Network Working Group	J.M. Jeganathan
Internet-Draft	M. Konstantynowicz
Intended status: Standards Track	H. Gredler
Expires: April 27, 2012	
	Juniper Networks
	October 25, 2011

2547 egress PE Fast Failure Protection
draft-minto-2547-egress-node-fast-protection-00

Abstract

This document specifies a mechanism for protecting RFC2547 based VPN service against egress node failure. The mechanism enables local repair to be performed immediately upon a egress node failure. In particular, the router at point of local repair (PLR) can redirect VPN traffic to a protector to repair in the order of tens of milliseconds, achieving fast protection that is comparable to MPLS fast reroute.

Status of this Memo

Copyright Notice

Table of Contents

*1. Introduction

*2. Specification of Requirements

*3. Terminology

*4. Reference topology

*5. Theory of Operation

*5.1. Protector and Protection Models

*5.1.1. Co-located protector

*5.1.2. Centralized protector

*5.2. Context Identifier and VPN prefixes.

*5.3. Context Identifier Advertisement by IGP

*5.3.1. Context-identifier advertised as stub router.

*5.3.1.1. ISIS context-node

*5.3.1.2. OSPF context-node

*5.4. Forwarding State on Protector PE

*5.4.1. Alternate egress PE for protected prefix.

*5.5. Bypass LSP

*5.5.1. RSVP-TE Signaled Bypass LSP and Backup LSP

*5.5.2. LDP Signaled Bypass LSP

*6. Egress node Failure

*7. Security Considerations

*8. Acknowledgements

*9. References

*9.1. Normative References

*9.2. Informative References

*Authors' Addresses

1. Introduction

This document specifies a mechanism for protecting RFC2547 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 designed on the basis of MPLS context specific label switching [RFC 5331]. 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 [RFC 4090]. This 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 can provide faster restoration for VPN traffic. However, they are intended to complement the global repair mechanisms, rather than replacing them in any way.

2. Specification of Requirements

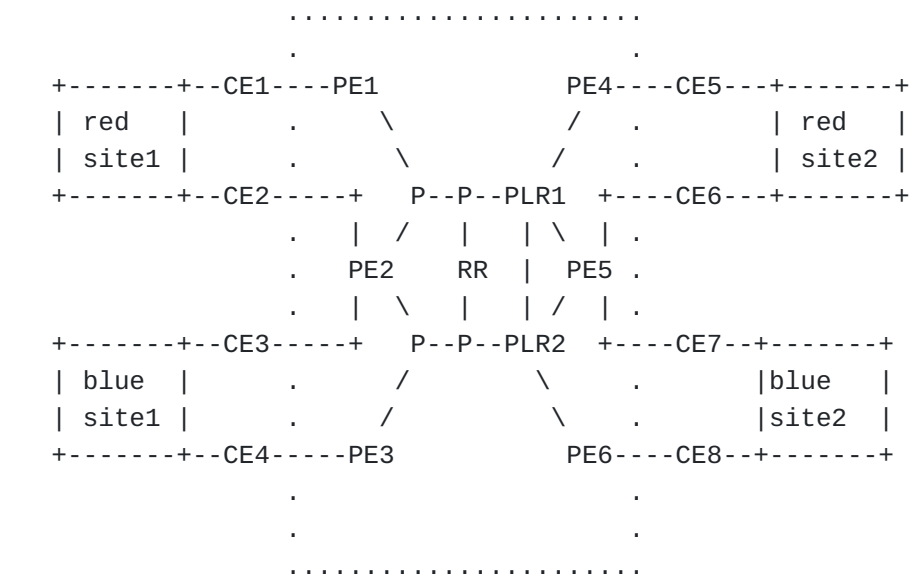
The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119.

