

Network Working Group
Internet-Draft
Intended status: Informational
Expires: May 4, 2016

F. Coras
A. Cabellos-Aparicio
J. Domingo-Pascual
Technical University of Catalonia
F. Maino
cisco Systems
D. Farinacci
lisppers.net
November 1, 2015

LISP Replication Engineering
draft-coras-lisp-re-08

Abstract

This document describes a method to build and optimize inter-domain LISP router distribution trees for locator-based unicast and multicast replication of EID-sourced multicast packets.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <http://datatracker.ietf.org/drafts/current/>.

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."

This Internet-Draft will expire on May 4, 2016.

Copyright Notice

Copyright (c) 2015 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. Code Components extracted from this document must

include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	2
2. Definition of Terms	3
3. Overview	5
4. Overlay Signaling	5
4.1. RTR Registration	6
4.2. ETR/RTR Subscription	6
4.3. ETR/RTR Unsubscription	7
5. Overlay Management	8
5.1. RLOC Failure and Unreachability	8
5.2. Other Overlay Management Considerations	8
5.3. Automated Computation of RTR Level	9
5.3.1. Algorithm for Computing Optimized Distribution Trees	9
6. Security Considerations	10
7. IANA Considerations	10
8. Acknowledgements	11
9. References	11
9.1. Normative References	11
9.2. Informative References	12
Appendix A. MADDBST heuristic	12
Authors' Addresses	12

1. Introduction

The Locator/Identifier Separation Protocol (LISP) [RFC6830] provides the mechanisms for the separation of Location and Identity semantics presently overloaded by IP addresses. The split results in the creation of two namespaces that unambiguously identify edge-site network objects, Endpoint IDs (EIDs), and core routing objects, Routing LOCators (RLOCs). Apart from aiding the scalability of the core routing infrastructure, the decoupling also enables the (re)implementation of new or existing inter-domain routing mechanisms.

One such mechanism is inter-domain IP source-specific multicast (SSM) [RFC4607]. In this sense, [RFC6831] defines the procedures carried out for delivering multicast packets from a source host in a LISP site to receivers residing in the same domain or in other LISP or non-LISP sites when an underlying multicast infrastructure exists. The signaling protocol it specifies for conveying (S-EID,G) state and building the distribution tree that connects the source ITR and the receiving ETRs is PIM [RFC4601]. An alternative method that uses

Map-Requests for propagating (S-EID,G) state from ETRs to the ITR is established in [I-D.farinacci-lisp-mr-signaling].

Although desirable to use multicast routing in the core network when available, a mismatch between the multicast capabilities of receiver ETRs and source ITR might impede their interconnection. In such a case, unicast RLOC encapsulation is necessary to deliver multicast packets directly to the ETRs. This however leads to high ITR head-end replication for large sets of ETRs. Therefore, to reduce the replication load of the ITR and scale the service with the number of multicast receivers, the ITR may choose to offload replication to a set of RTRs.

The current document describes how multicast RTRs can be used to build an inter-domain distribution tree rooted at the ITR that can perform unicast and/or multicast encapsulated replication of multicast packets. This concept, of distributing the replication load from ITR to other RTRs downstream on the core overlay distribution tree, is known as Replication Engineering or LISP-RE. Since unicast replication in such overlays can be suboptimal when compared to the underlay network, methods to optimize packet delivery over the distribution tree are also presented.

This specification does not define the mechanisms used to build (S-EID,G) state in source and receiver domains, nor does it describe the messages used to propagate such state from receiver ETRs to source ITR. What it defines is how (S-EID,G) state is built in the ITR, RTRs and ETRs participating in the overlay distribution tree.

2. Definition of Terms

The terminology in this document is consistent with the definitions in [RFC6830] and [RFC6831] however, it is extended to account for LISP-RE concepts:

Delivery Group (DG): This is the outer destination address of a packet when LISP encapsulating a multicast packet with an EID source within a multicast packet.

Re-encapsulating Tunnel Router (RTR): An RTR is a router that implements the re-encapsulating tunnel function detailed in Section 8 of the main LISP specification [RFC6830]. Such router performs packet re-routing by chaining ETR and ITR functions, whereby they first remove the LISP header of ingressing packets and then prepend a new one prior to forwarding them.

Unicast Replication: Is the notion of replicating a multicast packet with an EID source address at an ITR or RTR by encapsulating it

into a unicast packet. That is, the oif-list of a multicast map-cache entry can not only have interfaces present for link-layer replication and multicast encapsulation but also for site-facing interfaces and unicast encapsulation.

Overlay Distribution Tree: A degree-constrained spanning tree that represents the path followed by unicast and/or multicast encapsulated multicast packets from the root (ITR) to the leaves (ETRs) through intermediary nodes (RTRs). The ITR and RTRs unicast and/or multicast replicate packets to their tree children.

LISP Replication Node: A router (either the ITR or an RTR) participating and replicating packets downstream in the distribution tree.

Multicast Ingress Tunnel Router (ITR): An ITR as specified in [RFC6830] that supports multicast and participates as the root in the distribution tree. In this document we use the term "ITR" to mean a multicast capable ITR.

Multicast Egress Tunnel Router (ETR): An ETR as specified in [RFC6830] that participates as a leaf in the distribution tree. ETR are the only members of the tree that do not unicast replicate. In this document we use the term "ETR" to mean a multicast capable ETR.

Multicast Re-encapsulating Tunnel Router (RTR): An RTR as specified in [I-D.farinacci-lisp-te] that participates as an intermediary node in the distribution tree. In this document we use the term "RTR" to mean a multicast capable RTR.

Replication Target: A multicast channel-id (S-EID,G) or a set of multicast channel-ids (S-EID-prefix,G).

Joining-OIF-list: Represents a collection of state per multicast routing table entry at an RTR or ETR that is created by received Map-Request/Join-Request.

Forwarding-OIF-list: Represents the outgoing RLOC list a multicast router stores per multicast routing table entry such that it knows to which RLOCs to replicate multicast packets. Although the Joining-OIF-list contains sufficient information to allow the forwarding of encapsulated multicast packets, using it is inefficient. Thereby, an RTR implementation may wish to build an efficient Forwarding-OIF-list. Ways of implementing a Forwarding-OIF-list are out of the scope of this document.

Upstream: Towards the root of the tree.

Downstream: Away from the root of the tree.

3. Overview

This document describes a method to diminish the ITR's replication load by using RTRs to build an inter-domain distribution tree. The tree is managed by the source domain's Map-Server. RTRs join the overlay due to either manual or automatic configuration and advertise to the Map-Server their availability to replicate traffic for a multicast channel (S-EID,G). Out of all the RTRs registering for the same multicast channel, the Map-Server builds one mapping and organizes the RLOCs in a multi-level hierarchy. The hierarchy is rooted at the ITR and computed based on the configured information RTRs register or by means of local policy and algorithms. ETRs always join the overlay as leaves and their attachment prompts the creation of a path, which traverses the RTR hierarchy, from the ITR. The path is built at receiver request by incrementally linking all distribution tree levels, starting at the joining ETR up to the source ITR.

The way the distribution tree is built has several benefits. First, it ensures that packets in the source domain do not reach the ITR if no ETR is joined. Second, it ensures that packets are forwarded from ITR to all ETRs without mapping database lookups since the state that defines the distribution tree, i.e., the replication hierarchy, is created prior to forwarding/replicating the packets. Finally, the multicast source is allowed to roam since a first level RTR, when informed of the roam event, can do a new database lookup to find the new ITR to join to.

It is worth pointing out that because of the receiver-initiated approach multicast employs to build distribution trees, whereby receivers join upstream sources, LISP-RE operates backwards from LISP point of view. That is, ETRs are the ones to send Map-Requests to discover potential upstream parents and the ITR answers with Map-Replies to joining downstream clients.

4. Overlay Signaling

This section describes the signaling the ITR, RTRs and ETRs use in order to participate in the overlay and build a distribution tree. The signaling messages used are described in [I-D.farinacci-lisp-mr-signaling] and [RFC6831].

4.1. RTR Registration

RTR participation in the overlay is condition by the configuration of a replication target, a multicast channel (S-EID,G) or set of channels (S-EID-prefix,G), the RTR is to perform replication for. Once configured, manually or by automated mechanisms, an RTR Map-Registers its replication target with merge-semantics to the appropriate Map-Server. In the registration it also provides its list of RLOCs to be used by overlay peers and a set of corresponding weights and priorities. If present, information about the level of the hierarchy where the RTR should attach is also conveyed by means of an Replication List Entry canonical address [I-D.ietf-lisp-lcaf].

Due to the merge-semantics, the Map-Server aggregates all RTR originated Map-Register messages in a single, per replication target mapping. If no level information is provided or if so configured, the Map-Server should use local policy to compute a hierarchy and associate a level within it to each entry in the list (more details in Section 5.3). It should be noted that the entries that are pointed to in the resulting mapping are not RLOCs but Replication List entries.

