Abstract

This document specifies a fast-protection mechanism for protecting [RFC2547] based VPN service against egress node failure. This mechanism enables local repair to be performed immediately upon a egress node failure. In particular, the routers upstream to egress node could redirect VPN traffic to a protector (a new role) to repair in the order of tens of milliseconds, achieving fast protection that is comparable to MPLS fast reroute.

Status of This Memo

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

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

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."

This Internet-Draft will expire on January 22, 2015.

Copyright Notice

Copyright (c) 2014 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 Regarding 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 .................................................. 2
2. Specification of Requirements ............................... 3
3. Terminology ................................................... 3
4. Reference topology ........................................... 3
5. Theory of Operation ........................................... 5
   5.1. Protector and Protection Models ....................... 6
       5.1.1. Co-located protector ........................... 6
       5.1.2. Centralized protector ........................ 7
       5.1.3. Hybrid protector ................................ 7
   5.2. Context Identifier and VPN prefixes ................ 7
   5.3. MPLS egress Fast reroute ............................ 8
       5.3.1. RSVP ........................................... 8
       5.3.2. LDP ............................................. 8
   5.4. Forwarding State on Protector PE ..................... 9
       5.4.1. Alternate egress PE for protected prefix ....... 9
6. Egress node Failure ........................................... 10
7. Deployment Considerations ................................... 10
   7.1. Discussion on deployment models ...................... 10
   7.2. Simple deployment model ............................. 11
   7.3. Deployment requirements ............................. 12
8. Security Considerations ..................................... 12
9. Acknowledgements ............................................ 12
10. References .................................................. 12
   10.1. Normative References ............................... 12
   10.2. Informative References ............................. 13
Authors’ Addresses ............................................. 14

1. Introduction

This document specifies a fast-protection mechanism for protecting RFC 2547 based VPN against egress PE failure. The procedures in this document are relevant only when a VPN site is multi-homed to two or more PEs. This is mainly designed based on MPLS context specific label switching[RFC5331]. This fast-protection refers to the ability to provide local repair upon a failure in the order of tens of milliseconds, which is comparable to MPLS fast-reroute [RFC4090]. This fast-protection is achieved by establishing local protection as close to a failure as possible. Compared with the existing global repair mechanisms that rely on control plane convergence, these procedures could provide faster and more deterministic restoration...
for VPN traffic. However, this is intended to complement the global repair mechanisms, rather than replacing them in any way.

2. Specification of Requirements

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.

3. Terminology

Protected PE: A PE which request fast-protection for set of VPN-IP prefixes.

Protected VPN-IP prefix: A multi-homed VPN-IP prefix that required protection in event of protected node goes down.

Protector: A router which protect one or more Protected VPN-IP prefix when a Protected node goes down.

BGP nexthop: A nexthop advertised in the BGP-Update for the VPN-IP prefix by a BGP speaker.

VPN label: A label advertised by a BGP speaker for set of VPN-IP prefixes. This label could be per-VRF label or per-nexthop label or per-prefix label.

Transport LSP: A MPLS LSP setup to BGP nexthop either by LDP or RSVP.

Alternative egress PE: A PE originates VPN-IP prefix with same IP prefix of the protected VPN-IP prefix in a same VPN.

Context MPLS table: A context-specific label space FIB. This table is populated with VPN labels advertised by the protected-PE for the protected VPN-IP prefix.

Context label: A label from protector provides context for context-specific label forwarding.

Context VRF: A IP FIB with alternate nexthop per context per site.

PLR: Point of Local Repair.

4. Reference topology

This document refers to the following topologies to describe various roles, procedures and solution.
In Figure 1 there are two VPNs red and blue with two multi-homed sites connecting to their PEs. Assume blue VPN site2 and red VPN site2 required egress protection in case of PE5 goes down. Then PE5 is protected PE for red VPN site2 for and blue VPN site2. VPN-IP prefixes originated by PE5 associated with red site2 and blue site2 are protected VPN prefixes. The MPLS label associated with VPN-IP prefix is VPN Label. The PE4 is an alternative egress PE for red site2 and PE6 is an alternative egress PE for blue site2. The protector role could be delegated to any existing router in the network. For example PE4 could act as protector for red VPN site2 and PE6 could acts as protector for blue VPN site2. This protector model is co-located model. Alternatively, RR or any other router participates in VPN-IP control plane and not connected to VPN sites could also act as protector for both red and blue VPN site2. This model is centralized model.
In Figure 2 there is a VPN red with four sites and all sites are multi homed to their PEs. Assume site2 and site4 require egress protection in case PE5 goes down. Then PE5 is the protected PE for site2 and site4. PE4 and PE6 are alternate PEs for site2 and site4 respectively. Here also the protector role could be delegated to any existing router in the network. For example PE4 could act as a protector for site2 and PE6 could act as a protector for site4. This is called the 'co-located model'. Also PE4 or PE6 could act as protector for both sites. This is called the 'hybrid model'.

