

PWE3
Internet-Draft
Intended status: Standards Track
Expires: January 2, 2010

Y(J). Stein
RAD Data Communications
July 1, 2009

Ethernet PW Congestion Handling Mechanisms
draft-stein-pwe3-ethpwcong-00.txt

Status of this Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of [BCP 78](#) and [BCP 79](#).

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at <http://www.ietf.org/ietf/1id-abstracts.txt>.

The list of Internet-Draft Shadow Directories can be accessed at <http://www.ietf.org/shadow.html>.

This Internet-Draft will expire on January 2, 2010.

Copyright Notice

Copyright (c) 2009 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to [BCP 78](#) and the IETF Trust's Legal Provisions Relating to IETF Documents in effect on the date of publication of this document (<http://trustee.ietf.org/license-info>). Please review these documents carefully, as they describe your rights and restrictions with respect to this document.

Abstract

Mechanisms for handling congestion in Ethernet pseudowires are presented. These mechanisms extend capabilities of the native service across the PSN, and require use of the PWE3 control word.

Internet-Draft

ethpwcong

July 2009

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC 2119](#) [[RFC2119](#)].

Table of Contents

1.	Introduction	3
2.	Control Word Format	3
3.	Drop Eligibility Indication	4
4.	Explicit Congestion Notification	5
5.	Security Considerations	6
6.	IANA Considerations	6
7.	References	7
7.1.	Normative References	7
7.2.	Informative References	7
	Author's Address	7

Internet-Draft

ethpwcong

July 2009

1. Introduction

Ethernet PWs do not presently have mechanisms for handling AC congestion. When the egress AC becomes congested, the egress PE will receive PAUSE (802.3x) frames or experiences back-pressure, denying it the capability of forwarding frames to the AC. This will result in the egress PE's output buffers filling up and eventually Ethernet frames will need to be discarded. Not only are such frames lost after precious PSN bandwidth has already been consumed, they are also discarded without regard to importance, priority, or fairness.

If the Ethernet frames being transported are carrying TCP/IP traffic, then TCP rate cut-back will limit the traffic volume to some extent. However, the early discard that triggers the rate cut-back also results in packet retransmission, adding additional Ethernet PW traffic to be transported. When the Ethernet frames are not carrying TCP/IP, but rather UDP/IP, or any other non-TCP/IP traffic that does not react to packet discard by cutting back the transmission rate, the situation is potentially worse.

The native Ethernet service handles congestion by causing the sender to stop sending frames. On full duplex links this is accomplished by the congested receiver sending PAUSE frames. On half-duplex networks this is accomplished by the congested receiver introducing back-pressure. In either case the effect is that the sender stops forwarding frames until the receiver is once again ready to process them, thus eliminating congestion.

Ethernet PWs do not transport received congestion indications across the PSN, nor do they generate congestion indications when the egress PE detects congestion.

It is possible to rectify this lack of functionality by adding indications in the PWE control word. The arbitrariness of the packets discarded can be alleviated by including a drop eligibility indication. The loss itself can be possibly avoided by mechanisms

that explicitly indicate forward and backward congestion. Such indications enable a PE to reflect the egress AC congestion status back towards the ingress AC, where steps can be taken to limit the ingress rate, thus avoiding buffer overflow.

2. Control Word Format

The mechanisms described herein are only available when the Ethernet PW employs the PWE3 control word. Thus, when congestion handling is support the control word MUST be included in the PW packet. The use of the control word is usually signaled using the PWE3 control

[Page 3]

July 2009

protocol [RFC4447]. There is no need to additionally signal the use of the mechanisms described herein, as the default actions suffice.

The format of the control word is given in Figure 1 and has been chosen to be compatible with that of [RFC 4619](#) [[RFC4619](#)].

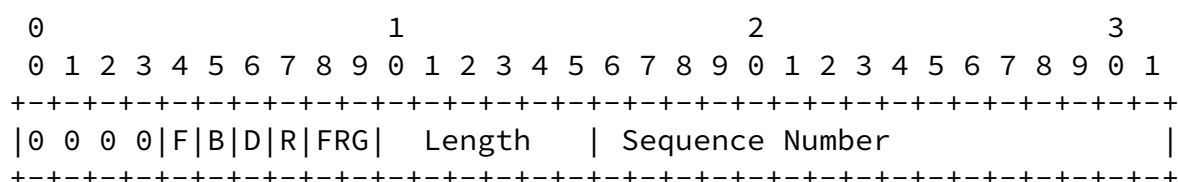


Figure 1. Control Word structure

Bits 0 to 3 In the above diagram, the first 4 bits MUST be set to 0.

F (bit 4) Forward Explicit Congestion Notification (FECN) bit.

B (bit 5) Backward Explicit Congestion Notification (BECN) bit.

D (bit 6) Discard Eligibility Indication (DEI) bit.

R (bit 7) RESERVED bit.

FRG (bits 8 and 9) described in RFC 4623 [RFC4623].

Length (bits 10 - 15) described in [RFC 4385](#) [RFC4385].

Sequence Number (bits 16 - 31) described in RFC 4385 [RFC4385] and

service specific encapsulation documents.

[3.](#) Drop Eligibility Indication

If drop eligibility is supported, then the ingress PE MUST set the Drop Eligibility Indicator (DEI) bit in the PWE3 control word, and during congestion the egress PE MUST preferentially discard Ethernet frames that arrived in PW packets with the DEI bit set.

When the ingress PE receives a Q-in-Q Ethernet frame from the AC, it MUST copy the DEI bit from the Ethernet frame into the DEI bit in the PWE3 control word.