4.2. ETR/RTR Subscription

When an ETR creates (S-EID,G) state from a site based multicast join, i.e., its oif-list goes non-empty, it must send an upstream Join request. If the ETR does not have multicast connectivity to its upstream and unicast replication must be performed, the ETR requests that a path from ITR to itself, over the RTR hierarchy be constructed. The following procedure is followed to build the path:

1. ETR sends a Map-Request/Join-Request for (S-EID,G) multicast channel to the mapping database system which further ensures its delivery to the authoritative Map-Server.
2. The Map-Server looks up the mapping associated to (S-EID,G) and, out of the distribution tree hierarchy encoded within, it selects a set of leaf RTRs, i.e., members of the level furthest away from the ITR, with spare replication capacity. The set of potential parents is encoded in a new (S-EID, G) mapping the Map-Server conveys to the ETR in a Map-Reply.
3. From the list it receives, the ETR selects the best upstream RTR RLOC according to local policy, taking into account the associated priorities and weights and sends to the owning RTR a Map-Request/Join-Request for (S-EID,G). If the ETR itself has multiple RLOCs it wishes to use in the overlay, it may convey them all to the upstream RTR encoded in the Map-Reply field of

the Map-Request/Join-Request together with associated priorities and weights.

4. The RTR stores the ETR's subscription information in the join-oif-list associated to (S-EID,G) and inserts the RLOC obtained after evaluating the priorities and weights in the oif-list for (S-EID,G). It then confirms the ETR's subscription with a Map-Reply.
5. If not already a member of (S-EID,G), the RTR initiates it's own attachment to the distribution tree by repeating the steps 1-4. An important difference at step 2, the Map-Server replies to a joining RTR with a list of RTRs in the adjacent upstream layer, as opposed to a list of leaf RTRs, like in the case of an ETR join. This procedure may recurse upstream up to when the ITR or an RTR already attached to the distribution tree is joined. On completion, there should exist a path from ITR to joining ETR.
6. If the ITR is already member of (S-EID,G) the process stops. Otherwise, the ITR sends a PIM join to the intra-domain multicast source ensuring the creation of a path from the multicast source to the receiver end-hosts.

If at any point, when creating a link between two adjacent layers, native multicast replication can be performed, instead of unicast replication, the router joining its upstream could set as source of the Map-Request/Join-Request a delivery group. However, group naming must be coordinated between the participating parties in this case, if core network replication is to be exploited.

4.3. ETR/RTR Unsubscription

When an ETR's oif-list goes empty a Map-Request/Leave-Request is sent to the upstream RTR which will result in the removal of the ETR's associated entry from the RTR's oif-list. The procedure is repeated by the RTR, and it may recurse upstream, if its own oif-list also goes empty.

When an RTR with active downstreams departs, it should first change the priority of the RLOCs it registers with the Map-Server to 255 and set its locators as unreachable in the RLOC-Probing replies it sends downstream. Finally, once all adjacent lower level members have sent Map-Request/Leave-Request messages the RTR can stop registering (S-EID,G) with the mapping database system and thus leave the overlay.

5. Overlay Management

5.1. RLOC Failure and Unreachability

RLOC failure is detected at control-plane level through RLOC-probing [RFC6830] by both upstream and downstream routers. When an RTR detects the failure of an downstream RLOC, it ceases replicating towards it. The affected RLOC is removed from the forwarding-oif-list and marked as unreachable in the join-oif-list. If a backup RLOC was provided by the downstream router in the Map-Request/Join-Request, it is instead inserted in the forwarding-oif-list and the failure results in no packet loss.

The routers downstream of a failed RTR RLOC, or who lost connectivity to said RLOCs, remove their Map-Request/Join-Request associated state and reperform the join procedure. Packet loss in this case must be solved by out-of-band mechanisms that are out of the scope of the current document.

5.2. Other Overlay Management Considerations

An overloaded RTR, i.e., one whose fan-out can not be increased, should change the priority of the RLOCs it registers with the mapping database system to 255. In such a situation, the Map-Server updates the associated mapping and informs all routers having requested it about the change through solicit Map Request (SMR) messages. Both new ETRs attaching to the distribution tree and those already connected but reperforming the join procedure must not use the RLOCs with a priority of 255 as specified in [RFC6830]. However, routers having performed Join-Requests prior to the change should not break their existing connections to the affected RTR.

All routers part of an (S-EID,G) multicast channel should re-evaluate their attachment point to the distribution tree whenever the Map-Server updates the associated mapping. This ensures the overlay member routers attach to the best suited parent when new RTRs join or previously attached ones stop being overloaded. Change of a parent should be done following a "make before break" procedure. Specifically, the router changing attachment point first connects to the new parent and only afterwards sends the Leave-Request.

When a downstream RTR subscribes to a set of channel-ids (S-EID-prefix,G) using multiple RLOCs in a load-balancing configuration, the upstream RTR may choose to load-split channel-ids (S-EID,G) over the given set of RLOCs.

5.3. Automated Computation of RTR Level

Operators wishing to automate the RTR joining procedure may wish to use an algorithm for computing an optimized distribution tree. The algorithm could be implemented in the Map-Server and its output should be used to associate to all RTRs a level in the distribution tree. Due to the centralized management, on-line switching between algorithms may be possible in accordance to the required distribution tree performance. However, their use of such algorithms is dependent on the presence of overlay topological information. Ways of obtaining topological information will be discussed in future versions of this document.

5.3.1. Algorithm for Computing Optimized Distribution Trees

The current document does not recommend an algorithm for computing optimized distribution trees. However, it provides as an example a low computation cost heuristic, which, in the scenarios simulated in [LCAST], can produce latencies between the ITR and the multicast receivers close to unicast ones. Its choice is to be influenced by operational requirements and the hardware constraints of the equipment in charge of running it. Future experiments might result in a recommendation.

In what follows, we use the term "distance" when referring to a relative length or amplitude of a metric, observed on a path connecting two points, but when the exact nature of the metric is of no interest.

Considering as goal the delivery of content for delay sensitive applications, the function the algorithm minimizes is the maximum distance (e.g. latency or number of AS hops) from a multicast receiver to the ITR source. Notice that the reference is the multicast receiver host and not an ETR. Thus, what matters in deciding a member's position in the distribution tree is not solely its distance to the ITR but also the number of multicast receivers it serves. Then, a router close to the source but serving few receivers might find itself lower in the distribution tree than another with a slightly higher distance to the source but with a larger receiver set. The algorithm optimizes the quality of experience for multicast receivers and not for tunnel routers.

The problem described above, that searches for a minimum average distance, degree-bounded spanning tree (MADDBST), can be formally stated as:

Definition: Given an undirected complete graph $G=(V,E)$, a designated vertex r belonging to V , for all vertices v in V , a degree bound

$d(v) \leq d_{\max}$, d_{\max} a positive integer, a vertex weight function $c(v)$ with positive integer values, and an edge weight function $w(e)$ with positive values, for all edges e in E . Let $W(r,v,T)$ represent the cost of the path linking r and v in the spanning tree T . Find the spanning tree T of G , routed at r , satisfying that $d(v) \leq d_{\max}$ and the distance to the source per multicast receiver is minimized.

The heuristic used to solve this problem works by incrementally growing a tree, starting at the root node r , until it becomes a spanning tree. For each node v , not yet a tree member, it selects a potential parent node u in the tree T , such that the distance per receiver to r , is minimized. At each step, the node with the smallest metric value is added to the tree and the parent selection is redone. The pseudocode of the heuristic is provided in Appendix A.

[SHI] and [BAN] have previously defined and solved similar optimization problems. Shi et al. [SHI] also prove that a particular instance of the problem, where all vertices have weight 1, is NP-complete for degree constraints $2 \leq d_{\max} \leq |V|-1$.

The algorithm can optimize an unicast overlay however, it should not be used to optimize multicast underlay delivery. As a result, if multicast is used as underlay between part of the overlay members, once one of the members of such Delivery Group is added to the distribution tree, the others should be marked as attached also. These nodes should receive multicast encapsulated multicast packets from the chosen node over the underlying multicast distribution tree.

Finally, since the RTRs do not replicate packets for multicast receiver hosts, prior to applying the MADDBST heuristic, a Minimum Spanning Tree (MST) algorithm should be used to compute the RTR distribution tree. In this case, the MADDBST heuristic should start attaching ETRs having as input the tree resulting from MST.

6. Security Considerations

Security concerns for LISP-RE the same as for [RFC6831] and [I-D.farinacci-lisp-mr-signaling].

7. IANA Considerations

This memo includes no request to IANA.

8. Acknowledgements

The authors would like to thank Noel Chiappa for his technical and editorial commentary.

