Congestion Management deepdive & Qdw considerations

Lihao Chen (lihao.chen@huawei.com)
Paul Congdon (paul.congdon@outlook.com)
Lily Lv (lvyunping@huawei.com)
Ramesh Sivakolundu (sramesh@cisco.com)

Qdw considerations

- The rapid acceleration of AI workloads has introduced substantial shifts in traffic patterns, scalability requirements, and control-plane expectations within modern Data Center Networks (DCNs). In this context, it is necessary to evaluate this project in terms of its advantages and applicability across diverse deployment models. This includes examining emerging trends such as edge<->network collaboration and functional partitioning, advances in congestion-management and flow-control mechanisms, and the evolving operational priorities expressed by both end-users and vendors.

Historical Decisions: Congestion management, QCN & PFC, and the DCB TG

- Direction: Forward vs. Backward notification?
- Intention: control the networking hardware vs. only provide feedback to the edge?
- Method: Rate control vs. On-Off (Pause)

These are based on the author's best-effort investigation and understanding. There remains a possibility of some degree of misinterpretation.

Direction: Forward vs. Backward notification?

- **Forward notification / FECN**
  - **Arguments**: High path accuracy (forward frames actually experience congestion) and support for congestion aggregation (switches can accumulate or update congestion markings), and consistency with IP ECN principles.
  - Links: http://www.ieee802.org/1/files/public/docs2006/au-jain-ecn-20061115.pdf , http://www.ieee802.org/1/files/public/docs2007/au-jain-fecn-20070124.pdf , http://www.ieee802.org/1/files/public/docs2007/au-jain-fecn-20070313.pdf , http://www.ieee802.org/1/files/public/docs2007/au-sim-kwan-ding-revised-prelim-fecn-20070329.pdf , etc.
- **Backward notification (goes to QCN eventually)**
  - **Arguments**: Direct and fast control loops that reduce response latency (more suitable for burst traffic and short-lived flows), no reliance on receivers, simple hardware implementation, quantized feedback and recovery mechanisms to address the original BCN deficiencies.
  - Links: http://www.ieee802.org/1/files/public/docs2007/au-bergamasco-bcn-ecn-comparison-jan-2007-interim-v0.1.pdf , http://www.ieee802.org/1/files/public/docs2007/au-prabhakar-qcn-description.pdf , http://www.ieee802.org/1/files/public/docs2007/au-bergamasco-qcn-problems-solutions-proposal-070905.pdf , http://www.ieee802.org/1/files/public/docs2007/au-pan-qcn-details-053007.pdf , http://www.ieee802.org/1/files/public/docs2007/au_prabhakar_qcn_overview_geneva.pdf , http://www.ieee802.org/1/files/public/docs2007/au-ZRL-prelim-QCN-r1.01.pdf , etc.
- Lessons learned:
  - Both Forward and Backward have their advantages.
  - ‘Backward’ was chosen in Qau for several other reasons: Engineering feasibility in controllable and homogeneous data center environments; clearly defined protocol boundaries, explicit interfaces, and pseudocode; and supports.

Intention: control the networking hardware vs. only provide feedback?

- **To control the hardware**
  > QCN messages carry quantized feedback (Fb) to the source. The source NIC then uses this feedback, alongside internal variables (such as RI/RD), to algorithmically calculate and enforce an explicit transmission rate limit.
- **Provide status feedback to the source and let it decide how to adjust transmission**
  > No discussion of this approach is found in 802.1Qau.

  > Interestingly, DCQCN — the de facto DCN standard used with RoCEv2 — retains QCN's core control-loop concepts but replaces its L2 signaling with L3 IP ECN (forward) and a receiver-generated CNP (backward). Unlike QCN's L2 rate controlling parameters defined in Qau, the CNP acts as a general explicit-congestion signal, giving end-stations greater flexibility to tune algorithmic parameters for different traffic profiles.

