Project for Improved Congestion Management

Feng Gao (Baidu),

Kevin Shen, Paul Congdon, Yolanda Yu (Huawei),

Carmi Arad (Marvell),

Barak Gafni (Mellanox)

IEEE 802.1 DCB

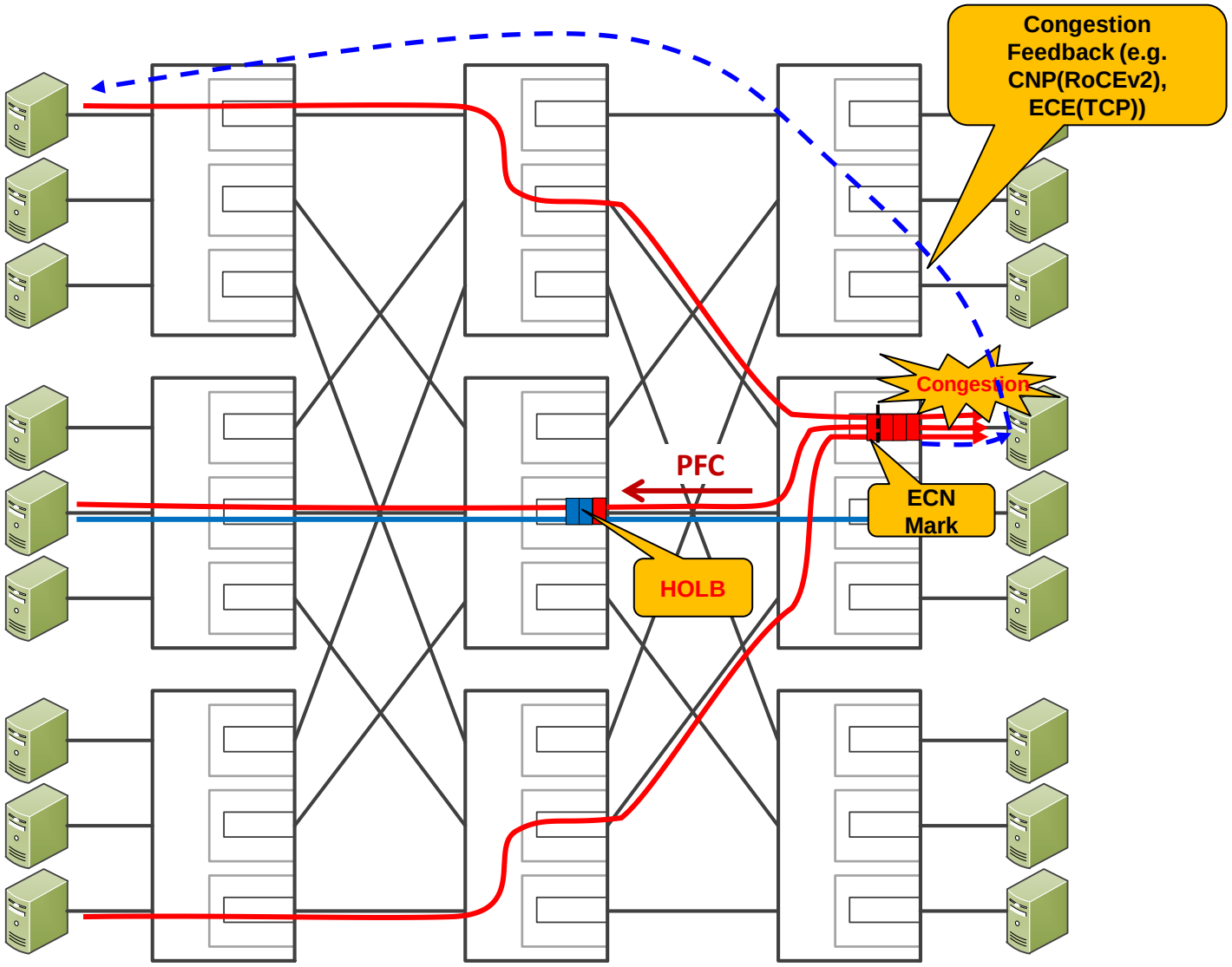
Orlando, FL

November 2017

Agenda

- Current Challenges
- Goals
- Discussion of Scope of effort
- Next Steps

Congestion Management DCN State-Of-The-Art



Congested Flow

Victim Flow

ECN Control Loop

- DCN is primarily an L3 network
- ECN used for end-to-end congestion control
- Congestion feedback can be protocol and application specific
- PFC used as a last resort to ensure lossless environment, or not at all in low-loss environments.
- Traffic classes for PFC are mapped using DSCP as opposed to VLAN tags

