Failure Transition Distance-Based Importance Sampling Schemes for theSimulation of Repairable Fault-Tolerant Computer Systems

Juan A. Carrasco

***Abstract*—Markov models are often used to evaluate dependability attributes of fault-tolerant computer systems. The use in practice of Markov models is, however, hampered by the well-known state space explosion problem. Simulation alleviates the problem. For Markov models of repairable fault-tolerant systems, standard simulation of dependability measures tends to be expensive due to the rarity of the system failure event. Importance sampling can speed up the simulation.**

This paper develops two importance sampling schemes, called *failure transition distance biasing* & *balanced failure transition distance biasing*, which exploit the failure transition distance concept in an attempt to improve the efficiency of two other schemes, failure biasing & and balanced failure biasing. The schemes require the computation of the so-called failure transition distances, and procedures to perform those computations are developed. The presentation is tied to a previously proposed measure-specific simulation method for the steady-state unavailability. An optimization method of the parameters of the importance sampling schemes is also developed. For the simulation of the steady-state unavailability, failure transition distance biasing has (as failure biasing) the bounded relative error property for balanced fault-tolerant systems & balanced failure transition distance biasing has (as balanced failure biasing) the bounded relative error property for both balanced & unbalanced fault-tolerant systems. It is proved that, for balanced fault-tolerant systems, both failure transition distance biasing & balanced failure transition distance biasing can indeed improve the efficiency of failure biasing & balanced failure biasing. In addition, numerical experiments seem to indicate that, for unbalanced fault-tolerant systems, balanced failure transition distance biasing can also improve the efficiency of balanced failure biasing. The application of the failure transition distance-based importance sampling schemes is, however, limited to systems not having too many minimal failure covers, or, at least, not having too many minimal failure covers of small cardinality. A minimal failure cover is a minimal bag of failure bags such that the failure of its components implies the failure of the system; a failure bag is any non-empty bag of component classes which can fail simultaneously.

***Index Terms*—Fault-tolerant computer systems, importance sampling, Markov models, rare event simulation, steady-state availability, variance reduction.**

Manuscript received August 1998; revised November 1999, October 2001, June 2003, August 2004, July 2005, September 2005, November 2005, January 2006, February 2006. This work was supported by the "Comisión Interministerial de Ciencia y Tecnología" (CICYT) of the Ministry of Science and Technology of Spain under the Research Grants TIC95-0707-C02-02 and TAP1999-0443-C05-05. Associate Editor: C. Alexopoulos.

The author is with the Departament d'Enginyeria Electrònica, Universitat Politècnica de Catalunya, Barcelona 08028, Spain (e-mail: carrasco@eel.upc.edu).

Digital Object Identifier 10.1109/TR.2006.874910

ACRONYM[1]

CTMC	homogeneous continuous-time Markov chain
DTMC	homogeneous discrete-time Markov chain
FB	failure biasing
BFB	balanced failure biasing
FTDB	failure transition distance biasing
BFTDB	balanced failure transition distance biasing
BRE	bounded relative error
LAN	local area network

NOTATION, AND DEFINITIONS

$a \subseteq b$	a is a subset of b
$a \subset b$	a is a proper subset of b, i.e. $a \subseteq b$ & $a \neq b$
Bag[2]	collection of possibly repeated elements from a given domain $\mathcal{E}$
$\#(x, b)$	number of instances of x in a bag b
$x \in b$	b being a bag, $x \in b$ iff $\#(x, b) > 0$
$c_1[n_1] \cdots c_k[n_k]$	with $k > 0$, $c_i \neq c_j$, $1 \leq i, j \leq k$, $i \neq j$, and $n_i > 0$, $1 \leq i \leq k$, bag b with domain $\mathcal{E} \supseteq \{c_1, \ldots, c_k\}$ such that $\#(c_i, b) = n_i$, $1 \leq i \leq k$ & $\#(c, b) = 0$, $c \in \mathcal{E} - \{c_1, \ldots, c_k\}$
$\emptyset$	empty bag: bag with no elements
$\|b\|$	b being a bag with domain $\mathcal{E}$, cardinality of b, i.e. $\sum_{x \in \mathcal{E}} \#(x, b)$
$a \subseteq b$	a & b being bags with domain $\mathcal{E}$, $a \subseteq b$ denotes that a is a subbag of b, i.e. that $\#(x, a) \leq \#(x, b)$, $x \in \mathcal{E}$
$a = b$	a & b being bags with domain $\mathcal{E}$, $a = b$ denotes that a & b are equal, i.e. that $\#(x, a) = \#(x, b)$, $x \in \mathcal{E}$
$a \neq b$	a & b being bags with domain $\mathcal{E}$, $a \neq b$ denotes that a & b are not equal, i.e. that $a = b$ does not hold
$a \subset b$	a & b being bags with domain $\mathcal{E}$, $a \subset b$ denotes that a is a proper subbag of b, i.e. that $a \subseteq b$, and $a \neq b$

[1]The singular and plural of an acronym are always spelled the same.

[2]A brief exposition of bag theory with the same notation as we follow can be found in [23].

$a + b$	bag sum: a & b being bags with domain $\mathcal{E}$, bag c with domain $\mathcal{E}$ such that $\#(x,c) = \#(x,a) + \#(x,b), x \in \mathcal{E}$
$a \cup b$	bag union: a & b being bags with domain $\mathcal{E}$, bag c with domain $\mathcal{E}$ such that $\#(x,c) = \max\{\#(x,a), \#(x,b)\}$, $x \in \mathcal{E}$
$a \cap b$	bag intersection: a & b being bags with domain $\mathcal{E}$, bag c with domain $\mathcal{E}$ such that $\#(x,c) = \min\{\#(x,a), \#(x,b)\}, x \in \mathcal{E}$
$a - b$	bag difference: a & b being bags with domain $\mathcal{E}$, bag c with domain $\mathcal{E}$ such that $\#(x,c) = \max\{\#(x,a) - \#(x,b), 0\}, x \in \mathcal{E}$
minimal bag	a bag b is minimal with respect to some condition $c(b)$ which is increasing with b ($c(b)$ satisfied implies $c(b')$ satisfied for $b' \supseteq b$) if $c(b)$ is satisfied, and, for no bag $b' \subset b$, $c(b')$ is satisfied. Condition $c(b)$ is satisfied iff $b \supseteq d$, where d is a minimal bag[3]
generalized structure function of a system made up of a bag of component classes E[4]	binary-valued function $\Psi(b)$, $b \subseteq E$ such that $\Psi(b) = 1$ iff the system is up when the bag of operational component classes of the system is b
increasing generalized structure function of a system made up of a bag of component classes E[4]	generalized structure function $\Psi(b)$, $b \subseteq E$ such that $\Psi(b) \geq \Psi(b')$, $b \supseteq b'$, $b, b' \subseteq E$
cut of a system made up of a bag of component classes E with increasing generalized structure function $\Psi(b)$, $b \subseteq E$[4]	bag $b \subseteq E$ such that $\Psi(E - b) = 0$, i.e. the system is down when the bag of failed component classes of the system is b
minimal cut of a system made up of a bag of component classes E with increasing generalized structure function $\Psi(b)$,$b \subseteq E$[4, 5]	cut b of the system such that no other $b' \subset b$ is a cut; $\Psi(E-c) = 0, c \subseteq E$ iff $c \supseteq b$ for some minimal cut b

[3]The result can be proved as follows. Assume there exists a minimal bag d such that $b \supseteq d$. Because c is increasing, $c(b)$ is satisfied. To prove the other implication, consider first the case where $c(\emptyset)$ is satisfied. Because c is increasing, it follows that $c(b)$ is satisfied for any bag b, and the result is trivial ($\emptyset$ is the only minimal bag). Assume $c(\emptyset)$ not Satisfied & $c(b)$ satisfied with $b \supset \emptyset$. To find a minimal bag d such that $d \subseteq b$ we proceed by steps, considering at the kth step a bag $b_k = d_{k,1}[n_{k,1}] \cdots d_{k,l_k}[n_{k,l_k}]$ & the l_k subbags of b_k, $b_{k,1} = b_k - d_{k,1}[1], \ldots, b_{k,l_k} = b_k - d_{k,l_k}[1]$. In the first step, $b_1 = b$. If none of $c(b_{1,1}), \ldots, c(b_{1,l_1})$ is satisfied, then, c being increasing, b_1 is a minimal bag, and we are done. Otherwise, to be considered in the next step, we take as b_2 any subbag $b_{1,j}$, $1 \leq j \leq l_1$ such that $c(b_{1,j})$ is satisfied, and proceed. Because $c(\emptyset)$ is not satisfied, at some step k we will have that $c(b_k)$ is satisfied & that none of $c(b_{k,1}), \ldots, c(b_{k,l_k})$ is satisfied, which, being c increasing, implies that b_k is a minimal bag.

[4]The definitions generalize the corresponding usual definitions for systems made up of distinguishable components [3], except that we allow cuts & minimal cuts for systems where not all component classes are relevant.

[5]The result concerning the characterization of an increasing generalized structure function $\Psi(b)$, $b \subseteq E$ in terms of the minimal cuts of the system follows by noting that condition $\Psi(E - c) = 0$ is increasing with c.

$X = \{X(t); t \geq 0\}$	irreducible finite CTMC modeling the given fault-tolerant system ($X(0) = r$)
Ω	finite state space of X
$\lambda_{x,y}, x, y \in \Omega, y \neq x$	transition rate of X from state x to state y
$\lambda_x = \sum_{y \in \Omega - \{x\}} \lambda_{x,y}$, $x \in \Omega$	output rate of X from state x
$h_x = 1/\lambda_x, x \in \Omega$	mean holding time of X in state x
C	bag of component classes with domain $\mathcal{C}$ making up the given fault-tolerant system
$\Phi(b), b \subseteq C$	increasing generalized structure function of the given fault-tolerant system; Φ satisfies $\Phi(C) = 1$ & $\Phi(\emptyset) = 0$
$F(s)$	bag of failed component classes in state $s \in \Omega$
U	subset of up states of X ($U \neq \emptyset$)
D	subset of down states of X ($D \neq \emptyset$)
r	regenerative state ($F(r) = \emptyset$ & $r \in U$)
Ω'	$\Omega - \{r\}$
U'	$U - \{r\}$
$\Pi = \{\Pi_n; n = 0, 1, 2, \ldots\}$	embedded DTMC of X ($\Pi_0 = r$)
$P_{x,y}$	transition probability of Π from state x to state y; $P_{x,y} = \lambda_{x,y}/\lambda_x$ for x, $y \in \Omega$, $y \neq x$ & $P_{x,x} = 0$ for $x \in \Omega$
T	set of transitions of X (Π); $T = \{(x,y) \in \Omega \times \Omega, y \neq x : \lambda_{x,y} > 0\} = \{(x,y) \in \Omega \times \Omega, y \neq x : P_{x,y} > 0\}$
failure transition	a transition (x,y) of X (Π) is a failure transition if $F(y) \supset F(x)$
repair transition	a transition (x,y) of X (Π) is a repair transition if $F(y) \subseteq F(x)$
T_F	set of failure transitions of X (Π)
T_R	set of repair transitions of X (Π)
$T_F(x)=\{(x,y) \in T_F\}$	set of failure transitions going out of state x
$T_R(x)=\{(x,y) \in T_R\}$	set of repair transitions going out of state x
failure bag	any bag b with domain $\mathcal{C}$, $b \subseteq C$, $b \neq \emptyset$ for which there exists a failure transition (x,y) in X (Π) with $F(y) - F(x) = b$
F_B	set of failure bags
$active(x)$	set of failure bags which are active in state x, i.e. failure bags f for which there exists a transition $(x,y) \in T_F$ such that $F(y) - F(x) = f$
failure path from state $x \in U$	any path in X (Π) including only failure transitions from x to D; formally $(s_0, s_1, \ldots, s_l)$, $s_0 = x$, $s_i \in U$, $0 \leq i < l$, $s_l \in D$, $(s_i, s_{i+1}) \in T_F$, $0 \leq i < l$; l is the length of the path

$r_{\min}$	$\min_{(x,y)\in T_R} \lambda_{x,y}$
$f_{\min}$	$\min_{(x,y)\in T_F} \lambda_{x,y}$
$f_{\max}$	$\max_{(x,y)\in T_F} \lambda_{x,y}$
ε	$f_{\max}/r_{\min}$
$r_{x,y}$	repair transitions $(x,y) \in T_R$ are modeled as $\lambda_{x,y} = r_{\min} r_{x,y}$, $r_{x,y} \geq 1$
$f_{x,y}$, $d_{x,y}$	failure transitions $(x,y) \in T_F$ are modeled as $\lambda_{x,y} = r_{\min} f_{x,y} \varepsilon^{d_{x,y}}$, $f_{x,y} \in (0,1]$, $f_{x,y} \gg \varepsilon$, $d_{x,y} \geq 1$
balanced fault-tolerant system	a fault-tolerant system is balanced if its failure transition rates can be modeled with $d_{x,y} = 1$, $(x,y) \in T_F$
$P'_{x,y}$	biased transition probability of Π from state x to state y
UA	steady-state unavailability of the given fault-tolerant system
T_i	time at which X makes its ith entry into r ($T_0 = 0$)
$\widetilde{W}_i = T_i - T_{i-1}$	length of the ith regenerative cycle of X
$1_A(x)$	1 if $x \in A$; 0 otherwise
$\widetilde{Z}_i = \int_{T_{i-1}}^{T_i} 1_D(X(t))dt$	down time during the ith regenerative cycle of X
$\mathcal{S}$	set of regenerative cycles of Π
τ	step at which Π hits r for first time ($\tau = \min\{n > 0 : \Pi_n = r\}$)
W	r.v. (random variable) $\sum_{n=0}^{\tau-1} h_{\Pi_n}$
Z	r.v. $\sum_{n=0}^{\tau-1} 1_D(\Pi_n) h_{\Pi_n}$
P	probability measure on the set of regenerative cycles of Π
P'	modified probability measure on the set of regenerative cycles of Π used by the importance sampling schemes
P°	P' as a function of the parameters of the importance sampling schemes
L	r.v. likelihood ratio ($L(\omega) = P\{\omega\}/P'\{\omega\}$)
L°	r.v. likelihood ratio as a function of the parameters of the importance sampling scheme ($L^\circ(\omega) = P\{\omega\}/P^\circ\{\omega\}$)
Z'	r.v. ZL
Z°	r.v. ZL°
$E_P\{V\}$	expectation under the probability measure P of the r.v. V
$\mathrm{Var}_P\{V\}$	variance under the probability measure P of the r.v. V
$\widehat{x}$	estimator of x
$\overline{V}$	sample mean obtained from a set of independent realizations of the r.v. V
$S^2(V)$	sample variance obtained from a set of independent realizations of the r.v. V
$o(\varepsilon^d)$	a function $f(\varepsilon)$ is said to be $o(\varepsilon^d)$ (written $f(\varepsilon) = o(\varepsilon^d)$), where $d > 0$, if $\lim_{\varepsilon\to 0} f(\varepsilon)/\varepsilon^d = 0$
$\Theta(\varepsilon^d)$	a function $f(\varepsilon)$ is said to be $\Theta(\varepsilon^d)$ (written $f(\varepsilon) = \Theta(\varepsilon^d)$), where $d > 0$, if $f(\varepsilon) = c\varepsilon^d + o(\varepsilon^d)$, for some constant $c > 0$
$\mathcal{S}_k$	subset of $\mathcal{S}$ including the regenerative cycles which hit D & include k failure transitions from states $x \in \Omega'$
$F_{\max}$	$\max_{f\in F_B} \lvert f\rvert$, i.e. maximum number of components which can fail simultaneously
$k_{\min}$	$\min\{k : \mathcal{S}_k \neq \emptyset\}$
$\widetilde{\mathbf{F}}$	matrix $(1_{T_F}((x,y)) f_{x,y})_{x,y\in\Omega'}$
$\lVert\mathbf{A}\rVert_\infty$	given a matrix $\mathbf{A} = (a_{i,j})_{1\leq i,j\leq n}$, maximum row sum matrix norm, i.e. $\lVert\mathbf{A}\rVert_\infty = \max_{1\leq i\leq n} \sum_{j=1}^{n} \lvert a_{i,j}\rvert$
$td(x)$	failure transition distance from state x (0, for $x \in D$ & minimum number of failure transitions which build up a failure path from state x for $x \in U$)
F_B^c	with $c \in \mathcal{C}$, $\{f \in F_B : \#(c,f) > 0\}$
$A(c,n)$	with $c \in \mathcal{C}$, and $n > 0$, set of minimal bags $f_1[n_1]\cdots f_k[n_k]$ with domain F_B^c satisfying $\sum_{i=1}^{n_1} f_1 + \cdots + \sum_{i=1}^{n_k} f_k \supseteq c[n]$
$n(f)$	with $f \in F_B$, maximum i such that $f[i]$ is an input atom of the generalized fault tree of the fictitious fault-tolerant system if the generalized fault tree of the fictitious fault-tolerant system has some input atom $f[i]$; 0 otherwise
C'	bag of component classes with domain F_B making up the fictitious fault-tolerant system
$\Phi'(b)$, $b \subseteq C'$	increasing generalized structure function of the fictitious fault-tolerant system; Φ' satisfies $\Phi'(C') = 1$ & $\Phi'(\emptyset) = 0$
$B(x)$	being $x \in \Omega$, bag with domain F_B defined by $\#(c[1], B(x)) = \#(c, F(x))$, $c \in \mathcal{C}$ & $\#(b, B(x)) = 0$, $b \in F_B - \{c[1], c \in \mathcal{C}\}$
$ex(f)$	being $f \in F_B$, bag with domain F_B defined by $\#(c[1], ex(f)) = \#(c, f)$, $c \in \mathcal{C}$ & $\#(b, ex(f)) = 0$, $b \in F_B - \{c[1], c \in \mathcal{C}\}$
failure cover	bag b with domain F_B such that $\Phi(C - \sum_{f\in F_B} \sum_{i=1}^{\#(f,b)} f) = 0$
minimal failure cover	failure cover b such that there does not exist any other failure cover $b' \subset b$
MFC	set of minimal failure covers; $MFC \neq \emptyset$
FB	parameter of the FB, BFB, FTDB, and BFTDB importance sampling schemes

DB	parameter of the FTDB & BFTDB importance sampling schemes
dominant	a failure transition (x, y) of X (Π) is dominant if $td(y) = td(x) - 1$; otherwise it is non-dominant
$T_D(x)$	set of dominant failure transitions from state x
$T_{ND}(x)$	set of non-dominant failure transitions from state x
$\Omega_D = \{x \in \Omega : T_D(x) \neq \emptyset \wedge T_{ND}(x) \neq \emptyset\}$	set of states of X (Π) with both outgoing dominant failure transitions & outgoing non-dominant failure transitions
$atd(x, f)$	failure transition distance from any state y reached from x through a failure transition having associated with it failure bag f
$AMFC_f = \{b' : b' = b - ex(f), b \in MFC, b \cap ex(f) \neq \emptyset\}$	set of after minimal failure covers associated with failure bag f
RL	$\min_{b \in MFC} \lvert b \rvert$
RL_f	assuming $AMFC_f \neq \emptyset$, $\min_{b \in AMFC_f} \lvert b \rvert$
$MFC(c) = \{b \in MFC : \lvert b \rvert = c\}$	set of minimal failure covers of cardinality c
$MFCS_q(c)$	set of selectors of cardinality q of $MFC(c)$; a selector of $MFC(c)$ is any non-empty bag b with domain F_B such that $b \subseteq b'$ for some $b' \in MFC(c)$
$MFCL_c(p)$	list of minimal failure covers of cardinality c including selector p
$AMFC_f(c) = \{b \in AMFC_f : \lvert b \rvert = c\}$	set of after minimal failure covers of cardinality c associated with failure bag f
$AMFCS_q(c)$	set of selectors of cardinality q of some $AMFC_f(c)$, $f \in F_B$; a selector of $AMFC_f(c)$ is any non-empty bag b with domain F_B such that $b \subseteq b'$ for some $b' \in AMFC_f(c)$
$AMFCL_{f,c}(p)$	list of after minimal failure covers of cardinality c associated with failure bag f including selector p

I. Introduction

INCREASING demand for system's reliability, understood in its broad sense, i.e. as the capability of the system to perform properly, has created increased interest in fault-tolerant systems. A fault-tolerant system is a system which, through the use of redundancy, has the capability to sustain correct operation in the presence of faults. Many fault-tolerant systems can be regarded as either performing properly (up), or not (down), and the reliability of those systems can be properly quantified by using simple dependability measures such as the steady-state unavailability, the mean time to failure, the instantaneous unavailability, the expected interval unavailability, the distribution of the interval unavailability, and the reliability.[6]

Homogeneous continuous-time Markov chains (CTMC) are a commonly used mathematical framework to model the dependability of fault-tolerant systems. Computation of dependability measures requires, then, the analysis of the CTMC modeling the behavior of the fault-tolerant system. CTMC 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 & 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 it limits the application in practice of numerical analysis techniques [24], [33]. Simulation is an approach which alleviates the problem. 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 are 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 both CTMC models [1], [2], [4], [6], [9], [10], [13]–[15], [28], [36] & and non-Markovian stochastic models [17], [19]–[21]. Failure biasing (FB) is an importance sampling scheme which was first proposed in [15], [36] for the simulation of the expected interval unavailability, and, in combination with transition forcing, for the simulation of the unreliability. The scheme has been adapted in [6], [9] for the simulation of the steady-state unavailability, in [28] for the simulation of the mean time to failure, and in [10] for the simulation of other dependability measures. Balanced failure biasing (BFB) [10], [29] & failure distance biasing [4] are other closely related importance sampling schemes. Theoretical properties dealing with the robustness of FB, BFB, failure distance biasing, and a balanced version of the latter for some measures & classes of failure/repair CTMC models have been investigated in [16], [18], [29], [30]. More recently, robust 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 & high redundancy degrees. Some of those techniques use the "shortest path to failure" concept, and seem to outperform BFB also for fault-tolerant systems with failure rates much smaller than repair rates. Importance sampling schemes which are

[6]We make a distinction between the reliability as a property of a system (capability of the system to perform properly) & the reliability as a dependability measure. The latter is defined as the probability that the system has not failed by time t.

robust & efficient when the model has high-probability cycles have also been developed [13], [14].

