

## Reliability Analysis of a Fault Tolerant Switch

Luisito Tabada<sup>1</sup> and Pierre Tagle<sup>2</sup>

<sup>1</sup> Northern Mindanao State Institute of Science and Technology, Ampayon, 8600 Butuan City

<sup>2</sup> Ateneo de Manila University, Loyola Heights, 1108 Quezon City, Philippines

**Abstract.** This study proposes an enhancement to the Tagle-Sharma network, a high performance self routing fault tolerant switch fabric which employs an enhanced scheme of the banyan network. The enhancement incorporates a shared-buffer in each switching element of the network to handle evolving traffic in converged networks. As a consequence to the addition of buffer in the system, its complexity is increased. The authors investigate the effect of time dependent reliability, mean time to failure, and steady-state availability as a function of buffer and network sizes of the buffered Tagle-Sharma network. The proposed network is compared with the parallel banyan and benes networks, each is also using the same internal buffering scheme. Numerical results reveal the superiority of this switch architecture than the said established Multistage Interconnection Network (MIN) designs. Reliability models are derived using reliability block diagrams and computational methods.

**Keywords:** multi-stage interconnection network (MIN), Tagle-Sharma Network, reliability block diagram (RBD), internal buffering, shared-buffer

### 1. Introduction

Multistage Interconnection Networks (MIN) possess features such as self-routing, fault tolerance, cost-effectiveness, low transmission delay and modular which are suitable for Very Large Scale Integration (VLSI) implementation. Banyan based networks possess these capabilities for implementing switches for converged networks. However, these networks can be internally blocking which consists of  $n = \log_2 N$  stages for a network size  $N$ . Internal blocking of packets occurs within a switch when more than one packet contends for the same internal resource that can bring severe performance limitations. One of the approach to solve this contention is to employ internal buffering. Buffering schemes that can be applied in these switching elements (SEs) include input, shared and output buffering. Studies have shown that shared buffering scheme outperform the other buffering schemes, since they allow a higher utilization of buffers [7].

For various implementations of MIN switches, reliability analysis is important for justifying the feasibility of the design. It is a crucial requirement in the industry of broadband communications where consequences of the system failures are very expensive. In the literature, Tagle and Sharma [5] have analyzed the buffer less baseline banyan, parallel banyan, Benes and the Tagle-Sharma switching networks using hierarchical composition [11]. Numerical results showed that the Tagle-Sharma network to be better than the other banyan-based switching networks. This study applies the reliability block diagram (RBD) technique to derive expressions of reliability, mean time to failure and steady-state availability. An RBD is a graphical structure with two types of nodes: blocks representing system components and nodes for connections between the components. Edges and nodes model the operational dependency of a system on its components. At any instant of time, if there exist a path in the system from the start node to the end node, then the system is considered operational; otherwise, the system is considered failed. RBD thus maps the operational dependency of a system on its components and not the actual physical structure of the system. Series system is one in which the entire system will fail if any one of its components fail. Parallel system is one that will fail only if all its components fail. In this model-type, the failures of the components are assumed to be independent. Expressions for the time dependent reliability ( $R(t)$ ), the mean time to failure (MTTF), and the availability ( $A(t)$ ) were obtained to find the network's suitability for real time applications. The buffered Tagle-Sharma network is compared with the parallel banyan and Benes networks.

## 2. Network Model

### 2.1. Multistage interconnection network designs

MIN is an intermediate class of networks between crossbar interconnection network, which is scalable in terms of performance but unscalable in terms of cost, and shared bus network, a converse of crossbar. The MIN is more scalable than the bus in terms of performance and more scalable than crossbar in terms of cost [9]. Common MIN designs developed are the baseline banyan, parallel banyan, benes and the omega networks to name a few. The network complexity of the MIN is  $O(N \log_2 N)$ . Survey of different MINs discussed in the succeeding items.

