First Experiences with Congestion Control in InfiniBand Hardware

Sven-Arne Reinemo
Simula Research Laboratory

Credits

- Ernst Gunnar Gran – Simula
- Magne Eimot – Simula
- Tor Skeie – Simula
- Olav Lysne – Simula
- Lars-Paul Huse – Sun/Oracle
- Bjør Dag Johnsen – Sun/Oracle
- Line Holen – Sun/Oracle
- Gilad Shainer – Mellanox/HPC Advisory Council

Presentation Outline

- A few words about Simula
- Introduction to network congestion
- Congestion control in InfiniBand
- Experiment results
- Summary
- Future work

Simula Research Laboratory

Established 1 January 2001.

Owned by the Norwegian government (80%), Norwegian Computing Center (10%), and SINTEF (10%).

Three subsidiaries:
- Simula Innovation AS (establ. 2004)
- Kalkulo AS (establ. 2006)
- Simula School of Research and Innovation AS (establ. 2007)

The SSRI is co-owned by Simula Research Laboratory (56%), StatoilHydro (21%), the Municipality of Bærum (14%), Telenor (7%), Norwegian Computing Center (1%), and SINTEF (1%).

Kalkulo and Simula Innovation are both fully owned by Simula Research Laboratory.

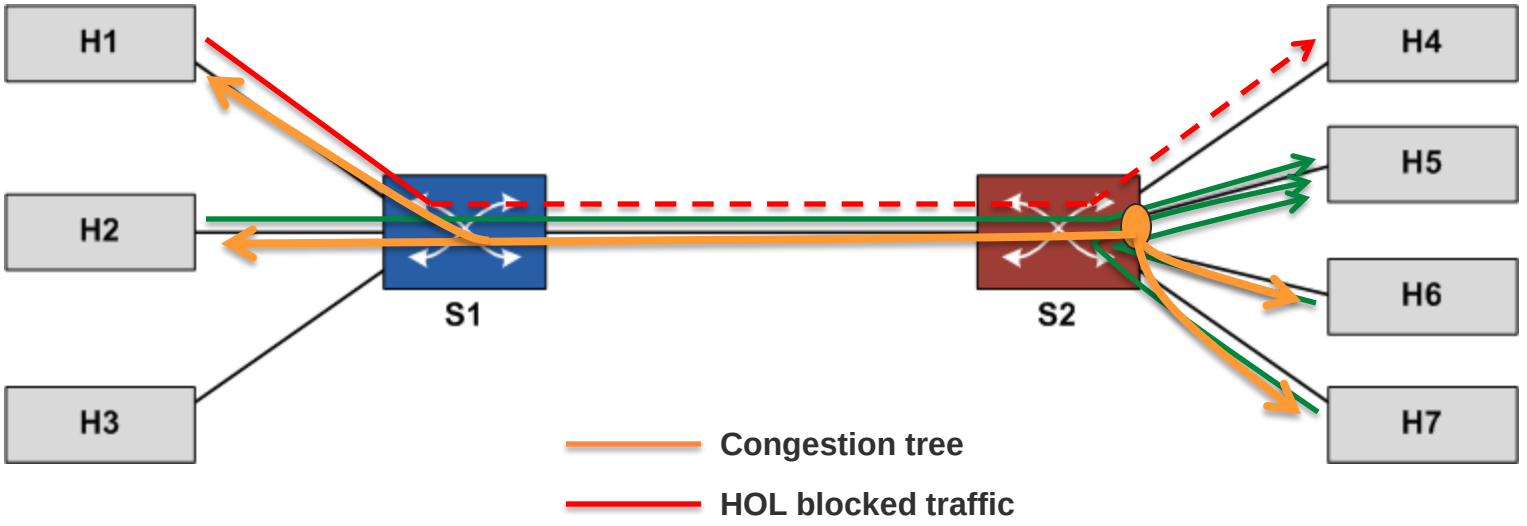
KEY FIGURES	
Operating revenue 2008	**MEUR 12**
Basic funding	**MEUR 7**
Employees year end 2008	**118**
Active PhD students year end 2008	**35**
PhD theses supervised to completion (2001-2008)	**36**
Master students supervised to completion (2001-2008)	**169**

Mission

Simula carries out basic research, explores ways to apply the research in both industry and the public sector, and educates master and PhD students.

