Network Working Group
Internet Draft
Intended status: Standards Track
Expires: August 2011

A. Bashandy
B. Pithawala
Cisco Systems

February 28, 2011

Scalable, Loop-Free BGP FRR using Repair Label
draft-bashandy-idr-bgp-repair-label-00.txt

Abstract

Consider a BGP free core scenario. Suppose the provider edge BGP speaker PE1, PE2,..., PEn know about a prefix P/p via the external routers CE1, CE2,..., CEm. If the PE router PEi loses connectivity to the primary path, whether it is another PE router or a CE router, it desirable to immediately restore traffic by rerouting packets arriving to PEi and destined to the prefix P/p to one of the other PE routers that advertised P/p, say PEj, until BGP re-converges. However if the loss of connectivity of PEi to the primary path also resulted in the loss of connectivity between PEj and CEj, rerouting a packet without before the control plane converges may result in a loop. In this document, we propose using a repair label for traffic restoration while avoiding loops. We propose advertising the ''repair'' label through BGP.

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

This document may contain material from IETF Documents or IETF Contributions published or made publicly available before November 10, 2008. The person(s) controlling the copyright in some of this material may not have granted the IETF Trust the right to allow modifications of such material outside the IETF Standards Process. Without obtaining an adequate license from the person(s) controlling the copyright in such materials, this document may not be modified outside the IETF Standards Process, and derivative works of it may not be created outside the IETF Standards Process, except to format it for publication as an RFC or to translate it into languages other than English.

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/ietf/1id−abstracts.txt

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

This Internet−Draft will expire on August 28, 2011.

Copyright Notice

Copyright (c) 2011 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

In a BGP free core, where traffic is tunneled between edge routers and edge routers assign labels to prefixes, BGP speakers advertise reachability information about prefixes and associate a local label with each prefix such as L3VPN [9], 6PE [10], and Softwire [8]. Suppose that a given edge router is chosen as the best next-hop for a prefix P/p. An ingress router that receives a packet from an external router and destined for the prefix P/p pushes the label advertised by the egress edge router and then ’’tunnels’’ the packet across the core to that egress router. Upon receiving the labeled packet from the core, the egress router uses the label on the packet to take the appropriate forwarding decision.

In modern networks, it is not uncommon to have a prefix reachable via multiple edge routers. One example is the best external path [7] Another more common and widely deployed scenario is L3VPN [9] with multi-homed VPN sites. As an example, consider the L3VPN topology depicted in Figure 1.

```
         +--------------------------+
         |                          |
         |      BGP free Core       |
         |                          |
         |      +------------------PE1----+
         |     /                    |      \
         |    /                     |       \
         |   /                      |        \
         |  /                       |         \
         | /                        |          *
        PE3                         |          CE....... VPN prefix
         | \                        |          *          (P/p)
         |  \                       |         /
         |   \                      |        /
         |    \                     |       /
         |     \                    |      /
         |      +------------------PE2----+
         |                          |
         |                          |
         +--------------------------+
```

Figure 1 VPN prefix reachable via multiple PEs

PE3 is the ingress PE. PE1 and PE2 are both egress PEs connected to CE. CE advertises one or more VPN prefixes, denoted by P/p. PE1 and PE2 advertise P/p as VPNv4 or VPNv6 routes to all ingress PEs, including PE3, and associates a label with each route.

Suppose that the ingress PE, PE3, chooses PE1 as the next-hop for the prefix P/p. In order to minimize traffic loss, it is highly desirable for PE1 to reroute all traffic destined to P/p to PE2 as soon as the connectivity to CE is lost and without waiting for the control plane (whether it is IGP or BGP) to re-converge and computes new the best path. In doing so, PE1 pushes the label advertised by PE2 for the prefix P/p, and then ’’tunnels’’ the packet to PE2. However if the loss of PE1-CE connectivity was due to CE crash, then PE2 will also reroute the traffic back to PE1, resulting in loop. Due to ultra scalability requirements, where there is a need to support thousands of peers and hundreds of thousands of prefixes, there is a need to support quick traffic restoration without waiting for the control plane to converge and without risking loops.

