AD-A212 164

A UNIFIED FRAMEWORK FOR SIMULATING MARKOVIAN MODELS
OF HIGHLY DEPENDABLE SYSTEMS

by

Ambuj Goyal, Perwez Shahabuddin, Philip Heidelberger,
Victor F. Nicola, Peter W. Glynn

TECHNICAL REPORT No. 39

July 1989

DEPARTMENT OF OPERATIONS RESEARCH
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

DTIC
ELECTE
SEP 12 1989
S   B   D

DISCLAIMER NOTICE
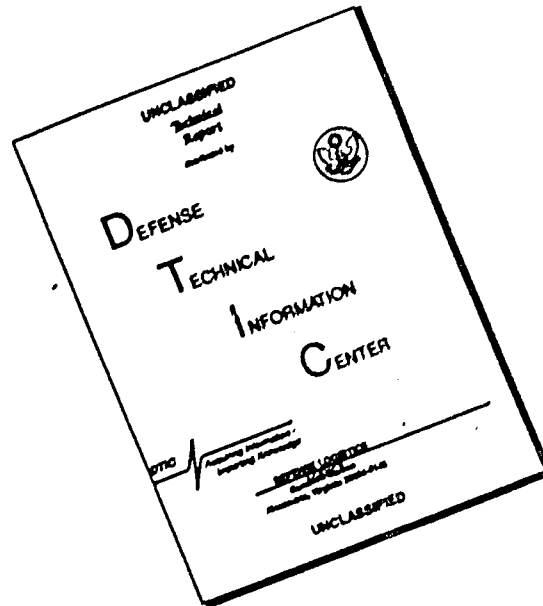
THIS DOCUMENT IS BEST QUALITY AVAILABLE. THE COPY FURNISHED TO DTIC CONTAINED A SIGNIFICANT NUMBER OF PAGES WHICH DO NOT REPRODUCE LEGIBLY.

A UNIFIED FRAMEWORK FOR SIMULATING MARKOVIAN

MODELS OF HIGHLY DEPENDABLE SYSTEMS

Ambuj Goyal *
Perwez Shahabuddin **[1]
Philip Heidelberger *
Victor F. Nicola *
Peter W. Glynn **[2]

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

** Department of Operations Research
Stanford University
Stanford, California 94305

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.

Keywords: simulation, reliability, importance sampling, conditioning, variance reduction

1. 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, 30], detailed fault and error handling models in [9] and detailed recovery hierarchies, operational and repair dependences in [15, 17]. Models which include degraded modes of operation have also been introduced [17, 40]. 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 [32]. Detailed surveys of these modeling techniques and the dependability measures calculated appear in [12, 31].

The most common stochastic models used in this context are continuous-time Markov chains (CTMCs). Typically, numerical methods are used to solve Markov chains. Although, many modeling packages have been built, e.g., [17] 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 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 [16,35] and state aggregation and bounding [1, 34] 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, 29]. 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 [17], 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 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 [20, 22] can be very effective in reducing the simulation run length significantly.

Importance sampling for rare event simulation has been used successfully in [6], [28], [42], [44] and [38]. 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], [42] and [44, 38] 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 [28] to select importance sampling distributions for reliability estimation in large models of machine repairman type.

In this paper, we review and extend the methods in [4], [18] and [41] 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 [17]. 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 2. We also give formal definitions of the dependability measures of interest in this section.

In Section 3 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 [18]) 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 a regenerative simulation), then significant variance reductions can be achieved using our importance sampling techniques (see also [41].) 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 3. The equations for optimal run-length allocation and asymptotic bias expansions are also given.

In Section 4 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., [28]) methods with importance sampling. In Section 5 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 6 we show the effectiveness of the above techniques in reducing the simulation run time of a large example. Typically, we obtain orders of magnitude reduction in the run time over standard simulation. 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 7 we give concluding remarks and suggest some directions for future research.

2. 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 non-specialist. 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.

2.1 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., [20]). 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) $f_\theta(x)$ 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 $f_\theta(x)$. Then

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

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

$$r(\theta) = \int_{-\infty}^{\infty} h(x) \left(\frac{f_\theta(x)}{p_{\theta_0}(x)} \right) p_{\theta_0}(x)\, dx = E_{\theta_0}[h(X)\, L(\theta, \theta_0, X)]. \tag{2.2}$$

where $L(\theta, \theta_0, x) \equiv f_\theta(x)/p_{\theta_0}(x)$ 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 $p_{\theta_0}(x) = h(x) f_\theta(x)/r(\theta)$ yields a zero variance estimator, since in this case the rv $h(X)\, L(\theta, \theta_0, X)$ 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.

We next specialize these results to rare event simulation, and show why importance sampling can be particularly effective for estimating the probability of rare events (this argument is basically taken from [44]). Suppose that $h(x) = g(x) \times 1_{\{x \in F\}}$ for some function g where $1_{\{x \in F\}}$ is the indicator of a set F, i.e., $1_{\{x \in F\}} = 1$ if $x \in F$ and $1_{\{x \in F\}} = 0$ if $x \notin F$. We assume that F is a so-called rare set in the sense that the probability that the rv $X \in F$ is close to 0. A single sample using standard simulation, i.e., sampling X from the pdf $f_\theta(x)$ and forming the estimate $g(X) \times 1_{\{X \in F\}}$, has variance $V(\theta) = E_\theta[g(X)^2 1_{\{X \in F\}}] - r(\theta)^2$ (note that $1^2_{\{X \in F\}} = 1_{\{X \in F\}}$). On the other hand, the variance obtained using importance sampling, i.e., sampling X from the pdf $p_{\theta_0}(x)$ and forming the estimate $g(X) \times 1_{\{X \in F\}} \times L(\theta, \theta_0, X)$ is $V(\theta_0) = E_{\theta_0}[g(X)^2 1_{\{X \in F\}} L(\theta, \theta_0, X)^2] - r(\theta)^2$. Notice that if the likelihood ratio $L(\theta, \theta_0, x) \ll 1$ for all $x \in F$, then

$$
\begin{aligned}
V(\theta_0) &\ll E_{\theta_0}[g(X)^2 1_{\{X \in F\}} L(\theta, \theta_0, X)] - r(\theta)^2 \\
&= E_\theta[g(X)^2 1_{\{X \in F\}}] - r(\theta)^2 = V(\theta)
\end{aligned}
\tag{2.3}
$$

where the first equality is true by Equation 2.2. Thus to achieve variance reduction, we need to make the likelihood ratio very small on the rare set F. Since F is a rare set, $f_\theta(x)$ is typically small for $x \in F$ and since $L(\theta, \theta_0, x) = f_\theta(x)/p_{\theta_0}(x)$, we obtain a small likelihood ratio by choosing the parameter θ_0 so that $p_{\theta_0}(x)$ is relatively large for $x \in F$, i.e., by making the rare set more likely to occur under the new measure defined by $p_{\theta_0}(x)$. This intuitively explains the basic notion behind using importance sampling in rare event simulation.

