A Unified Framework for Simulating Markovian Models of Highly Dependable Systems

Ambuj Goyal, *Member, IEEE*, Perwez Shahabuddin, Philip Heidelberger, *Member, IEEE*, Victor F. Nicola, *Member, IEEE*, and Peter W. Glynn

Abstract—**In this paper we present a unified framework for simulating Markovian models of highly dependable systems. Since the failure event is a rare event, the estimation of system dependability measures using standard simulation requires very long simulation runs. We show that a variance reduction technique called Importance Sampling can be used to speed up the simulation by many orders of magnitude over standard simulation. This technique can be combined very effectively with regenerative simulation to estimate measures such as steady-state availability and mean time to failure. Moreover, it can be combined with conditional Monte Carlo methods to quickly estimate transient measures such as reliability, expected interval availability, and the distribution of interval availability. We show the effectiveness of these methods by using them to simulate large dependability models. We also discuss how these methods can be implemented in a software package to compute both transient and steady-state measures simultaneously from the same sample run.**

Index Terms— **Dependability measures, highly available systems, importance sampling, Markovian models, rare event simulation, variance reduction.**

I. Introduction

The requirements for highly dependable systems, such as air traffic control and transaction processing systems, increase the importance of dependability prediction at a design stage. Typically, stochastic models are used to analyze such systems. A system is considered to be a collection of components which can fail and possibly get repaired. The system is considered operational if at any given moment the operational components satisfy some minimum system operational requirements. Many details of failure and repair behavior of the components have been introduced in such models: common-mode failures in [2], [5], and [31], detailed fault and error handling models in [9] and detailed recovery hierarchies, operational and repair dependences in [16] and [18]. Models which include degraded modes of operation have also been introduced ([18], [42]). Different measures are used to evaluate the modeled systems depending upon whether they are mission oriented systems or continuously operating systems. Some of the dependability measures of interest are steady-state availability, reliability, mean time to failure, expected interval availability, and the complementary distribution of interval availability (i.e., the probability that a system would achieve a higher interval availability than a specified value between 0 and 1). Similar measures have also been constructed for degradable systems, e.g., steady-state performance and distribution of performance over a time interval ([33]). Detailed surveys of these modeling techniques and the dependability measures calculated appear in [12] and [32].

The most common stochastic models used in this context are continuous-time Markov chains (CTMC's). Typically, numerical methods are used to solve Markov chains. Although, many modeling packages have been built, e.g., [18] and [9], which incorporate numerical methods capable of computing steady-state as well as transient state probabilities of Markov chains with thousands of states, the size of the system modeled is typically small because the number of states in the system increases exponentially with the number of components. Techniques like state lumping and unlumping ([17], [36]) and state aggregation and bounding ([1], [35]) can reduce the size of the state space substantially. However, large systems with a large number of redundant components are still out of the range of the solution capabilities of current numerical methods, primarily due to storage or computational limitations.

An alternate approach for the solution of large models is Monte Carlo simulation, which is the subject of this paper. Simulation is especially useful for those models for which the transition rate matrices exceed the available storage. By nature, this approach has the immediate advantage of having relatively small storage requirements. On the other hand, since the failure events are rare events, it is apparent that the analysis by simulation of large models with a high degree of redundancy will require many regenerative cycles or many long independent replications in order to attain reasonable confidence intervals ([12], [30]). Our goal is to obtain variance reduction methods that are applicable to a broad class of models. Specifically, we are interested in models defined by the reliability and availability modeling language described in [18], so that the techniques can be implemented in a software package and made available to designers in an automatic and transparent fashion. A typical system contains multiple component types with redundant units for each component type. Failure of these systems is usually caused by exhaustion of redundancy or by a combination of component failures leading to a system failure. Failed components may be repaired. If all components are repairable and component failures

Manuscript received July 29, 1989; revised February 29, 1990.

A. Goyal, P. Heidelberger, and V. F. Nicola are with IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598.

P. Shahabuddin and P. W. Glynn are with the Department of Operations Research, Stanford University, Stanford, CA 94305. This work was performed while P. Shahabuddin was visiting IBM Research and in part when P. W. Glynn was a visiting scientist at IBM Research.

IEEE Log Number 9103031.

0018-9340/92$03.00 © 1992 IEEE

are exponential, then a regenerative state for the system (see, e.g., [8]) is the state where all units of all component types are operational. If, in addition, all repair times are exponentially distributed, then the system can be modeled by a continuous time Markov chain. For highly reliable and highly available systems, it is usual for the repair/recovery rates of components to be orders of magnitude larger than the failure rates, and in these circumstances the use of importance sampling variance reduction techniques [21], [23] can be very effective in reducing the simulation run length significantly.

Importance sampling for rare event simulation has been used successfully in [6], [29], [45], [47], and [40]. Proper selection of the importance sampling distribution makes the rare events more likely to occur; this results in a variance reduction. The key, of course, is to choose a good importance sampling distribution. The theory of large deviations was used in [6], [45], and [47], [40] to select an effective distribution for problems arising in Markov chains with "small increments," random walks, and queueing networks, respectively. Effective heuristics were used in [29] to select importance sampling distributions for reliability estimation in large models of machine repairman type, and in [13] for acyclic reliability models.

In this paper, we review and extend the methods in [4], [19], and [43] and present a comprehensive and unified framework for simulating a broad class of models, specifically models defined by the reliability and availability modeling language described in [18]. The language is used to describe failure and repair behavior of components as well as operational, repair, recovery and common-mode failure dependencies among them. The language is also used to describe conditions (e.g., reliability block diagram or fault-tree) under which the system is considered operational. The model described by the language is simulated assuming that all failure, repair, and recovery distributions are exponential. We estimate both steady-state and transient measures simultaneously from the simulation. Importance sampling techniques are used to estimate these measures; these techniques are orders of magnitude faster than ordinary simulation. The basic idea behind importance sampling is described in Section II. We also give formal definitions of the dependability measures of interest in this section.

In Section III we present our methods for estimating dependability measures, such as the steady-state availability and the mean time to failure (MTTF). The estimators are based on combining regenerative simulation ([8]) with importance sampling. The concepts of dynamic importance sampling (DIS, see [4]) and measure specific dynamic importance sampling (MSDIS, see [19]) are explained using a very simple three state example. Direct application of these techniques does not yield a significant variance reduction for the MTTF. However, when the MTTF is formulated as a ratio of two expectations (both are estimated using regenerative simulation), then significant variance reductions can be achieved using our importance sampling techniques (see also [43]). Therefore, while the MTTF is, in fact, a transient measure, we can estimate it using a regenerative simulation; this is the reason why we have considered its estimation with other steady-state measures in Section III. The equations for optimal run-length allocation and asymptotic bias expansions are also given.

In Section IV we present our methods for estimating the transient measures, such as reliability, interval availability, and distribution of interval availability. The estimators are obtained by independently replicating observations based on combining "conditioning" (e.g., [10], [11]) or "forcing" (e.g., [29]) methods with importance sampling. In Section V we show how we implemented both regenerative simulation and the independent replications so that steady-state and transient measures can be computed simultaneously from the same sample run. Some implementation issues as well as theoretical issues in using importance sampling are also described in this section.

In Section VI we show the effectiveness of the above techniques in a large example. Typically, we obtain orders of magnitude reduction in variance over standard simulation, which translates into large reductions in simulation run times. We also perform coverage experiments on the confidence intervals and compute the bias values for the estimates of the steady-state availability. Finally, in Section VII we give concluding remarks and suggest some directions for future research.

II. Background and Notation

In this section, we review the basic ideas of importance sampling. We include this background material to make the paper self contained and more accessible to the nonspecialist. A continuous-time Markov chain (CTMC) model of systems is then introduced and the associated measures that are of primary interest in evaluating highly available systems are defined.

A. Review of Importance Sampling

The basic notion behind importance sampling can be illustrated using a simple example: estimating the expected value of a function of a random variable (see, e.g., [21]). Suppose that θ is an input parameter to the simulation., e.g., a failure rate. Associated with each θ is a probability density function (pdf) $p_\theta(x)$ for $-\infty < x < \infty$. Suppose we wish to estimate $r(\theta) = E_\theta[h(X)]$ for some function h where the subscript θ indicates that the random variable (rv) X is sampled from the pdf $p_\theta(x)$. Then

$$r(\theta) = E_\theta[h(X)] = \int_{-\infty}^{\infty} h(x)p_\theta(x)dx. \tag{2.1}$$

Now assuming that $p'_{\theta_0}(x)$ is another probability density function such that $p'_{\theta_0}(x) > 0$ for all x. Equation (2.1) can be written as

$$\begin{aligned} r(\theta) &= \int_{-\infty}^{\infty} h(x)\left(\frac{p_\theta(x)}{p'_{\theta_0}(x)}\right)p'_{\theta_0}(x)dx \\ &= E_{\theta_0}[h(X)L(\theta,\theta_0,X)] \end{aligned} \tag{2.2}$$

where $L(\theta,\theta_0,x) \equiv p_\theta(x)/p'_{\theta_0}(x)$ is called the likelihood ratio. Equation (2.2) thus provides a means to produce an unbiased estimate of $r(\theta)$ by simulation using the different probability density function $p'_{\theta_0}(x)$. This switch of the probability density function is called a change of measure; the resulting simulation algorithm is called importance sampling.

For our purposes, the goal of this change of measure is to produce an estimate with lower variance. In fact, if $h(x) > 0$ for all x, then choosing $p'_{\theta_0}(x) = h(x)p_\theta(x)/r(\theta)$ yields a zero variance estimator, since in this case the r.v. $h(X)L(\theta, \theta_0, X)$ takes on the constant value $r(\theta)$ with probability one. In practice, however, this is not a feasible change of measure since it requires knowing $r(\theta)$, the unknown measure to be estimated. Nevertheless, we will find the zero variance transformation useful, since it can be calculated for some simple examples and forms the basis of a heuristic for simulating more complex examples of highly available systems.

Walrand [47] shows why importance sampling can be particularly effective for estimating the probability of rare events. Basically, good variance reduction is achieved by making the likelihood ratio very small on the rare set. This corresponds to choosing the parameter θ_0 so that $p'_{\theta_0}(x)$ is relatively large for x in the rare set, i.e., by making the rare set more likely to occur under the new measure defined by $p'_{\theta_0}(x)$.

B. Markov Chains and Associated Dependability Measures

We assume that the system can be represented by a CTMC $\boldsymbol{Y} = \{Y_s, s \geq 0\}$ with finite state space E and infinitesimal generator matrix $\boldsymbol{Q} = \{q(i,j), i,j \in E\}$. We let $q(i) = -q(i,i)$ denote the total rate out of state i (see, e.g., [24]). We further assume that E can be partitioned into two subsets: $E = O \cup F (O \cap F = \phi)$ where O is the set of up states, i.e., the set of states for which the system is operational, and F is the set of down, or failed states. We assume that the system starts out in the state for which all components are operational; we label this state as state 0. For any set of states A, let α_A denote the time the Markov chain first enters the set A. Of particular interest are α_0, which is the first return time to state 0, and α_F, which is the first entrance time into the subset F of failed sates.

We will be interested in two types of dependability measures associated with the CTMC $\boldsymbol{Y}$: transient measures and so-called steady-state measures. Considering the transient measures first, the interval availability $A(t)$ is defined by

$$A(t) = \frac{1}{t}\int_{s=0}^{t} 1_{\{Y_s \in O\}} ds \tag{2.3}$$

where $1_{\{Y_s \in O\}}$ is the indicator of the event $\{Y_s \in O\}$, i.e., $1_{\{Y_s \in O\}} = 1$ if $Y_s \in O$ and $1_{\{Y_s \in O\}} = 0$ if $Y_s \notin O$. This is the fraction of time that the system is operational in the time interval $(0, t)$. We let

$$I(t) = E[A(t)] \tag{2.4}$$

be the expected interval availability and let

$$A(t,x) = \mathrm{P}\{A(t) \leq x\} \tag{2.5}$$

denote the distribution of availability. The reliability of the system is defined to be the probability that the system does not fail in the interval $(0, t)$:

$$R(t) = \mathrm{P}[\alpha_F > t] = E[1_{\{\alpha_F > t\}}]. \tag{2.6}$$

For steady-state measures we assume that $\boldsymbol{Y}$ is irreducible, in which case $Y_s \Rightarrow Y$ as $s \to \infty$, where $\Rightarrow$ denotes convergence in distribution and Y is a r.v. having the steady-state distribution $\underline{\pi} = \{\pi_i, i \in E\}$ ($\underline{\pi}$ solves the equations $\underline{\pi}\boldsymbol{Q} = \boldsymbol{0}$). Notice that steady-state measures are independent of the starting state of the system; however, we will choose the fully operational state (i.e., state 0) to define a regenerative state for the system. By regenerative process theory (see [46] or [8]), steady-state measures take the form of a ratio of two expected values:

$$\begin{aligned} r \equiv [f(Y)] &= \lim_{t\to\infty} E[f(Y_t)] \\ &= \lim_{t\to\infty} \frac{1}{t}\int_{s=0}^{t} f(Y_s)ds = \frac{E[\int_{s=0}^{\alpha_0} f(Y_s)ds]}{E[\alpha_0]} \end{aligned} \tag{2.7}$$

where f is a real valued function on E. If $f(i) = 1_{\{i \in O\}}$, then $E[f(Y)]$ is the long run fraction of time the system is operational and is called the steady-state availability, which we denote by $A = \lim_{t\to\infty} E[A(t)]$. We will sometimes find it convenient to consider the expected unavailability $U(t) = 1 - I(t) = 1 - E[A(t)]$ and the steady-state unavailability, $U = 1 - A$. The problem of steady-state estimation thus reduces to one of estimating the ratio of two expected values.