The various protector models and deployment guidance are spelled-out in Section 5.1 and Section 7.

5. Theory of Operation

Each (egress) PE attached to a given multi-homed site originates VPN-IP route(s) associated with the destination(s) within that site. Each such route should have its own Route Distinguisher, and its own next-hop, although all these routes have the same Route Target(s). Each (ingress) PE attached to other sites within the same VPN, import these route(s) into VRF creating more than one possible path to multi-homed sites. When an egress PE goes down, all VPN traffic destined to the multi homed sites attached to the downed egress PE gets rerouted to alternate egress PE(s) attached to same multi-homed site by ingress PE(s) after it detects the egress PE down. Until ingress PE(s) reroute the VPN traffic, the traffic that used to go through the failed PE get dropped in penultimate hop router. Even though connectivity of multi-homed site is not bound to an egress PE, the VPN traffic gets dropped in the P router as a result of the
downed transport LSP that binds to that egress PE. This document specifies a mechanism that repairs VPN traffic at the point of failure (typically a P router which is penultimate hop of the transport LSP) and still keep P router unaware of the VPN information with the help of a protector. Section 5.1 explain the details. The penultimate hop router(s) of the transport LSP to egress PE (PLR) reroutes VPN traffic to protector through a bypass LSP in the event of egress PE failure. Protector forwards VPN traffic received from PLR in the bypass LSP to the alternative egress PE until the ingress PE reroute traffic to alternate egress PE.

5.1. Protector and Protection Models

Protector, a new role, could be delegated to a router which participates in VPN-IP control plane for VPN-IP prefixes that requires egress node protection. In a network, protector could be the alternate egress PE of a egress protected multi homed site (precisely: the egress protected VPN-IP prefixes), or any other PE or stand-alone router for egress protection.

This specification defines three types of protector:

- co-located
- centralized
- hybrid

Its designation is dependent on the protector having direct links to the alternate site for a given VPN. A network MAY use either protection model or a combination depending on the requirements and actual network topology.

5.1.1. Co-located protector

In this model, the protector role is delegated to the alternate egress PE for a protected VPN site. Protector is co-located with the alternate PE for the protected VPN site, and it has a direct connection to the multi-homed site that originates the protected VPN-IP prefix. In the event of an egress node failure, the protector receives traffic from the PLR, and forwards VPN traffic to the multi-homed site. In the Figure 1 co-located protector could be PE4 red VPN site2 and PE6 could be the co-located protector for blue VPN site2.
5.1.2. Centralized protector

In this model, the protector serves as a centralized protector and does not have a direct connection to egress protected multi-homed sites. This model can be played by existing PEs or a dedicated protector. In the event of an egress PE failure, protector MUST forwards the traffic to an alternate egress PE with the VPN label advertised by the alternate egress PE for the VPN-IP prefix, which in turn forwards the traffic to the multi-homed site. In the Figure 1RR could act as protector for red’s site2 and blue’s site2 or PE6 could act as protector for red’s site2 and PE4 acts as protector for VPN blue’s site2. This is centralized protector model (A PE protecting VPN(s) and not connected to any protected VPN site).

5.1.3. Hybrid protector

In this model, the protector is co-located for some egress protected sites and centralized for other egress protected sites. These protected egress sites could be in the same VPN or in different VPN. In the Figure 2 either PE4 or PE6 could act as hybrid protector. Figure 1PE6 could act as hybrid protector for VPNs red site2 and blue site2.

5.2. Context Identifier and VPN prefixes.

Context-identifier is an IP address that is either globally unique or unique in the private address space of the routing domain. A context-identifier is shared between protected PE and protector(s) and it provides forwarding context for protected PE and protector. In the Protected PE each VPN-IP prefix is assigned to a context-identifier. The granularity of a context identifier is (Egress PE, VPN-IP prefix) tuple. However, a given context identifier MAY be assigned to one or multiple VPN-IP prefixes. A given context identifier MUST NOT be used by more than one protected PE and should never used for setting up BGP sessions or any control plane sessions.

