Marina Garcia
22 August 2013

OFAR-CM: Efficient Dragonfly Networks with Simple Congestion Management

M. Garcia, E. Vallejo, R. Beivide, M. Valero and G. Rodríguez

Document number

Table of contents

Introduction 1

Routing in Dragonfly networks 2

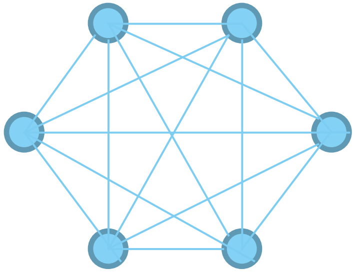
OFAR-CM 3

Performance results 4

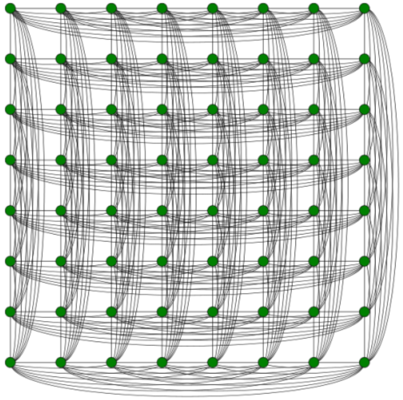
Conclusions 5

Introduction

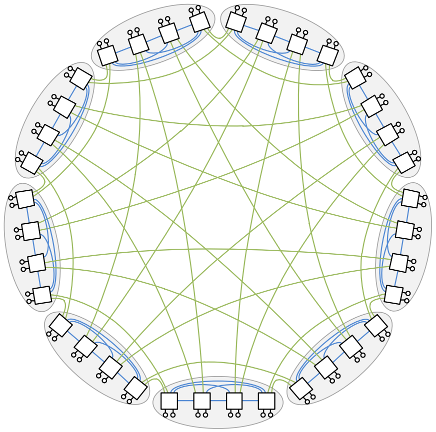
- System networks for **exascale computing** will require **low power and latency.**
  - This implies: **low diameter** and **average distance.**
- Traditional HPC networks employ low-radix routers (few ports).
  - 3D or 5D torus in IBM BlueGene, 3D Torus in Cray XE-series.
- High-radix routers are the norm today [1].
- Frequent direct networks recently proposed for high-radix routers:



All-to-all topology
(complete graph)



Flattened Butterfly
(Hamming graph, rook's graph, …)
Kim, ISCA'07

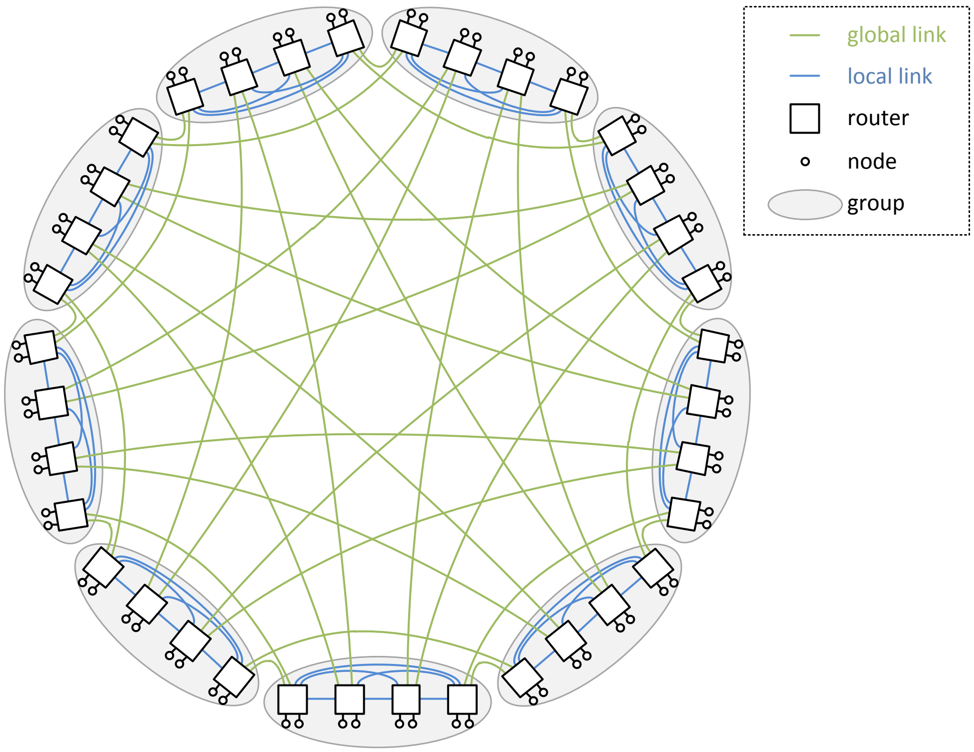


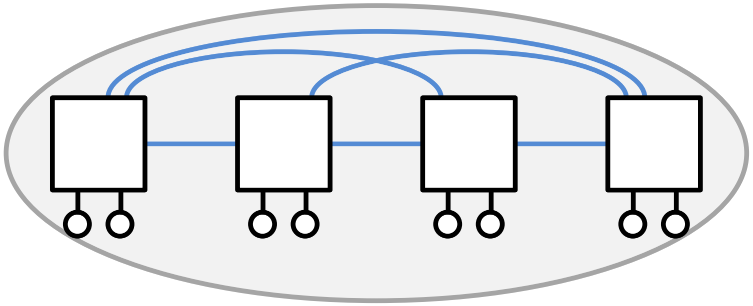
Dragonfly
(2-level direct network…)
Kim, ISCA'08

[1] Kim et al, "Microarchitecture of a high-radix router," ISCA'05

Introduction: Dragonfly interconnection network

- Dragonfly: Hierarchical direct network.
  - High-radix routers forming groups.
  - Cheap & scalable system-level network.
  - Low diameter.
- Inter-group connectivity:
  - Cheap electrical cables (local links).
  - All-to-all topology.
- Intra-group connectivity:
  - Optical cables (More $$$, global links).
  - All-to-all topology.
- Parameters
  - ***a***: Routers per group
  - ***p***: Nodes per router
  - ***h***: Global links per router
  - *"Well balanced": a = 2p =2h*





Introduction: Traffic patterns

- Uniform Traffic Pattern (**UN**)
  - Destination node randomly chosen.
  - Balanced use of the network links.

- Adversarial Traffic Pattern + N (**ADVG+N**)
  - Source node in group *i*, router *j.*
  - Destination node randomly chosen among those in group *i+N.*
  - Only one link connecting each pair of groups → Unbalanced use of network links.
  - Less adversarial → N=1
  - Most adversarial → N=h

