Fault Tolerance in Optically Interconnected Multiprocessor Networks

Poornima Lalwaney and Israel Koren
Department of Electrical and Computer Engineering
University of Massachusetts
Amherst, MA 01003

Abstract

With the increasing demand on the interprocessor communication bandwidth, optical interconnects are being considered as alternatives to electronic interconnects in high-performance systems. In this paper, we investigate the fault-tolerance properties of optical networks in the presence of link failures. The principles of implementing a network topology using optical links differ from those of its electronic counterpart. We propose two schemes to tolerate link failures in fiber-optic networks based on wavelength-division multiplexing. The performance of these schemes is compared with the rerouting scheme, which is commonly used to tolerate link failures in multiprocessors with electronic interconnects.

1: Introduction

Optical interconnects are being considered as alternatives to electronic interconnects in large multiprocessor systems [7]. Besides the advantages of high bandwidth and low wire density, they support high data rate communication with lower power requirements than electronic interconnects [6] . The advantages of optics in local area networks and wide area networks are well established. Due to advances in semiconductor optoelectronic device technology over the past decade, devices with increased optical-to-electrical conversion efficiency and vice-versa , low power requirements, and small physical dimensions are currently available [4, 8]. Optical interconnects are therefore being considered in multiprocessing and distributed processing environments. Some recent efforts in this direction include using the high bandwidth offered by optics to lower wiring density in a Connection machine prototype [9] and for demonstarting system scalability in Intel's Touchtone supercomputer [5].

In optical link implementations, every unidirectional link requires a transmitter at the source node and a receiver at the destination node, transmitting and receiving data at a common wavelength over a communication medium. In the fiber-optic domain, wavelength division multiplexing (WDM) using passive star couplers is widely employed in local area networks to realize network interconnections. In WDM based networks, the bandwidth of the optical fiber is split into many wavelength channels, each channel carrying data at a particular wavelength [12]. The logical connectivity is obtained by assigning wavelengths to the system's transmitters and receivers. Reconfiguring the interconnection network to a different topology is a simple matter of wavelength reassignment if the transmitters and/or receivers are tunable over the entire range of wavelengths used. The passive star coupler is one of the devices that may be used to realize network configurations using wavelength-division multiplexing. The transmitters and receivers from the network nodes are connected to the input and output ports of the star coupler, respectively. The signal power at each input port is equally divided among the output ports. Thus the wavelengths from all the transmitters appear at each output port. Demultiplexing is performed by the receivers connected to the output ports to recover the desired transmitter wavelengths from the combined wavelength signals at each output port. As the distances involved in multiprocessor interconnects do not exceed a few meters, we assume that a transmitter wavelength is directly coupled to a fiber which in turn is connected to the input port of the coupler. Thus every fiber connected to an input port carries a single wavelength. On the output side of the coupler, a fiber from the ouput port feeds a single receiver at a node.

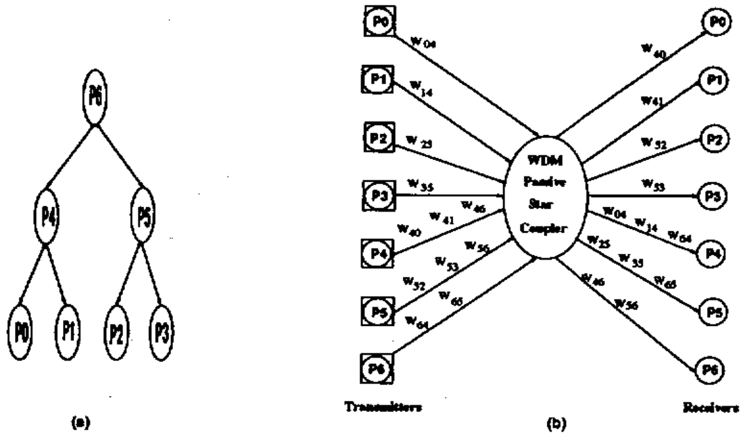

Figure 1: **(a) A binary tree with seven nodes. (b) WDM star embedding of the tree. W_{ij} is the wavelength assigned for communicating from node i to node j. W_{ij} is assigned to the transmitter at processor i and the receiver at node j.**

