Qdw– Why we need to think again

Mick Seaman
mickseaman@gmail.com

What is in the PAR

"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 errors and omissions in the description of existing IEEE Std 802.1Q functionality."

Let's look at these points one by one, and then at the whole picture.

Message conversion poor & unnecessary

"... to allow bridges at the edge of the network to intercept and convert signaling messages to existing Priority-based Flow Control (PFC) frames. ..."

- Whatever the capability of the signaling protocol across the network, this conversion changes it to bang-bang (on-off) control.[1]
- Alternatively, if the edge bridge simply stops or rate limits forwarding frames from the end station, then a PFC will be sent iff necessary.[2]

1. Bang-bang control is notoriously poor when there are significant control loop (round-trip) delays (see Wikipedia if you don't know this). Worse, in this case, the delays are variable, depending on other flows. Stopping one flow can result in an increased rate of other flows at other recipients. Control stability suspect unless the network is always lightly loaded with limited hops.
2. Relying on the existing PFC loss prevention mechanism for flow control between edge bridge and end station (for which the round trip delay can be measured, necessary buffer headroom calculated) means the cross-network signaling can be independently designed and as sophisticated as needed. No need for changes to the existing PFC spec.

Use of ‘existing stream filters’

“... specifies the optional use of existing stream filters to allow bridges at the edge of the network to intercept and convert ...”

Cross-network congestion control protocol participation:

1. End station and edge bridge do not understand it.
2. End station understands it, edge bridge does not.
3. End station does not understand it, edge bridge does.
4. End station and edge bridge understand it.

Only in case 3 does the edge bridge have to take any exceptional action. It knows[1] for any unicast packet through which interface it will forward the frame, so all it has to do is recognize the protocol and ask itself “do(es) the station(s) reach through this interface support this protocol”. If not the edge bridge should take whatever action is appropriate to limit the flow from the interface, which if the PFC headroom is placed at risk by that limiting will cause a PFC to be sent.

Use of DCBX

"... specifies enhancements to the Data Center Bridging Capability (DCBX) protocol to advertise the new capability ..."

Unclear from the PAR who sends and receives these advertisements, The only useful advertisement is from an end station (that understands the cross-network congestion control protocol) to its edge bridge[1] so the edge bridge will forward the protocol and not proxy for it.

In this case the 'DCBX protocol' part is just a packet format. But DCBX:

- Capabilities much reduced in the course of its development, but are still significantly overstated in 802.1Q-2022.
- Ties a whole set of capabilities together in a way that is ignored.
- Abuses the use of LLDP by changing it, for PFC, into an acknowledged protocol (using the 'Willing' bit).
- Has a current YANG module that needs revision.

We have been here before (802.1Qau)

802.1Q Clause 30, QCN (Quantized Congestion Notification)

Started as a naive protocol to manage congestion in the network.

Extensive analysis and simulation from multiple teams over several years

Known stability (control loop delay) bounds.

Controls rate, not on/off.[1]

Effectively uses higher order terms (derivative) in rate setting.[2]

1. Rapid transmission of alternating stop/go signaling or stop with rapid expiry is not a substitute when delays are variable. Per network tuning is cost prohibitive and vendor dependent tweaks mean that we don't have a standard.
2. See [B1] in 802.1Q-2022.

But this time is different?

- Could just start with QCN and add an encapsulation run over IP etc.
- But the traffic flows differ, and that could be important
- If the problem is final hop incast, with networks of a known structure, and consistent flow profiles duration then that:
  a) Simplifies analysis and simulation
  b) Allows use of additional information that significantly simplifies the problem, leading to useful answers e.g. knowing the total acceptable rate for a set of competing incast flows, knowing how many flows contribute to that.
  c) Means that a congestion control solution which does take advantage of additional information will supersede one that does not.

A ‘standard’ that is simply a marketing tool that does not facilitate multi-vendor interoperability is no standard.

A multi-vendor solution that requires extensive per-deployment management is unlikely to succeed.

A good standard takes considerable time and on-going effort beyond the drafting of amendment text. If something better is going to come along in the next 3 years we are wasting our time.