Fault Injection Simulation:

A Variance Reduction Technique For Studying The Performance Consequences Of Rarely Occurring Failures In Communication Networks

Aad P.A. van Moorsel
Boudewijn R. Haverkort
Ignas G. Niemegeers

University of Twente
Department of Computer Science, Tele-Informatics and Open Systems group
P.O. Box 217, 7500 AE Enschede, The Netherlands
tel. +31 53 893767; fax: +31 53 333815; e-mail: moorsel@cs.utwente.nl.

Abstract

In this short paper a new technique, called Fault Injection Simulation (FIS), is discussed that is suited for deriving results for steady-state measures in discrete event dynamic systems which are strongly influenced by rarely occurring events. FIS is based on division of the observations in those that are affected and those that are not affected by these rare events. If methods are available FIS can be used as a (partly) analytical technique, else as a pure fast simulation technique. Under intuitively appealing assumptions FIS gives an unbiased estimator and a variance reduction. In this short paper we discuss FIS from a practical point of view; when can FIS be used and how should FIS be used. Furthermore, we show how to adjust FIS during the simulation to benefit maximally from its variance reduction capabilities.
KEYWORDS: Discrete event simulation, variance reduction techniques, rare events, performability evaluation, real-time systems.

1 Introduction

Fault tolerant distributed computer and communication systems are often used in environments in which they have to satisfy certain real-time constraints. Especially in safety critical applications the dependability characteristics of a system have to support the real-time performance. To establish the real-time characteristics it is necessary to integrate performance and dependability modelling and evaluation; in other words, to consider the *performability* of a system [1].

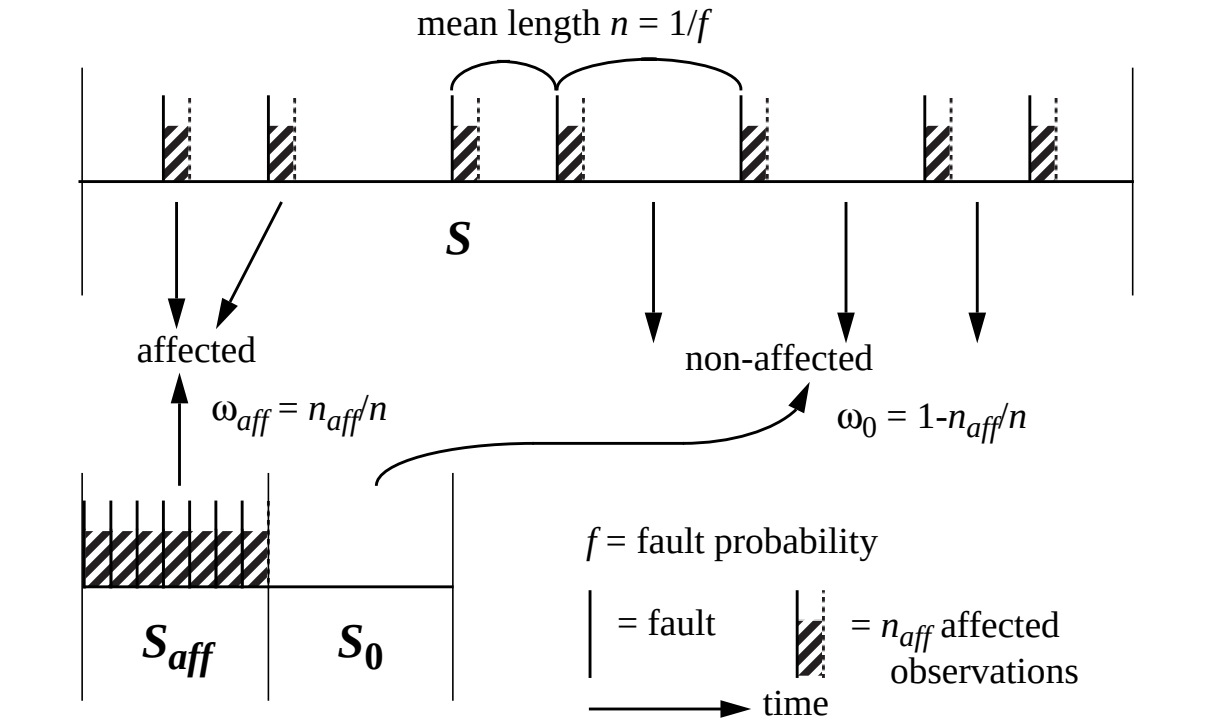
Performability modelling has been most successfully applied for gracefully degradable computer systems by using Markov reward models [2]. In this approach the relation of the performability model with the underlying performance and dependability models is exploited. For many systems and for many measures of interest this type of modelling is not suited, so another performability modelling approach is necessary. In this paper we discuss *Fault Injection Simulation* (FIS) [3], a modelling approach which can be used to evaluate highly

