ANALYSIS AND IMPROVEMENT OF VALIANT ROUTING IN LOW-DIAMETER NETWORKS

Mariano Benito
Pablo Fuentes
Enrique Vallejo
Ramón Beivide

UNIVERSIDAD DE CANTABRIA

With support from:

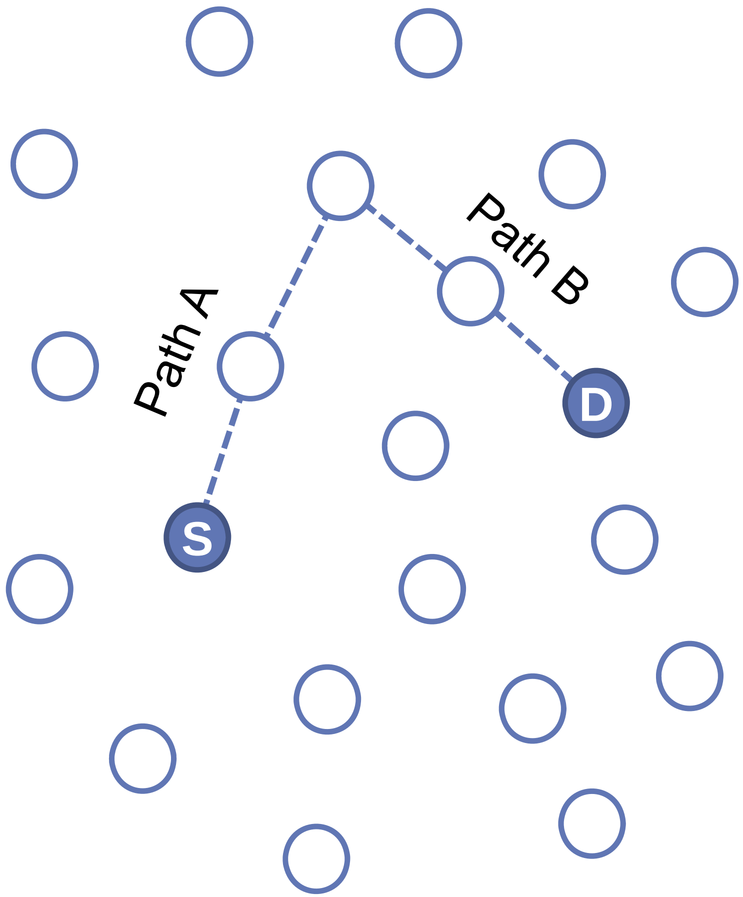
4th IEEE International Workshop of High-Perfomance Interconnection Networks in the Exascale and Big-Data Era (HiPINEB)
Vienna, Austria, 24-Feb, 2018.

Index

1. Background and motivation
2. Improvements to Valiant routing
3. Performance evaluation
4. Discussion and conclusions

1. Background and motivation

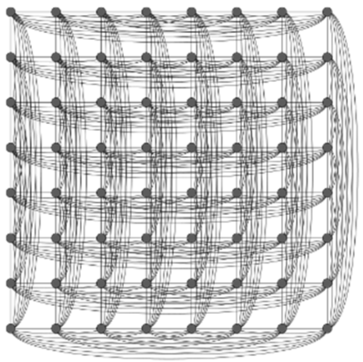
- Valiant routing
  - Randomized Routing mechanism originally proposed by Leslie Valiant for Hypercubes in [1] and square mesh, d-way shuffle and shuffle-exchange graphs networks in [2] .
  - Diverts traffic to an intermediate router
  - Double path length on average wrt minimal routing
  - Bounded worst-case permutation time
  - *Oblivious*



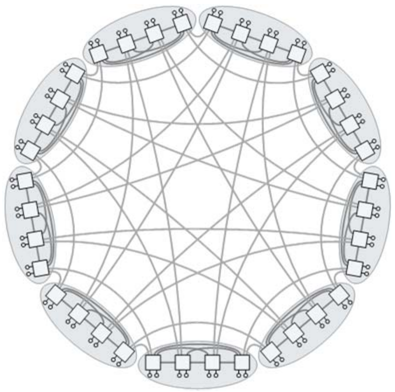
[1] L. Valiant, “A scheme for fast parallel communication,” SIAM journal on computing, vol. 11, p. 350, 1982
[2] L. G. Valiant, “Optimality of a two-phase strategy for routing in interconnection networks,” IEEE Trans. Comput., vol. 32, no. 9, pp. 861–863, Sep. 1983.

1. Background and motivation

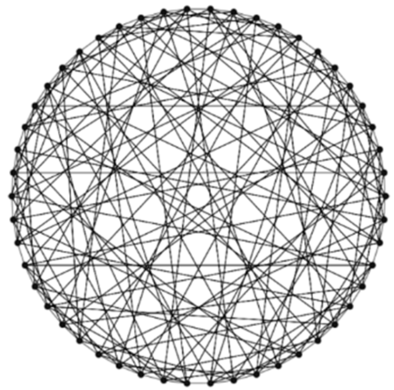
- Valiant has been used in low-diameter system networks
- Highly-scalable, low-cost topologies



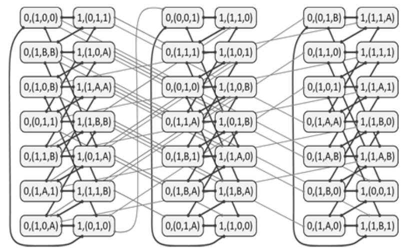
Flattened-Butterfly [3]



Dragonfly [4]



Slim Fly [5]



Projective Network [6]

- Low diversity of minimal paths
- Concentration (multiple nodes per switch)

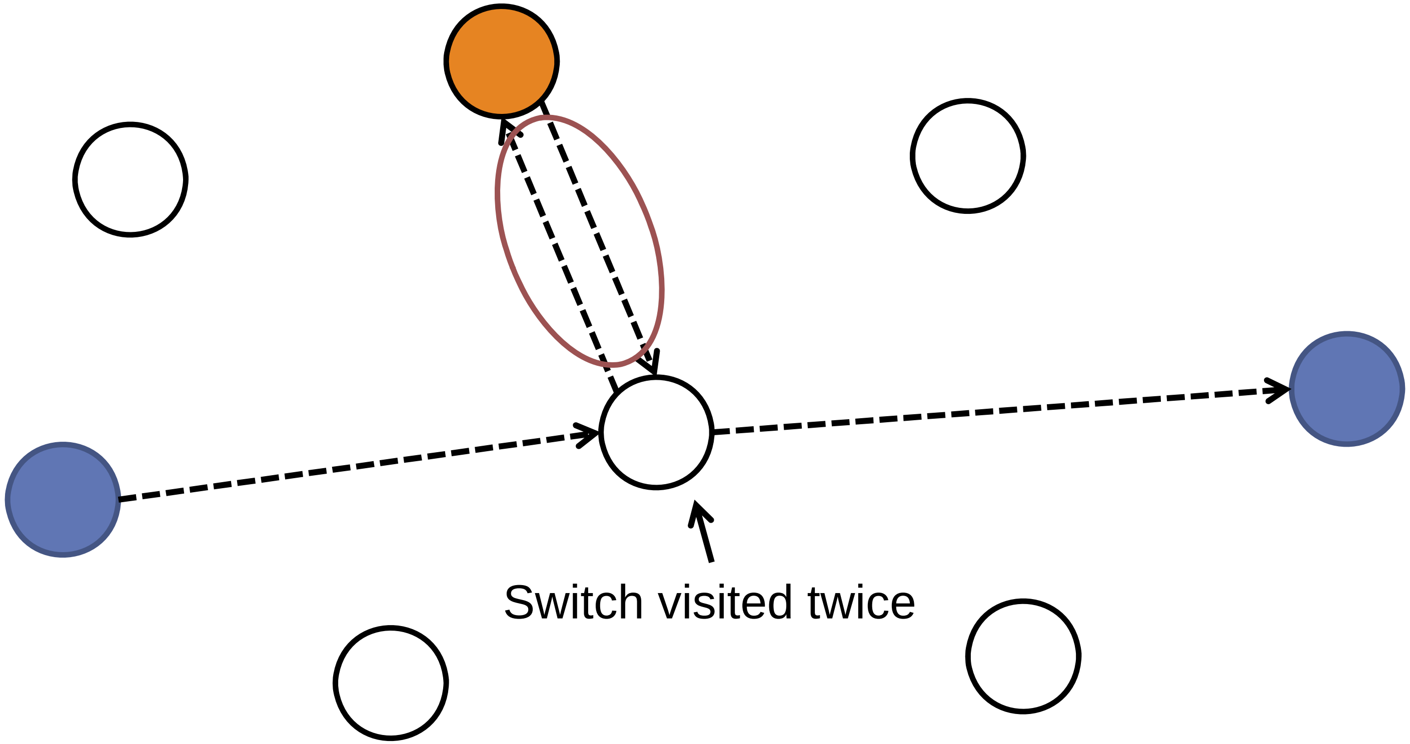
Congestion-prone

- Valiant routing avoids such patterns of congestion
  - Often implemented as part of an adaptive routing mechanism.