2.2 Markov Chains and Associated Dependability Measures

We assume that the system can be represented by a CTMC $Y = \{ Y_s, s \geq 0 \}$ with finite state space E and infinitesimal generator matrix $Q = \{ q(i,j), i, j \in E \}$. We let $q(i) = - q(i,i)$ denote the total rate out of state i (see, e.g., [23]). We further assume that E can be partitioned into two subsets: $E = O \cup F$ 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 Y: transient measures and so-called steady-state measures. Considering the transient measures first, the interval availability, $A(t)$, is defined by

$$A(t) = \frac{1}{t} \int_{s=0}^{t} 1_{\{Y_s \in O\}} \, ds \, . \tag{2.4}$$

This is the fraction of time that the system is operational in the time interval $(0,t)$. We let

$$I(t) = E[A(t)] \tag{2.5}$$

be the expected interval availability and let

$$F(t,x) = P\{A(t) \le x\} \tag{2.6}$$

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 $(0,t)$:

$$R(t) = P\{\alpha_F > t\} = E[1_{\{\alpha_F > t\}}]. \tag{2.7}$$

For steady-state measures we assume that Y is irreducible, in which case $Y_s \Rightarrow Y$ as $s \to \infty$, where $\Rightarrow$ denotes convergence in distribution and Y is a rv having the steady-state distribution $\underline{\pi} = \{ \pi_i, i \in E \}$ ($\underline{\pi}$ solves the equations $\underline{\pi} Q = 0$). 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 [43] or [8]), steady-state measures take the form of a ratio of two expected values:

$$r \equiv E[f(Y)] = \lim_{t \to \infty} E[f(Y_t)] = \lim_{t \to \infty} \frac{1}{t} \int_{s=0}^{t} f(Y_s) ds = \frac{E[\int_{s=0}^{\alpha_0} f(Y_s) ds]}{E[\alpha_0]} \tag{2.8}$$

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. A ratio representation for $E[\alpha_F]$ is found to be particularly useful. To derive this ratio, we write

$$\tau_F = \tau_I 1_{\{\tau_I < \tau_0\}} + (\tau_0 + (\tau_I - \tau_0)) 1_{\{\tau_0 < \tau_I\}}$$
$$= \min(\tau_I, \tau_0) + (\tau_I - \tau_0) 1_{\{\tau_0 < \tau_I\}}. \tag{2.9}$$

Now, applying the Markov property at time τ_0 shows that, on the set $\{\tau_0 < \tau_I\}$, $(\tau_I - \tau_0)$ is conditionally independent of $1_{\{\tau_0 < \tau_I\}}$ and furthermore has the same distribution as τ_I. Therefore, taking expected values of Equation 2.9 and rearranging terms yields the ratio formula

$$E[\tau_I] = \frac{E[\min(\tau_I, \tau_0)]}{P\{\tau_I < \tau_0\}}. \tag{2.10}$$

Thus we can view estimating $E[\tau_F]$ as a ratio estimation problem, where both the numerator and the denominator are estimated using a regenerative simulation. Therefore, in Section 3 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.

3. Estimating Steady-State Measures

In this section, we discuss the estimation of steady-state measures of CTMCs. We begin by reviewing how this problem can be reduced to estimating associated steady-state measures in discrete-time Markov chains (DTMCs) and describe the regenerative method of simulation (see, e.g., [8]). We next describe the application of importance sampling to DTMCs. 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.

3.1 Discrete Time Conversion of CTMCs

In [21] 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 $X = \{X_n, n \geq 0\}$ denote the embedded DTMC of the CTMC Y: X has transition matrix $P = \{p(i,j)\}$ 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

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

To emphasize the dependence of this ratio on the transition matrix P, we write Equation 3.1 explicitly as $r = E_P[G]/E_P[H]$ where $G = \sum_{k=0}^{\tau_0-1} g(X_k)$ and $H = \sum_{k=0}^{\tau_0-1} h(X_k)$. In [21], 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 CTMCs (see, e.g., [8]). One simulates (using the transition matrix $\mathbf{P}$) m (say) independent and identically distributed (iid) replicates of the random vector (G, H), yielding the iid random vectors $\{(G_j, H_j): j = 1, \ldots, m\}$. Each replication involves simulating the DTMC $\mathbf{X}$ (with the initial condition $X_0 = 0$) to time τ_0; these replications are known, in the simulation literature, as regenerative cycles. Let $r_m(\mathbf{P}) = \sum_{j=1}^{m} G_j / \sum_{j=1}^{m} H_j$. Then, because the cycles are iid, $\lim_{m \to \infty} r_m(\mathbf{P}) = r$ with probability one and $\sqrt{m}(r_m(\mathbf{P}) - r) \Rightarrow N(0, \sigma^2(\mathbf{P})/E_{\mathbf{P}}[H_j]^2)$, where $N(0, \sigma^2)$ denotes a normally distributed random variable with mean zero and variance σ^2, and $\sigma^2(\mathbf{P}) = \mathrm{Var}_{\mathbf{P}}[G_j - rH_j]$.

3.2 Importance Sampling for DTMCs

We next extend the change of measure transformation of Equation 2.1 to DTMCs. Let τ be any stopping time of the DTMC $\mathbf{X}$ and let Y be a rv defined on $\mathbf{X}$ up until time τ. Informally, τ is a stopping time if the event $\{\tau = n\}$ is determined by $X_n = (X_0, \ldots, X_n)$. The rv Y is then a (measurable) function of $X_\tau = (X_0, \ldots, X_\tau)$ (see [23] 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_j are stopping times. Let A_n denote the set of all possible sample paths up until time n, i.e., $A_n = \{s_n = (s_0, \ldots, s_n): s_j \in E\}$. For any $s_n \in A_n$, let

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

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 $\mathbf{P}$, is finite with probability one and let Y be a (measurable) function of X_τ for which $E_{\mathbf{P}}[|Y(X_\tau)|] < \infty$. Let $\mathbf{P}'$ be any other measure such that, under $\mathbf{P}'$, τ is finite with probability one and for any $s_n \in B_n$, $\mathbf{P}'(s_n) \neq 0$ whenever $Y(s_n)\mathbf{P}(s_n) \neq 0$. Then

$$E_{\mathbf{P}}[Y(X_\tau)] = E_{\mathbf{P}'}[Y(X_\tau) L'_\tau(X_\tau)] \tag{3.3}$$

where for any n, $L'_\tau(X_n) = \mathbf{P}(X_n)/\mathbf{P}'(X_n)$.

