Evaluation of Performability Measures for Replicated Banyan Networks *

Amiya Bhattacharya  Ramesh R. Rao  Ting-Ting Y. Lin
CSE Department  ECE Department  ECE Department
University of California, San Diego
La Jolla, CA 92093

Abstract

Performability, a composite measure that evolved from the synergy between performance and reliability, can capture the cumulative performance of a degradable multistage interconnection network over its operational life. In this paper, we present a technique for conservative analysis of performability measures for the replicated banyan network. Assuming uniformly distributed message generation at unblocked sources, we make a reward assignment for synchronous circuit-switched operation that translates the cumulative performance to total number of messages routed. The analytical results and simulated values for the average level of the cumulative performance of a replicated banyan are compared. The analysis technique is capable of handling non-Markovian model of component failure. The effect of perturbation from the Markovian model on the cumulative performance has been studied.

1 Introduction

Multistage interconnection networks (MIN), a low-cost alternative to full crossbar, have been used to set up connections between two groups of modules in a wide range of computer and communication systems. The shared memory multiprocessor system is the most common example, where every memory access for a processor takes place through a MIN placed between the processors and memory modules. Banyan networks [13] (also referred to as delta networks by some authors [3]), a class of blocking MINs with unique path property, have received the widest publicity for their self-routing capability. A simple multipath banyan can be constructed by replicating the single-path pro-

*Supported by ONR under grant N0001491-1-J1017 and NSF under grant NCR-8904029

totype and coupling them at the input and output. Each layer of a replicated banyan inherits the self-routing and blocking properties, and together they achieve higher permutation capability and fault tolerance. The fault-free MINs offer better performance and reliability, but the performance-reliability trade-off plays an important role when the system operates in a degraded mode as a consequence of component failure.

The traditional performance analysis of fault-free multipath banyans [5, 9] left out reliability issues. On the other hand, the degraded modes of operation were ignored by structural reliability analysis techniques [11]. The gap was bridged by the introduction of a composite measure called performability [1] and its wide acceptance [18]. This measure is essentially the probability that the system reaches an accomplishment level over a mission time. Under this framework, the system configuration under successive faults is characterized by a stochastic process. Even if the system is potentially repairable, the designer is concerned with the transient behavior of the system over the mission. So the quantification reduces to the evaluation of the distribution or moments of some accumulated performance metric (reward) defined on this stochastic process.

The closed form solution to performability can be obtained by conditioning on the state trajectories (sample paths) [4, 6]. This involves a number of integral evaluations that grows linearly with the number of trajectories. An exponential growth of trajectories is quite common. The tradition is to do performability analysis under a Markov reward model [4, 8, 10, 14, 16], where the problem reduces to the transient analysis of the underlying Markov chain. However, due to the memoryless property of Markov process, it implies an exponential distribution of sojourn time, and a constant hazard rate. The valid-
ity of this assumption is oftentimes questionable. In practice, system components have been better characterized by varying hazard rate over the lifetime, often modelled by Weibull distribution [12, 19]. The use of exponential approximation to non-exponential distribution is known to be a cause of considerable error in transient analysis [15].

The method outlined in this paper is a very reasonable trade-off in this regard. In [21], we developed a technique based on order statistics for exact analysis of performability measures for an unstructured multi-component system. It imposes no assumption on the component life other than mutual independence, and we deduced the distributions of the sojourn times from the distributions of individual component lifetimes. Here we present a technique that creates a conservative bound for performability of replicated banyans based on a reduced state-space, and evaluates it under a similar analytical framework. For completeness, the order statistics formulation has been partially reproduced.

For illustration, we consider the synchronous circuit switched operation of a non-repairable replicated MIN in all analyses and simulations throughout this paper. However, there is absolutely no loss of generality, because factors such as switching, timing, protocol and source behavior that affect the performance, only change the reward rates. We formulate the performance models of replicated banyans for all configurations using a reduced state-space. Subsequently, a meaningful delay oriented reward rate assignment is chosen that conforms to the bounding argument used in state-space reduction. Unlike some previous works which drop requests at the point of conflict [3], we analyse blocking and its effect on delay under resubmission as in [20]. This leads to a bound on mean delay at each degraded state under consideration, that is conservative in estimating the accumulated reward over successive levels of degradation.

