vFtree – A fat-tree routing algorithm using virtual lanes to alleviate congestion

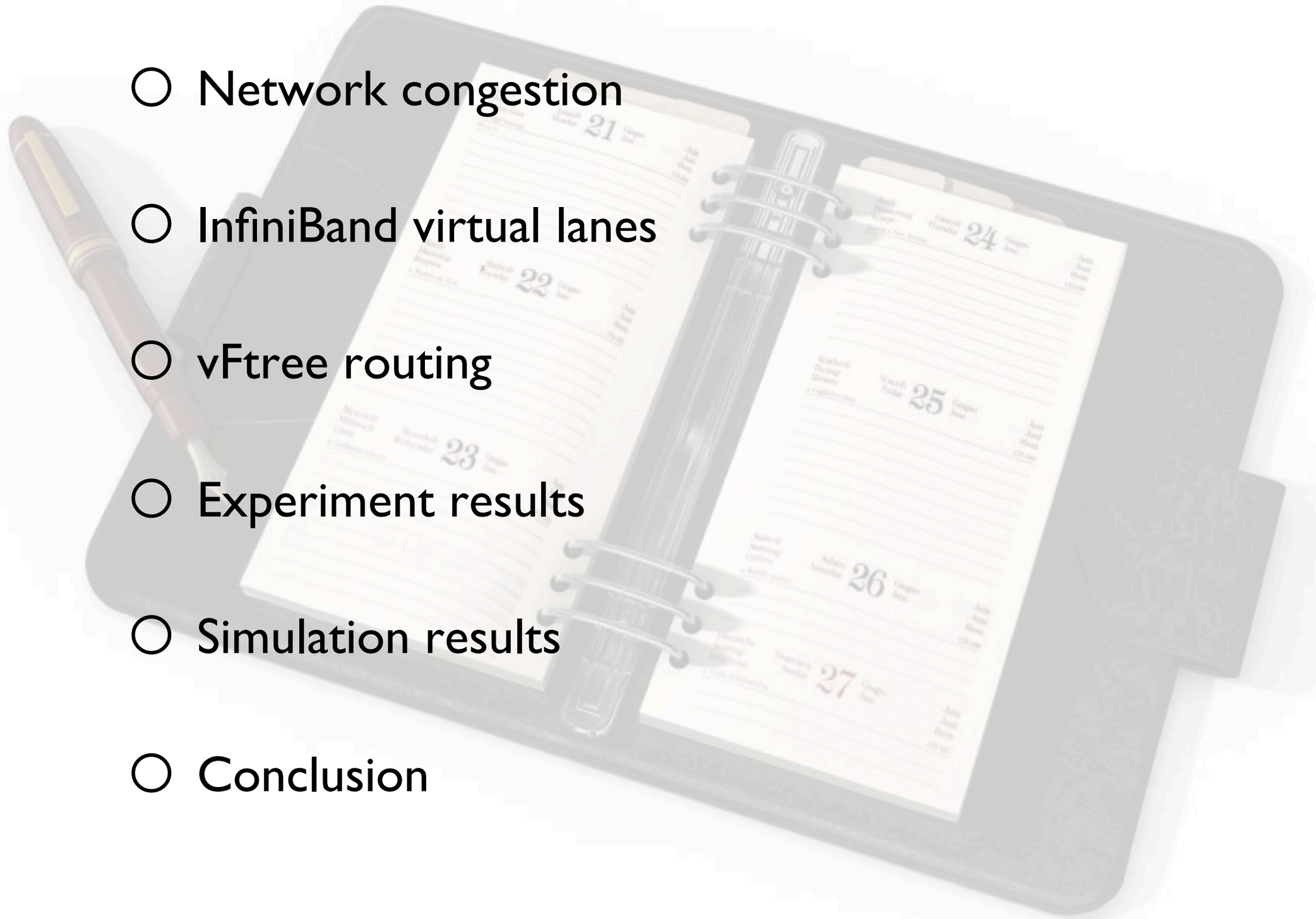
Sven-Arne Reinemo
svenar@simula.no
Simula Research Laboratory