3. Terminology

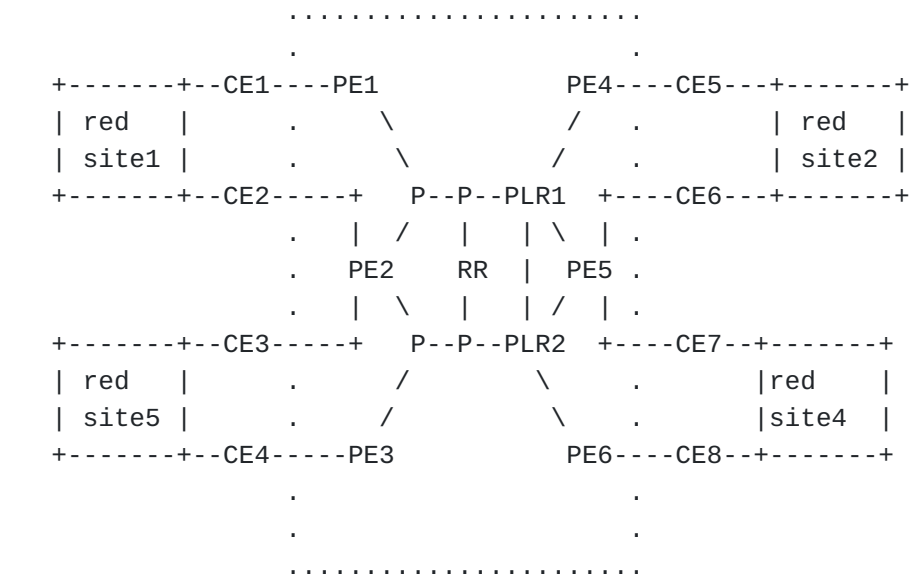
Protected PE: A PE which request protection for minimum one VPN prefix.
Protected prefix: A VPN prefix that required protection in case of Protected PE goes down.
Protector: A router which protect one or more VPN prefix when a Protected PE goes down.
BGP nexthop: A nexthop advertised in the BGP-Update for the VPN prefix by a BGP speaker.
VPN label: A label advertised by a BGP speaker for set of VPN prefixes. This label can be per-VRF label or per nexthop label or per prefix label.
Transport LSP: A LSP setup to BGP nexthop either by LDP or RSVP.
Alternative egress PE: A PE originates same IP prefix as Protected prefix in a same VPN.
VPN transport LSP: A Transport LSP that carries VPN traffic.
Context table: A context-specific label space routing table. This table is per is populated with VPN labels advertised by the protected-PE.
Context node: A stub router advertised into IGP by protected PE for a context-identifier.

4. Reference topology

This document refers to the following topologies to describe various roles and solution.



In Topology-1 two VPNs red and blue with two sites multihomed with PEs. Let assume blue and red VPN site2 prefixes required egress protection in case of PE5 goes down. PE5 is protected PE for site2 prefixes for both VPN. PE4 is alternate PE for red site2 prefixes. PE6 is alternate PE for blue site2 prefixes. For PE4 could act as protector for red VPN site2 and PE6 could acts as protector for blue VPN site2. This model is co-located protector model. RR could act as protector for both red and blue VPN site2. This is Centralized protector model (A PE protecting set of VPNs and not connected to any VPN site).



In Topology-2 has a VPNs red with four sites and multihomed with PEs. Let assume red VPN site2 and site4 prefixes required egress protection in case of PE5 goes down. PE5 is protected PE for site2, site4 prefixes for red VPN. PE4 is alternate PE for site2 prefixes. PE6 is alternate PE for site4 prefixes. Either PE4 or PE6 could act as protector. This is a slight variation of the co-located model.

5. Theory of Operation

The Egress PEs attached to multi-homed site export VPN prefixes with different route distinguisher, different nexthop but with same route target. The other PEs attached to other sites with same VPN import these route into VRF creates more than one path to multi-homed sites. When one egress PE goes down all VPN traffic towards the multihomed site moved to alternate egress PEs attached to the multi-homed site. This is done by ingress PE. The VPN traffic going via failed PE get dropped in penultimate hop router until ingress PE reroute VPN traffic. Even though connectivity of multi-homed site is not bound to an egress PE the transport LSP bind to egress PE. As result of down transport LSP VPN traffic getting dropped in P router. This document specifies a mechanism that repair VPN traffic at point of failure (typically a P router which penultimate hop of the transport LSP) and still keep P router unaware of the VPN information with the help of protector (a new role). The PLR (point of local repair) send VPN traffic to protector through bypass LSP incase of egress PE failure. This protector send VPN traffic received from PLR to the alternative egress PE until the ingress reroute traffic to alternate egress PE.

