Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: October 29, 2015

K. Patel
S. Boutros
J. Liste
Cisco Systems
B. Wen
Comcast
J. Rabadan
Alcatel-Lucent
April 27, 2015

# Extensions to BGP Signaled Pseudowires to support Flow-Aware Transport Labels
## draft-keyupate-l2vpn-fat-pw-bgp-04.txt

Abstract

[RFC6391] describes a mechanism that uses an additional label (Flow Label) in the MPLS label stack that allows 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 [RFC4447].

This draft defines protocol extensions required to synchronize flow label states among PEs when using the BGP-based signaling procedures defined in [RFC4761]. These protocol extensions are equally applicable to point-to-point L2VPNs defined in [RFC6624].

Table of Contents



## 1. Introduction

A pseudowire (PW)[RFC3985] is normally transported over one single network path, even if multiple 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. 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], [RFC4928].

Using the principles defined in [RFC6391], this draft augments the BGP-signaling procedures of [RFC4761] and [RFC6624] to allow an OPTIONAL mode that may be employed when the use of ECMPs is known to be beneficial to the operation of the PW.

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.

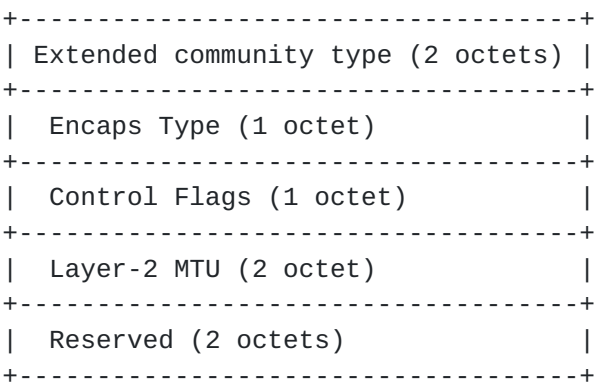
Furthermore, following the procedures for Inter-AS scenarios described in [RFC4761] section 3.4, the Flow label should never be handled by the ASBRs, only the terminating PEs on each AS will be responsible for popping or pushing this label. This is equally applicable to Method B [section 3.4.2] of [RFC4761] 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 Layer 2 Info Extended Community

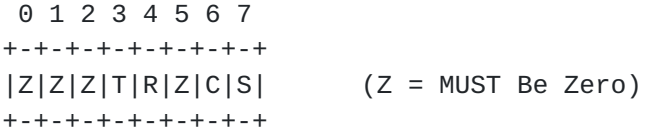
The Layer 2 Info Extended Community is used to signal control information about the pseudowires to be setup. The extended community format is described in [RFC4761]. The format of this extended community is described as:




Layer 2 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.




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.

Z  Must be set to Zero.

T  When the bit value is 1, the PE is requesting 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 Layer 2 Info Extended Community is advertised in the VPLS BGP NLRI. This draft recommends that the Control Flags field of this extended community be used to synchronize the flow label states amongst PEs for a given L2VPN.

A PE that wishes to send a flow label in a Pseudowire packet MUST include in its VPLS BGP NLRI a Layer 2 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 Layer 2 Info Extended Community using Control Flags field with R = 1.

A PE that receives a VPLS BGP NLRI containing a Layer 2 Info Extended Community with R = 0 NUST 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.

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

## 7. Security Considerations

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

## 8. References

### 8.1. Normative References

[I-D.ietf-l2vpn-vpls-multihoming]
Kothari, B., Kompella, K., Henderickx, W., Balus, F., Uttaro, J., Palislamovic, S., and W. Lin, "BGP based Multi-homing in Virtual Private LAN Service", draft-ietf-l2vpn-vpls-multihoming-06 (work in progress), July 2013.

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

[RFC3985] Bryant, S. and P. Pate, "Pseudo Wire Emulation Edge-to-Edge (PWE3) Architecture", RFC 3985, March 2005.

[RFC4271] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.

[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.

[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.

[RFC4761] Kompella, K. and Y. Rekhter, "Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling", RFC 4761, January 2007.

[RFC4928] Swallow, G., Bryant, S., and L. Andersson, "Avoiding Equal Cost Multipath Treatment in MPLS Networks", BCP 128, RFC 4928, June 2007.

[RFC6391] Bryant, S., Filsfils, C., Drafz, U., Kompella, V., Regan, J., and S. Amante, "Flow-Aware Transport of Pseudowires over an MPLS Packet Switched Network", RFC 6391, November 2011.

## 8.2. Informative References

[RFC2842] Chandra, R. and J. Scudder, "Capabilities Advertisement with BGP-4", RFC 2842, May 2000.

[RFC2858] Bates, T., Rekhter, Y., Chandra, R., and D. Katz, "Multiprotocol Extensions for BGP-4", RFC 2858, June 2000.

[RFC6624] Kompella, K., Kothari, B., and R. Cherukuri, "Layer 2 Virtual Private Networks Using BGP for Auto-Discovery and Signaling", RFC 6624, May 2012.

Authors' Addresses

Keyur Patel
Cisco Systems
170 W. Tasman Drive
San Jose, CA  95134
USA

Email: keyupate@cisco.com

Sami Boutros
Cisco Systems
170 W. Tasman Drive
San Jose, CA  95134
USA

Email: sboutros@cisco.com

Jose Liste
Cisco Systems
170 W. Tasman Drive
San Jose, CA  95134
USA

Email: jliste@cisco.com

Bin Wen
Comcast
1701 John F Kennedy Blvd
Philadelphia, PA  19103
USA

Email: bin_wen@cable.comcast.com

Jorge Rabadan
Alcatel-Lucent
777 E. Middlefield Road
Mountain View, CA  94043
USA

Email: jorge.rabadan@alcatel-lucent.com