Considerations for congestion notification in FCoE networks

GILLES CHEKROUN
SNIA Europe Technology Chair, Cisco Systems

Since the earliest days of data networking, congestion control and management have been a major effort to ensure the best throughput. Years ago we had technologies such as Frame Relay Forward Explicit Congestion Notification (FECN) and Backward Explicit Congestion Notification (BECN), and others to achieve simple and effective methods of congestion notification and avoidance.

As TCP/IP networks grew across the globe, tools such as Weighted Random Early Discard – Explicit Congestion Notification (WRED-ECN) and others became increasingly popular mechanisms for congestion notification and avoidance. These mechanisms offered methods to selectively drop packets to control the bandwidth of TCP flows by controlling the TCP window size.

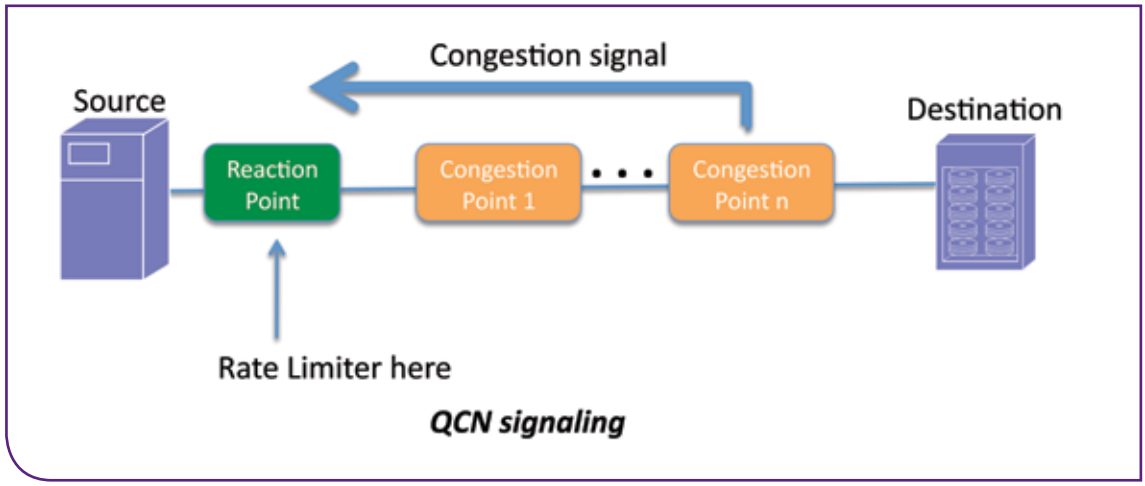
The TCP window size allows transmission of a certain number of frames without acknowledgement. This window size grows dynamically until packets are dropped and acknowledgments are not received so TCP will reduce the window size and the process repeats. When Fibre Channel traffic is carried over Ethernet via FCoE, there is no TCP/IP layer and so another mechanism for congestion control and management should be applied. As in traditional Fibre Channel deployments, data loss inside the FCoE fabric is a highly undesirable mechanism for controlling congestion notification and avoidance.

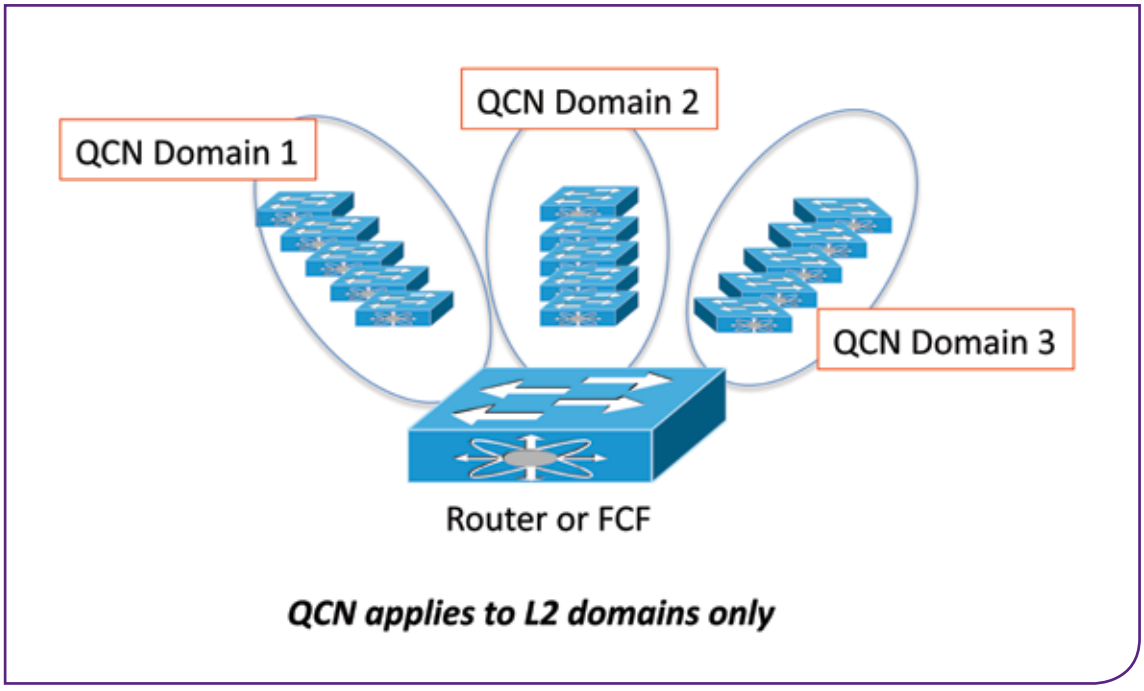
This document will review some considerations for congestion notification and avoidance in FCoE networks.