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

$$\mathrm{E_P}[Z(X_\tau)] = \sum_{n=0}^{\infty} \sum_{s_n \in B_n} \mathbf{P}(s_n)\, Z(s_n) = \sum_{n=0}^{\infty} \sum_{s_n \in B_n} \mathbf{P}'(s_n)\, Z(s_n)\, L'_1(s_n) = \mathrm{E_{P'}}[Z(X_\tau)L'_1(X_\tau)] \quad (3.4)$$

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

Versions of this proposition have appeared elsewhere, e.g., in [44], [38] or [14]. Note that there can exist a sample path s_n such that $\mathbf{P}(s_n) > 0$ even though $\mathbf{P}'(s_n) = 0$, provided that $Z(s_n) = 0$. We emphasize however that the measure $\mathbf{P}'$ 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 $\mathbf{P}'$ not to be Markovian. The general form that we will consider for $\mathbf{P}'$ is

$$\mathbf{P}'(s_n) = \mathbf{P}'(s_0)\, \mathbf{P}'(s_1 \mid s_0)\, \mathbf{P}'(s_2 \mid s_0, s_1) \dots \mathbf{P}'(s_n \mid s_0, \dots, s_{n-1}). \quad (3.5)$$

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 Equation 3.1 yields the following procedure. A total of m iid regenerative cycles of the DTMC X are simulated using the DIS distribution $\mathbf{P}'$. 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 $\hat{r}_m(\mathbf{P}') = \sum_{j=1}^{m} G_j L'_{1j} / \sum_{j=1}^{m} H_j L'_{1j}$. Then, as in the case without importance sampling, we have $\lim_{m \to \infty} \hat{r}_m(\mathbf{P}') = r$ with probability one and $\sqrt{m}\,(\hat{r}_m(\mathbf{P}') - r) \Rightarrow N(0, \sigma^2(\mathbf{P}')/\mathrm{E_P}[H_j]^2)$ where

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

Now from the form of $\sigma^2(\mathbf{P}')$, it is seen that selecting a good DIS distribution $\mathbf{P}'$ involves taking three terms into consideration. For example, selecting a $\mathbf{P}'$ 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 trade-off.

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 denomi-

nator 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 3.3 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 $\mathbf{P}'$ and $\mathbf{P}''$ 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

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

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

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

The optimal run length allocation between the numerator and the denominator will be considered in Section 3.4.

3.3 A Three State Example

In this section, we consider a simple availability example, namely a three state Birth and Death process (see [23]). 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 $\lambda_i, i = 0, 1$ and the death rates are $\mu_i, i = 1, 2$. 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 non-zero entries: $p(0,1) = p(2,1) = 1$, $p(1,2) = \varepsilon \equiv \lambda_1/(\lambda_1 + \mu_1)$ and $p(1,0) = (1 - \varepsilon)$. 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 $h_0 = \Theta(1/\varepsilon)$, $h_1 = \Theta(1)$ and $h_2 = \Theta(1)$ (we follow Knuth's [24] 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 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, $P\{n_f = i\} = (1 - \varepsilon)\varepsilon^i$ for $i \geq 0$, so that $E_P[n_f] = \varepsilon/(1 - \varepsilon)$ and $Var_P[n_f] = \varepsilon/(1 - \varepsilon)^2$. Thus $E_P[G] = h_2\varepsilon/(1 - \varepsilon)$ and $E_P[H] = (h_0 + h_1) + (h_1 + h_2)\varepsilon/(1 - \varepsilon)$. Straightforward calculations show that $r = \Theta(\varepsilon^2)$ and that the asymptotic squared coefficient of variation of $\hat{r}_m(P)$ (obtained from the central limit theorem) is

$$\frac{Var_P[G - rH]}{m r^2 E_P[H]^2} = \Theta(1/m\varepsilon). \tag{3.9}$$

The dominant term in Equation 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 $m\varepsilon \to \infty$. This demonstrates the potentially large sample size required for rare event simulations (in which $\varepsilon \approx 0$).

Let $P(s_i)$ be the probability of a regeneration sample path s_i, then $P(s_i) = (1 - \varepsilon)\varepsilon^i$, $i \geq 0$.

The optimal zero variance importance sampling distribution $P'(s_i)$, $i \geq 0$, for estimating $E_P[G]$ is computed from explicit enumeration of all sample paths. First, we write $E_P[G] = \sum_{\forall i} G(s_i)P(s_i) = \sum_{\forall i}[G(s_i)L(s_i)]P'(s_i) = E_{P'}[G L]$, with $L(s_i) = P(s_i)/P'(s_i)$. Now, the optimal $P'(s_i)$, for all $i \geq 0$, can be computed (similar to what is described in Section 2.1):

$$\mathbf{P'}^*(\mathbf{s}_i) = \frac{\mathbf{P}(\mathbf{s}_i)G(\mathbf{s}_i)}{\mathbb{E}_P[G]} = (1 - \varepsilon)^2 \varepsilon^{i-1} i, \quad i \ge 0, \tag{3.10}$$

since then $G(\mathbf{s}_i)/L(\mathbf{s}_i)$, for all $i \ge 0$, is a constant equal to $\mathbb{E}_P[G]$.

Similarly, the optimal zero variance importance sampling distribution for estimating $\mathbb{E}_P[H]$ is given by

$$\mathbf{P''}^*(\mathbf{s}_i) = \frac{\mathbf{P}(\mathbf{s}_i)H(\mathbf{s}_i)}{\mathbb{E}_P[H]} = \frac{\left[(h_0 + h_1) + (h_1 + h_2) i \right] (1 - \varepsilon)^2 \varepsilon^i}{h_0 + h_1 - \varepsilon (h_0 - h_2)}, \quad i \ge 0. \tag{3.11}$$

From Equation 3.10, $\mathbf{P'}^*(\mathbf{s}_0) = 0$, $\mathbf{P'}^*(\mathbf{s}_1) = (1 - \varepsilon)^2$, $\mathbf{P'}^*(\mathbf{s}_2) = 2\varepsilon(1 - \varepsilon)^2$ and so on. Now let $p'^*(1,0|n_F = i)$ 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 $p'^*(1,0|n_F = i) = \mathbf{P'}^*(\mathbf{s}_i)/(\sum_{j \ge i} \mathbf{P'}^*(\mathbf{s}_j))$ and thus, from Equation 3.10, $p'^*(1,0|n_F = 0) = 0$ and

$$p'^*(1,0|n_F = i) = \frac{(1 - \varepsilon)^2}{1 - \varepsilon (i - 1) / i}, \quad i \ge 1. \tag{3.12}$$

Therefore, each successive time the simulation enters state 1, the probability of returning to state 0 changes (under both $\mathbf{P'}^*(\mathbf{s})$ and $\mathbf{P''}^*(\mathbf{s})$). Thus the optimal changes of measure for both the numerator and the denominator of Equation 2.1 are dynamic. In particular note that while $p'^*(1,0|n_F = 0) = 0$, $p'^*(1,0|n_F = 1) = (1 - \varepsilon)^2 \approx (1 - 2\varepsilon) \approx (1 - \varepsilon) = p(1,0)$ for $\varepsilon \approx 0$. Also, $\lim_{i \to \infty} p'^*(1,0|n_F = i) = (1 - \varepsilon) = p(1,0)$. 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 2.1 as well as Walrand's suggestion in [44, 38] (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 Equation 3.11 has

$$p''^*(\mathbf{s}_0) = p''^*(1,0|n_F = 0) = \frac{(1 - \varepsilon)^2}{1 - \varepsilon (h_0 - h_2) / (h_0 + h_1)} \approx (1 - \varepsilon) = p(1,0) \tag{3.13}$$

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

3.4 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

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

where $\sigma_n^2 = \mathrm{Var}_{\mathbf{P}'}[G_j L'_{1j}]$ and $\sigma_d^2 = \mathrm{Var}_{\mathbf{P}''}[H_j L''_{1j}]$. This result is obtained by applying results from renewal theory (see [43]). 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 Equation 2.10, 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 5, 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.

