QUICK SIMULATION OF A PERFORMABILITY MODEL*

W. Douglas Obal II
Electrical and Computer Engr. Dept.
University of Arizona
obal@ece.arizona.edu

William H. Sanders
CRHC, Coordinated Science Laboratory
University of Illinois
whs@crhc.uiuc.edu

Abstract

We consider the problem of estimating via simulation the tails of the performability of a system. Traditional simulation is inefficient in this case because the number of samples required to obtain an adequate characterization of the tails is very large. The system we consider is the M/M/1/K system subject to server breakdowns and repairs. This system is simple, but has the characteristics shared by more complex systems, namely boundary conditions and multiple structural states. We propose a solution to this problem that combines results from large deviations theory and importance sampling. Specifically, we use a large deviations analysis to derive an effective strategy for importance sampling. Using this strategy, we obtain results that indicate a large improvement in simulation efficiency relative to the traditional approach.

1 Introduction

Performability [3] is an effective "bottom-line" measure of an imperfect system's ability to meet user demands. Usually, the performability of a system is presented as the probability distribution function of some reward accumulated over a given interval of time. The accumulated reward might be the total number of customers served over some interval of time, for example.

Simulation is one method for evaluating the performability of a system. However, if one is interested in the tails of performability, straightforward simulation is very inefficient, due to the large number of samples required to adequately characterize the tails of the distribution. The performability tails of a subsystem can be important because of the impact of such events on other elements in the overall system. For example, since the departure process from one system is often the arrival process to another, it can be important to know the probabilities associated with excessive over- or under-performance of the first system, in order to evaluate the probabilities of the second system blocking or becoming idle.

Motivated by such thoughts, in this extended abstract we consider the evaluation of the performability of a queueing system subject to server breakdowns and repairs. We look at the total number of departures from the system over an interval of time, and seek the probability that this number exceeds or falls below threshold values far from the mean. We begin with the required background results from importance sampling. Then we carry out a large deviations analysis of our model and derive an importance sampling strategy for evaluating the performability tails. We present the results of our importance sampling simulation and compare them with results from naive simulation. Finally, we offer a few concluding remarks and some directions for further research.

*This work was supported, in part, by NASA Grant NAG 1-1782.

2 Importance Sampling

Importance sampling is a variance reduction technique that has been applied successfully to the problem of estimating rare events. For a good survey of results for dependability models and blocking probabilities, see [2] and the references therein. To characterize a simulation, let $(\Omega, \mathcal{F}, P)$ be the probability triple corresponding to the space of sample paths over the time interval $[0, T]$. Let $Y(\omega)$, $\omega \in \Omega$, be a random variable on this triple, and let Q be another probability measure on $(\Omega, \mathcal{F})$ such that for all $A \in \mathcal{F}$, $P(A) = 0$ if $Q(A) = 0$. The key to importance sampling is the transformation

$$\begin{aligned} E_P[Y] &= \int_\Omega Y(\omega) dP(\omega) \\ &= \int_\Omega Y(\omega) \frac{dP(\omega)}{dQ(\omega)} dQ(\omega) \\ &= E_Q[YL]. \end{aligned}$$

The function $L(\omega) = dP(\omega)/dQ(\omega)$ is called the Radon-Nikodym derivative, or likelihood ratio, and satisfies $P(A) = \int_A L(\omega) dQ(\omega)$, for all events $A \in \mathcal{F}$.

Using standard simulation, $E_P[Y]$ is estimated via $\xi(P) = \frac{1}{n}\sum_{i=1}^{n} Y(\omega_i)$, with ω_i sampled from $\Omega(T)$ according to P. To carry out an importance sampling simulation, $E_P[Y]$ is estimated via $\xi(Q) = \frac{1}{n}\sum_{i=1}^{n} Y(\omega_i)L(\omega_i)$ and the ω_i are sampled according to Q. Since $E_Q[Y^2L^2] = E_P[Y^2L]$ and $E_Q[YL]^2 = E_P[Y]^2$, variance reduction is achieved when Q is selected so that $E_P[Y^2L] < E_P[Y^2]$. For Y nonnegative, the optimal choice for Q is $dQ(\omega) = Y(\omega)dP(\omega)/E_P[Y]$, because then $E_P[Y^2L] = E_P[Y]^2$ and $\mathrm{VAR}[\xi(Q)] = 0$. Clearly the optimal choice can not be determined in advance, since its calculation requires $E_P[Y]$, knowledge of which obviates the simulation.

