Topology Optimization of Interconnection Networks

Amit K Gupta and William J Dally
Computer Systems Laboratory, Stanford University
{agupta, billd}@cva.stanford.edu

***Abstract*— This paper describes an automatic optimization tool that searches a family of network topologies to select the topology that best achieves a specified set of design goals while satisfying specified packaging constraints. Our tool uses a model of signaling technology that relates bandwidth, cost and distance of links. This model captures the distance-dependent bandwidth of modern high-speed electrical links and the cost differential between electrical and optical links. Using our optimization tool, we explore the design space of hybrid Clos-torus (C-T) networks. For a representative set of packaging constraints we determine the optimal hybrid C-T topology to minimize cost and the optimal C-T topology to minimize latency for various packet lengths. We then use the tool to measure the sensitivity of the optimal topology to several important packaging constraints such as pin count and critical distance.**

I. Introduction

Interconnection networks are today used in a variety of applications such as switch and router fabrics [6], and memory-processor interconnect [11]. Large scientific computers such as ASCI White [1], have thousands of processors and large internet routers, such as the Avici TSR, are scalable to thousands of ports. These applications, and many others, demand networks that can be incrementally expanded from small (less than ten node) configurations to large (many thousands of nodes) configurations.

Changes in signaling and packaging technologies over the past decade have led to a qualitative change in the ideal topology for an interconnection network. New high-speed electrical links [4] enable router chips with total pin bandwidth approaching 1Tb/s [9] at very low cost over short distances. Optical links [7] can be used to signal over long distances, but at significantly increased cost. As these signaling technologies change, the optimal topology for a given application changes to make the most cost effective use of current signaling technologies. To capture the effect of signaling technologies on interconnection network toplogy, we introduce a distance-bandwidth-cost model for electrical and optical interconnects. This model captures the reduction in electrical signaling rate with increased distance and the cost differential between electrical and optical signaling.

To solve the problem of picking the proper topology for a given set of signaling and packaging technologies, we have developed an automated tool for topology optimization. Our tool enumerates a family of possible topologies and selects the optimal topology from this family given a set of technology constraints. The technology constraints and the measure of optimality are specified by the user. Currently our tool explores the family of hybrid Clos-torus (C-T) topologies. This includes all possible torus networks, all possible Clos networks, and all possible Clos-torus hybrids. The tool can easily be extended to include other topology families.

Traditionally, network topology was determined in a largely ad-hoc manner after examining only a few alternative topologies. The topologies generated were often suboptimal and often failed to account for key technology constraints - e.g., critical signaling distance. Our automated approach to topology design evaluates a large set of possible topologies, accurately models technology constraints, and selects the optimum topology, from the family considered, to maximize a specified figure of merit.

The remainder of the paper explains the optimization process and presents results. Section II describes the packaging assumptions and our model of the signaling technology. Section III describes the various components of the optimization process and Section IV shows some of the results we have obtained from the topology optimization program.

II. Technology Model

In this paper we use a parameterized model of technology, and assume a hierarchical packaging of the network. Routers are packaged on integrated circuit chips which are in turn packaged, possibly with terminal nodes, on printed-circuit boards; a cabinet may contain several boards, and a large system is composed of several cabinets. The properties of this assembly are largely determined by the properties of the packaging technology in terms of size, density, pinout, and pin cost as listed in Table I. Here *size* is the footprint of one packaging level at the next packaging level. For example, the size of a board is the 30cm × 3cm connector footprint on the cabinet backplane, not its 30cm × 40cm surface area.

TABLE I
Size, density, and pinout of network packaging technologies.

Level	Size	Density	Pinout
Chip	4cm × 4cm	N/A	1024 signals
Board	30cm × 3cm	16 chips	2048 signals
Cabinet	60cm × 60cm	16 boards	10,000 signals
System	N/A	256 cabinets	N/A

The selection of the optimal topology depends strongly on the cost of channel bandwidth. We model the cost of a unit of bandwidth as a function of distance as shown in Figure 1. Electrical signaling bandwidth decreases with distance due to frequency-dependent attenuation. For board and backplane signals (where attenuation is dominated by

Manuscript submitted: 6 May 2005. Manuscript accepted: 6 July 2005. Final manuscript received: 13 July 2005.

dielectric absorption) bandwidth is constant up to a critical distance d_c and then proportional to the reciprocal of distance. Thus, the cost per unit bandwidth of electrical signaling is constant for distance $d \leq d_c$ and then increases linearly for $d > d_c$. For example, if we have a 40Gb/s signaling technology with a critical distance of 20cm, then at 20cm providing 40Gb/s of bandwidth requires one signal pair, while at 40cm providing the same bandwidth requires two signal pairs. At distance d_o, the cost of a unit of bandwidth has increased to the point d_o/d_c where it equals the cost of an optical signal. At distances $d > d_o$ cost is again constant because the critical distance for optical signaling is much larger than the maximum size of a system.

