Congestion Management in Datacenter Networks

Manoj Wadekar (Intel)

May 12, 2005

Contributors

- Davide Bergamasco (Cisco)
- Paul Congdon (HP)
- Gopal Hegde (Intel)
- Hugh Barrass (Cisco)

Agenda

- Framework for CM
- Service Differentiation and 802.1p
- Congestion Management in L2 network
- Summary

Framework for Congestion Management

1. Increase Available Bandwidth

2. Service Differentiation: 802.1p can provide "template" identification

3. MIBs and Configuration: (SBM, RFC-MAP, etc.)

4. Congestion Management
   - Oversubscription Congestion: manage @ source
   - Transient Congestion: 802.3x link level flow control

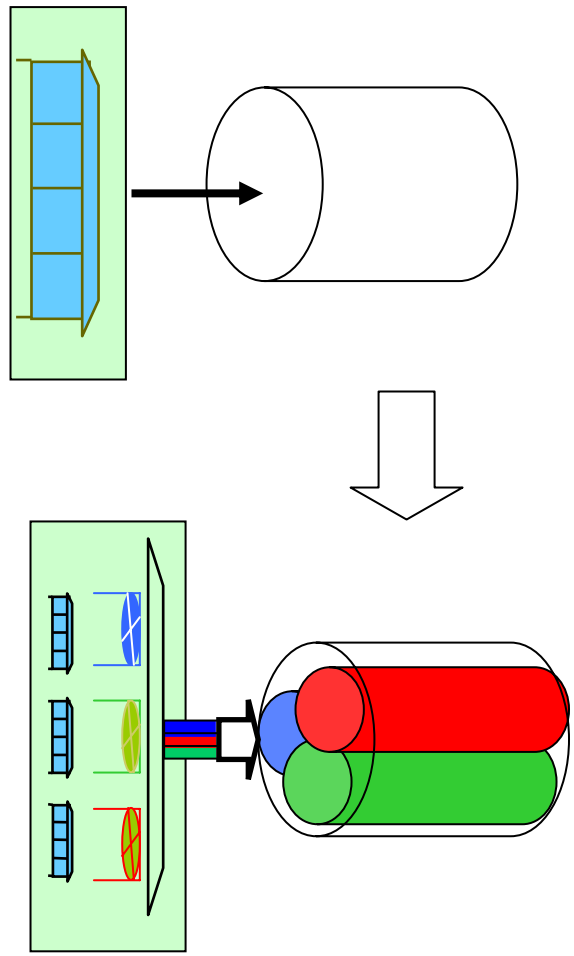
Enhancements for Datacenter Networks

1. Increase Available Bandwidth
   - Shortest Path Bridging
2. Service Differentiation: 802.1p can provide "template" identification
   - Queue draining behavior – more than Strict Priority
3. MIBs and Configuration: (SBM, RFC-MAP, etc.)
   - Standard MIBs and configuration
4. Congestion Management
   - Oversubscription Congestion: manage @ source
     - E.g. Backward Congestion Notification
   - Transient Congestion: 802.3x link level flow control
     - Buffer extension across a link
     - Needs per-priority flow control

Enhancements needed for Congestion Management in Datacenter

Service Differentiation and 802.1p

- 802.1p enables possible solution
  - Go beyond standard behavior – i.e. strict priority
  - Need to associate resources at bridges & end stations
- 802.1p provides 8 code points
  - Adequate for Datacenter applications
  - Discard Eligibility (PCP) can reduce number of available code points



Configuration in L2 networks

- SBM: provides RSVP based provisioning protocol for IEEE 802-style networks
- RFC2815: (RFC-MAP) defines mapping Integrated Services on IEEE 802 network
- How to:
  - Configure handling of “aggregate flow bundles”
  - Configure simplified specification (E.g. percentage rather than absolute value?)
- More comprehensive presentation from Paul Congdon (HP)

Congestion Management

- Oversubscription Congestion:
  - Congestion needs to be pushed to the ingress node for effective solution
  - L2 CM: notification from network, ingress rate control @ source
    - Backward Congestion Notification - Presentation from Davide Bergamasco (Cisco)
- Transient Congestion: Emergency insurance only
  - Link level flow control as insurance against packet drop
  - 802.3x causes all “templates” to be affected
  - Per-priority flow control can provide necessary granularity
    - Increase granularity in 802.3x PAUSE
    - Define Per-priority flow control in 802.1

Summary

- Specify “queue draining behavior” for 802.1p
  - configuration mechanisms, MIBs etc.
- Provide end-to-end L2 congestion management mechanism
- Granular link level flow control
  - Discussion in 802.3 and 802.1 required