CM: Reaching "Good Enough" From Below

Are we there yet…?

M. Gusat and C. Minkenberg
IBM ZRL GmbH

802.1 Interim
Los Gatos, January 2008

Outline

- Outside .1au: Top 5 req'ts compiled from DC & HPC feedback
- Inside .1au: CM evolution in 802
- Conclusion: Are we there yet – Do we have confidence to draft CM std. ?

Outside .1au: Top 5 DC & HPC req'ts
Req'ts, sizes and wklds are evolving - 1

- Collected feedback: DC industry's Top 5 from users' perspective

1. **Cost** (of Management)
    1. Manageable and visible (no black box operation; signal when *upgrades needed*)
    2. Low TCO: incremental deployment (1-100GigE), heterogenous nets, multiple vendors
    3. Tunable: at least guidelines for parms settings per topology, wkld and objective.
    4. Civil: Mix'n Match with **other apps** & protocols
        1. Conflict-free mixes of multiple protocols (load balancing/LB, adaptive routing/AR, NetFlow, virtualization)

2. **Speed**
    1. Dominant hotspot model: "meteor showers"
        1. Many spurious HS's < 100ms (partly leveled thru load balancing and adaptive routing)
        2. Few persistent HS's: Misconfiguration, under-provisioning, app bugs must be exposed (1.1)
    2. Layer 2 req't: 10x faster than any other L3..7 alternative.
        1. Indep. of its choice of CM algorithm, 802.1 is expected to provide the fastest and most accurate load sensor to higher layer apps

Outside .1au: Top 5 DC & HPC req'ts
Req'ts, sizes and wklds are evolving - 2

3. **Scalability** to large DCs: size & wkload.
    3. Queuing delay (PPP=On) dominates multihop DCNs
    4. Flows span RTT=[5us, 10ms] and Bw=[1,100] Gbps
        => bandwidth-delay product (BDP) range >> PAR=5Mb

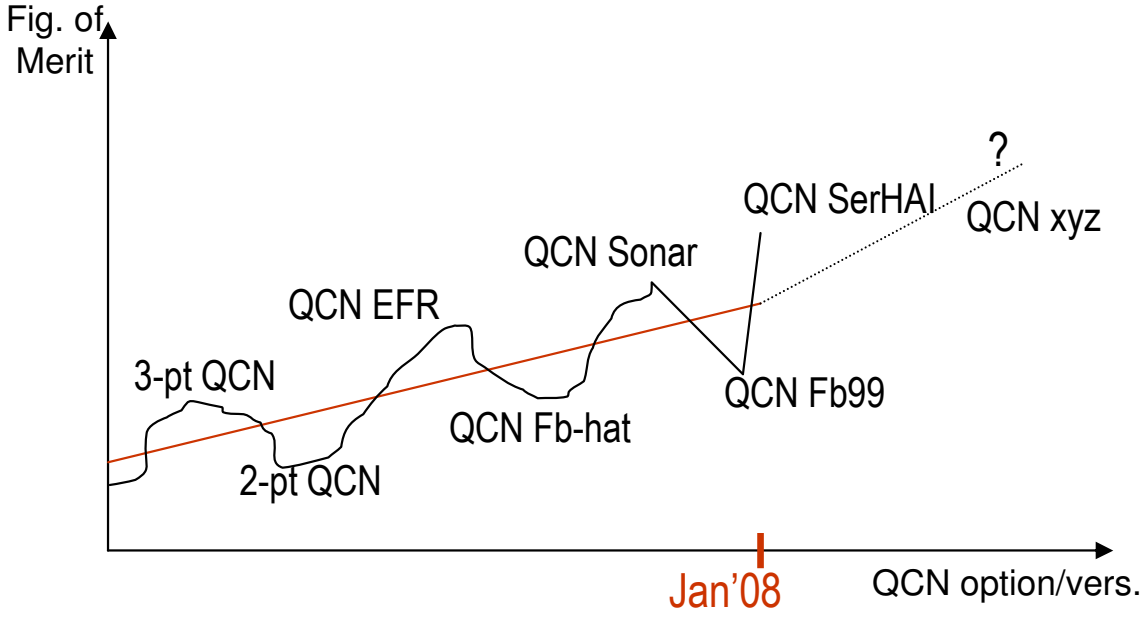
4. **TCP**
    3. Stacking: Over L2 CM
    4. Sharing: With L2 CM
    5. Using: the L2 feedback as pre-congestion notification (PCN) – see http://www.ietf.org/html.charters/pcn-charter.html

