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

A. Bashandy
M. Konstantynowicz
N. Kumar
Cisco Systems

October 8, 2012

BGP FRR Protection against Edge Node Failure Using Table Mirroring with Context Labels
draft-bashandy-bgp-frr-mirror-table-00.txt

Abstract

Consider a BGP free core scenario. 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 is desirable for a core router "P" carrying traffic to the failed edge router PEi to immediately restore traffic by re-tunneling 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. This draft proposes a BGP FRR scheme that relies on having the repairing edge router mirror the protected edge router forwarding table. The repairing edge router uses a locally allocated context label to identify the correct mirrored table.

Status of this Memo

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 April 8, 2013.

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 other edge routers but not to core routers. 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 [11], 6PE [12], and Softwire [10]. Suppose that a given edge router is chosen as the best next-hop for a prefix P/m by an ingress router. The ingress router that receives a packet from an external router and destined to the prefix P/m "tunnels" the packet across the core 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 tunneling the packet to 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 or the label pushed on the packet.

In modern networks, it is not uncommon to have a prefix reachable via multiple edge routers. One example is the best external path [9]. Another more common and widely deployed scenario is L3VPN [11] with multi-homed VPN sites. As an example, consider the L3VPN topology depicted in Figure 1.

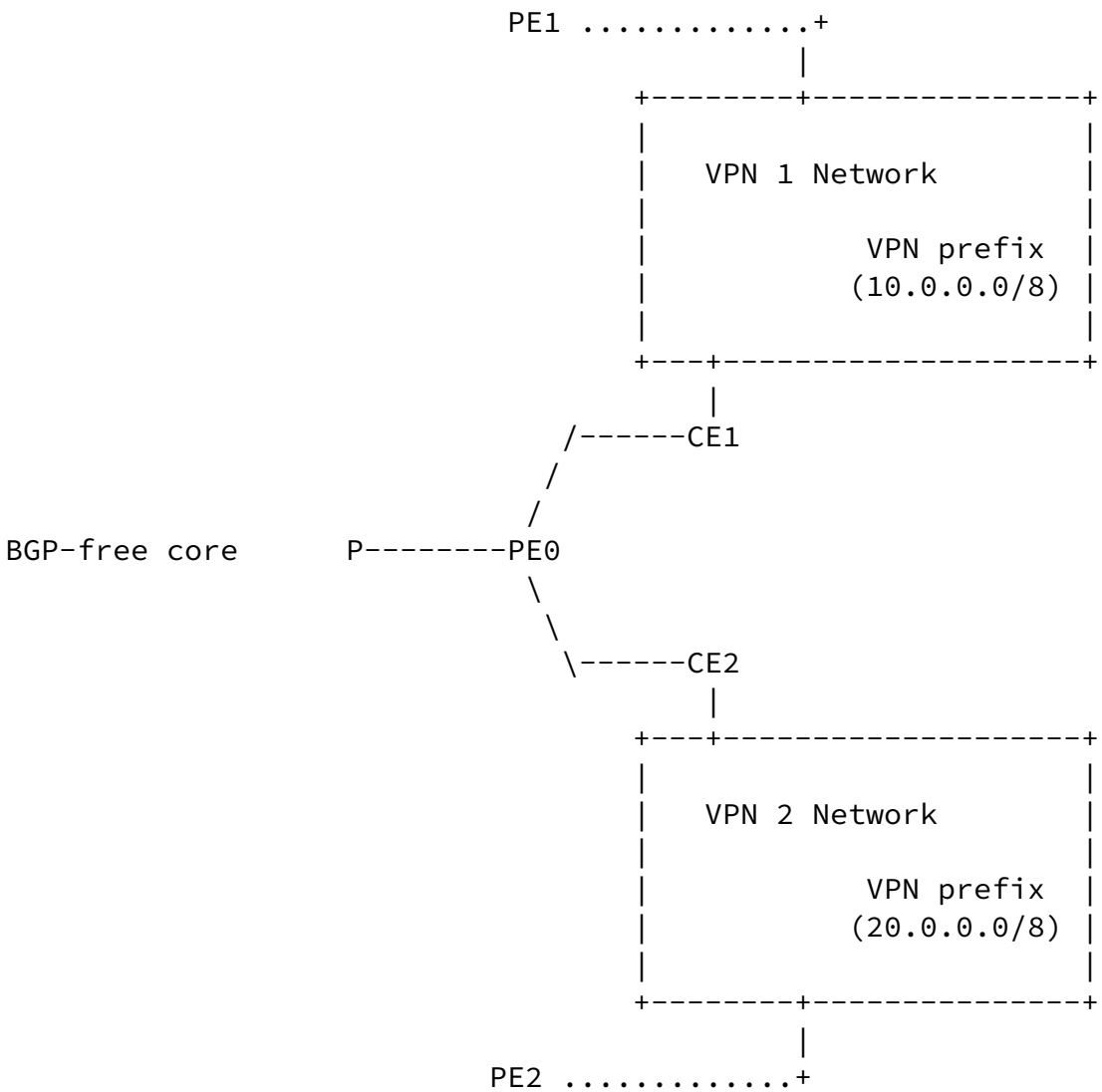

Figure 1 VPN prefix reachable via multiple PEs

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. On the failure of the edge router PE0, it is highly desirable for the core router P to re-route traffic for VPN 1 and VPN 2 to PE1 and PE2, respectively, without waiting for IGP or BGP to re-converge. This document proposes a scheme by which the egress and core routers participate to enable a core router to re-route traffic to the correct backup edge router when the primary edge router fails while keeping the core BGP-free