Table of contents

Introduction 1

Routing in Dragonfly networks 2

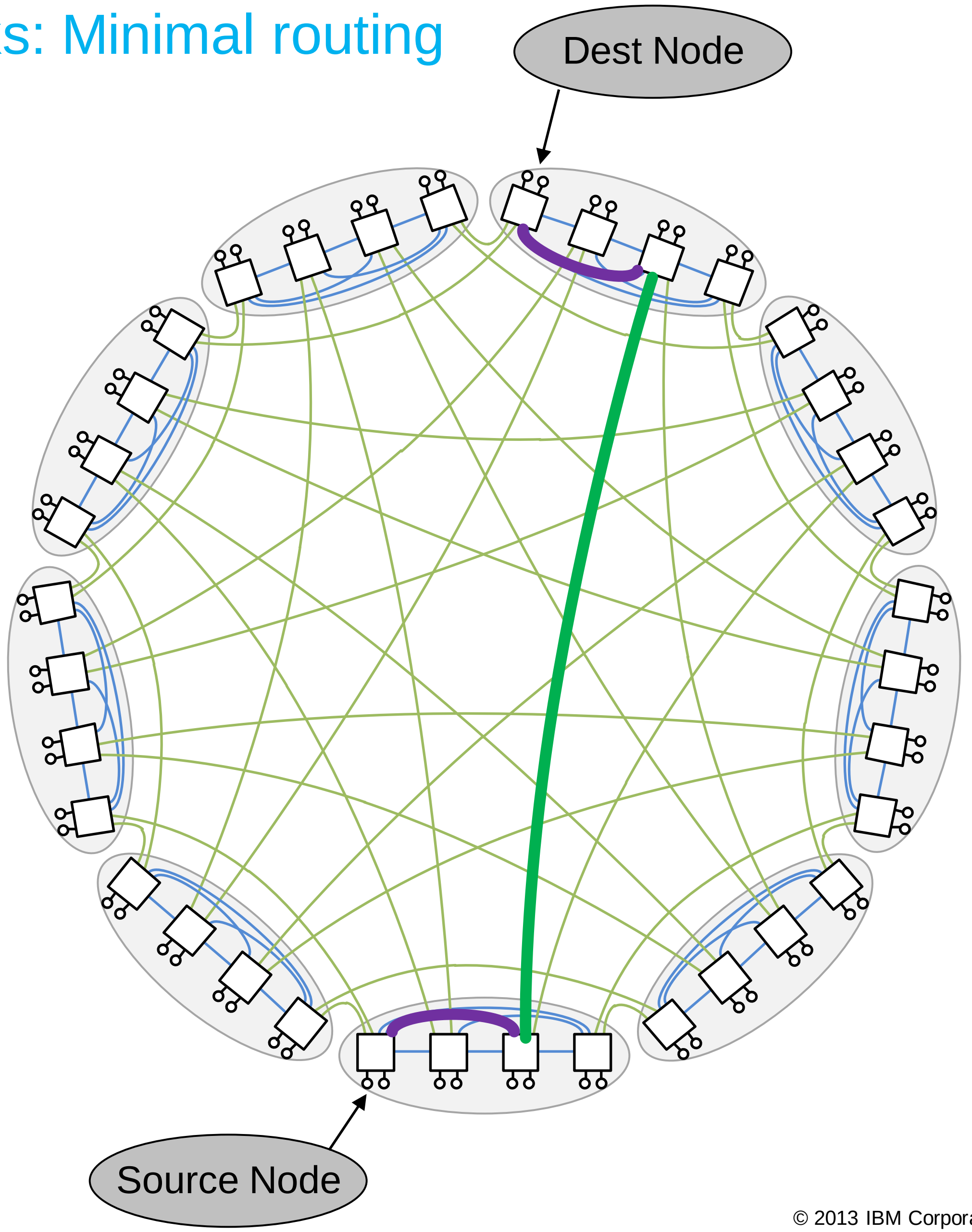
OFAR-CM 3

Performance results 4

Conclusions 5

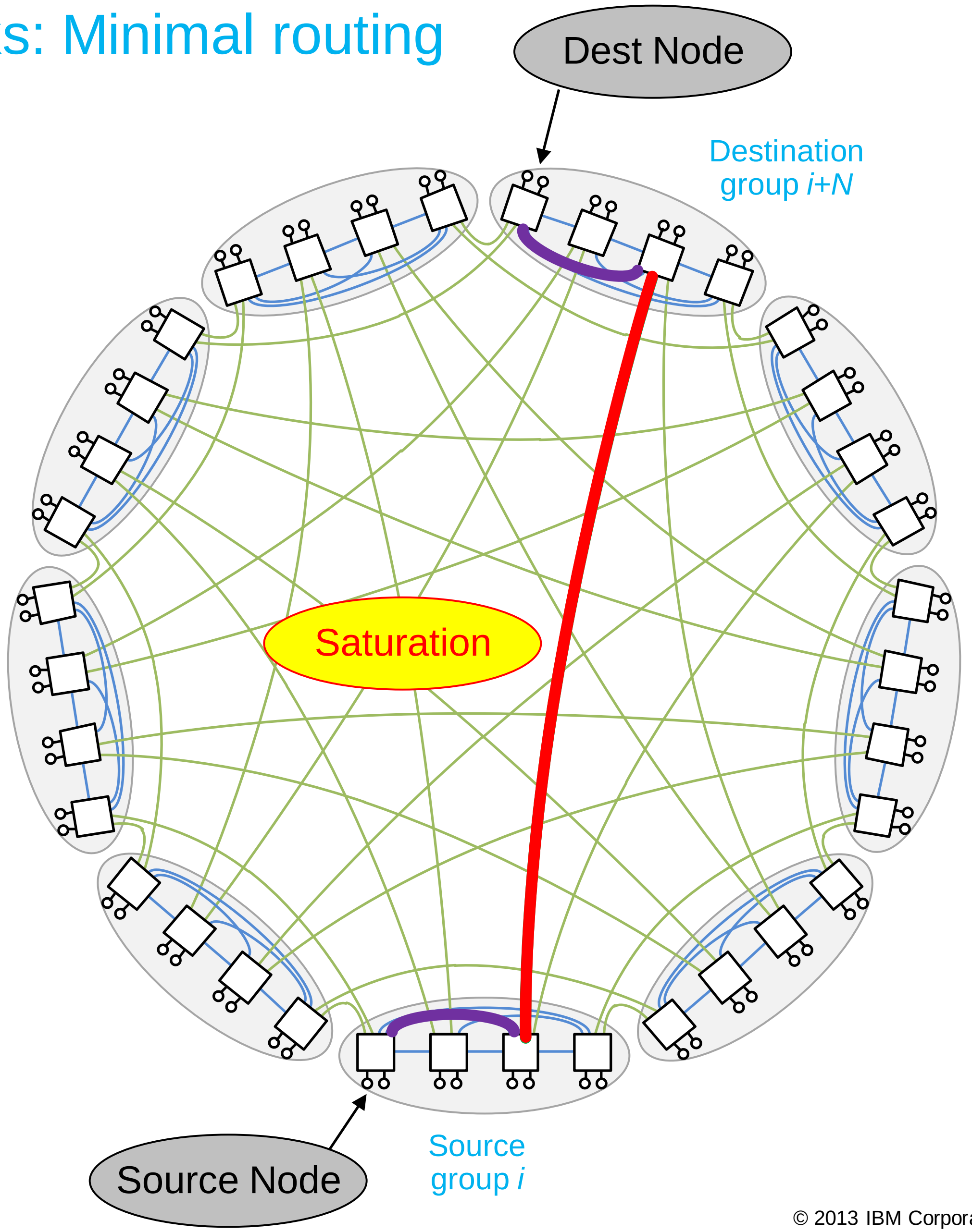
Routing in Dragonfly networks: Minimal routing

- Minimal Routing
  - Longest path 3 hops:
    - ➢ local – global – local
  - Deadlock avoidance:
    - ➢2 VC per local port + 1 VC per global port (2/1)
- Good performance under UN.
- Saturation of the global link with ADVG+N.



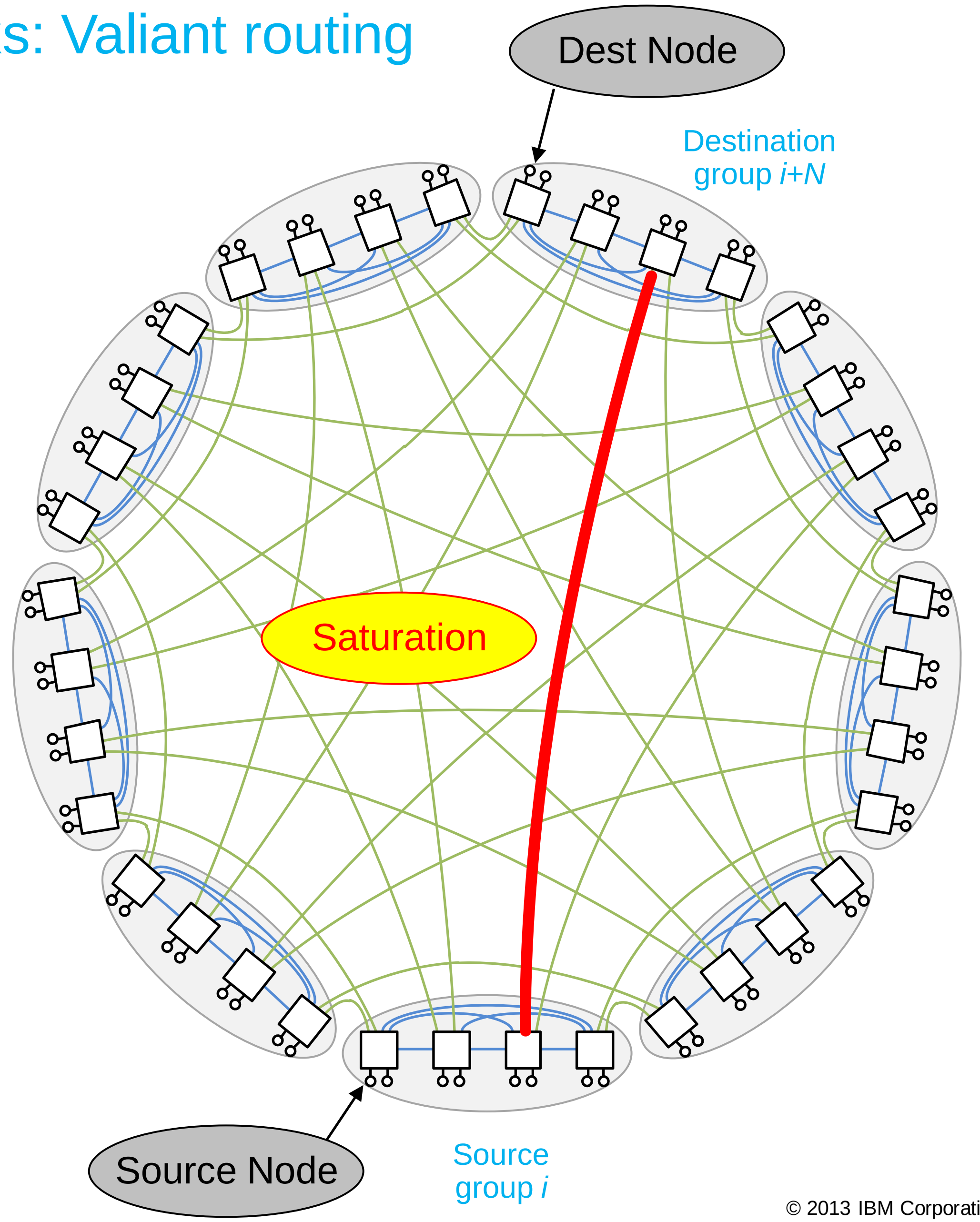
Routing in Dragonfly networks: Minimal routing

- Minimal Routing
  - Longest path 3 hops:
    - ➢ local – global – local
  - Deadlock avoidance:
    - ➢2 VC per local port + 1 VC per global port (2/1)
- Good performance under UN.
- Saturation of the global link with ADVG+N.