5. **Location… Location… Location**.
    3. CPID is the single most requested feature…!
    4. Most DC and HPC apps, mgnt. tools and protocols elicit timely CPID feedback

Inside .1au: Incumbent CM's evolution in 802 - Iterations

QCN '06-'08 timeline

1. 2-point QCN (base)
2. 3-pt. QCN
3. QCN EFR
4. QCN HAI
5. QCN Fb-hat
6. QCN Sonar
7. QCN Fb99
8. QCN Serial HAI



- 6-8 QCN options&versions in 14 months
  - Oct.'06: Paper on BCN w/ 3 self-recovery options
  - 1st QCN alpha proposal
    - ✓ Large conclusions
    - ✓ - The proposed BCN is pretty robust
    - ✓ - It gives consistent performance, measured in FCT and fairness
    - ✓ • Even at high loads
    - ✓ • And even when switch-signaled increase is turned off
  - self-recovery also introduced here;

- Latest algorithm: QCN-SerHAI (last p-code release Jan. 17th, 2008)

.1au's R&D pattern since Fall'06: QCN's Preemptive Spiral - Loops
T-Model: You can have your Ford in any color - as long as it's black.

RepeatLoopWhile(!BreakCondition) {

1. New QCN option/version proposed
    1. Conceptually appealing
    2. Scarce and intriguing initial results
2. Preempt ongoing QCN modeling work
    1. More solid previous results discarded as obsolete
    2. Advanced BMRKs not performed //large topos not exercised
3. Simulation teams asked to refocus on new QCN
4. New QCN pseudo-code issued
5. Basic benchmarking restarted
    1. Heavy simulations fired up at various sites
6. New results issued
    1. Positive and negative findings published on Adhoc calls
    2. Suggestions and new param. settings from teams
7. P-code further tweaked
8. // Next QCN rev. issued: goto (1)}

Current status: Catching up w/ Latest QCN

- QCN serial HAI: current proposal
- Some results already available (not shown here)
- Our view: QCN's maturing
  - ✓ We see marked improvements
    - Recovery time
    - Simple hotspots are well-controlled
    - Better trade-off between stability and transient response (basic BMRKs)
  - ✓ Open
    - Blackbox operation: open loop recovery w/ pre-loaded 'spring' (past critique)
    - *OG*; fairness/starvation; stability; scalability; hetero networks; multiple HS…
- BMRK-ing and fat-tree topos: ramping-up
  - ✓ unless a new QCN version preempts the ongoing effort…

Are we there yet – Do we have our CM std.?

- DC: Problem and Solution are Evolving

- Outside 802.1
  - DCNs are growing in hops and nodes
  - Apps assume QoS and tight Tput/L bounds
  - Wklds are still to be characterized
- Trend: DC = HPC + 7-yr. (lag decreasing)
  - Today's largest installed supercomputer exceed 0.2M cores
  - DC's are lagging behind HPC by 1-2 OM (except largest few DC's)

- Inside 802.1: QCN has evolved, while
  - Preempting serious benchmarking attempts
  - Reaching from below for "good enough", QCN tends to over-simplify and under-solve the problem of CM
  - Disregarding the TOE alternatives and IETF efforts
  - Aiming for a "*closed*" solution shouldn't *lock* into all-or-nothing deal:
    - ✓ We must avoid a disabling effect on DC apps and protocols that rely on the presence of L2 feedback @ RP, while not using the .1Qau SRF (RL).

Our View on .1au's CM Selection

1. Must work in basic scenarios w/o hurting the common case
    1. QCN-SerHAI almost there…
2. Datacenters are growing in size and speed; workloads grow in complexity => BDP range exceeds PAR.
    1. RTT, HSV and HSD ranges are not predictable today.
3. Reaching 'from above' reduces confidence.
    1. Reducing the PAR objectives and BMRK-ing intensity to meet QCN's current capabilities leads to effort replication in other fori.
4. 802 standard must close on an algorithm - without preventing the natural growth of DC infrastruct. and apps.
    1. QCN-SerHAI may provide the core alg.

Q: How to achieve closure while still disagreeing on DC apps, metrics, scale and tuning?

1. Finish and close the algorithmic work using QCN-SerHAI as base.
2. Unlock signaling => open thru judicious options. Enable new apps and vendor differentiation.
    - execute The Stockholm Agreement

Conclusions: Are we there yet – Do we have our CM std. ?

- Nearly…
  - QCN-SerHAI is *almost* acceptable

- We propose to answer the remaining concerns thru open signaling:
  - **CPID** (tag/probing) **options** according to **The Stockholm Agreement**

- Only thus we can agree today on QCN-SerHAI as CM algorithm.

BKUP