$

We define a new deterministic system

$$\lambda'(k+1) - \Lambda = a\,(\lambda'(k) - \Lambda) + by(k) \tag{14}$$

$$x'(k+1) = x'(k) + \lambda'(k) - \mu \tag{15}$$

Let

$$V(T, y(0), \ldots, y(T-1)) = \sum_{k=0}^{T-1} h(y(k)) \tag{16}$$

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 constraints

$$x'(0) = \lambda'(0) = 0 \tag{17}$$

$$x'(T) = N \tag{18}$$

In Cottrell *et al* (1983), it is shown that the solution of this optimal control problem with an infinite horizon (i.e. $t \rightarrow \infty$) defines the mean trajectory of the optimally efficient importance sampling simulation system.

For simplicity, we will assume that $w(\cdot)$ has a gaussian distribution, with unit variance. We will also assume that N is large, and hence that T is also large. Then (16) can be rewritten:

$$V(T, y(0), \ldots, y(T-1)) = \sum_{k=0}^{T-1} \frac{1}{2} y^2(k) \qquad (19)$$

This optimization problem would be a garden variety discrete time linear optimal control problem if:

1. the system $(x(\cdot), \lambda(\cdot))$ defined by (15) were controllable; and
2. $\mu = 0$.

Given that neither of these conditions holds, and that the standard solution for the optimal control problem is therefore not available, we will use the method of lagrange multipliers to solve this optimal control problem. We define the lagrangian:

$$\mathcal{L} = \sum_{k=0}^{T-1} \frac{1}{2} y^2(k) - g \times \left[\sum_{k=0}^{T-1} (\lambda'(k)) - T\mu - N \right] \qquad (20)$$

Let $y^\star(\cdot)$ indicate the optimal value of $y(\cdot)$. Then it can be shown that

$$y^\star(k) = 2(\mu - \Lambda)(1 - a)\left(1 - a^{T-k}\right) \qquad (21)$$

Letting $T \rightarrow \infty$, we have

$$y^\star(k) = 2(\mu - \Lambda)(1 - a) \qquad (22)$$

Let $\lambda^\star(\cdot)$ be the arrival process in the optimal simulation system, and $x^\star(\cdot)$ the number of customers resident in the queue. Then:

$$\lambda^\star(k+1) - \Lambda^\star = a(\lambda^\star(k) - \Lambda^\star) + bw(k) \qquad (23)$$

$$x^\star(k+1) = x^\star(k) + \lambda^\star(k) - \mu \qquad (24)$$

where

$$\Lambda^\star = 2\mu - \Lambda \qquad (25)$$

It should be noted that this analysis deals only with a single video source, and that it does not take into account interactions between sources that may have a large impact on the loss rate. The application of this analysis to systems involving multiple video sources is being investigated currently.

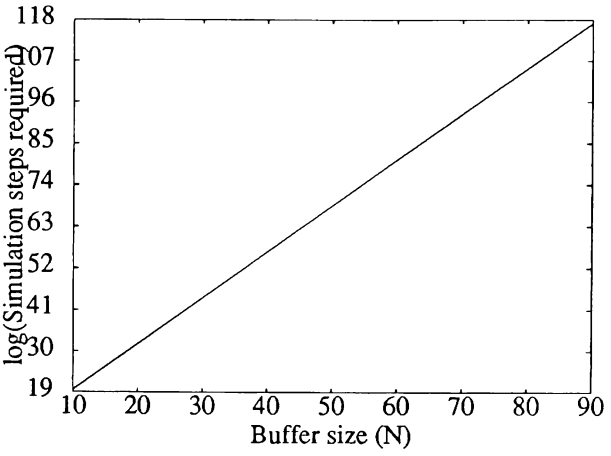

Figure 1: Simulation time vs buffer size for direct simulation.

6 SIMULATION RESULTS

In order to verify the above results, a number of simulations have been carried out. The results are described below. For convenience, in all cases the virtual service rate of the switch was taken to be 1. In addition, no attempt was made to model the fact that the number of cells generated in a video frame is always an integer, or multiple arrival processes feeding a single switch. Also, in a real network, the arrival of cells and their service is spread throughout a video frame. Here, we assumed that all cells arrive at the beginning of a video frame, and that all cells that are to be serviced during that frame are serviced immediately. All of these effects would need to be taken into account in any simulation study designed to estimate network performance.

