Next Generation of CEE Flow Control: Problem Identification

M.Gusat, K.Keshav, C.Minkenberg, D.Crisan, J.Lynch and Renato Recio

IBM Corporation

July 18th, 2013

DCB 802 Plenary Geneva

Target: Problems...

1. Identification, e.g.
    1. Why lossless DCN/SDN?
    2. PFC: Buffer design; BDP-dependency; self-clocking (lack thereof)
    3. Multi-hop fabrics: HOL-blocking; Deadlocks
    4. QCN: Complexity; tuning; multi-flow HS; fairness
2. Issues' prioritization
3. ONE PROBLEM STATEMENT
4. Solution space

Problem (weak) Statement(s)

- In multihop fabrics HOL-blocking within a shared priorities
  - Need for finer flow control granularity: Currently the 8 priorities are insufficient... even for 10G CEE
    - ⇒ Massive HOL-blocking is possible within any single priority, shared by many flows - much more for 100G, 400G and 1T.
    - ⇒ New flow control ‘lane’ identification mechanisms required, e.g. S-VLAN.

- Correctness, Performance and 1/2 Buffer size (cost, power) =>
  - Correctness, aka **Lossless** operation: PFC requires large 'skid' RX buffers, which depend on the BDP = Bw * RTT. Beyond BDPmax (longer or faster link), PFC is no longer lossless. OTOH, a cdt scheme is ALWAYS lossless.
  - Performance: The PFC buffer elicits 2x size = (1x Overflow 'skid' buffer for correct lossless operation) + (1x Underflow protection for work conservation to provide 100% downstream link utilization during STOP/GO traffic)
  - Shift from PAUSE (grant On/Off) to e.g., credits would save 50% buffers, by overlapping Overflow and Underflow areas. (yet, this ain’t about credits...)

Market Segmentation: Focal Points

1. Switch ASIC / OEM vendors: Buffer, HOL, VOQ, tuning...

2. DCN/Cloud: Fat-tree topos, HOL, VOQ, e2e...

3. HPC, BA, embedded CPU/switch: Direct topo, deadlocks, HOL, routing...

4. SDN / Virtualization: HOL, e2e, L2 support...

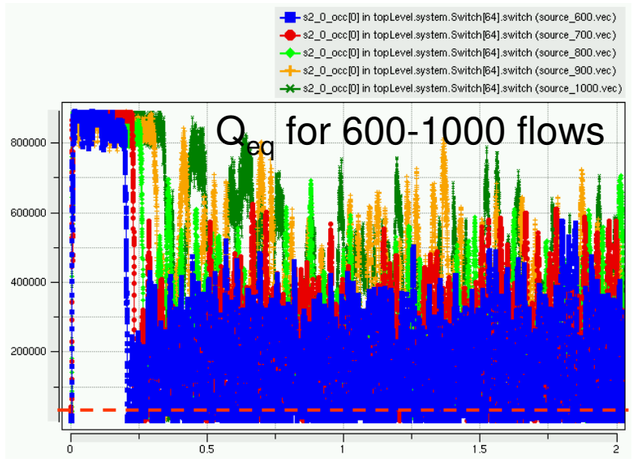
5. MAN/WAN/inter-DC: transport, e2e...

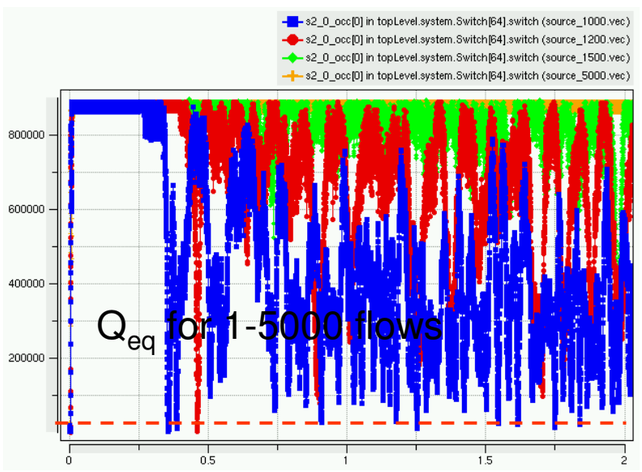
Why HOL-blocking in lossless DCNs... Wasn’t this fixed already by QCN?

(some) QCN Challenges

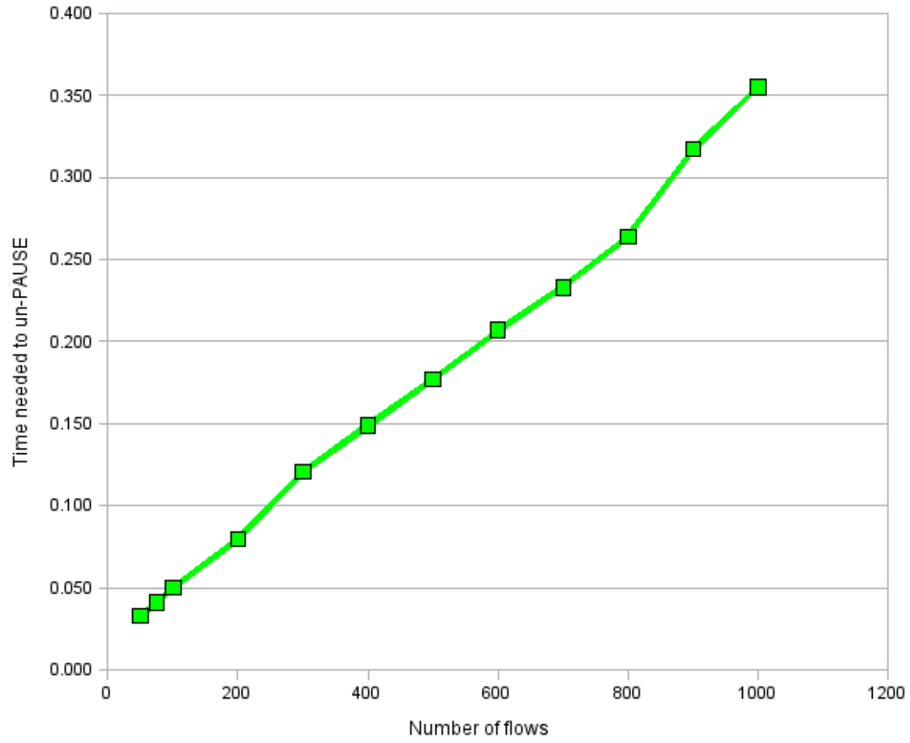
Complexity; Many Flows; Unfairness.

QCN Buffer Control f(# flows): Q_{eq}=33KB





QCN control lag: Until PFC=0 and within Q_{eq} limits



Unfairness: An Extreme Case of 1 Winner + (N-1) Loosers