

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

A. Bashandy  
B. Pithawala  
K. Patel  
Cisco Systems  
July 16, 2012

Scalable BGP FRR Protection against Edge Node Failure  
draft-bashandy-bgp-edge-node-frr-03.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 PE<sub>i</sub> crashes or becomes totally disconnected from the core, it is desirable for a core router "P" carrying traffic to the failed edge router PE<sub>i</sub> to immediately restore traffic by re-tunneling packets originally tunneled to PE<sub>i</sub> and destined to the prefix P/m to one of the other edge routers that advertised P/m, say PE<sub>j</sub>, until BGP re-converges. In doing so, it is highly desirable to keep the core BGP-free while not imposing restrictions on external connectivity. Thus (1) a core router should not be required to learn any BGP prefix, (2) the size of the forwarding and routing tables in the core routers should be independent of the number of BGP prefixes, (3) provisioning overhead should be kept at minimum, (4) re-routing traffic without waiting for re-convergence must not cause loops, and (4) there should be no restrictions on what edge routers advertise what prefixes. For labeled prefixes, (6) the label stack on the packet must allow the repair PE<sub>j</sub> to correctly forward the packet and (7) there must not be any need to perform more than one label lookup on any edge or core router during steady state

## 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-Draft

BGP FRR For Edge Node Failure

July 2012

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/lid-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 January 16, 2013.

## Copyright Notice

Copyright (c) 2012 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

<a href="#">1.</a>	Introduction.....	<a href="#">3</a>
<a href="#">1.1.</a>	Conventions used in this document.....	<a href="#">4</a>
<a href="#">1.2.</a>	Terminology.....	<a href="#">5</a>
<a href="#">1.3.</a>	Problem definition.....	<a href="#">6</a>

2.	Overview of the solution in an MPLS Core.....	7
2.1.	Control Plane operation for Automated pNH Assignment.....	7
2.2.	Control Plane operation for Configured pNH.....	10
2.3.	Forwarding behavior at Steady State (When pPE is reachable)	11
2.4.	Forwarding behavior when pPE Fails.....	12
3.	Overview of the solution in a Pure IP Core.....	13
3.1.	Control Plane operation.....	13

3.2.	Forwarding Behavior at Steady State (while pPE is reachable) .....	13
3.3.	Forwarding Behavior at Failure (when pPE is not reachable)	14
4.	Example.....	15
4.1.	Control Plane.....	16
4.2.	Forwarding Plane at Steady State (When PE0 is reachable).	16
4.3.	Forwarding Plane at Failure (When PE0 is not reachable)..	17
5.	Inter-operability with Existing IP FRR Mechanisms.....	19
6.	Security Considerations.....	19
7.	IANA Considerations.....	19
8.	Conclusions.....	19
9.	References.....	20
9.1.	Normative References.....	20
9.2.	Informative References.....	21
10.	Acknowledgments.....	21
Appendix A.	How to protect Against Misconfigured pNH.....	22
Appendix B.	Alternative Approach for advertising (pNH,rNH) to iPE	23
Appendix C.	Modification History.....	24
A.1.1.	Changes from Version 02.....	24
A.1.2.	Changes from Version 01.....	24

## 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 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 [10], 6PE [11], and Software [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 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 [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.