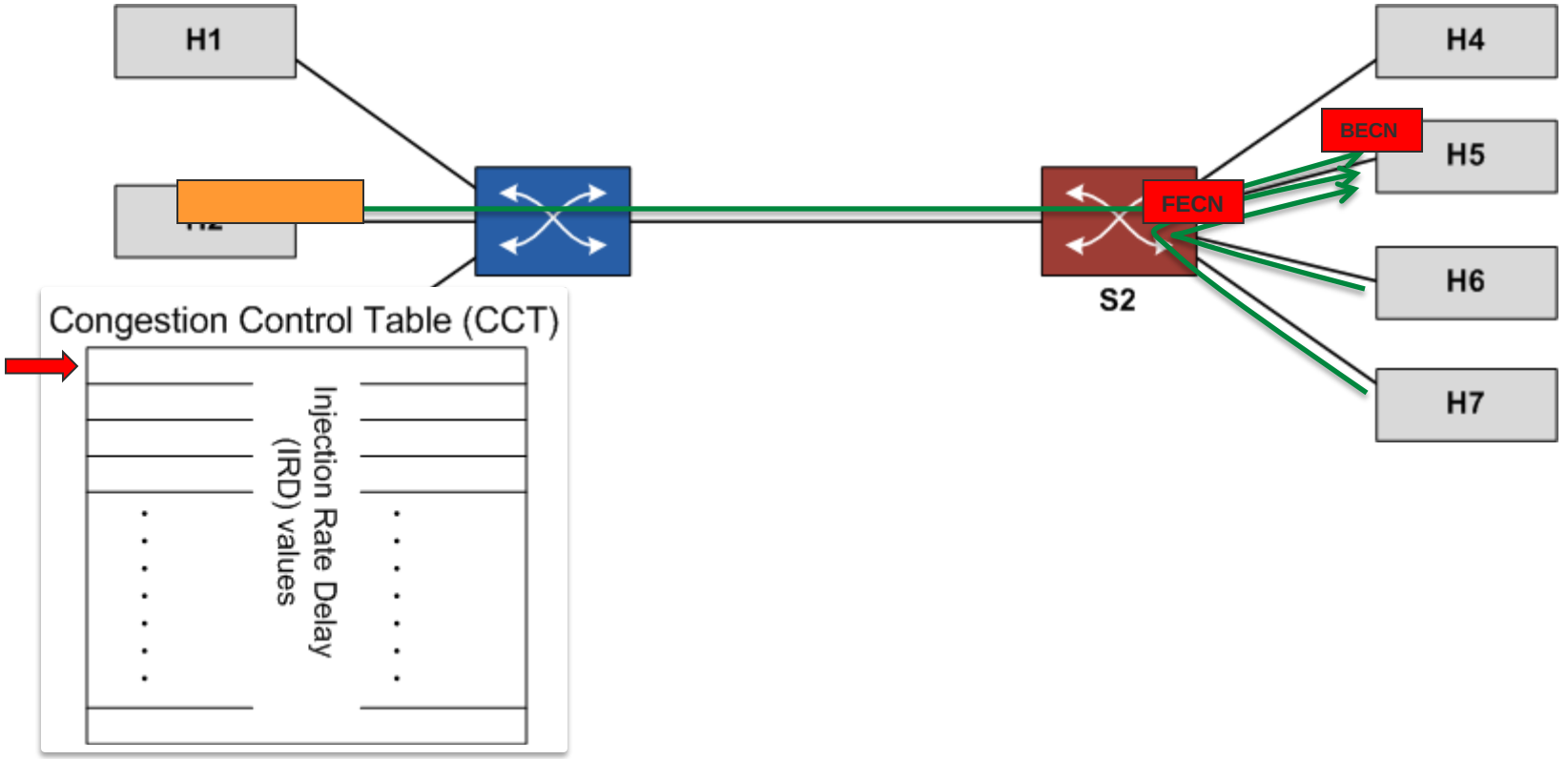
Three autonomous research departments:
- Networks and Distributed Systems
- Scientific Computing
- Software Engineering

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



To avoid performance degradation, the HOL blocking must be removed.