It is noteworthy to mention that the behavior specified in this draft requires supporting more than one BGP path. Methods, such as

[9], [17], and [18], may be needed to satisfy the multi-path requirement in certain scenarios such as the case were MED [2] or local preference [2] is used to determine the best path. The mechanism(s) by which a router supports BGP multi-path is beyond the scope of this document.

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 defines the terms used in this document. For ease of use, we will use terms similar to those used by L3VPN [11]

- BGP-Free core: A network where BGP prefixes are only known to the edge routers and traffic is tunneled between edge routers

- External prefix: It is a prefix P/m (of any AFI/SAFI) that a BGP speaker has an external path for. The BGP speaker may learn about the prefix from an external peer through BGP, some other

protocol, or manual configuration. The external prefix is advertised to some or all of the internal peers.

- Protectable prefix: It is an external prefix P/m of any AFI/SAFI) that a BGP speaker has an external path to and is eligible to have a repair path.

- Protected prefix: It is an external prefix P/m (of any AFI/SAFI) that a BGP speaker has an external path to and also has a repair path to.

- Primary Egress PE, "ePE": It is an IBGP peer that can reach the prefix P/m 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 by some ingress PEs when there is no failure. Referring to Figure 1, PE0 is an egress PE.

- Protected egress PE, "pPE" (Protected PE for simplicity): It is an egress PE for which there exists a repair path for some or all of the prefixes to which it has an external path. Referring to Figure 1, PE0 is a protected egress PE.

- Protected edge router: Any protected egress PE.

- Protected next-hop (pNH): It is an IPv4 or IPv6 host address belonging to the protected egress PE. Traffic tunneled to this IP address will be protected via the mechanism proposed in this document.

- CE: It is an external router through which an egress PE can reach a prefix P/m. The routers "CE1" and "CE2" in Figure 1 are examples of such CEs.

- Ingress PE, "iPE": 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.

- Repairing P router "rP" (Also "Repairing core router" and "repairing router"): A core router that attempts to restore

traffic when the primary egress PE is no longer reachable without waiting for IGP or BGP to re-converge. The repairing P router restores the traffic by rerouting the traffic (through a tunnel) towards the pre-calculated repair PE when it detects that the primary egress PE is no longer reachable. Referring to Figure 1, the router "P" is the repairing P router.

- Repair egress PE "rPE" (Repair PE for simplicity): It is an egress PE other than the primary egress PE that can reach the protected prefix P/m through an external neighbor. The repair PE is pre-calculated via other PEs prior to any failure. Referring to Figure 1, PE1 is the repair PE for 10.0.0.0/8 while PE2 is the repair PE for 20.0.0.0/8.

- Repair next-hop (rNH): It is an IPv4 or IPv6 address belonging to the repair egress PE. If the protected prefix is advertised via BGP, then the repair next-hop SHOULD be the next-hop attribute in the BGP update message [2][3].

- BGP nexthop (bgpNH): This is the usual next-hop attribute for route advertisements as specified in [2] and [3].

- Context Label (cL): It is an MPLS label allocated by the repairing PE (rPE) to identify the mirrored forwarding table of the protected PE (pPE). An rPE must allocate a locally distinct context label for each mirrored forwarding table. Context labels on different rPEs may overlap

- Repair path (Also Repair Egress Path): It is the repair next-hop.

- Primary tunnel: It is the tunnel from the ingress PE to the primary egress PE

- Repair tunnel: It is the tunnel from the repairing P router to the repair egress PE

1.3. Problem definition

The problem that we are trying to solve is as follows

- Even though multiple prefixes may share the same egress router, they have different repair edge router. On losing connection to the edge router, a core router "P" detecting the loss of connection MUST reroute traffic towards the *correct* repair

edge router that can reach prefixes that were reachable via the failed edge router without waiting for IGP or BGP to re-converge and update the routing tables.

- The repairing core router P MUST NOT be forced to learn about the BGP prefixes on any of the edge router. The same applies for all core routers.

- The size of the routing table on any core router MUST be independent of the number of BGP prefixes in the network.

- Rerouting traffic without waiting for IGP and BGP to re-converge after a failure MUST NOT introduce loops.

- For labeled prefixes, when a packet gets re-routed to the repair PE, the label stack on the packet MUST ensure correct forwarding.