Tagle-Sharma network as shown in Fig. 1a was initially developed for Asynchronous Transfer Mode (ATM) switching. Because of its versatile design, it can be applied to other current broadband or converged networks. Tagle-Sharma Network is a high performance self routing fault tolerant switch fabric which employs an enhanced scheme of the banyan network. It consists of two banyan networks with links provided at every stage to allow packet transfer to and from each banyan plane, thereby offering multiple paths between each input-output pair and giving it a high degree of fault tolerance and overcoming the single path limitation of banyan networks. Study proves that the network offer better performance than other MINs in terms of throughput with or without the presence of faults in the network [5]. Related study on this network reported that increasing the number of planes in the network, the improvements seen using standard transmission are apparent only up to 3 planes, that is, increasing the number of planes above 3 yield no considerable benefit [2].

In this study, the Tagle-Sharma network is compared with the parallel banyan and benes networks. In the parallel banyan switch, a packet has to go through  $n = \log_2 N$  stages before it reaches an output port. The available extra plane gives it a certain degree of fault tolerance. The number of planes can be increased further but would make it unscalable in terms of hardware cost. However, this network is still susceptible to faults that may occur within a particular plane while the packet is traversing the given plane. A Benes switch requires  $(2n - 1)$  stages. Among the important characteristics of this network is that it is only fault tolerant up to stage  $n$ . Once in stage  $n$  a packet has only one path to traverse to reach a particular output line. In fact, stages  $n$  to  $2n - 1$  are actually a baseline banyan network in reverse [5].

### 2.2. Buffering approach

All different applications like voice, video and data, especially in the converged networks, have different requirements in terms of packet loss or throughput and transmission delay. If the level of throughput and delay of these applications are not met, quality of service suffers. One of the limitations of buffer less MIN architectures is that packet losses are evident. Blocking is a problem with which every switch design must deal. Blocking can happen internally when cells contend for the same resource. It is therefore imperative that buffers be installed in the switching network to support those applications. Buffers provide physical storage to hold packets that cannot yet be sent to their desired output ports. Buffers can be located at the input ports, at the output ports, internal to the switch fabric, or any combination. Buffer size and location affects system performance.

Several efforts have been reported to study the performance of MINs with internal buffering [10] [7] [4]. Internal buffering strategy installs buffer inside the switching fabric. Studies have agreed [4] that the shared buffering strategy has the best performance among all other approaches. In addition, shared buffer provides high buffer utilization and needs least amount of buffer space. It is on these reasons that shared buffer strategy is used in this study. Basically, routing of packets is not affected in this buffering scheme. Routing is discussed below.

### 2.3. Routing

Routing of packets in the Tagle-Sharma network still follow the basic algorithm presented in Fig. 1c. The difference is that when the packet enters the input terminal, the packet is then routed to the buffer before it is sent to the desired output terminal of the switching element (SE) depending on its destination address. Assume that network inputs are labeled  $S = s_{n-1} s_{n-2} \dots s_0$  and outputs as  $D = d_{n-1} d_{n-2} \dots d_0$ . The algorithm relies solely on the destination address where for each stage  $i$ , where  $0 \leq i \leq n-1$ , bit  $i$  of the destination address is

examined to determine whether the upper path should be taken (if  $d_i = 0$ ) or the lower (if  $d_i = 1$ ) within a particular plane  $p$ . A routing tag,  $RT$ , is the destination output  $D$  with  $a = 0$  appended to the front, i.e. bit  $n$ . While the packet is traversing plane  $0$  and it encounters a fault, it is deflected to plane 1 to bypass the fault in plane  $0$ . Thus, the cell can bounce to and from different planes to bypass faults. The number of stages traversed in the faulty network is still  $\log_2 N$  since the packets always move forward and the sequence of packets is maintained.

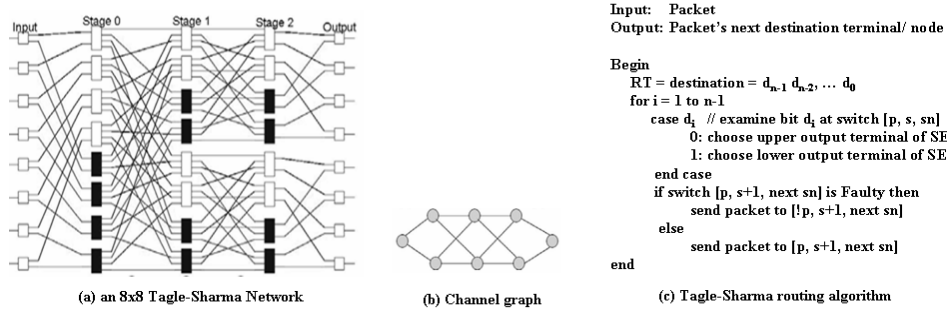


Fig. 1. The Tagle-Sharma switching network.