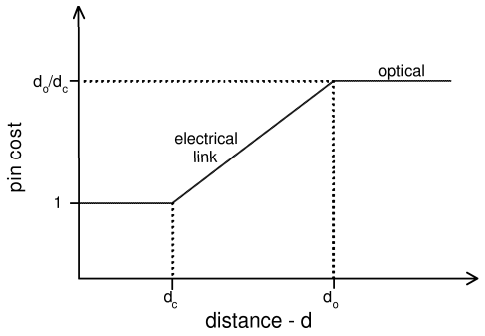

Fig. 1. Cost vs. distance for a fixed bandwidth link ($d_c = 20$cm).

III. Topology Optimization

Our tool performs a depth-first search to explore all possible topologies in the family being considered. The root of the search tree represents a single network terminal. Other nodes represent partial networks. At each node, child nodes are created for all possible topologies (in the family being considered). Each child node represents a new network constructed from multiple copies of the parent node's network with like nodes in each copy connected by the child node's topology. At each level, channel bandwidth is selected to meet the network's bandwidth specification and the cost of these channels is calculated using our technology model. The channels induced by each level of the tree are constrained to fit entirely in one level of packaging. Several levels of the tree may share a packaging level. Also, the topologies are defined so that a single ring, for example, may span several packaging levels – by composing sub-rings at each level.

In enumerating possible topologies we use *slicing* and *concentration* [5] to decouple the range of possible radices from pinout constraints. Concentration combines the traffic of several terminals into a single network node. For low degree network nodes (such as torus nodes), concentration allows full utilization of the pins available on the network node. This in turn reduces the number of nodes required and decreases the network cost. Slicing – the inverse of concentration – distributes the traffic from one terminal across several network nodes. Slicing is particularly useful in Clos networks, as it allows higher degree nodes to be used reducing the number of stages in the network. This reduces pin cost and hop count, but at the expense of additional serialization latency from the narrower channels.

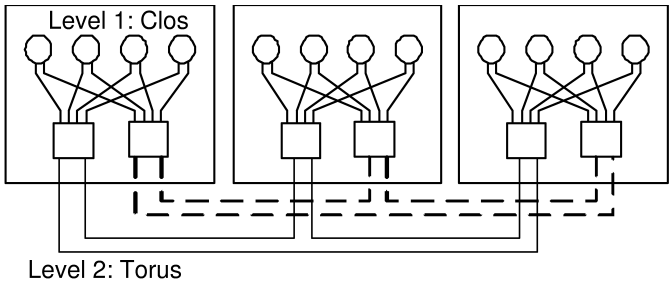

Fig. 2. A hybrid network – a 3-ring of 2-sliced 4-Clos subnetworks

For example, for the hybrid Clos-torus topology family, at the first level of the tree we create nodes for every possible crossbar (every possible radix, slicing, and concentration) and every possible ring (every possible radix, slicing, and concentration) that fits on a circuit board. Multiple rings can be composed to create tori [3] and multiple crossbars can be combined to form Clos[1] [2] networks. At the next level we consider every possible ring and crossbar to connect these subnetworks that still fits on a circuit board and also every possible ring and crossbar that fits at the cabinet level. The tree is expanded until the target network size is achieved. Leaf nodes are then evaluated for cost, latency, and other figures of merit. At each step, the topology is mapped to the packaging, and checked for feasibility. Topologies that cannot satisfy the physical constraints of density or pinout are pruned.

For example, the topology associated with a second-level node that represents a radix-3 torus with no slicing packaged on a backplane and that has a parent representing a radix-4 Clos with a slicing of 2 packaged on a board is shown in Figure 2.

IV. Results

A. Optimizing for Cost

Our optimizer accepts 15 parameters as input. A list of these parameters and their default values is shown in Table II.

Running the optimizer with the default parameters, varying the $\max N$ parameter from 1 to 1024, and optimizing for cost gives the phase diagram of Figure 3. The horizontal axis of this figure shows the value of $\max N$, the maximum number of nodes the network can scale to. The vertical axis shows the number of nodes packaged at a particular level of the tree. Each region of the phase diagram shows the topology selected at a particular level. For example, for $16 < \max N \leq 40$, the first 10 nodes are packaged as a 5×2 torus on a board. A ring of up to four boards is then used to scale from 10 to 40 nodes.

The figure shows that, for this set of parameters, the optimal network for less than 40 nodes is a torus. A 2-D single-board torus is used for less than 16 nodes, and a 3-D torus (up to $5 \times 2 \times 4$) is used for up to 40 nodes. At these small

[1]The Clos network is folded to simplify packaging and allow the network to take advantage of locality. These networks are also referred to as fat trees [10].