- At steady state, when pPE is reachable, paths taken by traffic before deploying the solution proposed in this document MUST NOT be impacted after deploying the solution proposed in this document unless desired by the operator.

- The solution MUST be incrementally deployable

- Minimize the number of nodes that need to be upgraded. Hence only egress PE's that participate in the solution (namely pPE's and rPE's) and protecting core routers (namely rP's) need to be upgraded.

Applying the problem to the topology in Figure 1 above, both 10.0.0.0/8 and 20.0.0.0/8 share the same primary egress router PE0,

the routing protocol(s) must identify that the node protecting repair node for 10.0.0.0/8 is PE1 while the node protecting repair node for 11.0.0.0/8 is PE2. On the failure of PE0, the core router P must reroute traffic for 10.0.0.0/8 towards PE1 and traffic for 11.0.0.0/8 towards PE2 without requiring the core router P to know about any BGP prefix.

2. Overview of BGP FRR using Mirrored Forwarding Table in an MPLS Core

The solution proposed in this document relies on the collaboration of egress PEs, and the repairing core router. This section gives an overview of how to the solution works for both labeled (AFI/SAFI 1/4,

2/4, 1/128, and 2/128) and unlabeled (AFI/SAFI 1/1, 2/1, 1/2, and 2/2) protected prefixes in a core where the tunnels between edge routers are LDP LSPs [7]. Specifications of the solution in IP core are provided in Section 3.

2.1. Control Plane operation

Control plan requires certain operating parameters to be assigned. This section explains how the parameters are assigned through configuration. Automatic determination of the operating parameters is explained in Appendix A.

1. Setting the Operating parameters on pPE

   a. Suppose the protectable prefixes on a given pPE are protected by the repair edge routers rPE1, rPE2,...

   b. For the set of prefixes protected by a given rPE, assign a distinct local next-hop pNH. The pNH is also advertised as the bgpNH when the pPE advertises the prefixes to other iBGP peers. This section assumes that pNH is assigned via configuration. pNH can be automatically calculated as described in Appendix A.

   c. pNH MUST be unique within a routing domain

   d. Because pNH is also used as bgpNH, then pNH MUST be advertised into IGP as usual

2. Setting the Operating parameters on the rPE

   a. Suppose the rPE can protect prefixes whose bgpNH is pNH1, pNH2,...

   b. The operator informs rPE about the bgp next-hops that it can protect. This task can be carried out through configuration. Appendix A outlines how rPE can automatically determine the BGP next-hops it can protect.

   c. rPE performs the following tasks for each pNH

      i. rPE allocates a "locally" distinct context label "cL" for each pNH that the rPE can protect

      ii. rPE advertises "pNH" as its own prefix into IGP but with (maximum metric - 1) so as not to affect the path taken by

the traffic flowing from iPE's to pPE's

iii. rPE advertises "cL" for pNH instead of implicit NULL to its neighboring LSRs. As explained in Appendix B, this behavior is necessary to ensure correct forwarding during the period starting from complete disconnect of pPE till all iPE stop using pPE as an exit point for BGP traffic.

iv. rPE allocates a separate "mirror" forwarding table for each pNH. The mirror forwarding table consists of a mirror IP table and a corresponding label table. The mirror table is identified by the context label "cL"

v. rPE assigns a local IP address rNH as the repair next-hop. rNH may be any local IP address on the rPE. "rNH" SHOULD be any next-hop attribute advertised by rPE when it announces reachability to the protected prefix P/m to minimize the number of prefixes advertised into IGP.

vi. rPE advertises the triplet (pNH,rNH,cL) to candidate repairing core routers. The syntax is TBD. For example, an LDP optional TLV can be used for this purpose

d. Remember that pNH1, pNH2,... are advertised as the BGP next-hop by pPE's. When rPE receives a prefix advertisement from an iBGP peer with bgpNH equal to one of the pNHs it can protect AND rPE has at least one "external" path for the received prefix:

i. If the prefix is labeled ((AFI/SAFI 1/4, 2/4, 1/128, and 2/128), insert the received label into the mirror label table corresponding to the pNH

ii. If the prefix is unlabeled, (AFI/SAFI 1/1, 2/1, 1/2, and 2/2), insert the prefix into the mirror IP table corresponding to the pNH

iii. The forwarding entry of the prefix or the label in the mirror table is to either send the packet to (one of) the external path(s) or drop the packet

iv. Remember that the external path MAY or MAY NOT be the best path. For example, if MED is used to decide the best path

and the best path happened to be the internal path, then techniques, such as [9], [17], [18], and [20] are needed to calculate and advertise (an) alternative external path(s).

