Scalable Opto-Electronic Network (SOENet)

Amit K. Gupta*, William J. Dally, Arjun Singh†, and Brian Towles‡
Computer Systems Laboratory,
Stanford University.
{agupta,billd,arjuns,btowles}@cva.stanford.edu

Abstract

In applications such as processor-memory interconnect, I/O networks, and router switch fabrics, an interconnection network must be scalable to thousands of high-bandwidth terminals while at the same time being economical in small configurations and robust in the presence of single-point faults. Emerging optical technology enables new topologies by allowing links to cover large distances but at a significant premium in cost compared to high-speed electrical links. Existing topologies do not cost-effectively exploit these optical links. In this paper we introduce **SOENet***, a family of topologies that exploits emerging high-speed optical and electrical links to provide cost effective scalability and graceful degradation in the presence of faults. We show that SOENet scales more economically than alternative topologies. For networks scalable to 32,000 nodes, a 32-node SOENet costs 4x less than a 3-D torus. Finally we investigate the fault tolerance properties of these networks and show that they degrade more gracefully in the presence of faults than alternative topologies.*

1 Introduction

Interconnection networks are widely used to connect processors and memories in multiprocessors [20], as switching fabrics for high-end routers and switches [9], and for connecting I/O devices [18]. Large scientific computers, e.g., ASCI White [1], have thousands of processors and large internet routers, e.g., the Avici TSR, are scalable to thousands of ports. These applications, and many others, demand a network that is *economically scalable*: a network that is inexpensive in small (less than ten node) configurations but can be incrementally expanded from this size to large (many thousands of nodes) configurations.

In addition to being economically scalable, an interconnection network must also provide service guarantees. In applications such as switches and routers, guaranteed bandwidth is essential because there is no backpressure and the router must deliver packets for arbitrary and even worst case traffic patterns. Bandwidth guarantees are also important in throughput-sensitive I/O networks.

For some applications, these service guarantees must be met even in the presence of equipment failure. When a network link or router fails, the performance of the network should degrade gracefully rather than fail abruptly.

Recent developments in high-speed electrical signaling [10] and parallel optical links [12] enable very high performance interconnection networks. These new technologies also change the design space of interconnection networks and greatly change the cost/performance equation. High-speed electrical links enable router chips with total pin bandwidth approaching 1Tb/s [13] at very low cost. However, these high-speed electrical links are limited to relatively short distances: about 1m over a backplane and about 10m over a cable. Parallel optical links extend the range of these high-speed signals up to 1 km. However, one pays dearly for these long links. The cost per unit bandwidth of an optical link is about 17× more expensive than for a high-speed electrical link (see Section 2).

To exploit this emerging high-speed signaling technology, we introduce *Scalable Opto-Electronic Network (SOENet)*, a family of network topologies that provides economical scalability and graceful degradation in the presence of faults while minimizing the number of long (and hence expensive) links.

A SOENet is constructed from many M-node local subnetworks, each of which is designed to be as large as practical without requiring long links. As shown in Figure 1, corresponding nodes of each local subnetwork are connected by long links to a global switch slice. The number of global switch slices is equal to the number of nodes per subnet-

*Supported by a grant from the Stanford Networking Research Center (SNRC) in the School of Engineering at Stanford University.

†Supported by the Richard and Naomi Horowitz Stanford Graduate Fellowship.

‡Supported by an NSF Graduate Fellowship with supplement from Stanford University and under the MARCO Interconnect Focus Research Center.

work. The global switch slices grow in size as subnetworks are added. A SOENet requires the smallest possible number of long links on a network of size N that must handle arbitrary traffic patterns with guaranteed throughput.

For networks that scale to 32K nodes, a SOENet costs about 4× less than a torus network in small configurations and about 2× less in large configurations. Our experiments show that a SOENet degrades gracefully in the presence of faulty channels. Failing 10% of the long channels in the network results in only a 13% degradation in throughput. Thus, SOENet can offer guranteed bandwidth in the presence of faults by providing a small amount of excess bandwidth.