The logical topology and the corresponding physical passive star implementation is illustrated in Figure 1 for a seven node binary tree with bidirectional links. As an example, to establish a unidirectional logical link from node P0 to P4, one of the transmitters at node P0 transmits at wavelength W_{04}. One of the receivers at node P4 is set to receive at wavelength W_{04}. In this manner, the data signal transmitted on wavelength W_{04} from node P0 is received at node P4. W_{04} is therefore the wavelength assigned to a transmitter and receiver at node P0 and P4, respectively. Reconfiguring the network to a different topology involves reassigning wavelengths to the transmitters and receivers. This can be achieved if the transmitters and/or receivers are tunable over a range of wavelengths used in the network. Some limitations in implementing reconfigurable networks using WDM technology include the tuning range and tuning time of the transmitters and receivers and the number of input and output ports on the star coupler [2, 4, 11].

Using the property of device tunability and the independence of the logical topology from the physical star topology, we propose two schemes for tolerating link failures in multiprocessor networks implemented with this technology. We then compare the performance of these schemes with the message rerouting scheme that is used in electronic networks to tolerate link failures [1]. In Section 2, we introduce two schemes that use the property of wavelength division multiplexing to tolerate link failures. The performance of these schemes is discussed in Section 3.

2: Fault-tolerance issues

In this section, we discuss the fault tolerance of optical networks in the presence of link failures. Due to the difference in implementing network configurations in the optical and electronic domain, schemes to tolerate link failures in electronic implementations may not be the most efficient ones for optical implementations. Rerouting messages along alternate paths in the network topology is commonly used in electronically implemented networks to tolerate link failures. We propose and analyze two schemes that use the properties of wavelength division multiplexing to tolerate link failures in optically interconnected multiprocessor networks. These are referred to as *wavelength reassignment* and *time division multiplexing*. The performance of these schemes is compared with the optical implementation of the message rerouting scheme.

We analyze the wavelength reassignment, time division multiplexing, and rerouting schemes by assuming that link failures are caused by transmitter failures. The analysis is equally applicable to receiver failures. The failure of the physical link may be modeled as a combination of transmitter or receiver failures. We estimate the performance of the schemes by considering the number of link failures that may be tolerated in the network, and the effect of a single link failure on the average distance and average delay of the network. In the analysis, we take into account the properties of the link implementation (for example, latency and bandwidth) and limitations of current optical technology (for example, number of tunable wavelengths, reconfiguration overheads, and passive star coupler limitations).

2.1: Rerouting

The commonly used technique to tolerate link failures in multiprocessor networks is to reroute messages between the source and destination along alternate paths so as to avoid the use of the failed links. These alternate paths may be present in the original network topology as with the case of the mesh topology. Many variations of network topologies that cannot provide alternate paths have been proposed by using additional links that allow for message rerouting. The tree topology is one such example. Variations of the tree topology to tolerate single link/single node failures include full-ringed, half-ringed, and leaf-ringed trees. All the above-mentioned variations use additional horizontal links at various levels of the binary tree topology. In a full-ringed binary tree, the adjacent nodes at a level are connected. In a half-ringed tree, alternate pairs of nodes at a level are connected. In a leaf-ringed tree, adjacent leaf nodes are connected.

Rerouting of messages affects both the average distance between nodes and the average delay between them. For the mesh and the three variations of the tree topology considered here, the average distance in the presence of a single link failure was derived [10].

Figure 2 shows the percent increase in average distance for the three variations of the tree in the presence of a single link failure as a function of the network size. The full-ringed tree has the maximum number of redundant links and is therefore least affected by the link failure. As can be seen from the figure, with increasing network size the effect of a single link failure diminishes in the full-ringed and half-ringed trees. In leaf-ringed trees, the alternate path involves traversing down the tree to the leaf level and up the tree from the leaf nodes to the destination. The sharp increase in network delay for the leaf-ringed trees for increasing network size can be seen in the figure.

