Performance and Reliability Analysis of New Fault-Tolerant Advance Omega Network

RITA MAHAJAN\textsuperscript{1}, Dr.RENU VIG\textsuperscript{2},
\textsuperscript{1}Department Of Electronics And Electrical Communication,
Punjab Engineering College, Chandigarh,
INDIA.
\texttt{rita_mahajan@rediffmail.com}
\textsuperscript{2}Department of Electronics,
UIET, Panjab University, Chandigarh,
INDIA.
\texttt{renuvig@hotmail.com}

Abstract:- The performance and fault tolerance are two very crucial factors in designing interconnection networks for a multiprocessor system. A new type of MIN, Fault Tolerant Advanced Omega network, which using 4×4 switches, is proposed on the basis of the idea of the Delta network and the Omega network. In this paper, it is proved that 4×4 switches have the better performance/cost ratios than 2×2 switches based on the current level of the VLSI technology. This paper expounds its topological properties and routing algorithm and makes performance/cost ratios comparisons. The mathematical analysis approach is used here to find the probability of acceptance and bandwidth with change in traffic. Furthermore reliability analysis of Fault-Tolerant Advanced Omega Network (FTAON) is discussed in detail, It is seen that FTAON is more reliable and cost effective than other previously proposed MINs of similar class. It has been also observed that it has fault tolerant and nonblocking capability in complex parallel systems.

Keywords:- Interconnection Network, Advanced Omega Network, Fault-Tolerance, Reliability.

1 Introduction

A myriad of tasks require the computational power and speed made possible by parallel processing. Several supercomputer architectures with multiple processors for scientific and commercial applications have been described in the literature [1]. With the progress of research on parallel processing, interconnection networks provide the communications needed in a parallel processing system between processors, between processors or memory modules benefits, enhancing the performance of the system, requiring lower hardware overhead. The interconnection scheme (MIN) banyan network including Omega networks, Delta networks have been favored for processor memory connection in large-scale shared memory machine because of its cost effectiveness [3]. The modest cost of unique path MINs (a single path between any source and destination), makes them attractive for large multiprocessor systems, but then lack of fault tolerance is a major drawback. A major component of these systems is the interconnection network that enables the processors to communicate among themselves or with memory units. The failure of a component in the interconnection network can frequently bring down the entire system or cause a severe degradation in performance, unless sufficient measures are taken to make the
network tolerant to such failures. The central part of the MPPs, interconnection network, has three important factors: topology, routing technique and flow control mechanism[13]. In order to reduce the cost of it and to improve its transmission performance and scalability, many types of interconnection network have been proposed. As a large subset, Multistage Interconnection Network (MIN) are well suited to communication among tightly coupled system components of UMA, and offer a good balance between cost and performance, such as Omega network, Generalized Cube network, Delta network, the extra stage network and the Clos network [2][15-19] are proposed to improve the transmission performance. Up to now these Multistage Interconnection Networks have been used in many types of parallel computers and in most cases have been realized with 2x2 switches, which is available and convenient to research. Among them, Delta network has quite high bandwidths and performance, but its scalability is inferior. Omega Networks are a famous subclass of blocking Multistage Interconnection Networks (MINs). They have been recently identified as an efficient interconnection network for a switching fabric of communication structures such as gigabit ethernet switch, terabit router, and ATM switch[8]. Omega network has the qualities of simple routing techniques and better scalabilities. In section 2 it has been proved that based on the current level of the VLSI technology, 4x4 switches have the better performance/cost ratios than 2x2 switches. Thus, this paper, on the basis of the idea of the Delta network and the Omega network, proposes the idea of a new type of multistage interconnection network using 4x4 switches, the Advanced Omega network. It has simple routing technique, can achieve many kinds of permutations, and has superior scalability. The evaluation of its performance shows that the Advanced Omega network has higher performance than Delta network, Omega network and Crossbar network.

