Global Misrouting Policies in Two-level Hierarchical Networks

M. García, E. Vallejo, R. Beivide, M. Odriozola, C. Camarero
University of Cantabria, Santander, Spain
{marina.garcia,enrique.vallejo,ramon.beivide,cristobal.camarero}@unican.es

M. Valero, J. Labarta
Universitat Politécnica de Catalunya and BSC, Spain
{mateo.valero,jesus.labarta}@bsc.es

G. Rodríguez
IBM Research GmbH Zürich Research Lab.
{rod}@zurich.ibm.com

ABSTRACT

Dragonfly networks are composed of interconnected groups of routers. Adaptive routing allows packets to be forwarded minimally or non-minimally adapting to the traffic conditions in the network. While minimal routing sends traffic directly between groups, non-minimal routing employs an intermediate group to balance network load.

A random selection of this intermediate group (denoted as RRG) typically implies an extra local hop in the source group, what increases average path length and can reduce performance. In this paper we identify different policies for the selection of such intermediate group and explore their performance. Interestingly, simulation results show that an eager policy (denoted as CRG) that selects the intermediate group only between those directly connected to the ongoing router causes starvation in some network nodes. On the contrary, the best performance is obtained by a "mixed mode" policy (denoted as MM) that adds a local hop when the packet has moved away from the source router.

Categories and Subject Descriptors

C.1.2 [**Multiprocessors**]: Interconnection architectures

1. INTRODUCTION

Dragonfly interconnection networks are two-layered hierarchical networks, introduced in [5] and employed in the IBM PERCS [1]. They are organized as groups of routers. Routers within a group are interconnected and behave as a larger-degree router. The inter and intra-group interconnection networks could be anyone. However, a complete graph which minimizes the number of hops has been considered in most of the previous works [5, 4, 1, 2]. In this work we will focus on that specific Dragonfly based on complete graphs.

Links connecting nodes within a group are denoted as *local* (l) while links connecting groups are denoted as *global* (g).

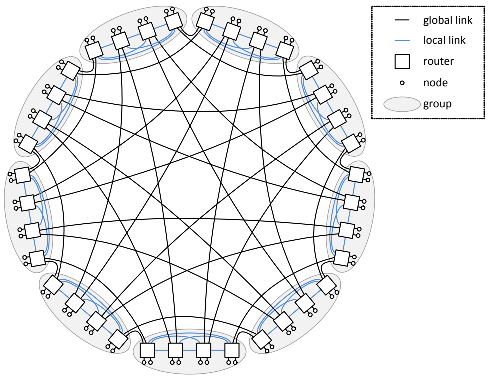


Figure 1: Example of a well-balanced Dragonfly size $h = 2$.

The Dragonfly topology is defined by three parameters a, p and h. a is the number of routers per group, p is the number of processing nodes per router and h is the number of global links per router. A *well-balanced* Dragonfly has been defined as the case with $a = 2p = 2h$ [5]. An example of such Dragonfly size $h = 2$ is shown in Figure 1. The diameter in a Dragonfly based on complete graphs is 3, that is, the longest path between any pair of nodes with minimal routing is 3 hops, 2 of them local and 1 global: $l - g - l$.

Several routing policies have been proposed for these networks. Under certain traffic patterns, the global link in the minimal path can suffer congestion, which makes non-minimal routing profitable [5]. For example, adversarial traffic $ADV + N$, in which every node in group i sends its traffic to any node in group $i + N$. Valiant routing [7] is a non-minimal routing that selects an intermediate group to misroute each packet. The packet is first sent minimally to the intermediate group, what is denoted as global misrouting. Then, it is sent minimally to the destination node. Global misrouting avoids congested global links, but increases the path length to, at most, 5 hops: $l - g - l - g - l$. Similarly, congestion in local channels was studied in [2], with local misrouting proposed to avoid congested local links in the intermediate and destination groups. Adaptive mechanisms choose between minimal or non-minimal routing to adapt routing to changing traffic conditions.

