Congestion Management Protocols
An Implementation Perspective

Guenter Roeck, Teak Technologies

Asif Hazarika, Fujitsu

May 29, 2007

Rationale

- Working group focus is on simulation and simulation results
- No or little focus on complexity or implementability

Outline

- Evaluate proposed CM protocols from an implementation perspective
  - Evaluation Criteria
  - Protocol Characteristics
  - Classify and evaluate protocols
- Identify issues with current protocol proposals
- Propose possible solutions

Protocol Evaluation Criteria

- Protocol should be
  - Simple
  - Elegant
  - Easy to implement
  - Flexible
  - Low overhead
  - No inherent (build-in) limitations/restrictions
- And of course it should do its job

4

Congestion Management Protocol Characteristics

- Data path handling
  - Tagging
    - Active tagging
    - Passive tagging
  - Non-tagging
- Feedback mechanisms
  - Forward Notification
  - Backward Notification

5

Tagging Protocols

- Protocol embedded into data flow (‘inband’)
  - Some protocol information is attached (tagged) to data packets
  - Higher forward path signaling overhead
    - ECM Tag for 10 packets: 14 * 10 = 140 bytes
    - Probe: 64 bytes
    - Applies only if all (congested) packets are tagged
  - Requires End-to-end protocol support
    - Endpoint has to understand tags
  - May require in-flow packet modification (‘active tagging’)
    - Requires checksum recalculation
    - May impact data packet latency
  - Must be missing something ...
- Examples
  - ECM, FECN, QCN

6

Non-tagging Protocols

- Protocol handling outside data flow ('Outband')
  - Unmodified data packets
  - Reduced forward path signaling overhead
  - End-to-end support not (necessarily) mandatory
  - Flows can benefit even if not completely within congestion controlled domain
  - Flow control (probe) packets can be sent at high priority
    - Improved reaction time
- Examples
  - E2CM

Feedback Mechanisms – Forward Notification

- Protocol covers entire flow data path
- Potential for reduced return path signaling overhead
- Endpoint calculations can improve protocol operation
- End-to-end support mandatory
- Endpoint implementation more complex
  - Needs to calculate and send response
  - May have to support per-flow status
- Slower reaction time
  - Feedback sent through reflection point (L2 endpoint)
  - Yet faster convergence (?)
- Examples
  - E2CM, FECN, QCN (partial)

Feedback Mechanisms – Backward Notification

- Covers part of flow data path
- May have higher return path signaling overhead
- No end-to-end support required
- Faster reaction time
  - Feedback sent directly to reaction point
  - Yet slower convergence (?)
- Examples
  - ECM, QCN (partial)

9

Protocol Class Specific Concerns

- Tagging Protocols
  - Reaction point and/or switch must know if entire flow is in CM domain
    - Reaction Point must classify using destination MAC address
      - How does it know if the destination is in the CM domain ?
- Forward Notification Protocols
  - RP and/or switch must know if entire flow is in CM domain
  - Impact of multi-speed links in path
  - Impact of congestion in reaction point and return path
- All
  - Filtering/handling requirements in CM domain edge switches
    - Incoming: Filter/handle packets w/ congestion managed CoS
    - Outgoing: Filter/handle CM control messages

Protocol Specific Concerns

- ECM
  - None besides generic tagging protocol concerns
- E2CM
  - Reflection point calculation complexity
    - Timestamp synchronization
    - Flow service rate calculation
  - Congestion not only determined by number of queued bytes
    - Switches can also be limited by number of packet descriptors
- FECN, QCN
  - Data packet tagging and modification
    - Requires data packet checksum re-calculation

