INTERNET-DRAFT
Intended status: Proposed Standard

Mingui Zhang
Peng Zhou
Donald Eastlake
Huawei
Russ White
IETF

Expires: September 20, 2018
March 21, 2018

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

Abstract

This document describes a method to be used by VPN (Virtual Private Network) 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 the basis of existing Interior Gateway Protocol (IGP) Fast ReRoute (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.

Distribution of this document is unlimited. Comments should be sent to the authors or the TRILL working group mailing list: trill@ietf.org.

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.

Table of Contents

1. Introduction

For the sake of reliability, ISPs often connect one CE (Customer Edge) device to multiple PE (Provider Edge devices. 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 50 msec switchover.

VPN (Virtual Private Network) 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 called "Label Sharing for Fast PE Protection".

1.1 Overview

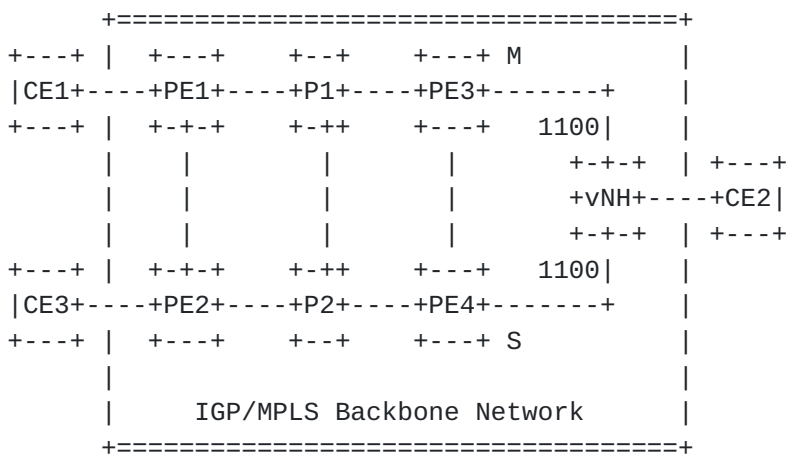


Figure 1.1: Egress PE routers share the same VPN route label 1100

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 others act as backups.

A vNH (virtual Next Hop) node is created in the backbone. The primary

PE allocates a loopback IP address to vNH (say 192.0.2.2). 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 a BGP/MPLS IP VPN [RFC4364] is deployed in the backbone and a Label Distribution Protocol (LDP) is used to distribute MPLS labels. The approach developed in this document confines changes to routers in an RG. Provider and PE routers outside 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 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

1.3 Terminology

CE: Customer Edge device, e.g., a host or network.

FEC: Forwarding Equivalency Class

FRR: Fast ReRoute [RFC7812]

LFA: Loop-Free Alternate [RFC6571]

LSP: Label Switched Path

PE: Provider Edge

PLR: Point of Local Repair

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

VRF: Virtual Routing and Forwarding table [RFC4364]

2. The Virtual Next Hop

A virtual router (the virtual Next Hop, vNH) is created in IGP to represent the Redundant Group (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 Designated Routed is selected in OSPF (see section 7.3 of [RFC2328]) or the Designated Intermediate System is selected in [IS-IS]. 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 IS-IS 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 of the 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 IS-IS, 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

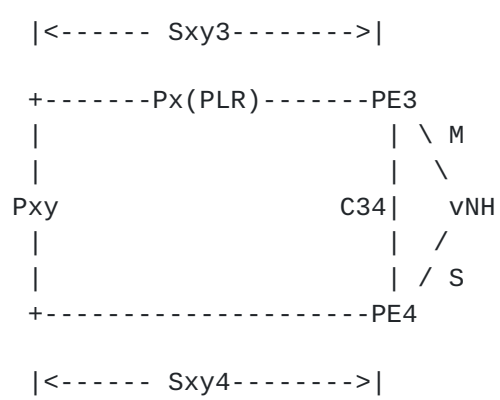


Figure 3: Illustration of equations.

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 [RFC6571] as the IGP (Interior Gateway Protocol) FRR (Fast ReRoute [RFC7812]) mechanism.

