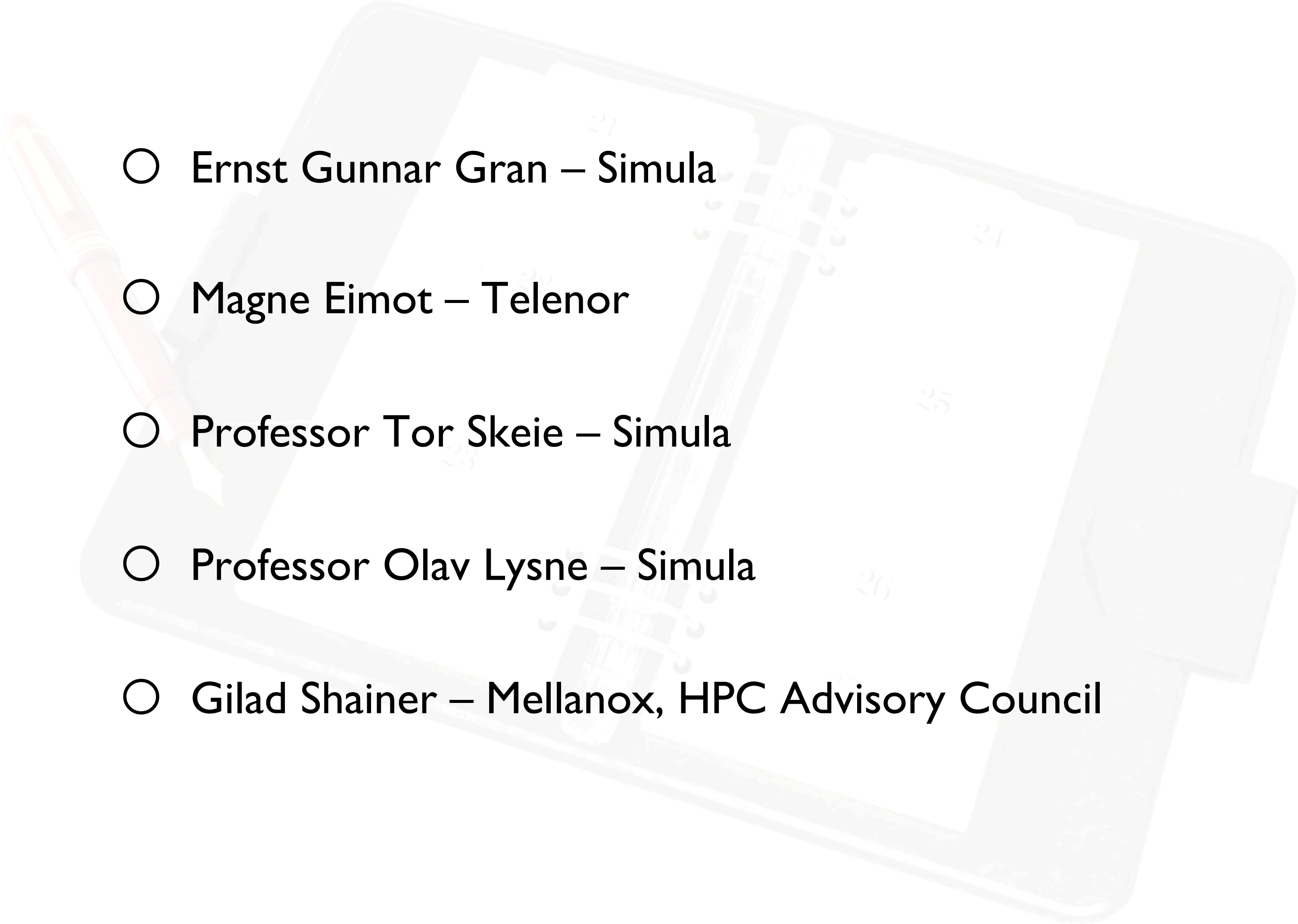
Congestion Control in InfiniBand

OpenFabrics Monterey Workshop 2011

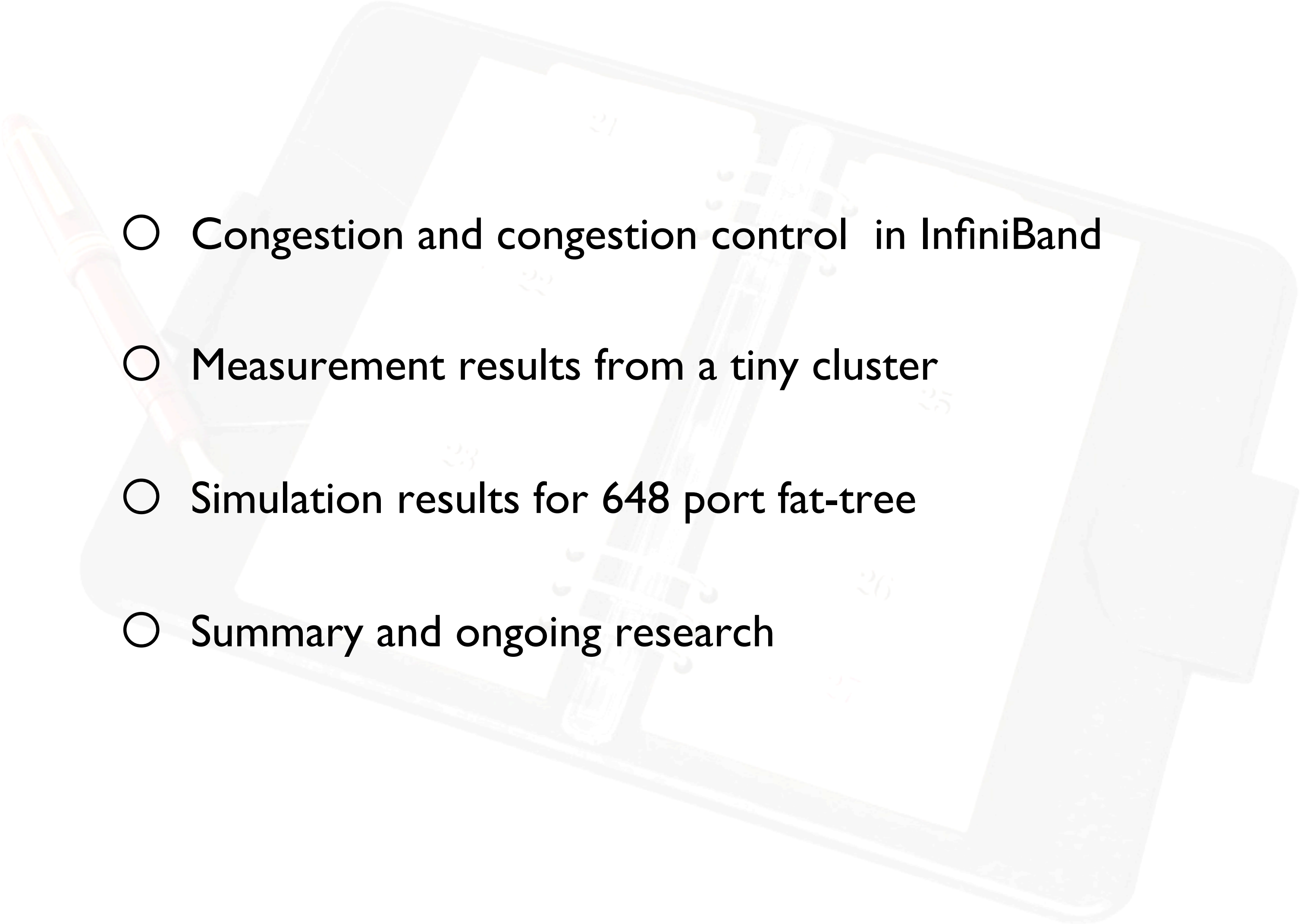
Sven-Arne Reinemo
svenar@simula.no
Simula Research Laboratory