9. References

9.1. Normative References

- [I-D.farinacci-lisp-mr-signaling]
Farinacci, D. and M. Napierala, "LISP Control-Plane Multicast Signaling", draft-farinacci-lisp-mr-signaling-06 (work in progress), February 2015.
- [I-D.farinacci-lisp-te]
Farinacci, D., Kowal, M., and P. Lahiri, "LISP Traffic Engineering Use-Cases", draft-farinacci-lisp-te-09 (work in progress), September 2015.
- [I-D.ietf-lisp-lcaf]
Farinacci, D., Meyer, D., and J. Snijders, "LISP Canonical Address Format (LCAF)", draft-ietf-lisp-lcaf-11 (work in progress), September 2015.
- [RFC4601] Fenner, B., Handley, M., Holbrook, H., and I. Kouvelas, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", RFC 4601, DOI 10.17487/RFC4601, August 2006, <<http://www.rfc-editor.org/info/rfc4601>>.
- [RFC4607] Holbrook, H. and B. Cain, "Source-Specific Multicast for IP", RFC 4607, DOI 10.17487/RFC4607, August 2006, <<http://www.rfc-editor.org/info/rfc4607>>.
- [RFC6830] Farinacci, D., Fuller, V., Meyer, D., and D. Lewis, "The Locator/ID Separation Protocol (LISP)", RFC 6830, DOI 10.17487/RFC6830, January 2013, <<http://www.rfc-editor.org/info/rfc6830>>.
- [RFC6831] Farinacci, D., Meyer, D., Zwiebel, J., and S. Venaas, "The Locator/ID Separation Protocol (LISP) for Multicast Environments", RFC 6831, DOI 10.17487/RFC6831, January 2013, <<http://www.rfc-editor.org/info/rfc6831>>.

9.2. Informative References

- [BAN] Banerjee, S., Kommareddy, C., Kar, K., Bhattacharjee, B., and S. Khuller, "Construction of an efficient overlay multicast infrastructure for real-time applications", INFOCOM , 2002.
- [LCAST] Coras, F., Cabellos, A., Domingo, J., Maino, F., and D. Farinacci, "Lcast: Software-defined inter-domain multicast", Computer Networks , 2014.
- [SHI] Shi, S., Turner, J., and M. Waldvogel, "Dimensioning server access bandwidth and multicast routing in overlay networks", NOSSDAV , 2001.

Appendix A. MADDBST heuristic

```

INPUT: G = (V,E); r; dmax; w(u,v); c(v); u, v in V
OUTPUT: T

  FOREACH v in V DO
    delta(v) = w(r,v)/c(v);
    p(v) = r;
  END FOREACH

  T takes (U = {r}, D={});
  WHILE U != V DO
    LET u in U-V be the vertex with the smallest delta(u);
    U = U U {u}; L = L U {(p(u),u)};
    FOREACH v in V-U DO
      delta(v) = infinity;
      FOREACH u in U DO
        IF d(u) < dmax and
           W{r,u,T} + w(u,v)/c(v) < delta(v) THEN
          delta(v) = W{r,u,T} + w(u,v)/c(v);
          p(v) = u;
        END IF
      END FOR
    END FOR
  END WHILE

```

Figure 1

Authors' Addresses

Florin Coras
Technical University of Catalonia
C/Jordi Girona, s/n
BARCELONA 08034
Spain

Email: fcoras@ac.upc.edu

Albert Cabellos-Aparicio
Technical University of Catalonia
C/Jordi Girona, s/n
BARCELONA 08034
Spain

Email: acabello@ac.upc.edu

Jordi Domingo-Pascual
Technical University of Catalonia
C/Jordi Girona, s/n
BARCELONA 08034
Spain

Email: jordi.domingo@ac.upc.edu

Fabio Maino
cisco Systems
Tasman Drive
San Jose, CA 95134
USA

Email: fmaino@cisco.com

Dino Farinacci
lispers.net

Email: farinacci@gmail.com

MBONED Working Group
Internet Draft
Intended status: BCP
Expires: April 27, 2015

Percy S. Tarapore
Robert Sayko
AT&T
Greg Shepherd
Toerless Eckert
Cisco
Ram Krishnan
Brocade
October 27, 2014

Multicasting Applications Across Inter-Domain Peering Points
draft-tarapore-mboned-multicast-cdni-07.txt

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <http://datatracker.ietf.org/drafts/current/>.

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."

This Internet-Draft will expire on April 27, 2015.

Copyright Notice

Copyright (c) 2014 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. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

This document may contain material from IETF Documents or IETF Contributions published or made publicly available before November 10, 2008. The person(s) controlling the copyright in some of this material may not have granted the IETF Trust the right to allow modifications of such material outside the IETF Standards Process. Without obtaining an adequate license from the person(s) controlling the copyright in such materials, this document may not be modified outside the IETF Standards Process, and derivative works of it may not be created outside the IETF Standards Process, except to format it for publication as an RFC or to translate it into languages other than English.

Abstract

This document examines the process of transporting applications via multicast across inter-domain peering points. The objective is to describe the setup process for multicast-based delivery across administrative domains and document supporting functionality to enable this process.

Table of Contents

1. Introduction.....	3
2. Overview of Inter-domain Multicast Application Transport.....	4
3. Inter-domain Peering Point Requirements for Multicast.....	5
3.1. Native Multicast.....	5
3.2. Peering Point Enabled with GRE Tunnel.....	7
3.3. Peering Point Enabled with an AMT - Both Domains Multicast Enabled.....	8
3.4. Peering Point Enabled with an AMT - AD-2 Not Multicast Enabled.....	9
3.5. AD-2 Not Multicast Enabled - Multiple AMT Tunnels Through AD-2.....	11
4. Supporting Functionality.....	13
4.1. Network Interconnection Transport and Security Guidelines	14
4.2. Routing Aspects and Related Guidelines.....	15
4.2.1 Native Multicast Routing Aspects.....	15
4.2.2 GRE Tunnel over Interconnecting Peering Point.....	16
4.2.3 Routing Aspects with AMT Tunnels.....	16
4.3. Back Office Functions - Billing and Logging Guidelines...	19
4.3.1 Provisioning Guidelines.....	19
4.3.2 Application Accounting Billing Guidelines.....	20
4.3.3 Log Management Guidelines.....	21
4.3.4 Settlement Guidelines.....	21
4.4. Operations - Service Performance and Monitoring Guidelines	22
4.5. Client Reliability Models/Service Assurance Guidelines...	24

5. Security Considerations.....	25
6. IANA Considerations.....	25
7. Conclusions.....	25
8. References.....	26
8.1. Normative References.....	26
8.2. Informative References.....	26
9. Acknowledgments.....	26

1. Introduction

Several types of applications (e.g., live video streaming, software downloads) are well suited for delivery via multicast means. The use of multicast for delivering such applications offers significant savings for utilization of resources in any given administrative domain. End user demand for such applications is growing. Often, this requires transporting such applications across administrative domains via inter-domain peering points.

The objective of this Best Current Practices document is twofold:

- o Describe the process and establish guidelines for setting up multicast-based delivery of applications across inter-domain peering points, and
- o Catalog all required information exchange between the administrative domains to support multicast-based delivery.

While there are several multicast protocols available for use, this BCP will focus the discussion to those that are applicable and recommended for the peering requirements of today's service model, including:

- o Protocol Independent Multicast - Source Specific Multicast (PIM-SSM) [RFC4607]
- o Internet Group Management Protocol (IGMP) v3 [RFC4604]
- o Multicast Listener Discovery (MLD) [RFC4604]

This BCP is independent of the choice of multicast protocol; it focuses solely on the implications for the inter-domain peering points.

This document therefore serves the purpose of a "Gap Analysis" exercise for this process. The rectification of any gaps identified - whether they involve protocol extension development or otherwise - is beyond the scope of this document and is for further study.

2. Overview of Inter-domain Multicast Application Transport

A multicast-based application delivery scenario is as follows:

- o Two independent administrative domains are interconnected via a peering point.
- o The peering point is either multicast enabled (end-to-end native multicast across the two domains) or it is connected by one of two possible tunnel types:
 - o A Generic Routing Encapsulation (GRE) Tunnel [RFC2784] allowing multicast tunneling across the peering point, or
 - o An Automatic Multicast Tunnel (AMT) [IETF-ID-AMT].
- o The application stream originates at a source in Domain 1.
- o An End User associated with Domain 2 requests the application. It is assumed that the application is suitable for delivery via multicast means (e.g., live streaming of major events, software downloads to large numbers of end user devices, etc.)
- o The request is communicated to the application source which provides the relevant multicast delivery information to the EU device via a "manifest file". At a minimum, this file contains the {Source, Group} or (S,G) information relevant to the multicast stream.
- o The application client in the EU device then joins the multicast stream distributed by the application source in domain 1 utilizing the (S,G) information provided in the manifest file. The manifest file may also contain additional information that the application client can use to locate the source and join the stream.

It should be noted that the second administrative domain - domain 2 - may be an independent network domain (e.g., Tier 1 network operator domain) or it could also be an Enterprise network operated by a single customer. The peering point architecture and requirements may have some unique aspects associated with the Enterprise case.

