Date: May 25, 1997 COST257(97)
Issue: 1

Two Approaches to Simulating Loss Networks

Pasi Lassila[1] Jorma Virtamo[2] Samuli Aalto[3]

Laboratory of Telecommunications Technology
Helsinki University of Technology

Abstract: For multiservice loss networks the calculation of the normalization constant is computationally impossible for realistic sized problems. This paper deals with two methods for simulating the loss probabilities in multiservice loss networks. The regenerative method is based on the regenerative nature of the process and the simulation is carried out by simulating the DTMC of the process. The second approach tries to combine the effectiveness of Monte Carlo simulation with the computational easiness of simulating the DTMC. Some considerations, on using importance sampling (IS) for increasing the efficiency of the simulations, are also given.

[1] Pasi.Lassila@hut.fi
[2] Jorma.Virtamo@hut.fi
[3] Samuli.Aalto@hut.fi

List of Common Variables

$s = 1, \ldots, S$	index of the traffic streams in the multiservice network
$m = 1, \ldots, M$	sample index during the simulations, i.e. denotes the m^{th} sample
$l = 1, \ldots, L$	index of arrivals for the path simulation; path length = L arrivals
$\{\mathbf{X}_i\}_{i\geq 0}$	the embedded DTMC of the simulated stochastic process
X_i	distribution for number of blockings in a cycle in chap. 4.2
$\mathbf{P}$	transition matrix of the DTMC
$\mathbf{P}^*$	the induced transition matrix in importance sampling
$\mathbf{n}$	state of the multiservice network
N	size of the single link Erlang model
n	state of the single link Erlang model
a	offered traffic to the single link Erlang model
a^*	offered traffic to the single link Erlang model when using importance sampling
K	r.v. for the number of arrivals during a regenerative cycle
A_i	arrival event (including blocked arrivals) at time i during a regeneration cycle
C_i	congestion (blocking) event at time i during a regenerative cycle
C_l	congestion event at the time of l^{th} arrival in combined simulation
B	blocking probability for the single link Erlang system
B_s	stream s blocking probability in the multiservice case
τ	r.v. for the length of a regenerative cycle
ω	a realization from a stochastic process
Λ	likelihood ratio in importance sampling

1. Introduction

The broadband networks have been designed to offer various types of services integrated into the same network. On the call level the system appears as a loss system, where each service type is characterized by a fixed quantity representing the amount of bandwidth the service requires. If the call can not be setup, the call is lost. The arrival process is assumed to be Poisson, but the service time distribution can be general by the so called insensitivity property [Ros95]. This is a natural extension from the classical Erlang model for teletraffic.

The steady state probabilities of the above system have a well known product form solution: Consider a network consisting of a number of independent traffic streams s = 1, …, S, having an offered load a_s. The state of the network is completely defined by a vector $\mathbf{n}$ whose elements n_s represent the number of stream s calls present in the network. Then the equilibrium state probabilities are given by:

(1) $$\pi(\mathbf{n}) = G(\Omega)^{-1} p(\mathbf{n})$$

where the values $p(\mathbf{n})$ are the unnormalized probabilities of each stream s

(2) $$p(\mathbf{n}) = \prod_{s=0}^{S} \frac{a_s^{n_s}}{n_s!} \mathrm{e}^{-a_s}$$

and $G(\Omega)$ is the normalizing constant

(3) $$G(\Omega) = \sum_{n \in \Omega} p(\mathbf{n})$$

Ω is the set of allowed states, i.e. all the states where the link occupancies do not exceed the capacity of the links of the network.

The blocking probability for stream s can then be expressed as

(4) $$B_s = \frac{G(\Omega_s)}{G(\Omega)}$$

where $G(\Omega_s)$ is the state sum over the states where an addition of one stream s call would result in blocking for that stream.

In the case of a single link, which is offered several types of calls, the blocking probabilities can be calculated recursively by for example the Kaufman algorithm [Kau81]. Recursive algorithms can also be found for special network topologies like the tree topology, but for a general topology network such methods do not exist. In fact it is known that the calculation of the normalization constant G(Ω) is an NP complete problem [Ros95] and thus the problem becomes computationally impossible to solve for realistic size problems.

In such a situation we are faced with two possible alternatives for tackling the problem: either to find analytical approximations or to simulate. Analytical approximations have been developed based on either the so called reduced load approximations, which reduce to problems requiring the solving of fixed point equations [Kel86] [Chu93], or the solution is based on a numerical solution to the z-transform of the occupancy distribution [Mit94] (single link case) [Dow96]. However,

in this paper the focus is on approaches dealing with effective simulation of such quantities.

In this paper, two approaches to the simulation problem are discussed:

1. The use of regenerative simulation: The idea is based on the theory of so called regenerative processes, which have the property of regenerating themselves probabilistically at certain time epochs during the processes lifetime, e.g. at the beginning of a new busy cycle. Thus we are able to break the (theoretically) infinite length steady-state simulation into distinct independent identically distributed finite length samples. This also allows us to draw statistical inferences about the derived simulation results.
2. Combination of traditional Monte-Carlo simulation and a fixed length path simulation, which tries to combine the efficiency of the traditional Monte Carlo-method and the computational easiness of simulating the Markov chain describing the underlying stochastic process. Here too, the result is i.i.d. samples of the stochastic process and we have the possibility to make statistical inferences on the results.