SOENet builds on a long history that includes Clos networks [6], Beneš networks [5], and fat trees [15]. Our work extends these previous hierarchical networks by (a) optimizing the topology to exploit the cost and distance properties of modern signaling technology, and (b) introducing the concept of economical scalability and evaluating topologies on this metric.

The remainder of this paper describes SOENet and their properties in more details. Section 3 describes the topology of SOENet, how these networks are incrementally extended. Section 4 compares the cost of SOENet to other networks, and investigates their fault tolerance. Related work is described in detail in Section 5.

2 Signaling Technology for Interconnection Networks

Electrical signaling is inexpensive but limited in range. Modern ASICs are capable of driving and receiving several hundred differential pairs each of which operates at data rates of 3.125 Gb/s for an aggregate pin bandwidth approaching 1Tb/s [13]. Counting the cost of ASICs, backplane connectors, and backplanes (amortized over all of the links carried by each), the total cost of a bidirectional electrical backplane link is $6.51 per 3.125Gb/s pair[1], or about $2.08 per Gb/s.

Unfortunately a typical backplane link is limited to about 1m in range (distance) by the frequency-dependent attenuation of transmission lines fabricated on the backplane and printed circuit cards. For network topologies that require channels that are longer than this 1m limit for electrical backplanes, parallel optical links are an attractive but expensive technology [12, 2, 3]. Optical modules are currently available that provide 12 or 36 bidirectional 3.125 Gb/s channels in a single package. Such an optical link costs $111.46 per bidirectional 3.125Gb/s channel[2] or $35.67 per Gb/s, a factor of 17 higher than an electrical link. Thus, there is a need for network topologies that are designed to make efficient use of these costly resources.

[1]This cost is derived from quotes for Teradyne VHDM-HSD connectors and historical cost data on printed-circuit boards and ASICs.

[2]This cost is derived from quotes for an Infineon Paroli link and historical cost data for printed-circuit boards and ASICs.

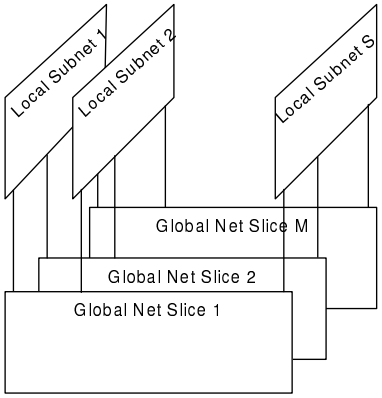

Figure 1. A SOENet consists of S subnetworks of M nodes connected by long links to M global network slices with S ports each.

3 SOENet and Routing

SOENets are designed to make efficient use of modern signaling technology by grouping nodes into *local subnetworks* that can be connected entirely using electrical links. Expensive optical links are only required to interconnect the subnetworks through a set of *global switch slices*.

3.1 Network Topology

As illustrated in Figure 1, a SOENet of size $N = SM$ consists of a set of S local subnetworks each containing M terminal nodes and M S-port global network slices which contain only switch nodes, no terminals. Each terminal node has one bidirectional connection to a global network slice. All other connections from the terminal node are to other terminal nodes within the same subnetwork. Each global network slice connects corresponding nodes across the subnetworks.

The subnetworks within a SOENet can be of any topology, with one link added to each node to connect to the global network slices. To realize the advantages of SOENets, the local subnetworks should be small enough that they can be realized using entirely short, electrical links. Long expensive links are only used for connections from the local subnetworks to the global network slices.

