SCOPIN: A New Scalable Optical Interconnection Network for Multiprocessor Systems

Ekpe Okorafor and Mi Lu
Department of Electrical Engineering
Texas A&M University
College Station, TX 77843
{ekpe,mlu}@ee.tamu.edu

Abstract

We present a design of a scalable interconnection network that possesses a hierarchical structure. We also present the network topology, features, and performance analysis. The proposed network is well suited to optical implementation and will be used in massively parallel systems. This network is called a SCalable OPtical Interconnection Network (SCOPIN). SCOPIN is a single-hop, scalable, flexible, fault-tolerant interconnection network that uses WDM (Wavelength Division Multiplexing) techniques to achieve high performance. Here single hop refers to the ability of each transmitter of an access node to directly reach the receivers of every access node by one of its wavelength-channels. The proposed network structure consists of a local ncube cluster decomposed into transmitter/receiver groups that are scalable and a global torus link network that connects these local ncube clusters. This paper also takes a look at the structural aspects, which include the transit nodes in the optical medium and the transmitter/receiver configuration. These in turn determine the physical and virtual topologies respectively.

1. Introduction

Interprocessor communication in large processor systems such as massively parallel processing (MPP) systems will increasingly become the bottleneck that limits the performance of such supercomputing systems [1-3]. There are often needs for these systems to be scalable for a wide range of applications. The interconnection network must therefore be designed to support communications to remote nodes at a bandwidth similar to the bandwidth of local nodes [1].

Our idea is to use a regular topology as a building block. These building blocks form a cluster in the local level and are scalable. The whole interconnection network, (local, and global) is coupled optically.

2. Topology of the SCOPIN network

The proposed SCOPIN architecture features a hierarchical structure consisting of *m* local clusters (hypercubes) and a global spanning network linking the clusters in a torus. The SCOPIN topology is a modified network of torus-connected hypercubes.

This approach for the topology stems from the fact that a higher percentage of multiprocessor communication occurs with neighboring nodes. Thus it is expected that the vast bulk of the communication will be at n-cube level.

Network representation

The set of vertices represents the set of clusters (*i.e.*, each cluster or hypercube of 2^n nodes is represented by a vertex in the undirected graph, $G_{SCOPIN} = (V, E)$) and the set of edges represents the global fiber links. Each cluster consists of 2^n nodes in a hypercube topology. In *m* clusters, there will be $m \times 2^n$ nodes. Each node is represented by a 3-turple (i, j, k), as shown in Figure 1.

$i =$ local index, where $i \in (0, 2^n - 1)$,

$j =$ row global index, where $j \in (0, m_r - 1)$,

$k =$ column global index, where $k \in (0, m_c - 1)$.

m_r is the total number of rows, and m_c is the total number of columns. Figure 1 shows the SCOPIN network. Each node is connected to two global links, horizontal and vertical links (fiber link), that correspond to the two-torus links, while also connected to its local nodes. In effect each node has

- *n* local links (free space or space invariant optical links)
- Two global links (fiber links with WDM implemented)

