Cost-Efficient Dragonfly Topology for Large-Scale Systems

John Kim
Northwestern University
Evanston, IL 60208
jjk12@northwestern.edu

William J. Dally
Stanford University
Stanford, CA 94305
dally@stanford.edu

Steve Scott
Cray Inc.
Chippewa Falls, WI 54729
sscott@cray.com

Dennis Abts
Google Inc.
dabts@google.com

ABSTRACT

Evolving technology and increasing pin-bandwidth motivate the use of high-radix routers to reduce the diameter, latency, and cost of interconnection networks. This migration from low-radix to high-radix routers is demonstrated with the recent introduction of high-radix routers and they are expected to impact networks used in large-scale systems such as multicomputers and data centers. As a result, a scalable and a cost-efficient topology is needed to properly exploit high-radix routers.

High-radix networks require longer cables than their low-radix counterparts. Because cables dominate network cost, the number of cables, and particularly the number of long, global cables should be minimized to realize an efficient network. In this paper, we introduce the *dragonfly* topology which uses a *group* of high-radix routers as a virtual router to increase the effective radix of the network. With this organization, each minimally routed packet traverses at most one global channel. By reducing global channels, a dragonfly reduces cost by 20% compared to a flattened butterfly and by 52% compared to a folded Clos network in configurations with $\geq$ 16K nodes.

The paper also introduces two new variants of global adaptive routing that enable load-balanced routing in the dragonfly. Each router in a dragonfly must make an adaptive routing decision based on the state of a global channel connected to a different router. Because of the *indirect* nature of this routing decision, conventional adaptive routing algorithms give degraded performance. We introduce the use of selective virtual-channel discrimination and the use of credit round-trip latency to both sense and signal channel congestion. The combination of these two methods gives throughput and latency that approaches that of an *ideal* adaptive routing algorithm.

1 Summary

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 – i.e. routers with a small number of ports. As a result, these networks used low-radix topologies such as 2-D or 3-D mesh or torus networks. Examples of machines employing such networks include the Cray T3D, T3E, and XT3. Earlier work [4, 2] showed that, for the packaging and technology constraints of the 80s and 90s, low-radix networks provide optimal latency for a given cost. This was true with the relatively low pin bandwidth available during the 80s and the early 90s. It is no longer the case.

Over the past 20 years, the pin bandwidth of router chips has increased by approximately an order of magnitude every 5 years (Figure 1) – a rate very similar to Moore's Law. This increase in bandwidth is a result of both an increase in the signaling rate and an 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 [11]. In contrast, low-radix routers divide the bandwidth into a smaller number of wide ports. The Cray BlackWidow system [1], one of the first systems to take advantage of high-radix routers, uses radix-64 routers and a variant of the high-radix folded-Clos topology [13].

Topology is a critical aspect of any interconnection network because it sets performance bounds for the network by establishing the network diameter as well as bisection bandwidth. The topology also largely determines the cost of the system. Existing topologies such as folded-Clos or fat-tree pay too high a penalty on load-balanced traffic (e.g. uniform random) to provide good performance on adversarial traffic pattern. In essence, they consume costly bandwidth to load balance traffic that is already balanced. A conventional butterfly network, on the other hand,

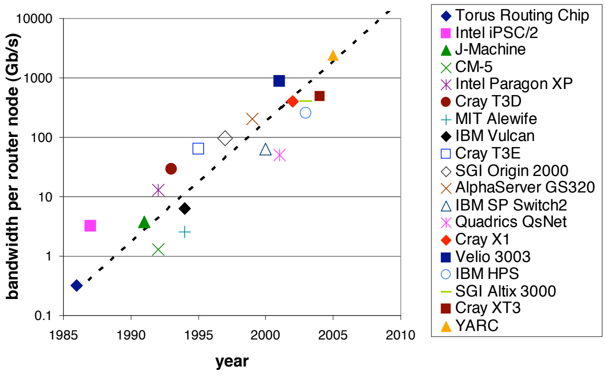

Figure 1: Router Bandwidth Scaling Relationship.