When the inter-group topology is a complete graph with just one link connecting every pair of groups, the selection

INA-OCMC '13, January 23, 2013, Berlin, Germany.
Copyright 2013 ACM 978-1-4503-1784-9/13/01 ...$15.00.

of an intermediate group is equivalent to the selection of one global link in the source group. We will refer to this selection process as the global misrouting policy. Since each router in a group only connects to some few other groups in the system, the selected global link can be directly connected to the current router, or to a different neighbor router in the group. Thus, the global misrouting policy conditions the overall length of the path followed by the packet, saving or imposing a local hop in the source group, what can affect the performance of the network. It is also important that the global misrouting policy balances the traffic properly across all the network groups. However, as far as we know, this selection policy has not been studied in detail before.

In this paper we make a detailed study of different global misrouting policies that select an intermediate group for non-minimal routing in a Dragonfly network. Specifically, the main contributions of this paper are: 1) We describe different global misrouting policies for Dragonfly networks, highlighting their impact on two key aspects: the path length and the traffic balance in the network. 2) We evaluate the performance of the different policies by simulation, observing that an eager policy that misroutes traffic using only the global links of the current router obtains the highest throughput. 3) We identify a starvation problem in the previous policy, which is solved by another misrouting strategy.

The rest of the paper is organized as follows: In Section 2 we give an overview of the previous work related to global misrouting in Dragonfly topologies. Section 3 describes several global misrouting policies for the selection of an intermediate group in Dragonflies. Section 4 describes the experimental setup and presents the results obtained. Finally, in Section 5 we discuss the main results and contributions.

2. RELATED WORK

Different routing mechanisms have been proposed for the Dragonfly. The simplest ones are oblivious to the network status. Minimal routing follows the shortest path between each pair of nodes. Valiant routing [7], proposed for the Dragonfly in [5], applies global misrouting to each packet regardless the network status. PERCS allows the programmer to specify the intermediate group for each packet [1].

Adaptive routing mechanisms select the path of each packet depending the network conditions. Two types of adaptive routing mechanisms have been proposed: source and in-transit (or on-the-fly) routing. Source routing mechanisms determine the path of each packet at injection time. Examples of source-routing mechanisms are UGAL, Piggybacking (PB) or CRT [4]. These mechanisms need to estimate the congestion in the global links of the group to select between a minimal or non-minimal routing for each packet. This estimation typically relies on indirect information (for example, the credit count in the outputs of other routers in the group). For this reason, they are relatively complex and slow in adapting to traffic changes.

In-transit adaptive routing mechanisms can decide to apply misrouting in each hop of a route. If the inter and intra-group topologies are complete graphs, this means that a global misrouting decision can be taken at injection time, or after a first local hop in the source group. Progressive Adaptive Routing (PAR, [4]) implements this solution for global misrouting, what increases the maximum path length in one hop, the first local hop in the source group before deciding to apply global misrouting. OFAR [2] supports both in-transit local and global misrouting, with maximum path length of 8 hops. The maximum path length is important for several reasons: the maximum base packet latency, the load of traffic that can accept the network, and the relation with the deadlock-avoidance mechanism used. All previously proposed deadlock-avoidance mechanisms are based on an idea proposed by Günther [3], which employs as many virtual channels as hops in the longest allowed path. On the contrary, OFAR employs a Hamiltonian ring with bubble flow-control added to the base Dragonfly network.

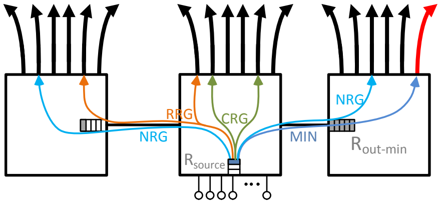


Figure 2: Global link selection models: *RRG*, *CRG*, *NRG*.

