PFC Management

Topics and assumptions to get straightened out before initial ballot.[1, 2]

What is required for conformance?

How do we tell PFC is working satisfactorily?

DCBX spec + LLDP MIB vs YANG model

Mick Seaman
mickseaman@gmail.com

[1] The whole point of P802.1Qdt is to make management of PFC easier. Automating buffer allocation calculation, whether by delay measurement or by using PTP together with realistic estimates of local delays in the possible configuration scenarios (unprotected data and PFC, MACsec protected data, MACsec protected data and PFC, Privacy protection) is just part of that.

[2] Initial draft in 36.11 P802.1Qdt/D0.4.

PFC Conformance

Currently, in Clause 5 Conformance:[1]

- f) of 5.4.1 ‘VLAN Bridge component options’ is support of PFC, referencing 5.11 (see below).
- aa) of 5.4.1 is support of Enhanced Transmission Selection (ETS) referencing 5.4.1.6 ‘ETS Bridge requirements’ which:
  - — mandates bandwidth allocation configuration with a granularity of 1% or finer, and actual allocation with a precision of 10%;
  - — mentions PFC only as a footnote;
  - — requires support of DCBX, referencing Clause 38.
- 5.4.1.7 ‘DCBX Bridge requirements’ is not referenced by any other text, but requires support of LLDP and the DCBX state machines with the following:
  - — ETS Configuration and ETS Recommendation TLVs;
  - — PFC Configuration TLV;
  - — Application Priority TLV and Application TLVs.

    All of which specify transmission of the TLV as an option ‘may’.
- 5.11 ‘System requirements for ... (PFC)’ references Clause 36 details, and also requires:
  - — ‘Enable use of PFC only in a domain controlled by DCBX (Clause 38).’

    Clause 38 does not include the word ‘domain’.None of the remaining 1,094 uses of the word ‘domain’ are associated with DCBX, so there is no explanation of how a system can determine that it is ‘in a domain controlled by DCBX’.
- 5.13.1 ‘MAC Bridge component options’ item k) is support of PFC (referencing 5.11) and item l) is support of DCBX, but DCBX requires support of the Application VLAN TLV.

[1] P802.1Q-Rev/D1.3 (unchanged, I believe w.r.t., the above points from IEEE Std 802.1Q-2022).

PFC Conformance (PICS—Annex A)

Currently in Annex A ‘PICS proforma—Bridge implementations’:

- A.5 ‘Major capabilities’ includes item:
  - — ‘PFC’ ‘Is Priority-based Flow Control implemented’, status ‘O’ (optional) not qualified by any reference to DCBX, with references to 5.11 and Clause 36 [no reference to A.33, or f) of 5.4.1].
- A.14 ‘Bridge management’ includes items:
  - — MGT-208 ‘PFC entities’, status ‘PFC:O’, i.e. optional conditional on the A.5 item, with reference 12.22.6 — which is the ‘SRP Stream Preload Table’. This is the only occurence of ‘PFC entities’ in 802.1Q.
  - — MGT-209 ‘ETS Control Entities’, status ‘ETS:M’, with reference 12.24 — which is ‘1:1 PBB-TE IPS managed objects’. This is the only occurence of ‘ETS Control Entities’ in 802.1Q.
  - — MGT-210 ‘PFC Control Entities’, status ‘PFC:M’, with reference 12.22.6 (see above). This is the only occurence of ‘PFC Control Entities’ in 802.1Q.

    I have checked prior 802.1Q amendments and revisions for the above references back to their introduction in 802.1Qaz and could find no time where they pointed to relevant text.
- A.24 ‘MIB” includes item:
  - — MIB-36 ‘Is the PFC MIB module fully supported (per its MODULE COMPLIANCE)? status PFC AND MIB:O.
- A.33 ‘Priority-based Flow Control (PFC)’ details mandatory features for PFC, including ‘DCBX’, but is not referenced by the A.5 entry. ‘DCBX’ is defined in A.5 ‘Does the implementation support configuration management via DCBX?’ with status ‘O’, reference ‘Clause 38’.
- A.35 ‘Data Center Bridging eXchange protocol (DCBX)’ requires support of the ETS Recommendation TLV and Application VLAN TLV. Is it really the intent that PFC cannot be implemented in systems unless they can recognize individual higher layer application flows?

