Simulation of Steady-state Availability Models of Fault-Tolerant Systems with Deferred Repair*

Juan A. Carrasco
Departament d'Enginyeria Electrònica
Universitat Politècnica de Catalunya
Diagonal 647, plta. 9
08028 Barcelona, Spain
carrasco@eel.upc.edu

Technical report DMSD_2006_1
January 19, 2006
complete version of paper appeared in *Proc. 39th IEEE Annual Symp.* with title "Adapted Importance Sampling Schemes for the Simulation of Dependability Models of Fault-Tolerant Systems with Deferred Repair"

Abstract

This paper targets the simulation of continuous-time Markov chain models of fault-tolerant systems with deferred repair. We start by stating sufficient conditions for a given importance sampling scheme to satisfy the bounded relative error property. Using those sufficient conditions, it is noted that many previously proposed importance sampling schemes such as failure biasing and balanced failure biasing satisfy that property. Then, we adapt the importance sampling schemes failure transition distance biasing and balanced failure transition distance biasing so as to develop new importance sampling schemes which can be implemented with moderate effort and at the same time can be proved to be more efficient for balanced systems than the simpler failure biasing and balanced failure biasing schemes. The increased efficiency for balanced and unbalanced systems of the new adapted importance sampling schemes is illustrated using examples.

Keywords: Fault-tolerant computer systems, Markov models, steady-state availability, rare event simulation, importance sampling, variance reduction.

*This work was supported by the Ministry of Education and Science of Spain under the research grant DPI2004–05077.

1 Introduction

Fault-tolerant systems with deferred repair have interesting applications, particularly systems for which the replacement of failed components is an expensive procedure, for instance, because the system is located at a remote site. The dependability of those systems can be analyzed by using continuous-time Markov chain (CTMC) models. CTMCs provide enough flexibility to accommodate characteristics that real fault-tolerant systems may have such as failure propagation, impact of system's operational configuration on failure and repair processes, and sophisticated repair policies. However, the size (number of states) of the resulting CTMC tends to increase fast with the complexity of the modeled fault-tolerant system. That behavior is known as *state space explosion* and limits the application in practice of numerical analysis techniques [18, 23]. Simulation is an approach which by nature is not limited by the size of the CTMC. However, for CTMC dependability models of repairable fault-tolerant systems, standard simulation tends to be expensive due to the rarity of the system failure event.

Importance sampling techniques can be used to speed up standard simulation when the measure under estimation is determined by rare events. The basic idea behind importance sampling [11] is to modify the sampling distributions so that the rare events be sampled with higher probabilities. It is a heuristic approach in which the modified sampling distributions are chosen using available high-level knowledge about the model at hand, and has been used successfully to estimate dependability measures using CTMC models [1, 2, 4, 6, 7, 9, 10, 13, 14, 15, 19, 24]. Failure biasing (FB) is an importance sampling scheme which was first proposed in [15, 24] for the simulation of the expected interval unavailability and, in combination with transition forcing, for the simulation of the unreliability, and it has been adapted in [7, 9] for the simulation of the steady-state unavailability, in [19] for the simulation of the mean time to failure, and in [10] for the simulation of other dependability measures. Balanced failure biasing (BFB) and failure distance biasing are other closely related importance sampling schemes proposed in, respectively, [10, 20] and [4]. Balanced likelihood ratio techniques have been developed [1, 2] which seem to be more efficient than BFB for fault-tolerant systems with failure rates not much smaller than repair rates and high redundancy degrees. Importance sampling schemes which are robust and efficient when the model has high-probability cycles have also been developed [13, 14]. Finally, in [6] two importance sampling schemes called failure transition distance biasing (FTDB) and balanced failure transition distance biasing (BFTDB) have been developed. Those importance sampling schemes are guaranteed to be more efficient than FB and BFB for balanced systems, i.e. systems with failure transition rates of the same order of magnitude. In addition, numerical experiments seem to indicate that, for unbalanced systems, BFTDB can also be significantly more efficient than BFB.

The robustness of the previously reviewed importance sampling schemes has also been investigated [2, 6, 13, 14, 16, 17, 20, 21]. With the exception of [13, 14], all that work has considered CTMC models of fault-tolerant systems with repair in every state with failed components. The robustness of the importance sampling schemes has been guaranteed by proving that the importance sampling schemes satisfy the so-called bounded relative error property, which establishes that the

relative error of the estimator remains bounded as failure rates become smaller compared with repair rates.

In this paper we target the efficient simulation of CTMC models of fault-tolerant systems with deferred repair. More specifically, we will consider the simulation of the steady-state unavailability, although the techniques we will develop can be easily adapted to the simulation of other dependability measures. We will start by deriving sufficient conditions for a given importance sampling scheme to satisfy the bounded relative error property for the simulation of the steady-state unavailability for CTMC models of fault-tolerant systems with deferred repair. Then, by using those conditions we will note that many previously proposed importance sampling techniques such as FB, BFB, FTDB, and BFTDB satisfy the bounded relative error property for the class of CTMC models considered in the paper. Then, we will adapt the importance sampling schemes FTDB and BFTDB so as to develop new importance sampling schemes which can be implemented with moderate effort and at the same time can be proved to be more efficient for balanced systems than the simpler FB and BFB importance sampling schemes. The increased efficiency for both balanced and unbalanced systems of the adapted importance sampling schemes will be illustrated using examples. Although quite general, the type of repair deferment we will consider is not completely general, i.e. we will consider systems in which repair is deferred till some condition on the collection of failed components is reached and, then, repair proceeds till the single state in which no component is failed is reached. Other types of repair deferment exist which yield high probability cycles in the CTMC model. Simulation of those CTMC models can be achieved robustly using the importance sampling schemes developed in [13, 14], which are more expensive than FB and BFB.

The rest of the paper is organized as follows. Section 2 describes the class of targeted CTMC models, reviews the basic simulation method for the steady-state unavailability which will be accelerated using importance sampling, and reviews the importance sampling schemes FB, BFB, FTDB, and BFTDB. Section 3 gives sufficient conditions guaranteeing for the considered class of CTMC models that a given importance sampling scheme satisfies the bounded relative error property. By using those conditions, it will be noted that FB and FTDB satisfy that property for balanced systems and BFB and BFTDB satisfy the property for both balanced and unbalanced systems. Section 4 will motivate and describe the new adapted importance sampling schemes, which will be called AFTDB (adapted failure transition distance biasing) and ABFTDB (adapted balanced failure transition distance biasing). Finally, numerical experiments will be reported in Section 5 illustrating that, for balanced systems, AFTDB can be significantly more efficient than FB and that, for unbalanced systems, ABFTDB can be significantly more efficient than BFB. Section 6 will conclude the paper.

Throughout the paper, we will use the following notation. A function $f(\varepsilon)$ will be said to be $o(\varepsilon^k)$ (written $f(\varepsilon) = o(\varepsilon^k)$), where k is an integer ≥ 0, if $\lim_{\varepsilon \to 0} f(\varepsilon)/\varepsilon^k = 0$; we will extend the notation to vectors and matrices to mean vectors and matrices all of which elements are $o(\varepsilon^k)$. Also, a function $f(\varepsilon)$ will be said to be $\Theta(\varepsilon^k)$ (written $f(\varepsilon) = \Theta(\varepsilon^k)$), where k is an integer ≥ 0, if $f(\varepsilon) = c\varepsilon^k + o(\varepsilon^k)$, for some constant $c \neq 0$; we will extend the notation to vectors and matrices to mean vectors and matrices all of which non-null elements are $\Theta(\varepsilon^k)$. We will also use the following notation regarding bags: $\#(c, B)$ will denote the number of instances in bag B of the

element c; $c_1[n_1]c_2[n_2]\cdots c_k[n_k]$ will denote the bag with exactly n_i, $n_i > 0$, instances of element c_i, $1 \leq i \leq k$; $B_1 \subseteq B_2$ will denote that B_1 is a subbag of B_2, i.e. that $\#(c, B_1) \leq \#(c, B_2)$, for all $c \in \mathcal{C}$, $\mathcal{C}$ being a common domain for B_1 and B_2; $B_1 \subset B_2$ will denote that B_1 is a strict subbag of B_2, i.e. that $B_1 \subseteq B_2$ and $B_1 \neq B_2$.

2 Preliminaries

This section includes preliminary material. We start by providing a complete, unambiguous description of the type of CTMC models which are the target of the new importance sampling schemes developed in the paper. Then, we will review the basic simulation method for the steady-state unavailability which will be accelerated using importance sampling. Finally, we will review the importance sampling schemes FB, BFB, FTDB, and BFTDB.

2.1 Class of CTMC models

We will consider fault-tolerant systems made up of a bag C of component classes with domain $\mathcal{C}$ which can be operational or failed. We assume that the up/down system's state is determined from the bag of operational component classes of the system by an increasing generalized structure function $\Phi(b)$, $b \subseteq C$ represented by a generalized fault tree such as those considered in [5]. To be specific, the generalized fault tree is assumed to be made up of AND and OR gates and to have as inputs atoms of the form $c[n]$, $c \in \mathcal{C}$, $0 < n \leq \#(c, C)$ which evaluate to 1 if and only if the bag b of failed component classes of the system is such that $b \supseteq c[n]$. $\Phi(b) = 0$ if and only if the output of the generalized fault tree evaluates to 1 when the bag of failed component classes is $C - b$. Because the generalized fault tree only includes AND and OR gates, when the bag of failed component classes is the empty bag all input atoms evaluate to 0, the output of the generalized fault tree evaluates to 0, and $\Phi(C) = 1$. Similarly, when the bag of failed component classes is C, all input atoms evaluate to 1, the output of the generalized fault tree evaluates to 1, and $\Phi(\emptyset) = 0$.

We will consider irreducible CTMC $X = \{X(t); t \geq 0\}$ with finite state space Ω modeling fault-tolerant systems with the characteristics described in the previous paragraph, in which each state $s \in \Omega$ has associated with it a bag of failed component classes $F(s) \subseteq C$. The CTMC X has two types of transitions: failure transitions (x, y), characterized by $F(y) \supset F(x)$ and repair transitions (x, y), characterized by $F(y) \subset F(x)$. There exists a single state r with $F(r) = \emptyset$. $F(s)$ determines through the generalized structure function ($\Phi(C - F(s))$) whether the system is up or down in state s. We will denote by U the subset of up states of X and by $D = \Omega - U$ the subset of down states of X. Let $\Psi(b)$, $b \subseteq C$ be some Boolean function with $\Psi(\emptyset) = 1$ and $\Psi(b) = 0$ for $\emptyset \subset b \subseteq C$ and $\Phi(C - b) = 0$, determining whether repair has to be deferred ($\Psi(b) = 1$) or not ($\Psi(b) = 0$) in a state s with $F(s) = b$. The state space Ω can be partitioned as $\Omega = E \cup G \cup E'$, $E, E', G \neq \emptyset$, where all states s in E and E' satisfy $\Phi(F(s)) = 1$ and all states s in G satisfy $\Phi(F(s)) = 0$. The subset E includes the states without repair, the subsets

G and E' includes the states with repair. Note that repair is not deferred in any down state and, therefore, $E \subset U$. We will denote by $\lambda_{x,y}$ the transition rate of X from state x to state y, by $T = \{(x,y) \in \Omega \times \Omega \ : \ y \neq x \ \wedge \ \lambda_{x,y} > 0\}$ the set of transitions of X, by T_F the set of failure transitions, by T_R the set of repair transitions, by $T_F(x) = \{(x,y) \in T_F\}$ the set of failure transitions going out of x, and by $T_R(x) = \{(x,y) \in T_R\}$ the set of repair transitions going out of x. Any bag of component classes $b \subseteq C$, $b \neq \emptyset$ such that there exists in X some $(x,y) \in T_F$ with $F(y) - F(x) = b$ will be called *failure bag* and we will denote by F_B the set of failure bags of the fault-tolerant system. A failure bag f will be said to be *active* in some state x if there exists some failure transition (x,y) having associated with it failure bag f, i.e. $F(y) - F(x) = f$. The set of failure bags which are active in state x will be denoted by $active(x)$. Let $\Omega' = \Omega - \{r\}$. We will make the following five assumptions:

A1) For each state $x \in G \cup E'$, $T_R(x) \neq \emptyset$ and for each $(x,y) \in T_R(x)$, $y \neq r$, $y \in G \cup E'$.

