Efficient Topologies for Large-scale Cluster Networks

John Kim
KAIST
Daejeon, Korea
jjk12@kaist.edu

William J. Dally
Stanford University
Stanford, CA
dally@stanford.edu

Dennis Abts
Google Inc.
Madison, WI
dabts@google.com

ABSTRACT

Increasing integrated-circuit pin bandwidth has motivated a corresponding increase in the degree or radix of interconnection networks and their routers. This paper describes the *flattened butterfly*, a cost-efficient topology for high-radix networks. On benign (load-balanced) traffic, the flattened butterfly approaches the cost/performance of a butterfly network and has roughly half the cost of a comparable performance Clos network. The advantage over the Clos is achieved by eliminating redundant hops when they are not needed for load balance. On adversarial traffic, the flattened butterfly matches the cost/performance of a folded-Clos network and provides an order of magnitude better performance than a conventional butterfly. In this case, global adaptive routing is used to switch the flattened butterfly from minimal to non-minimal routing — using redundant hops only when they are needed. Different routing algorithms are evaluated on the flattened butterfly and compared against alternative topologies. We also provide a detailed cost model for an interconnection network and compare the cost of the flattened butterfly to alternative topologies to show the cost advantages of the flattened butterfly.

1 Introduction

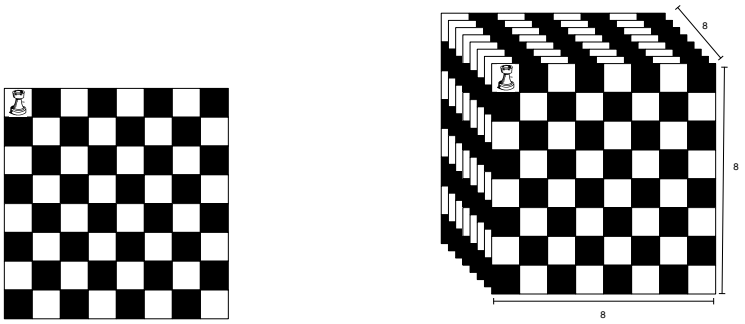
The cost of power and its associated delivery and cooling are a major factor in the total cost of ownership in a large-scale cluster. The interconnection network plays a critical role in the cost and performance of a scalable multiprocessor. Previous interconnection networks have been built with low-radix routers – e.g. routers with a small number of ports. As a result, the topologies used were also low-radix such as 2-D or 3-D mesh or torus networks and examples of such networks include the Cray T3D, T3E, and XT3. Earlier work [4, 2] showed that low-radix networks provide optimal latency for a given cost. Given the relatively low pin bandwidth available during the 80s and the early 90s low-radix routers were suitable.

However, over the past 20 years, the pin bandwidth of router chips has increased by approximately an order of magnitude every 5 years [8] – a rate very similar to Moore's Law. The increase in bandwidth is a result of both the increase in the signaling rate as well as the increase in the number of signals. *High*-radix routers have been shown to take advantage of this increasing bandwidth by dividing the bandwidth into larger number of narrow ports instead of low-radix routers where the bandwidth is divided into a smaller number of wide ports [8]. The Cray BlackWidow system [1] is one of the first systems to take advantage of high-radix routers. The topology in the BlackWidow system [10] which is a variant of the high-radix folded-Clos topology is a significant departure from previous low-radix networks that employed 2-D or 3-D torus topologies.