  > DCQCN: https://conferences.sigcomm.org/sigcomm/2015/pdf/papers/p523.pdf

  > NSCC (Network Signal-based Congestion Control) in UEC shares a similar design philosophy with DCQCN. https://ultraethernet.org/wp-content/uploads/sites/20/2025/10/UE-Specification-1.0.1.pdf

- Lessons learned:
  > A theoretically elegant model may mismatch real-world requirements — particularly regarding backward compatibility, hardware implementation complexity, and the essential need for **users to retain control and flexibility at the network edge (end-stations)**.

Method: Rate control vs. On-Off (Pause)

- Based on backward notification, QCN adopts **rate control** at the source using rate limiters.
  - > There are also practical limitations to consider, as well as trade-offs between complexity and performance. E.g., the finite number of rate limiters lead to multiple flows get mapped into a same rate limiter.
  - > Links: http://www.ieee802.org/1/files/public/docs2007/au-wadekar-practical-limitations-RL-v1.pdf , http://www.ieee802.org/1/files/public/docs2007/au-wadekar-cm-nic-perspective-rev-1.0.pdf , http://www.ieee802.org/1/files/public/docs2008/au-bestler-endstationrps-0708-05.pdf
- To **pause (on-off control)** a neighbor or a remote port (e.g., the source)
  - > No discussion of ‘remote pause’ is found in 802.1Qau.
  - > However, there was a clear requirement for priority-based pause, as the previous port-based PAUSE was too coarse-grained, leading to severe congestion spreading and performance degradation. And this requirement led to the specification of PFC in 802.1Qbb (2008–2011).
  - > Links: http://www.ieee802.org/1/files/public/docs2006/au-Brunner-Hazarika-Priority-Pause-considerations-111406.pdf , http://www.ieee802.org/1/files/public/docs2007/au-ZRL-Ethernet-LL-FC-requirements-r03.pdf

- Lessons learned:
  - > QCN handles ms-level congestion mitigation, while PFC provides µs-level pause capability for lossless networking.
  - > The intent is for **QCN and PFC to coexist and complement each other**. While QCN can reduce the use of PFC, PFC is still needed as a ‘last ditch’ effort to avoid packet loss.

Summary of Historical Decisions

- Direction: Forward vs. Backward notification?
  > Decision: QCN (802.1Qau) ultimately adopted **Backward** notification.

  > Arguments that lead to the decision: Faster reaction time, no dependency on receiver, quantized feedback for finer precision.

- Intention: control the networking hardware vs. only provide feedback to the edge?
  > Decision: QCN **control**s the source NIC transmission hardware.