TABLE II

DEFAULT VALUES OF INPUT PARAMETERS TO THE OPTIMIZER PROGRAM

Parameter	Value	Units
Network BW and size targets		
minN	1	
maxN	256	
node_BW	4000	Gb/s
packet_length	128	bits
cost_weight	100	100 for cost optimal 0 for latency optimal
Technology constraints		
pins_chip	1024	
pins_connector	2048	
chips_board	16	
boards_backplane	16	
Technology parameters		
signal_rate	40	Gb/s
critical_distance	20	cm
optical_cost	10	
Latency = fixed + $8_{factor} \times log_8(p)$		
clock_cycle	0.8	ns
delay_fixed	4	clock cycles
delay_8factor	1	clock cycles

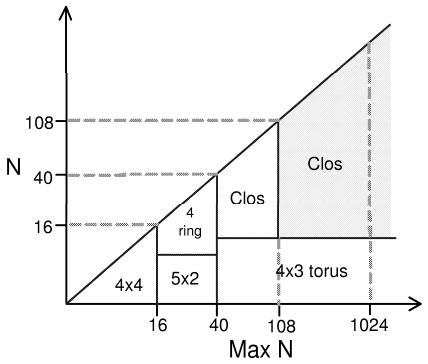

Fig. 3. Optimal network topology for various network sizes when the network is optimized for pin cost. Unshaded regions represent electrical links, whereas shaded regions represent optical links.

values of $\max N$ a torus best exploits the inexpensive high-speed electrical signaling on the board and over the backplane. For $\max N > 40$, a 4×3 single-board torus is used at the bottom level with these 12-node subnetworks connected by a central Clos network. A Clos with electrical signaling is used for $\max N \leq 108$ and an optical Clos is used for networks with $\max N > 108$. Beyond 108 nodes there is no backplane-level connection, the boards are directly connected to a central switch using optical links.

Figure 4 compares the cost of the optimal hybrid Clos-torus network selected by our optimizer with the cost of the optimal torus and the optimal Clos network as network size N is varied. As discussed above, for small N, the cost-optimal network is a torus. Above 40 nodes, the optimal hybrid C-T network shows a cost advantage. At 1K nodes, the optimal Clos network is 16% more expensive and the all-torus network is 39% more expensive than the hybrid C-T network.

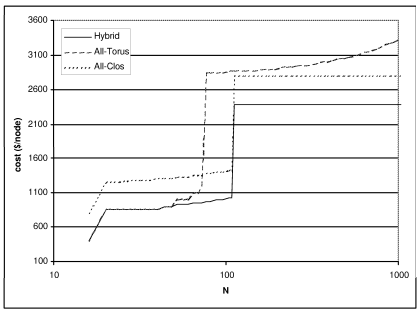

Fig. 4. Comparing the cost of the optimal network determined by the optimizer to networks using only torus or only Clos topologies.

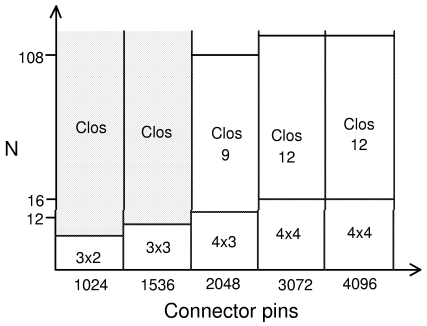

Fig. 5. Optimal network topology for a network of 108 nodes when the network is optimized for pin cost as the number of pins on the board connector varies.

B. Varying the pin constraints

Our topology optimizer can also be used to measure the sensitivity of topology to technology parameters. Pin constraints – off-chip, off-board (connector pins), and off-cabinet – are key parameters that determine how much of a network can be packaged in a given module (board or cabinet). Figure 5 shows how the cost-optimal network topology changes as we vary the board-pin constraint from 1024 to 4096 signals per board. With less than 2K connector pins, only a small 6- or 9-node torus can be packaged on a single board and the remainder of the network is implemented as a Clos with optical signaling. This low density drives up the required signaling length, making electrical signaling expensive and further stressing the pin constraint. Above 2K connector pins per board, it is more cost-effective to use electrical signals to implement the Clos that connects 12- or 16-node tori on each board.

The pin constraint also affects the degree of the Clos network. At 2K pins/board, the Clos is limited to degree 9, above 3K pins a degree-12 Clos is feasible. Above 3K connector pins, there is no change in the optimal topology. At this point the network is no longer connector-pin constrained.