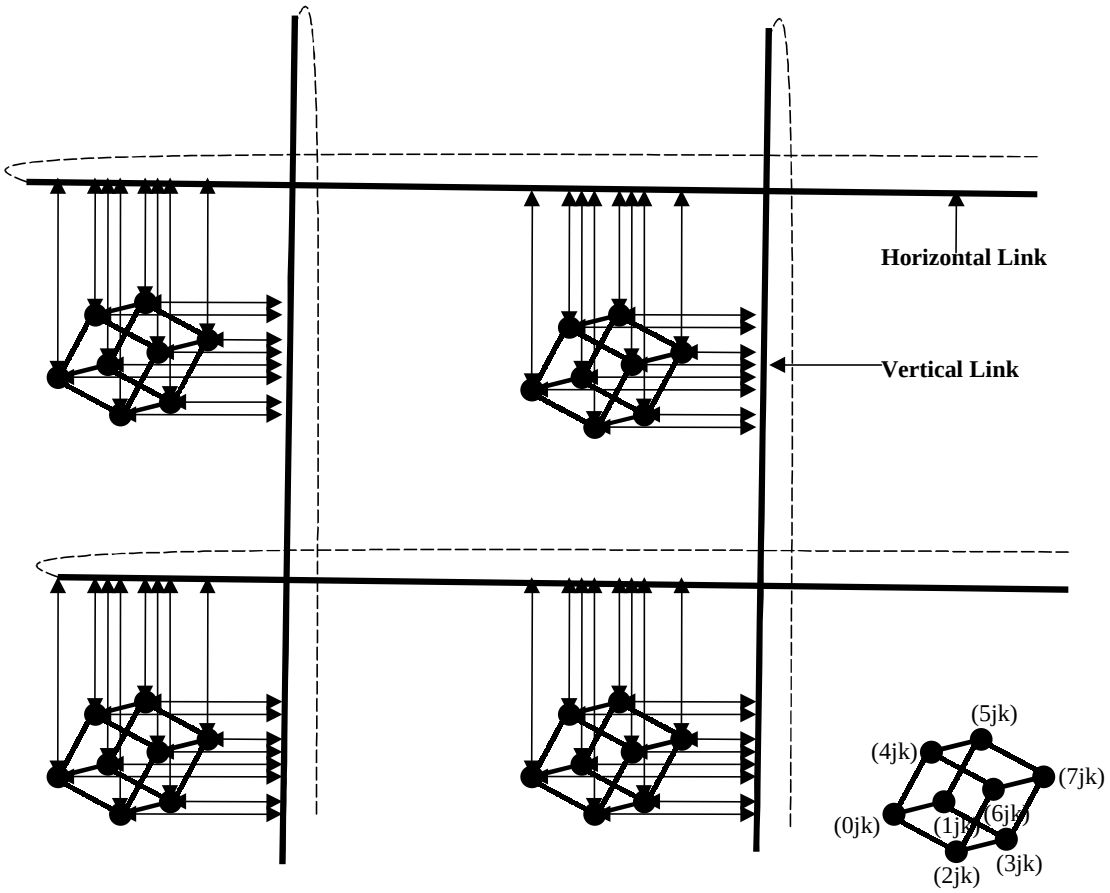

Figure 1. SCOPIN network logical layout

Partitioning

The idea behind partitioning stems from the fact that it is desirable to design a hierarchical structure using a cluster-based approach. This becomes quite obvious since it is intended that the interconnection network should be scalable, flexible and implemented with wavelength division multiplexing techniques.

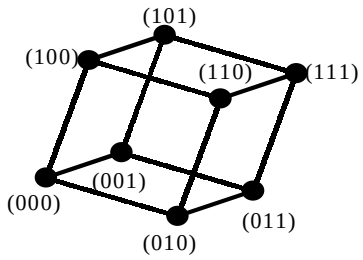

Figure 2: Hypercube Showing Binary Notation.

Partitioning involves using consecutive binary inverted partitions (CBIP). This process involves the partitioning of the set of transmitters and receivers into four (4) sets of transmitter-groups and receiver-groups. A close look at the ncube shows that four transmitter groups can be found such that in each group there are $^{2^n}/_4 = 2$ transmitters whose binary indexes have all their bits different. This is called an inverted partition, since the bit notation representing a transmitter in each group is inverted to get its neighbors' notation. The receiver groups are formed consecutively. In this case "consecutively" means that the elements in a group have their least significant bit (LSB) in binary indexes different. The idea is shown in Figure 3.

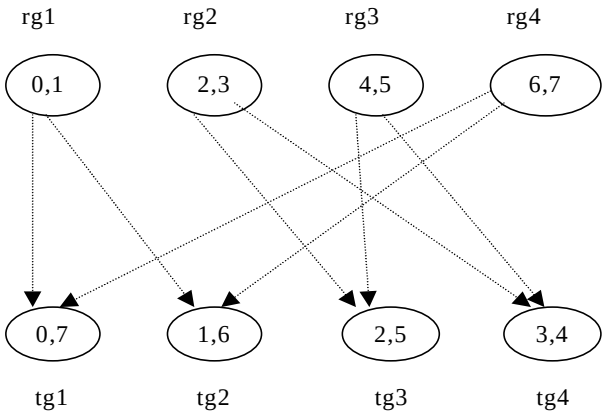

Figure 3. Hypercube decomposition into transmitter/receiver groups.

In our partitioning scheme, the local cluster is scalable by itself while still retaining the regular structure. The partition mechanism is scalable in the number of transmitters per transmitter-group and the number of receivers per receiver-group. The mechanism is also flexible in the creation of the network components, which provides the flexibility of the interconnection pattern. As will be shown, a number of bijections between the transmitter-groups and the receiver-groups are possible.