PFC Conformance (PICS—Annex B)

Currently in Annex B ‘PICS proforma—End station implementations’:

- B.5 ‘Major capabilities’ includes item:
  - ‘PFC’ ‘Is Priority-based Flow Control implemented’, status ‘O’ (optional) not qualified by any reference to DCBX, with references to 5.11,[1] Clause 36, and B.12.
- B.12 , B.13, and B.14 are just copies of A.33, A.34, and A35 respectively.

[1] The intent of 5.11 ‘System requirements for … PFC’ seems to have been to cover both Bridges and end-systems, where as 5.4.1.6 is Bridge specific. The split into Annex A and Annex B is artificial when dealing wih protocols for which Bridges and end-systems are peers, and leads to staright-forward duplication. Need to study whether this is a common problem, in which case cross-referencing would seem to be desirable rather than copying. Or perhaps a common category is desirable.

Annex D TLV Conformance

- D.1 ‘Requirements of the IEEE 802.1 Organizationally Specific TLV sets’:
  - — ‘TLVs may be supported’
  - — ‘If any … TLV set is supported, all … TLVs that are … members of that TLV set shall be supported’. So implementing PFC requires implementation of the ETS, Application Priority, and Application VLAN TLVs, and hence flow recognition and detailed bandwidth sharing between traffic classes.
- D.2 Organizationally Specific TLV definitions
  - — All TLVs in the dcbxSet (D.2.8, D.2.9, D.2.11, D.2.14) ‘may be transmitted’ so it is not clear what ‘shall be supported’ in D.1, ETS-5 in A.34 and B.13, and A.35 and B.14 actually require.

Clause 12. ‘Bridge management’

Explicit consideration of PFC is limited to 12.23: LinkDelayAllowance and counts of PFC Requests and PFC Indications (neither by PFC priority). Table 12-21 claims that these are required for Conformance, but that is only stated in a footnote to Table 12-21 (footnotes to Tables can be normative) there is no Clause 5 Conformance requirement for that in the absence of MIB support.

Effectiveness of PFC cannot be assessed, unless all priorities are subject to PFC (in which case PAUSE could have been used):

- 12.1.12 ‘Fault Management’ item b) 6) is ‘Determine whether a flow of data frames can be sent without error from a specific location within a network to a station or stations specified by the DA field of each frame.’
- 12.1.3 ‘Performance Management’ considers frame discard counts.
- 12.6.1.1.3 item e) counts discards by transmission port, but not by priority, and does not cover inbound discard prior to a forwarding decision.

Clause 17 MIBs

IEEE8021-PFC-MIB just covers LinkDelayAllowance and PFCRequests and PFCIndications counters.

17.3.17 does not clarify the relationship of PFC to the DCBX MIB, and appears to add no value. Its largely a repeat of non-PFC specific boilerplate, and ends ‘The relationship between IETF RFC 2863 and IETF RFC 3418 interfaces and ports is also described in previous subclauses of 17.3’ which does not refer to any specific previous subclauses, thus making the prior 9 lines redundant.

DCBX and LLDP

DCBX (Clause 38) uses LLDP (802.1AB) and the clause is written around the LLDP MIB module (38.4 … "DCBX state machine transitions are based on the DCBX objects in the LLDP MIB module").

The state machines are based on the use of three values for each independent DCBX attribute, to paraphrase:

— An Operational Parameter value, which is set to:

— A Local Admin Parameter value, if LLDP is not in receipt of a value from the Remote Peer or is not 'Willing' to accept the Remote Peer's Parameter value, or

— The Remote Peer's Parameter value (that peer's Local Admin value).[1]

The 802.1 LLDP extension MIB module version 2 includes

— …dcbxConfigPFCTable with per port …ConfigPFCTxEnable for PFC Configuration TLV tx

— …dcbxLocPFCTable with Willing, MBC, and per priority enable bits
corresponds to the Operational Parameter values (above, I believe)

— …dcbxRemPFCTable with Willing, MBC, and per priority enable bits[2]
corresponds to the Remote Parameter values (above)

— …dcbxAdminPFCTable with Willing, MBC, and per priority enable bits
corresponds to the Local Admin Parameter values (above)

So the MIB matches the Clause 38 description, and although support for LLDP for DCBX attributes is mandatory, It seems that PFC can be used without LLDP advertisements by setting the Admin parameters.