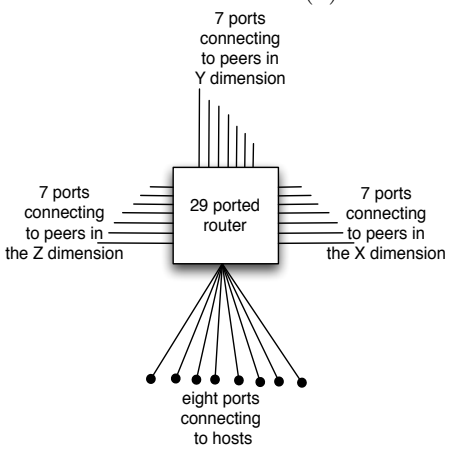
Topology is a critical aspect of any interconnection network as it establishes performance bounds for the network since topology determines the network diameter as well as bisection bandwidth. The topology also largely determines the cost (both capital cost and power consumption, or operating expense) of the system as well. In this paper, we analyze topologies of high-radix networks and describe the flattened butterfly topology – a cost-efficient topology for high-radix networks [7]. Existing topologies such as folded-Clos or fat-tree pay too much penalty on load-balanced traffic (e.g. uniform random) to provide good performance on adversarial traffic pattern. Other topologies such as the conventional butterfly network can exploit high-radix routers to provide low cost but because of lack of path diversity, the performance is severely limited on adversarial traffic pattern. The flattened butterfly attempts to approach the cost of a conventional butterfly network while providing the cost/performance of a folded-Clos topology.

The flattened butterfly topology can be easily derived by starting with a conventional butterfly network and combining or *flattening* all the routers in each row of the network. In terms of the inter-router connections, the same connections that existed in the conventional butterfly are maintained in the corresponding flattened butterfly. As a result, if only minimal routing is used on the flattened butterfly, the topology behaves identically to a conventional butterfly. However, by utilizing non-minimal adaptive routing (such as the global adaptive routing [11]), the full benefits of the flattened butterfly can be explored. For load-balancing, non-minimal routing can be utilized to achieve path diversity that approaches the path diversity provided in a folded-Clos network. Although non-minimal routing can imply higher hop count, non-minimal routing on the flattened butterfly does not incur higher hop count compared to a comparable folded-Clos. In a folded-Clos, all the packets first need to be routed up to the near common ancestor before routing back down to its destination. These extra hops are not required with minimal routing on the flattened butterfly but are added for non-minimal routing to achieve load-balancing. The added extra hops in the folded-Clos, compared to the flattened butterfly, also results in approximately 2× increase in cost as it requires one set of links to connect to the middle stages and another set of links to connect to the destination router – whereas in the flattened butterfly, the routers are directly connected to each other without connecting through intermediate routers.

An example illustrating the benefit of flattened butterfly is shown in Figure 1. A 64 node 2-D flattened butterfly can be described with how a rook moves in a conventional chessboard with each square representing a router (Figure 1(a)). The rook can reach any square within its column or row – similarly, the direct connections in a flattened butterfly allows a packet to traverse to any other router within each dimension with a single hop. A 512 node 3-D flattened butterfly can be described with



(a) 8×8 chess board (b) stack of chess boards



(c) 29-ported switch chip which is required to build an 8-ary 4-flat.

Figure 1: A packet traverses an 8-ary 3-flat (with 512 nodes) in the same way that a rook moves on a chess board (a), we scale the network by adding another dimension (b) making a stack of chessboards. Each square on the chess board represents a switch chip. A 29-ported router (c) is required to build the 8-ary 4-flat with 4k nodes.

a stack of chessboard (Figure 1(b)) with a diagram of the 29-port router needed to scale to 512 nodes (Figure 1(c)). By reducing the diameter of the network, high-radix networks (and the flattened butterfly topology) are advantageous for lower latency and reduced power as the number of intermediate routers is greatly reduced.

The flattened butterfly is similar to the generalized hypercube [3] that was proposed in the early 80s. However, one main difference is that the flattened butterfly utilizes concentration which significantly reduces the wiring complexity of the topology and can more efficiently exploit high-radix routers. Other differences between the two topologies are explained more in detail in [7].

Minimal routing or non-minimal routing with oblivious load-balancing does not provide good performance across different traffic patterns on the flattened butterfly. Global adaptive routing is essential to maximize the performance of the flattened butterfly. Compared to the conventional butterfly, on benign traffic pattern, the flattened butterfly achieves similar performance but provides 2× increase in throughput compared to a folded-Clos of similar cost. On adversarial traffic pattern, the flattened butterfly provides similar performance as the folded Clos but provides over a magnitude increase in performance compared to the conventional butterfly.