A2) For each state $x \in E$, $T_R(x) = \emptyset$.

A3) $c[1] \in F_B$ for each $c \in \mathcal{C}$.

A4) For each $f \in F_B$, $f' \in F_B$ for each $f' \subset f$, $f' \neq \emptyset$.

A5) For every $x \in U$, $active(x) = \{f \in F_B \ : \ f \subseteq C - F(x)\}$.

Informally, A5 states that from every up there are failure transitions associated with all possible failure bags (those for which there are operational components building up the failure bag), and A3 and A4 state reasonable conditions that the set of failure bags must satisfy (for instance, the conditions are satisfied under any model in which the failure of a component can be propagated to others with probabilities between 0 and 1). Note that $F(r) = \emptyset$ implies $\Phi(C - F(r)) = \Phi(C) = 1$ and $r \in U \neq \emptyset$. Also, because $\Phi(\emptyset) = 0$, assumptions A3 and A5 imply: 1) $D \neq \emptyset$, 2) the existence in X of a path made up of only failure transitions from every state $x \in U$ to D (we will call such paths *failure paths*).

Let $r_{\min} = \min_{(x,y) \in T_R} \lambda_{x,y}$ denote the minimum repair transition rate of X and let $f_{\max} = \max_{(x,y) \in T_F} \lambda_{x,y}$ denote the maximum failure transition rate of X. Let $\varepsilon = f_{\max}/r_{\min}$. The ε parameter can be regarded as a "rarity" parameter measuring how small failure transition rates are with respect to repair transition rates. We will assume that failure transition rates are much smaller than repair transition rates, i.e. $\varepsilon \ll 1$. This corresponds to fault-tolerant systems made up of highly reliable components. To give results regarding the robustness of the importance sampling schemes, we will model repair transition rates as constants $\lambda_{x,y} = r_{\min} r_{x,y}$, $r_{x,y} \geq 1$ and will model failure transition rates as $\lambda_{x,y} = r_{\min} f_{x,y} \varepsilon^{d_{x,y}}$, $f_{x,y} \in (0,1]$, $f_{x,y} \gg \varepsilon$, $d_{x,y} \geq 1$. A fault-tolerant system will be called *balanced* if $d_{x,y} = 1$, $(x,y) \in T_F$. Otherwise, the fault-tolerant system will be called *unbalanced*. Informally, a fault-tolerant system is balanced if failure transition rates differ among them much less than failure transition rates differ from repair transition rates, i.e. calling $f_{\min} = \min_{(x,y) \in T_F} \lambda_{x,y}$, if $f_{\min}/f_{\max} \gg \varepsilon = f_{\max}/r_{\min}$.

2.2 Review of the Simulation Method for the Steady-state Unavailability

The steady-state unavailability UA is defined as the steady-state probability that the system is down. Formally,

$$UA = \lim_{t\to\infty} P\{X(t) \in D\}\,.$$

Because X is irreducible and finite, UA is independent of the initial probability distribution of X and we can assume without loss of generality $X(0) = r$. A formulation for UA can be obtained in terms of random variables W, Z defined on the set of regenerative cycles with regenerative state the state r of the embedded homogeneous discrete-time Markov chain (DTMC) $\Pi = \{\Pi_n; n = 0, 1, 2, \ldots\}$ of X. Π has the same state space and initial probability distribution as X and transition probabilities $P\{\Pi_{n+1} = y \mid \Pi_n = x\} = P_{x,y} = \lambda_{x,y}/\lambda_x$, $x, y \in \Omega$, $y \neq x$ and $P\{\Pi_{n+1} = x \mid \Pi_n = x\} = P_{x,x} = 0$, $x \in \Omega$, where $\lambda_x = \sum_{y\in\Omega-\{x\}} \lambda_{x,y}$ is the output rate of X from state x. Letting $\tau = \min\{n > 0 \,:\, \Pi_n = r\}$, W and Z are defined as

$$W = \sum_{n=0}^{\tau-1} h_{\Pi_n}\,,$$

$$Z = \sum_{n=0}^{\tau-1} 1_D(\Pi_n) h_{\Pi_n}\,,$$

where I_c denotes the indicator function returning the value 1 if condition c is satisfied and the value 0 otherwise and $h_x = 1/\lambda_x$ denotes the mean holding time of X in state x, and we have

$$UA = \frac{E_P[Z]}{E_P[W]}\,, \tag{1}$$

where the subscript P in $E_P[Z]$ and $E_P[W]$ makes explicit the probability measure with respect to which the expectation is defined. Formally, letting $T = \{(x, y) \in \Omega \times \Omega, y \neq x \,:\, P_{x,y} > 0\}$ [1] the set of transitions of Π (it coincides with the set of transitions of X), denoting by $\mathcal{S}$ the set of regenerative cycles of Π, i.e.

$$\mathcal{S} = \{(s_0, s_1, \ldots, s_l) \,:\, s_0 = r \,\wedge\, s_i \neq r, 0 < i < l \,\wedge\, s_l = r \,\wedge\, (s_i, s_{i+1}) \in T, 0 \leq i < l\}\,,$$

denoting by $\mathcal{A}$ the σ-algebra of all subsets of $\mathcal{S}$, the probability space $(\mathcal{S}, \mathcal{A}, P)$ is defined by

$$P\{(s_0, s_1, \ldots, s_l)\} = \prod_{i=0}^{l-1} P_{s_i,s_{i+1}}\,, (s_0, s_1, \ldots, s_l) \in \mathcal{S}\,. \tag{2}$$

The standard regenerative simulation method to estimate UA is based on (1).

Estimation of UA by the regenerative simulation method tends to be inefficient. Intuitively, this is because, being the system failure often a rare event, it may happen that the vast majority of regenerative cycles do not contain down states. Importance sampling techniques can be used to speed up the simulation. This would involve obtaining sample pairs (W_i', Z_i'), $i = 1, 2, \ldots, n$ of

[1] The fact that X and Π have same state space and same set of transitions allows us to apply definitions such as T_F, T_R, $T_F(x)$, and $T_R(x)$ to both X and Π and we will do so throughout the paper.

the random variables $W' = WL$ and $Z' = ZL$, where $L(\omega) = P\{\omega\}/P'\{\omega\}$ is the likelihood ratio, by sampling $\mathcal{S}$ under a modified probability measure P' such that $P'\{(s_0, s_1, \ldots, s_l)\} > 0$, $(s_0, s_1, \ldots, s_l) \in \mathcal{S}$, where P' is constructed so that the system failure event becomes more likely and the variance of Z' is smaller than the variance of Z. However, changing the probability measure may result in a variance of W' larger than the variance of W, which tends to be relatively very small. This has motivated the development of a measure-specific simulation method for UA [9, 21]. That method is the one that we will use. We review it next.

In the measure-specific simulation method for UA, $n = \widetilde{n}k$ samples of W, W_i, $i = 1, 2, \ldots, n$, are obtained by sampling $\mathcal{S}$ under the probability measure P, and $m = \widetilde{m}k$ independent samples of $Z' = ZL$, where L is the likelihood ratio, Z'_i, $i = 1, 2, \ldots, m$, are obtained by sampling $\mathcal{S}$ under a modified probability measure P' such that $P'\{(s_0, s_1, \ldots, s_l)\} > 0$, $(s_0, s_1, \ldots, s_l) \in \mathcal{S}$. The estimator for UA is

$$\widehat{UA} = \frac{\overline{Z'}}{\overline{W}},$$

where $\overline{Z'}$ and $\overline{W}$ are, respectively, the sample means of Z' and W, i.e.

$$\overline{Z'} = \frac{1}{m}\sum_{i=1}^{m} Z'_i = \frac{1}{m}\sum_{i=1}^{m} Z_i L_i,$$

$$\overline{W} = \frac{1}{n}\sum_{i=1}^{n} W_i.$$

The corresponding $100(1-\alpha)$ percent confidence interval for UA is given by

$$\widehat{UA} \pm z_\alpha \frac{\overline{Z'}}{\overline{W}} \left(\left(\frac{1}{\sqrt{m}} \frac{\sqrt{S^2(Z')}}{\overline{Z'}} \right)^2 + \left(\frac{1}{\sqrt{n}} \frac{\sqrt{S^2(W)}}{\overline{W}} \right)^2 \right)^{1/2}, \tag{3}$$

where $S^2(Z')$ and $S^2(W)$ are the sample variances of, respectively, Z' and W, i.e.

$$S^2(Z') = \frac{1}{m-1}\sum_{i=1}^{m} \left(Z'_i - \overline{Z'}\right)^2 = \frac{1}{m-1}\sum_{i=1}^{m} \left(Z_i L_i - \overline{Z'}\right)^2, \tag{4}$$

$$S^2(W) = \frac{1}{n-1}\sum_{i=1}^{n} \left(W_i - \overline{W}\right)^2,$$

and z_α is the $1 - \alpha/2$ quantile of the standard normal distribution. That confidence interval is obtained by applying the central limit theorem with independent, identically distributed random variables (see [2])

$$V_i = \frac{1}{\widetilde{m}} \sum_{j=(i-1)\widetilde{m}+1}^{i\widetilde{m}} Z'_j - UA \left(\frac{1}{\widetilde{n}} \sum_{j=(i-1)\widetilde{n}+1}^{i\widetilde{n}} W_j \right), \ i = 1, 2, \ldots, k,$$

and, then, the goodness of the confidence interval depends on $E\{V_i^2\} < \infty$ and k being sufficiently large. Being $E_{P'}\{Z'\} = E_P\{Z\} < \infty$, $E_P\{W\} < \infty$, and $E_P\{W^2\} < \infty$ [12], $E\{V_i^2\} < \infty$ if and only if $E_{P'}\{Z'^2\} < \infty$. Thus, when choosing P' care should be taken that $E_{P'}\{Z'^2\} < \infty$.