3. Determining the Operating Parameters on Protecting Core router "rP"

a. rP receives the triplet (pNH,rNH,cL) from rPE

b. rP installs the following entry for pNH in its forwarding table

i. if pNH is reachable, forward the packet as usual

ii. If pNH is not reachable

1. Swap the label bound to pNH with "cL"

2. tunnel the traffic towards rNH

4. Operating parameters on the rest of the routers

a. Other than pPE, rPE, and rP, the rest of the routers can remain totally agnostic to the BGP FRR scheme proposed in this document

b. Because rPE advertises pNH with (maximum-metric - 1), all the routers will prefer pPE when sending traffic to the IP address pNH. Hence as long as pPE is reachable, there is no change in traffic patterns

2.2. Forwarding behavior at Steady State (while pPE is reachable)

When pPE is reachable, there is no change in behavior due to deploying the scheme proposed in this document

2.3. Forwarding behavior when pPE Fails

The repairing router "rP" directly connected to a failure detects that pNH is no longer reachable. The following steps are applied.

1. Repairing router "rP"

a. Receives packet with top label bound to pNH

b. pNH is not reachable

c. Pop the label of pNH and swap it with the context label cL received in the triplet (pNH,rNH,cL) from rPE

d. Push the label corresponding to rNH

e. Send the packet towards rNH

2. Penultimate hop of rPE performs the usual penultimate hop popping

3. Repair PE (rPE)

a. Because its penultimate hop performed penultimate hop popping, rPE receives a packet with the top label being the context label "cL"

b. rPE uses "cL" to identify the correct mirror forwarding table

c. rPE pops the context label "cl"

d. if the packet underneath "cL" is labeled, lookup the top label in the mirror label table corresponding to cL

e. If the packet underneath "cL" is unlabeled, lookup the destination address of the packet in the mirror IP table corresponding to cL

f. Forward the packet to an external neighbor or drop it based on the mirror table lookup

4. Ingress PEs (iPEs)

a. An ingress PE that has not yet learnt about the disappearance of pPE will continue to send traffic towards pNH and hence will be re-routed towards rPE by rP and forwarded correctly

b. An ingress PE that learns about the disappearance of pPE will calculate a new best path for traffic previously destined to pNH

5. The rest of the core routers

a. A core router that has not yet learnt that pPE is no longer reachable will continue send traffic destined to pNH towards pPE. This traffic will be intercepted by rP and re-routed towards rPE

b. A core router that has learnt that pPE is no longer reachable will send traffic towards rPE because rPE advertises pNH with (maximum-metric - 1).

i. Because rPE advertises the label "cL" for rNH instead of the usual implicit NULL, a packet originally destined towards pPE that gets re-routed towards rPE will arrive at rPE with "cL" at the top

ii. Hence rPE will process it as described in step 3.

c. Eventually all iPEs learn that pPE is unreachable and hence no traffic will be sent to any of the pNHs advertised by pPE that has just disappeared

The next section presents the solution in an IP core.

3. Overview of the BGP FRR using Mirrored Forwarding Table in IP Core

This section describes the BGP FRR using mirrored tables solution in an IP core for both labeled (AFI/SAFI 1/4, 2/4, 1/128, and 2/128) and unlabeled (AFI/SAFI 1/1, 2/1, 1/2, and 2/2) protected prefixes.

The primary difference between a MPLS core and an IP core is that the tunnels between edge routers are IP based such as [5][6][7]. We assume that rP is capable of handling MPLS labels

3.1. Control plane modification for IP core

When using IP tunnels instead of MPLS tunnels between edge routers, there is one small modification at the repair edge router rPE. For the MPLS core, the correct mirror table at rPE is identified by the context label "cL". For the IP core, the correct mirror table must be indentified by either the context label "cL" or the protected next-hop "pNH". As explained in Appendix B, this behavior is necessary to ensure correct forwarding during the period starting from complete disconnect of pPE till all iPE stop using pPE as an exit point for BGP traffic.

3.2. Forwarding behavior at Steady State (while pPE is reachable)

When pPE is reachable, there is no change in behavior due to deploying the scheme proposed in this document

3.3. Forwarding plane at Failure (when pPE is unreachable)

1. iPE is not yet aware of the failure so its behavior remains the same

2. rP

   a. Decapsulates the tunnel header towards pNH

   b. Pushes the context label "cL"

   c. Encapsulates the packet into a tunnel header with destination address rNH and forwards the packet towards rPE

3. rPE

   a. If the tunnel packet arrives with desitination address "rNH"

      i. Decapsulates the tunnel header. This exposes the context label "cL"

   b. Otherwise (i.e. the destination address is "pNH")

      i. Decapsulate the tunnel header and associate the exposed packet with the mirror table based on "pNH"

   c. The rest of the behavior is identical to the MPLS core outlined in Section 2.3.

4. Rules for Choosing and Managing the Repair path

This section specifies rules governing how the repair path is chosen and installed in the forwarding plan. Other than the rules in this section, the method of choosing the repair path is beyond the scope of this document.

1. A repair PE MUST be another edge router that advertises the same prefix to the protected edge router pPE via IBGP peering.

2. If a repairing core router "rP" determines that the path taken by the repair tunnel to a repair edge router rPE passes through the protected edge router pPE, then the repairing router "rP" MUST NOT install this repair path in its forwarding plane. Instead, the repairing "p" router MAY use other paths that do not pass through pPE or use existing core FRR mechanisms such as [13], [14], and

[15].