3.5 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. References for this type of bias expansion may be found in Section 27 of [7], Chapter 2 of [27] or [13]. They are derived using Taylor series expansions and multidimensional central limit theorems. Let $\{C_n = (C_n(1), \ldots, C_n(d)), n \geq 1\}$ be a sequence of iid vectors of length d and let $\mu = (\mu_1, \ldots, \mu_d)$ where $E[C_n] = \mu$. Suppose we are interested in estimating $g(\mu)$ for some function g. In the case of ratio estimation $\mu = (\mu_1, \mu_2)$ and $g(\mu) = \mu_1/\mu_2$. Let $\overline{C}_m = (1/m) \sum_{n=1}^{m} C_n$. Then, under appropriate technical conditions on g and the moments of C_n,

$$E[g(\overline{C}_m)] = g(\mu) + \frac{1}{2m} \sum_{i=1}^{d} \sum_{j=1}^{d} \sigma_{ij} g_{ij} + o(1/m) \qquad (3.15)$$

where $\sigma_{ij} = \mathrm{Cov}[C_n(i), C_n(j)]$ and $g_{ij} = \frac{\partial^2}{\partial x_i \partial x_j} g(x) |_{x=\mu}$. In our case, $C_n = (G_n L'_{1n}, H_n L''_{1n})$, $\sigma_{11} = \mathrm{Var}_{P'}[G_n L'_{1n}]$, $\sigma_{22} = \mathrm{Var}_{P''}[H_n L''_{1n}]$ 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 $g_{11} = 0$, $g_{22} = 2\mu_1/\mu_2^3 = 2r/\mu_2^2$ and $g_{12} = -1/\mu_2^2$. Since $\sigma_{12} = 0$, the value of g_{12} does not enter into the MSDIS bias expansion. Therefore

$$E[g(\overline{C}_m)] = g(\mu) + \frac{r \mathrm{Var}_{P''}[H_n L''_{1n}]}{m \mu_2^2} + o(1/m). \qquad (3.16)$$

For the measures of interest in availability modeling, $r \geq 0$, $H_n \geq 0$ and $\mu_2 > 0$ so that, asymptotically, $E[g(\overline{C}_m)] \geq g(\mu)$. Furthermore, this asymptotic bias expansion is independent of the importance sampling distribution P' chosen for simulation of the numerator.

For the steady-state unavailability we select $P'' = P$. By the results of Section 3.3, in the three state example $r = \Theta(\epsilon^2)$, $\mathrm{Var}_P[H_n] = \Theta(\epsilon)$ and $\mu_2 = E_P[H_n] = \Theta(1/\epsilon)$. 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

$$E[g(\overline{G}_m, \overline{H}_m)] = g(\mu) + \frac{r \mathrm{Var}_P[H_n]}{m \mu_2^2} - \frac{\mathrm{Cov}_P[G_n, H_n]}{m \mu_2^2} + o(1/m). \qquad (3.17)$$

For the three state example, $\text{Cov}_P[G_n, H_n] = h_2(h_1 + h_2)\text{Var}_P[n_j] = \Theta(\varepsilon)$ so the dominant term in the bias expansion is $\Theta(\varepsilon^3/m)$ which is significantly larger than the $\Theta(\varepsilon^5/m)$ bias obtained using MSDIS. Moreover, using standard simulation, the relative bias (bias/r) is only $\Theta(\varepsilon/m)$. These observations are consistent with the experimental results described in Section 6.

For the MTTF, r is large and $\mu_2 = P\{\alpha_T < \alpha_0\}$ is small, which potentially makes the leading bias term large. However, as seen from Equation 3.16, choosing an importance sampling distribution P'' for the purpose of reducing variance also has the beneficial effect of reducing bias.

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 non-negligible way so that, e.g., $E[g(\overline{C}_m)] \leq g(\mu)$. 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 [33]).

4. Estimating Transient Measures

Simulation of the CTMC Y consists of two parts: simulating the sequence of states visited by the embedded DTMC X with transition matrix 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 $t_n = (t_0, \dots, t_n)$ denote the first $n + 1$ holding times of the CTMC. Given that $X_n = (X_0, \dots, X_n)$, the likelihood of t_n is therefore

$$f(t_n \mid X_n) \equiv q(X_0)\, e^{-q(X_0)t_0} \dots q(X_n)\, e^{-q(X_n)t_n} \tag{4.1}$$

and thus the likelihood of the sample path (X_n, t_n) is

$$Q(X_n, t_n) \equiv P(X_n)\, f(t_n \mid X_n) \tag{4.2}$$

where $P(X_n)$ was defined in Equation 3.2. Equation 4.2 gives the likelihood at the times of the jumps of the embedded DTMC.

We basically adapt the development in [14] in order to extend Proposition 3.1. Define $T_0 = 0$ and $T_n = t_0 + \dots + t_{n-1}$ for $n \geq 1$. Then T_n is the time at which state X_n is entered, i.e., the time of the n'th transition. Let $Y_t = (Y_s, 0 \leq s \leq t)$. 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 $(X_0, t_0), \dots, (X_n, t_n)$. We let Q' denote another measure for generating sequences $\{(X_n, t_n), n \geq 0\}$. We will specifically assume that $Q'(X_n, t_n) = P'(X_n) f'(t_n \mid X_n)$. With this factorization, the form of the contribution of the holding times to the likelihood, $f'(t_n \mid X_n)$, 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 Y for which $\tau = n$.

Proposition 4.1:
Let τ be an integer valued stopping time which, under Q, is finite with probability one. Define $\alpha \equiv T_\tau$ and let Z be a (measurable) function of Y_α for which $E_Q[|Z(Y_\alpha)|] < \infty$. Let Q' be another measure of the form $Q'(X_n, t_n) = P'(X_n) f'(t_n \mid X_n)$ such that, under P', τ is finite with probability one and for any $(s_n, t_n) \in B_n$, $P'(s_n) f'(t_n \mid s_n) \neq 0$ whenever $Z(Y_{T_n})P(s_n) f(t_n \mid s_n) \neq 0$. Then

$$E_Q[Z(Y_\alpha)] = E_{Q'}[Z(Y_\alpha)\, L'_1(X_\tau)\, L'_2(X_\tau, t_\tau)] \tag{4.3}$$

where for any n, $L'_2(X_n, t_n) = f(t_n \mid s_n)/f'(t_n \mid s_n)$.

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., $\alpha = \tau_F$. 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 $Z = 1_{\{T_\tau \leq t\}}$.