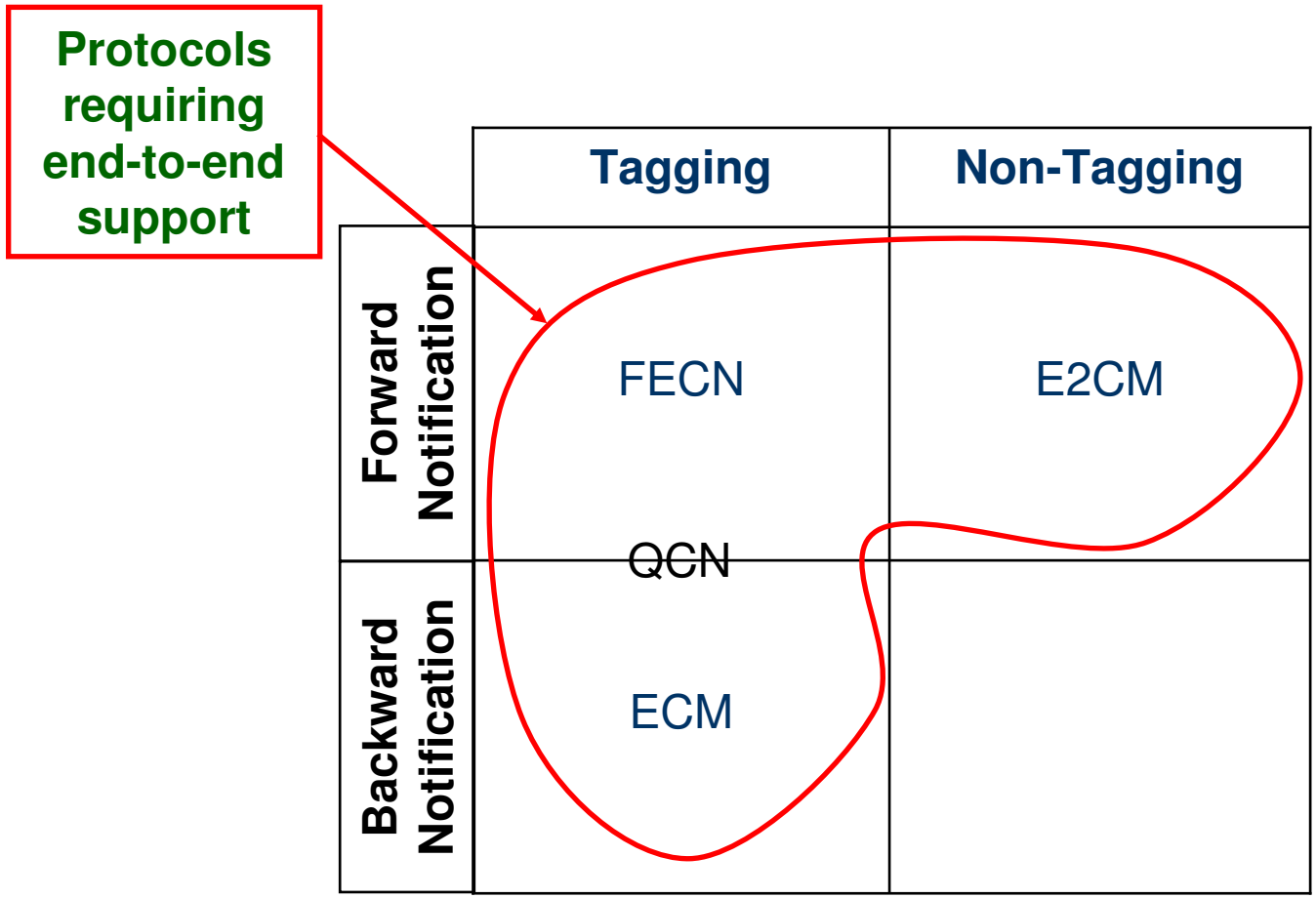
11

Protocol Characteristics vs. Evaluation Criteria

	Tagging	Non-tagging	Forward Notification	Backward Notification
Simple	O	X	O	X
Elegant	X	X	X	X
Ease of implementation	O	X	O	X
Flexible	O	X	O	X
No inherent limitations	-	X	-	X
Low overhead	O	O	X	O
Does its job	X	X	X	X

X: Advantage
O: Neutral
- : Disadvantage

Protocol Classification



13

Conclusion & Proposed Solution

- Conclusion
  - All currently proposed protocols have potential implementation issues
  - All currently proposed protocol proposals require endpoint (reflection point) support
  - Currently no proposal for a non-tagging protocol w/ Backward Notification
- Proposed solution
  - Modify existing protocols to remove tagging and use backward notification
    - Additional modifications to address implementation concerns
  - Validate/compare result against existing protocols
  - At the very least, address concerns in protocol specifications

- Use probes instead of tagged packets to solicit feedback for rate limited flows

15

Solution Details – E2CM

- Reaction point sends probes to congested switch
- CP responds to probes with CP queue length (in bytes)
- Reaction point uses CP queue length (instead of amount of queued data in network) to calculate new rate
- Possible variants
  - Send probes to ‘longest distance switch’ or ‘last switch in CM domain’
    - Need to determine address of this switch
  - Intermediate switches can add queue length to probe packets
  - Use elapsed time (probe sent -> response received) for rate calculations
    - Send probes in-band with regular data frames
    - Send replies as high priority packets to reduce return path latency

Solution details – FECN

- Send explicit probe packets instead of tagging data packets
- Send probe packets to congested switch (or to last switch in CM domain)
  - Still need to determine “last switch in CM domain”

17

Solution details - QCN

- Congested switch sends congested messages to reaction point, with Fb set to level of congestion
  - Similar to ECM / E2CM
- Reaction point performs rate changes as with current QCN
- Reaction point sends probes to congestion point
- Congestion point responds to Reaction Point with updated Fb
- Intermediate switches MAY update Fb
- Variants
  - Send probes to “longest distance” switch
    - Would require adding hop count into control packets

Solution details – Summary

- All modified protocols would use probe packets to solicit feedback
- Protocols vary in
  - Feedback parameters & calculation
  - Flow data rate calculation
- Possibility for converged protocol
  - Needs agreement on feedback parameters and rate calculation
  - Must standardize packet format and feedback parameters
  - Rate calculation can be implementation dependent
    - May be desirable to improve flexibility and enable vendor differentiation
    - May be undesirable to avoid unfairness

19

Issues

- More protocol variants to deal with
- Simulation coverage

20

Thank you

21

Backup Slides

Probe packet destination address

- Unicast
  - MAC address of congested switch
  - Multiple congestion points
    - Either send to worst congestion point, or to congestion point with max number of hops
    - Would require hop count in probe and feedback packets
  - Or send to L2 endpoint, and have edge switch filter for CM packet type
- Multicast
  - Create new “CM Probe” Multicast address
  - Packet contains MAC address of congested switch or flow endpoint
  - Switch packet forwarding rules
    - If embedded address (EA) is local address, or Congestion Point ID is local, terminate and handle packet
    - If EA next hop is outside domain or not a switch, terminate and handle packet
    - Otherwise forward packet to EA next hop port. If necessary/appropriate, update packet contents

23