The InfiniBand CC mechanism relies on a closed loop feedback control systems to remove the congestion tree.



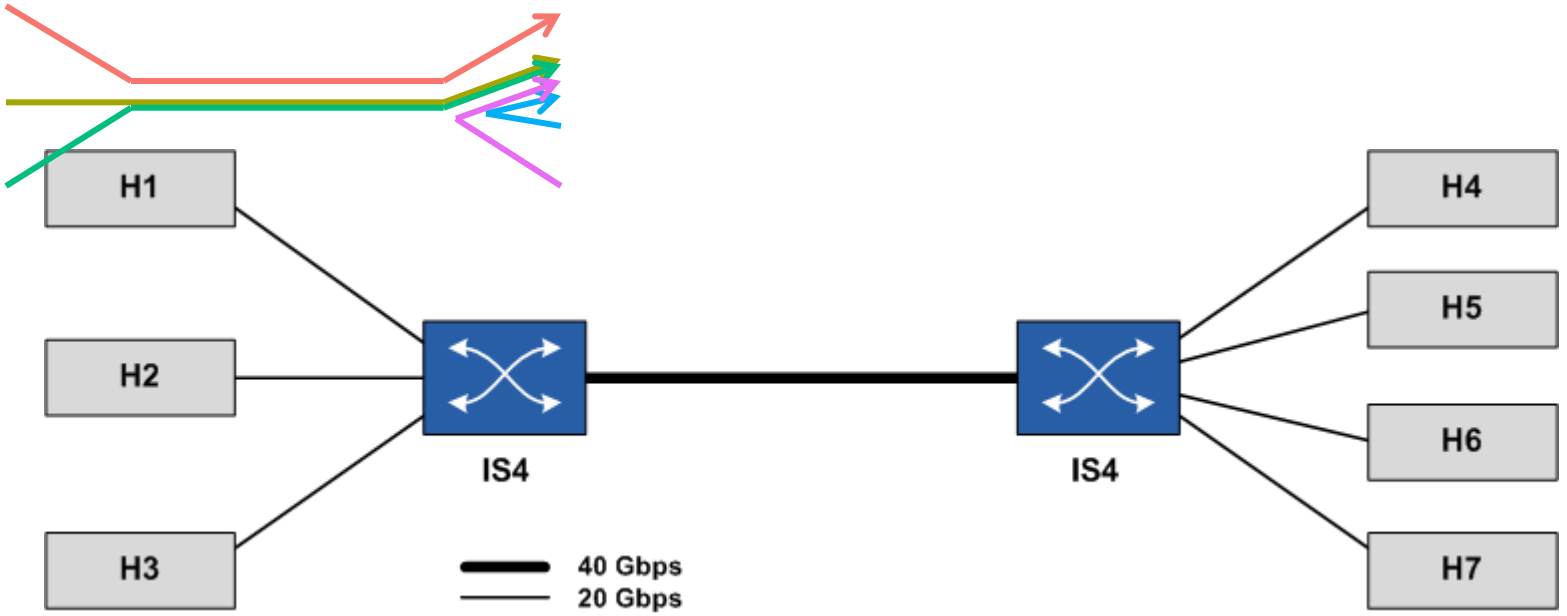
The InfiniBand CC mechanism is configured using several parameters

- Switch
  - Threshold
  - Marking rate
  - Packet size
- Host
  - Congstion Control Table (CCT)
  - CCT Index Increase
  - CCT Index Limit
  - CCT Index Min
  - CCT Index Timer