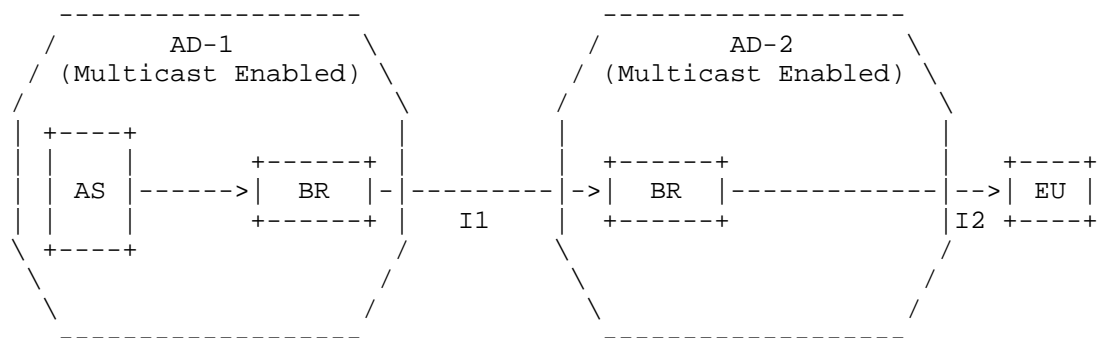
The Use Cases describing various architectural configurations for the multicast distribution along with associated requirements is described in section 3. Unique aspects related to the Enterprise network possibility will be described in this section. A comprehensive list of pertinent information that needs to be exchanged between the two domains to support various functions enabling the application transport is provided in section 4.

3. Inter-domain Peering Point Requirements for Multicast

The transport of applications using multicast requires that the inter-domain peering point is enabled to support such a process. There are three possible Use Cases for consideration.

3.1. Native Multicast

This Use Case involves end-to-end Native Multicast between the two administrative domains and the peering point is also native multicast enabled - Figure 1.



AD = Administrative Domain (Independent Autonomous System)
AS = Application (e.g., Content) Multicast Source
BR = Border Router
I1 = AD-1 and AD-2 Multicast Interconnection (MBGP or BGMP)
I2 = AD-2 and EU Multicast Connection

Figure 1 - Content Distribution via End to End Native Multicast

Advantages of this configuration are:

- o Most efficient use of bandwidth in both domains
- o Fewer devices in the path traversed by the multicast stream when compared to unicast transmissions.

From the perspective of AD-1, the one disadvantage associated with native multicast into AD-2 instead of individual unicast to every EU in AD-2 is that it does not have the ability to count the number of End Users as well as the transmitted bytes delivered to them. This information is relevant from the perspective of customer billing and operational logs. It is assumed that such data will be collected by the application layer. The application layer mechanisms for generating this information need to be robust enough such that all pertinent requirements for the source provider and the AD operator are satisfactorily met. The specifics of these methods are beyond the scope of this document.

Architectural guidelines for this configuration are as follows:

- o Dual homing for peering points between domains is recommended as a way to ensure reliability with full BGP table visibility.
- o If the peering point between AD-1 and AD-2 is a controlled network environment, then bandwidth can be allocated accordingly by the two domains to permit the transit of non-rate adaptive multicast traffic. If this is not the case, then it is recommended that the multicast traffic should support rate-adaption.
- o The sending and receiving of multicast traffic between two domains is typically determined by local policies associated with each domain. For example, if AD-1 is a service provider and AD-2 is an enterprise, then AD-1 may support local policies for traffic delivery to, but not traffic reception from AD-2.
- o Relevant information on multicast streams delivered to End Users in AD-2 is assumed to be collected by available capabilities in the application layer. The precise nature and formats of the collected information will be determined by directives from the source owner and the domain operators.

3.2. Peering Point Enabled with GRE Tunnel

The peering point is not native multicast enabled in this Use Case. There is a Generic Routing Encapsulation Tunnel provisioned over the peering point. In this case, the interconnection I1 between AD-1 and AD-2 in Figure 1 is multicast enabled via a Generic Routing Encapsulation Tunnel (GRE) [RFC2784] and encapsulating the multicast protocols across the interface. The routing configuration is basically unchanged: Instead of BGP (SAFI2) across the native IP multicast link between AD-1 and AD-2, BGP (SAFI2) is now run across the GRE tunnel.

Advantages of this configuration:

- o Highly efficient use of bandwidth in both domains although not as efficient as the fully native multicast Use Case.
- o Fewer devices in the path traversed by the multicast stream when compared to unicast transmissions.
- o Ability to support only partial IP multicast deployments in AD-1 and/or AD-2.
- o GRE is an existing technology and is relatively simple to implement.

Disadvantages of this configuration:

- o Per Use Case 3.1, current router technology cannot count the number of end users or the number bytes transmitted.
- o GRE tunnel requires manual configuration.
- o GRE must be in place prior to stream starting.
- o GRE is often left pinned up

Architectural guidelines for this configuration include the following:

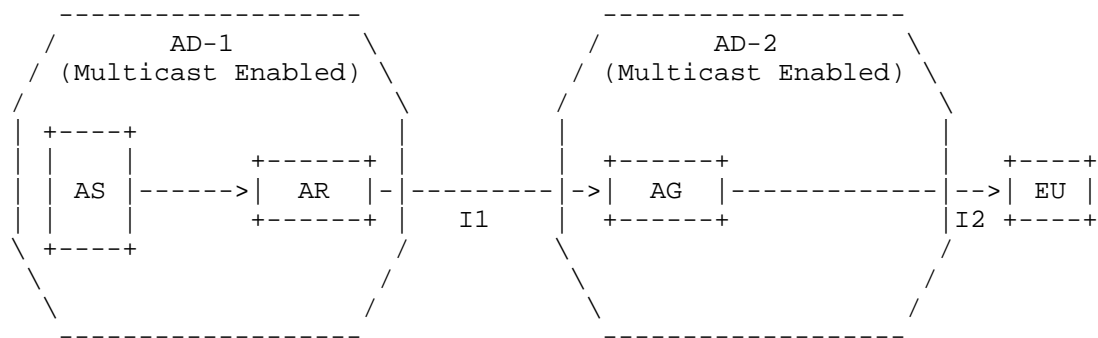
Guidelines (a) through (d) are the same as those described in Use Case 3.1.

- o GRE tunnels are typically configured manually between peering points to support multicast delivery between domains.

- o It is recommended that the GRE tunnel (tunnel server) configuration in the source network is such that it only advertises the routes to the application sources and not to the entire network. This practice will prevent unauthorized delivery of applications through the tunnel (e.g., if application - e.g., content - is not part of an agreed inter-domain partnership).

3.3. Peering Point Enabled with an AMT - Both Domains Multicast Enabled

Both administrative domains in this Use Case are assumed to be native multicast enabled here; however the peering point is not. The peering point is enabled with an Automatic Multicast Tunnel. The basic configuration is depicted in Figure 2.



AR = AMT Relay
 AG = AMT Gateway
 I1 = AMT Interconnection between AD-1 and AD-2
 I2 = AD-2 and EU Multicast Connection

Figure 2 - AMT Interconnection between AD-1 and AD-2

Advantages of this configuration:

- o Highly efficient use of bandwidth in AD-1.

- o AMT is an existing technology and is relatively simple to implement. Attractive properties of AMT include the following:
 - o Dynamic interconnection between Gateway-Relay pair across the peering point.
 - o Ability to serve clients and servers with differing policies.

Disadvantages of this configuration:

- o Per Use Case 3.1 (AD-2 is native multicast), current router technology cannot count the number of end users or the number bytes transmitted.
- o Additional devices (AMT Gateway and Relay pairs) may be introduced into the path if these services are not incorporated in the existing routing nodes.
- o Currently undefined mechanisms to select the AR from the AG automatically.

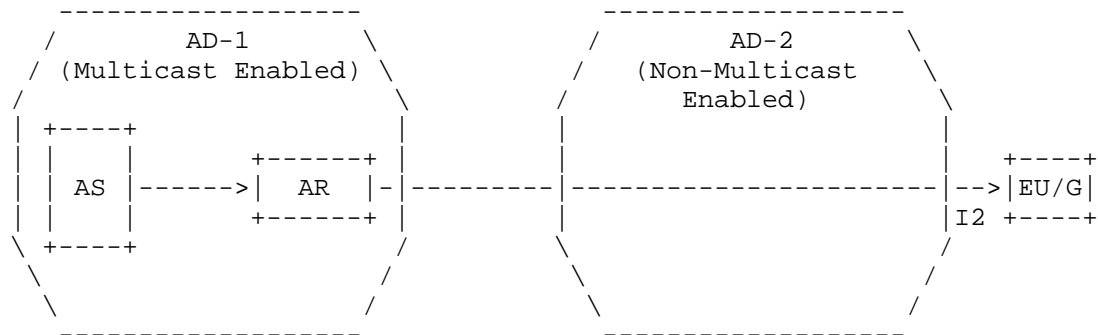
Architectural guidelines for this configuration are as follows:

Guidelines (a) through (d) are the same as those described in Use Case 3.1.

- e. It is recommended that AMT Relay and Gateway pairs be configured at the peering points to support multicast delivery between domains. AMT tunnels will then configure dynamically across the peering points once the Gateway in AD-2 receives the (S, G) information from the EU.