dependable systems with very small failure rates. In this respect one can think for instance of the very small token loss probabilities in the high speed FDDI token ring.
FIS decomposes a realization of a discrete event dynamic system [4] in a fault affected and a non-affected part. For these two parts steady-state measures are derived and weighted to calculate the desired overall steady-state measure. The individual measures can be obtained by analytical or numerical means or by simulation. Often analytical or numerical methods will not be available in which case FIS is a pure fast simulation technique.
When one wants accurate simulation results in a reasonable time span for systems that are strongly influenced by rarely occurring events, the use of variance reduction techniques (e.g. [4], [5], [6]) is inevitable. Otherwise far too much time is needed before enough of these *rare events* are simulated. Besides the possibility to use FIS as a partly analytical or numerical method, first results indicate that FIS is also competitive with known fast simulation techniques in ease of applicability and amount of variance reduction. An important reason for that is the possibility to tune FIS during the simulation such that it reduces variance most.
In this short paper it is not our intention to discuss the theoretical background of FIS in full detail (therefor see [3]) but merely to discuss the practical impact of the method. This paper is organized as follows. First we briefly introduce FIS in Section 2. In Section 3 we discuss the use of FIS; when to use it and how to make optimal use of it. Therefore we discuss a method to adjust FIS during the simulation in order to obtain maximum variance reduction. This is illustrated by a simple example of an M/M/1 queueing model with rarely occurring service breakdowns. Section 4 concludes this short paper.

2 Fault Injection Simulation

Fault Injection Simulation (FIS) is a form of decomposition of a discrete event system [4] which is visualized in Figure 1. We discuss FIS in the context of rarely occurring faults in computer and communication networks and use the corresponding terminology. FIS is based on the idea that the effect of a fault on the observations [7] fades out. After a fixed number of

Figure 1 Fault Injection Simulation



observations, denoted n_{aff}, the observations are considered to be no longer affected by the occurred fault and to be identical to those in a fault free system. This forms the basis of the following three steps of FIS:

A. Division of observations:

When S is the realization of a system with probability f that a fault occurs between two successive observations, then we divide the observations in S in two sets:

1. *affected observations* which are considered to be affected by an occurred fault. These are the n_{aff} observations directly following a fault occurrence;
2. *non-affected observations* which are considered not to be affected by a fault.

B. Separate simulation:

The two sets are simulated separately from each other by the simulation run S_{aff} and S_0 such that:

1. the *injected simulation* S_{aff} gives a realization of a system with a fault occurring (*injected*) deterministically every n_{aff} observations;
2. the *fault free simulation* S_0 gives a realization of a system without fault occurrences.

Both S_{aff} and S_0 are steady-state simulations.

C. Weighting:

The results of the simulation runs S_{aff} and S_0 are weighted by the *weighting factors* ω_{aff} and ω_0. Let $\hat{Y}_{aff}$ and $\hat{Y}_0$ respectively be the estimators obtained by S_{aff} and S_0, then the estimator $\hat{Y}_{FIS}$ obtained by FIS is:

$$\hat{Y}_{FIS} = \omega_{aff}\hat{Y}_{aff} + \omega_0\hat{Y}_0. \quad (1)$$

The weighting factors represent the fraction of the overall number of observations that are affected respectively non-affected. This means:

$$\omega_{aff} = \frac{n_{aff}}{n} \text{ and } \omega_0 = 1 - \frac{n_{aff}}{n}, \quad (2)$$

with $n = 1 / f$, the mean number of observations between two successive faults.

