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

SIMULTANEOUS AND EFFICIENT SIMULATION OF HIGHLY DEPENDABLE SYSTEMS WITH DIFFERENT UNDERLYING DISTRIBUTIONS

Philip Heidelberger
Victor F. Nicola
Perwez Shahabuddin

IBM T.J. Watson Research Center
P.O. Box 704
Yorktown Heights, New York 10598

ABSTRACT

Importance sampling is a well known technique that can be used for either variance reduction or obtaining performance estimates at multiple input parameter settings from a single simulation run ("what if" simulations). However, in queueing simulations, there is an essentially unique asymptotically efficient importance sampling distribution for estimating the probability of certain rare events (e.g., buffer overflows). Furthermore, this unique distribution depends critically on the inputs of the model, thereby making it difficult to obtain good "what if" estimates from a single run. (An example of this is using a single run to estimate the mean time until buffer overflow at multiple arrival rates.) In this paper, we show that a single importance sampling distribution can effectively be used for both variance reduction and "what if" simulation of certain rare events in models of highly dependable systems.

1 INTRODUCTION

Importance sampling (see, e.g., Hammersley and Handscomb (1964) or Glynn and Iglehart (1989)) is a technique that has been used for either variance reduction or for simultaneous estimation of output performance measures at multiple input parameter settings from a single set of simulation runs. Such simultaneous estimation has been called the "what if" approach in Rubinstein (1986) and Arsham et al. (1989). (An extension of this approach for use in optimization is described in Rubinstein (1991).) In this paper, we combine the variance reduction and "what if" aspects of importance sampling in a particularly effective manner for simulating rare events in a class of models of highly dependable computing systems. More specifically, from a single set of runs, we are able to simultaneously obtain very accurate estimates of the system failure time distribution corresponding to many different input failure rates and/or distributions. This is possible because the importance sampling method used has the following properties:

- It is independent of the input failure distributions.
- It is guaranteed to produce accurate estimates even as the input failure rates approach zero.

Contrast this situation to that of estimating the probability of rare events in queueing systems, e.g., in estimating the mean time until buffer overflow for a large buffer size. In such problems, the theory of large deviations has been used to select a good importance sampling change of measure; see Frater, Lenon and Anderson (1991), Parekh and Walrand (1989), Sadowsky (1991), and the references therein. The asymptotically optimal, and in a certain sense the only efficient, change of measure depends explicitly on the input parameters. As a simple example, Parekh and Walrand (1989) considered estimating the mean time until a stable M/M/1 queue first exceeds queue length N for large values of N. Suppose the queue has arrival rate λ and service rate μ. Then, for large N, the optimal change of measure interchanges λ and μ, i.e., an unstable system with arrival rate μ and service rate λ is simulated. While this is highly effective for a single given value of λ, suppose one were interested in estimating this quantity for a number of different values of λ, say $\lambda_1, \ldots, \lambda_k$. According to the above theory, one should simulate k different systems with k different service rates $\lambda_1, \ldots, \lambda_k$, since simultaneous estimation from a single set of runs is inefficient. Similarly, Arsham et al. (1989) and Rubinstein (1986) have found that, in simple queueing systems, the "what if" approach can be applied to estimate quantities such as expected queue lengths, but the variance typically increases as the parameter range increases.

However, in the highly dependable systems setting, we are able to obtain extremely accurate estimates of probabilities that span ten orders of magnitude (corresponding to two orders of magnitude change in input values) from a single set of replications.

The rest of the paper is organized as follows. In Section 2 we review the concepts of importance sampling and accurate estimation of rare events. In Section 3 we describe a class of models of highly dependable systems and a method of importance sampling, called exponential transformation, for simulating such models. In Section 4 the results of experiments combining variance reduction and "what if" simulation on several different models are described. Section 5 summarizes the results of the paper.

