INTERNET-DRAFT
Intended Status: Proposed Standard
Expires: December 18, 2014

Mingui Zhang
Peng Zhou
Huawei
Russ White
IETF
June 16, 2014

Label Sharing for Fast PE Protection
draft-zhang-l3vpn-label-sharing-02.txt

Abstract

This document describes a method to be used by VPN Service Providers to provide multi-homed CEs with fast protection of egress PEs. Egress PEs in a redundant group always share the same label in distribution of VPN routes of a VRF. A virtual Next Hop (vNH) in the IGP/MPLS backbone is created as the common end of LSP tunnels which would otherwise terminate at each egress PE. Primary and backup LSP tunnels ended at the vNH are set up by MPLS on basis of existing IGP FRR mechanisms. If the primary egress PE fails, the backup egress PE can recognize the "shared" VPN route label carried by the data packets. Therefore, the failure affected data packets can be smoothly rerouted to the backup PE for delivery without changing their VPN route label.

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/1id-abstracts.html

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

Copyright and License Notice

Table of Contents

1. Introduction

For the sake of reliability, ISPs often 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 a 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 advertise 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 (Point of Local Repair) 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".

1.1. Overview

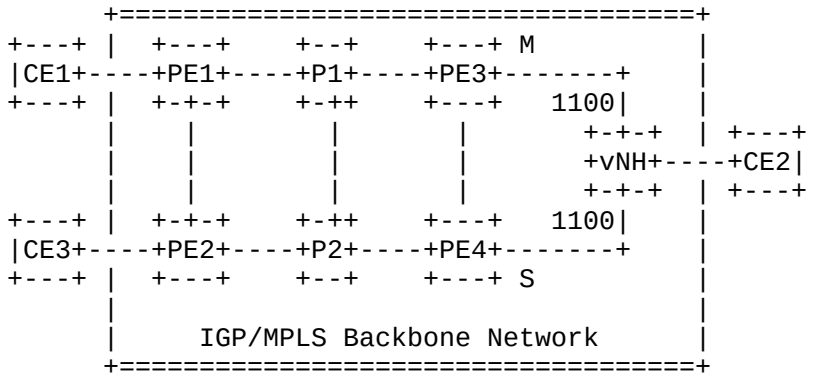


Figure 1.1: Egress PE routers share the same VPN route label 1100.

An example topology is shown in Figure 1.1. Let PE1 and PE2 be ingress routers, and let PE3 and PE4 be egress routers. CE2 is connected to both PE3 and PE4 so they form an Redundant Group (RG). Usually, egress PEs may be configured to be in the same RG or discover each other from the CE routes learning process which can be a dynamic routing algorithm or a static routing configuration [RFC4364]. Suppose PE3 is the primary while PE4 is the backup. For topologies with more than two egress PEs in an RG, one PE acts as the primary while other act as backups.

A vNH node is created in the backbone. The primary PE allocates a loopback IP address to vNH (say 2.2.2.2). Instead of the egress PEs, vNH acts as the common end node of LSP tunnels which otherwise end at

egress PEs. The metrics ('M' and 'S') for the links between egress PEs and vNH is set up in a way that the primary and backup LSP tunnels to vNH respectively use PE3 and PE4 as the penultimate hop.

Egress PEs in an RG MUST advertise the same VPN route label for each VPN connected to this RG. When a route is learn from CE2 (say 10.9.8/24), PE3 and PE4 will distribute this route to other PEs sharing the same label (say 1100). In this way, when the primary PE fails, the VPN route label carried with the rerouted data packets need not be changed. It can be recognized by the backup PE as well.

This document supposes BGP/MPLS IP VPN [RFC4364] is deployed in the backbone and Label Distribution Protocol (LDP) is used to distribute MPLS labels. The approach developed in this document confines changes to routers in an RG. P and PE routers out of this RG are totally oblivious to these changes.

1.2. 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].