The optimal Q is out of reach, but there is still hope for obtaining a large variance reduction through a good choice of Q. The problem is that even P is not directly available to us, since it is induced by the stochastic elements of the simulation model. To produce Q, we must first decide what we want it to look like, and then we need to decide how best to induce it by altering the stochastic elements of the model. For this task, we utilize the theory of large deviations.

3 Large Deviations of Performability

Large deviations theory is focused primarily on the analysis of rare events that come about through the combined effects of a long sequence of minor anomolies. The portion of the theory we focus on is designed to yield estimates of the probability that a Markov process follows a path that is far from the "average" path, in a sense precisely defined below. This relates to our performability evaluation in the sense that we have an interest in sample paths that lead to endpoints far from the expected value of the reward accumulated over some interval of time. Shwartz and Weiss [5] have given a very accessible treatment of the theory of large deviations for Markov processes.

Let $\vec{x} = (x_1, x_2, x_3)$ denote the state of a queueing system subject to server breakdowns and repairs. All distributions are exponential, with arrival rate a, service rate b, server failure rate c, and server repair rate d. In addition, we limit the system to a finite capacity of K customers. We assign $x_1(t)$ the number in the system at time t, $x_2(t)$ the number served in the interval $[0,t]$, and we let $x_3(t) = 1$ if the server is up at time t, and 0 otherwise. Thus $(x_1, x_2) \equiv \vec{x}^{\,p}$ captures the performance part of the model, while x_3 corresponds to the structural part of the model. In the following analysis, when we speak of the process following a path, we will be speaking of the performance part of the process. The structural part of the process is utilized in the calculation of the local rate function, as described below.

The dynamics of the system are described in terms of the jump directions and the corresponding transition rates. We follow notation in Shwartz and Weiss. At each structural level m we have $k(m)$ jump directions denoted $\vec{e}_i(m)$. Associated with each jump direction is a transition rate $\lambda_i(m)$. The jump directions and rates for the queue with server breakdowns are shown in Table 1. There are three different operating zones at each of the two server states. The queue can be empty, occupied, or full. The empty queue and full queue boundaries pose a problem, because at these points the service or arrival rate abruptly drops to zero. However, within each of these operating zones, our model fits into the "finite levels" framework of Shwartz and Weiss, and their theory may be used to analyze the behavior of a boundary-free version of our model during a large deviation. The results on this "free" version of the process are useful in the derivation of the rate function for the restricted version of the process.

To begin with, we consider large deviations of the free queue with server breakdowns. For this model, we remove the queueing process barriers at zero and K.

Table 1: Jump directions and rates for performability model.

$m = 0$ (Server up)		
$\vec{e}_1(0) = (1,0,0)$	$\lambda_1(0) = a$	$x_1 < K$
	0	$x_1 = K$
$\vec{e}_2(0) = (-1,1,0)$	$\lambda_2(0) = b$	$x_1 > 0$
	0	$x_1 = 0$
$\vec{e}_3(0) = (0,0,-1)$	$\lambda_3(0) = c$	
$m = 1$ (Server down)		
$\vec{e}_1(1) = (1,0,0)$	$\lambda_1(1) = a$	$x_1 < K$
	0	$x_1 = K$
$\vec{e}_2(1) = (0,0,1)$	$\lambda_2(1) = d$	

Then the model is scaled to obtain $\vec{z}_n(t) \equiv \frac{1}{n}\vec{x}(nt)$. The next step is to examine the so-called fluid limit of the process, $\vec{z}_\infty(t)$, which for finite levels models can be shown to solve

$$\frac{d}{dt}\vec{z}_\infty(t) = \sum_{m=0}^{1} \phi(m) \sum_{i=1}^{k(m)} \lambda_i(m) \cdot \vec{e}_i^{\,p}(m),$$

where $\phi(m)$ solves

$$\sum_{m=0}^{1} \phi(m) \sum_{i=1}^{k(m)} \lambda_i(m) \cdot \vec{e}_i^{\,s}(m) = 0.$$

As $n \to \infty$, the rate of jumps increases, while the distance moved at each jump decreases. The stochastic process $\vec{z}_n(t)$ approaches the deterministic process $\vec{z}_\infty(t)$, which can be described by a differential equation. Plugging in our model parameters, we find $\phi(0) = d/(c+d)$ and

$$\frac{d}{dt}\vec{z}_\infty(t) = \frac{d}{c+d}(a(1,0) + b(-1,1)) + \frac{c}{c+d}(a(1,0)),$$

which simplifies to

$$\frac{d}{dt}\vec{z}_\infty(t) = (a - \frac{bd}{c+d}, \frac{bd}{c+d}), \vec{z}_{\infty,1}(t) > 0. \qquad (1)$$

The numeral one in the subscript of $\vec{z}$ refers to the first component of the process, which is the queue length.