4.1 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 $L'_1(X_\tau)$) and the second is for the state holding times given the DTMC's sample path (corresponding to $L'_2(X_\tau, t_\tau)$). Lewis and Böhm [28] 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 towards 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

$$f(t_{n+1} \mid X_n, t_n) = \frac{\lambda_0 e^{-\lambda_0 t_{n+1}}}{1 - e^{-\lambda_0(t - T_n)}} , \qquad 0 \leq t_{n+1} \leq t - T_n, \qquad (4.4)$$

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 [37] 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 $Z = P\{T_{\tau_F} \leq t \mid X_{\tau_F}\}$. 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 [39]. 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 non-zero states are randomly sampled, but no holding times are sampled for state 0. Let β denote the total holding time in states other than 0: $\beta = \sum_{k=0}^{\tau_F} t_k \times 1_{\{X_k \neq 0\}}$. Let n_0 denote the number of visits to state 0 and let γ be a rv

denoting the total holding time in state 0. Given n_0, y has an Erlang distribution with shape parameter n_0 and scale parameter λ_0 and we write $P\{y \leq s | n_0\} = E_{n_0}(s, \lambda_0)$. We then set

$$Z = P\{\tau_F \leq t | X_{\tau_F}, \beta\} = P\{y \leq t - \beta | n_0\} = E_{n_0}(t - \beta, \lambda_0) . \tag{4.5}$$

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 $\mathrm{Var}[E[X | Y]] \leq \mathrm{Var}[X]$, see, e.g., page 12 of [36]), the conditional approach plus failure biasing is always guaranteed to reduce the variance over just failure biasing. To see this, notice that

$$\mathrm{Var}_{P'}[L'_1(X_{\tau_F}) 1_{\{\tau_F \leq t\}}] \geq \mathrm{Var}_{P'}[E[L'_1(X_{\tau_F}) 1_{\{\tau_F \leq t\}} | X_{\tau_F}, \beta]]$$
$$= \mathrm{Var}_{P'}[L'_1(X_{\tau_F}) E_{n_0}(t - \beta, \lambda_0)] . \tag{4.6}$$

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 [19]). A discussion of approaches to using uniformization in simulations, including discrete conversions, may be found in [foxglynn1]. 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.

4.2 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 $r(t) = E[\int_{s=0}^{t} f(Y_s) ds]$. 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 [28], uses failure biasing and forcing as described in Section 4.1. 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 I visits to state 0 at some value of I for which $F_I(t, \lambda_0)$ 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]:

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

where $g(i) = f(i)/q(i)$ and the expectation is with respect to the transition matrix $\mathbf{P}$. 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 Equation 4.7 would also require importance sampling on the number of state 0 pseudo transitions in order to be effective. To combine forcing with Equation 4.7, we write

$$r(t) \;=\; E[\; \sum_{k=0}^{N(t)-1} g(X_k) \,|\, t_0 > t\,]\, P\{t_0 > t\} \;+\; E[\; \sum_{k=0}^{N(t)-1} g(X_k) \,|\, t_0 \leq t\,]\, P\{t_0 \leq t\}$$

$$=\; E[\; \sum_{k=0}^{N(t)-1} g(X_k) \,|\, t_0 \leq t\,]\, P\{t_0 \leq t\} \qquad\qquad (4.8)$$

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 Equation 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 4.1, define $n_0(k) = \sum_{j=0}^{k} 1_{\{X_j = 0\}}$ and $\beta_k = \sum_{j=0}^{k} t_j 1_{\{X_j \neq 0\}}$. With these definitions, $n_0(k)$ is the number of visits to state 0 and β_k is the total holding time in the non-zero states at the k'th transition. Then, by Equation 4.7

$$r(t) \;=\; E\Big[\sum_{k=0}^{\infty} g(X_k)\, 1_{\{k \le N(t)-1\}}\Big] \;=\; \sum_{k=0}^{\infty} E\big[g(X_k)\, 1_{\{N(t)\ge k+1\}}\big]$$

$$\;=\; \sum_{k=0}^{\infty} E\big[\, E[g(X_k)\, 1_{\{N(t)\ge k+1\}} \mid X_k, \beta_k]\,\big] \;=\; E\Big[\sum_{k=0}^{\infty} g(X_k)\, E_{n_0(k)}(t - \beta_k, \lambda_0)\Big] \;. \tag{4.9}$$

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

To apply Equation 4.9 requires determining a stopping criterion. We could simulate until $\beta_k \ge t$ at which point $E_{n_0(k)}(t - \beta_k, \lambda_0) = 0$. However, since repairs are fast, β_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 L'th visit to state 0, begin sampling the state 0 holding times and adding them to β_k. Very quickly, β_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 L'th time. For $k \ge N_0(L) + 1$, let $\widetilde{\beta}_k = \beta_{N_0(L)} + \sum_{j=N_0(L)+1}^{k} t_j$. Then, arguing as above,

$$r(t) \;=\; E\Big[\sum_{k=0}^{N_0(L)} g(X_k)\, E_{n_0(k)}(t - \beta_k, \lambda_0)\Big] \;+\; E\Big[\sum_{k=N_0(L)+1}^{\infty} g(X_k)\, E_L(t - \widetilde{\beta}_k, \lambda_0)\Big] \;. \tag{4.10}$$

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

5. 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 [15, 17] 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 3.1). 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 $L'_1(X_\tau)$ and $L'_2(X_\tau, t_\tau)$. 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, $L'_2(X_\tau, t_\tau)$ 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 3.3, 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 [28]. We assign a combined probability *bias*1 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 6. We call this the *Bias1 Balancing* method. The second is to give a higher combined probability *bias*2 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 6. We call this the *Bias1 Bias2* method.

For the steady-state availability each regenerative cycle corresponds to a sample. We use either the DIS or the MSDIS method given in Section 3.2 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.

6. 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 Figure 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.
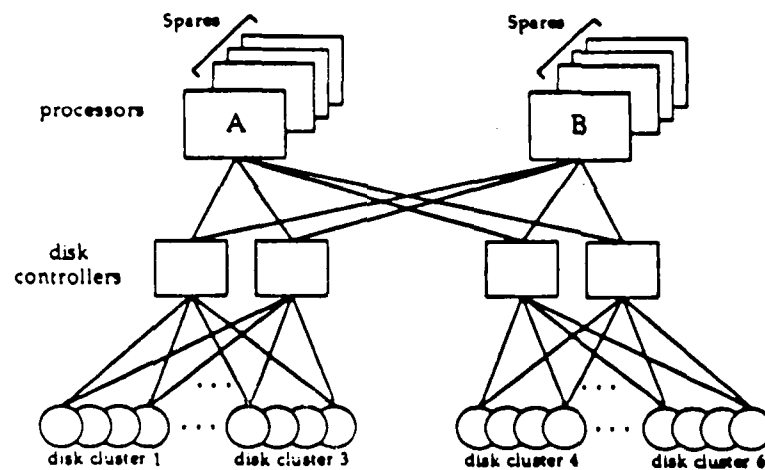
Figure 1. A block diagram of the computing system modeled.

For a balanced system we select two sets of processors with 2 processors per set, two sets of controllers with 2 controllers per set, and 6 clusters of disks, each consisting of 4 disk units. In a disk cluster, data is 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 Figure 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 is 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.

6.1. 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 (non-simulation) results for these measures were obtained using the SAVE package [17]. 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 [34]. 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 3. 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 1 and Table 2 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 1 and 1,000,000 events for each entry in Table2, respectively. For the MSDIS entries, we assigned 10% of the total events to estimate the denominator (numerator) for unavailability (MTTF) as suggested in Section 3.4. Based on impirical results obtained in [4], [18] and [41], the values for $bias1$ and $bias2$ were selected as follows: for DIS, .5 and .5, and for MSDIS, .9 and .9.