Acknowledgements



- Ernst Gunnar Gran – Simula
- Magne Eimot – Telenor
- Professor Tor Skeie – Simula
- Professor Olav Lysne – Simula
- Gilad Shainer – Mellanox, HPC Advisory Council

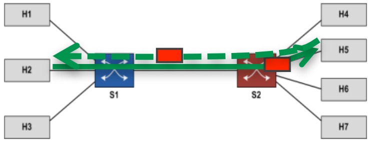
Presentation Outline



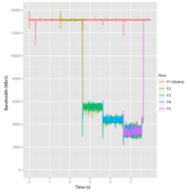
- Congestion and congestion control in InfiniBand
- Measurement results from a tiny cluster
- Simulation results for 648 port fat-tree
- Summary and ongoing research

Presentation Outline

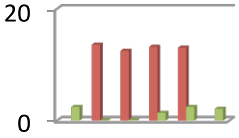
CC Collaboration history



Congestion and Congestion Control in InfiniBand



Experiment results (IPDPS)



Simulation results (M9)

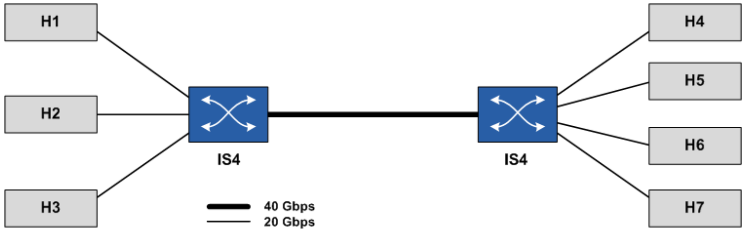
Future

The IB CC connection between Simula and Mellanox

- Fall 2008, Sun Microsystems forwards a request from Mellanox (Gilad Shainer) to Simula about a possible collaboration on IB CC research/proof of concept.

- Early 2009 we received two IS4 switches from Mellanox/HPC Advisory Council, and later CC capable firmware and CC management tools.

- Summer 2009 we started extensive testing of IB CC in our IS4 based test bed at Simula

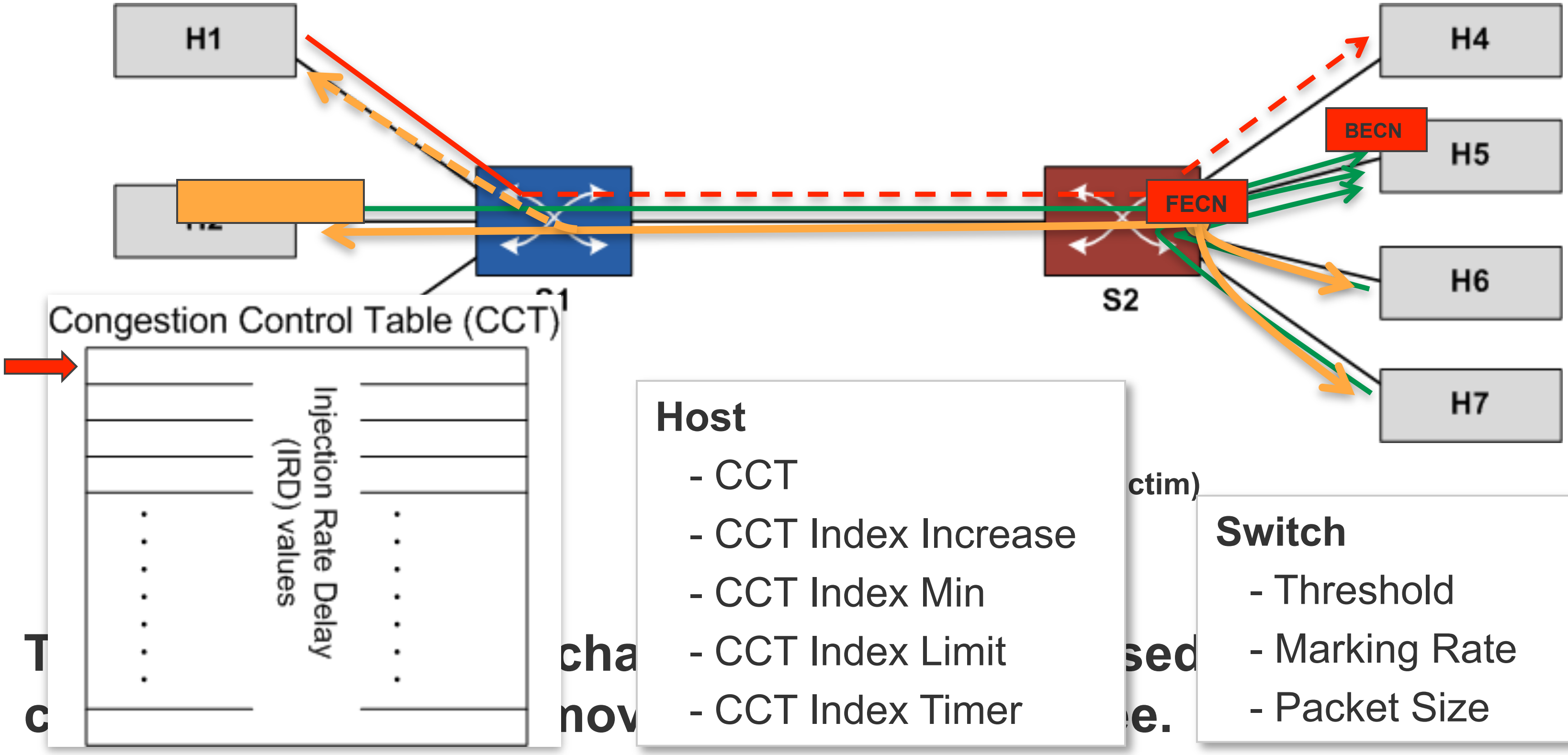


- Preliminary results presented at SC 2009, Portland.