A detailed cost model comparing alternative topologies is also presented in [7]. The cost model takes into account the main components of an interconnection networks, the routers and the cables, and incorporates the packaging hierarchy of the network into the cost model. As expected, the cost of interconnection networks is dominated by the cables, accounting for as much as 80% of the total network cost. The flattened butterfly which reduces the number of cables in the network provides significant savings over the folded Clos network. However, instead of the expected 2× reduction in cost, for a network of size $1K$, the cost savings for a flattened butterfly is only approximately 43%. Although the flattened butterfly reduces the number of *global* cables, it does not reduce the number of local cables such as the backplane traces from the processors to their routers. However, for a system of size $4K$, the cost savings of the flattened butterfly exceeds 50%. The additional savings come from *packaging locality* that the flattened butterfly can exploit – e.g. the routers in the lowest dimensions are fully connected and can be packaged locally where as in a folded-Clos, all of the routers need to be connected to a set of centralized routers corresponding to the middle stages.

2 Significance

Processor and memory technology have advanced according to Moore's Law – with the number of available transistors growing exponentially with time. This scaling has made interconnection networks more critical as wire density has scaled at a slower rate and wire delay has remained constant over time. Hence, the network is a major factor in determining the overall performance and cost of the system. It has been shown high-radix networks can exploit the increasing pin-bandwidth to provide a lower latency and a lower cost network [8]. This paper presents the flattened butterfly topology, a cost-efficient topology for high-radix

networks that not only exploit high-radix routers but provides a much more cost-efficient topology compared to alternative topologies.

The migration towards high-radix networks has been shown with the development of the Cray BlackWidow system [1, 10] – one of the first high-radix networks built. The BlackWidow network employs a variant of the high-radix folded Clos network and reduces the network cost by employing *sidelinks* to directly connect neighboring subtrees for some configurations. The flattened butterfly topology extends the concept of *sidelinks* to provide a topology where all of the inter-router links are sidelinks. This removes the need for intermediate routers and thus not only reduce latency but also the cost of the network. As a result, the flattened butterfly provides approximately 2× cost improvement compared to a folded-Clos topology on benign traffic while achieving similar cost/performance ratio on adversarial traffic patterns.

The flattened butterfly topology exploits high-radix routers while taking advantage of *packaging locality*. Since all of the routers in the lowest dimension are fully connected to each other, they can be packaged locally with short cables and the network cost can be significantly reduced.

Another significant impact of this work is that the flattened butterfly topology can be extended to on-chip networks for multicore processors [5]. The use of flattened butterfly for on-chip networks provides a new approach to the design of on-chip networks, compared to a conventional 2-D mesh network. Existing on-chip networks such as a ring or a crossbar will not scale and the 2-D mesh network will not scale efficiently in terms of power or performance. By utilizing concentration and direct links between routers, the topology reduces intermediate routers – resulting in not only lower latency but lower power consumption as well.

By reducing the diameter of the network, high-radix networks are advantageous both for latency and power as well as cost. However, high-radix networks are difficult to realize on-chip in a cost-efficient implementation. The flattened butterfly with two dimensions maps well to a VLSI layout and results in a topology where any two terminals are separated by only two hops – thus, providing a network that approaches the *ideal* latency for an on-chip network [9]. The on-chip flattened butterfly also exploits the on-chip constraints such that by using the bypass channels, non-minimal routing can be achieve without traveling non-minimal physical distance.