Routing in Dragonfly networks: Valiant routing

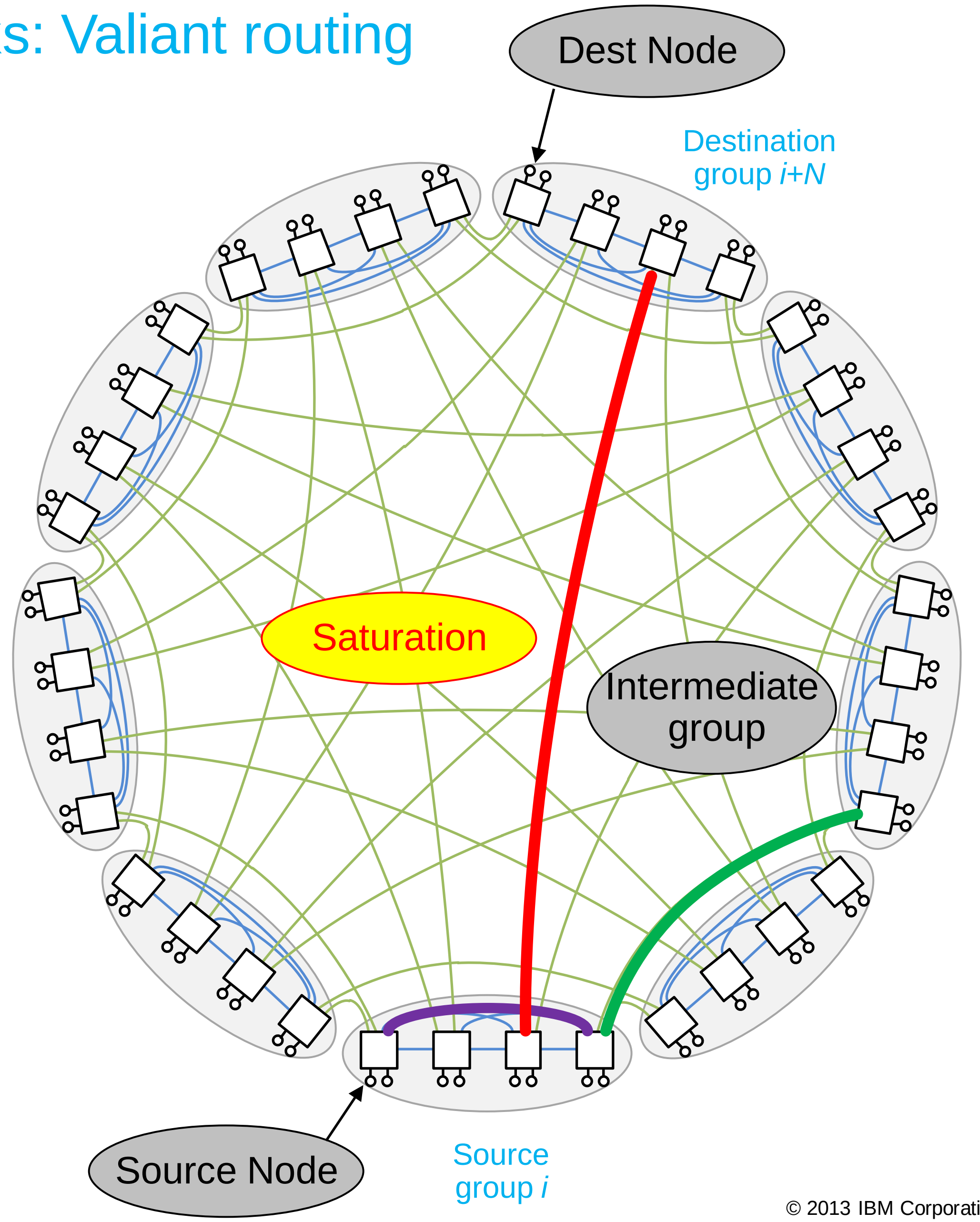
- Valiant Routing [4]
  - Misroutes packets to a random intermediate group.
  - Balances use of links
  - Doubles latency and halves throughput
  - Longest path 5 hops:
    - ➢ local – global – local global – local
  - Deadlock avoidance:
    - ➢ 3 VC per local port + 2 VC per global port (3/2)



▪[2] L. Valiant, "A scheme for fast parallel communication," SIAM journal on computing, vol. 11, p. 350, 1982.

Routing in Dragonfly networks: Valiant routing

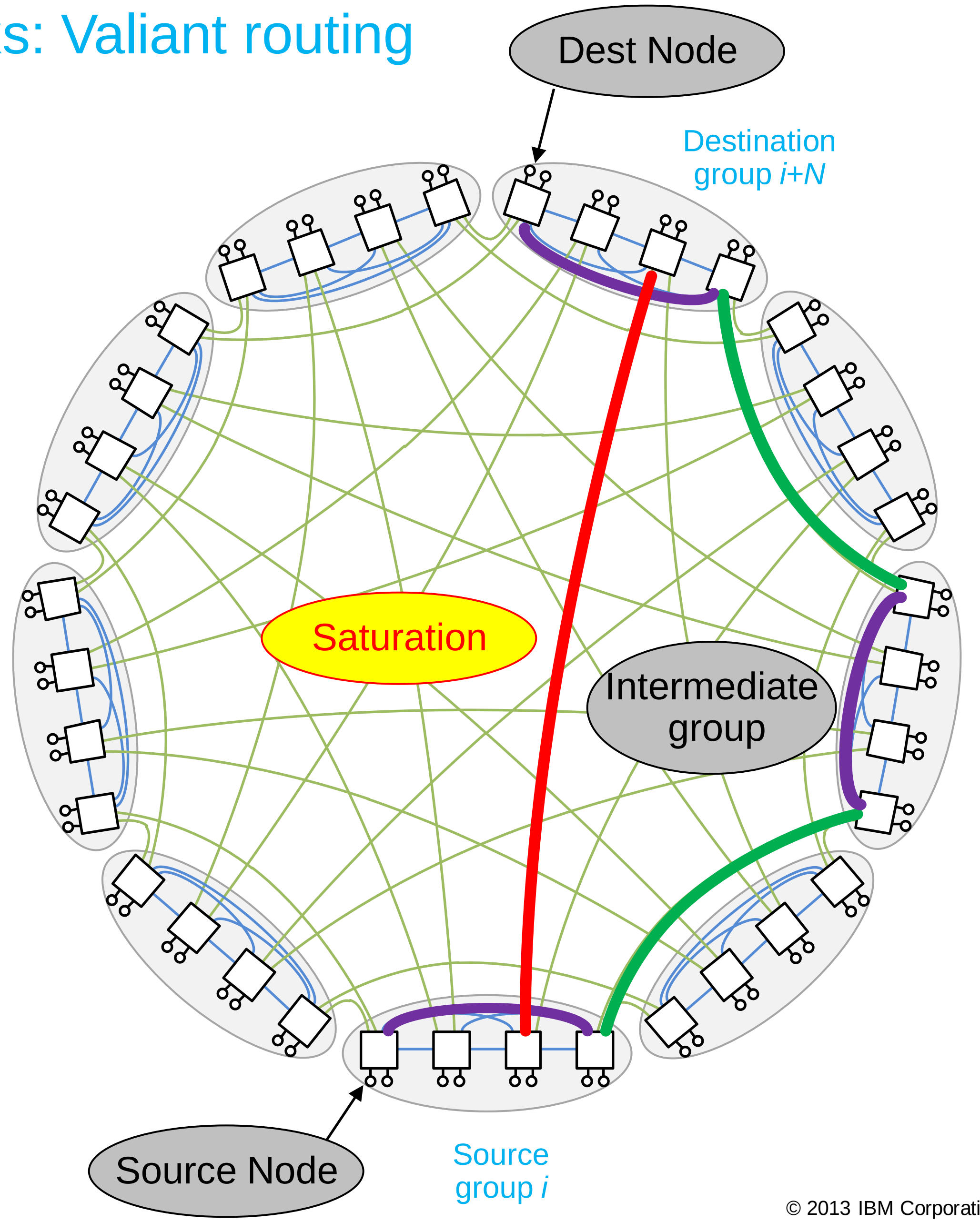
- Valiant Routing [4]
  - Misroutes packets to a random intermediate group.
  - Balances use of links
  - Doubles latency and halves throughput
  - Longest path 5 hops:
    - ➢ local – global – local global – local
  - Deadlock avoidance:
    - ➢ 3 VC per local port + 2 VC per global port (3/2)



▪[2] L. Valiant, “A scheme for fast parallel communication," SIAM journal on computing, vol. 11, p. 350, 1982.