1.3. Terminology

VRF: Virtual Routing and Forwarding table [RFC4364]

FRR: Fast ReRouting

PLR: Point of Local Repair

LFA: Loop-Free Alternate [LFA]

RG: Redundant Group. A Redundant Group of Provider Edge nodes (PEs) to which a set of CEs are multi-homed.

2. The Virtual Next Hop

A virtual router (the virtual Next Hop, vNH) is created in IGP to represent the RG in the Service Provider's backbone. For other routers in the backbone, the vNH acts as the common egress PE connecting a set of CEs. Multiple vNHs may be created for one RG. Then multiple paths can be computed from ingress PEs to the vNHs. Ingress PEs can choose from these paths to achieve load balance for the CEs.

Service Providers may configure one PE to be the primary when an RG is created. The primary PE may also be automatically elected out of

the RG in the same way the DR is selected (see section 7.3 of [RFC2328]), or the DIS is selected [ISIS]. Other PEs in the RG will act as backup ones. This primary PE determines the loopback IP address for the vNH. This loopback IP address can be configured manually or assigned automatically. The SystemID of the vNH under ISIS is composed based on this loopback IP address. The primary PE generates the router link state information (LSA/LSP) on behalf of the vNH. Links to each PE and each CE in the group are included in router link state information PDUs of the PE and 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. (See Appendix A and B for the detail set up of LSAs/LSPs.)

3. Link Costs Set Up for IGP FRR

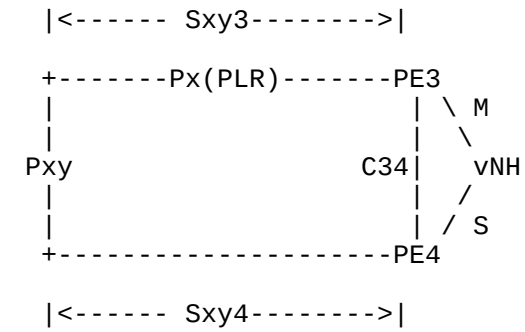


Figure 2.2: The illustration of equations.

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 the other PE appears on the backup path, the PLR can make use of the multiple egress PEs to achieve fast failure protection. Link weights can be set up according to the following rule in order to leverage the well supported [LFA] as the IGP FRR mechanism.

1. This document supposes bidirectional link weights are being used. As illustrated in Figure 2.2, 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.

eq1: Sxy4+S < Sxy3+M

eq2: C34+S > M

The eq1 guarantees that Pxy is safe, i.e., no loop occurs, to be used as the next hop by the PLR for bypass. The eq2 is designed to insure that the primary path does not go through the primary egress PE and backup egress PE in series.

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] can be applicable regardless of how the link weights are set up.

4. The LSP Tunnels

Egress PEs use the IP address of the vNH to identify the FEC. Its LSPs on basis of IGP routes with vNH as the last hop are set up using LDP:

- The primary LSP tunnel follows the IGP route from ingress PEs to the vNH;

- The backup LSP tunnel is set up according to existing IGP FRR calculation, such as [MRT] and [LFA].

Data packets are tunneled through the backbone using a "tunnel label" at the top of the label stack. Egress PE will not really transmit a packet to the tunnel end node vNH. Rather, they need locally deliver the packet. It can be interpreted that at the egress PE, the packet’s next hop is the egress PE itself (see Section 3.10 of [RFC3031]). The tunnel label will be popped at the egress PE. The indication for popping is got from the tunnel label at the top of the stack since this is a label assigned to the FEC identified by the PE’s loopback IP address. Next, there will be a pop of the VPN route label followed by an address lookup in the VRF. Section 5 will explain how to set the VPN route label in order to leverage these LSP tunnels to achieve the egress PE protection.

5. The VPN Route Label

5.1. Sharing the VPN Route Label

In [RFC4364], egress PEs separately allocate and distribute the label for the route to an address prefix they learn from CEs. In this document, it's REQUIRED that backup PE(s) in an RG always advertises the label already advertised by the primary PE for the address prefix in question. The primary PE RG SHOULD distribute the same label for any address prefix in an attached VPN. This is per VRF label sharing. Others granularities, such as per address family per VRF label sharing, are also feasible.

