IEEE 802.1Qau January 2007 Interim Meeting Minutes

Recording Secretary – Manoj Wadekar

Wednesday 01/24/07

1. Pat Thaler: Agenda Discussion

2. Pat Thaler: Review of CN Objectives and Schedule
    a. Objectives discussion:
        i. Fairness, need to support multiple speeds, how long should congestion persist - there is no objective currently
    b. Schedule discussion:
        i. Chair concerned about schedule. Simulation team working hard - but plenty work TBD. Message and Tag formats yet to come.

3. Manoj Wadekar: Simulation AdHoc Report
    a. Need to add topology and workload description for "required sims"
4. Prof. Raj Jain: FECN Proposal and Simulation
    a. http://www.ieee802.org/1/files/public/docs2007/au-jain-fecn-20070124.pdf
    b. Q: Is there any specific quantity in mind for fairness, quick convergence?
    c. A: Will come up during discussion
    d. Q: Does Rx need to associate rx flow to tx flow for returning information to source?
    e. A: Yes, if there is no packet waiting to go back to source, then control packet to be generated
    f. Source always sends RLT with rate=-1 (infinity). Switch inserts appropriate rate.
    g. Q:Alpha works for steady rate, but for bursty traffic alpha can bring second degree effects: A: Scheme works for bursty traffic very well
    h. Tau=T is measuring sample time, not RTT. Few mS.
    i. Results for TCP only traffic: CM from TCP and FECN are working simultaneously.
    j. No PAUSE is enabled in the network.
    k. Foils need to be updated to the website.
    l. Q: What is Tps? A: Transactions per seconds : Need to be reflected back to foils
    m. Link latencies used in simulation = .5 uS; Switch and node delay is 1uS.
    n. Receiver may need longer delay to turn around. A: Few uS will not make much difference.
    o. Convergence is quick.
    p. Discussion: Dependency on initial rate R0. A: Yes. Administrator will set N. Switch starts with R0 = C/N.
    q. Notification is received before causing damage (Congestion).
    r. Smaller T leads to faster convergence. Simulations have used T=.1mS
    s. There is no packet by packet computation. Every T: compute is done and packets are stamped.

t. Q: Is 32 bit field being used for getting rate information back? A: Think so. Not quite sure. It is possible to use some scaled information for rate feedback.
u. Foil 23: Each source starts at rate R0 - configured my network administrator. For simulations same rate is used at switches as well.
v. Q: What R0 is used? A: C/200
w. Higher R0 - will need more buffering to avoid packet drops.
x. Slide 25 & 27: No PAUSE required: Clarification from Raj: This is for the given scenario where appropriate R0 is used.
y. Slide 28: PAUSE should not be used as primary mechanism - but only as emergency mechanism. Comment: Scheme needs to be working in case where PAUSE is in effect in part of the network
z. Tau=1mS; Sources
aa. Discussion: When does a flow return to R0? There is sampling every Tau secs.
    i. NICs don't know when does the flow start.
    ii. A: Yes, we need to address this ambiguity of timer in next rev.
    iii. Tau could be made T/2
    iv. It will be interesting to capture number of active sources vs time and also number of changes vs. time
bb. Q: Foil 32 - Behavior when sources change from R0 - why are they not going to rate declared by switch - they seem to be changing with DeltaR. A: Will check with student and get back.
cc. Typo on Foil 34: should be changed 0.05mS to 0.05 S etc.
dd. Capacity is measured from the idle time and bits transmitted
ee. Q: Need for no PAUSE - is it qualified statement?: A: Yes, R0 and Rdelta needs to be cfg'ed. But should be easy enough to provide generic configuration that applies to most deployment.
ff. More discussion tomorrow..