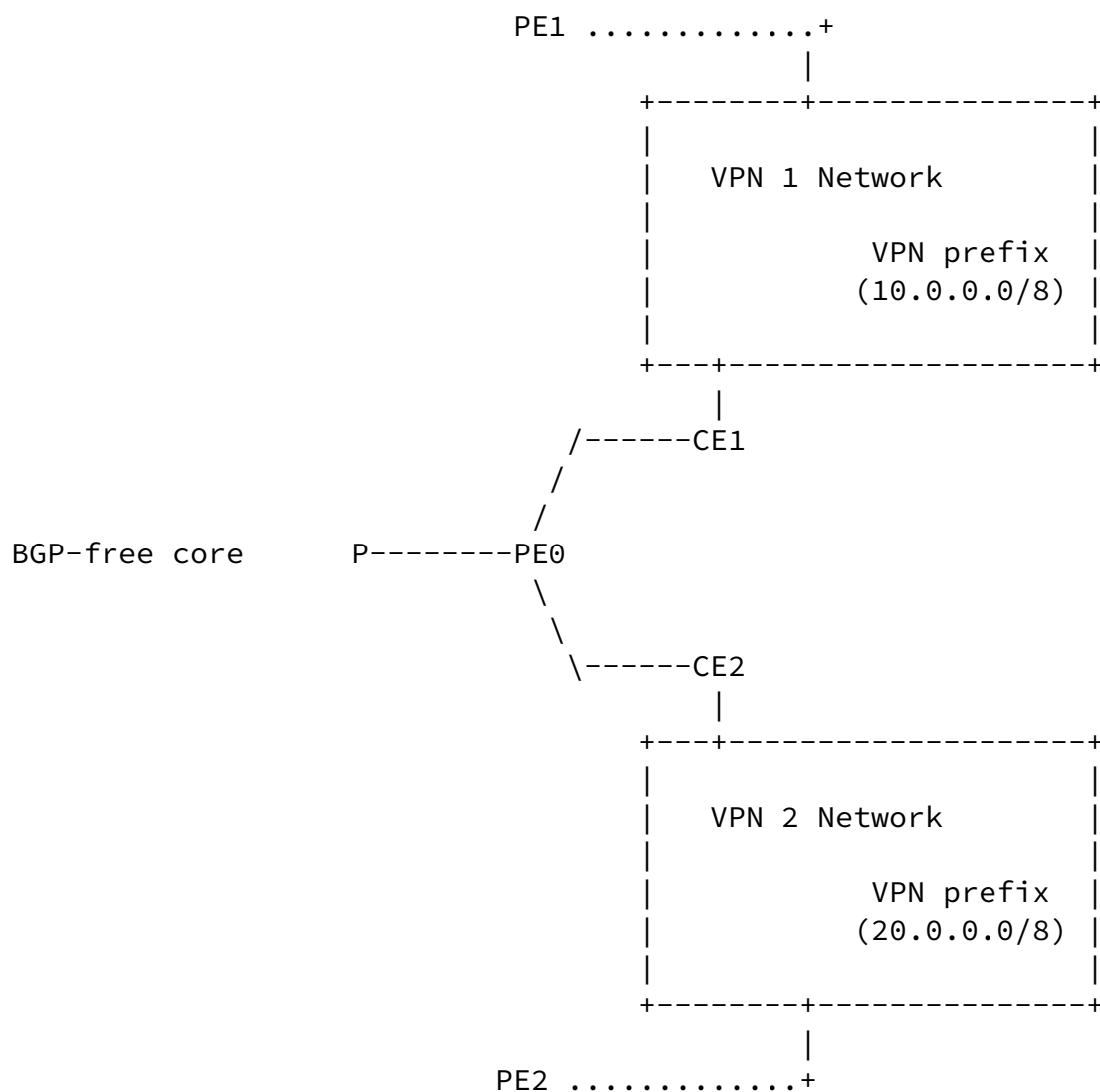


Figure 1 VPN prefix reachable via multiple PEs

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.

Bashandy

Expires January 16, 2013

[Page 4]

---

Internet-Draft

BGP FRR For Edge Node Failure

July 2012

### 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 [10]

- o BGP-Free core: A network where BGP prefixes are only known to the edge routers and traffic is tunneled between edge routers
- o 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 protected prefix is advertised to some or all of the internal peers.
- o 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.
- o 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.
- o Protected egress PE, "pPE" (Protected PE for simplicity): It is

an egress PE that has or eligible to have 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.

- o Protected edge router: Any protected egress PE.
- o 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. Note that the protected next-hop MUST be different from the next-hop attribute in the BGP update message [2][3].
- o 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.
- o 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.

- o 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.
- o 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 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.
- o Underlying Repair label (rL): The underlying repair label is the label that will be pushed so that the repair PE can forward repaired traffic correctly. A repair label is defined for labeled protected prefixes only.
- o Repair next-hop (rNH): It is an IPv4 or IPv6 host 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].

- o Repair path (Also Repair Egress Path): It is the repair next-hop. If an underlying repair label exists, the repair path is the repair next-hop together with the underlying repair label.
- o Primary tunnel: It is the tunnel from the ingress PE to the primary egress PE
- o 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

- o Even though multiple prefixes may share the same egress router, they have different repair edge router. In Figure 1 above, both 10.0.0.0/8 and 20.0.0.0/8 share the same primary next hop 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

- o On loosing connection to the edge router, the core router "P" MUST reroute traffic towards the \*correct\* repair edge router without waiting for IGP or BGP to re-converge and update the routing tables. On the failure of PE0 illustrated in Figure 1, the core router P needs to reroute traffic for 10.0.0.0/8 towards PE1 and traffic for 11.0.0.0/8 towards PE2
- o 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.
- o The size of the routing table on any core router MUST be independent of the number of BGP prefixes in the network.
- o Rerouting traffic without waiting for IGP and BGP to re-converge after a failure MUST NOT cause loops.

