

Quick Simulation of ATM Buffers with On-off Multiclass Markov Fluid Sources ¹

G. Kesidis

E & CE Dept, University of Waterloo, Waterloo, Ontario, N2L 3G1, Canada.

J. Walrand

EECS Dept, University of California, Berkeley, CA94720.

ACM TOMACS, Vol. 3, No. 3, pp. 269-276, July, 1993.

Abstract

The problem we address is how to quickly estimate by simulation the loss in a buffer with multiclass on-off Markov fluid sources. We generate the Markov fluids with the altered rate matrices given in [11], instead of the originals, to speed up the simulation. Likelihood ratios are used to recover an estimate of the loss for the original traffic parameters.

1 Introduction

The widely agreed upon standard for future broadband, integrated services digital networks is Asynchronous Transfer Mode (ATM). ATM carries statistically multiplexed streams of 53-byte cells called virtual circuits. ATM's use of statistical multiplexing results in a flexible and efficient utilization of the bandwidth by the bursty traffic. Buffering is used in the network's switches in order to be able to keep a virtual circuit's cell loss probability extremely small.

In this note, we address the problem of evaluating a switch design under simulated traffic conditions. We assume that the traffic sources are modeled as Markov fluids, the model parameters are known, and that the performance of a switch is determined by the probability of overflow in a busy cycle (Φ) of the switch buffers. A busy cycle is an interval $[S_k, S_{k+1})$, $k \in \mathbf{Z}$, where the $\{S_k\}$ are successive points in time such that the buffer occupancy $X(S_k-) = 0$ and $X(S_k) > 0$.

Φ is extremely small (for a nominal switch) so that a way to quickly estimate it is needed. Using a form of important sampling deduced from large deviations, the traffic parameters are altered (or "tilted") so that Φ is much larger and can be more accurately estimated in a reasonable amount of simulation time ("quick simulation" [16],[4], [2],[13], [15]). Likelihood ratios are used to recover an estimate for the original traffic parameters.

This note is organized as follows. In section 2, we describe the conjectured optimal choice for the tilted parameters reported in [11] and our quick simulation approach. A multiclass simulation example is reported in section 3. Conclusions are drawn in section 4.

2 Quick Simulation Approach

To speed up the simulation to estimate Φ , we alter or "tilt" the parameters of the arrival process so that the *deviant behavior* we are interested in (buffer overflow) becomes *typical*

¹Supported by: NSERC of Canada, Pacific Bell, Micro Grant of the State of California

(see [4], [2],[13]). We assume that the sources of the buffer are independent Markov fluids [1]. That is, the time-derivative of each source is a function of a continuous-time Markov chain on a finite state space. If the arrival process to a buffer with deterministic service rate is a superposition of independent Markov fluids, then the buffer occupancy has piecewise-linear trajectories with random slope. Let c cell/s be the buffer's deterministic service rate and B cells be the buffer's size.

A two-state or "on-off" Markov fluid source of type i , $i = 1, \dots, K$, is defined to have a Markovian time-derivative with state-space $\Lambda^i = (\Lambda_1^i, \Lambda_2^i)$ and transition rate matrix Q^i . We assume $\Lambda_1^i < \Lambda_2^i < \infty$ for all i . Let π^i be the invariant of Q^i ($\pi^i Q^i = 0$) and N_i be the number of sources of type i .

In [12] and [5], we heuristically argued that Φ has the following form

$$\Phi = \exp(-BI(N, c) + o(B)) \quad (1)$$

where $N = (N_1, \dots, N_K)$,

$$I(N, c) = \inf_{\sum N_i m_i > c} \frac{\sum_{i=1}^K N_i H_i(m_i)}{\sum_{i=1}^K N_i m_i - c},$$

and $\exp(-TH_i(m_i))$ is approximately the probability that a type i source generates $m_i T$ arrivals over large time T . Thus, H_i is determined by the large deviations for the empirical distribution of a type i source:

$$H_i(m_i) := \inf_{[\mu^i, \Lambda^i] = m_i} \inf_{P: \mu^i P = 0} G(P; Q^i) \quad (2)$$

where $[\mu^i, \Lambda^i] = \sum_j \mu_j^i \Lambda_j^i$, and the infima are taken over the space Σ_{m_i} of distributions μ^i on Λ^i and over the space of transition rate matrices P on Λ^i . G is the relative entropy rate between continuous-time Markov chains [10]:

$$G(P; Q) := \sum_{k=1}^2 \mu_k \sum_{j=1, j \neq k}^2 \left(P_{k,j} \log \frac{P_{k,j}}{Q_{k,j}} + Q_{k,j} - P_{k,j} \right)$$

where μ is the invariant of P ($\mu P = 0$).