Egress PEs continue to locally allocate VPN route labels so that the proposal need not modify existing forwarding processes of L3VPN egress PEs. At the backup egress PE, the allocated label and the distributed label would be inconsistent. The following two options arise to address this issue.

5.1.1. Option A: Reserved Label Ranges per RG

PEs in an RG are physically connected to the same set of CEs. It's viable for them allocate the same VPN route label per VPN. For each VPN served by an RG, the backup egress PE always allocates the same label as the primary PE. It acts as a 'compromised' network entity which always listens to the label advertised by the primary then allocates and also distributed the same label. By doing this, they are intimating the VPN route label allocation of the virtual node, vNH.

For this option, PEs in an RG are REQUIRED to reserve the same label range(s) for allocation at the management plane. PEs with h/w disjoint label ranges are not qualified for this option. This option SHOULD only be used in well managed and highly monitored networks. It's not intended to be applicable when the RG spans more than one administrative domain. It ought not to be deployed on or over the public Internet.

Note that if one PE participates in multiple RGs, a label range reserved for one RG can't be used by another RG on this PE. It increases the consumption of labels on this PE. So this option should be deployed with care in this case.

The architecture of the label sharing method allows a 'higher-layer' entity to allocate labels for all PEs across all RGs. This document leaves this choice as for future study.

5.1.2. Option B: The Label Swapping Table

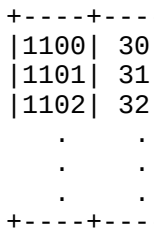


Figure 2.3: The label 'swapping' table

In the inter-AS L3VPN Option B defined in Section 10 of [RFC4364], when an ASBR distributes a VPN route to an ASBR in another AS, it need perform a label swap for this route. Similarly, the backup PE in this proposal uses a label swapping table to record the mapping between advertised labels and locally assigned labels for VPN routes. Obviously, the backup PE need maintain one such table per RG. Whenever a data packet to a route in a VPN attached to the RG arrives at the backup PE, the locally assigned label (e.g., 30) got from the swapping will be used in the VPN route label lookup followed by an address lookup.

5.2. Binding to LSP Tunnels

When the VPN route with a shared label is distributed to other PEs by the primary PE and backup PEs, the BGP next hop is set to the IP address of the vNH. As defined in Section 4, LSP tunnels are set up for the FEC identified also by the IP address of the vNH. By doing this, the VPN route is bound to these LSP tunnels. When data packets to this VPN route are tunneled through the backbone, these LSP tunnels will offer the protection.

6. Examples To Walk Through

Two examples are included in this section. Figure 1.1 is referred. The first one describes how to distribute VPN route label to peers. It's westbound in the control plane. The second one interprets how egress PE act in the case of the primary PE failure. It's eastbound in the data plane.

6.1. Label Distribution Procedure

Assume PE3 is elected as the primary while PE4 is the backup. The loopback IP address of vNH is 2.2.2.2.

1) PE3 learns the VPN route to address prefix 10.9.8/24 from CE2. It allocates the VPN route label 1100 and distributes it in BGP with 2.2.2.2 as the BGP Next Hop. (prefix = 10.9.8/24|label = 1100|BGP

Next Hop = 2.2.2.2)

2) PE4 also learns the VPN route to address prefix 10.9.8/24 and allocate the VPN route label 30. It then waits for the primary PE3 to advertise the VPN route label for this prefix.

3) PE4 monitors the VPN route label 1100 from PE3 for the prefix 10.9.8/24. The mapping from 1100 to 30 is inserted to the swapping table.

4) PE4 distributes the VPN route using the monitored label 1100. (prefix = 10.9.8/24|label = 1100|BGP Next Hop = 2.2.2.2)

6.2. Protection Procedure

Suppose the label for the primary LSP tunnel to vNH is 2100 while the backup LSP tunnel to vNH is 3100. P1 is the PLR.