Reasons for new work consideration

- Lossless, low-latency Data Centers are desirable
  - Congestion is the primary cause of loss and delay (often due to naturally occurring Incast)
  - Retransmission penalties for storage and RDMA (RoCEv2) are debilitating
- Over-use of Priority-based Flow Control without appropriate congestion control is problematic
  - Too course grain, create head-of-line blocking for victim flows sharing traffic class
  - Congestion spreading impacts non-congested flows
  - Low adoption of PFC because of configuration complexities and above side-effects
- Adoption of existing 802.1 Congestion Notification (802.1Qau) is low
  - DCN Fabric is L3, not large scale multi-hop L2
  - Limited or no use of non-IP based flows (e.g. FCoE, RoCEv1)
  - L3/L4 Explicit Congestion Notification (ECN) used to trigger reduced data rates
- Increasing switch buffering is hard and expensive
  - Amount of buffering per-port-per-Gbps is shrinking
  - Implementation challenges at higher speed and greater densities
- Identifying IP flows in current switch silicon is not a problem
  - Precedence has been set by 802.1CB for IP based flow identification
  - Design learning from early OpenFlow support
  - Security filtering and QoS treatment have long been flow aware

Goals

- Support larger, faster data centers (Low-Latency, High-Throughput)
- Support lossless transfers
- Improve performance of TCP and UDP based flows. Especially the use of RDMA protocols.
- Reduce pressure on switch buffer growth
- Reduce the frequency of relying on PFC for a lossless environment

- Eliminate or significantly reduce HOLB caused by over-use of PFC

Congestion Isolation is a viable approach

Definition: An approach to isolate flows causing congestion to avoid head-of-line blocking.

The approach involves:

1. Identifying the flows creating congestion (e.g. perhaps already done for QCN and/or ECN)
2. Using implementation specific approaches to dynamically adjust the traffic class of offending flows without packet re-ordering
3. Signaling upstream indications that the flows have isolated

CI is not like PFC or QCN

but we don't have to reinvent all the wheels

- However CI can leverage and build-upon changes done for 802.1Qau.
  - Architecture for injecting CIP messages and fitting into the baggy pants
  - Congestion detection and triggering mechanism for sending CNM
  - Congestion message format
- Additional components will be needed
  - Table keeping track of upstream peer MAC address (possibly elsewhere specified?)
  - Congested Flow identification and management (possible leverage from 802.1CB)
  - State machines for processing CIP messages
- Simplifications over Qau
  - No congestion domains to discover or defend against
  - No need to support CN-Tag
  - CI is hop-by-hop, so no issue within the PBB domain
  - No need to specify a reaction point