In this paper, we develop two importance sampling schemes, called *failure transition distance biasing* (FTDB) & *balanced failure transition distance biasing* (BFTDB), which exploit the failure transition distance concept in an attempt to achieve more efficient simulations than when using FB & BFB. The implementation of FTDB & BFTDB requires the computation of the so-called failure transition distances. Procedures to perform those computations are developed. We also develop a method for the optimization of the importance sampling schemes. The presentation of the importance sampling schemes is tied to the measure-specific simulation method for the steady-state unavailability described in [9] & to failure/repair CTMC models with repair in every state with failed components. FTDB has (as FB) the bounded relative error (BRE) property for balanced fault-tolerant systems, i.e. fault-tolerant systems with failure transition rates differing much less than how failure transition rates differ from repair transition rates, whereas BFTDB has (as BFB) the BRE property for both balanced & unbalanced fault-tolerant systems. Furthermore, theoretical results are available, motivating FTDB & BFTDB for balanced fault-tolerant systems & suggesting that, for those systems, FTDB & BFTDB can be significantly more efficient than both FB & BFB. Although no theoretical results are available suggesting that BFTDB may indeed improve the efficiency of BFB for unbalanced fault-tolerant systems, experimental results seem to indicate that this is the case.

The rest of the paper is organized as follows. Section II describes the considered class of CTMC models, and reviews both importance sampling theory & the measure-specific simulation method developed in [9] for the estimation of the steady-state unavailability. Section III reviews FB & BFB, motivates theoretically for balanced fault-tolerant systems FTDB & BFTDB, and describes FTDB & BFTDB with their implementations. Section IV deals with the optimization of the simulation method for the steady-state unavailability described in [9]. There, we describe two versions of the simulation method: one (NOPT) in which the parameters of the importance sampling schemes are not optimized, and another (OPT) in which they are. Section V gives implementation details. Section VI presents numerical results comparing the performances of FB, BFB, FTDB, and BFTDB; analyses the overhead resulting from the computation of failure transition distances in FTDB & BFTDB; and relates FTDB & BFTDB to the importance sampling schemes described in [1], [2]. We also show in that section how FTDB & BFTDB can be adapted when only minimal failure covers of up to a given cardinality are known, and investigate to what extent this degrades the performance of the importance sampling schemes. The motivation for introducing the adapted FTDB & BFTDB is to extend the applicability of those importance sampling schemes, which, in their pure form, are limited to fault-tolerant systems not having too many minimal failure covers. The Appendix includes the proofs.

II. Preliminaries

A. Class of 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 & 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 iff the bag b of failed component classes of the system is such that $b \supseteq c[n]$. $\Phi(b) = 0$ iff 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 & 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 $F(y) \supset F(x)$ & repair transitions (x, y), characterized by $F(y) \subseteq F(x)$. 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 & by $D = \Omega - U$ the subset of down states of X. We will also 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 $active(x)$. The simulation method for the steady-state unavailability we will consider will use r as a regenerative state. Let $\Omega' = \Omega - \{r\}$. We will make the following four assumptions:

- A1) From each state $x \in \Omega'$, $T_R(x) \neq \emptyset$.
- A2) $c[1] \in F_B$ for each $c \in \mathcal{C}$.
- A3) For each $f \in F_B$, $f' \in F_B$ for each $f' \subset f$, $f' \neq \emptyset$.
- A4) For every $x \in U$, $active(x) = \{f \in F_B : f \subseteq C - F(x)\}$.

Informally, A1 states that there is repair from every state with failed components (every state except state r), A4 states that

from every up state there are failure transitions associated with all possible failure bags (those for which there are operational components building up the failure bag), and A2 & A3 state reasonable conditions that the set of failure bags must satisfy. Note that $F(r) = \emptyset$ implies $\Phi(C - F(r)) = \Phi(C) = 1$ & $r \in U \neq \emptyset$. Also, because $\Phi(\emptyset) = 0$, assumptions A2 & A4 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 which achieve high dependability with moderate redundancy levels. To motivate theoretically FTDB & BFTDB, and to give results regarding the robustness of those 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 $f_{\min}/f_{\max} \gg \varepsilon = f_{\max}/r_{\min}$.

B. Review of Importance Sampling Theory

Let $(\mathcal{S}, \mathcal{A}, P)$ be a probability space with discrete sample space S with $P\{\omega\} > 0, \omega \in \mathcal{S}$, and let Y be a nonnegative r.v. defined on $\mathcal{S}$. The expectation of Y under the probability measure P is $E_P\{Y\} = \sum_{\omega\in\mathcal{S}} Y(\omega) P\{\omega\}$. Assume $E_P\{Y\} > 0$ & that we want to estimate $E_P\{Y\}$. The standard simulation method for doing that consists in generating n independent samples of Y, Y_i, $i = 1, \ldots, n$, and using the sample mean

$$\overline{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i,$$

which is an unbiased estimator of $E_P\{Y\}$. Assuming $E_P\{Y^2\} < \infty$ & n sufficiently large, a confidence interval for $E_P\{Y\}$ can be derived using the central limit theorem. The variance of $\overline{Y}$, $\mathrm{Var}\{\overline{Y}\}$, is related to the variance of Y under P, $\mathrm{Var}_P\{Y\}$, by $\mathrm{Var}\{\overline{Y}\} = \mathrm{Var}_P\{Y\}/n$. The variance $\mathrm{Var}_P\{Y\}$ is estimated by the sample variance

$$S^2(Y) = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \overline{Y})^2,$$

and a $100(1-\alpha)$ confidence interval for $E_P\{Y\}$ is

$$\overline{Y} \pm z_\alpha \frac{\sqrt{S^2(Y)}}{\sqrt{n}},$$

where z_α is the $1 - \alpha/2$ quantile of the standard s-normal distribution. Thus, the confidence interval halfwidth is roughly proportional to $\sqrt{\mathrm{Var}_P\{Y\}/n}$. If $\sqrt{\mathrm{Var}_P\{Y\}}$ is large relative to $E_P\{Y\}$, a very large number of samples n will be necessary to obtain a confidence interval of reasonable quality.

Let P' be another probability measure on $\mathcal{S}$ with $P'\{\omega\} > 0$, $\omega \in \mathcal{S}$. We have

$$\begin{aligned} E_P\{Y\} &= \sum_{\omega\in\mathcal{S}} Y(\omega) \frac{P\{\omega\}}{P'\{\omega\}} P'\{\omega\} \\ &= \sum_{\omega\in\mathcal{S}} Y(\omega) L(\omega) P'\{\omega\} \\ &= E_{P'}\{YL\}, \end{aligned} \tag{1}$$

where $L(\omega) = P\{\omega\}/P'\{\omega\}$, $\omega \in \mathcal{S}$ is the likelihood ratio. Based on (1), importance sampling is a technique in which an estimate for $E_P\{Y\}$ is obtained by considering the probability space $(\mathcal{S}, \mathcal{A}, P')$ & sampling with probability measure P' the r.v. $Y' = YL$. Being Y'_i, $i = 1, 2, \ldots, n$, the samples of Y', this yields the estimator

$$\overline{Y}' = \frac{1}{n} \sum_{i=1}^{n} Y'_i$$

with variance $\mathrm{Var}\{\overline{Y}'\} = \mathrm{Var}_{P'}\{Y'\}/n$. If $\mathrm{Var}_{P'}\{Y'\}$ is significantly smaller than $\mathrm{Var}_P\{Y\}$ & the computational effort of sampling Y' under P' is similar to the computational effort of sampling Y under P, this will result in a significantly more efficient simulation method for estimating $E_P\{Y\}$ than the standard one.

Further insight on how to apply the importance sampling technique can be gained by noting that, given a target confidence interval halfwidth, the required number of samples using importance sampling is roughly proportional to $\mathrm{Var}_{P'}\{Y'\}$, and that

$$\begin{aligned} &\mathrm{Var}_{P'}\{Y'\} \\ &\quad = E_{P'}\{Y'^2\} - E_{P'}\{Y'\}^2 \\ &\quad = E_{P'}\{Y'^2\} - E_P\{Y\}^2 \\ &\quad = \sum_{\omega\in\mathcal{S}} Y'(\omega)^2 P'\{\omega\} - E_P\{Y\}^2 \\ &\quad = \sum_{\omega\in\mathcal{S}} Y(\omega)^2 L(\omega)^2 P'\{\omega\} - E_P\{Y\}^2 \\ &\quad = \left(\sum_{\omega\in\mathcal{S}} \left(\frac{Y(\omega)P\{\omega\}}{E_P\{Y\}} \frac{1}{P'\{\omega\}} \right)^2 P'\{\omega\} - 1 \right) E_P\{Y\}^2, \end{aligned}$$

where $Y(\omega)P\{\omega\}/E_P\{Y\}$ can be interpreted as the relative contribution of ω to $E_P\{Y\}$. For $P'\{\omega\} =$

$Y(\omega)P\{\omega\}/E_P\{Y\}$, $\mathrm{Var}_{P'}\{Y'\} = 0$. Accordingly, importance sampling theory suggests exploiting any heuristic insight we may have about the distribution of Y to choose the probability measure P' so that $P'\{\omega\}, \omega \in \mathcal{S}$ be "close" to their relative *importances* (relative contributions of ω to $E_P\{Y\}$). In first approximation, more "important" sample points ω should be sampled more often than less "important" sample points.

C. Measure-Specific 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 & finite, UA is independent of the initial probability distribution of X, and we can assume without loss of generality that $X(0) = r$. Let $T_i, i = 0, 1, 2, \ldots$ be the time at which X makes its ith entry into r ($T_0 = 0$). Then, X can be regarded as a regenerative process with regeneration epochs T_i. Let $\widetilde{W}_i = T_i - T_{i-1}$ denote the length of the ith regenerative cycle of X, and let $\widetilde{Z}_i = \int_{T_{i-1}}^{T_i} 1_D(X(t))dt$ denote the down time during the ith regeneration cycle of X. $\widetilde{W}_i, i = 1, 2, \ldots$ are independent identically distributed r.v.; similarly, $\widetilde{Z}_i, i = 1, 2, \ldots$ are independent identically distributed r.v. Let $\widetilde{W}$ denote any of the r.v. $\widetilde{W}_i$, and let $\widetilde{Z}$ denote any of the r.v. $\widetilde{Z}_i$. Then, noting that $P\{X(t) \in D\} = E\{1_D(X(t))\}$, we have [27]

$$UA = \frac{E\{\widetilde{Z}\}}{E\{\widetilde{W}\}}. \tag{2}$$

Based on (2), a standard method to estimate UA is the regenerative simulation method [35]. In the method, sample pairs are obtained of the random variables $\widetilde{W}$, $\widetilde{Z}$, and the estimator

$$\widehat{UA} = \frac{\overline{\widetilde{Z}}}{\overline{\widetilde{W}}}$$

is used, where $\overline{\widetilde{Z}}$ & $\overline{\widetilde{W}}$ are, respectively, the sample means of $\widetilde{Z}$ & $\widetilde{W}$. An alternative formulation to (2) can be obtained in terms of random variables W, Z defined on the set of regenerative cycles 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 & initial probability distribution as X & transition probabilities $P\{\Pi_{n+1} = y | \Pi_n = x\} = P_{x,y} = \lambda_{x,y}/\lambda_x$, $x, y \in \Omega$, $y \neq x$, and $P\{\Pi_{n+1} = x | \Pi_n = x\} = P_{x,x} = 0$, $x \in \Omega$, 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 & 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 $h_x = 1/\lambda_x$ is the mean holding time of X in state x, and we have

$$UA = \frac{E_P\{Z\}}{E_P\{W\}}, \tag{3}$$

where the subscript P in $E_P\{Z\}$ & $E_P\{W\}$ makes explicit the probability measure with respect to which the expectation is defined. Formally, letting $T = \{(x, y) \in \Omega \times \Omega, y \neq x : P_{x,y} > 0\}$[7] 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

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

The regenerative simulation method based on (3) is guaranteed [12] to be more efficient than the regenerative simulation method based on (2).

Estimation of UA by the regenerative simulation method tends to be inefficient. Intuitively, this is because, system failure being 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 the r.v. $W' = WL$ & $Z' = ZL$, where L 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 & 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], [30]. We review next that method.

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 & $m = \widetilde{m}k$ 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

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

[7]The fact that X & Π have same state space & 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 & Π, and we will do so throughout the paper.

where $\overline{Z'}$ & $\overline{W}$ are, respectively, the sample means of Z' & 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

$$\widehat{UA} \pm z_\alpha \frac{\overline{Z'}}{\overline{W}} \left(\left(\frac{1}{\sqrt{m}} \frac{\sqrt{S^2(Z')}}{\overline{Z'}} \right)^2 + \left(\frac{1}{\sqrt{n}} \frac{\sqrt{S^2(W)}}{\overline{W}} \right)^2 \right)^{1/2}, \tag{4}$$

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

$$S^2(Z') = \frac{1}{m-1}\sum_{i=1}^{m} \left(Z'_i - \overline{Z'}\right)^2 = \frac{1}{m-1}\sum_{i=1}^{m} (Z_i L_i - \overline{Z'})^2,$$
$$S^2(W) = \frac{1}{n-1}\sum_{i=1}^{n} (W_i - \overline{W})^2, \tag{5}$$

& z_α is the $1-\alpha/2$ quantile of the standard s-normal distribution. That confidence interval is obtained by applying the central limit theorem with i.i.d. r.v. (see [2])

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

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

III. Importance Sampling Schemes

In this section, we review the importance sampling schemes FB & BFB & describe FTDB & BFTDB, the importance sampling schemes developed in the paper. FTDB & BFTDB will be motivated theoretically for balanced fault-tolerant systems. 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 $P'_{x,y} > 0$ iff $P_{x,y} > 0$. The implementation of both FTDB & BFTDB requires the computation of the failure transition distances from the last sampled state x & its successors (states $y \in \Omega$ such that $P_{x,y} > 0$) through failure transitions, and procedures for performing those computations will also be described. Those procedures are based on a formalization of the failure transition distances in terms of the set of minimal cuts of a fictitious fault-tolerant system whose generalized fault tree can be obtained easily from the generalized fault tree of the given fault-tolerant system. That formalization is obtained before actually deriving the procedures for the computation of failure transition distances.

Some previous literature [30] has made a distinction between simple failure biasing (simple balanced failure biasing) & dynamic failure biasing (dynamic balanced failure biasing), the difference being that, in the dynamic versions, biasing is turned-off as soon as D is hit. In this paper, by FB & BFB we mean the dynamic versions. In FTDB & BFTDB, biasing is also turned-off when D is hit. Then, for all FB, BFB, FTDB, and BFTDB, the probability measure P' under importance sampling is

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

where

$$l_D(s_0, s_1, \ldots, s_l) = \max\{k \le l-1 : s_0, s_1, \ldots, s_k \in U\}.$$

The importance sampling schemes only differ in the way $P'_{x,y}$ are computed from $P_{x,y}$.

A. Review of FB & BFB

In FB, when a state has both outgoing failure transitions & outgoing repair transitions, the probabilities associated with failure transitions & 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 $1 - FB$. This is done so that regenerative cycles hitting D are sampled with significant probabilities. 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, letting $U' = U - \{r\}$, for the considered class of CTMC models, the biased transition probabilities in FB are

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

& the biased transition probabilities in BFB are

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

B. Failure Transition Distance-Based Schemes

We start by motivating, for balanced fault-tolerant systems, FTDB & BFTDB. 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,

$$\begin{aligned} P_{r,y} &= \frac{r_{\min} f_{r,y} \varepsilon}{\sum_{(r,z)\in T_F(r)} r_{\min} f_{r,z} \varepsilon} = \Theta(1), \\ P_{x,y} &= \frac{r_{\min} f_{x,y} \varepsilon}{\sum_{(x,z)\in T_R(x)} r_{\min} r_{x,y} + \sum_{(x,z)\in T_F(x)} r_{\min} f_{x,z} \varepsilon} \\ &= \Theta(\varepsilon), \quad x \in \Omega', (x,y) \in T_F(x), \end{aligned}$$

and

$$\begin{aligned} P_{x,y} &= \frac{r_{\min} r_{x,y}}{\sum_{(x,z)\in T_R(x)} r_{\min} r_{x,y} + \sum_{(x,z)\in T_F(x)} r_{\min} f_{x,z} \varepsilon} \\ &= \Theta(1), \quad x \in \Omega', (x,y) \in T_R(x). \end{aligned}$$

Also,

$$\begin{aligned} P_{x,y} &= \frac{r_{\min} f_{x,y} \varepsilon}{\sum_{(x,z)\in T_R(x)} r_{\min} r_{x,y} + \sum_{(x,z)\in T_F(x)} r_{\min} f_{x,z} \varepsilon}, \\ &\leq f_{x,y} \varepsilon, \quad x \in \Omega', (x,y) \in T_F(x), \\ h_x &= \frac{1}{\sum_{(x,z)\in T_R(x)} r_{\min} r_{x,y} + \sum_{(x,z)\in T_F(x)} r_{\min} f_{x,z} \varepsilon} \\ &= \Theta(1), \quad x \in \Omega', \end{aligned}$$

and

$$h_x \leq \frac{1}{r_{\min}}, \quad x \in \Omega'.$$

We will find it 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 $x \in \Omega'$. 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 following theorem gives results regarding the relative values, for balanced fault-tolerant systems, of the contributions to $E_P\{Z\}$ of the subsets $\mathcal{S}_k$ & the individual regenerative cycles in $\mathcal{S}_{k_{\min}}$.

Theorem 1: 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: See the Appendix. Preceded by Lemma 1, Theorem 5, Corollary 1, Proposition 1, Proposition 2, and Theorem 6, all in the Appendix. □

According to Theorem 1, for $\varepsilon \to 0$, importance sampling theory suggests to sample regenerative cycles so that $P'\{\mathcal{S}_{k_{\min}}\}$ be close to 1, and each $\omega \in \mathcal{S}_{k_{\min}}$ is assigned a $\Theta(1)$ probability.

