TRILL Working Group
INTERNET-DRAFT
Intended status: Informational

Donald Eastlake
Huawei
Manoj Wadekar
QLogic
Anoop Ghanwani
Dell
Puneet Agarwal
Broadcom
Tal Mizrahi
Marvell

Expires: July 1, 2013

January 2, 2013

TRILL: Support of IEEE 802.1 Priority-based Flow Control and Enhanced Transmission Selection <draft-eastlake-trill-pfc-ets-00.txt>

Abstract

This document briefly explains the IEEE 802.1 Priority-based Flow Control and Enhanced Transmission standards and discusses the support of these standards in TRILL switches (devices that implement the IETF TRILL protocol standard).

Status of This Memo

Table of Contents

1. Introduction.....3
1.1 Overview of PFC and ETS.....4
1.2 Terminology.....4

2. Priority-Based Flow Control.....6
3. Enhanced Transmission Selection.....7
4. The DCB Exchange Protocol.....8

5. Management Considerations.....9
6. IANA Considerations.....9
7. Security Considerations.....9
8. References.....10
8.1 Normative References.....10
8.2 Informative References.....10

1. Introduction

IEEE 802.1 has developed various standards as part of its Data Center Bridging (DCB) activity. The intent of these standards is (1) to efficiently minimize data loss due to queue overflow for selected classes of traffic within Local Area Networks (LANs) meeting certain conditions and (2) to provide limited means to allocate the available bandwidth to different classes of traffic. Those standardes are Priority Based Flow Control (the IEEE [802.1Qbb] standard), Enhanced Tramission Selection (the IEEE [802.1Qaz] standard), and the Congestion Notification (CN) feature in the IEEE [802.1Q] standard. Intended uses include the support of loss sensitive services, such as Fiber Channel over Ethernet [FCoE], in data centers.

Because they are primarily implemented at the port level, no changes in the TRILL protocol are required to support PFC or ETS which are discussed in this document. CN support by TRILL may be considered in a separate document.

The existing optional PAUSE feature of IEEE 802.3 (Annex 31B of [802.3]) can, with appropriate engineering, also provide Ethernet service without loss of frames due to queue overflow. However, PAUSE has problems as follows:

1. Traffic for some protocols, for example TCP [RFC793], requires frame losses to signal congestion for flow control. Elimination of frame drops due to congestion would prevent TCP flow control, unless some other mechanism were added.

2. Some traffic consists of time critical network control frames, for example IS-IS Hellos [IS-IS]. PAUSE is relatively indiscriminant and pauses such frames, except for some MAC Control frames such as PAUSE control frames themselves, along with less critical traffic. Pausing such critical network control frames can compromise transport connectivity.

3. PAUSE can result in intermittent waves of spreading traffic paralysis, crippling network throughput, as follows: When a switch S1 receives a PAUSE on a port P1 and can no longer transmit frames out that port it is likely that output queues to P1 will fill up quickly. As soon as one output queue to P1 is full or almost so then, to avoid frame loss, S1 must send PAUSE frames out on each of its ports that might receive a frame for output to P1. For example, it might have to PAUSE input on P2 through P9, unnecessarily blocking traffic between any pair of those ports, to be sure it will not receive input on any of them for P1. This can repeat in switches connected to S1, switches connected to switches connected to S1, etc.

1.1 Overview of PFC and ETS

Overviews of the PFC and ETS standards covered herein are given below. IEEE 802.1 has specified these standards and the behavior needed to support them in bridges and end stations. This document discusses the support of these standards in TRILL switches [RFC6325].

IEEE [802.1Qbb], Priority-based Flow Control (PFC), provides a frame priority based refinement of the Ethernet PAUSE feature as described in Section 2. To the extent that a switch implements separate queues for different priorities at each port, this can eliminate the first and second of the PAUSE problems listed above. Traffic requiring frame drops due to congestion can be assigned a priority for which PFC is not enabled. PFC is not normally enabled for the two highest priorities, 6 and 7, which are typically used for time sensitive control frames. PFC also reduces the third problem as any congestion spreading would affect only priorities with PFC enabled.

IEEE [802.1Qaz] is a standard covering two things: One, Enhanced Transmission Selection (ETS), allocates bandwidth between traffic class groups indicated by priority. It is described in Section 3. Second, [802.1Qaz] contains the specification of the Data Center Bridging Exchange Protocol (DCBX) for discovering and configuring the three standards that this document covers, as described in Section 4.

PFC and ETS may be implemented independently or in any combination except that implementation of either of them implies implementation of DCBX, specified in IEEE [802.1Qaz].

1.2 Terminology

The following acronyms are used in this document in addition to those defined in [RFC6325].

AVB - Audio-Visual Bridging

CN - Congestion Notification [802.1Q]

DCB - Data Center Bridging [802.1Qaz]

DCBX - DCB Exchange protocol [802.1Qaz]

ETS - Enhanced Transmission Selection [802.1Qaz]

FCoE - Fiber Channel over Ethernet [FCoE]

LLDP - Link Layer Discovery Protocol (IEEE 802.1AB)

PFC - Priority-based Flow Control [802.1Qbb] [802.3bd]

RBridge - "Routing Bridge", an alternative name for a TRILL switch [RFC6325]

TRILL Switch - A device implementing the TRILL protocol [RFC6325]

2. Priority-Based Flow Control

IEEE [802.1Qbb], Priority-Based Flow Control (PFC), refines the IEEE [802.3] PAUSE feature to permit separately requesting the pausing and unpausing the traffic of each of the eight available frame priority levels. The actual priority-based pause control frame is specified in IEEE [802.3bd].