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

This section outlines the terms used in this document. For ease of use, we will use terms similar to those used by L3VPN [9]

o Protected prefix: It is a prefix P/p (of any AFI) that a BGP speaker has an external path to. The BGP speaker may learn about the prefix from an external peer through BGP, some other protocol, or manual configuration. The protected prefix is advertised to some or all the internal peers.

o Primary egress PE: It is an IBGP peer that can reach the protected prefix P/p through an external path and advertised the prefix to the other IBGP peers. The primary egress PE was chosen as the best path by one or more internal peers. In other words, the primary egress PE is an egress PE that will normally be used when there is no failure. Referring to Figure 1, PE1 is a primary egress PE.

o CE: It is an external router through which an egress PE can reach a prefix P/p. The router ’’CE’’ in Figure 1 is an example of such CE

o Ingress PE: It is a BGP speaker that learns about a prefix through another IBGP peer and chooses that IBGP peer as the next−hop for the prefix. PE3 in Figure 1 is an example of an ingress PE

o Repairing PE: It is the PE that attempts to restore traffic when the primary path is no longer reachable ’’without’’ waiting for BGP to re−converge. The repairing PE restores the traffic by rerouting the traffic (through a tunnel) towards the pre−calculated repair PE when it detects that the primary path longer reachable. The primary path may be a CE or another egress PE. Referring to Figure 1, if PE3 chooses PE1 as the primary egress PE and PE1 decides to reroute traffic to PE2 on losing reachability with CE, then PE1 is a repairing PE. If PE3 chooses PE1 as a primary path and PE3 decides to use PE2 as a repair path when it loses reachability to PE2, then PE3 is a repairing PE.

o Primary label: It is the label advertised by the primary egress PE to be used for normal traffic forwarding.

- Repair egress PE: It is an egress PE other than the primary egress PE that can reach the protected prefix P/p through an external neighbor. The repair PE is pre-calculated via other PEs prior to any failure

- Repair label: It is the label that will be pushed on the packet when the repairing PE reroutes the traffic (through a tunnel) towards the repair egress PE. Section 2. discusses how the repair label is used. Section 3. discusses semantics of and methods for disseminating repair label information.

- Repair path: It is the repair egress PE and the repair label.

2. Protocol Operation

This section explains the operation of the control and forwarding planes of routers participating BGP-free core traffic restoration.

2.1. Control plane Operation

1. As usual, each PE allocates a local label for each prefix it can reach through an external neighbor CE. This is the primary label used for normal traffic forwarding.

2. To provide repair path information to all PEs, the PE also allocates a repair label to the prefix if it can reach that prefix via an external neighbor. Different repair label allocation schemes are proposed in Section 3.

3. If repair label advertisement is used (Sections 3.1.), the PE advertises both the primary and repair labels to all IBGP peers.

4. When a PE receives the label advertisement from egress PEs, it calculates a primary egress PE and a repair egress PE based on its internal path selection criteria. Note that the method of choosing the repair path is beyond the scope of this document.

5. In the end, for some of the prefix advertised by more than one PE, an egress PE will have

    - a primary path

    - a repair path consisting of a repair PE and a repair label advertised by or agreed upon with the chosen repair PE.

6. A PE "never" protects a repair label. Hence on any PE, a repair label only has paths towards the CE. However a primary label may have a repair path towards a chosen repair PE

2.1.1. Additional Rules for allocating and advertising a Repair label

- A repairing PE MUST NOT advertise a repair for a prefix if it does NOT have an external path to the prefix
- A repairing PE MUST NOT associate an internal path with a repair label
- Repair labels SHOULD be advertised with labeled address families only. That is AFI/SAFI 1/4, 2/4, 1/128, and 2/128.

2.2. Forwarding Plane Operation on Losing Primary path Reachability

As soon as a PE loses reachability to the primary path of the protected prefix P/p, the forwarding plane processes arriving traffic as follows:

1. If the repair label is an advertised label
   a. If the repairing PE is an egress PE, the packet arrives at the repairing PE with the primary label at the top because the packet is ’’tunneled’’ from the ingress PE(s). In that case, the repairing swaps the incoming label stack with the "repair label stack" advertised by the repair egress PE. Section 3.1.2. specifies all the details
   b. If the repairing PE is an ingress PE, it pushes the "repair label stack" advertised by the repair egress PE. Section 3.1.2. specifies all the details
2. If the repair label is an agreed upon service label
   a. If the repairing PE is an egress PE, it swaps the incoming label with the normal label advertised by the repair PE. Otherwise it pushes the primary label advertised by the repair PE.
   b. The repairing PE pushes the repair label on top of the label stack.
   c. Section Error! Reference source not found.specifies the details.
3. The repairing PE tunnels the packet to the repair PE

4. At the repair PE, the packet arrives with the repair label at the top. If the repair label is a service label, the repair PE pops the service label and uses the rest of the label stack. Otherwise the repair PE uses the incoming label stack