The importance sampling schemes FTDB & BFTDB are designed so that the simulation effort can be concentrated into the regenerative cycles in $\mathcal{S}_{k_{\min}}$. This is achieved by: 1) assigning higher probabilities to the failure transitions which take Π "closer" to D until D is hit, so that D is hit with high probability after a minimum number of failure transitions; and 2) turning biasing off when D is hit, so that with high probability all transitions sampled from that point on are repair transitions. The closeness to D is established in terms of failure transition distances. The failure transition distance, $td(x)$, 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 $td(y) = td(x) - 1$ & *non-dominant* otherwise ($td(y) = td(x)$). Both FTDB & BFTDB use two biasing parameters. The first one, FB, $0 < FB < 1$, plays a similar role as FB in FB & 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 & outgoing non-dominant failure transitions ($\Omega_D = \{x \in \Omega : T_D(x) \neq \emptyset \wedge T_{ND}(x) \neq \emptyset\}$), for the considered class of CTMC models, the biased transition probabilities in FTDB are shown in the first equation at bottom of the next page. BFTDB differs from FTDB in that the probability assigned to each subset of failure transitions is evenly distributed among the transitions in the subset, as shown in the second equation at bottom of the next page.

The subset $\mathcal{S}_{k_{\min}}$ includes just the regenerative cycles Which, before hitting D, include $k_{\min}+1$ dominant failure transitions (one from state r & $k_{\min}$ from states $x \in \Omega'$), and, after hit of D, include only repair transitions. It is easy to check that, for $\varepsilon \to 0$, under both FTDB & BFTDB $P'\{\mathcal{S}_{k_{\min}}\} \to FB^{k_{\min}}DB^{l}$, $0 \le l \le k_{\min}+1$, which is close to 1 for FB & DB close to 1. Furthermore, for balanced fault-tolerant systems, all regenerative cycles in $\mathcal{S}_{k_{\min}}$ would be sampled with $\Theta(1)$ probabilities under both FTDB & BFTDB. On the other hand, $P'\{\mathcal{S}_{k_{\min}}\}$ could be small under both FB & BFB. As an example, consider a fault-tolerant system made up of a bag of 200 components of 100 different classes, with two components per class, which is up if at least one component of each class is operational. Components of class c_i, $1 \le i \le 100$, fail with rate $\lambda_i = (1+(101-i)/100)\times 10^{-6}\mathrm{h}^{-1}$, and there is a single repairman to repair failed components at rate $1\mathrm{h}^{-1}$, with random selection among the set of failed components. For that system, $k_{\min}=1$. A simple calculation shows that, under FB, $P'\{\mathcal{S}_1\} \approx FB(\sum_{i=1}^{100}\lambda_i^2)/(2(\sum_{i=1}^{100}\lambda_i)^2) = FB/193$, and, under BFB, $P'\{\mathcal{S}_1\} \approx FB/100$, which for the usual selection $FB=0.5$ give $P'\{\mathcal{S}_{1,0}\}=0.00259$ under FB, and $P'\{\mathcal{S}_{1,0}\}=0.005$ under BFB. Thus, we can conclude that, for balanced fault-tolerant systems, simulation under FTDB & BFTDB can be much more efficient that under FB & BFB, provided that the computational cost of generating a regenerative cycle under FTDB & BFTDB is comparable to the computational cost of generating a regenerative cycle under FB & BFB. The latter requires that failure transition distances be computed efficiently, an issue that we will address in Section III-D.

The relative error of the estimator $\widehat{UA}$ is defined as the expected relative halfwidth of the confidence interval. From (4), the relative error is (for large m & n)

$$\left(\left(\frac{z_\alpha}{\sqrt{m}}\frac{\sqrt{\mathrm{Var}_{P'}\{Z'\}}}{E_{P'}\{Z'\}}\right)^2+\left(\frac{z_\alpha}{\sqrt{n}}\frac{\sqrt{\mathrm{Var}_{P}\{W\}}}{E_{P}\{W\}}\right)^2\right)^{1/2} = \left(\left(\frac{z_\alpha}{\sqrt{m}}\frac{\sqrt{\mathrm{Var}_{P'}\{Z'\}}}{E_{P}\{Z\}}\right)^2+\left(\frac{z_\alpha}{\sqrt{n}}\frac{\sqrt{\mathrm{Var}_{P}\{W\}}}{E_{P}\{W\}}\right)^2\right)^{1/2}.$$

The component attributable to the importance sampling scheme with which Z' is sampled is

$$\frac{z_\alpha}{\sqrt{m}}\frac{\sqrt{\mathrm{Var}_{P'}\{Z'\}}}{E_P\{Z\}},$$

and the importance sampling scheme is said to have the BRE property if $\sqrt{\mathrm{Var}_{P'}\{Z'\}}/E_P\{Z\}$ remains bounded as $\varepsilon \to 0$.

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

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

This ensures that, for small enough ε, $\mathrm{Var}_{P'}\{Z'\}$ (and, therefore, $E_{P'}\{Z'^2\}$) will be finite, and the simulation method will be robust. It also implies that, as $\varepsilon \to 0$, the performance of the importance sampling schemes will not deteriorate pathologically. Under the additional assumption that either $\{(r,y) \in T_F(r), y \in D\} = \emptyset$ or $d_{r,y} > 1$ for every $(r,y) \in T_F(r)$, $y \in D$, it has been proved in [30] that $E_P\{Z\} = \Theta(\varepsilon^r)$, $r > 0$ & that, under any importance sampling scheme for which $P'_{x,y}$ is independent of ε, $E_{P'}\{Z'^2\} = \Theta(\varepsilon^{2r})$, and both imply the BRE property. The additional assumption was taken in [30] because, when that assumption is not satisfied, the set of regenerative cycles hitting D is not rare. It is easy to check that, when the assumption is not satisfied, i.e. $\{(r,y) \in T_F(r), y \in D\} \neq \emptyset$ & $d_{r,y} = 1$ for every $(r,y) \in T_F(r)$, $y \in D$, $E_P\{Z\} = \Theta(1)$ & $E_{P'}\{Z'^2\} = \Theta(1)$ under any importance sampling scheme for which $P'_{x,y}$ is independent of ε, and both facts imply the BRE property. FTDB (as FB) yields $P'_{x,y}$ independent of ε for balanced fault-tolerant systems. BFTDB (as BFB) yields $P'_{x,y}$ independent of ε for both balanced & unbalanced fault-tolerant systems. Thus, FTDB has the BRE property for balanced fault-tolerant systems, and BFTDB has the BRE property for both balanced & unbalanced fault-tolerant systems.

FTDB & BFTDB can be regarded as elaborations of the failure distance biasing & balanced failure distance biasing importance sampling schemes described in, respectively, [4] & [18], in which the biasing parameters DB & CB of those schemes are replaced by a single biasing parameter DB, and focus of the simulation effort towards the regenerative cycles in $\mathcal{S}_{k_{\min}}$ is precisely controlled through the biasing parameters FB & DB.

C. Formalization of Failure Transition Distances

In this section, we will formalize failure transition distances in terms of the set of minimal cuts of a fictitious fault-tolerant system whose generalized fault tree can be easily obtained from the generalized fault tree of the given fault-tolerant system.

A *failure cover* is defined to be any bag b with domain F_B such that $\Phi(C - \sum_{f \in F_B} \sum_{i=1}^{\#(f,b)} f) = 0$. Informally, a failure cover is any bag of failure bags such that the failure of all the components in the failure cover implies the failure of the system. A failure cover b is a *minimal failure cover* if there does not exist any other failure cover $b' \subset b$. Let MFC denote the set of minimal failure covers.

The fictitious fault-tolerant system is made up of a certain bag of component classes C' with domain F_B, and has a generalized fault tree which can be obtained from the generalized fault tree of the given system by replacing each input atom $c[n]$, $c \in \mathcal{C}$ by the AND/OR logic

$$\bigvee_{f_1[n_1] \cdots f_k[n_k] \in A(c,n)} f_1[n_1] \wedge \cdots \wedge f_k[n_k],$$

where

$$A(c,n) = \{\text{minimal bags } f_1[n_1] \cdots f_k[n_k] \text{ with domain } F_B^c : \sum_{i=1}^{n_1} f_1 + \cdots + \sum_{i=1}^{n_k} f_k \supseteq c[n]\},$$
$$F_B^c = \{f \in F_B : \#(c,f) > 0\}.$$

For $f \in F_B$, shown in the equation at bottom of page, C' is defined as

$$C' = \sum_{f \in F_B : n(f) > 0} f[n(f)],$$

so that for any input atom $f[n]$, $f \in F_B$ of the generalized fault tree of the fictitious fault-tolerant system $0 < n \leq \#(f, C')$. Fig. 1 illustrates the relationships between the generalized fault tree of the given fault-tolerant system & F_B on one hand, and the generalized fault tree of the fictitious fault-tolerant system & C' on the other hand.

Let $\Phi'(b)$, $b \subseteq C'$ be the generalized structure function of the fictitious fault-tolerant system ($\Phi'(C' - b) = 0$ iff the output of the generalized fault tree of the fictitious fault-tolerant system is implied at 1 when any input with input atom $f[n]$ is implied at 1 iff $f[n] \subseteq b$). Because the generalized fault tree of the fictitious fault-tolerant system only includes AND & OR gates, as for the generalized structure function of the given fault-tolerant system, $\Phi'(b)$, $b \subseteq C'$ is increasing. Furthermore, $\Phi'(C') = 1$ & $\Phi'(\emptyset) = 0$. Because Φ' is increasing, $\Phi'(C' - b) = 0, b \subseteq C'$ iff $b \supseteq c$, where c is a minimal cut of the fictitious fault-tolerant system. The following key Theorem relates MFC to the set of minimal cuts of the fictitious fault-tolerant system.

Theorem 2: MFC is the set of minimal cuts of the fictitious fault-tolerant system & $MFC \neq \emptyset$.

Proof: See the Appendix. Preceded by Proposition 3. □

For $x \in \Omega$, let $B(x)$ be the bag with domain F_B defined by $\#(c[1], B(x)) = \#(c, F(x))$, $c \in \mathcal{C}$, and $\#(b, B(x)) = 0$, $b \in F_B - \{c[1], c \in \mathcal{C}\}$, i.e. $B(x)$ includes just as many instances

$$n(f) = \begin{cases} \max\{i : f[i] \text{ is an input atom of the generalized fault tree of the fictitious} \\ \qquad \text{fault} - \text{tolerant system}\} \\ \quad \text{if the generalized fault tree of the fictitious fault} - \text{tolerant system has} \\ \quad \text{some input atom} f[i], \\ 0 \qquad \text{otherwise.} \end{cases}$$

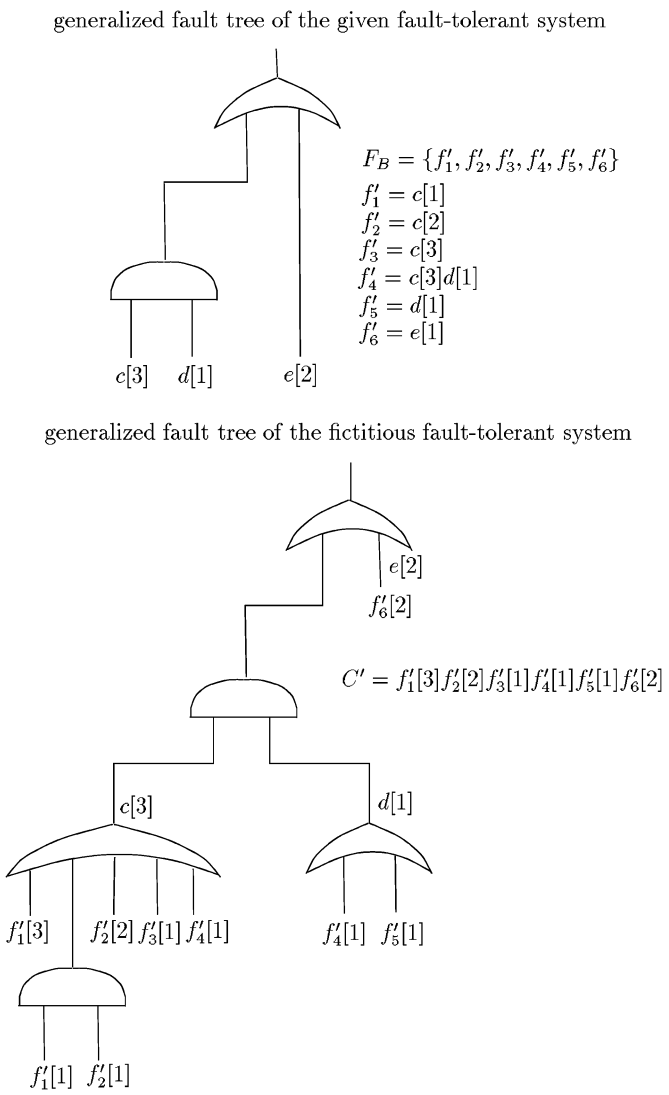

Fig. 1. Relationships between the generalized fault tree of the given fault-tolerant system & F_B, on one hand, and the generalized fault tree of the fictitious fault-tolerant system & C', on the other hand.

of each failure bag $c[1]$ as instances of component class c are failed in state x. The formalization of failure transition distances in terms of MFC is given by the following theorem.

Theorem 3: For $x \in U$,

$$td(x) = \min_{b \in MFC} |b - B(x)|.$$

Proof: See the Appendix. □

D. Procedures for the Computation of Failure Transition Distances

The implementation of both FTDB & BFTDB requires the knowledge of the failure transition distances from the states x in the regenerative cycle under construction & the successors of x reached through failure transitions, while the regenerative cycle has not hit D. To be specific, assuming that D has not been hit when state x is generated, and denoting by $atd(x, f)$ the failure transition distance from any state y reached from x through a failure transition (x, y) having associated with it failure bag f, what is required is $td(x)$ & $atd(x, f)$, $f \in active(x)$. Let $ex(f)$, $f \in F_B$ be the bag with domain F_B defined by $\#(c[1], ex(f)) = \#(c, f)$, $c \in \mathcal{C}$ & $\#(b, ex(f)) = 0$, $b \in F_B - \{c[1], c \in \mathcal{C}\}$, i.e. $ex(f)$ includes as many instances of each failure bag $c[1]$, $c \in \mathcal{C}$ as instances of c includes f. Assuming $B(x)$ & MFC known, $td(x)$ could be computed using Theorem 3; and, because, for each state y reached from x through a failure transition (x, y) having associated with it failure bag f, $B(y) = B(x) + ex(f)$, $atd(x, f)$, $f \in active(x)$ could be computed using

$$atd(x, f) = \min_{b \in MFC} |b - (B(x) + ex(f))|. \tag{6}$$

However, use of Theorem 3 & (6) would be computationally expensive when $|MFC|$ is large. In this section, we will describe typically much more efficient procedures, which can be embedded in the simulation, for computing $td(x)$ & $atd(x, f)$, $f \in active(x)$.

Those procedures are based on a reformulation of (6) & two observations. The reformulation of (6) is based on the concept of after minimal failure cover. The *after minimal failure cover* associated with a minimal failure cover b & a failure bag f such that $b \cap ex(f) \neq \emptyset$ is $b' = b - ex(f)$. Let $AMFC_f$ denote the set of after minimal failure covers associated with failure bag f, i.e. $AMFC_f = \{b' : b' = b - ex(f), b \in MFC, b \cap ex(f) \neq \emptyset\}$. Then, the reformulation of (6) is given by the following theorem.

Theorem 4:

$$atd(x, f) = \begin{cases} td(x) & \text{if } AMFC_f = \emptyset \\ \min\{td(x), \min_{b \in AMFC_f} |b - B(x)|\} & \text{if } AMFC_f \neq \emptyset. \end{cases}$$

Proof: See the Appendix. □

The first observation is that, b' being a minimal failure cover or an after minimal failure cover & b being a bag with domain F_B, for $|b' - b| < ub$, it is required $|b'| < ub + |b|$, because otherwise we would have $|b' - b| \geq |b'| - |b| \geq ub$. The second observation is that, b' being a minimal failure cover or an after minimal failure cover & b being a bag with domain F_B, for $|b' - b| < ub$, it is required $|b' \cap b| \geq |b'| - ub + 1$, because otherwise we would have $|b' - b| = |b'| - |b' \cap b| > ub - 1$.

The first procedure, $Compute\_td(b, td)$, takes as input a bag b with domain F_B, and has as output

$$td = \min_{b' \in MFC} |b' - b|. \tag{7}$$

The second one, $Compute\_atd(b, S_F, td, atd(f))$, takes as inputs a bag b with domain F_B, a set of failure bags $S_F \subseteq F_B$, and $td = \min_{b' \in MFC} |b' - b|$, and has as outputs shown in (8) at the bottom of the next page. It is assumed that the procedures have access to

$$RL = \min_{b \in MFC} |b|$$

and

$$RL_f = \min_{b \in AMFC_f} |b|,\ f \in \{f' \in F_B : AMFC_{f'} \neq \emptyset\},$$

which are computed before the simulation starts.

Using those procedures, for each sampled regenerative cycle, for all states y preceding the state through which the regenerative cycle hits D for first time, $td(y)$ & $atd(y,f)$, $f \in active(y)$ are computed as follows. For $y = r$, we have $td(y) = RL$, $atd(y,f) = RL$, $f \in \{f' \in active(r) : AMFC_{f'} = \emptyset\}$, and $atd(y,f) = \min\{RL, RL_f\}$, $f \in \{f' \in active(r) : AMFC_{f'} \neq \emptyset\}$. For $y \neq r$, let x be the state before y in the regenerative cycle, and assume that $td(x)$ & $atd(x,f)$, $f \in active(x)$ are available. If y was reached from x through a failure transition (x,y) having associated with it a failure bag f', $td(y)$ is obtained as $atd(x,f')$, and $atd(y,f)$, $f \in active(y)$ are obtained by making the call $Compute\_atd(B(y), active(y), td(y), atd(y,f))$. If y was reached from x through a repair transition (x,y) having associated with it an empty bag, then, $td(y) = td(x)$, $atd(y,f) = atd(x,f)$, $f \in active(y) = active(x)$. Otherwise, $td(y)$ is computed by making the call $Compute\_td(B(y), td(y))$, and after that, $atd(y,f)$, $f \in active(y)$ are computed by making the call $Compute\_atd(B(y), active(y), td(y), atd(y,f))$. We will describe & justify next each procedure. The procedures depend on a parameter $R \geq 1$.