Such queue pausing occurs within the transmission logic associated with a port and requires no changes to the TRILL protocol, which is implemented above such port logic, as described in [RFC6325]. LLDP/DCBX is used in PFC discovery and agreement with peers as described in Section 4. A TRILL switch implementing the PFC standard should implement DCBX, signaling PFC support and configuration. Guarantee of lossless handling of frames with a particular priority in a TRILL campus requires implementation and enablement of PFC for that priority at all end stations that originate frames and all TRILL switches and bridges in that campus as well as meeting the PFC engineering requirements in [802.1Qbb].

The PFC control frames specified in [802.3bd] are MAC control frames that are not VLAN tagged. Their transmission normally bypasses the output queue at a port so they are transmitted immediately, or as soon as the frame currently being transmitted is sent, so as to meet the timing requirements of PFC.

3. Enhanced Transmission Selection

Enhanced Transmission Selection (ETS), specified in IEEE [802.1Qaz], allocates bandwidth, between traffic classes, through each of the ports of a switch or end station. (To be more precise, it modifies the algorithm used to select, from multiple priority-based output queues at a port, the next frame to transmit. Provision is made for proprietary algorithms and 802.1 has also specified an algorithm in connection with precise frame timing (AVB), but we are only concerned with the default algorithm.)

Transmission selection occurs within the logic associated with a port and requires no changes to the TRILL protocol, which is implemented above such port logic, as described in [RFC6325]. A TRILL switch implementing the ETS standard should implement DCBX (see Section 4) signaling of ETS support and configuration. For ETS to be effective, traffic in different ETS groups cannot share an output queue.

4. The DCB Exchange Protocol

The DCB Exchange Protocol (DCBX) is specified in IEEE [802.1Qaz], which also specifies ETS as described in Section 3.

DCBX is built on the Link Layer Discovery Protocol (LLDP), which is specified in IEEE [802.1AB]. DCBX is used for the discovery of DCB capabilities of peer switches, for the detection of inconsistent configuration of DCB features between peer switches, and for the propagation of DCB features to switches configured to accept configuration via DCBX. For purposes of TRILL protocol peering, TRILL switches ignore intervening bridges, but for the purposes of LLDP and DCBX all stations, including TRILL switches, 802.1 bridges, and end stations are considered peers.

TRILL switches implementing PFC or ETS should also implement DCBX.

5. Management Considerations

---TBD---

6. IANA Considerations

This document requires no IANA actions. This section should be deleted by the RFC Editor before publication.

7. Security Considerations

See [RFC6325] for general RBridge Security Considerations.

---more TBD---

8. References

Normative and informational references for this document are given below.

8.1 Normative References

As this is an informational document, there are no normative references.

8.2 Informative References

[IS-IS] - ISO/IEC, "Intermediate system to Intermediate system routeing information exchange protocol for use in conjunction with the Protocol for providing the Connectionless-mode Network Service (ISO 8473)", ISO/IEC 10589:2002.

[802.1AB] - IEEE, "IEEE Standard for Local and metropolitan area networks / Station and Media Access Control Connectivity Discovery", IEEE 802.1AB-2009, 17 September 2009.

[802.1Q] - IEEE, "IEEE Standard for Local and metropolitan area networks / Virtual Bridged Local Area Networks", IEEE 802.1Q-2011, May 2011.

[802.1Qaz] - IEEE, "Draft Standard for Local and Metropolitan Area Networks / Virtual Bridged Local Area Networks / Amendment XX: Enhanced Transmission Selection for Bandwidth Sharing Between Traffic Classes", IEEE Std 802.1Qaz-2011, June 2011.

[802.1Qbb] - IEEE, "Draft Standard for Local and Metropolitan Area Networks / Virtual Bridged Local Area Networks / Amendment: Priority-based Flow Control", IEEE Std 802.1Qbb-2011, June 2011.

[802.3] IEEE, "IEEE Standard for Information technology / Telecommunications and information exchange between systems / Local and metropolitan area networks / Specific requirements Part 3: Carrier sense multiple access with collision detection (CSMA/CD) access method and physical layer specifications", IEEE 802.3-2008, 26 December 2008.

[802.3bd] - IEEE 802.3, "Draft Standard for Information technology / Telecommunications and information exchange between systems / Local and Metropolitan Area Networks / Specific requirements Part 3: Carrier Sense Multiple Access with Collision Detection (CSMA/CD) Access Method and Physical Layer Specifications / Amendment: MAC Control Frame for Priority-based Flow Control", IEEE Std 802.3bd-2011, June 2011.

[FCoE] - http://fcoe.com/

[RFC793] - Postel, J., "Transmission Control Protocol", STD 7, RFC 793, September 1981

[RFC6325] - Perlman, R., Eastlake 3rd, D., Dutt, D., Gai, S., and A. Ghanwani, "Routing Bridges (RBridges): Base Protocol Specification", RFC 6325, July 2011.

Authors' Addresses

Donald Eastlake 3rd
Huawei Technologies
155 Beaver Street
Milford, MA 01757 USA

Tel: +1-508-333-2270
Email: d3e3e3@gmail.com

Manoj Wadekar
QLogic Corporation
26650 Aliso Viejo Pkwy
Aliso Viejo, CA 92656 USA

Tel: +1-949-389-6000
Email: manoj.wadekar@qlogic.com

Anoop Ghanwani
Dell
350 Holger Way
San Jose, CA 95134 USA

Phone: +1-408-571-3500
Email: anoop@alumni.duke.edu

Puneet Agarwal
Broadcom
3975 Freedom Circle
Santa Clara, CA 95054 USA

Phone: +1-949-926-5000
Email: pagarwal@broadcom.com

Tal Mizrahi
Marvell
6 Hamada Street
Yokneam, 20692 Israel

Email: talmi@marvell.com

Copyright, Disclaimer, and Additional IPR Provisions