- Results from the experiments were submitted to IPDPS 2010 during fall 2009, and later accepted.

- Cited by e.g. HPC Advisory Council (Workshop in Switzerland, 2010) and Mellanox (OpenFabrics 2010 Sonoma Workshop).

Shared network resources could lead to network congestion and head-of-line (HOL) blocking.


H1
H2
S1
S2
H4
H5
H6
H7
BECN
FECN
Congestion Control Table (CCT)
Injection Rate Delay (IRD) values

Host

- CCT
- CCT Index Increase
- CCT Index Min
- CCT Index Limit
- CCT Index Timer

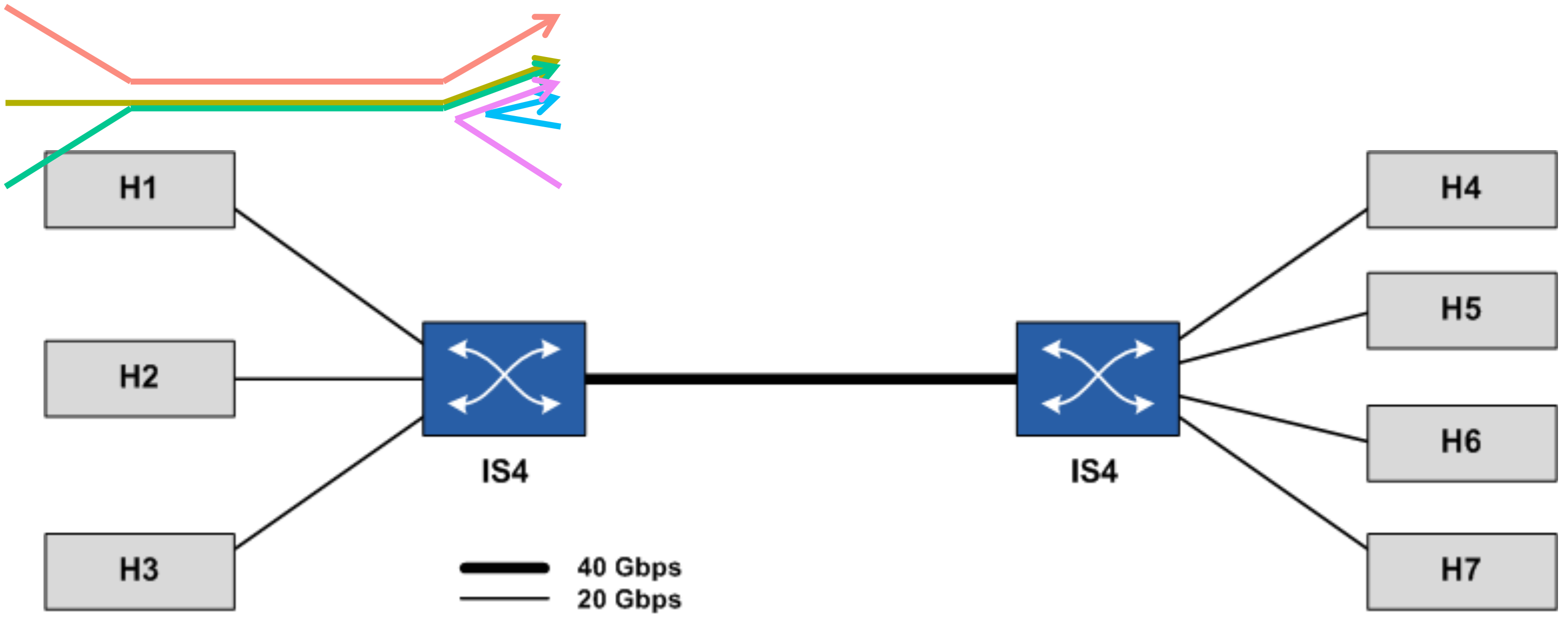
Switch

- Threshold
- Marking Rate
- Packet Size

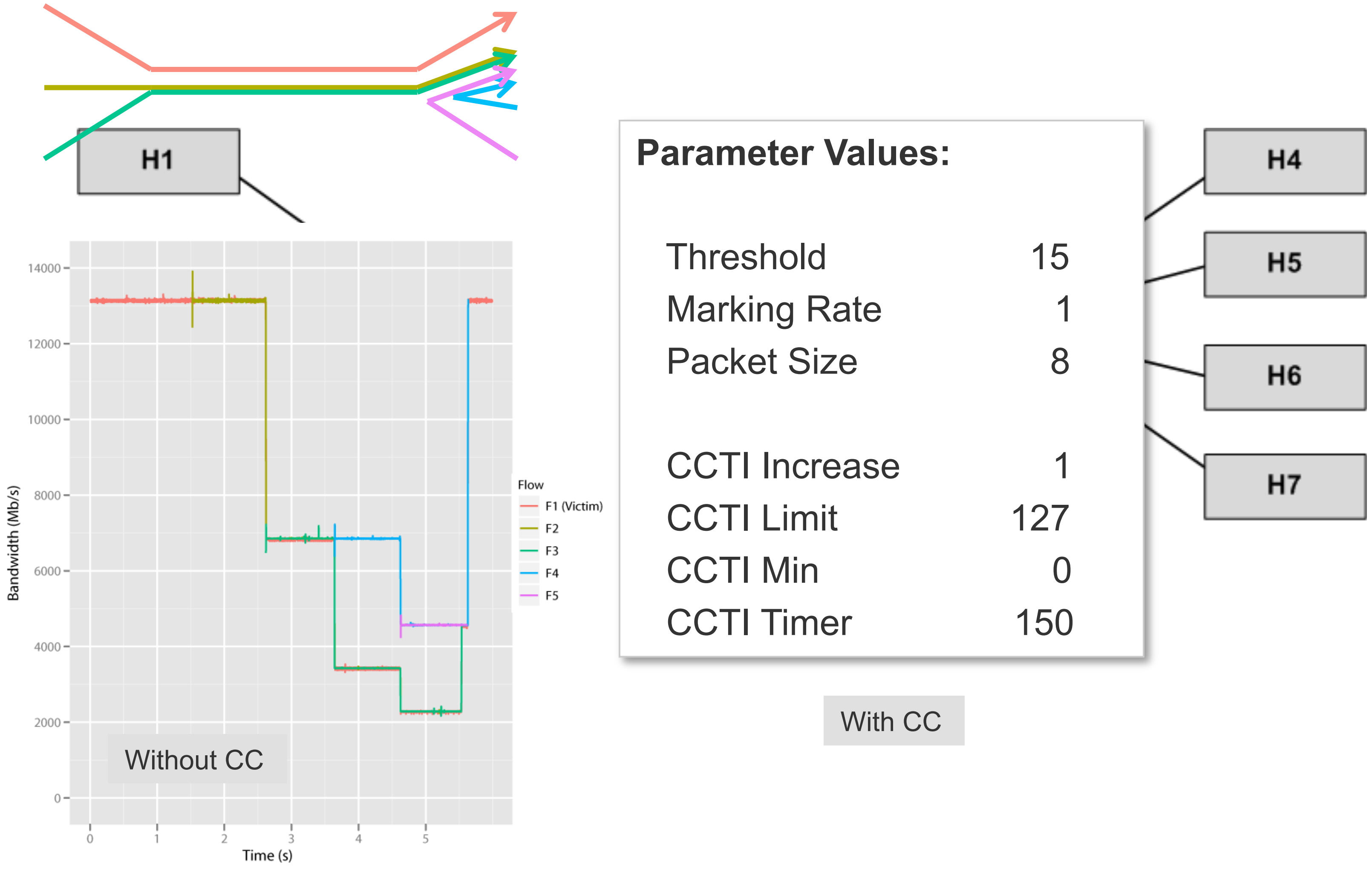
ctim)

