Design Scheme and Performance Evaluation of a new Fault-tolerant Multistage Interconnection Network

Karamjit Kaur Cheema, Rinkle Aggarwal

Department of Computer Sc. & Engineering
Thapar University, Patiala, Punjab, 147004 (India)

Summary
The effectiveness of a parallel or distributed system is often determined by its communication network. In order to operate more efficiently a network is required to provide low latency and be able to handle large amount of traffic. This paper introduces a new fault-tolerant multistage interconnection networks (MIN) named as Modified Augmented Baseline Network (MABN). Performance of the proposed network is analysed in terms of permutation passibility, reliability and cost. The performance comparison of the MABN with ABN shows the significant improvement of MABN over existing Augmented Baseline Network (ABN).

Key words:
Fault-tolerance, multistage interconnection networks, reliability, performance evaluation, baseline network, augmented baseline networks, mod.

1. Introduction

A number of tachniques have been proposed to oncrease the reliability and falt-tolereance of MINs, a survey of the fault-tolerant attributes of these networks can be found in [1]. The modest cost of unique path MINs makes them attractive for large multi-processor systems, but their lack of fault-tolerance is a major drawback. To mitigate this problem, three hardware options are available : replicate the entire network, add extra stages, and/or add additional links.

The general goals for the design of fault-tolerant MINs are high reliability, good performance even in presence of faults, low cost and simple control. However, most fault-tolerant MINs proposed in the literature cannot achieve all of these goals at the same time. Some of the networks fail to tolerate faults in the first and/or last stages. Some others can tolerate faults in any stage, but they are, in general, too costly.

In this paper, we present methods of increasing fault-tolerance of an network by increasing the size of de-multiplexers. Hence have doubled the number of paths available between each source and destination, as compared to existing network ABN. The proposed Modified Augmented Baseline Network(MABN) is an augmented baseline network(ABN) [4] with increased size of de-multiplexers. In an MABN, there are four possible paths between any source-destination pair, whereas ABN has only two. As we will see, MABNs can achieve general goals for the design of fault-tolerant networks i.e. hign reliability, good performance even in presence of faults, simple control.

In the following section structure and topology of existing network ABN and proposed network MABN is described. The routing procedure of MABN is given in section 3. Performance of MABN for various parameters is explained in section 4. Finally, some concluding remarks are given in section 5.

2. Structure and topology of Networks

2.1 Constuction of ABNs

To construct an ABN of size N i.e. N sources and N destinations, two identical groups of N/2 sources and N/2 destinations need to be formed first. Each group consists of a multiple path modified baseline network of size N/2 [4]. The modified baseline network is a baseline network with one less stage and feature links among switches belonging to the same stage and forming several loops of switches. The switches in the last stage are of size 2 x 2 and the remaining switches in stages 1 through n-3(n=log2N) are of size 3 x 3. In each stage, the switches can be grouped into conjugate subsets, where a conjugate subset is composed of all switches in a particular stage that lead to the same subset of destinations. Switches which communicate through the use of auxiliary links are called a conjugate loop. The conjugate loops are formed in such a way that the two switches which form a loop have their respective conjugate switches in a different loop. These pair of loops is called conjugate loops.

A redundancy graph offers a convenient way to study the properties of a multi-path MIN, such as the number of faults tolerated or the type of rerouting possible [6]. A redundancy graph depicts all the available paths between a source and a destination in a MIN. It consists of two distinguished nodes-the source S and the destination D-and

Manuscript received September 5, 2009
Manuscript revised September 20, 2009

the rest of the nodes correspond to the switches that lie along the paths between S and D. An ABN of size 16X16 and the redundancy graph of ABn is shown in figure 1.