Also, the use of importance sampling for increasing the efficiency of the simulation will be discussed to some extent. Especially when performing path simulation with a deterministic path length, the use of importance sampling produces interesting phenomena.

Throughout this paper the classical single link Erlang model will be used as an example. Extensions to the multiservice network case are indicated when possible.

2. Regenerative Simulation

The regenerative method has been developed in the late 70's and early 80's for the analysis of so called regenerative stochastic processes, see for example [Cra77], [Igl80]. Heuristically a regenerative process is a stochastic process, which starts probabilistically afresh at certain points in time, in other words regenerates itself. In our case, the system is the multiservice loss network, for which every state represents a regeneration point, because of the Markov property of the process.

2.1. Method Formulation

Let us define the regenerative method now more formally using the notation of [Goy92] and [Dev93]. Let $\{\mathbf{X}_i\}_{i\geq 0}$ denote the embedded discrete time Markov chain of the stochastic process with finite state space and transition matrix **P**. Assume that $\{\mathbf{X}_i\}_{i\geq 0}$ has a steady state distribution and converges in distribution to **X**. The goal is to estimate the expectation $\mathrm{E}[f(\mathbf{X})]$ of some function $f(\mathbf{X}) = h(\mathbf{X})/g(\mathbf{X})$. Now let **r** be a regeneration state and ω denote a sample path in the evolution of the system under study. Let $H = \sum_{i=0}^{\tau-1} h(\mathbf{X}_i)$ and $G = \sum_{i=0}^{\tau-1} g(\mathbf{X}_i)$, where $\mathbf{X}_0 = \mathbf{r}$ and τ is the random variable representing the number of steps (transitions) in a regeneration cycle, i.e. a path starting from state **r** and ending in state **r**. Then the expectation of $f(\mathbf{X})$ can be written as

(5) $$\mathrm{E}[f] = \frac{\mathrm{E}[H]}{\mathrm{E}[G]}$$

From which we get the estimator

(6) $$\hat{\mathrm{E}}[f] = \frac{1/M \sum_{m=1}^{M} H_m}{1/M \sum_{m=1}^{M} G_m} = \frac{\sum_{m=1}^{M} H_m}{\sum_{m=1}^{M} G_m} = \frac{\hat{H}_M}{\hat{G}_M}$$

where H_m and G_m are i.i.d. observations of H and G during the m^{th} simulation run. $\hat{H}_M$ and $\hat{G}_M$ denote the M-cycle estimators of H and G.

Example 1:

In our case, if we wish to simulate for example the blocking probability of the classical single link Erlang system, the H and G functions would be defined as follows: Let A_i denote the event of an arrival into the system (including blocked arrivals) and C_i the event of a blocking (congestion). Then we have that

(7) $$H = \sum_{i=0}^{\tau-1} 1_{A_i \cap C_i}$$

(8) $$G = \sum_{i=0}^{\tau-1} 1_{A_i}$$

The regenerative method for simulation would then be as follows: One simulates (using the transition matrix **P**) M independent samples of the random vector (G, H) thus obtaining the i.i.d. random vectors $\{(G_m, H_m) : m = 1, \ldots, M\}$. Then the blocking probability B is estimated using (6):

(9) $$\hat{B}_M = \frac{\sum_{m=1}^{M} H_m}{\sum_{m=1}^{M} G_m}$$

One of the most important benefits from regenerative simulation is that now the M simulated cycles are i.i.d. Thus we can make statistical inferences on the obtained results. Additionally we need the fact that the estimator (9) is strongly consistent, i.e. $\lim_{M\to\infty} \hat{B}_M = B$ with probability one. Then we have by the central limit theorem, for example, the following result

(10) $$\frac{\sqrt{M}\left(\hat{B}_M - B\right)}{\sigma / \mathrm{E}[G]} \to N(0,1)$$

where $\sigma = \sqrt{\mathrm{Var}[H - BG]}$. See e.g. [Cra77] for the derivation of this result.

The extension of this method for simulation of a multiservice loss network is straight forward: One needs to have separate indicator functions calculating the blockings and arrivals separately for each traffic stream in the system. However, one problem that will probably appear in this context is that the regeneration cycles may become extremely long.