HPC Advisory Council Switzerland Workshop
March 23, 2011

ACKNOWLEDGEMENTS

- Wei Lin Guay
- Bartosz Bogdanski
- Tor Skeie
- Professor Olav Lysne

AGENDA



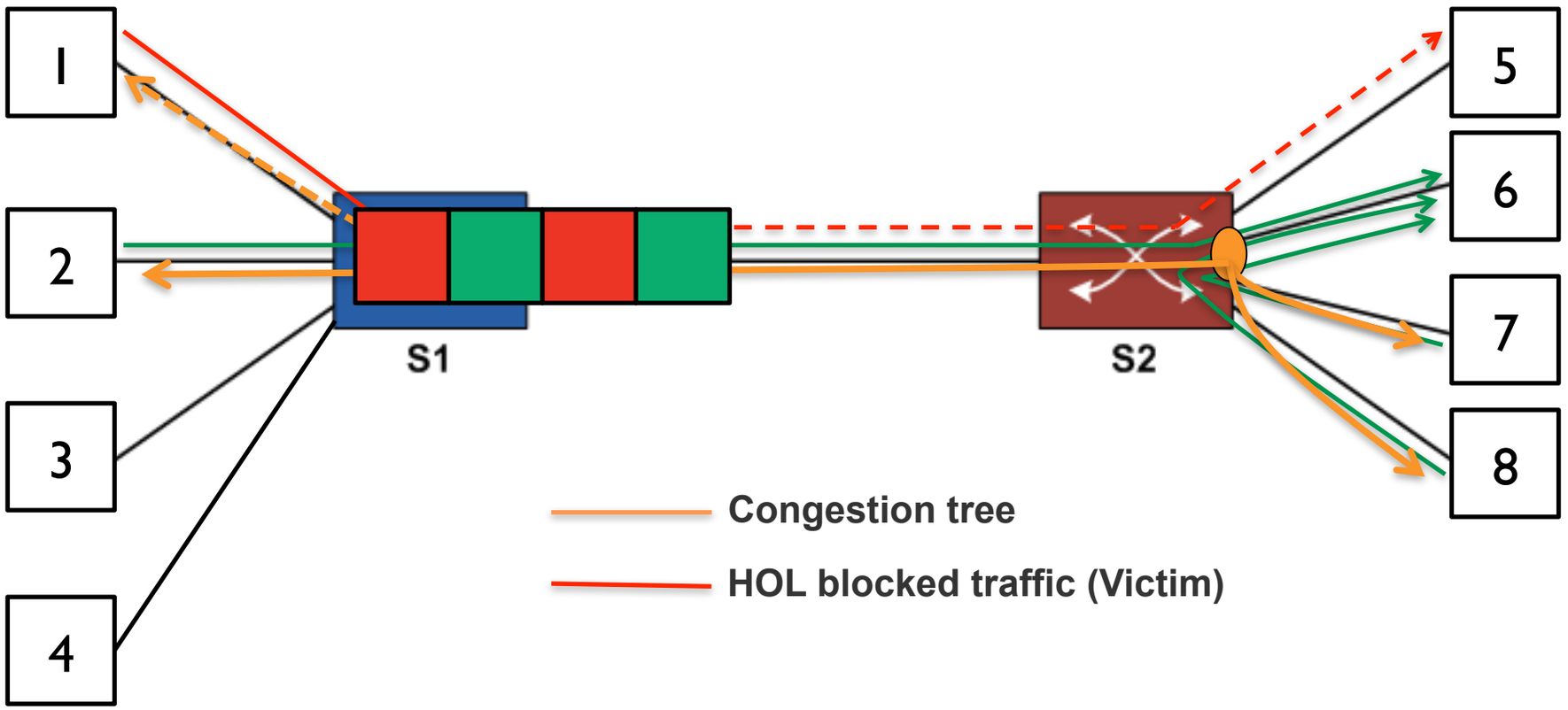
- Network congestion
- InfiniBand virtual lanes
- vFtree routing
- Experiment results
- Simulation results
- Conclusion

CONGESTION

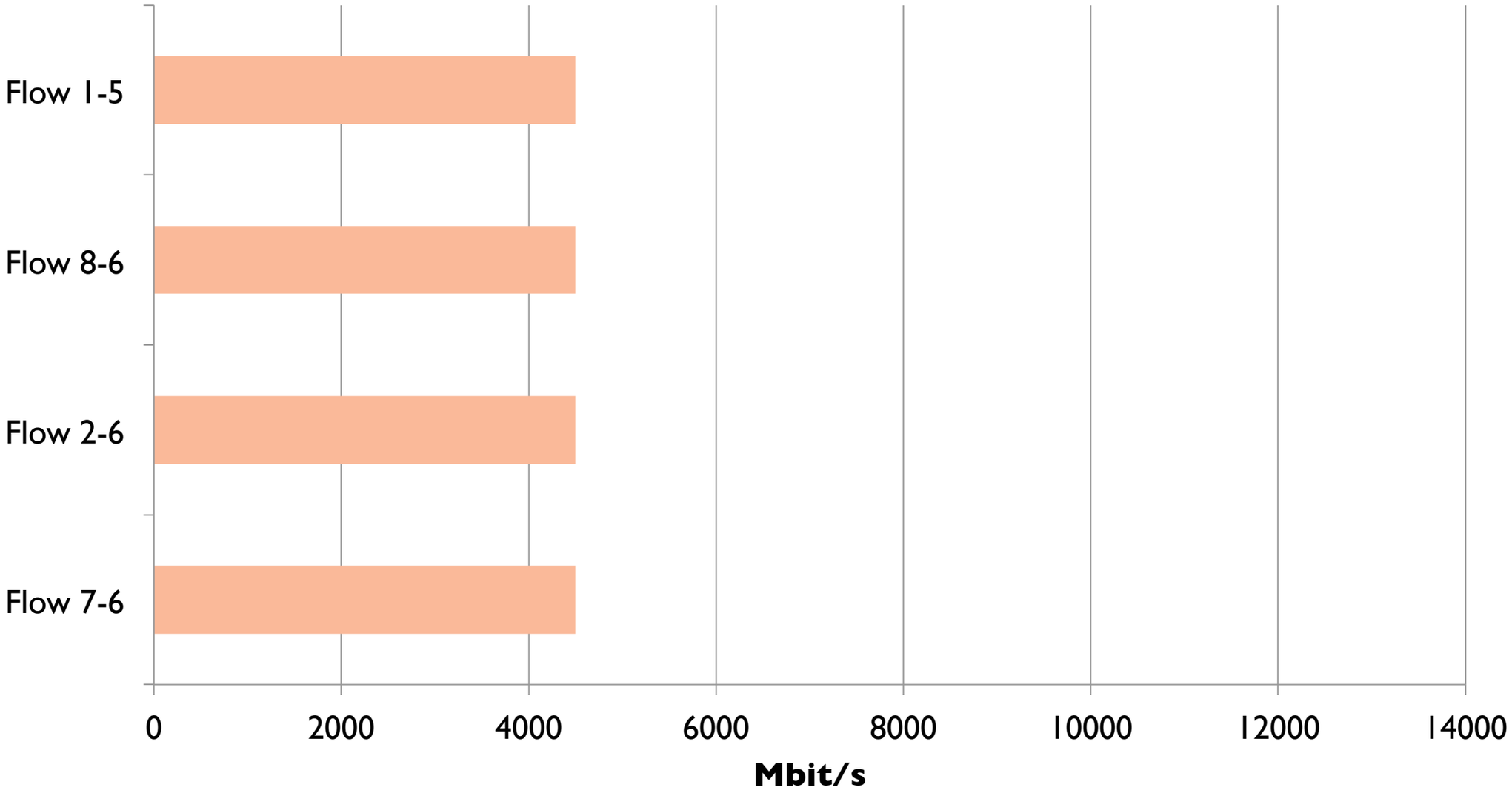
- Shared network resources can lead to network congestion due to hot-spots.
- In loss-less interconnects such as InfiniBand congestion leads to head-of-line (HOL) blocking.
- HOL blocking can lead to reduced performance for innocent flows, i.e. flows not accessing hot-spots.
- The InfiniBand congestion control mechanism[1] can solve this in many situations, but is not always available.