2 Performability and Related Measures

Let \( \{X_t\}_{t \geq 0} \) be a continuous-time discrete-space stochastic process representing the state of the degrading system. The state-space \( Q \) is finite for any practical system, however large it may be. Each state \( q \in Q \) represents a possible configuration of the system with some components failed, and has an associated reward rate \( \rho_q = \rho(q) \), a non-negative real number reflecting the level of performance per unit time, when the system is in state \( q \). Performability of the system over a mission time \( t \) is defined as the probability density function (pdf) of the accumulated reward (cumulative performance)

\[
Y_t = \int_0^t \rho(X_s) \, ds. \tag{1}
\]

The cumulative distribution function (cdf) of \( Y_t \) is called the performability distribution function. The mean value of \( Y_t \), which we would call mean reward in mission (MRIM), is a useful characterization. In this paper we focus our attention on the accumulated reward until the complete system failure, and consider the value of \( Y_t \) as \( t \) approaches infinity. We use the abbreviation mean reward until failure (MRUF) for \( E[Y_{\infty}] \).

For a nonrepairable system reentry to a state is not feasible. The sojourn time or the residence time \( \tau_q \), which is the total time spent by the system in configuration \( q \) over the mission time \( t = \sum_{q \in Q} \tau_q \) is therefore contiguous. As \( t \to \infty \), equation (1) reduces to

\[
Y_{\infty} = \sum_{q \in Q} \rho_q \tau_q, \tag{2}
\]

which also provides a simple expression for MRUF

\[
E[Y_{\infty}] = \sum_{q \in Q} \rho_q E[\tau_q]. \tag{3}
\]

One implicit assumption in formulating the composite performance-reliability measure as above is that the performance and failure rates are functions of only the system state. This means that increasing load or stress during the degradation due to component failure would not affect the lifetime distribution of the surviving components. This is not only a feature which makes the analysis tractable, but also a desired characteristic in many implementations.

3 Multipath Banyan Networks

A \( N \times N \) (where \( N = m^K \)) symmetric MIN is constructed by arranging \( m \times m \) crossbar switches in \( K \) stages, with \( N/m \) switches in each stage. We adhere to \( m = 2 \) henceforth, as \( 2 \times 2 \) switches are most common. Switches of adjacent stages are connected by an interconnection pattern, so that a path can be established between any pair of input and output [3, 13]. MINs with unique paths between any input-output pair form the class of banyan or delta networks, that includes well-known topologies such as flip, omega and baseline networks. The unique path gives rise to their self-routing capability — switches at successive stages can
complete routing over the network by locally checking the successive bits of the destination addresses. However, paths between disjoint input-output pairs share links, and two propagating message headers may conflict. The situation is usually resolved by blocking one of the messages randomly. As a result, these networks are called blocking networks.

An \( r \)-replicated banyan (also known as layered banyan) consists of \( r \) identical distinct copies of the single-path prototype \([5, 9]\). Figure 1 illustrates the construction of a \( 8 \times 8 \) multistage shuffle-exchange (or omega) network, each of the layers shown as rectangles in (b) is individually a single-path omega as shown in (a), coupled at the inputs and outputs by the distributors and collectors. The MIN is considered corrupted if there exists a single fault in every one of the layers, because at this point complete biconnectivity cannot be guaranteed.

4 Reward Model

Let us recall that the reward rate reflects the steady level of performance per unit time, while the MIN is under some specified configuration. The assumption that the configuration does not change before this steady state is reached is critical in decomposing the performability problem into subproblems of performance evaluation. However, the solvability and complexity of these subproblems dictate the reward assignment so that the performability measure is both meaningful and tractable.

4.1 Choice of reward

In a shared-memory multiprocessor environment, the initial distribution of traffic is governed by the memory reference pattern imposed by the running processes. We assume a uniform distribution that has become a standard in the literature \([3, 5]\). Moreover, we adhere to strict data-dependency in memory reference within a processor. This ensures that blocked requests are not dropped and the net effect of their resubmission is taken into account. We proposed a resubmission protocol earlier \([20]\) that extends the analysis of probability of success \([3, 5]\) to an upper bound analysis for delay, provided it is assumed that the destinations are also re-randomized during resubmission. We have observed that the average delay in any active configuration settles down to a value very close to this analytical bound.

