

Network Working Group
Internet Draft

Category: Standard Track

L. Yong
W. Hao
D. Eastlake
Huawei
A. Qu
J. Hudson
Brocade

Expires: December 2014

June 12, 2014

IS-IS Protocol Extension For Building Distribution Trees
draft-yong-isis-ext-4-distribution-tree-02

Abstract

This document proposes an IS-IS protocol extension for automatically building bi-directional distribution trees to transport multi-destination traffic in an IP network.

Status of this document

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/ietf/lid-abstracts.txt>.

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

This Internet-Draft will expire on December 12, 2014.

Copyright 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.

Table of Contents

1. Introduction.....	3
1.1. Conventions used in this document.....	4
2. IS-IS Protocol Extension.....	5
2.1. RTADDR sub-TLV.....	5
2.2. RTADDRV6 sub-TLV.....	6
2.3. The Group Address Sub-TLV.....	7
3. Procedures.....	8
3.1. Distribution Tree Computation.....	8
3.2. Parent Selection.....	8
3.3. Parallel Local Link Selection.....	9
3.4. Tree Selection for a Group.....	10
3.5. Pruning a Distribution Tree for a Group.....	10
3.6. Reverse Path Forwarding Check (RPFC).....	10
3.7. Forwarding Using a Pruned Distribution Tree.....	11
3.8. Local Forwarding at Edge Router.....	11
3.9. Distribution Tree across different IGP Levels.....	12
4. Mobility Support.....	14
4.1. Listener moves from one edge router to another.....	14
4.2. Source host moves from one edge router to another.....	14
5. Backward Compatibility.....	14
6. Interworking with PIM.....	14
7. Security Considerations.....	14
8. IANA Considerations.....	14
9. Acknowledgements.....	15
10. References.....	15
10.1. Normative References.....	15
10.2. Informative References.....	15

1. Introduction

Computer virtualization and cloud applications motivate the DC network virtualization technology [NVO3FRWK]. This technology decouples the end-points networking from the DC physical infrastructure network in terms of address space and configuration [NVO3FRWK].

DC network virtualization solutions are required to carry all types of traffic in today's DC physical networks including multi-destination traffic. It is also desirable to use an IP network as the DC underlying network for the overlay virtual networks [NVO3FRWK].

IP network technology does not yet support multi-destination traffic forwarding. A variety of Protocol Independent Multicast (PIM) solutions [RFC4601] [RFC5015] are designed to carry IP multicast traffic over IP networks. However DC infrastructure for multi-tenancy application is simple IGP domain where using PIM for multicast transport has several drawbacks. This is because the PIM use their own hello protocol and hop-to-hop Join/Leave message so each router does not have global information about the receivers; in the PIM, the data packets could be forwarded unnecessarily to the Rendezvous Point(RP), and then get dropped there when no receiver at all or the sender and receivers for a multicast group are on the same branch towards the RP. This can unnecessarily consume network resources. Furthermore PIM solutions maintain a lot of soft-state, have intensive CPU utilization, and have additional convergence time, besides the IGP's, under a failure condition.

Although the PIM protocol is mature and has been deployed in IP networks, applying PIM to DC IP network that supports the Network Virtualization Overlays can be an extremely challenging [MCASTISS] [DCMCAST]. For example, VXLAN [VXLAN] solutions require multicast support in the underlying network to simulate overlay L2 broadcast capability, where every edge node in an overlay virtual network (VN) is a multicast source and receiver. An overlay VN topology may be sparse and dynamic compared to the underlying IP network topology. Also a large number of overlay VNs may exist in a DC, which PIM solutions can't scale to.

Furthermore IP Overlay based network virtualization technology has been adopted by network vendors to create a VN automatically, self-healing, multi-service fabric to achieve the goal of a SDN capable fabric which is open, programmable, and elastic. Within the fabric, it is a closed IP network carrying all types of traffic, hence

having one control plane protocol to support both uni-destination and multi-destination forwarding.

This is the motivation to extend IGP protocol in support multicast transport so one IGP protocol can support both unicast and multicast transport. This document uses extensions to the IS-IS protocol to build a distribution tree for multi-destination traffic transport in an IP network. A router uses either a Router Capabilities TLV or an MT Router Capabilities TLV to announce the tree root address and the multicast groups associated to the tree. With this information, routers in the IGP can compute rooted distribution trees by using the link state information, i.e. LSDB, and shortest path algorithm. Edge routers include information in their LSPs to announce their multicast group-memberships. Routers perform distribution tree pruning for each multicast group based on other router's group membership announcements. A router forwards the multi-destination traffic along the pruned tree.