3. If the repair PE "rPE" advertises one or more protected next-hops, then the repair next-hop "rNH" MUST be different from any protected next-hop "pNH" advertised by rPE

If rules (1) and (2) are not applied, then the tunnel to the repair edge router rPE does not provide protection against the failure of the edge node pPE. Rule (5.) ensures that there is no ambiguity about the primary and repair next-hops

5. Inter-operability with Existing IP FRR Mechanisms

Current existing IP FRR mechanisms can be divided into two categories: core protection and edge protection. Core protection techniques, such as [13], [14], and [15], provide protection against internal node and/or link failure. Thus the technique proposed in this document is not related to existing IP FRR mechanisms. If the failure of an internal node or link results in completely disconnecting a protectable edge node, then an administrator MAY configure the repairing router to prefer the technique proposed in this document over existing IP FRR mechanisms.

Edge protection techniques, such as [16] provide protection against the failure of the link between PE and CE routers. Thus existing PE-CE link protection can co-exist with the techniques proposed in this document because the two techniques are independent of each other.

6. Example

We will use and LDP core as an example. Consider the diagram depicted in Figure 2 below.

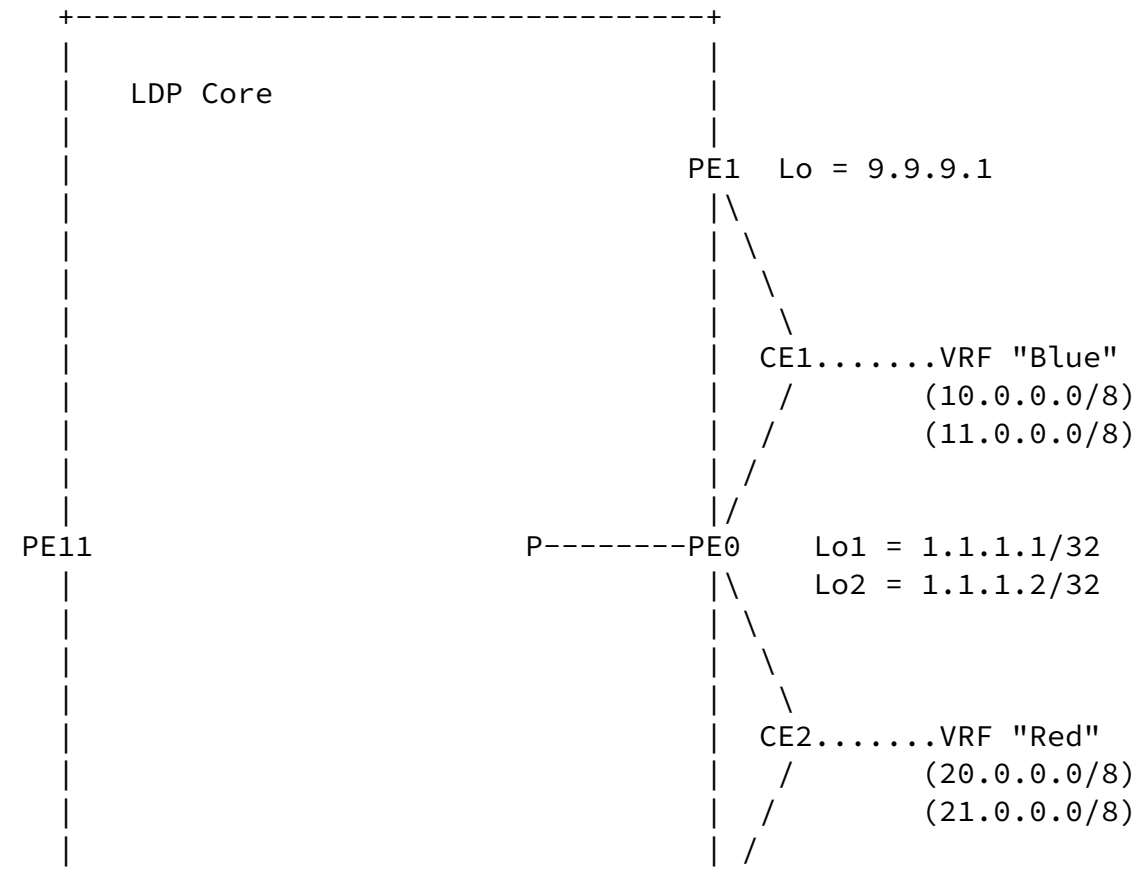

Figure 2 : Edge node BGP FRR in LDP core

- In Figure 2, PE0 is the pPE for VRFs "Blue" and "Red". PE1 and PE2 are the rPEs for VRFs "Blue" and "Red", respectively. VRF Blue has 10.0.0.0/8 and 11.0.0.0/8 and VRF Red has 20.0.0.0/8 and 21.0.0.0/8