The key assumption that the reward rate must be a function of only the system state means that the performance metric to be used as reward should attain a steady state value between state changes. Since the network cycle time is extremely small as compared to component lifetime, resubmissions when ignored at the end of the sojourn time has no effect on the mean delay. All requests generated can thus be considered successful over the long run, and the network becomes lossless. Under this condition, the choice of \( Y_t = \sum_{t=1}^N (N_r(t))/\delta_s \tau_t \) represents the total number of requests routed during the mission time, where \( N \) is the number of sources, \( \delta_s \) is the delay (in network cycles) and \( \tau_t \) is the network cycle time. This leads to an obvious choice of reward rate \([22]\)

\[
\rho_s = \frac{N}{\delta_s \tau_t}.
\]  

4.2 Delay analysis

A resubmission protocol proposed in \([20]\) has been used to derive an upper bound on the mean delay.
To enforce the uniform traffic distribution in successive network cycles, this protocol dictates the following modification in source behavior. Blocked sources, not being allowed to generate a new request in the next sweep, resubmit the unserviced request with some probability \( q \). However, they are assumed to re-randomize the choice of destination. Unblocked sources have a probability \( p \) for generating a new request \((p \leq q)\). For the analysis, however, they are also assumed to submit a new request independently with probability \( q \), effectively loading the network with spurious requests. This leads to an overestimation of the mean delay, and the upper bound has been found to be

\[
\bar{\delta} = 1 + \frac{1}{q \nu} - \frac{1}{q},
\]

(5)

where \( \nu \) is the probability of success, defined as the probability that a request survives at the output of the MIN, given it is submitted at the input.

Computation of \( \nu \) for replicated banyan can be simplified if we make the independence approximation as in [5]. For a single layer, define \( \{q_i, 0 \leq i < K\} \) to be the probability that an input port of an \( i \)th stage switch in that layer carries a request, while \( q_K \) is the probability that a request appears at any of the output ports of the \( K \)th stage (i.e., the output of that layer). Initialize \( q_0 = q \) when requests are multiply loaded. Success through any of the layers results in a success, and as a result \( \nu \) is given by

\[
\nu = 1 - (1 - \frac{q_k}{q})^r,
\]

where

\[
q_{i+1} = 1 - (1 - \frac{q_i}{2})^2, \quad 0 \leq i \leq K - 1.
\]

Clearly, \( \bar{\delta} \) in equation (5) is a function of \( q \), a design parameter to be chosen within the range \([p, 1]\).

4.3 Reduction of state-space

Substituting the value of \( \nu \) in the equation (5) we have \( \bar{\delta} \) as an upper bound for the average delay of the MIN. Since the result has been derived based on a uniform request distribution over uniformly replicated banyan [20], \( \nu \) needs to be recomputed for all levels of replication or dilution, and we have the reward assignment for certain configurations only. However, we can argue that the reward rates for any configurations in between can be conservatively estimated by the reward rates of one of those known configurations. As a result, we consider only those configurations for which an exact value of the assigned reward is known, and all other configurations can be identified to be in between two of those. This leads to the development of the order statistics formulation, described next.

5 Order Statistics Formulation of Performability

Our goal is to have a modelling scheme that can potentially handle more general distribution of component lifetime than the traditional Markov reward model. The Markovian transient analysis essentially involves formulation of a system of differential difference equations from the directed acyclic graph representing the degradation. The system of equations are usually solved in a transformed domain. However, if the degradation graph is a chain, it is easy to represent the order statistics of lifetimes in time domain and relate them to the sojourn times. Closed form expressions for the marginal and joint distributions of the order statistics are known in terms of the distribution of component lifetimes with any arbitrary distribution. We exploit this generality and attempt to embed by conditioning a number of such degradation chains into the reduced degradation graph of the replicated MIN.

5.1 State-space

A \( r \)-replicated banyan network has \( r \) unique-path banyan layers. Each layer has \( S = (N \log_2 N)/2 \) switches and \( L = N(\log_2 N + 1) \) links. The degradation of the MIN can be conservatively decomposed as follows. A layer gets corrupted by a single failure of any of its switches or links, in the sense that at least one pair of source and destination cannot be connected through that layer. For a conservative bound on performability, we ignore the contribution of such a defective layer, and assume that the network has as many layers as the perfect ones. Thus the MIN reduces to an \( r \)-component system, where each layer can be considered as a component. In what follows, we show how the analysis technique for multicomponent systems as developed in [21], can directly be applied to derive a closed form integral expression for the bound.

