Use of Ethernet Control Word RECOMMENDED

draft-bryant-pals-ethernet-cw-00

Stewart Bryant, Andy Malis & Ignas Bagdonas

History of the Design

- When PWs were first deployed, some equipment of commercial significance was unable to process the Ethernet Control Word.
- At that time no Ethernet MAC address had been issued by the IEEE Registration Authority Committee (RAC) that started with 0x4 or 0x6,
- Considered safe to deploy Ethernet PWs without the CW.
- Thus the CW for an Ethernet PW is OPTIONAL.

What Has Happened Since?

- The IEEE RAC has since issued Ethernet MAC addresses start with 0x4 or 0x6 invalidating the assumption that in practical networks there would be no confusion between an Ethernet PW packet without the CW and an IP packet.
- Possibly through the use of unauthorized Ethernet MAC addresses, this assumption has been unsafe for a while.
- Some equipment implement more complex, proprietary, methods to discriminate between Ethernet PW packets and IP packets. Such mechanisms rely on heuristics to determine the payload type but these can be unreliable.
- Operators see misordering and drops of packets in the network and have expressed concern.

Why Does the Misordering Happen?

- Network Operators configure load balancing (ECMP) in their networks to distribute the load across a number of links to minimise congestion.
- A popular method load balancing is to (illegally in protocol terms) look below the MPLS label stack, verify that the payload is IP (from the first nibble which is the IP version number) and then identify the flow based on the “five tuple”.
- When the packet is not IP, but starts with 0x4 or 0x6, it can be misidentified as IP. The packet may then be load balanced on whatever fields are in the same place as the five tuple elements would have been.
- The packet is then ECMPed on a “random”, possibly inconsistent, ECMP path.

Why is misordering a problem?

- An invariant of Ethernet was that packets would arrive in order – before we invented bridges, Ethernet was simple a wire.
- Some protocols assume this ordering invariant.
- Those protocols do not work as well is the invariant is not upheld.

How does the PW CW prevent this?

- The first nibble of the PW CW is zero.
- An MPLS payload starting with a PW CW can never be confused with an IP packet. This disambiguation was exactly what the PW CW was designed to do.

Why don’t we mandate the CW?

- The PWE3 WG looked at this in the past, but was concerned about the deployed equipment.
- There is a huge deployment of PW PEs.
- Mandating that the PW CW is used is impractical due to scale and expense.
- … but we can start the migration process.

Technical 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 therefore:

This document updates RFC4448 [RFC4448] to state that where both the ingress PE and the egress PE support the Ethernet pseudowire control word, then the CW MUST be used.

Question for the WG

Should we say:

> where both the ingress PE and the egress PE support the Ethernet pseudowire control word, then the CW MUST be used.

Or:

> where both the ingress PE and the egress PE support the Ethernet pseudowire control word, then it is RECOMMENDED that the CW be used.

Side Effects of this Recommendation

- CW increases packet size by 4 octets. There are deployments where MTU size is tightly controlled and has unmovable upper bounds.

- There are middleboxes that analyse transit PW traffic, are not aware of PW setup signalling and make decisions about packet format and structure based on the assumption that there is no CW at all.

What if we actually want to ECMP a PW?

- Get the PW to provide the flow identification to the underlay via:
  - FAT PWs – (RFC6391) or
  - Having the ingress PE insert an LSP Entropy Label (RFC6790)

Process Recommendation

- An experienced document shepherd is selected.
- We start the formal adoption process as soon as the meeting ends (the draft is not perfect, but it is good enough to be taken into WG control).
- Aim for this to be with the IESG by Singapore.

Why the rush?

- A number of operators have raised the concern on NOG mailing lists.
- The IEEE liaison raised it with the IESG.
- Our AD has asked us to fix the problem.
- The fix is simple, the time will be taken by description of the context in the RFC and the RFC process.