[1]E.G.Gran et al. First Experiences with Congestion Control in InfiniBand Hardware. IPDPS 2010.

CONGESTION



CONGESTION



HOL blocking reduces the utilisation of the link between S1 and S2 to 1/3 of its bandwidth reducing performance for flow 1-5.

VIRTUAL LANES

- Virtual Lanes (VLs) are logical channels on the same physical link, but with separate buffering and flow-control resources.
- The most obvious use of VLs are for service differentiation.
- Our proposed vFtree routing uses VLs for enhanced routing and improved network performance.

VIRTUAL LANES

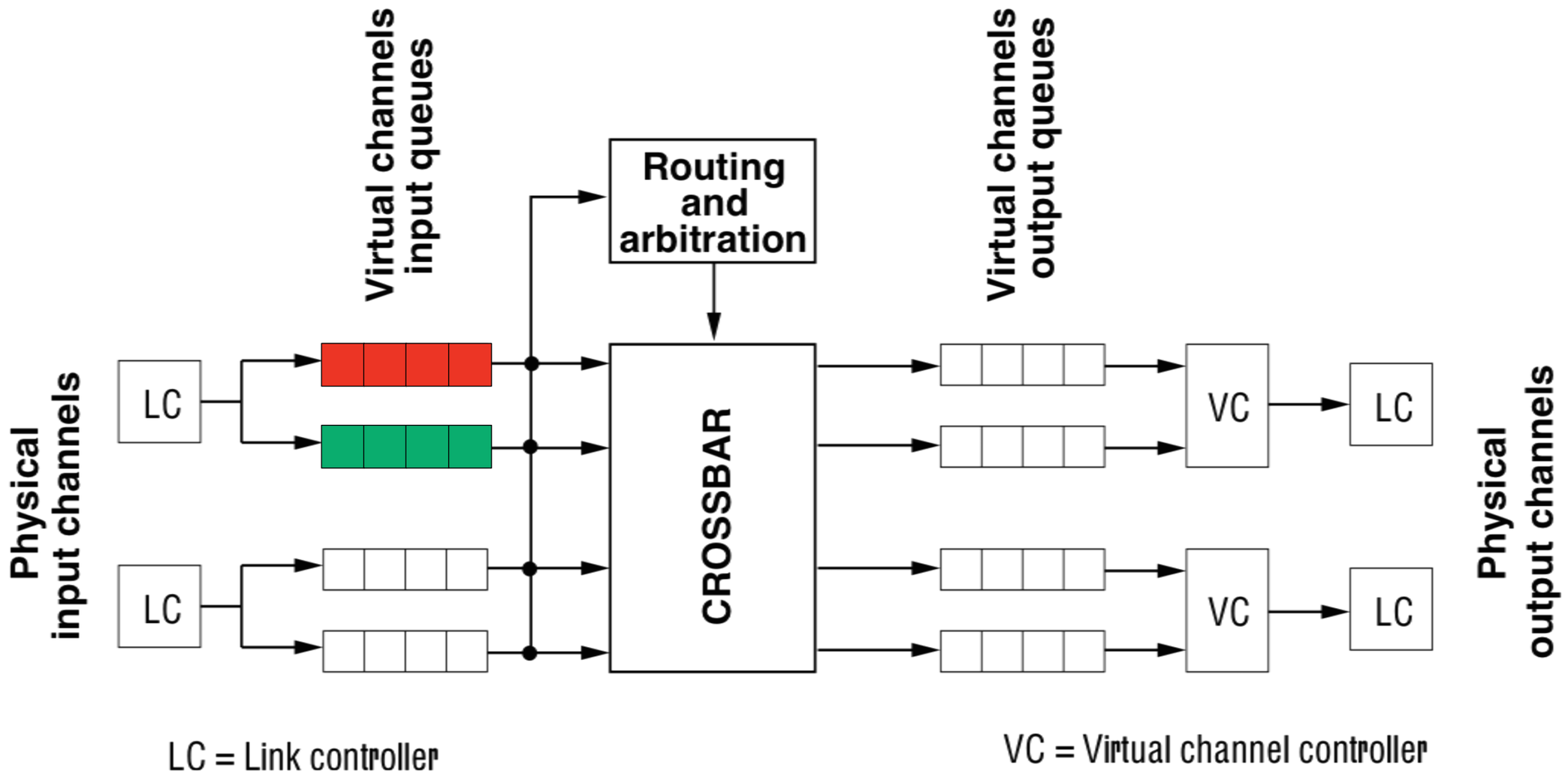
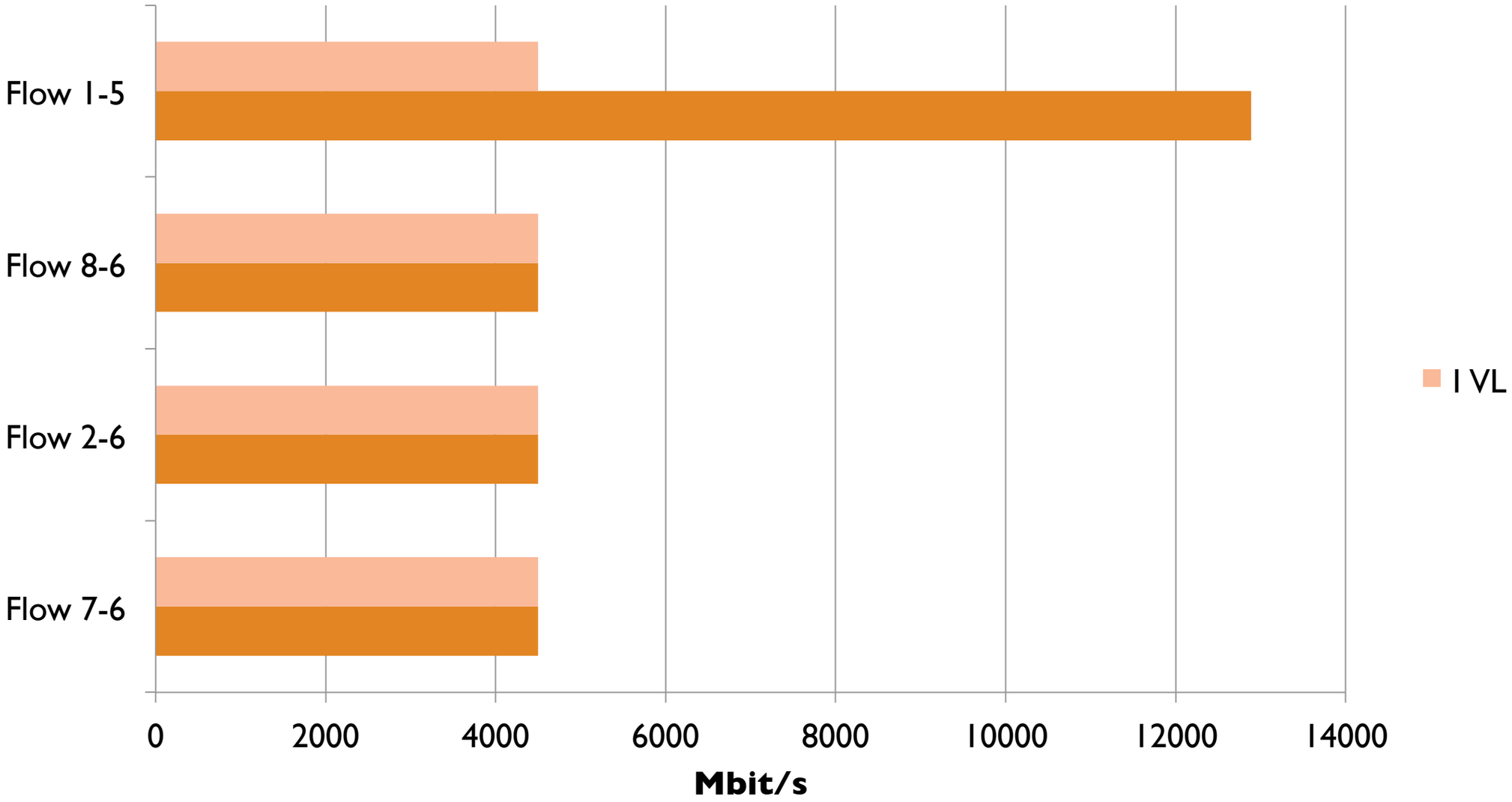


Figure from the book "Interconnection networks: an engineering approach".
By José Duato, Sudhakar Yalamanchili, Lionel M. Ni.