\[
F(t), \text{ the cdf of lifetime of the layer, can be obtained using the argument that it would require all } S \text{ switches and } L \text{ links to be fault-free at time } t \text{ so that the layer does not fail until time } t. \text{ i.e.}
\]

\[
F(t) = 1 - (1 - F_s(t))^S(1 - F_l(t))^L
\]

(7)

\[
f(t) = \{1 - F_s(t)\}^S L(1 - F_l(t))^{L-1} f_l(t)
\]

\[+ S \{1 - F_s(t)\}^{S-1} (1 - F_l(t))^L f_s(t)
\]

(8)
where $F_0(t)$ and $F_l(t)$ are the cdf's of the lifetime of the individual switches and links; $f_0(t)$ and $f_l(t)$ are the corresponding pdf's.

Let $q_k$ denote the configuration of $k$ faulty and $(r - k)$ perfect layers. The state space consists of $\{q_k, k = 0, 1, \ldots, r\}$. For reward assignment, we can choose

$$\rho_k = \frac{N}{\delta_{r-k} t_k}$$  \hspace{1cm} (9)

where $\delta_{r-k}$ is the upper bound on mean delay for a $(r - k)$-replicated banyan. It should be noted that $\rho_r = 0$ in a conservative way.

The boundaries of the sojourn times can be identified with the time of successive failure of the layers. These are precisely the order statistics of the random variables representing lifetime of the layers. In what follows, we first introduce the order statistics, and show how their distributions relate to the distributions of the component lifetimes. Subsequently the closed form expressions for the performability distribution and MRUF are derived in terms of the distributions of the order statistics of component lifetimes.

5.2 Distribution of order statistics

If a family of random variables $T_1, T_2, \ldots, T_r$ are arranged in ascending order of magnitude and renamed as $T_{1:r} \leq T_{2:r} \leq \cdots \leq T_{r:r}$, the new random variable $T_{k:r}$ is called the $k$th order statistic for $k = 1, 2, \ldots, r$. Suppose $T_i$, $i = 1, 2, \ldots, r$ represents the lifetime of the $i$-th layer. They are independent and identically distributed (iid) if we make the assumption of independent failure. Let $F(t) = Pr[T \leq t]$ be the common cdf of component lifetime, and $f(t)$ be the corresponding pdf.

An extensive treatment of the distributions and moments of order statistics can be found in [2] and [17]. The cdf of the $k$th order statistic $T_{k:r}$ is given by

$$F_{k:r}(t) = \sum_{i=k}^r \binom{r}{i} \{F(t)\}^i \{1 - F(t)\}^{r-i},$$  \hspace{1cm} (10)

The corresponding pdf (marginal) is given by

$$f_{k:r}(t) = \frac{r!}{(r-1)!(r-k)!} \{F(t)\}^{k-1} \{1 - F(t)\}^{r-k} f(t).$$  \hspace{1cm} (11)

The marginal pdf is a special case of the general expression for the joint pdf of any $k$ order statistics out of the $r$. For $1 \leq r_1 < r_2 < \cdots < r_k \leq r$ and $t_1 \leq t_2 \leq \cdots \leq t_k$,

$$f_{r_1,\ldots,r_k}(t_1,\ldots,t_k) = \frac{r!}{(r_1-1)!(r_2-r_1-1)! \cdots (r-r_k)!} \{F(t_1)\}^{r_1-1} f(t_1) \{F(t_2) - F(t_1)\}^{r_2-r_1-1} f(t_2) \cdots \{1 - F(t_k)\}^{r-r_k}$$

5.3 Conservative analysis

We are now ready to evaluate the performability measures of the $r$-layer banyan. For simplicity, let us first start with the assumption that the distributors and collectors holding the $r$ layers never fail. The system degrades strictly through the state sequence $q_k, k = 0, \ldots, r$. Clearly $T_{k:r}, k = 1, \ldots, r$ are the sojourn time boundaries. Defining $T_{0:r} = 0$, the sojourn time $T_k$ at state $q_k$ is

$$T_k = T_{k+1:r} - T_{k:r}, \hspace{1cm} k = 0, 1, \ldots, r - 1$$  \hspace{1cm} (13)