We start by considering the procedure $Compute\_td$. Let $MFC(c)$ be the subset of MFC including the minimal failure covers of cardinality c. Then, a *selector* of $MFC(c)$ is any non-empty subbag of some $b' \in MFC(c)$. The procedure $Compute\_td$ assumes that, for each minimal failure cover cardinality c, the sets $MFCS_{q'}(c)$ of selectors of $MFC(c)$ of cardinalities $q' = 1, 2, \ldots, \min\{R, c\}$ have been obtained before the simulation starts. It also assumes that, for each minimal failure cover cardinality c & each selector $p \in MFCS_{q'}(c)$, $q' = 1, 2, \ldots, \min\{R, c\}$, a list $MFCL_c(p)$ linking the minimal failure covers of cardinality c including p has been obtained before the simulation starts. Let ub be a known upper bound for $\min_{b' \in MFC} |b' - b|$ satisfying $1 \leq ub \leq RL = \min_{b' \in MFC} |b'|$. According to (7), we can compute td by initializing td to ub & updating td by considering only the minimal failure covers b', $b' \cap b \neq \emptyset$ for which $|b' - b|$ can be $< ub$. Then, the first observation allows us to limit the set of minimal failure covers which have to be considered to those having a cardinality $|b'| < ub + |b|$. Furthermore, exploiting the second observation, from the minimal failure covers b' of cardinality c, only those with $|b' \cap b| \geq c - ub + 1$ have to be considered, and those minimal failure covers have to include some non-empty subbag of b of cardinality $\geq c - ub + 1 \geq \min\{R, c - ub + 1\} = q(c)$. Notice that $ub \geq 1$ implies $q(c) \leq \min\{R, c\}$. Also, $ub \leq \min_{b' \in MFC} |b'|$ implies $ub \leq c$, $c - ub + 1 \geq 1$, and, with $R \geq 1$, $q(c) \geq 1$. Therefore,

```
Compute_td(b, td)
Inputs: b
Outputs: td
td = RL;
for (increasing minimal failure cover cardinality c while c < td + |b|){
    q(c) = min{R, c − td + 1};
    for (each subbag p of b of cardinality q(c))
        if (p ∈ MFCS_q(c)(c))
            for (each b' ∈ MFCL_c(p)) td = min{td, |b' − b|};
}
```

Fig. 2. Description of procedure $Compute\_td$.

the $MFCS_{q(c)}(c)$ are available for all possible c. Then, for each minimal failure cover cardinality $c < ub + |b|$, we can generate all subbags p of b of cardinality $q(c)$, and, for each $p \in MFCS_{q(c)}(c)$, process only the minimal failure covers b' in $MFCL_c(p)$. As upper bound ub, we can use initially RL, and update ub to $|b' - b|$, if the latter is smaller, each time a minimal failure cover b' is processed. Note that, for all failure covers b', $|b' - b|$ is guaranteed to be ≥ 1, because x is not a down state, and, therefore, as required, $ub \geq 1$ throughout the processing. The discussion justifies the procedure $Compute\_td$ described in Fig. 2, where ub is held in the output parameter td.

Let $AMFC_f(c)$ be the subset of $AMFC_f$ including the after minimal failure covers of cardinality c associated with failure bag f. Then, a *selector* of $AMFC_f(c)$ is any non-empty subbag of some $b' \in AMFC_f(c)$. The procedure $Compute\_atd$ assumes that, for each after minimal failure cover cardinality c, the sets $AMFCS_{q'}(c)$ including all selectors of some $AMFC_f(c)$, $f \in F_B$ of cardinalities $q' = 1, 2, \ldots, \min\{R, c\}$ have been obtained before the simulation starts. It also assumes that, for each after minimal failure cover cardinality c, each selector $p \in AMFCS_{q'}(c)$, $q' = 1, 2, \ldots, \min\{R, c\}$, and each failure bag $f \in F_B$, a list $AMFCL_{f,c}(p)$ linking the after minimal failure covers of cardinality c associated with failure bag f and including p has been obtained before the simulation starts. The $atd(f)$, $f \in \{f' \in S_F : AMFC_{f'} = \emptyset\}$ are equal to td. Note that $S_F = active(x)$ for some state $x \in U$. Some successor of x, y, through a failure transition having associated with it some failure bag $f \in active(x)$ will have $td(y) = td(x) - 1$. According to Theorem 4, this can only happen if $AMFC_f \neq \emptyset$. Therefore, we can assume $\{f' \in S_F : AMFC_{f'} \neq \emptyset\} \neq \emptyset$. Let $f \in \{f' \in S_F : AMFC_{f'} \neq \emptyset\}$ & let ub_f be a known upper bound for $\min\{td, \min_{b' \in AMFC_f} |b' - b|\}$ satisfying $0 \leq ub_f \leq \min\{td, RL_f\}$. According to (8), $atd(f)$, $f \in \{f' \in S_F : AMFC_{f'} \neq \emptyset\}$ can be computed by initializing $atd(f)$ to ub_f, and updating $atd(f)$ by considering only the after minimal failure covers b', $b' \cap b \neq \emptyset$ associated with failure bag f for which $|b' - b|$ can be $< ub_f$. Letting $ub = \max_{f \in \{f' \in S_F : AMFC_{f'} \neq \emptyset\}} ub_f$, it is enough to con-

$$atd(f) = \begin{cases} td & \text{for } f \in \{f' \in S_F : AMFC_{f'} = \emptyset\}, \\ \min\left\{td, \min_{b' \in AMFC_f} |b' - b|\right\} & \text{for } f \in \{f' \in S_F : AMFC_{f'} \neq \emptyset\} \end{cases}. \quad (8)$$

```
Compute_atd(b, S_F, td, atd(f))
Inputs: b, S_F, td
Outputs: atd(f), f ∈ S_F
for (each f ∈ S_F : AMFC_f = ∅) atd(f) = td;
for (each f ∈ S_F : AMFC_f ≠ ∅) atd(f) = min{td, RL_f};
ub = max_{f∈{f'∈S_F : AMFC_f'≠∅}} atd(f);
for (increasing after minimal failure cover cardinality c ≥ 1 while c < ub + |b|){
    q(c) = min{R, c, max{1, c − ub + 1}};
    for (each subbag p of b of cardinality q(c))
        if (p ∈ AMFCS_q(c)(c))
            for (each f ∈ S_F : AMFC_f ≠ ∅)
                for (each b' ∈ AMFCL_f,c(p))
                    atd(f) = min{atd(f), |b' − b|};
}
```

Fig. 3. Description of procedure $Compute\_atd$.

sider after minimal failure covers b' for which $|b' - b|$ can be $< ub$. Then, the first observation allows us to limit the set of after minimal failure covers b' which have to be considered to those having a cardinality $|b'| < ub + |b|$. As will be discussed later, after minimal failure covers b' with $|b'| = 0$ do not have to be considered. Furthermore, exploiting the second observation, from the after minimal failure covers b' of cardinality $c \geq 1$, only those with $|b' \cap b| \geq \min\{c, \max\{1, c - ub + 1\}\}$ have to be considered (the maximum is introduced because there is no guarantee that $c - ub + 1$ will be ≥ 1; the minimum is introduced because for $ub = 0$, $c - ub + 1$ would be equal to $c+1$), and those after minimal failure covers have to be included in some non-empty subbag of b of cardinality $\geq \min\{c, \max\{1, c - ub + 1\}\} \geq \min\{R, c, \max\{1, c - ub + 1\}\} = q(c)$. Notice that $q(c) \leq \min\{R, c\}$. Also, $R \geq 1$ & $c \geq 1$ implies $q(c) \geq 1$. Therefore, the $AMFCS_{q(c)}(c)$ are available for all possible $c \geq 1$. Then, for each minimal failure cover cardinality c, $1 \leq c < ub + |b|$, we can generate all subbags p of b of cardinality $q(c)$, and, for each $p \in AMFCS_{q(c)}(c)$, process only the after minimal failure covers b' in $AMFCL_{f,c}(p)$, $f \in \{f' \in S_F : AMFC_{f'} \neq \emptyset\}$. As upper bounds ub_f, $f \in \{f' \in S_F : AMFC_{f'} \neq \emptyset\}$ satisfying $0 \leq ub_f \leq \min\{td, RL_f\}$ we can use $\min\{td, RL_f\}$, which yields $ub = \max_{f \in \{f' \in S_F : AMFC_{f'} \neq \emptyset\}} \min\{td, RL_f\}$. The discussion justifies the procedure $Compute\_atd$ described in Fig. 3. It remains to discuss the case in which $AMFC_f \neq \emptyset$ & $AMFC_f$ includes $\emptyset$. In that case, $RL_f = 0$, and the procedure will return $atd(f) = 0$, which is correct.

The parameter R controls the degree of selection. As R increases, $q(c)$ may increase, decreasing the number of minimal failure covers or after minimal failure covers which are processed, but increasing the cost of generating the subbags of b. In our experiments, we have found $R = 2$ to be a good selection.

IV. Optimization and Simulation Control

In this section, we deal with the optimization of the simulation. Two issues can be considered. The first one is to optimize the distribution of the simulation effort between the generation of the m samples of Z' & the generation of the n samples of W. The second one is the optimization of the parameters of the importance sampling scheme. We will deal first with the first issue. This will result in a version (NOPT) of the simulation method in which the importance sampling scheme is not optimized. Then, we will deal with the second issue, resulting in a version (OPT) of the simulation method in which the importance sampling scheme is optimized.

Assume $E_{P'}\{Z'\}$, $\mathrm{Var}_{P'}\{Z'\}$, $E_P\{W\}$, and $\mathrm{Var}_P\{W\}$ known. Then, for large m & n, the relative halfwidth of the confidence interval for UA is approximately equal (4) to

$$\left(\left(\frac{\beta}{\sqrt{m}}\right)^2 + \left(\frac{\gamma}{\sqrt{n}}\right)^2\right)^{1/2},$$

with $\beta = z_\alpha \sqrt{\mathrm{Var}_{P'}\{Z'\}}/E_{P'}\{Z'\}$ & $\gamma = z_\alpha \sqrt{\mathrm{Var}_P\{W\}}/E_P\{W\}$. Given a target relative halfwidth for the confidence interval, δ, and using the number of generated regenerative cycles as a measure of the simulation effort, the problem of distributing optimally the simulation effort between obtaining the m samples Z'_i & obtaining the n samples W_i can be stated as, given $\beta > 0$, $\gamma > 0$, and $\delta > 0$, find the values, $m' > 0$ & $n' > 0$, of, respectively, m & n, which minimize $m + n$, under the restriction

$$\left(\left(\frac{\beta}{\sqrt{m}}\right)^2 + \left(\frac{\gamma}{\sqrt{n}}\right)^2\right)^{1/2} = \delta.$$

The solution of that minimization problem is

$$n' = \left(\frac{\gamma}{\delta}\right)^2 \left(1 + \left(\frac{\beta}{\gamma}\right)\right),$$
$$m' = \left(\frac{\beta}{\gamma}\right) n'.$$

Inputs: r_rhw, α, max_rc, K, FB, DB
Outputs: $\widehat{UA}$, rhw
$l\_m = 0$; $l\_n = 0$; $m = K$; $n = K$;
do{
  for ($i = l\_m + 1$; $i \leq m$; i++)
    Generate a regenerative cycle under P' and collect Z_i, L_i;
  for ($i = l\_n + 1$; $i \leq n$, i++)
    Generate a regenerative cycle under P and collect W_i;
  $l\_m = m$; $l\_n = n$;
  $\overline{Z'} = (\sum_{i=1}^{m} Z_i L_i)/m$; $\overline{W} = (\sum_{i=1}^{n} W_i)/n$;
  if ($\overline{Z'} == 0$){ m += K; $done$ = NO; }
  else{
    $S^2(Z') = (\sum_{i=1}^{m}(Z_i L_i - \overline{Z'})^2)/(m-1)$; $S^2(W) = (\sum_{i=1}^{n}(W_i - \overline{W})^2)/(n-1)$;
    $rhw = z_\alpha\left(\left(\sqrt{S^2(Z')/m}/\overline{Z'}\right)^2 + \left(\sqrt{S^2(W)/n}/\overline{W}\right)^2\right)^{1/2}$;
    if ($rhw \leq r\_rhw$ || $m + n \geq max\_rc$) $done$ = YES;
    else{
      $done$ = NO;
      $\widehat{\beta} = z_\alpha\sqrt{S^2(Z')}/\overline{Z'}$; $\widehat{\gamma} = z_\alpha\sqrt{S^2(W)}/\overline{W}$;
      if ($\widehat{\gamma} == 0$) n += K;
      else {
        $r\_n = (\widehat{\gamma}/r\_rhw)^2(1 + (\widehat{\beta}/\widehat{\gamma}))$; $r\_m = (\widehat{\beta}/\widehat{\gamma})r\_n$;
        if ($r\_m + r\_n > max\_rc$){
          $factor = max\_rc/(r\_m + r\_n)$; $r\_m = factor * r\_m$; $r\_n = factor * r\_n$;
          $r\_m = (\lfloor r\_m/K \rfloor + 1)K$; $r\_n = (\lfloor r\_n/K \rfloor + 1)K$;
          if ($r\_m + r\_n < max\_rc$) r_m += K;
        }
        else { $r\_m = (\lfloor r\_m/K \rfloor + 1)K$; $r\_n = (\lfloor r\_n/K \rfloor + 1)K$; }
        $m = \min\{2m, \max\{m, r\_m\}\}$; $n = \min\{2n, \max\{n, r\_n\}\}$;
      }
    }
  }
}
while (!$done$);
$\widehat{UA} = \overline{Z'}/\overline{W}$;

Fig. 4. Version (NOPT) of the simulation method in which the importance sampling scheme is not optimized.

Note that m'/n' is independent of δ. In addition, the following relation between δ & $m' + n'$ can be easily found.

$$\delta = \frac{1}{\sqrt{m' + n'}}\beta\left(1 + \left(\frac{\gamma}{\beta}\right)\right). \tag{9}$$

For the theory supporting the confidence interval to be valid (see Section II-C), m & n have to be multiples of some not too small integer k. Furthermore, $E_{P'}\{Z'\}$, $\mathrm{Var}_{P'}\{Z'\}$, $E_P\{W\}$, and $\mathrm{Var}_P\{W\}$ are unknown, and have to be estimated using sample means & variances $\overline{Z'}$, $S^2(Z')$, $\overline{W}$, and $S^2(W)$ collected during the simulation. Taking all this into account, Fig. 4 describes a reasonable implementation of the version (NOPT) of the simulation method in which the importance sampling scheme is not optimized. To be specific, Fig. 4 assumes that the importance sampling scheme is either FTDB or BFTDB. The implementation has as inputs a required relative halfwidth r_rhw for the confidence interval, the probability α of the confidence interval, a maximum allowed number of regenerative cycles max_rc, a parameter K such that the k parameter on which the confidence interval given by (4) is based is guaranteed to be not smaller than K, and the values of the parameters of the importance sampling scheme. The implementation has as outputs the estimate of UA, $\widehat{UA}$, and the relative halfwidth of the finally obtained confidence interval, rhw. The implementation uses two simulation streams: one to obtain the samples $Z_i' = Z_i L_i$, and the other to obtain the samples W_i. The streams advance at control steps until the required relative confidence interval halfwidth is achieved, or a number of regenerative cycles equal or slightly larger than max_rc have been generated, with the lengths (number of regenerative cycles) of both streams being equal to K at the first control step, and at most doubling between consecutive control steps. At each control step, estimates r_m & r_n of the optimally distributed m & n which would be required to achieve the requested relative confidence interval halfwidth are obtained using estimates $\widehat{\beta}$ & $\widehat{\gamma}$ of, respectively, β & γ. If $r\_m + r\_n \leq max\_rc$, the streams are allowed to progress with scheduled lengths equal to r_m & r_n, corrected so that they are multiples of K. However, if $r\_m + r\_n > max\_rc$, the scheduled lengths are obtained by scaling r_m, r_n so that $r\_m + r\_n$

is approximately equal to $max\_rc, \geq max\_rc$ (if the latter condition were not satisfied, the implementation could "hang"), and both r_m & r_n are multiples of K. That scaling yields an optimal distribution of the simulation effort because, as pointed out previously, m'/n' is independent of the target relative confidence interval halfwidth, δ. To avoid divisions by 0, the cases $\overline{Z'} = 0$ & $\widehat{\gamma} = 0$ receive special treatments.

We deal next with the optimization of the importance sampling schemes. To be specific, we will discuss with detail the OPT version of the simulation method for the importance sampling schemes FTDB & BFTDB. For large m, the component of the relative halfwidth of the confidence interval attributable to the samples Z'_i is approximately equal (4) to

$$z_\alpha \frac{\sqrt{\mathrm{Var}_{P'}\{Z'\}}}{\sqrt{m}E_{P'}\{Z'\}} = z_\alpha \frac{\sqrt{\mathrm{Var}_{P'}\{Z'\}}}{\sqrt{m}E_P\{Z\}}.$$

Then, a natural goal is to minimize $\mathrm{Var}_{P'}\{Z'\}$, which depends on the values of the parameters of the importance sampling scheme. To remark that dependence, we will denote the generic probability measure (as a function of the biasing parameters) by P°, will denote the quantities referred to that generic probability measure using "$\circ$", will reserve P' to denote the probability measure actually used, and will continue using "$'$" to denote the quantities referred to P'. With that notation, the goal is to minimize $\mathrm{Var}_{P^\circ}\{Z^\circ\}$. Using $E_{P^\circ}\{Z^\circ\} = E_{P'}\{Z'\} = E_P\{Z\}$, we get

$$\begin{aligned}\mathrm{Var}_{P^\circ}\{Z^\circ\} &= E_{P^\circ}\{Z^{\circ 2}\} - E_{P^\circ}\{Z^\circ\}^2 \\ &= E_{P^\circ}\{Z^{\circ 2}\} - E_P\{Z\}^2,\end{aligned}$$

and, similarly,

$$\mathrm{Var}_{P'}\{Z'\} = E_{P'}\{Z'^2\} - E_P\{Z\}^2,$$

which combined give

$$\mathrm{Var}_{P^\circ}\{Z^\circ\} = \mathrm{Var}_{P'}\{Z'\} + E_{P^\circ}\{Z^{\circ 2}\} - E_{P'}\{Z'^2\}. \quad (10)$$

Using $Z^\circ = ZL^\circ$, $L^\circ(\omega) = P\{\omega\}/P^\circ\{\omega\}$, and $L(\omega) = P\{\omega\}/P'\{\omega\}$, we get

$$\begin{aligned}E_{P^\circ}\{Z^{\circ 2}\} &= E_{P^\circ}\{Z^2 L^{\circ 2}\} \\ &= \sum_{\omega\in\mathcal{S}} Z(\omega)^2 L^\circ(\omega)^2 P^\circ\{\omega\} \\ &= \sum_{\omega\in\mathcal{S}} Z(\omega)^2 L(\omega)^2 \frac{P'\{\omega\}^2}{P^\circ\{\omega\}^2} P^\circ\{\omega\} \\ &= \sum_{\omega\in\mathcal{S}} Z(\omega)^2 L(\omega) \frac{P\{\omega\}}{P'\{\omega\}} \frac{P'\{\omega\}^2}{P^\circ\{\omega\}} \\ &= \sum_{\omega\in\mathcal{S}} Z(\omega)^2 L(\omega) \frac{P\{\omega\}}{P^\circ\{\omega\}} P'(\omega) \\ &= \sum_{\omega\in\mathcal{S}} Z(\omega)^2 L(\omega) L^\circ(\omega) P'\{\omega\},\end{aligned}$$

proving that $E_{P^\circ}\{Z^{\circ 2}\}$ is the expected value under P' of the r.v.

$$N = Z^2 L L^\circ.$$

Then, the sample mean of N under P', $\overline{N}$, is an unbiased estimator of $E_{P^\circ}\{Z^{\circ 2}\}$. Denoting by Q the r.v. $Q = Z'^2 = Z^2 L^2$, by $\overline{Q}$ the sample mean of Q under P', by $S^2(Z')$ the sample variance of Z' under P', and using (10), an unbiased estimator of $\mathrm{Var}_{P^\circ}\{Z^\circ\}$ is

$$\widehat{\mathrm{Var}_{P^\circ}\{Z^\circ\}} = \left(S^2(Z') + \overline{N} - \overline{Q}\right), \quad (11)$$

where $S^2(Z')$ is given by (5), and $\overline{N}$ & $\overline{Q}$ are given by

$$\overline{N} = \frac{1}{m}\sum_{i=1}^m N_i = \frac{1}{m}\sum_{i=1}^m Z_i^2 L_i L_i^\circ, \quad (12)$$

$$\overline{Q} = \frac{1}{m}\sum_{i=1}^m Q_i = \frac{1}{m}\sum_{i=1}^m Z_i^2 L_i^2, \quad (13)$$

L_i° being the sample of L° corresponding to the ith regenerative cycle.