### 3. Time Dependent Reliability, MTTF and Steady-State Availability

#### 3.1. Buffered switching elements

The output terminals of the buffered SE1x2 can be modelled as two nodes in parallel since a packet has two paths to go. The input terminals of the SE2x1 can be viewed as two components in series as one of these components fails, the SE will fail. Failure means that the SE failed to route the packet from its input terminal to the input terminal of the next SE. The lower bound of SE2x2 which is used by parallel banyan and lower bound of benes networks, can be modelled as a series of all nodes because the packet can traverse in only one path inside the SE. The output terminals of the upper bound SE2x2 can be viewed as two nodes in parallel because the packet can traverse in any of the output terminals. However, its input terminals are logically connected in series since any failure of the input terminals will cause the SE to fail. The packets in the output terminals of SE4x4 can go in two paths, that is, 2 in the upper terminals or 2 in the lower terminals. However, the SE4x4 will fail if both nodes of either upper or lower terminals fail. Thus, the output terminals can be logically connected as a series of two parallel components. The buffer is modelled as a series of  $m$  nodes, where  $m$  is the buffer size, since it follows the first-in-first-out (FIFO) policy. Expressions for the derivation of steady-state availability are based on expressions derived by Blake and Trivedi: equations (1) and (2) are for series RBD ( $A_s$ ) with  $h$  components and the 2-component parallel RBD ( $A_p$ ), respectively. The imperfect coverage  $c$  is the probability that the system successfully reconfigures given that a component fault occurs. Imperfect coverage is important in reliability of MINs since as their network size increases, the number of components increases, and the potential for an uncovered fault occur increases, as well [11]. The study considers the repair rate  $\mu$  and repair transition rate  $\gamma$ .

$$A_s = \frac{\gamma \mu}{\gamma (h \lambda + \mu) + h \lambda \mu (1 - c)} \quad (1)$$

$$A_p = \frac{1 + \frac{2 \lambda}{\mu}}{1 + \frac{2 \lambda}{\mu} + \frac{2 \lambda (1 - c)}{\gamma} + \frac{2 \lambda^2}{\mu^2}} \quad (2)$$

#### 3.2. Buffered switching networks

The RBD of the buffered Tagle-Sharma network is illustrated in Fig. 2a. At its worst case, this network mimics a parallel banyan network as shown in Fig. 2b. Therefore, its lower bound expression will be similar to the parallel banyan network. At any stage of the network, the Tagle-Sharma switch, with the exception of the multiplexer stage, a packet has two SEs to go. Hence, it can be represented as two parallel SEs in series. The variable  $k=(N/2)\log_2 N$  is substituted in equations (3) and (4). Likewise, expressions for reliability, MTTF and steady state availability of the Tagle-Sharma network are derived as presented in equations (3) to (5) where,  $\mu=0$  and  $c=1$ . However, for steady-state availability expression,  $\mu$  is considered but perfect coverage ( $c=1$ ) is still assumed. These assumptions are similar to all the other networks. Any fault in the

SE1x2 and SE2x1 of the parallel banyan will cause the system to fail, which can be represented as  $N$  devices in series, where  $N$  is the network size. Each of the planes can be represented as  $(N/2)\log_2 N$  SE2x2LB in series. Reliability, MTTF and steady state availability expressions for buffered parallel banyan network are presented in equations (6) to (8). Assumptions used for Tagle-Sharma network is also used in this network. The variables  $a=2n(m+3)$  and  $b=(N/2)(\log_2 N)(m+4)$  in equations (6) and (7).