CBIP-component:

In this stage we formed a bijection by associating one transmitter-group exclusively with one receiver-group. An association formed by this bijection, is called a CBIP-component. The idea is shown in Figure 4.

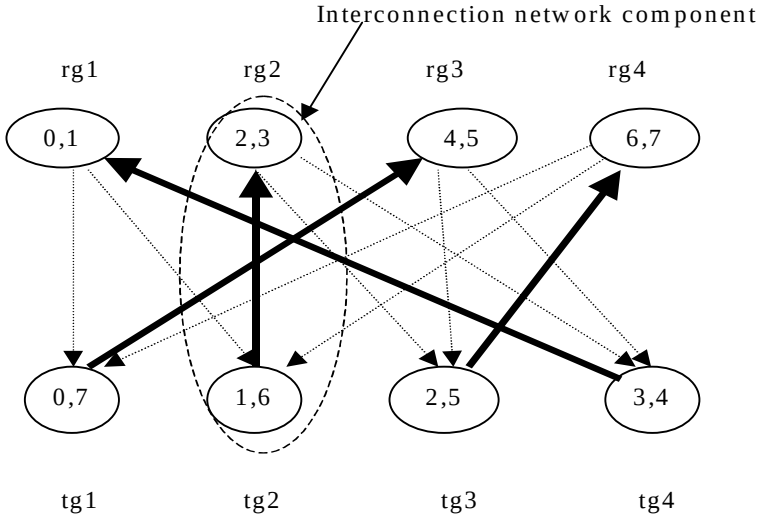

Figure 4. Interconnection network component

3. Optical Implementation

In order to justify our choice of scheme, we take a look at the proposed wavelength assignment. Our wavelength assignment scheme follows closely the decomposition mechanism developed in [18]. The objective is to minimize the tunability (number of wavelengths a transmitter can tune to) requirements of the transmitters while maintaining a high performance and still achieve the goal of single hop communication to all other nodes.

An extreme solution will be to have *N* unique wavelengths for each node in the network, with both the receivers and transmitters able to tune to all n wavelengths.

Our approach is to define a minimum set of nodes for a local structure. De-couple the transmitters and receivers into separate entities. Assign a unique wavelength to a transmitter. Apply our decomposition mechanism to obtain the interconnection network components. Identify, for each transmitter, the minimum set of additional wavelengths needed to achieve communication with every other node in the local cluster and hence all the nodes in the network

In our implementation, we assign wavelengths such that the tunability for the transmitters is minimum and optimal while the receivers must be able to tune to the maximum number of wavelengths used in the entire network. This is important to simplify the case where broadcast is needed as in the case in massively parallel computing.

An overview of the logical topology is shown in Figure 5.

