PALS Working Group
Internet-Draft
Updates: 4448 (if approved)
Intended status: Standards Track
Expires: February 3, 2018

S. Bryant
A. Malis
Huawei
I. Bagdonas
Equinix

August 02, 2017

Use of Ethernet Control Word RECOMMENDED draft-bryant-pals-ethernet-cw-01

Abstract

The pseudowire (PW) encapsulation of Ethernet, as defined in RFC4448, specifies that the use of the control word (CW) is optional. In the absence of the CW an Ethernet pseudowire packet can be misidentified as an IP packet by a label switching router (LSR). This in turn may lead to the selection of the wrong equal-cost-multi-path (ECMP) path for the packet, leading in turn to the mis-ordering of packets. This problem has become more serious due to the deployment of equipment with Ethernet MAC addresses that start with 0x4 or 0x6. The use of the Ethernet PW CW addresses this problem. This document recommends the use of the Ethernet pseudowire control word in all but exceptional circumstances.

This document updates RFC4448.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

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."

This Internet-Draft will expire on February 3, 2018.

Copyright Notice

Copyright (c) 2017 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 (http://trustee.ietf.org/license-info) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction

The pseudowire(PW) encapsulation of Ethernet, as defined in RFC4448, specifies that the use of the control word (CW) is optional. It is common for label switching routers (LSRs) to search past the end of the label stack to determine whether the payload is an IP packet, and if the payload is an IP packet, to select the next hop based of the so called "five-tuple" (IP source address, IP destination address, protocol/next-header, transport layer source port and transport layer destination port). In the absence of a PW CW an Ethernet pseudowire packet can be misidentified as an IP packet by a label switching router (LSR) selecting the ECMP path based on the five-tuple. This in turn may lead to the selection of the wrong equal-cost-multi-path (ECMP) path for the packet, leading in turn to the mis-ordering of packets. Further discussion of this topic is published in [RFC4928].

Flow misordering can also happen in a single path scenario when traffic classification and differential forwarding treatment mechanisms are in use. This occurs when a forwarder incorrectly assumes that the packet is IP and applies forwarding policy based on fields in the PW payload.

This problem has recently become more serious for a number of reasons. Firsly due to the deployment of equipment with Ethernet MAC addresses that start with 0x4 or 0x6 assigned by the IEEE RAC. Secondly, concerns over privacy have led to the use of MAC address randomization which assigns local MAC addresses randomly for privacy. Random assignmen produce addresses starting with one of the two values about 1/8 of the time.

The use of the Ethernet PW CW addresses this problem.

This document recommends the use of the Ethernet pseudowire control word in all but exceptional circumstances.

2. Specification of Requirements

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].

3. Background

Ethernet pseudowire encapsulation is specified in [RFC4448]. In particular the reader is drawn to section 4.6, part of which is quoted below for the convenience of the reader:

> "The control word defined in this section is based on the Generic PW MPLS Control Word as defined in [RFC4385]. It provides the ability to sequence individual frames on the PW, avoidance of equal-cost multiple-path load-balancing (ECMP) [RFC2992], and Operations and Management (OAM) mechanisms including VCCV [RFC5085].
>
> "[RFC4385] states, "If a PW is sensitive to packet misordering and is being carried over an MPLS PSN that uses the contents of the MPLS payload to select the ECMP path, it MUST employ a mechanism which prevents packet misordering." This is necessary because ECMP implementations may examine the first nibble after the MPLS label stack to determine whether the labeled packet is IP or not. Thus, if the source MAC address of an Ethernet frame carried over the PW without a control word present begins with 0x4 or 0x6, it could be mistaken for an IPv4 or IPv6 packet. This could, depending on the configuration and topology of the MPLS network, lead to a situation where all packets for a given PW do not follow the same path. This may increase out-of-order frames on a given PW, or cause OAM packets to follow a different path than actual traffic (see Section 4.4.3, "Frame Ordering").
>
> "The features that the control word provides may not be needed for a given Ethernet PW. For example, ECMP may not be present or active on a given MPLS network, strict frame sequencing may not be required, etc. If this is the case, the control word provides little value and is therefore optional. Early Ethernet PW implementations have been deployed that do not include a control word or the ability to process one if present. To aid in backwards compatibility, future implementations MUST be able to send and receive frames without the control word present."

At the time when pseudowires were first deployed, some equipment of commercial significance was unable to process the Ethernet Control Word. In addition, at that time it was considered that no Ethernet MAC address had been issued by the IEEE Registration Authority Committee (RAC) that starts with 0x4 or 0x6, and thus it was thought to be safe to deploy Ethernet PWs without the CW.

Since that time the RAC has issued Ethernet MAC addresses start with 0x4 or 0x6 and thus the assumption that in practical networks there would be no confusion between an Ethernet PW packet without the CW and an IP packet is no longer correct.

Possibly through the use of unauthorized Ethernet MAC addresses, this assumption has been unsafe for a while, leading some equipment

vendors to implement more complex, proprietary, methods to discriminate between Ethernet PW packets and IP packets. Such mechanisms rely on the heuristics of examining the transit packets in trying to find out the exact payload type of the packet and cannot be reliable due to the random nature of the payload carried within such packets.