A lower bound $\hat{Y}_\infty$ on the accumulated reward is given by

$$\hat{Y}_\infty = \sum_{k=0}^r \rho_k T_k$$

so that

$$Pr[\hat{Y}_\infty \leq \gamma] = Pr[\sum_{k=0}^r (\rho_k - 1) T_{k:r} \leq \gamma] = \int \cdots \int f_{T_{1:r},\ldots,T_{r:r}}(t_1,\ldots,t_r) dt_1 \cdots dt_r$$

Similarly, a lower bound on MRUF is found to be

$$E[\hat{Y}_\infty] = E[\sum_{k=0}^r (\rho_k - 1) T_{k:r}]$$

5.4 Critical fault

Until now we have not considered system failure caused by faulty distributors or collectors, which act as the critical components. The effect of such a component fault is an abrupt system failure — the graceful degradation gets truncated at that point. To incorporate the effect of critical components, we have to change the expression for the sojourn time by

$$T_k = 0, \hspace{1cm} T_{k+1:n} \leq T_k \leq T_{k:n}$$

Therefore, the system degrades strictly through the state sequence $q_k, k = 0, \ldots, r$. Clearly $T_{k:n}, k = 1, \ldots, r$ are the sojourn time boundaries. Defining $T_{0:n} = 0$, the sojourn time $T_k$ at state $q_k$ is

$$T_k = T_{k+1:n} - T_{k:n}, \hspace{1cm} k = 0, 1, \ldots, r - 1$$  \hspace{1cm} (17)
where $T_c$ denotes the lifetime of the critical component, distributed with cdf $F_c(t)$ and pdf $f_c(t)$.

It is quite straightforward to handle multiple critical components. One can lump the effect onto a single critical component of effective lifetime $T_c$ with cdf $F_c(t)$, by considering the hypothetical critical component to be active until the first real critical component failure. As there are 2N distributor/collectors with individual cdf $F_d(t)$, the effective distribution of critical component is given by

$$F_c(t) = 1 - (1 - F_d(t))^{2N}$$  \hspace{1cm} (18)

$$f_c(t) = 2N(1 - F_d(t))^{2N-1}f_{max}(t)$$  \hspace{1cm} (19)

Given that $T_{mr} < T_c \leq T_{m+1:r}$, $\tilde{Y}_c$ reduces to

$$\tilde{Y}_c = \sum_{k=0}^{m} \rho_k T_{kr} + \rho_m T_c$$

$$= \sum_{k=0}^{m} \rho_k (T_{kr} - T_{m+1:r}) + \rho_m (T_r - T_{m+1:r})$$

$$= \sum_{k=1}^{m} (\rho_k - \rho_m) T_{kr} + \rho_m T_c.$$  \hspace{1cm} (20)

As $T_{mr} < T_c \leq T_{m+1:r}$ are disjoint, by the axiom of probability

$$\Pr[\tilde{Y}_c \leq y]$$

$$= \Pr[(\tilde{Y}_c \leq y) \cap \{T_c \leq T_{1:1}\}]$$

$$+ \sum_{k=1}^{m-1} \Pr[(\tilde{Y}_c \leq y) \cap \{T_{mr} < T_c \leq T_{m+1:1}\}]$$

$$+ \Pr[(\tilde{Y}_c \leq y) \cap \{T_c < T_r\}]$$

$$= \Pr[(\rho_k T_c \leq y) \cap \{T_c \leq T_{1:1}\}]$$

$$+ \sum_{k=1}^{m-1} \Pr[(\rho_k - \rho_m) T_{kr} + \rho_m T_c \leq y]$$

$$\cap \{T_{mr} < T_c \leq T_{m+1:1}\}]$$

$$+ \Pr[(\rho_k - \rho_m) T_{kr} \leq y] \cap \{T_r < T_c\}]$$

$$= \int t \cdot \int f_{f_{1:1,r}}(t_1) f_c(t_c) dt_1 dt_c + \sum_{k=1}^{m-1} \int \left( \int f_{f_{1:1,r}}(t_1) f_{f_{k:1,r}}(t_k) dt_1 \cdot \sum_{t_m}^{t_{m+1}} f_c(t_c) dt_c \right)$$

$$\cdot \left( \int t \cdot \int f_{f_{1:1,r}}(t_1) f_{f_{k:1,r}}(t_k) dt_1 \cdot \sum_{t_m}^{t_{m+1}} f_c(t_c) dt_c \right)$$

$$\cdot \left( \int t \cdot \int f_{f_{1:1,r}}(t_1) f_{f_{k:1,r}}(t_k) dt_1 \cdot \sum_{t_m}^{t_{m+1}} f_c(t_c) dt_c \right)$$

$$\cdot \left( \int t \cdot \int f_{f_{1:1,r}}(t_1) f_{f_{k:1,r}}(t_k) dt_1 \cdot \sum_{t_m}^{t_{m+1}} f_c(t_c) dt_c \right)$$