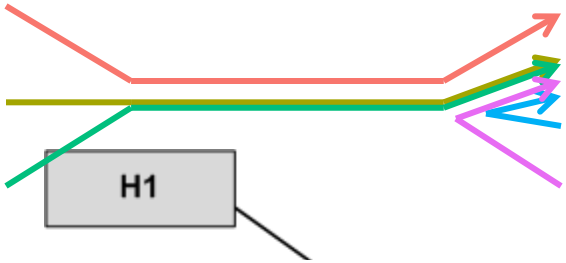
Parameter Values

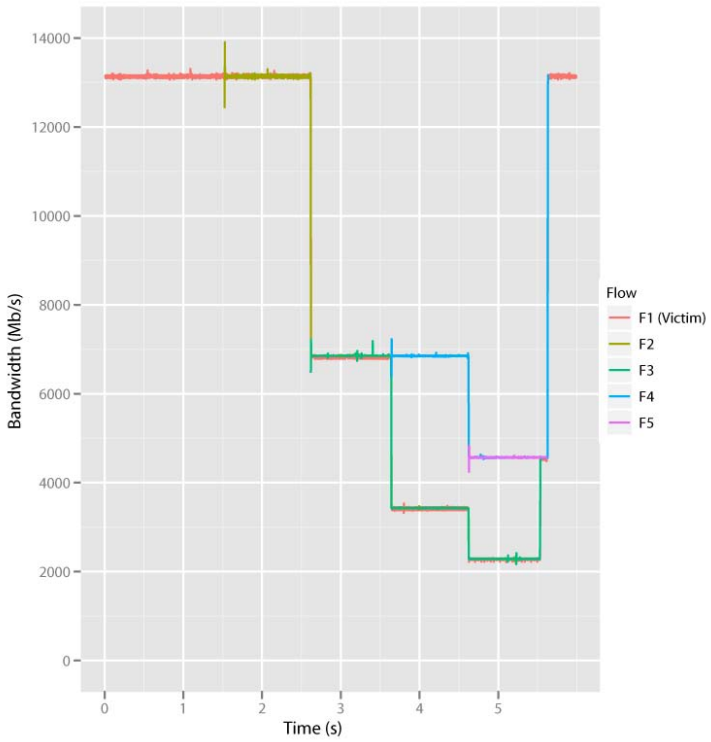
Threshold	15
Marking Rate	1
Packet Size	8
CCTI Increase	1
CCTI Limit	127
CCTI Min	0
CCTI Timer	150

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

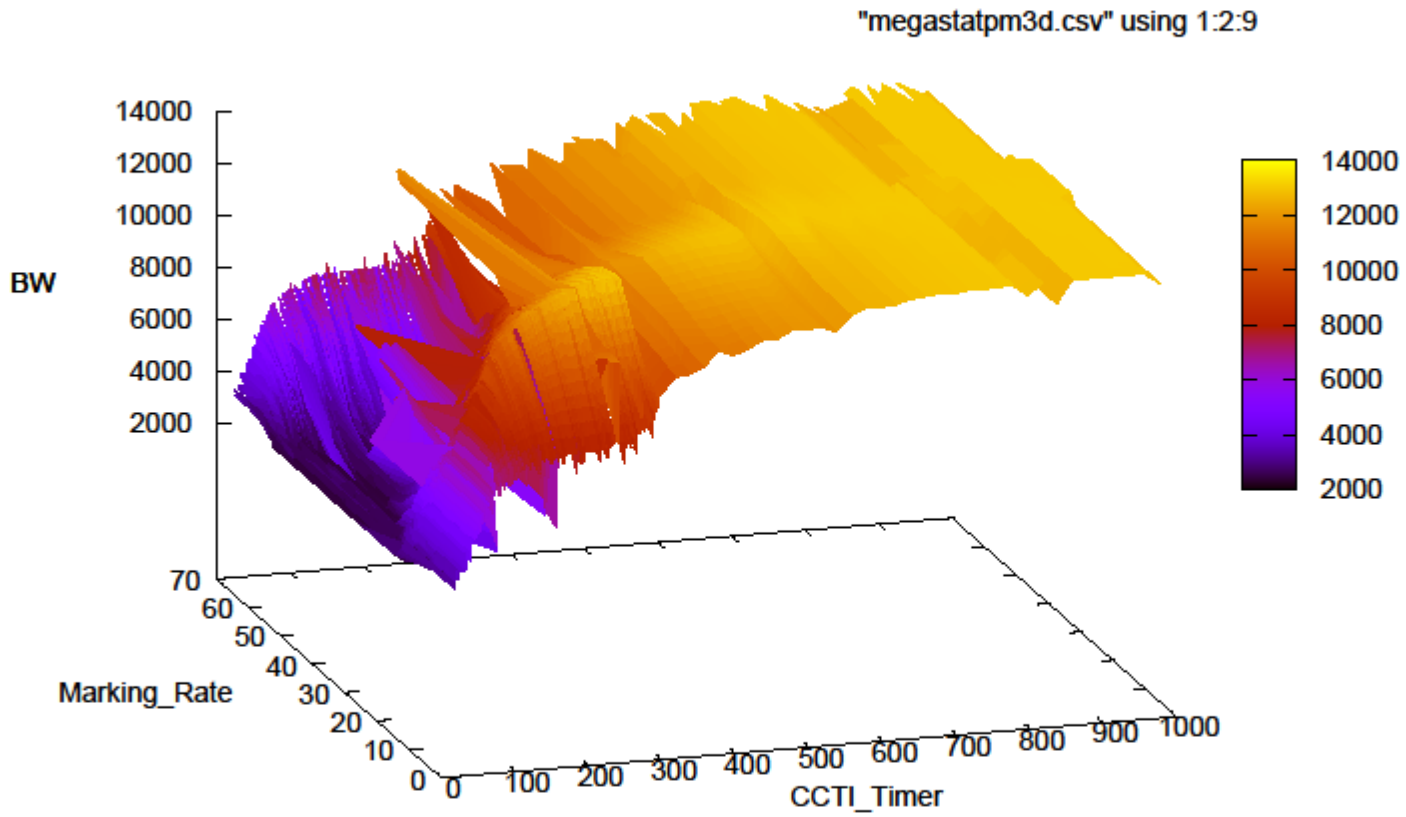
Threshold	15
Marking Rate	1
Packet Size	8
CCTI Increase	1
CCTI Limit	127
CCTI Min	0
CCTI Timer	150