C. Varying signaling technology

Figure 6 shows how the cost-optimal technology changes as we vary the critical distance d_c from 10cm to 40cm. With a critical distance of 10cm, the topology is a 16-node torus on the board and a global Clos with optical signaling. With this short d_c, the cost and pin-count required for electrical

backplane interconnection becomes high and an optical Clos network is required. With d_c =20cm (our default case), the number of pins required for electrical signaling is reduced to the point that it becomes less costly than optical signaling. However, the lower signaling density (more pins required for the same bandwidth) forces the use of a smaller (12-node) torus on each board. As the critical distance increases a larger degree Clos (at d_c = 30cm) and a larger torus (at d_c = 40cm) become possible.

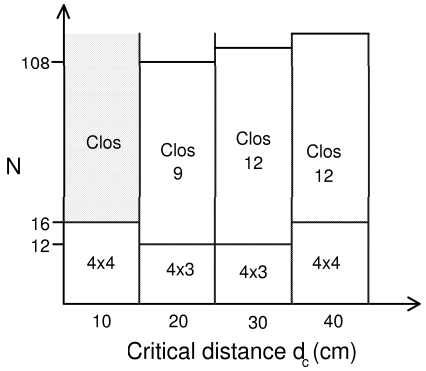

Fig. 6. Optimal topology for a network of 108 nodes when optimized for pin cost as the critical length varies.

D. Optimizing for latency

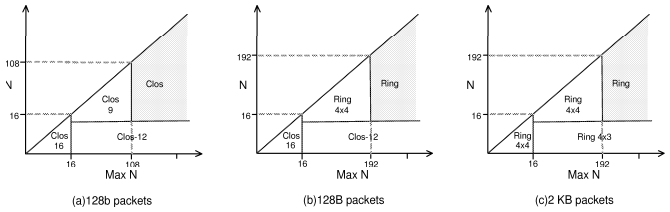

Fig. 7. Latency-optimal topology vs network size for various packet sizes.

Figure 7 shows the latency-optimal topology as a function of network size for various packet lengths. We assume uniform random traffic and the use of the randomized routing algorithm described in [8]. When optimizing for latency, packet size greatly affects the choice of topology. For small (128 bit) packets, Clos networks are favored for their low diameter. At very large (2K Byte) packet sizes torus networks are favored because their low degree enables wide channels and hence low serialization latency. At an intermediate point of 128 Byte packets, Clos networks are used to connect the nodes on a board with rings used to connect the boards together. This is a reversal of the cost-optimal design that uses tori locally and Clos globally. For all cases off-board optical links are used above a critical size.

V. Conclusions and Future Work

We have described an automated tool that determines the optimal topology from a topology family given technology parameters and design goals. Our tool incorporates a signaling cost model that captures the reduction in electrical signaling bandwidth with distance and the cost premium of optical signaling. The tool computes the optimal topology (from the family considered), the embedding of this topology into packaging levels, and the selection of technology (optical or electrical) for each channel. Using our tool we have calculated the cost-optimal and latency-optimal topologies for a representative set of technology constraints. We have also explored the sensitivity of topology to particular constraints such as board connector pin count and critical signaling distance.

This approach promises to rationalize the design of interconnection network topologies by permitting large spaces of possible topologies to be quickly explored and quantitatively compared in terms of cost and latency for a required throughput. The automated exploration of topologies also gives a tool to assess the importance of new packaging and signaling technologies by measuring the sensitivity of topology (and cost) to technology parameters.

We have only scratched the surface of automated topology exploration with this paper. Using our existing tool, many more analyses can be performed. For example we can optimize for minimum cost with a maximum latency constraint. We can also optimize for specific traffic patterns or sets of traffic patterns - rather than for flat bandwidth. While our current tool is limited to exporing hybrid C-T topologies, it can easily be extended to handle other topologies, such as butterflies, Cayley graphs, and hybrids using these topologies. The tool can easily be expanded to consider new topologies, new technologies, and new analyses as the need arises. Ultimately we expect this concept of automated design-space exploration to be applied to other areas of computer architecture beyond interconnection networks.

References

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

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

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

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

[5] W. Dally and B. Towles, *Principles and Practices of Interconnection Networks*. Morgan Kaufmann, 2003.

[6] W. Dally, P. Carvey, and L. Dennison, "Architecture of the Avici terabit switch/router," in *Proceedings of Hot Interconnects Symposium VI, August 1998*, 1998, pp. 41–50.

[7] D. R. Engebretsen, D. M. Kuchta, R. C. Booth, J. D. Crow, and W. G. Nation, "Parallel fiber-optic SCI links," *IEEE Micro*, vol. 16, no. 1, pp. 20–26, 1996.

[8] A. K. Gupta, W. J. Dally, A. Singh, and B. Towles, "Scalable opto-electronic network (soenet)," in *proceedings of Hot Interconnects (HotI) X*, Stanford, California, USA, August 2002.

[9] F. Heaton et al, "A single-chip terabit switch," in *Proceedings of Hot Chips Symposium XIII*, 2001.

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

[11] 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.