Network Working Group
Internet Draft
Intended status: Standards Track
Expires: February 2013

A. Bashandy
Cisco Systems

August 31, 2012

IS-IS Extension for BGP FRR Protection against Edge Node Failure
draft-bashandy-isis-bgp-edge-node-frr-01.txt

Abstract

Consider a BGP free core scenario where traffic is tunneled between edge routers. Suppose the edge BGP speakers PE1, PE2,..., PEn know about a prefix P/m via the external routers CE1, CE2,..., CEm. If the edge router PEi crashes or becomes totally disconnected from the core, it desirable for a core router "P" that is carrying traffic to the failed edge router PEi to immediately restore traffic by re-routing packets originally tunneled to PEi and destined to the prefix P/m to one of the other edge routers that advertised P/m, say PEj, until BGP re-converges. If the packets originally flowing to the failed edge router PEi are labeled, then the repairing core router P router may need to swap, push, or pop the label advertised by the failed edge router PEi with another label before re-routing the packet through an LSP terminating PEj so that PEj can correctly forward the packet. The document proposes an extension to IS-IS protocol to inform core routers about the repair edge router PEj and, for labeled packets, the label that needs to be pushed/swapped before sending the packet into the tunnel terminating on PEj

Status of this Memo

This Internet-Draft will expire on February 31, 2013.

Copyright Notice

Table of Contents

1. Introduction

In a BGP free core, where traffic is tunneled between edge routers, BGP speakers advertise reachability information about prefixes to edge routers only. For labeled address families, namely AFI/SAFI 1/4, 2/4, 1/128, and 2/128, an edge router assigns local labels to prefixes and associates the local label with each advertised prefix such as L3VPN [10], 6PE [11], and Softwire [9]. Suppose that a given edge router is chosen as the best next-hop for a prefix P/m. An ingress router that receives a packet from an external router and destined for the prefix P/m sends the packet through a tunnel to that egress router. If the prefix P/m is a labeled prefix, the ingress router pushes the label advertised by the egress router before sending the packet into the tunnel terminating on the egress router. Upon receiving the packet from the core, the egress router takes the appropriate forwarding decision based on the content of the packet and/or the label pushed on the packet.

In modern networks with redundancy in place, it is not uncommon to have a prefix reachable via multiple edge routers. One example is the best external path [8]. Another more common and widely deployed scenario is L3VPN [10] with multi-homed VPN sites. As an example, consider the L3VPN topology depicted in Figure 1.

