Proceedings of the 1989 Winter Simulation Conference
E.A. MacNair, K.J. Musselman, P. Heidelberger (eds.)

QUICK SIMULATION OF RARE EVENTS IN NETWORKS

Roman D. Fresnedo
University of California, Berkeley
Berkeley, CA 94720, U.S.A

ABSTRACT

We study the problem of how to simulate the occurrence of rare events on networks of queues; an interesting application is to obtain the expected time until the network buffers fill up.

We show that the unique *optimal* (minimum variance) change of measure (*importance sampling*) to simulate an event is given by the law of the process conditioned on the event (rare or not).

Some theory is needed to circumvent the fact that knowledge of the conditional laws implies knowledge of the solution. We present two ways to handle the problem. Boundary theory of Markov Chains provides the theoretical framework. The method sheds light on the way that rare events happen; this in turn explains why some large deviations ("LD" in what follows) heuristics (Walrand and Parekh) fail for important combinations of parameter values (optimal buffer allocation).

A compactness argument and a scale-down version of the model are used to simulate successfully the chances of excessive backlogs for many M/M/1 queues in tandem and for any combination of parameters.

Alternatively, we can build on the LD heuristics and, using the notion of convex combination of harmonics, we successfully treat the optimal buffer allocation case.

1. SETUP AND THEORY

Let P_0 be the governing measure of a positive recurrent Markov Process defined on the non-negative quadrant Z_+^d of Z^d started at 0 ("the empty system" in queueing applications) and let E_0 be the corresponding expectation operator. Let B be a boundary for Z_+^d that delimits a bounded region R containing 0 (see Figure 1). For example, one can think of the state space of a Markov Process representing queue sizes of an open Jackson Network (Walrand (1987), Kelly (1979)). See the concluding remarks on the relevancy of the relatvely simple Markovian case.

1.1 The Problem

Let A be the event that the chain hits the boundary before coming back to 0, and let $\alpha = P_0(A)$. Suppose we want to find α, but it is hard to do it analytically, and the boundary is so far away that the event A is rare, and therefore direct simulation is impractical: it would require not only too much computer time, but could also exhaust the random number generator by using it beyond its period.

In most cases, the interest in finding α comes from the fact that, if T is the time until the rare event happens, a reasonable good estimator of $E_0(T)$ is $\frac{1}{\alpha}E_0(T_0)$ where T_0 is the time to return to 0.

For a stable system, $E_0(T_0)$ can be estimated using direct simulations. In non-Markovian contexts, the usual idea of regenerative simulation can be used (Ripley (1987) and references therein).

1.2 Criteria and Method

The criteria by which we select our estimators are unbiasedness and minimum variance; they correspond to the two main ingredients in the method: *Importance Sampling and Conditioning*.

The idea is to change the measure under which the chain evolves and use likelihood ratios to estimate α. (Siegmund (1976), Cottrel, Fort, and Malgouyres (1983), Parekh and Walrand (1989)).
Denote by E_0 the expected value under P_0 and by $\tilde{E}_0$ the expected value under $\tilde{P}_0$, where $\tilde{P}_0$ is such that P_0 is absolutely continuous with respect to $\tilde{P}_0$. Then, if we denote by 1_A the indicator of the event A:

$$\alpha = E_0(1_A) = \tilde{E}_0(1_A \tilde{L}), \quad \textit{where} \quad \tilde{L} = \frac{dP_0}{d\tilde{P}_0} \qquad (1)$$

A direct simulation would consist on running the process under P_0 a convenient number of times n, stop it upon hitting B or 0 and use the relative frequency of A to estimate α. Alternatively, one could run the process under $\tilde{P}_0$, for some $\tilde{P}_0$ conveniently chosen (*Importance Sampling*), and use $\tilde{\alpha}_n = \frac{1}{n}\sum_{i=1}^{i=n} 1_A \tilde{L}$ as an unbiased estimator.

In terms of variance the estimator $\tilde{\alpha}_n$ will be more economical than α_n if

$$n.\tilde{\sigma}_n^2 = \tilde{E}_0(1_A \tilde{L}^2) - \alpha^2 < E_0(1_A) - \alpha^2 \qquad (2)$$

and much more economical if $\tilde{L} \ll 1$, or in intuitive terms, if A is much more likely under $\tilde{P}$ than under P; hence the idea of *conditioning*. If we write the variance in the usual form,

$$\tilde{Var}(1_A \tilde{L}) = \tilde{E}_0 (1_A \tilde{L} - \alpha)^2$$

we can see another important fact: *We should try to find the change of measure that achieves (almost) constant likelihood ratio on A*. This is a common theme in all importance sampling applications (Ripley (1987)).

Fix the boundary B, let h be the function $h(i) = P_i(A)$, and let P^h be the transition matrix $P^h_{i,j} = P_{i,j} h(j)/h(i)$. Then, P^h is the transition matrix associated to the chain P_0 conditioned on A:

$$P_i(\textit{jump to } j/A) = \frac{P_i(\textit{jump to } j; A)}{P_i(A)} = \frac{P_{i,j} h(j)}{h(i)}$$

Write L^h for $\tilde{L}$ when $\tilde{P}$ is P^h. We note that P^h has a 'potential property':

1.2.1 Proposition: Upon reaching the boundary, L^h is constant ($= \alpha$) independently of the path.

The proof is easy, just write:

$$\text{if } \omega \in A, \textit{ then} \quad L^h = \prod_{(i \to j) \in \omega} \frac{P_{i,j} h(i)}{P_{i,j} h(j)} = \frac{h(0)}{h(b)} = \alpha.$$

In the above telescopic product, b is some point in B depending on the path and $h(b) = 1$. The previous proposition is very elementary; however, it tells the researcher where to look for an optimal change of measure. Note also that the property is exact, there is nothing asymptotic involved in it. However, some asymptotic analysis is always needed to simplify things. If we are interested in simulation of rare events that happen because the boundary is far away, then there is a natural limit process that results from boundaries going to infinity. This will give us the opportunity to use limit theorems and simplify the analysis.