- o For labeled prefixes, when a packet gets re-routed to the repair PE, the label stack on the packet MUST ensure correct forwarding.
- o Provisioning overhead must be kept at minimum. In addition, misconfiguration should be detectable.
- o At steady state, when pPE is reachable, a path taken by traffic flow must not be impacted by enabling the solution proposed in this document on some or all routers

## 2. Overview of the solution in an MPLS Core

The solution proposed in this document relies on the collaboration of egress PE, ingress PE, penultimate hop routers, and repairing router. This section gives an overview of how the solution works for labeled and unlabeled protected prefixes in an MPLS core.

### 2.1. Control Plane operation for Automated pNH Assignment

This section outlines the solution for the case where the protected next hop "pNH" is automatically calculated instead of being assigned by an operator.

1. Each egress router that is capable of handling repaired traffic assigns each protectable labeled prefix a repair label: "rL". "rL" is advertised as optional path attribute. "rL" MUST be per-CE or per-VRF for good BGP attribute packing and forwarding simplicity. For unlabeled prefix, no repair label is needed. A router that is capable of handling repaired traffic is called a repair PE "rPE". The semantics of the repair label "rL" is:

- a. pop *\*two\** labels
  - b. If "rL" is per-CE, then and send the packet to the appropriate CE
  - c. If "rL" is per-VRF, forward the packet based on the contents under the two popped labels
2. If an Egress PE knows that a P/m to which it has an external path is also reachable via another PE and that other PE advertises a repair label "rL" for P/m,



- a. It chooses the other PE as a repair PE. Let's call the chosen repair PE "rPE". The ePE chooses an IP address "rNH" local to or advertised by rPE.
  - i. "rNH" SHOULD be the 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.
  - ii. if rPE also advertised a protected next-hop (pNH) for any BGP prefix that rPE can protect, then rNH MUST NOT be any protected next-hop (pNH) advertised by rPE.
- b. Allocates a local IP address corresponding to the chosen rPE, say "pNH". "pNH" represents the protected NH. I.e. Traffic tunneled to "pNH" will be protected against edge node failure via the BGP FRR mechanism proposed in this document
- c. A separate pNH is needed for every rPE (for a given protected PE). Each pNH must be unique within a single BGP-free core.
- d. Now that "ePE" has a repair path for P/m, it becomes a protected PE "pPE".
- e. Advertise pNH as a prefix into IGP
- f. Re-advertise the protected prefix P/m to other iBGP peers with "pNH" as optional non-transitive attribute
- g. pPE advertises the mapping (pNH,rNH) separately to all ingress PEs. A method analogous to how tunnel information is advertised [\[4\]](#) can be used to advertise this mapping (pNH,rNH) to ingress PE's.
- h. Once iPE receives the pNH for each prefix and the mapping (pNH,rNH), the iPE can retrieve "rL" for P/m from the advertisement of rPE for P/m.

- i. "pPE" advertises the pair (pNH,rNH) to candidate repairing core routers.
- j. "pPE" advertises the protected next-hop "pNH" to the penultimate hops to indicate that traffic flowing through the tunnel to the tail end "pNH" is protected against the failure of the node "pPE" and requires special processing by the

penultimate hop as will be described in the next few steps

- k. pPE advertises an explicit label for pNH instead of the usual implicit NULL. This way pPE can carry out the special label popping behavior (described in the next section if the penultimate hop cannot perform this task

### 3. Ingress PE "iPE"

- a. iPE receives the protected prefix P/m with "pNH" as an optional attribute
- b. iPE also receives the mapping (pNH,rNH) from pPE
- c. When iPE receives "rL" with P/m from rPE, then iPE can associate "rL" with P/m as described in [Section 2.1](#).