5. Mitch Gusat: Zurich HS Benchmark: BCN Sensitivity Analysis II
    a. http://www.ieee802.org/1/files/public/docs2007/au-sim-ZRL-ZHB-sim-results-Monterey-r04.pdf
    b. Sensitivity analysis : Ps, Gd and Gi are interesting - in that order
    c. Q: How long was PAUSE generated? A: Graph is not capturing clearly. Will start capturing stats next time.
    d. Q: What is partitioned memory? A: Shared memory was partitioned to maintain stats for input port utilization.
    e. Q: Part of the PAR was not to change output queued model. A: Queueing is not changed, only bookkeeping. But, agreed that concept of purely output queued model is polluted.
    f. Foil 33: What is the flow that is seen between Switch 2 & 1? A: Looks like mistake. Will correct.
    g. For difficult traffic pattern like moving congestion (Sweeping Hotspot): overdamping with Gd does not seem to help.
    h. There is no bursty traffic added to simulations yet.
    i. Q: What was the desired Qeq? A: 150K
    j. Need to simulate large no. of sources.
    k. Need agreement on priority of metric: Aggregate throughput, fairness, FCT etc.
    l. In DC, individual flows don't matter as much as aggregate throughput.
    m. Comment: If flows didn't matter - then disparity will be great.
    n. Need to agree on baseline routing algorithm - could be simpler STP - but need to agree on it.

    o. IBM research team is trying to break BCN for last 6 months. So far no inherent flaws found. Protocol overhead is negligible.
    p. Possible enhancements: Sampling, parameter setting.
    q. Q: Few changes on foils compared to Web. Can you re-upload? A: Yes.

6. Mick Seaman:
    a. Few goals:
        i. Try to operate w/o pause
        ii. Exactly what switch needs to do, how easy it can be expressed
        iii. How functions are distributed across Source, Switch, Destination
    b. If switch samples: Since it does not have flow information, damage needs to be done before getting information back to source
    c. Packet goes fwd, picks up congestion info, goes back to source.
    d. Dest can send rate report frame back that switch can stamp with rate information - assuming symmetrical path, explicitly labeled paths.
    e. Q: Do you send it to everyone? A: No, only to sources that talk to you.
    f. This allows switch not to generate a packet. It just stamps packets going from dest to source.
    g. Packets never get stamped in forward direction. Does not touch data packets. Only control packets.
    h. Switch looks at packet which port came from, uses priority value from the packet, and stamps with congestion information for associated queue for reverse direction on the same port.
    i. Q: What do you not like about tagging in fwd path? A: Not much, but avoiding will be good.
    j. Comment: Symmetrical assumption is problematic. Will have problems with link aggregation.
    k. Q: would it work if instead of tagging in fwd direction send control packet in fwd? A: It will be much more interesting if single bit in fwd direction triggers backward control packet
    l. This is not finer polished scheme. Will send more text later to describe in details.
    m. Q: Raj's scheme has large protocol overhead. How does yours improve it? A:Not using more than .5%
    n. Q: How does source associate Rate to flows? A: based on source of the control packet received.
    o. Q: How about priorities? Other dimension? A: For these types of traffic will be run only on 1 or 2 priorities. Could be multiple packets for each priority or single packet with priority values.
    p. Q: How about L2 multipath? A: There are multiple options. Sweeping over destinations, following multiple, deterministic paths.

Here is detailed proposal from Mick:

Rate Reports

The basic idea is that,

as part of a Congestion Avoidance algorithm comprising three sets of algorithms for

Sources, Bridges, and Destinations, the Destination originates and transmits regular Rate Report (RR) frames to each active Source. Each RR traces the reverse path from the Destination to the Source and carries an advertised rate for use by the Source in transmitting to that Destination. The RR originally carries a rate set by the destination to be its receiving link speed. At each Bridge Port, if the rate that that port wishes to advertise for the S->D direction is lower than the RR rate, the latter is replaced in the RR frame

by that lower rate.

The principle purpose of this mechanism is to improve control loop feedback, by ensuring that feedback is received regularly for all destinations, and to allow the feedback to be provided as a potential conversation starts, instead of relying on a statistical chance of sampling a forward going frame to trigger the feedback. The latter (Bridge sampling) naturally means that traffic has to be sent at a high rate simply to improve the chance of feedback, which essentially means increasing the chance of congestion and loss is necessary to gain the feedback to avoid it. This does not seem optimal. The overall goal is to reduce the overall control loop delay and provide early feedback to the point that no additional link level mechanism (especially PAUSE) is required to achieve acceptably low loss probability.