2.3 Review of FB, BFB, FTDB, and BFTDB

In this section we review the importance sampling schemes FB, BFB, FTDB, and BFTDB. In all those schemes, $\mathcal{S}$ is sampled by sampling realizations of Π until state r is hit using either the transition probabilities $P_{x,y}$ or biased transition probabilities $P'_{x,y}$ such that $P'_{x,y} > 0$ if and only if $P_{x,y} > 0$. The biased transition probabilities are used up to the step in which D is hit. The unbiased transition probabilities are used after that point. Then, we have:

$$P'\{(s_0, s_1, \ldots, s_l)\} = \prod_{i=0}^{l_D(s_0,s_1,\ldots,s_l)} P'_{s_i,s_{i+1}} \prod_{i=l_D(s_0,s_1,\ldots,s_l)+1}^{l-1} P_{s_i,s_{i+1}}\,, \quad (s_0, s_1, \ldots, s_l) \in \mathcal{S}\,, \quad (5)$$

where

$$l_D(s_0, s_1, \ldots, s_l) = \max\{k \leq l \,:\, s_0, s_1, \ldots, s_k \in U\}\,.$$

In FB, when a state has both outgoing failure transitions and outgoing repair transitions, the probabilities associated with failure transitions and the probabilities associated with repair transitions are scaled so that the sum of the probabilities associated with failure transitions becomes FB, $0 < FB < 1$, and, consequently, the sum of the probabilities associated with repair transitions becomes $1 - FB$. BFB differs from FB in that the probability assigned to failure transitions (1, if the state does not have outgoing repair transitions) is evenly distributed among those transitions. Formally, denoting by Ω_{FR} the set of states having both outgoing failure transitions and outgoing repair transitions, the biased transition probabilities in FB are

$$P'_{x,y} = \begin{cases} \dfrac{P_{x,y}}{\sum_{z\,:\,(x,z)\in T_F(x)} P_{x,z}}\, FB & \text{if } x \in \Omega_{FR} \wedge (x,y) \in T_F(x)\,, \\ \dfrac{P_{x,y}}{\sum_{z\,:\,(x,z)\in T_R(x)} P_{x,z}}\,(1 - FB) & \text{if } x \in \Omega_{FR} \wedge (x,y) \in T_R(x)\,, \\ P_{x,y} & \text{if } x \notin \Omega_{FR}\,. \end{cases}$$

and the biased transition probabilities in BFB are

$$P'_{x,y} = \begin{cases} \dfrac{FB}{|T_F(x)|} & \text{if } x \in \Omega_{FR} \wedge (x,y) \in T_F(x)\,, \\ \dfrac{P_{x,y}}{\sum_{z\,:\,(x,z)\in T_R(x)} P_{x,z}}\,(1 - FB) & \text{if } x \in \Omega_{FR} \wedge (x,y) \in T_R(x)\,, \\ \dfrac{1}{|T_F(x)|} & \text{if } x \notin \Omega_{FR}\,. \end{cases}$$

The importance sampling schemes FTDB and BFTDB exploit the failure transition distance concept. The failure transition distance from a state x, $td(x)$, is defined to be 0 for $x \in D$, and is defined for $x \in U$ as the length of the shortest failure path from x. A failure transition (x, y) is said to be dominant if $td(y) = td(x) - 1$ and non-dominant otherwise ($td(y) = td(x)$). Both FTDB and BFTDB use two biasing parameters. The first one, FB, $0 < FB < 1$, plays a similar role as FB in FB and BFB and biases failure transitions with respect to repair transitions. The second

one, DB, $0 < DB < 1$, biases dominant failure transitions with respect to non-dominant failure transitions. Formally, denoting by $T_D(x)$ the set of dominant failure transitions from state x, by $T_{ND}(x)$ the set of non-dominant failure transitions from state x, and by Ω_D the set of states having both outgoing dominant failure transitions and outgoing non-dominant failure transitions, the biased transition probabilities in FTDB are

$$P'_{x,y} = \begin{cases} \dfrac{P_{x,y}}{\sum_{z\,:\,(x,z)\in T_D(x)} P_{x,z}}\, FB \times DB & \text{if } x \in \Omega_{FR} \cap \Omega_D \ \wedge\ (x,y) \in T_D(x)\,, \\ \dfrac{P_{x,y}}{\sum_{z\,:\,(x,z)\in T_{ND}(x)} P_{x,z}}\, FB(1-DB) & \text{if } x \in \Omega_{FR} \cap \Omega_D \ \wedge\ (x,y) \in T_{ND}(x)\,, \\ \dfrac{P_{x,y}}{\sum_{z\,:\,(x,z)\in T_F(x)} P_{x,z}}\, FB & \text{if } x \in \Omega_{FR} - \Omega_D \ \wedge\ (x,y) \in T_F(x)\,, \\ \dfrac{P_{x,y}}{\sum_{z\,:\,(x,z)\in T_D(x)} P_{x,z}}\, DB & \text{if } x \notin \Omega_{FR} \ \wedge\ x \in \Omega_D \ \wedge\ (x,y) \in T_D(x)\,, \\ \dfrac{P_{x,y}}{\sum_{z\,:\,(x,z)\in T_{ND}(x)} P_{x,z}}\, (1-DB) & \text{if } x \notin \Omega_{FR} \ \wedge\ x \in \Omega_D \ \wedge\ (x,y) \in T_{ND}(x)\,, \\ P_{x,y} & \text{if } x \notin \Omega_{FR} \ \wedge\ x \notin \Omega_D\,, \\ \dfrac{P_{x,y}}{\sum_{z\,:\,(x,z)\in T_R(x)} P_{x,z}}\, (1-FB) & \text{if } x \in \Omega_{FR} \ \wedge\ (x,y) \in T_R(x)\,. \end{cases}$$

BFTDB differs from FTDB in that the probability assigned to each subset of failure transitions is evenly distributed among the transitions in the subset. Then, in BFB the biased transition probabilities are

$$P'_{x,y} = \begin{cases} \dfrac{FB \times DB}{|T_D(x)|} & \text{if } x \in \Omega_{FR} \cap \Omega_D \ \wedge\ (x,y) \in T_D(x)\,, \\ \dfrac{FB(1-DB)}{|T_{ND}(x)|} & \text{if } x \in \Omega_{FR} \cap \Omega_D \ \wedge\ (x,y) \in T_{ND}(x)\,, \\ \dfrac{FB}{|T_F(x)|} & \text{if } x \in \Omega_{FR} - \Omega_D \ \wedge\ (x,y) \in T_F(x)\,, \\ \dfrac{DB}{|T_D(x)|} & \text{if } x \notin \Omega_{FR} \ \wedge\ x \in \Omega_D \ \wedge\ (x,y) \in T_D(x)\,, \\ \dfrac{1-DB}{|T_{ND}(x)|} & \text{if } x \notin \Omega_{FR} \ \wedge\ x \in \Omega_D \ \wedge\ (x,y) \in T_{ND}(x)\,, \\ \dfrac{1}{|T_F(x)|} & \text{if } x \notin \Omega_{FR} \ \wedge\ x \notin \Omega_D\,, \\ \dfrac{P_{x,y}}{\sum_{z\,:\,(x,z)\in T_R(x)} P_{x,z}}\, (1-FB) & \text{if } x \in \Omega_{FR} \ \wedge\ (x,y) \in T_R(x)\,. \end{cases}$$

The implementation of FTDB and BFTDB requires the computation of the failure transition distances from the the currently sampled state and their successors through failure transitions. Efficient procedures which can be embedded into the simulation for computing them are described in [6].

Those procedures require the computation of the minimal cuts of the generalized structure function of the system. An algorithm for computing those minimal cuts is described in [5].

3 Robustness Results

Throughout the section, we will denote by $\mathbf{P}$ the transition probability matrix of Π. Also, being $\mathbf{A}$ a matrix, $\mathbf{A}_{B,C}$ will denote the restriction of $\mathbf{A}$ to the pairs of indices (i,j), $i \in B$, $j \in C$, and, being $\mathbf{x}$ a vector, $\mathbf{x}_B$ will denote the restriction of $\mathbf{x}$ to the indices $i \in B$. Let

$$\begin{aligned} T_C \;=\; \big\{(x,y) \in T \,:\, & x \in \Omega' \,\wedge\, (x,y) \in T_R \\ & \vee\, x \in \Omega' - \Omega_{FR} \,\wedge\, (x,y) \in T_F \,\wedge\, d_{x,z} \geq d_{x,y} \;\text{ for all }\; z \,:\, (x,z) \in T_F \big\}\,. \end{aligned}$$

Given a CTMC (DTMC), a *cycle* of the CTMC (DTMC) is defined to be any subset of the transitions of the CTMC (DTMC) (with non-null rate (probability)) such that, properly sorted, can be put in the form $(x_0,x_1),(x_1,x_2),\ldots,(x_{n-1},x_n)$ with $x_n = x_0$ and $x_j \neq x_i$ for $0 \leq i,j < n$, $j \neq i$. Because of the assumed properties for X, it follows that the transitions in T_C do not build any cycle in X. Then, we will prove that any importance sampling scheme in which biasing is turned off when D is hit and in which biasing is done by using biased transition probabilities $P'_{x,y}$ satisfies the bounded relative error provided the following conditions hold:

C1) $P'_{x,y} > 0$ if and only if $P_{x,y} > 0$, $x \in U$.

C2) $P'_{x,y} = \Theta(1)$, $x \in U$.

For the simulation method for the steady-state availability we consider, the bounded relative error property asserts that $\sqrt{\mathrm{Var}_{P'}[Z']}/E_P[Z]$ ($\mathrm{Var}_P[Z]$ denotes the variance of the random variable Z under the probability measure P) remains bounded as $\varepsilon \to 0$. The property supports both the robustness and the efficiency of an importance sampling scheme. The first follows from the fact that, being $E_P[Z]$ finite, for sufficiently small ε, $\sqrt{\mathrm{Var}_{P'}[Z']}$ and $E_{P'}[Z'^2]$ will be finite. The second follows from the fact that $\sqrt{\mathrm{Var}_{P'}[Z']}/E_P[Z]$ cannot become pathologically large for small ε. The strategy to prove the bounded relative error property is similar to the strategy used in [21] and is to show that if the previous conditions hold then $E_{P'}[Z'^2] = \Theta(\varepsilon^{2\rho})$ and $E_P[Z] = \Theta(\varepsilon^{\rho})$ for some $\rho \geq 0$, which, using $\mathrm{Var}_{P'}[Z'] = E_{P'}[Z'^2] - E_P[Z]^2$, imply $\sqrt{\mathrm{Var}_{P'}[Z']}/E_P[Z] = c + o(\varepsilon)$, $c \geq 0$ and, therefore, either $\sqrt{\mathrm{Var}_{P'}[Z']}/E_P[Z] = \Theta(1)$ or $\sqrt{\mathrm{Var}_{P'}[Z']}/E_P[Z] = o(\varepsilon)$, both of which imply the bounded relative error property.