3.4. Peering Point Enabled with an AMT - AD-2 Not Multicast Enabled

In this AMT Use Case, the second administrative domain AD-2 is not multicast enabled. This implies that the interconnection between AD-2 and the End User is also not multicast enabled as depicted in Figure 3.



AS = Application Multicast Source

AR = AMT Relay

EU/G = Gateway client embedded in EU device

I2 = AMT Tunnel Connecting EU/G to AR in AD-1 through Non-Multicast Enabled AD-2.

Figure 3 - AMT Tunnel Connecting AD-1 AMT Relay and EU Gateway

This Use Case is equivalent to having unicast distribution of the application through AD-2. The total number of AMT tunnels would be equal to the total number of End Users requesting the application. The peering point thus needs to accommodate the total number of AMT tunnels between the two domains. Each AMT tunnel can provide the data usage associated with each End User.

Advantages of this configuration:

- o Highly efficient use of bandwidth in AD-1.
- o AMT is an existing technology and is relatively simple to implement. Attractive properties of AMT include the following:
 - o Dynamic interconnection between Gateway-Relay pair across the peering point.
 - o Ability to serve clients and servers with differing policies.
- o Each AMT tunnel serves as a count for each End User and is also able to track data usage (bytes) delivered to the EU.

Disadvantages of this configuration:

- o Additional devices (AMT Gateway and Relay pairs) are introduced into the transport path.
- o Assuming multiple peering points between the domains, the EU Gateway needs to be able to find the "correct" AMT Relay in AD-1.

Architectural guidelines for this configuration are as follows:

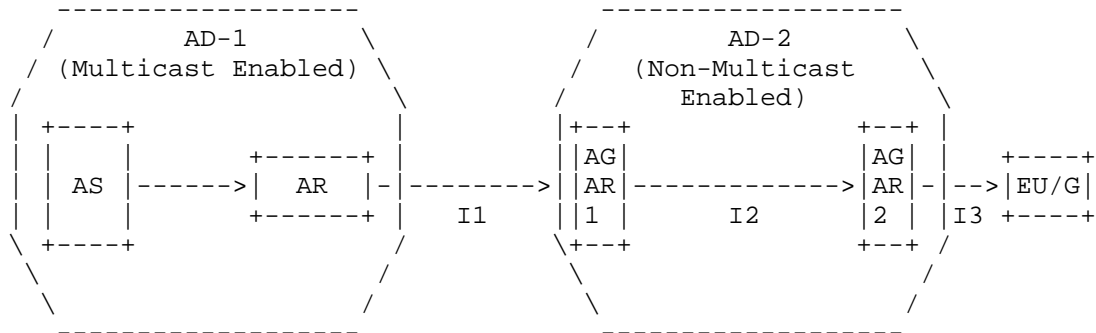
Guidelines (a) through (c) are the same as those described in Use Case 3.1.

d. It is recommended that proper procedures are implemented such that the AMT Gateway at the End User device is able to find the correct AMT Relay in AD-1 across the peering points. The application client in the EU device is expected to supply the (S, G) information to the Gateway for this purpose.

e. The AMT tunnel capabilities are expected to be sufficient for the purpose of collecting relevant information on the multicast streams delivered to End Users in AD-2.

3.5. AD-2 Not Multicast Enabled - Multiple AMT Tunnels Through AD-2

This is a variation of Use Case 3.4 as follows:



(Note: Diff-marks for the figure have been removed to improve viewing)

AS = Application Source
 AR = AMT Relay in AD-1
 AGAR1 = AMT Gateway/Relay node in AD-2 across Peering Point
 I1 = AMT Tunnel Connecting AR in AD-1 to GW in AGAR1 in AD-2
 AGAR2 = AMT Gateway/Relay node at AD-2 Network Edge
 I2 = AMT Tunnel Connecting Relay in AGAR1 to GW in AGAR2
 EU/G = Gateway client embedded in EU device
 I3 = AMT Tunnel Connecting EU/G to AR in AGAR2

Figure 4 - AMT Tunnel Connecting AD-1 AMT Relay and EU Gateway

Use Case 3.4 results in several long AMT tunnels crossing the entire network of AD-2 linking the EU device and the AMT Relay in AD-1 through the peering point. Depending on the number of End Users, there is a likelihood of an unacceptably large number of AMT tunnels - and unicast streams - through the peering point. This situation can be alleviated as follows:

- o Provisioning of strategically located AMT nodes at the edges of AD-2. An AMT node comprises co-location of an AMT Gateway and an AMT Relay. One such node is at the AD-2 side of the peering point (node AGAR1 in Figure 4).
- o Single AMT tunnel established across peering point linking AMT Relay in AD-1 to the AMT Gateway in the AMT node AGAR1 in AD-2.
- o AMT tunnels linking AMT node AGAR1 at peering point in AD-2 to other AMT nodes located at the edges of AD-2: e.g., AMT tunnel

I2 linking AMT Relay in AGAR1 to AMT Gateway in AMT node AGAR2 in Figure 4.

- o AMT tunnels linking EU device (via Gateway client embedded in device) and AMT Relay in appropriate AMT node at edge of AD-2: e.g., I3 linking EU Gateway in device to AMT Relay in AMT node AGAR2.

The advantage for such a chained set of AMT tunnels is that the total number of unicast streams across AD-2 is significantly reduced thus freeing up bandwidth. Additionally, there will be a single unicast stream across the peering point instead of possibly, an unacceptably large number of such streams per Use Case 3.4. However, this implies that several AMT tunnels will need to be dynamically configured by the various AMT Gateways based solely on the (S,G) information received from the application client at the EU device. A suitable mechanism for such dynamic configurations is therefore critical.

Architectural guidelines for this configuration are as follows:

Guidelines (a) through (c) are the same as those described in Use Case 3.1.

d. It is recommended that proper procedures are implemented such that the various AMT Gateways (at the End User devices and the AMT nodes in AD-2) are able to find the correct AMT Relay in other AMT nodes as appropriate. The application client in the EU device is expected to supply the (S, G) information to the Gateway for this purpose.

e. The AMT tunnel capabilities are expected to be sufficient for the purpose of collecting relevant information on the multicast streams delivered to End Users in AD-2.

4. Supporting Functionality

Supporting functions and related interfaces over the peering point that enable the multicast transport of the application are listed in this section. Critical information parameters that need to be exchanged in support of these functions are enumerated along with guidelines as appropriate. Specific interface functions for consideration are as follows.

4.1. Network Interconnection Transport and Security Guidelines

The term "Network Interconnection Transport" refers to the interconnection points between the two Administrative Domains. The following is a representative set of attributes that will need to be agreed to between the two administrative domains to support multicast delivery.

- o Number of Peering Points
- o Peering Point Addresses and Locations
- o Connection Type - Dedicated for Multicast delivery or shared with other services
- o Connection Mode - Direct connectivity between the two AD's or via another ISP
- o Peering Point Protocol Support - Multicast protocols that will be used for multicast delivery will need to be supported at these points. Examples of protocols include eBGP, BGMP, and MBGP.
- o Bandwidth Allocation - If shared with other services, then there needs to be a determination of the share of bandwidth reserved for multicast delivery.
- o QoS Requirements - Delay/latency specifications that need to be specified in an SLA.
- o AD Roles and Responsibilities - the role played by each AD for provisioning and maintaining the set of peering points to support multicast delivery.

From a security perspective, it is expected that normal/typical security procedures will be followed by each AD to facilitate multicast delivery to registered and authenticated end users. Some security aspects for consideration are:

- o Encryption - Peering point links may be encrypted per agreement if dedicated for multicast delivery.
- o Security Breach Mitigation Plan - In the event of a security breach, the two AD's are expected to have a mitigation plan for shutting down the peering point and directing multicast traffic

over alternated peering points. It is also expected that appropriate information will be shared for the purpose of securing the identified breach.

4.2. Routing Aspects and Related Guidelines

The main objective for multicast delivery routing is to ensure that the End User receives the multicast stream from the "most optimal" source [INF_ATIS_10] which typically:

- o Maximizes the multicast portion of the transport and minimizes any unicast portion of the delivery, and
- o Minimizes the overall combined network(s) route distance.

This routing objective applies to both Native and AMT; the actual methodology of the solution will be different for each. Regardless, the routing solution is expected to be:

- o Scalable
- o Avoid/minimize new protocol development or modifications, and
- o Be robust enough to achieve high reliability and automatically adjust to changes/problems in the multicast infrastructure.

For both Native and AMT environments, having a source as close as possible to the EU network is most desirable; therefore, in some cases, an AD may prefer to have multiple sources near different peering points, but that is entirely an implementation issue.

4.2.1 Native Multicast Routing Aspects

Native multicast simply requires that the Administrative Domains coordinate and advertise the correct source address(es) at their network interconnection peering points(i.e., border routers). An example of multicast delivery via a Native Multicast process across two administrative Domains is as follows assuming that the interconnecting peering points are also multicast enabled:

- o Appropriate information is obtained by the EU client who is a subscriber to AD-2 (see Use Case 3.1). This is usually done via an appropriate file transfer - this file is typically known as the manifest file. It contains instructions directing the EU

client to launch an appropriate application if necessary, and also additional information for the application about the source location and the group (or stream) id in the form of the "S,G" data. The "S" portion provides the name or IP address of the source of the multicast stream. The file may also contain alternate delivery information such as specifying the unicast address of the stream.