T cha sed c nov e.

Experiments show that the HOL blocking leads to performance degradation when CC is not activated.



The InfiniBand CC mechanism is able to remove both the HOL blocking and the parking lot problem.

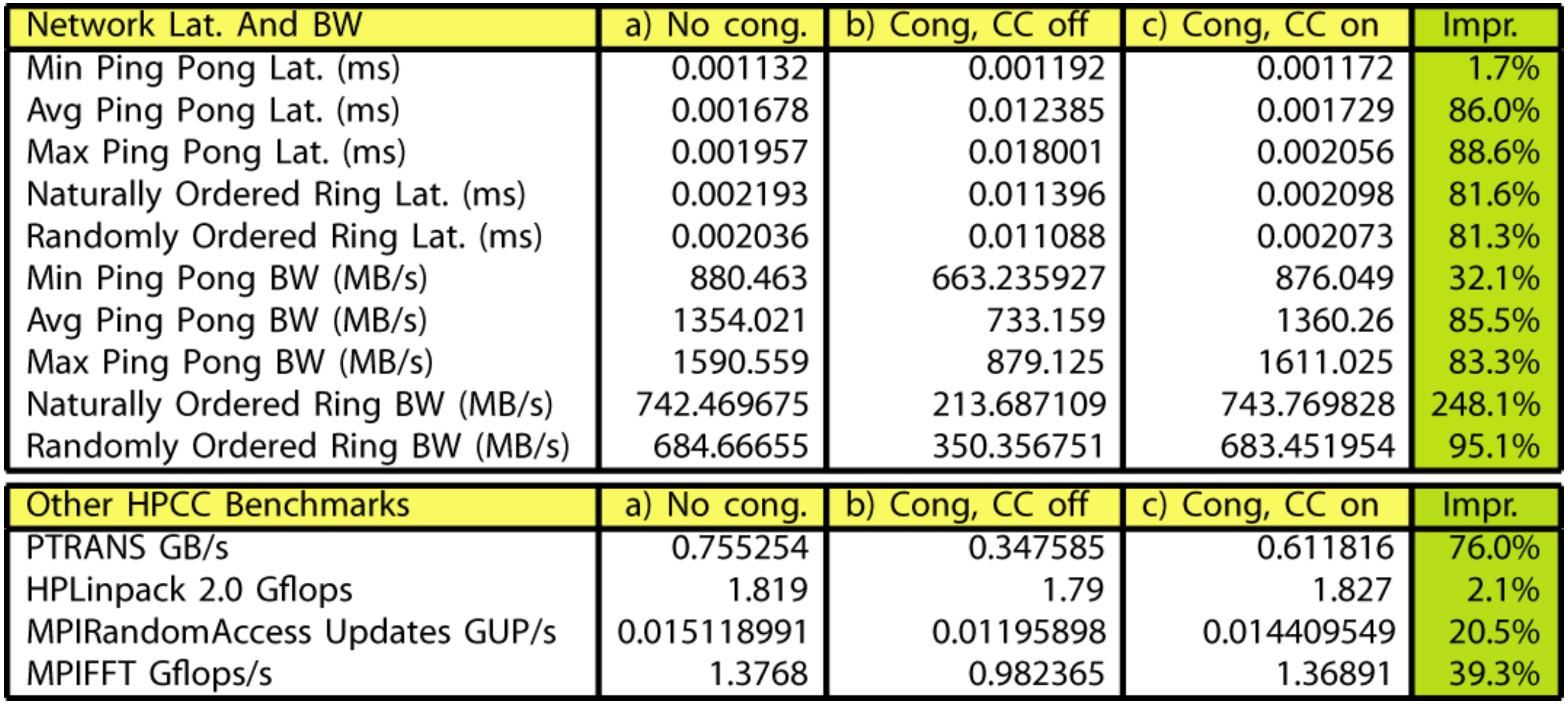


Parameter Values:

Threshold	15
Marking Rate	1
Packet Size	8
CCTI Increase	1
CCTI Limit	127
CCTI Min	0
CCTI Timer	150

