On the use of contention information for adaptive routing

Pablo Fuentes, Enrique Vallejo,
Marina García, Ramón Beivide [1] [2]

Computer Science and Electronics Department, University of Cantabria, Avda. Los Castros SN, 39005, Santander, Spain

ABSTRACT

In-network congestion is generated by output-port contention, in which packets from multiple flows of traffic coincide in the same network link. Adaptive routing mechanisms react to network congestion in order to improve latency and throughput, by sending traffic using alternative uncongested paths. Such approach has some drawbacks, such as a relatively high reaction time on traffic changes, or requiring that some traffic fills the slow and congested queue.

In this document we introduce the idea of reacting to network contention, rather than network congestion, and present early performance results of a mechanism based on "contention counters", which track the traffic demand for each output port of a router. We evaluate the idea using a dragonfly network, which has only one minimal path between nodes but multiple longer non-minimal paths. Early results show that an implementation based on contention counters provides optimal latency under minimal traffic and competitive performance under adversarial traffic conditions.

KEYWORDS: Adaptive routing; Contention counters; dragonfly network

1 Introduction

Certain communication traffic patterns lead to network congestion, in which some areas of the network accumulate multiple packets that fill the router queues, forming congestion trees that stop transmission and limit performance. Adaptive routing mechanisms adapt the path followed by traffic to the changing network conditions to avoid such congested areas. Nonminimal routing can employ longer paths to avoid minimal congested ones. Dragonfly networks [KDSA08] are one example of a network which requires of such nonminimal adaptive routing. Dragonflies are direct networks composed of multiple groups of routers, with a complete graph as the inter-group and intra-group topology. Multiple computing nodes are connected to each router. Under uniform traffic, all network links are used in a balanced way. However, when multiple nodes communicate with nodes in the same destination group, the global link between these groups saturates and congestion appears, lead-

[1] E-mail: {pablo.fuentes, enrique.vallejo, marina.garcia, ramon.beivide}@unican.es

[2] This work has been supported by the Spanish Ministry of Science under TIN2010-21291-C02-02, the European HiPEAC Network of Excellence. The research leading to these results has received funding from the European Research Council under the EU FP7 (FP/2007-2013) / ERC Grant Agreement ERC-2012-Adg-321253-RoMoL.

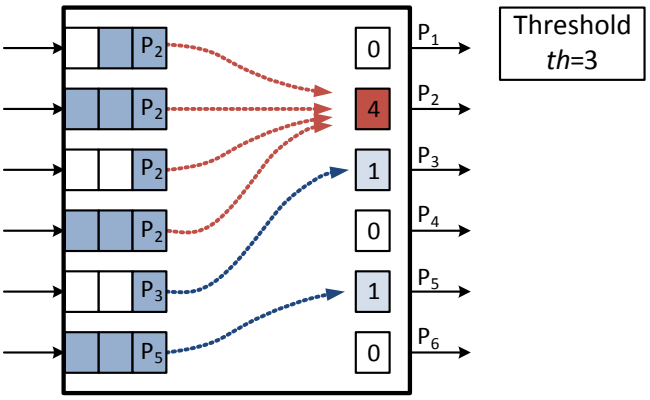

Figure 1: Example of use of contention counters. One counter is associated to each output port. Packets at the head of an input queue increase the counter corresponding to their minimal output. When this counter exceeds a given threshold (in this example, $th = 3$) traffic is diverted nonminimally.

ing to very poor performance. Valiant routing [Val82] randomizes traffic by sending traffic to a random intermediate node before forwarding it to its final destination. Adaptive routing selects between minimal or Valiant paths, either at the source (such as Piggybacking routing [JKD09]) or in-transit (such as OLM [GVB+13]).

Congestion-detection typically relies in the credit count of the output ports: a small amount of available credits is an indicative of congestion, whereas a high amount means that there is no congestion. This traditional approach requires a relatively large amount of traffic in the slow and congested path. Such traffic suffers a larger latency due to increased drain time, and this mechanism increases the detection time after traffic changes since it requires to wait for the queues to fill before triggering non-minimal routing.

The alternative approach presented in this document is to trigger misrouting based on network contention, this is, the simultaneity of flows converging in the same network path. In fact, in-network congestion is a consequence of contention, so this approach responds on the source of the problem, hopefully leading to faster response time on transient traffic changes. The general idea is presented in Section 2, and early evaluation is presented in Section 3 and some conclusions and future work ideas are presented in Section 4.

2 Contention counters

Contention counters are a set of counters tracking the extent of the demand of each output port, from the flows of traffic in the input queues. Figure 1 shows an example of the use of contention counters. There exists a counter associated with each output port. A packet reaching the head of an input buffer will increase the contention counter associated to its *minimal path*, this is, the path without misrouting. Congestion is detected when such counter exceeds a given threshold ($th = 3$ in the figure), and the traffic is sent nonminimaly using an alternative output port whose counter is below the limit. Such port can be selected randomly, according to the rules of each topology, but its counter is not incremented by this packet.

When a packet leaves completely an input buffer, its corresponding contention counter is decremented. This tries to avoid small values in the counters when packet headers are not received concurrently, that would lead to excessive incorrect estimations.

3 Early performance results

This section presents early performance results of the contention counter mechanisms. The proposed mechanism has been implemented in a cycle-accurate network simulator. We model a balanced dragonfly with 129 groups of 16 routers each, using a complete graph for the inter- and intra- group topologies. 8 computing nodes connect to each router, which also has 8 global links to different groups and 15 local links to the remaining routers of the group. Overall, the modelled system connects 16,512 computing nodes. Local links have a latency of 10 cycles and global links 100 cycles, and we model the latency of credit management.