Figure 1 shows how the simulation time required for direct simulation increases with buffer size (N) for a switch with virtual service rate 1, fed by an AR(1) process, as described above. The parameters of this process are $\Lambda = 0.5$, $a = 0.5$, $b = 0.3$, and $\Lambda^\star = 1.5$. These results were obtained using an importance sampling simulation. Clearly, with simulation times ranging from 10^{19} to 10^{115}, these simulations are impossible to perform using direct simulation.

We note that for the series of simulations shown in Figure 1, there is an exponential relationship between the required simulation time and the buffer size. In Frater (1990), a similar relationship was observed to hold for M/D/1 queues.

Table 1: Comparison of results.

	$\widehat{\alpha}$	
N	Direct	Fast
10	0.11	0.104
20	0.079	0.081
30	0.075	0.075
40	0.066	0.061

Table 1 shows a comparison of results obtained for direct and fast simulation for a number of different buffer sizes in a queue with virtual service rate 1, and arrival process defined by $\Lambda = 0.7$, $a = 0.7$, $b = 0.8$. In each case, the standard deviation of the estimate $\widehat{\alpha}$ is $\sigma = 0.05$. In the case of the direct simulation, this corresponds to 95 % confidence that the error is less than 10 %. For the fast simulation, because the likelihood ratio depends on the trajectory followed during each cycle, it is not possible to easily establish such a confidence interval. However, the largest difference between the results obtained by the different techniques is approximately 7 %, which suggests that the two methods are consistent.

7 CONCLUSION

Further work needs to be done to extend these results to higher order autoregressive processes. However, the work described here demonstrates the feasibility of using this approach in estimating the statistics of buffer overflows in queueing systems with autoregressive arrival processes.

Clearly, before practical application of this technique can be made to broadband networks, several refinements will be required. These include:

- allowing for the fact that a large number of sources will feed a single switch in the network, and that losses may occur as a result of simultaneous bursts in several switches;
- noting that there is an upper limit on the rate of a source, caused by the data rate of link fed by the source.

Extension of the analysis to a network of switches would also be advantageous.

REFERENCES

Cottrell, M., J. C. Fort, and G. Malgouyres (1983). Large deviations and rare events in the study of stochastic algorithms. *IEEE Trans. Automatic Control*, AC-28(9):907–918.

Frater, M. R. (1990). *Fast Estimation of the Statistics of Rare Events in Data Communications Systems.* PhD thesis, Australian National University.

Frater, M. R. and B. D. O. Anderson (1989). Fast estimation of the statistics of excessive backlogs in tandem networks of queues. *Australian Telecommunication Research*, 23(1):49–55.

Frater, M. R., J. Walrand and B. D. O. Anderson (1990). Optimally Efficient Simulation of Buffer Overflows in Queues with Deterministic Service Times. *Australian Telecommunication Research*, 24(1):1–8.

Frater, M. R., T. M. Lennon and B. D. O. Anderson (1991). Optimally Efficient Estimation of the Statistics of Rare Events in Queueing Networks. *IEEE Trans. Automatic Control*, 36(12):1395–406.

Heath, D. and P. Sanchez (1986). On the adequacy of pseudo-random number generators (or: How big a period do we need?). Operations Research Letters, 5(1):3–6.

Maglaris, B., D. Anastassiou, P. Sen, G. Karlsson and J. D. Roberts. 1988. Performance models of statistical multiplexing in packet video communications. *IEEE Trans. Communications*, 37(7):834–843.

Parekh, S. and J. Walrand (1989). A quick simulation of excessive backlogs in networks of queues. *IEEE Trans. Automatic Control*, 34(1):54–66.

AUTHOR 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, and the Ph. D. degree from the Australian National University in 1991. Since 1991, he has held the position of lecturer in the Department of Electrical Engineering, University College, Australian Defence Force Academy. His research interests include the efficient simulation of rare events in telecommunications systems, and the application of modern signal processing techniques to practical problems, such as oceanographic measurements.