With CC

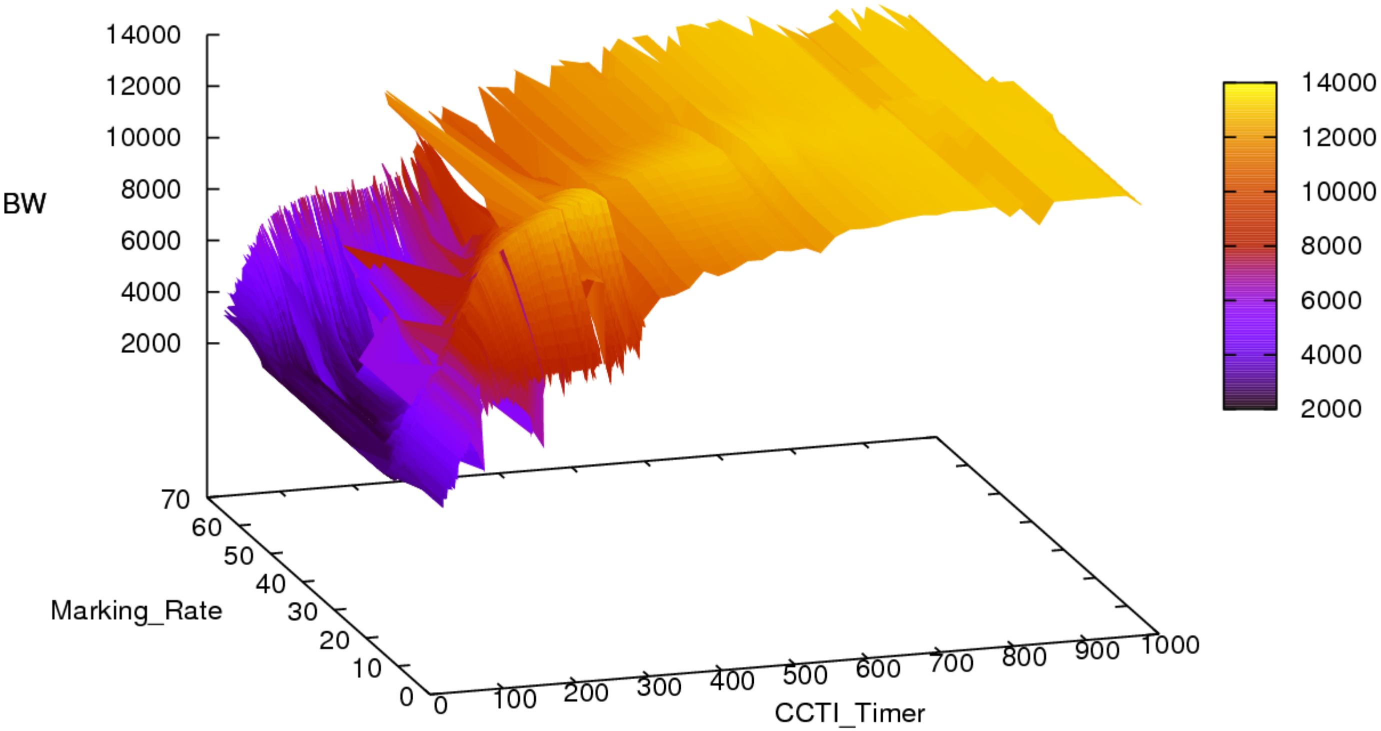
The experiments repeated with the HOL blocked victim flow replaced by the HPCC benchmark.



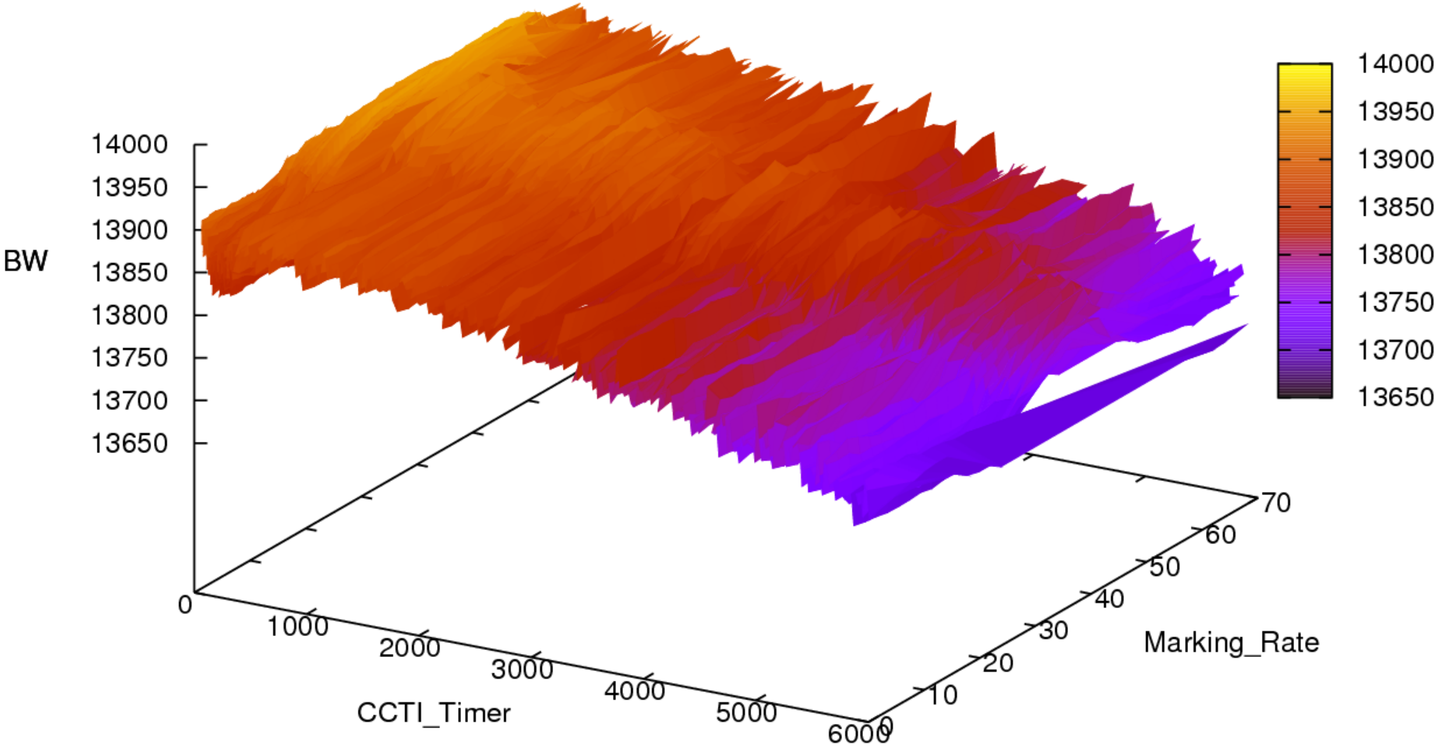
Network Lat. And BW	a) No cong.	b) Cong, CC off	c) Cong, CC on	Impr.
Min Ping Pong Lat. (ms)	0.001132	0.001192	0.001172	1.7%
Avg Ping Pong Lat. (ms)	0.001678	0.012385	0.001729	86.0%
Max Ping Pong Lat. (ms)	0.001957	0.018001	0.002056	88.6%
Naturally Ordered Ring Lat. (ms)	0.002193	0.011396	0.002098	81.6%
Randomly Ordered Ring Lat. (ms)	0.002036	0.011088	0.002073	81.3%
Min Ping Pong BW (MB/s)	880.463	663.235927	876.049	32.1%
Avg Ping Pong BW (MB/s)	1354.021	733.159	1360.26	85.5%
Max Ping Pong BW (MB/s)	1590.559	879.125	1611.025	83.3%
Naturally Ordered Ring BW (MB/s)	742.469675	213.687109	743.769828	248.1%
Randomly Ordered Ring BW (MB/s)	684.66655	350.356751	683.451954	95.1%

Other HPCC Benchmarks	a) No cong.	b) Cong, CC off	c) Cong, CC on	Impr.
PTRANS GB/s	0.755254	0.347585	0.611816	76.0%
HPLinpack 2.0 Gflops	1.819	1.79	1.827	2.1%
MPIRandomAccess Updates GUP/s	0.015118991	0.01195898	0.014409549	20.5%
MPIFFT Gflops/s	1.3768	0.982365	1.36891	39.3%

The average throughput of the *victim* flow as a function of the Marking_Rate (sw) and the CCTI_Timer (host).



