On the Use of Contention Information for Adaptive Routing

Pablo Fuentes, Enrique Vallejo, Marina García, Ramón Beivide
University of Cantabria, Spain

Introduction

- **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.
- Under network congestion, some areas of the network accumulate multiple packets that fill the router queues, forming congestion trees that stop transmission and limit performance.
- **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 work proposes a mechanism for **congestion prevention** through the use of contention counters.
- We evaluate our proposal in a Dragonfly network.

Dragonfly networks

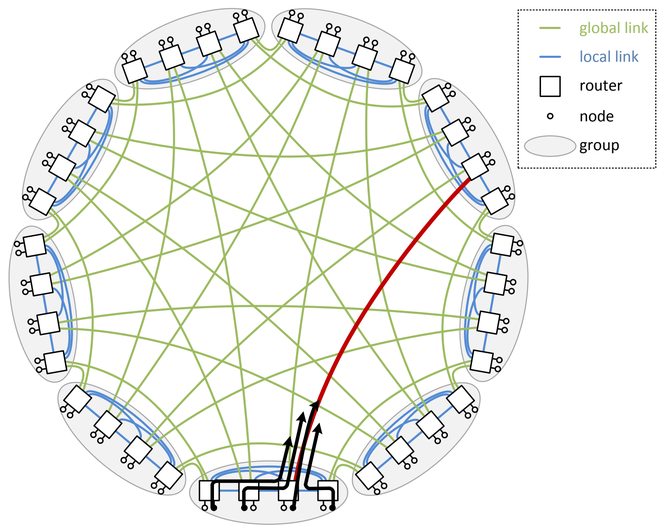


Figure 1: Dragonfly network displaying an example of network contention that limits throughput.

- A ***Dragonfly network*** is hierarchically divided into 2 levels: the first level is composed of multiple groups of routers, with routers connected all-to-all within a group; and the second level connects the different groups in an all-to-all topology. This network has been proposed to achieve high scalability in order to face exascale computation needs.
- **Minimal routing** in this network only permits one possible route between any pair of given nodes. Non-minimal routing is achieved through **Valiant misrouting**: packets are sent to a group randomly selected, and they are routed minimally from there towards the destination.
- Throughput in this network is highly sensitive to traffic pattern, and its network global links rapidly saturate under adversarial traffic. This requires an **adaptive routing** mechanism that switches between the minimal route and a non-minimal Valiant path.