[3] Kim, Dally, Abts. *Flattened Butterfly : A Cost-Efficient Topology for High-Radix Networks.* ISCA'07
[4] Kim, Dally, Scott, Abts. *Technology-Driven, Highly-Scalable Dragonfly Topology*. ISCA '08
[5] Besta, Hoefler. *Slim Fly: A Cost Effective Low-Diameter Network Topology.* SC'14.
[6] Camarero, Martínez, Vallejo, Beivide. *Projective networks: Topologies for large parallel computer systems.* TPDS'17

1. Background and motivation

- Yébenes *et al*. [7] identified the *turnaround problem* when using Valiant routing in Slim Fly networks

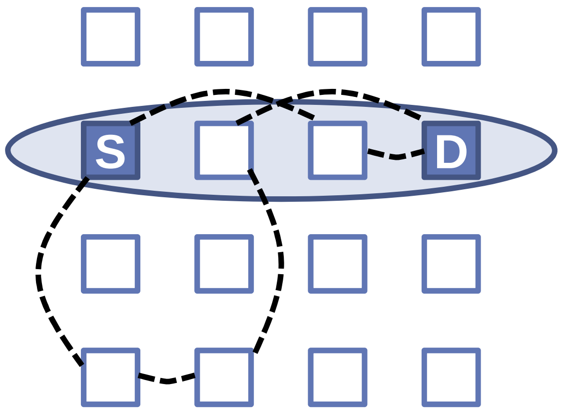


[7] P. Yébenes, J. Escudero-Sahuquillo, P. J. García, F. J. Quiles, and T. Hoefler, "Improving non-minimal and adaptive routing algorithms in slim fly networks," in HOTI'17

Index

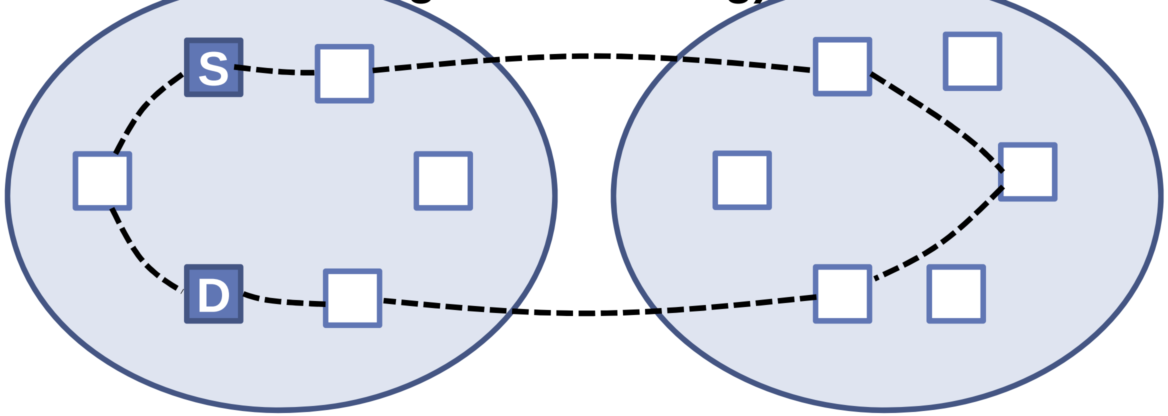
2. Improvements to Valiant Routing
2.1 Intermediate router selection

- A variant of the turn-around problem may occur without packets visiting switches twice.
  - Packets leave and return to a given network partition.



Flattened Butterfly

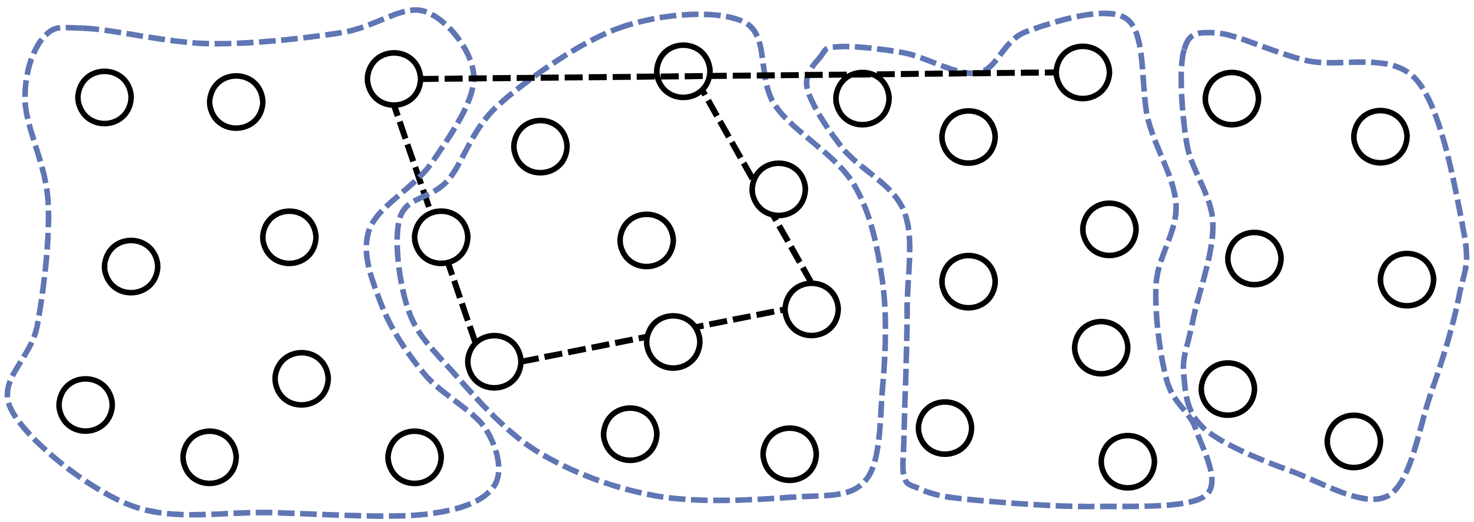
Dragonfly (with global trunking)



2. Improvements to Valiant Routing

2.1 Intermediate router selection

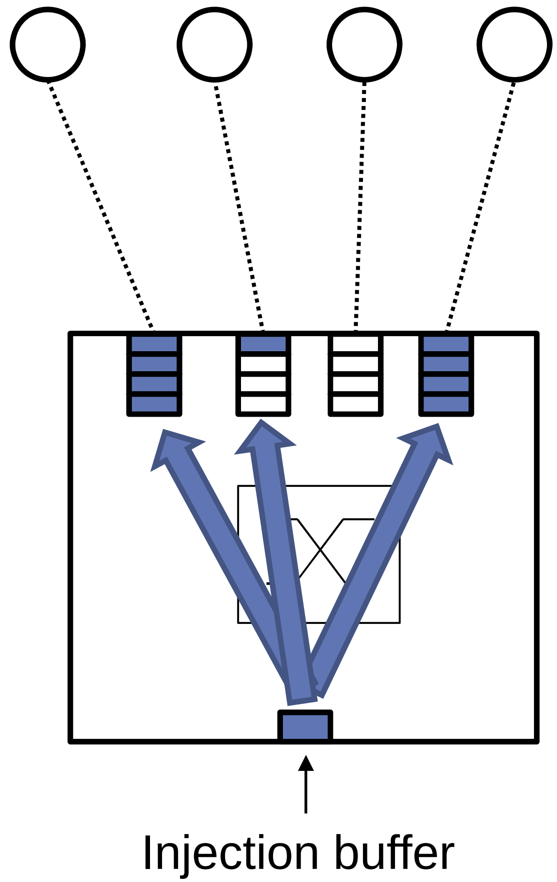
1. Determine network partitions
2. When both source & dest. nodes belong to the same partition
   1. Select intermediate node inside the partition
3. Otherwise
   1. Select intermediate node anywhere in the network.



