EFFICIENT ESTIMATION OF THE MEAN TIME BETWEEN FAILURES
IN NON-REGENERATIVE DEPENDABILITY MODELS

Peter W. Glynn
Department of Operations Research
Stanford University
Stanford, CA 94305

Philip Heidelberger
Victor F. Nicola *
Perwez Shahabuddin
IBM T.J. Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598

ABSTRACT

In this paper we discuss fast simulation techniques for estimating the steady-state mean time between failures (MTBF) in non-Markovian models of highly dependable systems. The key is to use a ratio representation of the MTBF, in which the denominator is closely related to the probability of a rare event and is therefore amenable to estimation using importance sampling. A simulation methodology based on splitting and batch means, used for steady-state estimation in non-regenerative systems, can then be employed. Experiments using this methodology yield good results.

1 INTRODUCTION

This paper is concerned with efficient simulation techniques for estimating the mean time between failures (MTBF) in availability models of highly dependable computing systems. The class of models considered are basically those that can be described by the System AValiability Estimator (SAVE) package (see Goyal and Lavenberg (1987)), except that the failure and repair times can be generally distributed whereas they are restricted to be exponentially distributed in SAVE. In this class of models, system failure events occur rarely, so it is natural to try to improve the efficiency of the simulation by using importance sampling (see, e.g., Hammersley and Handscomb (1964) or Glynn and Iglehart (1989)). Since recent surveys (with numerous references) on the application of importance sampling to availability and reliability models are given in Nicola, Shahabuddin and Heidelberger (1993) and Heidelberger (1993), we will mention only a few of the most relevant references. (The latter paper also surveys the use of importance sampling for rare event simulation in queueing models.) Lewis and Bohm (1984) initiated work on fast simulation of

Markovian availability and reliability models. They introduced an importance sampling technique called failure biasing in which component failure events are accelerated with respect to component repair events so as to make system failures occur more frequently. Additional papers on failure biasing for Markovian models include Shahabuddin (1990 and 1991) and Goyal et al. (1992). In particular, Shahabuddin (1990 and 1991) proved that a form of failure biasing, called balanced failure biasing, is provably efficient (in the sense that the resulting estimates have bounded relative error) as the component failure rates approach zero. Efficient simulation techniques for estimating steady-state quantities in Markovian models rely on the regenerative method.

For non-Markovian models, an importance sampling approach based on rescheduling failure events is given in Nicola et al. (1991). Several provably efficient importance sampling heuristics (i.e., with bounded relative error) based on uniformization for estimating the system failure time distribution are given in Nicola, Heidelberger and Shahabuddin (1992), Heidelberger, Nicola and Shahabuddin (1992) and Heidelberger, Shahabuddin and Nicola (1993). Extensions of these techniques for estimating the steady-state unavailability in non-Markovian models are described in Nicola, Shahabuddin, Heidelberger and Glynn (1993). Although the regenerative method is not easily applicable in non-Markovian models, a ratio formula for steady-state measures, similar to the familiar ratio formula used in the regenerative method, can be exploited to devise efficient importance sampling schemes.

This paper describes how to adapt these importance sampling heuristics for estimating the MTBF. At first glance, importance sampling would not appear to be effective for estimating the MTBF; the MTBF is typically quite large whereas failure biasing produces unusually short failure times. However, a ratio formula for the MTBF also holds. Since one of the terms in the ratio is closely related to a rare event probability, failure biasing can be effectively applied.

*Victor F. Nicola is currently on leave from IBM in the group of Tele-Informatics and Open Systems, Department of Computer Science, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.
This approach is similar to that used for estimating the mean time to first system failure in Markovian models, in which a ratio formula is also exploited for effective importance sampling.

The rest of the paper is organized as follows. The type of dependability models we are considering and the ratio formula for the MTBF are described in Section 2. A description of the importance sampling methodology is described in Section 3, and empirical results are given in Section 4.

2 DEPENDABILITY MODELS AND THE RATIO FORMULA FOR THE MTBF

We start with a brief description of the class of dependability models under consideration. There are \( r \) types of components, with \( n_i \) components of type \( i \). Components are subject to failure and repair, thus affecting the availability of the system as a whole. Each component type is assigned a repairman class. There are a fixed number of repairmen in each repairman class. One or more types of components may be assigned to the same repairman class and the repairman fixes the component types according to some preemptive resume or non-preemptive priority rule on the component types; within a priority class First-Come-First-Served (FCFS) is used. A failure of a component may cause the simultaneous failure of other components with certain probabilities; this is called failure propagation. The techniques described in this paper can also be used for systems with operational/repair dependencies (the operation/repair of a component depends on some other components being up), spares for a component type, etc. However, for ease of presentation we will not consider these features in this paper. Let \( X(t) = (X_1(t), \ldots, X_r(t)) \) where \( X_i(t) \) is the number of components of type \( i \) that are up. Whether the system is considered up or down at time \( t \) depends only on \( X(t) \).