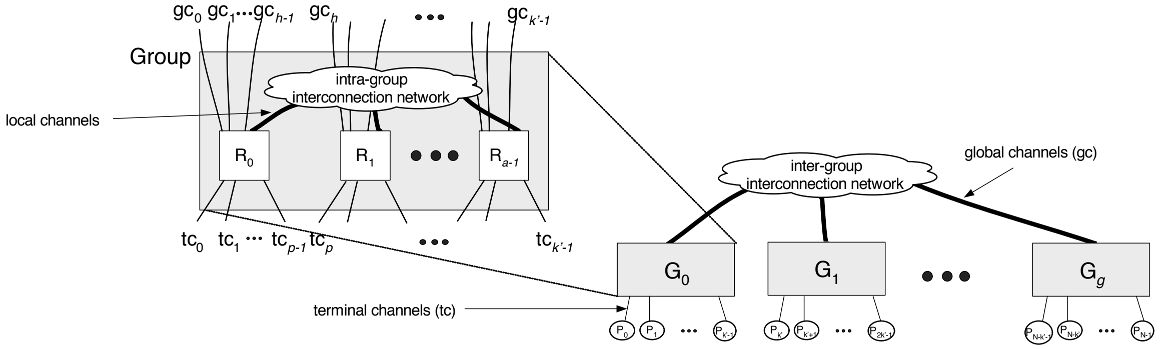
gc$_0$ gc$_1$…gc$_{h-1}$ gc$_h$ … gc$_{k'-1}$
Group
intra-group interconnection network
local channels
R$_0$
R$_1$
R$_{a-1}$
tc$_0$ tc$_1$ … tc$_{p-1}$ tc$_p$ … tc$_{k'-1}$
terminal channels (tc)
inter-group interconnection network
global channels (gc)
G$_0$
G$_1$
G$_g$
P$_0$ P$_1$ … P$_{k'-1}$
P$_{k'}$ P$_{k'+1}$ … P$_{2k'-1}$
P$_{N-k'-1}$ P$_{N-k'}$ … P$_{N-1}$

Figure 2: A high-level block diagram of a dragonfly topology and a diagram of a group or a virtual router.

gives significantly lower cost (approximately half) than a folded-Clos on balanced traffic. However, because a conventional butterfly has no path diversity, its performance is severely limited on adversarial traffic patterns.

The recently proposed flattened butterfly [10] approaches the cost of a conventional butterfly network on balanced traffic while matching the cost/performance of a folded-Clos topology on adversarial traffic. However, the scalability of the flattened butterfly is limited by the radix of a single router where as the proposed dragonfly topology is able to effectively increase the radix through the use of a *virtual* router or a collection of routers. In addition, the cost of a network is dominated by channels – especially, the long, global channels. The flattened butterfly requires each packet to traverse multiple global channels which increases the network cost. In this paper, we introduce the dragonfly topology [9] which reduces the number of global channels traversed per packet to one when minimal routing is used.

To achieve this unity *global diameter*, very high-radix routers, with a radix of $\sim 2\sqrt{N}$ (where N is the size of the network) are required. [1] While radix 64 routers have been introduced [13], much higher radices are needed to build machines that scale to 8K - 1M nodes with unity global diameter. To achieve the benefits of a very high radix, the paper proposes using a *group* of routers connected into a subnetwork as one very high-radix *virtual router* as shown in Figure 2. This very high effective radix in turn allows us to build a network in which all minimal routes traverse at most one global channel. The high effective radix also allows the dragonfly topology to provide high scalability – with radix-64 routers, the topology can scale to over 256k nodes with a network diameter of only three hops.

The high-radix topology, especially the dragonfly topology, increases the physical length of the global channels but by exploiting emerging optical signaling technology, the impact of long global channel lengths can be reduced. Historically, many networks have been proposed using optical signaling but because of its high cost, it has not been used in large-scale systems. However, the recent advent of economical optical signalling [12, 7] enables topologies with long channels but they are still more expensive than electrical channels. The proposed dragonfly results in a hierarchical topology that exploits the economical, optical signalling for the global channels while using the cheap, electrical channels for the short, local communication. Many hierarchical topologies have been previously proposed but the dragonfly topology is significantly different as it increases the effective radix through the local network. Previously proposed hierarchical topologies have often been built as tree structures which introduce a bandwidth bottleneck and increase hop count as the packets traverse up the hierarchy.

For the dragonfly topology, a critical aspect to fully exploit the benefits of the topology is the need for adaptive routing. The topology provides a high path diversity but non-minimal global adaptive routing is needed to properly exploit them. Achieving good performance on a wide range of traffic patterns on a dragonfly topology requires a routing algorithm that can effectively balance load across the global channels. Global adaptive routing (UGAL) [14], 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 paper 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.

The paper also provides a cost comparison of the dragonfly topology to alternative topologies using a detailed cost model. By reducing global channels, a dragonfly reduces cost by 20% compared to a flattened butterfly and by 52% compared to a folded Clos network in configurations with $\geq$ 16K nodes. Compared to a 3-D torus topology which only requires relatively short electrical cables, the dragonfly still provides a cost savings of up to 60% since the number cables (or channels) required is significantly reduced. The reduction of network cost in the dragonfly also translates to reduction of power as shown in prior work [10]. In the cost comparisons, we implicitly normalize the throughput (or the performance) of the alternative topologies to provide an accurate cost comparisons. The flattened butterfly has been shown to reduce the zero-load latency compared to alternative topologies [10] and in this paper, we also qualitatively and quantitatively compare the characteristics of the flattened butterfly to the dragonfly to highlight the benefits of the dragonfly topology such as reduced hop count.

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. The cost of a network (in terms of capital cost as well as energy consumption) is dominated by the topology – thus, it is critical to use a cost-efficient topology in the network. This paper presents the dragonfly topology for large-scale systems which exploits recent development of high-radix routers as well as recent advances in economical, optical signaling to create a cost-efficient, hierarchical topology.

[1] A fully connected topology with a concentration of $\sqrt{N}$ is assumed.