By evaluating $I(N, c)$ we can approximate Φ by $\exp(-BI(N, c))$; but this approximation may not be accurate when the " $o(B)$ " terms are significant (see the simulations reported in [5]). Therefore, we base our change of parameters on an analysis of $I(N, c)$. A (quick) simulation is conducted to take the possible significance of the " $o(B)$ " terms into account.

Define the vector $\hat{M} := \{\hat{m}_1, \dots, \hat{m}_K\}$:

$$\hat{M} = \arg \inf_{\sum N_i m_i > c} \frac{\sum_{i=1}^K N_i H_i(m_i)}{\sum_{i=1}^K N_i m_i - c},$$

i.e., $\sum N_i \hat{m}_i$ is the most likely slope that the buffer occupancy will take to overflow from empty, for sufficiently large buffer size B . To minimize the variance of an estimate of Φ , we postulate that the optimum choice of tilted rate matrix for a type i source is

$$\hat{Q}^i = \arg \inf_{[\mu, \Lambda^i] = \hat{m}_i} G(P; Q^i) \quad (3)$$

where the infimum is taken over rate matrices P with invariants μ having mean \hat{m}_i .

We can obtain a closed form expression for H_i when the Markov fluid is of the on-off type:

$$H_i(m_i) = \frac{1}{\Lambda_2^i - \Lambda_1^i} \left(\sqrt{q_1^i(\Lambda_2^i - m_i)} - \sqrt{q_2^i(m_i - \Lambda_1^i)} \right)^2$$

where $q_1^i = Q_{1,2}^i$ and $q_2^i = Q_{2,1}^i$.

A more manageable expression for \hat{M} is obtained by using the concept of effective bandwidths ([8], [7], [12]). If $\alpha_i(\cdot)$ is defined to be the effective bandwidth of a call of type i , then α_i depends only on the type i parameters (i.e., Q^i and Λ^i) and [12]:

$$I(N, c) \geq \delta > 0 \Leftrightarrow \sum_{i=1}^K N_i \alpha_i(\delta) \leq c. \quad (4)$$

Thus, for large B , the constraint $\Phi \leq e^{-B\delta}$ is equivalent to a linear constraint on the effective bandwidths of the buffer sources.

We can show that [9],

$$\hat{m}_i = \arg \inf_{m_i > \alpha_i(I(N, c))} \frac{H_i(m_i)}{m_i - \alpha_i(I(N, c))} \quad (5)$$

and

$$\sum_{i=1}^K N_i \alpha_i(I(N, c)) = c.$$

By equation (3), the entries of \hat{Q}^i are

$$\hat{Q}_{1,2}^i = \sqrt{\frac{q_1^i q_2^i}{\kappa_i}} \quad \text{and} \quad \hat{Q}_{2,1}^i = \kappa_i \hat{Q}_{1,2}^i. \quad (6)$$

where $\kappa_i = (\Lambda_2^i - \hat{m}_i)/(\hat{m}_i - \Lambda_1^i)$.

In [7], a closed form expression is obtained for the effective bandwidths of (multiclass) Markov fluids sharing buffer with deterministic service rate (here we use the notation of [12]):

$$\alpha_i(\delta) = \frac{1}{2} \left(-a_i(\delta) + \sqrt{a_i^2(\delta) - 4b_i(\delta)} \right)$$

where

$$a_i(\delta) = \frac{q_1^i + q_2^i}{\delta} - \Lambda_2^i - \Lambda_1^i \quad \text{and} \quad b_i(\delta) = \Lambda_2^i \Lambda_1^i - \frac{q_1^i \Lambda_2^i + q_2^i \Lambda_1^i}{\delta}.$$

By equation (5) and direct calculation,

$$\hat{m}_i = \frac{q_1^i \Lambda_1^i (\Lambda_2^i - \alpha_i(I(N, c)))^2 + q_2^i \Lambda_2^i (\alpha_i(I(N, c)) - \Lambda_1^i)^2}{q_1^i (\Lambda_2^i - \alpha_i(I(N, c)))^2 + q_2^i (\alpha_i(I(N, c)) - \Lambda_1^i)^2}.$$

Thus, given $I(N, c)$, we can directly compute \hat{m}_i and, using equation (6), \hat{Q}^i as well, for all i . What remains is to numerically evaluate $I(N, c)$ using the fact that $I(N, c)$ is the only positive solution δ of the equation:

$$\sum_{i=1}^K N_i \alpha_i(\delta) - c = 0. \quad (7)$$

In [12] it was shown that the left hand side of equation (7) is, in general, nondecreasing in δ . We can use Newton's method to solve equation (7).