An expression for MRUF can be obtained by conditioning on the same set of disjoint events as

$$E[\tilde{Y}_c]$$

$$= E[\tilde{Y}_c | T_c \leq T_{1:1}] \cdot \Pr[T_c \leq T_{1:1}]$$

$$+ \sum_{k=1}^{m-1} E[\tilde{Y}_c | T_{mr} < T_c \leq T_{m+1:1}] \cdot \Pr[T_{mr} < T_c \leq T_{m+1:1}]$$

$$+ E[\tilde{Y}_r | T_r < T_c] \cdot \Pr[T_r < T_c]$$

$$= E[\rho_k T_c | T_c \leq T_{1:1}] \cdot \Pr[T_c \leq T_{1:1}]$$

$$+ \sum_{k=1}^{m-1} E[\rho_k T_{kr} + \rho_m T_c | T_{mr} < T_c \leq T_{m+1:1}] \cdot \Pr[T_{mr} < T_c \leq T_{m+1:1}]$$

$$+ E[\sum_{k=1}^{m} (\rho_k - \rho_m) T_{kr} | T_r < T_c] \cdot \Pr[T_r < T_c]$$

$$= \rho_0 \int t \cdot \int f_{f_{1:1,r}}(t_1) f_c(t_c) dt_1 dt_c$$

$$\cdot \left( \int t \cdot \int f_{f_{1:1,r}}(t_1) f_c(t_c) dt_1 dt_c \right)$$

$$+ \sum_{k=1}^{m-1} \left( (\rho_k - \rho_m) \cdot \int \left( \int f_{f_{1:1,r}}(t_1) f_c(t_c) dt_1 \cdot \sum_{t_m}^{t_{m+1}} f_c(t_c) dt_c \right) dt_k \cdot \sum_{t_m}^{t_{m+1}} f_c(t_c) dt_c \right)$$

$$+ \rho_m \int \left( \int f_{f_{1:1,r}}(t_1) f_c(t_c) dt_1 \cdot \sum_{t_m}^{t_{m+1}} f_c(t_c) dt_c \right) dt_k \cdot \sum_{t_m}^{t_{m+1}} f_c(t_c) dt_c$$

$$+ \sum_{k=1}^{m-1} \left( (\rho_k - \rho_m) \cdot \int \left( \int f_{f_{1:1,r}}(t_1) f_c(t_c) dt_1 \cdot \sum_{t_m}^{t_{m+1}} f_c(t_c) dt_c \right) dt_k \cdot \sum_{t_m}^{t_{m+1}} f_c(t_c) dt_c \right)$$

$$+ \rho_m \int \left( \int f_{f_{1:1,r}}(t_1) f_c(t_c) dt_1 \cdot \sum_{t_m}^{t_{m+1}} f_c(t_c) dt_c \right) dt_k \cdot \sum_{t_m}^{t_{m+1}} f_c(t_c) dt_c$$

$$= \rho_0 \int t \cdot \int f_{f_{1:1,r}}(t_1) f_c(t_c) dt_1 dt_c$$

$$+ \sum_{k=1}^{m-1} \left( (\rho_k - \rho_m) \cdot \int \left( \int f_{f_{1:1,r}}(t_1) f_c(t_c) dt_1 \cdot \sum_{t_m}^{t_{m+1}} f_c(t_c) dt_c \right) dt_k \cdot \sum_{t_m}^{t_{m+1}} f_c(t_c) dt_c \right)$$

$$+ \rho_m \int \left( \int f_{f_{1:1,r}}(t_1) f_c(t_c) dt_1 \cdot \sum_{t_m}^{t_{m+1}} f_c(t_c) dt_c \right) dt_k \cdot \sum_{t_m}^{t_{m+1}} f_c(t_c) dt_c$$

$$+ \sum_{k=1}^{m-1} \left( (\rho_k - \rho_m) \cdot \int \left( \int f_{f_{1:1,r}}(t_1) f_c(t_c) dt_1 \cdot \sum_{t_m}^{t_{m+1}} f_c(t_c) dt_c \right) dt_k \cdot \sum_{t_m}^{t_{m+1}} f_c(t_c) dt_c \right)$$

$$+ \rho_m \int \left( \int f_{f_{1:1,r}}(t_1) f_c(t_c) dt_1 \cdot \sum_{t_m}^{t_{m+1}} f_c(t_c) dt_c \right) dt_k \cdot \sum_{t_m}^{t_{m+1}} f_c(t_c) dt_c$$