The quantities $S^2(Z')$ & $\overline{Q}$ of (11) are constants, but $\overline{N}$ depends on the biasing parameters. The optimization of the biasing parameters is done using a symbolic estimate of $\mathrm{Var}_{P^\circ}\{Z^\circ\}$ based on a symbolic expression $\overline{N}(FB, DB)$ for $\overline{N}$. Denoting by $P^\circ_{x,y}$ the value of the biased probabilities under P°, and $(s_0(i), s_1(i), \ldots, s_{l(i)}(i))$ being the ith sampled regenerative cycle, L_i° has the expression

$$L_i^\circ = \prod_{j=0}^{l_D\left(s_0(i), s_1(i), \ldots, s_{l(i)}(i)\right)} \frac{P_{s_j(i), s_{j+1}(i)}}{P^\circ_{s_j(i), s_{j+1}(i)}},$$

with $l_D(s_0, s_1, \ldots, s_l) = \max\{k \leq l-1 : s_0, s_1, \ldots, s_k \in U\}$. But, in both FTDB & BFTDB, each $P^\circ_{s_j(i), s_{j+1}(i)}$ has a form

$$\begin{aligned}P^\circ_{s_j(i), s_{j+1}(i)} &= A_{i,j} \times FB^{n_1(i,j)} \\ &\quad \times (1-FB)^{n_2(i,j)} DB^{n_3(i,j)} (1-DB)^{n_4(i,j)},\end{aligned}$$

with $A_{i,j} > 0$ & $n_k(i,j)$, $1 \leq k \leq 4$ integers equal to 0 or 1. This implies that L_i° has a form

$$L_i^\circ = \frac{A_i}{FB^{n_1(i)}(1-FB)^{n_2(i)} DB^{n_3(i)} (1-DB)^{n_4(i)}},$$

with $A_i > 0$ & $n_k(i)$, $1 \leq k \leq 4$ integers ≥ 0. Each term of the summation of (12) has, then, a form

$$Z_i^2 L_i L_i^\circ = \frac{B_i}{FB^{n_1(i)}(1-FB)^{n_2(i)} DB^{n_3(i)} (1-DB)^{n_4(i)}},$$

with $n_k(i)$ as above & $B_i > 0$, and a symbolic expression $\overline{N}(FB, DB)$ for $\overline{N}$ can then be obtained by grouping terms $Z_i^2 L_i L_i^\circ$ with same $n_k(i)$, $1 \leq k \leq 4$, resulting in an expression

$$\overline{N}(FB, DB) = \frac{G(FB, DB)}{m}, \quad (14)$$

$$G(FB,DB) = \sum_{l=1}^{p} \frac{C_l}{FB^{n_1(l)}(1-FB)^{n_2(l)}DB^{n_3(l)}(1-DB)^{n_4(l)}}, \quad (15)$$

with $C_l > 0$ & different 4-tuples $(n_1(l), n_2(l), n_3(l), n_4(l))$ for different l. The symbolic estimate of $\mathrm{Var}_{P^\circ}\{Z^\circ\}$ is defined by (11) with $\overline{N}$ replaced by $\overline{N}(FB,DB)$, (5), (14), (15), and (13).

The optimization of the importance sampling scheme can be embedded into the NOPT simulation algorithm by doing two things: 1) building $\overline{Q}$ & the symbolic expression $\overline{N}(FB,DB)$, using initially some values FBI & DBI for, respectively, FB & DB, and 2) at each control step, estimating the optimum values of the biasing parameters by minimizing the symbolic estimate of $\mathrm{Var}_{P^\circ}\{Z^\circ\}$, deciding whether it would be beneficial to use thereafter the estimated optimum values of the biasing parameters, and changing FB & DB to those values if that is estimated to be beneficial. In making the decision, it has to be taken into account that, if the values of the biasing parameters are changed, the collected Z_i & L_i before the change will have to be thrown away.[8] The function $G(FB,DB)$ is "cleaned" $(p = 0)$ when the values of the biasing parameters are changed. Fig. 5 gives a detailed description of the resulting OPT version of the simulation method. The method has as inputs the required confidence interval halfwidth, r_rhw, the probability of the confidence interval, α, the maximum number of regenerative cycles, max_rc, a parameter K with the same meaning as for the NOPT version, and the initial values for the biasing parameters. The method has as outputs the estimate, $\widehat{UA}$, of UA & the achieved relative confidence interval halfwidth, rhw. The method uses the procedure $Decide\_change$ described in Fig. 6, which determines whether it is beneficial or not to substitute the currently used values of FB & DB by the respective estimated optimum ones, FB_o & DB_o. The variable d_m holds the number of regenerative cycles of the biased simulation stream which have been thrown away. The anomalous, but possible, case $V \leq 0$ receives a special treatment.

The procedure $Decide\_change$ has as inputs the estimated required lengths, r_m & r_n (r_m_o & r_n_o), of, respectively, the biased stream & the unbiased stream, not taking into account the limitation on the number of regenerative cycles, for the current (estimated optimum) values of the biasing parameters; the number of regenerative cycles of the biased stream, d_m, which have been thrown away; the number of regenerative cycles, m, of the biased stream which have been generated with the current values of the biasing parameters; the number of regenerative cycles, n, of the unbiased stream which have been generated so far; the estimates $\widehat{\beta}$ & $\widehat{\gamma}$ of, respectively, the β & γ constants, for the current values of the biasing parameters; an estimate of the β constant, $\widehat{\beta}_o$, under the estimated optimum values of the biasing parameters; and the allowed maximum number of regenerative cycles, max_rc. The procedure has as output a boolean variable $change$ conveying the taken decision. Ignoring the limitation on the number of regenerative cycles, the number of such cycles which would be consumed to achieve the required relative confidence interval halfwidth can be estimated as $a = \max\{m, r\_m\} + \max\{n, r\_n\} + d\_m$, if the values of the biasing parameters are not changed, & as $b = r\_m_o + \max\{n, r\_n_o\} + d\_m + m$, if the values are changed. The procedure deals first with the case $a \leq max\_rc$. In that case, it is estimated that the required relative confidence interval halfwidth will be achieved if the change is not made, and the change should be made only if it is estimated that this would allow us to achieve the required relative confidence interval halfwidth with a smaller number of regenerative cycles, i.e. if $b < a$. The procedure deals next with the case $a > max\_rc$. In that case, it is estimated that, if the change is not made, max_rc regenerative cycles will be generated, and, using (9), the achieved relative confidence interval can be estimated as

$$est\_rhw = \frac{1}{\sqrt{max\_rc - d\_m}}\widehat{\beta}\left(1 + \left(\frac{\widehat{\gamma}}{\widehat{\beta}}\right)\right) > r\_rhw.$$

If $b < max\_rc$, it can be estimated that the required relative confidence interval halfwidth will be achieved if the change is made, the change should be made, and, because, according to (9), the achieved relative confidence interval halfwidth is decreasing with the number of regenerative cycles, we have

$$c = \frac{1}{\sqrt{max\_rc - d\_m - m}}\widehat{\beta}_o\left(1 + \left(\frac{\widehat{\gamma}}{\widehat{\beta}_o}\right)\right) < r\_rhw.$$

For $b \geq max\_rc$, max_rc cycles will be generated if the change is made, the achieved relative confidence interval halfwidth can be estimated by c, and the change should be made only if $c < est\_rhw$. Taking all this into account, for the case $a > max\_rc$, the procedure $Decide\_change$ directs us to make the change if $c < est\_rhw$, covering well both the case $b < max\_rc$ & the case $b \geq max\_rc$.

The OPT version of the simulation method requires the solution of a minimization problem. Because $S^2(Z')$ & $\overline{Q}$ are constants, minimization of $\widehat{\mathrm{Var}_{P^\circ}\{Z^\circ\}}$ is equivalent to minimization of $\overline{N}(FB,DB)$, and we minimize $\overline{N}(FB,DB)$. To guarantee the BRE properties, the biasing parameters FB & DB have to be apart from both 0 & 1 by a quantity larger than some constant, and then, $\overline{N}(FB,DB)$ is minimized in a domain $[PINT, 1-PINT]^2$ with $0 < PINT < 0.5$. In our implementation, we use $PINT = 0.1$. Theorem 7 in the Appendix establishes that $\overline{N}(FB,DB)$ is convex in $[PINT, 1-PINT]^2$. Because the domain $[PINT, 1-PINT]^2$ is convex, any local minimum of the function in that domain is also a global minimum of the function in the domain [22]. Furthermore, being the function differentiable & convex, any stationary point in the domain is a global minimum (see, for instance, [26, Theorem 8.12]). Then, the steepest-descent method (see, for instance, [8]) is guaranteed to find the global minimum of the function in the constrained domain. The cost of the optimization is roughly proportional to the number of different terms p in the symbolic expression of $\overline{N}(FB,DB)$ & to the number of moves of the minimization method. We start the optimization from the point in the domain $[PINT, 1-PINT]^2$ corresponding to the currently used values of the biasing parameters & limit the number of moves to 1,000. That approach has worked well in our experiments.

[8]Trying to use the samples obtained before the change does not seem feasible. This is because of the dependence of the samples after the change on the samples obtained before the change, which makes very difficult, if at all possible, the derivation of a confidence interval for $\widehat{UA}$.

Inputs: r_rhw, α, max_rc, K, FBI, DBI
Outputs: $\widehat{UA}$, rhw
$l\_m = 0$; $l\_n = 0$; $m = K$; $n = K$; $d\_m = 0$; $FB = FBI$; $DB = DBI$;
make an empty expression $G(FB, DB)$;
do{
  for ($i = l\_m + 1; i \le m; i$++)
    Generate a regenerative cycle under P', collect Z_i, L_i and update $G(FB, DB)$;
  for ($i = l\_n + 1; i \le n; i$++)
    Generate a regenerative cycle under P and collect W_i;
  $l\_m = m$; $l\_n = n$;
  $\overline{Z'} = (\sum_{i=1}^{m} Z_i L_i)/m$; $\overline{Q} = (\sum_{i=1}^{m} Z_i^2 L_i^2)/m$; $\overline{W} = (\sum_{i=1}^{n} W_i)/n$;
  if ($\overline{Z'} == 0$){ m += K; $done$ = NO; }
  else{
    $S^2(Z') = (\sum_{i=1}^{m}(Z_i L_i - \overline{Z'})^2)/(m-1)$; $S^2(W) = (\sum_{i=1}^{n}(W_i - \overline{W})^2)/(n-1)$;
    $rhw = z_\alpha\left(\left(\sqrt{S^2(Z')/m}/\overline{Z'}\right)^2 + \left(\sqrt{S^2(W)/n}/\overline{W}\right)^2\right)^{1/2}$;
    if ($rhw \le r\_rhw$ || $m + n + d\_m \ge max\_rc$) $done$ = YES;
    else{
      $done$ = NO;
      $\widehat{\beta} = z_\alpha\sqrt{S^2(Z')}/\overline{Z'}$; $\widehat{\gamma} = z_\alpha\sqrt{S^2(W)}/\overline{W}$;
      if ($\widehat{\gamma} == 0$) n += K;
      else {
        $r\_n = (\widehat{\gamma}/r\_rhw)^2(1 + (\widehat{\beta}/\widehat{\gamma}))$; $r\_m = (\widehat{\beta}/\widehat{\gamma})r\_n$;
        Find the values, FB_o and DB_o, of, respectively, FB and DB
        which minimize $\widehat{\mathrm{Var}_{P^\circ}\{Z^\circ\}} = S^2(Z') + G(FB, DB)/m - \overline{Q}$;
        Let V be the value of $\widehat{\mathrm{Var}_{P^\circ}\{Z^\circ\}}$ for $FB = FB_o$ and $CB = CB_o$;
        if ($V \le 0$) $change$ = NO;
        else{
          $\widehat{\beta}_o = z_\alpha\sqrt{V}/\overline{Z'}$; $r\_n_o = (\widehat{\gamma}/r\_rhw)^2(1 + (\widehat{\beta}_o/\widehat{\gamma}))$; $r\_m_o = (\widehat{\beta}_o/\widehat{\gamma})r\_n_o$;
          $Decide\_change(r\_m, r\_n, r\_m_o, r\_n_o, d\_m, m, n, \widehat{\beta}, \widehat{\gamma}, \widehat{\beta}_o, max\_rc, change)$;
        }
        if ($change$){
          $FB = FB_o$; $DB = DB_o$;
          d_m += m; $l\_m = 0$; $r\_m = r\_m_o$; $r\_n = r\_n_o$; clean $G(FB, DB)$;
        }
        if ($d\_m + r\_m + r\_n > max\_rc$){
          $factor = (max\_rc - d\_m)/(r\_m + r\_n)$;
          $r\_m = factor * r\_m$; $r\_n = factor * r\_n$;
          $r\_m = (\lfloor r\_m/K \rfloor + 1)K$; $r\_n = (\lfloor r\_n/K \rfloor + 1)K$;
          if ($d\_m + r\_m + r\_n < max\_rc$) r_m += K;
        }
        else { $r\_m = (\lfloor r\_m/K \rfloor + 1)K$; $r\_n = (\lfloor r\_n/K \rfloor + 1)K$; }
        if ($change$) $m = \min\{2m, r\_m\}$; else $m = \min\{2m, \max\{m, r\_m\}\}$;
        $n = \min\{2n, \max\{n, r\_n\}\}$;
      }
    }
  }
}
while (!$done$);
$\widehat{UA} = \overline{Z'}/\overline{W}$;

Fig. 5. Version (OPT) of the simulation method in which the importance sampling scheme is optimized.

The method for optimizing the biasing parameters used in the OPT version of the simulation method has connections with optimization methods for discrete-event systems (see, for instance, [25]), in which a symbolic estimate for the gradient of the objective function with respect to the optimization parameters is obtained from samples of the system, and used to estimate the values of the parameters which maximize or minimize the objective function. The difference is that, due to the special structure of our problem, we are able to derive a symbolic expression for an estimate of the objective function itself, and use that estimate directly in combination with a standard optimization method.

```
Decide_change(r_m, r_n, r_m_o, r_n_o, d_m, m, n, β̂, γ̂, β̂_o, max_rc, change)
Inputs: r_m, r_n, r_m_o, r_n_o, d_m, m, n, β̂, γ̂, β̂_o, max_rc
Outputs: change
a = max{m, r_m} + max{n, r_n} + d_m;
if (a ≤ max_rc)
    if (r_m_o + max{n, r_n_o} + d_m + m < a) change = YES;
    else change = NO;
else{
    est_rhw = (1/√(max_rc − d_m)) β̂(1 + (γ̂/β̂));
    c = (1/√(max_rc − d_m − m)) β̂_o(1 + (γ̂/β̂_o));
    if (c < est_rhw) change = YES;
    else change = NO;
}
```

Fig. 6. Description of procedure $Decide\_change$.

Although the NOPT & OPT versions of the simulation method have been formally described for the FTDB & BFTDB importance sampling schemes, it should be clear that the versions can be easily adapted to accommodate the FB & BFB schemes, the only difference being that in those cases there is a single biasing parameter, FB.

V. Implementation Details

A prototype software package has been developed based on the METFAC tool offering both the OPT & the NOPT versions of the simulation method for the steady-state unavailability. The user can select FB, BFB, FTDB, or BFTDB as the importance sampling scheme. Model specification is done in METFAC using a strongly typed production rule-based modeling language. Production rules can be either *simple* or *with responses*. A simple production rule specifies a simple action describing a way in which the state of the CTMC may change & at which rate. Production rules with responses describe actions with responses, where each response describes a way in which the state of the model may change. In actions with responses, a rate is associated with the action, and a probability, with a default value of 1, is associated with each response. The transition rates corresponding to the state changes are then obtained by multiplying the action rate by the response probability. A model specification also includes the description of a "start" state from which the CTMC can be generated & a reward rate specification which associates reward rates with the states of the CTMC. A comment is in order. It could happen that the model specification is such that different action/response pairs lead from a given current state x to the same state y. Our implementation of the importance sampling schemes treats those pairs as leading to different states, because doing it otherwise would require us to generate from every current state x the descriptions of all its successors, and to compare them, which would be expensive. The strategy that our implementation follows is equivalent to simulating a CTMC X^* with expanded state space, Ω^*, with respect to the true CTMC X. This strategy only affects the performance under BFB & BFTDB, and does not rule out the theoretical properties of those schemes, because Ω^* is finite if, as we assume, Ω is.[9]

The METFAC tool allows the specification of arbitrary finite rewarded CTMC models. The simulation methods described in this paper cover however a particular class of CTMC models. Furthermore, they require high-level information. That high-level information is provided using three auxiliary files & labels optionally associated with actions & responses. The first file specifies the bag of component classes C making up the modeled system; the second file specifies the failure bags; the third file gives the minimal failure covers. Actions without responses are required to have a label with a predefined syntax which identifies "failure" & "repair" actions & tells the bag of component classes involved in the action. In actions with responses, that information is given in the responses' labels. Using the labels, it is possible to check that all transitions of the CTMC are either of "failure" type or of "repair" type, and to identify during the simulation the failure bag associated with a failure transition & the bag of components affected in a repair transition. The "start" state of the model specification is required to be the state r without failed components. Because the bag of component classes affected by each transition of the model is known, it is possible to compute along the simulation the bag of failed component classes $F(x)$ in each generated state x, and to compute $B(x)$ from $F(x)$.

The model specification is preprocessed, and a model-specific simulator interface is automatically generated including a function giving the "start" state, a function returning the action/response pairs which are active in a given state, two functions returning respectively the rate of an action & the probability of a response in a given state, another function returning the description of the state reached from a given state through an action/response pair, and a function giving the reward rate associated with a given state. A reward rate 1 is interpreted as indicating that the state is "down". A reward rate 0 is interpreted as indicating that the state is "up". Some assumed characteristics

[9] Ω^* can be obtained from Ω by considering in some arbitrary order the states $x \in \Omega$ having transition rates to them due to more than one action/response pair, and substituting the state under consideration x by as many copies as necessary so that each copy is reached from every predecessor with transition rates due to only an action/response pair. Because the number of transformation steps is finite & at each step the number of state copies is finite, Ω^* will be finite.

of the CTMC are checked during the simulation. This includes checking that reward rates are either 0 or 1.

Regarding the procedures for computing failure transition distances described in Section III-D, the prototype software package uses a hash table to hold the subsets of selectors $MFCS_q(c)$ with a key including c & the selector description, allowing an efficient test of inclusion in some subset $MFCS_q(c)$ of a potential selector. The subsets of selectors $AMFCS_q(c)$ are held in an independent similar hash table.

A potential problem of the simulation methods, as described in Figs. 4 & 5, is that, when the lengths of the simulation streams are large, the storage requirements are large, due to the space required to hold Z_i, L_i, $1 \leq i \leq m$, and W_i, $1 \leq i \leq n$. That storage can be avoided (and our actual implementation of the simulation methods does avoid it) by noting that $\overline{Z'}$, $\overline{Q}$, $\overline{W}$, $S^2(Z')$, and $S^2(W)$ can be expressed as

$$\begin{aligned}
\overline{Z'} &= \frac{S_1}{m}, \\
\overline{Q} &= \frac{S_2}{m}, \\
\overline{W} &= \frac{S_3}{n}, \\
S^2(Z') &= \frac{S_2}{m-1} - \frac{S_1^2}{m(m-1)}, \\
S^2(W) &= \frac{S_4}{n-1} - \frac{S_3^2}{n(n-1)},
\end{aligned}$$

with $S_1 = \sum_{i=1}^{m} Z_i L_i$, $S_2 = \sum_{i=1}^{m} Z_i^2 L_i^2$, $S_3 = \sum_{i=1}^{n} W_i$, $S_4 = \sum_{i=1}^{n} W_i^2$. Then, it is enough to use variables holding S_1, S_2, S_3, and S_4, and update them as regenerative cycles are obtained.

VI. Experimental Results and Discussion

For balanced fault-tolerant systems, there is theoretical evidence that simulation of the steady-state unavailability under FTDB & BFTDB can be indeed significantly more efficient than under FB & BFB. However, that theoretical evidence is for small enough ε. How small does ε have to be in practice? How does BFTDB, which has sound theoretical properties for unbalanced fault-tolerant systems, perform in relation to BFB for such systems? How large can the overheads due to failure transition distances computation be expected to be? In this section we try to provide answers to those questions using numerical examples. We also analyse the performance of the optimization method for the values of the biasing parameters, and to what extent optimization can help to deal with hard cases, i.e. cases in which the simulation should not be focused alone on the regenerative cycles in $\mathcal{S}_{k_{\min}}$. Finally, we will analyse how FTDB & BFTDB can be adapted to take into account limited knowledge of minimal failure covers, i.e. minimal failure covers of cardinality $\leq M$, and to what extent the adaptation affects the performance of FTDB & BFTDB. The experimental evidence provided by the analysis performed in this section is, obviously, limited.