2. Improvements to Valiant Routing

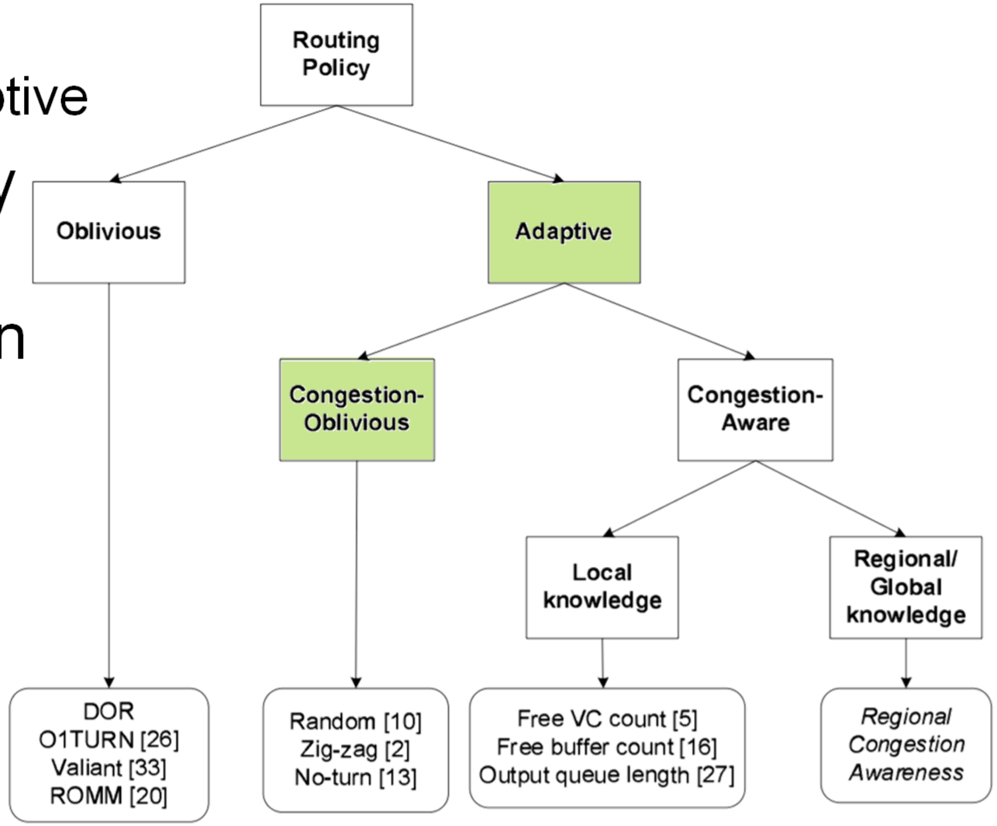
2.2 Recomputation

- Congestion situations may appear despite the randomization mechanism
  - Particularly when Valiant is used as part of an *adaptive* routing mechanism
- *Valiant with recomputation* makes a new intermediate node selection when the output port is not available



2. Improvements to Valiant Routing
2.2 Recomputation

- Valiant with Recomputation (VAL-Recomp) is no longer oblivious
  - But it is not completely adaptive
- According to the taxonimy by P. Graz *et al* in [8], Valiant with recomputation is *adaptive, congestion-oblivious*



[8] P. Gratz, B. Grot, and S. W. Keckler, “Regional congestion awareness for load balance in networks-on-chip,” HPCA’08

Index

3. Performance evaluation
3.1 Simulation setup

- Dragonfly [4] network modelled using FOGSim [9]

Parameter	Value
Router size	23 ports (h=6 global, p=6 injection, 11 local)
Group size	12 routers, 72 computing nodes
System size	73 groups, 876 routers, 5,256 computing nodes
Latency (ns)	40/400 (local/global links), 200 (router pipeline)
Buffer size (KB)	100 KB (transit queues), 200 (injection buffers)
Router	2x frequency speedup, Virtual Cut-Through, iterative input-first separable allocator
Routing mechanisms	Minimal (MIN) Valiant (VAL) Restricted Valiant (RVAL) Valiant-Recomp (VAL-Recomp) Restricted Valiant-Recomp (RVAL-Recomp)

Topology

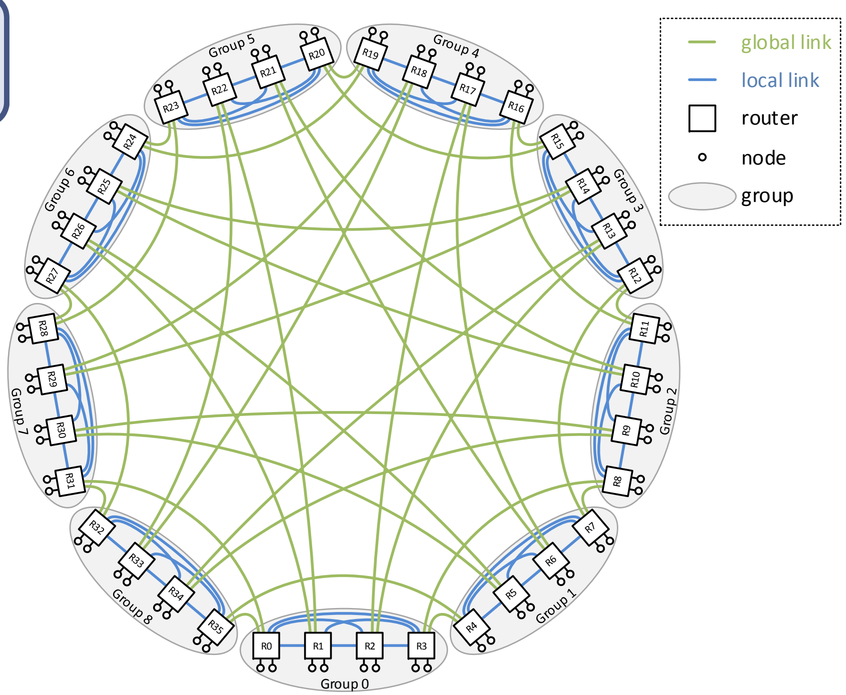
Uniform traffic

Adversarial traffic

Adv-local traffic

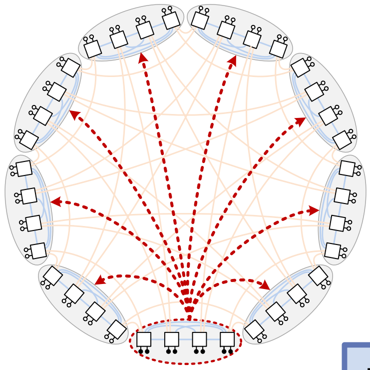
Hot-Region traffic

Permutation traffic

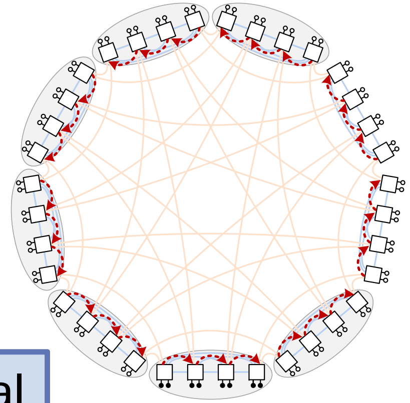


Strict permutation: No endpoint congestion
Random: pattern differs for each simulation, but is fixed during each simulation