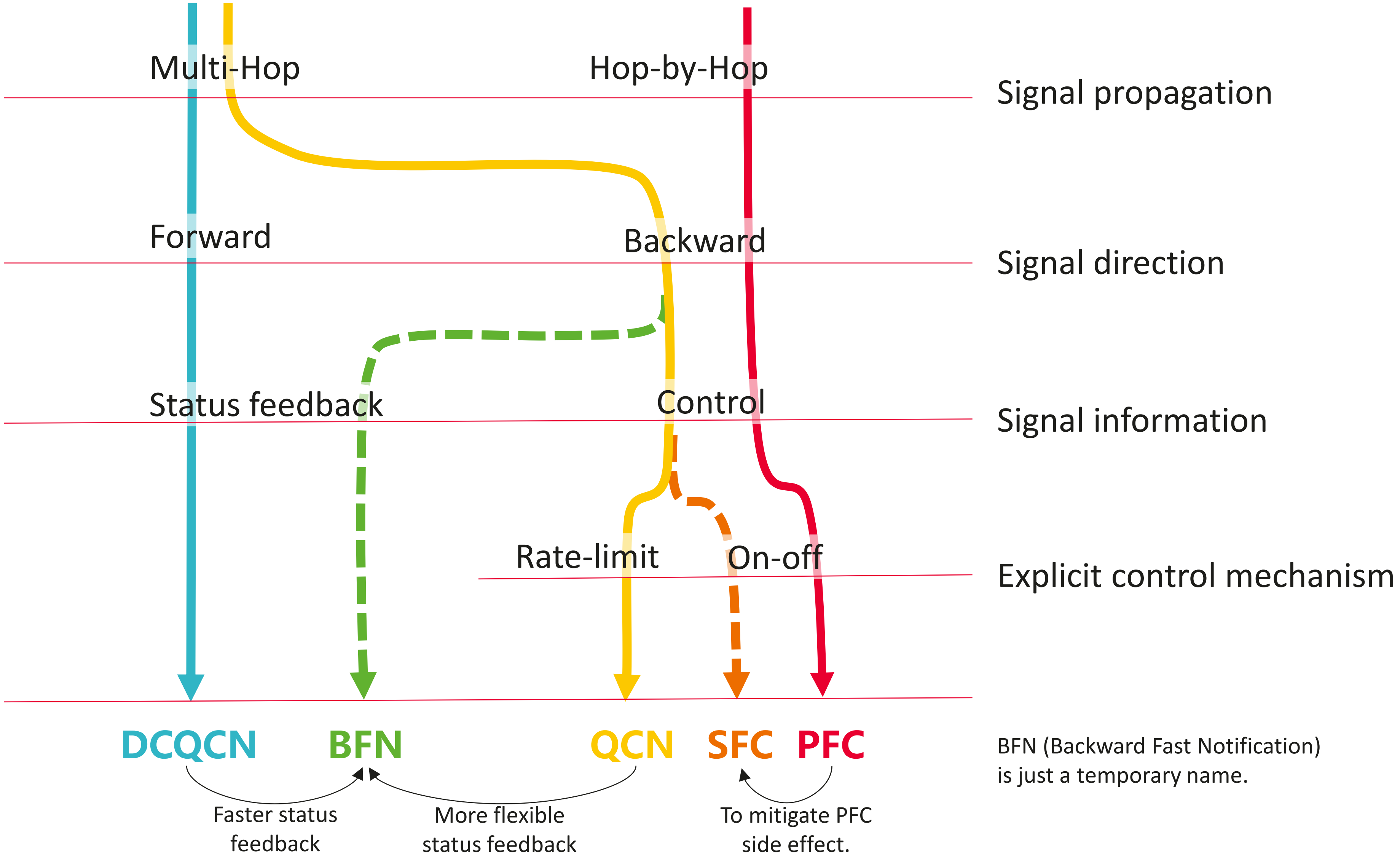
  > Mechanism: QCN messages carry quantized feedback (Fb) to the source. The source NIC then uses this feedback, alongside internal variables (such as RI/RD), to algorithmically calculate and enforce an explicit transmission rate limit.

- Method: Rate control vs. On-Off (Pause)
  > Decision: QCN performs **Rate Control**.

  > The intent is for QCN and PFC to coexist and complement each other. While QCN can reduce the use of PFC, PFC is still needed as a ‘last ditch’ effort to avoid packet loss.

These are based on the author’s best-effort investigation and understanding. There remains a possibility of some degree of misinterpretation.

