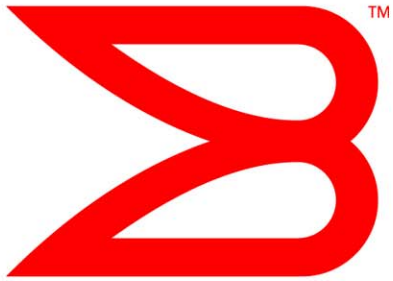
BROCADE



Making Flow ID Optional in 802.1Qau

au-ghanwani-flowid-1008-v1

Anoop Ghanwani
anoop@brocade.com

October 2008

Overview

- How did we get here?
- FlowID uses
- Problems introduced by the FlowID
- Proposal for optional FlowID

How did we get here?

- RPIDs were introduced in P802.1Qau/D1.2
  - Based on <au-nfinn-RPID-0508-v03.pdf> presented in May ‘08
  - Mainly needed for dealing with link aggregation
    - Avoiding fate sharing in the network
    - Processing of CNMs at the RP
  - If we find problems with it, we revisit the decision
- In July ’08 the issue was discussed extensively
  - RPID renamed to FlowID
  - Some other uses of the FlowID were discussed
- QCN was particularly attractive because it didn’t require any frame format changes
  - Should 802.1Qau use a new tag for data? (y/n/a – 0/14/7)
  - See <au-luijten-thaler-straw-polls-0705.pdf>

Uses for the Flow ID

- Multi-NIC link aggregation
  - In the absence of this, the CNM may have to be processed in host software adding to the reaction time
- Determining the RP that a CNM corresponds to without parsing the CNM
  - Original frame may have changed in transit due to additional headers put on by the network
  - When using fragmentation at the host, a fragment may not contain original transport port numbers
  - When doing things like, e.g., IPv6 over IPv6 tunneling the CNM may not contain enough data (depending on how much data is sent back)
- Determining which “flow” within an RP is the real cause of congestion
  - Need to extract information as above
  - At this point, seems to be the top reason cited in favor of FlowIDs

- Do these problems justify a new tag for *all* deployments?

Problems introduced by the FlowID

- 4 additional bytes to all frames
- Need to standardize a method of hashing based on FlowID so that switches and NICs agree on which FlowIDs are used on a member
  - This is needed to solve the multi-NIC link aggregation problem
- Need to modify end station link aggregation to deal with flow to FlowID assignments
  - May not be easy depending on OS
- Need to worry about stripping these tags off at the edges of CNDs
- In order to take advantage of FlowIDs, we will probably have more than one RP at the end station
  - Recall that 802.1Qau relies on sampling, so it will, on average take more CNMs to hit the right RP

FlowIDs are not the ultimate answer

- Does not solve the problem for NICs with stateful offload
  - When performing link aggregation, reverse direction of the flow must come to the same NIC
  - This is an unsolved problem even when using FlowIDs
    - Unless we come up with a way to negotiate FlowIDs between end stations and agree on a hashing algorithm
    - Increases the scope and complexity of the FlowID discussions

Proposal for an optional FlowID

- End station may or may not choose to tag frames with a FlowID
- In the CNM, the bridge returns ‘n’ bytes in the frame following the VLAN tag
  - If the FlowID is present, it will appear in the CNM (2 bytes FlowID Etype + 2 bytes FlowID)
  - More later on ‘n’
- Allows seamless interoperation of systems that need the FlowID and those that don’t
- Does not burden all implementations with FlowID

Without the FlowID…

- Multi-NIC link aggregation can still be made to work
  - The CNM may need to be forwarded to the correct NIC via software causing some increase in the processing time
- Identifying the correct RP using the returned frame
  - A single RP for the priority, so the CNM's contents do not even need to be parsed
    - We are doing much, much better than just PFC
- When we have realized some benefit of CN, we can worry about adding FlowIDs

How big should ‘n’ be?

- ‘n’ is the number of bytes of the original frame following the VLAN tag returned in the CNM
- This information is useful even if we have a FlowID for network management purposes
- All frames could potentially have a FlowID (4 bytes) in addition to the Ethertype/Length (2 bytes)
- In addition to those 6 bytes, we would need
  - TCP or UDP over IPv4
    - 20 (IPv4) + 20 (TCP or UDP) = 40 bytes
  - TCP or UDP over IPv6
    - 40 (IPv6) + 20 (UDP) = 60 bytes
  - FCoE
    - 14 (Ver, Resd, SOF) + 24 bytes (FC header) = 38 bytes
  - FCoE with IFR
    - 14 (Ver, Resd, SOF) + 24 bytes (Encap header) + 8 bytes (IFR header) + 24 bytes (FC header) = 70 bytes
- In order to cover the common the cases, 80 bytes should suffice

Summary

- The FlowID is useful and solves some problems
- There are a large number of deployments that can benefit from CN without a FlowID
  - A single RP will suffice for many deployments
- Allow for end station implementations that don't use FlowIDs