It can be shown [3] that under intuitively attractive assumptions the FIS estimator in (1) gives an unbiased estimator of the measure of interest in the original system.

3 The use of FIS

In this section we discuss on when and how to use FIS. First we describe in Section 3.1 for what kind of systems FIS can be used, then we discuss in Section 3.2 how to use FIS in an optimal way. In Section 3.3 we illustrate the variance reduction possibilities of FIS by an M/M/1-queue with rarely occurring service breakdowns.

3.1 When to use FIS

If one wants to apply FIS the considered system has to have some characteristics which we discuss in this section.

- First, n_{aff} has to be chosen such that the effect of a fault on the measure of interest has faded out within n_{aff} observations. Moreover, because we inject a fault in S_{aff} every n_{aff} observations the system has to be in steady-state of the fault free system. Theoretically this will never completely be the case. The faced problem has similarities with that of finding an appropriate length for the initial transient phase of a simulation (see Pawlikowski [8] for a survey of study in this area as well as several rules of thumb).
- The faults have to occur in steady-state of the fault free system. It is also possible to use FIS for systems with bursty fault occurrences provided that the first fault of a burst occurs in steady-state.

- The fault probability f has to be known to make it possible to derive the appropriate weighting factors. The fault probability is defined as the probability that a fault occurs between any pair of successive observations. FIS as discussed in Section 2 is not directly applicable when the fault probability depends on the concrete values of the observations.

3.2 How to use FIS

When the system under consideration is suited to apply FIS, it is of course advisable to apply FIS in such a way that it reduces the variance most. Most reduction is achieved when analytical or numerical methods are used to obtain results for S_{aff} or S_0. When these are not available the amount of reduction depends on the fraction of the simulation time dedicated to S_{aff} and S_0 respectively. An optimal fraction depends on the variance and the mean of the observations in both simulations. We discuss this in some detail.

Variance reduction possibilities of FIS

Let us make use of FIS as a pure simulation technique and divide the observations in both S_{aff} and S_0 in batches of size n_{aff} which we consider to be independent. In other words, we use the *method of batch means* [4] to derive confidence intervals. Let the results of the batches in S_{aff} be distributed such that they have variance $\boldsymbol{V}_{aff}^2$ and mean $\boldsymbol{E}_{aff}$ and let the batches in S_0 have variance $\boldsymbol{V}_0^2$ and mean $\boldsymbol{E}_0$. The variance and the mean of both simulations can be estimated using the collected simulation results. When a simulation of m batches is carried out, we call the fraction β ($0 < \beta < 1$) of batches that are simulated for S_{aff} the *allocation fraction*. So βm batches are dedicated to S_{aff} and $(1 - \beta)m$ to S_0. Now it can be derived [3] how the achieved amount of variance reduction depends on the allocation fraction β. We give some results in this paper, based on results about Stratified Sampling [9] which partly can be found in [3].
Define A, B and C as follows, with the weighting factors ω_{aff} and ω_0 as defined in (2):

$$A = (\omega_{aff}\boldsymbol{V}_{aff} + \omega_0\boldsymbol{V}_0)^2;\ B = \omega_{aff}\omega_0(\boldsymbol{V}_{aff} - \boldsymbol{V}_0)^2;\ C = \omega_{aff}\omega_0(\boldsymbol{E}_{aff} - \boldsymbol{E}_0)^2;$$

and consider the following allocation fractions:

Optimal allocation fraction: $$\beta_{opt} = \frac{\omega_{aff}\boldsymbol{V}_{aff}}{\omega_{aff}\boldsymbol{V}_{aff} + \omega_0\boldsymbol{V}_0};$$

Proportional allocation fractions: $$\beta_{prop}^{(1)} = \omega_{aff} \text{ and } \beta_{prop}^{(2)} = \frac{\omega_{aff}\boldsymbol{V}_{aff}^2}{\omega_{aff}\boldsymbol{V}_{aff}^2 + \omega_0\boldsymbol{V}_0^2}.$$

In Table 1 the variance per batch is given for these allocation fractions as well as for direct simulation of the original system.