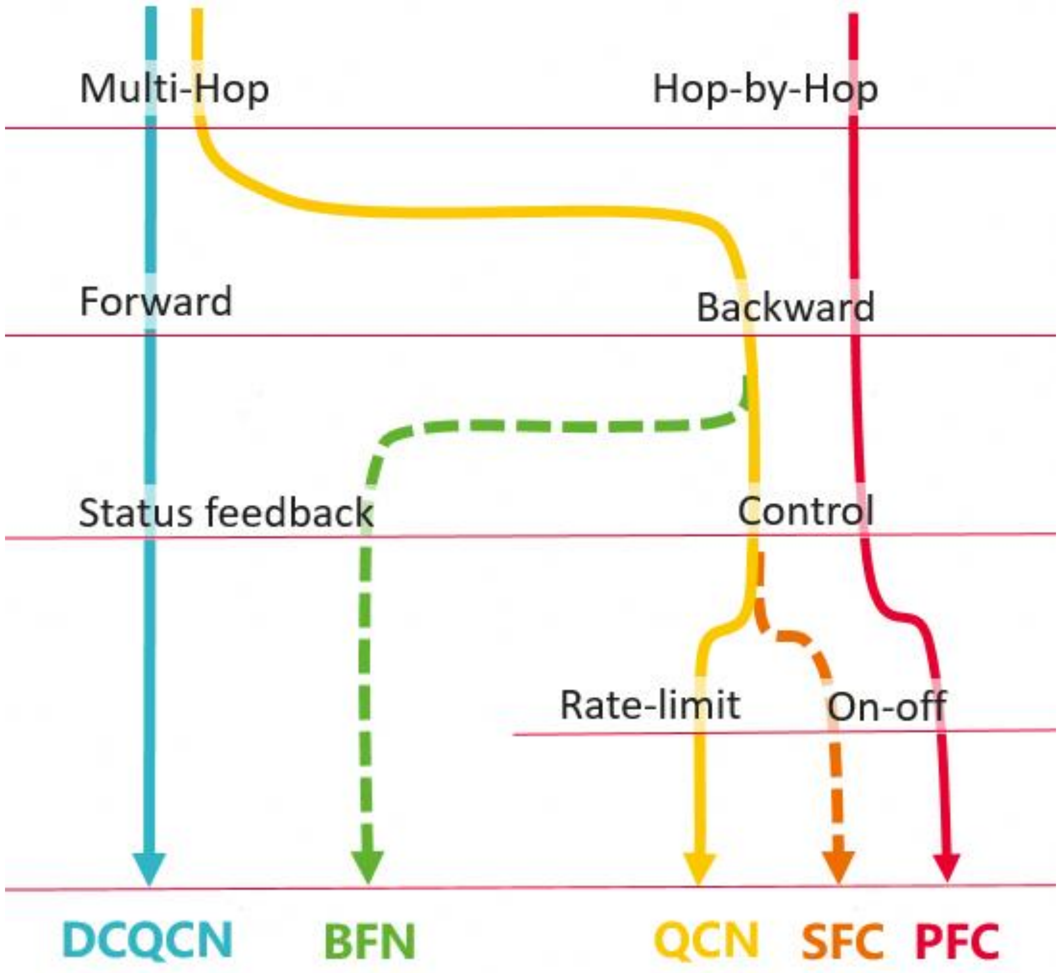
Taxonomy of discussed solutions



BFN (Backward Fast Notification) is just a temporary name.