The average combined throughput of the *contributors* as a function of the Marking_Rate and the CCTI_Timer.



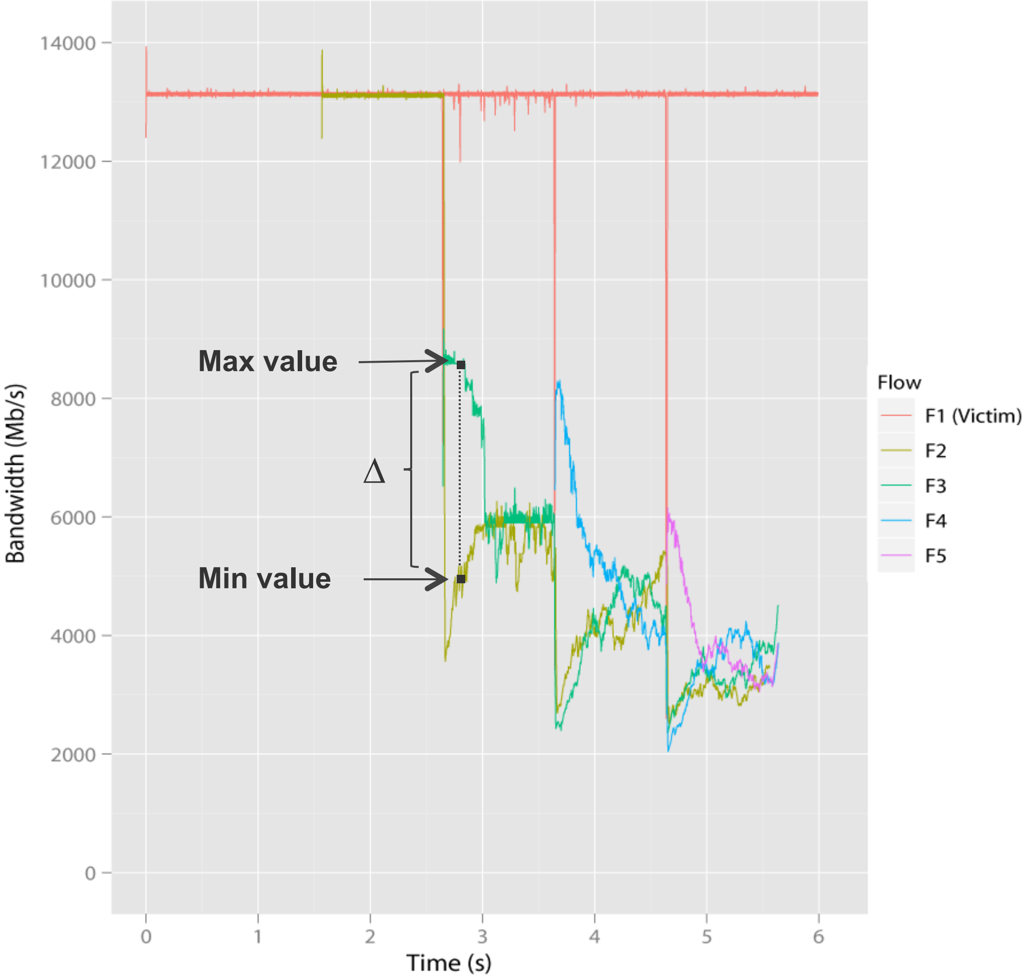
Contributors may experience unfairness if an unfortunate CCTI_Timer value is chosen

If the ”wrong” timer is used the contributors experience unfairness for an extended periode of time after a new contributer is added.

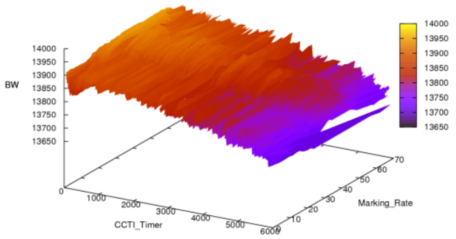
Δ = (max value) – (min value)

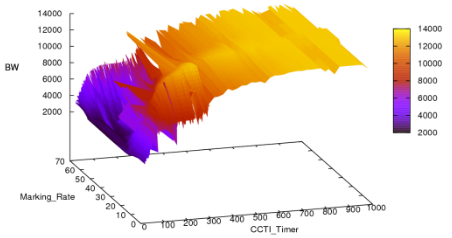
The “treatment variation variable”:

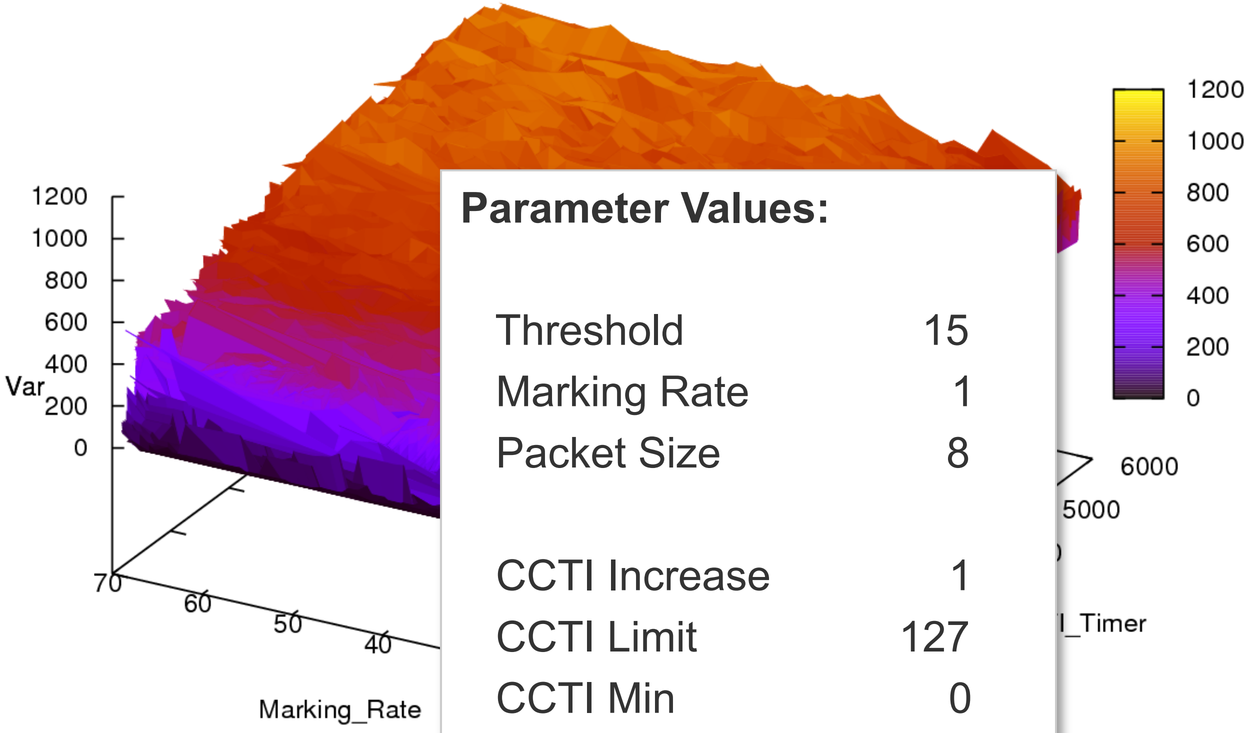
TVV = Var(Δ_1, Δ_2, ..., Δ_n)



The “treatment variation variable” rules out a large part of the parameter space.



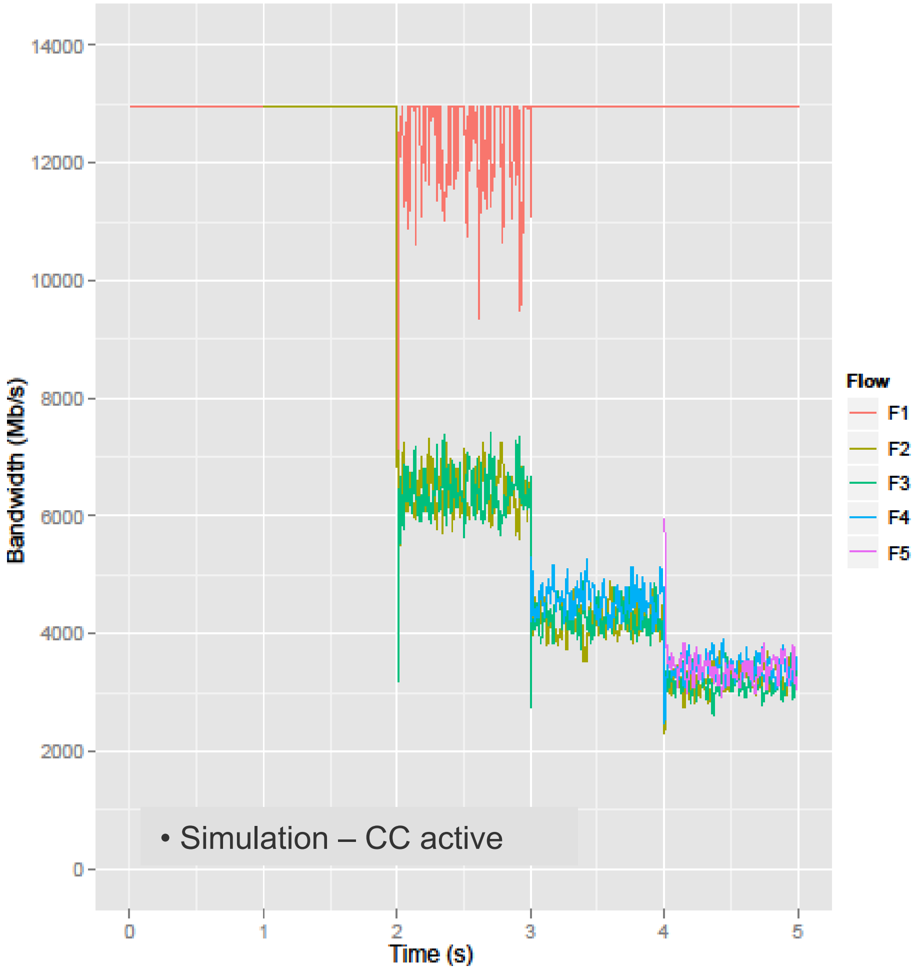




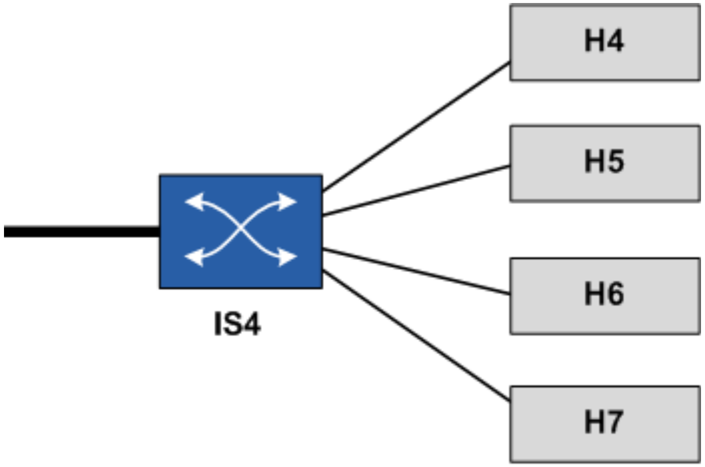
Parameter Values:

Threshold	15
Marking Rate	1
Packet Size	8
CCTI Increase	1
CCTI Limit	127
CCTI Min	0
CCTI Timer	150

The InfiniBand CC mechanism is modeled in OMNet++ to study larger networks.



