Network Working Group
Internet Draft

Category: Standard Track

L. Yong
W. Hao
D. Eastlake
Huawei
A. Qu
MetiaTek
J. Hudson
Brocade
U. Chunduri
Ericsson

Expires: May 2015 November 9, 2014

IGP Multicast Architecture

draft-yong-rtgwg-igp-multicast-arch-01

Abstract

This document specifies Interior Gateway Protocol (IGP) network architecture to support multicast transport. It describes the architecture components and the algorithms to automatically build a distribution tree for transporting multicast traffic and provides a method of pruning that tree for improved efficiency.

Status of this document

Table of Contents

1. Introduction....3
   1.1. Motivation....3
   1.2. Conventions used in this document....4
2. IGP Architecture for Multicast Transport....4
3. Computation Algorithms in IGP Multicast Domain....5
   3.1. Automatic Tree Root Node Selection....5
   3.2. Distribution Tree Computation....5
      3.2.1. Parent Selection....6
      3.2.2. Parallel Local Link Selection....6
   3.3. Multiple Distribute Trees for a Multicast Group....7
   3.4. Pruning a Distribution Tree for a Group....7
4. Router Forwarding Procedures....8
   4.1. Packet Forwarding Along a Pruned Distribution Tree....8
   4.2. Local Forwarding at Edge Router....8
      4.2.1. Overlay Multicast Transport....9
   4.3. Multi-homing Access Through Active-active MC-LAG....10
   4.4. Reverse Path Forwarding Check (RPFC)....11
5. Security Considerations....12
6. IANA Considerations....12
7. Acknowledgements....12
8. References....12
   8.1. Normative References....12
   8.2. Informative References....12

1. Introduction

This document specifies Interior Gateway Protocol (IGP) network architecture to support multicast transport. It describes the architecture components and the algorithms to automatically build a distribution tree for transporting multicast traffic and provides a method of pruning that tree for improved efficiency.

An IGP network is built to transport unicast traffic. Traditionally, transporting multicast traffic relies on a protocol independent mechanism and a different protocol, i.e. PIM [RFC4601] [RFC5015]. The PIM protocol builds on top of IGP network and maintains its own states, which results longer convergence time for multicast traffic

Data Center infrastructure and advanced systems for cloud applications are looking for an IGP network to transport both unicast and multicast packets in a simpler and more efficient way than use of a separate protocol beyond IGP protocol. (see Section 1.1 for motivation)

This draft proposes the architecture and algorithms for an IGP based multicast transport. The architecture and algorithms automatically build a bi-directional distribution tree and pruned bi-directional tree for a multicast group without use of PIM. IGP protocol extension for this architecture is addressed in the [ISEXT].

1.1. Motivation

Network-as-a-service technically can be achieved by decoupling network IP space from service IP space as with a VxLAN [RFC7348] based network overlay. Decoupling network IP space from service IP address space also provides network agility and programmability to applications in a Data Center environment. To support all service applications, such IP network fabric must support both unicast and multicast. If network IP space is decoupled from service IP space, the network itself no longer needs manual configuration; automatically forming an IP network fabric can be done. The resulting "plug and play" can greatly simplify network operation.

With the goal of automation in forming a network fabric and support of any type of forwarding behavior the service applications require, IGP protocol should be extended to support:

1. Network formation

2. Multi destination distribution tree computation

Using external PIM prohibits the "automatic" nature requirement and results a longer convergence time of multicast transport than unicast transport because the convergence time for PIM is added to the basic IGP unicast route convergence time.

IGP based multicast reduces the number of protocols, states, and convergence time for multicast, which means a simpler underlay IP network that supports both unicast and multicast transport.

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].

2. IGP Architecture for Multicast Transport

An IGP multicast domain defined in this document contains edge routers and transit routers. Multicast source(s) and receiver(s) in a service space locally attach to edge routers or connect to edge routers through a layer 2 or layer 3 network that belongs to the same service space. When an ingress edge router receives a multicast packet from a multicast source in the service space, it replicates it along a pruned tree in the IGP domain. When an egress edge router receives a multicast packet from the IGP domain, it forwards the packet to the L2 or L3 service network that the receivers on and replicates the packet along the pruned tree in the domain. When a transit router receives a multicast packet from another router in the domain, it replicates the packet to its neighbor router(s) in the domain along a pruned tree.

An IGP multicast domain is used to carry L2 or L3 multicast traffic in a service (tenant) space in multi-tenant environment. Upon receiving a multicast packet from a source, the edge router first encapsulates the packet, adds its IP address as the source address and the corresponding underlay multicast IP address as the destination address on the encapsulated packet, then replicates it along a pruned tree. Egress edge router(s) decapsulate the packet before sending toward the receiver(s).

In an IGP multicast domain, each router has a unique IP address and the router IP address is advertised as a host address by IGP protocol. An IGP domain can be an IGP multicast domain if all

routers support the multicast capability described in this document; a subset of an IGP domain can be an IGP multicast domain where only some edge routers and transit routers have IGP multicast capability described in this draft and the draft [ISEXT]. In the case where the IGP multicast domain is subset of an IGP domain, a router in an IGP multicast domain must have at least one adjacency (next hop) to another router that is in the IGP multicast domain, that is, the IGP multicast domain must be connected. Configuring an IP tunnel between two routers in an IGP multicast domain can achieve this. How to configure such tunnel is outside the scope of this document.

