

INTERNET-DRAFT
Intended Status: Proposed Standard
Expires: January 13, 2014

Mingui Zhang
Peng Zhou
Huawei
July 12, 2013

Label Sharing for Fast PE Protection
draft-zhang-l3vpn-label-sharing-00.txt

Abstract

This document describes a method to be used by Service Providers to provide fast protection of VPN connections for a CE. Egress PEs in a redundant group always assign the same label for VPN routes from a VRF. These egress PEs create a BGP virtual Next Hop (vNH) in the domain of the IP/MPLS backbone network as an agent of the CE router. Primary and backup tunnels terminated at the vNH are set up by the BGP/MPLS IP VPN based on IGP FRR. If the primary egress PE fails, the backup egress PEs can recognize the "shared" VPN route label and deliver the failure affected packets accordingly.

Status of this Memo

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

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/lid-abstracts.html>

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

Copyright and License Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal

Provisions Relating to IETF Documents
(<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	3
1.1. Conventions used in this document	3
1.2. Terminology	3
2. The Label Sharing Method	3
2.1. The Virtual Next Hop	4
2.1.1. Generating OSPF LSAs	5
2.1.2. Generating ISIS LSPs	7
2.2. Link Costs Set Up for IGP FRR	9
2.3 Label Assignment and Processing	10
2.3.1. The VPN Route Label	10
2.3.2. The Tunnel Label	10
3. Security Considerations	10
4. IANA Considerations	11
5. References	11
5.1. Normative References	11
5.2. Informative References	11
Author's Addresses	12

1. Introduction

For the sake of reliability, ISPs usually connect one CE to multiple PEs. When the primary egress PE fails, a backup egress PE continues to offer VPN connectivity to the CE. If local repair is performed by the upstream neighbor of the primary egress PE on the data path, it's possible to achieve 50msec switchover.

VPN routes learnt from CEs are distributed by egress PEs to ingress PEs that need to know these VPN routes. Egress PEs in a redundant group (RG) MUST allocate the same VPN route label for routes of the same VPN. When the primary egress PE fails, data packets are redirected to a backup egress PE by the PLR router, the backup PE can recognize the VPN route label in these data packets and deliver them correctly. The method developed in this document is so called "Label Sharing for Fast PE Protection". This method requires only software update on egress PE routers while their data plane remains unchanged.

This document supposes BGP/MPLS IP VPN is deployed on the backbone and Label Distribution Protocol (LDP) is used as the tunneling technology. Through generating virtual LSAs/LSPs in OSPF/ISIS, egress PEs in an RG create a virtual router (the vNH) in the IP/MPLS backbone to represent the CE router. When the VPN route is distributed, those egress PEs use vNH as the "BGP next hop". The vNH will be treated as the egress point of the tunnel by other routers. Metrics for the virtual links attached to the vNH are set up in a way that the IGP FRR mechanism defined in [LFA] can be leveraged to achieve local protection.

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 [RFC2119].

1.2. Terminology

VRF: Virtual Routing and Forwarding table
FRR: Fast ReRouting
PLR: Point of Local Repair
LFA: loop-free alternate

2. The Label Sharing Method

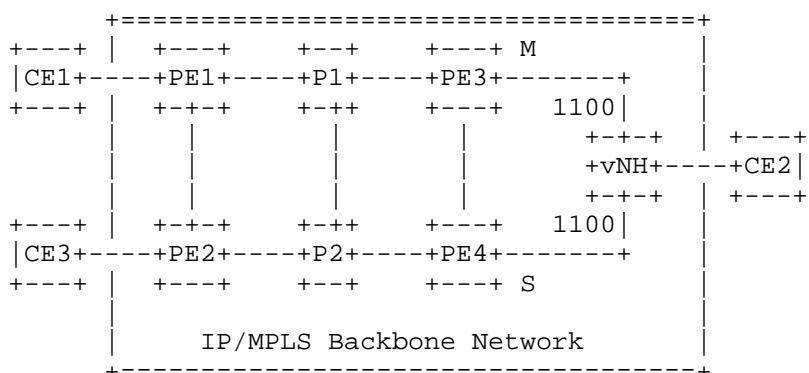


Figure 2.1: Egress PE routers share the same VPN route label.