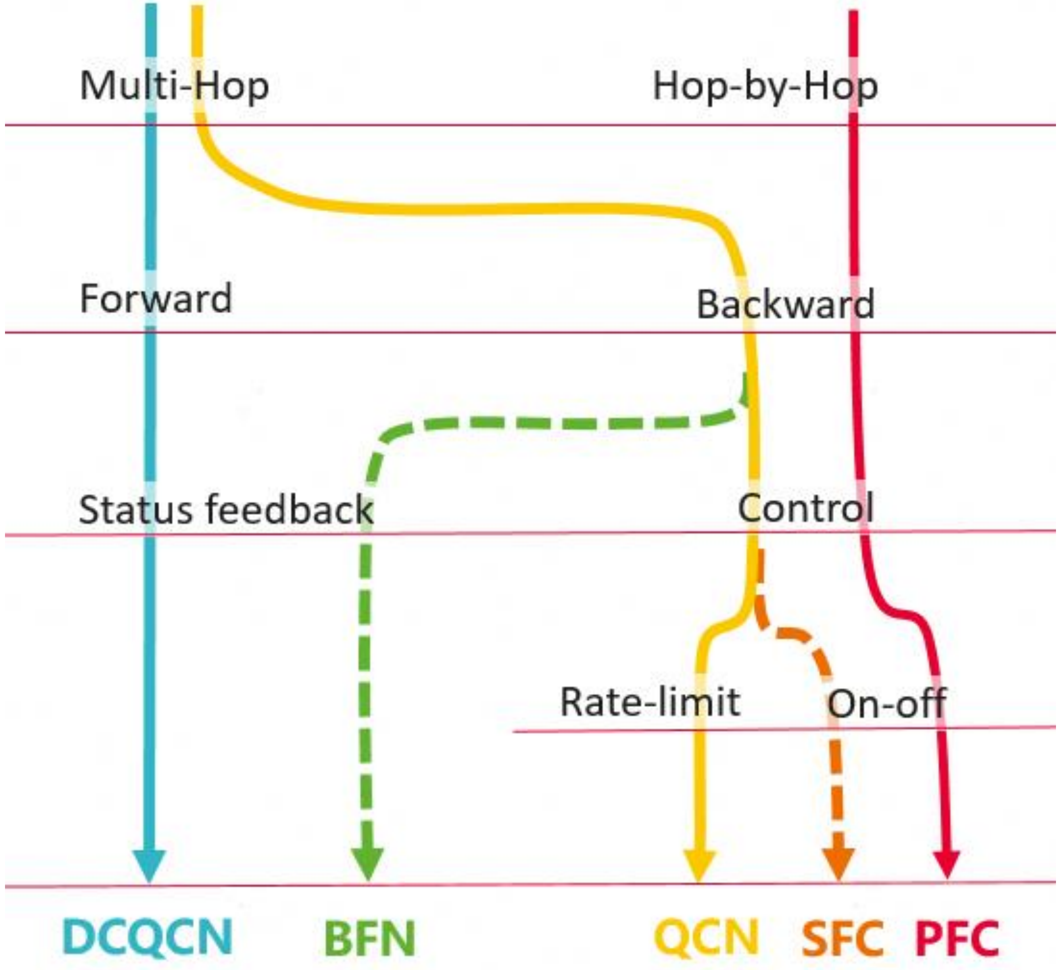
Industry Practices

- Current deployment status: Hardware implementations favor the simpler link-layer PFC for lossless networking, while rate control is typically offloaded to transport or application layers. For example, **PFC+DCQCN**.
- New directions:
  - > 1) SFC (Source Flow Control). The **remote PFC** is an instance of SFC in practice.
    - https://support.huawei.com/enterprise/en/doc/EDOC1100518851/a390b2b1/understanding-rpfc
  - > 2) BFN (Backward Fast Notification): Resembles QCN's telemetry but does not mandate hardware-based algorithms, granting end-stations the freedom to interpret congestion signals and make autonomous decisions. Related approaches include **FastCNP** and **BTS with packet trimming**.
    - https://docs.broadcom.com/doc/BCM78919-PB
    - https://blogs.cisco.com/datacenter/ultra-ethernet-for-scalable-ai-network-deployment#:~:text=Smarter%20congestion%20recovery,congestion%20and%20improving%20tail%20latency.