One important definition at this point is that of a FCoE Fiber Channel Forwarder (FCF) which is a FC forwarding and services application instance within the fabric, much like a fabric switch is within traditional Fibre Channel fabrics.

What is QCN?

QCN stands for Quantize Congestion Notification.
QCN is an effort started in IEEE 802.1Qau to basically introduce a concept of end-to-end congestion notification in Layer 2 Networks as part of the larger Data Center Bridging





effort. The fundamental QCN goal is to be sure that congestion is controlled in a dynamic fashion from participating switches and notifications are sent to the hosts. To have a successful implementation, switches implement a portion of algorithm and hosts implement another portion of algorithm.

Along the network, switches run the Congestion Point (CP) algorithm where the switch attached to an oversubscribed link samples incoming packets and generates feedback message addressed to the source of the sampled packet.
The feedback message contains information about the extent of congestion at the Congestion Point and is propagated to the Reaction Point (RP) is where rate limiting will be applied. The QCN signaling is here to allow the sender to dynamically adapt to the bandwidth available in the Layer 2 network.

QCN is applicable within a given Layer 2 domain, meaning that if a MAC address is re-written, for example when traversing a Layer 3 switch or a Fibre Channel Forwarder (FCF), QCN will be terminated and will not enter a new domain. It also requires multiple shaping queues on every QCN-enabled host/switch, making hardware more complex and therefore more expensive.

QCN Implementation

Switches in the network are running a Congestion Detection algorithm and are sending notification to the edge devices so action can be taken to reduce and eliminate the congestion.

All QCN frames are specifically tagged and need to have support on all switches. This is a very important point and it is particularly difficult to implement since ALL devices participating in QCN will need to be updated at once.

This has proven to be a very challenging operational model in computer networks. An easier deployment model is when you can upgrade hosts without having to upgrade multiple switches and vice-versa. Quality of Service is a good example of a simple deployment model, as it can be applied on a specific link when needed.

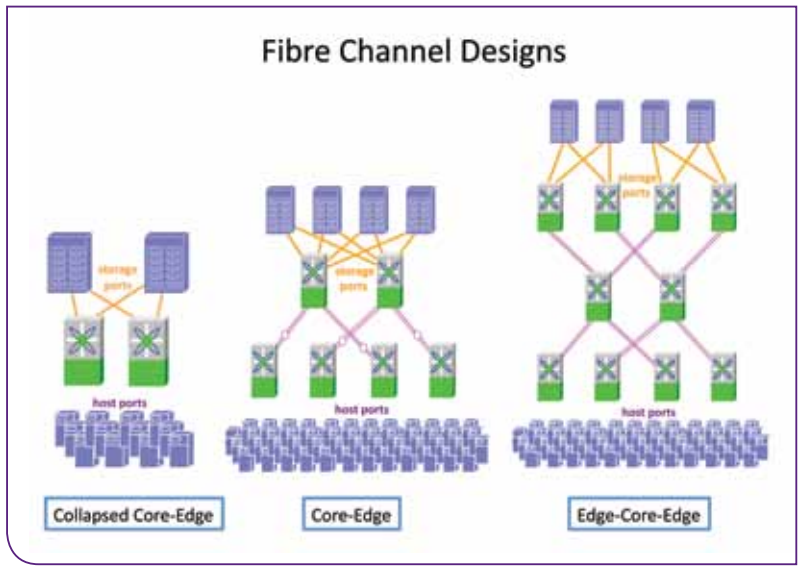
QCN has to be implemented in hardware on switches and all hosts since trying to respond to congestion management messages will be resource intensive and the current view is doing QCN in software will be too slow. It is extremely unlikely that existing hardware within the industry will be able to support QCN. There will be a need to change host interfaces and switches line cards before QCN will become deployable.