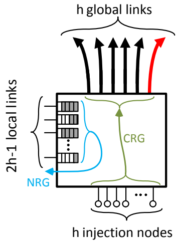


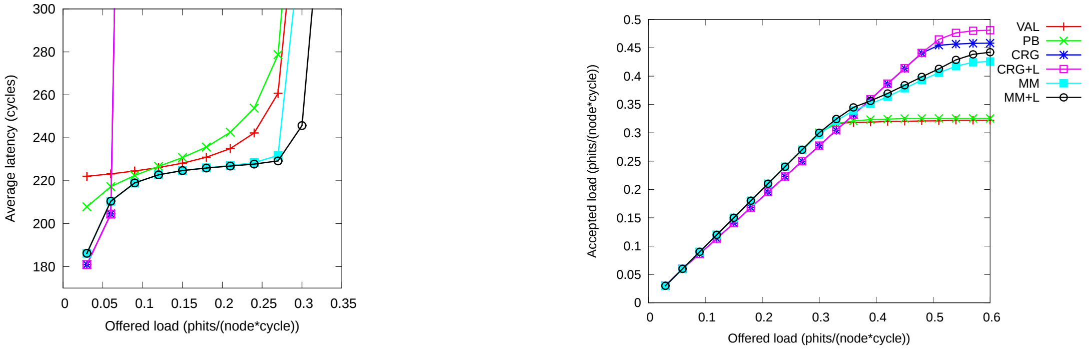
Figure 3: *MM* policy: *CRG* applied to packets in the h injection queues, and *NRG* to those in the $2h-1$ local queues.

Regarding the selection of the specific global link used for global misrouting, previous works have used two alternatives. Most proposals (including PB and PAR) have selected the link randomly among all global links of the group, except the minimal one. This is the policy that we will denote as "Random router global", or *RRG*. OFAR, by contrast, makes a different selection depending on the location of the packet: at injection time, global misrouting is only allowed on the global outputs of the source router; after a first hop, global misrouting is allowed on any global port of the group, *except* the ones in the current router. We will later denote this policy as "Mixed mode" (*MM*).

3. GLOBAL MISROUTING POLICIES

In this section we will study the possible policies that can be applied for global misrouting selection. We will consider two different figures of merit: path length and load balance. We start by the simplest case; source-routing mechanisms that determine at injection time if global misrouting is applied. When the packet is still in an injection port of the source router, once decided that a packet will be misrouted, the path length to its destination node will be the same no matter through which global link of the current router the packet is misrouted. However, the selection of a global link connected to a neighbor router implies a path length increase of one local hop. Three different policies can be considered:

- **Random-router Global (*RRG*)**: The global link is



(a) Latency

(b) Throughput

Figure 4: Latency and throughput under adversarial +2 traffic (ADV+2)

chosen randomly among any global link in any router at the source group.

- **Current-router Global (*CRG*)**: The global link is chosen randomly among those in the current router.
- **Neighbor-router Global (*NRG*)**: A neighbor router in the group is randomly chosen; the global link is randomly chosen among the ones in that neighbor. Unlike RRG, this explicitly forbids the global links of the current router.

Figure 2 shows the behavior of the mechanisms RRG, CRG and NRG. We have denoted as $R_{out-min}$ the router that contains the global link that corresponds to the minimal path. Global misrouting with CRG would only allow the global links in the source router R_{source}. RRG allows any global link in the group (including the ones in R_{source} and $R_{out-min}$) while NRG allows only neighbor routers.

These policies balance traffic across all global links in the group. This is obvious for RRG, in which every router in a group can use all global links in the group to misroute packets. When all the routers in the network use NRG or CRG for misrouting, all the global links will be used for misrouting with the same probability. However, NRG requires a local hop in the source group before reaching the global link to misroute the packet, and RRG almost always does. By contrast, CRG saves this local hop. Thus, CRG appears as the best option for source routing.

