# 802.1Qau Plenary Meeting March 2007

Tuesday, 3/13/2007

Attendees:

| | |
|---|---|
| Manoj Wadekar | Intel |
| Pat Thaler | Broadcom |
| Uri Cummings | Fulcrum |
| Jan Bialkowski | Infinera |
| Claudio Desanti | Cisco |
| Ravi Shenoy | Emulex |
| Mike Ko | IBM |
| Craig W. Carlson | Qlogic |
| Joe Pelissier | Brocade |
| Balaji Prabhakar | Stanford University |
| Asif Hazarika | Fujitsu |
| Robert Brunner | Ericsson |
| Guenther Roeck | Teak |
| Hiroshi Ohta | NTT |
| Tae-eun Kim | Extreme Networks |
| Menu Menuchehry | Marvell Semiconductors |
| Diego Crupnicoff | Mellanox |
| Davide Bergamasco | Cisco |
| Bruce Kwan | Broadcom |
| Chien-Hsien Wu | Broadcom |
| Anoop Ghanwani | Brocade |
| Mark Gravel | HP |
| Raj Jain | Washington University |
| Mitch Gusat | IBM |
| Mick Seaman | |
| Osama Aboul-Magd | Nortel |

Recording Secretary: Manoj Wadekar

Minutes:

1. Pat Thaler: Agenda, patent policy etc.
    a. http://www.ieee802.org/1/files/public/docs2007/au-thaler-agenda-0703.pdf
    b. Agenda discussed and approved.
    c. Patent policy shared and discussed in TF.
    d. Review of CN Objectives and schedule:
    e. Objectives:
        i. Should there be any objective that defines "limit" for deployment network?
        ii. CM/No-CM traffic classes requires Transmission Selection different than Strict Priority. However, it is not part of .1au project. We need presentation to discuss: Need, 5 criteria etc.
2. Manoj Wadekar: Simulation Ad-hoc report
    a. http://www.ieee802.org/1/files/public/contrib/au-sim-wadekar-adhoc-report-031307-v1.pdf
3. Davide Bergamasco: Ethernet Congestion Manager
    a. http://www.ieee802.org/1/files/public/docs2007/au-bergamasco-ethernet-congestion-manager-070313.pdf

   b. Slide 5: What is unit: this is scaling factor for Timestamp (multiplying factor)
   c. What about ECM-MAX? - Foils do not cover it. ECM-MAX generates maximum feedback when Qmc is exceeded. At RP - no distinction between normal ECM messages and ECM-MAX.
      i. Listed on slide 9
   d. Over-sampling is not mentioned in slide 9
   e. Notation on slide 6 is not clear on Sr - it is uniform distribution between 0, 20 KB
4. Prof. Raj Jain: FECN for Datacenter Ethernet Networks
   a. http://www.ieee802.org/1/files/public/docs2007/au-jain-fecn-20070313.pdf
   b. Quick, fairness are important features
   c. Single parameter for whole deployment (small network/large network etc.)
   d. Q:Is tag in each packet? A:No only on each N-th packet. (typically every 1mS)
   e. Source has multiple RL queues. RL Q-id identifies that.
   f. Packet is being reflected at dest - any analysis of impact of delay of reflection at destination?
   g. There are two tags to the packet: Forward tag (collects rate information) and reflected tag (this is just copy of forward tag being sent back to source by the destination/reflector)
   h. Requires rate limiter per destination - in the current proposal.
      i. So, if two flows (for different destinations ) are hashed in a single rate limiter (due to limitations on number of rate limiters) - current proposal does not address it.
   i. Q: How can capacity be calculated if link is flow controlled? A: Will be addressed in following foils. To be tabled.
   j. Q: Variable capacity will result due to multi-class traffic A: not addressed currently. Focusing on single traffic class - in line with other simulations
   k. Slide 9: did not see difference in additive and multiplicative feedback mechanisms
      i. However - agree in principle that additive is right direction
   l. Slide 9: f(q) is calculated with q. Typical implementation will have a table.
      i. Queue length is instantaneous queue length. (Not average)
   m. Slide 9: a, b, c are fixed, hard coded values
      i. However, there is dependency on Qeq - work TBD
   n. Slid 12: Q: inconsistency with slide 9 - A: needs to be fixed
   o. Slide 12: Q: Ri and Rho-I are two 32 bits numbers and they are being divided - difficult to do in HW
      i. A: SW implementation will be ok since this is done at relatively slow speed - per port once in each 1mS
   p. Goal is not to use PAUSE - appropriate rate is reached quickly - avoids packet drops
   q. Q: Given sampling interval - how can one handle no-drop? A: all sources start with small rate and probe to go higher.
   r. Q: How no-drop can be achieved (w/o PAUSE) if sampling is happening at 1mS? A: Each source starts R0 not at line rate (R0 = C/N0)