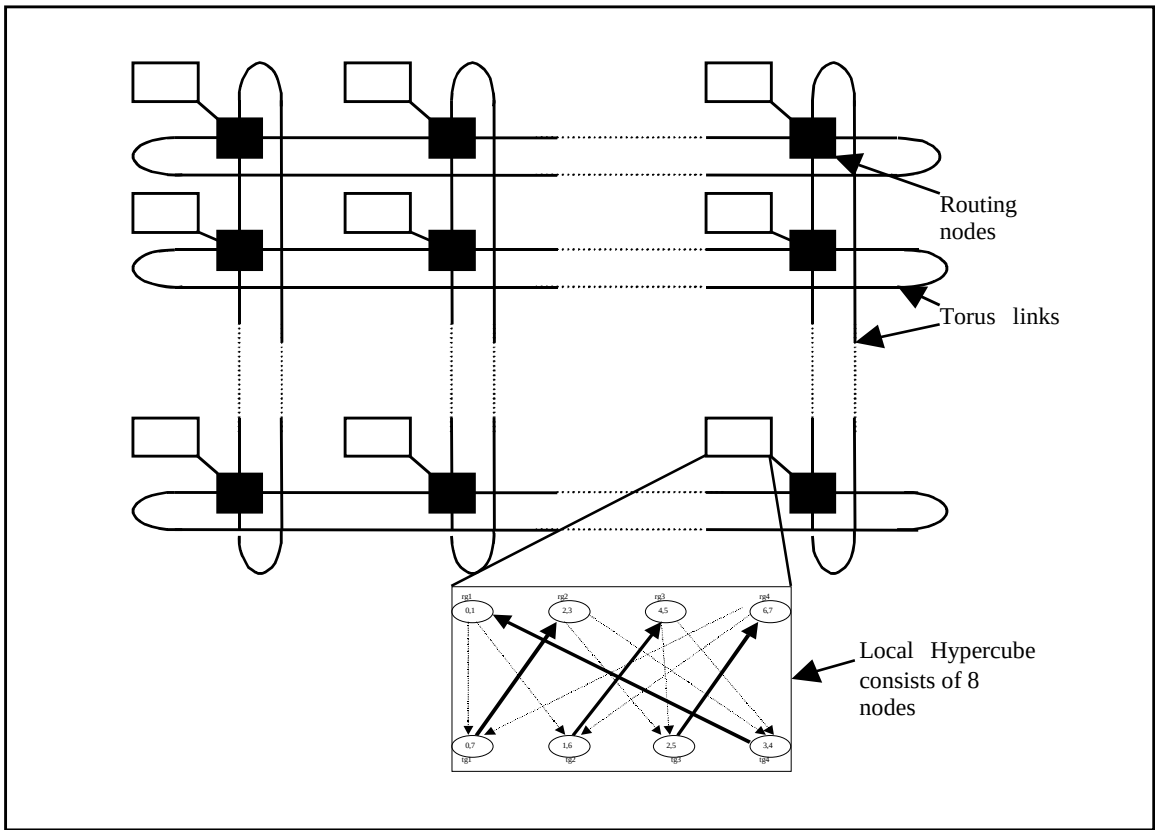

Figure 5: Diagram Showing Logical Topology of the Proposed SCOPIN Network.

For our design, each cluster has eight access nodes (eight transmitters and eight receivers). The physical topology is shown in Figure 6. Both transmitters and receivers are tunable. Depending on the number of access nodes in each cluster and on the design, it is expected that the optical interconnection be incorporated in the board design.

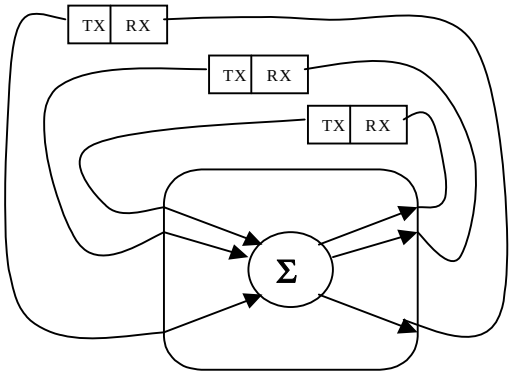

Figure 6: Diagram Showing Physical Topology of the Local Network.

SCOPIN local cluster optical implementation

In our implementation, we propose using a fast tuning laser with limited tuning range. As with high-bandwidth communications in parallel computers, the tuning latency must be very low. For the receivers, the tuning range should be higher, in view of the fact that we aim at implementing a broadcast-and-select type network. Tuning speeds must remain high while the latencies must be in the order of nanoseconds. In a broadcast-and-select network, all receivers always have all transmitted channels available on the incoming fiber. Also a broadcast-and-select network does not need any wavelength selective devices in the network. Instead, the receivers decide (on the basis of the protocol used) when to receive and/or which channel to tune in.

MQW-DBR [19] lasers are chosen for the transmitter. Biasing the active region provides laser oscillation. The phase and DBR regions are biased below threshold. Adjusting the current in the DBR region, which subsequently alters the light emitted, to the desired frequency, does the tuning. The receiver also uses MQW-DBR

lasers but biased below threshold. When exposed to a light source of the correct frequency, photodetection occurs in the active layer. This causes an external voltage across the forward bias diode.

The use of the same structure for both the transmitter and receiver is strategic. This will greatly simplify the coupling of the local structure. Each node can switch channels (frequencies) during execution by dynamically changing the injection current to the laser. Additionally, transmission and reception can be performed on different channels (because nodes are not assigned to a fixed channel). The bandwidth of the filter is divided into many high-speed data channels or virtual busses. Data is transfered serially over the medium using a message-based protocol and CSMA/CD arbitration. By using CSMA/CD, multiple processes operating on different processing nodes can coexist.