2.2. Consistency and Bias of the Ratio Estimator

By the strong law of large numbers, the M-cycle estimators $\hat{H}_M$ and $\hat{G}_M$ are statistically unbiased estimates of E[H] and E[G] as $M \to \infty$, i.e. strongly consistent. Then from the theory of regenerative processes the ratio estimator (6) is also strongly

consistent. [Igl80] However, for finite M their ratio yields a biased estimate of E[f], because $E\left[\hat{H}_M / \hat{G}_M\right] \neq \mathrm{E}\left[\hat{H}_M\right] / \mathrm{E}\left[\hat{G}_M\right]$

In general, this bias is a result of the strong correlation of the results to the choice of the regeneration state **r**. For our example, the single link Erlang model, this can be shown easily: In this system every state n is a regeneration point. Let the size of the system be N and let us choose the state N as the regeneration state. Event A_1 denotes the occurrence of an arrival when the system is started from state N, i.e. the first event in the path is an arrival. This event is also a blocking event and thus in the case of a blocking the regeneration cycle has length 1. Then the H and G functions become

$$\begin{cases} H = 1_{A_1} \\ G = 1_{A_1} + 1_{\bar{A}_1} K \end{cases}$$

where K is an integer valued random variable, whose distribution is the number of arrivals during a cycle, when the cycle begins with a customer departure event. Then the estimator $\hat{B} = \mathrm{H}/\mathrm{G}$ has expected value

$$E\left[\hat{B}\right] = E\left[E\left[\frac{1_{A_1}}{1_{A_1} + 1_{\bar{A}_1} K} \mid K\right]\right] = E\left[P\{A_1\}\frac{1}{1+0\cdot K} + \left(1 - P\{A_1\}\right)\frac{0}{0+1\cdot K}\right] = P\{A_1\}$$

Now in our example $P\{A_1\}$ is simply the success probability of a Bernoulli variable of either moving upwards (blocking) from state N or downwards. Thus $P\{A_1\} = a/(a+N)$, which is (obviously) not the same as the blocking probability B for all $N > 1$.

The next interesting question is then what is the convergence rate of the ratio estimator, since it is strongly consistent.

2.3. Convergence Rate of The Ratio Estimator

It has been shown in [Cra75] that, if the stochastic process has several states, which serve as regeneration states, then asymptotically as $t \to \infty$ the obtained confidence intervals for simulations starting from different states are equally large. This implies that the asymptotic efficiency of the estimator does not depend on the chosen regeneration state.

But this is only an asymptotic result, which applies at the limit $t \to \infty$ and as such this result does not tell very much about the convergence rate of the estimators. In general, analysis of the convergence rate for estimators starting from arbitrary regeneration points is very difficult due to the handling of the correlation between the estimator's numerator and denominator.

However, for our example, this problem can be nicely avoided by a suitable choice of the regeneration state: Let the system size be N, and we shall choose the regenerative state to be state N just after the previous blocking event. The regeneration cycle will be defined as a path starting from state N until the next blocking event occurs. Thus a regeneration cycle will always have exactly one blocked customer and the estimator for blocking will be of the form (given M simulation cycles)

$$\hat{B} = \frac{M}{K_1 + K_2 + \ldots + K_M} \tag{11}$$

where M is a deterministic number and K_i are i.i.d. integer valued r.v. describing the total number of arrived customers during a regeneration cycle.

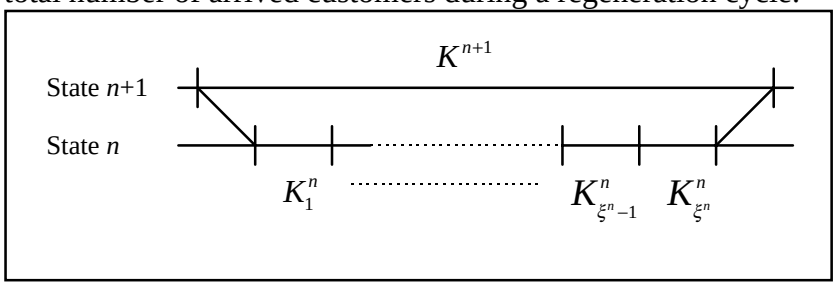

Figure 1: The meaning of the r.v. K^n in state n

Now let us examine more closely the distribution of K, i.e. a single cycle (see fig. 1). At each state n the probability of moving up in state n is $p_{\uparrow}^n = a/(a+n)$ and down $p_{\downarrow}^n = n/(a+n)$. At each state n the number of arrivals before the path goes to state n+1, denoted with r.v. K^{n+1}, consists of ξ^n independent cycles starting in state n and ending in state n. Furthermore, the number of these cycles ξ^n is geometrically distributed, with success probability $p_{\uparrow}^n$. Thus if we start the analysis by looking at the situation at state N just after a blocking event has occurred, we note that the cycle lengths at each state has the following recursive structure

$$K^{N+1} = 1 + \left(K_1^N + \ldots + K_{\xi^N}^N\right) \quad , \xi^N \sim \mathrm{Geo}\left(p_{\uparrow}^N\right)$$

$$K^N = 1 + \left(K_1^{N-1} + \ldots + K_{\xi^{N-1}}^{N-1}\right) \quad , \xi^{N-1} \sim \mathrm{Geo}\left(p_{\uparrow}^{N-1}\right)$$

$$\vdots$$

$$K^1 = 1$$

Thus the probability generating function for K^{N+1} becomes

$$\begin{aligned} K^{N+1}(z) &= \mathrm{E}\left[z^{K^{N+1}}\right] \\ &= z\,\mathrm{E}\left[\mathrm{E}\left[z^{A_1^N + \ldots + A_{\xi^N}^N} \middle| \xi^N\right]\right] \\ &= z\,\mathrm{E}\left[\mathrm{E}\left[z^{K^N}\right]^{\xi^N}\right] \\ &= z\,\mathrm{E}\left[\left(K^N(z)\right)^{\xi^N}\right] \\ &= z\,\xi^N\left(K^N(z)\right) \end{aligned}$$