1) In normal case, P1 sends data packets with tunnel label 2100 to PE3. When PE3 fails, P1 redirects data packets to the backup LSP tunnel (say P2-PE4-vNH) using tunnel label 3100.

2) PE4 will receive a packet with two levels of labels. It pops the outer label 3100 and use this label to identify a swapping table.

3) PE4 pops the VPN route label and looks up the swapping table. The VPN route label 1100 is mapped to 30.

4) The VPN route label 30 is looked up in the VPN route label table followed by an address lookup in the VRF.

7. Operations

7.1. Label Space Management for Option A

A label range should be reserved before an RG comes to operate. Operators need set a large label sharing space for label ranges reservation. When an RG is created, the operator needs reserve a unused label range for it. The label range should be reserved in a manner of ’enough is enough’. If a label range of an RG is being used out, the operator can reserve a new range from the unused label sharing space. The newly reserved range is then appended to the one being used out.

If a backup PE is partitioned from the primary PE, it continues to work with those allocated labels for the RG. However, it MUST NOT allocate any more labels in the reserved ranges. A label in a reserved range can only be allocated by a backup PE when it monitors

that the primary PE has distributed this label.

When a primary PE resumes from a failure, its reserved label ranges come to work again. It SHOULD conserve the labels it allocated for each range.

7.2. Backup LSP Tunnel Exceptions

The label sharing method requires that the backup LSP tunnel is set up as specified in Section 4, following the IGP route. However, Service Providers are allowed to have exceptions. For instance, an operator may use BGP Local_Pref to give a higher degree of preference to the route advertised by the primary PE. For another instance, the operator may have the primary PE advertise a more specific prefix. Take Figure 1.1 for example, the backup tunnel will actually goes through PE4->PE3->CE2 for both instances. When the VPN route is bound to this tunnel, it does not protect the primary egress PE. An alarm should be generated to notify the operator that such kind of configuration will jeopardize the VPN route’s resilience to egress PE node failure.

8. Security Considerations

This document raises no new security issues.

9. IANA Considerations

This document requires no IANA actions. RFC Editor: please remove this section before publication.

Acknowledgements

Authors would like to thank the comments and suggestions from Bruno Decraene, Eric Rosen, Eric Gray, Jakob Heitz, James Uttaro, Jeff Tantsura, Loa Andersson, Nagendra Kumar, Robert Raszuk, Stewart Bryant, Shunwan Zhuang, Wim Henderickx and Zhenbin Li.

10. References

10.1. 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.

[ISIS] ISO, "Intermediate system to Intermediate system routeing 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.

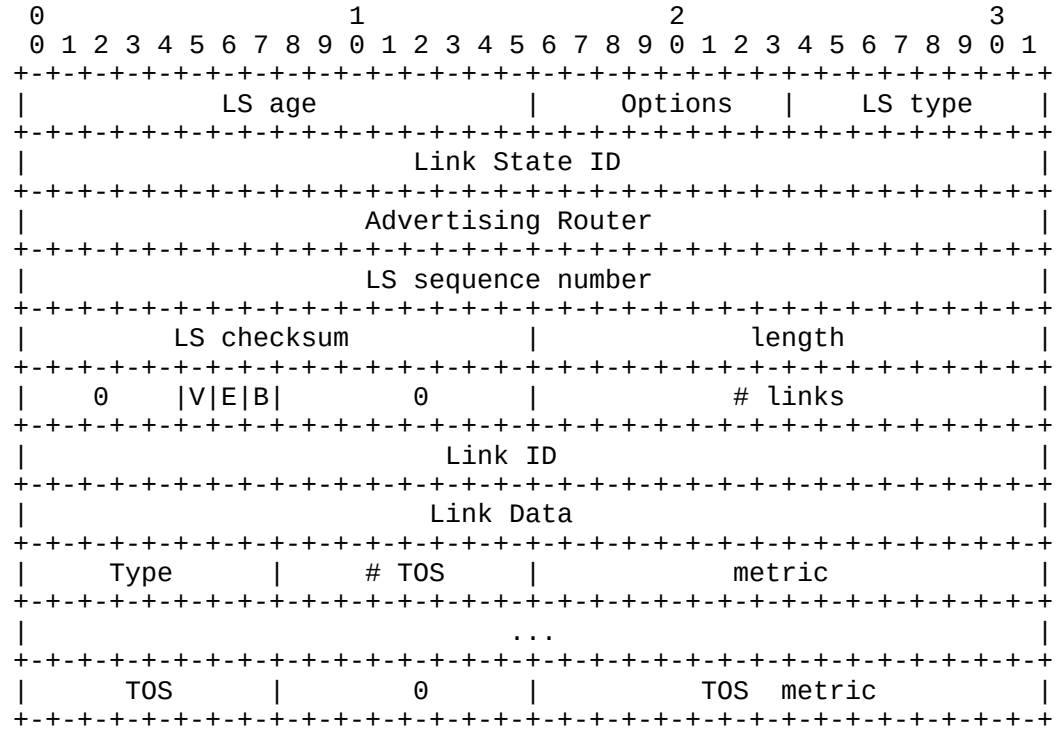
[RFC1213] McCloghrie, K. and M. Rose, "Management Information Base for Network Management of TCP/IP-based internets:MIB-II", STD 17, RFC 1213, March 1991.