CONGESTION



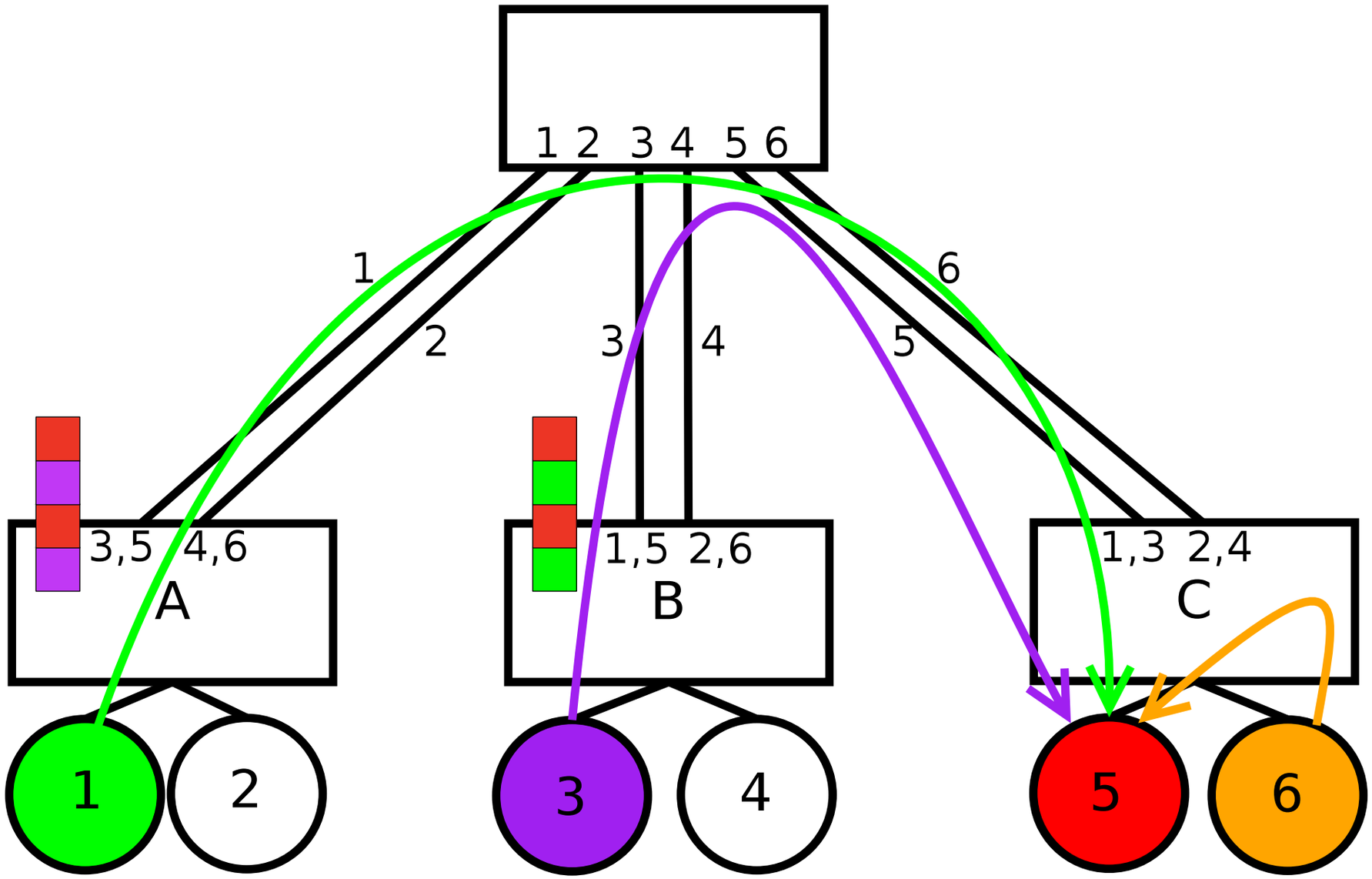
Putting flow 1-5 on a separate VL removes the negative effect of HOL blocking on this flow and fully utilises the link between S1 and S2.

VFTREE ROUTING

- It is a well known fact that using multiple VLs can improve network performance, but this has seen limited use in existing installations.
- Most InfiniBand installations are based on equipment that supports 8 VLs, but use only 1 VL. The remaining VLs are left idle.
- With vFtree[2] routing we make use of the idle VLs to improve performance and alleviate congestion.
- The basic idea is that the routing algorithm distributes source/destination pairs across the available VLs.