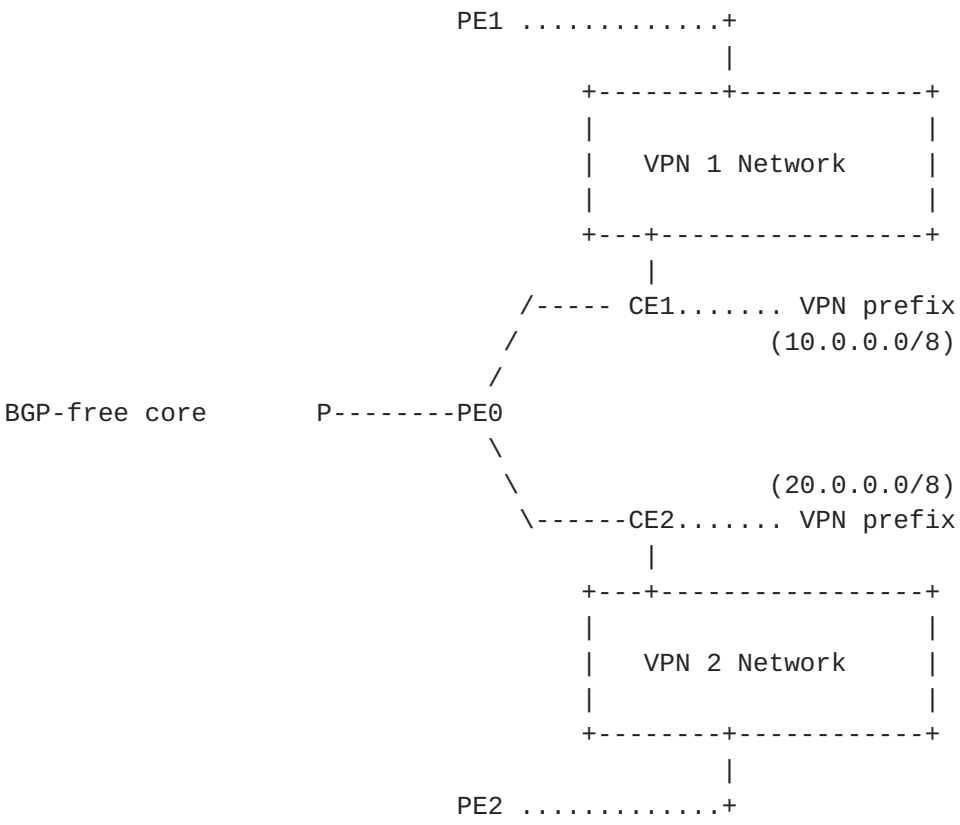

Figure 1 VPN prefix reachable via multiple PEs

As illustrated in Figure 1, the edge router PE0 is the primary NH for both 10.0.0.0/8 and 20.0.0.0/8. At the same time, both 10.0.0.0/8 and 20.0.0.0/8 are reachable through the other edge routers PE1 and PE2, respectively.

1.1. Conventions used in this document

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 [1].

In this document, these words will appear with that interpretation only when in ALL CAPS. Lower case uses of these words are not to be interpreted as carrying RFC-2119 significance.

1.2. Terminology

Refer to [7].

1.3. Problem definition

The general problem for the example shown in Section 1. is specified in [7]. The objective of this document is to specify an IS-IS [2] [3][4][5] extension to let the primary egress PE inform repairing core router(s) about the repair path [7] in a BGP-free core for both labeled and unlabeled protected prefixes. Other problems, such as determining the repair PE or the repair path or failure detection, are beyond the scope of this document.

2. The Proposed IS-IS Extension

This document specifies two new TLVs, namely "IPv4 Repair Egress Path" and "IPv6 Repair Egress Path". The new IPv4 and IPv6 Repair Egress Path TLVs identify the primary egress next-hop and its corresponding "Repair Egress Path" specified in [7] for IPv4 primary next-hop [7] and IPv6 primary next-hop [7], respectively

The encoding of the proposed TLVs is as follows

TLV Type (Value TBD):

1 octet identifying the IPv4 or IPv6 Repair Egress Path TLV code point. The code point value is assigned by IANA from the IANA "IS-IS TLV Code point Registry".

The code point for "IPv4 Repair Egress Path" means the "Primary next-hop" sub-field contains an IPv4 address. The code point for "IPv6 Repair Egress Path" means "Primary next-hop" sub-field contains an IPv6 address.

Length:

The length of the value field in multiples of 1 octet

Value (variable length):

The value specifies the IPv4 Repair Egress Path. Details in Section 3.

3. Operation of the Repair Egress Path TLVs

3.1. Structure of the Repair Egress Path TLVs

The "Value" field of the proposed TLVs contains more than one "repair tuple". Each "repair tuple" consists of the following sub-fields in the following order

o L bit

If set, then the repair path contains the underlying repair label

- P bit

  If set, then the label in the "Underlying Repair label" sub-field MUST be pushed instead of swapped. More details about this field in Section 3.2.

- AF-different Bit (D bit for simplicity)

  If set, the "Primary next-hop" sub-field contains an IPv4 address while the "Repair Next-hop" contains an IPv6 address or vice versa.

- MT bit (M bit for simplicity)

  If set, then the sub-fields " MTID-Num" and "MT-List" exist

- Reserved (Mandatory)

  This is a 4 bits field that MUST be zero by transmitter(s) and ignored by receiver(s)

- Primary next-hop (Mandatory)

  This is either a 4 octet IPv4 address or 16 octet IPv6 address representing the protected primary next-hop as defined in [7].

- Repair next-hop (Mandatory)

  This is the repair next-hop as defined in [7]. It has the same syntax as "Primary next-hop" sub-field

- Underlying Repair label (optional)

  If the L bit is set, then this field MUST contain the underlying repair label as defined in [7]. The length of this field is 3 octets.