Then because ξ^n is geometrically distributed

$$(12) \qquad K^{N+1}(z) = z\frac{p_{\uparrow}^N}{1 - p_{\downarrow}^N K^N(z)}$$

Thus we are able to find the p.g.f. for the distribution of the number of arrivals during a cycle starting from the blocking state N of an Erlang system with arbitrary size. The expectation of the one cycle estimator for the blocking probability then becomes

$$(13) \qquad \mathrm{E}\left[\hat{B}\right] = \mathrm{E}\left[\frac{1}{K^{N+1}}\right]$$

This and the second moment can be quite easily calculated with the help of some additional results relating to the properties of the probability generating function (p.g.f.): Given the p.g.f. $K^{N+1}(z) = \sum_{j=1}^{\infty} p_j^{N+1} z^j$, then the following holds

$$(14) \qquad \mathrm{E}\left[\frac{1}{K^{N+1}}\right] = \sum_{j=1}^{\infty} p_j^{N+1} \frac{1}{j} = \int_0^1 \frac{K^{N+1}(z)}{z}\, dz$$

$$(15) \qquad \mathrm{E}\left[\left[\frac{1}{K^{N+1}}\right]^2\right] = \sum_{j=1}^{\infty} p_j^{N+1} \frac{1}{j^2} = \int_0^1 \frac{1}{z}\left(\int_0^z \frac{K^{N+1}(u)}{u}\, du\right) dz$$

Now, if we wish to examine the rate of convergence of the estimator, we must be able to evaluate (14) for increasing number of cycles. If we simulate *M* cycles, then the number of arrivals distribution becomes

$$K_\Sigma^{N+1} = K_1^{N+1} + \ldots + K_M^{N+1}$$

Since the cycles are i.i.d. the p.g.f. is simply

$$(16) \qquad K_\Sigma^{N+1}(z) = \left(K^{N+1}(z)\right)^M$$

The estimator then has the following form, which can be evaluated using (14)

$$(17) \qquad \mathrm{E}\left[\hat{B}\right] = \mathrm{E}\left[\frac{M}{K_\Sigma^{N+1}}\right] = M \int_0^1 \frac{K_\Sigma^{N+1}(z)}{z}\, dz$$

Example 2:

For example, if the size of the system $N = 2$, then the single cycle distribution for the number of arrivals, when starting from state 2 and ending in a blocking event, is

$$K^3(z) = z \frac{\frac{a}{a+2}}{1 - \frac{2}{a+2} z \frac{\frac{a}{a+1}}{1 - \frac{1}{a+1} z}}$$

Figure 2 shows a graph representing the estimator's (17) mean m_K and its standard deviation s_K as a function of increasing number of cycles *M*.

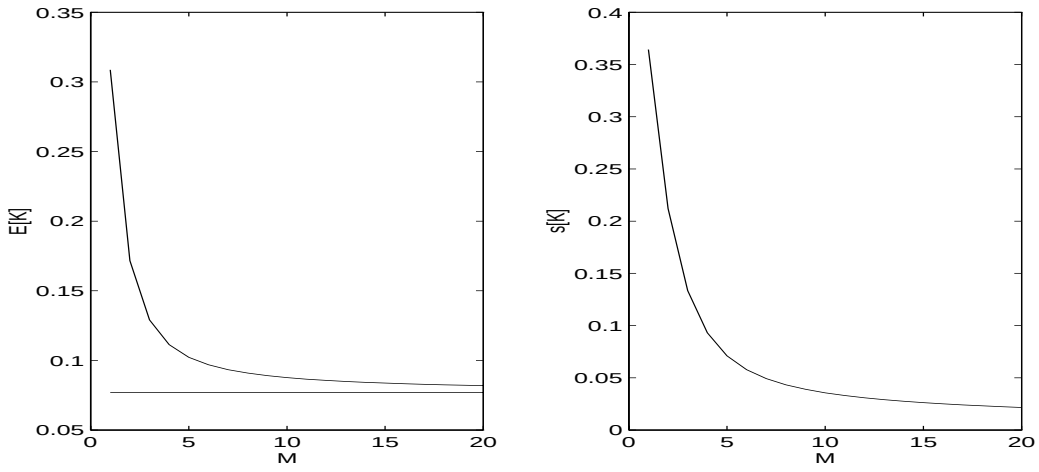

Figure 2: Mean and standard deviation of estimator (17) for $N = 2$ and $a = 0.5$

From the figure it can be seen that the estimator converges nicely when the amount of cycles increases. However, it still might take a long time to have a confident estimate, because now we have chosen the blocking state as the regeneration state and thus the

regeneration cycles themselves can be very long depending on how rare the blocking event is.