5.1. Protector and Protection Models

Protector, is a new role for the egress PE failure local repair. This protector role could be played by a PE(alternate egress PE) or any other nodes which participates in VPN control plane for VPN prefixes that required egress node protection. Hence, there are two protection models based on the location and role of a protector.

5.1.1. Co-located protector

In this model, the protector is a alternate egress PE for a protected prefix. It is co-located with the alternate PE for the protected prefix, and it has a direct connection to the multi-homed site that originate the protected prefix. In the event of an egress node failure, the protector receives traffic from the PLR, and sends the traffic to the multi-homed site. In the topology-1 PE4 could act as protector for red VPN site2 and PE6 could acts as protector for blue VPN site2. This model is co-located protector model. RR could act as protector for both red and blue VPN site2. This is Centralized protector model (A PE protecting set of VPNs and not connected to any VPN site).

A slight variant of this model is, protector is not alternate PE for a protected prefix but has same VRF. In the topology-2 either PE4 or PE6 could act as protector. This is example for the above model.

5.1.2. Centralized protector

In this model, the protector serves as a centralized protector MAY NOT have a direct connection to multi-homed site. This model can be played by existing PEs or other PEs. In the event of an egress PE failure, protector MUST send the traffic to a alternate egress PE with VPN label advertised alternate egress PE for the prefix which in turn sends the traffic to the multi-homed site. In the topology-1 RR could act as protector for both red and blue VPN site2. This is Centralized protector model (A PE protecting set of VPNs and not connected to any VPN site).
A network MAY use either protection model or a combination of both, depending on requirements.

5.2. Context Identifier and VPN prefixes.

The context-identifier is an IP address that is either globally unique or unique in the private address space of the routing domain. In Egress PE each VPN prefix is assigned to context-identifier. The granularity of a context identifier is {Egress PE, VPN prefix} tuple. However, a given context identifier MAY be assigned to one or multiple VPN prefix. Possible context identifier assignments are

- *Unique context-identifier for all VPN prefixes, both VPN-IPv4 and VPN-IPv6.
- *Unique context-identifier per address family.
- *Unique context-identifier per site for all VPN prefixes, both VPN-IPv4 and VPN-IPv6.
- *Unique context-identifier per site per address family.
- *Unique context-identifier per CE address (nexthop).
- *Unique context identifier for each prefix.

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 widely, as not all of them may be of practical usage. A given context identifier MUST NOT be used by more than one protected PE. The egress PE that required protection for a VPN prefix MUST put context-identifier as nexthop in BGP update. This context-identifier as nexthop indicates to protector that this prefix need protection. For

e.g. In topology 1 PE5(protected PE) advertise VPN prefixes with context-identifier as BGP nexthop.

5.3. Context Identifier Advertisement by IGP

IGP MUST advertise context identifiers to allow computation of TE paths for bypass LSPs and VPN transport LSPs that are destined for context identifiers. Context identifiers MUST be advertised a stub router in IGP and TE. Advertised as a stub router allow operator to deploy egress protection without upgrading all P routers.
A protected PE MUST advertise a context identifier as a stub router to TE domain and in IGP. Also Protected PE MUST advertise a link to the stub router.
A protector MUST advertise link to stub router advertised by protected PE in IGP and TE.