In our examples, we use torus networks (k-ary n-cubes) for our local subnetworks. An n-dimensional torus subnetwork with a terminal bandwidth of λ on each node requires a terminal node with degree $\delta_t = 2n + 1$, $2n$ bidirectional

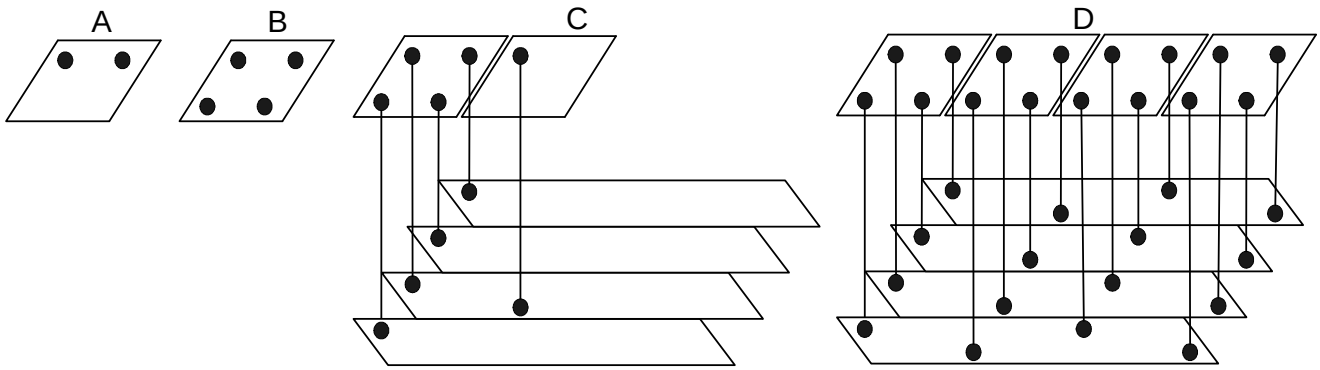

Figure 2. Scaling of a SOENet. For small configurations all nodes are in a single local subnetwork and no long links are required. As the node count increases, additional subnets start to be populated, and some long links are required. The global slices are populated incrementally as nodes are added.

links with bandwidth $s\lambda$ (where s is the link speedup) for the torus, and a single bidirectional link with bandwidth λ to the global network slice.

A radix k torus requires that each local link have a speedup of $s = k/4$ to handle worst-case traffic. That is, the bandwidth of each local link must be at least $b = s\lambda = \lambda k/4$, to handle worst-case traffic.

The global fabric slices can also be realized with an arbitrary topology and may themselves be SOENets for large configurations. In our examples we use torus networks or SOENets constructed from torus networks for the global network slices. Each global network slice node requires a degree of $\delta_s = 2n+2$, $2n$ links with bandwidth $s\lambda$, a single link with bandwidth λ the level of the hierarchy above this one, and a single link with bandwidth λ to the level of the hierarchy below this one (possibly the local subnetworks).

Routing in the SOENet is performed with a randomized algorithm similar to the one used in the CM-5[17, 16]. Unlike the CM-5, which has only *long* links in its network, our algorithm utilizes extra bandwidth on the short links to load balance the long links.

3.2 Scaling of SOENet

Figure 2 illustrates how a SOENet torus network scales from a single node to maximum capacity. For clarity the figure shows only four-node subnetworks (2×2 torus). Each node in the figure represents a 4×4 array of nodes in the following example in which each subnetwork and switch slice is an 8-ary 2-cube (8×8 torus).

A torus with $k = 8$ requires the speedup on the local links to be $s = k/4 = 2$. For networks smaller than 64-nodes, nodes are added one at a time to a single subnetwork as shown in the left-most panel of the figure. For these small configurations that fit in a single local subnetwork, no global network slices or long links are required.

Once the node count exceeds 64, the next 64 nodes are added to a second subnetwork as shown in the middle panel of Figure 2. At this point a backplane is provided for each of the M global switch slices and a single switch-only node (no terminals) is added to the appropriate global slice for each populated switch node in the local subnetworks. After the second subnetwork is fully populated additional nodes are added to a third subnetwork and so on. As each terminal node is added to a subnetwork, the corresponding switch node is added to a global network slice. The economic scalability of the network derives from the fact that the nodes of the global switch slices are added incrementally as the local subnetworks are populated. Small networks are not burdened with their cost.