5.5 Over-approximation: Critical failure

Let us recall that we ignored the contribution of a corrupted layer in the lower bound analysis. Oftentimes the cause of this layer corruption is only one or two switch or link faults. The resulting performance deterioration is imperceptible, and one can over-approximate all $\rho_k$ other than $\rho_r$ by $\rho_0$. This in turn results in the following over-approximation for MRUF

$$E[\tilde{Y}_c] \approx \rho_0 \left\{ \int t \cdot \int f_{f_{1:1,r}}(t_1) f_c(t_c) dt_1 dt_c$$

$$+ \sum_{k=1}^{m-1} \int \left( \int f_{f_{1:1,r}}(t_1) f_c(t_c) dt_1 \cdot \sum_{t_m}^{t_{m+1}} f_c(t_c) dt_c \right) dt_k \cdot \sum_{t_m}^{t_{m+1}} f_c(t_c) dt_c \right\}$$

$$+ \int t \cdot \int f_{f_{r:r}}(t) f_c(t_c) dt_c dt_c.$$  \hspace{1cm} (23)

Evaluation of the approximate expression is computationally less complex as it has $O(r)$ dominating terms involving triple integrals, whereas the lower bound has $O(r^2)$ dominating terms with quadruple integrals. Moreover, the reward value and the integral factor in the approximation characterize the performance and the reliability components respectively.

6 Results

It remains now to observe how well the methodology developed here can model the operation of a replicated MIN under arbitrary distribution of component lifetime. Particularly, it is necessary to pay close attention to components with increasing and decreasing
hazard rates and their impact on the system performance. For illustration, we assume Weibull distributed lifetimes for all components,

\[ \Pr[T \leq t] = 1 - e^{-(\lambda t)^\alpha}, \text{MTTF} = E[T] = \frac{\Gamma(1 + \frac{1}{\alpha})}{\lambda}. \]

With \( \alpha = 1.0 \), Weibull reduces to the exponential distribution that corresponds to the Markovian model. In contrast, we study the effect of the increasing and decreasing hazard rates by using two other values of \( \alpha \), viz. 0.8 and 1.2 respectively. However, instead of using the same \( \lambda \) values in those three cases, we use the same values for the component MTTFs. The reason for this choice is that component MTTFs are more direct results of reliability measurement. While MTTFs of components remain unchanged, \( \alpha > 1 \) depicts an increasing hazard rate that causes more faults to occur later as compared to the constant hazard rate (\( \alpha = 1 \)). \( \alpha < 1 \) stands for a decreasing hazard rate that has the opposite effect, i.e. the failures happen earlier. Of course, we assign different MTTFs for different components — 300 days for the links and 1000 days for both the switches and the distributors/collectors. We study a 4-replicated banyan of size 64 \( \times \) 64 implemented with 2 \( \times \) 2 switches, operating at 100 million network cycles per second. In practice, this can reflect a 64 processor 64 module shared memory multiprocessor environment (one of a moderate size), with exactly four distinct paths between any PE-MM pair.

An event-driven simulator has been developed to study the MRUF as the replicated banyan degrades under component failure. A pattern generator creates a fault pattern by randomly choosing the life of the components so that they follow any given distribution. A fault-injector creates the state-trajectory of network degradation under such a specific pattern. The state-space is not the reduced state-space used for analysis, i.e. all possible degraded configurations of the MIN can be in the trajectory. The fault that causes the MIN to break down is always the last fault in the pattern. Each state corresponds to a reward assigned in equation (4). The transition into a new state is an event. At the core, there is a network cycle emulator which is called upon any such event. The emulator runs until the average delay settles to a steady value, and then uses that value to compute the reward rate for that state until the next event. The accumulated reward for each fault pattern is computed, which contributes to the MRUF as a large number of patterns are generated and observed. The number of network cycles simulated on each fault, and the number of random fault patterns can be controlled for the simulation runs. The effect of these two simulation parameters on the convergence of MRUF has been studied in [22]. The simulation results in this paper have been obtained by averaging over 1000 fault patterns, and 1000 network cycles simulated on each fault. The event-driven design is a necessity, as there is a huge order of magnitude difference between the network cycle time and the time between faults (a few nanoseconds vs. a few days).