2 RARE EVENTS AND "WHAT IF" SIMULATIONS

Consider the problem of estimating the probability of a rare event A, i.e., $\alpha(\theta) \equiv E_{P_\theta}(I(A))$ where θ is some parameter, $I(A)$ is an indicator random variable corresponding to the event A and $E_{P_\theta}()$ denotes expectation when sampling under the distribution P_θ. First consider the estimation of $\alpha(\theta)$ for a single value of θ. The standard way of doing this is to use P_θ to generate n samples of $I(A)$ and form the estimator $\hat{\alpha}(\theta) = \sum_{i=1}^{n} I(A)^{(i)}/n$. (For any random variable, say Z, $Z^{(i)}$ denotes its i-th realization.) The variance of this estimator is given by $(\alpha(\theta) - \alpha^2(\theta))/n \approx \alpha(\theta)/n$ when $\alpha(\theta)$ is small. The relative half width (or relative error, $RE(\theta)$) corresponding to a $100 \times (1-\delta)\%$ confidence interval is given by $RE(\theta) \approx z_{\delta/2}/\sqrt{n\alpha(\theta)}$ where $z_{\delta/2}$ is a multiplier from the normal distribution. As the event becomes rarer, i.e., as $\alpha(\theta) \rightarrow 0$, then $RE(\theta) \rightarrow \infty$. This means that if $\alpha(\theta)$ is small, then n must be very large in order to accurately estimate $\alpha(\theta)$.

One way to make the simulation more efficient is to use importance sampling. Importance sampling makes use of the identity $\alpha(\theta) = E_{P'}(I(A)L_\theta)$ where P' is some other measure (P' must be chosen so that P_θ is absolutely continuous with respect to P') and L_θ is the likelihood ratio (Radon-Nikodym derivative). If P_θ and P' have densities p_θ and p', respectively, then $L_\theta(X) = p_\theta(X)/p'(X)$ where X represents a random sample. In this case we simulate the system using P' and form the new estimator $\alpha'(\theta) = \sum_{i=1}^{n} I(A)^{(i)} L_\theta^{(i)}/n$, the variance of which is given by $\sigma^2(\theta)/n \equiv [E_{P'}(I(A)L_\theta^2) - \alpha^2(\theta)]/n$. The main problem in applying importance sampling for variance reduction is to find an easily implementable P' such that $E_{P'}(I(A)L_\theta^2) \ll \alpha(\theta)$, i.e., the new variance is significantly less than the original variance. In estimating certain rare events in queueing systems and highly dependable systems, there exists a P' that yields orders of magnitude reduction in variance.

If the importance sampling distribution is such that the relative error, $\sigma(\theta)/\alpha(\theta)$, remains bounded as $\alpha(\theta) \rightarrow 0$, then it is said that the method satisfies the "bounded relative error" property. In effect, this property implies that only a fixed sample size is required to get an accurate estimate of $\alpha(\theta)$, no matter how rare the event is. In the case of highly dependable systems with exponential failure and repair distributions, an importance sampling heuristic described in Goyal et al. (1992) was shown to have the bounded relative error property for certain performance measures in Shahabuddin (1990) and Shahabuddin and Nakayama (1992). Similarly, Nakayama (1991) showed that certain derivative estimates obtained using this heuristic have bounded relative error. This heuristic was extended to such systems with more general failure and repair time distributions in Nicola, Heidelberger and Shahabuddin (1992), and the bounded relative error property in this case was established in Heidelberger, Nicola and Shahabuddin (1992). (See the above papers, and Juneja and Shahabuddin (1992) for more references.)

Now suppose we wish to estimate $\alpha(\theta)$ for several values of θ. Note that if the P' which is required for efficient simulation is independent of θ, then to perform "what if" simulations, we need only do one simulation run using P' and use likelihood ratios to get accurate estimates of $\alpha(\theta)$ at all values of θ of interest (there is one likelihood ratio for each θ). In many rare event simulations, efficient changes of measure are very much dependent on the parameters of the input distribution. This is particularly true, as mentioned in the Introduction, for the estimation of rare events in queueing systems. However, this is not true for changes of measure that are used to simulate highly dependable systems. We will describe such a change of measure in the next section.

3 EXPONENTIAL TRANSFORMATION

In this section we will give a brief review of the type of highly dependable systems we are considering, and the change of measure that is used to simulate them. The type of highly dependable systems for which this change of measure is efficient are basically those which can be modeled by the System Availability Estimator (SAVE) described in Goyal and Lavenberg (1987) except that now the component failure times and repair times are generally distributed, instead of exponentially distributed. For ease of presentation we will describe a simpler class of systems. Consider a system in which there are N types of components with N_i components of each type. Component number j of type i will sometimes be referred to as component (i, j). Each component can be in either of two states: up or down. The system is considered down when certain combinations of components of each type are down. When a component of type i fails, it may affect components of other types (with some probability) causing them to fail as well. There are different classes of repairmen (each component type is assigned a repairman class) and each repairman class repairs components with some priority discipline which can be fairly general. The problem is as follows: given a fixed time t, estimate the unreliability $U(t)$ which is defined to be the probability that the time to system failure, T_F, is less than t.