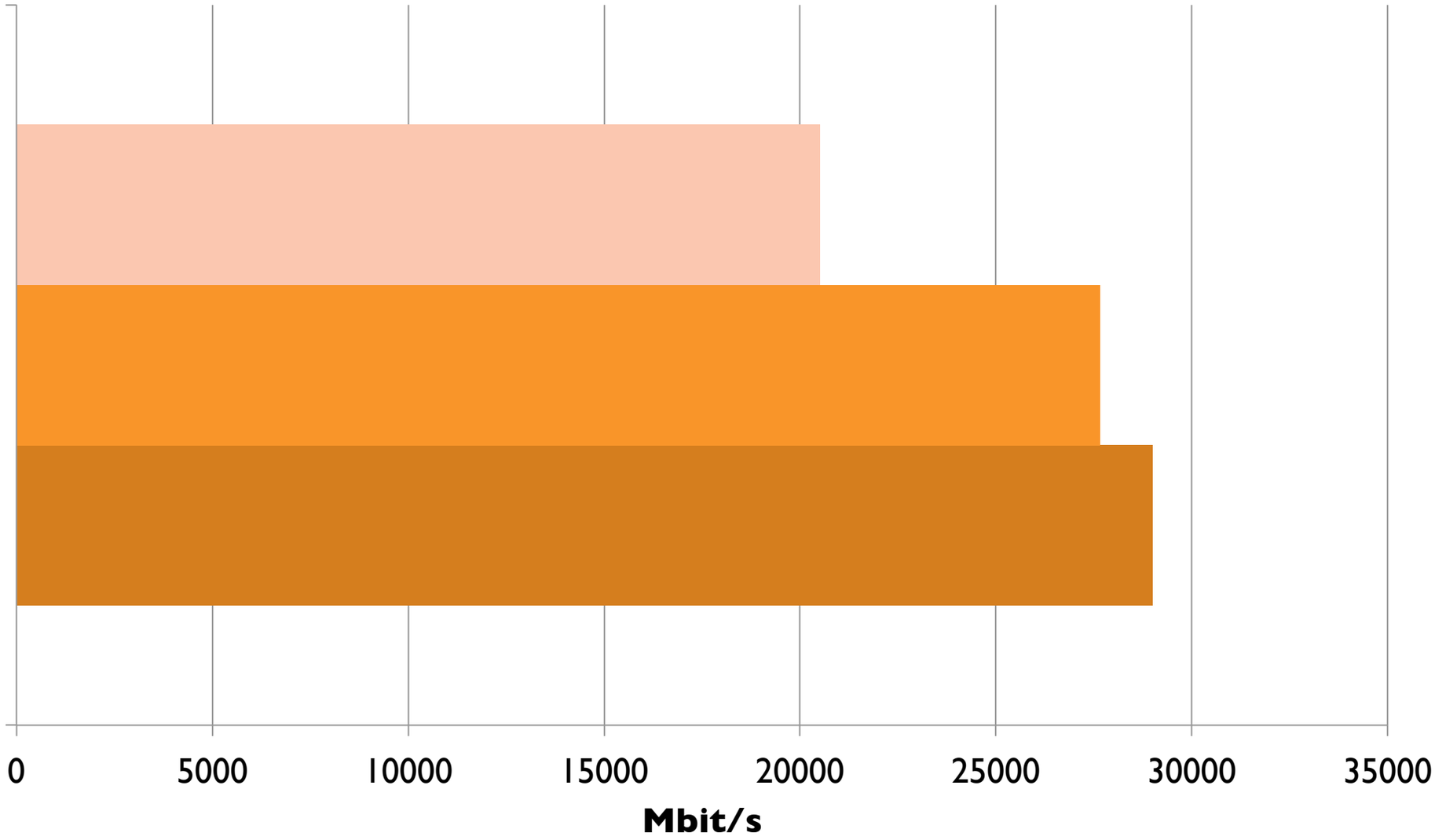
[2]W.L. Guay et al. vFtree – A Fat-tree Routing Algorithm using Virtual Lanes to Alleviate Congestion. To appear at IPDPS 2011.

MEASUREMENT SCENARIO



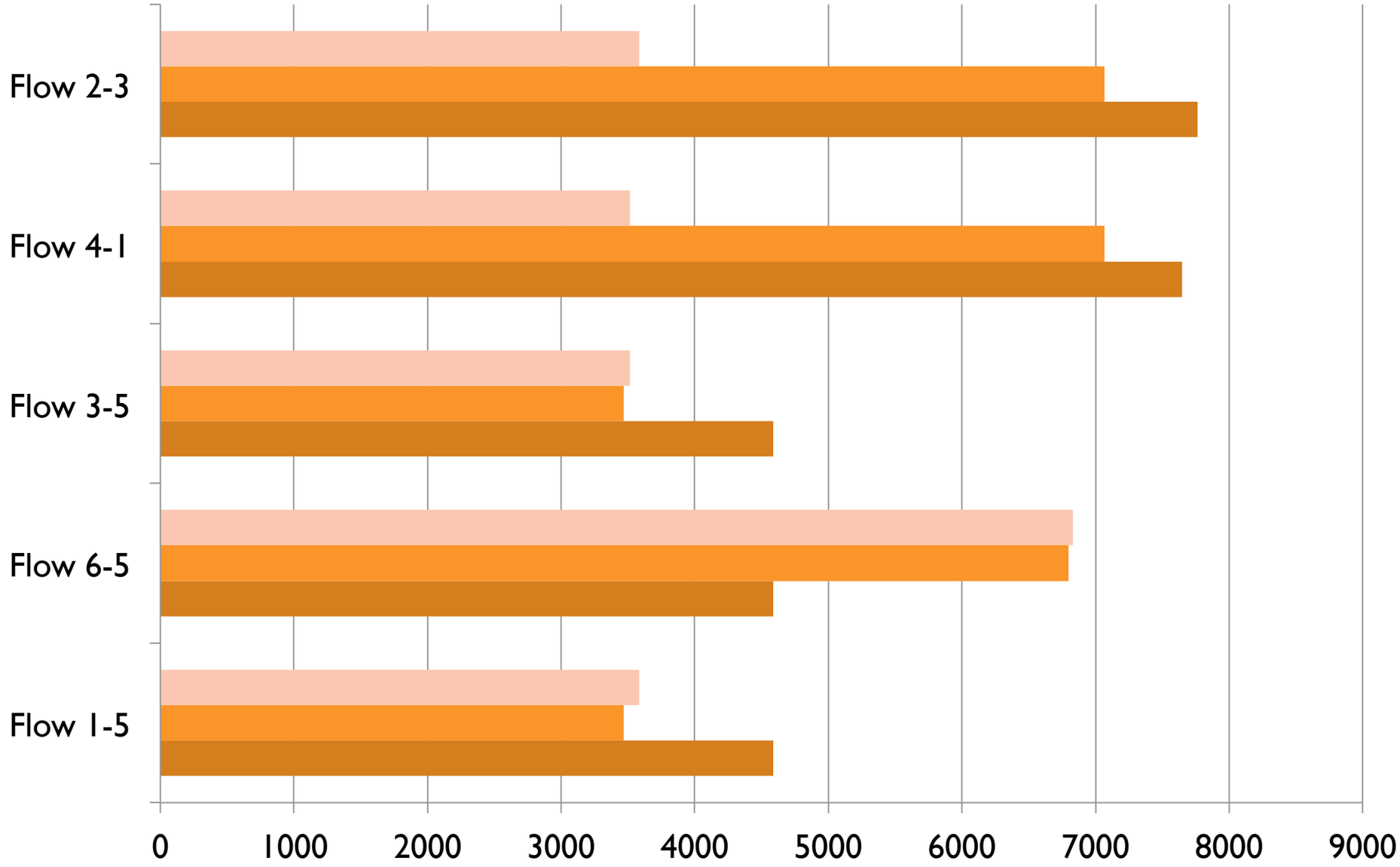
Node 1,3, and 6 are accessing the hot receiver 5. Node 2 and 4 are accessing 3 and 1, respectively. These flows are the victims of congestion.

