BCN 2.0: Issues and Questions

Bruce Kwan (Broadcom Corp)

IEEE 802.1 Interim Meeting
Beijing, China
May 16, 2006

BROADCOM
Connecting everything

Overview

- Goals
- Protocol Issues
  - Congestion Detection
    - Frame sampling
    - Queue depth units
    - BCN Message Generation
    - Congestion Detection Trigger
  - Response Function
    - RL queue management
    - Handling multiple congestion points
    - Slow Recovery from Congestion
  - Frame Formats
    - RL Tag
    - BCN Message
  - BCN Parameters
- Simulation Issues
- Further Performance Studies

BROADCOM
Connecting everything

Goals

- Identify open questions regarding the BCN 2.0 protocol
- Identify issues and questions surrounding the simulation results
- Identify further work for performance verification
- Intent is to motivate further clarification of detailed operation of the BCN 2.0 protocol and verification of its performance

BROADCOM
Connecting everything®

Overview

- Goals
- Protocol Issues
  - Congestion Detection
    - Frame sampling
    - Queue depth units
    - BCN Message Generation
    - Congestion Detection Trigger
  - Response Function
    - RL queue management
    - Handling multiple congestion points
    - Slow Recovery from Congestion
  - Frame Formats
    - RL Tag
    - BCN Message
  - BCN Parameters
- Simulation Issues
- Further Performance Studies

BROADCOM
Connecting everything

Frame Sampling
Sampling Overhead

- 1 in 100 sampling overhead isn't bad on a link with max size packets but it seems high for heavy min size packet loads:
  - Time between samples (@10Gbps)*
    - 1500 byte packets: 120 us
    - 64 byte packets: 5.12 us
  - BCN Message Loading (on 10Gbps link)
    - Network load 1500 byte packets: 0.043%
    - Network load 64 byte packets: 1%
- Should a mechanism be put in place that allows the user to specify a maximum bandwidth overhead that would be dedicated to BCN messages (i.e. a time since last sample test)?
  - May impact control loop delay
  - Requires further study.

** Assumes congestion is present such that BCN messages are being sent*

BROADCOM
Connecting everything

Frame Sampling
Meaning of the Sampling

- BCN 2.0 proposal states that incoming frames are sampled with a probability P
- Does it really mean this?
  - Sampling occurs as if for each incoming frame a random number is generated with uniform distribution from 0 to 99. If 99 is generated, the frame is sampled.
  - The implication is that samples can occur as close together as back to back frame arrivals (= P^2)
- Or, is the intent something else such as a jittering of the arrivals between samples with an average value of 100 but a smaller standard deviation.

BROADCOM
Connecting everything

Frame Sampling
Effects of Sampling on Q_delta

- Over what period is Q_delta measured?
  - Change since the last probabilistic sample
  - Change during the past n microsecond
  - Change over the last n packet arrivals
- Change since the last sample is easiest to implement but makes the value difficult to interpret since it can be measured over very different time periods (and depending on the sampling probability operation this could be over very different number of arrivals)
  - A small delta value may mean the queue level is changing slowly or that time since the last sample was short (or even very short).
  - A large delta value may mean that the queue level is changing quickly or that time since the last sample was long
- Change since last sample and change over n arrivals mean it is a derivative with respect to the arrival rate and not with respect to time.
- The BCN 2.0 analysis from September appears to treat this as a derivative with respect to time

BROADCOM
Connecting everything

Overview

- Goals
- Protocol Issues
  - Congestion Detection
    - Frame sampling
    - Queue depth units
    - BCN Message Generation
    - Congestion Detection Trigger
  - Response Function
    - RL queue management
    - Handling multiple congestion points
    - Slow Recovery from Congestion
  - Frame Formats
    - RL Tag
    - BCN Message
  - BCN Parameters
- Simulation Issues
- Further Performance Studies

BROADCOM
Connecting everything

Queue Depth Units

- As an indication of congestion, the BCN 2.0 proposal defines queue utilization at the congestion point in units of ***packets*** as a measure of congestion.

- Most switch architectures for this market *do not* allocate buffers proportional to packet size (i.e. not fixed space per packet)
  - On a 10Gbps link in 10ms
    - If packet size = 64 bytes, can receive ~149,000 packets
    - If packet size = 1518 bytes, can receive ~8,130 packets
  - Buffer Allocation Block = Max Size Packet
    - If buffer is allocated in max packet chunks, buffering to hold 10 ms of 64 byte packets would require ~215 Mbytes of storage (149,000 * 1518 bytes per packet buffer)
  - Buffer Allocation Block = 16-128 bytes
    - If buffer is allocated in small units, buffering to hold 10 ms of 64 byte packets would require ~9-18 Mbytes
  - Therefore, buffer space is often allocated in small chunks to allow efficient utilization under small packet loads.

BROADCOM
Connecting everything

Queue Depth Units

- **Ethernet switch implementations use varying architectures for buffer allocation.**
- **Protocol should support accurate report of queue utilization across a variety of implementations.**

Queue depth in packets is not a good measure of of buffer utilization.

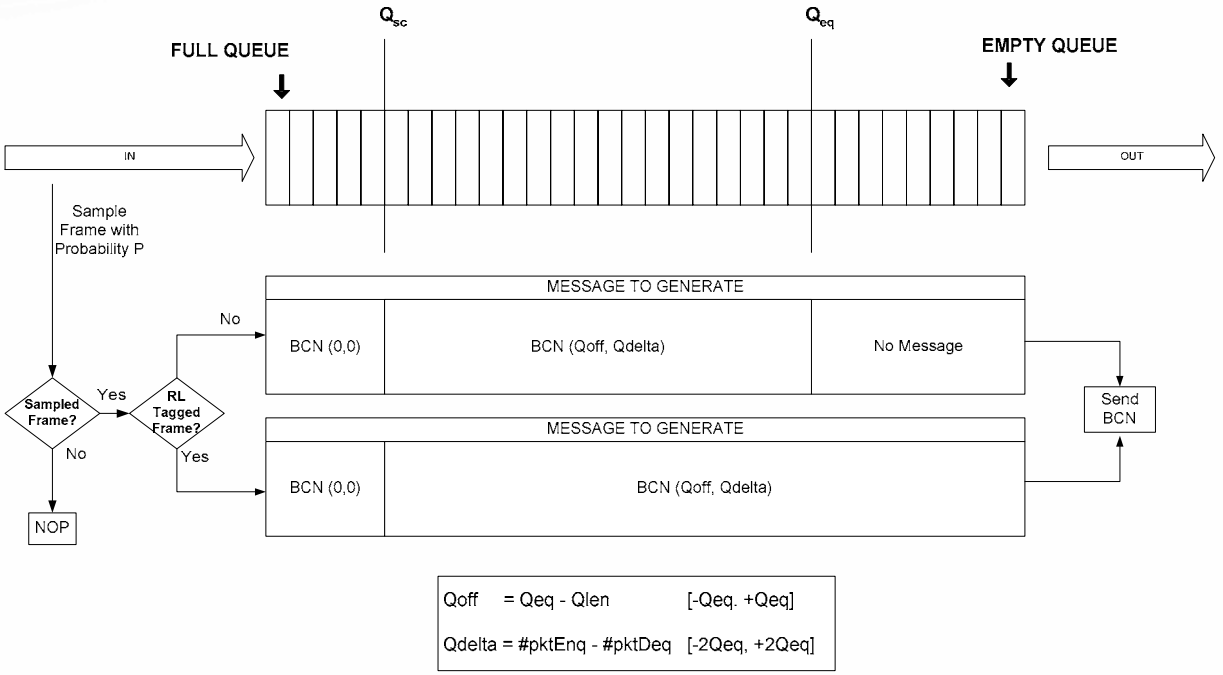
BROADCOM
Connecting everything®

Overview

- Goals
- Protocol Issues
  - Congestion Detection
    - Frame sampling
    - Queue depth units
    - BCN Message Generation
    - Congestion Detection Trigger
  - Response Function
    - RL queue management
    - Handling multiple congestion points
    - Slow Recovery from Congestion
  - Frame Formats
    - RL Tag
    - BCN Message
  - BCN Parameters
- Simulation Issues
- Further Performance Studies

BROADCOM
Connecting everything

BCN Message Generation



- At the congestion point (CP) when an RL tagged frame is sampled, does the CP only consider generating a BCN message if the RL tagged frame has a CPID equal to the local CPID?
- What happens when an RL tagged frame is sampled at a CP with a different CPID but the queue is larger than Q_eq?

BROADCOM
Connecting everything

Overview

- Goals
- Protocol Issues
  - Congestion Detection
    - Frame sampling
    - Queue depth units
    - BCN Message Generation
    - Congestion Detection Trigger
  - Response Function
    - RL queue management
    - Handling multiple congestion points
    - Slow Recovery from Congestion
  - Frame Formats
    - RL Tag
    - BCN Message
  - BCN Parameters
- Simulation Issues
- Further Performance Studies

BROADCOM
Connecting everything

Congestion Detection Trigger

- Two goals of congestion detection are as follows:
  - Fast detection of the onset of congestion
  - Avoid false detects
- In the current BCN 2.0 proposal, the initial trigger for generating BCN messages is based solely on the queue length and not the rate at which the queue grows.
- Using Q_eq
  - The choice of Q_eq affects how quickly congestion may be detected.
  - Current protocol definition drives the setting for Q_eq to be small in order to provide fast detection but it may lead to false detects
- Using Q_delta
  - During the onset of congestion, Q_delta will rise and F_b will become negative and lead to a congestion response (multiplicative decrease) before Q_eq is exceeded
  - The use of Q_delta should be considered to aid in faster detection of congestion

BROADCOM
Connecting everything

Overview

- Goals
- Protocol Issues
  - Congestion Detection
    - Frame sampling
    - Queue depth units
    - BCN Message Generation
    - Congestion Detection Trigger
  - Response Function
    - RL queue management
    - Handling multiple congestion points
    - Slow Recovery from Congestion
  - Frame Formats
    - RL Tag
    - BCN Message
  - BCN Parameters
- Simulation Issues
- Further Performance Studies

BROADCOM
Connecting everything

RL Queue Management

- When is the RL queue at the reaction point no longer used to rate limit the traffic that contributed to congestion?
  - Rate at the RL Queue == link speed &&
  - RL Queue is empty
- How is multicast handled?
- When there are not sufficient RL Queues at the Reaction Point to handle all of the current congestion points, what are the rules for managing these resources?

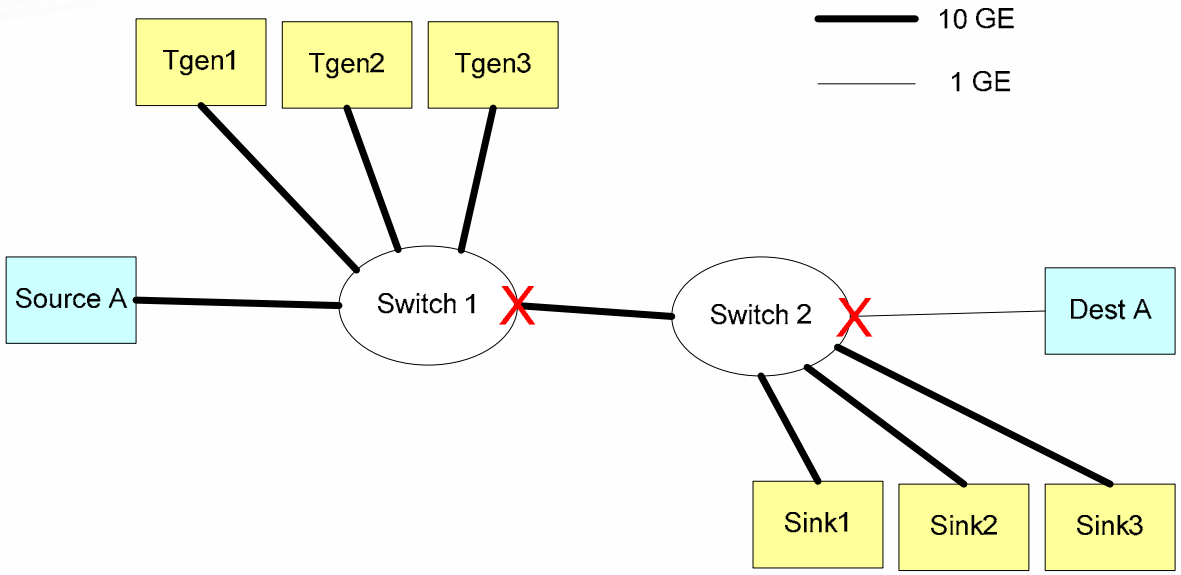
BROADCOM
Connecting everything

Overview

- Goals
- Protocol Issues
  - Congestion Detection
    - Frame sampling
    - Queue depth units
    - BCN Message Generation
    - Congestion Detection Trigger
  - Response Function
    - RL queue management
    - Handling multiple congestion points
    - Slow Recovery from Congestion
  - Frame Formats
    - RL Tag
    - BCN Message
  - BCN Parameters
- Simulation Issues
- Further Performance Studies

BROADCOM
Connecting everything

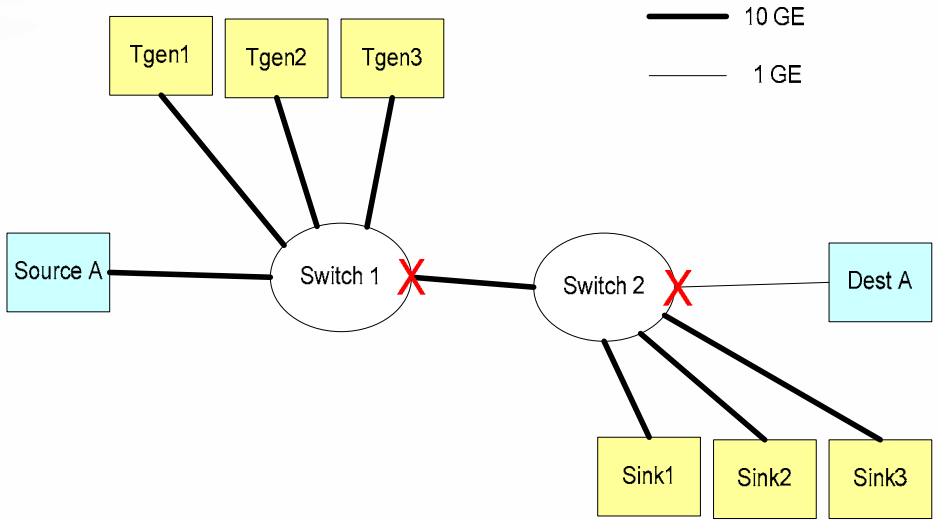
Handling Multiple Congestion Points



- TgenN sends to SinkN (N=1,2,3) in a bursty on-off pattern
- SourceA sends to Dest A
- SourceA is contributing to both congestion points using a single flow
- Ideally, SourceA would be able to detect the worst case congestion across the multiple congestion points and respond accordingly.

BROADCOM
Connecting everything

Handling Multiple Congestion Points



- **How should the Reaction Point in SourceA handle the case where it is receiving multiple BCN messages from different CP's for the same flow?**
  - **Which CPID goes in the RLTag?**
  - **Congested CP handling for a packet with a foreign CPID in RLTag?**
  - **Reaction point may receive conflicting BCN messages from the two different CPID's**

BROADCOM
Connecting everything

Overview

- Goals
- Protocol Issues
  - Congestion Detection
    - Frame sampling
    - Queue depth units
    - BCN Message Generation
    - Congestion Detection Trigger
  - Response Function
    - RL queue management
    - Handling multiple congestion points
    - Slow Recovery from Congestion
  - Frame Formats
    - RL Tag
    - BCN Message
  - BCN Parameters
- Simulation Issues
- Further Performance Studies

BROADCOM
Connecting everything

Slow Recovery from Congestion

- The rate at a Rate Limited (RL) queue in the reaction point may become very small.
  - Effect varies depending on status of solicit bit (discussed later)

- Rate may fall to 0
  - Occurs when hitting a Severe Congestion state in the CP
  - When this occurs, how does the rate limited queue become unfrozen given it will not be receiving any BCN messages to increase its rate?
  - If a timeout mechanism is used, what is the rate of the RL queue when the timeout expires?

BROADCOM
Connecting everything

Slow Recovery from Congestion

- **Example:**
  - Suppose the rate at an RL queue is 8Mbps
  - For 1500 byte frames, sending 100 frames requires ~150ms
  - This implies that after congestion subsides, it would take at least 150ms (on avg) to receive an initial BCN message directing the RL queue to additively increase its rate.
  - This seems like a long time

- **Is there support in the protocol to mitigate these effects?**

BROADCOM
Connecting everything

Overview

- Goals
- Protocol Issues
  - Congestion Detection
    - Frame sampling
    - Queue depth units
    - BCN Message Generation
    - Congestion Detection Trigger
  - Response Function
    - RL queue management
    - Handling multiple congestion points
    - Slow Recovery from Congestion
  - Frame Formats
    - RL Tag
    - BCN Message
  - BCN Parameters
- Simulation Issues
- Further Performance Studies

BROADCOM
Connecting everything

RL Tag Solicit Bit

- **Status of the Solicit Bit is unclear. It appeared in the appendix of the 9/2005 BCN 2.0 presentation.**
  - **Appeared to address the issue of slow recovery from congestion**
  - **Solicit bit isn't shown in RLT Tag format**

- **If Supported:**
  - **Will there be rules on how often the solicit bit can be turned on to limit abuse?**
  - **There are times when many flows coming together result in many solicit bit frames even if transmitters use it responsibly. Should there be limits on how often solicit generates a BCN to protect reverse direction bandwidth?**
  - **How is R_solicit defined?**
  - **Does the solicit bit override the Fb sign test (non-linearity)?**

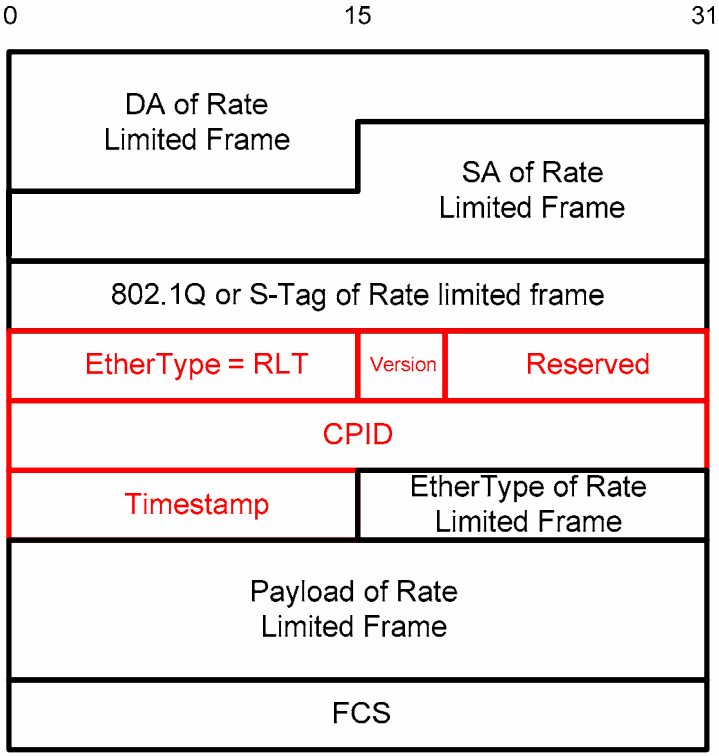
If R < R_solicit, set solicit bit in RL Tag

If solicit bit is on, CP will attempt to send BCN messages to the endpoint regardless of sampling

BROADCOM
Connecting everything

RL Tag Format (Alignment Issue)

- Viewing a frame as an array of 4-byte words, usually Ethertype is the first two bytes of a word.
- RLT Tag format puts the Ethertype of the rate limited frame in the last two bytes of the word which implies a two-shift of word alignment for the whole frame.
- This is awkward for some implementations inserting and deleting tags. The format should maintain 4-byte alignment of encapsulated frame.



RL Tag Format

BROADCOM
Connecting everything

RL Tag Fields

- There is a timestamp field in the RL tag.
  - What is this?
  - How is it used?

Flow Identification

- In the BCN Message, the first N bytes of the sampled frame is to be placed into the BCN message.
  - Issue
    - This implies that flows are defined as being resolved using (SA, DA, VLAN tag)
    - Fixed processing requirements to resolve the flow
  - Potential Fix/Enhancement
    - Could add a flow tag field to the RL tag that the end node can generate and use to identify the flow associated with the RL tag.
    - Reduces processing to associate BCN with rate limited queue vs. using SA/DA/VLAN tag

BROADCOM
Connecting everything

Overview

- Goals
- Protocol Issues
  - Congestion Detection
    - Frame sampling
    - Queue depth units
    - BCN Message Generation
    - Congestion Detection Trigger
  - Response Function
    - RL queue management
    - Handling multiple congestion points
    - Slow Recovery from Congestion
  - Frame Formats
    - RL Tag
    - BCN Message
  - BCN Parameters
- Simulation Issues
- Further Performance Studies

BROADCOM
Connecting everything

BCN Message CPID

- **How is the uniqueness of the CPID achieved?**
  - **Single administrative domain MUST ensure this?**

BCN Message
Sign of Qoff v.s. Qdelta

- Current definitions of Qoff and Qdelta
  - *Qoff = Qeq-Qlen, [-Qeq, +Qeq]*
  - *Qdelta = #pktArrival-#pktDeparture, [-2Qeq, +2Qeq]*
- With these definitions, Qoff and Qdelta have the opposite sense -
  - Qoff
    - Positive when the queue is underfilled,
    - Negative when the queue is overfilled.
  - Qdelta
    - Negative when the queue is emptying (Qoff is increasing)
    - Positive when the queue is filling (Qoff is decreasing)
- Qdelta is the inverse of the derivative of Qoff.
- It would be more intuitive to invert the sense of Qoff (adjusting Fb calculation for the change) so that they have the same sense.

BROADCOM
Connecting everything

Overview

- Goals
- Protocol Issues
  - Congestion Detection
    - Frame sampling
    - Queue depth units
    - BCN Message Generation
    - Congestion Detection Trigger
  - Response Function
    - RL queue management
    - Handling multiple congestion points
    - Slow Recovery from Congestion
  - Frame Formats
    - RL Tag
    - BCN Message
  - BCN Parameters
- Simulation Issues
- Further Performance Studies

BROADCOM
Connecting everything

Parameter Selection

- Are the BCN parameters link speed dependent?
- Q_eq may be challenging to choose optimally
  - May be implementation dependent as a function of the queuing capabilities of the CP
  - Qeq is significant to performance since it bounds Qoff and Qdelta
    - |Qoff| ≤ Qeq
    - |Qdelta| ≤ 2*Qeq
    - For W = 2, |Fb| ≤ 5 * Qeq
  - Congestion Points (CP's) with a large Q_eq trigger a larger response (slows down more quickly) at the reaction point relative to CP's with a small Q_eq
  - Jain suggests using Gd = 0.0124 to prevent negative R in his March 2006 presentation
    - This works for Qeq ≤ 16
  - However, the end node won't know Qeq value for CP's and shouldn't have to adjust calculation per CP configuration or implementation value

BROADCOM
Connecting everything

Overview

- Goals
- Protocol Issues
  - Congestion Detection
    - Frame sampling
    - Queue depth units
    - BCN Message Generation
    - Congestion Detection Trigger
  - Response Function
    - RL queue management
    - Handling multiple congestion points
    - Slow Recovery from Congestion
  - Frame Formats
    - RL Tag
    - BCN Message
  - BCN Parameters
- Simulation Issues
- Further Performance Studies