A basic assumption is that the system is composed of highly reliable components, so that the failure rates are much smaller than the repair rates. In this case, the event $\{T_F \leq t\}$ is rare and $U(t)$ is very small. In mathematical terms, the assumptions that is used is as follows. Let $h_i(x)$ denote the hazard rate of component type i when the age of the component is x. Then there exists a small but positive parameter ϵ such that $h_i(x) \leq \lambda_i \epsilon^{b_i}$ where $0 < \lambda_i < \infty$ and $b_i > 0$. The mean component repair times are considered to be of order one. Under this and some more minor assumptions it can be shown that $P(T_F < t)$ (and thus $VAR_P(I(T_F < t))$) is $\Theta(\epsilon^r)$ for some positive constant r. (A function $f(\epsilon)$ is $\Theta(\epsilon^r)$ if there exist two constants, K_1 and K_2 such that $K_1 \epsilon^r \leq f(\epsilon) \leq K_2 \epsilon^r$, for all sufficiently small $\epsilon > 0$.)

The change of measure which we use for highly dependable systems is called exponential transformation and was presented in Nicola, Heidelberger, and Shahabuddin (1992). A sample path consists of a sequence of component failure and component repair events. Exponential transformation applies a change of measure to the failure distributions of components. Repairs are sampled from their original distribution. Sample paths are generated as follows. Let t_n, $n \geq 0$, denote the time of the nth event (failure or repair) in the system. Suppose the event at time t_{n-1} has just taken place. Let R_n be the time of the next scheduled repair event after time t_{n-1}. We generate an exponentially distributed random variable X_n with rate α_n. If $t_{n-1} + X_n > R_n$ then the next event is a repair event. In that case we set $t_n = R_n$, schedule any repairs (that may have been enabled by the freeing of a repairman) and continue as before. However, if $t_{n-1} + X_n < R_n$ then the next event is a failure event and we set $t_n = t_{n-1} + X_n$. Let $A(s)$ denote the set of components that are operational at time s. Hence $A(t_n^-)$ is the set of components that are operational just before the nth event. In case of a failure event at time t_n, component $(i, j) \in A(t_n^-)$ is chosen as the failing component with some positive probability $q_{ij}(n)$ where $\sum_{(i,j) \in A(t_n^-)} q_{ij}(n) = 1$. The failing component may affect other components, in which case the failure propagation probabilities are sampled from their given distributions. Then any repairs that may have become possible due to preemption are scheduled and the process continues.

The likelihood ratio expression for this change of measure is given in Nicola, Heidelberger and Shahabuddin (1992). Assume that there exist finite positive constants $\underline{\alpha}$, $\overline{\alpha}$, $\underline{q}$, $\overline{q}$ such that $\underline{\alpha} \leq \alpha_n \leq \overline{\alpha}$ for all n and $\underline{q} \leq q_{ij}(n) \leq \overline{q}$ for all i, j and n. The first inequality in the first assumption states that, for small ϵ, the component failure event rates are much higher than before, i.e., the rate of component failure events has been accelerated. It can be shown that if the above properties are satisfied, then (under certain additional technical assumptions) estimates of $U(t)$ have the bounded relative error property.

There is considerable flexibility in choosing α_n and $q_{ij}(n)$ as long as all these quantities are greater than or equal to a constant. In actual implementation, the first failure time is sampled using an α_0 such that there is a significant probability f of this event occurring before the time horizon ends. This is called approximate forcing. For a given time horizon, $f = 0.8$ has been found to be quite good, and this value of f was used in our experiments. Once a component fails, then repair events are scheduled. In cases where there are on-going repairs in the system, α_n is chosen such that, unlike in the original system, there is a significant probability of a failure event happening before the next repair event. This is a version of what is called failure biasing. (Both forcing and failure biasing were introduced in the context of Markovian systems by Lewis and Böhm (1984).) In the experiments we describe, all components had exponentially distributed repair times with the same rate μ, and we set $\alpha_n = 0.5\mu$ whenever a repair is on-going. In the event that all components are up following a repair, we, in effect, "turn off" the acceleration of additional failures. More specifically, the next failure event is sampled from an exponential distribution whose rate equals the sum (over all components) of the individual maxima of the hazard rates. In these experiments, $q_{ij}(n)$ was selected as follows. Suppose at time t_{n-1} there were $M_i(n)$ components of type i up and $M(n)$ different types of components with at least one component up. Then $q_{ij}(n) = [1/M(n)] \times [1/M_i(n)]$. This is a version of what is called balanced failure biasing.