The fluid limit is the asymptotic average path. Large deviations theory tells us that the probability of paths far from this average is very small. In fact, it is characterized by

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathrm{P}_{\vec{x}}(\vec{z}_n \in F) \leq$$
$$-\inf\{I_0^T(\vec{r}) : \vec{r} \in F, \vec{r}(0) = \vec{x}\}.$$

The set F is a closed set in $D^d[0,T]$, the space of R^d-valued functions of a parameter $t \in [0,T]$ that are

right continuous with left limits. The Skorohod metric is used to measure the distance between two functions in this space, which means that paths which differ slightly in terms of their jump times, but have similar shapes, are still considered close together. I is called the rate function, and is defined for paths $\vec{r} \in D^d[0,T]$ with $\vec{r}\,'(t) = \frac{d}{dt}r(t)$ as

$$I_0^T(\vec{r}) \equiv \int_0^T l(\vec{r}\,'(s))\, ds$$

where

$$l(\vec{y}) \equiv \sup_{\vec{\theta}} \left(\langle \vec{\theta}^{\,p}, \vec{y}^{\,p} \rangle - \max_m \sum_{i=1}^{k(m)} \lambda_i(m)(e^{\langle \vec{\theta}, \vec{e}_i(m) \rangle} - 1) \right).$$

The angle brackets stand for inner product, and the p superscript refers to the components of $\vec{y}$ and $\vec{\theta}$ corresponding to the performance part of the process $(\vec{x}^{\,p})$. The function l is called the local rate function, because it measures the "cost" of moving in a direction other than the average. It can be shown [5] that $l(\vec{y}) \geq l(\frac{d}{dt}\vec{z}_\infty(t))$ and that $l(\frac{d}{dt}\vec{z}_\infty(t)) = 0$. In essence, the large deviations principle states that the probability that the scaled process $\vec{z}_n$ remains close to some continuous path $\vec{r} \in F$ decreases exponentially with rate $nI_0^T(\vec{r})$ unless F contains $\vec{z}_\infty$, in which case the probability goes to one as n goes to infinity.

The particular value of $\vec{\theta}$, say $\vec{\theta}^{\,*}$, that solves the optimization problem $l(\vec{y})$ defines a change in the probability measure that makes the direction $\vec{y}$ the average direction of travel. This is the result that is useful in importance sampling, as first pointed out in [1], in the context of evaluating the stability of ALOHA. The focus is on the direction of travel, rather than on the path itself, because the direction of travel is determined by the generator of the process. The generator is derived directly from the stochastic parameters of the model, whereas we usually can not obtain the probability measure on the sample paths. Our approach to importance sampling is to change the generator so that at each point, the average direction becomes the direction that moves the process along the most likely path to the rare event. The path that minimizes I_0^T is the most likely path to the rare event.

For systems where the jump rates are independent of the performance part of the process $(\vec{x}^{\,p})$, Schwarz and Weiss have shown that for all paths $\vec{r}$, $I_0^T(\vec{r}) \geq T \cdot l(\frac{\Delta\vec{r}}{T})$, so we only need to consider straight line paths in order to calculate the minimum cost. If the rate function is not strictly convex, then there may be minimum cost paths that are not straight lines, but there will be a straight line path that attains the minimum cost. For straight line paths, calculating the rate function I reduces to calculating the local rate function. In the next section we calculate the local rate function for the performability of the queue with server breakdowns and repairs.