[4] Kim, Dally, Scott, Abts. *Technology-Driven, Highly-Scalable Dragonfly Topology*. ISCA '08
[9] García *et al.*, *FOGSim Interconnection Network Simulator*, http://fuentesp.github.io/fogsim/



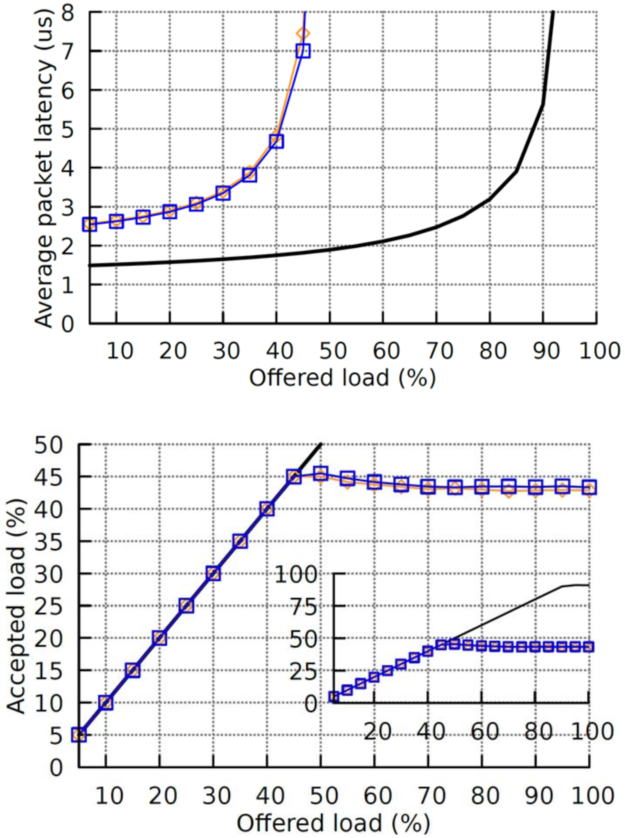
3.2 Restricted Valiant (RVAL)

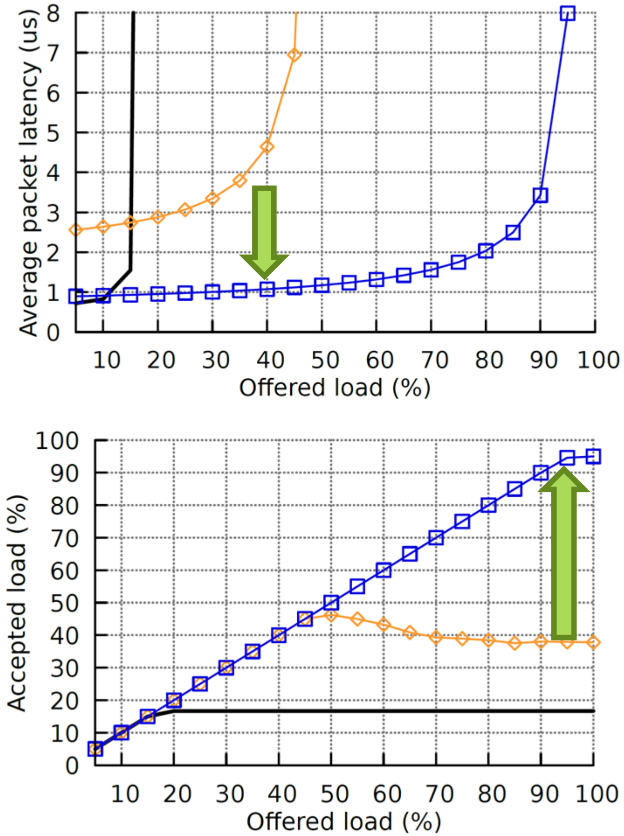


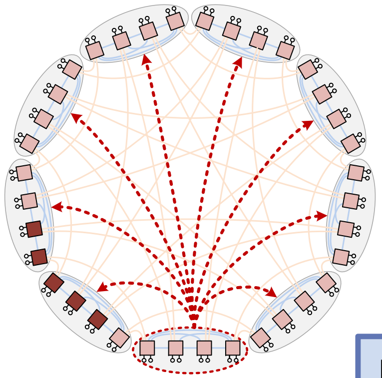
MIN VAL RVAL

Random Uniform

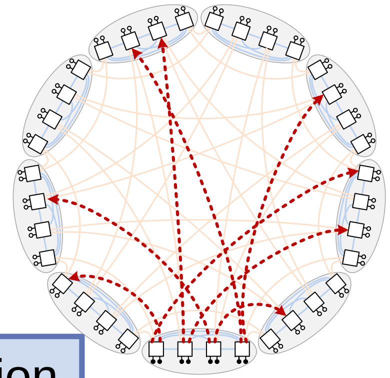
Adversarial-local







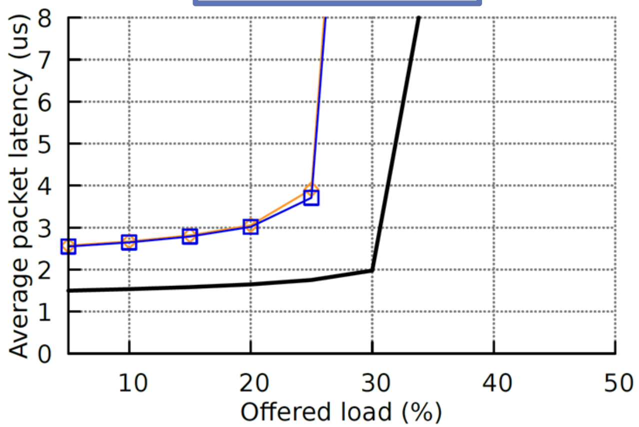
3.2 Restricted Valiant (RVAL)