SCOPIN global link optical implementation

Consider the proposed SCOPIN network shown in Figure 7. Each local cluster (logically, each node) is linked to two global links. The global network is a torus network implemented using multi-mode fiber optic links. The key implementation here is to be able to connect remote clusters (nodes). Each cluster, in our design consists of eight transmitters and receivers linked to eight access nodes or PEs. Each cluster has access to the global links through a routing node. Figure 7 shows a logical arrangement of the local cluster and its connection to the global link.

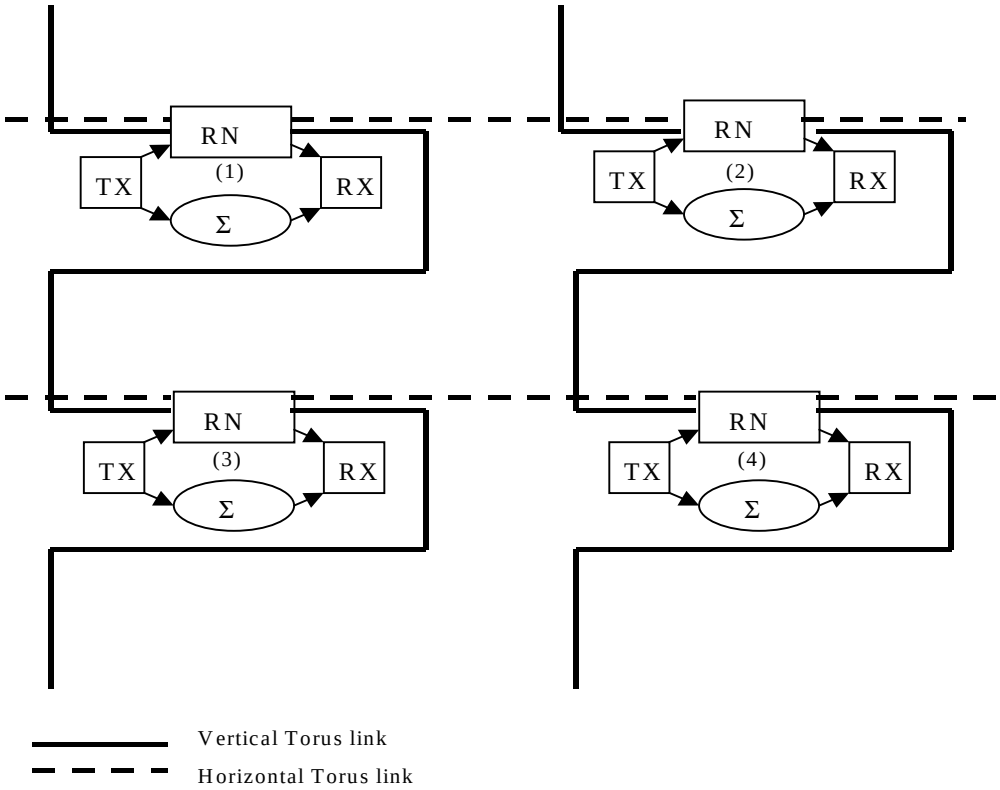

Figure 7: Illustration of SCOPIN Link Connections.

4. Bit Error Rate (BER) Estimation and Power Analysis

In this section, we present some insight to the power requirements and delay obtained in our design. We show some calculations to investigate the bit-error rate (BER) capabilities, optical losses, power budget and the delay.

Bit-Error Rate (BER) Estimation

Optical receivers will usually convert optical signals to electrical signals. An obvious important issue will be to determine the minimum power required at the receiver to maintain signal (information) integrity. For interprocessor optical interconnection networks, a BER of 10^{-15} or greater is largely agreed to be a current standard. The Signal-to-Noise (SNR) is used in calculating the BER. SNR is defined as the ratio of signal power to noise at the threshold point. It gives an indication of the expected BER of the system.

We assume a signal current I_s of 8μA, an MQW-DBR laser (biased below threshold as receiver) with quantum efficiency η of 80%. The total capacitance is 3pF, data rate of 2Gbits/s, operating wavelength of 1.55μm at 300^{o}K. From this we can achieve a minimum SNR of 392 corresponding to a BER in the order of 10^{-22}. This is well below the standard acceptable BER for digital communication of 10^{-15}. The minimum optical power at the receiver to achieve a BER of 10^{-15} is 3μW (-25dBm).