5.3.1. **Context-identifier advertised as stub router.**

Context-identifier advertised as stub router required two parts. A node representation (context-node) and links to the node. The protected PE and protector advertise link to context-node and protected PE advertise context-node.
The protected PE will advertise context-node in to IGP. The router-id of the context-node is context-identifier. The system-ID is derived from the context-identifier with BCD encoding. The resulting system-ID MUST be unique with in IGP routing domain. Context-node advertised with two unnumbered transit links with MAX routable link metric to protected PE and protector. For TE these unnumbered links advertised with zero bandwidth and MAX TE metric. Other TE characteristic of TE links could be configured to advertise. The router-ID or system-ID of the protector could be dynamically learned from the IGP link state database or could be configured in protected PE.
Protected PE MUST advertise unnumbered transit link with metric 1 and TE metric 1 to context-node. Protector MUST advertise unnumbered transit link with maximum routable link metric and maximum TE metric to the context-node. Other TE characteristic of the links could be configured and advertised in to TE.

5.3.1.1. **ISIS context-node**

Only zeroth fragment of the context-node is only valid. All Other fragments SHOULD be ignored. Zeroth fragment MUST include area address TLV and MAY include hostname TLV.
The set of area addresses advertised MUST be a subset of the set of Area Addresses advertised in the protected LSP number zero at the corresponding level. Preferably, the advertisement SHOULD be syntactically identical to that included in the normal LSP number zero at the corresponding level. The hostname could be set as <context-identifier+ protected PE hostname>.

The Overload (OL) MUST be set to 1. The Attached (ATT), and Partition Repair (P) bits MUST be set to 0.

5.3.1.2. OSPF context-node

The advertising router and Link State ID of router LSA MUST be context-identifier. All options bits in router LSA MUST be set to zero. The number of links MUST be 2.

5.4. Forwarding State on Protector PE

A protector maintain the forwarding state in context-specific label spaces on a per protected PE basis. In particular, the protector MUST learn the VPN label by participating the VPN routing and also MUST keep all routes associated with VPN it required to protect.
For each VPN label with an associated context-identifier protector MUST map the context identifier to a context-specific label space [RFC 5331], and program the VPN label in that label space in forwarding plane. For each VPN prefix that required protection programmed in the forwarding plane with BGP nexthop to alternate egress PE. This VPN label in the context-specific label space identify the layer-3 forwarding table that need to lookup to send it alternate egress PE. The protector MAY maintain only VPN prefix originated with-in the multi-homed site for given {egress PE, VPN}. These VPN labels in context table and VPN context table will not be used in forwarding after ingress reroute the traffic to alternative PE. Protector MUST delete VPN label and the VPN context table after ingress reroute the traffic. This shall be achieved with a timer. This timer default value is 180 seconds.

5.4.1. Alternate egress PE for protected prefix.

Any route with BGP nexthop which has the following properties

- *Exact matching route-target set (RD may be different)
- *Exact matching Prefix part (not RD)

will be eligible as alternate egress PE for prefix.

5.5. Bypass LSP

An LSP MUST be either manually or automatically provisioned on a PLR to enable the PLR to send traffic to a protector, in the event of an egress PE failure. This LSP is referred to as a bypass LSP. The bypass LSP MUST be a LSP with ultimate hop popping (UHP) [RFC 3031]. That is, the protector MUST assign an un-reserved label to the bypass LSP. When the protector PE receives VPN packets on the bypass LSP from a PLR, it MUST rely on the bypass LSP's UHP label to determine the context-specific label space to forward the packets.

5.5.1. RSVP-TE Signaled Bypass LSP and Backup LSP

If a bypass LSP is an RSVP-TE signaled LSP, its destination MUST be the context identifier of the protected VPN prefix. The path taken by the bypass LSP MAY be statically configured or dynamically computed by CSPF. The signaling of the bypass LSP MUST be triggered by the "local protection desired" and "node protection desired" bits in SESSION_ATTRIBUTE of Path message of the transport LSP [RFC 2205, RFC 3209, RFC 4090]. After the bypass LSP is established, the PLR MUST set the "local protection available" and "node protection" bits in RRO of Resv message of the transport LSP. The protector MUST terminate the backup LSP as an egress. Once the local repair takes effect, the PLR MUST set the "local protection in use" bit in RRO of Resv message of the transport LSP.