MEASUREMENT RESULTS



The total network throughput increases as we make use of more virtual lanes, because the negative effect of HOL blocking is reduced.

MEASUREMENT RESULTS



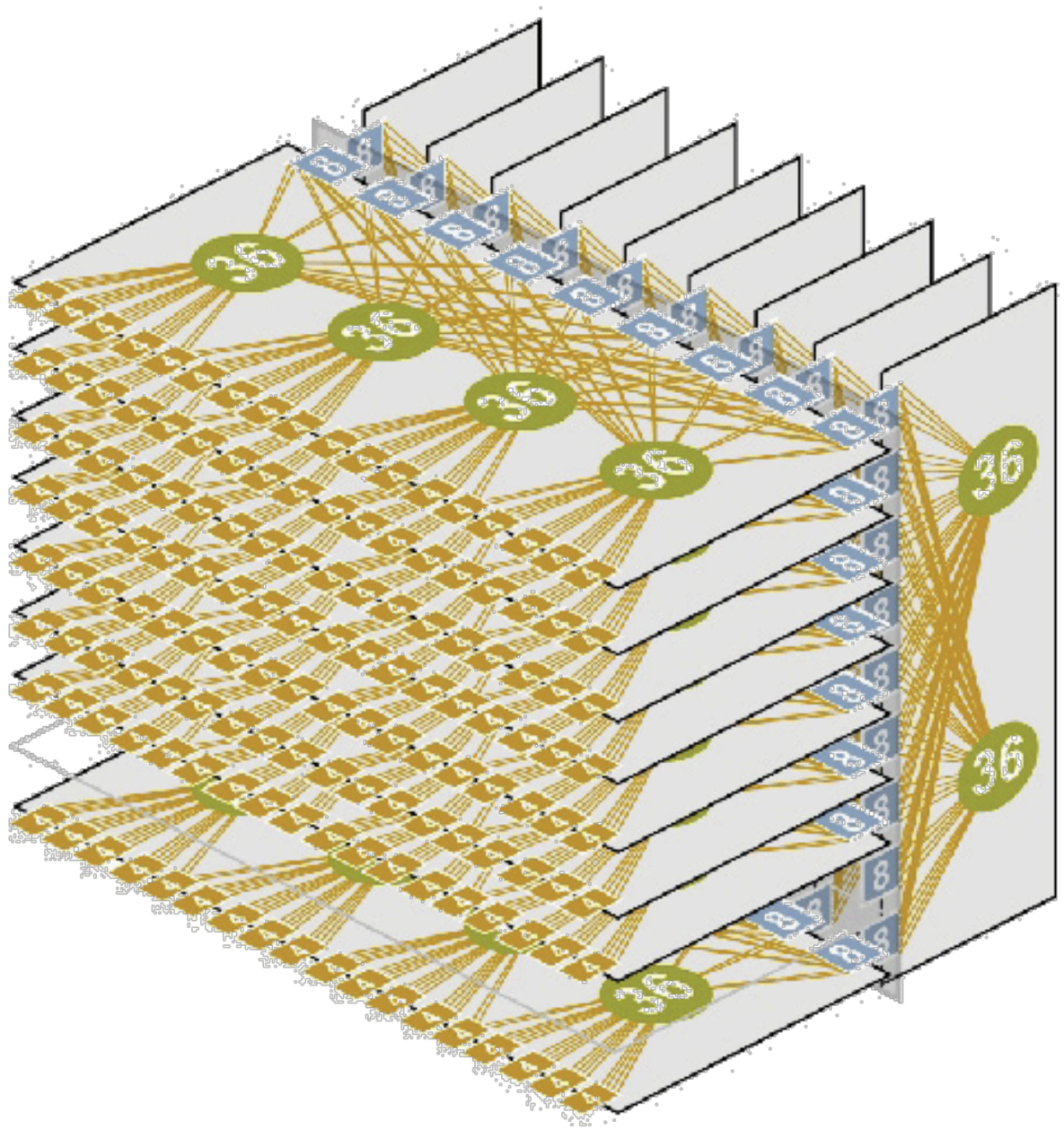
The per flow throughput shows that the victims see a 100% improvement.