The mean time to failure (MTTF), $E[\alpha_F]$, is typically thought of as a transient measure, since it depends on the starting state of the system (state 0) which is assumed to be the fully operational state. The same measure is sometimes referred to as the mean time to first failure (MTFF). A ratio representation for $E[\alpha_F]$ is found to be particularly useful. To derive this ratio, we write

$$\begin{aligned} \alpha_F &= \alpha_F 1_{\{\alpha_F < \alpha_0\}} + (\alpha_0 + (\alpha_F - \alpha_0))1_{\{\alpha_0 < \alpha_F\}} \\ &= \min(\alpha_F, \alpha_0) + (\alpha_F - \alpha_0)1_{\{\alpha_0 < \alpha_F\}}. \end{aligned} \tag{2.8}$$

Now, applying the Markov property at time α_0 shows that, on the set $\{\alpha_0 < \alpha_F\}, (\alpha_F - \alpha_0)$ is conditionally independent of $1_{\{\alpha_0 < \alpha_F\}}$ and furthermore has the same distribution as α_F. Therefore, taking expected values of (2.8) and rearranging terms yields the ratio formula

$$E[\alpha_F] = \frac{E[\min(\alpha_F, \alpha_0)]}{\mathrm{P}\{\alpha_F < \alpha_0\}}. \tag{2.9}$$

Thus, we can view estimating $E[\alpha_F]$ as a ratio estimation problem, where both the numerator and the denominator are estimated using a regenerative simulation. Therefore, in Section III we consider the estimation of the mean time to failure (MTTF) together with steady-state measures which are also (and more commonly) estimated using regenerative simulations.

III. Estimating Steady-State Measures

In this section, we discuss the estimation of steady-state measures of CTMC's. We begin by reviewing how this problem can be reduced to estimating associated steady-state measures in discrete-time Markov chains (DTMC's) and describe the regenerative method of simulation (see, e.g., [8]). We next describe the application of importance sampling to DTMC's. In particular, we note that the importance sampling transformation selected for actual simulation can be dynamic in the sense that it need not correspond directly to a time homogeneous DTMC. We also note that, since the problem is one of estimating the ratio of two expected values, there is no need to use the same importance sampling transformation for estimating both the numerator and denominator of this ratio, i.e., the importance sampling transformations can be measure specific. Analysis of a three state example emphasizes the benefit of both dynamic and measure specific importance sampling and serves as the basis for heuristics for larger, more complex system availability models. The optimal allocation of CPU time to estimation of the numerator and denominator is then discussed. The section concludes by considering asymptotic bias expansions of the estimators.

A. Discrete Time Conversion of CTMC's

In this it is shown how one can estimate steady-state measures of an irreducible CTMC by simulating only the embedded DTMC (and not generating random holding times). Let $\boldsymbol{X} = \{X_n,\ n \geq 0\}$ denote the embedded DTMC of the CTMC Y: $\boldsymbol{X}$ has transition matrix $\boldsymbol{P} = \{p(i,j)\}$ where $p(i,i) = 0$ and $p(i,j) = q(i,j)/q(i)$ for $j \neq i$. Let $h(i) = 1/q(i)$ be the mean holding time in state i and let $g(i) = f(i)/q(i)$. Let τ_A be the first entrance time of the DTMC into the set A and let τ_0 be the first return time of the DTMC to state 0. Then

$$E[f(Y)] = \lim_{t\to\infty} E[f(Y_t)] = \frac{E[\int_{s=0}^{\alpha_0} f(Y_s)ds]}{E[\alpha_0]} = \frac{E[\sum_{k=0}^{\tau_0 - 1} g(X_k)]}{E[\sum_{k=0}^{\tau_0 - 1} h(X_k)]}. \tag{3.1}$$

To emphasize the dependence of this ratio on the transition matrix $\boldsymbol{P}$, we write (3.1) explicitly as $r = \mathrm{E}_{\boldsymbol{P}}[G]/\mathrm{E}_{\boldsymbol{P}}[H]$ where $G = \sum_{k=0}^{\tau_0-1} g(X_k)$ and $H = \sum_{k=0}^{\tau_0-1} h(X_k)$. In this, it is shown that this discrete time conversion is always guaranteed to produce a variance reduction over simulation of the original CTMC. Fox and Glynn [10] have extended this result to simulation of semi-Markov processes.

Equation (3.1) forms the basis for the regenerative method of simulation for CTMC's (see, e.g., [8]). One simulates (using the transition matrix $\boldsymbol{P}$) m (say) independent and identically distributed (iid) replicates of the random vector (G, H), yielding the iid random vectors $\{(G_j, H_j) : j = 1, \cdots, m\}$. Each replication involves simulating the DTMC $\boldsymbol{X}$ (with the initial condition $X_0 = 0$) to time τ_0; these replications are known as regenerative cycles. Let $\hat{r}_m(\boldsymbol{P}) = \sum_{j=1}^m G_j / \sum_{j=1}^m H_j$. Then, because the cycles are iid, $\lim_{m\to\infty} \hat{r}_m(\boldsymbol{P}) = r$ with probability one and $\sqrt{m}(\hat{r}_m(\boldsymbol{P}) - r) \Rightarrow N(0, \sigma^2(\boldsymbol{P})/\mathrm{E}_{\boldsymbol{P}}[H_j]^2)$, where $N(0,\sigma^2)$ denotes a normally distributed random variable with mean zero and variance σ^2, and $\sigma^2(\boldsymbol{P}) = \mathrm{Var}_{\boldsymbol{P}}[G_j - rH_j]$.

B. Importance Sampling for DTMC's

We next extend the change of measure transformation of (2.1) to DTMC's. Let τ be any stopping time of the DTMC $\boldsymbol{X}$ and let Y be a r.v. defined on $\boldsymbol{X}$ up until time τ. Informally, τ is a stopping time if the event $\{\tau = n\}$ is determined by $\boldsymbol{X}_n \equiv (X_0, \cdots, X_n)$. The r.v. Y is then a (measurable) function of $\boldsymbol{X}_\tau = (X_0, \cdots, X_\tau)$ (see [24] for a more detailed and precise treatment of stopping times). The first entrance time to a state, or a set of states, is a stopping time. In particular, both τ_0 and τ_F are stopping times. Let A_n denote the set of all possible sample paths up until time n, i.e., $A_n = \{\boldsymbol{s}_n = (s_0, \cdots, s_n) : s_j \in E\}$.

For any $\boldsymbol{s}_n \in A_n$, let

$$\boldsymbol{P}(\boldsymbol{s}_n) \equiv p(s_0)p(s_0, s_1)p(s_1, s_2)\cdots p(s_{n-1}, s_n) \tag{3.2}$$

where $p(s_0)$ is the probability that the initial state is s_0. Let $B_n \subset A_n$ be the set of sample paths for which $\tau = n$.

Proposition 3.1: Let τ be a stopping time which, under the transition matrix $\boldsymbol{P}$, is finite with probability one and let Z be a (measurable) function of $\boldsymbol{X}_\tau$ for which $\mathrm{E}_{\boldsymbol{P}}[|Z(\boldsymbol{X}_\tau)|] < \infty$. Let $\boldsymbol{P}'$ be any other measure such that, under $\boldsymbol{P}'$, τ is finite with probability one and for any $\boldsymbol{s}_n \in B_n$, $\boldsymbol{P}'(\boldsymbol{s}_n) \neq 0$ whenever $Z(\boldsymbol{s}_n)\boldsymbol{P}(\boldsymbol{s}_n) \neq 0$. Then

$$\mathrm{E}_{\boldsymbol{P}}[Z(\boldsymbol{X}_\tau)] = \mathrm{E}_{\boldsymbol{P}'}[Z(\boldsymbol{X}_\tau)L_1'(\boldsymbol{X}_\tau)] \tag{3.3}$$

where for any n, $L_1'(\boldsymbol{X}_n) = \boldsymbol{P}(\boldsymbol{X}_n)/\boldsymbol{P}'(\boldsymbol{X}_n)$.

Proof: Since, under $\boldsymbol{P}$, τ is finite with probability one and since Z has a finite absolute first moment, we can write

$$\begin{aligned}\mathrm{E}_{\boldsymbol{P}}[Z(\boldsymbol{X}_\tau)] &= \sum_{n=0}^{\infty} \sum_{\boldsymbol{s}_n \in B_n} \boldsymbol{P}(\boldsymbol{s}_n)Z(\boldsymbol{s}_n) \\ &= \sum_{n=0}^{\infty} \sum_{\boldsymbol{s}_n \in B_n} \boldsymbol{P}'(\boldsymbol{s}_n)Z(\boldsymbol{s}_n)L_1'(\boldsymbol{s}_n) \\ &= \mathrm{E}_{\boldsymbol{P}'}[Z(\boldsymbol{X}_\tau)L_1'(\boldsymbol{X}_\tau)]\end{aligned} \tag{3.4}$$

where the last equality follows since τ is finite with probability one under $\boldsymbol{P}'$. □

Versions of this proposition have appeared elsewhere, e.g., in [47], [40], or [15]. Note that there can exist a sample path $\boldsymbol{s}_n$ such that $\boldsymbol{P}(\boldsymbol{s}_n) > 0$ even though $\boldsymbol{P}'(\boldsymbol{s}_n) = 0$, provided that $Z(\boldsymbol{s}_n) = 0$. We emphasize, however, that the measure $\boldsymbol{P}'$ does *not* have to correspond to a time homogeneous Markov chain, nor even that it corresponds to a Markov chain. Indeed we will see that it is highly advantageous in many circumstances for $\boldsymbol{P}'$ not to be Markovian. The general form that we will consider for $\boldsymbol{P}'$ is

$$\begin{aligned}\boldsymbol{P}'(\boldsymbol{s}_n) = {}&\boldsymbol{P}'(s_0)\boldsymbol{P}'(s_1|s_0) \\ &\cdot \boldsymbol{P}'(s_2|s_0, s_1)\cdots \boldsymbol{P}'(s_n|s_0, \cdots, s_{n-1}).\end{aligned} \tag{3.5}$$

With this formulation, we have the freedom to, e.g., adjust the transition probabilities to depend upon the number of visits the chain has made to a set of states (say the failed states) or simply to "turn off" the importance sampling whenever the likelihood ratio gets too small, thereby avoiding numerical problems. We term the use of such an importance sampling distribution Dynamic Importance Sampling (DIS).

Applying DIS to estimating the ratio of (3.1) yields the following procedure. A total of m iid regenerative cycles of the DTMC $\boldsymbol{X}$ are simulated using the DIS distribution $\boldsymbol{P}'$. Let G_j, H_j, and L'_{1j} be the samples of G, H, and L'_1, respectively, from cycle j. Define the point estimate $\hat{r}_m(\boldsymbol{P}') = \sum_{j=1}^m G_j L'_{1j} / \sum_{j=1}^m H_j L'_{1j}$. Then, as in the case without importance sampling, we have $\lim_{m\to\infty} \hat{r}_m(\boldsymbol{P}') = r$ with probability one and $\sqrt{m}(\hat{r}_m(\boldsymbol{P}') - r) \Rightarrow N(0, \sigma^2(\boldsymbol{P}')/\mathrm{E}_{\boldsymbol{P}}[H_j]^2)$ where

$$\begin{aligned}\sigma^2(\boldsymbol{P}') &= \mathrm{Var}_{\boldsymbol{P}'}[(G_j - rH_j)L'_{1j}] \\ &= \mathrm{Var}_{\boldsymbol{P}'}[G_j L'_{1j}] - 2r\mathrm{Cov}_{\boldsymbol{P}'}[G_j L'_{1j}, H_j L'_{1j}] \\ &\quad + r^2 \mathrm{Var}_{\boldsymbol{P}'}[H_j L'_{1j}]. \end{aligned} \tag{3.6}$$

From the form of $\sigma^2(\boldsymbol{P}')$, it is seen that selecting a good DIS distribution $\boldsymbol{P}'$ involves taking three terms into consideration. For example, selecting a $\boldsymbol{P}'$ to reduce the variance of the estimate of the ratio's numerator may actually increase the variance of the estimate of the denominator, or vice versa. In addition, the effect on the covariance term will generally be difficult to control, or even predict. Thus, selection of a single importance sampling distribution for both the numerator and denominator involves a tradeoff.

This suggests that, since we are really trying to estimate two different quantities, we should use different changes of measures to estimate each quantity. Estimating the numerator and denominator independently allows one to tailor the importance sampling distributions to the particular measure being estimated, without having to be concerned about the covariance term. We call this Measure Specific Dynamic Importance Sampling (MSDIS). Section III-C provides further motivation for the use of DIS, as well as MSDIS. In fact for the example given, the two optimal, i.e., zero variance, changes of measure are opposites in the sense that numerator's optimal change of measure brings the system very quickly to the failed state, whereas the denominator's optimal change of measure is approximately the same as the original measure and thus brings the system only very slowly to the failed state.