When the node count reaches 4,096, each switch slice is fully-populated with 64 switch nodes as shown in the right-most panel of Figure 2. At this point further scaling requires making each of the switch slices itself a SOENet, adding a third layer to the hierarchy. Just as the jump from 64 to 65 nodes required adding a backplane for every global switch slice and one switch node for every terminal node, the jump from 4,096 to 4,097 nodes requires adding a second layer of global switch slice backplanes and a second switch node for every terminal node. Every terminal node has a corresponding node on each level of the hierarchy.

4 Results

4.1 Cost Comparison to Other Topologies

Figure 3 compares the cost of scalable networks as a function of the number of nodes. The figure compares three SOENets to a torus, a crossbar, and a Clos network (with $m = n$). The torus and SOENets scale to 32K nodes. The

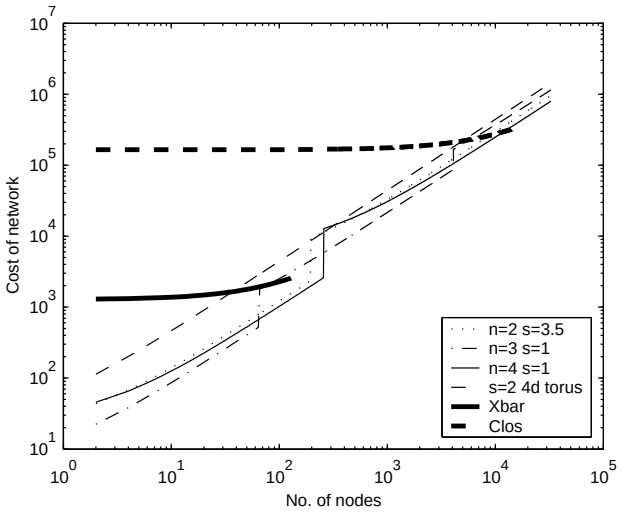

Figure 3. Cost of SOENet, torus, crossbar and Clos networks

Topology	Cost
Xbar	$N_{\max}G + NG$
Clos	$N_{\max}G + m \cdot \lceil \frac{N}{m} \rceil \cdot (G+1)$
4D Torus	$(2s+2G) \cdot N + \text{backplanecost}$
SOENet	$N(ns + \frac{G}{2} + l \cdot (ns+G)) + \text{backplanecost}$

Table 1. Equations for network cost

crossbar scales only to 128 nodes and the Clos scales only to 16K nodes[3]. Our cost model assigns half of the cost of a link (electrical or optical) to the node at each end of the link. Backplane costs include the cost of the printed-circuit board and the backplane portion of all connectors.

We assume that each of the subnets (whether local or global) has the same topology with paremeters n, the number of dimensions, and s, the local channel speedup. Using this model, the cost of a SOENet as a function of the number of nodes N and parameters s, n, and G is given by:

$$N \cdot ((ns + \frac{G}{2}) + l \cdot (ns+G)) + \text{backplanecost}$$

Here the first term gives the cost of electrical links in the local subnetworks and the cost of the local subnetwork side of the long link to the global network slice. The second term is the number of levels of hierarchy in the global network slices, l,multiplied by the cost per level. Each level of the global slice hierarchy has a cost of ns for the local links within the level, and a cost of G for the long links to the next higher and lower levels of the hierarchy. The number of levels in the SOENet is given by:

$$l = \lceil \frac{\log N}{\log M} - 1 \rceil$$

where M, the number of nodes in a subnet, is given by

$$M = (4 \cdot s)^n$$

The cost equations for each network are shown in Table 1. The results assume a cost ratio of $G = 10$ between the pins for optical and electrical links. The effect of varying the value of G is studied at the end of this section.