In case that edge router needs to get the host membership of a multicast group, edge routers may use IGMP query messages [RFC3376] to inform the attached hosts and the hosts use IGMP report message to response with their interested multicast group(s).

In cases where the solution described in this document applies to the underlying network that transports overlay virtual networks [NVO3FRWK], mapping between an overlay multicast group and a underlying multicast group is necessary. Edge routers further need to perform packet encapsulation/decapsulation.[NVO3FRWK]

The benefits of this solution are 1) protocol convergence: use single protocol for both unicast and multicast traffic transport and get the same convergence time for unicast and multicast traffic. 2) multi-destination transport simplification: rely on the LSDB for computing a distribution tree and not run PIM hello protocol. 3) forwarding efficiency: no need to always forward the traffic to the RP; 4) better scalability: no need to maintain heavy PIM soft states. TRILL [RFC6325] has used IS-IS for both single destination and multi-destination packet transport, which proves the protocol capability of doing both.

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].

2. IS-IS Protocol Extension

2.1. RTADDR sub-TLV

This is a sub-TLV that is used in either a Router Capabilities TLV or an MT Capabilities TLV. Each RTADDR sub-TLV contains a root IPv4 address and multicast group addresses that associate to the tree. A router may use multiple RTADDR sub-TLVs to announce multiple root addresses and associated multicast groups with each root. RTADDR sub-TLV format is below.

```

+-----+
|subType=RTADDR | (1 byte)
+-----+
| Length | (1 byte)
+-----+
| Root IPv4 Address |
+-----+
|S| RESV | (1 byte)
+-----+
| Tree Priority | (1 byte)
+-----+
|Num of Groups | (1 byte)
+-----+
| Group Address (1) |
+-----+
| Group Mask (1) |
+-----+
~
+-----+
| GROUP Address (N) |
+-----+
| Group Mask (N) |
+-----+

```

Where:

subType: RTADDR (TBD)

Length: variable depending on the number of associated groups

Root IPv4 Address: IPv4 Address for a root

S bit: If set, the rooted tree for single area only. Otherwise, the rooted tree crosses multiple areas.

RESV: 3 reserved bits. MUST be sent as zero and ignored on receipt.

Tree Priority: An eight bit unsigned integer where larger magnitude means higher priority. Zero means no priority.