5. If the repair egress PE can reach the CE, the repair PE forwards the packet towards the CE.

6. If the repair CE cannot reach the CE, the traffic will be dropped because a PE never protects a repair label

2.3. Example

Consider the L3VPN [9] topology depicted in Figure 2 where two PEs are connected to the same PE. Assume that the core is LDP. We will be using an advertised repair label.

```
                                   PE1
                                      \
                                       \
                                        \
                                         \
               LDP core                  CE....... VPN prefix
                                         /          (10.0.0.0/8)
                                        /
                                       /
                                      /
                                    PE2
```

Figure 2 : L3VPN Example

PE1: Repairing egress PE
PE2: repair PE
Primary VPN label advertised by PE1 to all PEs 4000
Repair VPN label advertised by PE1 to all PEs 5000
Primary VPN label advertised by PE2 to all PEs: 2000
Repair VPN label advertised by PE2 all PEs: 3000

LDP label for PE2 on PE1 is 1234
LDP label for PE1 on PE2 is 4567

Before failure
'''''''''''''''

PE1 has the following FIB entries

```
4000 -----> CE (unlabeled)
     -----> PE2, swap 4000 with 3000 and then push 1234
```

```
   5000 -----> CE (unlabeled)

   PE2 has the following
   2000 -----> CE (unlabeled)
        -----> PE1, swap 2000 with 5000 and then push 4567
   3000 ------> CE (unlabeled)

   After the CE crashes
   ''''''''''''''''''''
   PE1 has the following entry:
   4000 -----> PE2, swap 4000 with 3000 and then push 1234
   5000 -----> Drop


   PE2 has the following
   2000 -----> PE1, swap 2000 with 5000 and then push 4567
   3000 ------> Drop
```

Because of the above routing entries, any traffic arriving from the core at PE1 and destined for 10.0.0/8, is rerouted towards PE2 using the repair VPN label 3000. PE2 will just drop it instead of looping it back towards PE1.

```
   After the link between PE and CE fails (CE did not crash)
   ''''''''''''''''''''''''''''''''''''''''''''''''''''''''''
   PE1 has the following entry:
   4000 -----> PE2, swap 4000 with 3000 and then push 1234
   5000 -----> Drop

   PE2 has the following
   2000 -----> CE (unlabeled)
        -----> PE1, swap 2000 with 5000 and then push 4567
   3000 ------> CE
```

Because of the above routing entries, any traffic arriving from the core at PE1 and destined for 10.0.0/8 is rerouted towards PE2 using the repair VPN label 3000. PE2 will forward the traffic towards CE.

3. How to Disseminate Repair Label Information

To ensure maximum flexibility, we specify two approaches to disseminate the repair label:

- o Advertise the repair label as an optional path attribute

- o An agreed upon service label

3.1. Advertising the repair label as an Optional Path Attribute

Advertising the repair label as an optional path attributes has some advantages:

- An egress PE can benefits from a scalable repair label allocation schemes such as per-CE repair label allocation

- Allows the repairing PE to share the same repair path among multiple protected prefixes. Since the repair path is shared by all labels sharing the path attribute, the repairing PE can optimize its RIB and FIB by sharing the same repair path data structure among a large number of protected prefixes.

- Reduces the BGP update message size. Instead of having to send additional labels per prefix, multiple prefixes can share the same repair label

- The number of labels used for traffic restoration does not depend on the number of protected prefixes

- Allows for incremental deployment because the attribute is optional

The main disadvantage of sharing the same repair path among multiple primary paths is loss of fine grain control. It is not possible to manage, control, or provide differentiated handling to traffic on per prefix basis until the network re-converges. The loss of fine grain control is limited to the BGP re-convergence period.

It is noteworthy to mention that per-CE repair label allocation has some advantages over per-prefix repair label allocation. First it results in using fewer labels. Second it allows for better packing in BGP messages. Third it does not require special handling in the forwarding plane at the repair PE. Fourth it maximizes the packet switching performance because the egress PE can take a forwarding decision with a single FIB lookup.

3.1.1. Structure of the Repair Label Path Attribute

This document defines the repair label attribute as an optional non-transitive path attribute [2] as follows:

Attribute name: REPAIR_LABEL

Type code: 129

Attribute Flags:

Optional bit: 1

Transitive bit: 0

Partial bit: 0

Extended Length bit: 0

Length of the attribute: It indicates the length in octets of the attribute

Attribute Value: The attribute value contains a stack of one or more labels. The encoding of the labels is identical to encoding of the ’’label’’ field in [4]. The value of the bottom of stack (BOS) bit is determined at traffic restoration time as specified in Section 3.1.2.