Optical Losses

It is necessary to determine if the required power will be available to the receiver. At a threshold current range of 10-12 mA, the output power reaches 100 mW at 500 mA active current. This corresponds to an optical power of 20dBm.

The SCOPIN routing occurs in two levels. For the local communication, the source and destination nodes have their transmitters and receivers coupled together using a free space star coupler. There will be losses incurred during the coupling of the beam emitted from the DBR laser into the optical waveguide. The insertion loss L_i, for a commercially available star coupler is taken to be -1 dB while the fiber-to-detector losses L_{fd} are –0.5 dB. We add another –3dB for engineering errors. If we assume n = 8, then the total losses for local communication is:

$$L_{local} = L_{split} + L_i + L_{fd} - 3 = -13.5dB$$

For remote communication, couplers are used to connect each local cluster via routing nodes to the global fiber link. These couplers are bi-directional *n x 1* star couplers. In the worst case a message may have to pass through all the routing nodes in any given horizontal and vertical ring of the torus system connecting source and destination. The coupling loss for the couplers is taken as –2dB. The total losses for the remote communication is:

$$L_{remote} = L_i + L_{fd} + L_c + k(L_{rn} + L_c) - 3$$

where L_c is the coupling loss between couplers and fiber, L_{rn} is the transmission loss through the routing nodes.

Power Budget

We have enumerated the losses that will be incurred for the two scenarios. For the local communication, the minimum power reaching the receiver is given by:

$$P_{out} = P_{in} + L_{split} + L_i + L_{fd} - 3,$$

hence $P_{out} = P_{in} + -13.5dB$.

If $P_{in} = 20dBm$, with the losses (-13.5dB) calculated above, $P_{out} = 4.5$mW (6.5dBm). This is way above the required received power for a BER of 10^{-15}, which is 3μW (-25dBm). For the remote communication, the minimum power reaching the receiver is given by:

$$P_{out} = P_{in} + L_i + L_{fd} + L_c + k(L_{rn} + L_c) - 3.$$

For a system of 2048 nodes, we require 256 routing nodes. If we assume that a message will travel through at most $\sqrt{256} = 16$ of these and the transmission losses in the routing nodes are negligible, then $P_{out} = P_{in} + -38.5dB = -18.5dBm$. This corresponds to a receiver power of 14μW, still above what is needed for a BER of 10^{-15}.

5. Comparative Evaluation of SCOPIN

Table 1. Comparison of 256 node networks

	Binary 8-cube	16-ary Torus	TESH (m, L, q)	SMLH (w, n)	SCOPIN (m, n, τ)
Node Degree	8	4	4	5	4
Diameter	8	30	21	5	1
Bisection width	256	32	8	32	4

Table 2. Comparison of 4096 node networks

	Binary 12-cube	64-ary Torus	TESH (m, L, q)	SMLH (w, n)	SCOPIN (m, n, τ)
Node Degree	12	4	4	5	4
Diameter	12	126	30	5	1
Bisection width	4096	128	128	128	16

A comparison of the various attributes of several known networks is given in tables 1 and 2 for 256 and 4096 nodes, respectively. The results show that the SCOPIN network has very good attributes and scales well.

6. Conclusion

This paper has introduced a new interconnection network, “Scalable OPtical Interconnection Network (SCOPIN)”, for multiprocessor systems. We have presented the design and initial analysis of the interconnection network. The architecture of the network is hierarchical, scalable, flexible, single-hop and amenable to optical

implementation. The network can be configured to a regular structure, which makes node addressing and message routing straightforward. The network has a constant node degree and a unity diameter. These attributes make the SCOPIN architecture suitable for massively parallel systems. We have also shown the feasibility of the proposed SCOPIN network. The network boosts a hierarchical structure involving implementing local clusters with free space star couplers and global links with fiber links. The results of the MQW-DBR laser indicate that a tunability of 17 nm is obtained around the wavelength of 1.56μm with 35 modes regularly spaced by 50 GHz. It was also shown that the BER achievable was considerably lower than the 10^{-15} BER standard. Analysis was also done with 2048 nodes. As the technology of the lasers and photodetectors improve, higher number of nodes will be possible without the use of optical amplifiers.