3 A Simulation Example

We simulated a buffer of size $B = 7$ cells with deterministic service rate $c = 1.9$ cells/s. The buffer had two independent on-off Markov fluid sources. Source 1 had the following traffic parameters:

$$\Lambda_1^1 = 0, \quad \Lambda_2^1 = 1, \quad q_1^{o1} = 0.7, \quad \text{and} \quad q_2^{o1} = 1.$$

Source 2 had the following traffic parameters:

$$\Lambda_1^2 = 0, \quad \Lambda_2^2 = 2, \quad q_1^{o2} = 0.5, \quad \text{and} \quad q_2^{o2} = 1.$$

We conducted a direct and a quick simulation as above. The amount of simulation time is roughly proportional to the number of transitions of the two-dimensional embedded Markov chain. Thus, for comparison, we took the “cost” of each simulation to be the number of transitions of the embedded Markov chain required to achieve a relative error of 10% of the estimate of Φ . The quick simulation required slightly more execution time per iteration to compute the likelihood ratio.

For the above example, three iterations of the Newton’s method algorithm were required to compute $I(N, c) \approx 1.110$ to an accuracy of ± 0.001 . Thus, this first step of the quick simulation program required a negligible amount of execution time.

Successive busy cycles are, in general, not independent and this complicates the business of obtaining an accurate estimate of a confidence interval for Φ . So, we assumed that successive groups of 10 consecutive busy cycles were independent. That is, for the direct simulation (with original traffic parameters Q^i) let

$$y_i = \mathbf{1}\{\text{an overflow occurs in the } i^{\text{th}} \text{ busy cycle}\},$$

and for the quick simulation (with tilted traffic parameters \hat{Q}^i) let

$$\tilde{y}_i = L_i \mathbf{1}\{\text{an overflow occurs in the } i^{\text{th}} \text{ busy cycle}\}$$

where L_i is the likelihood ratio for the i^{th} busy cycle. Also, for $n \geq 1$, let

$$Y_n = \frac{1}{10} \sum_{i=10n-9}^{10n} y_i \quad \text{and} \quad \tilde{Y}_n = \frac{1}{10} \sum_{i=10n-9}^{10n} \tilde{y}_i.$$

So, we assumed that $\{Y_n\}$ and $\{\tilde{Y}_n\}$ were i.i.d. sequences.

During our simulations, we updated the estimates of the mean and variance of the sequences Y_n and \tilde{Y}_n every 10 busy cycles (i.e., after every new sample generated) using a recursive formula found in [14], p. 98. If, after $10n$ busy cycles, the mean was μ_n and the variance σ_n^2 , then the estimate of Φ at this point would be μ_n and the relative error of Φ would be $\sqrt{\sigma_n^2/n}/\mu_n$.

The direct and quick simulations were executed three times each and the results of these simulations are given in Tables 1 and 2.

Table 1: Direct Simulation Results

Trial	Cost	Estimate of Φ
1	2.30×10^6	2.44×10^{-4}
2	2.28×10^6	2.42×10^{-4}
3	2.60×10^6	2.12×10^{-4}

Table 1: Quick Simulation Results

Trial	Cost	Estimate of Φ
1	2617	2.16×10^{-4}
2	2894	2.28×10^{-4}
3	2480	2.20×10^{-4}

We see that a speed-up of about of about 1000 times is achieved by using the quick simulation approach. Note that $\exp(-BI(N, c)) = 4.2 \times 10^{-4}$.

Distribution at the Beginning of Busy Cycles

Care should be taken to properly initialize the likelihood ratio to the appropriate value at the start of a busy cycle. Let μ and $\tilde{\mu}$ be the joint distributions of the source Markov processes at the beginning of a busy cycle for the original and tilted parameters respectively. If the Markov processes are in state \mathbf{x} at the start of a busy cycle, then we should initialize the likelihood ratio for that cycle to $\mu(\mathbf{x})/\tilde{\mu}(\mathbf{x})$. An estimate of μ (respectively, $\tilde{\mu}$) was obtained by simulating the buffer with the original (respectively, tilted) traffic parameters. Clearly, the cost of these simulations degrade the performance of the quick simulation.

One could save the cost of estimating $\tilde{\mu}$ by initializing the likelihood ratios to 1.0 and estimating $\tilde{\mu}(\mathbf{x})$ and $\phi_{\mathbf{x}}$ in parallel where $\phi_{\mathbf{x}}$ is defined to be the sum of cycle likelihood ratios for busy cycles beginning in state \mathbf{x} that resulted in buffer overflows. Our estimate of Φ at the end of this quick simulation would therefore be

$$\frac{1}{\beta} \sum_{\mathbf{x}} \frac{\mu(\mathbf{x})}{\tilde{\mu}(\mathbf{x})} \phi_{\mathbf{x}}$$

where β is the total number of busy cycles simulated. The drawback is that we cannot obtain an accurate estimate of the sample standard deviation of our estimate of Φ if we do things this way.