The procedure for MSDIS can be described more completely as follows. Let $\boldsymbol{P}'$ and $\boldsymbol{P}''$ denote the DIS distributions for the numerator and denominator, respectively. A total of m cycles are simulated. Assume that βm cycles of the numerator are simulated and $(1-\beta)m$ cycles of the denominator are independently simulated where $0 < \beta < 1$ (for notational simplicity, assume that βm is an integer). Define

$$\hat{r}_m(\boldsymbol{P}', \boldsymbol{P}'') = \frac{\sum_{j=1}^{\beta m} G_j L'_{1j}/(\beta m)}{\sum_{j=1}^{(1-\beta)m} H_j L''_{1j}/([1-\beta]m)}. \tag{3.7}$$

Note that $G_j L'_{1j}$ is actually independent of $H_j L''_{1j}$ even though they have the same subscript. Then, as before, $\lim_{m\to\infty} \hat{r}_m(\boldsymbol{P}', \boldsymbol{P}'') = r$ with probability one and $\sqrt{m}(\hat{r}_m(\boldsymbol{P}', \boldsymbol{P}'') - r) \Rightarrow N(0, \sigma^2(\boldsymbol{P}', \boldsymbol{P}'')/\mathrm{E}_{\boldsymbol{P}}[H_j]^2)$ where

$$\sigma^2(\boldsymbol{P}', \boldsymbol{P}'') = \frac{\mathrm{Var}_{\boldsymbol{P}'}[G_j L'_{1j}]}{\beta} + r^2 \frac{\mathrm{Var}_{\boldsymbol{P}''}[H_j L''_{1j}]}{(1-\beta)}. \tag{3.8}$$

The optimal run length allocation between the numerator and the denominator will be considered in Section III-D.

C. A Three State Example

In this section, we consider a simple availability example, namely a three state birth and death process (see [24]). Because of its simple structure, the optimal zero variance importance sampling distributions can be derived in closed form. The optimal changes of measure for the numerator and denominator are quite different. These results would be of no significance except that the three state example serves as a paradigm for more complex models and thus strongly suggests a basic form for effective importance sampling distributions in more complex availability models.

The state space is $E = \{0, 1, 2\}$, the birth rates are λ_i, $i = 0, 1$ and the death rates are μ_i, $i = 1, 2$. In the reliability context, this models a system with two identical components which can fail and be repaired. We assume that births correspond to failures and deaths correspond to repairs so that state i corresponds to having i failed components. We consider the system to be operational in states 0 and 1, but failed in state 2.

The embedded DTMC has the following nonzero entries: $p(0,1) = p(2,1) = 1$, $p(1,2) = \epsilon \equiv \lambda_1/(\lambda_1 + \mu_1)$ and $p(1,0) = (1-\epsilon)$. Letting h_i denote the mean holding time in state i, then $h_0 = 1/\lambda_0$, $h_1 = 1/(\lambda_1 + \mu_1)$ and $h_2 = 1/\mu_2$. We assume that failure rates are much less than repair rates, specifically we assume that $h_0 = \Theta(1/\epsilon), h_1 = \Theta(1)$ and $h_2 = \Theta(1)$ (we follow Knuth's [25] usage of $f(x) = \Theta(g(x))$ if there exist constants C_1 and C_2 such that $\forall x,\ 0 < C_1 g(x) < f(x) < C_2 g(x)$).

The steady-state measure r of interest is the stationary probability of being in state 2, the steady-state unavailability. This can be estimated using regenerative simulation with function values $g(0)$, $g(1)$, and $g(2)$ equal to 0, 0, and h_2, respectively, and function values $h(0)$, $h(1)$, and $h(2)$ equal to h_0, h_1, and h_2, respectively. Assume state 0 is the regenerative state. We first compute the variance of the estimator using standard regenerative simulation. Let n_F be the number of visits to the failed state, state 2, during a regenerative cycle and let $\boldsymbol{s}_i$ denote the (unique) sample path of a regeneration cycle of the DTMC for which $n_F = i$. Then $G = n_F h_2$ and $H = h_0 + h_1 + n_F(h_1 + h_2)$. Furthermore, n_F has a geometric distribution, $\mathrm{P}\{n_F = i\} = (1-\epsilon)\epsilon^i$ for $i \geq 0$, so that $\mathrm{E}_{\boldsymbol{P}}[n_F] = \epsilon/(1-\epsilon)$ and $\mathrm{Var}_{\boldsymbol{P}}[n_F] = \epsilon/(1-\epsilon)^2$. Thus, $\mathrm{E}_{\boldsymbol{P}}[G] = h_2\epsilon/(1-\epsilon)$ and $\mathrm{E}_{\boldsymbol{P}}[H] = (h_0 + h_1) + (h_1 + h_2)\epsilon/(1-\epsilon)$. Straightforward calculations show that $r = \Theta(\epsilon^2)$ and that the asymptotic squared coefficient of variation of $\hat{r}_m(\boldsymbol{P})$ (obtained from the central limit theorem) is

$$\frac{\mathrm{Var}_{\boldsymbol{P}}[G - rH]}{mr^2 \mathrm{E}_{\boldsymbol{P}}[H]^2} = \Theta(\frac{1}{m\epsilon}). \tag{3.9}$$

The dominant term in (3.9) is due to contribution of the numerator. Thus, to obtain a confidence interval with a relative width (width divided by the point estimate) that goes to zero requires that the sample size m be large enough so that $m\epsilon \to \infty$. This demonstrates the potentially large sample size required for rare event simulations (in which $\epsilon \approx 0$).

Let $\boldsymbol{P}(\boldsymbol{s}_i)$ be the probability of a regenerative cycle sample path $\boldsymbol{s}_i$, then $\boldsymbol{P}(\boldsymbol{s}_i) = (1-\epsilon)\epsilon^i$, $i \geq 0$. The optimal zero variance importance sampling distribution $\boldsymbol{P}'(\boldsymbol{s}_i)$, $i \geq 0$, for estimating $\mathrm{E}_{\boldsymbol{P}}[G]$ is computed from explicit enumeration of all sample paths. First, we write $\mathrm{E}_{\boldsymbol{P}}[G] = \sum_{\forall i} G(\boldsymbol{s}_i)\boldsymbol{P}(\boldsymbol{s}_i) = \sum_{\forall i}[G(\boldsymbol{s}_i)L(\boldsymbol{s}_i)]\boldsymbol{P}'(\boldsymbol{s}_i) = \mathrm{E}_{\boldsymbol{P}'}[GL]$, with $L(\boldsymbol{s}_i) = \boldsymbol{P}(\boldsymbol{s}_i)/\boldsymbol{P}'(\boldsymbol{s}_i)$. Now, the optimal $\boldsymbol{P}'(\boldsymbol{s}_i)$, for all $i \geq 0$, can be computed (similar to what is described in Sectio II-A):

$$\boldsymbol{P}'^{*}(\boldsymbol{s}_i) = \frac{\boldsymbol{P}(\boldsymbol{s}_i)\mathrm{G}(\boldsymbol{s}_i)}{\mathrm{E}_{\boldsymbol{P}}[G]} = (1-\epsilon)^2\epsilon^{i-1}i, \qquad i \geq 0, \quad (3.10)$$

since then $G(\boldsymbol{s}_i)L(\boldsymbol{s}_i)$, for all $i \geq 0$, is a constant equal to $\mathrm{E}_{\boldsymbol{P}}[G]$.

Similarly, the optimal zero variance importance sampling distribution for estimating $\mathrm{E}_{\boldsymbol{P}}[H]$ is given by

$$\begin{aligned}\boldsymbol{P}''^{*}(\boldsymbol{s}_i) &= \frac{\boldsymbol{P}(\boldsymbol{s}_i)\mathrm{H}(\boldsymbol{s}_i)}{\mathrm{E}_{\boldsymbol{P}}[H]}\\ &= \frac{[(h_0+h_1)+(h_1+h_2)i](1-\epsilon)^2\epsilon^i}{h_0+h_1-\epsilon(h_0-h_2)},\\ &\qquad\qquad i \geq 0. \quad (3.11)\end{aligned}$$

From (3.10), $\boldsymbol{P}'^{*}(\boldsymbol{s}_0) = 0$, $\boldsymbol{P}'^{*}(\boldsymbol{s}_1) = (1-\epsilon)^2$, $\boldsymbol{P}'^{*}(\boldsymbol{s}_2) = 2\epsilon(1-\epsilon)^2$ and so on. Now let $p'^{*}(1,0 \mid n_F = i)$ denote the probability of going from state 1 to state 0 given that the chain is in state 1 and that the failed state has already been visited i times. Then $p'^{*}(1,0 \mid n_F = i) = \boldsymbol{P}'^{*}(\boldsymbol{s}_i)/(\sum_{j\geq i}\boldsymbol{P}'^{*}(\boldsymbol{s}_j))$ and thus, from (3.10), $p'^{*}(1,0 \mid n_F = 0) = 0$ and

$$p'^{*}(1,0 \mid n_F = i) = \frac{(1-\epsilon)^2}{1-\epsilon(i-1)/i}, \quad i \geq 1. \qquad (3.12)$$

Therefore, each successive time the simulation enters state 1, the probability of returning to state 0 changes (under both $\boldsymbol{P}'^{*}(\boldsymbol{s})$ and $\boldsymbol{P}''^{*}(\boldsymbol{s})$). Thus, the optimal changes of measure for both the numerator and the denominator of (2.1) are dynamic. In particular note that while $p'^{*}(1,0 \mid n_F = 0) = 0$, $p'^{*}(1,0 \mid n_F = 1) = (1-\epsilon)^2 \approx (1-2\epsilon) \approx (1-\epsilon) = p(1,0)$ for $\epsilon \approx 0$. Also, $\lim_{i\to\infty} p'^{*}(1,0 \mid n_F = i) = (1-\epsilon) = p(1,0)$. This suggests that, for more complex models, the importance sampling distribution for the numerator should be chosen to move the system very quickly to the set of failed states F, but that once F is entered, the importance sampling should be turned off so that the system quickly returns to state 0. This should hold true for systems in which the probability of two or more failures in a regenerative cycle is at least an order of magnitude less likely than the probability of one failure in a cycle. This is also consistent with the argument given in Section II-A as well as Walrand's suggestion in [47] and [40] (which was derived using large deviation results) to interchange λ and μ for estimating the probability of buffer overflow in the M/M/1 queue.

For the denominator, on the other hand, the largest contribution to the expected value comes from the sample path on which $n_F = 0$, which is not a rare event. This suggests using standard simulation., i.e., not using importance sampling, to estimate the denominator. Indeed, the optimal change of measure of (3.11) has

$$\begin{aligned}p''^{*}(\boldsymbol{s}_0) &= p''^{*}(1,0 \mid n_F = 0)\\ &= \frac{(1-\epsilon)^2}{1-\epsilon(h_0-h_2)/(h_0+h_1)}\\ &\approx (1-\epsilon) = p(1,0) \qquad (3.13)\end{aligned}$$

so that there is very little difference between $\boldsymbol{P}''^{*}(\boldsymbol{s})$ and $\boldsymbol{P}(\boldsymbol{s})$ for the most likely sample path.

D. Optimal Run Length Allocation

Equation (3.8) gave the form of the asymptotic variance when a fixed number m of cycles are simulated of which βm are devoted to simulation of the numerator and $(1-\beta)m$ are allocated to the denominator. Since the expected amount of CPU time to simulate a sample of the numerator and denominator may be different, a more practical run length allocation model can be formulated as follows. Let the total CPU time be T and assume that βT is allocated to the numerator and $(1-\beta)T$ is allocated to the denominator. Let c_n (c_d) denote the expected CPU time to simulate a sample of the numerator (denominator). Then, for large T, approximately $\beta T/c_n$ cycles of the numerator and $(1-\beta)T/c_d$ cycles of the denominator are obtained. The asymptotic variance of the resulting point estimate is

$$\frac{1}{T\mathrm{E}_{\boldsymbol{P}}[H_j]^2}\left(\frac{\sigma_n^2 c_n}{\beta} + r^2\frac{\sigma_d^2 c_d}{1-\beta}\right) \qquad (3.14)$$

where $\sigma_n^2 = \mathrm{Var}_{\boldsymbol{P}'}[G_j L'_{1j}]$ and $\sigma_d^2 = \mathrm{Var}_{\boldsymbol{P}''}[H_j L''_{1j}]$. This result is obtained by applying results from renewal theory (see [46]). Minimization of (3.14) with respect to β yields $\beta^* = \delta/(1+\delta)$ where $\delta = \sigma_n\sqrt{c_n}/(r\sigma_d\sqrt{c_d})$.