A CE router is usually connected to multiple PE routers of the IP/MPLS backbone network for the sake of reliability. Figure 2.1 shows such a scenario. In this document, PE1 and PE2 are defined as ingress routers and PE3 and PE4 are defined as egress routers. Suppose PE3 is the primary PE while PE4 is the backup egress PE. In this document, we suppose there are two PEs in one RG. It's possible to expand the method to support more than two PEs in one RG, though it is out the scope of this document.

Those egress PE routers may discover each other as in the same RG from the CE routes learning process which can be a dynamic routing algorithm or a static routing configuration [RFC4364].

2.1. The Virtual Next Hop

Egress PEs create a vNH router in IGP to represent the set of CEs dual-homed to the same egress PEs in the Service Provider's backbone. The PE with the highest priority in the RG determines the loopback IP address for the vNH. This loopback IP address can be configured manually or automatically. The SystemID of the vNH under ISIS is composed based on this loopback IP address. The router LSA/LSP for the vNH is generated by the egress PE with the highest priority. This router LSA/LSP also includes the the outgoing links of the vNH. For the incoming links of the vNH, all egress PEs need include these P2P adjacencies in their router LSAs/LSPs.

Egress PEs may create multiple vNHs for one CE. Then multiple tunnels can be set up from ingress PEs to the vNHs. Ingress PEs can choose from these tunnel routes to achieve load balance for the CE.

The overload mode MUST be set so that the rest routers in the network will not route transit traffic through the vNH. In OSPF, the overload

mode can be set up through setting the link weights from the vNH to egress PEs to the maximum link weight which is 0xFFFF. In ISIS, this overload mode is realized as setting the overload bit in the LSP of the vNH.

2.1.1.1. Generating OSPF LSAs

The following Type 1 Router-LSA is flooded by the egress PE with the highest priority. As defined in [RFC2328], this LSA can only be flooded throughout a single area.

0										1										2										3									
0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1								
LS age										Options										LS type																			
Link State ID																																							
Advertising Router																																							
LS sequence number																																							
LS checksum																				length																			
0					V E B					0					# links																								
Link ID																																							
Link Data																																							
Type										# TOS										metric																			
...																																							
TOS										0										TOS metric																			
Link ID																																							
Link Data																																							
...																																							

LS age

The time in seconds since the LSA was originated. (Set to 0x708 by default.)

Options

As defined in [RFC2328], options = (E-bit).

LS type
1

Link State ID
Same as the Advertising Router

Advertising Router
The Router ID of the vNH.

LS sequence number
As defined in [RFC2328].

LS checksum
As defined and computed in [RFC2328].

length
The length in bytes of the LSA. This includes the 20 byte LSA header. (As defined and computed in [RFC2328].)

VEB
As defined in [RFC2328], set its value to 000.

#links
The number of router links described in this LSA. It equals to the number of Egress PEs in the RG.

The following fields are used to describe each router link connected to an egress PE. Each router link is typed as Type 1 Point-to-point connection to another router.

Link ID
The Router ID of one of the egress PEs in the RG.

Link Data
It specifies the interface's MIB-II [RFC1213] ifIndex value. It ranges between 1 and the value of ifNumber. The ifNumber equals to the number of the PEs in the RG. The PE with the highest priority sorts the PEs according to their unsigned integer Router ID in the ascend order and assigns the ifIndex for each.

Type
Value 1 is used, indicating the router link is a point-to-point connection to another router.

TOS
This field is set to 0 for this version.

Metric

It is set to 0xFFFF.

The fields used here to describe the virtual router links are also included in the Router-LSA of each egress PEs. The Link ID is replaced with the Router ID of the vNH. The Link Data specifies the interface's MIB-II [RFC1213] ifIndex value. The "Metric" field is set as defined in Section 2.2.