The following routing mechanisms have been implemented:

- **Minimal (MIN):** A packet is forwarded minimally to the destination group, and then to the destination node. This routing requires 2 virtual channels (VCs) in local ports and 1 in global ones (denoted 2/1) for deadlock avoidance. This routing is oblivious.

- **Valiant (VAL):** A packet is sent minimally to a random intermediate group, and then minimally to the final destination. It employs 3/2 VCs. This routing is oblivious.

- **Piggybacking (PB):** A source-adaptive routing mechanism which relies on congestion information from the global links of the neighbor routers of the group, [JKD09]. This mechanism is adaptive and requires 3/2 VCs.

- **Opportunistic Local Misrouting (OLM):** An in-transit adaptive routing mechanism which supports global misrouting to an intermediate group, or local misrouting within groups to avoid congestion, and requires only 3/2 VCs as the previous mechanisms.

- **Contention-counters:** The proposal introduced in Section 2, using in-transit adaptive routing triggered by the values of the contention counters. We empirically set a threshold of $th = 5$, which is a performance tradeoff between the evaluated traffic patterns.

Two traffic patterns have been modelled, representative of the Dragonfly extreme cases:

- **Uniform (UN):** The destination node is selected randomly among all the possible nodes in the network. Misrouting is unnecessary under this traffic since it is naturally balanced. MIN is the reference for this traffic, because it never employs misrouting.

- **Adversarial+1 (ADV+1):** The destination node is selected randomly among all nodes in the following group (+1). In this case, Valiant routing is the reference since it avoids the single link between the source and destination groups, which would become a bottleneck.

Performance results are presented in Figure 2. Contention counters provide optimal latency under uniform traffic, significantly better than the adaptive mechanisms based on congestion estimation. However, its throughput drops after reaching the maximum level. In adversarial traffic, the latency is competitive, though it is slightly higher for low loads, in which there are not enough packets in the input queues to reach the threshold level. However, for larger loads the contention counters approach is very competitive and it reaches the maximum throughput.

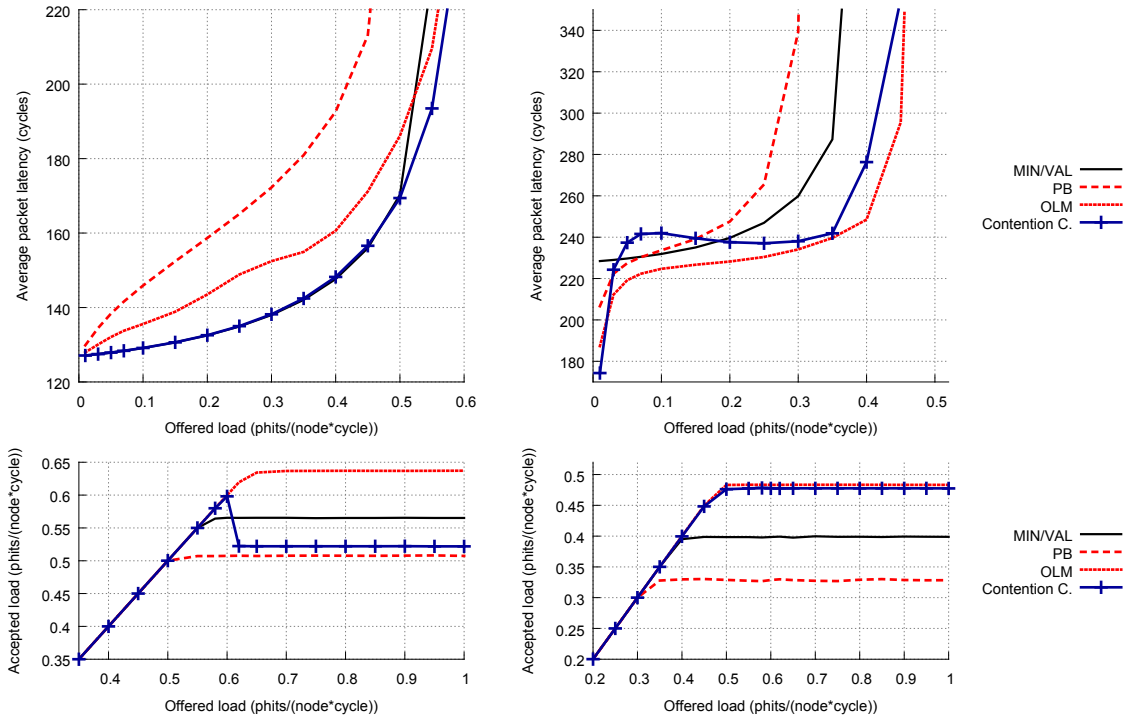

Figure 2: Latency (top) and throughput (down) results under Uniform (left) and Adversarial+1 (ADV+1, right) traffic patterns.

4 Conclusions and future work

The contention counters presented in this work are an appealing alternative for misrouting trigger in nonminimal adaptive routing mechanisms. Early evaluation results show that they obtain the optimal latency under UN and throughput under ADV+1 traffic patterns. Ongoing work implies a more detailed evaluation of the mechanism under transient and mixed traffic conditions, and alternative implementations which also consider congestion information.

References

[GVB+13] Marina García, Enrique Vallejo, Ramón Beividen Beivide, Miguel Odriozola, and Mateo Valero. Efficient routing mechanisms for dragonfly networks. In *The 42nd International Conference on Parallel Processing (ICPP-42)*, 2013.

[JKD09] Nan Jiang, John Kim, and William J Dally. Indirect adaptive routing on large scale interconnection networks. In *ISCA '09: 36th International Symposium on Computer Architecture*, pages 220–231, 2009.

[KDSA08] J. Kim, W.J. 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.

[Val82] L.G. Valiant. A scheme for fast parallel communication. *SIAM journal on computing*, 11:350, 1982.