Reliability block diagram of the entire Benes network is complicated. The final half of the network can be viewed as a baseline banyan or parallel banyan in reverse as a single fault in any of these SEs will cause the system to fail. As shown in Fig. 2d, this can be represented as two parallel devices each having  $(N/2)(\log_2 N - 2)$  SEs in series. This network can be considered as the lower bound of the Benes network. In the first half of this network, when a packet enters through these stages, it has two paths to go. This can be considered as two SEs in parallel. But as these SEs are shared among the input-output connections, these can be seen as two parallel SEs connected in series as shown in Fig. 4c. This is the upper bound of the Benes network. Equations (9) to (14) present the reliability, MTTF and steady state availability expressions for buffered Benes network, for both the lower and upper bounds. The variables  $r=(N/2)(1+\log_2 N)(m+3)$ ,  $s=(N/2)(\log_2 N - 2)(m+4)$ , and  $p=(N/2)(\log_2 N - 2)$  are substituted in equations (9) to (13).

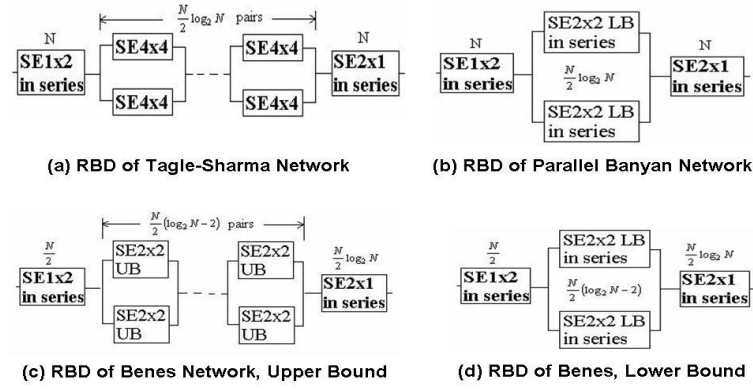


Fig. 2. RBD of the switching networks

$$R_{TS}(t) = \sum_{i=0}^N \binom{N}{i} \sum_{j=0}^k \binom{k}{j} \sum_{v=0}^j \binom{j}{v} \sum_{w=0}^v \binom{v}{w} \sum_{x=0}^{j-v} \binom{j-v}{x} \sum_{y=0}^{k-j} \binom{k-j}{y} \sum_{z=0}^y \binom{y}{z} \sum_{f=0}^{k-j-y} \binom{k-j-y}{f} {}_1(j-w+y-z) {}_2(4k+3f+2v-3w-x-4y) {}_5(j-v-x) e^{-((r-i)\lambda_1 + (m(2k-w-x+y-2z) + 15k-j-2f-2v-6w-7x+9y-16z)\lambda_2)t} \quad (3)$$

$$MTF_{TS} = \frac{\sum_{i=0}^N \binom{N}{i} \sum_{j=0}^k \binom{k}{j} \sum_{v=0}^j \binom{j}{v} \sum_{w=0}^v \binom{v}{w} \sum_{x=0}^{j-v} \binom{j-v}{x} \sum_{y=0}^{k-j} \binom{k-j}{y} \sum_{z=0}^y \binom{y}{z} \sum_{f=0}^{k-j-y} \binom{k-j-y}{f} {}_1(j-w+y-z) {}_2(4k+3f+2v-3w-x-4y) {}_5(j-v-x)}{(r-i)\lambda_1 + (m(2k-w-x+y-2z) + 15k-j-2f-2v-6w-7x+9y-16z)\lambda_2} \quad (4)$$

$$A_{TS} = \left( \frac{\mu}{\gamma((m+1)\lambda_1 + \mu)} * \frac{1 + \frac{2\lambda_1}{\mu}}{1 + \frac{2\lambda_1}{\mu} + \frac{2\lambda_1^2}{\mu^2}} * \frac{\mu}{\gamma((m+3)\lambda_1 + \mu)} \right)^N \times \left( 2 * \left( \frac{\mu}{\gamma((m+4)\lambda_4 + \mu)} \left( \frac{1 + \frac{2\lambda_4}{\mu}}{1 + \frac{2\lambda_4}{\mu} + \frac{2\lambda_4^2}{\mu^2}} \right) - \left( \frac{\mu}{\gamma((m+4)\lambda_4 + \mu)} \left( \frac{1 + \frac{2\lambda_4}{\mu}}{1 + \frac{2\lambda_4}{\mu} + \frac{2\lambda_4^2}{\mu^2}} \right) \right)^2 \right)^{\frac{N}{2} \log_2 N} \quad (5)$$