Suppose that $c_n \approx c_d$ and that $\sigma_n \approx \sigma_d$ (we are equally effective in reducing the variance of the numerator and denominator). Then, for estimating the steady-state unavailability, r is small and $\beta^* \approx 1$, i.e., the bulk of the effort should be applied to estimating the numerator, which in this case is a rare event simulation using importance sampling. On the other hand, for estimating the MTTF using the ratio formula given in (2.9), r is large and $\beta^* \approx 0$, i.e., the bulk of the effort is devoted to the denominator. However, for the MTTF, the denominator also corresponds to a rare event simulation using importance sampling (moreover, as will be discussed in Section V, the same importance sampling distribution can be used to estimate both measures). Thus, in either case, the optimal allocations are consistent in the sense that they allocate most of the effort to rare event simulation.

In practice, we always devote a minimum percentage, say 10%, of the effort to standard simulation even though the optimal allocation usually suggests devoting much less time to standard simulation. This permits stable variance estimation and the loss in asymptotic efficiency from the optimal allocation is small.

E. Bias Expansions

We now consider bias expansions of ratio estimators of steady-state measures. Because the numerator and denominator are simulated independently, some specific conclusions can be drawn from these expansions. In particular, we show that effective application of MSDIS for variance reduction has the added benefit of reducing bias. References for this type of bias expansion may be found in Section 27 of [7], Chapter 2 of [28], or [14]. They are derived using Taylor series expansions and multidimensional central limit theorems. Let $\{\boldsymbol{C}_n = (C_n(1), \cdots, C_n(d)),\ n \geq 1\}$ be a sequence of iid vectors of length d and let $\boldsymbol{\nu} = (\nu_1, \cdots, \nu_d)$ where $E[\boldsymbol{C}_n] = \boldsymbol{\nu}$. Suppose we are interested in estimating $g(\boldsymbol{\nu})$ for some function g. In the case of ratio estimation $\boldsymbol{\nu} = (\nu_1, \nu_2)$ and $g(\boldsymbol{\nu}) = \nu_1/\nu_2$. Let $\bar{\boldsymbol{C}}_m = (1/m)\sum_{n=1}^{m} \boldsymbol{C}_n$. Then, under appropriate technical conditions on g and the moments of $\boldsymbol{C}_n$,

$$E[g(\bar{\boldsymbol{C}}_m)] = g(\boldsymbol{\nu}) + \frac{1}{2m}\sum_{i=1}^{d}\sum_{j=1}^{d}\sigma_{ij}g_{ij} + \mathrm{o}(1/m) \quad (3.15)$$

where $\sigma_{ij} = \mathrm{Cov}[C_n(i), C_n(j)]$ and $g_{ij} = \frac{\partial^2}{\partial x_i \partial x_j} g(\boldsymbol{x})|_{\boldsymbol{x}=\boldsymbol{\nu}}$. In our case, $\boldsymbol{C}_n = (G_n L'_{1n}, H_n L''_{1n})$, $\sigma_{11} = \mathrm{Var}_{\boldsymbol{P}'}[G_n L'_{1n}]$, $\sigma_{22} = \mathrm{Var}_{\boldsymbol{P}''}[H_n L''_{1n}]$ and $\sigma_{12} = 0$ (since the numerator and denominator are simulated independently). Note that in the above we assume for simplicity that m cycles of both the numerator and denominator are simulated. Differentiation of g yields $g_{11} = 0$, $g_{22} = 2\nu_1/\nu_2^3 = 2r/\nu_2^2$ and $g_{12} = -1/\nu_2^2$. Since $\sigma_{12} = 0$, the value of g_{12} does not enter into the MSDIS bias expansion. Therefore,

$$E[g(\bar{\boldsymbol{C}}_m)] = g(\boldsymbol{\nu}) + \frac{r\mathrm{Var}_{\boldsymbol{P}''}[H_n L''_{1n}]}{m\nu_2^2} + \mathrm{o}(1/m). \quad (3.16)$$

For the measures of interest in availability modeling, $r \geq 0$, $H_n \geq 0$ and $\nu_2 > 0$ so that, asymptotically, $E[g(\bar{\boldsymbol{C}}_m)] \geq g(\boldsymbol{\nu})$. Furthermore, this asymptotic bias expansion is independent of the importance sampling distribution $\boldsymbol{P}'$ chosen for simulation of the numerator.

For the steady-state unavailability we select $\boldsymbol{P}'' = \boldsymbol{P}$. By the results of Section III-C, in the three state example $r = \Theta(\epsilon^2)$, $\mathrm{Var}_{\boldsymbol{P}}[H_n] = \Theta(\epsilon)$ and $\nu_2 = \mathrm{E}_{\boldsymbol{P}}[H_n] = \Theta(1/\epsilon)$. Therefore, the leading term in the bias expansion is of order ϵ^5/m (relative bias is of order ϵ^3/m) which is typically quite small. With standard simulation the bias expansion, which now includes the effect of correlation between the numerator and denominator, becomes

$$E[g(\bar{G}_m, \bar{H}_m)] = g(\boldsymbol{\nu}) + \frac{r\mathrm{Var}_{\boldsymbol{P}}[H_n]}{m\nu_2^2} - \frac{\mathrm{Cov}_{\boldsymbol{P}}[G_n, H_n]}{m\nu_2^2} + \mathrm{o}(1/m). \quad (3.17)$$

For the three state example, $\mathrm{Cov}_{\boldsymbol{P}}[G_n, H_n] = h_2(h_1 + h_2)\mathrm{Var}_{\boldsymbol{P}}[n_F] = \Theta(\epsilon)$ so the dominant term in the bias expansion is $\Theta(\epsilon^3/m)$ which is significantly larger than the $\Theta(\epsilon^5/m)$ bias obtained using MSDIS. Moreover, using standard simulation, the relative bias (bias/r) is only $\Theta(\epsilon/m)$. These observations are consistent with the experimental results described in Section VI. Notice that one could also simulate the numerator and the denominator independently, without using importance sampling. In this case, the covariance term drops out in (3.17). For the three state example, the dominant term in the bias expansion becomes $\Theta(\epsilon^5/m)$, which is the same order as that with MSDIS. However, as seen from (3.16), choosing an importance sampling distribution $\boldsymbol{P}''$ for the purpose of reducing variance also has the beneficial effect of reducing bias.

For the MTTF, notice that r is large and $\nu_2 = \mathrm{P}\{\alpha_F < \alpha_0\}$ is small, which potentially makes the leading bias term large.

In practice, m may have to be very large in order for these asymptotic expansions to be valid. In particular, for small values of m the higher order terms may contribute in a nonnegligible way so that, e.g., $E[g(\bar{\boldsymbol{C}}_m)] \leq g(\boldsymbol{\nu})$. If bias turns out to be of significant concern, then a bias reducing technique such as jackknifing may be used to remove the leading term of order $1/m$ in the bias expansion (see [34]).

IV. Estimating Transient Measures

Simulation of the CTMC $\boldsymbol{Y}$ consists of two parts: simulating the sequence of states visited by the embedded DTMC $\boldsymbol{X}$ with transition matrix $\boldsymbol{P}$, and simulating the holding times in each of the states. We let t_i denote the holding time in state X_i. Thus, given that $X_i = j$, t_i has an exponential distribution with mean $1/q(j)$ and the (conditional) likelihood of t_i is simply $q(j)e^{-q(j)t_i}$. We let $\boldsymbol{t}_n = (t_0, \cdots, t_n)$ denote the first $n+1$ holding times of the CTMC. Given that $\boldsymbol{X}_n = (X_0, \cdots, X_n)$, the likelihood of $\boldsymbol{t}_n$ is therefore

$$\boldsymbol{f}(\boldsymbol{t}_n \mid \boldsymbol{X}_n) \equiv q(X_0)e^{-q(X_0)t_0} \ldots q(X_n)e^{-q(X_n)t_n} \quad (4.1)$$

and thus the likelihood of the sample path $(\boldsymbol{X}_n, \boldsymbol{t}_n)$ is

$$\boldsymbol{Q}(\boldsymbol{X}_n, \boldsymbol{t}_n) \equiv \boldsymbol{P}(\boldsymbol{X}_n)\boldsymbol{f}(\boldsymbol{t}_n \mid \boldsymbol{X}_n) \quad (4.2)$$

where $\boldsymbol{P}(\boldsymbol{X}_n)$ was defined in (3.2). Equation (4.2) gives the likelihood at the times of the jumps of the embedded DTMC.

We basically adapt the development in [15] in order to extend Proposition 3.1. Define $T_0 = 0$ and $T_n = t_0 + \cdots + t_{n-1}$ for $n \geq 1$. Then T_n is the time at which state X_n is entered, i.e., the time of the nth transition. Let $\boldsymbol{Y}_t = (Y_s, 0 \leq s \leq t)$. Let τ be an integer valued stopping time with respect to the sequence of pairs $\{(X_n, t_n),\ n \geq 0\}$, i.e., the event $\{\tau = n\}$ can be determined by $(X_0, t_0), \cdots, (X_n, t_n)$. We let $\boldsymbol{Q}'$ denote another measure for generating sequences $\{(X_n, t_n),\ n \geq 0\}$. We will specifically assume that $\boldsymbol{Q}'(\boldsymbol{X}_n, \boldsymbol{t}_n) = \boldsymbol{P}'(\boldsymbol{X}_n)\boldsymbol{f}'(\boldsymbol{t}_n \mid \boldsymbol{X}_n)$. With this factorization, the form of the contribution of the holding times to the likelihood, $\boldsymbol{f}'(\boldsymbol{t}_n \mid \boldsymbol{X}_n)$, is almost arbitrary (the restrictions will be discussed below), but the sequence of states selected does not depend upon the holding times. Let B_n be the subset of the sample paths of $\boldsymbol{Y}$ for which $\tau = n$.

Proposition 4.1: Let τ be an integer valued stopping time which, under $\boldsymbol{Q}$, is finite with probability one. Define $\alpha \equiv T_\tau$ and let Z be a (measurable) function of $\boldsymbol{Y}_\alpha$ for which $\mathrm{E}_{\boldsymbol{Q}}[|Z(\boldsymbol{Y}_\alpha)|] < \infty$. Let $\boldsymbol{Q}'$ be another measure of the form $\boldsymbol{Q}'(\boldsymbol{X}_n, \boldsymbol{t}_n) = \boldsymbol{P}'(\boldsymbol{X}_n)\boldsymbol{f}'(\boldsymbol{t}_n|\boldsymbol{X}_n)$ such that, under $\boldsymbol{P}'$, τ is finite with probability one and for any $(\boldsymbol{s}_n, \boldsymbol{t}_n) \in B_n$, $\boldsymbol{P}'(\boldsymbol{s}_n)\boldsymbol{f}'(\boldsymbol{t}_n \mid \boldsymbol{s}_n) \neq 0$ whenever $Z(\boldsymbol{Y}_{T_n})\boldsymbol{P}(\boldsymbol{s}_n)\boldsymbol{f}(\boldsymbol{t}_n \mid \boldsymbol{s}_n) \neq 0$. Then

$$\mathrm{E}_{\boldsymbol{Q}}[Z(\boldsymbol{Y}_\alpha)] = \mathrm{E}_{\boldsymbol{Q}'}[Z(\boldsymbol{Y}_\alpha)L'_1(\boldsymbol{X}_\tau)L'_2(\boldsymbol{X}_\tau, \boldsymbol{t}_\tau)] \quad (4.3)$$

where for any n, $L'_2(\boldsymbol{s}_n, \boldsymbol{t}_n) = \boldsymbol{f}(\boldsymbol{t}_n \mid \boldsymbol{s}_n)/\boldsymbol{f}'(\boldsymbol{t}_n \mid \boldsymbol{s}_n)$.

The proof of this proposition is essentially the same as that of Proposition 3.1. Notice that if the stopping time τ of

Proposition 4.2 is $\tau = \tau_F$, then the α of Proposition 4.1 is first time to failure, i.e., $\alpha = \alpha_F$. Measures defined over a fixed time interval $(0, t)$ (e.g., the expected interval availability) are handled in this formulation by defining $\tau = N(t) + 1$ where $N(t) = \max\{n : T_n \leq t\}$. The reliability $R(t)$ is handled by setting $\tau = \min(\tau_F, N(t) + 1)$ (since, with this definition, at time T_τ either a failure has occurred or simulated time has surpassed t) and setting $Z = 1_{\{T_\tau \leq t\}}$.

A. Estimating the Reliability

By Proposition 4.1, there are two importance sampling distributions to construct, corresponding to two likelihood ratios. The first distribution is for the embedded DTMC [corresponding to $L_1'(\boldsymbol{X}_\tau)$] and the second is for the state holding times given the DTMC's sample path [corresponding to $L_2'(\boldsymbol{X}_\tau, \boldsymbol{t}_\tau)$]. Lewis and Böhm [29] presented a technique for estimating the reliability. They apply "failure biasing" to the embedded DTMC; this causes failures to occur with higher probability and therefore quickly moves (biases) the DTMC toward the set of failed states. They also apply "forced transitions" to the holding time in state 0 (the state with all components operational). This forces the next component failure to occur before time t. Specifically, if $X_n = 0$ and $T_n < t$, then the next holding time, t_{n+1} is forced to be between zero and $t - T_n$ by selecting t_{n+1} from the conditional density given by

$$\boldsymbol{f}'(t_{n+1} \mid \boldsymbol{X}_n, \boldsymbol{t}_n) = \frac{\lambda_0 e^{-\lambda_0 t_{n+1}}}{1 - e^{-\lambda_0 (t - T_n)}}, \qquad 0 \leq t_{n+1} \leq t - T_n, \tag{4.4}$$

where λ_0 is the total failure rate in state 0. The simulation continues until time $\tau = \min(\tau_F, N(t) + 1)$.

