Network Working Group
Internet Draft

L. Yong
W. Hao
D. Eastlake

Category: Standard Track

Huawei

Expires: January 2014

July 8, 2013

# ISIS Protocol Extension For Building Distribution Trees
# draft-yong-isis-ext-4-distribution-tree-00

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/1id-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 January 8, 2014.

Table of Contents

## 1. Introduction

The 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 necessary to carry all types of traffic in today's DC physical networks including multi-destination traffic. It is also desirable to use 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 variant of Protocol Independent Multicast (PIM) solutions [RFC4601] [RFC5015] are designed to carry IP multicast traffic over IP networks. However the PIM solutions 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 solution, 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, which consumes network resources. Furthermore PIM solutions maintain a lot of soft-state, have intensive CPU utilization, and have additional convergence time besides IGP's under a failure condition.

Although the PIM protocol is mature and has been deployed in IP networks, applying PIM to the IP network that supports the Network Virtualization can be an extreme challenge [MCASTISS]. For example, VXLAN [VXLAN] solutions requires 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 large number of overlay VNs may exist in a DC, which PIM solutions can't scale to.

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 Router Capability message 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

router's group membership announcement. A router forwards the multi-destination traffic along the pruned tree.

In this solution, edge routers use IGMP query messages to inform the attached hosts and the hosts use IGMP report message to response with their interested multicast group(s). The edge routers announce interested multicast groups in their LSPs so they are flooded to whole network.

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 protocol for both single destination and multi-destination packet transport, which proves the protocol capability for doing both.

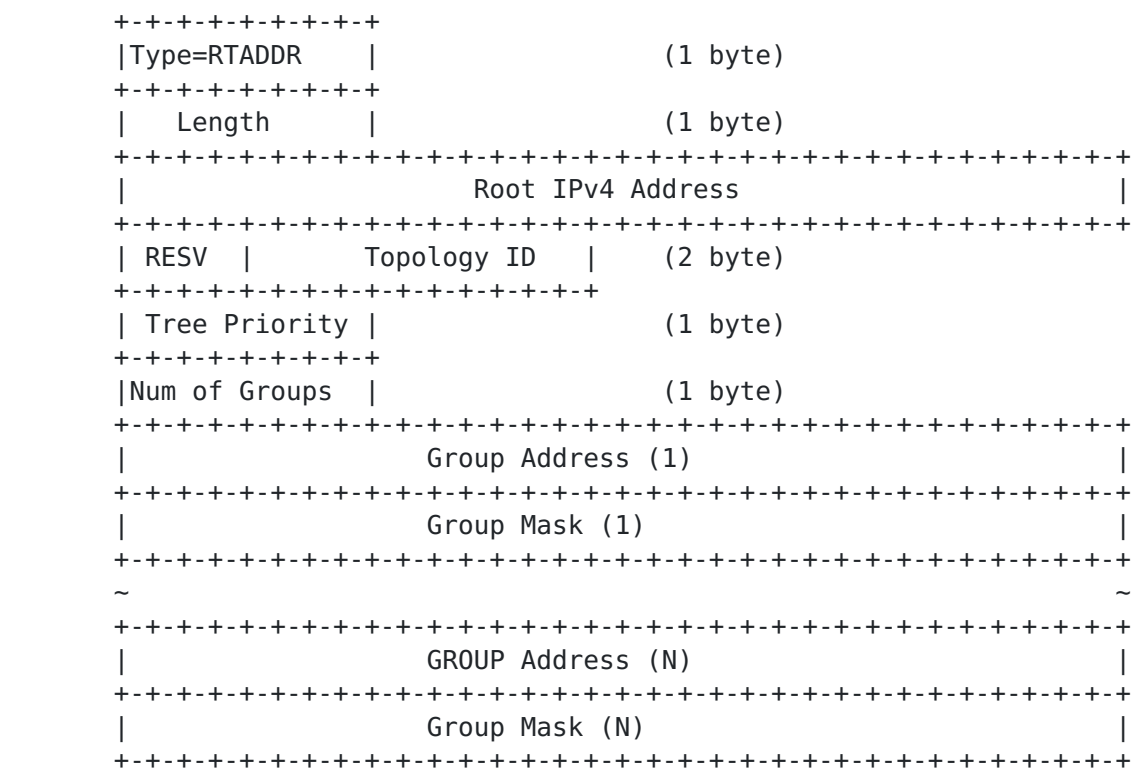
### 1.1. Conventions used in this document

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 the sub-TLV of Router Capability 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.




Where:

Type: sub-TLV of Router Capability for RTADDR (TBD)

Length: variable depending on the number of associated groups

Topology ID: This field carries a topology ID [RFC5120] or zero if topologies are not in use.

Root IP Address: IPv4 Address for a root

Tree Priority: high number means higher priority. Zero means no priority.