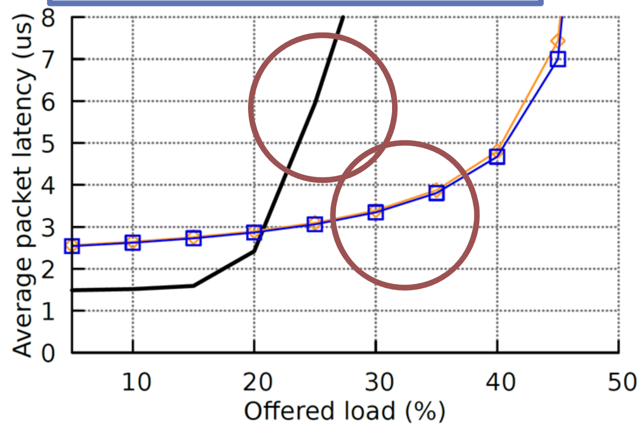


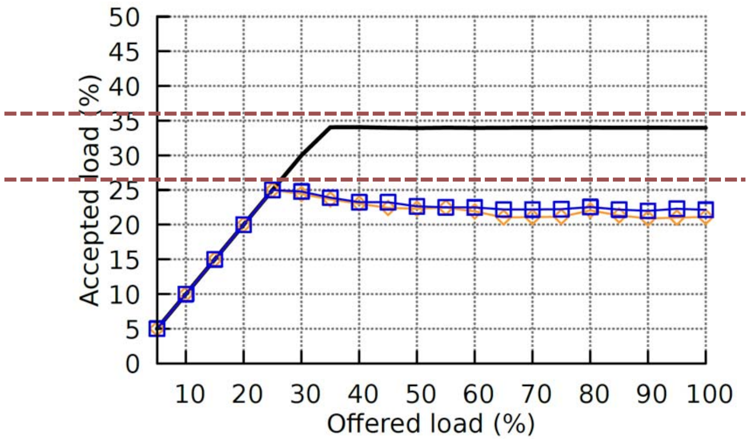
MIN VAL RVAL

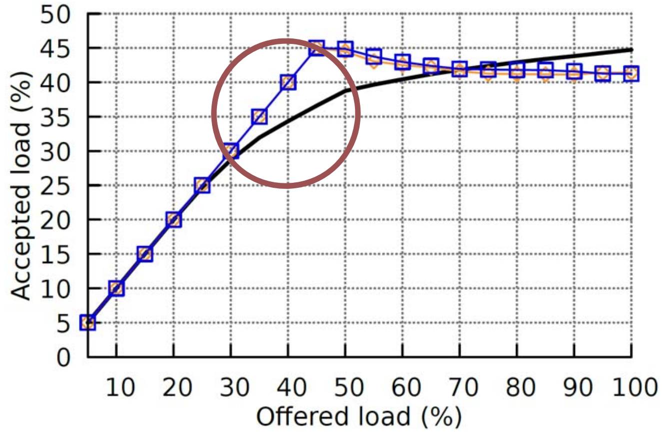
Hot-Region

Random permutation



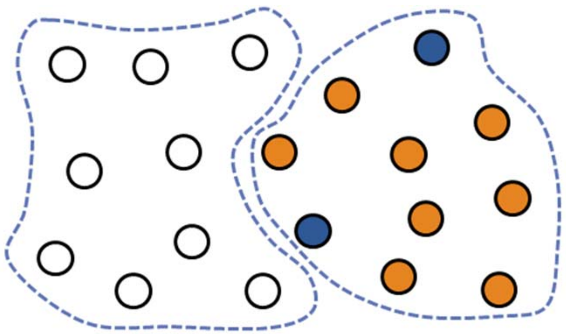


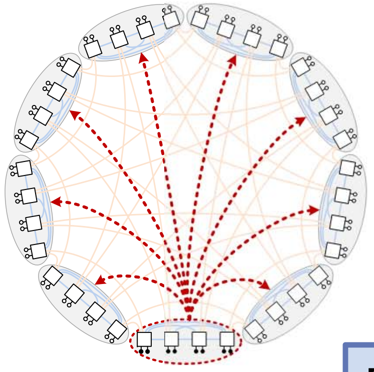




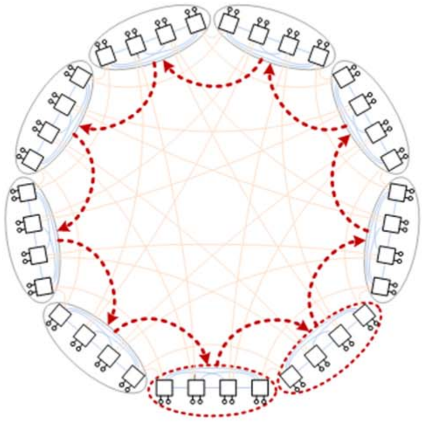
3.2 Restricted Valiant (RVAL)

- Partial conclusions:
  - **Restricted Valiant** in the Dragonfly is highly beneficial for intra-group traffic (Adversarial-local)
  - Very small benefit (no penalty) in other cases (~1%).