Routing in Dragonfly networks: Valiant routing

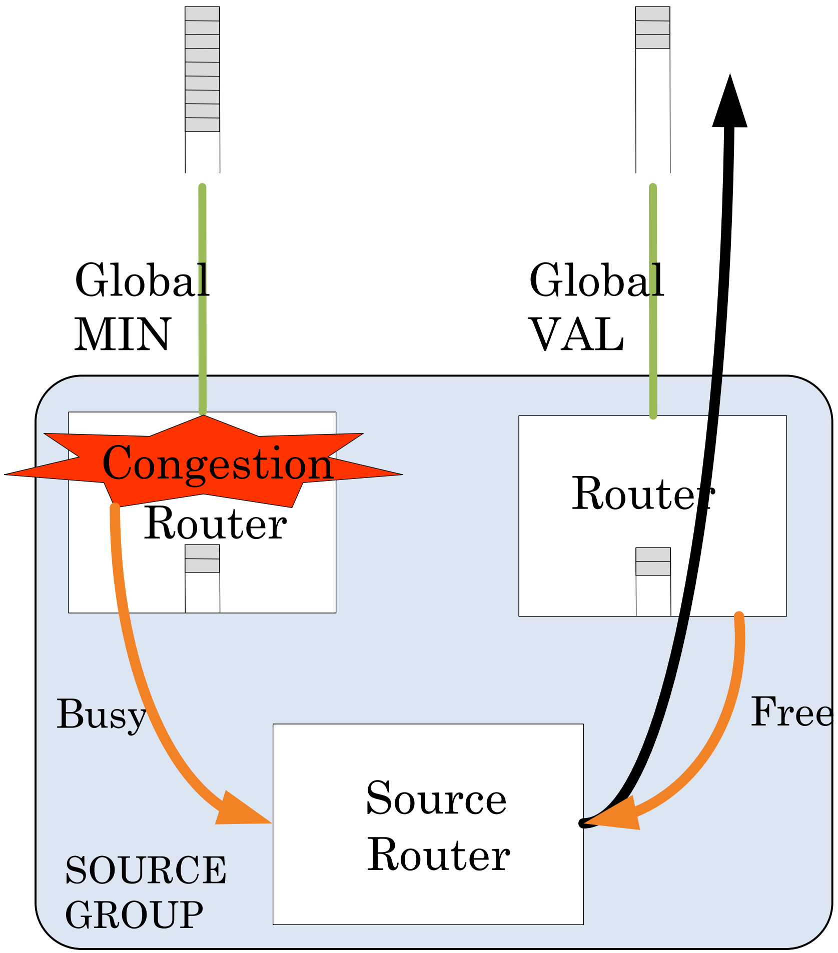
- Valiant Routing [4]
  - Misroutes packets to a random intermediate group.
  - Balances use of links
  - Doubles latency and halves throughput
  - Longest path 5 hops:
    - local – global – local global – local
  - Deadlock avoidance:
    - 3 VC per local port + 2 VC per global port (3/2)



- [2] L. Valiant, “A scheme for fast parallel communication," SIAM journal on computing, vol. 11, p. 350, 1982.

Routing in Dragonfly networks: Valiant routing

- Adaptive Routing
  - Maximizes performance.
  - Chooses between minimal and non-minimal routing.
  - Relies on the information about the state of the network.
- Piggybacking Routing (PB) [5]
  - Each router flags if a global queue is congested.
  - Broadcast information about queues
  - Source routing → Chooses between minimal and Valiant.
  - Deadlock Avoidance: 3 VC per local port + 2 VC per global port (3/2)



[5] Jiang, Kim, Dally. *Indirect adaptive routing on large scale interconnection networks. ISCA '09.*

Table of contents

Introduction 1

Routing in Dragonfly networks 2

OFAR-CM 3

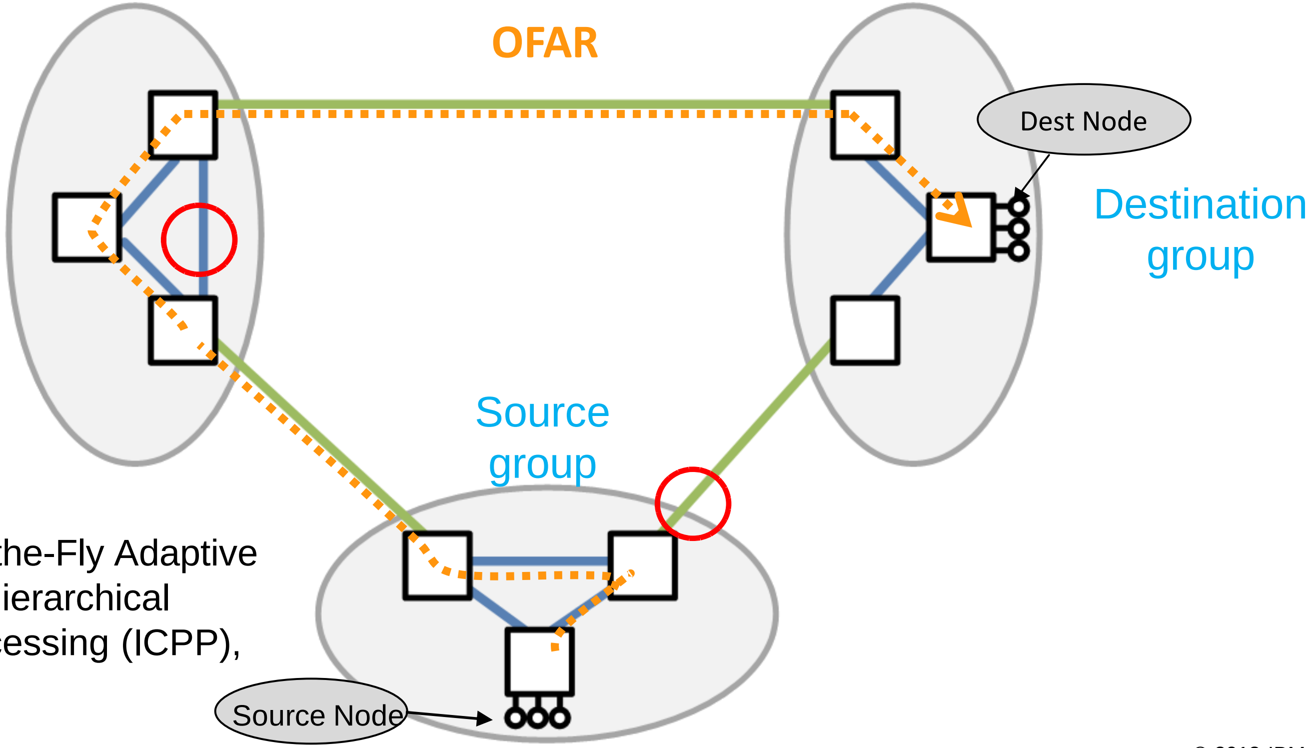
Congestion management 4

Performance results 5

Conclusions 6

OFAR-CM: Congestion management

- OFAR [6] revisits on each hop if a packet must be routed minimally or not
- Permits **local misrouting:** 2 local hops within a group to circumvent congested local link.
- Long routes: ***local*** **–** ***local*** **–** ***global*** **–** ***local*** **–** ***local*** **–** ***global*** **–** ***local*** **–** ***local*****: *8 hops***
- Naïve deadlock avoidance: 6 VC per local port + 2 VC per global port. (6/2)



[6] M. Garcia et al. "On-the-Fly Adaptive Routing in High-Radix Hierarchical Networks," Parallel Processing (ICPP), 2012.

OFAR-CM: Escape subnetworks

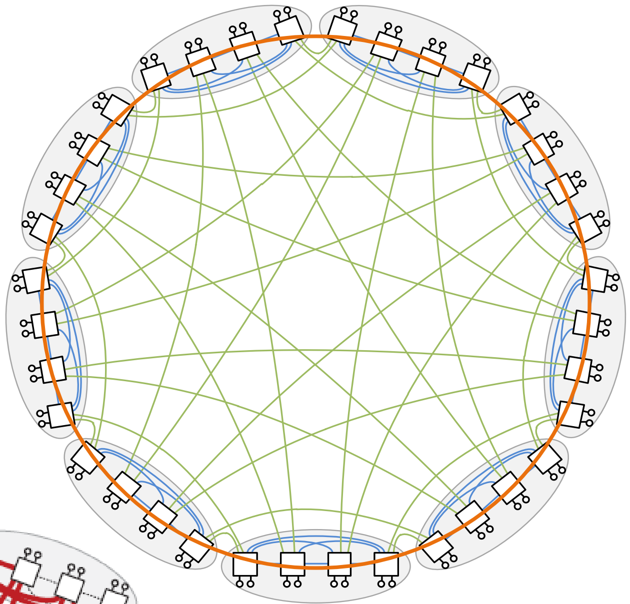
▪ OFAR implements a fully adaptive network without requiring virtual channels.

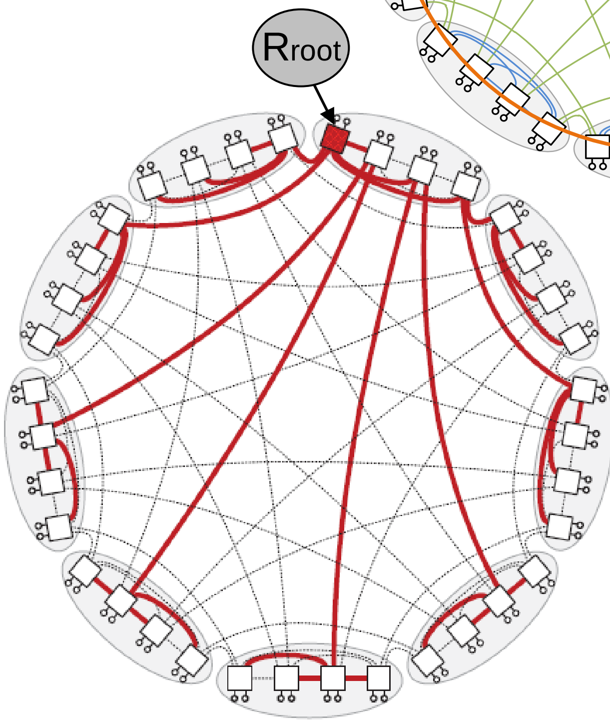
- It is deadlock-prone.