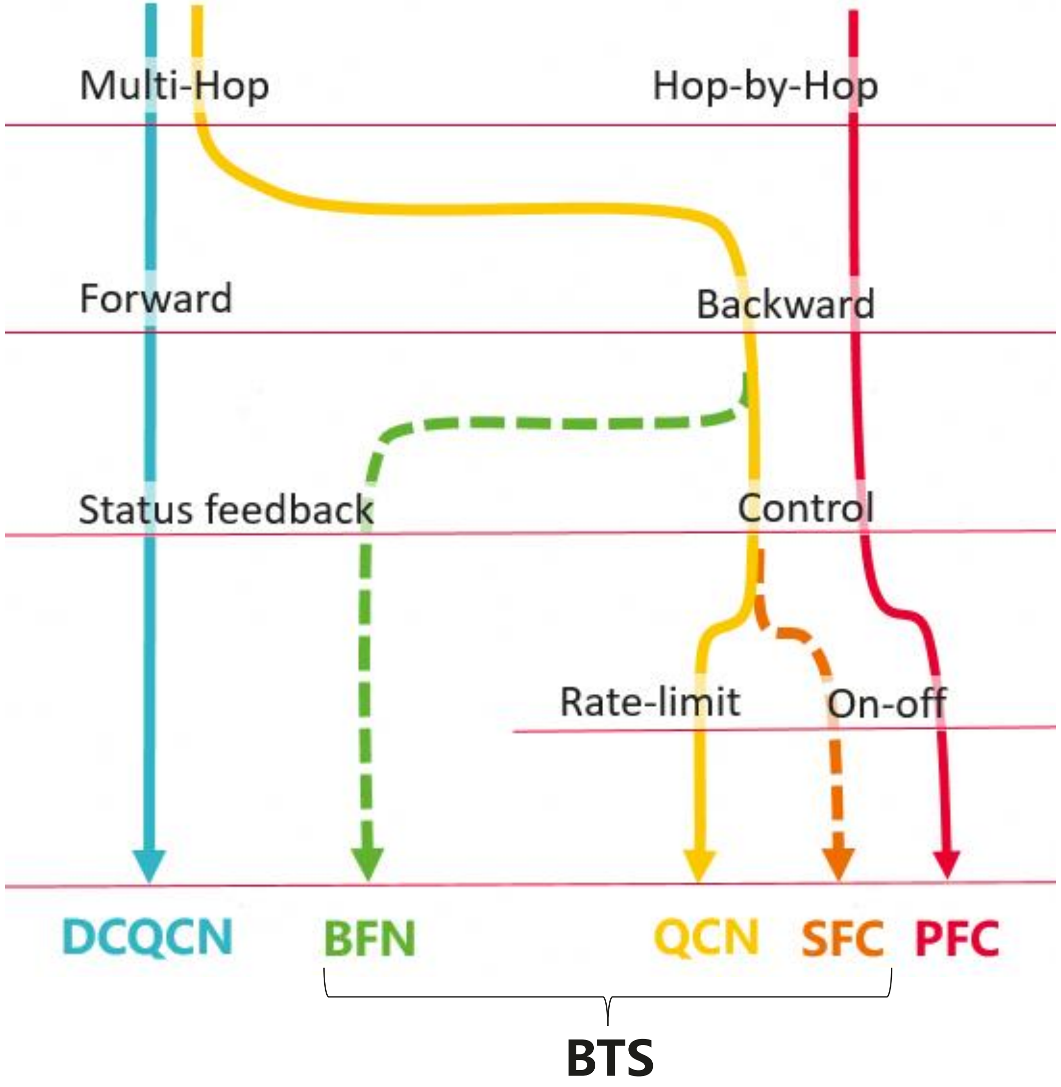
Qdw considerations

- **A generic 'Back-To-Sender' protocol** that carries information, e.g., congestion signals, trimmed packets, pause requests.
  > To provide congestion information on the network to help end-stations to make informed decisions.
  > To trigger faster retransmissions.
  > To mitigate PFC side effects.
- This protocol can operate in both lossless AND lossy (best effort) environments.
  > While current AI and HPC DC deployments predominantly rely on lossless, there is a growing trend toward lossy that prioritizes scalability.
- **The core scope of P802.1Qdw** — 'This amendment specifies procedures, managed objects, and a YANG data model for the signaling and remote invocation of flow control at the source of transmission in a data center network' — **remains valid**.
- The author's idea is to **address prevailing industry concerns and attempt to consolidate all BTS-like concepts into a unified protocol**, thereby avoiding future deployment incompatibilities and redundant efforts.



PAR proposal

- **A generic 'Back-To-Sender' protocol** that carries information, e.g., congestion signals, trimmed packets, pause requests.
  > To provide congestion information on the network to help end-stations to make informed decisions.
  > To trigger faster retransmissions.
  > To mitigate PFC side effects.
- BFN (Backward Fast Notification) can be the core of this generic BTS protocol. However the BFN can also be used — e.g., by end-stations or edge bridges — for PFC (SFC) or rate limiters (QCN).
- Proposed Scope:
  > This amendment specifies procedures, managed objects, and a YANG data model for the signaling of congestion indicators towards the source of transmission in a data center network. This standard makes provisions for backward compatibility to IEEE Std 802.1Q.