A crucial quality for MIN serving communications needs of large-scale multiprocessor systems is fault-tolerance. This paper describes the fault-tolerance methods used in the Advanced Omega network. The Fault-Tolerant Advanced Omega network consists of an Advanced Omega network with one additional stage at the input allow the bypass, when desired, of the extra stage or the output stage. Thus, it has a relatively low incremental cost over the Advanced Omega network and achieves better reliability. The extra stage provides some additional paths from each source to each destination. It makes the Fault-Tolerant Advanced Omega network as a non blocking network. Four paths are available between each source destination pair if no fault exists. The distributed control through the use of routing tags is available in the fault-tolerant Advanced Omega network. Therefore, the fault-tolerant Advanced Omega network is an answer to the need for reliable communications in parallel/distributed systems.

2  Element size: a factor of MIN’s performance /cost ratio

From the prospect of the art of VLSI technique, the expanding of system scale has made it difficult to implement many interconnection network configurations [5][10]. Systematic assembling is limited by the pin limitation problem of a LSI chip. This technique in assembling is the final decisive factor in the interconnection network configuration [13]. Only when the network configuration, under this condition, effectively takes advantage of the systematic assembling characteristics can it gain high performance. Even using the current art of technology, it is difficult to solve the pin limitation problem of a LSI chip. Assuring the MIN discussed in this paper have the property that there is only one route between every pair of source node and destination node in the network and the number of switches of every
switches stage in the network is fixed. Then, according with the property mentioned above, we can know that the number of switches needed to construct the MIN has nothing to do with the shuffle way of the interconnection structure. Suppose the delay has a great effect on the systematic performance and the cross points is a vital factor to measure the manufacture cost. So we can analysis the relationship between the number of cross points and the number of switches stages in the MIN in order to approximately find out which switch has the better performance/cost ratio. As the two parameters, the scale of the MIN and that of the switch, decided the number of switches in the MIN, We can use formula showed in equation (1) to calculate the number of cross point of MIN (we can consider a cross point as single channel ) :

\[ S = m \cdot N \cdot \log_{m}N, \]

(1)

where \( S \) is the number of cross point of MIN, \( N \) is the scale of MIN, \( m \) is the scale of switches and \( \log_{m}N \) is the number of switches stages. In this case, consider \( m=4 \), \( N=4^k \), \( k=1,2,3,... \). The results are shown in Table1, Figure1 and Figure 2.

<table>
<thead>
<tr>
<th>Scale of MIN</th>
<th>2x2</th>
<th>4x4</th>
<th>8x8</th>
<th>16x16</th>
<th>Crossbar</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>32</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>16</td>
<td>64</td>
<td>256</td>
<td>1024</td>
</tr>
<tr>
<td>8</td>
<td>16</td>
<td>64</td>
<td>256</td>
<td>1024</td>
<td>16384</td>
</tr>
<tr>
<td>16</td>
<td>64</td>
<td>256</td>
<td>1024</td>
<td>16384</td>
<td>65536</td>
</tr>
<tr>
<td>32</td>
<td>256</td>
<td>1024</td>
<td>4096</td>
<td>16384</td>
<td>65536</td>
</tr>
<tr>
<td>64</td>
<td>1024</td>
<td>4096</td>
<td>16384</td>
<td></td>
<td></td>
</tr>
<tr>
<td>128</td>
<td>4096</td>
<td>16384</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>256</td>
<td>16384</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table1. Comparison in the number of cross point of MIN using different switches and crossbar

From Figure1 and Figure 2, we can see that the number of cross points of MIN using 4x4 switches is the same as that using 2x2 switches, but its number of switches stages is less. That is to say, compared with the same scale MIN using 2x2 switches, we can improve the performance of MIN by using 4x4 switches without anymore cost in terms of latency. We can say that 4x4 switches have the better performance/cost ratios.

![Fig. 1: Comparison in the number of cross points of MIN using different switches and crossbar.](image1)

Fig. 1: Comparison in the number of cross points of MIN using different switches and crossbar.