For the simulation example described above, there were two possible states for the Markov processes at the beginning of a busy cycle: $\mathbf{x} \in \{(0, 1), (1, 1)\}$ where, for example, “(0, 1)” is the state in which Source 1 is in state 0 and Source 2 is in state 1. After 1000 iterations (i.e., transitions of the embedded Markov chain) of both the direct and tilted simulations, we obtained the following estimates: $\mu(0, 1) \approx 0.6$, $\mu(1, 1) \approx 0.4$, $\tilde{\mu}(0, 1) \approx 0.3$, and $\tilde{\mu}(1, 1) \approx 0.7$. The cost of these simulations (2000 iterations) was included in the cost of the quick simulation quoted above.

4 Discussion and Conclusions

The same approach described above can be used to for a buffer with discrete-time Markov sources. That is, the number of cells that arrive at time n from a source of type i is a (discrete-time) Markov chain with *transition probability* matrix Q^i and state space Λ^i . The only alteration to the analysis of section 2 is that the relative entropy between

Markov transition probability matrices P and Q is [6]:

$$G_d(P; Q) := \sum_k \mu_k \sum_j P_{k,j} \log \frac{P_{k,j}}{Q_{k,j}}$$

where μ is the invariant of P ($\mu P = \mu$). The optimality of this choice of tilted parameters for the discrete-time buffer was proved recently in [3]. A quick simulation approach which assumes equation (1) holds but does not rely on knowledge of the traffic parameters is described in [5].

In summary, for a buffer with deterministic service rate and multiclass independent Markov fluid sources, we have derived an expression for tilted rate matrices in order to estimate the probability of buffer overflow in a busy cycle quickly. Simulations were reported to demonstrate the speed-up obtained by using this quick simulation method for on-off Markov fluids.

Acknowledgements

We'd like to thank P. Glasserman and P. Heidelberger for their many helpful suggestions.

References

- [1] D. Anick, D. Mitra, and M. M. Sondhi. Stochastic theory of a data-handling system with multiple sources. *Bell Sys. Tech. J.*, 61 No.8:1871–1894, 1982.
- [2] S. Asmussen. Conjugate processes and the simulation of ruin problems. *Stoch. Proc. and their Appl.*, 20:213–229, 1985.
- [3] C.-S. Chang, P. Heidelberger, S. Juneja, and P. Shahabuddin. Effective bandwidths and fast simulation of ATM networks. *IBM Research Report RC 18586, to appear in Performance'93, Rome, Italy, ...*.
- [4] M. Cottrell, J-C. Fort, and G. Malgouyres. Large deviations and rare events in the study of stochastic algorithms. *IEEE Trans. on Auto. Control*, 28 No. 9:907–920, 1983.
- [5] C. Courcoubetis, G. Kesidis, A. Ridder, J. Walrand, and R. Weber. Admission control and routing in ATM networks using inferences from measured buffer occupancy. *to appear in IEEE Trans. Comm., ...*, 1991.
- [6] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. John Wiley and Sons, Inc., New York, NY, 1991.
- [7] R.J. Gibbens and P.J. Hunt. Effective bandwidths for multi-type UAS channel. *Queueing Systems*, 9:17–28, 1991.
- [8] F.P. Kelly. Effective bandwidths of multi-class queues. *Queueing Systems*, 9:5–15, 1991.

- [9] G. Kesidis. Cell loss estimation in high-speed digital networks. *Ph.D. Dissertation, EECS Dept, U.C. Berkeley, .*., 1992.
- [10] G. Kesidis and J. Walrand. Relative entropy between Markov transition rate matrices. *to appear in IEEE Trans. Info. Th., .* .
- [11] G. Kesidis and J. Walrand. Quick simulation of ATM buffers. *Proc. IEEE CDC, Tucson, AZ, Vol. 1:1018,1019.*, 1992.
- [12] G. Kesidis, J. Walrand, and C.-S. Chang. Effective bandwidths for multiclass Markov fluids and other ATM sources. *to appear in IEEE Trans. Comm., .* ., May 1992.
- [13] S. Parekh and J. Walrand. A quick simulation method for excessive backlogs in networks of queues. *IEEE Trans. on Auto. Control*, 34 No. 1:54–66, 1989.
- [14] S. M. Ross. *A Course in Simulation*. Macmillan, New York, NY, 1990.
- [15] J. S. Sadowsky. Large deviations and efficient simulation of excessive backlogs in a GI/G/m queue. *IEEE Trans. Auto. Contr.*, 36:1383–1394, 1991.
- [16] D. Siegmund. Importance sampling in the Monte Carlo study of sequential tests. *The Annals of Statistics*, 4:673–684, 1976.