2.1.2. Generating ISIS LSPs

The primary egress PE generates the following level 1 LSP to describe the vNH node.

	No. of octets
+-----+ Intradomain Routeing Protocol Discriminator	1
+-----+ Length Indicator	1
+-----+ Version/Protocol ID Extension	1
+-----+ ID Length	1
+-----+ R R R PDU Type	1
+-----+ Version	1
+-----+ Reserved	1
+-----+ Maximum Area Address	1
+-----+ PDU Length	2
+-----+ Remaining Lifetime	2
+-----+ LSP ID	ID Length + 2
+-----+ Sequence Number	4
+-----+ Checksum	2
+-----+ P ATT LSPDBOL IS Type	1
+-----+ : Variable Length Fields :	Variable
+-----+	

Intradomain Routeing Protocol Discriminator - 0x83 (as defined in [ISIS])

Length Indicator - Length of the Fixed Header in octets

Version/Protocol ID Extension - 1

ID Length - As defined in [ISIS]

PDU Type (bits 1 through 5) - 18

Version - 1

Reserved - transmitted as zero, ignored on receipt

Maximum Area Address - same as the primary egress PE

PDU Length - Entire Length of this PDU, in octets, including the header.

Remaining Lifetime - Number of seconds before this LSP is considered expired. (Set to 0x384 by default.)

LSP ID - the system ID of the source of the LSP. It is structured as follows:

+-----+	
Source ID	6
+-----+	
Pseudonode ID	1
+-----+	
LSP Number	1
+-----+	

Source ID - SystemID of the vNH

Pseudonode ID - Transmitted as zero

LSP Number - Fragment number

Sequence Number - sequence number of this LSP (as defined in [ISIS])

Checksum - As defined and computed in [ISIS]

P - Bit 8 - 0

ATT - Bit 7-4 - 0

LSDBOL - Bit 3 - 1

IS Type - Bit 1 and 2 - bit 1 set, indicating the vNH is a Level 1 Intermediate System

In the Variable Length Field, each link outgoing from the vNH to an egress PE is depicted by a Type #22 Extended Intermediate System Neighbors TLV [RFC5305]. The egress PE is identified by the 6 octets SystemID plus one octet of all-zero pseudonode number. The 3 octets metric is set as that in Section 2.2. None sub-TLVs is used by this version, therefore the value of the one octet length of sub-TLVs is 0. The Type #22 TLV requires 11 octets.

The Type #22 TLV is also included in the LSP of each egress PE to depict the incoming link of the vNH. Only the 6 octets SystemID is replaced with the SystemID of the vNH.

2.2. Link Costs Set Up for IGP FRR

Tunnel LSPs are set up based on IGP routes through LDP signaling. If the IGP costs for the links between egress PEs and the vNH can be set up in a way that one egress PE appears on the primary path while other PE(s) appears on the backup path, the PLR can make use of the multiple egress PEs to achieve fast failure protection. Suppose [LFA] is being used as the IGP FRR mechanism, the link weights can be set up according to the following rule.

1. This document supposes bidirectional link weights are being used. Assume the weight for the link between PE3 and vNH is "M" and the weight for the link between PE4 and vNH is "S". The weight for the link between PE3 and PE4 is C34.

2. Px is a neighbor of PE3. This Px will act as the PLR. Suppose Pxy is Px's neighbor with the shortest path to PE4, after PE3 is removed from the topology. The cost of this path is Sxy4.

3. Add PE3 back to the topology. The cost of the path from Pxy to PE3 is Sxy3.

4. "M" and "S" can be set up as long as the following two equations hold.

$$\text{eq1: } Sxy4+S < Sxy3+M$$

$$\text{eq2: } C34+S > M$$

Although this document designs the method based on [LFA] which is widely deployed, other IGP FRR mechanisms can also be utilized to

achieve the protection. For example, [MRT] is applicable regardless of how the link weights are set up.