As a result of the above steps, the following nodes store the following information

- o Ingress PE (iPE)
  - o Receives from pPE NLRI advertisement for the protected labeled prefix P/m containing the usual next-hop attribute and the optional information "pNH". iPE also receives that mapping (pNH, rNH).
  - o iPE retrieves "rL" from the advertisement of rPE for the protected prefix P/m.
  - o Assume that iPE chooses pPE as the primary NH. Then the iPE will use pNH as the tunnel tail end to pPE instead of the usual BGP next-hop
- o Penultimate Hop
  - o Receives the "pNH" from pPE
  - o As such, it knows that traffic destined to pNH needs certain special forwarding treatment as described in the next few steps

- o Penultimate hop advertises "pNH" as its own prefix but with one of the following conditions

- . For link-state IGPs, "pNH" MAY be advertised with \*maximum metric\* so as not to affect the path taken by the traffic flowing from iPE's to pPE's
- . For distance vector IGPs, the penultimate hop MAY advertise the metric of "pNH" as follows

PHP-metric(pNH) =

pPE-metric(pNH) + metric-From-PHP-to-pPE

That is the metric advertised by the penultimate hop for pNH equals the metric advertised by pPE for pNH plus the metric from the penultimate hop to pPE

- . This way the advertisement of pNH by the penultimate hop does not impact the path taken by the traffic from iPE's to pPE's
- o Repairing core router "rP" (which may also be the penultimate hop)
    - o Receives the pair (pNH,rNH) from pPE
    - o Installs the following forwarding entry for pNH
      - . If pNH is not reachable, re-tunnel traffic to rNH

## 2.2. Control Plane operation for Configured pNH

In [Section 2.1](#), the pPE assigned pNH to a protected prefix P/m based on the chosen rPE. The result of this behavior is the need to re-advertise the protected prefix P/m with the associated "pNH". In this section, we outline the procedure by which the operator can pre-assign pNH to protected prefixes and hence avoid the need to re-advertise protected prefixes.

### 1. Protected PE "pPE"

- a. The operator groups prefixes such that two prefixes belong to the same group if the operator knows that the two prefixes are protected by the same rPE
- b. The operator assigns a distinct protected next-hop "pNH" for every group of prefixes. The assignment occurs even a repair path for P/m is not yet known.

- c. pPE advertises "pNH" as an optional non-transitive attribute with the protected prefix P/m *\*all the time\** even if no other PE advertises P/m
- d. When pPE receives an advertisement for P/m from another PE
  - i. pPE chooses the other PE as rPE
  - ii. pPE advertises the mapping (pNH,rNH) separately to all ingress PEs. rNH SHOULD be the next-hop attribute advertised by rPE. A method analogous to how tunnel information is advertised [4] can be used to advertise this mapping (pNH,rNH) to ingress PE's.
- e. The rest of the behavior is identical to what specified in [Section 2.1](#).

## 2. How to Protect the network against misconfigured pNH?

See [Appendix A](#).

What is left is to outline the forwarding behavior before and after "pPE" failure.

### 2.3. Forwarding behavior at Steady State (When pPE is reachable)

This section outlines the packet forwarding procedure when pPE is still reachable

1. Ingress PE (iPE) receives a packet matching P/m and reachable via pPE
2. The iPE pushes three labels:
  - o Bottom label: VPN label advertised by pPE
  - o Second label: rL
  - o Top label: IGP label towards pNH (not the BGP next-hop attribute)
3. Penultimate Hop
  - a. Receives a packet with top label bound to pNH
  - b. Pops *\*two\** labels *\*all the time\**.

Internet-Draft

BGP FRR For Edge Node Failure

July 2012

- c. Sends packet to pNH

#### 4. Protected PE (pPE)

- a. Receives a packet with top label as VPN label
- b. Forwards the packet as usual
- c. For unlabeled packets, the iPE only pushes the rL and the IGP label of pNH and the pPE uses the IP header for forwarding.

Thus the packet can be delivered correctly to its destination.

#### 2.4. Forwarding behavior when pPE Fails

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

##### 1. Repairing core router "rP"

- a. Receives packet with top label bound to pNH
- b. pNH is not reachable
- c. Swap the top label with the label of rNH
- d. Send packet towards rPE

In effect, the repairing router re-tunnels the packet towards the repair PE

##### 2. Penultimate hop of rPE

- a. rNH is not a protected NH for rPE
- b. Thus the penultimate hop employs the usual penultimate hop popping and then forwards the packet to rPE

##### 3. Repair PE (rPE)

- a. Receives packet with top label rL (which rPE advertised) and underneath it the regular VPN label advertised by the protected PE "pPE"

b. Make a lookup on "rL"

c. rL per CE

i. Pop \*two\* labels.