To put things in a Generalized Semi-Markov Process (GSMP) framework, let \( Z(t) \) denote the vector of ages of the ongoing component failure and repair processes at time \( t \) and let \( Q(t) \) denote the priority order of components waiting for each repairman class. Define the state of the system at time \( t \) to be \( S(t) = (X(t), Z(t), Q(t)) \). Then \( \{S(t) : t \geq 0 \} \) is a Markov process, or more precisely, a GSMP.

The quantity of interest in this paper is the MTBF, the definition of which we will now make more precise. Let \( N(t) \) be the r.v. (random variable) denoting the number of system failures in the interval \((0, t)\). Then under suitable regularity conditions that ensure ergodicity, there exists a constant \( \mu \) such that

\[
\lim_{t \to \infty} \frac{t}{N(t)} = \mu \tag{1}
\]

with probability one. The quantity \( \mu \) is defined to be the MTBF.

Let \( A \) be the set of states of \( \{S(t) : t \geq 0 \} \) when all components are up and one component has just finished repair. Define an \( A \)-cycle to be the process between two successive instants when \( \{S(t) : t \geq 0 \} \) enters the set \( A \). Let \( \pi \) be the steady-state distribution of \( S(t) \) conditioned on \( S(t) \) entering the set \( A \). Let \( N \) be the (random) number of system failures during an \( A \)-cycle and let \( \tau \) be the (random) duration of an \( A \)-cycle. Finally, let \( N_A(t) \) be the number of \( A \)-cycles completed during time \( t \). Then (assuming the limits exist)

\[
\mu = \lim_{t \to \infty} \frac{t}{N(t)} = \lim_{t \to \infty} \frac{t/N_A(t)}{N(t)/N_A(t)} \tag{2}
\]

\[
= \frac{E_{\pi, \phi}(\tau)}{E_{\pi, \phi}(N)} \tag{3}
\]

where the first subscript \( \pi \) in the expectation denotes the steady-state distribution of \( S(t) \) at the beginning of an \( A \)-cycle and \( \phi \) is the probability dynamics governing the sample path of \( \{S(t) : t \geq 0 \} \), from which \( \tau \) and \( N \) are obtained.

The standard procedure to estimate \( \mu \) would be to first run enough \( A \)-cycles so that the process is approximately in steady-state, i.e., to ensure that the distribution of \( S(t) \) is close to \( \pi \) at the beginning of the successive \( A \)-cycles. Then samples of \( \tau \) and \( N \) are collected from each successive \( A \)-cycle. An estimator of \( E_{\pi, \phi}(\tau) \) may be obtained as the sample mean of the \( \tau \)'s and an estimator of \( E_{\pi, \phi}(N) \) may be obtained as the sample mean of the \( N \)'s. Thus the ratio of the two sample means gives a natural estimator for \( \mu \) and the method of batch means (that takes into account the dependency between the \( A \)-cycles) can be used to obtain confidence intervals. However, although estimation of \( E_{\pi, \phi}(\tau) \) is easy, most samples of \( N \) are zero (as system failures are rare), and thus it is hard to estimate \( E_{\pi, \phi}(N) \). We describe an importance sampling based procedure to efficiently estimate \( E_{\pi, \phi}(N) \), and thus to efficiently estimate \( \mu \).

3 SIMULATION METHODOLOGY

Let \( \phi' \) be another probability dynamics (to be used as a change of measure for importance sampling) on the sample path of \( \{S(t) : t \geq 0 \} \). Then, using importance sampling, \( E_{\pi, \phi}(N) = E_{\pi, \phi'}(NL) \) where \( L \) is the likelihood ratio; roughly speaking, for any sample path \( \omega \), \( L \) is the ratio of the original probability of the sample path, \( \phi(d\omega) \), to the new probability of the sample path, \( \phi'(d\omega) \). The \( \phi' \) that we use is analogous to failure biasing that is used in Markovian systems. Typically, in systems with highly dependable components, each \( A \)-cycle consists of a component
failure event (that may cause the instantaneous failure of other components due to failure propagation) followed by component repair events (of the failed components). This is because the repair processes of the failed components take place at a much faster rate than the failure processes of the remaining operational components. In most cases the number of components that fail in a single failure event are not enough to cause system failure. The basic idea behind failure biasing is to accelerate the component failure processes with respect to the component repair processes, so that enough components fail in an A-cycle to cause system failure. In certain systems, the rates of different component failure processes may also differ by orders of magnitude; these are termed "unbalanced" systems. In a version of failure biasing called balanced failure biasing (Shahabuddin (1990), Goyal et al. (1992)), in addition to failure biasing, all component failure processes are made to occur at approximately the same rate.