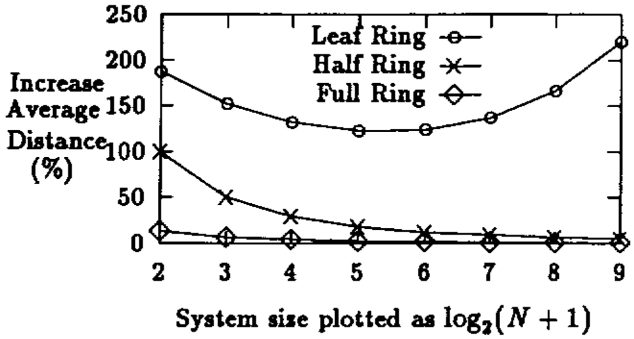

Figure 2: Percent increase in average distance for increasing tree sizes in the presence of a single link failure using rerouting in three variations of the binary tree topology. (System size $N = 2^l - 1$.)

2.2: Time-division multiplexing

In this scheme, we assume that every node has at least one tunable transmitter that can tune to the transmitting wavelengths of all the outgoing links at that node. In the event of a transmitter failure at the node, the tunable transmitter switches periodically between its normal operating wavelength and the wavelength of the failed transmitter. As the tunable transmitter time multiplexes between the outgoing communication in two directions, we refer to this scheme as time-division multiplexing (TDM).

Upon the occurence of a transmitter failure, the messages in two directions are likely to be affected (that is, the direction of the faulty link and the direction of the tunable link). Assuming that the time slot C is equally divided for communication along the two directions, we note that if the number of messages that need to be transmitted in either direction is less than $\lceil ((C/2) - \alpha - t_r)/\beta \rceil$, there is no increase in the time to transfer the message due to the link failure. In the expression for this threshold load, α and β represent the link setup time and the time for a 32-bit floating-point word transfer, respectively. The time to switch between the two wavelengths is represented by t_r, the reconfiguration overhead. Note that the parameters of the link implementation, α, β, and t_r, and the length of the communication time slot C determine the maximum number of messages for which the link transfer time is not affected by the presence of the link failure. This limit on the network load is independent of the network topology.

The average delay between two nodes in a network depends on the network topology and the routing scheme when the network load is beyond the threshold value. For an N node mesh configured in a $\sqrt{N} \times \sqrt{N}$ array, and an l level binary tree with N nodes, where $N = 2^l - 1$, the average delay using the TDM scheme for a uniform load of x messages per link was derived [10] and the results are summarized below.

$$D_{TDM}^{mesh} = \frac{2}{3} \cdot \sqrt{N} \cdot \lceil \frac{\beta \cdot x}{C - \alpha} \rceil \cdot C + \frac{1}{3N} \cdot (\lceil \frac{\beta \cdot x}{(C/2) - \alpha - t_r} \rceil - \lceil \frac{\beta \cdot x}{C - \alpha} \rceil) \cdot C$$

$$D_{TDM}^{tree} = \theta \cdot \lceil \frac{\beta \cdot x}{C - \alpha} \rceil \cdot C + \frac{\Delta d}{3 \cdot l \cdot N_p} \cdot (\lceil \frac{\beta \cdot x}{(C/2) - \alpha - t_r} \rceil - \lceil \frac{\beta \cdot x}{C - \alpha} \rceil) \cdot C$$

where $\Delta d = 2^{2l+2} \cdot \left(1 + (1/4)^l - 2 \cdot (1/2)^l\right) + 3 \cdot 2^{l+1} \cdot \left(1 - (1/2)^l\right) - 3 \cdot l \cdot 2^{l+1} + 2 \cdot l$. The average distance in the fault-free binary tree is denoted by θ and is given by $\theta = \left[(2l-6) \cdot 2^{2l} + 6 \cdot 2^l + l \cdot 2^{l+1}\right]/N_p$. Note that $N_p = (2^l - 1) \cdot (2^l - 2)$ denotes the number of possible source destination pairs for the tree topology. The first term represents average delay in the network with the defined routing scheme in the absence of a link failure. The second term represents the increase in average delay due to the presence of a single link failure.

Comparison of average network delays for the mesh and tree topologies for the TDM and rerouting schemes is presented in Section 3.

