INTERNET-DRAFT
Intended Status: Standard Track
Updates: 4761

K. Patel
Arrcus
S. Boutros
VMware
J. Liste
Cisco
B. Wen
Comcast
J. Rabadan
Nokia

Expires: September 3, 2018
March 2, 2018

Extensions to BGP Signaled Pseudowires to support Flow-Aware Transport Labels

draft-ietf-bess-fat-pw-bgp-04

Abstract

This draft defines protocol extensions required to synchronize flow label states among Provider Edges PE(s) when using the BGP-based signaling procedures. These protocol extensions are equally applicable to point-to-point Layer2 Virtual Private Networks (L2VPNs). This draft updates RFC 4761 by defining new flags in the Control Flags field of the Layer2 Info Extended Community.

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/1id-abstracts.html

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

Copyright and License Notice

Copyright (c) 2018 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 mechanism described in [RFC6391] uses an additional label (Flow Label) in the MPLS label stack to allow Label Switch Routers to balance flows within Pseudowires at a finer granularity than the individual Pseudowires across the Equal Cost Multiple Paths (ECMPs) that exists within the Packet Switched Network (PSN).

Furthermore, [RFC6391] defines the LDP protocol extensions required to synchronize the flow label states between the ingress and egress PEs when using the signaling procedures defined in the [RFC8077].

A pseudowire (PW) [RFC3985] is transported over one single network path, even if Equal Cost Multiple Paths (ECMPs) exist between the ingress and egress PW provider edge (PE) equipment. This is required to preserve the characteristics of the emulated service.

This draft introduces an optional mode of operation allowing to transport a PW over ECMPs, for example when the use of these is known to be beneficial to the operation of the PW. This specification uses the principles defined in [RFC6391], and augments the BGP-signaling procedures of [RFC4761] and [RFC6624]. The use of a single path to preserve the packet delivery order remains the default mode of operation of a PW, and is described in [RFC4385] and [RFC4928].

High bandwidth Ethernet-based services are a prime example that benefits from the ability to load-balance flows in a PW over multiple PSN paths. In general, load-balancing is applicable when the PW attachment circuit bandwidth and PSN core link bandwidth are of same order of magnitude.

To achieve the load-balancing goal, [RFC6391] introduces the notion of an additional Label Stack Entry (LSE) (Flow label) located at the bottom of the stack (right after PW LSE). Label Switching Routers (LSRs) commonly generate a hash of the label stack in order to discriminate and distribute flows over available ECMPs. The presence of the Flow label (closely associated to a flow determined by the ingress PE) will normally provide the greatest entropy.

Furthermore, following the procedures for Inter-AS scenarios described in [RFC4761] section 3.4, the Flow label should never be handled by the ASBRs, only the terminating PEs on each AS will be responsible for popping or pushing this label. This is equally applicable to Method B [RFC4761] section 3.4.2 where ASBRs are responsible for swapping the PW label as traffic traverses from ASBR to PE and ASBR to ASBR directions. Therefore, the Flow label will remain untouched across AS boundaries.

1.1 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].

2. Modifications to Layer2 Info Extended Community

The Layer2 Info Extended Community is used to signal control information about the pseudowires to be setup. The extended community format is described in [RFC4761]. The format of this extended community is described as:

```
+------------------------------------+
| Extended community type (2 octets) |
+------------------------------------+
|  Encaps Type (1 octet)             |
+------------------------------------+
|  Control Flags (1 octet)           |
+------------------------------------+
|  Layer-2 MTU (2 octet)             |
+------------------------------------+
|  Reserved (2 octets)               |
+------------------------------------+
```

Figure 1: Layer2 Info Extended Community

Control Flags:

This field contains bit flags relating to the control information about pseudowires. This field is augmented with a definition of 2 new flags field.

```
 0 1 2 3 4 5 6 7
+-+-+-+-+-+-+-+-+
|Z|Z|Z|Z|T|R|C|S|      (Z = MUST Be Zero)
+-+-+-+-+-+-+-+-+
```

Figure 2: Control Flags Bit Vector

With reference to the Control Flags Bit Vector, the following bits in the Control Flags are defined; the remaining bits, designated Z, MUST be set to zero when sending and MUST be ignored when receiving this Extended Community.

T   When the bit value is 1, the PE announce the ability to send a Pseudowire packet that includes a flow label. When the bit value is 0, the PE is indicating that it will

not send a Pseudowire packet containing a flow label.

R	When the bit value is 1, the PE is able to receive a Pseudowire packet with a flow label present. When the bit value is 0, the PE is unable to receive a Pseudowire packet with the flow label present.
C	Defined in [RFC4761].
S	Defined in [RFC4761].

3. Signaling the Presence of the Flow Label

As part of the Pseudowire signaling procedures described in [RFC4761], a Layer2 Info Extended Community is advertised in the VPLS BGP NLRI.

A PE that wishes to send a flow label in a Pseudowire packet MUST include in its VPLS BGP NLRI a Layer2 Info Extended Community using Control Flags field with T = 1.

A PE that is willing to receive a flow label in a Pseudowire packet MUST include in its VPLS BGP NLRI a Layer2 Info Extended Community using Control Flags field with R = 1.