Proposition: P_0^h is the unique Markovian $\tilde{P}_0$ such that:

$$\tilde{P}_0(A) = 1 \quad \text{and} \quad \tilde{L}(\omega) = \alpha \quad \text{for } \textit{every } \omega \in A \quad \text{(H)}$$

Proof (sketch): the likelihood of every path under $\tilde{P}_0$ is determined by the conditions (H) and P_0.

Then, we **propose the estimator**:

$$\hat{\alpha} \equiv \alpha_n^{\hat{P}} = \frac{1}{n}\sum_{i=1}^{i=n} 1_A \hat{L}, \qquad \hat{L} = \frac{dP_0}{d\hat{P}_0}$$

where $\hat{P}_0$ is some estimation of P^{h_0}. Hopefully $\hat{P}$ can be obtained either analytically, or, in non-markovian cases, with not very expensive simulations and some statistical techniques, or a combination of all of them. In the experiments described in Section 3 we exploit the fact that P^h depends only on the ratio of values of $h(.)$.

We note that, in principle, it is possible to obtain an estimate of P^h as close to it as desired, and that $\hat{L} = \hat{L}(\omega, \hat{P})$ is a continuous function of $\hat{P}$ for any of the usual distances in the space of $\hat{P}$. This points to the central issue of the project: how much computational effort should one devote to the estimation of P^h? Of course, the ideal would be to be able to decide the optimal change of measure analytically, as was partially done in Parekh and Walrand (1989); we say partially because even there a numerical minimization was required.

1.3 A Simple Case

It is elementary to check that for X_n, the M/M/1 queue with service rate μ and input rate λ (N is the boundary), the function $h(i)$ is

$$h(i) = \frac{\rho^i - 1}{\rho^N - 1}, \quad \textit{where} \quad \rho = \frac{\mu}{\lambda} \qquad (3)$$

Therefore the law P^h is, roughly, the one corresponding to the M/M/1 queue with service and input rates interchanged:

$$P^h_{i,i+1} = P_{i,i+1} \frac{h(i+1)}{h(i)} = \lambda \frac{(1-\rho^{i+1})}{(1-\rho^i)} \approx \mu \quad \text{for } i \textit{ large}$$

1.4 Theory

Say that a sequence of real functions $\{f_n\}$ defined on Z_+^d *converges* if the functions converge pointwise on any finite set. Say that a sequence of boundaries $\{B_n\}$ "increases to infinity" if the associated sequences $\{R_n\}$ eventually covers every point in Z_+^d.

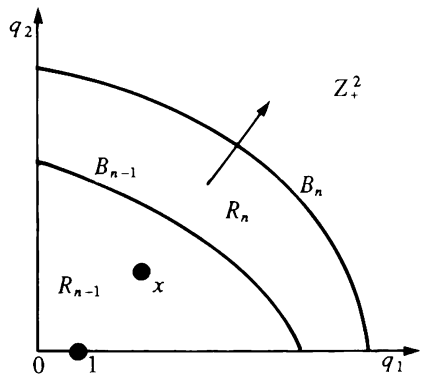

Figure 1: State Space and Boundaries

Consider the set $\mathbf{B} = \{ B : B \text{ is a boundary} \}$ where *boundary* has the same meaning as above. Then, if we let $1 \in Z_+^d$ be a reference point and use the notation $h_B(.)$ to stress the dependence on the boundary B, we have:

1.4.1 Proposition: The set $\left\{ \frac{h_B(.)}{h_B(1)} : B \in \mathbf{B} \right\}$ has the sequential compactness property: for every sequence of boundaries $\{B_n\}$ there is a converging subsequence of the ratios $\left\{ \frac{h_B(.)}{h_B(1)} \right\}$.

Proof: By the diagonal principle, it is enough to show that the functions can be uniformly bounded at each point:
Let $x \in R$ and $c(x)^{-1} = P_1(\textit{hit } x \textit{ before } 0)$. Then,

$$h_B(1) \geq P_1(\textit{hit } x \textit{ before } 0) P_x(A) \quad \textit{or} \quad \frac{h_B(x)}{h_B(1)} \leq c(x) .$$

Using this proposition, we can estimate successfully α for many of the multidimensional birth and death chains encountered in queuing applications.

1.4.2 Definitions: A real function h(.) defined on the state space of a Markov Chain with transition matrix P is super-harmonic, harmonic or sub-harmonic depending on whether $h \geq Ph$, $h = Ph$, *or* $h \leq Ph$.

1.4.3 Proposition: Provided the sequence $\{B_n\}$ increases to infinity, the limit of a converging subsequence of $\left\{ \frac{h_B(.)}{h_B(1)} \right\}$ is harmonic.

Note: the increasing condition is not needed; if not true we obtain a super-harmonic limit.