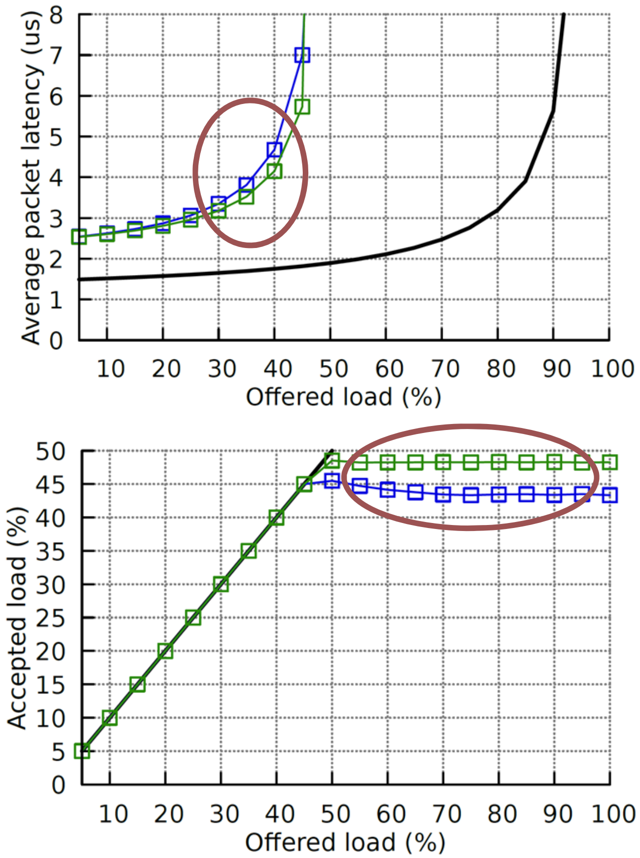


3.3 Valiant with recomputation

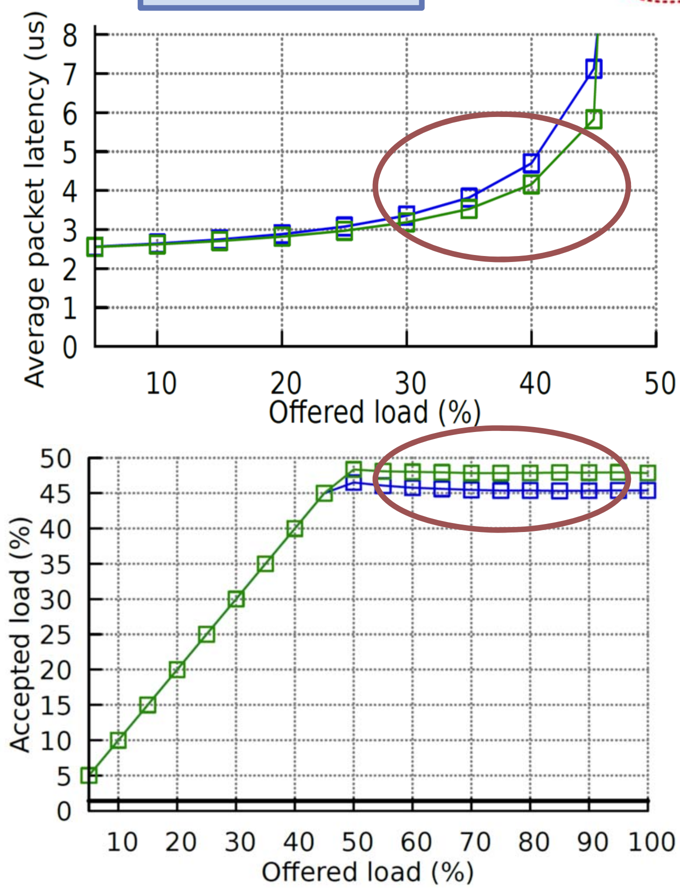


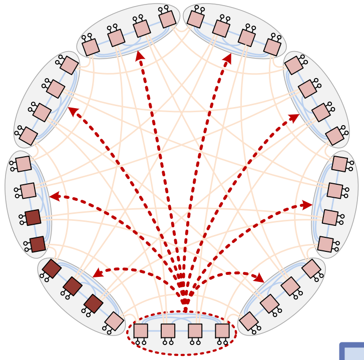
MIN RVAL RVAL-Recomp

Random Uniform

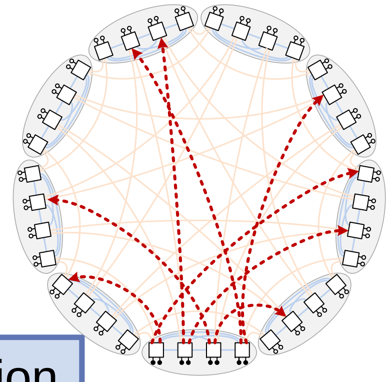


Adversarial



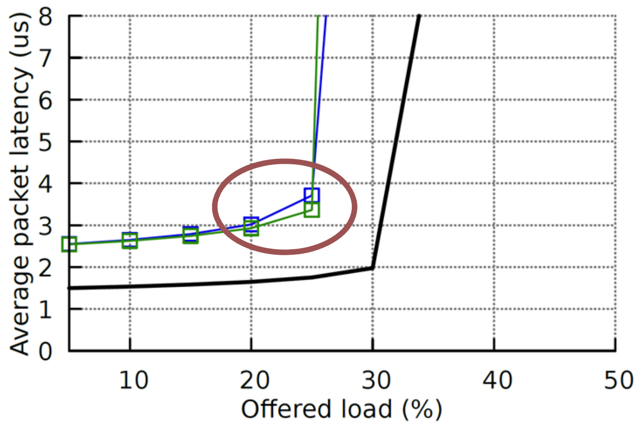


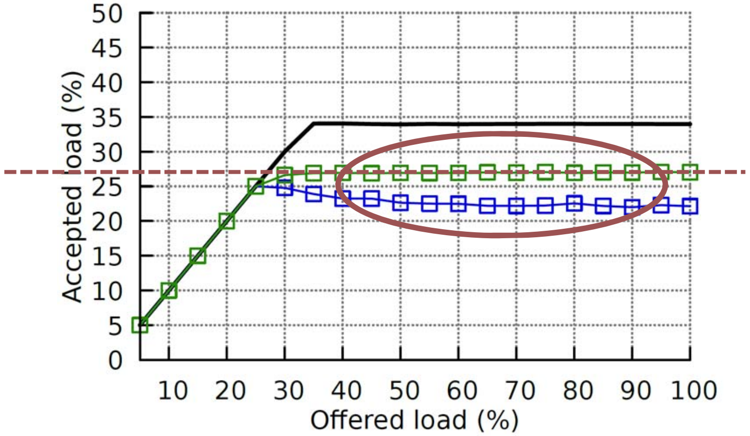
3.3 Valiant with recomputation



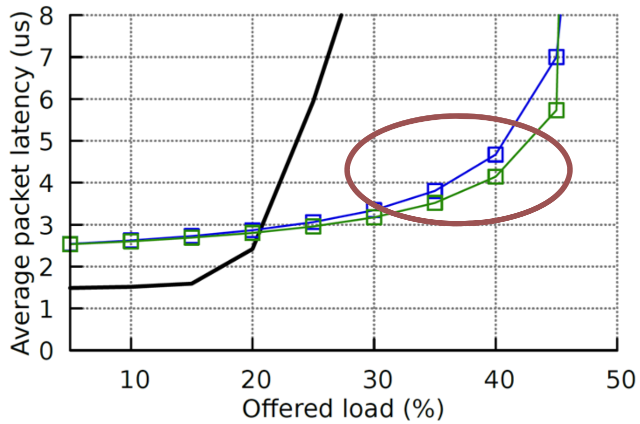
MIN RVAL RVAL-Recomp

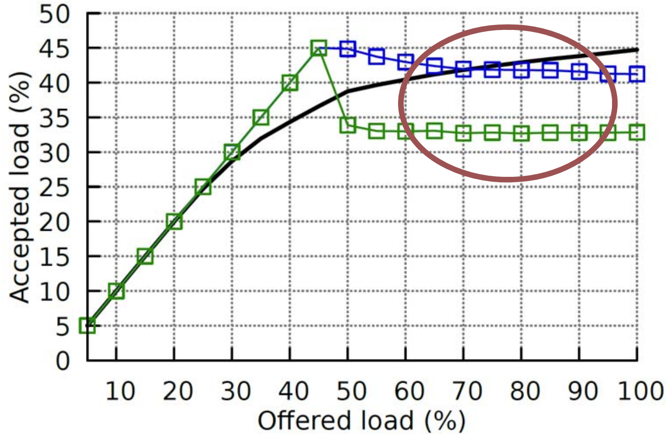
Hot-Region





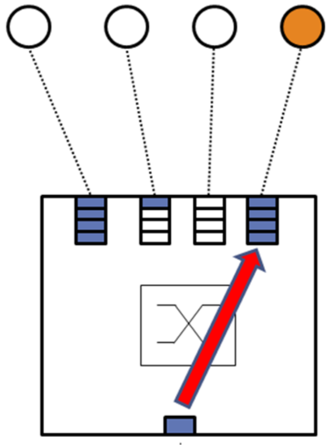
Random permutation





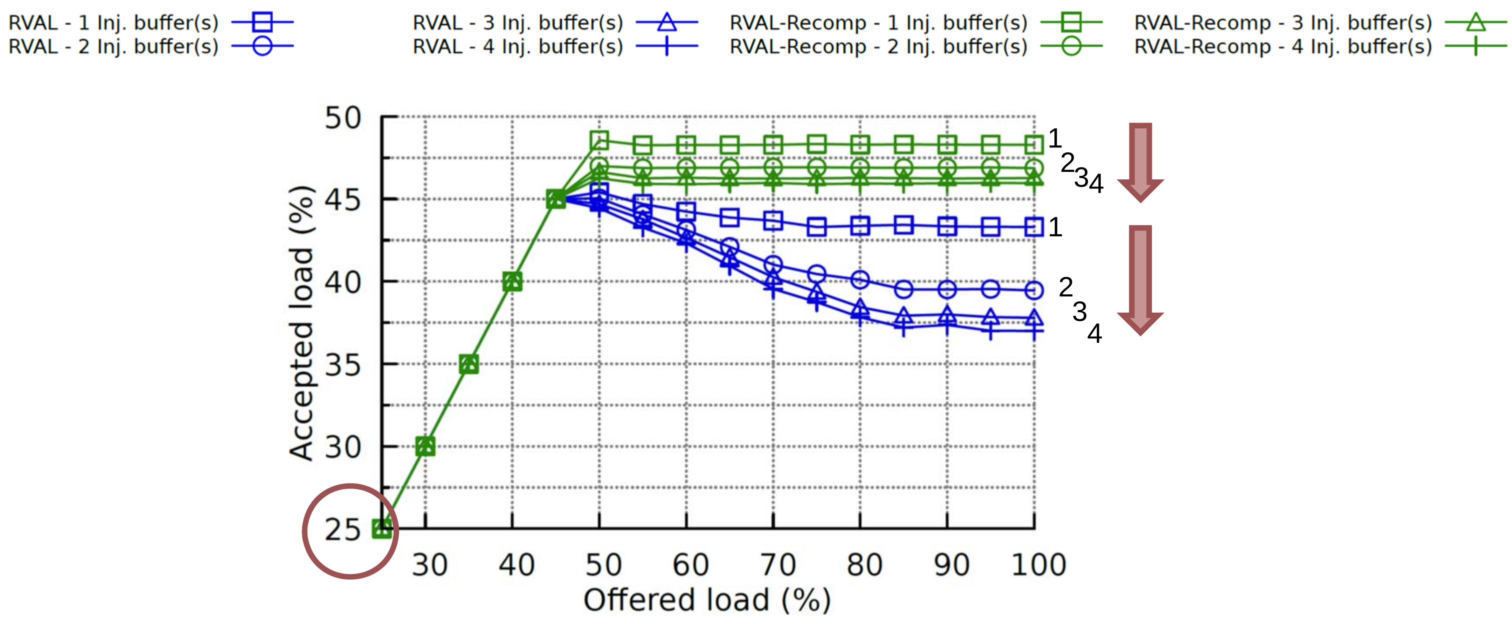
3.3 Valiant with recomputation

- Partial conclusions:
  - Valiant with recomputation improves:
    - Stability of the results (much less oscilations)
    - Latency before saturation
    - Peak throughput
  - The recomputation mechanism is negative for random permutations of traffic in the saturation regimen
    - It increases congestion issues after saturation