1. This document supposes bidirectional link weights are being used. As illustrated in Figure 3, 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 to be used as the next hop by the PLR for bypass, i.e., no loop occurs. 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 Loop Free

Alternative (LFA [RFC6571]) which is widely deployed, other IGP FRR

[RFC7812] mechanisms can also be utilized to achieve the protection. For example, maximally redundant trees [RFC7812] 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 are set up using LDP on basis of IGP routes with vNH as the last hop:

- 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 [RFC7812] calculation, such as maximally redundant trees [RFC7812] or LFA [RFC6571].

Data packets are tunneled through the backbone using a "tunnel label" at the top of the label stack. Egress PEs will not really transmit a packet to the tunnel end node vNH. Rather, they need to 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 tunnel label at the top of the stack indicates popping 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 to use 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 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 to 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 hardware-set 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 that 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 for future study.

5.1.2 Option B: The Label Swapping Table

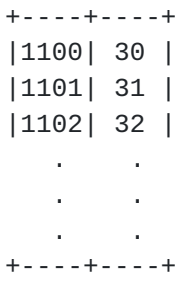


Figure 5.1.2: The label swapping table

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 needs to 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 needs to 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) obtained 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 specified in Section 4, LSP tunnels are set up for the FEC also identified 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 protection.

6. Examples To Walk Through

Two examples are included in this section using the topology in Figure 1.1. The first one describes how to distribute a VPN route label to peers. It's westbound in the control plane. The second one interprets how an egress PE acts 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 192.0.2.2.

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 192.0.2.2 as the BGP Next Hop. (prefix = 10.9.8/24|label = 1100|BGP Next Hop = 192.0.2.2)

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 = 192.0.2.2)

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 is made operational. Operators need to set a large label sharing space to reserve for label ranges. 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 becomes exhausted, the operator can reserve a new range from the unused label sharing space. The newly reserved range is then appended to the one being exhausted.

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 are again available to it. 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. In Figure 1.1, for example, the backup tunnel 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 configuration will jeopardize the VPN route's resilience to egress PE node failure.

8. Security Considerations

TBD

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.

Normative References

[IS-IS] 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.

[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.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[RFC2328] Moy, J., "OSPF Version 2", STD 54, RFC 2328, DOI 10.17487/RFC2328, April 1998, <https://www.rfc-editor.org/info/rfc2328>.

[RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol Label Switching Architecture", RFC 3031, DOI 10.17487/RFC3031, January 2001, <https://www.rfc-editor.org/info/rfc3031>.

[RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.

[RFC5305] Li, T. and H. Smit, "IS-IS Extensions for Traffic Engineering", RFC 5305, DOI 10.17487/RFC5305, October 2008, <https://www.rfc-editor.org/info/rfc5305>.

[RFC6571] 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.