- Simulation – CC active



- Hardware – CC active

Ongoing Research: InfiniBand Congestion Control in Fat-trees

!HS, !CC

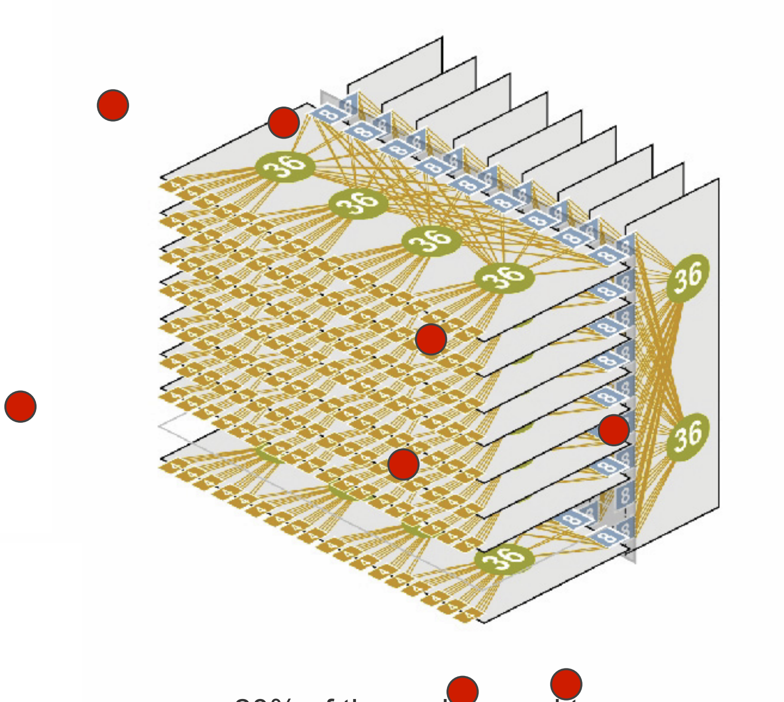
HS, !CC

HS, CC☹

HS, CC☺

HS, CC – QP ☺

!HS, CC



- 20% of the nodes send to *everyone*
- 80% of the nodes send to 8 hotspots

Further simulation studies:

- Different traffic patterns
- Other topologies
- Application traces

Switch figure from: *SUN™ DATACENTER INFINIBAND SWITCH 648 ARCHITECTURE AND DEPLOYMENT.* White paper, June 2009.

HS, C

The CC features in IB works…

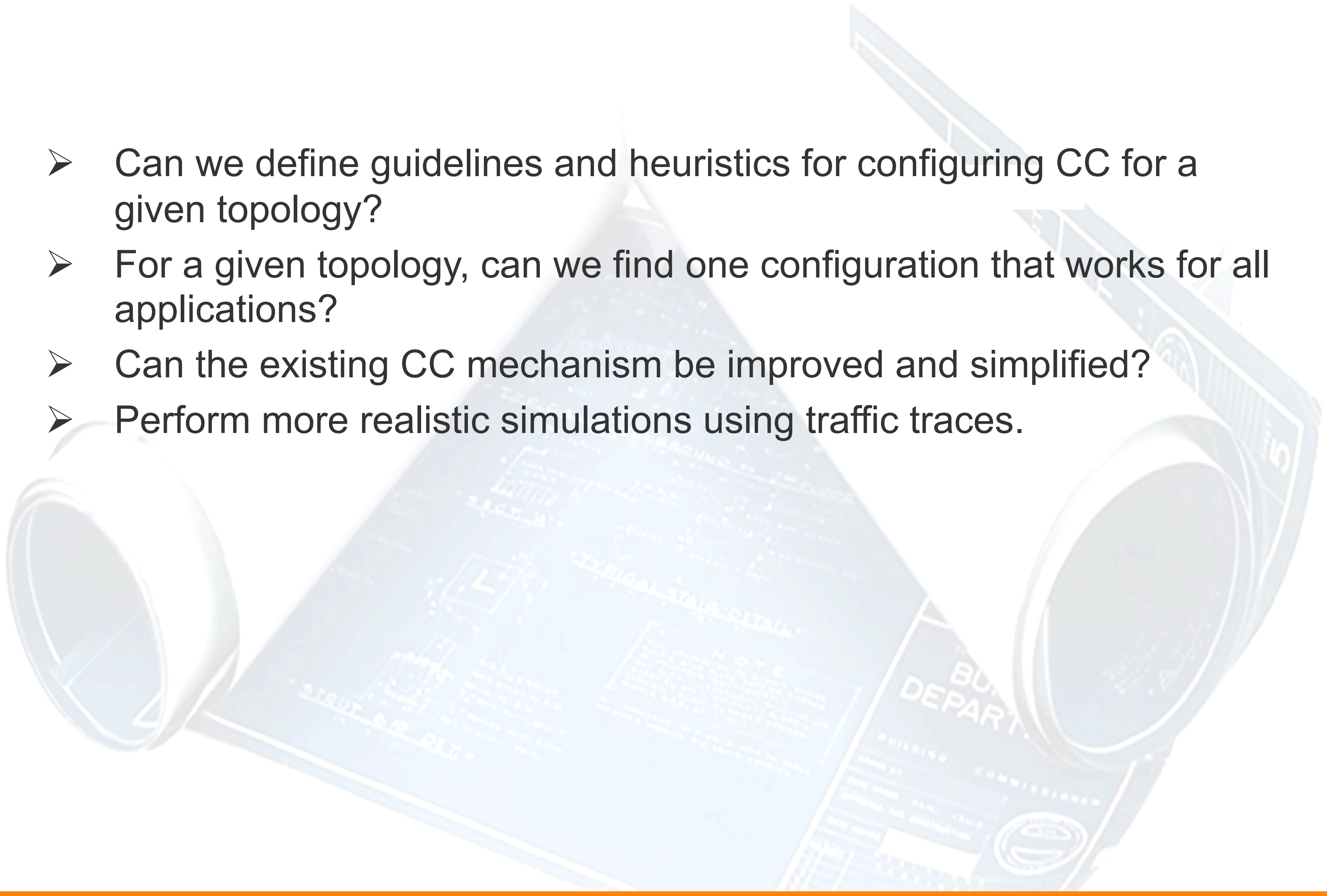
- ✓ It removes the HOL problem, which can be severe without CC.
- ✓ As a bonus it also removes the parking lot problem for the congested flows.
- ✓ It has a negligible negative effect on throughput when no congestion is present.

but…

- ➢ It must be properly configured and this can be time consuming.
- ➢ It is not well understood how to properly configure a given cluster.
- ➢ The real world scenarios where CC is beneficial is not well understood.

Ongoing research

- Can we define guidelines and heuristics for configuring CC for a given topology?
- For a given topology, can we find one configuration that works for all applications?
- Can the existing CC mechanism be improved and simplified?
- Perform more realistic simulations using traffic traces.



References

1. Ernst Gunnar Gran et al. *First experiences with Congestion Control in InfiniBand Hardware*. In the proceedings of IPDPS 2010.
2. Ernst Gunnar Gran et al. *InfiniBand Congestion Control – Modeling and Verification*. In the proceedings of SIMUTools 2011.
3. Ernst Gunnar Gran et al. *On the Relation Between Congestion Control, Switch Arbitration, and Fairness*. To appear in the proceedings of CCGrid 2011.