When in-transit routing mechanisms are considered (PAR and OFAR), global misrouting can be also selected after a first minimal hop, when the packet is in a local input queue of $R_{out-min}$. In this case, the same three policies might be applied. However, depending on the traffic pattern not all routers in the group will receive the same traffic from other routers. Specifically, under adversarial traffic patterns (ADV), $R_{out-min}$ is the same one for all packets in the group. In this case, applying CRG to in-transit traffic would concentrate all the traffic in the global links of $R_{out-min}$. Such problem would not occur when using RRG or NRG, as the packets would be misrouted through global links connected to other routers in the group instead of through $R_{out-min}$. But again, that would require additional local hops. However, we do not necessarily have to use the same misrouting policy for in-transit traffic and packets in injection queues. Specifically, we will define the following:

- **Mixed-Mode (*MM*)**: CRG is applied to packets at the injection queues and NRG to packets at the transit queues.

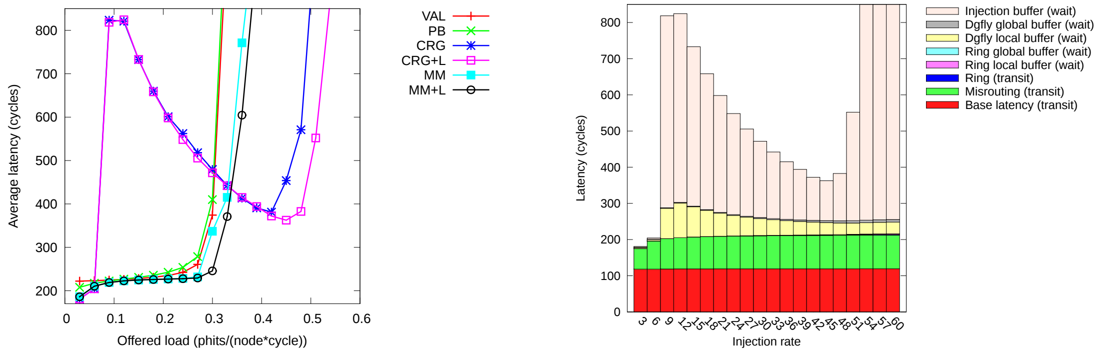
This MM policy is depicted in Figure 3. It balances traffic across all global links in the group, since both CRG in the source router and NRG in the $R_{out-min}$ router balance traffic across all links. Even when we consider adversarial traffic, the packets that are first sent minimally to $R_{out-min}$ and then misrouted do not compete for the global nonminimal queues against the packets directly injected in $R_{out-min}$: NRG is applied to packets in the local queues moving them away from $R_{out-min}$, while CRG is applied to those in the injection queues.

4. SIMULATION RESULTS

In this section we present performance results obtained by simulating a Dragonfly network with 5256 computation nodes and 876 routers of 24 ports (parameter $h = 6$), with 12 routers and 72 nodes per group. We employ an in-house developed single cycle simulator, which accurately models Valiant routing (VAL) as defined in [5], Piggybacking routing (PB) as defined in [4], and OFAR routing as defined in [2]. Both VAL and PB employ the RRG misrouting policy; in OFAR we have implemented both CRG and MM to evaluate their performance with an in-transit adaptive routing mechanism. Also, we have evaluated both models with and without support for local misrouting; denoted as CRG and MM (without local misrouting), and CRG+L and MM+L (with local misrouting). We model an input FIFO buffered Virtual Cut-through (VCT) router with latencies of 10 cycles in local links and 100 cycles in global ones, similarly to previous works [6, 2]. Packet size is 8 phits, and the size of local and global queues is 32 and 256 phits respectively.

We measure the average latency and throughput over a long period, after 20,000 network warm-up cycles. Each point in the plots shows the measured value for a given offered load in $phits/(node \cdot cycle)$. Figures 4a and 4b show latency and throughput results under adversarial traffic ADV+2. With this traffic, every node in group i sends its traffic to a node randomly chosen among all nodes in group $i + 2$.