2.3: Wavelength reassignment

In this scheme, we assume that redundancy is introduced in the network by selectively incorporating spare transmitters and receivers at network nodes. In the event of a link failure due to a failed transmitter, the outdegree at the affected node decreases by 1. This node could function as a node of lower degree in the network topology considered. A node of lower degree with unused spare transmitters, having as many functioning transmitters as required at the position of the failed node, could be used to replace the failed node. This change of logical connectivity between the failed node and its replacement can be achieved by wavelength reassignment. For a topology with nonuniform degree, spares may be placed at lower degree nodes, so as to logically change its functionality with that of a higher degree node with a failed link. In the case of a topology in which all nodes have the same degree, additional spare transmitters are added at all nodes to provide for link redundancy. In either case, the number of link failures that may be tolerated depends on the number of available spare transmitters.

In WDM implementations, the amount of redundancy that may be placed in the network depends on the number of available spare wavelengths and spare input ports available on the star coupler. For small networks, all the network links may be realized on a single passive star coupler. As there are limitations on the number of available wavelengths and the number of input and output ports on a star coupler, multiple star couplers are required to implement large networks. The network has to be partitioned and a set of nodes mapped onto a coupler. The mapping scheme affects the amount of redundancy that may be efficiently utilized. Further discussion on these issues can be found in [10].

Note that in the electronic implementations, the use of redundant spares is prevalant to cover single link failures [1, 3]. However, the method of reconfiguration and restoring the network to its full functioning capacity is not as easily accomplished. In optical WDM implementation, the logical topology may be different from the physical star topology. In the event of a link failure and the availability of spare wavelengths and transmitters, the wavelength reassignment phase restores the network to its fault-free state. Unlike electronic interconnects, where spare links can be used only locally, in optical interconnects spare links provide global redundancy. Thus a few spare links (transmitters/receivers) can provide a high level of fault tolerance.

3 Comparison of the three schemes

In this section we compare the wavelength reassignment, time-division multiplexing, and the rerouting schemes by considering the number of link failures that may be tolerated and the effect of the link failure on network performance. The performance measure considered is the average distance and the average delay.

Using the wavelength reassignment scheme, the number of single link failures that may be tolerated depends on the number of available spare wavelengths per star and the mapping scheme used. There is no effect on the average distance or the average delay. The overhead in implementing this scheme is the wavelength reassignment phase. The tuning time of the interface devices varies from nanoseconds to milliseconds, depending on the method used to achieve tuning [4]. A maximum overhead of a few milliseconds will be incurred for every link failure.

In the event of a link failure that cannot be covered by wavelength reassignment, either the TDM or the rerouting scheme may be used. In the TDM scheme, if the network load is below a threshold value, there is no increase in network delay. This threshold is topology independent but depends on link implementation parameters. Beyond this threshold load, the average delay increases depending on the network size and topology. In the rerouting scheme both the average distance and average delay increase due to the presence of a link failure.

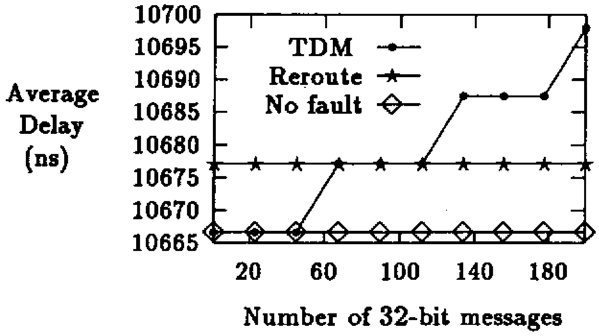

Figure 3: **Variation in average delay on a 64-node mesh when using the TDM and rerouting schemes in the presence of a single link failure. Average delay in the absence of failures plotted for comparison. $\alpha = 400$ ns, $\beta = 8$ ns, $t_{rec} = 100$ ns, and $C = 2$ μs.**