- Assuming PE0 uses per prefix label allocation, PE0 assigns the VPN labels 4100, 4200, 4300, and 4400 to 10.0.0.0/8, 11.0.0.0/8, 20.0.0.0/8, and 21.0.0.0/8, respectively. PE0 advertises the prefixes 10.0.0.0/8, 11.0.0.0/8, 20.0.0.0/8, and 21.0.0.0/8 using MP/BGP as usual

6.1. Control Plane

1. Configuring the pNHs on PE0

The operator assigns 1.1.1.1 (the IP address of Loopback0) as the bgpNH for prefixes belonging to vrf "Blue" and 1.1.1.2 (The IP address of Loopback1) as the bgpNH for prefixes belonging to vrf "Red"

2. Configuring protection parameters on rPEs

a. The operator informs PE1 that it can protect all traffic with bgpNH=1.1.1.1. Accordingly

i. PE1 advertises 1.1.1.1 with (maximum-metric - 1) into IGP

ii. PE1 allocates a distinct mirror table for prefixes with bgpNH=1.1.1.1

iii. PE1 allocates the context label cL=1100 for the mirror table of bgpNH=1.1.1.1

iv. When advertising the FEC 1.1.1.1 to its neighboring LSRs,

PE1 associates the label 1100

v. PE2 advertises the mapping (1.1.1.1, 9.9.9.1, 1100) to candidate repair router

vi. When PE1 receives a prefix advertisement from any peer with bgpNH=1.1.1.1, PE1 inserts the VPN labels in the mirror table identified by cL=1100. Hence PE1 inserts the VPN labels 4100 and 4200 in the mirror table. The forwarding entries for both labels is to either pop the label and send the packet to an external neighbor or drop the packet

b. The operator informs PE2 that it can protect all traffic with bgpNH=1.1.1.2. Accordingly

i. PE2 advertises 1.1.1.2 with (maximum-metric - 1) into IGP

ii. PE2 allocates a distinct mirror table for prefixes with bgpNH=1.1.1.2

iii. PE2 allocates the context label cl=1200 for the mirror table of bgpNH=1.1.1.2

iv. When advertising the FEC 1.1.1.2 to its neighboring LSRs, PE2 associates the label 1200

v. PE2 advertises the mapping (1.1.1.2, 9.9.9.2, 1200) to candidate repair router

vi. When PE2 receives a prefix advertisement from any peer with bgpNH=1.1.1.2, PE2 inserts the labels into the mirror table identified by cL=1200. Hence PE inserts the VPN labels 4300 and 4400 in the mirror table. The forwarding entries for both labels is to either pop the label and send the packet to an external neighbor or drop the packet

3. Enabling BGP FRR on the penultimate hop router "P"

a. If not enabled by default, the operator enables edge node protection on the router "P"

b. Acting as a rP, the core router "P" receives the advertisements (bgpNH,rNH,cL)=(1.1.1.1, 9.9.9.1,1100) and (bgpNH,rNH,cL)=(1.1.1.2, 9.9.9.2,1200) from PE1 and PE2,

respectively.

c. "rP" creates the following forwarding state for 1.1.1.1 and 1.1.1.2

i. If 1.1.1.1 is not reachable

1. Push the context label 1100

2. Send the packet through the LSP terminating on 9.9.9.1

ii. If 1.1.1.2 is not reachable

1. Push the context label 1200

2. Send the packet through the LSP terminating on 9.9.9.2

6.2. Forwarding Plane at Steady State (When PE0 is reachable)

No change in forwarding behavior when PE0 is reachable.

6.3. Forwarding Plane at Failure (When PE0 is not reachable)

1. Repairing core router "P"

a. Traffic for VRF "Blue"

i. Receives a packet with the top label being the LDP label for 1.1.1.1

ii. 1.1.1.1 is not reachable

iii. Pop the LDP label of 1.1.1.1.

iv. Push the context label 1100

v. Push the LDP label for 9.9.9.1 and forward the packet towards PE1

b. Traffic for VRF "Red"

i. Receives a packet with the top label being the LDP label for 1.1.1.2

ii. 1.1.1.2 is not reachable

iii. Pop the LDP label of 1.1.1.2.

iv. Push the context label 1200

v. Push the LDP label for 9.9.9.2 and forward the packet towards PE2

2. The repair Router "PE1"

a. The penultimate hop of PE1 performs the usual penultimate hop popping

b. PE1 receives a packet with the top label equals the context label 1100

c. PE1 makes a lookup for 1100 in its label table. The lookup yields the mirror table of the bgpNH=1.1.1.1

d. Pop the cL=1100. This exposes the VPN label 4100 or 4200.

e. Lookup VPN label 4100 or 4200 in the mirror table corresponding to cL=1100. The lookup results in popping the VPN label 4100 or 4200 and forwarding the packet natively to CE2

3. The repair Router "PE2"

a. The penultimate hop of PE2 performs the usual penultimate hop popping

b. PE2 receives a packet with the top label equals the context label 1200

c. PE2 makes a lookup for 1200 in its label table. The lookup yields the mirror table of the bgpNH=1.1.1.2

d. Pop the cL=1200. This exposes the VPN label 4300 or 4400

e. Lookup the VPN label 4300 or 4400 in the mirror table. The lookup results in popping the VPN label 4300 or 4400 and forwarding the packet natively to CE2

7. Security Considerations

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

8. IANA Considerations

No requirements for IANA

9. Conclusions

This document proposes a method that allows fast re-route protection against edge node failure or complete disconnected from the core in a BGP-free core. The method does not require support of LFA FRR [13][14][15] and most of the provisioning effort can be automated at the expense of the possible need to re-advertise prefixes as described in Appendix A.