Ross and Schechner [39] propose an alternative approach in which some, or all, of the holding times are conditioned out. If all holding times are conditioned out then no holding times are sampled and we set $Z = \mathrm{P}\{T_{\tau_F} \leq t \mid \boldsymbol{X}_{\tau_F}\}$. Calculation of Z requires computing the convolution of exponentially distributed random variables with different means. For a sequence of n states, this can be done in $\Theta(n^2)$ time using the recursions in [41]. Using failure biasing, τ_F will typically be small so that carrying out this computation is, in principle, not an obstacle. However, an effective and much simpler approach is to only condition out the total holding times in state 0 (which typically represents the bulk of the time anyway). The embedded DTMC is simulated until the set of failed states is entered, i.e., until time τ_F. Holding times in the nonzero states are randomly sampled, but no holding times are sampled for state 0. Let ϕ denote the total holding time in states other than 0: $\phi = \sum_{k=0}^{\tau_F} t_k \times 1_{\{X_k \neq 0\}}$. Let n_0 denote the number of visits to state 0 and let γ be a r.v. denoting the total holding time in state 0. Given n_0, γ has an Erlang distribution with shape parameter n_0 and scale parameter λ_0 and we write $\mathrm{P}\{\gamma \leq s \mid n_0\} = E_{n_0}(s, \lambda_0)$. We then set

$$\begin{aligned} Z &= \mathrm{P}\{\alpha_F \leq t \mid \boldsymbol{X}_{\tau_F}, \phi\} = \mathrm{P}\{\gamma \leq t - \phi \mid n_0\} \\ &= E_{n_0}(t - \phi, \lambda_0). \end{aligned} \tag{4.5}$$

Unlike Ross and Schechner, we apply the conditional Monte Carlo approach in addition to some form of failure biasing. By the variance reducing property of conditional expectations, (i.e., since $\mathrm{Var}[\mathrm{E}[X \mid Y]] \leq \mathrm{Var}[X]$, see, e.g., [38, p. 12]), the conditional approach plus failure biasing is always guaranteed to reduce the variance over just failure biasing. To see this, notice that

$$\begin{aligned} &\mathrm{Var}_{\boldsymbol{P}'}[L_1'(\boldsymbol{X}_{\tau_F}) 1_{\{\alpha_F \leq t\}}] \\ &\quad \geq \mathrm{Var}_{\boldsymbol{P}'}[\mathrm{E}[L_1'(\boldsymbol{X}_{\tau_F}) 1_{\{\alpha_F \leq t\}} | \boldsymbol{X}_{\tau_F}, \phi]] \\ &\quad = \mathrm{Var}_{\boldsymbol{P}'}[L_1'(\boldsymbol{X}_{\tau_F}) \mathrm{E}[1_{\{\alpha_F \leq t\}} | \boldsymbol{X}_{\tau_F}, \phi]] \\ &\quad = \mathrm{Var}_{\boldsymbol{P}'}[L_1'(\boldsymbol{X}_{\tau_F}) E_{n_0}(t - \phi, \lambda_0)] \end{aligned} \tag{4.6}$$

where the last equality follows from (4.5).

While no such analytic result exists for comparing conditioning with forcing, the conditioning approach has several advantages over the forcing approach. First, with forcing, different holding times must be generated for each value of t for which $R(t)$ is to be estimated. Because of sampling errors, the estimates of $R(t)$ may not be monotonic in t. Using the conditional approach, simultaneous, monotonic estimates of $R(t)$ are obtained. Second, with forcing, different conditional holding time distributions are used and different likelihood ratios must be maintained for each value of t for which $R(t)$ is to be estimated. This is not necessary in the conditional approach. Thus, it has computational time advantage when $R(t)$ is computed for multiple values of t simultaneously.

Another approach would be to use the technique of uniformization (see [20]). A discussion of approaches to using uniformization in simulations, including discrete conversions, may be found in [11]. In our context, failure rates are much less than repair rates and therefore $\lambda_0 \ll \bar{\lambda}$ where $\bar{\lambda} = \max\{q(i)\}$ is the maximum state exit rate. The number of transitions in the uniformized chain before exiting state 0 (sometimes called "pseudo transitions") is geometrically distributed with success parameter $\lambda_0 / \bar{\lambda} \approx 0$. Therefore, effective estimation of these rare event measures requires using some sort of importance sampling on the number of state 0 pseudo transitions. This, in turn, is very similar to using forced transitions.

B. Estimating the Expected Interval Availability

In this section we present two methods to estimate quantities, such as the expected interval availability, which take the form $r(t)/t$, with $r(t) = E[\int_{s=0}^{t} f(Y_s) ds]$. We assume that $\lambda_0 t$ is small so that very few failures are expected by time t. The first method, due to Lewis and Böhm [29], uses failure biasing and forcing as described in Section IV-A. The simulation ends at time $T_{N(t)+1} = t_0 + \cdots + t_{N(t)}$. With this notation, $t_{N(t)}$ is the holding time that crosses the threshold t. A practical implementation of this method typically turns off the forcing after L visits to state 0 at some value of L for which $E_L(t, \lambda_0)$ is extremely small; without this modification $N(t)$ may grow to be quite large and, furthermore, the simulation may generate extremely unlikely sample paths having an unusually large number of visits to state 0 in the interval $(0, t)$.

To apply the conditional Monte Carlo approach to $r(t)$, we begin with an important result from Fox and Glynn [11]:

$$r(t) = E[\sum_{k=0}^{N(t)-1} g(X_k)] \tag{4.7}$$

where $g(i) = f(i)/q(i)$ and the expectation is with respect to the transition matrix $\boldsymbol{P}$. Now, as suggested in [11], we could generate holding times for the sole purpose of determining $N(t)$, and then ignore these holding times by using $\sum_{k=0}^{N(t)-1} g(X_k)$ to estimate $r(t)$. However, we would still have to use conditioning or some sort of importance sampling, such as forcing, on the holding times in state 0 since otherwise $N(t) = 0$ with high probability. Similarly, uniformization implementations based on (4.7) would also require importance sampling on the number of state 0 pseudo transitions in order to be effective. To combine forcing with (4.7), we write

$$\begin{aligned} r(t) &= E[\sum_{k=0}^{N(t)-1} g(X_k) \mid t_0 > t]\mathrm{P}\{t_0 > t\} \\ &\quad + E[\sum_{k=0}^{N(t)-1} g(X_k) \mid t_0 \le t]\mathrm{P}\{t_0 \le t\} \\ &= E[\sum_{k=0}^{N(t)-1} g(X_k) \mid t_0 \le t]\mathrm{P}\{t_0 \le t\} \end{aligned} \tag{4.8}$$

since if $t_0 > t$, then $N(t) = 0$ and therefore $\sum_{k=0}^{N(t)-1} g(X_k) = 0$. Equation (4.8) can also be combined with failure biasing.

We next extend (4.7) in a way that allows us to condition out the state 0 holding times. While the development below is in terms of the original embedded DTMC, the results extend directly to using importance sampling as described in Proposition 3.1. We also present the method in terms of conditioning out only the holding times in state 0, although the method also applies more generally. Analogous to the approach in Section IV-A, define $n_0(k) = \sum_{j=0}^{k} 1_{\{X_j=0\}}$ and $\phi_k = \sum_{j=0}^{k} t_j \times 1_{\{X_j \ne 0\}}$. With these definitions, $n_0(k)$ is the number of visits to state 0 and ϕ_k is the total holding time in the nonzero states at the kth transition. Then, by (4.7)

$$\begin{aligned} r(t) &= E[\sum_{k=0}^{\infty} g(X_k) 1_{\{k \le N(t)-1\}}] \\ &= \sum_{k=0}^{\infty} E[g(X_k) 1_{\{N(t) \ge k+1\}}] \\ &= \sum_{k=0}^{\infty} E[E[g(X_k) 1_{\{N(t) \ge k+1\}} \mid \boldsymbol{X}_k, \phi_k]] \\ &= E[\sum_{k=0}^{\infty} g(X_k) E_{n_0(k)}(t - \phi_k, \lambda_0)]. \end{aligned} \tag{4.9}$$

The key step in the above derivation follows since on $\{N(t) \ge k+1\} = \{t_0 + \ldots + t_k \le t\}$. The exchanges of expectation and summation are easily justified for finite state spaces by using the dominated convergence theorem (see [3]).

To apply (4.9) requires determining a stopping criterion. We could simulate until $\phi_k \ge t$ at which point $E_{n_0(k)}(t - \phi_k, \lambda_0) = 0$. However, since repairs are fast, ϕ_k grows slowly and therefore an excessive number of transitions may have to be simulated. The summation could be truncated at some finite value. However, this introduces bias error. While the error is easily bounded, we prefer unbiased estimates, particularly for quantities such as the interval unavailability which itself is quite small. A simple unbiased estimate can be constructed in a reasonable amount of time as follows: after the Lth visit to state 0, begin sampling the state 0 holding times and adding them to ϕ_k. Very quickly, ϕ_k will exceed t and the sample is then complete. More formally, let $N_0(L)$ be the (discrete) time at which state 0 is entered for the Lth time. For $k \ge N_0(L)+1$, let $\tilde{\phi}_k = \phi_{N_0(L)} + \sum_{j=N_0(L)+1}^{k} t_j$. Then, arguing as above,

$$\begin{aligned} r(t) =& E[\sum_{k=0}^{N_0(L)} g(X_k) E_{n_0(k)}(t - \phi_k, \lambda_0)] \\ &+ E[\sum_{k=N_0(L)+1}^{\infty} g(X_k) E_L(t - \tilde{\phi}_k, \lambda_0)]. \end{aligned} \tag{4.10}$$

The estimators for the distribution of interval availability can be formulated in a similar way. We derive these estimators in the Appendix for the sake of completeness.

V. Implementation Issues

In this section we consider the implementation of the different variance reduction techniques described in the previous sections. These techniques have been implemented in the SAVE package [16], [18] so that large models can be simulated. One salient feature of our implementation is that we use one simulation run for estimating all the measures. Regenerative simulation is used with the "all components operational" state as the regeneration state. The event generator simulates only the embedded Markov chain (DTMC formulated in Section III-A). For the steady-state measures we accumulate functions of the mean holding times in the various states, and for the transient measures we accumulate functions of the sample holding times (from exponential distributions) in the various states. In the following paragraphs we describe the implementation of the importance sampling technique for the various measures.

Recall that we formulated the likelihood ratios for the transient measure in Proposition 4.1 as the product of two likelihood ratios $L_1'(\boldsymbol{X}_\tau)$ and $L_2'(\boldsymbol{X}_\tau, \boldsymbol{t}_\tau)$. The first likelihood ratio corresponds to the embedded Markov chain and it is needed for the steady-state as well as the transient measures as indicated in Propositions 3.1 and 4.1. On the other hand, $L_2'(\boldsymbol{X}_\tau, \boldsymbol{t}_\tau)$ corresponds to the holding times, given a sample path on the embedded DTMC; this likelihood ratio is needed only for transient measures and is different for different transient estimators.

The importance sampling for the embedded Markov chain is based on the following heuristics. As suggested in Section III-C, we need to move the system very quickly to the set of failed states F, and once F is entered, the importance sampling should be turned off so that the system quickly returns to state 0, the "all components operational" state. We achieve this by increasing the probability of failure transitions over repair transitions. This has been called "failure biasing" in [29]. We assign a combined probability *bias1* to the failure transitions in all the states where both failure and repair transitions are feasible. Individual failure and repair transitions are selected in the ratio of their rates given that a failure or a repair is selected, respectively. We call this the *Bias1/Ratio* method, or simply *Bias1* method. We have found two other methods useful for selecting individual failure transitions, given that a failure has occurred. The first is to use a uniform distribution on the failure transitions which has very good performance for "unbalanced systems" as shown in Section VI. We call this the *Bias1/Balancing* method. The second is to give a higher combined probability *bias2* to those failure transitions which correspond to component types which have at least one component of their type already failed. This exhausts the redundancy quickly and has much better performance for "balanced systems" as shown in Section VI. We call this the *Bias1/Bias2* method. Another method for failure biasing in acyclic reliability models is found in [13].

For the steady-state availability each regenerative cycle corresponds to a sample. We use either the DIS or the MSDIS method given in Section III-B to estimate the steady-state availability. For the mean time to failure, a sample ends when either the regeneration occurs or the system enters one of the system failed states from the set F. In the latter case, we continue to simulate the embedded Markov chain until the regeneration occurs before starting a new sample. This wastes only a few events as typically a regenerative cycle has a very few events (approximately twice the average redundancy which is typically 2 or 3). Once again, we use either the DIS or the MSDIS method to estimate the mean time to failure. For the transient measures, multiple regenerative cycles may be contained in a sample. Moreover, a sample typically ends either when a failure occurs or when the time interval expires, which is usually in the middle of some regenerative cycle. As in the mean time to failure case, we continue to simulate the embedded Markov chain until the next regeneration occurs before starting a new sample. Separate accumulators for the appropriate likelihood ratios are maintained for each transient estimator and for each time horizon of interest. Thus, all measures can be estimated simultaneously from a single simulation run.