There have been different implementations of the failure biasing and the balanced failure biasing idea, in the context of non-Markovian systems. The one used in this paper is based on the uniformization approach, and was introduced in Nicola, Heidelberg and Shahabuddin (1992) and Heidelberg, Shahabuddin and Nicola (1993). In Nicola, Shahabuddin, Heidelberg and Glynn (1993), this approach was used for the estimation of steady-state unavailability.

The simulation methodology that we use is based on a "splitting" approach. As before, we run enough A-cycles with the original measure \( \phi \), so that \( S(t) \) has a distribution close to \( \pi \) whenever it enters the set \( A \). At this point we run two parallel cycles; one with the original probability dynamics \( \phi \) and the other with the importance sampling probability dynamics \( \phi' \). Note that both cycles start from the same state, that has (approximately) the steady-state distribution \( \pi \). The cycle using \( \phi' \), called a "biased A-cycle," is used to obtain a sample of \( N \) and \( L \), say \( N_j \) and \( L_j \). The cycle using \( \phi \), called an "original A-cycle," is used to obtain a sample of \( \tau \), say \( \tau_j \). It is also used to obtain a starting point (having approximately the steady-state distribution \( \pi \)) for the next pair of original and biased A-cycles. By repeating this procedure \( n \) times, we obtain \((\tau_1, N_1, L_1), (\tau_2, N_2, L_2), \ldots, (\tau_n, N_n, L_n)\). Assuming the process actually is started in steady-state, then \( \sum_{i=1}^{n} \frac{\tau_i}{n} \) is an unbiased estimate of \( E_{\pi, \phi}(\pi) \) and \( \sum_{i=1}^{n} \frac{N_i L_i}{n} \) is an unbiased estimate of \( E_{\pi, \phi'}(NL) = E_{\phi}(N) \). The ratio of the first estimator to the second one yields an estimate of \( \mu \). As the successive cycles are dependent, the method of batch means can be used to obtain confidence intervals. We will briefly review the procedure in the context of ratio estimation.

Divide the \( n \) samples into \( b \) batches, with \( k = n/b \) samples in each batch (assume that \( n \) is chosen such that \( n/b \) is an integer). Form the samples \( \delta_1, \delta_2 \ldots \delta_j, \gamma_1, \gamma_2 \ldots, \gamma_b \) as given below:

\[
\delta_j = \frac{1}{k} \sum_{i=(j-1)k+1}^{jk} \tau_i \tag{4}
\]

\[
\gamma_j = \frac{1}{k} \sum_{i=(j-1)k+1}^{jk} N_i L_i. \tag{5}
\]

Note that \( \tilde{\mu} = \sum_{j=1}^{b} \delta_j / \sum_{j=1}^{b} \gamma_j \) is the same estimator as would be obtained without batching. For a sufficiently large batch size \( k \), \( \{(\delta_j, \gamma_j), j \geq 1\} \) can be considered to constitute an uncorrelated sequence. In that case, for sufficiently large \( b \), \( \sqrt{b}(\tilde{\mu} - \mu) \) is approximately normally distributed with mean zero and variance \( \sigma^2 \) where (analogous to the regenerative method),

\[
\sigma^2 = \frac{\text{Var}_\phi(\delta_j) - 2 \mu \text{Cov}_\phi(\delta_j, \gamma_j) + \mu^2 \text{Var}_\phi'(\gamma_j)}{E_{\phi}^2(\gamma_j)}. \tag{6}
\]

(The subscript \( \pi \), indicating the steady-state distribution, is implicitly understood but has been dropped from Equation 6.) The normal approximation and Equation 6 can then be used to construct confidence intervals.

In applications, more simulation effort is typically required to obtain accurate estimates of \( E_{\pi, \phi}\{N\} \), so \( m \geq 1 \) biased \( A \)-cycles can be generated for each original \( A \)-cycle. Let \( N_{il} \) and \( L_{il} \) be the samples obtained from the \( l \)-th biased \( A \)-cycle (\( i = 1, \ldots, m \)) corresponding to the \( i \)-th original \( A \)-cycle (after reaching steady-state), i.e., \( N_{il}, L_{il} \) and \( \tau_i \) have the same starting state. Then a new sample for the \( j \)-th batch \( \gamma_j \) can be defined to be

\[
\gamma_j = \frac{1}{km} \sum_{i=(j-1)k+1}^{jk} \sum_{l=1}^{m} N_{il} L_{il}. \tag{7}
\]

The other equations relating to this procedure remain unchanged.