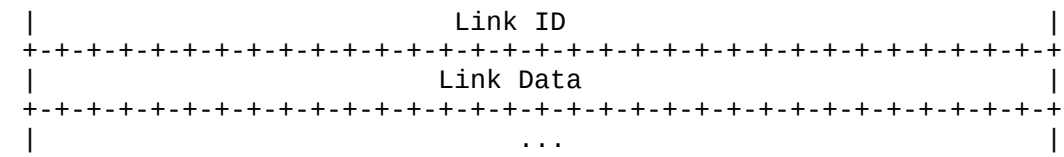
10.2. Informative References

[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, work in progress.

Appendix A: 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.





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.

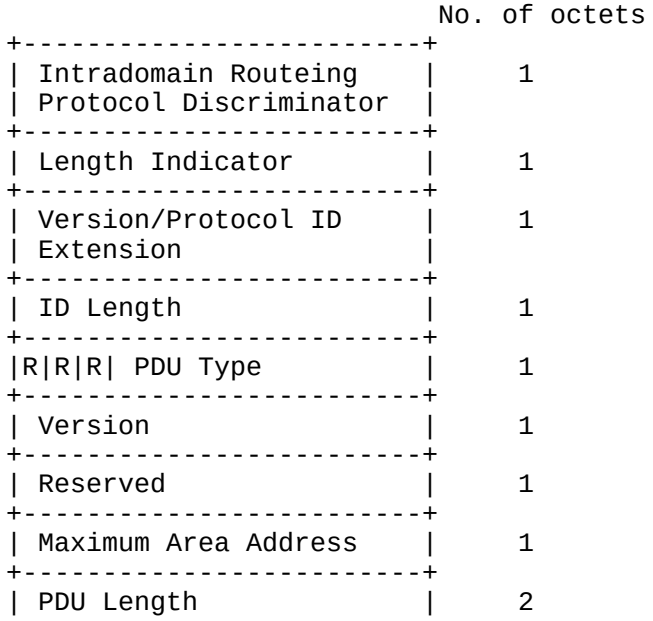
TOS
This field is set to 0 for this version.

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 3.

Appendix B: Generating ISIS LSPs

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



```
+-------------------------+
| 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 3. 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.

Author's Addresses

Mingui Zhang
Huawei Technologies
No.156 Beiqing Rd. Haidian District,
Beijing 100095 P.R. China

Email: zhangmingui@huawei.com

Peng Zhou
Huawei Technologies
No.156 Beiqing Rd. Haidian District,
Beijing 100095 P.R. China

Email: Jewpon.zhou@huawei.com

Russ White
Verisign
12061 Bluemont Way
Reston, VA 20190
USA

Email: russw@riw.us