Proof : the function h_B is super-harmonic and harmonic except at B. By the Markov property, if $i \in B^c \cap R$ then $h(i) = \sum_j P_{ij} h(j)$ and if $i \in B$ then $1 = h(i) \geq \sum_j P_{ij} h(j)$.

1.5 Results from Boundary Theory

In this section, we sketch the main ideas from transient boundary theory of Markov Chains that are relevant for our purposes. See Freedman (1983) for notation and details. The following theorems are a guiding light in our search for convenient transformations of the chains we study.

Let K be the normalized Green function, k the Martin Kernel, p an initial distribution and h any P-super-harmonic function. Let Ω^* be the space of finite paths on Z_+^2 and let $(Z_+^2)^\infty$ have the product σ–*field* **B**. Consider in $(Z_+^2)^\infty$ the set Ω^∞ of paths that visit every state finitely often, with the inherited sigma field. Let $\Omega = \Omega^* \cup \Omega^\infty$. Our processes will live in Ω: they will be transient because they disappear (Ω^*) or because they leave forever any finite set (Ω^∞).

(*CT*) For all h, $K(i, X_n)$ converges P_{ph}^h a.s., and h is extreme iff

$$P_{ph}^h \{ \lim_n K(i, X_n) = h \} = 1.$$

(RT) There exist an invariant set F such that, for every h, there exist a unique probability m on **F** with $\int_F k dm = h$, *namely*, P_{ph}^h *restricted to* **F**.

The first theorem gives a formal presentation of the idea that h-transformed processes correspond to particular ways for the chain to "go to ∞ ". The second, gives a representation of super-harmonic functions in terms of the extreme harmonics obtained from the Martin Kernel. It is clear that these theorems are relevant to our problem.

1.6 A Simple Case: Continuation

Consider the p-up q-down random walk W_n on Z with $p < \frac{1}{2}$ and, the harmonic function g given by:

$$g(i) = R(i,j) = \frac{P_i(hit\ j)}{P_1(hit\ j)} = \frac{(\frac{p}{q})^{j-i}}{(\frac{p}{q})^{j-1}} = (\frac{q}{p})^{i-1} \quad \text{for } 1 \le i \le j.$$

Then P^g is the p-down q-up random walk. This is the transformation we obtained asymptotically in the presence of a reflecting boundary. We learn from this example that the transformed process P^h of X_n (see 1.3) agrees, away from the boundary, with the transformed process of the similar process W_n *without* the reflecting boundary. We will be guided by this when, in dealing with process in Z_+^d, we consider the harmonic transformations corresponding to the free process in Z^d hoping that, asymptotically and away from the boundaries, they will give the right answer.

1.7 How we use Boundary Theory

Queues in Parallel. As our first example, we study the case of two M/M/1 queues in parallel with input and service rates λ_i and μ_i, i=1,2. Assume that $\lambda_i < \mu_i$, i=1,2 for stability. We will use the notation $MP_{(\lambda_1,\lambda_2,\mu_1,\mu_2)}$ for such a system, (see Figure 2).

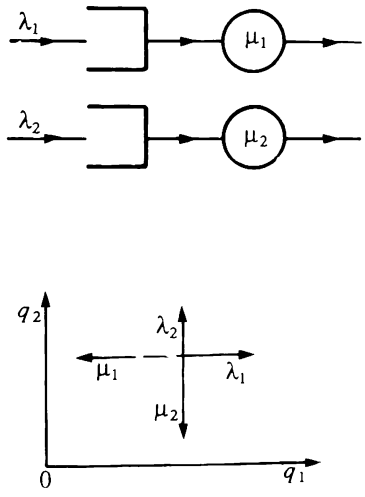

Figure 2: Queues in Parallel

This case was considered in Parekh (1986) and Parekh and Walrand (1989). Let P_0^* be the probability on **B** corresponding to the jumping chain started at (0,0). Let P_0 be obtained from P_0^* by means of the following alterations: the transitions from (0,1) to (0,0) and from (1,0) to (0,0) have now probability 0, instead of $\mu_2(\lambda_1+\lambda_2+\mu_2)^{-1}$ and $\mu_1(\lambda_1+\lambda_2+\mu_1)^{-1}$ respectively.

Then, by recurrence of P_0^*, the Green function of P_0 is finite and eventually the chain is going to disappear after reaching the set $\{(0,1),(1,0)\}$: $P_0(\Omega^*) = 1$. The evolution of the chain under P_0^* until it hits 0 is the same as under P_0 until it disappears. Since we want to understand how the original Process P_0^* goes far away before coming back to 0, it is clear that we need to understand and obtain those h for which $P_0^h(\Omega^\infty) = 1$.

Furthermore, (RT) is the type of representation we are looking for. Our limit of normalized hitting functions is harmonic; by (RT), it is a convex combination of extreme harmonics. A complete research program would include the search for a complete set of extreme harmonics of P_0, but we will not pursue this objective now.