![Fig. 2: Comparison in the number of switch stages of MIN using different switches.](image2)

Fig. 2: Comparison in the number of switch stages of MIN using different switches.

3 The Advanced Omega network
The Omega network has \(N=2^n\) inputs, termed as sources (S) and \(N\) outputs, termed as destinations. There exist \(\log_2 N\) stages and each stage has \(N/2\) switching elements. There is a unique path between each source-destination pair. The network complexity, defined as the total number of switching elements in the MIN is \((N/2)^*(\log_2 N)\). In Omega network, the routing of a message from a given source to a given destination is based on the destination address. This is termed as Self-routing. Each bit in the destination address can be used to route the message through one stage and if the bit in the destination address controlling the routing in a given stage is 0, then the message is routed to the upper output of the switch otherwise is routed to the lower output of the switch. Figure 3 shows 8 x 8 Omega network. Some paths are also shown for destination 2, 3 and 5. Routing tags are 010, 011, 101 respectively whatever the source is. One problem associated with the Omega network is the high probability of blocking. Although any input may be connected to any output, it is desirable to be able to connect other inputs to other outputs. Blocking occurs when the second and subsequent connections cannot be made.

![Fig. 3: 8 x 8 Omega Network](image)

On the basis of the idea of the Delta network and the Omega network, a new type of MIN, Advanced Omega network, which using 4x4 switches, is proposed. This section expounds its topological properties and routing algorithm and makes performance/cost ratios comparisons.

### 3.1 Definition of the Advanced Omega network

The Advanced Omega network is a kind of blocking multistage interconnection network using 4x4 switches. The interior structure of every switch is a 4x4 crossbar. The Advanced Omega Networks’ scale is \(N\) x \(N\) and \(N = 4^k\) (\(k = 2, 3, 4\ldots\)). The physical names of the switching components (stage, element and link) are assigned as follows. The stages are labeled in a sequence from 0 to \((\log_4 N)−1\) with 0 for the leftmost stage and \((\log_4 N)−1\) for the rightmost stage. Similarly, the levels of links are labeled in a sequence from 0 to \(\log_4 N\). In each stage, each switching element is named by a codeword of \(n=(\log_4 N)−1\) tetradic digits, \(p_n\ p_{n−1}\ldots\ p_1\). Each link in each level is named by a codeword of \(n+1\) tetradic digits \(p_n\ p_{n−1}\ldots\ p_0\). The first \(n\) leftmost digits \(p_n\ p_{n−1}\ldots\ p_1\) are the same as the tetradic representation of the switching element to which the link is connected at one of four right terminals of the switching element; the last digit, \(p_0\), is equal to 0 if the link is connected to the uppermost terminal of the switching element and \(p_0\) is equal to 1, 2, 3 in a sequence from 1 to 3 if the link is connected to the below terminals. The topology describing rules of the Advanced Omega network is defined by

For link \((p_n\ p_{n−1}\ldots\ p_0)\) of switch \((p_{n−1}\ldots\ p_1\) is connected to link \((p_{n−1}\ldots\ p_0\) \(p_n\)) of switch \((p_{n−1}\ldots\ p_0\) \(p_n\)) for each stage.

The topology of the Advanced Omega network can be generated in an advanced way. Figure 4 shows 16 x 16 Advance Omega network.
3.2 The Advanced omega network’s routing techniques

The routing algorithm uses a destination-tag routing method [4], by which the Advanced omega network can set up any mapping connection from one terminal on one side of the network to another terminal on the other side. Consider switching input number $S$ to output number $D$. Let $D = d_n d_{n-1}... d_1 d_0 [n=(\log_4 N)-1 \text{ and } 0 \leq d_i \leq 3, 0 \leq s_i \leq 3]$ be the destination tag, and let $S = S_n S_{n-1}... S_1 S_0$ be the source tag. The Routing algorithm assigns $c = d_n - i$ as the routing control signal of the switching element in stage $i$ to control its state. It is set to switch the input to the no. $c+1$ output of the switching element. It is easy to see that at any stage $i$, an input that has been switched to position $p_n p_{n-1}... p_1 p_0$ is then switched into position $p_n p_{n-1}... p_1 d_i$ by the switching element, and then goes through the shuffle and ends up at $p_n p_{n-1}... d_i p_{i+1}... P_1$. Thus, after $n+1$ operations, the original input must be connected to output $d_n d_{n-1}... d_1 d_0$. So it will certainly be able to get to the destination properly. The Advanced Omega network’s simple destination-tag routing method makes it efficient and lower cost in routing.