Bashandy

Expires January 16, 2013

[Page 12]

---

Internet-Draft

BGP FRR For Edge Node Failure

July 2012

ii. Send to correct CE

d. rL per VRF

i. Pop \*two\* labels.

ii. Make IP lookup in appropriate VRF

iii. Send to the CE

e. rL is assigned to unlabeled prefix

i. Pop "rL"

ii. Send the packet to the correct CE

### [3.](#) Overview of the solution in a Pure IP Core

This section provides an overview of the solution when operating in a pure IP core where core routers only understand IPv4 or IPv6 protocols. Thus traffic between PEs is transported using IP tunnels such as [\[4\]](#)[\[6\]](#)[\[7\]](#).

#### 3.1. Control Plane operation

The control plane behavior in an IP core is identical to its behavior in an MPLS core.

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

1. Ingress PE (iPE) receives a packet matching P/m and reachable via pPE

2. Ingress PE:

o For labeled traffic, Pushes two labels

. Bottom label: VPN label advertised by pPE

- . Second label: rL

- o For unlabeled traffic, just push "rL"

- o Encapsulates the packet into the IP tunnel header towards the pNH

### 3. Penultimate Hop

Bashandy

Expires January 16, 2013

[Page 13]

---

Internet-Draft

BGP FRR For Edge Node Failure

July 2012

- o No special behavior is needed from the penultimate hop while pPE is reachable

### 4. Protected PE

- a. Receives an IP packet encapsulated in an IP tunnel header with destination address pNH
- b. Decapsulate the IP tunnel header and the label right under it (which will be the repair label "rL")
- c. For labeled traffic, the VPN label is exposed. So pPE makes a lookup using the VPN label. Otherwise the usual IP forwarding is applied
- d. Forwards the packet as usual

#### 3.3. Forwarding Behavior at Failure (when pPE is not reachable)

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

##### 1. Repairing router "rP"

- a. Receives IP packet with a tunnel header destined to pNH
- b. pNH is not reachable
- c. Replace the tunnel header with a tunnel header with destination address rNH

- d. Forward the packet to rNH
2. Repair PE (rPE)
- a. Receives IP packet with a tunnel header destined to rNH
  - b. Decapsulate the tunnel header to expose the repair label "rL"
  - c. The rest of the behavior is identical to the behavior in an MPLS Core.