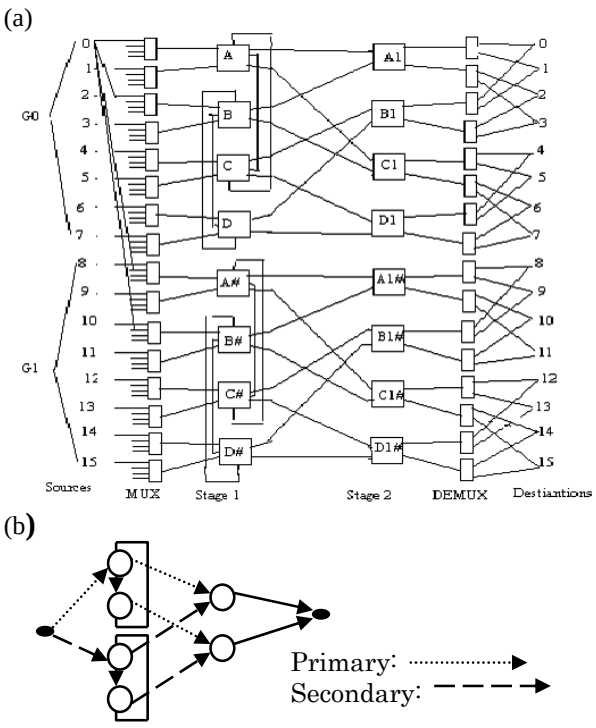


Fig. 1. (a) An ABN of size 16 X 16. (b) The redundancy graph.

2.2 Construction of MABNs

To construct an MABN of size N i.e. N sources and N destinations, two identical groups of N/2 sources and N/2 destinations need to be formed first. Each group consists of a multiple path modified baseline network of size N/2. Each source is linked to both the groups via multiplexers. There is one 4 x 1 MUX for each input link of a switch in stage 1 and one 1 x 4 DEMUX for each output link of a switch in stage n-2. Each group consisting of a modified baseline network of size N/2 plus its associated MUXs and DEMUXs is called a sub-network. Thus an MABN consists of two identical sub-networks which are denoted by G^i. For example, in Figure 1, switches A, B, C, D belonging to stage 1 of a sub-network (G^i) form a conjugate subset, switches A and B form a conjugate pair, and switches A and C form a conjugate loop.

A source selects a particular sub-network (G^i) based upon the most significant bit of the destination. As there are four available paths between a source-destination pair, so each source is connected to two switches (primary and secondary) in primary sub-network, and to two switches (primary# and secondary#) in secondary sub-network. The sources are connected to the switches of stage 1 as follows:

Let the source S and destination D be represented in binary code as:

$S=s_0, s_1,\ldots,s_{n-2},s_{n-1}$

$D=d_0, d_1,\ldots,d_{n-2},d_{n-1}$

(i) Source S is connected to the ($s_1,\ldots,s_{n-2}$) primary switch in both the sub-networks through the multiplexers.

(ii) Source S is also connected to the $[\{(s_1,\ldots,s_{n-2})+1\} \bmod N/4]$ secondary switch in both the sub-networks through the multiplexers.

Thus an MABN of size N consists of N 4 x 1 MUXs, N 1 x 4 DEMUXs, and n-2 stages of N/2 switches each (the last stage has 2 x 2 switches and the remaining stages have 3 x 3 switches). An MABN of size 16X16 is shown in figure 2(a).