When the system is simulated as described above, the importance sampling distribution is independent of the different input failure distributions, thus allowing efficient "what if" simulations.

A method related to exponential transformation, based on combining uniformization with importance sampling, has also been described in Nicola, Heidelberger and Shahabuddin (1992). This approach also has bounded relative error property under similar conditions as those for exponential transformation.

4 EXPERIMENTAL RESULTS

In this section, we report on the results of simulation experiments to test the efficiency of our importance sampling techniques when used in "what if" simulations. We consider two test models: one simple and one complex.

The first model consists of two types of components sharing a single repairman. There are three components of type one and two components of type two. The system is considered operational if there is at least one component of each type operational. In this example, all failure and repair times are assumed to be independent and exponentially distributed. The repairman gives preemptive priority to type two components. The failure rate of component type i is denoted by λ_i and the repair rate for both types of components is fixed at $\mu = 1$. We parameterize the model by ϵ and consider several variations of this basic model. In a "balanced" system, the failure rates are $\lambda_1 = \lambda_2 = \epsilon$, while in an "unbalanced" system, the failure rates are $\lambda_1 = \epsilon$ and $\lambda_2 = \epsilon^2$. (A system is considered balanced if the failure rates of different types of components are of the same order of magnitude.) In addition, when a component of type two fails, it affects, or causes, two components of type one to fail simultaneously with probability a. We consider two cases of such "failure propagation": $a = 0$, in which case we say the model is without failure propagation, and $a = 0.25$, in which case we say the model has failure propagation. This model falls within the class of systems that can be modeled and solved numerically by SAVE.

A state space diagram of this model, without failure propagation, is shown in Figure 1. Referring to Figure 1, consider determining the most likely path (sequence of states) leading to a system failure before some fixed time t. For a balanced system and small ϵ, this most likely path is $P_b = \{(3,2) \rightarrow (3,1) \rightarrow (3,0)\}$ which represents two failures of component type two (with no repairs). The probability of this path is of order ϵ^2, and any other path leading to a system failure state is $o(\epsilon^2)$. Now consider an unbalanced system. In this case, the most likely path to failure is $P_u = \{(3,2) \rightarrow (2,2) \rightarrow (1,2) \rightarrow (0,2)\}$ which has probability of order ϵ^3. Notice that the balanced system's most likely failure path, P_b, has probability of order ϵ^4 in the unbalanced system, and is thus much less likely to occur than P_u. Similarly, in the balanced system, P_u is much less likely to occur than P_b.

The simulator described in Nicola, Heidelberger and Shahabuddin (1992) was modified to handle multiple failure distributions as input. We used the exponential transformation importance sampling method that was described in Section 3 to simultaneously simulate both the balanced and unbalanced systems (without failure propagation) for eight values of ϵ spanning two orders of magnitude (from 0.0001 to 0.0128). Thus estimates for 16 different systems were obtained from the same set of runs. A total of 256,000 replications were performed.

Figure 2 plots the point estimates for $U_\epsilon(100)$, the probability that the system fails before time $t = 100$. Notice that these point estimates span approximately 8 orders of magnitude. In addition, the most likely failure paths for the balanced and unbalanced systems are completely different, yet all point estimates were obtained, in a single pass, from the same set of replications. As will be seen in Table 1, the relative errors of these point estimates do not increase as ϵ decreases, in agreement with the bounded relative error property. The intuitive explanation for this behavior is as follows. With balanced failure biasing, all fairly direct paths to the set of failure states, including the most likely path(s), are traversed a reasonable number of times. This is enough to ensure good variance reduction. In this example, when path P_b is simulated, it contributes significantly to the likelihood ratio corresponding to the balanced system but contributes little to the likelihood ratio corresponding to the unbalanced system. In effect, this sample is used for the balanced system but is wasted for the unbalanced system. Similarly, when path P_u is simulated, it contributes significantly to the likelihood ratio corresponding to the unbalanced system but contributes little to the likelihood ratio corresponding to the balanced system.