We have studied the effect of variation in the rate of new request generation on MRUF. The value of resubmission rate \( q \) has been chosen to be unity, i.e. unsuccessful requests are resubmitted in the following network cycle. The value of \( E[Y_{\infty}] \) shows the maximum message transfer capability of the MIN. Along with this we plotted the lower bound \( E[Y_{\infty}] \) from equation (22) and the over-approximation from equation (23). This is shown in figure 2. Table 1 shows the percentage difference of the lower bound and the over-approximation with respect to the simulation results.

The most important observation of all is the high sensitivity of MRUF towards the shape parameter \( \alpha \). An increase of only 0.2 in \( \alpha \) from unity almost triples MRUF, whereas an equal amount of decrease reduces it to less than one-fourth. Markovian approximation can thus be too costly in terms of error margin. The flatness of the curves is a consequence of blocked user model. Variation of resubmission rate also has little effect, so that the analytical upper bound in [20] is a good estimate for mean delay.
<table>
<thead>
<tr>
<th>p</th>
<th>α</th>
<th>0.8</th>
<th>1.0</th>
<th>1.2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>l.b.</td>
<td>app.</td>
<td>l.b.</td>
<td>app.</td>
</tr>
<tr>
<td>0.1</td>
<td>-32.3</td>
<td>5.4</td>
<td>-28.7</td>
<td>4.2</td>
</tr>
<tr>
<td>0.2</td>
<td>-34.6</td>
<td>5.2</td>
<td>-31.0</td>
<td>4.2</td>
</tr>
<tr>
<td>0.3</td>
<td>-32.7</td>
<td>10.0</td>
<td>-29.5</td>
<td>8.1</td>
</tr>
<tr>
<td>0.4</td>
<td>-34.8</td>
<td>7.3</td>
<td>-28.0</td>
<td>11.3</td>
</tr>
<tr>
<td>0.5</td>
<td>-33.4</td>
<td>9.8</td>
<td>-29.0</td>
<td>10.0</td>
</tr>
<tr>
<td>0.6</td>
<td>-31.9</td>
<td>12.2</td>
<td>-28.3</td>
<td>11.1</td>
</tr>
<tr>
<td>0.7</td>
<td>-32.9</td>
<td>10.2</td>
<td>-25.7</td>
<td>15.0</td>
</tr>
<tr>
<td>0.8</td>
<td>-31.3</td>
<td>12.6</td>
<td>-26.6</td>
<td>13.2</td>
</tr>
<tr>
<td>0.9</td>
<td>-30.3</td>
<td>13.9</td>
<td>-24.9</td>
<td>15.6</td>
</tr>
<tr>
<td>1.0</td>
<td>-27.1</td>
<td>18.7</td>
<td>-27.1</td>
<td>12.0</td>
</tr>
</tbody>
</table>

Table 1: Percentage difference in MRUF

7 Conclusions

In this paper, we have proposed a reward structure which shows that the total number of messages routed over a MIN under fault can serve as a performability measure. We have also developed a methodology based on order statistics to analytically evaluate the mean reward for a replicated banyan network. We question the traditional Markovian model of component failure that assumes memorylessness. Our approach does not call for making assumptions about the sojourn time distributions, whether memoryless or otherwise. The key assumption is that of the independence of component lifetime, which is more tangible to a system designer. Lifetime of a component can have any general (known) distribution, possibly to be supplied by the manufacturer. Simulation results obtained using the proposed reward rate justifies our concern showing that perturbation from Markovian lifetime has pronounced effect on the cumulative measure that cannot be ignored.

Performability analysis is complex even for the Markovian model because of an exploding state-space — and it becomes harder when we relax the assumption. The order statistics approach makes use of state-space reduction and stochastic ordering to analyze a conservative bound, that has been compared against the simulation results. Detail knowledge of network operation helps in designing an over-approximation that is computationally much less expensive. Specifically for the replicated MIN, the approximate value decomposes into two distinct performance and reliability factors. However, this behavior cannot be expected of every system in general.

References


273