5.5.2. LDP Signaled Bypass LSP

If it is LDP LSP then LDP FEC for this LSP MUST be the context identifier of the protected segment. Prefix LFA with node protection can be used for bypass LSP computation.

6. Egress node Failure

This section summarizes the procedures egress protection described above section for completeness. A Egress PE and a protector both advertise the context identifier of a protected prefixes in IGP as a stub link or stub router, with the egress PE advertising a lower metric and protector with maximum metric. The PLR establishes a UHP bypass LSP to the protector. The destination address of the bypass LSP is the context identifier. 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 VPN context table. The context table 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. Then forwards protector the packet based on 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 E.g. In topology-1 RR is act as Protector and PE5 required protection for red, blue site2 prefixes. As red, blue site2 VPN prefixes advertised with context-identifier, the protector set up the forwarding table for prefixes from site2 with alternative egress PE as nexthop. When PLR detects PE5 failure it send to protector through

bypass LSP. In protector the top label identify the context space table. VPN label in the context table identify the VPN layer-3 forwarding table with contains site2 prefixes with alternate PE as nexthop. A Layer-3 lookup gives mpls path to alternate egress PE and protector forward packet to alternate egress PE and reach to the site2.

7. Security Considerations

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

8. Acknowledgements

This document leverages work done by Yakov Rekhter and several others on LSP tail-end protection. Thanks to Nischal Sheth, Nitin Bahadur, Yimin shen for their contribution.

9. References

9.1. Normative References

[RFC5331]	Aggarwal, R., Rekhter, Y. and E. Rosen, "MPLS Upstream Label Assignment and Context-Specific Label Space", RFC 5331, August 2008.
[RFC4364]	Rekhter, Y. and E. Rosen, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.
[RFC5036]	Andersson, L., Minei, I. and B. Thomas, "LDP Specification", RFC 5036, October 2007.
[RFC2205]	Braden, B., Zhang, L., Berson, S., Herzog, S. and S. Jamin, "Resource ReSerVation Protocol (RSVP) -- Version 1 Functional Specification", RFC 2205, September 1997.
[RFC3209]	Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V. and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.
[RFC4090]	Pan, P., Swallow, G. and A. Atlas, "Fast Reroute Extensions to RSVP-TE for LSP Tunnels", RFC 4090, May 2005.
[RFC3471]	Berger, L., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Functional Description", RFC 3471, January 2003.
[RFC3031]	Rosen, E., Viswanathan, A. and R. Callon, "Multiprotocol Label Switching Architecture", RFC 3031, January 2001.
[LDP-UPSTREAM]	Aggarwal, R and J. L. Le Roux, "MPLS Upstream Label Assignment for LDP", Internet-Draft draft-ietf-mpls-ldp-upstream, 2011.
[RSVP-NON-PHP-OOB]	Ali, A, Swallow, Z and R Aggarwal, "Non PHP Behavior and out-of-band mapping for RSVP-TE LSPs", Internet-

	Draft draft-ietf-mpls-rsvp-te-no-php-oob-mapping, 2011.

9.2. Informative References

[RFC5920]	Fang, L., "Security Framework for MPLS and GMPLS Networks", RFC 5920, July 2010.
[RFC5286]	Atlas, A and A Zinin, "Basic Specification for IP Fast Reroute: Loop-Free Alternates", RFC 5920, September 2008.
[RFC5714]	Shand, M and S Bryant, "IP Fast Reroute Framework", RFC 5714, January 2010.

Authors' Addresses

Jeyananth Minto Jeganathan Jeganathan Juniper Networks 1194 N Mathilda Avenue Sunnyvale, CA 94089 USA EMail: minto@juniper.net

Maciek Konstantynowicz Konstantynowicz Juniper Networks 1194 N Mathilda Avenue Sunnyvale, CA 94089 USA EMail: maciek@juniper.net

Hannes Gredler Gredler Juniper Networks 1194 N Mathilda Avenue Sunnyvale, CA 94089 USA EMail: hannes@juniper.net

Juniper Networks