4 Calculating the Rate Function

For the boundary-free process,

$$\begin{aligned} l(\vec{y}) &= \sup_{\vec{\theta}}(\theta_1 y_1 + \theta_2 y_2 - a(e^{\theta_1} - 1) \\ &- \max\{b(e^{-\theta_1+\theta_2} - 1) + c(e^{\theta_3} - 1), \\ &\quad d(e^{-\theta_3} - 1)\}), \end{aligned}$$

as long as the system is not empty or blocked, corresponding to Equation (1). With a simple change of variable, $\theta = -\theta_1 + \theta_2$, the problem separates into two independent optimizations:

$$\begin{aligned} l_a(\vec{y}) &= \sup_{\theta_1}\left(\theta_1(y_1 + y_2) - a(e^{\theta_1} - 1)\right) \\ l_s(y_2) &= \sup_{\theta,\theta_3}(\theta y_2 - \max\{b(e^{\theta} - 1) + c(e^{\theta_3} - 1), \\ &\quad d(e^{-\theta_3} - 1)\}), \end{aligned}$$

with

$$l(\vec{y}) = l_a(\vec{y}) + l_s(y_2). \tag{2}$$

This result is appealing since the arrival process and service process are independent. While l_a has a simple analytic solution, l_s is complicated by the max function. However, a little thought shows that l_s can be solved using the method of Lagrange multipliers, since at a finite optimal point,

$$b(e^{\theta} - 1) + c(e^{\theta_3} - 1) - d(e^{-\theta_3} - 1) = 0. \tag{3}$$

As proof, assume there exists an optimal point (θ^*, θ_3^*) where (3) does not hold. But then we could improve the objective by adjusting θ_3^*, which contradicts the assumption of optimality. Letting

$$\begin{aligned} f(\theta,\theta_3) &= \theta y_2 - b(e^{\theta} - 1) - c(e^{\theta_3} - 1) \\ g(\theta,\theta_3) &= b(e^{\theta} - 1) + c(e^{\theta_3} - 1) - d(e^{-\theta_3} - 1), \end{aligned}$$

we solve the system

$$\begin{aligned} \nabla f(\theta,\theta_3) &= \nabla\lambda g(\theta,\theta_3) \\ g(\theta,\theta_3) &= 0, \end{aligned}$$

where ∇ is the gradient operator. The solution satisfies

$$\theta = \ln\left[\frac{y_2}{b}\left(1 + \frac{c}{d}e^{2\theta_3}\right)\right],$$

and $u = e^{\theta_3}$ solves

$$u^3 + \frac{d}{y_2}u^2 - \frac{d}{cy_2}(b + c - d - y_2)u - \frac{d^2}{cy_2} = 0.$$

The solutions of such a cubic equation are well known [6]. Using the definitions

$$\begin{aligned} Q &= \frac{3\frac{d}{cy_2}(y_2 - b - c + d) - \left(\frac{d}{y_2}\right)^2}{9}, \\ R &= \frac{9\frac{d^2}{cy_2^2}(y_2 - b - c + d) + 27\frac{d^2}{cy_2} - 2\left(\frac{d}{y_2}\right)^3}{54}, \\ S &= (R + \sqrt{Q^3 + R^2})^{1/3}, \\ T &= (R - \sqrt{Q^3 + R^2})^{1/3}, \end{aligned}$$

the discriminant is $D = Q^3 + R^2$. For the case $D > 0$ only one root is real, and it is given by $u_1 = S + T - d/(3y_2)$. When $D \leq 0$ there may be more than one real root, in which case we choose the positive one. In our experiments, no system has produced zero or multiple positive real roots. It would be nice to identify in terms of the model parameters when this situation might arise, and find an interpretation for such an event, but we have not worked on this yet. In the next section, all the analysis of the free model is extended to cover the boundary conditions of the original model.

5 Boundaries

In this section we derive the rate function for the model with boundaries. First the fluid limit is derived for the empty queue situation. The rate function is obtained using the contraction principle from large deviations theory, as described below.

Equation (1) describes the process when it is not at the boundary. In the actual system, when the queue becomes empty there can be no departures until the next arrival. If the system is stable, this boundary is attractive, meaning that the process rarely wanders far away, and always returns. To calculate the behavior at the boundary, we first calculate π_0, the steady state probability that the queue is empty, with the service rate modified by the server availability.