The egress PE that requires protection for a VPN-IP prefix MUST set context-identifier as the BGP nexthop for VPN-IPv4 and IPv4-Mapped context-identifier for VPN-IPv6. This context-identifier as nexthop indicates to the protector that a particular VPN-IP prefix need protection. For example in Figure 1 PE5 (protected PE) advertises VPN-IP prefixes with context-identifier as BGP nexthop. The context identifier MUST also be advertised in the IGP and in LDP if LDP is used to establish transport LSP.

Possible context identifier assignments are
o Unique context-identifier for all VPN-IP prefixes, both VPN-IPv4 and VPN-IPv6. Here all the VRFs on a PE share same context-identifier.

o Unique context-identifier per address family. Here all the VRFs on the PE share the same context-identifier for given address family.

o Unique context-identifier per site for all VPN-IP prefixes, both VPN-IPv4 and VPN-IPv6. Here every VRFs has different context-identifier.

o Unique context-identifier per site per address family. Here every VRFs has different context-identifiers for a given address family.

o Unique context-identifier per CE address (nexthop). Here every CE in a VRF has a different context-identifier.

o Unique context identifier for each VPN-IP prefix. Here every VPN-IP has a different context-identifier.

The first one is coarsest granularity of a context identifier and the last one is finest granularity of a context identifier. While all of the above options are possible in principle, their practical usage is likely to vary, as not all of them may be of practical usage.

5.3. MPLS egress Fast reroute

A Protector should be able to receive the traffic from PLR in the event of an egress PE failure with forwarding context that enables protector to repair VPN traffic.

5.3.1. RSVP

If RSVP LSP is used for transport then protector and primary MUST follow procedures specified in [rsvp-egress-frr]. The context-identifier will be used as destination address of the protected LSP and the protector will be backup egress node of the protected LSP. PLR MUST follow [rsvp-egress-frr] procedure if alias method is used.

5.3.2. LDP

If LDP is used for transport then LDP FEC MUST be the context identifier. The protector for the context identifier and context label could be learned through IGP which is beyond the scope of the document. The node protecting bypass path could be computed either by remote LFA or LFA for the context identifier to protector. This
5.4. Forwarding State on Protector PE

A Protector MUST maintain multiple forwarding tables. Protector maintains the forwarding state in context-specific label space on per context-identifier basis. It also maintains context specific IP forwarding table, context VRF, populated by extracting IP from VPN-IP prefix with nexthop to alternate egress PE for egress protected prefixes. In particular, the protector MUST learn VPN labels associated with VPN-IP prefixes by participating in VPN routing and MUST keep routes and labels associated with VPN(s) site(s) that required protection. For each VPN label with an associated context-identifier, the protector MUST map the context identifier to a context-specific label space [RFC5331], and programs the VPN label in that label space into its forwarding plane. The VPN label in the context-specific label space identifies the IP forwarding table, that need to be looked up to send it alternate egress PE.

The protector MAY maintain only VPN-IP prefix originated with-in the multi-homed site for given (egress PE, VPN) tuple. These VPN labels in context table and context VRF will not be used in forwarding after the ingress PE reroutes the traffic to the new best PE. Protector MUST delete VPN label and the VPN context table after ingress reroute the traffic. This SHOULD be achieved with a timer. This timer default value is 180 seconds, allowing to be able to sustain large reroute events.

Note that if the protected PE does advertise a distinct label per VPN-IP prefix, as an optimization, the protector PE does not need to create an context VRF as the MPLS lookup on the VPN label is enough to identify the outgoing PE and label.

5.4.1. Alternate egress PE for protected prefix.

Any route with BGP nexthop which has the following properties

- Exact matching route-target set
- Exact matching Prefix part (excluding the RD)

will be eligible as alternate egress PE for prefix.
6. Egress node Failure

This section summarizes the procedure for egress protection as described in the above section for completeness. A Egress PE, Protector, PLR follows the methods described in Section 5.3. The protector programs forwarding state in such a way that packets received on the bypass LSP will be forwarded based on VPN label in the context table, and prefix lookup in context VPN table. The context table is identified by the UHP label of the bypass LSP, i.e. the context identifier.

When the penultimate Hop router receives a VPN packet from the MPLS network, if the egress PE is down, the PLR tunnels the packet through the bypass LSP to the protector. The protector PE identifies the forwarding context of the egress PE based on the top label of the packet which is the UHP label of the bypass LSP. The protector further performs a second label lookup in the protected PE’s context label space followed by layer-3 lookup in the VPN context table. These UHP label, context table label and layer-3 lookup results in forwarding the packet to the site or send it to alternate egress PE based on protector model.