Table 1 Variance versus allocation fraction β

Simulation	*Variance per batch*
$\beta = \beta_{opt}$	A,
$\beta = \beta_{prop}^{(1)}$ or $\beta = \beta_{prop}^{(2)}$	$A + B$,
direct simulation	$A + B + C$.

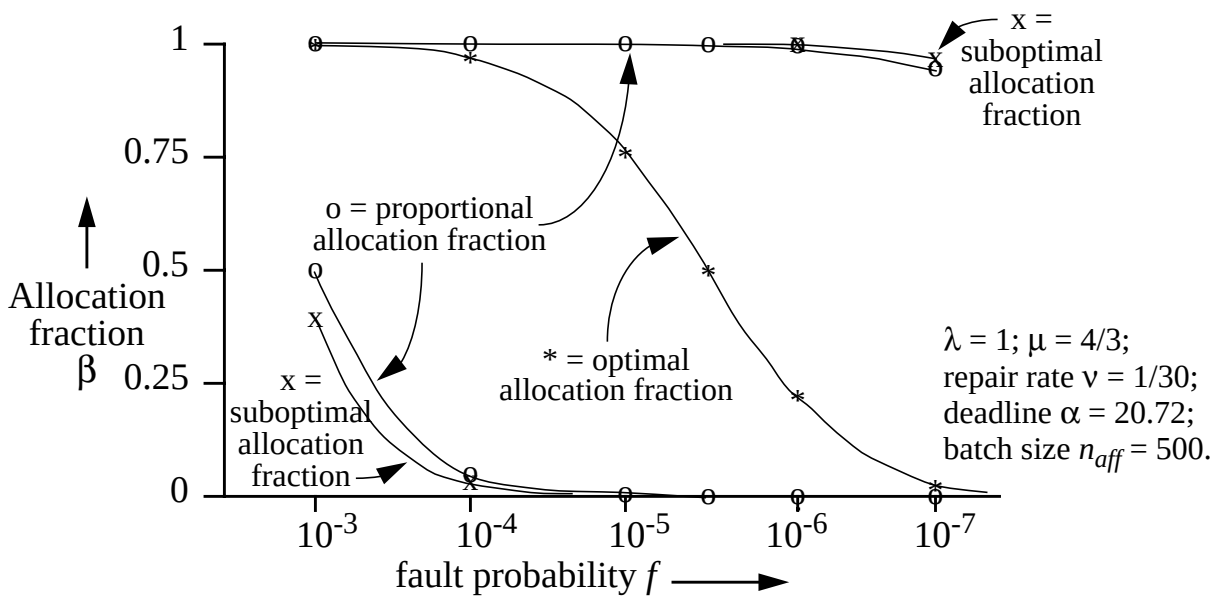
In [3] it is shown that the variance reduction that can be achieved using FIS is most when FIS is carried out with allocation fraction $\beta = \beta_{opt}$. The variance reduction obtained by FIS with $\beta = \beta_{prop}^{(1)} = \omega_{aff}$ is due to the fact that the probabilistic nature of the fault behavior is removed; the fault rate remains equal but faults occur deterministically. Notice that applying

FIS with the proportional allocation fraction $\beta = \omega_{aff}$, whose value is known in advance, always leads to reduction of the obtained variance.
There exist two allocation fractions $\beta_{sub}^{(1)}$ and $\beta_{sub}^{(2)}$ ($\beta_{sub}^{(1)} < \beta_{opt} < \beta_{sub}^{(2)}$) for which the variance of a batch in FIS equals the variance of a batch in the direct simulation. These *suboptimal allocation fractions* are the borders of the interval of allocation fractions for which FIS is useful, i.e. reduces variance.
To benefit maximally from the variance reduction capabilities of FIS, one needs an indication of the value of the optimal allocation fraction. This value is not known in advance but can be estimated from the first simulation data. During the simulation the used allocation fraction can be changed such that the ultimate variance reduction is maximal. This procedure can for instance go as follows. Let a simulation start with 100 batches of both S_{aff} and S_0. From the collected data the optimal allocation fraction β_{opt} is estimated as well as the number of batches m, necessary to get results that are 95% certain within 10% of the FIS estimate. Then the FIS simulation continues with simulating $\beta_{opt}m$ affected batches and $(1 - \beta_{opt})m$ non-affected batches. This procedure can be applied iteratively until the desired accuracy is obtained and might even be improved by taking a maximum for the number of observations that can be made without calculating a new optimal allocation fraction.