Table 1 lists the values for $U_\epsilon(100)$ as calculated numerically by SAVE, the point estimates for $U_\epsilon(100)$ obtained from the simulation, and the relative half widths of 99% confidence intervals (expressed as a percentage). For example, in the balanced system without failure propagation and with $\epsilon = 0.0001$, SAVE calculates that $U_\epsilon(100) = 1.980 \times 10^{-6}$, the simulation point estimate is 1.982×10^{-6}, and the half width of a 99% confidence interval is 2.4% of the point estimate. Throughout the table, notice the close agreement between the numerical and simulation results. Notice also the stability of the estimates as represented by the relative confidence interval half widths. This is especially true for small values of ϵ, say $\epsilon \leq 0.0032$. In this case the relative errors in the balanced system are all less than 2.4%, while in the unbalanced system they are all less than 7.0%. There is some increase in the relative errors as ϵ increases, but this is not worrisome since in that case the event being estimated isn't particularly rare. (Actually, for $\epsilon = 0.0128$ the importance sampling results in a slight increase in variance over standard simulation.) For small ϵ, the importance sampling results in many orders of magnitude improvement in variances as compared to standard simulation.

As seen in Table 1, similar results were obtained for the systems with failure propagation. In this case, failure propagation does not change the order of magnitude of $U_\epsilon(100)$, but simply increases it over the value of the corresponding system without failure propagation. The results of Table 1 were generated from two sets of simulations, one set with failure propagation and one set without failure propagation.

The second example we consider is the computing system model that was considered in Goyal et al. (1992) and Nicola, Heidelberger and Shahabuddin (1992). The system has two sets of processors with two processors in each set, two sets of disk controllers with two controllers per set, and six disk clusters with four disks per cluster. Each set of disk controllers is attached to three different disk clusters and each set of processors is attached to both sets of disk controllers. Data on the disks are replicated in such a way that one disk in each cluster can fail without causing data loss. The system is defined to be available if at least one processor from each processor set has access to all data. This implies that at least one processor per set, one controller per set, and three disks per cluster must be operational. There is a single repairman who repairs components according to a FCFS discipline, and all repair times are assumed to be exponentially distributed with rate $\mu = 1$.

We considered four different failure distributions for the components with two different sets of means for each distribution, corresponding to a total of eight different sets of input distributions. The distributions were Erlang with two stages (with a coefficient of variation, CV, equal to 0.707), a Weibull with shape parameter equal to 1.25 (CV = 0.805), an Exponential (CV=1.0), and a Hyperexponential with two stages (CV = 2.0). For the Hyperexponential, the branching probabilities were 0.2727 and 0.7273 and the mean of the first stage was 12 times longer than the mean of the second stage. For a given system, all components were assumed to have the same type of distribution (with possibly different means), e.g. all Weibull. The two sets of means were as follows:

Set I: processors = 200,000; controllers = 200,000; disks = 600,000

Set II: processors = 20,000; controllers = 20,000; disks = 60,000

Table 2 reports on the results of simulating this model with the eight sets of input distributions and means. 256,000 replications were performed to estimate $U(t)$, the probability that the system fails before time t, for three different values of t: $t = 5$, $t = 50$ and $t = 100$. For approximate forcing, we used a parameter such that the first component fails before time $t = 100$ with probability 0.8. With this forcing parameter, the first component fails before time $t = 50$ about 55% of the time, while it fails before time $t = 5$ only about 8% of the time. Thus about 92% of the replications are, in effect, wasted for the estimates corresponding to $t = 5$. As seen in Table 2, the estimates span 13 orders of magnitude with a maximum relative error of 21%, and yet were obtained from the same set of sample paths using the same importance sampling change of measure. The largest errors, corresponding to the $t = 5$ estimates, are explained by the fact that at least 92% of the samples are wasted with the chosen forcing parameter. The estimates corresponding to $t = 50$ and $t = 100$ span ten orders of magnitude with a maximum relative error of 10%. Notice that, for a fixed value of t, changing the failure distribution (while keeping the mean fixed) can change $U(t)$ by as many as ten orders of magnitude.