References

[1] T. M. Pinkston, "Design Considerations for Optical Interconnects in Parallel Computers," In *Proceedings of the First International Workshop on Massively Parallel Processing Using Optical Interconnects*, April 1994, pp 306-322.

[2] H. J. Siegel, "Interconnection Networks for Large Scale Parallel Processing," McGraw-Hill, 1990.

[3] K. Hwang, "*Advanced Computer Architecture: Parallelism, Scalability, Programmability*," New York: McGraw-Hill, 1993.

[4] P. B. Berra, A. Ghafoor, M. Guizana, S. J. Marcinkowski, and P. A. Mitkas, "Optics and Supercomputing," *Proc. IEEE*, vol. 77, pp. 1787-1815, 1989.

[5] A. Guha, J. Bristow, C. Sullivan, and A. Husain, "Optical Interconnections for Massively Parallel Architectures," *Applied Optics*, vol. 29, pp. 1077-1093, March 1990.

[6] A. D. McAulay, *Optical Computing Architectures: The Application of Optical Concepts to Next Generation Computers*. New York: Wiley, 1991.

[7] B. E. A. Saleh and M. C. Teich, *Fundamentals of Photonics*. New York: Wiley-Insterscience, 1991.

[8] M. Yang and L. M. Ni, "Incremental design of Scalable Interconnection Networks Using Basic Building Blocks," IEEE *Trans. Parallel and Distributed Computing*, vol. 11, no. 11, Nov. 2000, pp. 1126-39.

[9] A. Louri, B. Weech, C. Neocleous, " A Spanning Multichannel Linked Hypercube: A Gradually Scalable Optical Interconnection Network for Massively Parallel Computing," *IEEE Trans. on Parallel and Distributed Systems*, vol. 9, no. 5, May 1998.

[10] Ahmed Louri and Rajdeep Gupta, "Hierarchical Optical Ring Interconnections (HORN): Scalable Interconnections Network for Multiprocessors and Multicomputers," in *Applied Optics*, vol.36,no.2, pp 430- 442, January 10, 1997

[11] K. A. Aly, A. W. Dowd, "A Class of Scalable Optical Interconnection Networks through Discrete Broadcast-select Multi-domain WDM," *in Proc. IEEE INFOCOM'94*, (Toronto, Ontario Canada), pp. 392--399, June 1994.

[12] P.W. Dowd, K. Bogineni, K.A. Aly, J.A. Perreault, "Design and Analysis of a Hierarchical Scalable Photonic Architecture," 1994.

[13] V.K. Jain, T. Ghirmai, S. Horiguchi, "TESH: A New Hierarchical Interconnection Network for Massively Parallel Computing," *in Proc. IEICE Trans. Info. & Syst. Vol. E80-D*, no. 9, September 1997.

[14] S. S. Wagner and H. Kobrinski, "WDM Applications in Broadband Telecommunication Networks," *IEEE Comm.*, vol. 27, no. 3, pp. 22-30, 1989.

[15] D. R. Cheriton, H. A. Goosen, and P. D. Boyle, "Paradigm: A Highly Scalable Shared-Memory Multicomputer Architecture," *Computer*, pp. 33-46, Feb. 1991.

[16] Z. G. Vranesic, M. Stumm, D. M. Lewis, and R. White," Hector: A Hierarchically Structures Shared-Memory Multiprocessor," *IEEE Computer*, pp. 72-79, Jan. 1991.

[17] J.H. Laarhuis, "Multichannel Interconnection in All-Optical Networks," [*CTIT Ph. D. -thesis series No. 95-07], ISSN 1381-3617 / ISBN 90-365-0762-6*, September 1995, 332 pages.

[18] E. Okorafor and M. Lu, "A Decomposition Mechanism for Scalable Optical Interconnection Networks," Manuscript.

[19] M. S. Borella et al., "Optical Components for WDM Lightwave Networks," Proceedings of IEEE, vol. 85, pp. 1274-1307, Aug. 1997.

[20] E. Iannona and R. Sabella, "Optical Path Technologies: A Comparison Among Different Cross-Connect Architectures," *IEEE Journal of Lightwave Technology*, vol. 14, pp. 2184-96, Oct. 1996.

[21] P. E. Green, *Fiber Optic Networks*. New-Jersey: Prentice-Hall, Inc., 1992.