$$R_{PB}(t) = \sum_i \binom{N}{i} (-1)^{N-i} (2)^i (2e^{-((a-i)\lambda_1 + b\lambda_2)t} - e^{-((a-i)\lambda_1 + 2b\lambda_2)t}) \quad (6)$$

$$MTF_{PB} = \sum_i \binom{N}{i} (-1)^{N-i} (2)^i \left( \frac{2}{(a-i)\lambda_1 + b\lambda_2} - \frac{1}{(a-i)\lambda_1 + 2b\lambda_2} \right) \quad (7)$$

$$A_{PB} = \left( \frac{\mu}{\gamma((m+1)\lambda_1 + \mu)} * \frac{1 + \frac{2\lambda_1}{\mu}}{1 + \frac{2\lambda_1}{\mu} + \frac{2\lambda_1^2}{\mu^2}} * \frac{\mu}{\gamma((m+3)\lambda_1 + \mu)} \right)^N \times \left( 2 * \left( \frac{\mu}{\gamma((m+4)\lambda_2 + \mu)} \right)^{\frac{N}{2} \log_2 N} - \left( \frac{\mu}{\gamma((m+4)\lambda_2 + \mu)} \right)^{N \log_2 N} \right) \quad (8)$$

$$R_{BLB}(t) = \sum_i \binom{N}{i} (-1)^{N-i} (2)^i (2e^{-((r-i)\lambda_1 + s\lambda_2)t} - e^{-((r-i)\lambda_1 + 2s\lambda_2)t}) \quad (9)$$

$$MTTF_{BLB} = \sum_i^N \binom{N}{i} (-1)^{N-i} (2)^i \left( \frac{2}{(r-i)\lambda_1 + s\lambda_2} - \frac{1}{(r-i)\lambda_1 + 2s\lambda_2} \right) \quad (10)$$

$$A_{BLB} = \left[ \frac{\gamma\mu}{\gamma(m+1)\lambda_1 + \mu} \times \frac{1 + \frac{2\lambda_1}{\mu}}{1 + \frac{2\lambda_1}{\mu} + \frac{2\lambda_1^2}{\mu^2}} \right]^{\frac{N}{2}} \times \left( \frac{\gamma\mu}{\gamma(m+3)\lambda_1 + \mu} \right)^{\frac{N}{2} \log_2 N} \times \left( 2 * \left( \frac{\gamma\mu}{\gamma(m+4)\lambda_2 + \mu} \right)^{\frac{N}{2} \log_2 N - 2} - \left( \frac{\gamma\mu}{\gamma(m+4)\lambda_2 + \mu} \right)^{N \log_2 N - 2} \right) \quad (11)$$

$$R_{BUB}(t) = \sum_{i=0}^N \binom{N}{i} \sum_{j=0}^p \binom{p}{j} \sum_{v=0}^j \binom{j}{v} \sum_{w=0}^v \binom{v}{w} \sum_{x=0}^{p-j} \binom{p-j}{x} (-1)^{\frac{N}{2} - i - j + p + v} 2^{i+2j+x} e^{-((r-i)\lambda_1 + (m(v-2w+2p-x) + 2v - 6w + 8p - j - 4x)\lambda_2)t} \quad (12)$$

$$MTF_{BUB} = \sum_{i=0}^N \binom{N}{i} \sum_{j=0}^p \binom{p}{j} \sum_{v=0}^j \binom{j}{v} \sum_{w=0}^v \binom{v}{w} \sum_{x=0}^{p-j} \binom{p-j}{x} (-1)^{\frac{N}{2} - i - j + p + v} 2^{i+2j+x} \left( \frac{1}{(r-i)\lambda_1 + (m(v-2w+2p-x) + 2v - 6w + 8p - j - 4x)\lambda_2} \right) \quad (13)$$