For P_0 the functions

$$h_i(q) = h_i(q_1,q_2) = (\frac{\mu_i}{\lambda_i})^{q_i} \qquad i=1,2 \qquad (4)$$

are harmonic, P^{h_1} is the network $MP_{(\mu_1,\lambda_2,\lambda_1,\mu_2)}$ and P^{h_2} is the network $MP_{(\lambda_1,\mu_2,\mu_1,\lambda_2)}$. These two switching of rates were discovered in Parekh and Walrand (1989) and they correspond to the chain going to infinity as q_1 or q_1 go to infinity respectively. We note also that the product

$$h_{12} = h_1 \times h_2 \qquad (5)$$

is harmonic and $P^{h_{12}}$ is the network $MP_{(\mu_1,\mu_2,\lambda_1,\lambda_2)}$. The presence of this "product form" harmonic is unusual and is explained by the lack of interaction between the queues. However, we will use this product together with h_1 and h_2 later. This lack of interaction between queues makes the network a rather uninteresting one. However, the analytical calculation of α for some boundaries to be considered later is possible but complicated, and simulation and/or numerical methods seem to be appropriate.

Queues in Tandem. Consider two M/M/1 queues in tandem with input rate λ and service rates μ_1,μ_2. We will refer to this model often and use the notation $MS_{(\lambda,\mu_1,\mu_2)}$ for such a system. We will always assume that $\lambda < \min\{\mu_1,\mu_2\}$ for stability.

Let P_1^* be the probability on **B** corresponding to the jumping chain started at (1,0). Let P_1 be obtained from P_1^* by means of the following alteration: the transition from (0,1) to (0,0) has now probability 0, instead of $\mu_2(\lambda+\mu_2)^{-1}$. Then, by recurrence of P_1^*, the Green function of P_1 is finite and eventually the chain is going to disappear after reaching (0,1): $P_1(\Omega^*) = 1$.

For P_1 (in fact for P_1^*), the function

$$h_1(q) = h_1(q_1,q_2) = (\frac{\mu_1}{\lambda})^{q_1} \qquad (6)$$

is an extreme harmonic, and P^h is the network $MS_{(\mu_1,\lambda,\mu_2)}$. Irrespective of whether $\mu_1<\mu_2$ or not, it corresponds to the chain conditioned to go to ∞ by sliding near the axis $AX_1 = \{q: q_2 = 0\}$ and therefore P^h would give the almost optimal change of measure for a boundary B that has the property that, with probability close to one, the process hits B near its intersection with the set AX_1. This approximation is good enough for models like $MS_{(0.2,0.3,0.5)}$. It is one of the switching of rates discovered in Parekh (1986). The function

$$h_2(q) = h_2(q_1,q_2) = (\frac{\mu_2}{\lambda})^{q_1+q_2} \qquad (7)$$

is harmonic everywhere except on the set AX_1, where it is sub-harmonic. We will use h_1 and h_2 in Section 2. We note here that, although h_2 is not harmonic, intuitively it makes sense to consider it since under P^{h_2} the chain hits AX_1 only finitely often.

Consider the homogeneous random walk on Z^2 that has the same transitions as $MS_{(\lambda,\mu_1,\mu_2)}$ does on the region $\{q_1 > 0 \cap q_2 > 0\}$. It is possible to show that, if $\lambda+\mu_1+\mu_2 = \lambda x + \mu_1\frac{y}{x} + \mu_2\frac{1}{y}$, then the family of functions

$$h_{x,y} = h_{x,y}(q_1,q_2) = x^{q_1}y^{q_2} \qquad (8)$$

gives all the extreme harmonics that take the value 1 at (0,0) corresponding to this "free" walk. By Choquet's theorem, they must be unbounded. Of course, this random walk and $MS_{(\lambda,\mu_1,\mu_2)}$ are essentially different; nevertheless, the family of functions (8) provides us with a first step in the search of convenient transformations of the process $MS_{(\lambda,\mu_1,\mu_2)}$.

It is important to point out that the presence of boundaries (axes) is a real problem in dimensions greater than one, and it is the major stumbling block in the development of exact expressions for the harmonic transformations for our birth and death process in Z_+^d. In our project, we opted for the use of rough transformations as explained in Section 2.

Even if exact harmonics were to be found, the simplicity of the exponential ones should be emphasized. An exact but hard to implement formula would not be of much help.

The difficulty of dealing with the axes is also present when one attempts to construct a rigorous large deviations approach (see Parekh (1986).

1.8 OPTIMAL BUFFER ALLOCATION

In Anantharam (1988) the following Rule is considered: Consider an open network of J exponential servers with service rates $\mu_1,\mu_2,.....,\mu_J$, Bernoulli routing with matrix $R^{(J+1)\times(J+1)} = ((r_{i,j}))$, and exogenous Poisson arrivals of rate γ from the outside world o. The network satisfies the usual independence assumptions.

The optimal buffer allocation problem is the problem of how best to distribute a fixed number of available buffer space among the nodes of the network so as to optimize some performance criterion. A natural performance criterion is to maximize the expected time to buffer overflow, and this translates into the minimization of the probability α.

Assume that the network is stable, namely that the solution of the flow balance equations:

$$\lambda_i = \gamma r_{o,i} + \sum_{j=1}^{J}\lambda_j r_{j,i} \qquad 1\le i\le J,$$

satisfy $\lambda_i < \mu_i$, $1\le i\le J$. Then, the following rule is suggested by Anantharam:

1.8.1 Rule: In allocating N buffers to "maximize the time to buffer overflow", one should allocate roughly a fraction $p_i N$ of the buffers to node i, where p_i is proportional to $\log(\mu_i\lambda_i^{-1})$.