In the remaining of this section we will consider probability spaces $(\mathcal{S}^x, \mathcal{A}^x, P_x)$ and $(\mathcal{S}^x, \mathcal{A}^x, P'_x)$, $\mathcal{S}^x = \{(s_0,s_1,\ldots,s_l) \,:\, s_0 = x \,\wedge\, s_i \neq r, 0 < i < l \,\wedge\, s_l = r \,\wedge\, (s_i,s_{i+1}) \in T, 0 \leq i < l\}$, which capture the set of paths of Π from an arbitrary state $x \in \Omega$ to state r. The probability measure P_x has the same expression as P in (2) with $\mathcal{S}$ replaced by $\mathcal{S}^x$. The probability measure P'_x has the same expression as P' in (5) with $\mathcal{S}$ replaced by $\mathcal{S}^x$ and

$$l_D(s_0,s_1,\ldots,s_l) = \begin{cases} \max\{k \leq l \,:\, s_0,s_1,\ldots,s_k \in U\} & \text{if } s_0 \in U \\ -1 & \text{if } s_0 \in D \end{cases}\,.$$

Let

$$v_x = E_{P_x}\left[\sum_{n=0}^{\tau_x-1} I_{\Pi_n^x \in D} h_{\Pi_n^x}\right], \quad x \in \Omega\,,$$

with $\tau_x = \min\{n \,:\, \Pi_n^x = r\}$. We have the following result:

Lemma 1. *$v_x = \Theta(1)$, $x \in D$.*

Proof. Let ν_y^x, $x \in D$, $y \in \Omega'$ be the number of visits to y before hitting r of the version of Π with initial state x, Π^x. Letting the column vector $\boldsymbol{\nu}^x = (E[\nu_y^x])_{y\in\Omega'}$, denoting by $\mathbf{e}^x$ the column vector with component associated with x equal to 1, and the remaining components, associated with states in $\Omega' - \{x\}$, equal to 0, and letting $\mathbf{I}$ an identity matrix of appropriate dimension, we have

$$(\mathbf{I} - \mathbf{P}_{\Omega',\Omega'}^T)\boldsymbol{\nu}^x = \mathbf{e}^x,$$

where the superscript ^{T} denotes the transpose of a matrix. Note that, being Π finite and irreducible, the states in Ω' of the DTMC Π^* obtained from Π by making absorbing r, which has transition probability matrix restricted to $\Omega' \times \Omega'$ $\mathbf{P}_{\Omega',\Omega'}$, are transient and, therefore (see, for instance, [3, Chapter 8, Lemma 3.20]), $(\mathbf{I}-\mathbf{P}_{\Omega',\Omega'}^T)^{-1}$ exists. We can decompose $\mathbf{P}_{\Omega',\Omega'}^T$ as $\mathbf{A}^T+\mathbf{C}^T$, $\mathbf{A} = \Theta(1)$, $\mathbf{C} = o(1)$, where $\mathbf{A}$ includes the transition probabilities associated with transitions in T_C and $\mathbf{C}$ includes the remaining transition probabilities. Then, we can write

$$(\mathbf{I} - \mathbf{A}^T - \mathbf{C}^T)\boldsymbol{\nu}^x = \mathbf{e}^x.$$

Let $\boldsymbol{\nu}^{x*}$ be the solution of

$$(\mathbf{I} - \mathbf{A}^T)\boldsymbol{\nu}^{x*} = \mathbf{e}^x.$$

Note that, being $\mathbf{A}$ the transition probability matrix restricted to $\Omega' \times \Omega'$ of the DTMC Π^{**} differing from Π^* in that transition probabilities associated with transitions not in T_C have been redirected to the absorbing state r and being all states $x \in \Omega'$ transient in Π^{**}, $(\mathbf{I} - \mathbf{A}^T)^{-1}$ exists. Then, we have (see, for instance, [8, Section 5.5.3]):

$$\boldsymbol{\nu}^x = \boldsymbol{\nu}^{x*} + \delta\boldsymbol{\nu}^x\,, \tag{6}$$

with

$$\begin{aligned}\|\delta\boldsymbol{\nu}^x\|_\infty &\leq \left\|(\mathbf{I} - \mathbf{A}^T)^{-1}\right\|_\infty \left\|\mathbf{C}^T\right\|_\infty \|\boldsymbol{\nu}^x\|_\infty \\ &\leq \left\|(\mathbf{I} - \mathbf{A}^T)^{-1}\right\|_\infty \left\|\mathbf{C}^T\right\|_\infty \left\|(\mathbf{I} - \mathbf{A}^T - \mathbf{C}^T)^{-1}\right\|_\infty \|\mathbf{e}^x\|_\infty\,.\end{aligned}$$

The quantities $\|(\mathbf{I} - \mathbf{A}^T)^{-1}\|_\infty$ and $\|\mathbf{e}^x\|_\infty$ are $\Theta(1)$ and $\|\mathbf{C}^T\|_\infty$ is $o(1)$. From $\lim_{\varepsilon\to 0}\|(\mathbf{I} - \mathbf{A}^T - \mathbf{C}^T)^{-1}\|_\infty = \|(\mathbf{I} - \mathbf{A}^T)^{-1}\|_\infty$ it is easy to prove that $\|(\mathbf{I} - \mathbf{A}^T - \mathbf{C}^T)^{-1}\|_\infty = \Theta(1)$. Then, $\|\delta\boldsymbol{\nu}^x\|_\infty = o(1)$, which implies that all components of $\delta\boldsymbol{\nu}^x$ are $o(1)$. This, together with $\boldsymbol{\nu}^{x*} = (\mathbf{I} - \mathbf{A}^T)^{-1}\mathbf{e}^x = \Theta(1)$ and (6), shows that the components of $\boldsymbol{\nu}^x$, $E[\nu_y^x]$, are either $\Theta(1)$ or $o(1)$. Furthermore, since, being Π^x initially at state $x \in D$, $E[\nu_x^x] \geq 1$, it follows that $E[\nu_x^x] = \Theta(1)$. The result follows, then, by noting that $v_x = \sum_{y\in D} E[\nu_y^x]$. □

Let $\mathbf{v}$ be the column vector $(v_x)_{x\in\Omega}$. Note that

$$E_P[Z] = \mathbf{P}_{\{r\},D}\mathbf{v}_D + \mathbf{P}_{\{r\},U'}\mathbf{v}_{U'}\,. \tag{7}$$

Let

$$\mathbf{d}_{U'} = \mathbf{P}_{U',D}\mathbf{v}_D\,. \tag{8}$$

We have:

$$\mathbf{v}_{U'} = \mathbf{P}_{U',U'}\mathbf{v}_{U'} + \mathbf{d}_{U'}\,,$$

and using the fact that $\sum_{n=0}^{\infty}\mathbf{P}^n_{U',U'}$ is finite, because $\mathbf{P}_{U',U'}$ is the restriction of $\mathbf{P}$ to $U'\times U'$, the DTMC Π is irreducible and U' is strictly contained in Ω (see, for instance, [3, Chapter 8, Lemma 3.20]), we have:

$$\mathbf{v}_{U'} = \sum_{n=0}^{\infty}\mathbf{P}^n_{U',U'}\mathbf{d}_{U'}\,. \tag{9}$$

Combining (7), (8) and (9) we get:

$$E_P[Z] = \left(\mathbf{P}_{\{r\},D} + \mathbf{P}_{\{r\},U'}\sum_{n=0}^{\infty}\mathbf{P}^n_{U',U'}\mathbf{P}_{U',D}\right)\mathbf{v}_D\,. \tag{10}$$

Let

$$q_x = E_{P_x}\left[\left(\sum_{n=0}^{\tau_x-1} I_{\Pi^x_n\in D}h_{\Pi^x_n}\right)^2\right]\,,$$

$$a_x = E_{P'_x}\left[\left(\sum_{n=0}^{\tau_x-1} I_{\Pi^x_n\in D}h_{\Pi^x_n}\right)^2\left(\frac{P\{(\Pi^x_0,\Pi^x_1,\ldots,\Pi^x_{\tau_x})\}}{P'\{(\Pi^x_0,\Pi^x_1,\ldots,\Pi^x_{\tau_x})\}}\right)^2\right]\,,$$

and let the column vectors $\mathbf{q} = (q_x)_{x\in\Omega}$, $\mathbf{a} = (a_x)_{x\in\Omega}$. We have the following result:

Lemma 2. *$q_x = \Theta(1)$, $x\in D$.*

Proof. Being $\mathbf{x} = (x_i)$ and $\mathbf{y} = (y_i)$ column vectors of the same dimension, we will denote by $\mathbf{x}\circ\mathbf{y}$ the column vector of the same dimension (x_iy_i). We will also denote $\mathbf{x}\circ\mathbf{x}$ by $\mathbf{x}^2$. Let the column vector $\mathbf{g} = (I_{x\in D}h_x)_{x\in\Omega'}$. We have:

$$\mathbf{q}_{\Omega'} = \mathbf{g}^2_{\Omega'} + 2\mathbf{g}_{\Omega'}\circ(\mathbf{P}_{\Omega',\Omega'}\mathbf{v}_{\Omega'}) + \mathbf{P}_{\Omega',\Omega'}\mathbf{q}_{\Omega'}\,. \tag{11}$$

Using the fact that $\sum_{n=0}^{\infty}\mathbf{P}^n_{U',U'}$ is finite, from $\mathbf{v}_{\Omega'} = \mathbf{g}_{\Omega'}+\mathbf{P}_{\Omega',\Omega'}\mathbf{v}_{\Omega'}$, we get $\mathbf{v}_{\Omega'} = \sum_{n=0}^{\infty}\mathbf{P}^n_{\Omega',\Omega'}\mathbf{g}_{\Omega'}$, which combined with (11), using again the fact that $\sum_{n=0}^{\infty}\mathbf{P}^n_{U',U'}$ is finite, gives:

$$\mathbf{q}_{\Omega'} = \left(\sum_{n=0}^{\infty}\mathbf{P}^n_{\Omega',\Omega'}\right)\left(\mathbf{g}^2_{\Omega'} + 2\mathbf{g}_{\Omega'}\circ\left(\mathbf{P}_{\Omega',\Omega'}\sum_{n=0}^{\infty}\mathbf{P}^n_{\Omega',\Omega'}\mathbf{g}_{\Omega'}\right)\right)\,. \tag{12}$$

Expression (12) was also obtained in [21].

Let $\mathbf{u} = \sum_{n=0}^{\infty}\mathbf{P}^n_{\Omega',\Omega'}\mathbf{g}_{\Omega'}$. Since $\sum_{n=0}^{\infty}\mathbf{P}^n_{\Omega',\Omega'}$ is finite (because Π is finite and irreducible), $\mathbf{u}$ is the solution of

$$\left(\mathbf{I} - \mathbf{P}_{\Omega',\Omega'}\right)\mathbf{u} = \mathbf{g}_{\Omega'}\,,$$

where $\mathbf{I}$ is an identity matrix of appropriate dimension. Then, that all components of $\mathbf{u}$ are either $\Theta(1)$ or $o(1)$ can be proved as it was proved in the proof of Lemma 1 that all components of $\boldsymbol{\nu}^x$ were either $\Theta(1)$ or $o(1)$, noting that $\mathbf{g}_{\Omega'} = \Theta(1)$.

Let

$$\mathbf{w} = \mathbf{g}_{\Omega'}^2 + 2\mathbf{g}_{\Omega'} \circ \left(\mathbf{P}_{\Omega',\Omega'}\mathbf{u}\right) . \tag{13}$$

We have (12):

$$\mathbf{q}_{\Omega'} = \sum_{n=0}^{\infty} \mathbf{P}_{\Omega',\Omega'}^n \mathbf{w} \tag{14}$$

and

$$\left(\mathbf{I} - \mathbf{P}_{\Omega',\Omega'}\right) \mathbf{q}_{\Omega'} = \mathbf{w} . \tag{15}$$

From the elements of $\mathbf{g}_{\Omega'}$ being 0 or $\Theta(1)$, the non-null elements of $\mathbf{P}_{\Omega',\Omega'}$ being $\Theta(\varepsilon^d)$, $d \geq 0$, and the elements of $\mathbf{u}$ being either $\Theta(1)$ or $o(1)$, it follows that the elements of $\mathbf{w}$ are $\Theta(1)$ or $o(1)$. Then, from (15) all elements of $\mathbf{q}_{\Omega'}$ will be $\Theta(1)$ or $o(1)$. But, by (14) the elements of $\mathbf{q}_{\Omega'}$ are greater than or equal to the elements of $\mathbf{w}$ and, by (13), the elements of $\mathbf{w}$ are greater than or equal to the elements of $\mathbf{g}_{\Omega'}^2$, and, then, $q_x = \Theta(1)$, $x \in D$. □