[1] At this, appropriate, level of description there is no effective diffrence between the Assymetric and Symmetric state machines, the latter just has to refine the notion of 'Willing' to tie-break (using MAC Address) for the case where both peers communicate 'Willing' in protocol.

[2] Confusingly described as containing information "on the local system" Q-2022-Rev-D1.3, pg 2236, line 37.

DCBX PFC priority selection (1)

Two coupled design choices:

1. LLDP, by design, does not explicitly acknowledge TLV receipt. PFC transmission for a given priority might be:
   a) made dependent on mutual agreement (exact match) of tx'd & rx'd PFC Config TLVs; or
   b) enabled for the current intersection of priorities in tx'd & rx'd PFC Config TLVs; or
   c) enabled for the priorities currently specified in tx'd PFC Config TLVs.
2. Does the PFC Config TLV reflect the intention of the transmitter:
   a) to send PFCs for the enabled priorities; or
   b) to respond to received PFCs, pausing transmission for the enabled priorities; or
   c) both a) and b).

Readers' assumptions may vary. Recommend explicit statement of 2.b). Not all network scenarios are symmetric between peers, but the PFC Config TLV is defined as using Symmetric attribute passing "the objective is that both peers use the same attribute value". 2.b) supports some asymmetric configuration (see algorithm details, to be provided).

Use of the Willing bit implies the possibility of configuration change and PFC Config TLV changes as LLDP transmission proceeds, PFC transmission has to recognize this possibility. The TLV transmitted "always carries the current local operational state" (38.4.2.2).
Assumption (needs to be stated): an LLDP participant will not change its transmitted PFC Config TLV as a result of LLDPDU reception unless it sets the Willing bit (W).
A tie-breaker is provided if both peers (PFC is constrained to a point-to-point link) set the Willing bit. If both set the bit, the configuration transmitted by the peer with the higher priority (numerically lower MAC Address takes precedence).

ieee802-dot1q-lldp-dcbx-tlv.yang

The IEEE 802.1/LLDP extension dcbxSet YANG model is based directly on the UML specified in IEEE Std 802.1AB and is summarized by the UML provided in D.6.2.3 of 802.1Q. It specifies just two sets of values (not three as per the extension MIB and the DCBX specification in Clause 38) — one for the local port, and one for (each) remote systems attached to and communicating with the port—so the DCBX logic in Clause 38 cannot be used without supplementary data or explanation.

The "/lldp:lldp/lldp:port" augmentation container "pfc-tlv-extension" are read-write. Assume they are the configured Local Admin Parameter values. Administrators can calculate the (possibly transient) Operational Parameter values, given Local Admin and Remote Parameter values (including whether the latter have been received), and Willing bit precedence rules.

All the TLVs in the dcbxSet are independently received. There are some cross-TLV dependencies: the number of classes for which PFC can be used is in the PFC TLV, but the impact on PFC priority use cannot be understood without receipt of the ETS TLV, which maps priorities to classes. This is not necessarily a deal breaker, but is worth noting. There should be rules as to what subset needs to be present if PFC use is to be dependent on receipt of any TLV.

DCBX use and operation

There are two ways we might proceed with DCBX:

1. Receipt of at least a Priority-based Flow Control Configuration TLV with a configuration that matches that of the receiver is required before PFC is used.[1]
2. PFC use is not in fact dependent on DCBX use.

[1] This matches comments in the Security WG to the effect that PFC should not/cannot be used before DCBX because of the risk of a ‘PFC storm’, in which PFCs are repeatedly and rapidly sent to a system that does not use PFC for the priority that has caused their transmission.