A. 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 & B, with two processors per set, two sets of disk controllers with two controllers per set, and six disk clusters with four disks per cluster. Each set of controllers commands three disk clusters. The system is up iff at least one processor in each set, one controller in each set, and at least three disks in each disk cluster are operational. 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, $f_{\min}/f_{\max} = 0.025$ & $\varepsilon = f_{\max}/r_{\min} = 7.2 \times 10^{-4}$, and, therefore, the instance can be considered a balanced fault-tolerant system. For instance II, $f_{\min}/f_{\max} = 2.5 \times 10^{-4}$ & $\varepsilon = f_{\max}/r_{\min} = 7.2 \times 10^{-4}$, 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 Fig. 7. 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 & the CU communicate using two local area networks (LAN), La, Lb, to which each DPU & 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" & "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 & 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}, and μ_{DPUh}. The restart rate of DPU in "soft" failed mode is denoted by μ_{DPUs}. The system is up iff there is an operational FE & 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 & 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. The front-ends can be conceptualized as

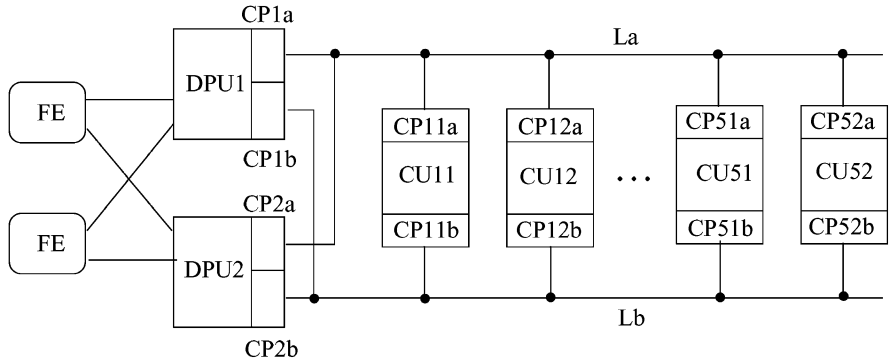

Fig. 7. Architecture of the fault-tolerant control system (FTC example).

TABLE I
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
$\mu_{\mathrm{DPU_S}}$	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

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 I. For set A, $f_{\min}/f_{\max} = 0.0556$ & $\varepsilon = f_{\max}/r_{\min} = 2.88\times10^{-4}$; for set B, $f_{\min}/f_{\max} = 0.0556$ & $\varepsilon = f_{\max}/r_{\min} = 2.88\times10^{-3}$; for set C, $f_{\min}/f_{\max} = 5.56\times10^{-3}$ & $\varepsilon = f_{\max}/r_{\min} = 5.76\times10^{-3}$; for set D, $f_{\min}/f_{\max} = 4\times10^{-3}$ & $\varepsilon = f_{\max}/r_{\min} = 5\times10^{-3}$. Thus, the fault-tolerant system can be considered balanced for sets A & B, and unbalanced for sets C & 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 FTDB & BFTDB in a hard scenario which defies the heuristic supporting those importance sampling schemes.

B. Comparison of Importance Sampling Schemes

In this section, we compare the performances of FTDB & BFTDB with those of FB & BFB in the simulation methods described in Section IV. We will call FB_o, BFB_o, FTDB_o, and BFTDB_o (FB_n, BFB_n, FTDB_n, and BFTDB_n) the OPT (NOPT) simulation method under, respectively, the importance sampling schemes FB, BFB, FTDB & BFTDB. For FTDB_n & BFTDB_n we use the values $FB = 0.9$ & $DB = 0.9$. For FB_n & BFB_n we use $FB = 0.5$. For FTDB_o & BFTDB_o we use the initial values for the biasing parameters $FBI = 0.8$ & $DBI = 0.8$. For FB_o & BFB_o we use $FBI = 0.8$. In addition, we use $K = 1{,}000$, and, for the methods using FTDB & BFTDB, we use $R = 2$. These choices are used in all the experiments reported in the paper. We will also give results for the standard regenerative simulation method, which will be denoted by REG. The simulation methods are run with a target 99% confidence interval of $\pm0.2\%$ & a maximum number of regenerative cycles $max\_rc = 10{,}000{,}000$. All CPU times are measured on a workstation with a Sun-Blade-1000 processor. Table II summarizes the obtained results. We give the estimate, number of regenerative cycles, and CPU times under BFTDB_o; and, for the other simulation method/importance sampling scheme combinations, we give the slow down factor, t_{sd}, defined as the ratio between the CPU times required under those simulation method/importance sampling scheme combinations & the CPU time required in BFTDB_o to achieve a confidence interval of same relative halfwidth. When the target confidence interval is not achieved, we compute t_{sd} 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. That rule is reasonable, because almost all of the computational effort is spent in obtaining the samples Z_i', and, for the method with optimization of the biasing parameters, the biasing parameters get stabilized very soon.

We will first comment on the impact of the optimization of the biasing parameters on the performance of the simulation method. In most cases, optimization has little effect on the performance of the simulation, the only exceptions being the FTC (D) example under BFB, FTDB, and BFTDB. The explanation for this is that, in most cases, the values for the biasing parameters used in the methods without optimization are close to the optimal ones. To illustrate the point, Table III gives the optimized values of FB & DB under the importance sampling scheme BFTDB for all examples. Significant differences with respect to the values $FB = 0.9$ & $DB = 0.9$ used by the methods without optimization only occur for the example FTC (D). For that example, the optimized value of DB has an intermediate value, assigning not too small sampling probabilities to regenerative cycles outside $\mathcal{S}_{k_{\min}}$, some of which have significant contributions to $E_P\{Z\}$.

Simulation under FTDB & BFTDB is in all the examples significantly more efficient than under FB & BFB, also for the examples in which the rarity parameter ε is not very small (examples FTC (B), FTC (C), and FTC (D)). For the examples core-

TABLE II
COMPARISON OF IMPORTANCE SAMPLING SCHEMES

	BFTDB_o		t_{sd}			
example	estimate	regenerative cycles CPU time (s)	REG BFTDB_n	FTDB_o FTDB_n	BFB_o BFB_n	FB_o FB_n
FTD (I)	1.814×10^{-8}	2,990,000	6,120	0.389	17.8	17.8
	$\pm 3.63 \times 10^{-11}$	112	0.944	0.311	20.3	19.3
FTD (II)	1.621×10^{-8}	3,999,000	4,248	0.242	12.5	6.99
	$\pm 3.24 \times 10^{-11}$	150	0.949	0.224	14.4	8.21
FTC (A)	2.694×10^{-10}	1,204,000	2.45×10^{5}	2.60	527	334
	$5.39 \pm \times 10^{-13}$	319	0.874	2.40	475	338
FTC (B)	2.695×10^{-8}	1,191,000	30,730	2.61	529	335
	$5.39 \pm \times 10^{-11}$	315	0.874	2.41	478	340
FTC (C)	1.969×10^{-8}	4,661,000	8,508	2.53	159	48.1
	$3.94 \pm \times 10^{-11}$	1,236	0.931	2.39	147	61.3
FTC (D)	1.107×10^{-7}	10,002,000	3,185	1.22	465	817
	$6.41 \pm \times 10^{-10}$	4,228	4.58	2.22	743	650

TABLE III
FINAL, OPTIMIZED VALUES OF THE PARAMETERS OF BFTDB IN THE BFTDB_O METHOD

example	FB	DB
FTD (I)	0.900	0.900
FTD (II)	0.900	0.900
FTC (A)	0.850	0.900
FTC (B)	0.850	0.900
FTC (C)	0.850	0.900
FTC (D)	0.900	0.618

sponding to balanced fault-tolerant systems, the explanation for this is that regenerative cycles in $\mathcal{S}_{k_{\min}}$ are rare within the set of regenerative cycles hitting D, and, therefore, those cycles are sampled with small probabilities in both FB & BFB. For all examples corresponding to unbalanced fault-tolerant systems, except example FTC (D), the explanation is that all regenerative cycles with important contributions to $E_P\{Z\}$ belong to $\mathcal{S}_{k_{\min}}$, and regenerative cycles in $\mathcal{S}_{k_{\min}}$ are rare within the set of regenerative cycles hitting D. For the FTC (D) example, the explanation is that, in spite of the fact that in that example some regenerative cycles outside $\mathcal{S}_{k_{\min}}$ have important contributions to $E_P\{Z\}$, those cycles are still sampled with significant probabilities in FTDB & BFTDB because $DB \leq 0.9$. The relative performance of FTDB & BFTDB varies from example to example. For the FTD examples, FTDB is more efficient than BFTDB, whereas, for the FTC examples, BFTDB is more efficient than FTDB. Because BFTDB has the BRE property for both balanced & unbalanced fault-tolerant systems, BFTDB should be preferred to FTDB. Furthermore, because the overhead due to the optimization of the biasing parameters is small, and for unbalanced fault-tolerant systems, it could be difficult to anticipate the optimal values of FB & DB, BFTDB should be used with optimization of the biasing parameters. The results for the FTC (D) example are illuminating in that sense. In that case, simulation under BFTDB with optimization of the biasing parameters is significantly more efficient than simulation under BFTDB without optimization of the biasing parameters.

C. Overheads due to Failure Transition Distances Computation

The algorithm for computation of failure transition distances proposed in Section III-D requires the knowledge of all minimal cuts of the fictitious fault-tolerant system. It has been shown in [34] that there does not exist any polynomial algorithm to obtain all minimal cuts of fault-tolerant systems defined by the generalized fault trees we consider. Because, when all failure bags have cardinality 1, the generalized fault tree of the fictitious fault-tolerant system is identical to the generalized fault tree of the given fault-tolerant system, there does not exist any polynomial algorithm to obtain all minimal cuts of the fictitious fault-tolerant system. This poses a theoretical limitation to the FTDB & BFTDB importance sampling schemes. Computing failure transition distances is NP-hard,[10] so, essentially, there is no better way to compute failure transition distances than the one we propose unless $\mathrm{P} = \mathrm{NP}$ [7]. Nevertheless, there currently exist algorithms [5] for computing all minimal cuts for fault-tolerant systems defined by the generalized fault trees we consider, which seem to be efficient unless the number of minimal cuts is very large. Thus, the algorithm described in [5] found the 512 minimal failure covers of the FTC examples in less than 0.5 seconds (the FTD examples have only 13 minimal failure covers, which were found in negligible time). Computation of failure transition distances introduces two CPU time overheads in the simulation. The first one is due to the construction of the data structures used by the algorithm for computation of failure distances. That overhead is essentially due to reading the minimal failure cover descriptions, building the hash table holding the selectors in the subsets $MFCS_q(c)$ & the hash table holding the selectors in the subsets $AMFCS_q(c)$, and building the lists linking the minimal failure covers & after minimal failure covers including some selector. For the FTC examples, that overhead was 0.05 seconds. The second overhead is

[10]It has been shown in [34] that computing failure distances is NP-hard, and that problem can be seen as a particular instance of the problem of computing failure transition distances when all failure bags have cardinality 1.

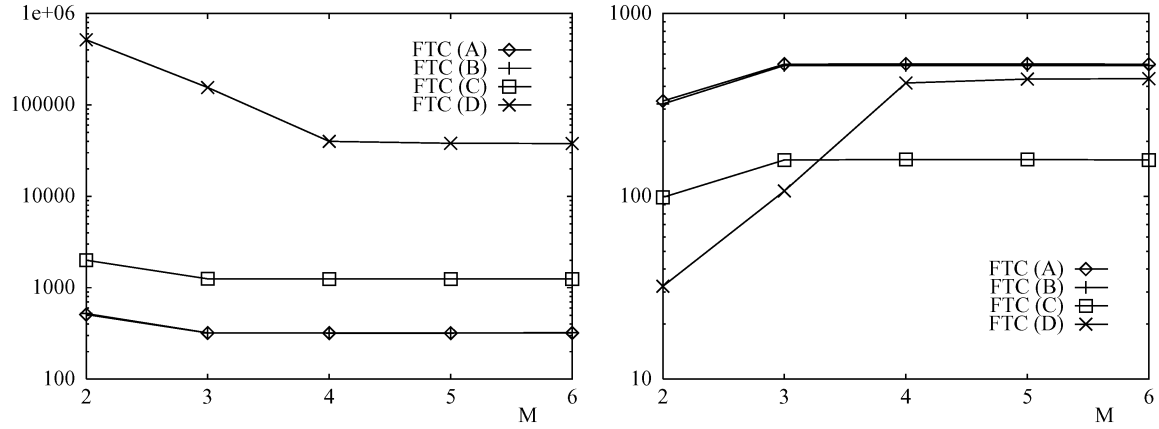

Fig. 8. CPU times in seconds required by BFTDB_o (left) & slow down factor of BFB_o with respect to BFTDB_o (right) as a function of M for the FTC examples.

TABLE IV
ANALYSIS OF THE CPU TIME OVERHEAD DUE TO COMPUTATION OF FAILURE DISTANCES

example	$\overline{T}$	total (%)	prop (%)
FTC (A)	0.550	3.0	0.3
FTC (B)	0.550	3.2	0.2
FTC (C)	0.550	3.2	0.3
FTC (D)	2.486	4.8	0.6

an increase in the simulation time per event due to actual computation of the failure transition distances. That overhead increases with the average number of minimal and after minimal failure covers "touched" (processed) per simulated transition of the biased stream, $\overline{T}$. To asses the importance of that second overhead, we profiled the execution of BFTDB_o for the examples FTC (A), FTC (B), FTC (C), and FTC (D). Table IV gives the obtained results. For each example, we give $\overline{T}$, the total percent CPU time overhead (total), and the portion of that CPU time overhead which is proportional to $\overline{T}$ (prop). We can see that $\overline{T}$ is extremely small, and, thus, the techniques used in the procedures for computing failure transition distances to reduce the number of minimal & after minimal failure cover "touches" seem to work very well. The CPU time overhead is small in all cases, and, more importantly, the part of that CPU time overhead which is proportional to $\overline{T}$, and, thus roughly proportional to the number of minimal failure covers, is very small, ranging from 0.2% to 0.6% in the worst case. This suggests that the CPU time overhead due to failure transition distance computations should remain reasonable even if the number of minimal failure covers of the system is much larger than 512, the number of minimal failure covers which the FTC examples have. Thus, we can conclude that BFTDB can still outperform significantly FB & BFB even if the number of minimal failure covers is of the order of tens of thousands. Memory consumption does not seem to be an issue. Thus, for the FTC examples, the memory overhead due to failure transition distance computation was about 0.9 MB; and, then, we can estimate a memory overhead of about 88 MB when the number of minimal failure covers is 50,000, which is affordable.

In some cases, the number of minimal failure covers can be reduced drastically by discarding the minimal failure covers of cardinality $> M$, where M is a moderate value $\geq RL$. Both FTDB & BFTDB can be easily adapted to that kind of partial knowledge by biasing failure transition probabilities as if the set of minimal failure covers of the given fault-tolerant system included the actual minimal failure covers of cardinality $\leq M$ & all bags with domain F_B of cardinality $M+1$ not including any of the former. Doing that will not affect the BRE properties of FTDB & BFTDB; neither will it affect the capability of FTDB & BFTDB to concentrate the simulation effort into $\mathcal{S}_{k_{\min}}$, and, therefore, should have a moderate impact on the performance of those importance sampling schemes for balanced fault-tolerant systems. The algorithms for computation of failure transition distances can easily be adapted to contemplate that change. Denoting by MFC the set of minimal failure covers of cardinality $\leq M$ & by $AMFC_f$ the corresponding sets of after minimal failure covers associated with failure bag f, the adaptation of the procedure $Compute\_td$ includes modifying the initial value of td to

$$td = \min\{RL, \max\{0, M+1-|b|\}\},$$

and exiting the procedure if that initial value is 0. The adaptation of the procedure $Compute\_atd$ includes modifying the initial values of $atd(f)$, $f \in \{f' \in S_F : AMFC_{f'} = \emptyset\}$ to

$$atd(f) = \min\{td, RL_f, \max\{0, M+1-|b|-|f|\}\}.$$

Fig. 8 (left) gives the CPU times required to achieve a target 99% confidence interval of $\pm 0.2\%$ as a function of M for M ranging from 2 to 6 for the BFTDB_o method, and the FTC examples. Fig. 8 (right) gives the slow down factor t_{sd} of BFB_o with respect to BFTDB_o as a function of M for M ranging from 2 to 6, also for the FTC examples. The examples have minimal failure covers of cardinalities 2, 3, 4, and 6. We can note that, even with $M = 2$, BFTDB_o is significantly faster than

BFB_o. For the examples FTC (A), FTC (B), and FTC (C), only a moderate improvement on the efficiency of BFTDB_o occurs beyond $M = 2$. For the example FTC (D), significant improvements on the efficiency of BFTDB_o occur until $M = 4$. That moderate improvements occur for the FTC (A) & FTC (B) examples beyond $M = 2$ can be explained by the fact that those examples correspond to balanced fault-tolerant systems, and, the minimum minimal failure cover cardinality being equal to 2, the selection $M = 2$ is enough to concentrate the simulation effort into the regenerative cycles in $\mathcal{S}_{k_{\min}}$. That moderate improvements occur for the FTC (C) example beyond $M = 2$ can be explained similarly, noting that all regenerative cycles with important contributions in that example belong to $\mathcal{S}_{k_{\min}}$. As previously commented, the FTC (D) example has regenerative cycles outside $\mathcal{S}_{k_{\min}}$ with significant contributions to $E\{Z\}$, and higher values of M are required to sample those regenerative cycles with significant probabilities.

Discarding minimal failure covers of cardinality $> M$ should allow us to deal with very large systems when M is small. Thus, for instance, for $M = 2$, the number of minimal cuts of cardinality $\leq M$ of a system having NC component classes with an instance of each class is roughly bounded from above by $NC^2/2$, and a budget of 50,000 minimal cuts would allow us to deal with systems with at least 316 components, and often many more. For $M = 3$, the number of minimal cuts with cardinality $\leq M$ is roughly bounded from above by $NC^3/6$, and a budget of 50,000 minimal cuts would allow us to deal with systems with at least 66 components, and often many more.

D. Related Work

Some of the balanced likelihood ratio techniques developed by Alexopoulos & Shultes [1], [2] exploit the "shortest path to failure concept" in a way similar to the way that FTDB & BFTDB do. In those techniques, it is assumed that the structure function of the system is specified by a graph in which each link has associated with it a given set of components, sets associated with different links being disjoint. The capacity of a link is defined as the number of operating components from those associated with the link, and the system is "up" if a flow of k units can be transmitted from a source node s to a terminal node t, i.e. if the capacity of the graph is $\geq k$. The "shortest path to failure concept" is exploited by identifying the set of links in "minimum cuts", a minimum cut being an $s - t$ cut of capacity equal to the capacity of the graph in a given state. An important advantage of that approach is that efficient *polynomial* algorithms exist for maintaining that set of links as the state of the system changes [31]. However, whereas every increasing structure function can be specified by a fault tree with only AND & OR gates, some increasing functions cannot be specified by a graph (a simple example is the function $c_1c_2 + c_1c_3 + c_1c_4 + c_2c_3c_4$[11]), so our framework is more general. Minimum cuts could be considered for systems made up of a set of components whose increasing structure functions are represented by ordinary fault trees including only AND & OR gates, a minimum cut being a minimal cut of minimum cardinality, and the techniques developed in [1], [2] based on the "shortest path to failure" concept could be adapted to that framework, which is a particularization of our framework. However, determining the components in minimum cuts is, in that context, NP-hard as is computing failure transition distances.[12]

[11]That such function cannot be specified by a graph can be easily checked by noting that the graph will have at most 4 links (because 4 components are involved), and building all graphs with at most 4 links & disjoint subsets associated with them with union $\{c_1, c_2, c_3, c_4\}$.