4 Analysis in the Advanced omega network performance

We evaluate the performance of the Advanced Omega network from three aspects: network’s pass rate, latency and performance/cost ratio. In accordance with the analysis model that has been made [16] [22], we can compare the performance of the Advanced Omega network with that of the existing Crossbar network, Delta network and Omega network.

Figure 5 shows the curves of performance/cost ratio. The number suggests that when $N \geq 16$, the performance/cost ratios of the Advanced Omega network is obviously superior to that of Crossbar, Delta network and Omega network[7]. Adopting the Advanced Omega network as the interconnection network of several multiprocessor systems is more attractive.

4.1 Bandwidth Analysis in the Advanced omega network

Assume a MIN of size $a^n \times b^n$ constructed from a x b crossbar switches and having $a^n$ Sources connected to $b^n$ destinations. The analysis of crossbar is applied to a x b crossbar switch and then extended to the complete MIN. The distinct destination digit (in base b) for setting of individual a x b switches controls each stage of MIN.
Since the destinations are independent and uniformly distributed, so are the destination digits. The probabilistic approach is used to analyze the MINs based on independent request assumptions[14]. Given the request rate p at each of the a inputs of an a x b crossbar module, the expected rate of requests on any one of the b output lines from any one input is given by p/b.

Probability of not getting the request is:

\[ 1 - \frac{p}{b} \]

Probability of not getting the request from all ‘a’ inputs is given by:

\[ (1 - \frac{p}{b})^a \]

Probability of one output getting the request from all ‘a’ inputs is given by:

\[ 1 - (1 - \frac{p}{b})^a \]

The total number of requests that it passes per unit time is given by

\[ b - b(1-p / b)^a \]

Since the output rate of a stage is the input rate of the next stage, output rate of any stage can be recursively calculated starting from stage 1. And the output rate of final stage n determines the bandwidth of MIN.

**Probability of acceptance** (**P_a**): This parameter is defined as the ratio of expected bandwidth to the expected number of request generated by a source per cycle. Thus, it is the ratio of requests matured to the total requests generated. The probability of acceptance of a request is given by:

\[ P_a = \frac{b^n p_n}{a^n p_0} \]  

(2)

Where \( p_0 \) is the request rate at each of the \( a^n \) inputs of an \( a^n \times b^n \) MIN and \( p_n \) is the expected rate of requests on any one of the \( b^n \) output lines.

**Bandwidth**: It is defined as the mean number of active memory modules in a transfer cycle of interconnection network. So BW is the total number of requests matured.

\[ BW = b^n x p_n \]  

(3)

The value of \( p_n \) is calculated recursively as

For omega network (\( N=16 \)) shown in figure 6 has five link stages.

\[ p_0 = p \]

\[ p_1 = 1 - (\frac{p_0}{2})^2 \]

\[ p_2 = 1 - (\frac{p_1}{2})^2 \]

\[ p_3 = 1 - (\frac{p_2}{2})^2 \]

\[ p_4 = 1 - (\frac{p_3}{2})^2 \]

\[ p_n = p_4 \]

For advance omega network (\( N=16 \)) shown in figure 4 has three link stages.

\[ p_0 = p \]

\[ p_1 = 1 - (\frac{p_0}{4})^4 \]

\[ p_2 = 1 - (\frac{p_1}{4})^4 \]

\[ p_n = p_2 \]

\[ P_a = \frac{p_n}{p_0} \] and \[ BW = P_a \times N \times p \]

The following graphs in figure 7 (a) and (b) show the comparative probability of acceptance and bandwidth w.r.t. congestion. From the results it can be inferred that as the request generation probability increases, probability of
acceptance and bandwidth is better in case of Advanced Omega network as compared to omega network. Hence the AON shows overall gain in performance.