The variation in average network delay with network load on a 64-node mesh for the TDM and rerouting schemes in the presence of a single link failure is shown in Figure 3. The average delay of the fault-free network is plotted for comparison. The link set-up time α is assumed to be 400 ns. This was the setup time reported in Intel's effort in incorporating fiber-optic interconnects in a Touchtone Delta Supercomputer prototype. We assume that the message size is 32 bits. Assuming a 4 Gbps fiber-optic link, a 32-bit word may be transferred in 8 ns. The message transfer time β is therefore 8 ns. The reconfiguration overhead t_r is assumed to be 100 ns and the length of the communication time slot C is assumed to be 2 μs. With these parameters, the maximum number of messages that can be transferred in time slot C over a fault-free link is 200. When using the TDM scheme, the threshold load per link, below which there is no increase in average delay, is 62 messages.

In Figure 4, we depict the variation in delay for the TDM scheme in the binary tree topology with increasing uniform network load in number of 32-bit messages per node for a 63-node tree. The average delay using the rerouting scheme in full-ring and half-ring trees is shown for comparison. As can be seen from the figure, the number of redundant paths between the source and destination affects the performance of the rerouting schemes. The full-ringed tree performs better than the half-ringed tree. As seen in Figure 2, the performance of the leaf-ringed tree is dependent on network size. For the network size considered, the delay in the leaf-ringed tree is much higher than that in the half-ringed tree and is therefore not included in the figure.

In Figure 5, we show the effect of network size on the relative increase in average network delay due to the link failure. As the network size increases, the effect of a single link failure on the network decreases as expected. This can be seen by the decrease in average delay with increasing system size. This decrease is seen for both the TDM and the rerouting schemes in the full-ringed and half-ringed trees. In the TDM scheme, the average delay is dependent on the network size and the network load. As seen in Figure 5, for small network loads, the network delay is not affected in the TDM scheme. This is seen from the curve labeled 'TDM x=50'. It represents the increase in delay due to a link failure over the average delay when the number of messages to be transferred per network link in a communication time slot is 50. The dependence of the performance of the TDM scheme on the network load can be seen by the upward shift in the relative delay curves for increasing loads. The TDM curve for a load of 100 messages per time slot per link falls between that of rerouting in full-ring and half-ring trees.

As seen from Figures 3, 4 and 5, for low network loads the TDM scheme is preferred over the rerouting scheme. The performance of the rerouting scheme compared to the TDM scheme for high network loads depends on the number of distinct paths available between source and destination. In other words, the number and distribution of the spare links in the rerouting network determine the performance. Thus the most efficient scheme to tolerate link failures can be selected based on the network load.

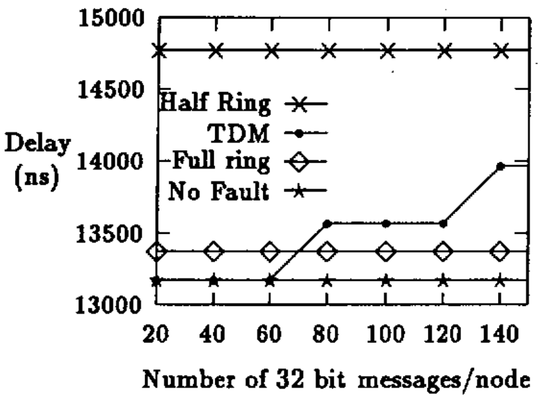

Figure 4: Variation in average delay for rerouting in full-ring and halfring binary trees in the presence of a single link failure compared with the TDM scheme. $\alpha = 400$ ns, $\beta = 8$ ns, $t_{rec} = 100$ ns, and $C = 2$ μs.

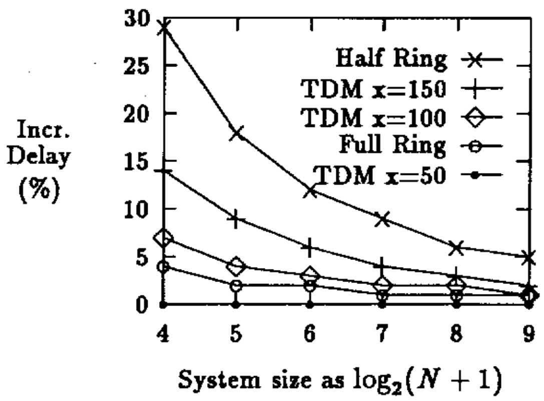

Figure 5: Percent increase in average delay for increasing tree sizes in the presence of a single link failure for TDM and rerouting schemes. $\alpha = 400$ ns, $\beta = 8$ ns, $t_{rec} = 100$ ns, and $C = 2$ μs.