10. References

10.1. 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] Malhotra, P. and Rosen, E., " The BGP Encapsulation Subsequent Address Family Identifier (SAFI) and the BGP Tunnel Encapsulation Attribute", RFC 5512, April 2009

[5] Lau, J., Ed., Townsley, M., Ed., and I. Goyret, Ed., "Layer Two Tunneling Protocol - Version 3 (L2TPv3)", RFC 3931, March 2005.

[6] Farinacci, D., Li, T., Hanks, S., Meyer, D., and P. Traina, "Generic Routing Encapsulation (GRE)", RFC 2784, March 2000.

[7] L. Andersson, I. Minei, B. Thomas, "LDP Specifications", RFC 5036' October 2007

[8] Perkins, C., "IP Encapsulation within IP", RFC 2003, October 1996.

10.2. Informative References

[9] Marques,P., Fernando, R., Chen, E, Mohapatra, P., Gredler, H., "Advertisement of the best external route in BGP", draft-ietf-idr-best-external-04.txt, April 2011.

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

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

[12] 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

[13] Atlas, A. and A. Zinin, "Basic Specification for IP Fast Reroute: Loop-Free Alternates", RFC 5286, September 2008.

[14] Shand, S., and Bryant, S., "IP Fast Reroute", RFC5714, January 2010

[15] Shand, M. and S. Bryant, "A Framework for Loop-Free Convergence", RFC 5715, January 2010.

[16] O. Bonaventure, C. Filsfils, and P. Francois. "Achieving sub-50 milliseconds recovery upon bgp peering link failures, " IEEE/ACM Transactions on Networking, 15(5):1123-1135, 2007

[17] D. Walton, E. Chen, A. Retana, J. Scudder, "Advertisement of Multiple Paths in BGP", draft-ietf-idr-add-paths-07.txt, June 2012

[18] R. Raszuk, R. Fernando, K. Patel, D. McPherson, K. Kumaki, "Distribution of diverse BGP paths", draft-ietf-grow-diverse-bgp-path-dist-08.txt, July 2012

[19] T. Bates, E. Chen, and R. Chandra, "BGP Route Reflection: An Alternative to Full Mesh Internal BGP (IBGP)", RFC4456, Apr 2006

[20] P. Mohapatra, R. Fernando, C. Filsfils, and R. Raszuk, "Fast Connectivity Restoration Using BGP Add-path", draft-pmohapat-idr-fast-conn-restore-02, October 2011

11. Acknowledgments

Special thanks to Clarence Filsfils, Eric Rosen, Stewart Bryant, and Pradosh Mohapatra for the valuable comments

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

Appendix A. Auto-determination of Operating Parameters on rPE and pPE

The main provisioning effort as outlined in Section 2 is the assignment of a domain-wide distinct pNH for each rPE-pPE pair and configuring the pNH on the correct pPE and rPE. This section outlines a method by which the assignment of pNH to rPE on a given pPE is automated thereby eliminating the need for any operator intervention except for configuring the range of IP addresses from which pNHs are taken. The automation comes at the expense of the need to re-advertise BGP prefixes under certain conditions as outlined below in this Section.

The objective of the automation is to

- Let the rPE determine which pPEs the rPE can protect and hence assign a local context label "cL" for each pPE and mirror the portion of the pPE routing table that rPE can protect (Remember that rPE can protect a prefix advertised by pPE if rPE has an external path for that prefix)

- Let the pPE determine which PEs can act as rPE's for some or all of its prefixes and hence automatically assign a pNH for each distinct rPE out of a preconfigured range of IP addresses

When PEs peer directly with each other, it is easy to determine the router ID of the advertising router. In the presence of a router reflector [19], it is not possible to directly determine the router ID of the advertising PE. Hence we introduce the "RID-attr" optional non-transitive attribute. The actual format of the "RID-attr" attribute is TBD. It contains the router ID of the advertising PE.

Each PE MUST have a distinct router ID within a routing domain. "RID-attr" MUST be advertised with each protectable prefix.