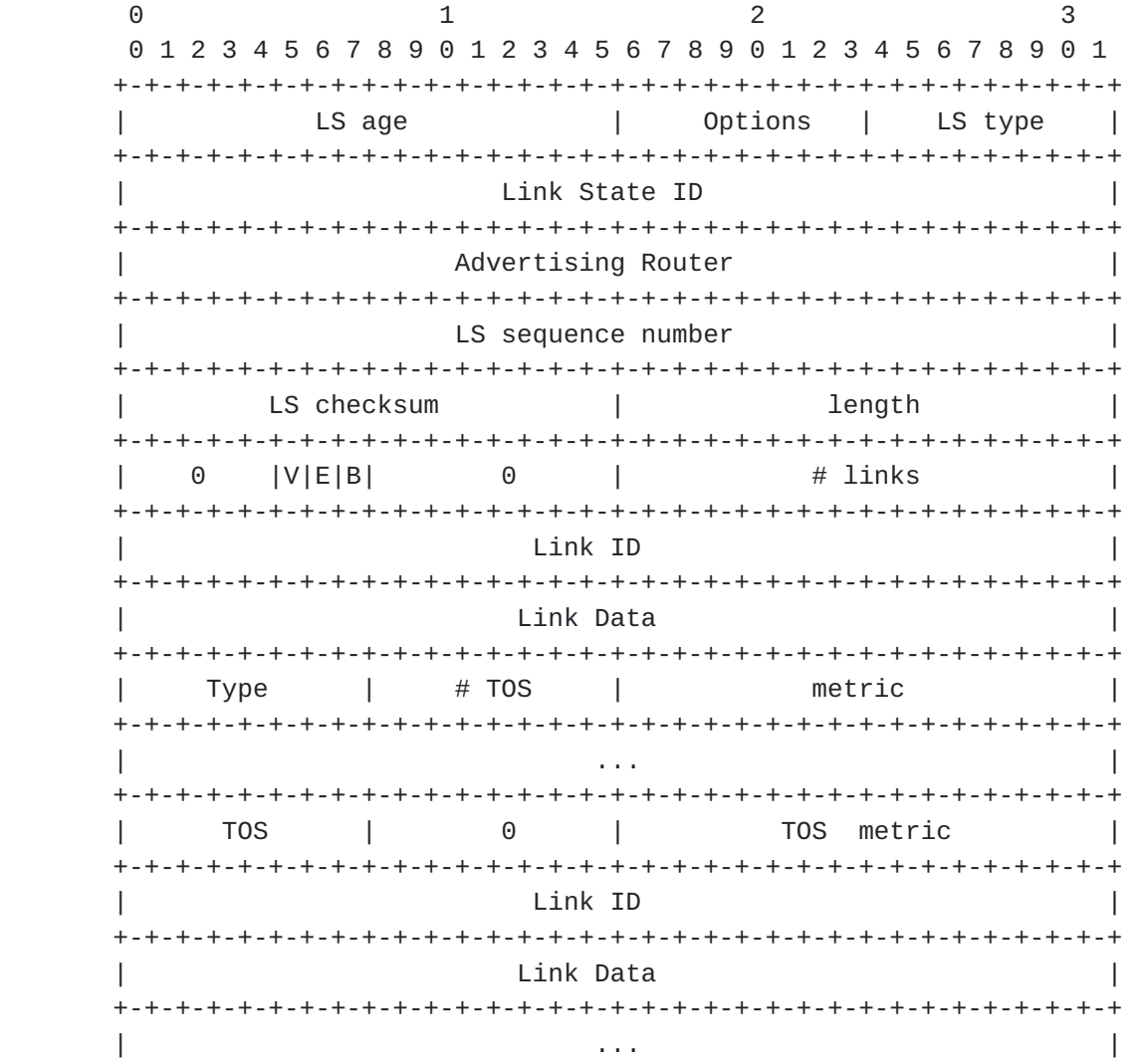
[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

Informative References

[RFC7812] Atlas, A., Bowers, C., and G. Enyedi, "An Architecture for IP/LDP Fast Reroute Using Maximally Redundant Trees (MRT-FRR)", RFC 7812, DOI 10.17487/RFC7812, June 2016, <https://www.rfc-editor.org/info/rfc7812>.

Appendix A: Generating OSPF LSAs

The following Type 1 Router-LSA is flooded by the egress PE with the highest priority. As specified 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 (a half an hour) 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 IS-IS 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 specified in [IS-IS])

Length Indicator - Length of the Fixed Header in octets

Version/Protocol ID Extension - 1

ID Length - As defined in [IS-IS]

PDU Type (bits 1 through 5) - 18

Version - 1

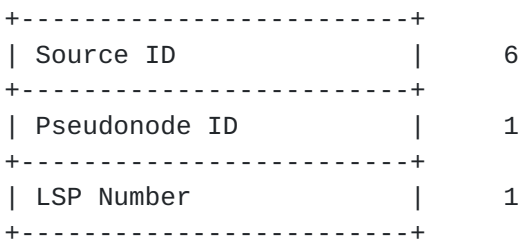
Reserved - transmitted as zero, ignored on receipt

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 (fifteen minutes) 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 [IS-IS])

Checksum - As defined and computed in [IS-IS]

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. No sub-TLVs are 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 but in this case the 6 octets SystemID is replaced with the SystemID of the vNH.

Authors' 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

Donald Eastlake, 3rd
Huawei Technologies
155 Beaver Street
Milford, MA 01757 USA

Email: d3e3e3@gmail.com

Russ White
Verisign
12061 Bluemont Way
Reston, VA 20190
USA

Email: russ@riw.us

Copyright, Disclaimer, and Additional IPR Provisions