Let $\mathbf{B} = (B_{x,y})_{x,y\in\Omega}$ be the matrix defined by (note that, by condition C1, $P'_{x,y} \neq 0$ if and only if $P_{x,y} \neq 0$):

$$B_{x,y} = \begin{cases} 0 & \text{if } P_{x,y} = 0 \\ \dfrac{P_{x,y}^2}{P'_{x,y}} & \text{if } x \in U \ \wedge \ P_{x,y} \neq 0 \\ P_{x,y} & \text{if } x \in D \ \wedge \ P_{x,y} \neq 0 \end{cases} .$$

We have the following result:

Lemma 3. *Assume that conditions C1 and C2 hold. Then, for all sufficiently small $\varepsilon > 0$, $\sum_{n=0}^{\infty} \mathbf{B}_{U',U'}^n$ is finite.*

Proof. By construction, $B_{x,y} \geq 0$ and $B_{x,y} > 0$ if and only if $P_{x,y} \neq 0$ ($P_{x,y} > 0$). Then, $\mathbf{B}_{U',U'}$ is a positive matrix with same non-null pattern as $\mathbf{P}_{U',U'}$. Also, non-null elements $B_{x,y}$ are $\Theta(\varepsilon^{2d'_{x,y}})$, $d'_{x,y} \geq 0$, where $P_{x,y} = \Theta(\varepsilon^{d'_{x,y}})$. Then, $\mathbf{B}_{U',U'}$ can be written as $\mathbf{A} + \mathbf{C}$, with $\mathbf{A} = (A_{x,y})_{x,y\in U'} = \Theta(1)$ and $\mathbf{C} = o(\varepsilon)$, and where $A_{x,y} > 0$ if and only if $P_{x,y} = \Theta(1)$, which implies that the non-null pattern if $\mathbf{A}$ includes precisely the transitions in T_C. Then, since transitions in T_C do not build up cycles in X, the states in Ω' can be sorted so that $\mathbf{A}$ is strictly upper triangular, which implies $\mathbf{B}_{U',U'}^{|\Omega'|} = o(\varepsilon)$, $\|\mathbf{B}_{U',U'}^{|\Omega'|}\|_\infty = o(\varepsilon)$, from which:

$$\left\| \sum_{n=0}^{\infty} \mathbf{B}_{U',U'}^n \right\|_\infty \leq \sum_{n=0}^{|\Omega'|-1} \left\|\mathbf{B}_{U',U'}\right\|_\infty \sum_{k=0}^{\infty} \left\|\mathbf{B}_{U',U'}^{|\Omega'|}\right\|_\infty^k = \sum_{n=0}^{|\Omega'|-1} \left\|\mathbf{B}_{U',U'}^n\right\|_\infty + o(\varepsilon) ,$$

which implies the result. □

Note that

$$E_{P'}[Z'^2] = E_{P'}[Z^2 L^2] = \mathbf{B}_{\{r\},D}\mathbf{q}_D + \mathbf{B}_{\{r\},U'}\mathbf{a}_{U'} \,. \tag{16}$$

Let

$$\mathbf{z}_{U'} = \mathbf{B}_{U',D}\mathbf{q}_D \,. \tag{17}$$

We have:

$$\mathbf{a}_{U'} = \mathbf{B}_{U',U'}\mathbf{a}_{U'} + \mathbf{z}_{U'} \,,$$

and using Lemma 3:

$$\mathbf{a}_{U'} = \sum_{n=0}^{\infty} \mathbf{B}_{U',U'}^n \mathbf{z}_{U'} \quad \text{for all sufficiently small } \varepsilon > 0 \,. \tag{18}$$

Combining (16), (17) and (18) we get:

$$E_{P'}[Z'^2] = \left(\mathbf{B}_{\{r\},D} + \mathbf{B}_{\{r\},U'} \sum_{n=0}^{\infty} \mathbf{B}_{U',U'}^n \mathbf{B}_{U',D} \right) \mathbf{q}_D \quad \text{for all sufficiently small } \varepsilon > 0 \,. \tag{19}$$

Using (10) and (19) we can prove the desired result:

Theorem 1. *Assume that conditions C1 and C2 hold. Then,* $E_{P'}[Z'^2] = \Theta(\varepsilon^{2\rho})$ *and* $E_P[Z] = \Theta(\varepsilon^{\rho})$*,* $\rho \geq 0$*.*

Proof. Assume $\varepsilon > 0$ sufficiently small for the equality in (19) to hold. Using conditions C1 and C2, $B_{x,y} = 0$ if and only if $P_{x,y} = 0$, $x \in U$ and, for $B_{x,y} \neq 0$, $B_{x,y} = \Theta(\varepsilon^{2d'_{x,y}})$, $x \in U$, $P_{x,y} = \Theta(\varepsilon^{d'_{x,y}})$, $d'_{x,y} \geq 0$. Let the row vectors $\mathbf{u} = (u_x)_{x \in D} = \mathbf{P}_{\{r\},D} + \mathbf{P}_{\{r\},U'} \sum_{n=0}^{\infty} \mathbf{P}_{U',U'}^n \mathbf{P}_{U',D}$ and $\mathbf{w} = (w_x)_{x \in D} = \mathbf{B}_{\{r\},D} + \mathbf{B}_{\{r\},U'} \sum_{n=0}^{\infty} \mathbf{B}_{U',U'}^n \mathbf{B}_{U',D}$. We have (10) $E_P[Z] = \mathbf{u}\mathbf{v}_D$ and (19) $E_{P'}[Z'^2] = \mathbf{w}\mathbf{q}_D$. Since matrices $\mathbf{P}$ and $\mathbf{B}$ are positive and have the same non-null pattern, $u_x \neq 0$ if and only if $w_x \neq 0$, $x \in D$. Also, for x such that $u_x \neq 0$, $w_x = \Theta(\varepsilon^{2d'_x})$ where $d'_x \geq 0$ such that $u_x = \Theta(\varepsilon^{d'_x})$. The result follows, then, using Lemmas 1 and 2. □

The importance sampling schemes FB, BFB, FTDB, and BFTDB satisfy condition C1. Also, for balanced systems, all FB, BFB, FTDB, and BFTDB satisfy condition C2, and, for unbalanced systems, BFB and BFTDB satisfy condition C2. Then, we can conclude that, for the class of models considered in the paper, FB, BFB, FTDB, and BFTDB satisfy the bounded relative error property for balanced systems, and BFB and BFTDB satisfy the bounded relative error property for unbalanced systems.

4 Adapted New Importance Sampling Schemes

We start by motivating, for balanced fault-tolerant systems, the adapted importance sampling schemes. Towards that end, we will rank, for balanced fault-tolerant systems, the regenerative cycles in $\mathcal{S}$ according to the importance of their contributions to $E_P[Z]$. Remember that a balanced fault-tolerant system is a fault-tolerant system in which failure rates can be assumed to have the form

$\lambda_{x,y} = r_{\min} f_{x,y}\,\varepsilon$, $f_{x,y} \in (0,1]$, $f_{x,y} \gg \varepsilon$, where $\varepsilon = \lambda_{\max}/r_{\min}$ is the rarity parameter measuring how small failure transition rates are compared to repair transition rates. Repair transition rates have the form $\lambda_{x,y} = r_{\min} r_{x,y}$, $r_{x,y} \geq 1$. For a balanced fault-tolerant system

$$P_{x,y} = \frac{r_{\min} f_{x,y}\,\varepsilon}{\sum_{(x,z)\in T_F(x)} r_{\min} f_{x,z}\,\varepsilon} = \Theta(1)\,, \quad x \notin \Omega_{FR}, (x,y) \in T_F(x)\,,$$

$$P_{x,y} = \frac{r_{\min} f_{x,y}\,\varepsilon}{\sum_{(x,z)\in T_R(x)} r_{\min} r_{x,y} + \sum_{(x,z)\in T_F(x)} r_{\min} f_{x,z}\,\varepsilon} = \Theta(\varepsilon)\,, \quad x \in \Omega_{FR}, (x,y) \in T_F(x)\,,$$

$$P_{x,y} = \frac{r_{\min} r_{x,y}}{\sum_{(x,z)\in T_R(x)} r_{\min} r_{x,y} + \sum_{(x,z)\in T_F(x)} r_{\min} f_{x,z}\,\varepsilon} = \Theta(1)\,, \quad x \in \Omega_{FR}, (x,y) \in T_R(x)\,.$$

Also,

$$P_{x,y} = \frac{r_{\min} f_{x,y}\,\varepsilon}{\sum_{(x,z)\in T_R(x)} r_{\min} r_{x,y} + \sum_{(x,z)\in T_F(x)} r_{\min} f_{x,z}\,\varepsilon} \leq f_{x,y}\,\varepsilon\,, \quad x \in \Omega_{FR}, (x,y) \in T_F(x)\,,$$

$$h_x = \frac{1}{\sum_{(x,z)\in T_R(x)} r_{\min} r_{x,y} + \sum_{(x,z)\in T_F(x)} r_{\min} f_{x,z}\,\varepsilon} = \Theta(1)\,, \quad x \in \Omega_{FR}\,,$$

and

$$h_x \leq \frac{1}{r_{\min}}\,, \quad x \in \Omega_{FR}\,.$$

We will find useful to consider some subsets of regenerative cycles $\mathcal{S}_k$. The subset $\mathcal{S}_k$ includes the regenerative cycles which hit D and include k failure transitions from states $x \in \Omega_{FR}$. Let $k_{\min} = \min\{k \,:\, \mathcal{S}_k \neq \emptyset\}$. Because only regenerative cycles which hit D contribute to $E_P\{Z\}$, we have

$$E_P\{Z\} = \sum_{k=k_{\min}}^{\infty} C(k)\,,$$

where

$$C(k) = \sum_{\omega \in \mathcal{S}_k} P\{\omega\} Z(\omega)$$

is the contribution of the regenerative cycles in $\mathcal{S}_k$ to $E_P\{Z\}$. The motivating theorem (Theorem 4) ranking the importances of the contributions of the regenerative cycles to $E_P[Z]$ is preceded by some results. In the proofs which follow, we will use the parameters $E_{\max} = \max_{x\in E} |F(x)|$ and $F_{\max} = \max_{f\in F_B} |f|$. Informally, $E_{\max}$ is the maximum number of failed components in states with deferred repair and $F_{\max}$ is the maximum number of components which can fail simultaneously. For most fault-tolerant systems $E_{\max}$ and $F_{\max}$ will have moderate values. The proofs are generalizations of similar proofs performed in [6] for the particular case $E_{\max} = 0$. An upper bound on the length of the regenerative cycles in $\mathcal{S}_k$ can be easily found in terms of $E_{\max}$ and $F_{\max}$:

Lemma 4. *Let $\omega = (s_0, s_1, \ldots, s_l) \in \mathcal{S}_k$. Then, $l \leq 2E_{\max} + (k+1)F_{\max} + 1$.*

Proof. Consider the transitions (s_i, s_{i+1}), $0 \leq i < l$ of the regenerative cycle $\omega = (s_0, s_1, \ldots, s_l)$. Let f be the sum of the cardinalities of the failure bags associated with the failure transitions and let p be the sum of the cardinalities of the bags of components repaired in the repair transitions.