PAR extension / modification (Mark mode)

- ~~Source Flow Control~~ Backward Fast Notification
- Scope
  > This amendment specifies procedures, managed objects, and a YANG data model for the signaling ~~and remote invocation of flow control at~~ of congestion indicators towards the source of transmission in a data center network. ~~This amendment specifies enhancements to the Data Center Bridging Capability (DCBX) protocol to advertise the new capability. This amendment specifies the optional use of existing stream filters to allow bridges at the edge of the network to intercept and convert signaling messages to existing Priority-based Flow Control (PFC) frames.~~ This standard makes provisions for backward compatibility to ~~amendment also addresses technical and editorial corrections to existing~~ IEEE Std 802.1Q ~~functionality~~.
- Need for the project
  > Congestion~~, in particular incast congestion,~~ is detrimental to network performance in the data center and most acutely affects the widely used Remote Direct Memory Access (RDMA) protocols, such as RDMA over Converged Ethernet (RoCE), in High-Performance Computing (HPC) and Artificial Intelligence (AI) data center networks. PFC is used to avoid packet loss from congestion, however, PFC applies flow control to a locally attached link resulting in problematic side effects at scale such as congestion spreading and head-of-line blocking. Higher layer end-to-end congestion control typically takes too long adjusting the source transmission rate to avoid buffer exhaustion~~, especially during incast congestion~~. ~~Source Flow Control (SFC) is needed to apply flow control directly to the source as quickly as possible, mitigating packet loss and the side effects of existing PFC at scale. Enabling an edge bridge at the source to intercept signaling messages and convert them to existing PFC supports early adoption and eases end-user migration while also allowing SFC implementations directly on server network adapters.~~ Backward Fast Notification (BFN) transmits rapid feedback towards the source, enabling timely actions such as transmission-rate adaptation and load-balancing decisions, which in turn improves link utilization and helps avoid congestion hotspots.

Discuss

PAR extension / modification

- P802.1Qdw Source Flow Control -> P802.1XX Backward Fast Notification

- Current scope

  > This amendment specifies procedures, managed objects, and a YANG data model for the signaling and remote invocation of flow control at the source of transmission in a data center network. This amendment specifies enhancements to the Data Center Bridging Capability (DCBX) protocol to advertise the new capability. This amendment specifies the optional use of existing stream filters to allow bridges at the edge of the network to intercept and convert signaling messages to existing Priority-based Flow Control (PFC) frames. This amendment also addresses technical and editorial corrections to existing IEEE Std 802.1Q functionality.

- Proposed change

  > This amendment specifies procedures, managed objects, and a YANG data model for the signaling of congestion indicators towards the source of transmission in a data center network. This standard makes provisions for backward compatibility to IEEE Std 802.1Q.

PAR extension / modification

- Current need for the project
  - > Congestion, in particular incast congestion, is detrimental to network performance in the data center and most acutely affects the widely used Remote Direct Memory Access (RDMA) protocols, such as RDMA over Converged Ethernet (RoCE), in High-Performance Computing (HPC) and Artificial Intelligence (AI) data center networks. PFC is used to avoid packet loss from congestion, however, PFC applies flow control to a locally attached link resulting in problematic side effects at scale such as congestion spreading and head-of-line blocking. Higher layer end-to-end congestion control typically takes too long adjusting the source transmission rate to avoid buffer exhaustion, especially during incast congestion. Source Flow Control (SFC) is needed to apply flow control directly to the source as quickly as possible, mitigating packet loss and the side effects of existing PFC at scale. Enabling an edge bridge at the source to intercept signaling messages and convert them to existing PFC supports early adoption and eases end-user migration while also allowing SFC implementations directly on server network adapters.

- Proposed change
  - > Congestion is detrimental to network performance in the data center and most acutely affects the widely used Remote Direct Memory Access (RDMA) protocols, such as RDMA over Converged Ethernet (RoCE), in High-Performance Computing (HPC) and Artificial Intelligence (AI) data center networks. PFC is used to avoid packet loss from congestion, however, PFC applies flow control to a locally attached link resulting in problematic side effects at scale such as congestion spreading and head-of-line blocking. Higher layer end-to-end congestion control typically takes too long adjusting the source transmission rate to avoid buffer exhaustion. Backward Fast Notification (BFN) transmits rapid feedback towards the source, enabling timely actions such as transmission-rate adaptation and load-balancing decisions, which in turn improves link utilization and helps avoid congestion hotspots.