![Fig. 7(a): Comparative Probability of acceptance](image)

**Space Redundancy:** Multiple paths are provided from each input port to each output port by using redundant hardware-extra stages of switches, extra switches in each stage and/or extra links, duplicated network in each switch. In the presence of faults, an alternate path can be chosen to continue providing access.

**Time Redundancy:** In this case, multiple passes through the network can be allowed to route data from an input to an output. The ability to route packets from input to output of the network in a finite number of passes is called dynamic full redundancy (DFA) property. Fault-tolerance is an important quality for MIN serving the communication needs of large-scale multiprocessor systems. Various methods have been discussed for fault tolerance techniques[6][21]. This section describes the fault-tolerance methods used in the Advanced Omega network and present the fault-tolerant Advanced Omega network that is capable of operating in SIMD, multiple-SIMD, and partitionable SIMD/MIMD environments. The fault-tolerant Advanced Omega Network consists of an Advanced Omega network with one additional stage at the input to allow the bypass, when desired, of the extra stage or the output stage. Thus, it has a relatively low incremental cost over the Advanced Omega network and achieves better reliability. The extra stage provides some additional paths from each source to each destination [18][19]. The distributed control through each of routing tags is available in the fault-tolerant Advanced Omega network. The fault-tolerant Advanced Omega network has all of the interconnection capabilities of the Advanced Omega network. Furthermore, the network can be controlled when it has a failure, using a simple modification of a routing-tag scheme proposed for the Advanced Omega network. The proposed routing algorithm is as follows:

1. **Step1:** Read the destination address \(d_n d_{n-1} d_{n-2} \ldots d_1 d_0\).
2. **Step2:** Apply tag 0 \(d_n d_{n-1} \ldots d_1 d_0\). If path free pass the message otherwise goto step3.
Step3: Apply tag 1 \( d_n \) \( d_{n-1} \ldots d_1 \) \( d_0 \). If path free pass the message otherwise goto step4.

Step4: Apply tag 2 \( d_n \) \( d_{n-1} \ldots d_1 \) \( d_0 \). If path free pass the message otherwise goto step5.

Step5: Apply tag 3 \( d_n \) \( d_{n-1} \ldots d_1 \) \( d_0 \). If path free pass the message otherwise wait for the average pass time and goto step2.

Normally, the network will be set so that stage e is disabled (bypassed) and stage n-1 is enabled. If the fault is in stage n-1, stage e is enabled and stage n-1 is disabled, i.e., stage e replaces the function of stage n-1. For a fault in a link or switch in stage 0 to n-2, both stage e and n-1 will be enabled. The fault-tolerant Advanced Omega network gets its fault-tolerant abilities by having redundant paths from any source to any destination. In the fault-tolerant Advanced Omega network with both stage e and n-1 enabled, there exist exactly 4 paths between any source and destination. The extra stage of the fault-tolerant Advanced Omega network does increase the likelihood of a fault, compared to the Advanced Omega network, due to the additional hardware and become a non blocking network due to the availability of alternate paths.

6 Reliability

A well-known criterion used to measure the reliability of fault tolerant multistage networks is full access i.e. the capability of the network that provides the connection from any of its input sources to any of its output destinations[11]. Under the criterion of full access, a network is assumed to be faulty if there is any source–destination pair that cannot be connected because of faulty components in the network. Under this criterion, reliability of a network can be measured in terms of MTTF (Mean Time to failure) of the network and/or the number of faults a network can tolerate[9].