Obviously, $f = p$. It is clear that $f \leq E_{\max} + (k+1)F_{\max}$. Also, the number of repair transitions in the regenerative cycle is not smaller than $l - E_{\max} - k - 1$ and, then, $p \geq l - E_{\max} - k - 1$. Then, $l - E_{\max} - k - 1 \leq E_{\max} + (k+1)F_{\max}$, implying the result. □

We will start with the following result:

Theorem 2. *For balanced fault-tolerant systems and each k such that $\mathcal{S}_k \neq \emptyset$, $C(k) = \Theta(\varepsilon^k)$. Furthermore, for each $\omega \in \mathcal{S}_k$, $P\{\omega\}Z(\omega) = \Theta(\varepsilon^k)$.*

Proof. Let $\omega = (s_0, s_1, \ldots, s_l) \in \mathcal{S}_k$. We have

$$P\{\omega\} = \prod_{i=0}^{l-1} P_{s_i,s_{i+1}} .$$

$P_{s_0,s_1} = P_{r,s_1} = \Theta(1)$. Of the remaining $l-1$ factors, k factors correspond to failure transitions from states $x \in \Omega_{FR}$ and, therefore, are $\Theta(\varepsilon)$, and the remaining $l-k-1$ factors correspond to either failure transitions from states not in Ω_{FR} or to repair transitions and, therefore, are $\Theta(1)$. All together, this implies $P\{\omega\} = \Theta(\varepsilon^k)$. On the other hand,

$$Z(\omega) = \sum_{i=0}^{l-1} I_{s_i \in D} h_{s_i} ,$$

where, according to Lemma 4, $l \leq 2E_{\max} + (k+1)F_{\max} + 1$, and because $s_i \in D$ for some i, $0 \leq i < l-1$ and, for $s_i \in D$, $h_{s_i} = \Theta(1)$, we have $Z(\omega) = \Theta(1)$ and $P\{\omega\}Z(\omega) = \Theta(\varepsilon^k)$. To prove $C(k) = \Theta(\varepsilon^k)$, $\mathcal{S}_k \neq \emptyset$, note that

$$C(k) = \sum_{\omega \in \mathcal{S}_k} P\{\omega\}Z(\omega)$$

and because each term is $\Theta(\varepsilon^k)$ and, being $l \leq 2E_{\max} + (k+1)F_{\max} + 1$, $|\mathcal{S}_k|$ is finite, $C(k) = \Theta(\varepsilon^k)$. □

Let $k_{\min} = \min\{k \,:\, \mathcal{S}_k \neq \emptyset\}$. Using Theorem 2 we have the following corollary.

Corollary 1. *For balanced fault-tolerant systems, $C(k_{\min}) = \Theta(\varepsilon^{k_{\min}})$.*

According to Theorem 2, every contribution to $E_P\{Z\}$ $C(k)$, $k > k_{\min}$ is, for $\varepsilon \to 0$, negligible compared to $C(k_{\min})$. This, however, does not ensure that $\sum_{k=k_{\min}+1}^{\infty} C(k)$ will be negligible compared to $C(k_{\min})$. That result is established by the following theorem.

Theorem 3. *For balanced fault-tolerant systems, $\sum_{k=k_{\min}+1}^{\infty} C(k) = o(\varepsilon^{k_{\min}})$.*

The proof of Theorem 3 will be preceded by two propositions. For k such that $\mathcal{S}_k \neq \emptyset$, let

$$E_P\{Z \mid \mathcal{S}_k\} = \frac{\sum_{\omega \in \mathcal{S}_k} P\{\omega\}Z(w)}{\sum_{\omega \in \mathcal{S}_k} P\{\omega\}} = \frac{C(k)}{P\{\mathcal{S}_k\}} .$$

We have $C(k) = P\{\mathcal{S}_k\}E_P\{Z \mid \mathcal{S}_k\}$. The first proposition gives an upper bound for $E_P\{Z \mid \mathcal{S}_k\}$. The second one gives an upper bound for $P\{\mathcal{S}_k\}$.

Proposition 1. *For balanced fault-tolerant systems and k such that $\mathcal{S}_k \neq \emptyset$, $E_P\{Z|\mathcal{S}_k\} \leq (2E_{\max} + (k+1)F_{\max})/r_{\min}$.*

Proof. We prove $Z(\omega) \leq (2|C| + (k+1)F_{\max} - 1)/r_{\min}$, $\omega \in \mathcal{S}_k$, implying the result. Let $\omega = (s_0, s_1, \ldots, s_l) \in \mathcal{S}_k$. We have

$$Z(\omega) = \sum_{i=0}^{l-1} I_{s_i \in D} h_{s_i} .$$

$s_0 = r \notin D$. Therefore, since $x \in D$ implies $x \in \Omega_{FR}$, $Z(\omega)$ is the sum of a number of h_x, $x \in \Omega_{FR}$ no greater than $l-1$. Each h_x is upper bounded by $1/r_{\min}$ and, according to Lemma 4, $l - 1 \leq 2E_{\max} + (k+1)F_{\max}$. Then,

$$Z(\omega) \leq \frac{2E_{\max} + (k+1)F_{\max}}{r_{\min}} . \quad \square$$

Let $\widetilde{\mathbf{F}} = (I_{x \in \Omega_{FR} \wedge (x,y) \in T_F(x)} f_{x,y})_{x,y \in \Omega'}$. The upper bound for $P\{\mathcal{S}_k\}$ is in terms of $F_{\max}$, $\|\widetilde{\mathbf{F}}\|_\infty$ and ε.

Proposition 2. *For balanced fault-tolerant systems and $k \geq k_{\min}$,*

$$P\{\mathcal{S}_k\} \leq 2^{2E_{\max} + F_{\max} - 1} \left(2^{F_{\max}} \|\widetilde{\mathbf{F}}\|_\infty \varepsilon\right)^k .$$

Proof. Let $\widetilde{\mathbf{P}} = (P(x,y))_{x,y \in \Omega'}$. We can partition $\widetilde{\mathbf{P}}$ as

$$\widetilde{\mathbf{P}} = \widetilde{\mathbf{R}} + \widetilde{\mathbf{\Lambda}} ,$$

where

$$\widetilde{\mathbf{R}} = (I_{x \notin \Omega_{FR} \vee x \in \Omega_{FR} \wedge (x,y) \in T_R(x)} P(x,y))_{x,y \in \Omega'}$$

collects failure transition probabilities from states not in Ω_{FR} and repair transition probabilities from states in Ω_{FR} and

$$\widetilde{\mathbf{\Lambda}} = (I_{x \in \Omega_{FR} \wedge (x,y) \in T_F(x)} P(x,y))_{x,y \in \Omega'}$$

collects failure transition probabilities from states in Ω_{FR}. According to the definition of $\widetilde{\mathbf{F}}$, we have $\widetilde{\mathbf{\Lambda}} \leq \varepsilon \widetilde{\mathbf{F}}$, where the inequality between matrices means inequality between every pair of corresponding elements of the matrices. Let $\mathbf{u}$ be the column vector $(P_{r,x})_{x \in \Omega'}$ and let $\mathbf{v}$ be the column vector $(P_{x,r})_{x \in \Omega'}$. Consider $\widetilde{\mathbf{P}}^n = (\widetilde{\mathbf{R}} + \widetilde{\mathbf{\Lambda}})^n$ and let $F(n,k)$, $n \geq k$ be the set of factors $\widetilde{\mathbf{A}}_m^{n,k}$, $1 \leq m \leq \binom{n}{k}$ of the expansion of $(\widetilde{\mathbf{R}} + \widetilde{\mathbf{\Lambda}})^n$ including exactly k times $\widetilde{\mathbf{\Lambda}}$ and $n-k$ times $\widetilde{\mathbf{R}}$. According to Lemma 4, regenerative cycles $w = (s_0, s_1, \ldots, s_l)$ in $\mathcal{S}_k$ include at most $2|C| + (k+1)F_{\max}$ transitions. The first transition is from r to a state $x \in \Omega'$, the following $l-2$ transitions are between states in Ω', the last (repair) transition is from a state $x \in \Omega'$ to r. Then, denoting by $\mathbf{u}^T$ the

transpose of $\mathbf{u}$, we have

$$\begin{aligned} P\{\mathcal{S}_k\} &= \sum_{n=k}^{2E_{\max}+(k+1)F_{\max}-1} \sum_{\widetilde{\mathbf{A}}_m^{n,k}\in F(n,k)} \mathbf{u}^T \widetilde{\mathbf{A}}_m^{n,k}\mathbf{v} = \sum_{n=k}^{2|C|+(k+1)F_{\max}-2} \sum_{\widetilde{\mathbf{A}}_m^{n,k}\in F(n,k)} \|\mathbf{u}^T \widetilde{\mathbf{A}}_m^{n,k}\mathbf{v}\|_\infty \\ &\leq \sum_{n=k}^{2E_{\max}+(k+1)F_{\max}-1} \sum_{\widetilde{\mathbf{A}}_m^{n,k}\in F(n,k)} \|\mathbf{u}^T\|_\infty \, \|\widetilde{\mathbf{A}}_m^{n,k}\|_\infty \, \|\mathbf{v}\|_\infty \,. \end{aligned}$$

Trivially, $\|\mathbf{u}^T\|_\infty = 1$. Also, $\|\mathbf{v}\|_\infty \leq 1$ and $\|\widetilde{\mathbf{A}}_m^{n,k}\|_\infty \leq \|\widetilde{\mathbf{R}}\|_\infty^{n-k} \, \|\widetilde{\mathbf{\Lambda}}\|_\infty^k$, and, because $\|\widetilde{\mathbf{R}}\|_\infty \leq 1$ and $\|\widetilde{\mathbf{\Lambda}}\|_\infty \leq \varepsilon \, \|\widetilde{\mathbf{F}}\|_\infty$, $\|\widetilde{\mathbf{A}}_m^{n,k}\|_\infty \leq \|\widetilde{\mathbf{F}}\|_\infty^k \, \varepsilon^k$. Then,

$$\begin{aligned} P\{\mathcal{S}_k\} &\leq \sum_{n=k}^{2E_{\max}+(k+1)F_{\max}-1} \sum_{\widetilde{\mathbf{A}}_m^{n,k}\in F(n,k)} \|\widetilde{\mathbf{F}}\|_\infty^k \, \varepsilon^k = \sum_{n=k}^{2E_{\max}+(k+1)F_{\max}-1} \binom{n}{k} \|\widetilde{\mathbf{F}}\|_\infty^k \, \varepsilon^k \\ &= \|\widetilde{\mathbf{F}}\|_\infty^k \, \varepsilon^k \sum_{n=k}^{2E_{\max}+(k+1)F_{\max}-1} \binom{n}{k} \leq \|\widetilde{\mathbf{F}}\|_\infty^k \, \varepsilon^k \, 2^{2|C|+(k+1)F_{\max}-2} \\ &= 2^{2E_{\max}+F_{\max}-1} \left(2^{F_{\max}} \, \|\widetilde{\mathbf{F}}\|_\infty \, \varepsilon\right)^k . \quad \square \end{aligned}$$

Proof of Theorem 3. We start from

$$\sum_{k=k_{\min}+1}^{\infty} C(k) = \sum_{k=k_{\min}+1}^{\infty} P\{\mathcal{S}_k\} E_P\{Z \mid \mathcal{S}_k\} \,.$$

Using Propositions 1 and 2:

$$\begin{aligned} \sum_{k=k_{\min}+1}^{\infty} C(k) &< \sum_{k=k_{\min}+1}^{\infty} \frac{2E_{\max}+(k+1)F_{\max}}{r_{\min}} \, 2^{2E_{\max}+F_{\max}-1} \left(2^{F_{\max}} \, \|\widetilde{\mathbf{F}}\|_\infty \, \varepsilon\right)^k \\ &= A_1 \sum_{k=k_{\min}+1}^{\infty} k(B\varepsilon)^k + A_2 \sum_{k=k_{\min}+1}^{\infty} (B\varepsilon)^k \end{aligned}$$

with

$$A_1 = \frac{2E_{\max}+F_{\max}}{r_{\min}} \, 2^{2E_{\max}+F_{\max}-1} \,,$$
$$A_2 = \frac{F_{\max}}{r_{\min}} \, 2^{2E_{\max}+F_{\max}-1} \,,$$
$$B = 2^{F_{\max}} \, \|\widetilde{\mathbf{F}}\|_\infty \,.$$