- MTID-Num (Optional, 1 octet)

  The number of elements in the sub-field "MT-List". If the "MT" bit is set, then this field MUST exist and contain a value greater than zero

- MT-List (optional, variable length)

The size of the field is multiple of 2 octets. It represents a list of topology IDs. Each entry in the list represents a topology ID and has the same format and semantics of the "R" bits and the "MT ID" field in TLVs 235 and 237 defined in [6]. The semantics of the "MT-List" are specified in Section 3.2. If the "MT" bit is set, then this field MUST exist and contain at least one entry.

The "value" field of the proposed "IPv4/IPv6 Repair Egress Path" TLV MAY contain more than one "repair tuple", each consisting of the sub-fields defined in this section. See Section 4. provides an example of how the "value" field may look like.

3.2. Semantics of the Repair Path TLV

The Repair egress Path TLV is an implementation of the repair path defined in [7]. This section explains the IS-IS specific use.

The "Primary next-hop" and "Repair next-hop" subfield in specified in this document identifies the exit point of the primary and repair tunnels [7], respectively.

The semantics of the "P" bit is identical to the semantics of the "Push" flag in [7].

The same values of "Primary next-hop" and "Repair next-hop" subfield MUST NOT appear more than once in the "IPv4/IPv6 egress repair path" TLVs in the same LSP

The "MT-LIST" represents a list of topology IDs to be used to calculate the path taken by the repair tunnel. The semantics of the "MT-LIST" sub-field is as follows. If the repairing router decides to calculate a repair tunnel towards the "Repair next-hop", then the path taken by the tunnel SHOULD be calculated according to one of the topologies specified in the list "MT-LIST". If the path taken by the repair tunnel does not satisfy the conditions specified in [7], then the repairing not SHOULD NOT install this repair tunnel in the forwarding plane.

The addresses specified in the "Primary next-hop" and "Repair next-hop" sub-fields SHOULD be covered by (possibly different) reachability TLVs. Furthermore, if the "MT-LIST" sub-field exists, then the prefix covering the "Repair next-hop" SHOULD be advertised in a TLV of type 235 or 237 and the "MT ID" sub-field value in the 235 or 237 TLV SHOULD be identical to one of the topology IDs in the "MT-LIST" sub-field defined in this document.

This document does NOT require that the address family of the primary and repair next-hop be identical. However an implementation MAY

require that the "Primary Next-Hop" and "Repair Next-hop" fields belong to the same address family. Thus a core P router MAY ignore the "IPv4/IPv6 Repair Egress Path" TLVs if the "AF-different" bit is set. Similarly, a primary egress PE MAY NOT advertise the "IPv4/IPv6 Repair Egress Path" TLVs with the field "AF-Different" set.

For a protected Primary BGP next-hop allocated according to [7], the TLVs defined in this document support no more than one repair egress path per repair tuple. However a protected PE MAY advertise more than one repair path for the same protected next-hop by advertising more than one "repair tuple" for the same primary NH but with different repair paths. If a repairing core router receives more than one repair path for the same protected next-hop, the repairing core router MAY choose one repair path. The method of choosing a repair path is beyond the scope of this document.

4. Example

Figure 2 illustrates an example for the "value" field "IPv4 Repair Egress Path".