VI. Examples and Discussions

In this section, we provide an example, based on a model of a computing system, to illustrate the effectiveness of the different variance reduction techniques discussed in the previous sections. A block diagram of the computing system considered is shown in Fig. 1. We use two different parameter sets to create a "balanced" and an "unbalanced" system. A balanced system is one in which each type of component has the same amount of redundancy, (i.e., same number of components of a type must fail in order that the system fail, e.g., 1-out-of-2 of a type has the same redundancy as 3-out-of-4 of another type); in addition, the components must have the same order of magnitude failure rates. A system that is not balanced is unbalanced.

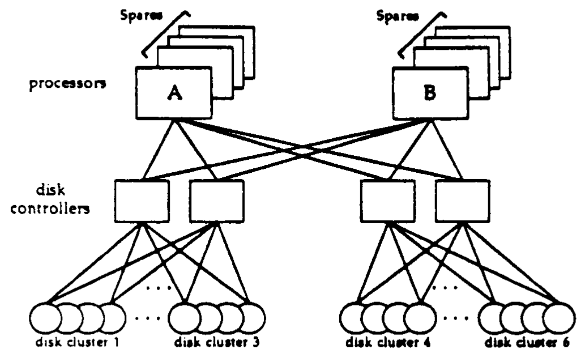


Fig. 1 A block diagram of the computing system modeled.

For a balanced system we select two sets of processors with two processors per set, two sets of controllers with two controllers per set, and six clusters of disks, each consisting of four disk units. In a disk cluster, data are replicated so that one disk can fail without affecting the system. The "primary" data on a disk is replicated such that one third is on each of the other three disks in the same cluster. Thus, one disk in each cluster can be inaccessible without losing access to the data. The connectivity of the system is shown in Fig. 1. We assume that when a processor of a given type fails, it has a 0.01 probability of causing the operating processor of the other type to fail. Each unit in the system has two failure modes which occur with equal probability. The failure rates of processors, controllers, and disks are assumed to be 1/2000, 1/2000, and 1/6000 per hour, respectively. The repair rates for all mode 1 and all mode 2 failures are 1 per hour and 1/2 per hour, respectively. Components are repaired by a single repairman who chooses components at random from the set of failed units. The system is defined to be operational if all data are accessible to both processor types, which means that at least one processor of each type, one controller in each set, and 3 out of 4 disk units in each of the 6 disk clusters are operational. We also assume that operational components continue to fail at the given rates when the system is failed.

We make minor changes to the above parameters setting in order to create an unbalanced system. We increase the number of processors of each type to 4, and double each processor's failure rate to 1/1000 per hour. We decrease the failure rates of all other components by a factor of ten. In this system, although a processor failure is more likely to occur in a failure transition, it is less likely to cause a system failure due to the high processor redundancy. This is typical behavior for an unbalanced system.

A. Steady-State Measures

In this section we discuss the results of our experiments for estimating the steady-state unavailability and the mean time to failure. Numerical (nonsimulation) results for these measures were obtained using the SAVE package [18]. Since the balanced system has a few hundred thousand states and the unbalanced system has close to a million states, only bounds could be computed ([35]). These bounds are very tight and typically do not differ from the exact results significantly. We simulate both the balanced and the unbalanced systems. The goal of the simulation experiments is to study the efficiency of the importance sampling methods, described in this paper, compared to standard simulation. We also experimented with the MSDIS technique described in Section III. It is shown that the *Bias1* method gives many orders of magnitude variance reduction over the standard Monte Carlo simulation. Moreover, further significant improvements can be obtained using the *Bias1/Bias2* method for the balanced systems and *Bias1/Balancing* method for the unbalanced systems. Further improvements are obtained when these methods are combined with MSDIS. Table I and Table II show the results obtained for the balanced and the unbalanced systems, respectively. We ran the simulation long enough so that the smallest entry in the tables for the percentage relative half-widths of the 90% confidence intervals was less than 5%. The percentage relative half-width of a confidence interval is defined to be 100% times the confidence interval half-width divided by the point estimate. This corresponds to approximately 100 000 events for each entry in Table I and 1 000 000 events for each entry in Table II, respectively. For the MSDIS entries, we assigned 10% of the total events to estimate the denominator (numerator) for unavailability (MTTF) as suggested in Section III-D. Based on empirical results obtained in [4], [19], and [43], the values for *bias1* and *bias2* were selected as follows: for DIS, 0.5 and 0.5, and for MSDIS, 0.9 and 0.9.

For the balanced system (Table I), the *Bias1/Bias2* method is most effective, which supports our intuition that it helps push the system quickly toward a likely path to failure. For the unbalanced system (Table II), the *Bias1/Balancing* is the most effective method, which also supports our intuition as follows. By making individual failures equally likely we are also increasing the failure probability of a more reliable but less redundant component, thus leading to a more likely path to failure.

Note that the percentage relative half-widths for both the steady-state unavailability (U) and the mean time to failure (MTTF) are approximately equal. This is because the estimate of U is approximately proportional to the estimate of 1/MTTF. To see this, using the notation of Section II, $\min(\alpha_F, \alpha_0) = \alpha_0$ with high probability when no importance sampling is used. Thus an individual sample r.v. in the numerator of the ratio for MTTF (2.9) is equal to the r.v. in the denominator of U (2.7) with high probability. Now for the three state model, a sample r.v. G in the numerator of U is $G = h_2 \times n_F$ where n_F is the number of times the failure state is entered. Using our importance sampling schemes, $G = h_2 1_{\{n_F=1\}}$ with high probability. Furthermore, $1_{\{n_F=1\}} = 1_{\{\alpha_F<\alpha_0\}}$ with high probability so an individual sample r.v. in the denominator of the MTTF ratio is proportional to the r.v. in the numerator of U with high probability. Thus, an estimate of U is approximately proportional to 1/MTTF. Finally, direct manipulation of the asymptotic variance (3.6) shows that the relative half-width of a ratio is equal to the relative half width of its reciprocal.

We next performed the coverage experiments (see e.g., [26]) to determine the validity of the confidence intervals that are formed based on the asymptotic central limit theorems described in Section III. Such studies are important since certain variance reduction techniques sometimes do not produce valid confidence intervals, except for very long run-lengths (see e.g., [26]). In such cases, the variance reduction technique cannot be relied upon to actually shorten simulation run lengths. We performed coverage experiments on estimates of the steady-state unavailability, U, in the above described balanced system as follows. We chose three run lengths corresponding to small, medium and large sample sizes and considered three ways of estimating U: standard simulation, the *Bias1/Bias2* method with DIS and the *Bias1/Bias2* method with MSDIS. For each method and run length we ran $R = 100$ replications and formed point estimates $\hat{U}_1, \cdots, \hat{U}_R$ and 90% confidence intervals. We then calculated the mean percent relative bias ($= 100\% \times (1/R)\sum_{i=1}^{R}(\hat{U}_i - U)/U$) and the standard deviation of this mean. Note that if an estimate is unbiased, then its mean percent relative bias should converge to zero as $R \to \infty$. We also calculated the 90% coverage which is the percentage of the (alleged) 90% confidence that actually contain the true value U. If the confidence interval is valid, then by definition, the 90% coverage should be close to 90%.

We also computed the mean percent relative half width of the 90% confidence intervals. For each replication, this relative value is computed using the point estimate and not the true value. The mean is computed over all replications with a nonzero point estimate. The results are listed in Table III. As anticipated from Section III-E, the standard estimate is significantly more biased than either the DIS-*Bias1/Bias2* or the MSDIS-*Bias1/Bias2* estimates and that its confidence intervals are at least an order of magnitude wider. Furthermore, for the small run length, its coverage drops significantly below 90%. In fact, there were no system failures in the runs corresponding to the 46% of the confidence intervals which did not contain U. Using our variance reduction techniques, all the coverages are close to the nominal 90% value except for the longest run using MSDIS which had a coverage of 81%. Changing the random number generator from the multiplicative linear congruential generator $I_{n+1} = (I_n \times 16807) \bmod 2^{31} - 1$ to the combined generator described in [27], and running $R = 200$ replications, increased the coverage to 85% which still represents a statistically significant departure from 90%. Despite a considerable effort, we have been unable to identify the source of this slight coverage problem. However, note that the first nonzero digit in U is in the sixth decimal place, so the problem (if any) is occurring in the eighth decimal place. In practice, such high precision may not be warranted, given inaccuracies in model parameters and distributional assumptions.

B. Transient Measures

In this section we discuss the results of our experiments for estimating reliability and expected interval availability. Recall that for transient measures we not only want the system to move quickly toward the set of system failed states F,

TABLE I
UNAVAILABILITY AND MTTF ESTIMATES AND RELATIVE HW IN A BALANCED SYSTEM

Numerical Results	Direct	Bias1 (0.5)	Bias1/Balancing (0.5)	Bias1/Bias2 (0.5/0.5)	MS-Bias1/Bias2 (0.9/0.9)
0.9309×10^{-5} Unavailability	1.0171×10^{-5} ± 27.1 %	0.9779×10^{-5} ± 7.6 %	0.9547×10^{-5} ± 6.2 %	0.9395×10^{-5} ± 2.7 %	0.9317×10^{-5} ± 1.0 %
$0.1637 \times 10^{+6}$ MTTF	$0.1524 \times 10^{+6}$ ± 25.7 %	$0.1581 \times 10^{+6}$ ± 7.0 %	$0.1631 \times 10^{+6}$ ± 5.7 %	$0.1626 \times 10^{+6}$ ± 2.5 %	$0.1633 \times 10^{+6}$ ± 1.0 %

TABLE II
UNAVAILABILITY AND MTTF ESTIMATES AND RELATIVE HW IN AN UNBALANCED SYSTEM

Numerical Results	Direct	Bias1 (0.5)	Bias1/Balancing (0.5)	Bias1/Bias2 (0.5/0.5)	MS-Bias1/Balancing (0.9)
0.6967×10^{-7} Unavailability	0.4164×10^{-7} ± 164.5 %	0.6644×10^{-7} ± 46.1 %	0.6976×10^{-7} ± 2.4 %	0.6375×10^{-7} ± 5.6 %	0.6810×10^{-7} ± 2.2 %
$0.2188 \times 10^{+8}$ MTTF	$0.4703 \times 10^{+8}$ ± 164.5 %	$0.2227 \times 10^{+8}$ ± 43.7 %	$0.2183 \times 10^{+8}$ ± 2.3 %	$0.2349 \times 10^{+8}$ ± 5.1 %	$0.2222 \times 10^{+8}$ ± 2.0 %

TABLE III
COVERAGE EXPERIMENTS FOR THE BALANCED SYSTEM

Events per Replication	Direct Simulation			Bias1/Bias2 (0.5/0.5)			MS - Bias1/Bias2 (0.9/0.9)		
	Relative Bias (Std. Dev.)	Relative HW	Coverage	Relative Bias (Std. Dev.)	Relative HW	Coverage	Relative Bias (Std. Dev.)	Relative HW	Coverage
2000	6.95 % (12.88 %)	144.40 %	54 %	0.74 % (1.21 %)	18.88 %	85 %	0.35 % (0.91 %)	6.70 %	91 %
20000	-3.94 % (3.43 %)	65.47 %	90 %	0.39 % (0.34 %)	5.99 %	90 %	0.11 % (0.22 %)	2.56 %	91 %
200000	1.29 % (1.09 %)	19.6 %	96 %	0.05 % (0.13 %)	1.90 %	90 %	0.03 % (0.073 %)	1.04 %	81 %

but also reach there before the observation period expires. Since these two issues are, in some sense, orthogonal, we use the same technique as in the steady-state case to bias the embedded Markov chain toward the system failed set, in addition to another independent technique (e.g., forcing or conditioning as discussed in Section V) to reduce the variance due to holding times in the various states. The likelihood ratios corresponding to these two aspects of simulation are independent and can be formulated as in Proposition 4.1. The goal of the simulation is to study the effect of the forcing and conditioning techniques. We considered only the balanced system. For each measure, we allowed each method to run for 400 000 events. Standard simulation was not considered as it is very ineffective for estimating transient measures. The results are given in Tables IV and V.

For all methods, the confidence intervals are smaller for some range of intermediate time periods and wider at the ends. To explain this, we recognize two key factors affecting the variance of the estimates; namely, the number of replications in a simulation run and the value of *bias1* used with importance sampling. For smaller time intervals, there are more replications in a simulation run than for larger time intervals (since we kept the total number of events fixed). This contributes to a larger variance for larger time intervals. Furthermore, for each time interval, there is an optimal value for *bias1* which maximizes the variance reduction. While *bias1* = 0.5 may be close to optimal for some intermediate range of time intervals, it departs from the optimal value for either smaller or larger time intervals.