▪ A **deadlock-free escape sub-network** is used to guarantee deadlock-freedom.

- It connects all the routers in the network with **extra channels** or **VC** (+1)
- Packets are injected when they cannot advance on the canonical Dragonfly.
- Hamiltonian **ring** with injection restriction (Bubble flow-control [7]).
- Spanning-**tree** with up/down routing.

Hamiltonian ring escape subnetwork





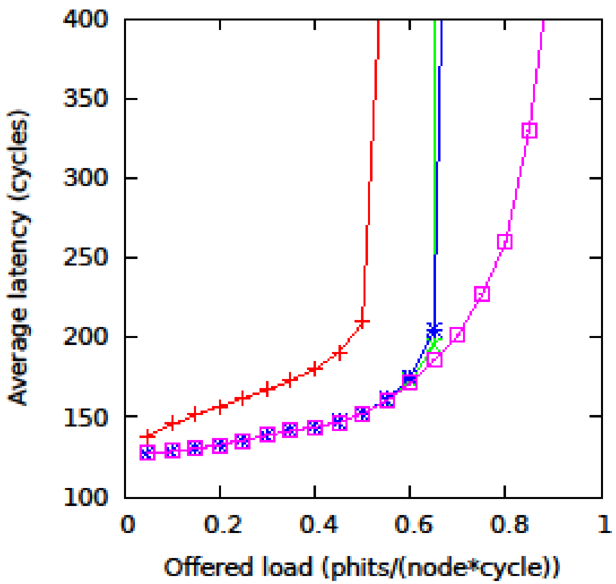
Tree escape subnetwork

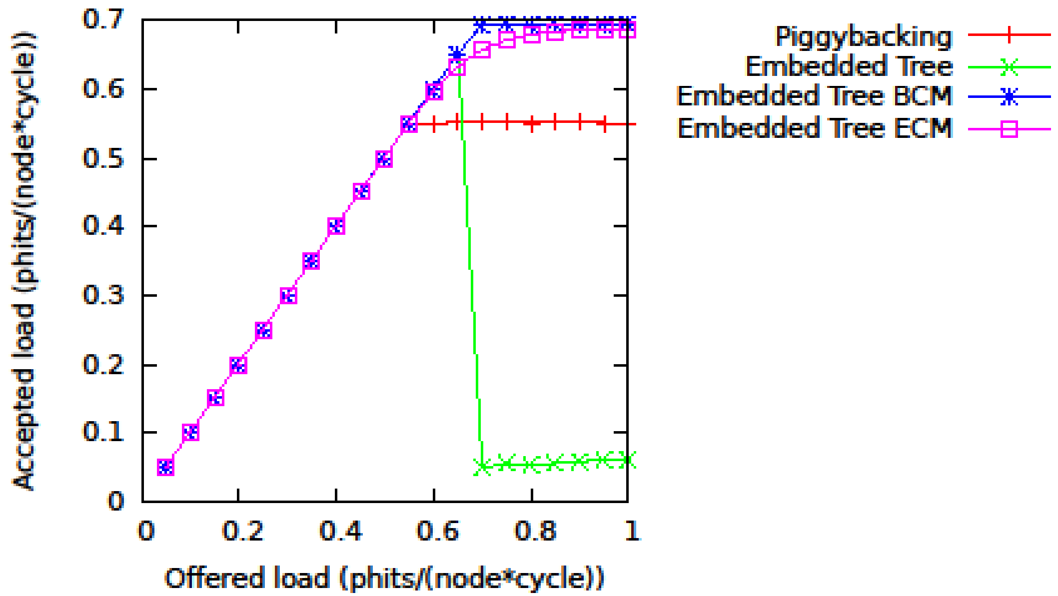
[7] C. Carrión, R. Beivide, J. Gregorio, and F. Vallejo, “A flow control mechanism to avoid message deadlock in k-ary n-cube networks," in HiPC, 1997.

OFAR-CM: Congestion management

- The capacity of the escape subnetwork is much lower than for the canonical Dragonfly → Possible significant drop of performance when all buffers are full.

- Latency an throughput depending on the congestion management employed
  - OFAR routing + Tree escape subnetwork
  - Uniform random traffic (UR)





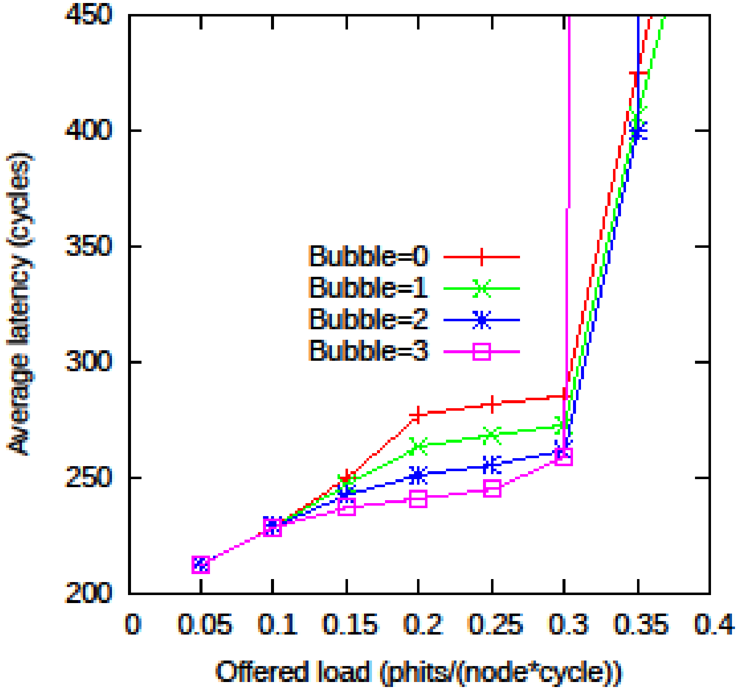
OFAR-CM: Congestion management

- Escape Congestion Management (**ECM**)
  - Employs the occupancy of the local buffers of the escape subnetwork as an indicator of congestion.
  - If the occupancy of all those buffers is higher than a given threshold. → Nodes will have to wait to a subsequent cycle to inject traffic.
  - The **threshold** size can range from 0% to 100% of the buffer size
  - The threshold is chosen empirically

- Base Congestion Management (**BCM**)
  - Forbids the injection of packets when the canonical (base) network is congested.
  - A packet can be injected in the network only if there is enough space in the next queue for one packet plus a given bubble.
  - The **bubble** size can range from 1 to the buffer size in packets minus 1
  - The bubble is chosen empirically to prevent over-throttling

OFAR-CM: Congestion management

- Throughput and latency depending on the Bubble size.
  - OFAR Ring 3/2(+1) virtual channels
  - Base Congestion Management **BCM**
  - Adversarial traffic (ADVG+2)