Using $\sum_{k=k_{\min}+1}^{\infty} a^k = a^{k_{\min}+1}/(1-a)$, $0 < a < 1$ and

$$\sum_{k=k_{\min}+1}^{\infty} k a^k = a^{k_{\min}+1} \left(\frac{k_{\min}+1}{1-a} + \frac{a}{(1-a)^2}\right) , \quad 0 < a < 1 \,,$$

which follows easily from (see, for instance, [22]) $\sum_{k=0}^{\infty} k a^k = a/(1-a)^2$, we have, for $\varepsilon \to 0$:

$$\sum_{k=k_{\min}+1}^{\infty} C(k) < A_1 (B\varepsilon)^{k_{\min}+1} \left(\frac{k_{\min}+1}{1-B\varepsilon} + \frac{B\varepsilon}{(1-B\varepsilon)^2}\right) + A_2 \frac{(B\varepsilon)^{k_{\min}+1}}{1-B\varepsilon} = o(\varepsilon^{k_{\min}}) \,. \quad \square$$

The sought result ranking the importances of the contributions of the regenerative cycles to $E_P[Z]$ is:

Theorem 4. *For balanced fault-tolerant systems,*

a) $C(k_{\min})/E_P\{Z\} = 1 + o(1)$,

b) $P\{\omega\}Z(\omega)/E_P\{Z\} = \Theta(1)$, $\omega \in S_{k_{\min}}$,

c) $\sum_{k=k_{\min}+1}^{\infty} C(k)/E_P\{Z\} = o(1)$.

Proof. According to Corollary 1, $C(k_{\min}) = a\,\varepsilon^{k_{\min}} + o(\varepsilon^{k_{\min}})$, $a > 0$. Using Theorem 3:

$$E_P\{Z\} = C(k_{\min}) + \sum_{k=k_{\min}+1}^{\infty} C(k) = a\,\varepsilon^{k_{\min}} + o(\varepsilon^{k_{\min}})\,.$$

Then,

$$\frac{C(k_{\min})}{E_P\{Z\}} = \frac{a\,\varepsilon^{k_{\min}} + o(\varepsilon^{k_{\min}})}{a\,\varepsilon^{k_{\min}} + o(\varepsilon^{k_{\min}})} = 1 + o(1)\,,$$

proving a). Using Theorem 2, $P\{\omega\}Z(\omega) = a_\omega \varepsilon^{k_{\min}} + o(\varepsilon^{k_{\min}})$, $a_\omega > 0$, $\omega \in \mathcal{S}_{k_{\min}}$. Then, for $\omega \in \mathcal{S}_{k_{\min}}$:

$$\frac{P\{\omega\}Z(\omega)}{E_P\{Z\}} = \frac{a_\omega \varepsilon^{k_{\min}} + o(\varepsilon^{k_{\min}})}{a\,\varepsilon^{k_{\min}} + o(\varepsilon^{k_{\min}})} = \Theta(1)\,,$$

proving b). Finally, using Theorem 3,

$$\frac{\sum_{k=k_{\min}+1}^{\infty} C(k)}{E_P\{Z\}} = \frac{o(\varepsilon^{k_{\min}})}{a\,\varepsilon^{k_{\min}} + o(\varepsilon^{k_{\min}})} = o(1)\,,$$

proving c). □

Importance sampling theory suggests that regenerative cycles should be sampled with probabilities close to the relative contributions of the cycles to $E_P[Z]$. Then, according to Theorem 4 the biased sampling probabilities should be chosen so that the probability of sampling cycles in $\mathcal{S}_{k_{\min}}$ be close to 1 and, furthermore, all cycles in $\mathcal{S}_{k_{\min}}$ be sampled with a probability $\Theta(1)$. A natural way of doing that would be to use modified FTDB and BFTDB importance sampling schemes in which failure transitions are considered dominant when they belong to paths which, starting from the given state x, would hit D after a minimum number of failure transitions from states in Ω_{FR}. Identification of those *dominant* failure transitions would require to consider both $F(x)$ and the functions $\Phi(b)$, $b \subseteq C$ and $\Psi(b)$, $b \subseteq C$. It seems doubtful that an efficient procedure for performing such identification can be devised. The alternative we propose in this paper is to adapt FTDB and BFTDB so that dominance biasing is only performed in the states $x \in \Omega_{FR}$. This provides focusing into the regenerative cycles which, after entering Ω_{FR} hit D after a minimum number of failure transitions from states in Ω_{FR}. For balanced fault-tolerant systems, this is better than both FB and BFB. The adapted FTDB and BFTDB importance sampling schemes can be implemented knowing the failure transition distances from the currently sampled state and all its successors through failure transitions, which can be computed efficiently using the procedures described in [6].

To clarify, the biased transition probabilities in the adapted FTDB scheme (AFTDB) would be:

$$P'_{x,y} = \begin{cases} \dfrac{P_{x,y}}{\sum_{z\,:\,(x,z)\in T_D(x)} P_{x,z}}\, FB \times DB & \text{if } x \in \Omega_{FR} \cap \Omega_D \ \wedge\ (x,y) \in T_D(x)\,, \\ \dfrac{P_{x,y}}{\sum_{z\,:\,(x,z)\in T_{ND}(x)} P_{x,z}}\, FB(1-DB) & \text{if } x \in \Omega_{FR} \cap \Omega_D \ \wedge\ (x,y) \in T_{ND}(x)\,, \\ \dfrac{P_{x,y}}{\sum_{z\,:\,(x,z)\in T_F(x)} P_{x,z}}\, FB & \text{if } x \in \Omega_{FR} - \Omega_D \ \wedge\ (x,y) \in T_F(x)\,, \\ P_{x,y} & \text{if } x \notin \Omega_{FR}\,, \\ \dfrac{P_{x,y}}{\sum_{z\,:\,(x,z)\in T_R(x)} P_{x,z}}\,(1-FB) & \text{if } x \in \Omega_{FR} \ \wedge\ (x,y) \in T_R(x)\,, \end{cases}$$

and the biased transition probabilities in the adapted BFTDB scheme (ABFTDB) would be:

$$P'_{x,y} = \begin{cases} \dfrac{FB \times DB}{|T_D(x)|} & \text{if } x \in \Omega_{FR} \cap \Omega_D \ \wedge\ (x,y) \in T_D(x)\,, \\ \dfrac{FB(1-DB)}{|T_{ND}(x)|} & \text{if } x \in \Omega_{FR} \cap \Omega_D \ \wedge\ (x,y) \in T_{ND}(x)\,, \\ \dfrac{FB}{|T_F(x)|} & \text{if } x \in \Omega_{FR} - \Omega_D \ \wedge\ (x,y) \in T_F(x)\,, \\ \dfrac{1}{|T_F(x)|} & \text{if } x \notin \Omega_{FR}\,, \\ \dfrac{P_{x,y}}{\sum_{z\,:\,(x,z)\in T_R(x)} P_{x,z}}\,(1-FB) & \text{if } x \in \Omega_{FR} \ \wedge\ (x,y) \in T_R(x)\,. \end{cases}$$

It should be clear that AFTDB satisfies the bounded relative error property for balanced fault-tolerant systems and ABFTDB satisfies the bounded relative error property for both balanced and unbalanced fault-tolerant systems.

5 Analysis

In this section we will compare the performances of the AFTDB and ABFTDB importance sampling schemes with those of FB and BFB. Our implementation of the simulation method optimizes the distribution of the regenerative cycles between the biased stream (used to estimate $E_P[Z]$) and the unbiased stream (used to estimate $E_P[W]$). It also optimizes the biasing parameters of the importance sampling schemes. That implementation is given in [6].

We will consider two examples. The first example (FTD) is a fault-tolerant database system similar to that described in [10]. The system contains two sets of processors, A and B, with three processors per set, two sets of disk controllers with two controllers per set, and six disk sets with four disks per set. Each set of controllers commands three disk sets. The system is up if and only if at

least one processor in each set, one controller in each set, and at least three disks in each disk set are operational. Repair is deferred in the states with no more than one failed processor in each set and no other component failed. In each processor set there is one operating processor, assuming that some processor is operational. Components do not fail when the system is down. When the operating processor of set A fails, it has a probability P_{P} of causing the operating processor of set B to fail. Each component in the system has two failed modes which occur with equal probabilities. Repair rates for all components are $1\ \mathrm{h}^{-1}$ in one mode and $1/2\ \mathrm{h}^{-1}$ in the other mode. Components are repaired by one repairman who chooses components at random from the set of failed components. Two instances of the example will be considered. In instance I, $P_{\mathrm{P}} = 0.10$, processors fail with rate $\lambda_{\mathrm{P}} = 10^{-5}\ \mathrm{h}^{-1}$, controllers fail with rate $\lambda_{\mathrm{C}} = 10^{-5}\ \mathrm{h}^{-1}$, and disks fail with rate $\lambda_{\mathrm{D}} = 10^{-5}\ \mathrm{h}^{-1}$. In instance II, $P_{\mathrm{P}} = 0.01$, processors fail with rate $\lambda_{\mathrm{P}} = 10^{-6}\ \mathrm{h}^{-1}$, controllers fail with rate $\lambda_{\mathrm{C}} = 10^{-6}\ \mathrm{h}^{-1}$, and disks fail with rate $\lambda_{\mathrm{D}} = 10^{-5}\ \mathrm{h}^{-1}$. For instance I, $f_{\min}/f_{\max} = 0.025$ and $\varepsilon = f_{\max}/r_{\min} = 4 \times 10^{-4}$ and, therefore, the instance can be considered a balanced fault-tolerant system. For instance II, $f_{\min}/f_{\max} = 2.5 \times 10^{-4}$ and $\varepsilon = f_{\max}/r_{\min} = 4 \times 10^{-4}$ and, therefore, the instance can be considered an unbalanced fault-tolerant system.