5 CONCLUSIONS

This paper considered efficient estimation of the system failure time distribution in models of highly dependable systems. We investigated an importance sampling heuristic, called exponential transformation, that is provably good (in the bounded relative error sense). This method of importance sampling is essentially independent of the underlying input failure rates and/or distributions of the model. This fact can be exploited to simultaneously generate estimates for many different failure rates and/or distributions from a single set of replications. Because of the bounded relative error property, each estimate so obtained is guaranteed to be accurate. Experimental results performed on several test models showed the method to work well in practice: highly accurate estimates spanning many orders of magnitude were obtained from a single set of runs. Assuming that the overhead to calculate the required likelihood ratios is small, parametric studies can be performed very efficiently using this approach.

APPENDIX A: FIGURES AND TABLES

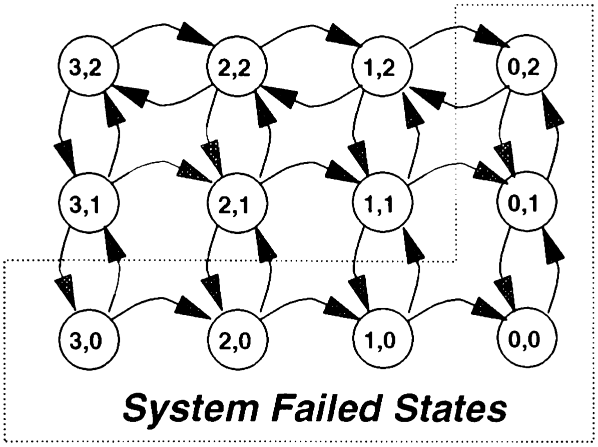

Figure 1: State Space Diagram of the Model With Two Types of Components (Without Failure Propagation)

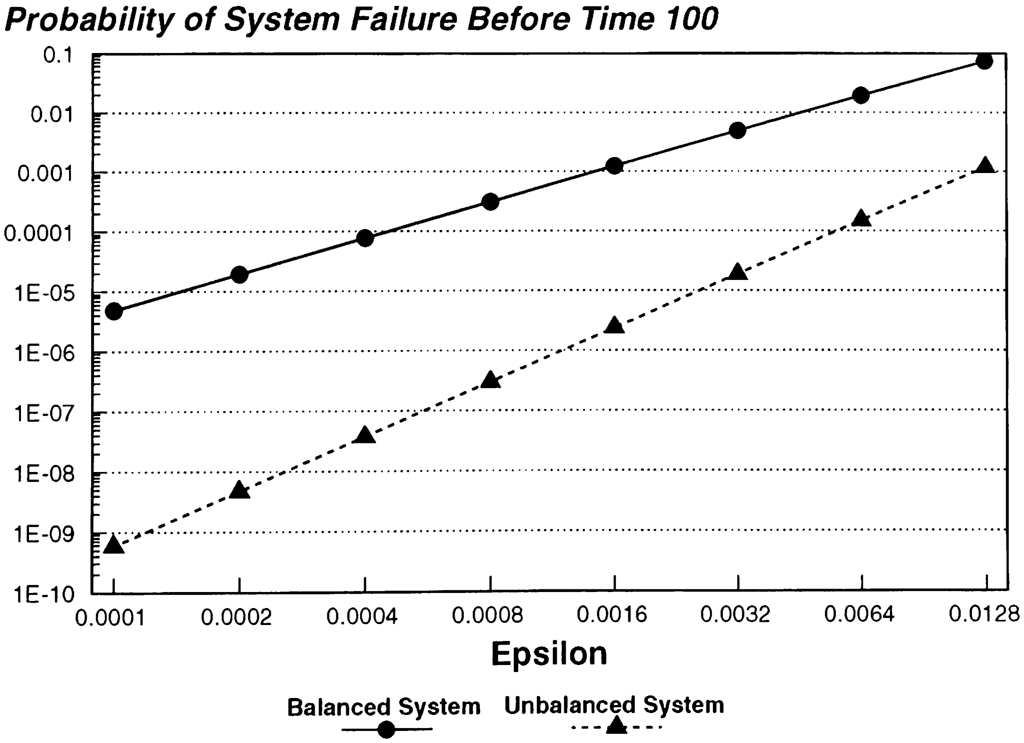

Figure 2: Point Estimates for the Model With Two Types of Components (Without Failure Propagation)

Table 1: Numerical and Simulation Results (Point Estimates and Relative Half Widths of 99% Confidence Intervals) for the Model With Two Types of Components