In Sections 2 and 3 we will consider what we call "symmetric examples", typically queues in tandem with equal service rate and equal buffers or queues in parallel with equal effective service rate $\mu_i\lambda_i^{-1}$ and equal buffers. Examples to keep in mind are $MS_{(0.2,0.4,0.4)}$ and $MP_{(0.2,0.2,0.3,0.3)}$ and the multidimensional generalizations. For queues in series, it is easy to see that the buffer sizes considered satisfy Rule 1.8.1. The routing matrix is very simple and the flow balance equations have the constant solution $\lambda_i = \gamma$, therefore the ratios $\mu_i\lambda_i^{-1}$ are constant. Similar results hold for queues in parallel.

This is the reason why we say that the problematic models considered in Section 2 are the most interesting ones. It is for them that the Large Deviations method described in Parekh and Walrand (1989) fail.

2 USING LD RESULTS

In order to test how well the notion of harmonic transformations and convex combinations of them describe the problem, we will consider the problematic models $MP_{(0.2,0.2,0.3,0.3)}$ and $MS_{(0.2,0.4,0.4)}$ and use the results obtained in Parekh and Walrand (1989) where, using large deviations, the authors found set(s) of "minimizing parameters" that hopefully give the optimal change of measure. It is for the cases where there are more than one set that it is hard to say which one to pick to do the change of measure. Using convex combinations h_β of the functions give by eqs (4), (5), (6), and (7) in Section 1 we will try to find a value of β such that P^{h_β} is roughly P^{h_B}. Here, $\beta = (\beta_1,\beta_2,\ldots,\beta_n)$ with $\beta_i \geq 0$ and $\sum_i \beta_i = 1$.

2.1 Queues in Parallel

The network $MP_{(0.2,0.2,0.3,0.3)}$ has three sets of minimizing parameters: (0.3,0.2,0.2,0.3), (0.2,0.3,0.3,0.2) and (0.3,0.3,0.2,0.2). They correspond, in the given order, to h_1, h_2 and h_{12} given by eqs (4) and (5) in Section 1.

Consider first convex combinations $\beta_1 h_1 + \beta_2 h_2$, with $\beta_1 + \beta_2 = 1$. For reasons of symmetry, it is clear that we can restrict our search for the optimal β to combinations that satisfy $\beta_1 = \beta_2 = 0.5$. With this in mind, we performed the experiments shown in Table 2.1, where we report the results of 13 experiments, each with 2,000 paths. Together with the estimator $\hat{\alpha}$ of α, we give the estimation $\hat{\sigma}$ of its standard deviation.

Comparable experiments done with $\beta_1 = 0$ *or* $\beta_2 = 0$ and 20,000 paths underestimated α by 20 percent or more. The results are shown in table 2.2. What about the inclusion of h_{12} in the convex combination? We run the process with several values of $\beta_1 = \beta_2 \geq \frac{1}{3}$ and obtained better answers with β_3 close to zero. It seems that a 10 percent mixture of h_{12} helps, as is reflected in the higher value of ρ, as compared to the results in Table 2.1.

We also show the numbers corresponding to three 2000-paths samples with $\beta_1 = \beta_2 = 0$, and one sample of 20,000 paths. Similar results were reported in Parekh and Walrand (1989) page 64.

$\alpha = 4.16\times10^{-04}$ *boundary*: $q_1 + q_2 = 25$ $\beta_1 = 0.5$ $\beta_2 = 0.5$ $\beta_3 = 0.0$			
# paths	$\hat{\alpha}\times10^5$	$\hat{\sigma}\times10^5$	$\rho = \hat{\alpha}\hat{\sigma}^{-1}$
2000	40.8	2.91	14.02
2000	41.1	3.04	13.50
2000	41.8	3.18	13.16
2000	39.3	2.81	13.99
2000	42.7	3.30	12.92
2000	40.3	3.16	12.76
2000	37.9	2.97	12.78
2000	42.7	3.29	12.98
2000	38.1	2.86	13.33
2000	39.9	3.11	12.82
2000	43.7	3.32	13.15
2000	43.3	3.80	11.41
2000	42.1	3.48	12.11
ave	4.105×10^{-4}	3.1×10^{-5}	
s.d.	1.89×10^{-5}		

Table 2.1

$\alpha = 4.16\times10^{-04}$ *boundary*: $q_1 + q_2 = 25$					
# paths	β_1	β_2	β_3	$\hat{\alpha}\times10^5$	$\hat{\sigma}\times10^5$	$\rho = \hat{\alpha}\hat{\sigma}^{-1}$
2000	0.00	0.00	1.00	34.1	4.93	6.9
2000	0.00	0.00	1.00	37.8	8.50	4.5
2000	0.00	0.00	1.00	47.4	8.90	5.3
20000	1.00	0.00	0.00	34.7	7.21	4.81
20000	0.00	1.00	0.00	31.5	5.43	5.80
20000	0.00	0.00	1.00	38.0	2.71	14.0
2000	0.33	0.33	0.33	40.6	3.08	13.2
2000	0.33	0.33	0.33	32.8	1.88	17.5
2000	0.33	0.33	0.33	35.0	3.26	10.7
2000	0.40	0.40	0.20	43.0	2.85	15.1
2000	0.40	0.40	0.20	39.8	2.96	13.5
2000	0.40	0.40	0.20	42.9	2.52	17.0
2000	0.45	0.45	0.10	39.9	1.60	24.9
2000	0.45	0.45	0.10	42.5	1.98	21.5
2000	0.45	0.45	0.10	39.6	1.76	22.5

Table 2.2

2.2 Queues in Series

We study now the model $MS_{(0.2,0.4,0.4)}$ with the aid of the harmonic function h_1 and the sub-harmonic h_2 given by equations (4) and (5) in chapter 1, and convex combinations $h_\beta = \beta h_1 + (1-\beta)h_2$.
P^{h_1} is $MS_{(0.4,0.2,0.4)}$ and P^{h_2} is $MS_{(0.4,0.4,0.2)}$, that is, we have the two rate switches proposed in Walrand and Parekh (1989).