Currently there is no mathematical theory, which would help in finding the 'right' regeneration state. It seems reasonable though, to suspect that the convergence is quite rapid as the number of regeneration cycles increases. In other words, the confidence of the estimate is dominated by the number of regeneration cycles obtained during a fixed simulation time. Therefore the heuristic of choosing the state, which gives the shortest expected return time, given in [Gly89], seems quite plausible. Clearly more investigation in this area is needed.

3. Combined Monte Carlo Simulation and Path Generation

3.1. Motivation for the Method

In pure Monte Carlo simulation, the idea is the following: From (1) and (2) we see that the steady state probability $\pi(\mathbf{n})$ of the system is known except for the normalization constant $G(\Omega)$. Then if the system's capacity is infinite, the traffic streams become independent of each other and behave as independent Poisson streams. The steady state probability of the system is then given exactly by (2). Then the system can be simulated by generating S independent random numbers and checking independently for each stream what state the obtained random number corresponds. The finite capacity system is then handled in exactly the same way, but after the total state **n** is known, it must be checked that this is indeed an allowed state, i.e. the state **n** does not violate the link capacity constraints. Thus we are able to calculate the normalization constant $G(\Omega)$ as the sum of the unnormalized probabilities for all the generated allowed states during the simulation.

Path simulation refers to the simulation of the associated embedded DTMC for estimating the steady state probabilities, i.e. from a given initial state, the process is simulated until a given number of arrivals has occurred.

In the combined method the basic idea is to combine the efficiency of Monte Carlo method and the computational easiness of simulating the associated embedded DTMC. The reason for this is twofold:

1. Obtaining samples with the pure Monte Carlo method becomes computationally quite heavy as the number of dimensions (traffic streams), $s = 1 \ldots S$, and the depth of each dimension, i.e. max. number of stream s calls $n^1_{\max} \ldots n^S_{\max}$ becomes larger. To generate a Monte Carlo sample has the complexity $O(SN^{\max})$, where $N^{\max}$ is $\max\{ n^1_{\max} \ldots n^S_{\max} \}$. On the other hand the generation of a transition has only complexity $O(S)$ and typically $N^{\max} >> 1$. Thus it seems plausible, that something could be gained by combining these two methods.
2. For pure Monte Carlo method, the use of importance sampling techniques for speeding up the simulation, does not provide very good results due to the static nature of the importance sampling methods available [Ros95, chap. 6]. On the other hand, for path generation several more dynamic methods are available through the use of large deviation results for Markov chains [Hee96, Buc90,

Par86]. Thus simulation speedup techniques can be implemented in both parts of this method to obtain the best of both methods.

3.2. The Method Formulation

The combination of a random starting state and a fixed length path, in terms of arrivals, for simulating the call blocking probability, rests on a certain property of the steady state distribution for multiservice loss networks with Poissonian arrivals: The steady state distribution $\pi(\mathbf{n})$ of the multiservice loss network represents the state of the system just prior to an arriving customer.

This suggests the following method for simulation:

1. Generate a starting state **n** using the previously described method.
2. Generate a path starting from state (**n**+1) , with this initial transition counted as the 1st arrival. Then simulate a fixed number of arrivals L and indexing the arrivals with $l = 1, \ldots, L$.
3. From each path calculate the number of blocking events for each stream, denoted with $\mathbf{C}_l$.
4. Simulate M samples.

Example 3

For example, for the single link Erlang model, the sample estimator for the blocking probability B becomes

$$\hat{B} = \frac{\sum_{l=1}^{L} 1_{C_l}}{L} \tag{18}$$

The extension of this to multiservice case is done by using separate indicator functions for each traffic stream.

For M simulated samples the estimator becomes

$$\hat{B} = \frac{\sum_{m=1}^{M}\sum_{l=1}^{L} 1_{C_l}}{ML} \tag{19}$$

Estimator (18) provides an unbiased estimator for the blocking probability. Intuitively this can be understood from the way the simulation process works: the estimator itself produces an estimate for the conditional blocking probability when starting from state n and with a given path length L. But since this state is chosen randomly from the true steady state distribution itself, the result is an unbiased estimate of the blocking probability B. This can be shown formally easily for $L \leq 2$, but for $L > 2$ it remains to be done. Note that when $L = 1$, the method reduces to the plain Monte Carlo method.

Also, the samples drawn from the process are i.i.d., because all the samples are drawn by first generating the starting state from the steady-state distribution and then the process is simulated for a fixed number of arrivals. From the central limit theorem we know then that

$$\frac{\hat{B} - B}{\sigma_{\hat{B}} / \sqrt{K}} \rightarrow N(0,1)$$

where σ is the sample standard deviation. Thus we are again able to make the estimate as accurate as desired by increasing the number of samples.