Contention Counters

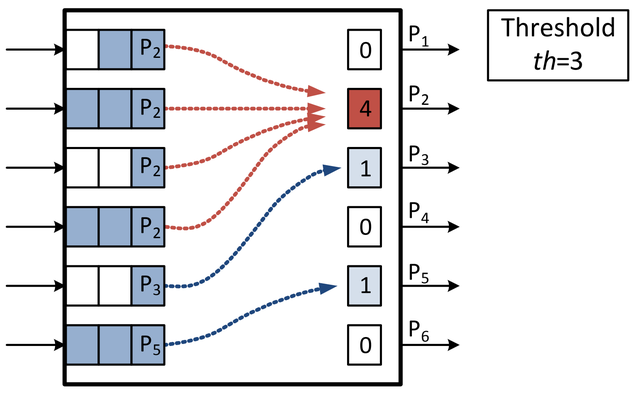
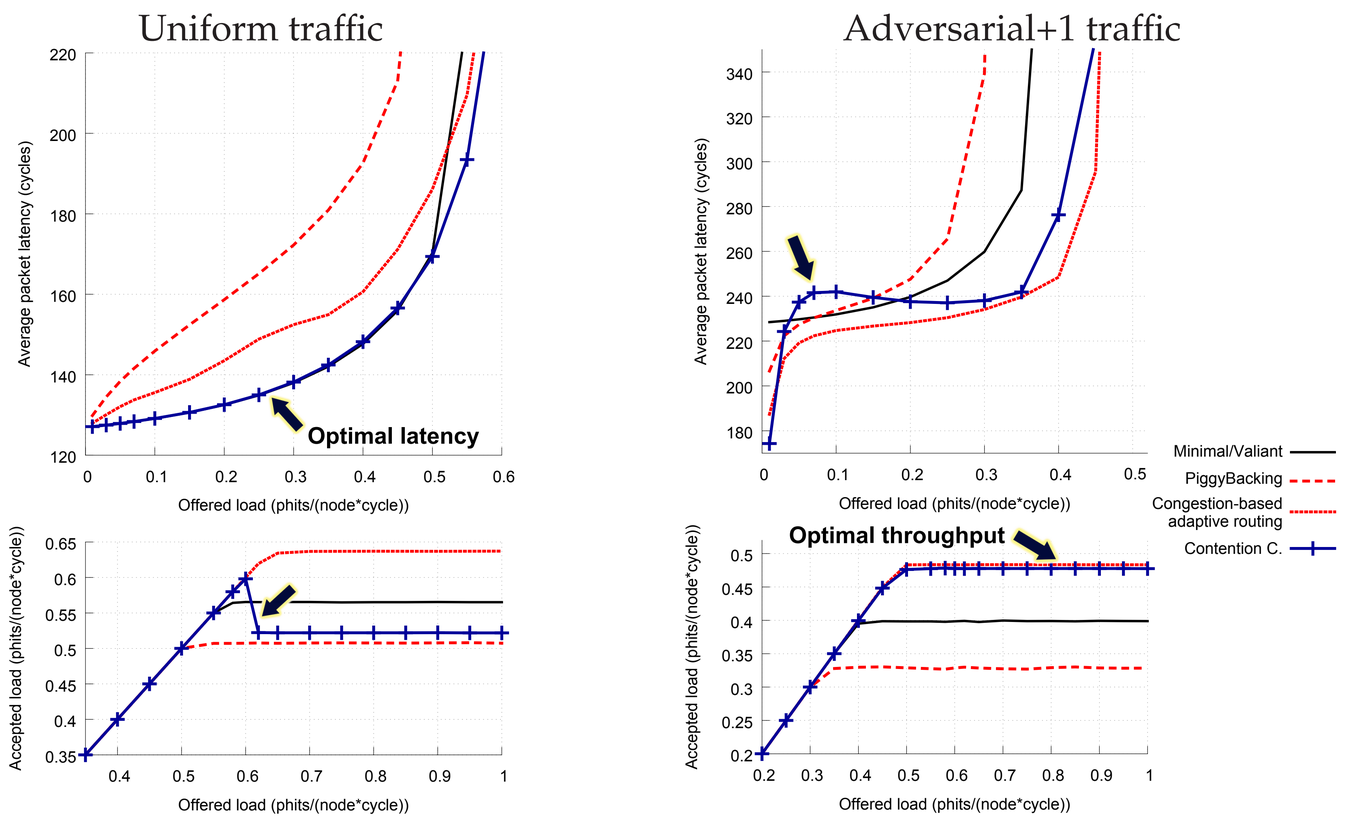


Figure 2: Example of use of contention counters. Packets at the head of an input queue increase the counter corresponding to their minimal output. When this counter exceeds a given threshold (in this example, $th = 3$) traffic is diverted nonminimally.

- **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. 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.
- **Contention 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.
- **Counters are decremented** when packets leave the input buffers completely. This tries to avoid small values in the counters when packet headers are not received concurrently, that would lead to excessive incorrect estimations.

Early Performance Results



- Contention counters provide **optimal latency** under uniform traffic, significantly better than the adaptive mechanisms based on congestion estimation.
- Its **throughput drops** under uniform traffic, after reaching the maximum level.
- **Competitive latency** in adversarial traffic. It is slightly higher for low loads, in which there are not enough packets in the input queues to reach the threshold level. This effect disappears for larger loads.
- It reaches the **maximum throughput** under adversarial traffic.

Ongoing Work and Conclusions

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.

Acknowledgments

This work has been supported by the Spanish Science and Technology Commission (CICYT) under contract TIN2010-21291-C02-02, the European Union's FP7 under Agreements ERC-321253 (RoMoL) and ICT-288777 (Mont-Blanc) and by the European HiPEAC Network of Excellence. This project has received funding from IBM Research GmbH and Barcelona Supercomputing Center under JSA no. 2013_119.