APPENDIX

In some of the proofs which follow, we will use the parameter $F_{\max} = \max_{f \in F_B} |f|$. Informally, $F_{\max}$ is the maximum number of components which can fail simultaneously. For most fault-tolerant systems $F_{\max}$ has a moderate value. The following lemma gives an upper bound on the length of the regenerative cycles in $\mathcal{S}_k$ in terms of $F_{\max}$.

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

Proof: Consider the transitions $(s_i, s_{i+1}), 0 \leq i < l$ of the regenerative cycle $\omega = (s_0, s_1, \ldots, s_l)$. Exactly $k + 1$ of those transitions are failure transitions (the transition from $s_0 = r$ & k of the remaining transitions, which are from states $x \in \Omega'$) & exactly $l - k - 1$ of those transitions are repair transitions. Let f be the sum of the cardinalities of the failure bags associated with the failure transitions & let p be the sum of the cardinalities of the bags of components repaired in the repair transitions. Obviously, $f = p$. Also, $f \leq (k + 1)F_{\max}$ & $p \geq l - k - 1$. Then, $l - k - 1 \leq (k + 1)F_{\max}$, implying the result. □

The path to the proof of Theorem 1 starts with the following result:

Theorem 5: 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

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

Note that $P_{s_0,s_1} = P_{r,s_1} = \Theta(1)$. Of the remaining $l-1$ factors, k factors correspond to failure transitions from states $x \in \Omega'$,

[12]Computing failure distances continues to be NP-hard for systems made up of a set of components whose increasing structure functions are represented by ordinary fault trees including only AND & OR gates [34]. Then, to show that the problem under consideration is NP-hard it is enough to transform polynomially into that problem the problem of computing failure distances for systems made up of a set of components whose increasing structure functions are represented by ordinary fault trees including only AND & OR gates. Such a transformation can be constructed as follows. Let ϕ be the given system with set of components C & fault tree $\mathcal{F}$ (including only AND & OR gates), and let F be the set of failed components in the state under consideration. First, consider the system ϕ_0 with set of components $C_0 = C - F$ & fault tree $\mathcal{F}_0$ representing ϕ conditioned to the components in F being failed ($\mathcal{F}_0$ can be obtained in polynomial time, and is not the trivial fault tree without inputs & output equal to 1, because ϕ is not failed when the components in F are failed), and determine for ϕ_0 the subset S_0 of components in minimum cuts. Then, pick up a component $c_0 \in S_0$, and consider the system ϕ_1 with set of components $C_1 = C_0 - \{c_0\}$ & fault tree $\mathcal{F}_1$ representing ϕ_0 conditioned to component c_0 being failed ($\mathcal{F}_0$ can be obtained in polynomial time). If $\mathcal{F}_1$ is the trivial fault tree with output equal to 1, the failure distance is 1. If not, determine for ϕ_1 the subset S_1 of components in minimum cuts. Pick up a component $c_1 \in S_1$, and consider the system ϕ_2 with set of components $C_2 = C_0 - \{c_1\}$ & fault tree $\mathcal{F}_2$ representing ϕ_1 conditioned to component c_1 being failed ($\mathcal{F}_2$ can be obtained in polynomial time). If $\mathcal{F}_2$ is the trivial fault tree with output equal to 1, the failure distance is 2, etc.

and, therefore, are $\Theta(\varepsilon)$ & the remaining $l-k-1$ factors correspond to repair transitions from states $x \in \Omega'$, and, therefore, are $\Theta(1)$. All together, this implies $P\{\omega\} = \Theta(\varepsilon^k)$. On the other hand,

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

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

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

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

Let $k_{\min} = \min\{k : \mathcal{S}_k \neq \emptyset\}$. Using Theorem 5, we have the following corollary.

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

According to Theorem 5, 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 6: For balanced fault-tolerant systems, $\sum_{k=k_{\min}+1}^{\infty} C(k) = o(\varepsilon^{k_{\min}})$.

The proof of Theorem 6 will be preceded by two propositions. For k such that $\mathcal{S}_k \neq \emptyset$, let

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

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

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

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

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

$s_0 = r \notin D$. Therefore, $Z(\omega)$ is the sum of a number of h_x, $x \in \Omega'$ no greater than $l-1$. Each h_x is upper bounded by $1/r_{\min}$. According to Lemma 1, $l-1 \leq (k+1)(F_{\max}+1)-1$. Then,

$$Z(\omega) \leq \frac{(k+1)(F_{\max}+1)-1}{r_{\min}}.$$

□

Let $\widehat{\mathbf{F}} = (1_{T_F}((x,y)) f_{x,y})_{x,y \in \Omega'}$. The upper bound for $P\{\mathcal{S}_k\}$ is in terms of $F_{\max}$, $\|\widehat{\mathbf{F}}\|_\infty$ & ε.

Proposition 2: For balanced fault-tolerant systems & $k \geq k_{\min}$,

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

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

$$\widetilde{\mathbf{R}} = (1_{T_R}((x,y)) P(x,y))_{x,y \in \Omega'}$$

collects repair transition probabilities &

$$\widetilde{\mathbf{\Lambda}} = (1_{T_F}((x,y)) P(x,y))_{x,y \in \Omega'}$$

collects failure transition probabilities. 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 $(P_{r,x})_{x \in \Omega'}$ & let $\mathbf{v}$ be the column vector $(P_{x,r})_{x \in \Omega'}$. Consider $\widetilde{\mathbf{P}}^n = (\widetilde{\mathbf{R}} + \widetilde{\mathbf{\Lambda}})^n$ & 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 $\widetilde{\mathbf{\Lambda}}$ & $n-k$ times $\widetilde{\mathbf{R}}$. According to Lemma 1, regenerative cycles $w = (s_0, s_1, \ldots, s_l)$ in $\mathcal{S}_k$ include at most $(k+1)(F_{\max}+1)$ 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

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

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

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

□

Proof of Theorem 6: We start from

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

Using Propositions 1 & 2,

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

with

$$A = \frac{(F_{\max}+1)2^{F_{\max}-1}}{r_{\min}}$$

&

$$B = 2^{F_{\max}+1}\|\widetilde{\mathbf{F}}\|_\infty.$$

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

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

which follows easily from (see, for instance, [32]) $\sum_{k=0}^{\infty} ka^k = a/(1-a)^2$, we have, for $\varepsilon \to 0$,

$$\begin{aligned}\sum_{k=k_{\min}+1}^{\infty} C(k) &< A(B\varepsilon)^{k_{\min}+1}\times \\ &\left(\frac{k_{\min}+1}{1-B\varepsilon} + \frac{B\varepsilon}{(1-B\varepsilon)^2} + \frac{1}{1-B\varepsilon}\right) = o(\varepsilon^{k_{\min}}),\end{aligned}$$

□

An immediate consequence of Corollary 1 & Theorem 6 is $E_P\{Z\} = \Theta(\varepsilon^{k_{\min}})$, which is in accordance with the results obtained in [30]. By combining Theorem 5, Corollary 1, and Theorem 6 we are finally able to prove Theorem 1.

Proof of Theorem 6: According to Corollary 1, $C(k_{\min}) = a\varepsilon^{k_{\min}} + o(\varepsilon^{k_{\min}})$, $a > 0$. Using Theorem 6,

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

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 5, $P\{\omega\}Z(\omega) = a_\omega\varepsilon^{k_{\min}} + o(\varepsilon^{k_{\min}})$, $a_\omega > 0$, $\omega \in \mathcal{S}_{k_{\min}}$. 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 6,

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

proving c). □

The proof of Theorem 2 is preceded by the following Proposition, which relates $\Phi'(b)$, $b \subseteq C'$ with $\Phi(b)$, $b \subseteq C$.

Proposition 3:

$$\Phi'\left(C' - f_1[n_1]\cdots f_k[n_k]\right) = \Phi\left(C - \sum_{i=1}^{k}\sum_{j=1}^{n_i} f_i\right).$$

Proof: $\Phi'(C' - f_1[n_1]\cdots f_k[n_k]) = 0$ iff implication at 1 of the input atoms of the generalized fault tree of the fictitious fault-tolerant system included in $f_1[n_1]\cdots f_k[n_k]$ implies the output of the generalized fault tree at 1. $\Phi(C - \sum_{i=1}^{k}\sum_{j=1}^{n_i} f_j) = 0$ iff implication at 1 of the input atoms of the generalized fault tree of the given fault-tolerant system included in $\sum_{i=1}^{k}\sum_{j=1}^{n_i} f_i$ implies the output of the generalized fault tree at 1. The result follows if the input nodes of the generalized fault tree of the given fault-tolerant system are implied at the same values as the corresponding nodes of the generalized fault tree of the fictitious fault-tolerant system. To prove that, consider, for instance, the node labeled $c[3]$ of the example of Fig. 1. Assume that $c[3]$ is implied at 1 in the generalized fault tree of the fictitious fault-tolerant system. It follows that $f_1[n_1]\cdots f_k[n_k] \supseteq f_1'[3]$, or $f_1[n_1]\cdots f_k[n_k] \supseteq f_1'[1]f_2'[1]$, or $f_1[n_1]\cdots f_k[n_k] \supseteq f_2'[2]$, or $f_1[n_1]\cdots f_k[n_k] \supseteq f_3'[1]$, or $f_1[n_1]\cdots f_k[n_k] \supseteq f_4'[1]$. Then, $\sum_{i=1}^{k}\sum_{j=1}^{n_i} f_i \supseteq f_1' + f_1' + f_1'$, or $\sum_{i=1}^{k}\sum_{j=1}^{n_i} f_i \supseteq f_1' + f_2'$, or $\sum_{i=1}^{k}\sum_{j=1}^{n_i} f_i \supseteq f_2' + f_2'$, or $\sum_{i=1}^{k}\sum_{j=1}^{n_i} f_i \supseteq f_3'$, or $\sum_{i=1}^{k}\sum_{j=1}^{n_i} f_i \supseteq f_4'$. Because $f_1' + f_1' + f_1' \supseteq c[3]$, $f_1' + f_2' \supseteq c[3]$, $f_2' + f_2' \supseteq c[3]$, $f_3' \supseteq c[3]$, and $f_4' \supseteq c[3]$, it

follows that $\sum_{i=1}^{k}\sum_{j=1}^{n_i} f_i \supseteq c[3]$, and that the input $c[3]$ is implied at 1 in the generalized fault tree of the given fault-tolerant system. Assume that the node $c[3]$ is implied at 0 in the generalized fault tree of the fictitious fault-tolerant system, and let $b = f_1[n_1]\cdots f_k[n_k]$. It follows that $\#(f_1', b) < 3$, either $\#(f_1', b) < 1$ or $\#(f_2', b) < 1$, $\#(f_2', b) < 2$, $\#(f_3', b) < 1$, and $\#(f_4', b) < 1$. Because $f_1'[3]$, $f_1'[1]f_2'[1]$, $f_2'[2]$, $f_3'[1]$, and $f_4'[1]$ are the minimal bags with domain in F_B^c "covering" $c[3]$, it follows that $\sum_{i=1}^{k}\sum_{j=1}^{n_i} f_i \not\supseteq c[3]$, and that the input $c[3]$ is implied at 0 in the generalized fault tree of the given fault-tolerant system. □

Proof of Theorem 2: From Proposition 3 & the definition of failure cover, $f_1[n_1]\cdots f_k[n_k] \subseteq C'$ will be a failure cover iff it is a cut of the fictitious fault-tolerant system. Therefore, the minimal failure covers $\subseteq C'$ will be the minimal cuts of the fictitious fault-tolerant system. But C' is a failure cover because $\Phi'(\emptyset) = 0$, and, according to Proposition 3, $\Phi(C - \sum_{f\in F_B}\sum_{i=1}^{\#(f,C')} f) = 0$ (it could happen that $\sum_{f\in F_B}\sum_{i=1}^{\#(f,C')} f \supset C$). This implies that all minimal failure covers are $\subseteq C'$, and that the minimal failure covers are the minimal cuts of the fictitious fault-tolerant system. By construction, the fictitious fault-tolerant system has cuts (C' is one). Then, being $\Phi'(b)$, $b \subseteq C'$ increasing, it has minimal cuts, and $MFC \neq \emptyset$. □

The following lemma collects results on bags which will be used in the proofs of Theorems 3 & 4.

Lemma 2: Let a, b, and c be bags with domain $\mathcal{E}$. Then,

a) $a + b \cap c = (a + b) \cap (a + c)$,
b) $(a \cap c + b) \cap c = (a + b) \cap c$,
c) $a = b \cap c$ implies $c - a = c - b$,
d) $a - b = c$ implies $a \subseteq b + c$,
e) $a - b - c = a - (b + c)$, and
f) $a \subseteq b$ implies $a + (b - a) = b$.

Proof:

a) Let $d = a + b \cap c$ & let $e = (a + b) \cap (a + c)$. Using the definitions of bag sum & bag intersection, for $x \in \mathcal{E}$,

$$\begin{aligned}\#(x,d) &= \#(x,a) + \min\{\#(x,b), \#(x,c)\}\\ &= \min\{\#(x,a) + \#(x,b), \#(x,a) + \#(x,c)\}\\ &= \#(x,e).\end{aligned}$$

b) Let $d = (a \cap c + b) \cap c$ & let $e = (a + b) \cap c$. Using the definitions of bag sum & bag intersection, for $x \in \mathcal{E}$,

$$\begin{aligned}\#(x,d) &= \min\{\min\{\#(x,a), \#(x,c)\} + \#(x,b), \#(x,c)\}\\ &= \min\{\#(x,a) + \#(x,b), \#(x,c)\} = \#(x,e).\end{aligned}$$

c) Using the definitions of bag difference & bag intersection, for $x \in \mathcal{E}$,

$$\begin{aligned}\#(x, c-a) &= \max\{0, \#(x,c) - \#(x,a)\}\\ &= \max\{0, \#(x,c) - \min\{\#(x,b), \#(x,c)\}\}\end{aligned} \tag{16}$$

&

$$\#(x, c-b) = \max\{0, \#(x,c) - \#(x,b)\}. \tag{17}$$

For $x \in \mathcal{E}$ such that $\#(x,c) \leq \#(x,b)$, (16) & (17) yield $\#(x, c-a) = 0$ & $\#(x, c-b) = 0$; for $x \in \mathcal{E}$ such that $\#(x,c) > \#(x,b)$, (16) & (17) yield $\#(x, c-a) = \#(x,c) - \#(x,b)$ & $\#(x, c-b) = \#(x,c) - \#(x,b)$.

d) Using the definition of bag difference, for $x \in \mathcal{E}$,

$$\#(x,c) = \max\{0, \#(x,a) - \#(x,b)\} \geq \#(x,a) - \#(x,b),$$

implying $\#(x,a) \leq \#(x,b) + \#(x,c)$, $x \in \mathcal{E}$, and $a \subseteq b + c$.

e) Let $d = a - b - c$ & let $e = a - (b + c)$. Using the definitions of bag difference & bag sum, for $x \in \mathcal{E}$,

$$\begin{aligned}\#(x,d) = \max\{0, \max\{0, \#(x,a) - \#(x,b)\}\\ -\#(x,c)\},\end{aligned} \tag{18}$$

$$\#(x,e) = \max\{0, \#(x,a) - \#(x,b) - \#(x,c)\}. \tag{19}$$

For $x \in \mathcal{E}$ such that $\#(x,a) - \#(x,b) - \#(x,c) \geq 0$, (18) & (19) yield $\#(x,d) = \#(x,a) - \#(x,b) - \#(x,c)$ & $\#(x,e) = \#(x,a) - \#(x,b) - \#(x,c)$. For $x \in \mathcal{E}$ such that $\#(x,a) - \#(x,b) - \#(x,c) < 0$ & $\#(x,a) - \#(x,b) \geq 0$, (18) & (19) yield $\#(x,d) = 0$ & $\#(x,e) = 0$. For $x \in \mathcal{E}$ such that $\#(x,a) - \#(x,b) < 0$, (18) & (19) yield $\#(x,d) = 0$ & $\#(x,e) = 0$.

f) Using $\#(x,a) \leq \#(x,b)$, $x \in \mathcal{E}$ & the definitions of bag sum & bag difference,

$$\begin{aligned}\#(x, a + (b-a)) &= \#(x,a) + \max\{0, \#(x,b) - \#(x,a)\}\\ &= \#(x,a) + \#(x,b) - \#(x,a) = \#(x,b).\end{aligned}$$

□

Proof of Theorem 3: Let $d = \min_{b\in MFC}|b - B(x)| > 0$, because otherwise $B(x) \supseteq b$ for some $b \in MFC$, $\Phi(C - F(x)) = \Phi(C - \sum_{c[1]\in B(x)}\sum_{i=1}^{\#(c[1],B(x))} c[1]) \leq \Phi(C - \sum_{f\in F_B}\sum_{i=1}^{\#(f,b)} f) = 0$, and $x \in D$. We start by showing $td(x) \leq d$. This is done by constructing a failure path from x of length $\leq d$. Let b' be any minimal failure cover such that

$$|b' - B(x)| = \min_{b\in MFC}|b - B(x)| = d$$

& let $b' - B(x) = f_1[n_1]f_2[n_2]\cdots f_k[n_k]$, $k \geq 1$. We have $\sum_{i=1}^{k} n_i = d$. Let the failure bags

$$\begin{aligned}f_1^1 &= f_1 \cap (C - F(x)),\\ f_1^2 &= f_1 \cap (C - F(x) - f_1^1),\\ &\vdots\\ f_1^{n_1} &= f_1 \cap \left(C - F(x) - \sum_{i=1}^{n_1-1} f_1^i\right),\\ f_2^1 &= f_2 \cap \left(C - F(x) - \sum_{i=1}^{n_1} f_1^i\right),\\ f_2^2 &= f_2 \cap \left(C - F(x) - \sum_{i=1}^{n_1} f_1^i - f_2^1\right),\\ &\vdots\end{aligned}$$

$$
\begin{aligned}
f_2^{n_2} &= f_2 \cap \left(C - F(x) - \sum_{i=1}^{n_1} f_1^i - \sum_{i=1}^{n_2-1} f_2^i \right), \\
&\vdots \\
f_k^1 &= f_k \cap \left(C - F(x) - \sum_{l=1}^{k-1} \sum_{i=1}^{n_l} f_l^i \right), \\
f_k^2 &= f_k \cap \left(C - F(x) - \sum_{l=1}^{k-1} \sum_{i=1}^{n_l} f_l^i - f_k^1 \right), \\
&\vdots \\
f_k^{n_k} &= f_k \cap \left(C - F(x) - \sum_{l=1}^{k-1} \sum_{i=1}^{n_l} f_l^i - \sum_{i=1}^{n_k-1} f_k^i \right).
\end{aligned}
$$

Some of the failure bags f_i^j could be empty. Note that, by construction, $F(x) + \sum_{i=1}^{m} \sum_{j=1}^{n_i} f_i^j + \sum_{j=1}^{l} f_{m+1}^j \subseteq C$, for $0 \leq m < k$, $0 \leq l \leq n_1$ if $m = 0$, and $1 \leq l \leq n_{m+1}$ if $0 < m < k$. Consider the non-empty failure bags in that set, f_i^j, $1 \leq i \leq k$, $1 \leq j \leq n_i'$. Note that $n_i' \leq n_i$, $1 \leq i \leq k$, & that some n_i', $1 \leq i \leq k$ could be 0. By construction, $f_i^j \subseteq f_i$, and, using Lemma 2, e), $f_i^j \subseteq C - F(x) - \sum_{l=1}^{i-1} \sum_{m=1}^{n_l'} f_l^m - \sum_{m=1}^{j-1} f_i^m = C - (F(x) + \sum_{l=1}^{i-1} \sum_{m=1}^{n_l'} f_l^m + \sum_{m=1}^{j-1} f_i^m)$. Then, by Assumptions A3 & A4, some path $(x, y_1^1, y_1^2, \ldots, y_1^{n_1'}, y_2^1, y_2^2, \ldots, y_2^{n_2'}, \ldots, y_k^1, y_k^2, \ldots, y_k^{n_k'})$ built up with failure transitions having associated with them, in that order, the failure bags $f_1^1, f_1^2, \ldots, f_1^{n_1'}, f_2^1, f_2^2, \ldots, f_2^{n_2'}, \ldots, f_k^1, f_k^2, \ldots, f_k^{n_k'}$ exist. Because $\sum_{i=1}^{k} n_i' \leq d$, it is enough to prove that $y_k^{n_k'}$ is a down state. We will start by proving

$$
F(x) + \sum_{i=1}^{k} \sum_{j=1}^{n_i} f_i^j = \left(F(x) + \sum_{i=1}^{k} \sum_{j=1}^{n_i} f_i \right) \cap C. \tag{20}
$$