Num of Groups: the number of group addresses

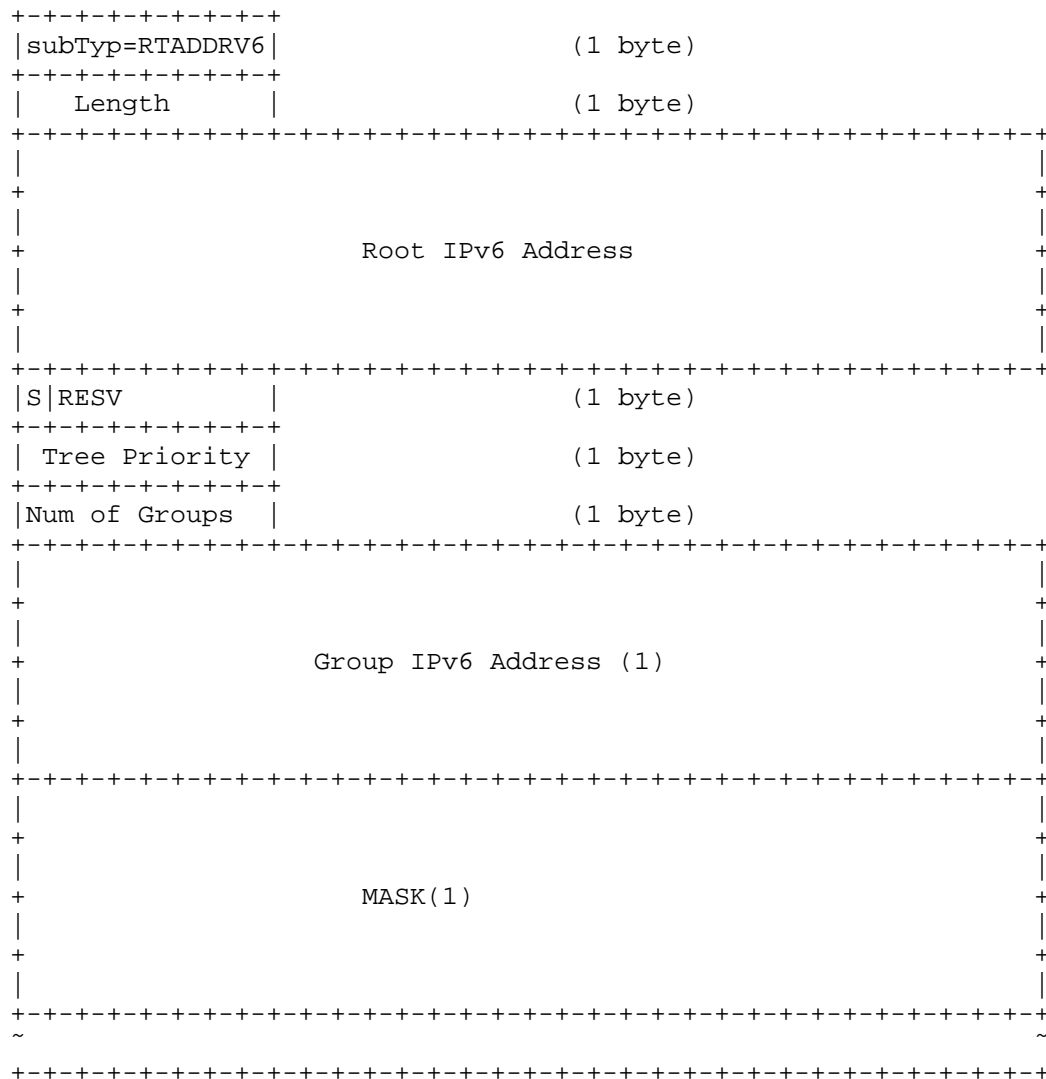
Group Address: IPv4 Address for the group

Group Mask: multicast group range

One router may be the root for multiple trees. Each tree associates to a set of multicast groups. In this case, a router encodes multiple RTADDR sub-TLVs to announce root addresses, one for each root, in either a Router Capabilities TLV or an MT Capabilities TLV. The group address/mask in different sub-TLVs can overlap. See section 3 for detail.

2.2. RTADDRV6 sub-TLV

This sub-TLV is used in an IPv6 network. It has the same format and usage except that the addresses are in IPv6.



2.3. The Group Address Sub-TLV

The Group Address TLV and a set of Group Address sub-TLVs are defined in RFC 7176 [RFC7176]. The GIP-ADDR and GIPV6-ADDR sub-TLVs are used in this solution. An edge router uses the GIP-ADDR sub-TLV or GIPV6-ADDR to announce its interested multicast groups. The GIP-

ADDR sub-TLV applies to an IPv4 network and GIPV6-ADDR sub-TLV for IPv6 network.

When using a GIP-ADDR or GIPV6-ADDR sub-TLV, the field VLAN-ID MUST set to zero and be ignored. Other field usage remains the same as [RFC7176]

3. Procedures

When an operator selects a router as a distribution tree root, he/she configures the tree root address and associated multicast groups on the router. A tree root address can be an interface address or router loopback address. After the configuration, the router will include a RTADDR sub-TLV, inside either a Router Capabilities TLV or an MT Capabilities TLV, where the tree root address and multicast groups are specified. If multiple trees are configured on the router, multiple RTADDR sub-TLVs are added in one or more Router Capabilities TLVs or MT Capabilities TLVs to specify individual tree roots. For IPv4 network, RTADDR sub-TLV is used. For IPv6, RTADDRV6 sub-TLV is used. Note that the rest of document specifies the processes for an IPv4 network only. The processes for an IPv6 network are the same.

Operators may associate one multicast group to more than one tree for the redundancy purposes and use the tree priority to specify the primary tree preference. Section 3.2 describes the primary tree selection.

3.1. Distribution Tree Computation

Upon receiving RTADDR sub-TLVs, routers track the tree roots and associated multicast groups. When the LSDB stabilizes, routers calculate all rooted trees according to the LSDB and shortest path algorithm.

One multicast group may associate to multiple trees. It is important that all the routers choose the same tree for a multicast group. Section 3.2 and 3.3 describes the tiebreaking rule for primary tree selection for a multicast group and parent selection in case of equal-cost to potential children.