$$A_{BUB} = \left[ \frac{\gamma\mu}{\gamma(m+1)\lambda_1 + \mu} \times \frac{1 + \frac{2\lambda_1}{\mu}}{1 + \frac{2\lambda_1}{\mu} + \frac{2\lambda_1^2}{\mu^2}} \right]^{\frac{N}{2}} \times \left( \frac{\gamma\mu}{\gamma(m+3)\lambda_1 + \mu} \right)^{\frac{N}{2} \log_2 N} \times \left( 2 * \left[ \left( \frac{\gamma\mu}{\gamma(m+2)\lambda_2 + \mu} \times \frac{1 + \frac{2\lambda_2}{\mu}}{1 + \frac{2\lambda_2}{\mu} + \frac{2\lambda_2^2}{\mu^2}} \right) - \left( \frac{\gamma\mu}{\gamma(m+2)\lambda_2 + \mu} \times \frac{1 + \frac{2\lambda_2}{\mu}}{1 + \frac{2\lambda_2}{\mu} + \frac{2\lambda_2^2}{\mu^2}} \right)^2 \right]^{\frac{N}{2} (\log_2 N - 2)} \right) \quad (14)$$

## 4. Numerical Results

Time dependent reliability for Tagle-Sharma, parallel banyan, and Benes (both upper bound and lower bound) networks is shown in Fig. 3. Perfect coverage and no repair are assumed because the purpose is to show the effect of fault tolerance as a function of buffer and network sizes. A failure rate of  $\lambda_2=10^{-6}$ ,  $\lambda_1=\lambda_2/2$ ,  $\lambda_4=3\lambda_2$  and repair rate  $\mu=10^4\lambda_2$ . Fig. 3a shows the effect of reliability as a function of time at network size = 256 with buffer size  $m=4$ . It can be observed from Fig. 3b that for small networks, Tagle-Sharma network demonstrates lower reliability because it has more components than the other networks. However, as the network size becomes larger, it demonstrates higher reliability. This is because the Tagle-Sharma switch provides fault tolerance at every stage of the network. Tagle-Sharma still provides higher reliability than the other networks. Fig. 3c presents the effect of reliability as a function of buffer size at network size = 32 and unit time = 1000. The reliability of Tagle-Sharma decreases as buffer size and network size is increased. This is also evident with other networks.

Likewise, expressions for MTTF assume perfect coverage and no repair for the same reason as the goal is to show the effect of fault tolerance as a function of buffer size. Same value for failure rates are used to calculate for the MTTF of each network. Table 1 presents the effect of MTTF as a function of buffer size. Large gap in the values are observed. This is because of the fault tolerance feature of the Tagle-Sharma switch. Benes network also show higher MTTF than the parallel banyan since the former is fault tolerant at the first half of the network while the latter is only fault tolerant at the first stage of the network.

Steady-state availabilities of the networks as a function of network size and buffer size are shown in Fig. 3d and 3e, respectively. Perfect coverage ( $c=1$ ) was assumed and repair is permitted with a repair transition rate  $\gamma = \mu/10$ . The same failure rates are used. For small network size ( $N=8$ ), Tagle-Sharma depicts smaller availability for increasing buffer size than the other networks. However, the network is increased, it demonstrates a higher availability. In all network designs, it is also observed that buffer size and network size have diminishing benefits as their sizes are increased. Furthermore, it is observed that reliability and availability decrease fast as buffer size is increased.

## 5. Conclusion

This paper deals with the derivation of expressions for the time dependent reliability, mean time to failure and steady-state availability of the buffered switched MIN architectures in order to establish its use to real time applications. Even with the addition of the shared-buffer in each switching element, the Tagle-Sharma network is found to be reliable, robust, and has promising application prospects. The buffered approach to the Tagle-Sharma network has been compared with similar buffered expressions of the parallel banyan and Benes networks.

Reliability analysis reveal that Tagle-Sharma switch is better than other MINs in terms of time dependent reliability, MTTF and availability. This is because the switch provides fault tolerance at every stage of the network. Benes network only offers fault tolerance up to the middle stage of the network while the Parallel Banyan give fault tolerance at the initial stage only. Reliability, MTTF and availability decrease as network size and buffer size increase. Moreover, reliability and availability decrease sharply as buffer size is increased.