The SOENets have very low cost in small configurations because they only need to provide enough bandwidth to connect a local subnetwork. There is a steep jump in the cost of the SOENet at the point where the network size exceeds one subnetwork. This happens at 64-nodes for the $n = 3, s = 1$ network and at 256-nodes for the $n = 4, s = 1$ network. The $n = 3, s = 1$ network incurs a second jump at 4K nodes where a third level of hierarchy must be added.

The torus network is more expensive than the SOENet in small configurations for two reasons. First, because it must scale to $N_{\max}$ = 32K nodes without adding levels of hierarchy, it requires twice the channel speedup, doubling the cost of each node. Second, and more importantly, it requires long, expensive links even in small configurations since long links are required for all channels in two of the four dimensions. In small configurations, the SOENet is about 3.93 times less expensive than the torus network because it entirely avoids long links. In large configurations the SOENet is about 1.81 times less expensive than the torus because it needs only a single long link per node with unit bandwidth compared to the torus which requires two long links per node with $s = 2$.

The crossbar network shown has an unfair advantage in that it only has to scale to $N_{\max} = 128$ nodes. Even so, it is very expensive in small configurations. This is because the entire switch side of the crossbar including half of each global link, $N_{\max}G$, has to be paid for with the first node. While the crossbar is more efficient than the torus when it is fully populated, this would not be the case for a torus sized to scale to only 128 nodes. The Clos network only scales to $N_{\max}$ =16K nodes and is similar to the crossbar in that the entire middle stage of the 3-stage Clos must be paid for up front resulting in very high per-node costs in small configurations. We have assumed a flat 3-stage Clos network with no speedup. The factor m in Table 1 refers to the number of inputs to one node of the first stage of the Clos network. These nodes can be added incrementally as increasing number of terminal nodes are needed. We have used $m = 128$ to obtain our results.

[3]The crossbar is limited to 128 nodes because this is the largest switch that can be implemented with existing crosspoint chips [13]. Similarly the Clos is limited to 16K nodes because this is the largest 3-stage Clos that can be implemented with 128-port crosspoints.

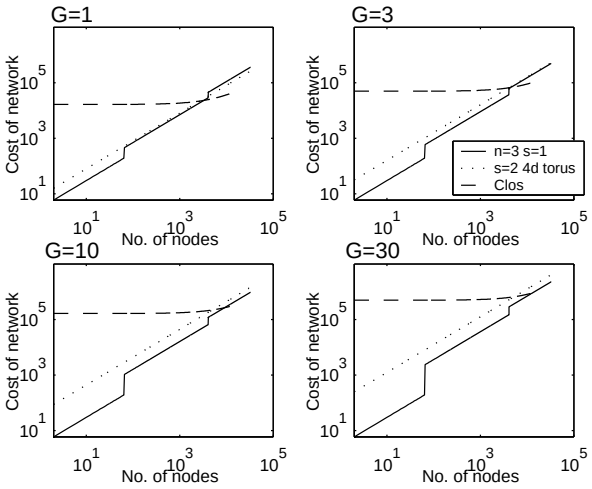

Figure 4. Network cost for various values of G.

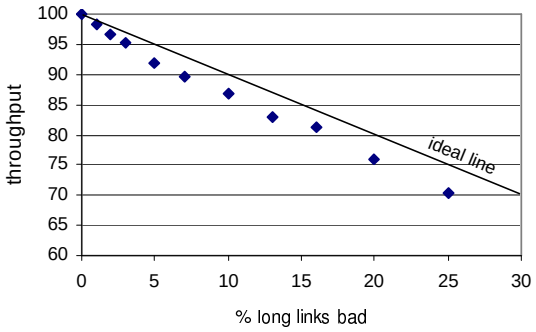

Figure 5. Fault tolerance: throughput of SOENets with some long links disabled

Figure 4 shows the cost of the various networks for different values of G. We note here that with $G = 1$, there is little advantage in using SOENet over torus networks. However, even with $G = 3$, we see that the cost of a torus network is larger than a SOENet for all network sizes. This advantage becomes more pronounced with increased values of G. The Clos network always has a very high startup cost for every value of G looked at.