3.3 An example: M/M/1-queue with server breakdown

In this section we show how this way of applying FIS works out for a simple M/M/1-queue with rarely occurring service breakdowns. The M/M/1-queue has arrival rate λ, service rate μ and an infinite buffer. After each service completion a service breakdown can occur with probability f, after which an exponentially distributed 'repair time' with mean $1/\nu$ follows. We are interested in the steady-state probability $\bar{F}(\alpha)$ of exceeding some response time (*deadline*) α. The response time is the waiting time of a client in the queue plus its service time.
We show results for the following parameter values: $\lambda = 1$, $\mu = 4/3$, repair time $\nu = 1/30$, fault probability $f = 10^{-4}$, deadline $\alpha = 20.72$ and batch size $n_{aff} = 500$. For each batch we keep track of the number of jobs exceeding the deadline.
The results for the two simulations in FIS are after 100 batches in both simulations: for S_{aff}: $\boldsymbol{E}_{aff} = 86.83$, $\boldsymbol{V}_{aff}^2 = 13595$ and for S_0: $\boldsymbol{E}_0 = 0.02$, $\boldsymbol{V}_0^2 = 0.04$. From this the optimal

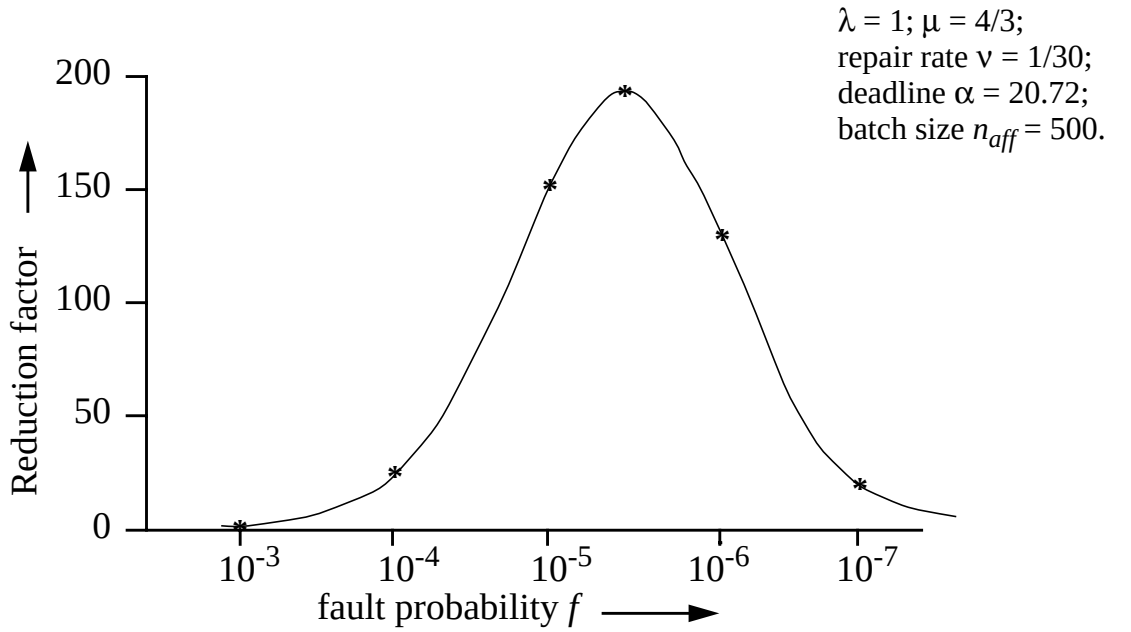
Figure 2 Allocation fractions



allocation fraction β_{opt} and the proportional allocation fractions can be derived. Furthermore it is possible to calculate the suboptimal allocation fractions from the variance that the direct estimator would have had (see above). This leads to Figure 2.