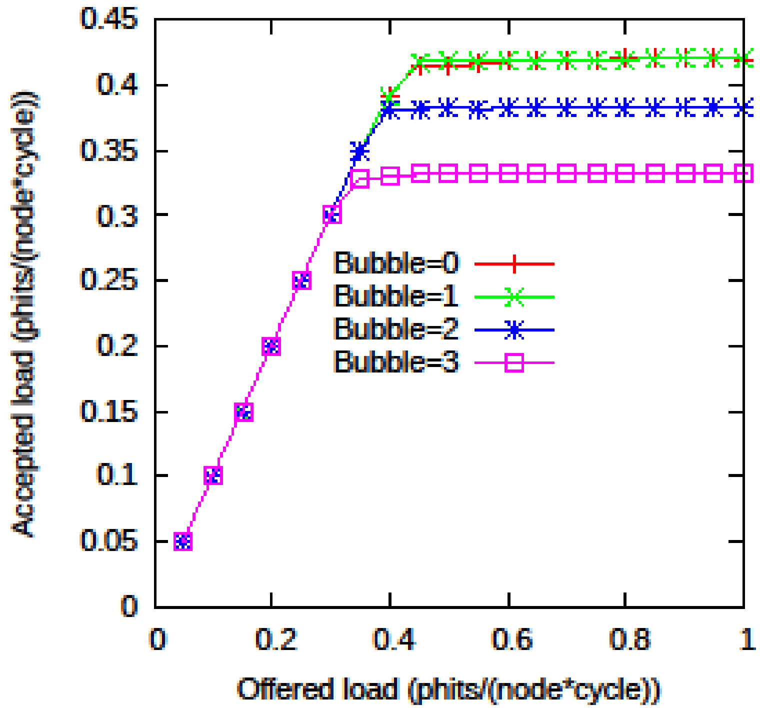


Table of contents

Introduction 1

Routing in Dragonfly networks 2

OFAR-CM 3

Performance results 4

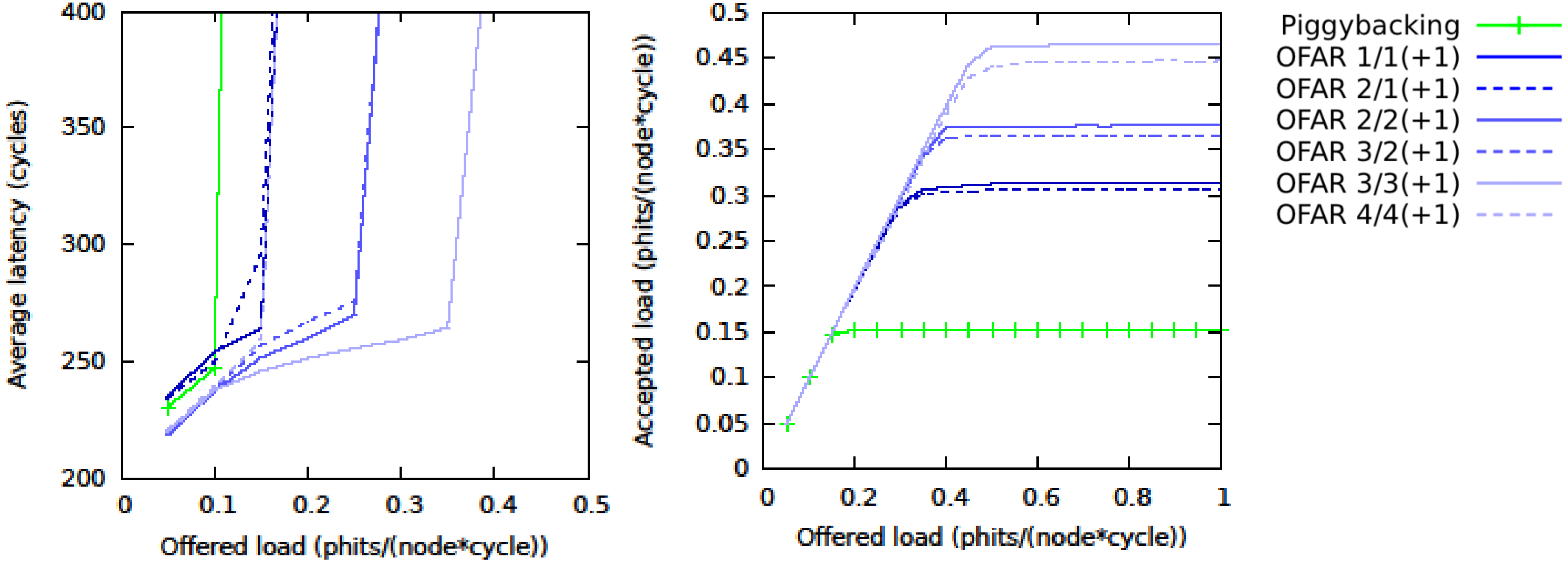
Conclusions 5

Performance results: Simulation setup

- **Dragonfly network simulator**
  - In-house developed time driven simulator
  - We model virtual cut-through input buffered routers with FIFO queues.
- Dragonfly with size:
  - **p** = 6 computing nodes per router.
  - **h** = 6 global ports per router.
  - **a** = 12 routers per group.
  - 5,256 computing nodes organized in 73 groups of 12 routers with 23 ports each.
  - Latencies are 10 cycles for local links and 100 for global links.
  - FIFO sizes are set to 32 phits for the local ones, and 256 phits for the global ones.
  - Packet length is 8 phits

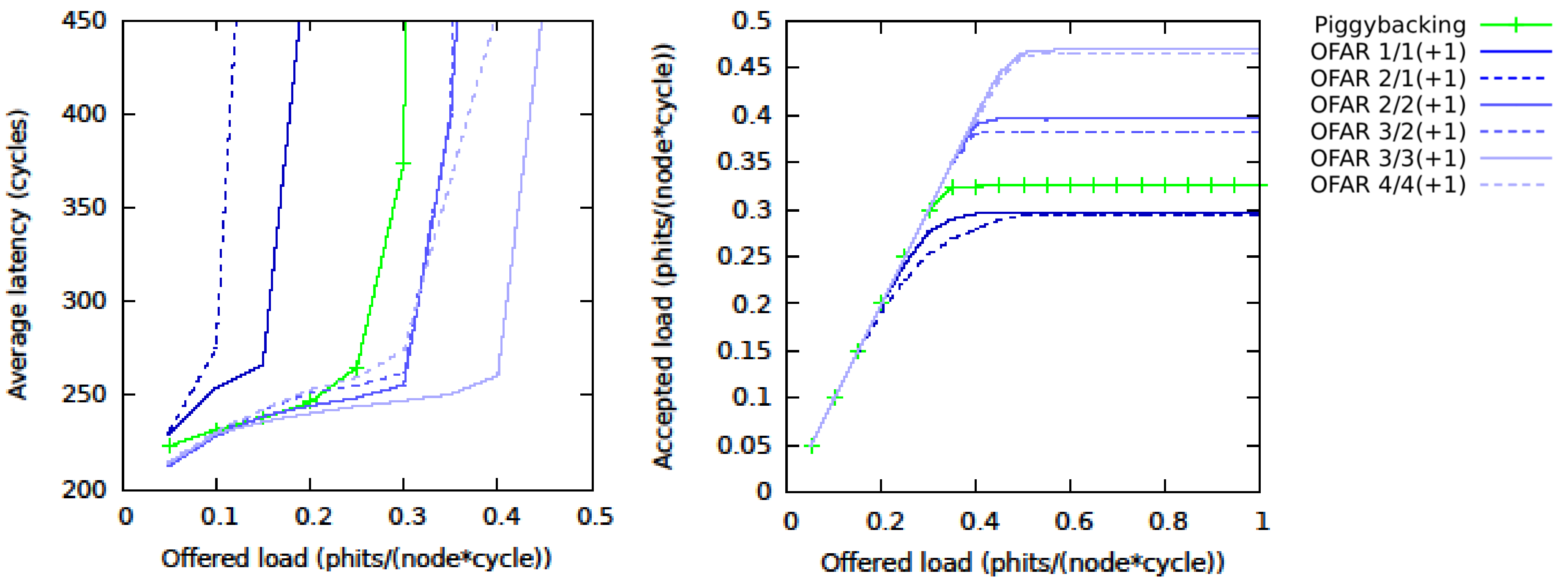
Performance results: Network resources

- Steady state adversarial global traffic + 6 (ADVG+h)
  - ADVG+6 is the most adversarial traffic in an h=6 Dragonfly
  - OFAR Ring. BCM bubble = 2
  - OFAR always outperforms BP



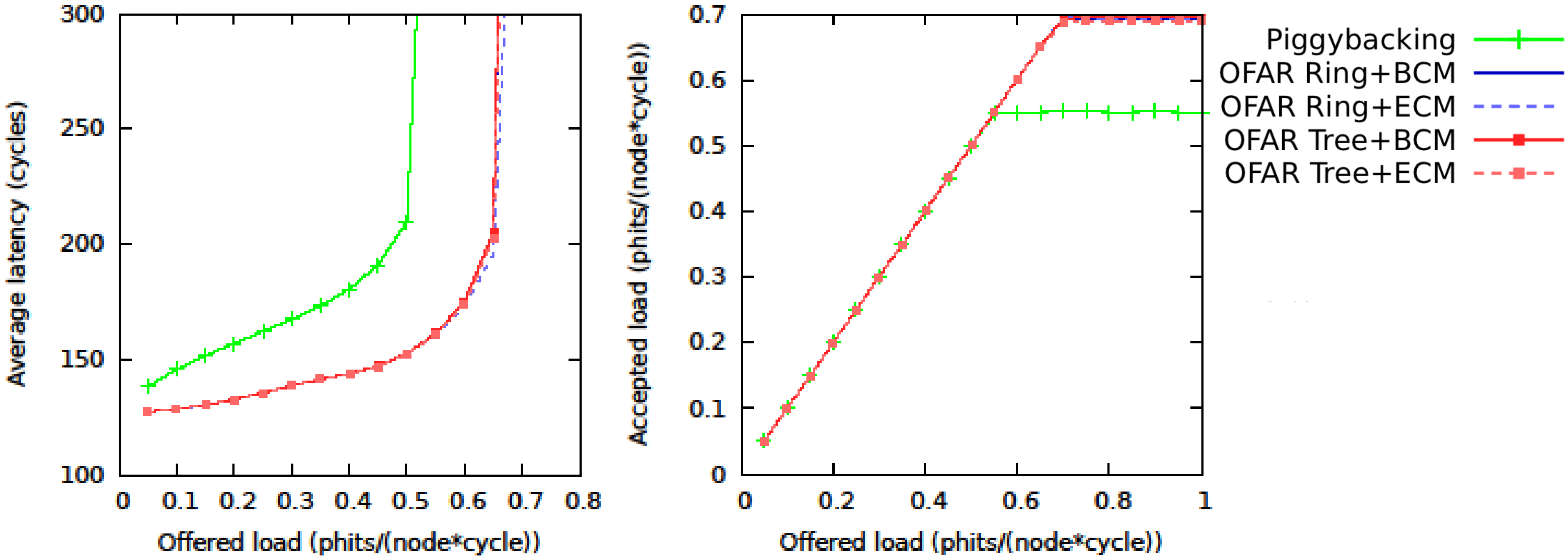
Performance results: Network resources

- Steady state adversarial global traffic + 2 (ADVG+2)
  - OFAR Ring. BCM bubble = 2.
  - OFAR with 2/1(+1) or less resources obtains worse performance than PB due to HoLB.
  - From now on we will use OFAR 2/1(+1) to study effects of congestion.