3.2. Parent Selection

It is important, when building a distribution tree, that all routers choose the same links for the tree. Therefore, when there are equal costs from a potential child node to possible parent nodes, all

routers need to use the same tiebreakers. It is also desirable to allow splitting of traffic on as many links as possible in such situations. TRILL [RFC6325] achieves this by defining multiple rooted trees and using the tiebreakers to enable nodes in these trees to choose different parents. This draft uses the same tiebreakers as TRILL ([RFC6325] as clarified and updated by section 3.4 and 3.5 of [RFC7180]), and states as follow:

If there are k distribution trees in the network, when each router computes these trees, the k trees calculated are ordered and numbered from 0 to $k-1$ in ascending order according to root IP addresses.

The tiebreaker rule is: When building the tree number j , remember all possible equal cost parents for router N . After calculating the entire "tree" (actually, directed graph), for each router N , if N has " p " parents, then order the parents in ascending order according to the 7-octet IS-IS ID considered as an unsigned integer, and number them starting at zero. For tree j , choose N 's parent as choice $(j-1) \bmod p$.

3.3. Parallel Local Link Selection

If there are parallel point-to-point links between two routers, say $R1$ and $R2$, these parallel links would be visible to $R1$ and $R2$, but not to other routers. If this bundle of parallel links is included in a tree, it is important for $R1$ and $R2$ to decide which link to use; if the $R1$ - $R2$ link is the branch for multiple trees, it is desirable to split traffic over as many link as possible. However the local link selection for a tree is irrelevant to other Routers. Therefore, the tiebreaking algorithm need not be visible to any Routers other than $R1$ and $R2$.

When there are L parallel links between $R1$ and $R2$ and they both are on K trees. L links are ordered from 0 to $L-1$ in ascending order of Circuit ID as associated with the adjacency by the router with the highest System ID, and K trees are ordered from 0 to $K-1$ in ascending order of root IP addresses. The tiebreaker rule is: for tree k , select the link as choice $k \bmod L$.

Note that if multiple distribution trees are configured in a network or on a router, better load balance among parallel links through the tie-breaking algorithm can be achieved. Otherwise, if there is only one tree is configured, then only one link in parallel links can be used for the corresponding distribution tree. However, calculating and maintaining many trees is resource consuming. Operators need to balance between two.

3.4. Tree Selection for a Group

Routers receive one or more possible multicast group-range-to-tree mappings. Each mapping specifies a range of multicast groups. It is possible that a group-range is associated with multiple trees that may have the same or different priority. When a multicast group-range associates with more than one tree, all routers have to select the same tree for the group-range. The tiebreaker rules specified in PIM [RFC4601] are used. They are:

- o Perform longest match on group-range to get a list of trees.
- o Select the tree with highest priority.
- o If only one tree with the highest priority, select the tree for the group-range.
- o If multiple trees are with the highest priority, use the PIM hash function to choose one. PIM hash function is described in section 4.1.1 in RFC4601 [RFC4601].

3.5. Pruning a Distribution Tree for a Group

Routers prune the distribution tree for each associated multicast group, i.e. eliminating branches that have no potential downstream receivers. Multi-destination packets SHOULD only be forwarded on branches that are not pruned. The assumption here is that a multicast source is also a multicast receiver but a multicast receiver may not be a multicast source.

Routers prune the trees based on the groups specified in GRADD-TLV from edge routers. Routers maintain a list of adjacency interfaces that are on the pruned tree for a multicast group. Among these interfaces, one interface may be toward the tree-root router and other are toward the egress routers.

3.6. Reverse Path Forwarding Check (RPFC)

The routing transients resulting from topology changes can cause temporary transient loops in distribution trees. If no precautions are taken, and there are fork points in such loops, it is possible for multiple copies of a packet to be forwarded. If this is a problem for a particular use, a Reverse Path Forwarding Check (RPFC) may be implemented.

In this case, the RPFC works by a router determining for each port, based on the source and destination IP address of a packet, whether the port is a port that router expects to receive such a packet. In other words, is there an edge router with reachability to the source IP address such that, starting at that router and using the tree indicated by the destination IP address, the packet would have arrived at the port in question. If so, it is further distributed. If not, it is discarded. An RPFC can be implemented at some routers and not at others.

3.7. Forwarding Using a Pruned Distribution Tree

Forwarding a multi-destination packet follows the pruned tree for the group that the packet belongs to. It is done as follows.