4.2 Fault Tolerance

A SOENet provides a large number of edge and vertex disjoint routes between all pairs of nodes enabling it to provide a high degree of fault tolerance and to gracefully degrade performance in the presence of faults. This path diversity leads us to expect that if 10% of the long links were to be disabled, we would get a loss of 10% in throughput. To test this, simulations were performed disabling a varying percentage of the long links in the SOENet, and measuring the observed throughput compared to the base case where all links are functional. The results obtained from these simulations are presented in Figure 5.

As expected, the performance of the network does degrade gracefully. The presence of the faults in a few of the long links leads to a small loss in network throughput, and does not cause a sudden failure of the network. For example, a loss of 10% of the links leads to about a 13% drop in performance in the network. This number is slightly higher than the expected 10%, but that can be explained by considering the load on the electrical channels in the local subnet. The traffic to the nodes connected to the 'defective' long links would be reduced, resulting in higher traffic in the other parts of the subnet. This added congestion results in lower throughput of the network. One possible solution to this problem is to overprovision the local links, thus removing that bottleneck from the system and allowing full use of the more expensive optical links.

5 Related Work

We developed SOENet as a topology that exploits inexpensive, short-range, electrical links in combination with expensive, long-range, optical links to provide economic scalability.

Other papers have looked at the problem of finding efficient topologies given technology cost models and at the problem of fitting a topology to the available packaging technology. A method for adapting fat-tree networks to the bandwidth available in a given packaging technology is described in [14]. In [7] networks are assumed to be limited by wire bisection and low-dimensional k-ary n-cube networks are found to offer minimum latency under this assumption. Reference [4] combines a wire bisection limit with a pin limit and concludes that a slightly higher dimensionality is optimal. Reference [8] adds express channels to a k-ary n-cube to balance wire and router delay and shows how channel bandwidth can be matched to the packaging technology. In [19] the impact of pipelining the channels in a k-ary n-cube is studied.

SOENets are tree-like networks. Many variations on hierarchical, tree topologies have been proposed over the years. Clos networks [6] and Beneš networks [5] are non-blocking hierarchical networks. An X-tree [11] is a tree architecture with all nodes at a given level of the tree connected by a set of channels. A fat tree [15] is a tree network where the width or number of channels at each level of the tree is increased to reduce congestion near the root of the

tree. A fat-tree without dilation at each level is isomorphic to a Beneš network [5].

Our work on *SOENet* extends these previous results on matching topology to technology and on tree-structred networks in four key ways. First, our work is the first to consider the case where networks are constructed from two types of channels with different parameters, and in particular where a high cost is paid for long channels. Second, our work is the first to address the problem of enabling scalability to large numbers of nodes while keeping the cost of small configurations low. Third, our hierarcical networks differ from previous tree-structured networks such as fat trees by using a torus network to implement each node of the tree which enables much larger tree nodes for a given set of technology constraints and introduces *short* links into the network. Finally, our topology is the first that optimizes the parameters of the tree (node size and depth) to exploit the cost difference between short and long links.

6 Conclusion

SOENet enables networks to scale to a large number of nodes while being very economical in small configurations. With a SOENet, the cost of scalability — adding long links and network switch slices — is incurred only when the network scales to this size. The only cost in small configurations is one extra port per node. This is in contrast to crossbar and Clos networks, where nearly half the cost of the fully populated network must be paid up front, and torus networks where the port bandwidth of each individual node must be increased to enable scalability to large sizes. For networks that scale to 32K nodes, the cost of a 16-node SOENet is about one quarter the cost of a 16-node torus network.