Num of Groups: the number of group addresses

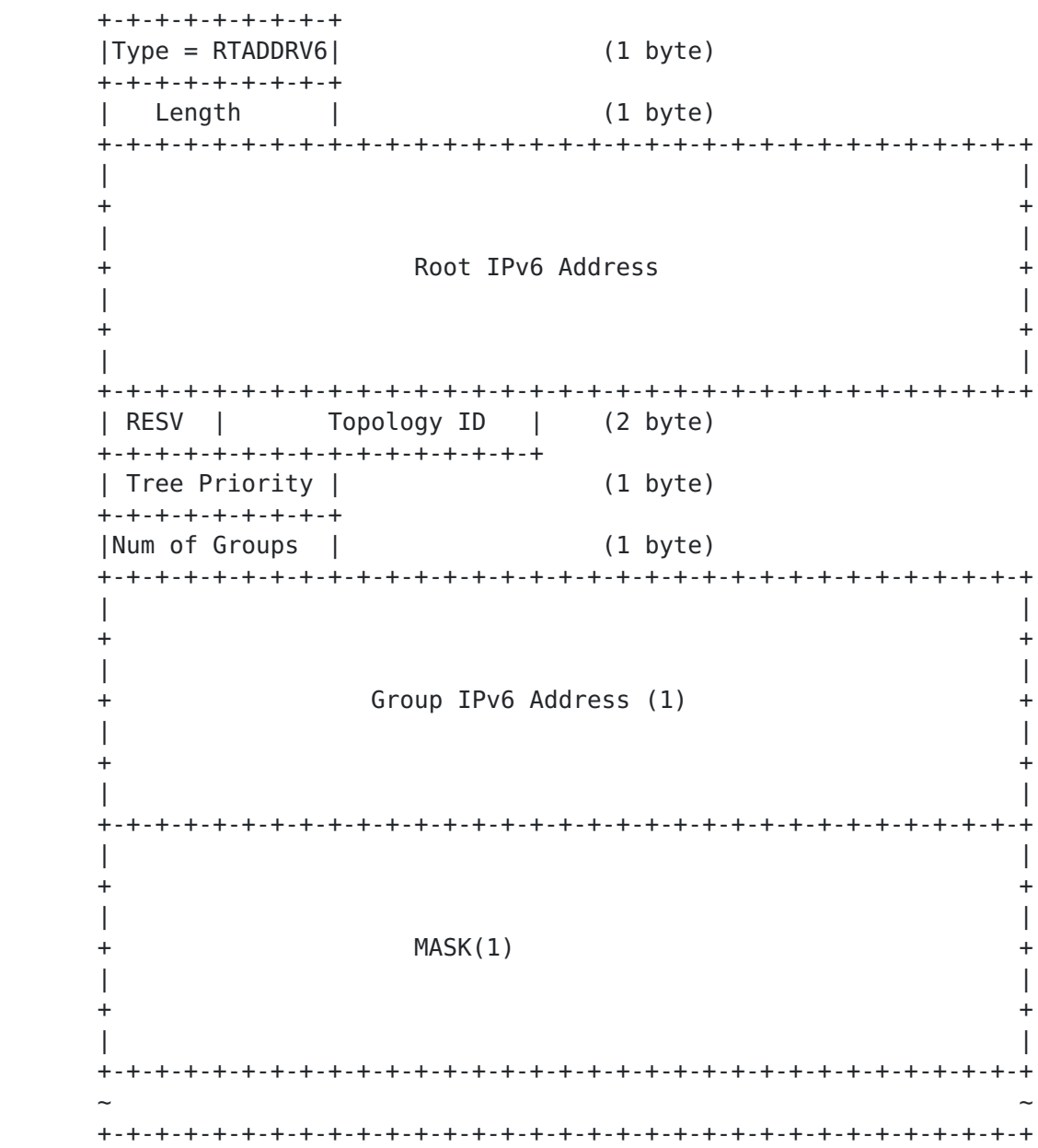
Group Address: IPv4 Address for the group

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 a router capability 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 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 RFC6326-bis [RFC6326BIS]. 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 [RFC6326-BIS]

## 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 a router capability 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 router capability TLV 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 and the processes for an IPv6 network is the same.

Operator may associate one multicast group to more than one tree for the redundancy purpose 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 these trees to choose different parents. This draft uses the same tiebreakers as TRILL [RFC6325].

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 mod p.

### 3.3. Parallel Local Link Selection

If there are parallel 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 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

C
i
r
c
u
i
t
D

I

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 mod 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 has to select the same tree for the group-range. The tiebreaker rules specified in PIM [RFC4601] are used. They are:

- Perform longest match on group-range to get a list of trees.
- Select the tree with highest priority.
- If only one tree with the highest priority, select the tree for the group-range.
- 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. RPF Mechanism

For the further study.

### 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.

- The router receives a multi-destination packet with group IP address that does not associated with any tree, the packet MUST be dropped.

- 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.

- Else perform RPF checking (section 3.5). If it fails, the packet SHOULD be dropped.

- 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 Router 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.

### 3.9. Distribution Tree across different IGP Levels

Coming soon.

## 4. 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. 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.

## 5. Security Considerations

Coming soon.

## 6. IANA Considerations

The document requires two new sub-TLVs, RTADDR and RTADDRV6 for the Router Capability TLV in IANA registry.

## 7. Acknowledgements

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

## 8. References

### 8.1. Normative References

[RFC3376] Cain B., etc, ''Internet Group Management Protocol, Version 3'', rfc4604, October 2002

[RFC4601] Fenner, B., etc, ''Protocol Independent multicast - Sparse Mode (PIM-SM): Protocol Specification'', rfc4601, August 2006

[RFC5015] Handley, M., etc, ''Bidirectional Protocol Independent Multicast (BIDIR-PIM'', rfc5015, October 2007

[RFC6325]  Perlman, R., et al, ''Routing Bridges (RBridges): Base Protocol Specification'', RFC6325, July 2011

[RFC6326]  Eastlake D, et al, ''  Transparent Interconnection of

Lots of Links (TRILL) Use of IS-IS'', RFC6326, July 2011

## 8.2. Informative References

[MCASTISS] Ghanvani, A., ''Multicast Issues in Networks Using NVO3'', draft-ghanwani-nvo3-mcast-issues-00, work in progress

[NVO3FRWK] Lasserre, M., ''Framework for DC Network Virtualization'', draft-ietf-nvo3-framework-02.txt, work in progress.

[RFC6326BIS] Eastlake, D., etc, ''Transparent Interconnection of Lots of Links (TRILL) Use of IS-IS'', draft-ietf-isis-rfc6326bis-01, 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