The ingress PE SHOULD perform the MEF bandwidth profile (token bucket) algorithm [[MEF10.1](#)]. Frames marked red MUST be discarded, and green and yellow frames MUST be encapsulated and forwarded. For yellow frames the ingress PE MUST set the DEI bit in the PWE control word.

Stein

Expires January 2, 2010

[Page 4]

Internet-Draft

ethpwcong

July 2009

Intermediate network elements MUST NOT clear the DEI bit.

Intermediate PW-aware network elements (e.g., S-PEs) MAY set the DEI bit upon experiencing congestion, if they run the MEF BW profile (token bucket) algorithm.

When the egress PE needs to discard an Ethernet frame, it MUST discard packets with the DEI bit set before discarding packets with the DEI bit cleared.

When the egress PE forwards Q-in-Q Ethernet frame to the AC, it MUST copy the DEI bit from the PWE control word into the DEI bit in the Ethernet frame.

[4.](#) Explicit Congestion Notification

If explicit congestion notification is supported, then the egress PE MUST make the ingress PE aware of the congestion experienced, and the ingress PE MAY make the egress PE aware of such congestion. An ingress PE being informed of congestion by the egress PE SHOULD take steps to alleviate this congestion.

If the egress PE receives PAUSE frames or detects Ethernet back-pressure or detects that its AC-bound queues pass a preconfigured threshold, then it MUST set the BECN bit in the PWE control word of all PW packets set in the opposite direction towards the ingress PE. If no packets are available for sending in the backward direction, the egress PE MUST send dummy BECN PW packets towards the ingress PE at a preconfigured rate (default is one per second). These dummy BECN packets have their BECN bit set, their length field set to zero, but contain no data.

When the egress PE PAUSE timer expires, or it detects that back-pressure that had been applied has been removed, or its AC-bound queues drop below a preconfigured threshold, it MUST clear the BECN bit of all PW packets set towards the ingress PE. If no packets are available for sending in the backward direction, the egress PE MUST send three dummy BECN PW packets towards the ingress PE at a preconfigured rate (default is one per second). These dummy BECN packets have their BECN bit cleared, their length field set to zero, but contain no data.

Intermediate network elements MUST NOT clear the BECN bit. Intermediate PW-aware network elements (e.g., S-PEs) upon experiencing congestion MAY set the BECN bit on packets forwarded in the opposite direction.

When the ingress PE receives packets with the BECN bit set (including

dummy BECN packets). it SHOULD perform one of the following operations to ameliorate the situation.

It SHOULD send PAUSE packets or apply backpressure towards the ingress AC.

If its Ethernet interface does not support PAUSE or back-pressure, it SHOULD apply the MEF bandwidth profile algorithm to frames received from the AC before sending them towards the PSN.

If the ingress PE has admission control functionality, it SHOULD refuse further connections with traffic that would be forwarded to the egress PE, and MAY withdraw low priority connections.

If the ingress PE detects that its output queues pass a preconfigured

threshold, then it SHOULD send PAUSE frames or apply back-pressure to the AC. It SHOULD also set the FECN bit in the PWE control word of all PW packets set towards the egress PE, in order to inform the egress PE to expect delays.

Intermediate network elements MUST NOT clear the FECN bit. Intermediate PW-aware network elements (e.g., S-PEs) MAY set the FECN bit upon experiencing congestion in the forward direction.

If packets with FECN set have been send, then when the ingress PE sees that its PSN-bound queues drop below a preconfigured threshold, it MUST clear the FECN bit of all PW packets sent towards the egress PE. If no packets are available for sending in the forward direction, the ingress PE MUST send three dummy FECN PW packets towards the egress PE at a preconfigured rate (default is one per second). These dummy BECN packets have their FECN bit cleared, their length field set to zero, but contain no data.

[5.](#) Security Considerations

The congestion handling mechanisms introduced here do not introduce significant security considerations above those present for PWs that do not use these mechanisms. For example, a denial of service attack based on forcing the ingress PE to slow down would require the ability to inject otherwise valid PW packets. A malicious entity that has attained that level has already breached the fundamental security of the PW infrastructure.

[6.](#) IANA Considerations

This document requires no IANA actions.

[7.](#) References

[7.1.](#) Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", [BCP 14](#), [RFC 2119](#), March 1997.

[RFC4385] Bryant, S., Swallow, G., Martini, L., and D. McPherson,

"Pseudowire Emulation Edge-to-Edge (PWE3) Control Word for Use over an MPLS PSN", [RFC 4385](#), February 2006.

[RFC4623] Malis, A. and M. Townsley, "Pseudowire Emulation Edge-to-Edge (PWE3) Fragmentation and Reassembly", [RFC 4623](#), August 2006.

[MEF10.1] "MEF Technical Specification MEF 10.1 - Ethernet Service Attributes Phase 2", Metro Ethernet Forum MEF 10.1, November 2006.

[7.2.](#) Informative References

[RFC4447] Martini, L., Rosen, E., El-Aawar, N., Smith, T., and G. Heron, "Pseudowire Setup and Maintenance Using the Label Distribution Protocol (LDP)", [RFC 4447](#), April 2006.

[RFC4619] Martini, L., Kawa, C., and A. Malis, "Encapsulation Methods for Transport of Frame Relay over Multiprotocol Label Switching (MPLS) Networks", [RFC 4619](#), September 2006.

Author's Address

Yaakov (Jonathan) Stein
RAD Data Communications
24 Raoul Wallenberg St., Bldg C
Tel Aviv 69719
ISRAEL

Phone: +972 3 645-5389
Email: yaakov_s@rad.com