Again, the idea is to find a value of β such that P^{h_β} is roughly P^{h_B}. This value of β must be determined experimentally, since we do not know which is the right mixture. We will use in our decision the sample variance as an estimate of the distance $d(P^{h_B}, P^{h_\beta})$. In all the experiments, we divide the work into the steps "find β" and "simulate using the chosen β". We decided to run the process for roughly 1000 paths for the first step, and we used them in two different ways according to the explanations below. The results of the experiments are summarized in tables 2.3 through 2.5. Next, we describe a few suggestions on how to pick β.

How to Find β. The *method of maxima* consists of running the process a number of times (referred to as "# paths" in the tables) under $P_0^{h_\beta}$ for a few (ten or so) values of β spaced uniformly in (0,1). For each β, we computed $\hat{\alpha}$, $\hat{\sigma}$, and $\rho = \frac{\hat{\alpha}}{\hat{\sigma}}$. The final simulation is done using the β that achieves the maximum of ρ (shown with a *).
The desire to have $\hat{\sigma}$ small is clear. We have observed that, in general, α is underestimated: large values of $\hat{\alpha}$ come together with large values of $\hat{\sigma}$ except for the right $\hat{P}$; and conversely, it is most common to have small values of $\hat{\sigma}$ associated with an underestimation of α. This seems to be true not only for our experiments but also for the ones done in Parekh and Walrand (1989). This method of finding β is shown in tables 2.3 and 2.4.

How to Find β. *Method of Few Values Near The Maxima* This is shown in table 2.5. We proceed as with the method of maxima explained above, but instead of picking the maxima, we choose a few (four in our case) values near it, and use an average of the estimates $\hat{\alpha}$ obtained by runing the process with these few values.

The overall conclusion seems to be that it may be risky to spend a large amount of resources on trying to find the exact β; the ρ curve seems to be rather peaked around it, and an average of near-values works well. This last method works as well as the one corresponding to the pure maxima, but we refer to Fresnedo (1989) where there is more information on the subject, as well a some other suggestions to find β that are too long to describe here.

2.3 Comments

In this Section we conjectured that the set of dominant tubes discovered by the Large Deviation method corresponds, roughly, to the set of dominant harmonics in the representation of h_B. An attempt to prove this should start with a rigorous explanation of the LD method and then with a rigorous connection between the two. We believe however that the experiments done give strong support to this conjecture and to the belief that there was a missing ingredient (the notion of convex combination) in the approach of Parekh and Walrand.

$\alpha = 1.81\times10^{-5}$ *boundary*: $q_1 + q_2 = 20$				
# paths	β	$\hat{\alpha}\times10^6$	$\hat{\sigma}\times10^6$	ρ
100	**0.00**	7.83	1.16	6.8
100	**0.17**	111	90.4	1.2
100	**0.33**	18.0	4.50	4.0
100	**0.50**	20.6	3.73	5.5
100	**0.67**	18.6	2.87	6.5
100	**0.83***	19.2	1.85	10.4
100	**1.00**	29.7	14.8	2.0
500	**0.83**	19.5	.78	24.8

Table 2.3

$\alpha = 3.54\times10^{-11}$ *boundary*: $q_1 + q_2 = 40$				
# paths	β	$\hat{\alpha}\times10^{12}$	$\hat{\sigma}\times10^{12}$	ρ
100	**0.00**	8.40	1.70	4.9
100	**0.10**	18.5	8.91	2.1
100	**0.20**	85.3	48.5	1.8
100	**0.30**	26.8	11.8	2.3
100	**0.40**	19.9	4.56	4.4
100	**0.50**	22.0	5.33	4.1
100	**0.60**	21.1	4.88	4.3
100	**0.70**	21.0	6.50	3.2
100	**0.80**	26.7	5.08	5.3
100	**0.90***	29.2	2.62	11.1
100	**1.00**	4.17	.67	6.2
2000	**0.90**	35.3	1.06	33.3

Table 2.4

$\alpha = 1.81\times10^{-5}$ *boundary*: $q_1 + q_2 = 20$				
# paths	β	$\hat{\alpha}\times10^6$	$\hat{\sigma}\times10^6$	ρ
100	**0.00**	18.0	8.14	2.21
100	**0.10**	18.8	6.41	2.94
100	**0.20**	37.8	16.1	2.34
100	**0.30**	25.3	6.03	4.20
100	**0.40**	12.2	2.58	4.74
100	**0.50**	14.9	2.21	6.76
100	**0.60**	15.4	2.85	5.38
100	**0.70**	14.3	1.90	7.53
100	**0.80***	21.5	2.20	9.79
100	**0.90**	13.4	1.63	8.22
100	**1.00**	7.1	1.60	4.46
250	**0.74**	18.7	1.78	10.53
250	**0.79**	16.9	1.15	14.73
250	**0.85**	19.9	1.77	11.27
250	**0.90**	18.4	1.65	11.16
total # paths =2000, ave of last 4 values = 1.84×10^{-5}				

Table 2.5

3. USING THEOREM 1.4.1

This Section describes the results of the "scaled-down model" method. It basically uses the compactness proposition proved in Section 1.