#### 4. Example

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

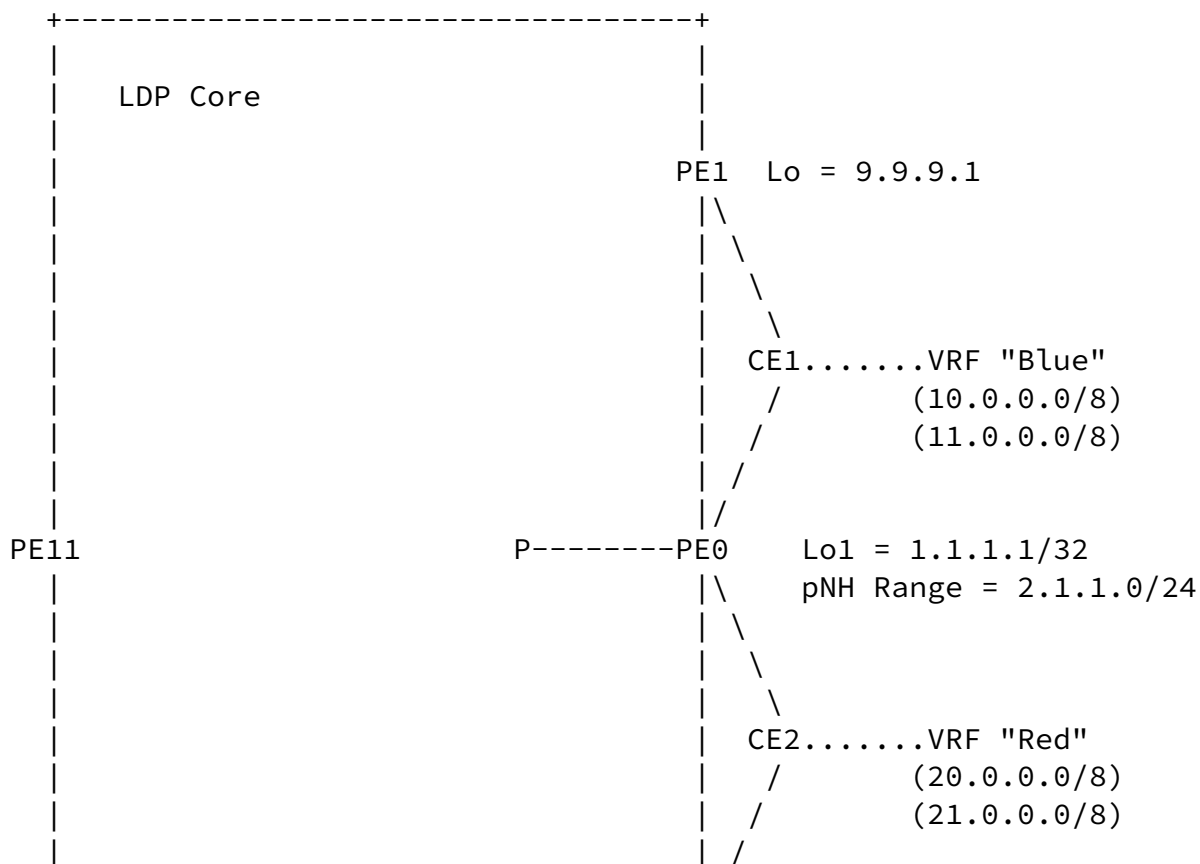






Figure 2 : Edge node BGP FRR in LDP core

- o In Figure 2, PE0 is the pPE for VRFs "Blue" and "Red" while 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
- o 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

#### 4.1. Control Plane

##### 1. rPEs Allocate and advertise Repair labels

- a. Acting as a rPE, PE1 allocates (on per-CE basis) and advertises a repair label rL1=3100 with the prefixes 10.0.0.0/8 and 11.0.0.0/8 to all iBGP peers
- b. Similarly, PE2 allocates and advertises the repair label rL2=3200 with the prefixes 20.0.0.0/8 and 21.0.0.0/8

##### 2. pPE calculates and advertises the pNHs

- a. For prefixes belonging to VRF "blue", PE0 allocates rNH1=2.1.1.1 because all of them are protected by PE1
- b. Similarly, for prefixes belonging to VRF "red", PE0 allocates rNH2=2.1.1.2 because VRF "red" is protected by PE2
- c. PE0 advertises (pNH1,rNH1)=(2.1.1.1, 9.9.9.1) and (pNH2,rNH2)=(2.1.1.2, 9.9.9.2) to the ingress PE PE1 and the repairing core router "P".

- d. PE0 re-advertises 10.0.0.0/8 & 11.0.0.0/8 with the optional attribute pNH=2.1.1.1, and 20.0.0.0/8 & 21.0.0.0/8 with pNH=2.1.1.2 to the ingress PE PE11
3. The ingress PE "PE11" creates the following forwarding state
- a. For prefixes 10.0.0.0/8 & 11.0.0.0/8: Push the VPN labels 4100 and 4200, respectively, followed by rL=3100 then tunnel the packet to 2.1.1.1
  - b. For prefixes 20.0.0.0/8 & 21.0.0.0/8: Push the VPN labels 4300 and 4400, respectively, followed by rL=3200; then tunnel the packet to 2.1.1.2

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

##### 1. Ingress PE PE11

###### a. Traffic for VRF "Blue"

- i. PE11 receives a packet for VRF Blue with destination address 10.1.1.1 from an external router.

- ii. PE11 pushes the following labels

- 1. The VPN label 4100

Bashandy	Expires January 16, 2013	[Page 16]
----------	--------------------------	-----------

---

Internet-Draft	BGP FRR For Edge Node Failure	July 2012
----------------	-------------------------------	-----------

- 2. The Repair label 3100

- 3. The LDP label for the pNH 2.1.1.1

###### b. Traffic for VRF "Red"

- i. PE11 receives a packet for VRF Red with destination address 20.1.1.1 from an external router

- ii. PE11 pushes the following labels

- 1. The VPN label 4300

- 2. The Repair label 3200

- 3. The LDP label for the pNH 2.1.1.2

##### 2. Penultimate Hop of PE0 (Which is also the rP "P")

- a. Receives a packet with top label for the protected next-hop 2.1.1.1 or 2.1.1.2
  - b. Pops \*2\* labels
  - c. Forwards the packet to pPE which is 1.1.1.1
3. Protected PE PE0
- a. Traffic for VRF "Blue"
    - i. PE0 receives traffic with the top label 4100.
    - ii. 4100 is the VPN label 10.1.1.1 belonging to VRF "Blue"
    - iii. PE0 pops the label 4100 and forwards the packet to CE1
  - b. Traffic for VRF "Red"
    - i. PE0 receives traffic with the top label 4300.
    - ii. 4300 is the VPN label for 20.1.1.1 belonging to VRF "Red"
    - iii. PE0 pops the label 4300 and forwards the packet to CE2
- 4.3. Forwarding Plane at Failure (When PE0 is not reachable)
1. The ingress PE PE11

Does not know about the failure yet and hence it does not change its behavior.