The Destination algorithm is to generate RRs every so many received bytes on an 'active' connection, and to generate an initial RR when a connection transitions from 'idle' to 'active'. By a connection in this sense I mean a particular {SA, DA} tuple, and such tuples are created a 'soft state' at Source and Destination in response to the frame flow. The overall RR generation algorithm is to be chosen to have an overhead of less than 1% bandwidth. As a first cut at the Destination algorithm, the 'idle' to 'active' transition occurs when two frames from the same Source are received within some number of mfts (maximum frame times - i.e. the time taken to transmit a maximum sized frame). The 'active' to 'idle' transition occurs after a time elapses with no reception , and the RR is sent about every 10 mf bytes (i.e. 15 Kbytes) when the connection is active. That's about a 0.5% overhead. I think it is likely that RRs could be sent less frequently but haven't tried lesser numbers yet.

The Source algorithm also treats connections as 'idle' or 'active', with an 'idle' connection being one for which no recent RR has been received. A low rate is associated with an 'idle' connection (perhaps 5 Mb/s = 1 max frame per 200 mfts on a 1 Gb/s link), and the rate is only updated when an RR is received. So a new or idle connection receives a low rate for the first few frames, which then stimulate the generation of an RR which reports the rate for the link. The Source rate is then increased towards that reported rate, with the RR rate diminishing as the new connection receives its share. The same RR rate is advertised to all sources, although any given source can behave as a number of (or as a fractional) virtual source.

I have only considered this sort of algorithm in terms of reporting rate feedback so far, though it is possible that the same idea of providing more predictable feedback per connection or conversation is applicable to feedback couched in other terms, and that the most important aspect - that of providing timely feedback on conversations that are just starting to be active so that they do not have to damage the network by injecting excess traffic to get congestion reports - may also be transferable.

The

Bridge Port

can calculate the rate to be placed in the RR packets (or other information as appropriate) periodically or upon some reasonably infrequent stimulus that does not require the RR to be updated with information that has only become available just as the RR is received. This simplifies the RR update process to one for checking the (reserved)

Ethertype for the RR, comparing the rate with that held by the Bridge Port, and overwriting if required. No new frames are injected into the stream of frames processed by this mechanism, and the precalculated rate held by the port can be that appropriate to the sum of the ports in a link aggregation, I believe that should simplify the Bridge/Bridge Port architecture as compared to an architecture that requires rapid injection of frames into the stream.

7. Bruce Kwan: Improving transient response time
    a. http://www.ieee802.org/1/files/public/docs2007/au-sim-kwan-transient-duration-012407.pdf
    b. Slide 8: BCN params same as baseline
    c. Comment: It will be good idea to capture overhead when increasing sampling rate.
    d. BCN-MAX kind of enhancements are very important for reducing transient duration
    e. If this parameter is important, we should take an objective for a number for consistent expectation.
    f. Characterizing transient in terms of RTT is a good idea. Can be used for other simulations as well.
8. Davide Bergamasco: Comparing BCN and B/FECN
    a. http://www.ieee802.org/1/files/public/docs2007/au-bergamasco-bcn-ecn-comparison-jan-2007-interim-v0.1.pdf
    b. Compared BCN and BECN results from Raj's presentation. BCN results are significantly worse in Raj's presentation than baseline param-based results in Davide's simulations
    c. Also BECN aggregate throughput is significantly lower due to overheads
    d. Comment: Raj commented that he proposes FECN and will not defend BECN.
    e. Discussion: Significant disconnect on how BECN worked. (Whether Broadcast was used for BECN - Raj clarified that it was not). Suggestion is for Davide to add a foil describing BECN simulated as "his understanding of Raj's scheme"
    f. Comment: BCN graphs are improved, but takeaway of low fairness stands. A: Yes. However, by increasing overhead for fair comparison - can show much improved fairness behavior.
    g. Comment: FECN proposal being presented is not correct - actual proposal was presented only today. So the presentation is not fair.
    h. Comment: Raj owes to group to provide baseline description of his protocol and parameters.
    i. Q: Is BCN protocol well documented? A: Baseline protocol and params are well understood and documented. There are few variants like BCN-MAX, oversampling to address transient congestion
    j. Comment: TF should discuss plan going forward.
    k. Q: Can Raj provide pseudo code for FECN today? A: No, still working on small details, will provide soon.
9. Davide Bergamasco:BCN in a large(r) topology
    a. http://www.ieee802.org/1/files/public/docs2007/au-sim-bergamasco-bcn-large-topologies-jan-2007-interim.pdf
    b. Goal was to build fat tree topology - but was aggressive. So, reduced to a tree topology with ES 16-256
    c. Emulating fat tree with switch with very fat pipes (imaginary) to avoid second hot spot.