$$\pi_0 a + (1 - \pi_0)(a - \frac{bd}{c+d}) = 0.$$

The behavior of the process on the boundary is then derived via the vector equation

$$\pi_0 \phi(1)(a, 0) + (1 - \pi_0)\phi(1)(a - b, b) + \phi(0)(a, 0) = 0.$$

The result is very simple;

$$\frac{d}{dt}\vec{z}_\infty(t) = (0, a), \vec{z}_{1,\infty}(t) = 0. \tag{4}$$

When the process is on the boundary, where the fluid limit behavior is described by Equation (4), the rate function is complicated by the fact that no departures can occur from an empty system, so in effect the service rate is not continuous in terms of the performance part of the model, a violation of the assumptions underlying the finite levels theory in [5]. But this difficulty may be circumvented using the contraction principle from the general theory of large deviations. Simply put, the contraction principle states that if there exists a continuous mapping, M, of a process w known to satisfy a large deviations principle with rate function I_w, to another process of interest, y, then y also satisfies a large deviations principle and its rate function may be written as $I_y(s) = \inf\{I_w(\vec{r}) : M(r) = s\}$. A precise statement and a proof for a version of this theorem are given in [5]. For our performability model, the reflection map, well known in the queueing literature, can play the role of M. The reflection principle is a continuous mapping from $D^d[0, T]$ to $D^d[0, T]$ that effectively deletes all downward transitions when the queue is empty. The only extension required for the performability model is to augment the process with the number of departures.

The task that remains is to find the minimum cost paths that map into interesting paths in the performability model. The first path we consider is one that stays on the boundary at $x_1 = 0$. As stated in Equation (4), the fluid limit behavior on the boundary is to output customers at the arrival rate. As a result, we expect that for performability rates below the effective service rate, the optimal paths will have a cost given by the local rate function $l_a(0, y_2)$, which is the cost of twisting the arrival rate to y_2 from a while remaining on the boundary. According to the fluid limit, this system will output y_2 customers per unit time.

To see that the cost of staying on the boundary is lower than that of operating just off the boundary, examine the rate function for the interior given in Equation (2). When the process is on the boundary and $y_1 = 0$, the first optimization problem is the cost of twisting the arrival rate. In order to maintain a zero drift in the queueing process away from the boundary, the effective service rate must also be modified, with the associated cost given by the second optimization problem. The rate function is nonnegative, and zero only when the rates are unchanged, so the cost of staying on the boundary is clearly less than staying slightly inside.

To tie everything in the last couple of paragraphs together, the full description of the local rate function is presented in Equation (5):

$$l(x_1, \vec{y}) = \begin{cases} l_a(\vec{y}) & \text{if } x_1 = 0 \text{ and } y_1 = 0 \\ & \text{and } 0 < y_2 \leq \frac{bd}{c+d}; \\ l_a(\vec{y}) + l_s(y_2) & \text{if } x_1 > 0 \text{ or } y_1 \neq 0 \\ & \text{or } y_2 > \frac{bd}{c+d}; \\ \infty & \text{if } x_1 < 0 \text{ or } y_2 \leq 0. \end{cases} \tag{5}$$

The local rate function for system parameters $a = 9$, $b = 10$, $c = 0.1$, and $d = 10$, and $\vec{y} = (0, 9.5)$ is plotted in Figure 1 for $x_1 = 0$ and $x_1 = 1$.

6 Results

The large deviations results were used to estimate the right and left tails of performability for the queue with server breakdowns and repairs. As shown in Table 2, we obtained close agreement between the importance sampling simulations and the numerical results. Despite the fact that large deviations is concerned with asymptotic results, we obtained good results for these transient simulations. To obtain the results we used the importance sampling simulator (ITSim) and a numerical solver (trs) in the *UltraSAN* modeling environment [4]. We ran 10,000 independent replications for each experiment, and calculated confidence intervals at a confidence level of 95%. The first experiment evaluated the probability that a system with the given parameters would output at least 440 customers within a time period of 40 units. To obtain the twisted rates for the simulation, we solved $l(0, 0, 11)$, and found $\vec{\theta}^* = (0.200671, 0.103423, -0.102544)$, which made

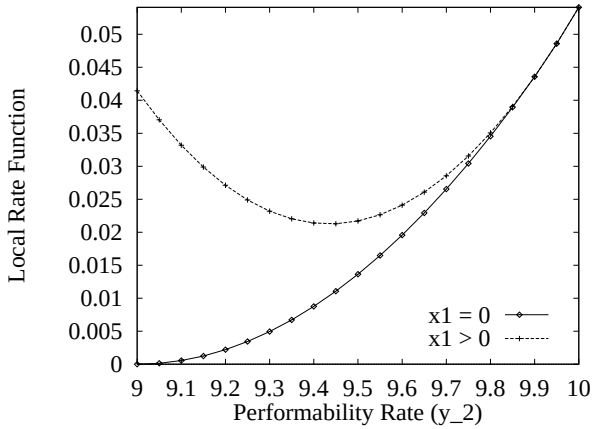

Figure 1: Local rate function for $a = 9$, $b = 10$, $c = 0.1$, $d = 10$, in the vicinity of $\vec{y} = (0, 9.5)$, plotted for the boundary and off-boundary cases.