For example in Figure 1 RR acts as Protector and PE5 requires protection for red, blue site2 VPN-IP prefixes. As red site2 and blue site2 VPN-IP prefixes are advertised with context-identifier, the protector sets up the forwarding table for VPN-IP prefixes from site2 with alternative egress PE as nexthop. When PLR detects PE5 failure it sends all the traffic that PLR used to forward directly to PE5 to protector through bypass LSP. In the protector the top label identifies the context specific table. The VPN label in the context table identifies the VPN layer-3 forwarding table which contains site2 VPN-IP prefixes with alternate PE as nexthop. A Layer-3 lookup gives mpls path to alternate egress PE and protector will forward the packet to alternate egress PE and reach to the site2.

7. Deployment Considerations

7.1. Discussion on deployment models.

As the context-identifiers are advertised in the IGP, they introduce additional states in the network and the forwarding tables. As such, in general, it’s desirable to keep their number limited. The granularity of context-identifier is also related to the protector model used. If a centralized or hybrid protector model is used, a unique context-identifier per egress PE is enough. If a co-located protector model is used, a context identifier per VPN or per CE may be needed.
The centralized protector model, using a single context identifier per protected PE, limits the number of additional states in the network (IGP, forwarding tables) but may add extra latency during the protection time. It also minimizes the configuration effort as zero configuration is achievable. On the contrary, the co-located mode, having a more granular context identifier, will minimize the latency during the protection time at the cost of adding more states in the network. It requires more configuration as the service provider will need to define the PE pairs (protected, protector). The hybrid model is expected to offer the best trade-off as the number of IGP states in the network can be minimized by using a single context identifier per protected PE, while the additional latency can be limited by geographically distributing the protector PE in the network.

7.2. Simple deployment model.

We propose the following simple deployment model:

- a single centralized Protector PE.
- a single context-identifier per protected PE, with all VPN routes advertised with this context-identifier as BGP next-hop.

It provides the following benefits:

- minimize the number of IGP states in the network.
- minimize the configuration required: no per VPN configuration on the protector PE.

Regarding the IGP states, no additional states are required if the PEs use the secondary loopback address as BGP next-hop for VPN-IP address family. Otherwise, one additional IP address per PE need is needed. However, the number of IP address used as BGP next-hop for the customer traffic is not increased, hence if the routers allow the prioritization of the prefix during FIB update, there is no impact on the IGP convergence time.

Regarding the configuration required on the network:

- The protected PE is configured once with an additional IP address which serves as a context identifier. The BGP Next-Hop of the BGP routes are set to this context-identifier.

- The centralized protector PE does not require per VPN configuration. But it should allow set of context-identifiers to control VPN or PE it need to protect. This will be useful in
multiple protectors in the network and set of PEs are protected by a given protector. The configured context-identifiers in protector protects subset of sites or PEs.

If one want to limit the protection to only a subset of VPN or a subset of PE (for lower VPN-SLA reasons, FIB capacities reasons on the protector, forwarding capacity reason during the protection time, for the hybrid model), one may not set context-identifier as a nexthop to the VPN-IP routes that required protection. VPN per protected PE configuration is required if user wants to limit egress protection for subset of sites. In this case protected be should allow user to not set the context-identifier as BGP nexthop for advertised VPN-IP prefixes.

7.3. Deployment requirements.

This solution does not mandate any protocol extension on any router. It does not mandate any additional feature on any routers except the new protector PE. In particular, it does not mandate implementation change on ingress nor egress PE, hence could works with legacy PE. In most topology, when LDP is used, the PLR will need to support the use of a LDP LSP as a targeted LFA. This is similar to R-LFA but the ability to configure a specific LSP to reach the protector PE may be specific.

8. Security Considerations

The security considerations discussed in RFC 5036, RFC 5331, RFC 3209, and RFC 4090 apply to this document.

9. Acknowledgements

This draft is based on the ideas originally developed by JL Le Roux, Bruno Decraene and Zubair Ahmad. This document leverages work done by Yakov Rekhter and several others on LSP tail-end protection. Thanks to Nischal Sheth, Nitin Bahadur, Yimin shen, Kaliraj Vairavakkalai and Maciek Konstantynowicz for their valuable contribution.

10. References

10.1. Normative References


10.2. Informative References


Authors’ Addresses

Jeyananth Minto Jeganathan
Juniper Networks
1194 N Mathilda Avenue
Sunnyvale, CA  94089
USA

Email: minto@juniper.net

Hannes Gredler
Juniper Networks
1194 N Mathilda Avenue
Sunnyvale, CA  94089
USA

Email: hannes@juniper.net

Bruno Decraene
France Telecom - Orange
38 rue du General Leclerc
Issy Moulineaux cedex 9  92794
France

Email: bruno.decraene@orange.com