When we apply FIS for some fault probability f with the corresponding optimal allocation fraction, the amount of variance reduction would be as given in Figure 3. This factor is equal to the reduction factor in the number of batches that needs to be simulated. Notice that both Figure 2 and Figure 3 show point estimates of the optimal allocation fraction and the amount of time saving respectively.

Figure 3 Variance reduction factors



We can calculate β_{opt} and the obtained simulation time reduction for a range of fault probabilities but we continue the simulation with the optimal allocation fraction belonging with $f = 10^{-4}$. We followed the procedure described above to use FIS in an optimal way, until the probability of exceeding the deadline for $f = 10^{-4}$ was 95% certain within 10% of the estimate. The final results for our simulation are shown in Table 2. It gives the estimated probabilities of missing the deadline α with 95% confidence intervals.

Table 2 $\overline{F}(\alpha)$ with confidence intervals

f	*point estimate with 95% confidence interval*
10^{-7}	$4.62\times10^{-5} \pm 6.61\times10^{-5}$
10^{-6}	$1.10\times10^{-4} \pm 0.77\times10^{-4}$
10^{-5}	$7.41\times10^{-4} \pm 1.03\times10^{-4}$
10^{-4}	$7.06\times10^{-3} \pm 0.69\times10^{-3}$
10^{-3}	$7.03\times10^{-2} \pm 0.69\times10^{-2}$

To conclude this example we note that it is possible to use known analytical results for the fault free part of the M/M/1-model. Because no fault free observations have to be made in that case, the same number of faults are simulated with FIS in a fraction n_{aff} / n of the time necessary with direct simulation. Another possibility is to use variance reduction techniques, such as

importance sampling ([4], [6]), within the two simulations of FIS. This will further decrease the necessary simulation time.

4 Conclusion

In this short paper we have emphasized on the practical aspects of Fault Injection Simulation (FIS), a method to study the performance consequences of rarely occurring failures in computer and communication networks. We have shown an example of how to adjust FIS during the simulation such that its variance reduction possibilities are exploited maximally. Besides the possibility to use FIS as a (partly) analytical technique, the obtained variance reduction in toy examples seem to make FIS also very attractive when it is used as a pure fast simulation method.

References

[1] J.F. Meyer, On Evaluating the Performability of Degradable Computing Systems, *IEEE Transactions on Computers, Vol.C-29, No.8, August 1980, pp.720-731.*

[2] R.M. Smith, K.S. Trivedi, A.V. Ramesh, Performability Analysis: Measures, an Algorithm and a Case Study, *IEEE Transactions on Computers, Vol.37, No.4, April 1988, pp.406-417.*

[3] A.P.A. van Moorsel, B.R. Haverkort, I.G. Niemegeers, Fault Injection Simulation: A Variance Reduction Technique for Systems with Rare Events, *Proceedings of the Second International Working Conference on Dependable Computing for Critical Applications, Tucson, February 1991, also as Memoranda Informatica 91-05, University of Twente, Enschede, The Netherlands, 1991.*

[4] G.S. Fishman, Principles of Discrete Event Simulation, *John Wiley & Sons, New York, 1978.*

[5] V.S. Frost, W.W. Larue Jr., K.S. Shanmugan, Efficient Techniques for the Simulation of Computer Communication Networks, *IEEE Journal on Selected Areas in Communications, Vol.6, No.1, January 1988, pp.146-157.*

[6] A. Goyal, P. Shahabuddin, Ph. Heidelberger, V.F. Nicola, P.W. Glynn, A Unified Framework for Simulating Markovian Models of Highly Dependable Systems, *IBM Research Report RC 14772, November 1989 (to appear in IEEE Transactions on Computers).*

[7] P.D. Welch, The Statistical Analysis of Simulation Results, C*omputer Performance Modeling Handbook, S.S. Lavenberg (Ed.), Academic Press, New York, 1983***.**

[8] K. Pawlikowski, Steady-State Simulation of Queueing Processes: A Survey of Problems and Solutions, *ACM Computing Surveys, Vol.22, No.2, June 1990, pp.123-170.*

[9] W.G. Cochran, Sampling Techniques, *John Wiley & Sons, New York, 1977.*