In an IGP multicast domain, a default distribution tree is established automatically (see Section of 3.1). Operators may configure other distribution trees with different priorities in the domain as well and specify the associated multicast groups carried by these configured trees. By default, all the multicast groups use the default distribution tree.

The distribution tree computation algorithm is described in Section 3.2. The tree pruning for a particular multicast group is described in Section 3.3. Section 3.4 describes multiple trees to support one multicast group. Section 4 describes router forwarding procedures.

3. Computation Algorithms in IGP Multicast Domain

3.1. Automatic Tree Root Node Selection

By default the tree root is the router with the largest magnitude Router ID, considering the Router ID, i.e. router IPv4 address, to be an unsigned integer. Note that the algorithms in following sections use Router ID for router identifier, i.e. unique IP address assigned to a router in a IGP multicast domain.

Operators may configure a default tree root node (based on the topology) that takes precedence over the default tree root auto-calculated. This configured tree root node would advertise its IP address as the default tree root for all multicast groups that are not assigned to a distribution tree in a IGP multicast domain.

3.2. Distribution Tree Computation

The Distribution Tree Computation Algorithm uses the existing IGP Link State Database (LSDB). Based on the LSDB and shortest path algorithm, all routers in an IGP multicast domain calculate the distribution tree that has the default tree root node and reaches all the edge routers.

If an operator configures other distribution tree roots on other routers, the operator specifies what multicast groups use those trees and the tree root routers will advertise themselves as the tree root for those multicast groups by use of the new RTADDR TLV [ISEXT]. All routers in the domain will track the tree root nodes and calculate the path toward to each configured tree root node by using the shortest path algorithm, which form multiple distribute trees.

It is important that all the routers calculate the identical branches in a distribution tree in an IGP multicast domain. Section 3.2.1 and 3.3.2 specifies the tiebreaking rules for parent router selection in case of equal-cost path and for the link selection in case of multiple local links. Because link costs can be asymmetric, it is important for all tree construction calculations to use the cost towards the root.

3.2.1. Parent Selection

When there are equal costs from a potential child router to more than one possible parent router, all routers need to use the same tiebreakers. It is desirable to allow splitting traffic on as many links as possible in such situations when multiple distribution trees presents. This document uses the following tiebreaker rules:

If there are k distribution trees in the domain, when each router computes these trees, the k trees calculated are ordered and numbered from 0 to k-1 in ascending order according by 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 System ID considered as an unsigned integer, and number them starting at zero. For tree j, choose N’s parent as choice (j-1) mod p.

3.2.2. 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 links 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 mod L.

Note that if multiple distribution trees are configured in a domain 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.

Another alternative is to use a lower level link aggregation protocol, such as [802.1AX-2011] on the parallel point-to-point links between R1 and R2. They will then appear to be a single link to the IGP and it will be the link aggregation protocol that spreads traffic across the actual lower level parallel links.

3.3. Multiple Distribute Trees for a Multicast Group

It is possible that a multicast group is associated with multiple trees that may have the same or different priority. When a multicast group associates with more than one tree, all routers have to select the same tree for the group. The tiebreaker rules specified in PIM [RFC4601] are used here. 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 RFC 4601 [RFC4601].

3.4. 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.

All routers in the domain receive LSP messages with GRADD-TLV [RFC7176] from the edge routers, which indicate which multicast group that an edge router is the receiver. According that, the routers prune the corresponding distribution tree for each multicast group and maintain a list of adjacency interfaces that are on the pruned tree for a multicast group. Among these interfaces, one interface will be toward the tree-root router (unless the router is the root) and zero or more interfaces will be toward some edge routers.

4. Router Forwarding Procedures

4.1. Packet Forwarding Along 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.

- If the router receives a multi-destination packet with group IP address that does not associated with any configured tree, the packet MUST be considered associated with the default tree.

- 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 optionally perform RPF checking (section 4.4). If the check is performed and it fails, the packet SHOULD be dropped.

- Else the packet is forwarded onto all the adjacency interfaces in the pruned tree for the group except the interface where the packet receive.

4.2. Local Forwarding at Edge Router

Upon receiving a multicast 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. Local forwarding table and multicast forwarding table in IGP domain should be stitched at each edge router. Local forwarding table can be generated using IGMP/PIM protocol running in the network between host and the edge router.

The local group database is needed to keep track of the group membership of attached hosts. Each entry in the local group database is a [group, host] pair, which indicates that the attached hosts belonging to the multicast group. When receiving a multicast packet, the edge router forwards the packet to the host that match the [group, host] pair in the local group database.

The local group database is built through the operation of the IGMPv3 [RFC3376]. An edge router sends periodic IGMPv3 Host Membership Queries to attached hosts. 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, host] pair. If at a later time Reports for Group A cease to be heard from the host, the entry is then deleted from the local group database. The edge router further sends the LSP message with GRADDR TLV to inform other routers about the group memberships in the local group database.