the new model parameters $a' = 11$, $b' = 11.0896$, $c' = 0.0902539$, and $d' = 11.0799$. Note that the upper bound calculated directly using the rate function is not tight. We expect this bound to tighten as n increases. The second experiment examined the left tail of the same model. It evaluated the probability that the system would output at most 280 customers in the 40 unit time frame. For this model, we solved $l_a(0, 7)$, and found $\vec{\theta}^* = (-0.251314, 0, 0)$. In this experiment, the large deviations upper bound was much closer to the computed probability. Finally, for the third experiment, a different model was chosen, with $a = 2$, $b = 5$, $c = .1$, $d = 10$, and the right tail probability that the performability over a 40 unit time period exceeds 120 was evaluated. To obtain the twisted rates, we solved $l_a(0, 3)$ and got $\vec{\theta}^* = (0.405465, 0, 0)$. We note that the upper bound is fairly tight in this experiment, too.

7 Conclusions

We have considered the evaluation via simulation of the performability of a finite queue with server breakdowns and repairs. We used large deviations theory to study the process away from the boundaries, and then made the extensions required to handle the boundary conditions. Solving the resulting optimization problems led to a good importance sampling strategy, which produced a large improvement in simulation efficiency for the problem of estimating both right and left tails of performability.

For the relatively simple model considered here, we obtained an analytic solution of the rate function. With the goal of gaining insight into the properties of this particular model, this approach made sense. We want to study much more complicated processes with more complicated structural components. The most important problem that must be solved to extend the applicability of the approach outlined here is the automatic numerical solution of the correspondingly more complicated optimization problems that arise from such models.

Table 2: Performability Estimates

$a = 9, b = 10, c = 0.1, d = 10, t = 40$	
$P(Y \geq 440)$	
Solver	Result
trs	$3.2 \times 10^{-8} \pm 1.00 \times 10^{-9}$
ITSim	$3.2 \times 10^{-8} \pm 5.4 \times 10^{-9}$
Bound	2.5×10^{-5}
$P(Y \leq 280)$	
Solver	Result
trs	$2.3007 \times 10^{-5} \pm 1.00 \times 10^{-9}$
ITSim	$2.0 \times 10^{-5} \pm 3.3 \times 10^{-6}$
Bound	6.6×10^{-5}
$a = 2, b = 5, c = 0.1, d = 10, t = 40$	
$P(Y \geq 120)$	
Solver	Result
trs	$1.1728 \times 10^{-5} \pm 1.00 \times 10^{-9}$
ITSim	$1.17 \times 10^{-5} \pm 6.7 \times 10^{-7}$
Bound	1.8×10^{-4}

Acknowledgement

We are grateful to Alan Weiss for his rapid and helpful responses to our questions about the material in [5].

References

[1] M. Cottrell, J.-C. Fort, and G. Malgouyres, "Large deviations and rare events in the study of stochastic algorithms," *IEEE Transactions on Automatic Control*, vol. AC-28, no. 9, pp. 907–920, September 1983.

[2] P. Heidelberger, "Fast simulation of rare events in queueing and reliability models," *ACM Transactions on Modeling and Computer Simulation*, vol. 5, no. 1, pp. 43–85, January 1995.

[3] J. F. Meyer, "On evaluating the performability of degradable computing systems," *IEEE Transactions on Computers*, vol. C-22, pp. 720–731, August 1980.

[4] W. H. Sanders, W. D. Obal, M. A. Qureshi, and F. K. Widjanarko, "The *UltraSAN* modeling environment," *Performance Evaluation*, vol. 24, no. 1, pp. 89–115, October–November 1995.

[5] A. Shwartz and A. Weiss, *Large Deviations for Performance Analysis: Queues, Communications, and Computing*, Stochastic Modeling Series, Chapman and Hall, New York, 1995.

[6] M. R. Spiegel, *Mathematical Handbook*, Shaum's Outline Series, McGraw-Hill, New York, 1968.