There are a few important points to note in the tables. For the balanced system, the $Bias1/Bias2$ method is most effective, which supports our intuition that it helps push the system quickly towards a likely path to failure. For the unbalanced system, the $Bisa1/Balancing$ 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.

Also 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 2, $\min(\alpha_f, \alpha_0) = \alpha_0$ with high probability when no importance sampling is used. Thus an individual sample r.v. in the numerator of the ratio for MTTF (Equation 2.10) is equal to the r.v. in the denominator of U (Equation 2.8) 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 (Equation 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 so called coverage experiments (see e.g., [25]) to determine the validity of the confidence intervals that are formed based on the asymptotic central limit theorems described in Section 3. 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., [25]). In such cases, the variance reduction technique cannot be relied upon to actually shorten simulation run lengths.

We performed 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 we 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 $\hat{U}_1, \ldots, \hat{U}_R$ 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. The results are listed in Table 3. Note that, as anticipated from Section 3.5, 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 only 81%.

This dip in coverage concerned us since, typically, coverages improve with run length. There are a number of possible explanations for this phenomenon: a random fluctuation, numerical errors in U or $\{\hat{U}_i, \hat{\sigma}_i\}$, problems with the random number generator, non-monotone convergence to normality, problems with the importance sampling scheme, or possibly a subtle bug in the computer codes (although the code has produced valid estimates for all examples tested). We recompiled the numerical solver and simulator to quadruple precision (from double) and obtained essentially identical results. We then changed random number generators from the multiplicative linear congruential generator $I_{n+1} = (I_n \times 16807) \bmod 2^{31} - 1$ to the combined generator described in [26]. With $R = 200$ replications, the coverage increased to 85% which still represents a statistically significant departure from 90% (although its confidence interval overlays with that of the 81% coverage). Thus, at present, and despite considerable effort, we have been unable to identify the source of this slight coverage problem. Since our importance sampling methods are designed to move towards the most likely failure states, it may be that we are missing secondary failure modes which would become significant at such high levels of precision. Note that the first non-zero digit in U is in the sixth decimal place, so the problem is occurring in the eight decimal place. Furthermore, in practice, such high precision is probably not warranted given inaccuracies in model parameters and distributional assumptions.

6.2. 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 towards the set of system failed states F, 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 towards the system failed set, in addition to another independent technique (e.g., forcing or conditioning as discussed in Section 5) 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 4 and 5.

For all methods, we notice that 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 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 Equation 4.7 with conditioning, but not with forcing. However, forcing can be similarly combined with Equation 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. This is intuitively reasonable because these methods correspond to importance sampling in the embedded Markov chain; this

sampling is independent of forcing and conditioning. For unbalanced systems, we expect that *Bias/ Balancing* will be consistently better than the other two methods; this is supported by a preliminary empirical investigation in progress.

7. 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. 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 non-exponential. 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. Nevertheless, it is likely that the failure biasing ideas that worked successfully in the Markovian setting can be suitably generalized to obtain effective sampling heuristics for non-Markovian models. However, it appears less likely that the conditioning techniques of this paper can be applied to non-Markovian models; our conditioning technique depends critically on the exponential nature of the holding times in Markovian models.

References

[1] Bobbio, A. and Trivedi, K. S. (1986). An Aggregation Technique for the Transient Analysis of Stiff Markov Chains. *IEEE Transactions on Computers* C-35, 803-814.

[2] Bouricius, W.G., Carter, W.C. and Schneider, P.R. (1969). Reliability Modeling Techniques for Self-Repairing Computer Systems. *Proceedings of ACM National Conference*. San Francisco, California, 295-309.

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

[4] Conway, A.E. and Goyal, A. (1987). Monte Carlo Simulation of Computer System Availability Reliability Models. *Proceedings of the Seventeenth Symposium on Fault-Tolerant Computing*. Pittsburgh, Pennsylvania, 230-235.

[5] Costes, A., Doucet, J.E., Landrault, C. and Laprie, J.C. (1981). SURF- A Program for Dependability Evaluation of Complex Fault-Tolerant Computing Systems. *Proceedings of the Eleventh Symposium on Fault-Tolerant Computing*. Portland, Maine, 72-78.

[6] Cottrell, M., Fort, J.C. and Malgouyres, G. (1983). Large Deviations and Rare Events in the Study of Stochastic Algorithms. *IEEE Transactions on Automatic Control* AC-28, 907-920.

[7] Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press, Princeton, New Jersey.

[8] Crane, M.A. and Iglehart, D.L. (1975). Simulating Stable Stochastic Systems, III: Regenerative Processes and Discrete Event Simulations. *Operations Research* 23, 33-45.

[9] Dugan, J.B., Trivedi, K.S., Smotherman, M.K. and Geist, R.M. (1986). The Hybrid Automated Reliability Predictor. *Journal of Guidance, Control, and Dynamics*. 9, 3, 319-331.

[10] Fox, B.L. and Glynn, P.W. (1986). Discrete-Time Conversion for Simulating Semi-Markov Processes. *Operations Research Letters* 5, 191-196.

[11] Fox, B.L. and Glynn, P.W. (1988). Discrete-Time Conversion for Finite-Horizon Markov Processes. Technical Report No. 790, School of Operations Research and Industrial Engineering, Cornell University, Itacha, New York.

[12] Geist, R.M. and Trivedi, K.S. (1983). Ultra-High Reliability Prediction for Fault-Tolerant Computer Systems. *IEEE Transactions on Computers* C-32, 1118-1127.

[13] Glynn, P.W. and Heidelberger, P. (1988). Bias Properties of Budget Constrained Monte Carlo Simulations. IBM Research Report RC 13581, Yorktown Heights, New York. To appear in *Operations Research*.

[14] Glynn, P.W. and Iglehart, D.L. (1989). Importance sampling for Stochastic Simulations. To appear in *Management Science*.

[15] Goyal, A., Carter, W.C., de Souza e Silva, E., Lavenberg, S.S., and Trivedi, K.S. (1986). The System Availability Estimator. *Proceedings of the Sixteenth Symposium on Fault-Tolerant Computing*. Vienna, Austria, 84-89.

[16] Goyal, A., Lavenberg, S.S., and Trivedi, K.S. (1987). Probabilistic Modeling of Computer System Availability. *Annals of Operations Research* 8, 285-306.

[17] Goyal, A. and Lavenberg, S.S. (1987). Modeling and Analysis of Computer System Availability. *IBM Journal of Research and Development*, 31 6, 651-664.

[18] Goyal, A., Heidelberger, P. and Shahabuddin, P. (1987). Measure Specific Dynamic Importance Sampling for Availability Simulations. *1987 Winter Simulation Conference Proceedings*. A. Thesen, H. Grant and W.D. Kelton (eds.). IEEE Press, 351-357.

[19] Gross, D. and Miller, D.R. (1984). The Randomization Technique as a Modeling Tool and Solution Procedure for Transient Markov Processes. *Operations Research* **32**, 343-361.