A recent posting on the Nanog email list has highlighted this problem:

https://mailman.nanog.org/pipermail/nanog/2016-December/089395.html

RFC EDITOR Please delete this paragraph.
Kramdown does not include references when they are only found in literal text so I include them here: [RFC4385] [RFC2992] [RFC5085] as a fixup.

4. Recommendation

The ambiguity between an MPLS payload that is a Ethernet PW and one that is an IP packet is resolved when the Ethernet PW control word is used. This document updates RFC4448 [RFC4448] to state that where both both the ingress PE and the egress PE support the Ethernet pseudowire control word, then the CW MUST be used.

5. Equal Cost Multi-path (ECMP)

Where the volume of traffic on an Ethernet PW is such that ECMP is required then one of two methods may be used:

o Flow-Aware Transport (FAT) of Pseudowires over an MPLS Packet Switched Network specified in [RFC6391], or

o LSP entropy labels specified [RFC6790]

RFC6391 works by increasing the entropy of the bottom of stack label. It requires that both the ingress and egress provider edge (PE)s support this feature. It also requires that sufficient LSRs on the LSP between the ingress and egress PE be able to select an ECMP path on an MPLS packet with the resultant stack depth.

RFC6790 works by including an entropy value in the LSP part of the label stack. This requires that the Ingress and Egress PEs support the insertion and removal of the entropy label (EL) and the entropy label indicator, and that sufficient LSRs on the LSP are able to preform ECMP based on the EL.

In both cases there considerations in getting Operations, Administration, and Maintenance (OAM) packets to follow the same path as a data packet. This is described in detail section 7 of [RFC6391], and section 6 of RFC6790. However in both cases the situation is improved compared to the ECMP behavior in the case where the Ethernet PW CW was not used, since there is currently no known method of getting a PW OAM packet to follow the same path as a PW data packet subjected to ECMP based on the five tuple of the IP payload.

6. Mitigations

Where it is not possible to use the Ethernet PW CW, the effects of ECMP can be disabled by carrying the PW over a traffic engineered path that does not subject the payload to load balancing (for example [RFC3209]. However such paths may be subjected to link bundle load balancing and of course the single LSP has to carry the full PW load.

7. Operational Considerations

CW presence on the PW is controlled by the configuration and may be subject to default operational mode of not being enabled. Care needs to be taken to ensure that software that implements this recommendation does not depend on existing configuration setting that prevents the use of control word. It is recommended that platform software emits a rate limited message indicating that CW can be used but is disabled due to existing configuration.

To remove this problem in the long term, and hence to reduce the operational cost of investigating problems associated with the incorrect forwarding of Ethernet packets over PWs not using the CW, it is RECOMMENDED that equipment that does not support the CW be phased out of operational use.

8. Security Considerations

This document expresses a preference for one existing and widely deployed Ethernet PW encapsulation over another. These methods have identical security considerations, which are discussed in [RFC4448]. This document introduces no additional security issues.

9. IANA Considerations

This document makes no IANA requests.

10. Acknowledgements

The authors thank Job Snijders for drawing attention to this problem. The authors also thank Pat Thaler for clarifying the matter of local MAC address assignment.

11. References

11.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <http://www.rfc-editor.org/info/rfc2119>.

[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, DOI 10.17487/RFC4385, February 2006, <http://www.rfc-editor.org/info/rfc4385>.

[RFC4448] Martini, L., Ed., Rosen, E., El-Aawar, N., and G. Heron, "Encapsulation Methods for Transport of Ethernet over MPLS Networks", RFC 4448, DOI 10.17487/RFC4448, April 2006, <http://www.rfc-editor.org/info/rfc4448>.

[RFC4928] Swallow, G., Bryant, S., and L. Andersson, "Avoiding Equal Cost Multipath Treatment in MPLS Networks", BCP 128, RFC 4928, DOI 10.17487/RFC4928, June 2007, <http://www.rfc-editor.org/info/rfc4928>.

[RFC6391] Bryant, S., Ed., Filsfils, C., Drafz, U., Kompella, V., Regan, J., and S. Amante, "Flow-Aware Transport of Pseudowires over an MPLS Packet Switched Network", RFC 6391, DOI 10.17487/RFC6391, November 2011, <http://www.rfc-editor.org/info/rfc6391>.

[RFC6790] Kompella, K., Drake, J., Amante, S., Henderickx, W., and L. Yong, "The Use of Entropy Labels in MPLS Forwarding", RFC 6790, DOI 10.17487/RFC6790, November 2012, <http://www.rfc-editor.org/info/rfc6790>.

11.2. Informative References

[RFC2992] Hopps, C., "Analysis of an Equal-Cost Multi-Path Algorithm", RFC 2992, DOI 10.17487/RFC2992, November 2000, <http://www.rfc-editor.org/info/rfc2992>.

[RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, DOI 10.17487/RFC3209, December 2001, <http://www.rfc-editor.org/info/rfc3209>.

[RFC5085] Nadeau, T., Ed. and C. Pignataro, Ed., "Pseudowire Virtual Circuit Connectivity Verification (VCCV): A Control Channel for Pseudowires", RFC 5085, DOI 10.17487/RFC5085, December 2007, <http://www.rfc-editor.org/info/rfc5085>.

Authors' Addresses

Stewart Bryant
Huawei

Email: stewart.bryant@gmail.com

Andrew G Malis
Huawei

Email: agmalis@gmail.com

Ignas Bagdonas
Equinix

Email: ibagdona.ietf@gmail.com>