2. Repair PE rP
- a. Traffic for VRF "Blue"
    - i. Receives a packet with the top label being the LDP label for 2.1.1.1
    - ii. 2.1.1.1 is not reachable
    - iii. Swap the LDP label for 2.1.1.1 with the LDP label of

#### 9.9.9.1

iv. Forward the packet towards 9.9.9.1

#### b. Traffic for VRF "Blue"

i. Receives a packet with the top label being the LDP label for 2.1.1.2

ii. 2.1.1.2 is not reachable

iii. Swap the LDP label for 2.1.1.1 with the LDP label of 9.9.9.2

iv. Forward the packet towards 9.9.9.2

#### 3. 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 repair label 3100, which was allocated on per-CE basis and points to CE1

c. PE1 pops \*2\* labels and forwards the packet to CE1

#### 4. The repair Router "PE2"

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

b. PE1 receives a packet with the top label equals the repair label 3200, which was allocated on per-CE basis and points to CE2

c. PE2 pops \*2\* labels and forwards the packet to CE2

#### [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 [12], [13], and [14], 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] and its practical implementation [15] 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. Security Considerations

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

## 7. IANA Considerations

No requirements for IANA

## 8. 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 proposed method has few advantages

- o Easy to apply protection policies. pPE is the router that chooses the rPE. Hence if an operator wants to control what prefixes/VRFs get to be protected or what router can be chosen as repair PE, the operator needs to apply the policy on the pPE only.
- o Simple forwarding plane. The only change in forwarding plane is the need to pop/push two labels on the iPE, rP, and rPEs.

- o Single label lookup even during failure. Forwarding decisions are taken based on a single label lookup on all routers all the time even during failure

- o Immunity to mis-configuration. The only required configuration is to choose non-overlapping address ranges on different pPEs. If an operator configures overlapping IP address ranges on two different pPEs, then one of the pPE will eventually allocate a pNH that is covered by the IP address range of another pPE and hence the mis-configuration can be detected
- o No Need for IP or TE FRR: Because the exit point of the repair tunnel from rP to rPE is different from the primary tunnel exit point
- o Works in both MPLS core and IP core
- o Works with per-CE, per-VRF, and per-prefix label allocation
- o Can be incrementally deployed. There is no flag day. Different routers can be upgraded at different times
- o Zero impact on the paths taken by traffic: Enabling/deploying the feature described in this document has no effect on the paths taken by traffic at steady state

## 9. References

### 9.1. 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] Perkins, C., "IP Encapsulation within IP", [RFC 2003](#), October 1996.

## 9.2. Informative References

- [8] 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.
- [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.
- [12] Atlas, A. and A. Zinin, "Basic Specification for IP Fast Reroute: Loop-Free Alternates", [RFC 5286](#), September 2008.
- [13] Shand, S., and Bryant, S., "IP Fast Reroute", [RFC5714](#), January 2010.
- [14] Shand, M. and S. Bryant, "A Framework for Loop-Free Convergence", [RFC 5715](#), January 2010.
- [15] Bashandy, A., Pithawala, P., and Heitz, J., "Scalable, Loop-Free BGP FRR using Repair Label", [draft-bashandy-idr-bgp-repair-label-02.txt](#), July 2011.
- [16] O. Bonaventure, C. Filssils, and P. Francois. "Achieving sub-50 milliseconds recovery upon bgp peering link failures," IEEE/ACM Transactions on Networking, 15(5):1123-1135, 2007.