The two tables indicate that forcing and conditioning are

TABLE IV
UNRELIABILITY ESTIMATES IN A BALANCED SYSTEM

Time (t) in Hours	Numerical Unreliability	Bias1 (0.5)			Bias1/Balancing (0.5)			Bias1/Bias2 (0.5/0.5)		
		Standard	Forcing	Conditioning	Standard	Forcing	Conditioning	Standard	Forcing	Conditioning
4	$.1528 \times 10^{-4}$	$.1395 \times 10^{-4}$ ± 25.0 %	$.1477 \times 10^{-4}$ ± 5.2 %	$.1435 \times 10^{-4}$ ± 7.5 %	$.1255 \times 10^{-4}$ ± 21.3 %	$.1531 \times 10^{-4}$ ± 4.0 %	$.1463 \times 10^{-4}$ ± 5.5 %	$.1583 \times 10^{-4}$ ± 7.0 %	$.1522 \times 10^{-4}$ ± 1.4 %	$.1537 \times 10^{-4}$ ± 1.6 %
16	$.8734 \times 10^{-4}$	$.8716 \times 10^{-4}$ ± 10.7 %	$.8569 \times 10^{-4}$ ± 4.7 %	$.8383 \times 10^{-4}$ ± 6.5 %	$.8855 \times 10^{-4}$ ± 8.4 %	$.8565 \times 10^{-4}$ ± 3.7 %	$.8301 \times 10^{-4}$ ± 4.8 %	$.8693 \times 10^{-4}$ ± 3.3 %	$.8699 \times 10^{-4}$ ± 1.3 %	$.8787 \times 10^{-4}$ ± 1.4 %
64	$.3804 \times 10^{-3}$	$.3748 \times 10^{-3}$ ± 6.2 %	$.3811 \times 10^{-3}$ ± 4.5 %	$.3738 \times 10^{-3}$ ± 5.5 %	$.3756 \times 10^{-3}$ ± 4.9 %	$.3680 \times 10^{-3}$ ± 3.5 %	$.3649 \times 10^{-3}$ ± 4.2 %	$.3801 \times 10^{-3}$ ± 1.8 %	$.3841 \times 10^{-3}$ ± 1.1 %	$.3833 \times 10^{-3}$ ± 1.1 %
256	$.1552 \times 10^{-2}$	$.1536 \times 10^{-2}$ ± 5.8 %	$.1587 \times 10^{-2}$ ± 4.7 %	$.1584 \times 10^{-2}$ ± 4.6 %	$.1579 \times 10^{-2}$ ± 4.9 %	$.1552 \times 10^{-2}$ ± 3.8 %	$.1546 \times 10^{-2}$ ± 3.6 %	$.1565 \times 10^{-2}$ ± 1.5 %	$.1578 \times 10^{-2}$ ± 1.0 %	$.1565 \times 10^{-2}$ ± 0.9 %
1024	$.6226 \times 10^{-2}$	1.1968×10^{-2} ± 73.3 %	$.7584 \times 10^{-2}$ ± 19.7 %	$.7017 \times 10^{-2}$ ± 14.1 %	$.6627 \times 10^{-2}$ ± 11.6 %	$.6667 \times 10^{-2}$ ± 11.2 %	$.6584 \times 10^{-2}$ ± 10.2 %	$.6275 \times 10^{-2}$ ± 4.9 %	$.6233 \times 10^{-2}$ ± 4.3 %	$.6206 \times 10^{-2}$ ± 3.0 %

TABLE V
EXPECTED INTERVAL UNAVAILABILITY ESTIMATES IN A BALANCED SYSTEM

Time (t) in Hours	Numerical Interval Unavailability	Bias1 (0.5)			Bias1/Balancing (0.5)			Bias1/Bias2 (0.5/0.5)		
		Standard	Forcing	Conditioning	Standard	Forcing	Conditioning	Standard	Forcing	Conditioning
4	$.3178 \times 10^{-5}$	$.3057 \times 10^{-5}$ ± 35.0 %	$.3110 \times 10^{-5}$ ± 6.8 %	$.3224 \times 10^{-5}$ ± 4.7 %	$.2572 \times 10^{-5}$ ± 27.5 %	$.3213 \times 10^{-5}$ ± 5.2 %	$.3189 \times 10^{-5}$ ± 3.7 %	$.3204 \times 10^{-5}$ ± 9.6 %	$.3207 \times 10^{-5}$ ± 2.0 %	$.3205 \times 10^{-5}$ ± 1.5 %
16	$.7322 \times 10^{-5}$	$.6945 \times 10^{-5}$ ± 15.8 %	$.7148 \times 10^{-5}$ ± 6.7 %	$.7401 \times 10^{-5}$ ± 3.9 %	$.7677 \times 10^{-5}$ ± 12.9 %	$.7462 \times 10^{-5}$ ± 5.5 %	$.7219 \times 10^{-5}$ ± 3.1 %	$.7489 \times 10^{-5}$ ± 4.8 %	$.7511 \times 10^{-5}$ ± 2.1 %	$.7383 \times 10^{-5}$ ± 1.2 %
64	$.8806 \times 10^{-5}$	$.8621 \times 10^{-5}$ ± 11.4 %	$.8866 \times 10^{-5}$ ± 8.3 %	$.9276 \times 10^{-5}$ ± 5.1 %	$.8651 \times 10^{-5}$ ± 8.4 %	$.9012 \times 10^{-5}$ ± 6.6 %	$.8822 \times 10^{-5}$ ± 4.0 %	$.8862 \times 10^{-5}$ ± 3.2 %	$.8966 \times 10^{-5}$ ± 2.4 %	$.8821 \times 10^{-5}$ ± 1.5 %
256	$.9178 \times 10^{-5}$	$.9882 \times 10^{-5}$ ± 19.7 %	$.9442 \times 10^{-5}$ ± 12.9 %	$.9187 \times 10^{-5}$ ± 10.5 %	1.1150×10^{-5} ± 14.9 %	1.0766×10^{-5} ± 13.9 %	1.0010×10^{-5} ± 9.3 %	$.9049 \times 10^{-5}$ ± 4.4 %	$.9133 \times 10^{-5}$ ± 4.2 %	$.9063 \times 10^{-5}$ ± 3.3 %
1024	$.9277 \times 10^{-5}$	$.9679 \times 10^{-5}$ ± 91.5 %	$.9677 \times 10^{-5}$ ± 91.5 %	1.3573×10^{-5} ± 103.5 %	$.3691 \times 10^{-5}$ ± 36.3 %	$.3707 \times 10^{-5}$ ± 36.1 %	$.6168 \times 10^{-5}$ ± 51.6 %	1.0760×10^{-5} ± 64.8 %	1.0728×10^{-5} ± 65.0 %	$.7822 \times 10^{-5}$ ± 21.3 %

most effective for short time intervals. This is intuitive because for a long interval enough transitions occur before the interval expires, and therefore, the embedded Markov chain has a chance to reach the system failed set using only failure biasing. This is not true for short intervals, and therefore, either forcing transitions to occur before the end of the period or conditioning the holding time out in state 0 has a significant effect. Both forcing and conditioning give similar results for unreliability, while conditioning is consistently better for the interval unavailability. Note that for interval unavailability we are using (4.7) with conditioning, but not with forcing. However, forcing can be similarly combined with (4.7) to possibly yield better results. Also, note that *Bias1/Bias2* method is consistently better than both the *Bias1* and the *Bias1/Balancing* methods. This is consistent with a similar conclusion with respect to the steady-state measures in a balanced system.

VII. Summary and Directions for Future Work

In this paper, we have developed a unified framework for simulation of Markovian models of highly dependable systems. Conventional numerical techniques are difficult to apply to this class of stochastic models because of the fact that the size of the state space of the Markovian model increases exponentially with the number of components in the system.

On the other hand, simulation algorithms tend to be relatively insensitive to the size of the state space of the simulated Markovian model, both in terms of storage and computational requirements. However, standard simulation is inefficient in our setting because the principle focus of interest; namely, system failures, occur so infrequently in highly dependable systems. As a consequence, few system failures, if any, would be observed if standard simulation methods were to be used in our problem context.

The emphasis in this paper has therefore centered on applying variance reduction techniques to improve the efficiency of the simulations associated with Markovian models of highly dependable systems. We have reviewed the basic theory of importance sampling in several elementary problem settings and then used this insight to develop sampling heuristics for the complex systems of interest here. Different variants of these ideas were developed for both transient and steady-state dependability measures. In addition, we have "fine-tuned" the importance sampling techniques to take advantage of the structure of highly dependable systems which are either balanced or unbalanced.

Our work has also shown that importance sampling may be fruitfully applied in conjunction with a variance reduction method known as conditioning. The basic idea here was to observe that highly dependable systems spend a significant fraction of time in the state in which all components are fully operational. Since the stochastic behavior of the time spent in the fully operational state was easy to calculate analytically, this permitted us to effectively integrate out the randomness in our importance sampling estimators due to the holding times in the fully operational state.

Our empirical investigation showed that the combined variance reduction obtained by using both conditioning and importance sampling is typically substantial. In fact, in all of our experiments, our methods yielded estimators in which the variance was decreased by several orders of magnitude. In a recent analytical work in this area ([44]), the heuristics used in this paper have been proven to be very efficient. Our empirical work also showed that the confidence intervals associated with our estimators typically provided acceptable levels of coverage. We view this as important, since the scientific representation of the accuracy of a simulation estimator is usually gauged through a confidence interval.

A number of possible directions for future research present themselves. One important issue relates to the fact that the importance sampling heuristics presented in this paper were basically developed for systems in which system dependability is achieved principally through high component reliability. However, another approach to obtaining high system dependability is through high levels of component redundancy. Importance sampling methods appropriate for the analysis of highly redundant systems differ from those presented here. Such techniques would likely have important ramifications for the simulation of certain highly dependable systems.

A second important research area involves the generalization of the ideas developed in this paper to stochastic models of highly dependable systems in which the underlying failure and repair distributions are nonexponential. Since the resulting stochastic process is typically no longer either Markovian or regenerative, many of the ideas presented in this paper cannot be implemented directly. Some work in this direction is reported in [37].

Appendix
Estimating the Distribution of Interval Availability

We will find it more convenient to estimate the distribution of the time in the set of failure states, $U(t,x) = \mathrm{P}\{U(t) \le x\}$, where $U(t) = \int_{s=0}^{t} 1_{\{Y_s \in F\}} ds$. Since $A(t) = 1 - U(t)/t$ we have $A(t,x) = \mathrm{P}\{A(t) \le x\} = 1 - U(t,(1-x)t)$. To derive an estimator for $U(t,x)$, write

$$U(t,x) = U_1(t,x) + U_2(t,x) \tag{A.1}$$

where $U_1(t,x) = \mathrm{P}\{U(t) \le x, Y_t \notin F\}$ and $U_2(t,x) = \mathrm{P}\{U(t) \le x, Y_t \in F\}$. Define $D_i = \sum_{j=0}^{i} t_j 1_{\{X_j \in F\}}$ and $C_i = \sum_{j=0}^{i} t_j 1_{\{X_j \ne 0, X_j \notin F\}}$, with γ_i, ϕ_i and $n_0(i)$ as defined previously. Note that $\phi_i = C_i + D_i$ and $T_{i+1} = \gamma_i + C_i + D_i$. Consider $U_1(t,x)$:

$$\begin{aligned} U_1(t,x) &= \mathrm{P}\{U(t) \le x, Y_t \notin F\} \\ &= \sum_{i=0}^{\infty} \mathrm{P}\{U(t) \le x, X_i \notin F, N(t) = i\} \\ &= \sum_{i=0}^{\infty} \mathrm{P}\{D_{i-1} \le x, X_i \notin F, T_i < t < T_{i+1}\} \\ &= \sum_{i=0}^{\infty} \mathrm{P}\{D_{i-1} \le x, X_i \notin F, \gamma_{i-1} \\ &\qquad + \phi_{i-1} < t < \gamma_i + \phi_i\}. \end{aligned} \tag{A.2}$$

Now if $X_i = 0$, then $\phi_{i-1} = \phi_i$. If $X_i \notin F$ and $X_i \ne 0$, then $n_0(i-1) = n_0(i)$ and therefore $\gamma_{i-1} = \gamma_i$. In either case, by conditioning on the sequence of states and the holding times in the nonzero states, we can write

$$\begin{aligned} U_1(t,x) = E\Bigg[&\sum_{i=0}^{\infty} 1_{\{D_{i-1} \le x, X_i \notin F\}} \big(E_{n_0(i-1)}(t - \phi_{i-1}, \lambda_0) \\ &- E_{n_0(i)}(t - \phi_i, \lambda_0) \big) \Bigg]. \end{aligned} \tag{A.3}$$

Now consider $U_2(t,x)$. If $Y_t \in F$ and $N(t) = i$, then $U(t) = D_{i-1} + t - T_i = t - \gamma_{i-1} - C_{i-1}$. Furthermore, on this set $n_0(i-1) = n_0(i)$, $\gamma_{i-1} = \gamma_i$ and $C_{i-1} = C_i$ so that

$$\begin{aligned} U_2(t,x) &= \sum_{i=0}^{\infty} E\Big[1_{\{D_{i-1} \le x, X_i \in F\}} 1_{\{t - \gamma_{i-1} - C_{i-1} \le x\}} \\ &\qquad 1_{\{\gamma_{i-1} + C_{i-1} + D_{i-1} < t < \gamma_i + C_i + D_i\}} \Big] \\ &= E\Big[\sum_{i=0}^{\infty} 1_{\{D_{i-1} \le x, X_i \in F\}} \\ &\qquad \big(E_{n_0(i)}(t - D_{i-1} - C_{i-1}, \lambda_0) \\ &\qquad - E_{n_0(i)}(t - C_i - \min(D_i, x), \lambda_0) \big) \Big]. \end{aligned} \tag{A.4}$$

This development provides a computationally attractive way to estimate $U(t, x)$ without sampling from the state 0 holding time distribution. It is easily combined with failure biasing. Practical implementations may require truncation or stopping rules as described in Section IV-B.