4. Importance Sampling

Importance sampling (IS) is a variance reduction technique that can be used to estimate the steady state probability of rare events. Rare in this context means events for which conventional simulation methods would require unacceptably long simulation runs to obtain reasonably accurate estimates of the interesting measures. For example, a typical rare event would be the probability of cell loss in an ATM network.

Importance sampling is based on the following observation: The notation is here as in chapter 2.1 before. Let $\mathbf{P}^*$ be an alternative sampling transition matrix with $P^*(\omega)$ the induced probability associated with path ω. Also, let $\mathrm{E}^*[\ ]$ denote the expectation operator with respect to the probability measure $P^*(\omega)$. Then we have that

(20) $\mathrm{E}[H] = \mathrm{E}^*[H\Lambda]$

where

(21) $$\Lambda(\omega) = \frac{P(\omega)}{P^*(\omega)} = \prod_{j=0}^{\tau-1} \frac{p\left(\mathbf{X}_j, \mathbf{X}_{j+1}\right)}{p^*\left(\mathbf{X}_j, \mathbf{X}_{j+1}\right)}$$

i.e. $\Lambda(\omega)$ is the cumulative product of likelihood ratios of the transition probabilities along the path during a regeneration cycle. The (ω) is included in the notation to stress the fact that Λ is a function of the generated path ω. The only thing that is required of the induced probability $P^*(\omega)$, is that $P^*(\omega) \neq 0$ when $HP(\omega) \neq 0$.

The terminology used above was based on the assumption of performing regenerative simulation. However, the results hold for the path generation with fixed number of arrivals also. In that case the choice of the path length becomes a critical issue as will be discussed later.

4.1. A Heuristic for Obtaining the IS Distribution

When using IS, the first question to be solved is, what is the induced distribution, which should be used for sampling. It is well known that the optimal choice of $P^*(\omega)$ would be to concentrate all the probability mass on the important event we are interested in, thus forcing the important (rare) event to happen with probability one. The desired probability would then be obtained through the likelihood ratio. Since this is impractical, because it requires the knowledge of the probability of the rare event itself, other methods must be used.

In [Hei95] it is shown that reducing the variance of the estimator by importance sampling corresponds to reducing the second moment of the estimator. This turns out to have the practical interpretation that, to reduce the variance, we want to make the likelihood ratio small for the rare event. To do this requires that we try to increase the probability of the rare event.

In [Par86], [Buc90] and [Hee96] large deviation results obtained for slow random walks have been used to obtain the IS distribution. However, here we present a more simple heuristic for the single link Erlang system.

In the infinite capacity Erlang system the steady state probability of being in state n the is given by $p_n = \frac{a^n}{n!}\mathrm{e}^{-a}$, where a is the offered traffic. The expected value of the state is E[n] = a. Thus to increase the probability of blocking events, we can move the probability mass of the distribution closer to the blocking state N of the link by choosing a new offered traffic $a^* = za$

$$\Rightarrow p_k^* = \frac{za^k}{k!}\mathrm{e}^{-za}$$

Then $\mathrm{E}^*[n] = za$. To make blockings more frequent we simply require

$$E^*[n] = N \Rightarrow z = \frac{N}{a} \Rightarrow a^* = N$$

Using this new a^*, we get the new transition matrix $\mathbf{P}^*$ for the Erlang model. This method can also be used for importance sampling in a pure Monte Carlo method.

4.2. Simple Analysis of the Proposed Heuristic

Let us look at the simple 1-server system, with offered traffic a and let a^* be the offered traffic in the system under importance sampling. The system has the following transition probabilities and their corresponding likelihood ratios

Transition	Prob.	Likelihood ratio
$0 \to 1$	1	$\Lambda_i = 1$
$1 \to 0$	$\frac{1}{1+a}$	$\Lambda_i = \frac{1+a^*}{1+a}$
$1 \to 1$	$\frac{a}{1+a}$	$\Lambda_i = \frac{1+1/a^*}{1+1/a}$

If we choose state 1 as the regeneration state, regeneration cycles have length 1. Additionally, there are only two kinds of realizations: 1->0->1 or 1->1 (with a blocking event). The probability of the first realization is the same for transition 1->0 and the likelihood ratio is also the same. For the second realization the corresponding measures are the same as given above for the 1->1 transition.

Now if X_i is the distribution for the number of blockings during a cycle then X_i~Bernoulli(a/(a+1)) and because the cycle length is always 1, the expected value of the estimator for the blocking probability is

$$\mathrm{E}\left[\hat{B}\right] = \mathrm{E}[X] = \frac{a}{a+1}$$

with variance

$$\mathrm{Var}[X] = \frac{a}{(1+a)^2}$$

Thus in this special case it is also an unbiased estimate of the true blocking probability B. This is a result of the fact that the path length is always exactly 1 arrival, hence the denominator in the ratio estimator (5) becomes deterministic and is equal to 1.