We first describe how the scaling is done and then report the results of experiments for a few ($n \leq 5$) queues in tandem, and different combinations of parameter values. The type of boundaries we consider are always of the cubic form $B = \{q : \max\{q_i : 1 \leq i \leq n\} = K_b\}$, for some integer K_b in the range 10 to 100.

About Notation. The subscripts s and b applied to hitting functions (h), regions (R), buffer sizes (K) and points in the region R (x or y), indicate that they refer to the small or big cube. Also, we identify regions R with the corresponding boundaries B: h_B is the same as h_R.

3.1 How the Simulation is Done

First Step. Given a big cube R_b with side size K_b, we choose a small cube, R_s, of side size K_s, usually ranging from $K_b/10$ to $K_b/2$. The theory underlying this choice is Proposition 1.4.1; however, we note here that it is probably hard to identify convergent subsequences and it is plausible that that a cube (equal side sizes K_s) will not do the best job. However, since our main concern is the symmetric case where all effective service rates are roughly the same, to simplify things and decrease the number of parameters in the experiments we just picked all buffers of the same size, both for R_b and R_s.
Then, starting with the function $1_{B_s}(.)$, we perform Gauss-Seidel relaxations (we used the methods in Greenberg and Vanderbei (1987)) to obtain h_{B_s}.

Second Step. We then calculate P^{h_s} for R_s. In the terminology of Section 1, P^{h_s} coupled with the mapping π described below, will give us an estimate $\hat{P}_0$ of $P_0^{h_b}$ corresponding to R_b.

Third Step. We run the process on R_b using (P^{h_s},π) as follows (see Figure 3). Imagine that we are at position x_b and we want to execute a jump according to $\hat{P}_0$. For the purpose of finding the next position y_b in R_b, we pretend that the particle is jumping in R_s. From $x_s = \pi(x_b)$, perform the jump according to P^h to obtain y_s and finally find y_b in R_b. For this procedure to work, a *mapping* π is needed: the one that says, for a point x_b in R_b, which is the corresponding point x_s in R_s. Note that the inverse mapping ψ from y_s to y_b is automatic: because of the nature of the problem, for a given transition in R_s, there is a unique transition in R_b. For example, if the transition in R_s indicates that there was a transfer from the second queue to the third, then the same thing happens in R_b. In other words, x_b, x_s and y_s determine y_b, and we stress that ψ is a function of three arguments. The following diagram tells the whole story:

$$x_b \xrightarrow{\ \pi\ } x_s \xrightarrow{\ jump\ } y_s \xrightarrow{\ \psi\ } y_b$$

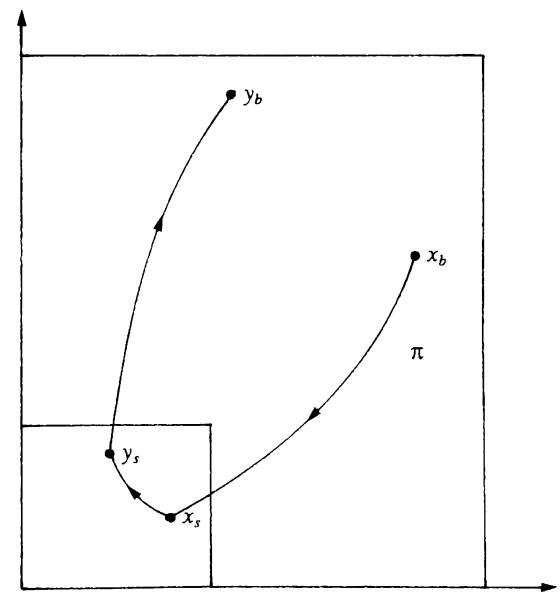

Figure 3: Mapping π from big cube to small cube

The Mapping π acts on each queue as follows: it has slope 1 near the axes and boundary and is linear in the middle. We refer to Fresnedo (1989) for details and the rationale behind this choice.

$\lambda = 0.20$, $\mu_1 = 0.4$, $\mu_2 = 0.4$ inner boundary = 10; outer boundary = 60 $\alpha = 3.465\times10^{-18}$		
#paths	$\hat{\alpha}\times10^{18}$	$\hat{\sigma}\times10^{18}$
1000	2.16	.75
1000	1.31	.34
2000	3.84	1.10
2000	3.67	1.38
3000	5.23	2.46
3000	3.99	1.16
weighted ave is 3.84×10^{-18}		

Table 3.1

$\lambda = 0.20$, $\mu_1 = 0.4$, $\mu_2 = 0.4$, inner boundary = 10; outer boundary = 30 $\alpha = 3.722\times10^{-9}$		
#paths	$\hat{\alpha}\times10^{10}$	$\hat{\sigma}\times10^{10}$
2000	40.4	2.19
2000	35.6	2.02
2000	35.7	1.98
2000	35.0	1.98
2000	40.9	2.27
2000	35.1	1.98
2000	38.1	2.08
ave is 3.726×10^{-9}		