SOENets are particularly well suited to contemporary electrical and optical interconnect technology. The topology exploits the inexpensive nature of short electrical links to construct the local subnetworks and switch slices. Expensive, but long-range, optical links are used only to connect levels of the hierarchy. The result is a network that uses fewer long, expensive links than alternative topologies, and hence a less expensive network for a given size and bandwidth. This efficient use of long links is reflected in the relative cost of different topologies in a maximal configuration. A 32K-node SOENet is about half the cost of a torus network of identical size.

While we have developed SOENet to exploit the properties of off-chip electrical and optical links, they can be applied to exploit the properties of other link types. In particular, they are well suited to networks that employ both short, on-chip interconnections and long, off-chip interconnections. The same techniques that make efficient use of long optical links can be used to best exploit the limited off-chip bandwidth of such networks.

A SOENet is fault tolerant, degrading gracefully in the presence of channel faults. Our simulations show that failing 10% of the channels results in a througput degradation of 13%. This graceful degradation enables the use of N+M redundancy in a SOENet.

References

[1] Asci White. http://www.llnl.gov/asci/platforms/white/.

[2] PAROLI Infineon Technologies Parallel Optical Link, Technical Description. http://www.infineon.com.

[3] Xanoptics XTM-72 Optical Transceiver Data Sheet. http://www.xanopics.com/xtm-72.pdf.

[4] A. Agarwal. Limits on interconnection network performance, 1991.

[5] V. Benes. Mathematical theory of connecting networks and telephone traffic, 1965.

[6] C. Clos. A Study of Non-Blocking Switching Networks. *The Bell System Technical Journal*, pages 406–421, 1953.

[7] W. Dally. Performance analysis of k-ary n-cube interconnection networks, 1990.

[8] W. Dally. Express cubes: Improving the performance of k-ary n-cube interconnection networks. *IEEE Transactions on Computers*, 40(9):1016–1023, Sept 1991.

[9] W. Dally, P. Carvey, and L. Dennison. Architecture of the Avici terabit switch/router. In *Proceedings of Hot Interconnects Symposium VI*, pages 41–50, 1998.

[10] W. Dally and J. Poulton. High performance electrical signaling. In *Proc.IEEE 5th International Conference on Massively Parallel Processing Using Optical Interconnects*, 1998.

[11] A. Despain and D. Patterson. X-tree: A tree structured multi-processor computer architecture, 1978.

[12] D. R. Engebretsen, D. M. Kuchta, R. C. Booth, J. D. Crow, and W. G. Nation. Parallel fiber-optic SCI links. *IEEE Micro*, 16(1):20–26, 1996.

[13] F. Heaton et al. A Single-Chip Terabit Switch. In *Proceedings of Hot Chips Symposium XIII, August 2001*.

[14] C. Leiserson. Vlsi theory and parallel supercomputing. In *Proceedings of the 1989 Decennial Caltech Conference on Advanced Research in VLSI*, pages 5–16.

[15] C. Leiserson. Fat-trees: Universal networks for hardware-efficient supercomputing. In *ICPP: 14th International Conference on Parallel Processing*, 1985.

[16] C. Leiserson et al. The Network Architecture of the Connection Machine CM-5. Proc. 1992 ACM Symposium on Parallel Algorithms and Architectures, 1992.

[17] C. E. Leiserson et al. The network architecture of the Connection Machine CM-5. *Journal of Parallel and Distributed Computing*, 33(2):145–158, 1996.

[18] G. Pfister. An Introduction to the InfiniBand Architecture. High Performance Mass Storage and Parallel I/O, IEEE Press, 2001.

[19] S. Scott and J. Goodman. Impact of pipelined channels on k-ary n-cube networks, 1994.

[20] S. Scott and G. Thorson. The Cray T3E network: adaptive routing in a high performance 3D torus. In *Proceedings of Hot Interconnects Symposium IV*, 1996.