Layer 2 Network Size

To take advantage of the QCN algorithm the Layer 2 network has to be of a minimum size. If we have two hosts connected to one switch, QCN will not bring ANY value. Mechanisms like PAUSE or Priority Flow Control can be used to control the traffic. On a large Layer 2 Network, PAUSE frames can create Head of Line Blocking if network traffic is many-to-many and with a mix of large and small frames. The network over-subscription design needs to be appropriate to avoid it and for Fibre Channel the fan-in ratio, i.e. how many hosts to a disk, will dictate it.

The average Network span where QCN will start to have advantages will be a minimum of 3 or 4 Layer 2 switch hops for a given flow. It is important to note, this means the 3 or

4 switches on the same Layer 2 domain, between the individual hosts, Layer 3 Switches or FCF devices.



Network Design and FCoE

Current Data Center networks do not have many Layer 2 switches in a row primarily due to spanning tree stability and broadcast/multicast domain size considerations. The moment a Layer 3 switch or an FCF is crossed the L2 domain is broken since MAC addresses are re-written and QCN is not effective across Layer 3 Switches, or FCF devices.

Current Data Center Fibre Channel fabrics implement forwarding and services functions on each node. Functionality such as Fabric Shortest Path First (FSPF) traffic routing, zoning, buffer crediting, FC link monitoring, are a few examples of these services enforced at each Fibre Channel fabric device today.

For FCoE networks, where the need for Fibre Channel services is required, an element like FCF will partition the L2 domain and cancel the value of QCN. Fibre Channel networks have been primarily used to transport SCSI Commands and Data blocks. The SCSI protocol is by nature an interlock protocol meaning that the Upper Level protocol sends a command and waits for a reply or and acknowledgement. SCSI is very sensitive to latency but most important a lossless delivery is mandatory.

Do we need QCN in FCoE Networks?

Fibre Channel Networks have been operating since day one without end-to-end congestion management but rather hop-by-hop. FCoE today provides today the same flow control as traditional Fibre Channel via Data Center Bridging:

- Priority Flow Control – providing the lossless fabric capability needed for Fibre Channel.
- Enhanced Transmission Selection – providing guaranteed bandwidth for FCoE flows.

QCN has been developed for effective congestion management in large Layer 2 Ethernet networks. Traditional Fibre Channel designs, however, are based on a hierarchy, with the forwarding and services on each node integrated into a given fabric. Three main topologies are currently used in storage fabrics:

- Core-Edge design – where storage arrays are in the core and hosts in the edge. The most common design in Data Centers today.
- Collapsed Core-Edge – where only one large switch is used to connect both hosts and disks.
- Edge-Core-Edge – to be able to create very large Fibre Channel networks.

QCN moves congestion from the center of the network to the edges. If we design FCoE networks in a similar way as we do with Fibre Channel networks, QCN will not be necessary since all switches (Edge and Core) will have Fibre Channel Forwarder (FCF) or FCoE Data Forwarder (FDF) function.

The Layer 2 domain is effectively now each link between these fabric nodes only. This will be valuable for exposing FC management (visibility, forwarding decisions, security, features, etc.) functionality in an identical manner that is used today.

If a new design paradigm was employed that only implemented the FCF functionality at the edges of large Layer 2 domains – and relied on lossless Ethernet and QCN underneath, many operational changes would be required – which contradicts the original design goals of FCoE to begin with.

In this case there is no real visibility to where the traffic is actually flowing on the Ethernet fabric without close coordination between the SAN and LAN administrators resulting in FSPF inefficiencies in forwarding decisions.

Conclusion

- QCN does in fact have value in specific large Layer 2 Ethernet Networks with specific traffic behavior. Small flows contribute very little to congestion.
- For FCoE networks, QCN is definitely NOT needed in a hop-by-hop FCF environment and will provide NO benefits in most of today's Data Center designs.
- Multi-hop devices interconnected with FCoE ISLs at 10 Gigabit Ethernet (or 40 or 100 in the future) will allow FCoE scalability and most probably will not required larger span compared to Fibre Channel networks we have today.
- From a network design perspective, where is the congestion going to happen and from a Fibre Channel point of view, is it going to happen?