Table 3.2

$\lambda = 0.1$, $\mu_1 = 0.2$, $\mu_2 = 0.4$, inner boundary = 10; outer boundary = 100 $\alpha = 1.051\times10^{-30}$		
#paths	$\hat{\alpha}\times10^{32}$	$\hat{\sigma}\times10^{32}$
50	108	4.38
100	104	.98
100	109	2.99
100	105	3.37
100	102	3.27
100	109	3.59
weighted ave is 1.06×10^{-30}		

Table 3.3

$\lambda = 0.2$, $\mu_1 = 0.3$, $\mu_2 = 0.5$, $\mu_3 = 0.5$, $\mu_4 = 0.5$ inner boundary = 10; outer boundary = 30 numeric evaluation of α was too costly		
#paths	$\hat{\alpha}\times10^{7}$	$\hat{\sigma}\times10^{7}$
2000	122	2.21
2000	124	2.20
2000	122	2.15
2000	118	2.11
2000	121	2.18
2000	121	2.16
2000	125	2.22
2000	121	2.16
ave is 1.22×10^{-5}		

Table 3.4

$\lambda = 0.2$, $\mu_1 = 0.4$, $\mu_2 = 0.5$, $\mu_3 = 0.5$, $\mu_4 = 0.4$,$\mu_5 = 0.4$ inner boundary = 4; outer boundary = 10 $\alpha = 2.992\times10^{-2}$		
#paths	$\hat{\alpha}\times10^{3}$	$\hat{\sigma}\times10^{3}$
500	29.6	2.68
500	31.6	2.64
1000	27.5	1.79
1000	29.7	1.82
1500	31.7	1.59
2000	28.9	1.27
3000	30.2	1.27
weighted ave is 2.99×10^{-2}		

Table 3.5

$\lambda = 0.2$, $\mu_1 = 0.4$, $\mu_2 = 0.5$, $\mu_3 = 0.5$, $\mu_4 = 0.4$ inner boundary = 7; outer boundary = 20 $\alpha = 1.07\times10^{-5}$		
#paths	$\hat{\alpha}\times10^{6}$	$\hat{\sigma}\times10^{6}$
100	28.3	12.7
1000	8.17	1.6
1000	9.74	1.6
1500	10.5	1.3
1500	12.5	2.0
1500	14.1	2.2
1500	11.7	1.8
1500	10.1	1.3
weighted ave is 1.13×10^{-5}		

Table 3.6

3.2 Comments

A numerical approach to obtain $\alpha = h(o)$ requires the solution of equations involving all components of the function h. Since we are interested only in the value of one of them, it is clear that simulations might have and advantage when the state space is very large, as is the case with networks. Here we have used a mixed method that solves a numerical problem similar to the original one but of a smaller dimension.

4. CONCLUSIONS

In Section 1 we stated a novel minima variance principle and showed the connection to a standard chapter in the theory of Markov Chains. In Section 2 the harmonics transformations corresponding to a boundary free process proved to be useful in dealing with the problem of balanced buffers. From the "convex combination" experiments we learned how to overcome the problems of a pure LD solution. In Section 3 a scaled down version of the model based on a compactness argument showed that a smaller size problem contained enough information to obtain reasonable changes of measure. We believe that these ideas will prove to be useful when dealing with networks of G/G/1 queues for which simulations are most useful, and refer to Fresnedo (1989) for some suggestions on this subject and many more experiments.

BIBLIOGRAPHY

Anantharam, V., "The Optimal Buffer Allocation Problem," Preprint, School of Electrical Engineering, Cornell University, Ithaca, NY, 1988.

Cottrell, M., Fort, J., Malgouyres, G., "Large Deviations and Rare Events in the Study of Stochastic Algorithms". IEEE Trans. A.C. vol. AC-28, No. 9

Freedman, David, "Markov Chains", Springer-Verlag, New York, 1983.

Fresnedo, R., "Quick Simulations of Rare Events in Networks", Ph.D. dissertation, Dept. of Statistics, Univ. Calif. Berkeley, 1989

Greenberg, A.G. and R.J. Vanderbei, "On Successive Approximation Methods for Stochastic Problems," Technical Memorandum, AT&T Bell Labs, Murray Hill, 1987.

Kelly, Frank, "Reversibility and Stochastic Networks", John Wiley & Sons, 1979.

Kemeny, John G., J. Laurie Snell, and Anthony W. Knapp, "Denumerable Markov Chains", Springer-Verlag, New York, 1976.

Parekh, S., "Rare Events in Networks", Ph.D. dissertation, Dep. EECS, Univ. Calif. Berkeley, 1986

Parekh, S., Walrand J., "A Quick Simulation Method for Excessive Backlogs in Networks of Queues", IEEE Transactions on Automatic Control, Vol. 34, No. 1, January, 1989

Ripley, Brian D., "Stochastic Simulation", John Wiley & Sons, New York, 1987.

Siegmund, D., "Importance Sampling in the Monte Carlo study of sequential tests". Annals of Statistics, 4, 673-684, 1976.

Walrand, Jean, "An Introduction to Queueing Networks", Prentice Hall, 1987

AUTHOR'S BIOGRAPHY

Roman D. Fresnedo studied Engineering in Uruguay, obtained his M.S. (Statistics) from CIENES (Chile, 1978) and his Ph.D. in Statistics from the University of California at Berkeley in 1989. His current research interests include the theory and practice of simulations and applications of stochastic processes and statistics to engineering problems.

Roman D. Fresnedo
University of California, Berkeley
Berkeley, CA 94720, U.S.A