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%

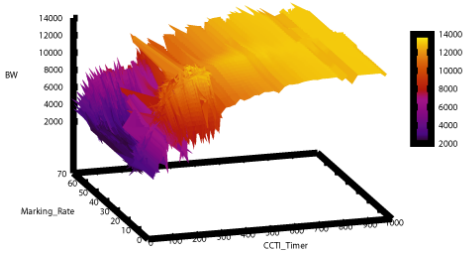
The parameter space is huge as this example shows.



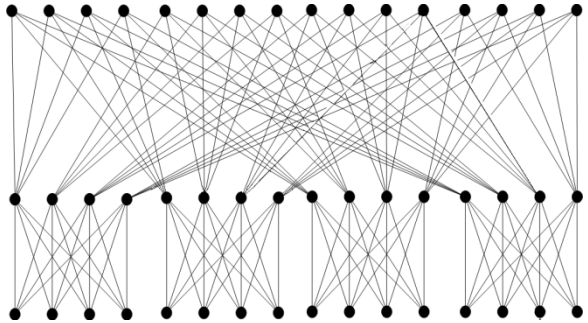
The InfiniBand CC mechanism works, but is hard to configured correctly. Bad configuration can reduce performance.

- IB CC works!
- But many parameters to choose from and their correlation is not well understood.
- A bad configuration can be worse than living with congestion.
- We need to understand the parameter space.
- And hopefully we can find a formula, heuristic or guideline to configure IB CC.
- More info: ”First Experiences with Congestion Control in InfiniBand Hardware” to appear at IPDPS, April 2010.

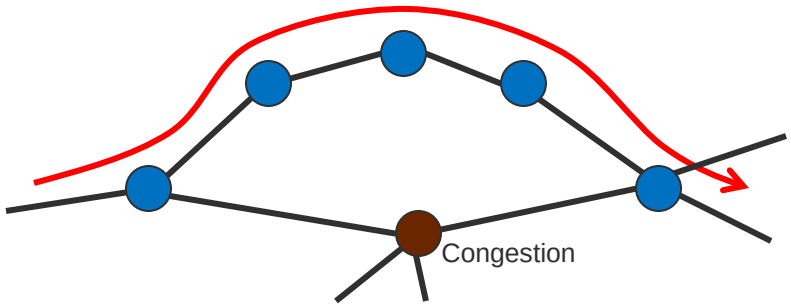
Ongoing research includes both further hardware experiments and simulation studies to:



...explore the CC parameter space



...study CC in larger topologies



...look at adaptive routing as a supplementary mechanism to CC

Questions?