PFC Headroom Measurement and Calculation Project Proposal

Lily Lv

Paul Congdon

Need for The Project

- PFC is used in low-latency Ethernet data center networks to avoid packet loss.
- Deploying PFC today can be difficult
  - Manual configuration is complex and is different for each vendor solution
  - Consistent settings across a large-scale data center network is tedious
  - Vendor provided default values waste buffer resource, and do not work in certain circumstances (e.g. long distance data center interconnection)
- A standard is needed to specify any wire protocols (e.g. capability exchange) and a headroom measurement mechanism.

See: https://www.ieee802.org/1/files/public/docs2021/new-lv-adaptive-pfc-headroom-and-PTP-0602-v03.pdf

Proposed Scope of Work

- Amendment to 802.1Q with limited changes are needed to support the PFC configuration mechanism
  - Update DCBX to discover the capability and auto-enable the feature
  - Specify timestamp points
  - PTP based procedure to measure delay
  - State machines and protocol description
  - Updates to DCBX MIBs and YANG
  - Enhanced descriptions in Annex M & N

See: https://www.ieee802.org/1/files/public/docs2021/new-congdon-a-pfc-h-Q-changes-0521-v01.pdf

Market Potential

- The data center market continues to grow very fast:
  - New high-performance applications (AI/ML).
  - Desire to converge high-performance computing and high-performance storage on Ethernet.
  - Desire to scale HPC networks as a Cloud Service.
  - A consolidated network saves operational and equipment costs, fueling more growth.
- RDMA over Converged Ethernet (RoCEv2) is widely used:
  - Requires low latency.
  - Requires lossless operation to avoid retransmission.
  - Is deployed within data centers and across data center interconnections.
- Automating the configuration of PFC makes Ethernet technology more applicable for data center environments.

Technical Feasibility

- Proposed mechanism includes 2 major parts
  - Capability notification
  - PFC delay measurement
- Capability notification can be supported by extension of DCBX
- PFC delay measurement considers roundtrip delay between participating systems, which can be based on PTP peer-to-peer delay measurement mechanism.
- Both DCBX and PTP are mature technologies, which are currently available in production.

Technical Solutions to Explore

- Based on PTP peer-to-peer delay measurement, 3 technical solutions have been explored. See: https://www.ieee802.org/1/files/public/docs2021/new-lv-adaptive-pfc-headroom-and-PTP-0602-v03.pdf
  - Option 1: new message timestamp reference plane + PTP Pdelay messages
  - Option 2: new message timestamp reference plane + new MAC control frames
  - Option 3: new procedure for internal processing delay + PTP Pdelay messages
- The details and decisions of the technical solution are work for the project.
  - Consideration on co-existence with PTP
  - Consideration on support of two-step procedure only or both one-step and two-step procedures
  - Consideration on option 3 accuracy good enough?
  - ……other
- Further contributions are welcome.

Proposed Next Step

- Propose a motion to develop a PAR and CSD for pre-circulation before the November plenary.
- Sample Motion Slide to follow:

Motion

- 802.1 authorizes the September 2021 Interim to generate PAR and CSD for pre-circulation to the EC for an amendment to IEEE Std 802.1Q to specify a mechanism for PFC headroom measurement and calculation.

- Proposed: Lily Lv
- Second: Paul Congdon

- In the WG (y/n/a): <y>, <n>, <a>