4.2.1. Overlay Multicast Transport

An IGP multicast domain may be used to carry overlay multicast traffic. [RFC7365] There are two architecture scenarios:

1) IGP multicast domain edge router separates with overlay network edge device [RFC7365]. Before multicast traffic is forwarded, Overlay network should trigger underlay multicast domain to construct multicast tree using IGMP protocol in beforehand. Group address in the protocol is underlay multicast group address. Outer layer traffic encapsulation is performed on the overlay network edge device, IGP multicast domain acts as pure underlay network.

2) IGP multicast domain edge router collapses with overlay network edge device. Before multicast traffic is forwarded, local connecting host should trigger underlay multicast domain to construct multicast tree using IGMP like protocol beforehand. Group address in the protocol is overlay multicast group address, edge router should map the group address into underlay multicast group address.

The IGP multicast domain can support both scenarios. To carry overlay multicast traffic, a (designated) edge router (see Section below on Multi-Homing Access) 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 a host or the underlay IGP network. Mapping between an overlay multicast group and a underlay multicast group can be manually configured, automatically generated by an

algorithm at a (designated) edge router. The same edge router MUST be selected as the Designated Forwarder for the overlay multicast group and underlying multicast group that are associated. If multiple overlay multicast groups attach to same edge router sets, these overlay multicast groups can be mapped to the same underlying multicast group to reduce underlay network multicast forwarding table size on each router. The mapping method is beyond the scope of this document.

4.3. Multi-homing Access Through Active-active MC-LAG

A multicast group receiver may attach to multiple edge routers through an active-active MC-LAG [802.1AX-2011] to enhance reliability.

When a remote edge router ingresses a multicast packet w/ multicast group address from local multicast source, if all egress routers in an MC-LAG forward the packet to the local host (receiver), the host will receive multiple copies of the multicast frame from the remote multicast source. To avoid duplicated packets received from the IGP domain to a local network, a Designated Forwarder (DF) mechanism is required. All the edge routers associated to a same MC-LAG use the same algorithm to select one DF edge router for a multicast group. Each MC-LAG should be assigned with a unique MC-LAG identifier in an IGP multicast domain, which may be manually configured or automatically provisioned. When an edge router in a MC-LAG receives a multicast group receiver joining message using IGMP/PIM like protocols, it announces its self MC-LAG ID and the multicast group correspondence to other routers in its IGP LSP. After network state reaches steady state, all edge routers in a MC-LAG elect the same router as DF for each multicast group. Upon receiving a multicast packet from the domain, only the DF edge router will forward the packet towards the receiver. All non-DF edge routers do not forward the packet towards the receiver.

All edge routers, including DF and non-DF, can ingress the traffic to IGP domain as usual. DF and non-DF state has influence only on the egress multicast traffic forwarding process.

If a multicast group source host attaches to multiple edge routers through an active-active MC-LAG, loop prevention, i.e. the packet sent by source host loops back to the source host via the edge routers in a MC-LAG, is necessary. The solutions for two scenarios are described below.

- When the multicast IGP domain edge routers separate with overlay network edge devices that carry overlay network traffic, these routers don't replace traffic source IP address when they inject the traffic into IGP domain. In this case, edge routers should acquire multicast source IP address in beforehand using a mechanism like IGMPv3 explicit tracking, and then the source IP addresses are synchronized among each edge routers in same MC-LAG. Then same split-horizon mechanism described in the above section can be used.

- When the multicast IGP domain edge routers collapse with overlay network edge devices, the edge router connecting to multicast source performs multicast encapsulation when it injects local multicast traffic into the IGP domain, source IP is the edge router's IP. Each edge router tracks the IP address(es) associated with the other edge router(s) with which it has shared MC-LAG. When the edge router receives a packet from an IGP domain, it examines the source IP address and filters out the packet on all local interfaces in the same MC-LAG. With this approach, local bias forwarding is required on the ingress edge router. It performs replication locally to all directly attached receivers no matter DF or non-DF state of the out interface connecting to each receiver.

4.4. 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 multicast 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.

5. Security Considerations

To come in future version

6. IANA Considerations

This document does not request any IANA action.

7. Acknowledgements

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

8. References

8.1. 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

[ISEXT] Yong, L., el al, "IS-IS Extension For Building Distribution Tree", draft-yong-isis-ext-4-distribution-tree, work in progress.

[802.1AX-2011] IEEE, "IEEE Standard for Local and metropolitan area networks - Link Aggregation", IEEE802.1AX, 2011

8.2. Informative References

[RFC7348] Mahalingam, M., Dutt, D., etc, "VXLAN: A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks", RFC7348, 2014

[RFC7365] Lasserre, M., "Framework for DC Network Virtualization", RFC7364, 2014.

Authors' Addresses

Lucy Yong
Huawei USA

Phone: 918-808-1918
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

Uma Chunduri

Ericsson Inc.
300 Holger Way,
San Jose, California 95134
USA

Phone: 408-750-5678
Email: uma.chunduri@ericsson.com