4 EXPERIMENTAL RESULTS

In this section we describe the results of experiments using the importance sampling methodology described in Section 3 to estimate the MTBF. The example models chosen are the same as those described in Nicola, Heidelberg, Shahabuddin and Glynn (1993) where they were used for estimating the steady-state unavailability.

After the initialization effects had dissipated, each model was run for 64,000 original \( A \)-cycles. For each
original $A$-cycle, $m = 4$ biased $A$-cycles were simulated with importance sampling using the same starting state as the corresponding original $A$-cycle. The method of batch means with $1,000$ batches was used to estimate the variance ($64$ original and $256$ biased $A$-cycles per batch). The tables list point estimates for the MTBF and the relative half-width (in percent) of $99\%$ confidence intervals.

The first example is a machine repairman model with two types of components and three components of Type I and two components of Type II. The system is considered operational if there is at least one component of each type operational. There is a single repairman who repairs components according to a preemptive-resume service discipline (with components of Type II having the highest priority). The repair time distribution is deterministic ($1.0$ hour) for Type I components and it is uniformly distributed between $0$ and $1.0$ hour for Type II components. Two kinds of failure distributions are considered: an Erlang and a Hyperexponential. These distributions have means $1/e$, where $e$ is a "reliability" parameter and $c = 1$, or $c = 1.5$. (Allowing $c = 1.5$ allows us to model unbalanced systems.) The Erlang distribution, denoted by $E_2(e)$, has two stages with rate $2e$ in each stage. This distribution has a coefficient of variation (CV) equal to $0.707$. The Hyperexponential distribution, denoted by $H_2(e)$, is equal (in distribution) to an exponential with rate $0.3342e$ with probability $0.2727$ and it is equal to an exponential with rate $4.01e$ with probability $0.7373$. The CV of this distribution is $2.0$. This model can either be with or without failure propagation. In the model with failure propagation, a failure of a Type II component causes two Type I components to fail with probability $0.25$.

The second example is a model of a fault-tolerant computing system. There are two sets of processors with two processors per set, six disk clusters with four disks per cluster, and two sets of disk controllers with two controllers per set. The system is considered available if there is at least one operational processor in each processor set, one operational controller in each controller set, and three operational disks in each disk cluster. There is a single repairman who repairs components according to a FCFS discipline. All repair times are exponentially distributed with mean one. Component failure time distributions could be either Erlang with two stages (CV=0.707), Weibull with a shape parameter equal to 1.25 (IFR with CV=0.805), Exponential, or Hyperexponential as described above (CV=2.0). Within a given experiment, all components had the same type of failure distribution, e.g. Weibull, but with possibly different means. Two sets of mean failure times were considered. In Set I, processors and controllers had a MTBF of 200,000 hours while disks had a MTBF of 600,000 hours. In Set II, the components were less reliable by a factor of ten, i.e., in Set II, processors and controllers had a MTBF of 20,000 hours while disks had a MTBF of 60,000 hours. Again, the model could be either with or without failure propagation. In the model with failure propagation, a failing processor causes a processor in the other set to fail with probability $0.1$.

For Markovian systems, failure biasing makes the probability that the next event is a failure equal to $p$ for some fixed $p$ whenever repairs are ongoing. (Without importance sampling this probability is usually very small.) Typically, a value of $p = 0.5$ yields good results. A uniformization-based importance sampling approach as described in Nicola, Shahabuddin, Heidelberger and Glynn (1993) was used for the biased $A$-cycles. In this approach, repair times are sampled from their given distributions and uniformization (with importance sampling) is used to sample failure times. Roughly speaking, at the $n$-th event, a Poisson process with rate $\beta_n$ is sampled; events in this Poisson process are accepted as failure events with probability $p_n$ and rejected with probability $(1 - p_n)$. Rejected events can be considered pseudo-events and have no effect on the "non-clock" part of the system state $X(t)$ and $Q(t)$. Similar to failure biasing, for non-Markovian systems we want to make the probability that the next event is a failure event approximately equal to $0.5$ when repairs are ongoing. Let $r_n$ denote the expected repair time of the component in repair on the $n$-th event. Thus making the product $\beta_n \times p_n = r_n$ has the desired effect. Guided by the results of previous experimentation, we fixed $\beta_n = 5 \times r_n$ and adjusted $p_n$ accordingly. If the event was a failure event, balancing was used to select the failing component, i.e., the failing component type was selected uniformly. In order to keep the estimates stable, if the actual time to the next repair event was much higher than the mean repair time (say, greater than five times as large), importance sampling of failures was turned off. Importance sampling was also turned off for the remainder of the cycle whenever the system failed. Uniformization was also used to sample failure times when no repairs were ongoing. In this case, the total probability of accepting an event was unchanged from the real system, however, balancing was used to select which component to fail given that the event is a failure event. (For the Weibull distribution, which has an unbounded failure rate, uniformization is, strictly speaking, not applicable. In this case the failure rate was truncated in order to sample using uniformization.)