This paper makes several novel contributions to the field of interconnection network design. The novel contributions of this paper include the introduction of the hierarchical, dragonfly topology, creating *virtual* routers in a network topology to effectively increasing the router radix, the use of non-minimal global adaptive routing to fully exploit path diversity in the dragonfly, and identifying the problem of indirect adaptive routing and using credits and virtual channel discrimination to overcome the limitations.

Recent research in high-radix routers [11] has had significant impact on industry with the migration from low-radix routers to high-radix routers. Based on our previously proposed router microarchitecture [11], radix-64 Cray YARC router [13] was one of the first high-radix routers implemented and used in the recently announced Cray BlackWidow system [1]. Mellanox has recently introduced high-radix routers and researchers at SUN Microsystems have investigated creating high-radix switches using proximity communication [5]. With the availability of high-radix routers, the topology that efficiently exploits high-radix routers will have to be a significant departure from previous low-radix networks that employed 2-D or 3-D torus topologies. The modified high-radix folded-Clos topology used in the Cray BlackWidow network [13] or the recently proposed flattened butterfly topology [10] take advantage of high-radix routers to reduce the network diameter but require each packet to traverse multiple global (and expensive) channels. The proposed dragonfly topology, by effectively increasing the router radix, only requires one global channel to be traversed with the use of minimal routing. As the result, the proposed dragonfly topology is significantly different from any previously proposed hierarchical topologies as we create a *virtual* high-radix routers by using a collection of router to form a very high-radix router.

Another significant, practical impact of the dragonfly topology is that the hierarchical nature of the topology is well matched to the *packaging hierarchy* of system components. By creating a group or a collection of routers, the local networks can be packaged within a cabinet through backplane or a few neighboring cabinets can form a group using short, electrical signaling. The global network connecting the multiple groups together can be connected using optical signaling. The global channels required are much longer, but the impact of long channel distance is reduced with the use of optical signaling, which can traverse longer distance compared to electrical signaling without having to regenerate the signal.

The long term impact of this paper on future networks will not only be the novel dragonfly topology contribution and the use of *virtual* routers but also the adaptive routing aspect of future interconnection networks, which include the use of non-minimal global adaptive routing on the dragonfly topology and identifying the limitations of *indirect* adaptive routing and providing a mechanism to overcome these limitations. The dragonfly topology, with only minimal routing or randomized routing, does not provide a significant benefit over alternative topologies. However, proper use of global adaptive routing enables significant advantage of the dragonfly topology over alternative topologies. In addition, indirectness in the topology and deep buffers do not provide stiff backpressure. This is the first work to address this issue in adaptive routing and presents a novel mechanism to provide stiffer backpressure by delaying the transmission of credits.

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 routers and networks will become even more significant over time. This work is relevant to the networks used in all types of large-scale systems – e.g., server clusters, internet routers, and storage-area networks as well as supercomputers.

One example application where this work will have impact is in data centers. Most computer architecture research (academia and industry) have focused on processor architecture and recently, multicore architectures. However, with the increasing importance of large-scale internet services and the large-scale systems required to support their requirements, computer architects also need to focus on "warehouse-sized computing systems, made up of thousands of computing nodes, their associated storage hierarchy and interconnection infrastructure" [3, 6]. This paper presents an efficient interconnection network for these large-scale systems. Studies show that the capital cost of a data center is matched by the energy (cooling) cost within the first 3 years of purchase [8]. Thus, having a cost- and energy-efficient topology and interconnection network will be critical in future data centers and the proposed dragonfly topology provides an efficient topology for the data centers. As the number of terminals (or nodes) in a data center increases, the topology needs to be highly-scalable and the proposed dragonfly topology, through the use of virtual routers, provides the scalability. Ultimately, the research on high-radix networks and the dragonfly topology will have significant impact on the interconnection network used in these large-scale systems as we expect to see dragonfly topology and different variations of the topology employed in future large-scale systems.

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] Luiz André Barroso, Jeffrey Dean, and Urs Hölzle. Web search for a planet: The google cluster architecture. *IEEE Micro*, 23(2):22–28, 2003.

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

[5] Hans Eberle, Pedro J. Garcia, Jos Flich, Jos Duato, Robert Drost, Nils Gura, David Hopkins, and Wladek Olesins. High-radix crossbar switches enabled by proximity communication. In *Supercomputing 08 (to appear)*, Austin, TX, 2008.

[6] Xiaobo Fan, Wolf-Dietrich Weber, and Luiz Andre Barroso. Power provisioning for a warehouse-sized computer. In *Proc. of the International Symposium on Computer Architecture (ISCA)*, pages 13–23, San Diego, CA, 2007.

[7] Intel Connects Cables. http://www.intel.com/design/network/ products/optical/cables/index.htm/.

[8] Kenneth G. Brill. The invisible crisis in the data center: The economic meltdown of moores law. *Uptime Institute White Paper*, 2007.

[9] 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.

[10] 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.

[11] 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.

[12] Luxtera Inc. White Paper: Fiber will displace copper sooner than you think, November 2005.

[13] 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.

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