ϵ	Balanced Without Failure Propagation	Balanced With Failure Propagation	Unbalanced Without Failure Propagation	Unbalanced With Failure Propagation
0.0001	1.980×10^{-6} $1.982 \pm 2.4\%$	4.936×10^{-6} $4.956 \pm 2.1\%$	5.886×10^{-10} $5.948 \pm 6.7\%$	8.356×10^{-10} $8.078 \pm 4.6\%$
0.0002	7.920×10^{-6} $7.928 \pm 2.3\%$	1.974×10^{-5} $1.980 \pm 2.1\%$	4.701×10^{-9} $4.756 \pm 6.6\%$	6.677×10^{-9} $6.464 \pm 4.5\%$
0.0004	3.168×10^{-5} $3.171 \pm 2.2\%$	7.888×10^{-5} $7.906 \pm 2.0\%$	3.758×10^{-8} $3.799 \pm 6.5\%$	5.337×10^{-8} $5.173 \pm 4.4\%$
0.0008	1.267×10^{-4} $1.268 \pm 2.1\%$	3.150×10^{-4} $3.152 \pm 2.0\%$	3.002×10^{-7} $3.031 \pm 6.4\%$	4.265×10^{-7} $4.138 \pm 4.3\%$
0.0016	5.068×10^{-4} $5.078 \pm 2.1\%$	1.256×10^{-3} $1.255 \pm 2.1\%$	2.394×10^{-6} $2.411 \pm 6.4\%$	3.403×10^{-6} $3.302 \pm 4.4\%$
0.0032	2.026×10^{-3} $2.034 \pm 2.1\%$	4.986×10^{-3} $4.986 \pm 2.2\%$	1.905×10^{-5} $1.916 \pm 7.0\%$	2.709×10^{-5} $2.615 \pm 4.8\%$
0.0064	8.086×10^{-3} $8.130 \pm 2.2\%$	1.958×10^{-2} $1.960 \pm 2.2\%$	1.506×10^{-4} $1.533 \pm 10.2\%$	2.146×10^{-4} $2.036 \pm 5.4\%$
0.0128	3.204×10^{-2} $3.182 \pm 3.2\%$	7.451×10^{-2} $7.444 \pm 4.4\%$	1.178×10^{-3} $1.189 \pm 15.5\%$	1.682×10^{-3} $1.521 \pm 8.4\%$

Table 2: Simulation Results (Point Estimates and Relative Half Widths of 99% Confidence Intervals) for the Computing System Model

Parameters	t	Erlang(2) (CV =0.707)	Weibull (CV = 0.805)	Exponential (CV=1.0)	Hyperexponential (CV = 2.0)
Set I	5	2.393×10^{-18} $\pm 21.0\%$	5.061×10^{-12} $\pm 13.3\%$	1.501×10^{-9} $\pm 11.7\%$	1.351×10^{-8} $\pm 11.8\%$
	50	3.552×10^{-15} $\pm 9.4\%$	2.118×10^{-10} $\pm 5.8\%$	1.943×10^{-8} $\pm 4.9\%$	1.749×10^{-7} $\pm 4.9\%$
	100	3.126×10^{-14} $\pm 8.9\%$	6.281×10^{-10} $\pm 5.3\%$	4.018×10^{-8} $\pm 4.4\%$	3.611×10^{-7} $\pm 4.4\%$
Set II	5	2.391×10^{-14} $\pm 21.0\%$	1.601×10^{-9} $\pm 13.2\%$	1.500×10^{-7} $\pm 11.7\%$	1.350×10^{-6} $\pm 11.7\%$
	50	3.530×10^{-11} $\pm 9.4\%$	6.697×10^{-8} $\pm 5.7\%$	1.945×10^{-6} $\pm 4.8\%$	1.749×10^{-5} $\pm 4.8\%$
	100	3.086×10^{-10} $\pm 8.9\%$	1.983×10^{-7} $\pm 5.2\%$	4.001×10^{-6} $\pm 4.2\%$	3.564×10^{-5} $\pm 4.0\%$

REFERENCES

Arsham, H., A. Fuerverger, D.L. McLeish, J. Kreimer, and R.Y. Rubinstein. 1989. Sensitivity analysis and the "what if" problem in simulation analysis. *Math. Comput. Modelling* 12: 193-219.

Frater, M.R., T.M. Lenon, and B.D.O. Anderson. 1991. Optimally efficient estimation of the statistics of rare events in queueing networks. *IEEE Transactions on Automatic Control* 36: 1395-1405.