s. Q: However if sources start light traffic - they get allocated C. If they start now using full C capacity - then CP needs large buffering to avoid packet drops? A: Simulations show later that this case is addressed.
t. Slide 14: Q: You may not even know the capacity. A: Mostly link capacity changes due to failure in link aggregation. One could also measure link capacity (dequeued packets/measurement time)
u. Slide 14: Q: Does this also apply for change in capacity due to multi-priority? A: Maybe, no analysis done currently.
v. Pseudo code for FECN will be posted today.
w. Q: Proposal is claiming single parameter for configuration. How about a,b,c, T? Performance typically depends upon such parameters. A: Tested with fixed param for many configs. However, it is possible that proposal will be stress tested for stability of proposal - now that pseudo code is being released. Testing was done for Qeq = 64 packets. It is possible that different a, b, c may be required for different Qeq.
    i. Single parameter is T.
    ii. N0 - needs to be calculated for (near) worst case - to avoid packet loss

5. Mitch Gusat: Extended Ethernet Congestion Management (E2CM) - Per Path ECM - A Hybrid Proposal
    a. [http://www.ieee802.org/1/files/public/docs2007/au-sim-IBM-ZRL-E2CM-proposal-r1.09.pdf](http://www.ieee802.org/1/files/public/docs2007/au-sim-IBM-ZRL-E2CM-proposal-r1.09.pdf)
    b. Slide 8: Q: "frequent use of BCN_MAX.." - does it imply only during transition? A: Yes, statement needs to be qualified.
    c. Dynamic range for BCN stability needs to be extended - concatenate queues and feedback (path probing)
    d. Slide 10: Scalability : beyond 100G, beyond 1 million flows
    e. Slide 10: Q: "per-flow" - does this allow aggregation of l2 flows at source? A: Yes, that is the plan.
    f. Q: What is SRF? A: Source Rate Function.
    g. Q: Why rate calculation at DST, instead of reflecting it? A: Will be answered in few foils.
    h. Q: Probe often or only once? A: Often while RL in installed
    i. Q:How does this compare with over-sampling? A: Looks better than over-sampling because sampling is increased a lot due to source probing
    j. Q: How often sampled for probe? A: More data waiting from simulation.
    k. Q: Why timestamping at dest? A: Want to calculate queue occupancy on fwd direction.
        i. Q: But is SRC and DST time-synchronized? A: Yes, that is the weak spot - but it is possible to get reasonable estimate of forward latency.
        ii. Q: Why not put reflection in high priority TC and get rid of queue occupancy in reverse path? A: Yes, we have tried it and it works pretty good as well.
        iii. Q: Per flow on DST Rx - is it required? A: Yes, this is a concern. Needs to be addressed as next step..
    l. Q: SRC knows its insertion size. Rate can be known if it knows latency. Why do you need service rate calculation at dest? A: Will look into this.

Wednesday, 3/14/2007

1. Prof. Balaji Prabhakar: Some ideas for simple CM

a. http://www.ieee802.org/1/files/public/docs2007/au-prabhakar-simple-ideas-0703.pdf

b. Rate allocation "over-solves" congestion management problem. Congestion Management tries to achieve smaller goal (subset solution) by "taxing" few larger flows.

c. Reflection point could be at the end station, at edge of the cloud or at each CP.

d. Sampling probability lower bound is used for fast recovery timer

e. Q: What if RTT is larger - in this case RP is reacting prematurely? A - this could be case in internet case - bic-tcp: binary self-increase TCP is addressing exactly this. In window based world - it is automatically self-timed. However in L2-CM networks - RTT is expected low. So, sampling will dominate RTT, rather than network latency.

f. Q: Reason for moving Qoff and Qdelta to RP was to allow use of "w" at RP based on RTT. So, why it helps here to move Fb calculation to CP? A: Gd can be used for similar purpose of tuning.