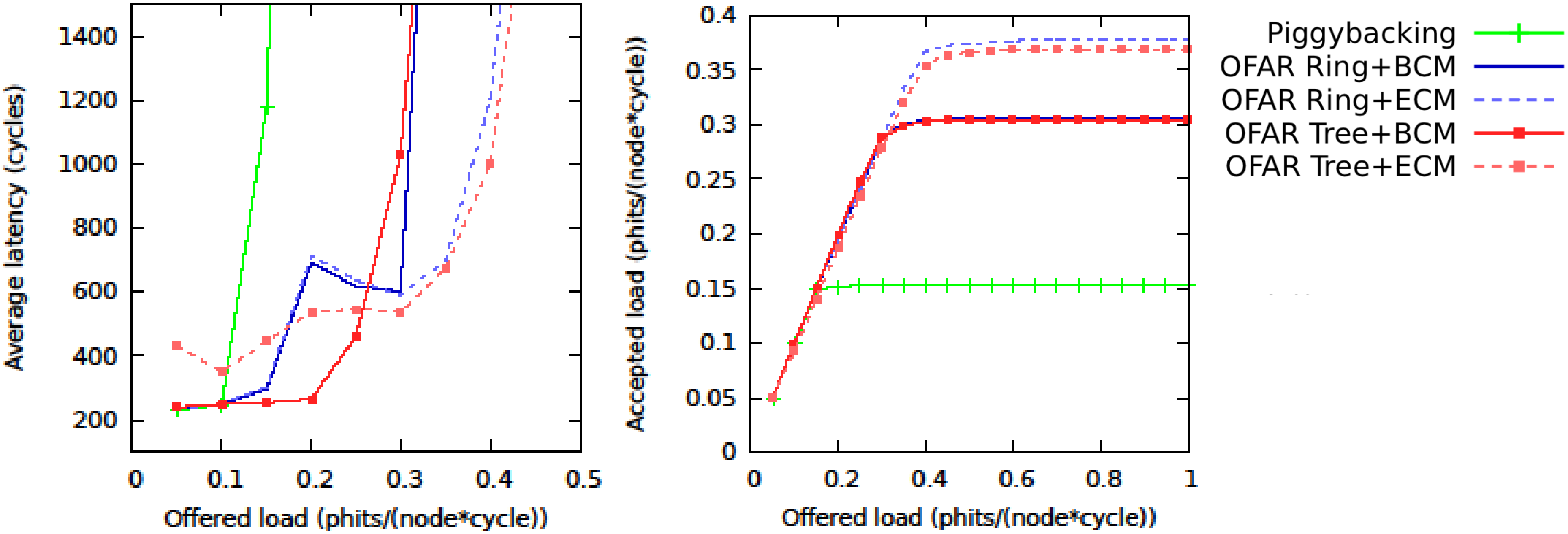
Performance results: Congestion management & escape subnetwork

- Steady state uniform random traffic (UR)
  - OFAR 2/1(+1)
  - All the configurations outperform PB



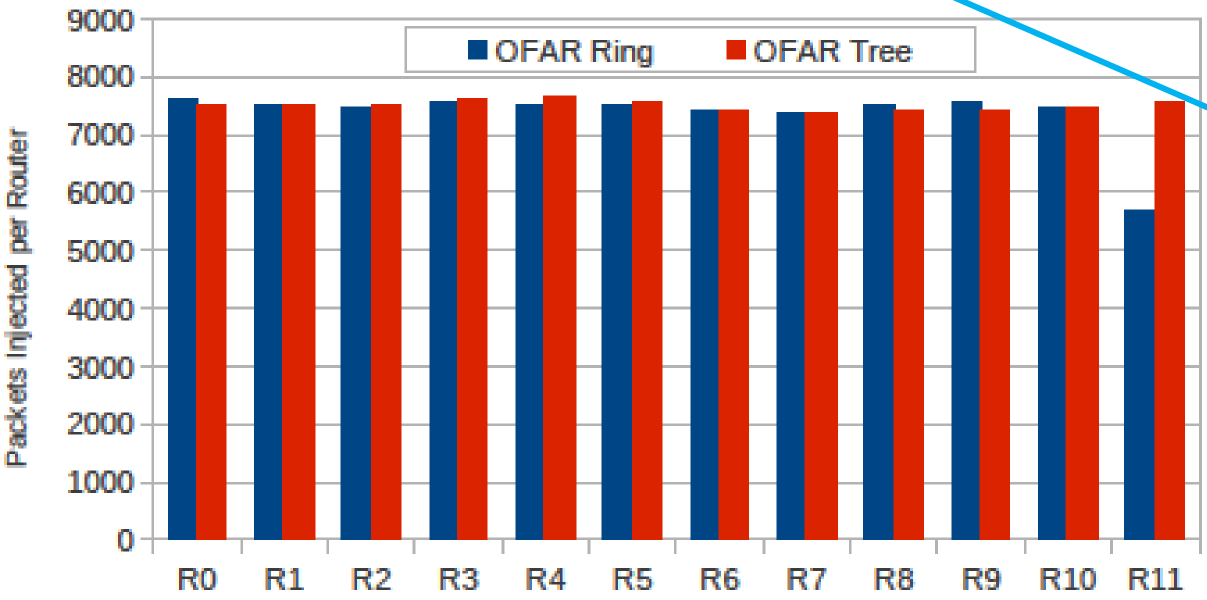
Performance results: Congestion management & escape subnetwork

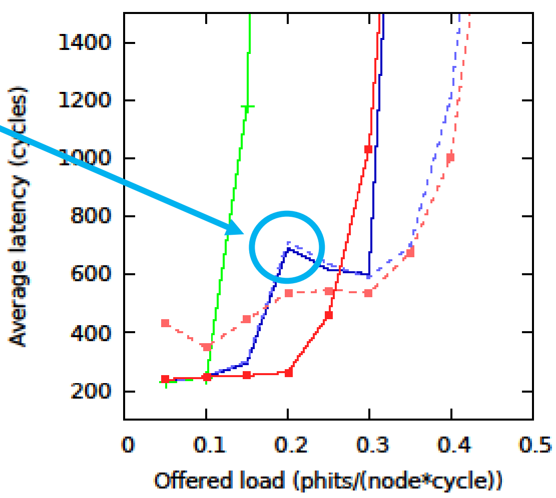
- Steady state adversarial global traffic + 6 (ADVG+h)
  - OFAR 2/1(+1)
  - All OFAR configurations outperform PB.
  - ECM provides better performance than BCM



Performance results: Network fairness

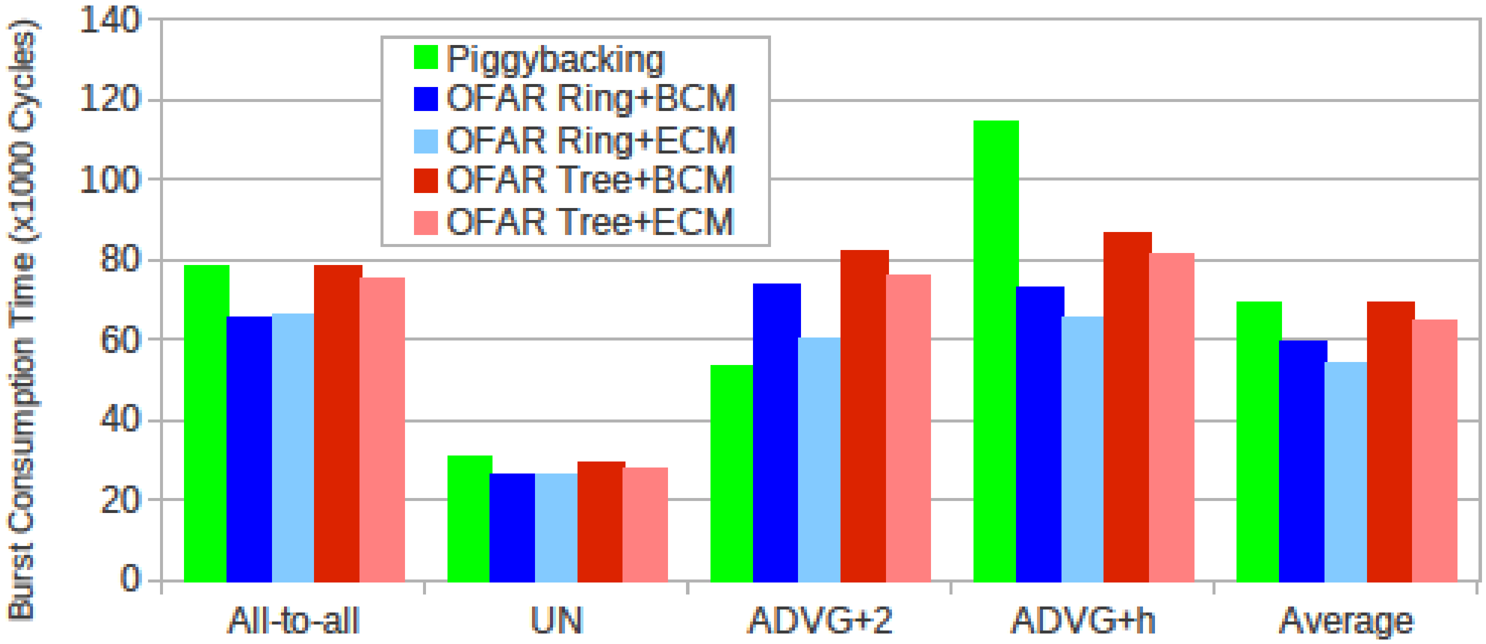
- Number of packets injected by each router in group 0
  - Offered traffic load of 0.2 phits/(node*cycle)
  - OFAR Ring: Escape traffic leave the group through R11. It injects 25% less packets than the rest of the routers in the group