REFERENCES

[1] A. Bobbio and K. S. Trivedi, "An aggregation technique for the transient analysis of stiff Markov chains," *IEEE Trans. Comput.*, vol. C-35, pp. 803–814, 1986.

[2] W. G. Bouricius, W. C. Carter, and P. R. Schneider, "Reliability modeling techniques for self-repairing computer systems," in *Proc. ACM Nat. Conf.*, San Francisco, CA, 1969, pp. 295–309.

[3] K. L. Chung, *Markov Chains With Stationary Transition Probabilities*, Second Edition. New York: Springer-Verlag, 1967.

[4] A. E. Conway and A. Goyal, "Monte Carlo simulation of computer system availability/reliability models," in *Proc. Seventeenth Symp. Fault-Tolerant Comput.*, Pittsburgh, PA, 1987, pp. 230–235.

[5] A. Costes, J. E. Doucet, C. Landrault, and J. C. Laprie, "SURF—A program for dependability evaluation of complex fault-tolerant computing systems," in *Proc. Eleventh Symp. Fault-Tolerant Comput.*, Portland, ME, 1981, pp. 72–78.

[6] M. Cottrell, J. C. Fort, and G. Malgouyres, "Large deviations and rare events in the study of stochastic algorithms," *IEEE Trans. Automat. Contr.*, vol. AC-28, pp. 907–920, 1983.

[7] H. Cramér, *Mathematical Methods of Statistics*. Princeton, NJ: Princeton University Press, 1946.

[8] M. A. Crane and D. L. Iglehart, "Simulating stable stochastic systems, III: Regenerative processes and discrete event simulations," *Oper. Res.*, vol. 23, no. 1, pp. 33–45, 1975.

[9] J. B. Dugan, K. S. Trivedi, M. K. Smotherman, and R. M. Geist, "The hybrid automated reliability predictor," *J. Guidance, Contr., Dynamics* vol. 9, no. 3, pp. 319–331, 1986.

[10] B. L. Fox and P. W. Glynn, "Discrete-time conversion for simulating semi-Markov processes," *Oper. Res. Lett.*, vol. 5, pp. 191–196, 1986.

[11] ——, "Discrete-time conversion for finite-horizon Markov processes," *SIAM J. Appl. Math.*, vol. 5, no. 5, pp. 1457–1473, 1990.

[12] R. M. Geist and K. S. Trivedi, "Ultra-high reliability prediction for fault-tolerant computer systems," *IEEE Trans. Comput.*, vol. C-32, pp. 1118–1127, 1983.

[13] R. M. Geist and M. K. Smotherman, "Ultrahigh reliability estimates through simulation," in *Proc. Annu. Reliability and Maintainability Symp.*, Atlanta, GA, 1989, pp. 350–355.

[14] P. W. Glynn and P. Heidelberger, "Bias properties of budget constrained simulations," *Oper. Res.*, vol. 38, no. 5, pp. 801–814, 1990.

[15] P. W. Glynn and D. L. Iglehart, "Importance sampling for stochastic simulations," *Management Sci.*, vol. 35, no. 11, pp. 1367–1392, 1989.

[16] A. Goyal, W. C. Carter, E, de Souza e Silva, S. S. Lavenberg, and K. S. Trivedi, "The system availability estimator," in *Proc. Sixteenth Symp. Fault-Tolerant Comput.*, Vienna, Austria, 1986, pp. 84–89.

[17] A. Goyal, S. S. Lavenberg, and K. S. Trivedi, "Probabilistic modeling of computer system availability," *Ann. Oper. Res.*, vol. 8, pp. 285–306, 1987.

[18] A. Goyal and S. S. Lavenberg, "Modeling and analysis of computer system availability," *IBM J. Res. Develop.*, vol. 31, pp. 651–664, 1987.

[19] A. Goyal, P. Heidelberger, and P. Shahabuddin, "Measure specific dynamic importance sampling for availability simulations," in *1987 Winter Simulation Conf. Proc.*, A. Thesen, H. Grant, and W.D. Kelton, Eds., IEEE Press, 1987, pp. 351–357.

[20] D. Gross and D. R. Miller, "The randomization technique as a modeling tool and solution procedure for transient Markov processes," *Oper. Res.*, vol. 32, pp. 343–361, 1984.

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

[22] A. Hordijk, D. L. Iglehart, and R. Schassberger, "Discrete time methods for simulating continuous time Markov chains," *Adv. Appl. Prob.*, vol. 8, pp. 772–788, 1976.

[23] H. Kahn and A. W. Marshall, "Methods of reducing sample size in Monte Carlo computations," *J. Oper. Res. Soc. Amer.*, vol. 1, no. 5, pp. 263–278, 1953.

[24] S. Karlin and H. M. Taylor, *A First Course in Stochastic Processes*, Second Edition. New York: Academic, 1975.

[25] D. E. Knuth, "Big omicron, big omega and big theta," *SIGACT News (ACM)*, vol. 8, no. 2, pp. 18–24, 1976.

[26] S. S. Lavenberg, T. L. Moeller, and C. H. Sauer, "Concomitant control variables applied to the regenerative simulation of queueing systems," *Oper. Res.* vol. 27, no. 1, pp. 134–160, 1979.

[27] P. L'Ecuyer, "Efficient and portable combined random number generators," *Commun. ACM*, vol. 31, no. 6, pp. 742–774. 1988.

[28] E. L. Lehmann, *Theory of Point Estimation*. New York: Wiley, 1983.

[29] E. E. Lewis and F. Böhm, "Monte Carlo simulation of Markov unreliability models," *Nuclear Eng. and Design*, vol. 77, pp. 49–62, 1984.

[30] C. A. Liceaga and D. P. Siewiorek, "Toward automatic Markov reliability modelling of computer architectures," NASA Tech. Memo. 89009, 1986.

[31] S. V. Makam, A. Avizienis, and G. Grusas, *UCLA ARIES 82 User's Guide*. Also Tech. Rep. CSD-820830, Univ. of California at Los Angeles, 1982.

[32] A. M. Johnson, Jr. and M. Malek, "Survey of software tools for evaluating reliability, availability, and serviceability," *ACM Comput. Surveys*, vol. 20, no. 4, pp. 227–269, 1988.

[33] J. F. Meyer, "On evaluating the performability of degradable computing systems," *IEEE Trans. Comput.*, vol. C-29, no. 8, pp. 720–731, 1980.

[34] R. G. Miller, "The Jackknife—A review," *Biometrika*, vol. 61, pp. 1–15, 1974.

[35] R. R. Muntz, E. de Souza e Silva, and A. Goyal, "Bounding availability of repairable computer systems," *IEEE Trans. Comput.*, vol. 38, no. 12, pp. 1714–1723, 1989.

[36] V. F. Nicola, "Lumping in Markov reward processes," IBM Res. Rep. RC 14719, Yorktown Heights, NY, 1989.

[37] V. F. Nicola, M. K. Nakayama, P. Heidelberger, and A. Goyal, "Fast simulation of dependability models with general failure, repair and maintenance processes," in *Proc. Twentieth Symp. Fault-Tolerant Comput.*, Newcastle upon Tyne, England, 1990, pp. 491–498.

[38] S. M. Ross, *Applied Probability Models with Optimization Applications*. San Francisco, CA: Holden-Day, 1970.

[39] S. M. Ross and Z. Schechner, "Using simulation to estimate first passage distribution," *Management Sci.*, vol. 31, no. 2, pp. 224–234, 1985.

[40] S. Parekh and J. Walrand, "A quick simulation method for excessive backlogs in networks of queues," *IEEE Trans. Automat. Contr.*, vol. 34, no. 1, pp. 54–66, 1989.

[41] R. A. Sahner and K. S. Trivedi, "Performance and reliability analysis using acyclic directed graphs," *IEEE Trans. Software Eng.*, vol. SE-13, no. 10, pp. 1105–1114, 1987.

[42] W. H. Sanders and J. F. Meyer, "METASAN: A performability evaluation tool based on stochastic activity networks," in *Proc. 1986 Fall Joint Comput. Conf.*, AFIPS, NY, 1986, pp. 807–816.

[43] P. Shahabuddin, V. F. Nicola, P. Heidelberger, A. Goyal, and P. W. Glynn, "Variance reduction in mean time to failure simulations," in *1988 Winter Simulation Conf. Proc.*, M. A. Abrams, P. L. Haigh, and J. C. Comfort, Eds., IEEE Press, 1988, pp. 491–499.

[44] P. Shahabuddin, "Simulation and analysis of highly reliable systems," Ph.D. dissertation, Dep. Oper. Res., Stanford Univ., 1990.

[45] D. Siegmund, "Importance sampling in the Monte Carlo study of sequential tests," *Ann. Statistics*, vol. 4, pp. 673–684, 1976.

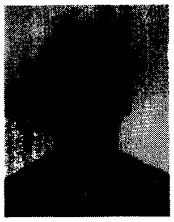
[46] W. L. Smith, "Regenerative stochastic processes," *Proc. Roy. Soc. A.*, vol. 232, pp. 6–31, 1955.

[47] J. Walrand, "Quick simulation of rare events in queueing networks," in *Proc. Second Int. Workshop Appl. Math. and Performance/Reliability Models of Comput./Commun. Syst.*, G. Iazeolla, P. J. Courtois, and O .J. Boxma, Eds. Amsterdam, North Holland, 1987, pp. 275–286.

Ambuj Goyal (S'81–M'82) received the B.Tech. degree from the Indian Institute of Technology, Kanpur, and the M.S. and Ph.D. degrees from the University of Texas at Austin,

He is a research staff member at the IBM Thomas J. Watson Research Center where he manages the Fault-Tolerant Systems Theory group. His research interests include fault-tolerant computing, computer architecture, communication networks, and performance evaluation.

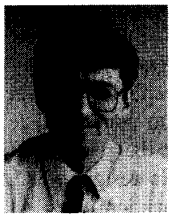
Dr. Goyal is a member of the IEEE Computer Society. He has served as the Program Committee Chairman of the Workshop on Reliability/Availability Modeling Tools and Their Applications.



Perwez Shahabuddin received the B.Tech. degree in mechanical engineering from the Indian Institute of Technology, Delhi, in 1984.

After working for a year in Engineers India Limited, India, he joined the Ph.D. program in Operations Research at Stanford University, Stanford, CA. After completing the Ph.D. degree in 1990, he joined the IBM T. J. Watson Research Center as a Research Staff Member. His research interests include simulation techniques, queueing theory, and performance analysis of computer systems.

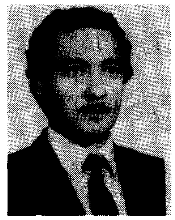
Dr. Shahabuddin is a member of the Operations Research Society of America.



Philip Heidelberger (M'82) received the B.A. degree in mathematics from Oberlin College, Oberlin, OH, in 1974 and the Ph.D. degree in operations research from Stanford University, Stanford, CA, in 1978.

He has been a Research Staff Member at the IBM Thomas J. Watson Research Center, Yorktown Heights, NY, since 1978. His research interests include modeling and analysis of computer performance and statistical analysis of simulation output.

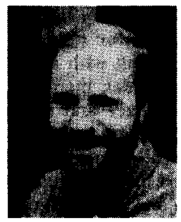
Dr. Heidelberger is an Associate Editor of *Operations Research* and was the Program Chairman of the 1989 Winter Simulation Conference, which is co-sponsored by the IEEE. He is a member of the Operations Research Society of America and the Association for Computing Machinery.



Victor F. Nicola (M'89) received the B.S. and M.S. degrees in electrical engineering from Cairo University, Egypt, and Eindhoven University of Technology, The Netherlands, respectively, and the Ph.D. degree in computer science from Duke University, Durham, NC, in 1986.

In 1979 he joined the scientific staff of the Measurement and Control Group at Eindhoven University. Currently, he is a Research Staff Member with the Department of Computer Sciences at IBM Thomas J. Watson Research Center. His research interests include performance and reliability modeling of computer systems, queueing theory, fault tolerance, and simulation.

Dr. Nicola is a member of the IEEE Computer Society.



Peter W. Glynn received the Ph.D. degree from Stanford University, Stanford, CA, in 1982.

He spent five years as an Assistant Professor in the Department of Industrial Engineering at the University of Wisconsin–Madison. He is currently an Associate Professor in the Department of Operations Research at Stanford University. His research interests include stochastic systems, computational probability, and simulation.