[20] Hammersley, J.M. and Handscomb, D.C. (1964). *Monte Carlo Methods*. Methuen, London.

[21] Hordijk, A., Iglehart, D.I. and Schassberger, R. (1976). Discrete Time Methods for Simulating Continuous Time Markov Chains. *Adv. Appl. Prob.* **8**, 772-788.

[22] Kahn, H. and Marshall, A.W. (1953). Methods of Reducing Sample Size in Monte Carlo Computations. *Journal of Operations Research Society of America* **1**, **5**, 263-278.

[23] Karlin, S. and Taylor, H.M. (1975). *A First Course in Stochastic Processes*, Second Edition. Academic Press, New York.

[24] Knuth, D.E. (1976). Big Omicron, Big Omega and Big Theta. *SIGACT News (ACM)* **8**, **2**, 18-24.

[25] Lavenberg, S.S., Moeller, T.I. and Sauer, C.H. (1979). Concomitant Control Variables Applied to the Regenerative Simulation of Queueing Systems. *Operations Research* **27**, **1**, 134-160.

[26] L'ecuyer, P. (1988). Efficient and Portable Combined Random Number Generators. *Comm. of the ACM* **31**, **6**, 742-774.

[27] Lehmann, E.L. (1983). *Theory of Point Estimation*. John Wiley and Sons, Inc., New York.

[28] Lewis, E.E. and Böhm, F. (1984). Monte Carlo Simulation of Markov Unreliability Models. *Nuclear Engineering and Design* **77**, 49-62.

[29] Liceaga, C.A. and Siewiorek, D.P. (1986). Towards Automatic Markov Reliability Modeling of Computer Architectures. *NASA Technical Memorandum* **89009**.

[30] Makam, S.V., Avizienis, A. and Grusas, G. (1982). *UCLA ARIES 82 User's Guide*. Also Technical Report CSD-820830, University of California at Los Angeles.

[31] Johnson, Jr. A. M. and Malek, M. (1988). Survey of Software Tools for Evaluating Reliability, Availability, and Serviceability. *ACM Computing Surveys* **20**, **4**, 227-269.

[32] Meyer, J.F. (1980). On Evaluating the Performability of Degradable Computing Systems. *IEEE Transactions on Computers* **C-39**, **8**, 720-731.

[33] Miller, R.G. (1974). The Jackknife - A Review. *Biometrika* **61**, 1-15.

[34] Muntz, R.R., de Souza e Silva, E. and Goyal, A. (1989). Bounding Availability of Repairable Computer Systems. *Proceedings of the ACM Sigmetrics and Performance'89*. Berkeley, California.

[35] Nicola, V.F. (1989). Lumping in Markov Reward Processes. IBM Research Report RC 14719, Yorktown Heights, New York.

[36] Ross, S.M. (1970). *Applied Probability Models with Optimization Applications*. Holden-Day, San Francisco.

[37] Ross, S.M. and Schechner, Z. (1970). Using Simulation to Estimate First Passage Distribution. *Management Science* **31**, **2**, 224-234.

[38] Parekh, S. and Walrand, J. (1989). A Quick Simulation Method for Excessive Backlogs in Networks of Queues. *IEEE Transaction on Automatic Control* **34**, **1**, 54-66.

[39] Sahner, R.A. and Trivedi, K.S. (1987). Performance and Reliability Analysis Using Acyclic Directed Graphs. *IEEE Transaction on Software Engineering* **SE-13**, **10**, 1105-1114.

[40] Sanders, W. H. and Meyer, J. F. (1986). METASAN: A Performability Evaluation Tool Based on Stochastic Activity Networks. *Proceedings of the 1986 Fall Joint Computer Conference.* AFIPS, New York, 807-816.

[41] Shahabuddin, P., Nicola, V.F., Heidelberger, P., Goyal A. and Glynn P.W. (1988). Variance Reduction in Mean Time to Failure Simulations. *1988 Winter Simulation Conference Proceedings.* M.A. Abrams, P.L. Haigh and J.C. Comfort (eds.). IEEE Press, 491-499.

[42] Siegmund, D. (1976). Importance Sampling in the Monte Carlo Study of Sequential Tests. *The Annals of Statistics* **4**, 673-684.

[43] Smith, W.I. (1955). Regenerative Stochastic Processes. *Proc. Roy. Soc. A.* **232**, 6-31.

[44] Walrand, J. (1987). Quick Simulation of Rare Events in Queueing Networks. *Proceedings of the Second International Workshop on Applied Mathematics and Performance/Reliability Models of Computer/Communication Systems.* G. Iazeolla, P.J. Courtois and O.J. Boxma (eds). North Holland Publishing Company, Amsterdam, 275-286.

Appendix A: 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, $U(t,x) = P\{U(t) \le x\}$, where $U(t) = \int_{s=0}^{t} 1_{\{Y_s \in F\}} ds$. Since $A(t) = 1 - U(t)/t$ we have $A(t,x) = P\{A(t) \le x\} = 1 - U(t,(1-x)t)$. To derive an estimator for $U(t,x)$, write

$$U(t,x) = U_1(t,x) + U_2(t,x) \qquad (A.1)$$

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

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

Now if $X_i = 0$, then $\beta_{i-1} = \beta_i$. 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 non-zero states, we can write

$$U_1(t,x) = E\Big[\sum_{i=0}^{\infty} 1_{\{F_{i-1} \le x, X_i \notin F\}} \Big(F_{n_0(i-1)}(t - \beta_{i-1}, \lambda_0) - F_{n_0(i)}(t - \beta_i, \lambda_0) \Big) \Big] . \qquad (A.3)$$

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

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

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 4.2.

Numerical Results	Direct	Bias1 (0.5)	Bias1-Balancing (0.5)	Bias1-Bias2 (0.5 0.5)	MS-Bias1-Bias2 (0.9/0.9)
0.9302×10^{-5}	1.0171×10^{-5}	0.9779×10^{-5}	0.9547×10^{-5}	0.9395×10^{-5}	0.9317×10^{-5}
Unavailability	$\pm\ 27.1\ \%$	$\pm\ 7.6\ \%$	$\pm\ 6.2\ \%$	$\pm\ 2.7\ \%$	$\pm\ 1.0\ \%$
$0.1637 \times 10^{+6}$	$0.1524 \times 10^{+6}$	$0.1581 \times 10^{+6}$	$0.1631 \times 10^{+6}$	$0.1626 \times 10^{+6}$	$0.1633 \times 10^{+6}$
MTTF	$\pm\ 25.7\ \%$	$\pm\ 7.0\ \%$	$\pm\ 5.7\ \%$	$\pm\ 2.5\ \%$	$\pm\ 1.0\ \%$