2.3 Label Assignment and Processing

2.3.1. The VPN Route Label

Egress PEs use BGP to distribute to ingress PEs the routes that they have learnt from CEs [RFC4364]. When egress PEs distribute the routes of the VPN that the CE is in, they MUST assign the same "VPN route label" for one VPN (per VRF label assignment). This label will become the first label of a data packet. The IP address of the vNH is used as the "BGP next hop". For example, in Figure 2.1, both PE3 and PE4 use 1100 as the VPN route label for the routes learnt from CE2.

Suppose PE3 fails and the packet with VPN route label 1100 is redirected to PE4, PE4 recognizes 1100 as the VPN route label it assigned for the VPN that the CE is in. As specified in Section 5 of [RFC4364], PE4 will be able to determine, the attachment circuit over which the packet should be transmitted (to the CE) as well as the data link layer header for that interface. It need to lookup the packet's destination address in the VRF identified by the VPN route label 1100.

When we speak of a PE fails, it may also means that a link to the PE on the primary tunnel fails. In general, we can say that a primary PE fails means that this PE becomes unreachable via its upstream neighbor on the primary tunnel.

The shared label may be manually configured or negotiated through signaling between egress PEs. In [LS-ICCP], application TLVs are defined for [ICCP] to achieve such kind of signaling.

2.3.2. The Tunnel Label

This document supposes Label Distribution Protocol is being used as the tunneling technology. The LDP LSP tunnel follows a IGP route from ingress PEs to the vNH. The backup path to vNH can be calculated according to IGP FRR mechanism, such as [MRT] and [LFA].

The ingress PE tunnels the data packet through the backbone network using the "tunnel label" as the second entry of the label stack. The "VPN route label" is not visible again until the MPLS packet reaches the egress PE. The egress PE need pop the second label and deliver the packet according to the "VPN route label".

3. Security Considerations

This document raises no new security issues.

4. IANA Considerations

No requirements for IANA.

5. References

5.1. Normative References

- [LFA] Filsfils, C., Ed., Francois, P., Ed., Shand, M., Decraene, B., Uttaro, J., Leymann, N., and M. Horneffer, "Loop-Free Alternate (LFA) Applicability in Service Provider (SP) Networks", RFC 6571, June 2012.
- [RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.
- [ICCP] L. Martini, S. Salam, et al, "Inter-Chassis Communication Protocol for L2VPN PE Redundancy", draft-ietf-pwe3-iccp-11.txt, work in progress.
- [ISIS] ISO, "Intermediate system to Intermediate system routing information exchange protocol for use in conjunction with the Protocol for providing the Connectionless-mode Network Service (ISO 8473)," ISO/IEC 10589:2002.
- [RFC2328] Moy, J., "OSPF Version 2", STD 54, RFC 2328, April 1998.
- [RFC1213] McCloghrie, K. and M. Rose, "Management Information Base for Network Management of TCP/IP-based internets:MIB-II", STD 17, RFC 1213, March 1991.
- [LS-ICCP] M. Zhang, P. Zhou, "ICCP Application TLVs for VPN Route Label Sharing", draft-zhang-pwe3-iccp-label-sharing-00.txt, work in progress

5.2. Informative References

- [MRT] A. Atlas, Ed., R. Kebler, et al, "An Architecture for IP/LDP Fast-Reroute Using Maximally Redundant Trees", draft-ietf-rtgwg-mrt-frr-architecture-02.txt, work in progress.

Author's Addresses

Mingui Zhang
Huawei Technologies Co., Ltd
Huawei Building, No.156 Beiqing Rd.
Z-park, Shi-Chuang-Ke-Ji-Shi-Fan-Yuan, Hai-Dian District,
Beijing 100095 P.R. China

Email: zhangmingui@huawei.com

Peng Zhou
Huawei Technologies Co., Ltd
Huawei Building, No.156 Beiqing Rd.
Z-park, Shi-Chuang-Ke-Ji-Shi-Fan-Yuan, Hai-Dian District,
Beijing 100095 P.R. China

Email: Jewpon.zhou@huawei.com