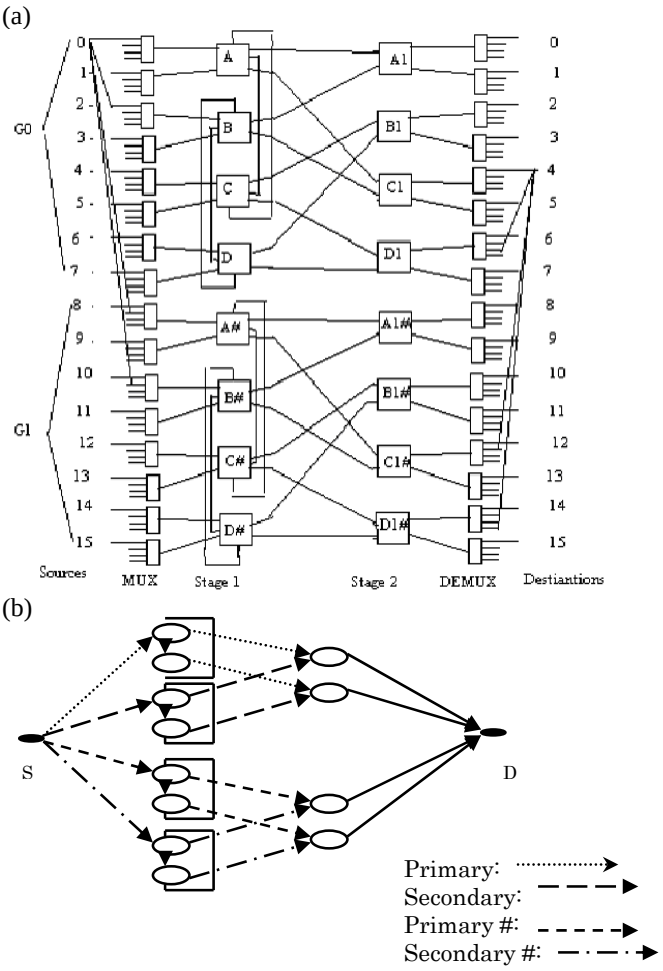


Fig. 2. (a) An MABN of size 16 X 16. (b) The redundancy graph.

Consider the redundancy graph of MABN as shown in Figure 2(b). Source is connected to four switches in the

network, two in each sub-network; which are in further attached to remaining switches of stage 1 via auxiliary links. Every request tries first the primary sub-network, within it the primary path then the secondary path is tried. If both fail then the same procedure is applied with secondary sub-network. If a switch becomes faulty, then the loop containing the faulty switch can be removed from the network and a replacement loop inserted. It is necessary to have a procedure for gracefully terminating the connections using the non-faulty switches in the loop before removing the loop. This all is possible only due to the presence of conjugate loops.

3. Routing Scheme

The routing scheme of MABN in the case that each source-destination pair tries to utilize only one path at a time is described below. This scheme assumes that sources and switches have the ability to detect faults in the switches to which they are connected. Several techniques of detecting faults have been discovered. A request from any source S to a given destination D is routed through the MABN as:

1. **For each source:**

The source S selects one of the sub-network G^i based on the most significant bit of the destination D ($i=d_0$).There are two parts, i.e. Primary and Secondary, between each source-destination pair in each sub-network. Each source attempts entry into the MABN via its primary path. If the primary path is faulty (i.e. either MUX or primary switch or both are faulty), then the request is routed to secondary path. If the secondary path is also faulty then the request is routed to the other sub-network of the MABN. Again, in the same sub-network source will attempt entry via primary# path. If the primary# path is faulty (i.e. either MUX or primary switch or both are faulty), then the request is routed to secondary# path. If the secondary# path is also faulty then the MABN fails.

2. **For each switch in stage n - 3**:(Requests for connection may arrive on any of the three input links.)

After the MUX, the routing of the request in the first (n-3) stage of the sub-network depends upon one tag bit, which depends on d_1d_2 destination address bits. Routing tag bit for stage 1 is calculated as follows:

If d1d2 = 00, then both conjugate pairs in the sub-network will have tag bit = 0.

If d1d2 = 01, then first conjugate pair (A/A, B/B#) will have tag bit = 1, and second conjugate pair(C/C#, D/D#) will have tag bit = 0.

If d1d2 = 10, then both conjugate pairs in the sub-network will have tag bit = 1.

If d1d2 = 11, then first conjugate pair(A/A#, B/B#) will have tag bit = 0,and second conjugate pair(C/C#, D/D#) will have tag bit = 1.

Use tag bit and route the request through the usual output link, if it is busy or if the successor switch (in the next stage) is faulty, route the request via the auxiliary output links to the other switch in the loop with the same tag bit .

If the auxiliary link is also unusable because it is busy or because of a fault, then try secondary path. If secondary path also have some fault, then try using auxiliary links. If all the possible paths in primary sub-network fail, then use the same tag bit and procedure stated above in secondary sub-network. If all the possible paths in secondary sub-network also fail, then drop the request.

3. **For each switch in stage n - 2**: (Requests for connection may arrive on any of the two input links.)

For a request at a switch in stage n-2, use bit d_{n-1} of the routing tag and route the request accordingly to one of the output links. If the required output link is busy, then repeat step two and three in the secondary sub-network. If again the required output link is busy in stage n-2, then drop the request.

4. **For each de-multiplexer at the output of stage n - 2:** (May receive a maximum of one request)

For routing a request through a DEMUX, following concept is used.

If destination and MUX are in same sub-network, then 1st MUX uses output line 00 and 2nd MUX uses output line 10.