    d. Difference from Baseline params: Sampling rate is being used as 2%, Qeq was increased to 48KB
    e. Pretty significant initial transient was experienced, hence BCN-MAX was used
    f. Q: What is larger topology? A: Number of nodes. This stresses BCN mechanism.
    g. Slide 7: Surprisingly
    h. throughput presented is average - need to reflect on slides
    i. Comment: Queue size is limited by PAUSE. Transient time is limited to ~50mS with use of BCN-MAX.
    j. Baseline params seem to be working really well. Need to simulate more to find the breaking point.
    k. Comment: Higher speed links will mean more aggregation of flows and can create challenge.
    l. Comment: HS degree should be increased to see whether any problem is being hidden. To make it faster for simulations - may be # of victims can be reduced.
    m. Request for all to participate in simulation ad-hoc and provide inputs.

10. CM capability exchange and discovery - Manoj Wadekar, Intel
    a. Not proposing a mechanism. Just letting us know one of the problems that we need to focus on.
    b. "CM cloud" is formed by CM-compliant bridges and end-stations.
    c. CM and non-CM traffic can co-exist on the same infrastructure differentiated by traffic class.
    d. CM cloud members need to ensure consistency of CM parameters within the cloud.
    e. CM frames/headers need to be filtered out while exiting the CM cloud.
    f. Need to think of BCN and FECN since they impose different requirements on this.
    g. Leverage LLDP for discovery.
    h. May create a plug and play challenge - perhaps look at how STP creates its regions.

<u>Thursday 01/25/2007</u>

11. Mitch Gusat: LLFC:
    a. http://www.ieee802.org/1/files/public/docs2007/au-ZRL-Ethernet-LL-FC-requirements-r03.pdf
    b. Q: Which application would like to have Reliable delivery with LL retransmission? A: Not in Storage, but with clustering applications where end-to-end retransmissions will be more expensive than link level retransmissions. However, there is no specific position being taken - merely stating possible requirements.
    c. Q:Foil16: Self-induced underflow - assumes shared buffer, correct? A: Yes. Canonical representation. Not one with better buffer schemes. Discussion: Proper credit based or Priority-PAUSE implementation can avoid this.
    d. Q: Foil 16: Push-Thru Blocking: What is IA? A: Input Adapter
        i. Q: Is this describing input queue blocking? A: Yes. Discussion: This is switch design issue. Not LLFC.
    e. Foil 17: Just queues per priority don't solve the problem - but better buffer design is required.
    f. Just increasing number of VLs or priorities does not solve HOL blocking

g. Foil 24: VOQ-Selective LL-FC: Comment: This is second order HOL blocking and switch does not have information about where the packet gets forwarded in the next hop neighbor. Hence it is second order problem.

h. Comment: Such protocols increase visibility one hop more than normal LLFC, but cost is very high due to large number of queues, database sync etc.

i. It is better to cover HOL blocking by End-to-end CM and solve transient congestion with LLFC.

j. TF has decided to support "controlled domain" and "not controlled domain" - so any LLFC mechanism needs to be domain-aware.

k. Lot of discussion on foil 32..