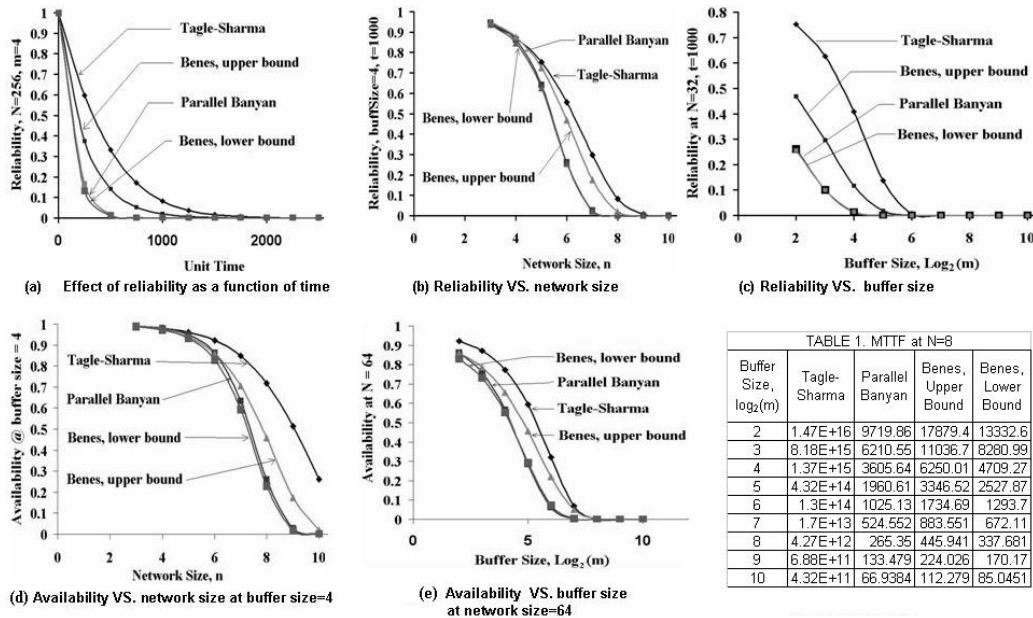


Fig. 3. Reliability analysis of buffered switch architectures.

## 6. References

- [1] R. Mahajan and R. Vig. Performance and Reliability Analysis of New Fault-Tolerant Advance Omega Network. WSEAS Transactions on Computers. ISSN: 1109-2750, Issue 8, Volume 7, August 2008.
- [2] C. Lagman and P.U. Tagle. Aggressive Transmission in the Interconnected Parallel Banyan Network. IEEE Computer Society – ISPAN. 02. 2002.
- [3] R. V. Boppana and C. Raghavendra. Designing efficient benes and banyan based input-buffered atm switches. ICC, Vancouver, B.C., Canada, 1999.
- [4] Y. Sayed. Performance analysis, Design and Reliability of Balanced Gamma Network. Ph.D. Thesis, Memorial University of Newfoundland, pp. 0-198, 1999.
- [5] P.U. Tagle and N.K. Sharma. Performance of fault tolerant atm switches. IEE Proceedings – Communication, Vol. 143, No. 5, Oct. 1996.
- [6] D. Patterson and J. Hannessy. Computer architecture: A quantitative approach. Morgan Kauffmann Publishers, Inc., San Francisco California, Second Edition, pp. 521-522, 1996.
- [7] S.L. Ng and B. Dewar. A fault tolerant load sharing replicated buffered banyan network. IEEE. 1995.
- [8] B. Zhou and M. Atiqzaman. Accurate analysis of multistage interconnection networks using finite output-buffered switching elements. IEEE, 1995.
- [9] V. Kumar, A. Grama, A. Gupta and G. Karypis. Introduction to parallel computing: Design and analysis of algorithm. The Benjamin Cummings Publishing Company, Inc., pp. 24-30, 1994.
- [10] Y. Mun and H.Y. Youn. Performance analysis of finite buffering multistage interconnection networks. IEEE Transactions on Computers. Vol. 43, No. 2, pp. 153-162, Feb. 1994.
- [11] J. Blake and K.S. Trivedi. Reliability analysis of interconnected networks using hierarchical composition. IEEE transactions on reliability, Vol. 38, No. 1, Apr. 1989.
- [12] K.S. Trivedi. Probability and statistics with reliability, queuing, and computer science applications. Prentice-Hall, Inc., Englewood Cliffs, NJ, USA, 1982.