If we wish to analyze the variance of the estimators for this system when using regenerative simulation or combined MC and path simulation, we can do this by doing two types of experiments:

1. regenerative method: the blocking probability is estimated as the mean of two individual regeneration cycles
2. combined method: the two cycles are taken as a single experiment consisting of two cycles

In the first case the estimator is $\hat{B}_r = \left(X_1 + X_2\right)/2$ and in the second case $\hat{B}_c = Y/2$ where Y is the joint distribution for number of blockings during two cycles. The expectation and variance of these yield identical results

$$\mathrm{E}\left[\hat{B}_r\right] = \mathrm{E}\left[\hat{B}_c\right] = \mathrm{E}[X]$$

$$\mathrm{Var}\left[\hat{B}_r\right] = \mathrm{Var}\left[\hat{B}_c\right] = \frac{\mathrm{Var}[X]}{2}$$

If we then use importance sampling and make the same experiments, we get quite interesting results. Let us denote with X^* and Y^* the previously defined r.v. X and Y under the importance sampling measure. Correspondingly the estimators $\hat{B}_r^* = \left(X_1^* + X_2^*\right)/2$ and $\hat{B}_c^* = Y^*/2$ are the estimators for the blocking probabilities under the importance sampling measure. Then the expectation of estimators becomes

$$\mathrm{E}\left[\hat{B}_r^*\right] = \mathrm{E}^*\left[\Lambda\hat{B}_r^*\right] = \mathrm{E}\left[\hat{B}_r\right] = \mathrm{E}[X]$$

$$\mathrm{E}\left[\hat{B}_c^*\right] = \mathrm{E}^*\left[\Lambda\hat{B}_c^*\right] = \mathrm{E}\left[\hat{B}_c\right] = \mathrm{E}[X]$$

However, the variances become quite different

$$\mathrm{Var}\left[\hat{B}_r^*\right] = \frac{1}{2}\left(\frac{a}{a+1}\right)^2\left(\frac{a^*+1}{a^*} - 1\right) = \mathrm{Var}\left[\hat{B}_r\right]\frac{a}{a^*}$$

$$\mathrm{Var}\left[\hat{B}_c^*\right] = \frac{1}{2}\frac{a}{(a+1)^2}\frac{a}{a^*}\left[\left(\frac{1+a^*}{1+a}\right)^2\left(2\frac{a^2}{a^*}+1\right) - 2a^*\right]$$

$$= \mathrm{Var}\left[\hat{B}_r\right]\frac{a}{a^*}\left[\left(\frac{1+a^*}{1+a}\right)^2\left(2\frac{a^2}{a^*}+1\right) - 2a^*\right]$$

$$= \mathrm{Var}\left[\hat{B}_r^*\right]\left[\left(\frac{1+a^*}{1+a}\right)^2\left(2\frac{a^2}{a^*}+1\right) - 2a^*\right]$$

The first result for the variance of the regenerative method is a manifestation of the known result for optimal importance sampling: when the estimated probability is known in advance, the variance of the importance sampling can be made zero, by forcing the important event (blocking) happen by probability 1, i.e. increasing a^* to infinity, and the desired estimate is then obtained from the likelihood ratio Λ. The second result is much more difficult to interpret: what can be seen at least is that when the simulated path length extends over several regeneration cycles, the relationship between the variance of the combined method and regenerative method becomes quite complex even for this simple example.

In the following figure the variances of both estimators $\hat{B}_r^*$ and $\hat{B}_c^*$ are drawn as a function of a^*. The variances are normalized to the variance of the estimator without importance sampling.

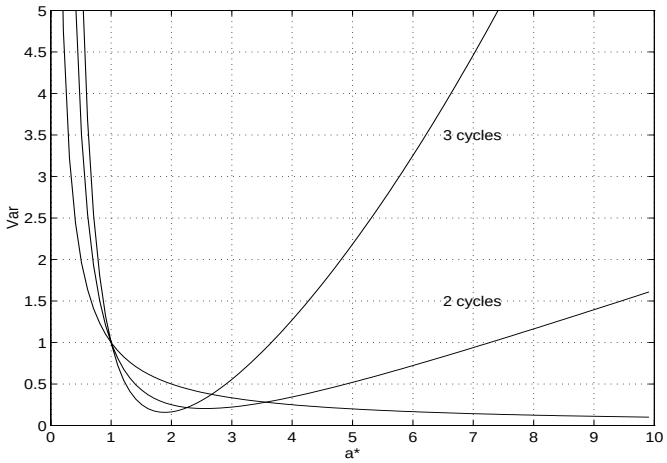

Figure 3: Variance of the estimators as a function a^*

As can be seen when the length of the simulation is increased, the more critical the choice of a^* becomes. But what is interesting also is that if a^* is fixed, there is a possibility to obtain an even smaller variance with combining cycles together than to interpret the cycles as individual samples !

This increase of variance, when the path length is increased, corroborates our experimental findings: when the path length was extended, the simulations produced consistently estimates, which were too small.