Architecture Fit Considerations/Thoughts

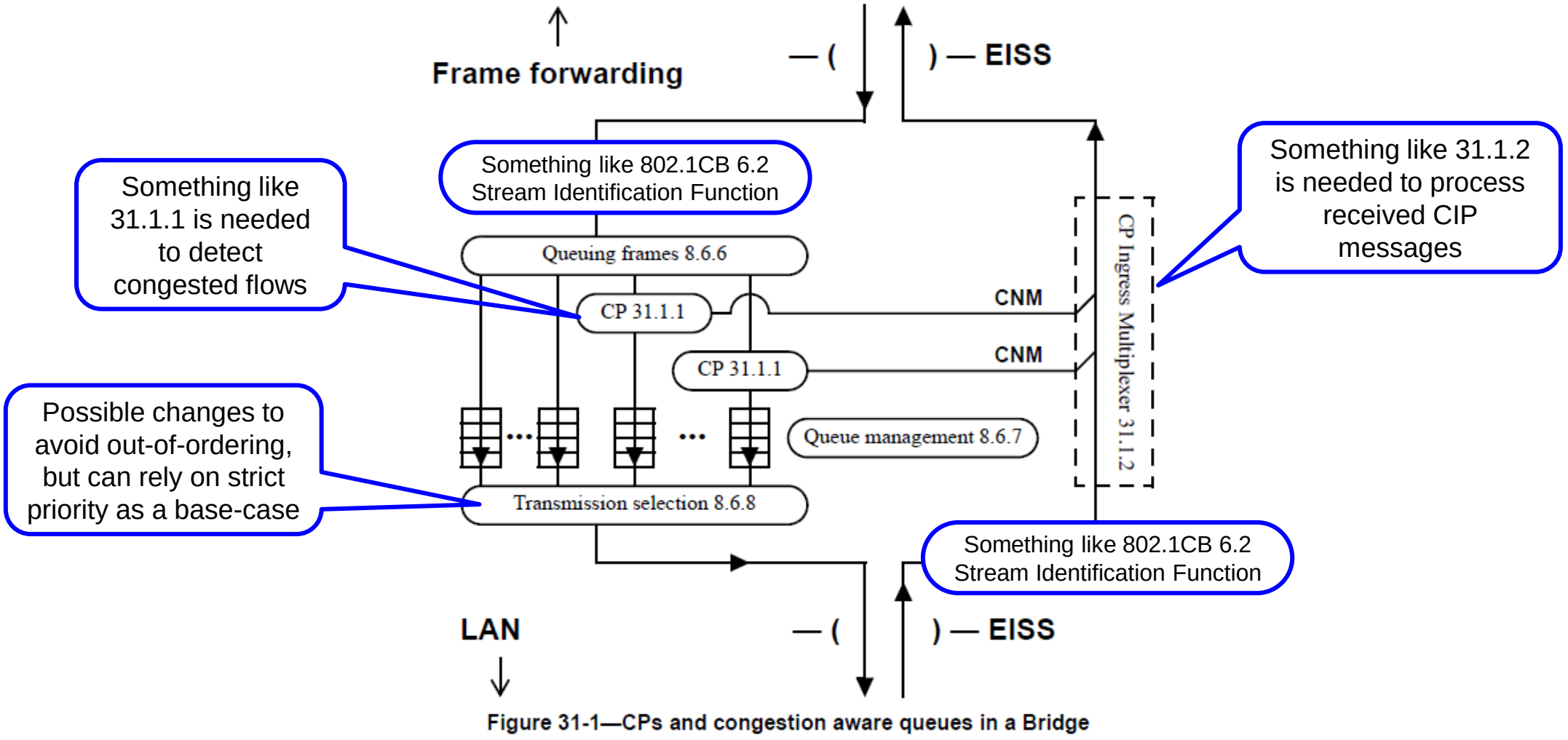


Figure 31-1—CPs and congestion aware queues in a Bridge

Early analysis of 802.1Q changes to support CI

- Many concepts and terminology from Qau could be leveraged if slightly modified
  - Congestion aware system
  - Congestion point
- Leverage Qau CNM decoding in Clause 6 – MAC Service
- Similar statements in Queueing Frames (8.6.6) regarding generating and injecting CNM/CIP messages
- Clause 12 requires CI Managed Objects to be define
- No Clause 17 (MIB), but will need YANG model
- Congestion Isolation TLV for LLDP
- New Main Clauses (Leverage aspects of Qau)
  - Principals of Congestion Isolation
  - Congestion Isolation Entity Operation
  - Congestion Isolation Protocol
  - Congestion Isolation PDU Encoding

Summary

- Congestion Isolation provides the following benefits:
  - Mitigates Head-of-Line blocking caused by PFC
  - Improves flow completion times
  - Reduces or eliminates the need for PFC on non-congested flow queues
- Congestion Isolation is a viable solution. Variations may also be viable.
  - Scope of changes to 802.1Q are manageable with some leverage of 802.1Qau
  - Improvements to 802.1 Congestion Management are needed to support current DCN L3 environment
- Next Steps
  - Motion for approval to draft PAR and 5C for interim review