If destination and MUX are in different sub-networks, then 1st MUX uses output line 01 and 2nd MUX uses output line 11.

A faulty DEMUX at the output of the MABN is regarded as a failure of its associated switch in stage n-2. This strategy essentially enables a switch to detect a failure of its successor switch and re-routes the request whenever possible.

5. **For each destination**: (Up to two requests may arrive) Accept request.

Multiple paths between S=0000 and D=0100 of an MABN are highlighted in Figure 3. For connections between S=0000 and D=0100, switch A and switch B of stage 1 belonging to sub-network G^0 act as the primary and the secondary switches respectively. The paths connecting, source and the primary switch is named as the primary path and the source and the secondary switch is named as the secondary path Switches A# and switch B# of stage 1 belonging to sub-network G^1 act as the primary# and the secondary# switches respectively. The paths connecting, source and the primary# switch is named as the primary#

path and the source and the secondary# switch is named as the secondary# path. As can be seen in the Figure 3(b), MABN has doubled the number of paths available in ABN. Availability of these multiple paths has hence increased fault-tolerance significantly.

A key issue in MABN is the manner in which rerouting or selection of alternate paths is achieved. The topology of a multi-path MABN allows rerouting to be done only at the source in the network. In MABN, a busy link, a faulty link, or a faulty switch encountered while setting up a path may necessitate backtracking to a stage 1 and then attempts to set up an alternate path from there. Backtracking MINs tend to have less hardware complexity than non-backtracking ones. But backtracking MINs are inconvenient to implement, since they may require bidirectional paths and reverse queues. Also, if backtracking is to be avoided in these MINs for performance enhancement, it becomes necessary to have global fault information.

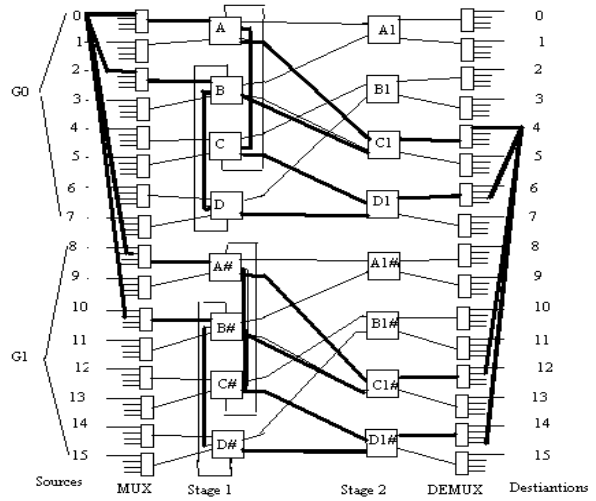


Figure 3: Routing in MABN

All available paths from S=0000 (0) to D=0100 (4) in MABN are as follows:

Primary path:
0->MUX (0) ->A->C1->DEMUX (4) ->4
0->MUX (0) ->A->C->D1->DEMUX (6) ->4

Secondary path:
0->MUX (2) ->B->C1->DEMUX (4) ->4
0->MUX (2) ->B->D->D1->DEMUX (6) ->4

Primary# path:
0->MUX (8) ->A#->C1#->DEMUX (12) ->4
0->MUX (8) ->A#->C#->D1#->DEMUX (14) ->4

Secondary# path:
0->MUX (10) ->B#->C1#->DEMUX (12) ->4
0->MUX (10) ->B#->D#->D1#->DEMUX (14) ->4

4. Perfomance analysis

Many performance parameters are applicable for MINs. Some of the important performance parameters are permutation passibility, reliability and cost. In this paper, proposed network MABN have been analyzed on the above said three parameters.

4.1 Permutation passibilty :

Permutation passibility is the measure which tells how many number of requests appearing at the source side has got successfully matured i.e. have reached the respective destinations successfully. Further both the cases have been considered when there is no faulty switch in the network and when there are faulty switches present in the network.