- o The client uses the join message with S,G to join the multicast stream [RFC2236].

To facilitate this process, the two AD's need to do the following:

- o Advertise the source id(s) over the Peering Points
- o Exchange relevant Peering Point information such as Capacity and Utilization (Other??)

4.2.2 GRE Tunnel over Interconnecting Peering Point

If the interconnecting peering point is not multicast enabled and both ADs are multicast enabled, then a simple solution is to provision a GRE tunnel between the two ADs - see Use Case 3.2.2. The termination points of the tunnel will usually be a network engineering decision, but generally will be between the border routers or even between the AD 2 border router and the AD 1 source (or source access router). The GRE tunnel would allow end-to-end native multicast or AMT multicast to traverse the interface. Coordination and advertisement of the source IP is still required.

The two AD's need to follow the same process as described in 4.2.1 to facilitate multicast delivery across the Peering Points.

4.2.3 Routing Aspects with AMT Tunnels

Unlike Native (with or without GRE), an AMT Multicast environment is more complex. It presents a dual layered problem because there are two criteria that should be simultaneously meet:

- o Find the closest AMT relay to the end-user that also has multicast connectivity to the content source and
- o Minimize the AMT unicast tunnel distance.

There are essentially two components to the AMT specification:

- o AMT Relays: These serve the purpose of tunneling UDP multicast traffic to the receivers (i.e., End Points). The AMT Relay will receive the traffic natively from the multicast media source and will replicate the stream on behalf of the downstream AMT Gateways, encapsulating the multicast packets into unicast packets and sending them over the tunnel toward the AMT Gateway. In addition, the AMT Relay may perform various usage and activity statistics collection. This results in moving the replication point closer to the end user, and cuts down on traffic across the network. Thus, the linear costs of adding unicast subscribers can be avoided. However, unicast replication is still required for each requesting endpoint within the unicast-only network.
- o AMT Gateway (GW): The Gateway will reside on an on End-Point - this may be a Personal Computer (PC) or a Set Top Box (STB). The AMT Gateway receives join and leave requests from the Application via an Application Programming Interface (API). In this manner, the Gateway allows the endpoint to conduct itself as a true Multicast End-Point. The AMT Gateway will encapsulate AMT messages into UDP packets and send them through a tunnel (across the unicast-only infrastructure) to the AMT Relay.

The simplest AMT Use Case (section 3.3) involves peering points that are not multicast enabled between two multicast enabled ADs. An AMT tunnel is deployed between an AMT Relay on the AD 1 side of the peering point and an AMT Gateway on the AD 2 side of the peering point. One advantage to this arrangement is that the tunnel is established on an as needed basis and need not be a provisioned element. The two ADs can coordinate and advertise special AMT Relay Anycast addresses with each other - though they may alternately decide to simply provision Relay addresses, though this would not be an optimal solution in terms of scalability.

Use Cases 3.4 and 3.5 describe more complicated AMT situations as AD-2 is not multicast enabled. For these cases, the End User device needs to be able to setup an AMT tunnel in the most optimal manner. Using an Anycast IP address for AMT Relays allows for all AMT Gateways to find the "closest" AMT Relay - the nearest edge of the multicast topology of the source. An example of a basic delivery via an AMT Multicast process for these two Use Cases is as follows:

- o The manifest file is obtained by the EU client application. This file contains instructions directing the EU client to an ordered list of particular destinations to seek the requested stream and, for multicast, specifies the source location and the group (or stream) ID in the form of the "S,G" data. The "S" portion provides

the URI (name or IP address) of the source of the multicast stream and the "G" identifies the particular stream originated by that source. The manifest file may also contain alternate delivery information such as the address of the unicast form of the content to be used, for example, if the multicast stream becomes unavailable.

- o Using the information in the manifest file, and possibly information provisioned directly in the EU client, a DNS query is initiated in order to connect the EU client/AMT Gateway to an AMT Relay.
- o Query results are obtained, and may return an Anycast address or a specific unicast address of a relay. Multiple relays will typically exist. The Anycast address is a routable "pseudo-address" shared among the relays that can gain multicast access to the source.
- o If a specific IP address unique to a relay was not obtained, the AMT Gateway then sends a message (e.g., the discovery message) to the Anycast address such that the network is making the routing choice of particular relay - e.g., closest relay to the EU. (Note that in IPv6 there is a specific Anycast format and Anycast is inherent in IPv6 routing, whereas in IPv4 Anycast is handled via provisioning in the network. Details are out of scope for this document.)
- o The contacted AMT Relay then returns its specific unicast IP address (after which the Anycast address is no longer required). Variations may exist as well.
- o The AMT Gateway uses that unicast IP address to initiate a three-way handshake with the AMT Relay.
- o AMT Gateway provides "S,G" to the AMT Relay (embedded in AMT protocol messages).
- o AMT Relay receives the "S,G" information and uses the S,G to join the appropriate multicast stream, if it has not already subscribed to that stream.
- o AMT Relay encapsulates the multicast stream into the tunnel between the Relay and the Gateway, providing the requested content to the EU.

Note: Further routing discussion on optimal method to find "best AMT Relay/GW combination" and information exchange between AD's to be provided.

4.3. Back Office Functions - Billing and Logging Guidelines

Back Office refers to the following:

- o Servers and Content Management systems that support the delivery of applications via multicast and interactions between ADs.
- o Functionality associated with logging, reporting, ordering, provisioning, maintenance, service assurance, settlement, etc.

4.3.1 Provisioning Guidelines

Resources for basic connectivity between ADs Providers need to be provisioned as follows:

- o Sufficient capacity must be provisioned to support multicast-based delivery across ADs.
- o Sufficient capacity must be provisioned for connectivity between all supporting back-offices of the ADs as appropriate. This includes activating proper security treatment for these back-office connections (gateways, firewalls, etc) as appropriate.
- o Routing protocols as needed, e.g. configuring routers to support these.

Provisioning aspects related to Multicast-Based inter-domain delivery are as follows.

The ability to receive requested application via multicast is triggered via the manifest file. Hence, this file must be provided to the EU regarding multicast URL - and unicast fallback if applicable. AD-2 must build manifest and provision capability to provide the file to the EU.

Native multicast functionality is assumed to be available in across many ISP backbones, peering and access networks. If however, native multicast is not an option (Use Cases 3.4 and 3.5), then:

- o EU must have multicast client to use AMT multicast obtained either from Application Source (per agreement with AD-1) or from AD-1 or AD-2 (if delegated by the Application Source).

- o If provided by AD-1/AD-2, then the EU could be redirected to a client download site (note: this could be an Application Source site). If provided by the Application Source, then this Source would have to coordinate with AD-1 to ensure the proper client is provided (assuming multiple possible clients).
- o Where AMT Gateways support different application sets, all AD-2 AMT Relays need to be provisioned with all source & group addresses for streams it is allowed to join.
- o DNS across each AD must be provisioned to enable a client GW to locate the optimal AMT Relay (i.e. longest multicast path and shortest unicast tunnel) with connectivity to the content's multicast source.

Provisioning Aspects Related to Operations and Customer Care are stated as follows.

Each AD provider is assumed to provision operations and customer care access to their own systems.

AD-1's operations and customer care functions must have visibility to what is happening in AD-2's network or to the service provided by AD-2, sufficient to verify their mutual goals and operations, e.g. to know how the EU's are being served. This can be done in two ways:

- o Automated interfaces are built between AD-1 and AD-2 such that operations and customer care continue using their own systems. This requires coordination between the two AD's with appropriate provisioning of necessary resources.
- o AD-1's operations and customer care personnel are provided access directly to AD-2's system. In this scenario, additional provisioning in these systems will be needed to provide necessary access. Additional provisioning must be agreed to by the two AD-2s to support this option.

4.3.2 Application Accounting Billing Guidelines

All interactions between pairs of ADs can be discovered and/or be associated with the account(s) utilized for delivered applications. Supporting guidelines are as follows:

- o A unique identifier is recommended to designate each master account.
- o AD-2 is expected to set up "accounts" (logical facility generally protected by login/password/credentials) for use by AD-1. Multiple

accounts and multiple types/partitions of accounts can apply, e.g. customer accounts, security accounts, etc.

4.3.3 Log Management Guidelines

Successful delivery of applications via multicast between pairs of interconnecting ADs requires that appropriate logs will be exchanged between them in support. Associated guidelines are as follows.