This problem is related to the property of the likelihood ratio in infinite length steady state estimation. In [Gly89] it is shown that as the simulation length approaches infinity, the likelihood ratio $\Lambda_\infty(\omega) \to 0$. Thus importance sampling fails in such situations. However, in [Gly89] it is also shown that the situation can be avoided by the use regenerative simulation, where the infinite horizon steady state behavior is reduced to a finite horizon analysis over regenerative cycles.

In our example here, when the simulation paths were made longer, we got more blocking events. The problem is that the number of new blocking events can only grow linearly as the path becomes longer, while the likelihood ratio becomes smaller at an exponential rate (the likelihood ratio < 1 for transitions producing blocking events). Thus the extra information gained from new blockings becomes increasingly smaller as the path becomes longer.

This simple analysis provided some explanations to the unexpected results obtained during experiments. More investigation is needed to examine the effects of the proposed heuristic when simulating larger systems.

5. Conclusions and Future Work

In this paper two approaches to simulating the blocking probability of multiservice networks were discussed. The regenerative method is a well known and theoretically quite thoroughly examined method. However, there is still room for some more analysis, e.g. relating to the choice of a suitable regenerative state.

The combined method is a new combination of previously used techniques and its analysis is only at its early stages at the moment. Preliminary tests on the single link Erlang model have proved that the method works but ,on the other hand, have not shown any exceptional performance over the traditional methods either. However, tests on the multiservice case have not been carried out yet.

In general, it seems that the simulation problem as approached in this paper, leaves room for quite many interesting puzzles to be solved. For example the following problems are subject to further study:

- analysis methods/heuristics for comparing the efficiency of the simulation methods
- convergence analysis from arbitrary starting state in regenerative simulation
- extend simulations to multiservice case
- investigate other variance reduction methods (RESTART)
- continue the analysis of the proposed importance sampling heuristic for larger size systems
- optimal choice for determining the path length in combined simulation
- research on other importance sampling distributions than the proposed heuristic

References

[Buc90] J. A. Bucklew; “Large Deviation Techniques in Decision, Simulation and Estimation”; Wiley; 1990; 259p.

[Chu93] S. Chung, K. W. Ross; “Reduced Load Approximations for Multirate Loss Networks”; IEEE Transactions on Communications; vol. 41, no. 8, Aug. 1993

[Cra75] M. A. Crane, D. L. Iglehart; “Simulating Stable Stochastic Systems: III. Regenerative Processes and Discrete Event Simulations”; Operations Research; vol. 23, no. 1, 1975

[Cra77] M. A. Crane, A. J. Lemoine; “An Introduction to the Regenerative Method for Simulation Analysis”; Springer-Verlag; 1977

[Dev93] M. Devitsikiotis, J. K. Townsend; “Statistical Optimization of Dynamic Importance Sampling Parameters for Efficient Simulation of Communication Networks”; IEEE/ACM Transactions on Networking; vol.1, no. 3, Jun. 1993

[Dow96] D. G. Down, J. T. Virtamo; “Blocking Probabilities in Multirate Circuit Switched Networks: Capturing Dependent Behaviour”; 13th Nordic Teletraffic Seminar; Aug 20-22, 1996; pp. 18-29

[Gly89] P. W. Glynn, D. L. Iglehart; “Importance Sampling for Stochastic Simulations”; Management Science; vol. 35, no. 11, Nov. 1989

[Goy92] A. Goyal, P. Shahabuddin, P. Heidelberger, V. F. Nicola, P. W. Glynn; “A Unified Framework for Simulating Markovian Models of Highly Dependable Systems”; IEEE Transactions on Computers; vol. 41, no. 1, Jan. 1992

[Hee96] P. E. Heegaard; “Adaptive Optimization of Importance Sampling for Multi-Dimensional State Space Models with Irregular Resource Boundaries”; Nordic Teletraffic Seminar-13; Aug. 1996

[Hei95] P. Heidelberger; “Fast Simulation of Rare Events in Queuing and Reliability Models”; ACM Transactions on Modeling and Computer Simulation; vol. 5, no. 1, Jan. 1995

[Igl80] D. L. Iglehart, G. S. Schedler; “Regenerative Simulation of Response Times in Networks of Queues”; Springer-Verlag; 1980

[Kau81] J. S. Kaufman; “Blocking in a Shared Resource Environment”; IEEE Transactions on Communications; vol. 29, no. 10, Oct. 1981

[Kel86] F. P. Kelly; “Blocking Probabilities in Large Circuit Switched Networks”; Advances in Applied Probability; vol. 18, 1986, pp. 473-505

[Mit94] D. Mitra, J. A. Morrison; “Erlang Capacity and Uniform Approximations for Shared Unbuffered Resources”; IEEE/ACM Transactions on Networking; vol. 2, no. 6, Dec. 1994

[Par86] S. Parekh, J. Walrand; “Quick Simulation of Excessive Backlogs in Networks of Queues”; IEEE Transactions on Automatic Control; vol. 34, no.1, Jan. 1995

[Ros95] K. W. Ross; “Multiservice Loss Models for Broadband Telecommunication Networks”, Springer-Verlag, 1995, 343 p.