4.1.1 ABN

Results of permutation passibility analysis done for existing regular network ABN gave us following results:

Total number of request appearing at source side =36
Total requests matured when no switch is failed =33
Total requests matured when switch is failed =24
Total path length when no switch is failed =73
Total path length when switch is failed =53
Average path length when no switch fails =73/33 = 2.21
Average path length when switch failed =53/24 = 2.208

4.1.2 MABN

Results of permutation passibility analysis done for proposed regular network MABN gave following results:

Total number of request appearing at source side =36
Total requests matured when no switch is failed =35
Total requests matured when switch is failed =33
Total path length when no switch is failed =80
Total path length when switch is failed =77
Average path length when no switch is failed
= 80/35 = 2.29
Average path length when switch is failed =77/33 = 2.33

From the data given above, it can be concluded that average path length of proposed network MABN is greater than ABN. But there is significant improvement in number of requests successfully maturing at the destination side in

case of MABN, in both the cases i.e. in presence and absence of faults. Main consideration in permutation passibility is how many requests get matured in presence of faults, the proposed regular network MABN gives better performance in this respect.

4.2 Reliablity

In this section the reliability of MABNs in terms of MTTF is analyzed. Mean Time to Failure (MTTF) is a well known criterion to measure reliability of fault-tolerant networks having full access [4]. Under this criterion, a network is faulty if there is any source-destination pair that cannot be connected because of faulty components in the network. MTTF of the network is defined as the expected time elapsed before some source is disconnected from some destination.Reliability equations of proposed network MABN derived for both lower bound and upper bound in terms of MTTF (Mean Time To Failure) are given below:

4.2.1 Upper bound (optimistic): To obtain an upper bound for the MABN, we observe that each source is connected to two multiplexers in each sub-network, and each switch has a conjugate. So if we assume that the MABN is operational as long as one of the two multiplexers attached to a source (in a particular sub-network) is operational and as long as a conjugate pair (loop or switch) is not faulty, then we will permit as many as one half of the components to fail and the MABN may still be operational. This permits a simple reliability block diagram of the optimistic (upper) bound as shown in Figure 4.

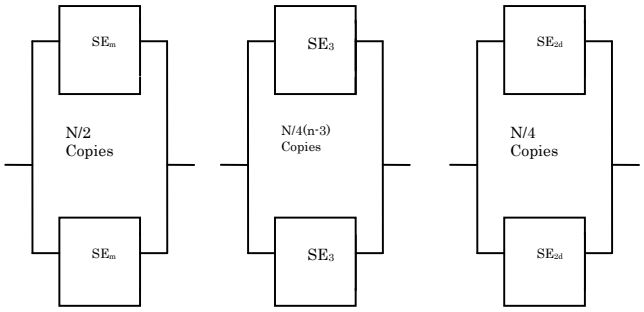


Figure 4: Reliability block diagram of MABN for MTTF upper bound.

The expression for the upper bound of the MABN reliability is:

$R_{MABN\text{-}ub}(t)$= f1 * f2 * f3

$$f1 = \left[1 - \left(1 - e^{-\lambda_m t}\right)^2\right]^{\left(N/2\right)}$$

$$f2 = \left[1 - \left(1 - e^{-\lambda_3 t}\right)^2\right]^{N/4(n-3)}$$

$$f3 = \left[1 - \left(1 - e^{-\lambda_{2d} t}\right)^2\right]^{\left(N/4\right)}$$

Where,

$\lambda_m=\lambda$, $\lambda_3=2.25\lambda$, $\lambda_{2d}=3\lambda$

$$MTTF_{MABN\text{-}ub} = \int_0^\infty R_{MABN\text{-}ub}(t).dt$$

Only difference in upper bound formula of ABN and MABN is the value of λ_{2d}, for ABN its value is 2λ and for MABN its 3λ. Difference in this value is due to the reason that for ABN λ_{2d} means one 2 x 2 switch and two 1 x 2 de-multiplexers ($\lambda+\lambda=2\lambda$), whereas for MABN λ_{2d} means one 2 x 2 switch and two 1 x 4 de-multiplexers ($\lambda+2\lambda=3\lambda$).

4.2.2 Lower bound (pessimistic): At the input side of the MABN, the routing scheme does not consider the multiplexers to be an integral part of a 3 x 3 switch. For example, as long as at least one of the two multiplexers attached to a particular switch is operational, the switch can still be used for routing. Hence, if we group two multiplexers with each switch in the input side and consider them a series system (SE3m), then we will have a conservative estimate of the reliability of these components. Their aggregate failure rate will be $\lambda 3m = 4.25\lambda$. Finally these aggregated components and the switches in the intermediate stages can be arranged in pairs of conjugate loops. To obtain the pessimistic (lower) bound on the reliability of MABN, we assume that the network is failed whenever more than one conjugate loop has a faulty element or more than one conjugate switch in the last stage fails. The reliability block diagram is shown in Figure 5.

