Recommendation to Use the Ethernet Control Word

Abstract

   The pseudowire (PW) encapsulation of Ethernet, as defined in
   RFC 4448, specifies that the use of the control word (CW) is
   optional. In the absence of the CW, an Ethernet PW packet can be
   misidentified as an IP packet by a label switching router (LSR).
   This may lead to the selection of the wrong equal-cost multipath
   (ECMP) path for the packet, leading in turn to the misordering of
   packets. This problem has become more serious due to the deployment
   of equipment with Ethernet Media Access Control (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 PW CW in
   all but exceptional circumstances.

   This document updates RFC 4448.

Copyright Notice

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 then, if it is, select the next hop based on 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 PW
   packet can be misidentified as an IP packet by a label switching
   router (LSR) selecting the ECMP path based on the five-tuple. This
   may lead to the selection of the wrong ECMP path for the packet,
   leading in turn to the misordering 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. These errors occur when a forwarder
   incorrectly assumes that the packet is IP and applies a forwarding
   policy based on fields in the PW payload.

   IPv4 and IPv6 packets start with the values 0x4 and 0x6,
   respectively. Misidentification can arise if an Ethernet PW packet
   without a CW is carrying an Ethernet packet with a destination
   address that starts with either of these values.

   This problem has recently become more serious for a number of
   reasons. First, the IEEE Registration Authority Committee (RAC) has
   assigned Ethernet MAC addresses that start with 0x4 or 0x6, and
   equipment that uses MAC addresses in these series has been deployed
   in networks. Second, concerns over privacy have led to the use of
   MAC address randomization, which assigns local MAC addresses randomly
   for privacy. Random assignment results in addresses starting with
   one of these two values approximately one time in eight.

   The use of the Ethernet PW CW addresses this problem.

   This document RECOMMENDS the use of the Ethernet PW CW in all but
   exceptional circumstances.

2. Specification of Requirements

   The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT",
   "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and
   "OPTIONAL" in this document are to be interpreted as described in
   BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all
   capitals, as shown here.

3. Background

   Ethernet PW encapsulation is specified in [RFC4448]. Of particular
   relevance is Section 4.6, part of which is quoted below for the
   convenience of the reader. Note that RFC 4448 uses the citation
   [PWE3-CW] to refer to [RFC4385] and the citation [VCCV] to refer to
   the document that was eventually published as [RFC5085].

      The control word defined in this section is based on the Generic
      PW MPLS Control Word as defined in [PWE3-CW]. 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 [VCCV].

      [PWE3-CW] 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.

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

Since that time, the RAC has issued Ethernet MAC addresses that start
with 0x4 or with 0x6. Therefore, 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 to
try 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 posting on the NANOG email list highlighted this problem:

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

4. Recommendation

The ambiguity between an MPLS payload that is an Ethernet PW and one
that is an IP packet is resolved when the Ethernet PW CW is used.
This document updates [RFC4448] to state that both the ingress
provider edge (PE) and the egress PE SHOULD support the Ethernet PW
CW and that, if supported, the CW MUST be used.

Where the application of ECMP to Ethernet PW traffic is required and
where both the ingress and the egress PEs support Entropy Label
Indicator / Entropy Label (ELI/EL) [RFC6790] and Flow-Aware Transport
of Pseudowires (FAT PW) [RFC6391], then either method may be used.
The use of both methods on the same PW is not normally necessary and
should be avoided unless circumstances require it. In the case of
multi-segment PWs, if ELI/EL is used, then it SHOULD be used on every
segment of the PW. The method by which usage of ELI/EL on every
segment is guaranteed is out of the scope of this document.

5. Equal-Cost Multipath (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 of Pseudowires over an MPLS Packet Switched
   Network, specified in [RFC6391], or

o Label Switched Path (LSP) entropy labels, specified in [RFC6790].

RFC 6391 works by increasing the entropy of the bottom-of-stack
label. It requires that both the ingress and egress PEs 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.

RFC 6790 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 EL and the ELI and that sufficient
LSRs on the LSP are able to perform ECMP based on the EL.

In both cases, there are considerations in getting Operations,
Administration, and Maintenance (OAM) packets to follow the same path
as a data packet. This is described in detail in 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.

The PW label is pushed before the LSP label. As the ELI/EL labels
are part of the LSP layer rather than part of the PW layer, they are
pushed after the PW label has been pushed.

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, RSVP-TE [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

In some cases, the inclusion of a CW in the PW is determined by
equipment configuration. Furthermore, it is possible that the
default configuration in such cases is to disable use of the CW.
Care needs to be taken to ensure that software that implements this
recommendation does not depend on existing configuration settings
that prevent the use of a CW. It is recommended that platform
software emit a rate-limited message indicating that the CW can be
used but is disabled due to existing configuration.

Instead of including a payload type in the packet, MPLS relies on the
control plane to signal the payload type that follows the bottom of
the label stack. Some LSRs attempt to deduce the packet type by MPLS

payload inspection, in some cases looking past the PW CW. If the
payload appears to be IP or IP carried in an Ethernet header, they
perform an ECMP calculation based on what they assume to be the
five-tuple fields. However, deduction of the payload type in this
way is not an exact science, and where a packet that is not IP is
mistaken for an IP packet, the result can be packets delivered out of
order. Misordering of this type can be difficult for an operator to
diagnose. When enabling capability that allows information gleaned
from packet inspection past the PW CW to be used in any ECMP
calculation, operators should be aware that this may cause Ethernet
frames to be delivered out of order despite the presence of the CW.

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 has no IANA actions.

10. References

10.1. 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,
            <https://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, <https://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,
            <https://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,
            <https://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,
              <https://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,
              <https://www.rfc-editor.org/info/rfc6790>.

   [RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC
              2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174,
              May 2017, <https://www.rfc-editor.org/info/rfc8174>.

10.2. Informative References

   [RFC2992] Hopps, C., "Analysis of an Equal-Cost Multi-Path
              Algorithm", RFC 2992, DOI 10.17487/RFC2992, November 2000,
              <https://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,
              <https://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, <https://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>