AD-2 needs to supply logs to AD-1 per existing contract(s). Examples of log types include the following:

- o Usage information logs at aggregate level.
- o Usage failure instances at an aggregate level.
- o Grouped or sequenced application access performance/behavior/failure at an aggregate level to support potential Application Provider-driven strategies. Examples of aggregate levels include grouped video clips, web pages, and sets of software download.
- o Security logs, aggregated or summarized according to agreement (with additional detail potentially provided during security events, by agreement).
- o Access logs (EU), when needed for troubleshooting.
- o Application logs (what is the application doing), when needed for shared troubleshooting.
- o Syslogs (network management), when needed for shared troubleshooting.

The two ADs may supply additional security logs to each other as agreed to by contract(s). Examples include the following:

- o Information related to general security-relevant activity which may be of use from a protective or response perspective, such as types and counts of attacks detected, related source information, related target information, etc.
- o Aggregated or summarized logs according to agreement (with additional detail potentially provided during security events, by agreement)

4.3.4 Settlement Guidelines

Settlements between the ADs relate to (1) billing and reimbursement aspects for delivery of applications, and (2) aggregation, transport, and collection of data in preparation for the billing and

reimbursement aspects for delivery of applications for the Application Provider. At a high level:

- o AD-2 collects "usage" data for AD-1 related to application delivery to End Users, and submits invoices to AD-1 based on this usage data. The data may include information related to the type of content delivered, total bandwidth utilized, storage utilized, features supported, etc.
- o AD-1 collects all available data from partner AD-2 and creates aggregate reports pertaining to responsible Application Providers, and submits subsequent reports to these Providers for reimbursements.
- o AD-1 may convey charging values or charging rules to the AD-2, proactively or in response to a query, especially in cases where these may change.
- o AD-2 may convey prices/rates to AD-1, proactively or in response to a query, especially in cases where these may change.
- o Usage data may be collected per end user or on an aggregated basis; the method of collection will depend on the application delivered and/or the agreements with the source provider. In all cases, usage volume is expected to be in terms of delivered packet bits or bytes.

4.4. Operations - Service Performance and Monitoring Guidelines

Service Performance refers to monitoring metrics related to multicast delivery via probes. The focus is on the service provided by AD-2 to AD-1 on behalf of all multicast application sources (metrics may be specified for SLA use or otherwise). Associated guidelines are as follows:

- o Both AD's are expected to monitor, collect, and analyze service performance metrics for multicast applications. AD-2 provides relevant performance information to AD-1; this enables AD-1 to create an end-to-end performance view on behalf of the multicast application source.
- o Both AD's are expected to agree on the type of probes to be used to monitor multicast delivery performance. For example, AD-2 may permit AD-1's probes to be utilized in the AD-2 multicast service footprint. Alternately, AD-2 may deploy its own probes and relay performance information back to AD-1.

- o In the event of performance degradation (SLA violation), AD-1 may have to compensate the multicast application source per SLA agreement. As appropriate, AD-1 may seek compensation from AD-2 if the cause of the degradation is in AD-2's network.

Service Monitoring generally refers to a service (as a whole) provided on behalf of a particular multicast application source provider. It thus involves complaints from End Users when service problems occur. EU's direct their complaints to the source provider; in turn the source provider submits these complaints to AD-1. The responsibility for service delivery lies with AD-1; as such AD-1 will need to determine where the service problem is occurring - its own network or in AD-2. It is expected that each AD will have tools to monitor multicast service status in its own network.

- o Both AD's will determine how best to deploy multicast service monitoring tools. Typically, each AD will deploy its own set of monitoring tools; in which case, both AD's are expected to inform each other when multicast delivery problems are detected.
- o AD-2 may experience some problems in its network. For example, for the AMT Use Cases, one or more AMT Relays may be experiencing difficulties. AD-2 may be able to fix the problem by rerouting the multicast streams via alternate AMT Relays. If the fix is not successful and multicast service delivery degrades, then AD-2 needs to report the issue to AD-1.
- o When problem notification is received from a multicast application source, AD-1 determines whether the cause of the problem is within its own network or within the AD-2 domain. If the cause is within the AD-2 domain, then AD-1 supplies all necessary information to AD-2. Examples of supporting information include the following:
 - o Kind of problem(s)
 - o Starting point & duration of problem(s).
 - o Conditions in which problem(s) occur.
 - o IP address blocks of affected users.
 - o ISPs of affected users.

- o Type of access e.g., mobile versus desktop.
- o Locations of affected EUs.
- o Both AD's conduct some form of root cause analysis for multicast service delivery problems. Examples of various factors for consideration include:
 - o Verification that the service configuration matches the product features.
 - o Correlation and consolidation of the various customer problems and resource troubles into a single root service problem.
 - o Prioritization of currently open service problems, giving consideration to problem impact, service level agreement, etc.
 - o Conduction of service tests, including one time tests or a series of tests over a period of time.
 - o Analysis of test results.
 - o Analysis of relevant network fault or performance data.
 - o Analysis of the problem information provided by the customer (CP).
- o Once the cause of the problem has been determined and the problem has been fixed, both AD's need to work jointly to verify and validate the success of the fix.
- o Faults in service could lead to SLA violation for which the multicast application source provider may have to be compensated by AD-1. Subsequently, AD-1 may have to be compensated by AD-2 based on the contract.

4.5. Client Reliability Models/Service Assurance Guidelines

There are multiple options for instituting reliability architectures, most are at the application level. Both AD's should work those out with their contract/agreement and with the multicast application source providers.

Network reliability can also be enhanced by the two AD's by provisioning alternate delivery mechanisms via unicast means.

5. Security Considerations

DRM and Application Accounting, Authorization and Authentication should be the responsibility of the multicast application source provider and/or AD-1. AD-1 needs to work out the appropriate agreements with the source provider.

Network has no DRM responsibilities, but might have authentication and authorization obligations. These though are consistent with normal operations of a CDN to insure end user reliability, security and network security

AD-1 and AD-2 should have mechanisms in place to ensure proper accounting for the volume of bytes delivered through the peering point and separately the number of bytes delivered to EUs.

If there are problems related to failure of token authentication when end-users are supported by AD-2, then some means of validating proper working of the token authentication process (e.g., back-end servers querying the multicast application source provider's token authentication server are communicating properly) should be considered. Details will have to be worked out during implementation (e.g., test tokens or trace token exchange process).

6. IANA Considerations

7. Conclusions

This Best Current Practice document provides detailed Use Case scenarios for the transmission of applications via multicast across peering points between two Administrative Domains. A detailed set of guidelines supporting the delivery is provided for all Use Cases.

For Use Cases involving AMT tunnels (cases 3.4 and 3.5), it is recommended that proper procedures are implemented such that the various AMT Gateways (at the End User devices and the AMT nodes in AD-2) are able to find the correct AMT Relay in other AMT nodes as appropriate. Section 4.3 provides an overview of one method that finds the optimal Relay-Gateway combination via the use of an Anycast IP address for AMT Relays.

8. References

8.1. Normative References

[RFC2784] D. Farinacci, T. Li, S. Hanks, D. Meyer, P. Traina, "Generic Routing Encapsulation (GRE)", RFC 2784, March 2000

[IETF-ID-AMT] G. Bumgardner, "Automatic Multicast Tunneling", draft-ietf-mboned-auto-multicast-13, April 2012, Work in progress

[RFC4604] H. Holbrook, et al, "Using Internet Group Management Protocol Version 3 (IGMPv3) and Multicast Listener Discovery Protocol Version 2 (MLDv2) for Source Specific Multicast", RFC 4604, August 2006

[RFC4607] H. Holbrook, et al, "Source Specific Multicast", RFC 4607, August 2006

8.2. Informative References

[INF_ATIS_10] "CDN Interconnection Use Cases and Requirements in a Multi-Party Federation Environment", ATIS Standard A-0200010, December 2012

9. Acknowledgments

Authors' Addresses

Percy S. Tarapore
AT&T
Phone: 1-732-420-4172
Email: tarapore@att.com

Robert Sayko
AT&T
Phone: 1-732-420-3292
Email: rs1983@att.com

Greg Shepherd
Cisco
Phone:
Email: shep@cisco.com

Toerless Eckert
Cisco
Phone:
Email: eckert@cisco.com

Ram Krishnan
Brocade
Phone:
Email: ramk@brocade.com

L3VPN Working Group
Internet Draft
Intended Status: Standards Track
Expires: November 12, 2014

Jeffrey Zhang
Lenny Giuliano
Juniper Networks, Inc.

Eric C. Rosen
Karthik Subramanian
Cisco Systems, Inc.

Dante J. Pacella
Verizon

Jason Schiller
Google

May 12, 2014

Global Table Multicast with BGP-MVPN Procedures

draft-zzhang-l3vpn-mvpn-global-table-mcast-04.txt

Abstract

RFC6513, RFC6514, and other RFCs describe protocols and procedures which a Service Provider (SP) may deploy in order offer Multicast Virtual Private Network (Multicast VPN or MVPN) service to its customers. Some of these procedures use BGP to distribute VPN-specific multicast routing information across a backbone network. With a small number of relatively minor modifications, the very same BGP procedures can also be used to distribute multicast routing information that is not specific to any VPN. Multicast that is outside the context of a VPN is known as "Global Table Multicast", or sometimes simply as "Internet multicast". In this document, we describe the modifications that are needed to use the MVPN BGP procedures for Global Table Multicast.