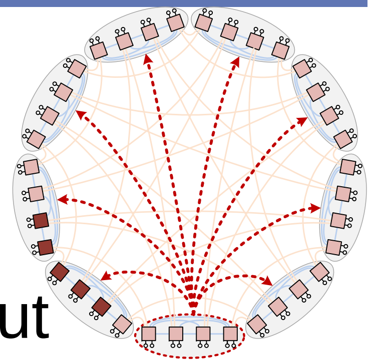


3.4 Number of injection buffers

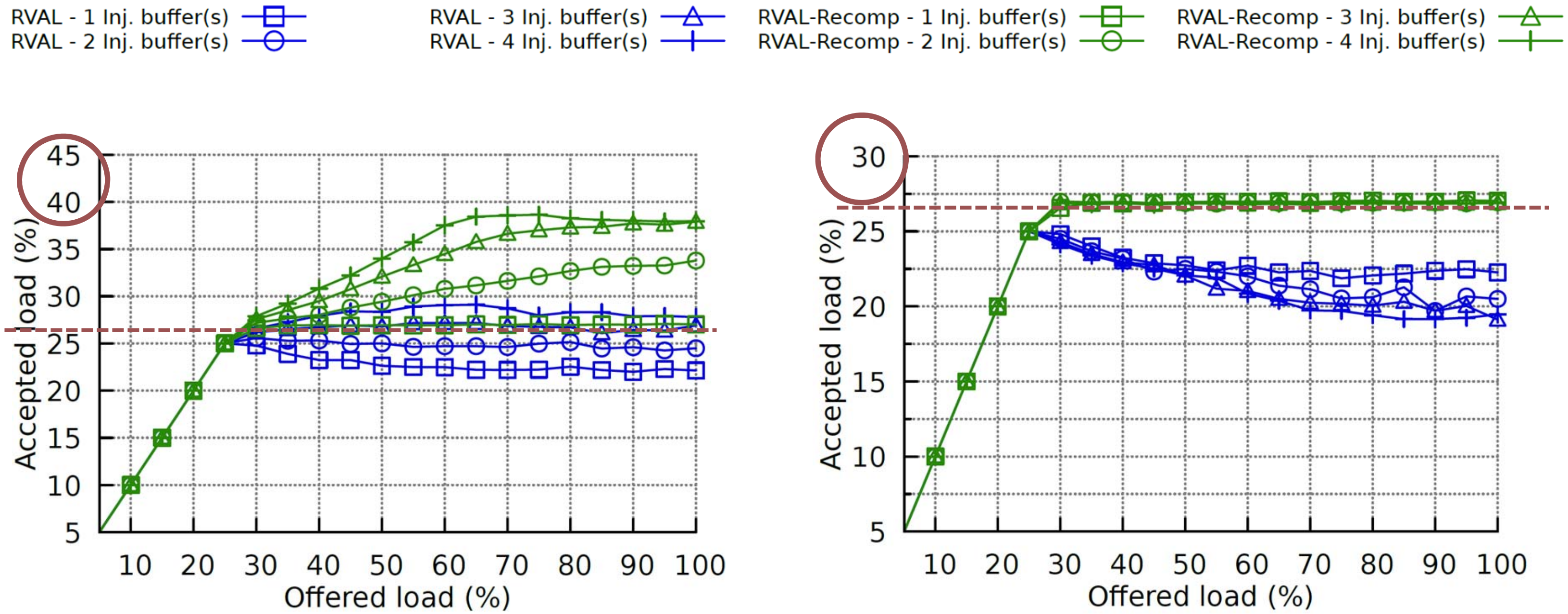
- When multiple buffers are used:
  - No significant difference for latency before saturation
  - Traffic injection after saturation increases with the number of buffers, what increases congestion
  - Typical behavior for 1-4 buffers under UN or ADV traffic:



3.4 Number of injection buffers



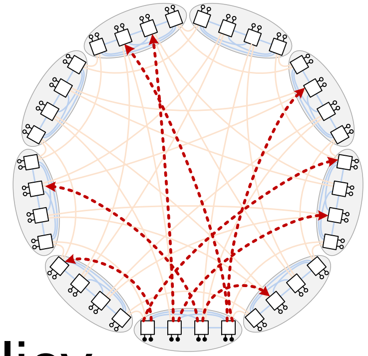
- Hot-región traffic: more buffers increase throughput with a DEST policy.



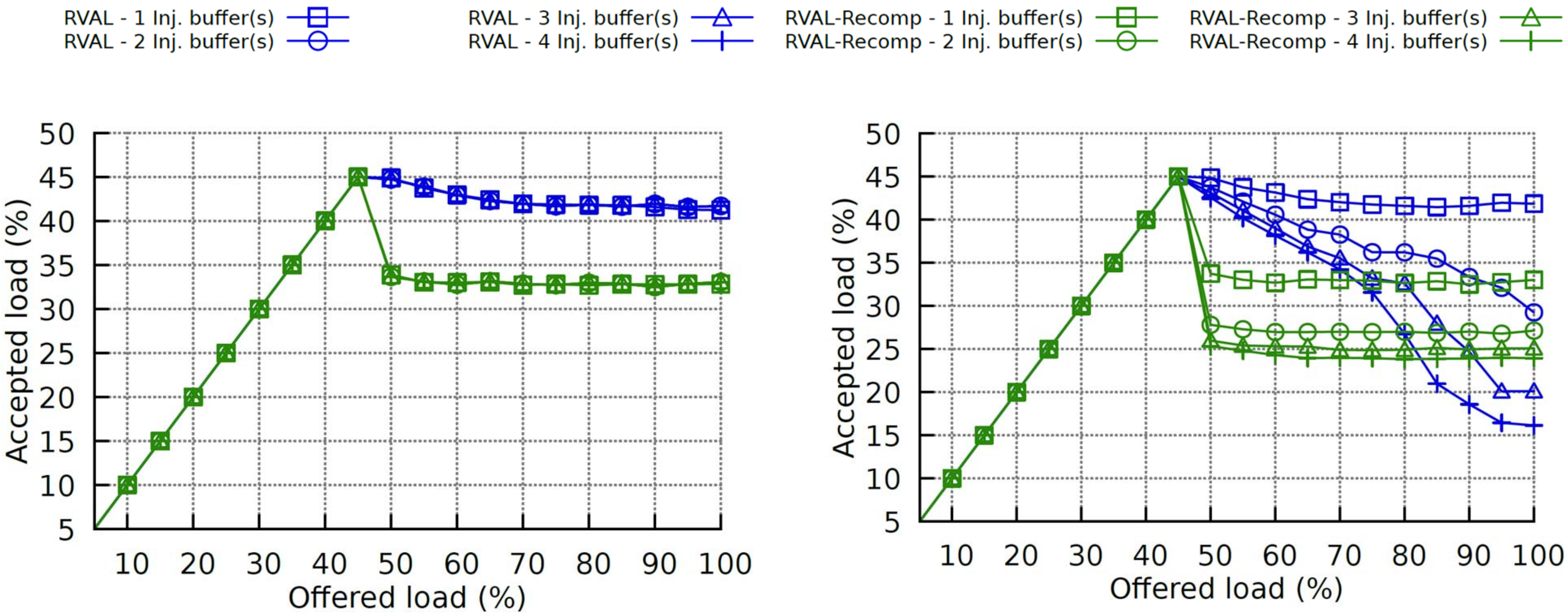
DEST policy: injection buffer selected by destination

RANDOM policy: injection buffer selected randomly, between available

3.4 Number of injection buffers



- **Random-permutation traffic**: more buffers severely increase congestion with a RANDOM policy



DEST policy: injection buffer selected by destination

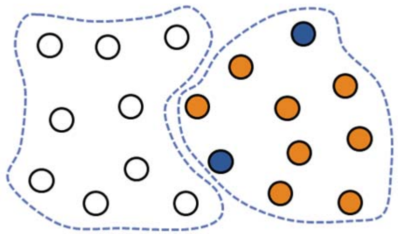
RANDOM policy: injection buffer selected randomly, between available

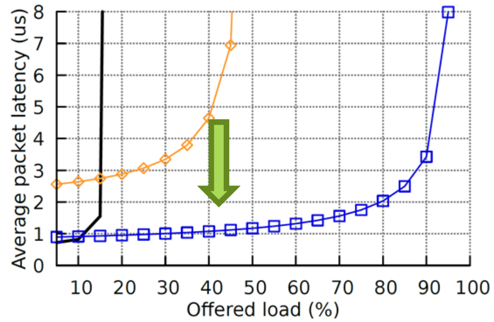
3.4 Number of injection buffers

- Partial conclusions:
  - Increasing the number of injection buffers increases the amount of injected traffic.
    - Increased congestion under UN, ADV and PERM
    - Reduces endpoint-congestion effect under HOT-REGION (traffic with endpoint congestion) with a per-destination buffer selection policy.