3.1.2. Semantics of the Repair Label Attribute

This document specifies the semantics of the repair label attribute when the attribute carries one repair label only. The semantics of more than one repair label is beyond the scope of this document.

Suppose a BGP speaker PE1 receives an update message with a repair label attribute containing the label ’’Lr2’’ from the IBGP peer PE2. Suppose the NLRI in the MP_REACH_NLRI attribute [3] contains the prefixes R1, R2,. . . , Rn each bound to a label L21, L22,. . . , L2n, respectively. This means the following:

1. PE2 will never attempt to repair a packet arriving with the label ’’Lr’’. Hence PE2 will either forward the packet to an external CE or drop the packet

2. PE2 expects the following from PE1:

   a. Case a: The route Ri on PE1 is bound to a local label ’’L1i’’ Suppose PE1 receives a packet with the label ’’L1i’’ at the top of the stack. If the PE1 loses the primary path for a prefix Ri and PE1 decides that PE2 is the repair PE for the prefix Ri, then PE1 has to swap the label ’’L1i’’ on the packet with the repair label ’’Lr2’’ and then tunnel the packet to PE2. The bottom of stack (BOS) bit MUST be copied from the label arriving on the packet to the label ’’Lr2’’

b. Case b: The router Ri on PE1 is not bound to any local label. If the PE1 loses the primary path for a prefix Ri and PE1 decides that PE2 is the repair PE for the prefix Ri, then PE1 MUST push the label ''Lr2'' and then tunnel the packet to PE2. The bottom of stack (BOS) bit in ''Lr2'' MUST be set as specified in[5].

3.1.3. Additional Rule when Forwarding Advertisements Containing the Repair Path Attribute

As specified in Section 3.1.1. the repair label attribute is a non-transitive attribute. However there may be cases, such as inter-AS option (b)[9], route reflectors [11], or confederation, [12], where a router may replace the advertised next-hop with its own before forwarding an advertisement. If a BGP speaker replaces the next-hop attribute with its own and the advertisement contains a repair label attribute with label stack ''Sr'', there are two options

- Option 1: The BGP speaker MUST NOT advertise the repair label attribute

- Option 2: The BGP speaker MUST replace the repair label stack ''Sr'' with a locally allocated label stack ''Sr1'' before advertising the route and then advertise the stack ''Sr1'' in the repair label attribute. For the forwarding plane, the BGP speaker MUST install a swap forwarding entry such that if the BGP speaker receives a packet with the label stack ''Sr1'', it swaps ''Sr1'' with the stack ''Sr''.

Note that advertising the repair label attribute by the router depends on whether the router understands the semantics of and supports the repair label attribute at the time of receiving an advertisement containing the repair label attribute.

4. Security Considerations

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

5. IANA Considerations

This document defines a new BGP path attribute. IANA maintains a list of the current BGP attribute typecodes in [6]. This document proposes defining a new typecode value of ''129'' for the REPAIR_LABEL path attribute

6. Conclusions

This document proposes using a repair label to allow restoring traffic prior to BGP convergence while avoiding loops

7. References

7.1. Normative References

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

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

[3] Bates, T., Chandra, R., Katz, D., and Rekhter Y., ''Multiprotocol Extensions for BGP'', RFC 4760, January 2007

[4] Rosen, E., Rekhter, Y., ''Carrying Label Information in BGP-4'', RFC 3107, May 2001

[5] Rosen, E., Tappan, D., Fedorkow, G., Rekhter, Y., Farinacci, D., Li, T. and A. Conta, "MPLS Label Stack Encoding", RFC 3032, January 2001.

7.2. Informative References

[6] BGP Parameters, http://www.iana.org/assignments/bgp-parameters/bgp-parameters.xhtml

[7] Marques,P., Fernando, R., Chen, E, Mohapatra, P., "Advertisement of the best external route in BGP", draft-ietf-idr-best-external-02.txt, April 2004.

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

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

[10] 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

[11] Bates, T., Chen, E., and Chandra, R., ''BGP Route Reflection: An Alternative to Full Mesh Internal BGP (IBGP)'', RFC 4456, April 2006

[12] Traina, P., McPherson, P., and Scudder, J., ''Autonomous System Confederations for BGP'', RFC 5065, August 2007

8. Acknowledgments

Special thanks to Keyur Patel, Robert Raszuk, and Eric Rosen for the valuable comments

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

Burjiz Pithawala
Cisco Systems
170 West Tasman Dr, San Jose, CA 95134
Email: bpithaw@cisco.com