4: Conclusions

Fault-tolerance schemes used in electronic implementations of multiprocessor networks may not be the best for optical implementations. The principles of realizing a network topology with optical interconnects differs from its electronic counterpart. We analyzed the fault tolerance properties of optical interconnects by considering fiber-optic networks based on wavelength-division multiplexing (WDM). Using the property of wavelength-division multiplexing, we presented two schemes, namely, wavelength reassignment and time-division multiplexing, for tolerating link failures in optical interconnects. In the wavelength reassignment scheme, redundant spares may be used to cover link failures with no degradation in network performance. In electronic networks, spare links are used to provide local redundancy. In optical implementations, the spares provide global redundancy, and so a few spare links can provide a high level of fault-tolerance. In the absence of spares, a time-division multiplexing scheme that uses the tunability of optical devices to implement logical links was presented. We observed that for low network loads, the TDM scheme does not affect the average delay or average distance in the network. Using the mesh topology and fault-tolerant variations of the tree topology, we estimated the effect on overall network performance for the TDM and rerouting schemes. For high network loads, the relative performance of the TDM and rerouting scheme depends on the number and length of the rerouting paths.

References

[1] M. S. Alam and R. G. Melhem, "Routing in Modular Fault-Tolerant Multiprocessor Systems," *Proceedings of the Twenty-Second IEEE International Symposium on Fault-Tolerant Computing*, pp. 185-193, July 1992.

[2] C. A. Brackett, "Dense Wavelength Division Multiplexing Networks: Principles and Applications," *IEEE Journal on Selected Areas in Communications*, Vol. 8, No. 6, pp. 948-964, August 1990.

[3] J. Bruck, R. Cypher and C. T. Ho, "Wildcard Dimensions, Coding Theory and Fault-Tolerant Meshes and Hypercubes," *Proceedings of the Twenty-Third IEEE International Symposium on Fault-Tolerant Computing*, pp. 260-267, June 1993.

[4] A. Cisneros and C. A. Brackett, "A large ATM Switch Based on Memory Switches and Optical Star Couplers," *IEEE Journal on Selected Areas in Communications*, Vol. 9, No. 8, pp. 1348-1360, October 1991.

[5] B. E. Floren et al, "Optical Interconnects in the Touchtone Supercomputer Program," *Integrated Optoelectronics for Communication and Processing*, Proc. SPIE 1582, pp. 46-54, 1991.

[6] J. W. Goodman, F. J. Leonberger, S. Y. Kung and R. A. Athale, "Optical Interconnections for VLSI Systems," *Proceedings of the IEEE*, Vol. 72, No. 7, July 1984.

[7] A. Guha, J. Bristow, C Sullivan and A. Husain, "Optical Interconnections for Massively Parallel Architectures," *Applied Optics*, Vol. 29, No. 8, March 1990.

[8] H. S. Hinton, "Architectural Considerations for Photonic Switching Networks," *IEEE Journal on Selected Areas in Communications*, August 1988.

[9] B.O. Kahle, E.C. Parish, T.A. Lane and J. A. Quam, "Optical Interconnects for Interprocessor Communications in the Connection Machine," *IEEE Conference on Computer Design*, Cambridge, MA, October 1989.

[10] P. A. Lalwaney and I. Koren, " Fault-tolerance issues in optical interconnection networks," *TR-94-CSE-12*, 1994.

[11] P. F. Moulton, "Tunable Solid State Lasers,"*Proceedings of the IEEE*, Vol. 80, No. 3, pp. 348-364, March 1992.

[12] Special Issue on "Dense Wavelength Division Multiplexing Techniques for High Capacity and Multiple Access Communication Systems," *IEEE Journal on Selected Areas in Communications*, August 1990.