A PE that receives a VPLS BGP NLRI containing a Layer2 Info Extended Community with R = 0 MUST NOT include a flow label in the Pseudowire packet.

Therefore, a PE sending a Control Flags field with T = 1 and receiving a Control Flags field with R = 1 MUST include a flow label in the Pseudowire packet. Under all other combinations, a PE MUST NOT include a flow label in the Pseudowire packet.

A PE MAY support the configuration of the flow label (T and R bits) on a per-service (e.g., VPLS VFI) basis. Furthermore, it is also possible that on a given service, PEs may not share the same flow label settings. The presence of a flow label is therefore determined on a per-peer basis and according to the local and remote T and R bit values. For example, a PE part of a VPLS and with a local T = 1, must only transmit traffic with a flow label to those peers that signaled R = 1. And if the same PE has local R = 1, it must only expect to receive traffic with a flow label from peers with T = 1. Any other traffic must not have a flow label. A PE expecting to receive traffic from a remote peer with a flow label MAY drop traffic that has no flow label. A PE expecting to receive traffic from a remote peer with no flow label MAY drop traffic that has flow label.

Modification of flow label settings may impact traffic over a PW as these could trigger changes in the PEs data-plane programming (i.e. imposition / disposition of flow label). This is an implementation specific behavior and outside the scope of this draft.

The signaling procedures in [RFC4761] state that the unspecified bits in the Control Flags field (bits 0-5) MUST be set to zero when sending and MUST be ignored when receiving. The signaling procedure described here is therefore backwards compatible with existing implementations. A PE not supporting the extensions described in this draft will always advertise a value of ZERO in the position assigned by this draft to the R bit and therefore a flow label will never be included in a packet sent to it by one of its peers. Similarly, it will always advertise a value of ZERO in the position assigned by this draft to the T bit and therefore a peer will know that a flow label will never be included in a packet sent by it.

Note that what is signaled is the desire to include the flow LSE in the label stack. The value of the flow label is a local matter for the ingress PE, and the label value itself is not signaled.

4 Acknowledgements

The authors would like to thank Bertrand Duvivier and John Drake for their review and comments.

5 Contributors

In addition to the authors listed above, the following individuals also contributed to this document:

Eric Lent

John Brzozowski

Steven Cotter

6. IANA Considerations

Although [RFC4761] defined a Control Flags Bit Vector as part of the Layer2 Info Extended Community, it did not ask for the creation of a registry.

This document requests that IANA creates a registry for this bit vector and that it be called the "Layer2 Info Extended Community Control Flags Bit Vector" registry.

This registry should be created here:

https://www.iana.org/assignments/bgp-extended-communities/bgp-extended-communities.xhtml.

Considering [RFC4761] and this document, the initial registry is as follows:

Value	Name	Reference
S	Sequenced delivery of frames	RFC4761
C	Presence of a Control Word	RFC4761
T	Request to send a flow label	This document
R	Ability to receive a flow label	This document

As per [RFC4761] and this document, the remaining bits are unassigned, and MUST be set to zero when sending and MUST be ignored when receiving the Layer2 Info Extended Community.

7. Security Considerations

This extension to BGP does not change the underlying security issues inherent in the existing [RFC4271] and [RFC4761].

8. References

8.1. 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>.

[RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <http://www.rfc-editor.org/info/rfc4271>.

[RFC4761] Kompella, K., Ed. and Y. Rekhter, Ed., "Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling", RFC 4761, DOI 10.17487/RFC4761, January 2007, <http://www.rfc-editor.org/info/rfc4761>.

[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>.

[RFC8126] M. Cotton, et al., "Guidelines for Writing an IANA

Considerations Section in RFCs", RFC 8126, DOI 10.17487/RFC6391, June 2017, <http://www.rfc-editor.org/info/rfc8126>.

8.2. Informative References

[RFC3985] Bryant, S., Ed. and P. Pate, Ed., "Pseudo Wire Emulation Edge-to-Edge (PWE3) Architecture", RFC 3985, DOI 10.17487/RFC3985, March 2005, <http://www.rfc-editor.org/info/rfc3985>.

[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>.

[RFC8077] Martini, L., and G. Heron, "Pseudowire Setup and Maintenance Using the Label Distribution Protocol (LDP)", RFC 8077, DOI 10.17487/RFC8077, February 2017, <http://www.rfc-editor.org/info/rfc8077>.

[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>.

[RFC6624] Kompella, K., Kothari, B., and R. Cherukuri, "Layer 2 Virtual Private Networks Using BGP for Auto-Discovery and Signaling", RFC 6624, DOI 10.17487/RFC6624, May 2012, <http://www.rfc-editor.org/info/rfc6624>.

Authors' Addresses

Keyur Patel
Arrcus
Email: keyur@arrcus.com

Sami Boutros
VMware
Email: sboutros@vmware.com

Jose Liste
Cisco
Email: jliste@cisco.com

Bin Wen
Comcast

Email: bin_wen@cable.comcast.com

Jorge Rabadan
Nokia
Email: jorge.rabadan@nokia.com