l. Foil 32: Example here is PCI: Posted Write need to be able to bypass Read.

m. Discussion: LLFC can create deadlocks - and it is important to elaborate and understand them completely.

n. For large RTT: Credit is more memory&latency efficient as compared to Grant mechanism

12. Discussion for March Plenary

a. TF members are required to submit presentation week before meeting at the latest.

b. TF will meet 2.5 days during Plenary meeting

13. Prof. Balaji Prabhakar: An Overview and a proposal

a. http://www.ieee802.org/1/files/public/docs2007/au-prabhakar-monterey-proposal-070124.pdf

b. Unit Step Response and Real Time Simulations - are two aspects of framework for analysis of any CM mechanism

c. Foil 21: 3rd equation has a typo for g(qk): needs to +ve. Should be:

i. Rk+1 = Rk +C - Ak **+** g(qk)

d. Question: Is the equation representing only one user? A: Does not matter. Coment-Raj/Discussion: It is important to have N in the equations.

e. Latest Raj's ECN (yesterday) was simulated and presented.

f. Slide 24: ECN (latest) shows oscillatory behavior loosing packets at top and link utilization at bottom.

g. Comment: Queue needs to be maintained at low value during long lived flows to consume bursty traffic.

h. Question: Why different params were used? A: Since no sensitivity to param is claimed why does it matter? Discussion: To avoid too many PAUSEs, specific params need to be used.

i. Qeq affects the stability of the queue. Comment-Raj: Need to define stability

j. Comment: For scientific study all the parameters need to be stated. Generic statement "loop is always stable" is not correct.

k. Comment: Buffer utilization, transport delay affect stability of control loop. It is important to know what are the boundaries.

l. 3 things different: sampling time reduced to 1 ms, Avg of rate over 2 intervals, switch at its discretion can limit its increase to a threshold (optional).

i. Nov. a,b,c parameters are used for these simulations (Comment-Raj: Need to use Jan parameters)

ii. Alpha = 0.5

m. Comparison of various Fairness schemes: ECN fits into Max-Min Fairness category

n. Max-min fairness algorithm is not practical due to need for global information: From J. Mo and Walrand (1998)

o. Measurement interval:
   i. Can't be too short - it will be too noisy, also can't signal too many flows
   ii. Cant be too long - Not responsive enough, needs more buffers

p. Less than 32 bits for rate information - quantization; Can lead to under/overutilization.

q. Comment: 32 bits is smaller than BCN's tag requirement. However, quantization effects need to be studied.

r. ECN scheme does not allow different sources to transmit at different rates.

s. Comment: Different rates for sources could be added by introducing Weights and RLQid Discussion: This will lead to flow level knowledge in network which will be unacceptable. Counter argument: It will be not required… Anyway since this is not a scheme on table..

t. Comment-Raj: T is not critical. But it is a parameter. Could be made dynamic in future rev.

u. Comment: All the implementations need to be interoperable. So, it will not be acceptable to have solutions that have different behaviors for different devices.

v. Comment: Need to differentiate between variable and parameter

w. Proposal for new simple algorithm - enhancement to BCN
   i. Reduce signaling traffic- compress and quantize
   ii. Let source increase rate - switch only sends decrease signals
   iii. Foil 41: Should be interesting to see how this works with no increase signals

14. Discussion for various proposals:
    a. Objectives discussion: Pat will post file to reflector
15. Discussion on Raj's presentation:
    a. Q: Is the presented proposal final? A: It could change to address issues raised? There is going to be a change for limited rate increase for sure. When there is a change - it will be announced on reflector/ad-hoc etc.
16. Ground rule should be to share information openly, clearly and quickly.
    a. Everybody agrees to start with agreed params and then experiments beyond.
    b. There should at least be a foil-set with description etc.
17. Davide will post a presentation with all the BCN information together.
    a. This will be tracked with version numbers
18. Topologies and workloads need also be well documented and tracked with version numbers
19. Should consider running AdHoc for 2 hours. Possibly start at 9.00am'
20. Meeting Adjourned.