Glynn, P.W. and D.L. Iglehart. 1989. Importance sampling for stochastic simulations. *Management Science* 35: 1367 - 1392.

Goyal, A. and S.S. Lavenberg. 1987. Modeling and analysis of computer system availability. *IBM Journal of Research and Development* 31: 651-664.

Goyal, A., P. Shahabuddin, P. Heidelberger, V.F. Nicola, P.W. Glynn. 1992. A unified framework for simulating Markovian models of highly reliable systems. *IEEE Transactions on Computers* C-41: 36-51.

Hammersley, J.M. and D.C. Handscomb. 1964. *Monte Carlo Methods.* London: Methuen.

Heidelberger, P., V.F. Nicola and P. Shahabuddin. 1992. Bounded relative error in estimating transient measures of highly dependable non-Markovian systems. In preparation.

Juneja, S. and P. Shahabuddin. 1992. Fast simulation of Markovian reliability/availability models with general repair policies. In *Proceedings of the Twenty-Second International Symposium on Fault-Tolerant Computing*, 150-159, IEEE Computer Society Press, Boston, Massachusetts.

Lewis, E.E. and F. Böhm. 1984. Monte Carlo Simulation of Markov Unreliability Models. *Nuclear Engineering and Design* 77: 49-62.

Nakayama, M.K.. 1991. *Simulation of highly reliable Markovian and non-Markovian systems.* Ph.D. Thesis, Department of Operations Research, Stanford University, California.

Nicola, V.F., P. Heidelberger, and P. Shahabuddin. 1992. Uniformization and exponential transformation: Techniques for fast simulation of highly dependable non-Markovian systems. In *Proceedings of the Twenty-Second International Symposium on Fault-Tolerant Computing*, 130-139, IEEE Computer Society Press, Boston, Massachusetts.

Parekh, S. and J. Walrand. 1989. A quick simulation method for excessive backlogs in networks of queues. *IEEE Transactions on Automatic Control* 34: 54-56.

Rubinstein, R.Y. 1986. The score function approach for sensitivity analysis of computer simulation models. *Mathematics and Computers in Simulation* 28: 351-379.

Rubinstein, R.Y. 1991. How to optimize discrete-event systems from a single sample path by the score function method. *Annals of Operations Research* 27: 175-212.

Sadowsky, J.S. 1991. Large deviations and efficient simulation of excessive backlogs in a GI/G/m queue. *IEEE Transactions on Automatic Control* 36: 1383-1394.

Shahabuddin, P. 1990. *Simulation and Analysis of Highly Reliable Systems.* Ph.D. Thesis, Department of Operations Research, Stanford University, California.

Shahabuddin, P. and M.K. Nakayama. 1992. Fast simulation for transient measures and their gradients in highly reliable Markovian systems. In preparation.

AUTHOR BIOGRAPHIES

PHILIP HEIDELBERGER received a B.A. in mathematics from Oberlin College in 1974 and a Ph.D. in Operations Research from Stanford University in 1978. He has been a Research Staff Member at the IBM T.J. Watson Research Center since 1978. He is an area editor of the ACM's *Transactions on Modeling and Computer Simulation*, and has served as an associate editor of *Operations Research*, program chairman of the 1989 Winter Simulation Conference, and program co-chairman of the ACM Sigmetrics/Performance '92 Conference.

VICTOR F. NICOLA holds the Ph.D. degree in computer science from Duke University, the B.S. and the M.S. degrees in electrical engineering from Cairo University, and Eindhoven University of Technology, respectively. From 1979 he held scientific and research staff positions at Eindhoven University and Duke University. Since 1987, he has been a Research Staff Member at the IBM T.J. Watson Research Center. His research interests include performance and reliability modeling of computer systems, queueing theory, fault-tolerance and simulation methodology.

PERWEZ SHAHABUDDIN received a B. Tech. in Mechanical Engineering from the Indian Institute of Technology, Delhi, in 1984. After working for a year in Engineers India Limited, India, he joined Stanford University from where he received a M.S. in Statistics in 1987 and a Ph.D. in Operations Research in 1990. Currently he is a Research Staff Member at the IBM T.J. Watson Research Center. His research interests include modelling and analysis of computer performance/availability, and simulation methodology. He won the first prize in ORSA's 1990 George E. Nicholson Student Paper Competition.