Latencies in Figure 4a show that, while both MM models behave as expected, CRG and $CRG + L$ suffer a huge increase in their average latency even for very low loads. Despite this problem, CRG obtains better throughput than



(a) Latency zoomout

(b) Dissection of the latency

Figure 5: Latency zoom under ADV+2 traffic and dissection of latency of CRG

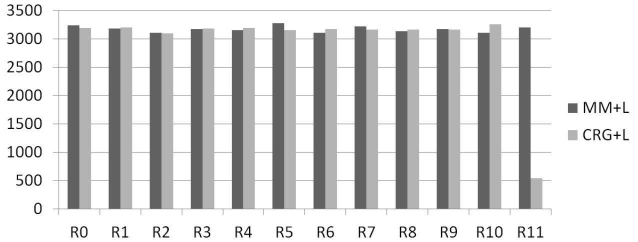


Figure 6: Packets injected by the nodes connected to each router in a group.

the corresponding MM models, as presented in Figure 4b. Both Valiant and PB provide worse latency and throughput than the MM models.

The latency problem of CRG can be clearly observed in Figure 5a. As explained in Section 3, when using an in-transit routing mechanism (we employ OFAR) CRG does not properly balance load across all global links. Specifically, the router denoted as $R_{out-min}$ (the one with the global port corresponding to the minimal path) receives most of the load. A relative starvation problem is present. Figure 6 shows the number of packets sent by nodes on a given group under load 0.20. Router 11, acting as $R_{out-min}$ in this case, can inject much less packets during the same time. Similarly, the latency of its packets (not depicted) is higher due to the long wait at the injection queues. Figure 5b confirms that the latency increase is due to the delays in local and injection queues.

Despite the congestion in $R_{out-min}$, CRG is the mechanism that obtains higher throughput in Figure 4b. Once $R_{out-min}$ is congested, most of the packets from other routers in the group will be sent non-minimally using global links of their own routers. Thus, most paths will have only 4 hops $g-l-g-l$, reducing contention in local links that can limit performance and decreasing average latency.

5. CONCLUSIONS

Our study has formally presented different policies for global misrouting in adaptive routing for Dragonfly networks. Such policies should focus on load balance and path length reduction. When in-transit routing is used, we have identified that the CRG policy introduces load imbalance which causes performance degradation in the nodes in some routers under adversarial traffic. By contrast, a mixed policy denoted as MM regains the balance by sending traffic out of the congested router, while saving the first local hop when misrouting is applied at injection time.

6. ACKNOWLEDGMENTS

This work has been supported by the Spanish Ministerio de Ciencia e Innovación, under projects TIN2010-21291-C02-02 and TIN-2007-60625, The HiPEAC Network of Excellence, the Consolider Project "Supercomputación y e-Ciencia" and the Spanish Ministerio de Educación, grant AP2010-4900.

7. REFERENCES

[1] B. Arimilli, R. Arimilli, V. Chung, S. Clark, W. Denzel, B. Drerup, T. Hoefler, J. Joyner, J. Lewis, J. Li, et al. The PERCS high-performance interconnect. In *2010 18th IEEE Symposium on High Performance Interconnects*, pages 75–82. IEEE, 2010.

[2] M. García, E. Vallejo, R. Beivide, M. Odriozola, C. Camarero, M. Valero, G. Rodríguez, J. Labarta, and C. Minkenberg. On-the-fly adaptive routing in high-radix hierarchical networks. In *The 41st International Conference on Parallel Processing (ICPP)*, 09 2012.

[3] K. Gunther. Prevention of deadlocks in packet-switched data transport systems. *Communications, IEEE Transactions on*, 29(4):512 – 524, apr 1981.

[4] N. Jiang, J. Kim, and W. J. Dally. Indirect adaptive routing on large scale interconnection networks. In *ISCA '09: 36th International Symposium on Computer Architecture*, pages 220–231, 2009.

[5] J. Kim, W. Dally, S. Scott, and D. Abts. Technology-driven, highly-scalable dragonfly topology. In *Proceedings of the 35th Annual International Symposium on Computer Architecture*, pages 77–88. IEEE Computer Society, 2008.

[6] J. Kim, W. Dally, S. Scott, and D. Abts. Cost-efficient dragonfly topology for large-scale systems. *Micro, IEEE*, 29(1):33–40, 2009.

[7] L. Valiant. A scheme for fast parallel communication. *SIAM journal on computing*, 11:350, 1982.