g. Q: However, w was biasing Qdelta. Why switch is better place to do this? A: Switch is the aggregator and hence good to place to decide on "w".

h. Comment: if there is large buffer at switch (due to longer delay link), then switch can use "w" to weight the feedback. If large RTT is due to multiple hops, then RP can use Gd to weight the feedback.

i. Can leverage significant work being done in BIC-TCP for validation of algorithm - already BIC-TCP is leading the pack for high speed TCP

j. BIC / TCP comparison will show more losses in BIC - but one should compare BIC with REM - as there is signal before actual packet loss.

k. RefP will collapse with CP - on CM cloud boundary. However, taking forward notification as far out in the cloud as possible helps improve performance by getting more feedback to the source in short time (less signaling as feedback is accumulated).

l. Per-path state may have to be carried if - this has to be enhanced for multi-pathing.

2. Davide Bergamasco: Discussion about metrics

   a. http://www.ieee802.org/1/files/public/docs2007/au-sim-bergamasco-on-metrics-070313.pdf

   b. Various metrics decided in Monterey meeting are difficult to implement in models

   c. Slide 4: bad flow - good flow: bucketize them.

3. Prof. Raj Jain: Simulation results

   a. http://www.ieee802.org/1/files/public/docs2007/au-jain-fecn-20070314.pdf (if this link does not work, please check contrib directory - file still may be left there).

   b. Rate, Explicit, forward are key parameters for the proposal.

   c. Slide 26: How is control loop delay is 400 uS achieved? A - By increasing link lengths in the topology.

   d. Q: what R0 should be used? A: Judicious decision is required based on buffering and estimate of how many sources may simultaneously clash.

   e. C: Since exact N is difficult to predict, so is R0. So, packet drop may result. A: Yes, in such cases PAUSE will be used. PAUSE will be used as emergency mechanism to avoid drops.

   f. Q: Slow start complexity - why is justified? A: To avoid frequent use of PAUSE.

   g. Q: PAUSE will influence accuracy of rate calculation and will affect convergence time. A: It takes 5-10 "T" intervals. => 5-10mS. Source gets "previous" right answer in one iteration.
   h. Slide 38: Change in capacity is known to be assumed. However, it can be measured. Not implemented in simulations yet.
   i. Slide 48: PAUSE is used to avoid packet losses, however PAUSE occurrences are reduced.

4. Bruce Kwan:
   a. FECN: http://www.ieee802.org/1/files/public/contrib/au-sim-kwan-ding-prelim-fecn-orlando-070314.pdf
   b. Slide : Need to resolve differences in Prof. Jain's and Bruce's results for low Qeq.
   c. BCN: http://www.ieee802.org/1/files/public/contrib/au-sim-kwan-ding-delay-effects-bcn-orlando-070314.pdf
   d. Associating w with Qeq seems to be improving results
   e. QCN (Jan 2007 version): http://www.ieee802.org/1/files/public/contrib/au-sim-kwan-matthews-qcn-orlando-070314.pdf
      i. Fast recovery mechanism enhancement is definitely required
      ii. Quantization effect can be seen in slide 11
      iii. Over sampling can improve convergence time and packet drops
5. Mitch Gusat:
   a. http://www.ieee802.org/1/files/public/docs2007/au-sim-IBM-ZRL-E2CM-proposal-r1.09.pdf
   b. Q: Is mice really reflecting low-rate flows? A: Yes. Not reflecting size of flow, but rate. (Comment: other analogy for this is: turtle-cheetah as compared to mice-elephant).
   c. Q: Signaling overhead? A: Yes, additional signaling on the top of baseline proposal. Don't have actual numbers yet.
6. Discussion:
   a. AdHoc Calls: Continue for now till people have chance to internalize solutions and discuss.
   b. End-node implementors should discuss how much practical proposals are.
   c. Comment/Q: How many RLs typical end-nodes will implement?
      i. Discussion: depends upon application. E.g. typical server sources will have less # of RLs than e.g. IPC server nodes.. 6-8 RLs seems reasonable.
      ii. General agreement that end nodes will not implement several 100s of RLs
   d. Interim Meeting:
      i. Presentations on implementation complexity
      ii. Presentations on User Complexity
   e. We should be using reflector more.
   f. No meeting tomorrow.
7. Meeting adjourned.