- o If the router receives a multi-destination packet with group IP address that does not associated with any tree, the packet **MUST** be dropped.
- o Else check if the link that the packet arrives on is one of the ports in the pruned distribution tree. If not, the packet **MUST** be dropped.
- o Else perform RPF checking (section 3.5). If it fails, the packet **SHOULD** be dropped.
- o Else the packet is forwarded onto all the adjacency interfaces in the list for the group except the interface where the packet receive.

3.8. Local Forwarding at Edge Router

Upon receiving a multi-destination packet, besides forwarding it along the pruned tree, an edge router may also need to forward the packet to the local hosts attached to it. This is referred to as local forwarding in this document.

The local group database is needed to keep track of the group membership of the router's directly attached network or host. Each entry in the local group database is a [group, network/host] pair, which indicates that the attached network has one or more hosts belonging to the multicast group. When receiving a multi-destination packet, the edge router forwards the packet to the network/host that match the [group, network/host] pair in the local group database.

The local group database is built through the operation of the IGMPv3 [RFC3376]. When an edge router becomes Designated Router on

an attached network, say N1, it starts sending periodic IGMPv3 Host Membership Queries on the network. Hosts then respond with IGMPv3 Host Membership Reports, one for each multicast group to which they belong. Upon receiving a Host Membership Report for a multicast group A, the router updates its local group database by adding/refreshing the entry [Group A, N1]. If at a later time Reports for Group A cease to be heard on the network, the entry is then deleted from the local group database. The Designated further sends the LSP message with GRADDR sub-TLV to inform other routers about the group memberships in the local group database. A router MUST ignore Host Membership Reports received on those networks where the router has not been elected Designated Router.

When the solution described in this document applies to the underlying network that transports overlay virtual networks [NVO3FRWK], A Designated Router further necessarily maintains the mapping between an overlay multicast group and a underlying multicast group, and performs packet encapsulation/descapsulation upon receiving a packet from host or the underlying network. Mapping between an overlay multicast group and a underlying multicast group can be manually configured, automatically generated by an algorithm, or dynamically informed at a Designated Router. The same edge router should be selected as the Designated Router for the overlay multicast group and underlying multicast group that are associated. The mapping method is beyond the scope of this document.

3.9. Distribution Tree across different IGP Levels

An IGP (Interior Gateway Protocol) network may be designed as a multi-area network for the scalability, faster-convergence. Multicast sources and listeners may be in the same or different areas. The former is a special case of the latter. To support multi-destination transport over multi-areas, it is necessary to build a distribution tree across areas and prune the tree based on the listener locations, i.e. interested edge routers that may reside in different areas.

For an IS-IS multi-area network, there are level1 and level2 routers as well as level1/2 (border) routers. A level1 router only has the router/topology information for its area. A level2 router has router/topology information for level2 area as well as reachability information for level1 areas. A border router participates in both level1 and level2 areas and has the router/topology information for

level2 and all directly attached level1 areas but maintains separate LSDBs for level2 and each attached level 1 area. Traffic from one

area to another area must traverse through a border router. It is possible to have more than one border router between two areas for resilience.

To build a distribution tree across mutli-areas, an operator can select a tree-root node for a set of multicast groups. The node can be in levell1 area or level2 area. All the nodes including border nodes in the area compute the distribution tree as described in section 3.1-3.4. Border routers automatically select a designated forwarder for the multicast groups associated to the tree (see below). The border router selected as designate forwarder (DF) announces itself as the tree root in the adjacent area if the S bit in the RTADDR TLV is clear. The nodes in the adjacent area will compute the distribution tree in the same way. Note that a border router may be the tree-root in the adjacent area for the multicast groups that may associate with different trees. If S bit in the RTADDR TLV is set, the rooted distribution tree is only built in the area where the root node resides.

The document specifies following additional rules for a border router that supports the multicast mechanism described here. The rules apply to the case of the distribution tree across multiple areas.

If a border router is selected as designated forwarder in adjacent area for a set of multicast groups, it should perform following:

- o It MUST track the group-memberships in its participated areas.
- o It MUST send a summary group membership of one area to the adjacent area as of an edge router.
- o It performs the pruning process in each area, respectively, based on the received group-membership LSPs from that area.
- o When receiving multicast traffic from one area, it forwards the packet along the pruned tree into the adjacent area.
- o Optionally performs reverse path forwarding check (RPFC)

If a border router is not selected as the designated forwarder for the multicast groups, the followings apply:

- o It SHOULD NOT propagate the group-membership information of one area to any other areas. It SHOULD remove the TLV before forwarding it.