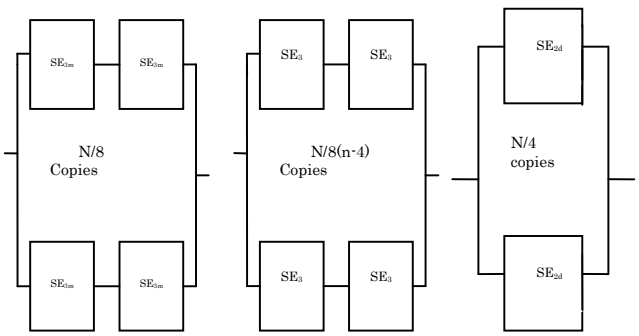


Figure 5: Reliability block diagram of MABN for MTTF lower bound

$R_{MABN\text{-}lb}(t)$= f1 * f2 * f3

$$f1 = \left[1 - \left(1 - e^{-2\lambda_{3m}t}\right)^2\right]^{(N/8)}$$
$$f2 = \left[1 - \left(1 - e^{-2\lambda_{3}t}\right)^2\right]^{N/8(n-4)}$$
$$f3 = \left[1 - \left(1 - e^{-\lambda_{2d}t}\right)^2\right]^{(N/4)}$$

Where,
$\lambda_{3m}=4.25\lambda$, $\lambda_3=2.25\lambda$, $\lambda_{2d}=3\lambda$

$$MTTF_{MABN\text{-}lb} = \int_0^\infty R_{MABN\text{-}lb}(t).dt$$

Again the only difference in lower bound formula of ABN and MABN is the value of λ_{2d}, for ABN its value is 2λ and for MABN its 3λ, due to the same reason stated above.

Table 1: MTTF Vs Log N

LogN	ABN_LB	ABN_UB	MABN_LB	MABN_UB
4	4.934369	5.141202	4.953677	5.184155
5	4.717386	4.923061	4.733636	4.959739
6	4.508375	4.71246	4.522473	4.74445
7	4.30494	4.507272	4.317438	4.535628
8	4.105551	4.306118	4.116806	4.331575
9	3.909194	4.108067	3.919452	4.131159
10	3.71518	3.912467	3.724616	3.933593

4.3 Cost-effectiveness

We can observe that ABNs and its variant (MABN) as proposed in this paper can provide higher or at least equal reliability compared to some other fault-tolerant networks. However, if such high reliability comes at the expense of high cost, it may have little value in practice. This section concerns the cost-effectiveness of ABN and the proposed network MABN.

To estimate the cost of a network, one common method is to calculate the switch complexity with an assumption that the cost of a switch is proportional to the number of gates involved, which is roughly proportional to the number of cross-points within a switch [3]. For example a 2 x 2 switch has four units of hardware cost, whereas a 3 x 3 switch has nine units. For the multiplexers and de-multiplexers, we roughly assume that each of m x 1 multiplexers or 1 x m de-multiplexers has m units of cost. Thus an ABN has the cost of $N/2(3\log_2 N+13)$ [4]. Similarly, MABN cost can be found.

Table 2: Cost Functions

MIN	Cost
ABN	$N/2(3\log_2 N+13)$
MABN	$N/2(5\log_2 N+9)$

Table 3: Log Cost Vs Log N

Log N	ABN	MABN	IABN
4	2.30103	2.365488	2.428135
5	2.651278	2.735599	2.770115
6	2.996512	3.096215	3.114611
7	3.337659	3.449633	3.459242
8	3.675412	3.797406	3.802363
9	4.0103	4.140634	4.143171
10	4.342738	4.480122	4.481414

Now a simple measure of the cost-effectiveness for reliability can be given by comparing MTTF and the cost of the network. Let the cost-effectiveness, η of a network for reliability be the ratio of MTTF to its cost [4, 8].