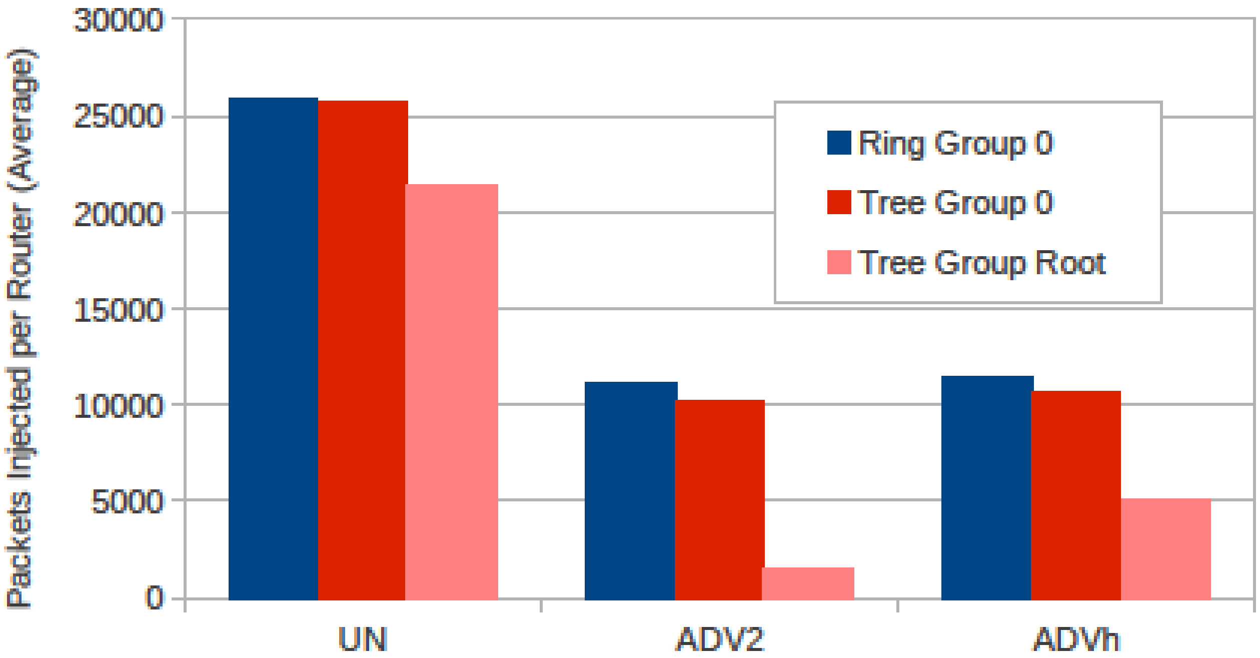
Performance results: Traffic consumption

- Traffic consumption
  - Cycles required to consume 2,000 packets/node at 1phit/(node*cycle) applied load.
  - Traffic patterns: All-to-all, UR, ADVG+1 and ADVG+h
  - OFAR Tree is slower than OFAR Ring consuming traffic



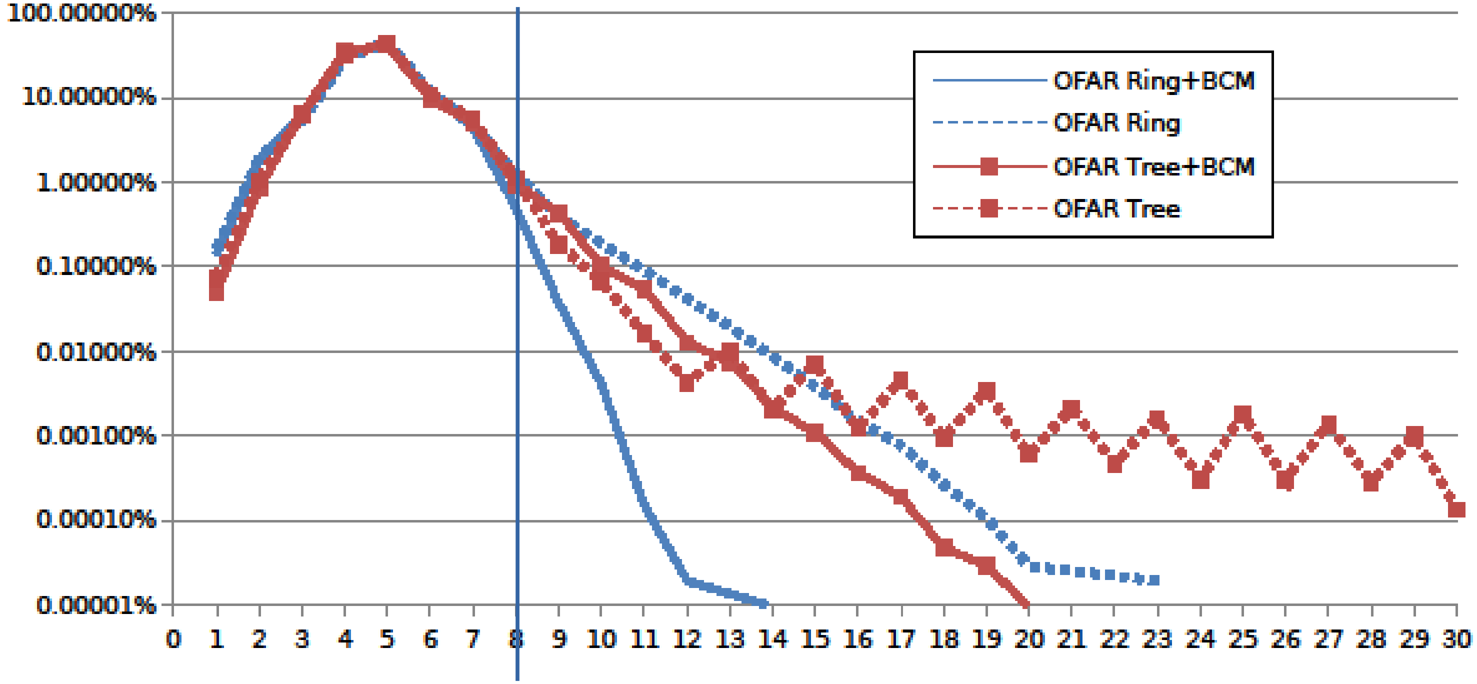
Performance results: Network fairness

- Total number of packets injected by nodes in group G_0 and G_{root} in 50,000 cycles
  - Congestion management: BCM
  - Traffic: UR, ADVG+2 and ADVG+h
  - G_{root} is saturated due to the concentration of traffic in that group → Routers in G_{root} prohibit packet injection, and average packet latency increases.



Performance results: Length of network paths

- ADVG+6 traffic
  - OFAR Ring provides shorter paths than OFAR Tree.
  - OFAR Tree: G_{root} is more prone to congestion and multiple injections are more likely
  - More than a **99.99%** of the packets need **less than 30 hops** to reach its destination



Performance results: Length of network paths

- In practice **unbounded paths do not occur when using congestion management**.
- Simple mechanism to limit the number of subnetwork injections and bound path lengths:
  - Packets need a **counter**, incremented on each escape subnetwork injection.
  - Once counter saturates (for example, 15 injections for a 4-bit counter) → Packet is forced to continue through the escape subnetwork until reaching its destination.
  - With congestion management, max. subnetwork injections = 12 times → Not significant impact on performance.

Table of contents

Introduction 1

Routing in Dragonfly networks 2

OFAR-CM 3

Performance results 4

Conclusions 5

Conclusions

- OFAR-CM combines OFAR with simple injection throttling.
  - Only relies on **local information**
  - Supports **local and global misrouting** without increasing the number of VCs
  - Achieves **higher performance** thanks to the higher routing freedom.
- With **similar cost** (VC), our proposal clearly **outperforms** alternatives such as **PB**.
- Implementations with **lower cost** might suffer **unfairness issues.** In such case, we have evaluated:
  - Two **congestion management** mechanisms, **BCM** and **ECM** that avoid network saturation that could lead to a performance drop.
  - Two **escape subnetwork** topologies, a **Hamiltonian ring** and **tree** a and how they affect network load imbalance and performance.
- Results show that, despite path lengths with OFAR-CM are unbounded in theory, they are relatively short in practice.

Thank you