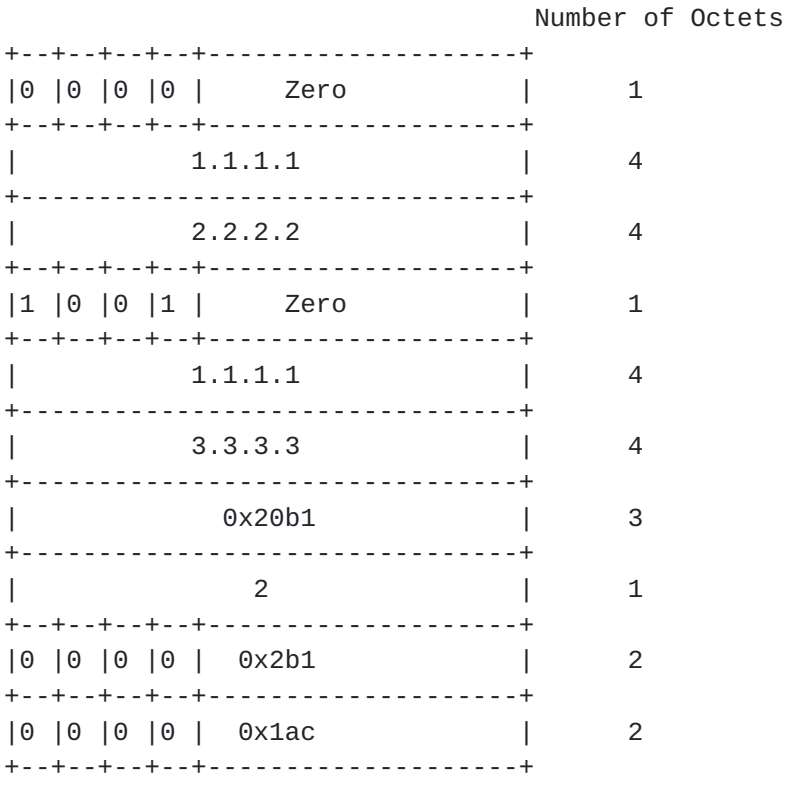

Figure 2 Example of "Value" field for "IPv4 Repair Egress Path" TLV

Figure 2 illustrates the case where "IPv4 Repair Egress Path" has two "repair tuples". The first one represents primary and repair path without MT support and without any label. The second repair tuple is the case where the repair path is labeled using the underlying repair label 0x20b1 and the repair next-hop belongs to two topologies.

5. Security Considerations

No additional security risk is introduced by using the mechanisms proposed in this document

6. IANA Considerations

This document introduces two new TLVs that require code point assignment by:

- IPv4 Repair Egress Path TLV type to be assigned from the IANA "IS-IS TLV Codepoints Registry".

- IPv6 Repair Egress Path TLV type to be assigned from the IANA "IS-IS TLV Codepoints Registry".

7. Conclusions

This document proposes an IS-IS extension that allows an egress PE to advertise a repair path consisting of another repair egress PE and possibly an underlying label to repairing core routers. Advertising this information to core routers allows core routers to provide FRR protection against primary egress PE node failure or complete disconnect from the core while keeping the core BGP-free.

8. References

8.1. Normative References

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

[2] International Organization for Standardization, "Intermediate system to Intermediate system intra-domain routing information exchange protocol for use in conjunction with the protocol for providing the connectionless-mode Network Service (ISO 8473)", ISO/IEC 10589:2002, Second Edition, Nov 2002.

[3] Callon, R., "Use of OSI IS-IS for routing in TCP/IP and dual environments", RFC 1195, December 1990.

[4] Li, T., and Smit, H., "IS-IS Extensions for for Traffic Engineering", RFC5305, October 2008

[5] Hopps, C. "Routing IPv6 with IS-IS", RFC5308, October 2008

[6] Przygienda, T., Shen, N., and N. Sheth, "M-ISIS: Multi Topology (MT) Routing in IS-IS", RFC 5120, February 2008.

8.2. Informative References

[7] Bashandy, A., Pithawala, B., Patel, P., "Scalable BGP FRR Protection against Edge Node Failure", draft-bashandy-bgp-edge-node-frr-02.txt, January 2012

[8] Marques,P., Fernando, R., Chen, E, Mohapatra, P., Gredler, H., "Advertisement of the best external route in BGP", draft-ietf-idr-best-external-05.txt, January 2012.

[9] Wu, J., Cui, Y., Metz, C., and E. Rosen, "Softwire Mesh Framework", RFC 5565, June 2009.

[10] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.

[11] De Clercq, J. , Ooms, D., Prevost, S., Le Faucheur, F., "Connecting IPv6 Islands over IPv4 MPLS Using IPv6 Provider Edge Routers (6PE)", RFC 4798, February 2007

9. Acknowledgments

Special thanks to Les Ginsberg, Keyur Patel, and Anton Smirnov for the valuable help.

This document was prepared using 2-Word-v2.0.template.dot.

Authors' Addresses

Ahmed Bashandy
Cisco Systems
170 West Tasman Dr, San Jose, CA 95134
Email: bashandy@cisco.com

Appendix A. Modification History

A.1. Changes from 00

Some editorial Changes