We extend the flattened butterfly topology to take advantage of additional packing locality and fewer optical links with the dragonfly topology [6]. The dragonfly topology introduces the critical need for adaptive routing to properly load balance; making the tradeoff of fewer links requiring more complex global adaptive routing to load balance the global channels. Global adaptive routing (UGAL) [11], can perform such load balancing if the load of the global channels is available at the source router, where the routing decision is made. With the dragonfly topology, however, the source router is most often not connected to the global channel in question. Hence, the adaptive routing decision must be made based on remote or *indirect* information. The indirect nature of this decision leads to degradation in both latency and throughput when conventional UGAL (which uses local queue occupancy to make routing decisions) is used. The work in [6] describes two modifications to the UGAL routing algorithm that overcome this limitation with performance results approaching an *ideal* implementation using global information. Adding selective virtual-channel discrimination to UGAL (UGAL_{VC_H}) eliminates bandwidth degradation due to local channel sharing between minimal and non-minimal paths. Using credit-round trip latency to both sense global channel congestion and to propagate this congestion information upstream (UGAL_{CR}) eliminates latency degradation by providing much stiffer backpressure than is possible using only queue occupancy for congestion sensing.

Over time, interconnection networks will become more critical to system performance and the size of networks will continue to increase. Hence this work on high-radix networks and topology will become even more significant over time. This work is relevant to the networks used in all types of digital systems – e.g., server clusters, internet routers, and storage-area networks as well as supercomputers. For example, many commercial high-performance interconnect such as Myrinet, InfiniBand, and Quadrics implement a fat-tree topology. However, by utilizing high-radix routers (radix-64 or higher) and the flattened butterfly topology, the cost of the networks in SAN or clusters can be reduced by 2× by reducing the number of cables.

REFERENCES

[1] Dennis Abts, Abdulla Bataineh, Steve Scott, Greg Faanes, James Schwarzmeier, Eric Lundberg, Tim Johnson, Mike Bye, and Gerald Schwoerer. The Cray BlackWidow: A Highly Scalable Vector Multiprocessor. In *Proc. of the International Conf. for High-Performance Computing, Network, Storage, and Analysis (SC'07)*, Reno, NV, November 2007.

[2] A. Agarwal. Limits on Interconnection Network Performance. *IEEE Trans. Parallel Distrib. Syst.*, 2(4):398–412, 1991.

[3] Laxmi N. Bhuyan and Dharma P. Agrawal. Generalized hypercube and hyperbus structures for a computer network. *IEEE Trans. Computers*, 33(4):323–333, 1984.

[4] W. J. Dally. Performance Analysis of k-ary n-cube Interconnection Networks. *IEEE Transactions on Computers*, 39(6):775–785, 1990.

[5] John Kim, James Balfour, and William J. Dally. Flattened Butterfly for On-Chip Networks. In *IEEE/ACM International Symposium on Microarchitecture (MICRO)*, Chicago, IL, December 2007.

[6] John Kim, Wiliam J. Dally, Steve Scott, and Dennis Abts. Technology-driven, highly-scalable dragonfly topology. *Proc. of the International Symposium on Computer Architecture (ISCA)*, pages 77–88, 2008.

[7] John Kim, William J. Dally, and Dennis Abts. Flattened Butterfly : A Cost-Efficient Topology for High-Radix Networks. In *Proc. of the International Symposium on Computer Architecture (ISCA)*, pages 126–137, San Diego, CA, June 2007.

[8] John Kim, William J. Dally, Brian Towles, and Amit K. Gupta. Microarchitecture of a High-Radix Router. In *Proc. of the International Symposium on Computer Architecture (ISCA)*, pages 420–431, Madison, WI, 2005.

[9] Amit Kumar, Li-Shiuan Peh, Partha Kundu, and Niraj K. Jhay. Express Virtual Channels: Towards the Ideal Interconnection Fabric. In *Proc. of the International Symposium on Computer Architecture (ISCA)*, pages 150–161, San Diego, CA, June 2007.

[10] Steve Scott, Dennis Abts, John Kim, and William J. Dally. The BlackWidow High-radix Clos Network. In *Proc. of the International Symposium on Computer Architecture (ISCA)*, pages 16–28, Boston, MA, June 2006.

[11] Arjun Singh. *Load-Balanced Routing in Interconnection Networks*. PhD thesis, Stanford University, 2005.