Three fault models are used for MINs: the stuck at fault model, the link-fault model, and the switch-fault model. In the stuck-at fault model, a failure causes a crossbar switch to remain in a particular state regardless of the control inputs given to it, thus affecting its capability to setup suitable connections. In the link-fault model, a failure affects an individual link of a switch, leaving remaining part of the switch operational. In the switch-fault model, the strongest of the three, a failure of a switch.
makes it totally non-operational. Switch-fault model is used for the analysis of Advanced Omega network.

The MTTF of a MIN is defined as the expected time elapsed before some source is disconnected from some destination. The analysis is based on the lower and upper bounds of the network reliability[12]. Following assumptions are used:

(i) The node reliability is homogenous with an exponential failure rate of 2λ for a 4x4 crossbar switch. (A reasonable estimate for λ is about 10^-6 per hour.)

(ii) Failure rate for a 2x1 multiplexer or 1x2 demultiplexer is λ/2. The failure rate for a kx1 multiplexer or 1xk demultiplexer is (kλ)/4.

(iii) Switches are statistically identical and they are fully operational or failed.

The bounds on reliability are computed by applying Series-Parallel model of the concerned network, considering that any of the switching elements, multiplexers or demultiplexers may fail.

The analysis is carried as under:

Let R_s(t), the reliability function of a component, be defined as the probability that a failure does not occur in the time period (0,t).

Then reliability for 2x2 switch is given by equation (4).

\[ R(t)_{2x2\text{switch}} = e^{-\lambda t} \] (4)

And for a 4x4 crossbar switch is given by equation (5).

\[ R(t)_{4x4\text{switch}} = e^{-2\lambda t} \] (5)

And for a k x 1 multiplexer or a 1 x k demultiplexer is given by equation (6).

\[ R(t)_{\text{mux}} = e^{-k\lambda t/4} \] (6)

For a system with reliability R_s(t), the MTTF is given by equation (7).

\[ MTTF = \int_0^\infty R_s(t) \, dt \] (7)

6.1 Omega Network and Advanced Omega Network

In these networks any single fault leads to a network failure under the full access criterion. Series reliability model is employed for unique path omega network and Advanced omega network. All switching elements can be considered in series.

\[ R(t)_{\text{omega}} = [e^{-\lambda t}]^N \log_2 N \] ......(8)

This MTTF is dependent on the network size N and on the real time operation 't'. By using 2x2 Switching element MTTF for Omega network is evaluated from the following expression (9)

\[ MTTF_{\text{omega}} = \int_0^\infty [e^{-2\lambda t}]^N \log_2 N \, dt \] ......(9)

By using 4x4 Switching element MTTF is given by expression (10)

\[ MTTF_{\text{AON}} = \int_0^\infty [e^{-2\lambda t}]^N \log_2 N \, dt \] ......(10)

Mean Time to Failure (MTTF) for Fault Tolerant Advance Omega Network is given by equation (11).
Comparison of MTTF of Omega Network, Advance Omega Network and Fault Tolerant Advance Omega Network is shown in figure 10. It is observed that MTTF of FTAON is maximum so it is most robust and reliable.

$$\text{MTTF}_{\text{FTAON}} = \int_0^x [1 - (1 - e^{-2\lambda t} + \frac{N}{N_{\text{log}_2 N}})^2] e^{-\lambda t N}\,dt \quad \ldots(11)$$

Fig. 10: MTTF for Omega and advance omega networks and FTAON

7 Conclusion

Based on the current level of the VLSI technology, MIN can achieve better performance/cost ratios by using 4x4 switches instead of 2x2 switches. The new proposed Advanced Omega Network is proposed, which has simple routing techniques and the better performance. It has been also proved that as congestion increases the probability of acceptance increases and hence the bandwidth. And a Fault-Tolerant Advanced Omega network is discussed in detail, which has a relatively low incremental cost over the Advanced Omega Network and achieves better reliability in terms of MTTF. The redundancy in paths improves both reliability and performance of the network. In conclusion, the fault-tolerant Advanced Omega network found to be robust and suits the current art of the VLSI technology and has promising application prospects.

References:


