Ethernet Congestion Management

Brad Booth, Intel

September 2004

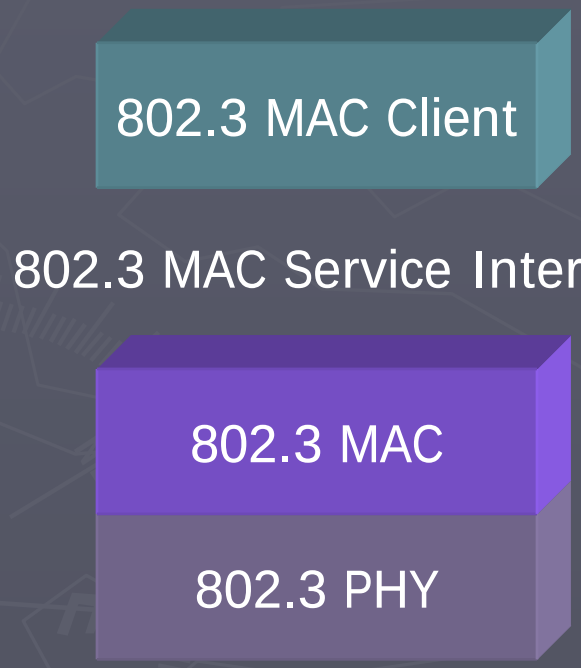
Issue

- “Ethernet not adequate for low latency applications”
- “Ethernet frame loss is inefficient”
- Markets impacted
  - Clustering and grid computing (RDMA, iWARP)
  - Storage (iSCSI)
  - Backplanes (802.3ap, ATCA)
  - Video (video over IP)
  - Telecom and voice (VoIP)
  - Others?

Market Need

- Decreased latency
  - Critical for storage and clustering
  - Reduces buffer requirements; therefore impacts cost of components
- Reduced frame loss
  - Important in all network applications
  - Prevent oversubscription with no latency impact
  - Improves performance of the system

Overview



- Oversubscription occurs in MAC Client
- 802.3x could assist, but halts all flows
- Preference is to keep 802.3 simple and rely on MAC Client to resolve

802.3x

- Pros
  - Proactive for oversubscription
- Cons
  - Halting all flows is not desirable
  - Removes control from upper layer protocols
  - Adds latency to all flows
- A feature no one uses

MAC Client

- Many varieties of MAC clients
  - 802.1 (bridging)
  - TCP/IP, UDP, etc.
- MAC clients are reactive
  - Wait for oversubscription to occur
  - Protocols force rate limiting when oversubscription occurs
  - Buffers used to prevent transient congestion from becoming oversubscription

The “Evil” Trade-off

- Buffers vs. frame loss
  - Frame loss is considered bad
  - MAC clients can control which frames are lost
  - Buffers decrease frame loss, add latency
- Trade-off
  - Increase latency and reduce frame loss
  - Or, reduce latency and increase frame loss
- No win situation for latency & frame loss sensitive applications

Solution

- Provide a means for MAC clients to be proactive
  - Decreases need for buffers
  - Reduces latency
- How?
  - 802.3x was close
  - Permit MAC clients to exchange congestion information via an 802.3 control messages

Value

- Opens up latency and frame loss sensitive markets to Ethernet
- Empower the Ethernet standards with support for improved congestion control in 802.3 L2 subnets
- Increase performance of MAC clients
  - Reduced frame loss decreases re-transmissions
- Decrease cost of Ethernet components
  - Reduction in buffer requirements has a direct correlation to cost of components and systems

Recommendation

- Change the “Objectives” to better align with this strategy
  - Thanks to Shimon Muller for his feedback on the “Objectives”

Current Objectives

- Focus solution to a single link only (hop-to-hop/end-to-end not specified)
- Specify a mechanism to limit the rate of transmitted data using a "pacing" algorithm (not a burst duty cycle)
- Specify the granularity of the rate limiter
- Specify a new MAC Control Opcode and parameter set to support exchange of rate control information
- Do not specify how the MAC Client generates these MA_CONTROL.requests nor how it responds to the reception of MA_CONTROL.indications
- Specify the response to the new MAC Control opcode's parameter set
- Work with other 802.3 activities on the "long standing inconsistency" between MA_DATA.requests and transmit_frame function call

Current Objectives

- Point-to-point links
- Specify a mechanism to limit the rate of transmitted data ~~using a "pacing" algorithm (not a burst duty cycle)~~
- ~~Specify the granularity of the rate limiter~~
- A mechanism to support exchange of congestion control information
- ~~Work with other 802.3 activities on the "long standing inconsistency" between MA_DATA.requests and transmit_frame function call~~

Proposed Objectives

- Support point-to-point links only
- Specify a mechanism to support the exchange of congestion control information
- Specify a mechanism to limit the rate of transmitted data
- *Preserve the MAC/PLS service interfaces*
- *Preserve the 802.3/Ethernet frame format at the 802.3 MAC Service Interface*
- *Support full duplex operation only*

Thank you!

Questions?