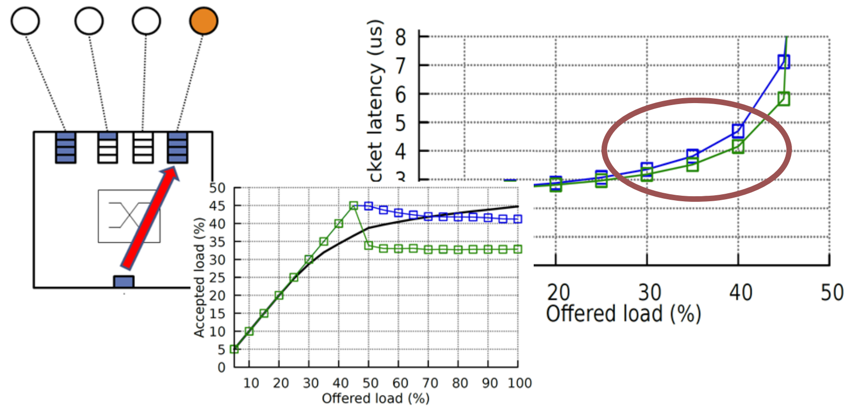
Index

4. Conclusions and future work

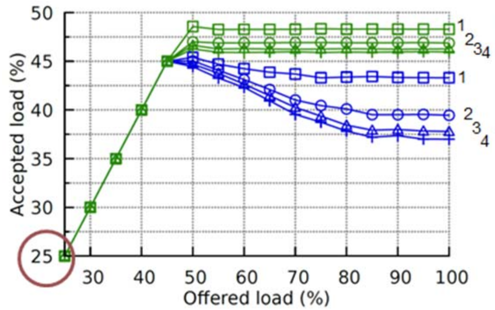


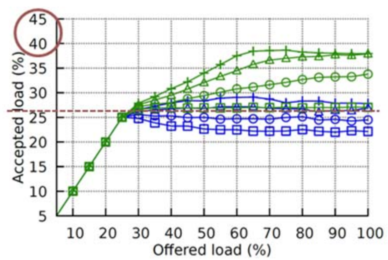


- **Restricted Valiant**
  - The performance improvement is dramatical for traffic internal to a partician (a Dragonfly group).



- **Valiant with recomputation** improves the stability of the results, latency and peak throughput.
  - More throughput also increases congestion

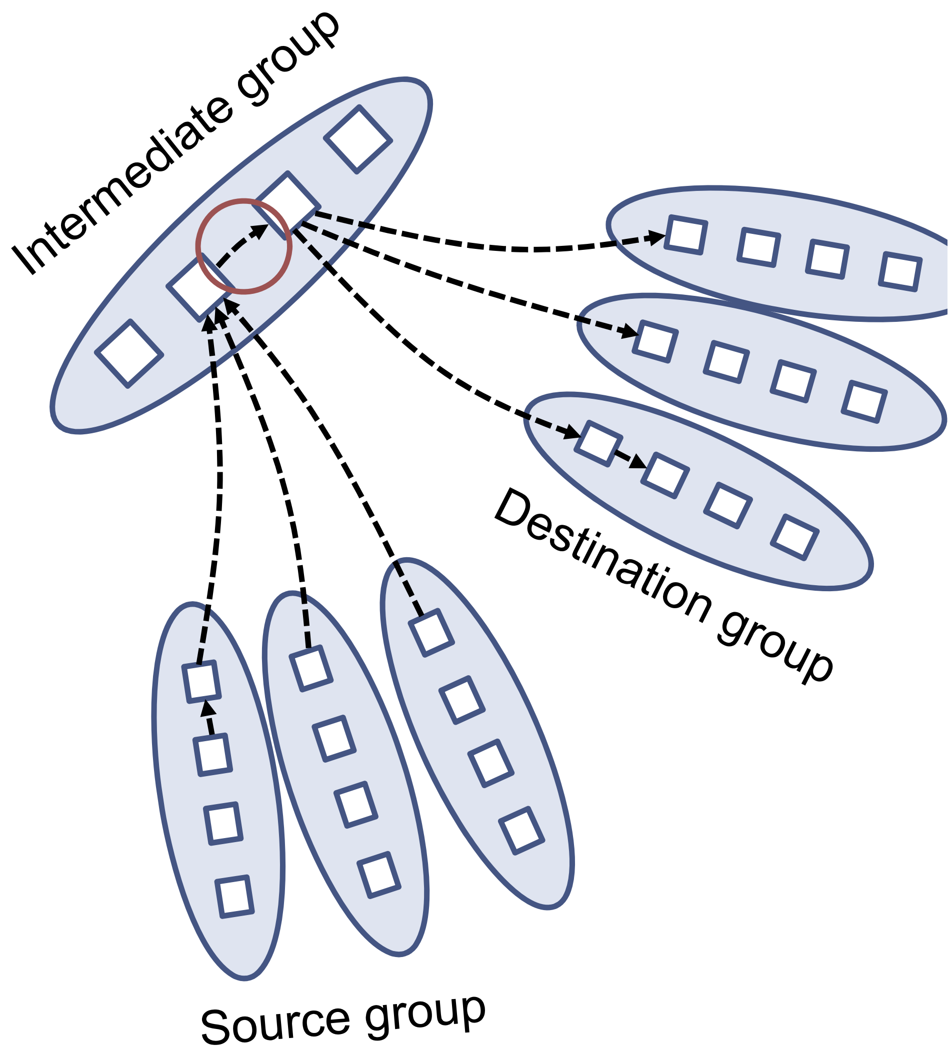




- **Number of virtual channels**:
  - More injection channels increase congestion
  - HoLB reduction is effective in cases of endpoint congestion (Hot-Region traffic)

4. Conclusions and future work

- Our proposal for Restricted Valiant relies on *network partitions.*
  - ***How to specify useful partitions for a given (nontrivial) topology?***
- How to define (proof) when the behavior of Restricted Valiant is "correct"?
  - Example: Restricted Valiant in the Dragonfly proposed by Kim *et al* in [4]
    - Denoted *Valiant-global* in [10]
    - Pathological performance under adversarial traffic identified in [11]
  - L. Valiant studies the *consumption time of a worst-case permutation.*
    - Should we use this analysis?
    - Is this equivalent to minimum throughput at saturation (per router)?



[4] Kim, Dally, Scott, Abts. *Technology-Driven, Highly-Scalable Dragonfly Topology*. ISCA '08
[10] J. Won, G. Kim, J. Kim, T. Jiang, M. Parker and S. Scott, "Overcoming far-end congestion in large-scale networks," HPCA'*15*
[11] M. García *et al*: "On-the-fly adaptive routing in high-radix hierarchical networks," IPDPS'12

4. Conclusions and future work

- **Other issues** we are exploring:
  - How does Restricted Valiant and Valiant with Recomputation behave when using **adaptive routing**?
  - How should we implement them in an interconnect that implements **table-based routing**?

ANALYSIS AND IMPROVEMENT OF VALIANT ROUTING IN LOW-DIAMETER NETWORKS

Mariano Benito
Pablo Fuentes
Enrique Vallejo
Ramón Beivide

UNIVERSIDAD DE CANTABRIA

With support from:

MONT BLANC

4th IEEE International Workshop of High-Perfomance Interconnection Networks in the Exascale and Big-Data Era (HiPINEB)
Vienna, Austria, 24-Feb, 2018.