Table 4: MTTF per unit log Cost Vs log N

LogN	ABN_LB	ABN_UB	MABN_LB	MABN_UB
4	2.152809	2.252971	2.333451	2.408275
5	1.785417	1.870697	1.962925	2.00602
6	1.509246	1.583324	1.677055	1.702432
7	1.293553	1.358925	1.450879	1.465078
8	1.120094	1.178528	1.26796	1.274538
9	0.977346	1.030137	1.117079	1.1182
10	0.857665	0.905786	0.99044	0.987526

5. Conclusion

In this paper, we proposed and analyzed a new fault-tolerant multi-stage interconnection networks named as Modified Baseline Network (MABN), which has achieved significant tolerance to faults and good performance with relatively low cost and simple control scheme.

The switch-fault model is used to analyze the reliability of MABNs. In our analysis, any switch, any multiplexer and any de-multiplexer in MABNs are assumed to have a possibility to fail. The analysis of the lower and upper bounds of MTTF showed that MABN is having better reliability than other related fault-tolerant

networks. However, if such reliability comes at the expense of high cost, it may have little value in practice. Our analysis on the cost of networks showed that MABNs is generally, more cost-effective than ABN.

The permutation passibility analysis shows that both in presence and absence of faults, number of requests maturing are more in MABN than ABN.

References

[1] Adams George B., Agrawal Dharma P. and Siegel Howard Jay, "A Survey and Comparison of Fault-Tolerant Interconnection Networks", IEEE Transactions on Computers, June 1987, pp. 14-27.

[2] Adams George B. and Siegel Howard Jay, "The Extra Stage Cube: A Fault-Tolerant Interconnection Network for Super systems", IEEE Transactions on Computers, vol. c-31, no. 5, May 1982, pp. 443-454.

[3] Bansal P.K, Singh Kuldeep and Joshi R.C., "Quad Tree: A Cost-Effective Fault-Tolerant Multistage Interconnection network**",** Proceeding of International Conference IEEE INFOCOM, 1992, pp. 6D.1.1-6D.1.7.

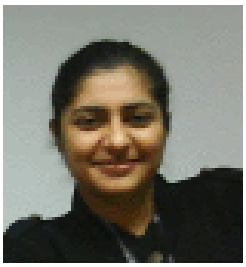
[4] Bansal P.K, Singh Kuldeep and Joshi R.C, " On Fault tolerant Multistage Interconnection Network", Conference on Computer Electrical Engineering, vol. 20, no.4, 1994, pp. 335-345.

[5] Bansal P.K, Singh Kuldeep and Joshi R.C, "Routing and path length algorithm for a cost-effective four-tree multistage interconnection network" International Journal of Electronics, vol. 73,no.1,1992, pp. 107-115

[6] Bhogavilli Suresh K. and Abu-Amara Hosame, "Design and Analysis of High Performance Multistage Interconnection Networks", IEEE Transactions on Computers, vol. 46, no. 1, January 1997, pp. 110 -117.

[7] Bhuyan Laxmi N., Yang Qing and Aggarwal P. Dharma, "Performance of Multiprocessor Interconnection Networks", Proceeding of IEEE, February 1989, pp. 25-37.

[8] Sadawarti Harsh and Bansal P.K., " Fault Tolerant Irregular Augmented Shuffle Network", Proceeding of the 2007 WSEAS International Conference on Computer Engineering and Applications, Australia, January 17-19,2007. pp. 7-12.



Er. Karamjit KkKaramjit Kaur Cheema received her Bachelor of Technology in Computer Science & Engineering from Punjab Technical University, Jalandhar in 2006 and Masters in Computer Science and Engineering from Thapar University, Patiala in 2008. Presently she is working as Lecturer (Computer Science) in Thapar University, Patiala. Her research interests include interconnection networks, parallel processing, computer architecture.

Rinkle Rani Aggarwal, B.Tech (Computer Science & Engg.), M.S. (Software Systems), is Senior Lecturer in Computer Science & Engineering Department at Thapar University, Patiala. She has more than 12 years of teaching experience and served academic institutions such as Guru Nanak Dev Engineering College, Ludhiana and S.S.I.E.T 'Derabassi. She has supervised 13 M.Tech. Dissertations and contributed 28 articles in Conferences and 12 papers in research Journals. Her areas of interest are Parallel Computing and Algorithms.