- o It SHOULD NOT forward multicast group traffic to another adjacent area. It SHOULD discard such traffic.

Selecting a border router as the designated forwarder of multicast group traffic may be done manually or automatically.

4. Mobility Support

4.1. Listener moves from one edge router to another

When listener moves from one edge router, say E1, to another, say E2. E1 will detect the host left and send IGMP query for (S, G). Upon the listener join E2, if E2 has not joined (S,G), E1 should announce itself as listener to the (S,G) tree.

4.2. Source host moves from one edge router to another

Multicast Tree reaches to every edge router, so source host mobility is supported naturally. If RPFC is used on a router, the port that router expects to receive packet may change. Thus, the notification on source host moves is necessary.

5. Backward Compatibility

If a router does not support the distribution tree function described in this document, distribution tree computation MUST NOT include this router. This may result the incomplete tree. An operator can build a tunnel between two routers, which allows a single rooted tree to be built. How to build the tunnel is outside scope of this document.

6. Interworking with PIM

It may be desirable for IS-IS multicast to interwork with PIM on the same network domain or different domains. The interworking solution is for further evaluation.

7. Security Considerations

For the further study.

8. IANA Considerations

IANA is requested to assign two new sub-TLV numbers for RTADDR and RTADDRV6 as specified in Sections 2.1 and 2.2. These sub-TLVs can be used under both the Router Capability (#242) and MT Capability (#144)

TLVs. To avoid confusion, each sub-TLV should be assigned the same sub-Type number under each of these two TLVs.

9. Acknowledgements

Authors like to thank Mike McBride and Linda Dunbar for their valuable inputs.

10. References

10.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC2119, March 1997.
- [RFC3376] Cain B., etc, "Internet Group Management Protocol, Version 3", rfc4604, October 2002
- [RFC4601] Fenner, B., et al, "Protocol Independent multicast - Sparse Mode (PIM-SM): Protocol Specification", rfc4601, August 2006
- [RFC5015] Handley, M., et al, "Bidirectional Protocol Independent Multicast (BIDIR-PIM)", rfc5015, October 2007
- [RFC5120] Przygienda, T., et al, "M-ISIS: Multi Topology (MT) Routing in Intermediate System to Intermediate Systems (IS-ISs)", rfc5120, February 2008
- [RFC6325] Perlman, R., et al, "Routing Bridges (RBridges): Base Protocol Specification", RFC6325, July 2011
- [RFC7176] Eastlake 3rd, D., Senevirathne, T., Ghanwani, A., Dutt, D., and A. Banerjee, "Transparent Interconnection of Lots of Links (TRILL) Use of IS-IS", RFC 7176, May 2014.

10.2. Informative References

- [DCMCAST] McBride, M., Lui, H., "Multicast in the Data Center Overview", draft-mcbride-armd-mcast-overview, 2012
- [MCASTISS] Ghanvani, A., "Multicast Issues in Networks Using NVO3", draft-ghanwani-nvo3-mcast-issues, work in progress

[NVO3FRWK] Lasserre, M., "Framework for DC Network Virtualization",
draft-ietf-nvo3-framework, work in progress.

[VXLAN] Mahalingam, M., Dutt, D., etc, "VXLAN: A Framework for
Overlaying Virtualized Layer 2 Networks over Layer 3
Networks", draft-mahalingam-dutt-dcops-vxlan, work in
progress

Authors' Addresses

Lucy Yong
Huawei USA
5340 Legacy Drive
Plano, TX 75025 USA

Phone: 469-277-5837
Email: lucy.yong@huawei.com

Weiguo Hao
Huawei Technologies
101 Software Avenue,
Nanjing 210012
China

Phone: +86-25-56623144
Email: haoweiguo@huawei.com

Donald Eastlake
Huawei
155 Beaver Street
Milford, MA 01757 USA

Phone: +1-508-333-2270
EMail: d3e3e3@gmail.com

Andrew Qu
MediaTek
San Jose, CA 95134 USA

Email: laodulaodu@gmail.com

Jon Hudson
Brocade
130 Holger Way

San Jose, CA 95134 USA

Phone: +1-408-333-4062

Email: jon.hudson@gmail.com