The proof is by induction. The base case is $F(x) + f_1^1 = (F(x) + f_1) \cap C$. Using the definition of f_1^1, Lemma 2, a) with $a = F(x), b = f_1$, and $c = C - F(x)$, and Lemma 2, f), taking into account $F(x) \subseteq C$,

$$
\begin{aligned}
F(x) + f_1^1 &= F(x) + f_1 \cap (C - F(x)) \\
&= (F(x) + f_1) \cap (F(x) + (C - F(x))) \\
&= (F(x) + f_1) \cap C.
\end{aligned}
$$

For the induction step, the induction hypothesis is $F(x) + \sum_{i=1}^{m} \sum_{j=1}^{n_i} f_i^j + \sum_{j=1}^{l} f_{m+1}^j = (F(x) + \sum_{i=1}^{m} \sum_{j=1}^{n_i} f_i + \sum_{j=1}^{l} f_{m+1}) \cap C$, $0 \leq m < k$, $1 \leq l < n_1$ if $m = 0$, and $0 \leq l < n_{m+1}$ if $0 < m < k$; and it has to be proved that $F(x) + \sum_{i=1}^{m} \sum_{j=1}^{n_i} f_i^j + \sum_{j=1}^{l+1} f_{m+1}^j = (F(x) + \sum_{i=1}^{m} \sum_{j=1}^{n_i} f_i + \sum_{j=1}^{l+1} f_{m+1}) \cap C$. Using, in that order, the definition of f_{m+1}^{l+1}; Lemma 2, a) with $a = F(x) + \sum_{i=1}^{m} \sum_{j=1}^{n_i} f_i^j + \sum_{j=1}^{l} f_{m+1}^j$, $b = f_{m+1}$, and $c = C - F(x) - \sum_{i=1}^{m} \sum_{j=1}^{n_i} f_i^j - \sum_{j=1}^{l} f_{m+1}^j$; Lemma 2, e); Lemma 2, f), taking into account $F(x) + \sum_{i=1}^{m} \sum_{j=1}^{n_i} f_i^j + \sum_{j=1}^{l} f_{m+1}^j \subseteq C$; the induction Hypothesis; and Lemma 2, b) with $a = F(x) + \sum_{i=1}^{m} \sum_{j=1}^{n_i} f_i + \sum_{j=1}^{l} f_{m+1}$, $b = f_{m+1}$, and $c = C$,

$$
\begin{aligned}
&F(x) + \sum_{i=1}^{m} \sum_{j=1}^{n_i} f_i^j + \sum_{j=1}^{l+1} f_{m+1}^j \\
&\quad = F(x) + \sum_{i=1}^{m} \sum_{j=1}^{n_i} f_i^j + \sum_{j=1}^{l} f_{m+1}^j + f_{m+1} \\
&\qquad \cap \left(C - F(x) - \sum_{i=1}^{m} \sum_{j=1}^{n_i} f_i^j - \sum_{j=1}^{l} f_{m+1}^j \right) \\
&\quad = \left(F(x) + \sum_{i=1}^{m} \sum_{j=1}^{n_i} f_i^j + \sum_{j=1}^{l} f_{m+1}^j + f_{m+1} \right) \\
&\qquad \cap \left(F(x) + \sum_{i=1}^{m} \sum_{j=1}^{n_i} f_i^j + \sum_{j=1}^{l} f_{m+1}^j \right. \\
&\qquad\quad \left. + \left(C - F(x) - \sum_{i=1}^{m} \sum_{j=1}^{n_i} f_i^j - \sum_{j=1}^{l} f_{m+1}^j \right) \right) \\
&\quad = \left(F(x) + \sum_{i=1}^{m} \sum_{j=1}^{n_i} f_i^j + \sum_{j=1}^{l} f_{m+1}^j + f_{m+1} \right) \\
&\qquad \cap \left(F(x) + \sum_{i=1}^{m} \sum_{j=1}^{n_i} f_i^j + \sum_{j=1}^{l} f_{m+1}^j \right. \\
&\qquad\quad \left. + C - \left(F(x) + \sum_{i=1}^{m} \sum_{j=1}^{n_i} f_i^j + \sum_{j=1}^{l} f_{m+1}^j \right) \right) \\
&\quad = \left(F(x) + \sum_{i=1}^{m} \sum_{j=1}^{n_i} f_i^j + \sum_{j=1}^{l} f_{m+1}^j + f_{m+1} \right) \cap C \\
&\quad = \left(\left(F(x) + \sum_{i=1}^{m} \sum_{j=1}^{n_i} f_i + \sum_{j=1}^{l} f_{m+1} \right) \cap C + f_{m+1} \right) \cap C \\
&\quad = \left(F(x) + \sum_{i=1}^{m} \sum_{j=1}^{n_i} f_i + \sum_{j=1}^{l+1} f_{m+1} \right) \cap C,
\end{aligned}
$$

completing the induction step. Using (20), we have $F(x) + \sum_{i=1}^{k} \sum_{j=1}^{n_i'} f_i^j = F(x) + \sum_{i=1}^{k} \sum_{j=1}^{n_i} f_i^j = (F(x) + \sum_{i=1}^{k} \sum_{j=1}^{n_i} f_i) \cap C$, which, using Lemma 2, c) with $a = F(x) + \sum_{i=1}^{k} \sum_{j=1}^{n_i'} f_i^j, b = F(x) + \sum_{i=1}^{k} \sum_{j=1}^{n_i} f_i$ & $c = C$ & Lemma 2, e), implies $C - F(x) - \sum_{i=1}^{k} \sum_{j=1}^{n_i'} f_i^j = C - F(x) - \sum_{i=1}^{k} \sum_{j=1}^{n_i} f_i$. But, $F(x) = \sum_{f \in B(x)} \sum_{i=1}^{\#(f,B(x))} f$. Also, by Lemma 2, d) with $a = b'$, $b = B(x)$, and $c = f_1[n_1] f_2[n_2] \cdots f_k[n_k]$, we have $b' \subseteq B(x) + f_1[n_1] f_2[n_2] \cdots f_k[n_k]$. Then, $C - F(x) - \sum_{i=1}^{k} \sum_{j=1}^{n_i'} f_i^j = C - F(x) - \sum_{i=1}^{k} \sum_{j=1}^{n_i} f_i = C - \sum_{f \in B(x)} \sum_{i=1}^{\#(f,B(x))} f - \sum_{i=1}^{k} \sum_{j=1}^{n_i} f_i \subseteq C - \sum_{f \in b'} \sum_{i=1}^{\#(f,b')} f$. b' being a failure cover & Φ being increasing, $\Phi(C - F(x) - \sum_{i=1}^{k} \sum_{j=1}^{n_i'} f_i^j) \leq \Phi(C - \sum_{f \in b'} \sum_{i=1}^{\#(f,b')} f) = 0$, i.e. $\Phi(C - F(x) - \sum_{i=1}^{k} \sum_{j=1}^{n_i'} f_i^j) = 0$, and, because the bag of

failed component classes in $y_k^{n'_k}$ is $F(x)+\sum_{i=1}^{k}\sum_{j=1}^{n'_i} f_i^j$, this implies that $y_k^{n'_k}$ is a down state.

We show next that $td(x) \geq d$. Assume there exists a failure path from x built up with failure transitions having associated with them failure bags $f'_1, f'_2, \ldots, f'_m$, $m < d$. Let b'' be the bag with domain F_B defined by $\#(f, b'') = |\{f'_i, 1 \leq i \leq m : f'_i = f\}|$. Because the final state of the path is a down state, $b''' = B(x)+b''$ is a failure cover. But $|b'''-B(x)| = |b''| = m$, which implies

$$\begin{aligned}\min_{b\in MFC}|b-B(x)| &= \min_{\text{all failure covers } b}|b-B(x)| \leq m < d\\ &= \min_{b\in MFC}|b-B(x)|,\end{aligned}$$

a contradiction. □

Proof of Theorem 4: Using (6) & $b-(B(x)+ex(f)) = b-ex(f)-B(x)$, which follows from Lemma 2, e), we have

$$atd(x,f) = \min_{b\in MFC}|b-ex(f)-B(x)|. \tag{21}$$

For the case $AMFC_f = \emptyset$, $b \cap ex(f) = \emptyset$ for all $b \in MFC$, $b - ex(f) = b$ for all $b \in MFC$, and, using Theorem 3,

$$atd(x,f) = \min_{b\in MFC}|b-B(x)| = td(x).$$

It remains to consider the case $AMFC_f \neq \emptyset$. In that case, $b \cap ex(f) \neq \emptyset$ for some $b \in MFC$. We will start by proving

$$atd(x,f) \geq \min\left\{td(x), \min_{b\in AMFC_f}|b-B(x)|\right\}$$

&

$$atd(x,f) \leq \min_{b\in AMFC_f}|b-B(x)|.$$

Two subcases will be considered: a) $b \cap ex(f) = \emptyset$ for some $b \in MFC$, b) $b \cap ex(f) \neq \emptyset$ for all $b \in MFC$. In subcase a), from (21),

$$\begin{aligned}atd(x,f) &= \min\left\{\min_{b\in MFC: b\cap ex(f)=\emptyset}|b-ex(f)-B(x)|,\right.\\ &\qquad\left.\min_{b\in MFC: b\cap ex(f)\neq\emptyset}|b-ex(f)-B(x)|\right\}\\ &= \min\left\{\min_{b\in MFC: b\cap ex(f)=\emptyset}|b-B(x)|,\right.\\ &\qquad\left.\min_{b\in AMFC_f}|b-B(x)|\right\},\end{aligned} \tag{22}$$

and it follows using Theorem 3 that

$$\begin{aligned}atd(x,f) &\geq \min\left\{\min_{b\in MFC}|b-B(x)|, \min_{b\in AMFC_f}|b-B(x)|\right\}\\ &= \min\left\{td(x), \min_{b\in AMFC_f}|b-B(x)|\right\}.\end{aligned}$$

Also, from (22),

$$atd(x,f) \leq \min_{b\in AMFC_f}|b-B(x)|.$$

In subcase b), from (21),

$$\begin{aligned}atd(x,f) &= \min_{b\in MFC: b\cap ex(f)\neq\emptyset}|b-ex(f)-B(x)|\\ &= \min_{b\in AMFC_f}|b-B(x)|,\end{aligned}$$

and it follows that

$$atd(x,f) \geq \min\left\{td(x), \min_{b\in AMFC_f}|b-B(x)|\right\}$$

&

$$atd(x,f) \leq \min_{b\in AMFC_f}|b-B(x)|.$$

To complete the case $AMFC_f \neq \emptyset$, it remains to prove $atd(x,f) \leq td(x)$. Using (6) & Theorem 3,

$$\begin{aligned}atd(x,f) &= \min_{b\in MFC}|b-(B(x)+ex(f))|\\ &\leq \min_{b\in MFC}|b-B(x)| = td(x).\end{aligned}$$

□

Lemma 3: Let $g(\mathbf{x})$ be a convex function in a convex set S & let $f(x)$ be an increasing convex function. Then, $f(g(\mathbf{x}))$ is convex in S.

Proof: Let $\mathbf{x}_1, \mathbf{x}_2 \in S$ & let λ, $0 \leq \lambda \leq 1$. Because $g(\mathbf{x})$ is convex in S,

$$g(\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2) \leq \lambda g(\mathbf{x}_1) + (1-\lambda)g(\mathbf{x}_2).$$

Because $f(x)$ is increasing,

$$f(g(\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2)) \leq f(\lambda g(\mathbf{x}_1) + (1-\lambda)g(\mathbf{x}_2)) \tag{23}$$

and, because $f(x)$ is convex,

$$f(\lambda g(\mathbf{x}_1) + (1-\lambda)g(\mathbf{x}_2)) \leq \lambda f(g(\mathbf{x}_1)) + (1-\lambda) f(g(\mathbf{x}_2)),$$

which with (23) implies

$$f(g(\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2)) \leq \lambda f(g(\mathbf{x}_1)) + (1-\lambda) f(g(\mathbf{x}_2)).$$

□

Theorem 7: The functions

$$f(x_1, x_2, \ldots, x_n) = \sum_{l=1}^{p} B_l \prod_{k=1}^{n} \frac{1}{x_k^{n_k(l)}(1-x_k)^{n'_k(l)}},$$

with $B_l > 0$ & $n_k(l)$, $n'_k(l)$ integers ≥ 0 are convex in $[a, 1-a]^n$, $0 < a < 0.5$.

Proof: The sum of convex functions is a convex function & the product of a positive constant by a convex function is also a convex function. Therefore, it suffices to prove that the functions

$$g(x_1, x_2, \ldots, x_n) = \prod_{k=1}^{n} \frac{1}{x_k^{n_k}(1-x_k)^{n'_k}}$$

with n_k, n'_k integers ≥ 0 are convex in $[a, 1-a]^n$. Because e^x is increasing & convex, according to Lemma 3, it suffices to prove that the function $h(x_1, x_2, \ldots, x_n) = \log(g(x_1, x_2, \ldots, x_n))$ is convex in $[a, 1-a]^n$. But $h(x_1, x_2, \ldots, x_n) = \sum_{k=1}^{n} \Psi_k(x_k)$, with $\Psi_k(x) = -n_k \log(x) - n'_k \log(1-x)$. Therefore, it suffices to prove that the function $\Psi_k(x) = -n_k \log(x) - n'_k \log(1-x)$ with n_k, n'_k integers ≥ 0 is convex in $[a, 1-a]$. The second derivative of $\Psi_k(x)$ is

$$\frac{d^2\Psi_k}{dx^2} = \frac{n_k}{x^2} + \frac{n'_k}{(1-x)^2},$$

which is > 0 for $x \in [a, 1-a]$. □

ACKNOWLEDGMENT

The author would like to thank the editor, Prof. C. Alexopoulos, and the anonymous reviewers for their comments, which have greatly contributed to the quality of the presentation; and V. Suñé for helping the author with the implementation of the algorithms for the computation of failure transition distances.

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*, 1998, pp. 1479–1486.

[2] ——, "Estimating reliability measures for highly-reliable Markov systems using balanced likelihood ratios," *IEEE Trans. on Reliability*, vol. 50, no. 3, pp. 265–280, September 2001.

[3] R. E. Barlow and F. Proschan, *Statistical Theory of Reliability and Life Testing. Probability Models*. : McArdle Press, 1981.

[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, pp. 31–41, March 1999.

[6] A. E. Conway and A. Goyal, "Monte Carlo simulation of computer system availability/reliability models," in *Proc. 17th IEEE Int. Symp. on Fault-Tolerant Computing*, 1987, pp. 230–235.

[7] M. R. Garey and D. S. Johnson, *Computers and Intractability. A Guide to the Theory of NP-completeness*. San Francisco, CA: W. H. Freeman and Company, 1979.

[8] P. E. Gill, W. Murray, and M. H. Wright, *Practical Optimization*. London: Academic Press, 1981.

[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, pp. 36–51, January 1992.

[11] J. M. Hammersley and D. C. Handscomb, *Monte Carlo Methods*. Methuen: , 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, pp. 772–788, 1976.

[13] S. Juneja and P. Shahabuddin, "Fast simulation of Markov chains with small transition probabilities," *Management Science*, vol. 47, no. 4, pp. 547–562, April 2001.

[14] ——, "Splitting-based importance-sampling algorithm for fast simulation of Markov reliability models with general repair-policies," *IEEE Trans. on Reliability*, vol. 50, no. 3, pp. 235–245, September 2001.

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

[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, pp. 52–88, 1994.

[17] ——, "Fast simulation methods for highly dependable systems," in *Proc. Winter Simulation Conf.*, 1994, pp. 221–228.

[18] ——, "General conditions for bounded relative error in simulations of highly reliable Markovian systems," *Advances in Applied Probability*, vol. 28, pp. 687–727, 1996.

[19] V. F. Nicola, P. Heidelberger, and P. Shahabuddin, "Uniformization and exponential transformation: techniques for fast simulation of highly dependable non-Markovian systems," in *Proc. 22nd IEEE Int. Symp. on Fault-Tolerant Computing*, 1992, pp. 130–139.

[20] V. F. Nicola, P. Shahabuddin, P. Heidelberger, and P. W. Glynn, "Fast simulation of steady-state availability in non-Markovian highly dependable systems," in *Proc. 23th IEEE Int. Symp. on Fault-Tolerant Computing*, 1993, pp. 38–47.

[21] V. F. Nicola, M. K. Nakayama, P. Heidelberger, and A. Goyal, "Fast simulation of highly dependable systems with general failure and repair processes," *IEEE Trans. on Computers*, vol. 42, no. 12, pp. 1440–1452, December 1993.

[22] C. H. Papadimitriou, *Combinatorial Optimization: Algorithms and Complexity*. Englewood Cliffs, New Jersey: Prentice-Hall Inc., 1982.

[23] J. L. Peterson, *Petri Net Theory and the Modeling of Systems*. : Prentice-Hall, 1981.

[24] A. Reibman and K. S. Trivedi, "Numerical transient analysis of Markov models," *Computers and Operations Research*, vol. 15, pp. 19–36, 1988.

[25] R. Rubinstein, "How to optimize discrete-event systems from a single sample path by the score function method," *Annals of Operations Research*, vol. 27, pp. 175–212, 1991.

[26] J. R. Schott, *Matrix Analysis for Statistics*. New York: John Wiley and Sons, 1997.

[27] R. F. Serfozo, "Point Processes," in *Handbooks in Operations Research and Management Science*, D. P. Heyman and M. J. Sobel, Eds. North-Holland: , 1990, vol. 2, Stochastic Models.

[28] P. Shahabuddin, F. Nicola, P. Heidelberger, A. Goyal, and P. W. Glynn, "Variance reduction in mean time to failure simulations," in *Proc. 1988 Winter Simulation Conference*, 1988, pp. 491–499.

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

[30] ——, "Importance sampling for the simulation of highly reliable Markovian systems," *Management Science*, vol. 40, no. 3, pp. 333–352, March 1994.

[31] B. C. Shultes, "Regenerative Techniques for Estimating Performance Measures of Highly Dependable Systems With Repairs," Ph. D. thesis, School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, USA, 1997.

[32] M. R. Spiegel, *Mathematical Handbook of Formulas and Tables*. New York: McGraw-Hill, 1968.

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

[34] V. Suñé, "Failure Distance Based Bounds of Dependability Measures," Ph. D. thesis, Universitat Politècnica de Catalunya, , 2000.

[35] P. D. Welch, "The statistical analysis of simulation results," in *Computer Performance Modeling Handbook*, S. S. Lavenberg, Ed. New York: Academic Press, 1983, pp. 268–329.

[36] T. Zhuguo and E. E. Lewis, "Component dependency models in Markov Monte Carlo simulation," *Reliability Engineering*, vol. 13, pp. 45–61, 1985.

Juan A. Carrasco (SM) has been "Profesor Titular" at the "Departament d'Enginyeria Electrònica" of UPC ("Universitat Politècnica de Catalunya") since 1988. His research is focused on fault-tolerant systems modeling. He received the Engineer degree (1981) in Industrial Engineering from UPC, the Master of Sc. degree in Computer Science (1987) from Stanford U., and the Doctor Engineer degree (1987) from UPC. He has been a member of the program committees of several international conferences. His research interests encompass numerical techniques, bounding methods, fast simulation methods for Markovian dependability & performability models, combinatorial methods for dependability evaluation, and applications. Most of the numerical techniques for the analysis of Markov models he has developed have been incorporated into METFAC-2.1 (http://dit.upc.es/qine/tools/metfac), a Markovian dependability & performability modeling tool developed under his direction. He is a Senior member of the IEEE.