HPCC RESULTS

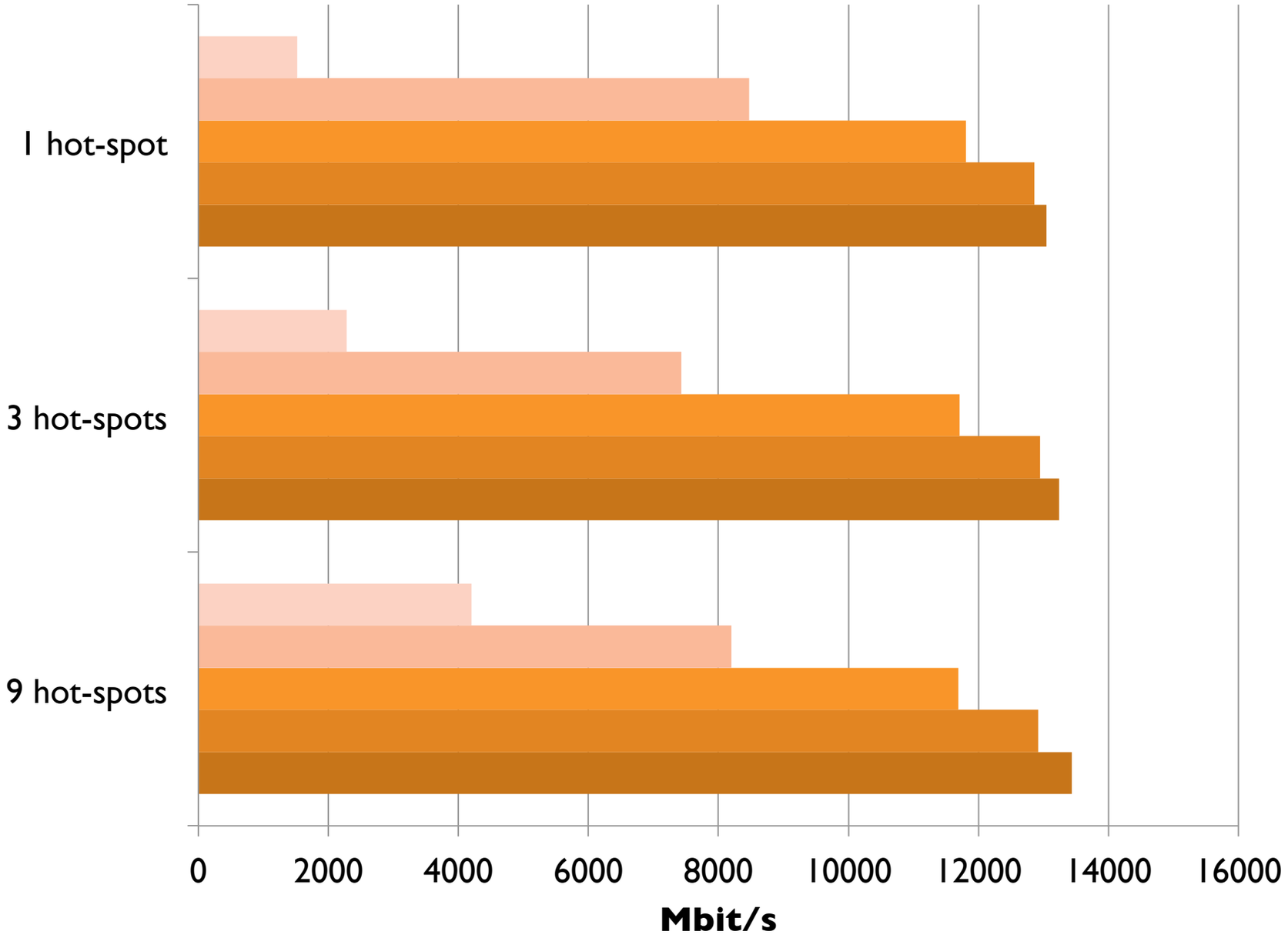
HPCC test	Ftree (1 VL)	vFtree (3 VLs)	Improvement in %
Max. ping pong latency (ms)	0.002116	0.002116	0.0
Avg. ping pong latency (ms)	0.022898	0.013477	41.14
Min. ping pong latency (ms)	0.050500	0.043005	14.84
Naturally ordered ring latency (ms)	0.021791	0.014591	33.04
Randomly ordered ring latency (ms)	0.024262	0.015826	**34.77**
Max. ping pong bandwidth (MB/s)	1593.127	1594.338	0.07
Avg. ping pong bandwidth (MB/s)	573.993	830.909	44.75
Min. ping pong bandwidth (MB/s)	94.868	345.993	264.71
Naturally ordered ring bandwidth (MB/s)	388.969246	454.236253	16.78
Randomly ordered ring bandwidth (MB/s)	331.847978	438.604531	**32.17**

SIMULATION SCENARIO

- 648-port 2-stage fat-tree.
- Largest 2-stage fat-tree using 36-port switches.
- Matches what several switch vendors provides today.
- Traffic pattern: 5% goes to predefined hot-spots, 95% goes to a random destination.



SIMULATION RESULTS



Average node throughput increases with the number of active VLs.

CONCLUSION

- vFtree is a simple and inexpensive method for reducing congestion in fat-tree networks.
- vFtree can be applied to existing InfiniBand networks as demonstrated by our OpenSM prototype.
- vFtree shows performance improvements from 38% to 757% depending on network size and traffic pattern.
- Future work includes research on dynamic allocation of virtual lanes during operation.

QUESTIONS?