Table 1. 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-Balancing (0.9)
0.6957×10^{-7}	0.4164×10^{-7}	0.6644×10^{-7}	0.6976×10^{-7}	0.6375×10^{-7}	0.6810×10^{-7}
Unavailability	$\pm\ 164.5\ \%$	$\pm\ 46.1\ \%$	$\pm\ 2.7\ \%$	$\pm\ 5.6\ \%$	$\pm\ 2.2\ \%$
$0.2189 \times 10^{+8}$	$0.4703 \times 10^{+8}$	$0.2227 \times 10^{+8}$	$0.2183 \times 10^{+8}$	$0.2349 \times 10^{+8}$	$0.2222 \times 10^{+8}$
MTTF	$\pm\ 164.5\ \%$	$\pm\ 43.7\ \%$	$\pm\ 2.3\ \%$	$\pm\ 5.1\ \%$	$\pm\ 2.0\ \%$

Table 2. Unavailability and MTTF Estimates and Relative HW
in an Unbalanced System

Events per Replication	Direct Simulation			Bias1/Bias2 (0.5/0.5)			MS - Bias1/Bias2 (0.9/0.9)		
	Relative Bias (Std. Dev.)	Relative IIW	Coverage	Relative Bias (Std. Dev.)	Relative IIW	Coverage	Relative Bias (Std. Dev.)	Relative IIW	Coverage
2000	6.95 % (12.88 %)	153.65 %	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 %

Table 3. Coverage Experiments for the Balanced System

Time (t)	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 4. Unreliability Estimates in a Balanced System

Time (t)	Numerical Interval Unavailability	Bias1 (0.5)			Bias1/Balancing (0.5)			Bias1/Bias-2 (0.5/0.5)		
		Standard	Forcing	Conditioning	Standard	Forcing	Conditioning	Standard	Forcing	Conditioning
4	$.3178 \times 10^{-5}$	$.3057 \times 10^{-5}$	$.3110 \times 10^{-5}$	$.3224 \times 10^{-5}$	$.2572 \times 10^{-5}$	$.3213 \times 10^{-5}$	$.3189 \times 10^{-5}$	$.3204 \times 10^{-5}$	$.3207 \times 10^{-5}$	$.3205 \times 10^{-5}$
		$\pm$ 35.0 %	$\pm$ 6.8 %	$\pm$ 4.7 %	$\pm$ 27.5 %	$\pm$ 5.2 %	$\pm$ 3.7 %	$\pm$ 9.6 %	$\pm$ 2.0 %	$\pm$ 1.5 %
16	$.7322 \times 10^{-5}$	$.6945 \times 10^{-5}$	$.7148 \times 10^{-5}$	$.7401 \times 10^{-5}$	$.7677 \times 10^{-5}$	$.7462 \times 10^{-5}$	$.7219 \times 10^{-5}$	$.7489 \times 10^{-5}$	$.7511 \times 10^{-5}$	$.7383 \times 10^{-5}$
		$\pm$ 15.8 %	$\pm$ 6.7 %	$\pm$ 3.9 %	$\pm$ 12.9 %	$\pm$ 5.5 %	$\pm$ 3.1 %	$\pm$ 4.8 %	$\pm$ 2.1 %	$\pm$ 1.2 %
64	$.8806 \times 10^{-5}$	$.8621 \times 10^{-5}$	$.8866 \times 10^{-5}$	$.9276 \times 10^{-5}$	$.8651 \times 10^{-5}$	$.9012 \times 10^{-5}$	$.8822 \times 10^{-5}$	$.8862 \times 10^{-5}$	$.8966 \times 10^{-5}$	$.8821 \times 10^{-5}$
		$\pm$ 11.4 %	$\pm$ 8.3 %	$\pm$ 5.1 %	$\pm$ 8.4 %	$\pm$ 6.6 %	$\pm$ 4.0 %	$\pm$ 3.2 %	$\pm$ 2.4 %	$\pm$ 1.5 %
256	$.9178 \times 10^{-5}$	$.9882 \times 10^{-5}$	$.9442 \times 10^{-5}$	$.9187 \times 10^{-5}$	1.1150×10^{-5}	1.0766×10^{-5}	1.0010×10^{-5}	$.9049 \times 10^{-5}$	$.9133 \times 10^{-5}$	$.9063 \times 10^{-5}$
		$\pm$ 19.7 %	$\pm$ 12.9 %	$\pm$ 10.5 %	$\pm$ 14.9 %	$\pm$ 13.9 %	$\pm$ 9.3 %	$\pm$ 4.4 %	$\pm$ 4.2 %	$\pm$ 3.3 %
1024	$.9277 \times 10^{-5}$	$.9679 \times 10^{-5}$	$.9677 \times 10^{-5}$	1.3573×10^{-5}	$.3691 \times 10^{-5}$	$.3707 \times 10^{-5}$	$.6168 \times 10^{-5}$	1.0760×10^{-5}	1.0728×10^{-5}	$.7822 \times 10^{-5}$
		$\pm$ 91.5 %	$\pm$ 91.5 %	$\pm$ 103.5 %	$\pm$ 36.3 %	$\pm$ 36.1 %	$\pm$ 51.6 %	$\pm$ 64.8 %	$\pm$ 65.0 %	$\pm$ 21.3 %

Table 5. Expected Interval Unavailability Estimates in a Balanced System

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION Unclassified	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION / AVAILABILITY OF REPORT Approved for public release; distribution unlimited.
2b. DECLASSIFICATION / DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) Technical Report No. 39	5. MONITORING ORGANIZATION REPORT NUMBER(S) ARO 25839.15-MA

6a. NAME OF PERFORMING ORGANIZATION Dept. of Operations Research	6b. OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION U. S. Army Research Office
6c. ADDRESS (City, State, and ZIP Code) Stanford, CA 94305-4022		7b. ADDRESS (City, State, and ZIP Code) P. O. Box 12211 Research Triangle Park, NC 27709-2211
8a. NAME OF FUNDING / SPONSORING ORGANIZATION U. S. Army Research Office	8b. OFFICE SYMBOL (If applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER DAAL03-88-K-0063
8c. ADDRESS (City, State, and ZIP Code) P. O. Box 12211 Research Triangle Park, NC 27709-2211		10. SOURCE OF FUNDING NUMBERS

10. SOURCE OF FUNDING NUMBERS			
PROGRAM ELEMENT NO	PROJECT NO	TASK NO.	WORK UNIT ACCESSION NO

11. TITLE (Include Security Classification)

A Unified Framework for Simulating Markovian Models of Highly Dependable Systems

12. PERSONAL AUTHOR(S)   Ambuj Goyal, Perwez Shahabuddin, Philip Heidelberger, Victor F. Nicola, Peter W. Glynn

13a. TYPE OF REPORT Technical	13b. TIME COVERED FROM _____ TO _____	14. DATE OF REPORT (Year, Month, Day) July 1989	15. PAGE COUNT 40

16. SUPPLEMENTARY NOTATION   The view, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documentation.

17.	COSATI CODES		18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	simulation, reliability, importance sampling, conditioning, variance reduction

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

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.

20. DISTRIBUTION / AVAILABILITY OF ABSTRACT ☐ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION Unclassified
22a. NAME OF RESPONSIBLE INDIVIDUAL	22b. TELEPHONE (Include Area Code) 22c. OFFICE SYMBOL

DD FORM 1473, 84 MAR   83 APR edition may be used until exhausted.   SECURITY CLASSIFICATION OF THIS PAGE
                                 All other editions are obsolete   UNCLASSIFIED