The second example (FTC) is ours and is the fault-tolerant control system whose architecture is depicted in Figure 1. A dual configuration of data processing units (DPU) command control subsystems located at remote sites. Each control subsystem comprises two redundant control units (CU) working in hot standby redundancy. The system can be accessed through two redundant front-ends (FE) connected to the DPU. The DPU and the CU communicate using two local area networks (LAN), La, Lb, to which each DPU and each CU has access through dedicated communication processors (CP). FE, DPU, CU, CP, and LAN fail with rates λ_{FE}, λ_{DPU}, λ_{CU}, λ_{CP}, and λ_{L}, respectively. Two failed modes are considered for DPU: "soft" and "hard". The first mode occurs with probability P_{S} and can be recovered by a restart; the second mode occurs with probability $1 - P_{\mathrm{S}}$ and requires hardware repair. Coverage is assumed perfect for all faults. There are three repairpersons. The first one repairs LAN and CP with preemptive priority given to LAN. The second one repairs FE, CU and DPU in "hard" failed mode, with preemptive priority given first to DPU, next to FE, and last to CU. The third one makes DPU restarts. Failed components with the same priority are chosen at random for repair/restart. The repair rates of LAN, CP, FE, CU and DPU in "hard" failed mode are denoted by, respectively, μ_{L}, μ_{CP}, μ_{FE}, μ_{CU}, μ_{DPUh}. The restart rate of DPU in "soft" failed mode is denoted by μ_{DPUs}. The system is up if and only if there is an operational FE and one operational DPU can communicate with at least one operational CU of each control subsystem. Different LAN can be used for communication between the DPU and the CU of each control subsystem, but the communication has to be direct, i.e. involving only one CP of the DPU, one CP of the CU and one LAN. Components do not fail when the system is down. Repair is deferred in the states with no more than one CP failed and no other component failed. The front-ends can be conceptualized as being instances of the same component class. However, the interconnection relationships make it mandatory to consider all the other components as unique representatives of different component classes. We use the four sets of model parameter values given in Table 1. For set A, $f_{\min}/f_{\max} = 0.0556$ and $\varepsilon = f_{\max}/r_{\min} = 2.88 \times 10^{-4}$; for set B, $f_{\min}/f_{\max} = 0.0556$ and $\varepsilon = f_{\max}/r_{\min} = 2.88 \times 10^{-3}$; for set C, $f_{\min}/f_{\max} = 5.56 \times 10^{-3}$ and

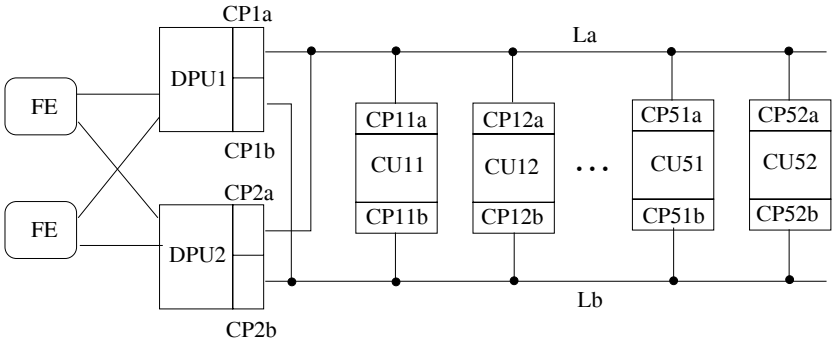

Figure 1: Architecture of the fault-tolerant control system (FTC example).

Table 1: Sets of model parameter values for the FTC example.

set	A	B	C	D
λ_{FE}	2×10^{-6}	2×10^{-6}	2×10^{-6}	2×10^{-6}
λ_{DPU}	10^{-5}	10^{-5}	2×10^{-5}	4×10^{-5}
λ_{CU}	2×10^{-6}	2×10^{-6}	4×10^{-7}	4×10^{-7}
λ_{L}	10^{-6}	10^{-6}	10^{-6}	10^{-6}
λ_{CP}	5×10^{-7}	5×10^{-7}	10^{-7}	10^{-4}
P_{S}	0.9	0.9	0.9	0.9
μ_{FE}	0.5	0.05	0.05	0.05
μ_{DPUh}	0.5	0.05	0.05	0.05
μ_{DPUs}	4	0.4	0.4	0.4
μ_{CU}	0.5	0.05	0.05	0.05
μ_{L}	0.2	0.02	0.02	0.02
μ_{CP}	0.5	0.05	0.05	0.05

$\varepsilon = f_{\max}/r_{\min} = 5.76 \times 10^{-3}$; for set D, $f_{\min}/f_{\max} = 4 \times 10^{-3}$ and $\varepsilon = f_{\max}/r_{\min} = 5 \times 10^{-3}$. Thus, the fault-tolerant system can be considered balanced for sets A and B and unbalanced for sets C and D. Furthermore, for set D, there are regenerative cycles outside $\mathcal{S}_{k_{\min}}$ with significant relative contributions to $E_P\{Z\}$, and, therefore, that set tests the behavior of AFTDB ABFTDB in a hard scenario which defies the heuristic supporting those importance sampling schemes.

For the examples corresponding to balanced fault-tolerant systems we will compare the performance of AFTDB with that of FB. For the examples corresponding to unbalanced fault-tolerant systems we will compare the performance of ABFTDB with that of BFB. The simulation method is run with a target 99% confidence interval of $\pm 0.2\%$ and a maximum number of regenerative cycles max_rc = 10,000,000. As initial value for FB in FB and BFB we take 0.5. As initial values for FB and DB in AFTDB and ABFTDB we take 0.8. For the K parameter described in [6] we took a value 1,000. All CPU times are measured on a workstation with a Sun-Blade-1000 processor. Table 2 summarizes the obtained results. We give the estimate, number of regenerative cycles and CPU times under AFTDB (ABFTDB), and the slow down factor of FB (BFB) defined as the ratio between the CPU times required under FB (BFB) and the CPU time required under AFTDB (ABFTDB) to achieve a confidence interval of same relative halfwidth. When the target confidence

Table 2: Comparison of importance sampling schemes.

example	estimate	cycles	CPU time in s	slow down factor
FTD (I)	$1.9634 \times 10^{-8} \pm 3.90 \times 10^{-11}$	4,331,000	158.8	24.5
FTD (II)	$1.6542 \times 10^{-8} \pm 3.30 \times 10^{-11}$	4,430,000	216.0	13.6
FTC (A)	$2.7550 \times 10^{-10} \pm 5.50 \times 10^{-13}$	7,032,000	4,295	66.1
FTC (B)	$2.7569 \times 10^{-8} \pm 5.50 \times 10^{-11}$	6,999,000	4,262	66.2
FTC (C)	$1.9720 \times 10^{-8} \pm 4.32 \times 10^{-11}$	10,001,000	8,652	48.3
FTC (D)	$3.9010 \times 10^{-7} \pm 3.18 \times 10^{-9}$	10,008,000	11,397	235

interval is not achieved, we compute the slow down factor using estimates for the CPU times which would be required to achieve it, based on the rule that CPU time is proportional to the inverse of the square of the relative confidence interval halfwidth. The results show that for balanced fault-tolerant systems AFTDB can speed up significantly FB and for unbalanced fault-tolerant systems ABFTDB can speed up significantly BFB. Overall, simulation under the new importance sampling schemes seems to be efficient making it possible to obtain highly accurate estimates in affordable CPU times.

6 Conclusions

We have proposed new importance sampling schemes for the efficient simulation of CTMC models of fault-tolerant systems with deferred repair. The new schemes have been proved to be robust. We have also proved that previously proposed importance sampling schemes, FB and BFB, are robust for the considered class of CTMC models. The new importance sampling schemes have been motivated theoretically for balanced fault-tolerant systems and, for those systems, are guaranteed to be more efficient than FB and BFB. Numerical analysis using representative examples has shown that the new importance sampling schemes can achieve significant speedups over FB and BFB for both balanced systems and unbalanced systems. Using the new importance sampling schemes it is possible to achieve highly accurate estimates in reasonable CPU times. Alternatively, under the new importance sampling schemes, it is possible to obtain estimates of reasonable accuracy in very small CPU times, opening the way to simulation based system optimization.

References

[1] C. Alexopoulos and B. C. Shultes, "The Balanced Likelihood Ratio Method for Estimating Performance Measures of Highly Reliable Systems," in *Proc. Winter Simulation Conference*, Piscataway, New Jersey, 1998.

[2] C. Alexopoulos and B. C. Shultes, "Estimating Reliability Measures for Highly-reliable Markov Systems, Using Balanced Likelihood Ratios," *IEEE Trans. on Reliability*, vol. 50, no. 3, September 2001, pp. 265-280.

[3] A. Berman and R. J. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*, SIAM, 1994.

[4] J. A. Carrasco , "Failure Distance-based Simulation of Repairable Fault-Tolerant Systems," in *Computer Performance Evaluation*, Elsevier, 1992, pp. 351–365.

[5] J. A. Carrasco and V. Suñé, "An Algorithm to Find Minimal Cuts of Coherent Fault Trees with Event Classes Using a Decision Tree," *IEEE Trans. on Reliability*, vol. 48, no. 1, March 1999, pp. 31–41.

[6] J. A. Carrasco, "Failure Transition Distance-Based Importance Sampling Schemes for the Simulation of Repairable Fault-Tolerant Computer Systems," May 2005, to appear in *IEEE Trans. on Reliability*.

[7] A. E. Conway and A. Goyal, "Monte Carlo Simulation of Computer Systems Availability/Reliability Models," in *Proc. 17th IEEE Int. Symp. on Fault-Tolerant Computing*, 1987, pp. 230–235.

[8] G. Dahlquist and Å. Björck, *Numerical Methods*, Prentice-Hall, 1974.

[9] A. Goyal, P. Heidelberger and P. Shahabuddin, "Measure Specific Dynamic Importance Sampling for Availability Simulations," in *Proc. 1987 Winter Simulation Conference*, A. Thesen, H. Grant and W. D. Kelton (eds.), 1987, pp. 351–357.

[10] A. Goyal, P. Shahabuddin, P. Heidelberger, V. F. Nicola, and P. W. Glynn, "A Unified Framework for Simulating Markovian Models of Highly Dependable Systems," *IEEE Trans. on Computers*, vol. 42, no. 1, January 1992, pp. 36–51.

[11] J. M. Hammersley and D. C. Handscomb, *Monte Carlo Methods*, Metheun, 1964.

[12] A. Hordijk, D. L. Iglehart, and R. Schassberger, "Discrete Time Methods for Simulating Continuous Time Markov Chains," *Advances in Applied Probability*, vol. 8, 1976, pp. 772-788.

[13] S. Juneja and P. Shahabuddin, "Fast Simulation of Markov Chains with Small Transition Probabilities," *Management Science*, vol. 47, no. 4, April 2001, pp. 547–562.

[14] S. Juneja and P. Shahabuddin, "Splitting-Based Inportance-Sampling Algorithm for Fast Simulation of Markov Reliability Models with General Repair-Policies," *IEEE Trans. on Reliability*, vol. 50, no. 3, September 2001, pp. 235–245.

[15] E. E. Lewis and F. Böhm, "Monte Carlo Simulation of Markov Unreliability Models," *Nuclear Engineering and Design*, vol. 77, 1984, pp. 49–62.

[16] M. K. Nakayama, "A Characterization of the Simple Failure Biasing Method for Simulations of Highly Reliable Markovian Systems," *ACM Trans. on Modeling and Computer Simulation*, vol. 4, 1994, pp. 52–88.

[17] M. K. Nakayama, "General Conditions for Bounded Relative Error in Simulations of Highly Reliable Markovian Systems," *Advances in Applied Probability*, vol. 28, 1996, pp. 687–727.

[18] A. Reibman and K. S. Trivedi, "Numerical Transient Analysis of Markov Models," *Computers and Operations Research*, vol. 15, 1988, pp. 19–36.

[19] P. Shahabuddin, V. F. Nicola, P. Heidelberger, A. Goyal, and P. W. Glynn, "Variance Reduction in Mean Time to Failure Simulations," in *Proc. 1988 Winter Simulation Conference*, M. Abrams, P. Haigh and J. Comfort (eds.), 1988, pp. 491–499.

[20] P. Shahabuddin, *Simulation and Analysis of Highly Reliable Systems*, Ph. D. thesis, Stanford University, 1990.

[21] P. Shahabuddin, "Importance Sampling for the Simulation of Highly Reliable Markovian Systems," *Management Science*, vol. 40, no. 3, March 1994, pp. 333–352.

[22] M. R. Spiegel, *Mathematical Handbooh of Formulas and Tables*, McGraw-Hill, 1968.

[23] W. J. Stewart, *Introduction to the Numerical Solution of Markov Chains*, Princeton University Press, 1994.

[24] T. Zhuguo and E. E. Lewis, "Component Dependency Models in Markov Monte Carlo Simulation," *Reliability Engineering*, vol. 13, 1985, pp. 45–61.