A.1. How rPE determines the Protected PE

Assuming that the "RID-attr" is advertised as an optional attribute with all protectable prefixes, the rPE applies the following steps to determine the pPE

1. rPE receives route advertisements from another peer and the advertisement includes the peer's RID in the optional attribute "RID-Attr"

2. If rPE has an external path for some or all of the received route advertisements and rPE advertises some or all these route advertisements (as best paths or otherwise such as [9], [17], and [18]), then it considers the peer as a pPE

   a. rPE allocates a distinct context "cL" label for the pPE

   b. rPE advertises the mapping cL-->RID all the time to all peers. The syntax is TBD for the time being but a method similar to advertising tunnel information [4] can be used

3. If rPE loses all external paths for all prefixes from the peer identified by "RID", then rPE withdraws the mapping "cL-->RID"

4. If rPE cannot protect all routes advertised by the pPE but can protect some of them, then rPE re-advertises the protectable prefixes it previously advertise but attaches the context label "cL" as a non-transitive optional attribute. The syntax of "cL" is TBD. This is one of the cases where prefixes previously advertised need to be re-advertised

5. rPE creates a mirror table for pPE. If rPE can protect a route received from pPE, then rPE mirrors that route into the mirror table for pPE

A.2. How pPE Determines its rPEs and Assigns pNH for each rPE

1. When pPE receives the mapping cL-->RID where RID is the router ID of the pPE, pPE assumes the router that advertised the mapping cL-->RID is an rPE

2. pPE allocates a distinct pNH for the rPE

3. The next step is for pPE to re-advertise some or all of its prefixes but use the pNH assigned to rPE as bgpNH. Let {P1/m1,...,Pk/mk} be the set prefix that rPE advertises to its peers (as best paths or otherwise such as [9], [17], and [18]) and, at the same time, pPE advertises as reachable prefixes in the the NLRI field. There are two cases

   a. Case 1: rPE advertises the mapping cL-->RID but rPE does not associate the context label "cL" as an optional attribute with any prefix {P1/m1,...,Pk/mk}

   b. Case 2: rPE advertises the mapping cL-->RID and rPE associates "cL" as an optional attribute with a *subset* of the prefixes {P1/m1,...,Pk/mk}

4. Case 1: rPE does not associate the context label "cL" with advertised prefixes. In that case, pPE assumes that rPE can protect all of the prefixes {P1/m1,...,Pk/mk}. Hence pPE re-advertises {P1/m1,...,Pk/mk} uses the pNH assigned for the rPE as bgpNH.

5. Case 2: rPE associates "cL" with a *subset* of {P1/m1,...,Pk/mk}. In that case, pPE assumes the rPE can only protect the subset of {P1/m1,...,Pk/mk} that has "cL". Hence rPE re-advertises this subset but uses the pNH assigned for the rPE as bgpNH.

6. Cases 1 and 2 are the second case where prefixes previously advertised are re-advertised without any topology changes

A.3. Detecting Mis-configuration

The auto assignment of pNH described in this appendix still requires the operator to configure a range of IP addresses from which a pPE allocates the protected next-hops "pNH". Because the pNH allocated by two different pPEs MUST NOT be identical, then the range of IP addresses on two different pPEs MUST NOT overlap. Hence the only possible misconfiguration is configuring overlapping IP ranges on two different pPE. This section describes how such misconfiguration can be detected. Suppose pPE1 and pPE2 where configured with overlapping IP ranges. Such misconfiguration can be detected as follows:

1. Because in case of misconfiguration the IP ranges on pPE1 and pPE2 overlap, then at one point in time, pPE1 will allocate a pNH that falls within the IP range configured on pPE2

2. As described in Section A.2 pPE1 re-advertises some or all of its prefixes and use the allocated pNH as the bgpNH attribute

3. When pPE2 receives an advertisement from another peer containing a bgpNH within pPE2's configured IP range, then pPE2 detects the misconfiguration

Appendix B. Ensuring correct forwarding at the edge routers

As mentioned in Section 2 both rPE and pPE advertise the protected next-hop "pNH" in the core. To ensure no impact on traffic engineering, rPE advertises "pNH" with (max-metric - 1). When the primary edge router pPE becomes totally disconnected from the core, some core routers may start to forward traffic originally destined to pPE to rPE. Thus it is possible that traffic originally destined to pPE arrives at rPE without "cL" appearing at the top of the label stack. The behavior explained in Section 2 for MPLS core and Section 3 for IP core ensures that traffic is forwarded correctly when arriving at rPE.

In an MPLS core, the rPE advertises the label "cL" for pNH. Hence traffic originally destined for pNH and re-routed by a core router towards rPE will arrive at rPE with "cL" at the top. Hence rPE can identify the correct mirror table and be able forward the packet correctly

In an IP core, rPE associates the IP address "pNH" with the mirror table. Hence if a core router re-routes traffic originally tunneled towards pPE to rPE, the tunnel packets arrive at rPE with the destination address "pNH". This allows rPE to identify the correct mirror table and be able to forward the packet correctly

Authors' Addresses

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

Maciek Konstantynowicz
Cisco Systems
170 West Tasman Dr, San Jose, CA 95134
Email: mkonstan@cisco.com

Nagendra Kumar
Cisco Systems
170 West Tasman Dr, San Jose, CA 95134
Email: naikumar@cisco.com