BROADCOM
Connecting everything

Simulation Setup Clarification

- What were the values for Q_eq for each simulation run?
- What was the value of R_u in the simulations shown in the latter half of the September 2005 BCN2.0 presentation?
- In the BCN simulations presented in 9/2005, does the model involve simulating the reaction point in the bridge or at the endpoints? My understanding is that it is in the bridge.
- For the long propagation delay scenarios, recreating the precise results appear challenging. Need to understand the reasons for this.

BROADCOM
Connecting everything

Simulation Setup Clarification

- What flavor of TCP is used?
  - Tahoe, Reno, New Reno, SACK
- What TCP parameters are assumed?
  - Window size
  - Slow Start
  - Quick Start
  - Nagel

BROADCOM
Connecting everything®

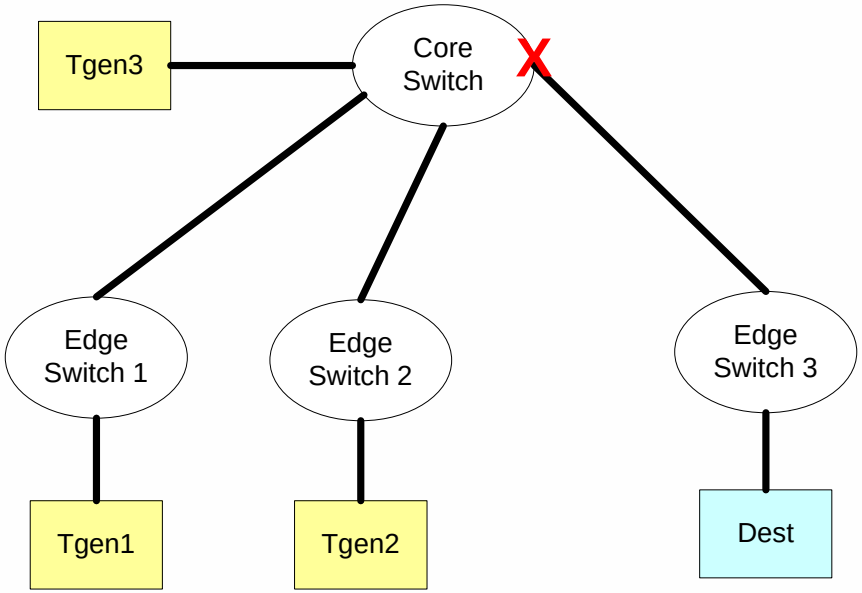
Overview

- Goals
- Protocol Issues
  - Congestion Detection
    - Frame sampling
    - Queue depth units
    - BCN Message Generation
    - Congestion Detection Trigger
  - Response Function
    - RL queue management
    - Handling multiple congestion points
    - Slow Recovery from Congestion
  - Frame Formats
    - RL Tag
    - BCN Message
  - BCN Parameters
- Simulation Issues
- Further Performance Studies

BROADCOM
Connecting everything

Performance Studies

- Asymmetric topology
  - Different RTT's
  - Different link rates on competing sources
- Mixed network speeds
- Multiple Congestion Points
- Fairness
  - 1 GE and 10GE sources competing for bandwidth at a congestion point
  - Fairness between sources at different distances from congestion point
  - Combination of both issues mentioned above



Example Asymmetric Topology

BROADCOM
Connecting everything

Performance Studies

- Processing Delays at End Node
  - Up to 500usecs
- Mixed packet sizes
- Effects of BCN message loss
- Effects of Severe Congestion events
- Bursty traffic scenarios
  - Transient bursts with no persistent congestion point. How much is system throughput degraded due to false detects?
- Effects of different Q_eq's at CP's
- Transient analysis of buffer utilization
- Consider alternate flavors of TCP
  - TCP Vegas
  - FAST
  - XCP

BROADCOM
Connecting everything