## 10. Acknowledgments

Special thanks to Eric Rosen, Clarence Filssils, Maciek Konstantynowicz, Stewart Bryant, Pradosh Malhotra, Nagendra Kumar, George Swallow, Les Ginsberg, and Anton Smirnov for the valuable comments

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

Internet-Draft

BGP FRR For Edge Node Failure

July 2012

[Appendix A.](#)

## How to protect Against Misconfigured pNH

[Section 2.2](#) outlines a method by which the operator can configure the protected next-hop "pNH". There is a possibility of a misconfiguration as follows

- o The operator configures the same pNH for two protected prefixes P1/m1 and P2/m2 but the two prefixes are protected by different rPEs
- o The operator configures two different pNH's for two protected prefixes P1/m1 and P2/m2 but the two prefixes are protected by same rPE

The second configuration does not cause a lot of harm. Either way, routers implementing the BGP FRR scheme proposed in this document can detect both misconfigurations.

Suppose the operator configures the same "pNH" for P1/m1 and P2/m2 but P1/m1 is protected by rPE1 and P2/m2 is protected by rPE2. In that case, the iPE and misconfigured pPE will detect this inconsistency because both will see that P1/m1 and P2/m2 are assigned the same pNH but are protected by two different rPEs. The reaction to the misconfiguration is beyond the scope of this document.

Similarly, iPE and pPE can detect that the operator configured different pNH's for P1/m1 and P2/m2 even though they are protected by the same rPE because both iPE and pPE will receive an advertisement for P1/m1 and P2/m2 from the same rPE. Reactions and remedy to the misconfiguration is beyond the scope of this document.



[Appendix B.](#)

## Alternative Approach for advertising (pNH,rNH) to i

In [Section 2.1](#), pPE re-advertises the protected prefixes with (pNH) as optional non-transitive attribute and advertises mapping (pNH,rNH) separately. Alternatively, iPE can re-advertise the protected prefix P/m to other iBGP peers with the mapping (pNH,rNH) as optional non-transitive attributes. Advertising (pNH) only with the prefixes has some advantages

- o Advertising pNH only with the prefixes can easily be used for configured pNH as described in [Section 2.2](#).
- o If the repair PE changes from one PE to another, there is no need to re-advertise all the prefixes. Only the mapping (pNH,rNH) needs to be re-advertised plus possibly some of the protected prefixes
- o Advertising pNH only with the prefix slightly reduces the BGP message size

Irrespective of whether (pNH,rNH) is advertised with the prefix or separately, (pNH,rNH) is better than advertising (pNH,rL) because there are many rL's for the same rNH. Hence advertising (pNH,rNH) yields better attribute packing

[Appendix C.](#)

## Modification History

[C.1.1.](#) Changes from Version 02

The whole scheme has been changed to a single next-hop per pPE-rPE. As a result, unlike version 00 and 01, there will be a need for behavioral changes in pPE, rP, iPE. The behavior for rPE remains almost unchanged

The second important change is requiring rP to advertise the pNH with maximum metric so that traffic does not get disrupted when the pPE disappears

[C.1.2.](#) Changes from Version 01

1. Use the term "underlying repair label" instead of just "repair label" to avoid confusion with the term "repair label" used in [\[15\]](#).
2. In version 01, it was assumed in many places that the repairing router is the penultimate hop P router. Although this would probably be the most common case, it is not always true. Hence in this version the repairing router may be any core router
3. Merged handling labeled and unlabeled prefixes into a single algorithm.
4. Allowed sending a repair label for unlabeled prefixes and added the "Push" flag. This ensures loop-free repair even for unlabeled prefixes in case that the repair PE has eiBGP paths as mentioned in Section Error! Reference source not found.
5. In Section Error! Reference source not found. discussing the rules

governing the choice of the underlying repair label for labeled prefix, we changed the wording so that the primary egress PE "SHOULD" instead of "MAY" use the repair label advertised according to [\[15\]](#) as an underlying repair label.

6. All occurrences of the term "backup" were replaced by "repair" as the term "repair" is the commonly used term in the IP FRR context such as [\[14\]](#) [\[13\]](#) [\[12\]](#)
7. Added the definition of primary and repair tunnels in [Section 1.2](#).
8. Added a definition of the term "Repair Next-hop" in [Section 1.2](#).
9. Modified the definition of "repair path" in [Section 1.2](#). to being the repair next-hop plus the underlying repair label instead of being the repair PE plus the underlying repair label.

Bashandy

Expires January 16, 2013

[Page 24]

---

Internet-Draft

BGP FRR For Edge Node Failure

July 2012

10. Outlined inter-operability with existing IP FRR techniques in [Section 5](#).
11. There were few editorial corrections.

#### 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

Keyur Patel  
Cisco Systems  
170 West Tasman Dr, San Jose, CA 95134  
Email: keyupate@cisco.com