Results for the machine repairmen models are given in Tables 1 and 2 for the cases without and with failure propagation, respectively. All relative errors are less than $\pm 10\%$. For $\epsilon = 0.01$, the variance reduction over standard simulation was typically less than
a factor of ten, but for $\epsilon = 0.0001$, no system failure events were observed in a standard simulation of the same number of cycles. Similar results, listed in Tables 3 and 4, were obtained for the computing system example. Note that, for fixed component mean failure times, the MTBF is little affected by the failure time distribution, especially in the models without failure propagation. These models are, in some sense, close to product form queueing networks that do exhibit such insensitivity to the form of the distribution (at infinite server stations).

In our experiments, we have found this approach to importance sampling to be effective for estimating the MTBF and system unavailability in highly dependable systems. However, a proof that it possesses the bounded relative error property has not been established.

REFERENCES


AUTHOR BIOGRAPHIES

PETER W. GLYNN After receiving his Ph.D. in Operations Research from Stanford University, Peter Glynn joined the faculty at the University of Wisconsin-Madison, where he held joint appointments in the Industrial Engineering, Computer Science, and Mathematics departments. In 1987, he returned to Stanford University, where he currently holds the Thomas Ford Faculty Scholar Chair in the Department of Operations Research. His research interests include discrete-event simulation, computational probability, queueing, and general theory for stochastic systems.

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 associate 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, North Carolina, the B.S. and the M.S. degrees in electrical engineering from Cairo University, Egypt, and Eindhoven University of Technology, The Netherlands, 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. He is currently on leave from IBM and holds the position of a Visiting Associate Professor in the group of Teleinformatics and Open Systems at the University of Twente. His research interests include performance and reliability modeling of computer and communication systems, queueing theory, fault-tolerance and simulation.

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 modeling and analysis of the performance/availability of computer and communication systems, rare event simulation and simulation output analysis. He won the first prize in the 1990 George E. Nicholson Student Paper Competition organized by the Operations Research Society of America.

**APPENDIX A: TABLES**

<table>
<thead>
<tr>
<th>Type</th>
<th>$E_2(e)$</th>
<th>$E_2(e^t)$</th>
<th>$H_2(e)$</th>
<th>$H_2(e^t)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>3.01 x 10^3</td>
<td>3.85 x 10^4</td>
<td>2.25 x 10^3</td>
<td>3.47 x 10^4</td>
</tr>
<tr>
<td>II</td>
<td>2.98 x 10^7</td>
<td>4.41 x 10^9</td>
<td>2.32 x 10^7</td>
<td>4.21 x 10^9</td>
</tr>
</tbody>
</table>

Table 2: Estimates of the MTBF in the machine repairman model (with failure propagation).

<table>
<thead>
<tr>
<th>Set</th>
<th>$E_2$</th>
<th>Weibull</th>
<th>Exponential</th>
<th>$H_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>2.52 x 10^6</td>
<td>2.51 x 10^6</td>
<td>2.53 x 10^6</td>
<td>2.56 x 10^6</td>
</tr>
<tr>
<td>II</td>
<td>2.54 x 10^7</td>
<td>2.57 x 10^7</td>
<td>2.48 x 10^7</td>
<td>2.53 x 10^7</td>
</tr>
</tbody>
</table>

Table 3: Estimates of the MTBF in the computing system example (without failure propagation).

<table>
<thead>
<tr>
<th>Set</th>
<th>$E_2$</th>
<th>Weibull</th>
<th>Exponential</th>
<th>$H_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>2.35 x 10^6</td>
<td>2.35 x 10^6</td>
<td>2.23 x 10^6</td>
<td>2.11 x 10^6</td>
</tr>
<tr>
<td>II</td>
<td>2.33 x 10^7</td>
<td>2.33 x 10^7</td>
<td>2.26 x 10^7</td>
<td>2.00 x 10^7</td>
</tr>
</tbody>
</table>

Table 4: Estimates of the MTBF in the computing system example (with failure propagation).