Status of this Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at <http://www.ietf.org/ietf/lid-abstracts.txt>.

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

Copyright and License Notice

Copyright (c) 2014 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. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1	Introduction	4
2	Adapting MVPN Procedures to GTM	6
2.1	Use of Route Distinguishers	7
2.2	Use of Route Targets	7
2.3	UMH-eligible Routes	9
2.3.1	Routes of SAFI 1, 2 or 4 with MVPN ECs	10
2.3.2	MVPN ECs on the Route to the Next Hop	11
2.3.3	Non-BGP Routes as the UMH-eligible Routes	12
2.3.4	Why SFS Does Not Apply to GTM	13
2.4	Inclusive and Selective Tunnels	14
2.5	I-PMSI A-D Routes	14
2.5.1	Intra-AS I-PMSI A-D Routes	14
2.5.2	Inter-AS I-PMSI A-D Routes	15
2.6	S-PMSI A-D Routes	15
2.7	Leaf A-D Routes	15
2.8	Source Active A-D Routes	15
2.8.1	Finding the Originator of an SA A-D Route	15
2.8.2	Optional Additional Constraints on Distribution	16
2.9	C-multicast Source/Shared Tree Joins	17
3	Differences from other MVPN-like GTM Procedures	18
4	IANA Considerations	19
5	Security Considerations	19
6	Additional Contributors	20
7	Acknowledgments	20
8	Authors' Addresses	21
9	References	22
9.1	Normative References	22
9.2	Informative References	22

1. Introduction

[RFC4364] specifies architecture, protocols, and procedures that a Service Provider (SP) can use to provide Virtual Private Network (VPN) service to its customers. In that architecture, one or more Customer Edge (CE) routers attach to a Provider Edge (PE) router. Each CE router belongs to a single VPN, but CE routers from several VPNs may attach to the same PE router. In addition, CEs from the same VPN may attach to different PEs. BGP is used to carry VPN-specific information among the PEs. Each PE router maintains a separate Virtual Routing and Forwarding table (VRF) for each VPN to which it is attached.

[RFC6513] and [RFC6514] extend the procedures of [RFC4364] to allow the SP to provide multicast service to its VPN customers. The customer's multicast routing protocol (e.g., PIM) is used to exchange multicast routing information between a CE and a PE. The PE stores a given customer's multicast routing information in the VRF for that customer's VPN. BGP is used to distribute certain multicast-related control information among the PEs that attach to a given VPN, and BGP may also be used to exchange the customer multicast routing information itself among the PEs.

While this multicast architecture was originally developed for VPNs, it can also be used (with a small number of modifications to the procedures) to distribute multicast routing information that is not specific to VPNs. The purpose of this document is to specify the way in which BGP MVPN procedures can be adapted to support non-VPN multicast.

Multicast routing information that is not specific to VPNs is stored in a router's "global table", rather than in a VRF; hence it is known as "Global Table Multicast" (GTM). GTM is sometimes more simply called "Internet multicast". However, we will avoid that term because it suggests that the multicast data streams are available on the "public" Internet. The procedures for GTM can certainly be used to support multicast on the public Internet, but they can also be used to support multicast streams that are not public, e.g., content distribution streams offered by content providers to paid subscribers. For the purposes of this document, all that matters is that the multicast routing information is maintained in a global table rather than in a VRF.

This architecture does assume that the network over which the multicast streams travel can be divided into a "core network" and one or more non-core parts of the network, which we shall call "attachment networks". The multicast routing protocol used in the attachment networks may not be the same as the one used in the core,

so we consider there to be a "protocol boundary" between the core network and the attachment networks. We will use the term "Protocol Boundary Router" (PBR) to refer to the core routers that are at the boundary. We will use the term "Attachment Router" (AR) to refer to the routers that are not in the core but that attach to the PBRs.

This document does not make any particular set of assumptions about the protocols that the ARs and the PBRs use to exchange unicast and multicast routing information with each other. For instance, multicast routing information could be exchanged between an AR and a PBR via PIM, IGMP, or even BGP. Multicast routing also depends on an exchange of routes that are used for looking up the path to the root of a multicast tree. This routing information could be exchanged between an AR and a PBR via IGP, via EBGp, or via IBGP ([RFC6368]). Note that if IBGP is used, the [RFC6368] "push/pop procedures" are not necessary.

The PBRs are not necessarily "edge" routers, in the sense of [RFC4364]. For example, they may be both be Autonomous System Border Routers (ASBR). As another example, an AR may be an "access router" attached to a PBR that is an OSPF Area Border Router (ABR). Many other deployment scenarios are possible. However, the PBRs are always considered to be delimiting a "backbone" or "core" network. A multicast data stream from an AR is tunneled over the core network from an Ingress PBR to one or more Egress PBRs. Multicast routing information that a PBR learns from the ARs attached to it is stored in the PBR's global table. The PBRs use BGP to distribute multicast routing and auto-discovery information among themselves. This is done following the procedures of [RFC6513], [RFC6514], and other MVPN specifications, as modified in this document.

In general, PBRs follow the same MVPN/BGP procedures that PE routers follow, except that these procedures are adapted to be applicable to the global table rather than to a VRF. Details are provided in subsequent sections of this document.

By supporting GTM using the BGP procedures designed for MVPN, one obtains a single control plane that governs the use of both VPN and non-VPN multicast. Most of the features and characteristics of MVPN carry over automatically to GTM. These include scaling, aggregation, flexible choice of tunnel technology in the SP network, support for both segmented and non-segmented tunnels, ability to use wildcards to identify sets of multicast flows, support for the Any Source Multicast (ASM), Single Source Multicast (SSM), and Bidirectional (bidir) multicast paradigms, support for both IPv4 and IPv6 multicast flows over either an IPv4 or IPv6 SP infrastructure, support for unsolicited flooded data (including support for BSR as RP-to-group mapping protocols), etc.

This document not only uses MVPN procedures for GTM, but also, insofar as possible, uses the same protocol elements, encodings, and formats. The BGP Updates for GTM thus use the same Subsequent Address Family Identifier (SAFI), and have the same Network Layer Reachability Information (NLRI) format, as the BGP Updates for MVPN.

Details for supporting MVPN (either IPv4 or IPv6 MVPN traffic) over an IPv6 backbone network can be found in [RFC6515]. The procedures and encodings described therein are also applicable to GTM.

The document [SEAMLESS-MCAST] extends [RFC6514] by providing procedures that allow tunnels through the core to be "segmented" at ABRs within the core. The ABR segmentation procedures are also applicable to GTM as defined in the current document. In general, the MVPN procedures of [SEAMLESS-MCAST], adapted as specified in the current document, are applicable to GTM.

The document [SEAMLESS-MCAST] also defines a set of procedures for GTM. Those procedures are different from the procedures defined in the current document, and the two sets of procedures are not interoperable with each other. The two sets of procedures can co-exist in the same network, as long as they are not applied to the same multicast flows or to the same multicast group addresses. See section 3 for more details.

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].

2. Adapting MVPN Procedures to GTM

In general, PBRs support Global Table Multicast by using the procedures that PE routers use to support VPN multicast. For GTM, where [RFC6513] and [RFC6514] talk about the "PE-CE interface", one should interpret that to mean the interface between the AR and the PBR. For GTM, where [RFC6513] and [RFC6514] talk about the "backbone" network, one should interpret that to mean the part of the network that is delimited by the PBRs.

A few adaptations to the procedures of [RFC6513] and [RFC6514] need to be made. Those adaptations are described in the following subsections.

2.1. Use of Route Distinguishers

The MVPN procedures require the use of BGP routes, defined in [RFC6514], that have a SAFI value of 5 ("MCAST-VPN"). We refer to these simply as "MCAST-VPN routes". [RFC6514] defines the Network Layer Reachability Information (NLRI) format for MCAST-VPN routes. The NLRI field always begins with a "Route Type" octet, and, depending on the route type, may be followed by a "Route Distinguisher" (RD) field.

When a PBR originates an MCAST-VPN route in support of GTM, the RD field (for those routes types where it is defined) of that route's NLRI MUST be set to zero (i.e., to 64 bits of zero). Since no VRF may have an RD of zero, this allows "MCAST-VPN" routes that are "about" GTM to be distinguished from MCAST-VPN routes that are about VPNs.

2.2. Use of Route Targets

The MVPN procedures require all MCAST-VPN routes to carry Route Targets (RTs). When a PE router receives an MCAST-VPN route, it processes the route in the context of a particular VRF if and only if the route is carrying an RT that is configured as one of that VRF's "import RTs".

There are two different "kinds" of RT used in MVPN.

- One kind of RT is carried only by the following MCAST-VPN route types: C-multicast Shared Tree Joins, C-multicast Source Tree Joins, and Leaf A-D routes. This kind of RT identifies the PE router that has been selected by the route's originator as the "Upstream PE" or as the "Upstream Multicast Hop" (UMH) for a particular (set of) multicast flow(s). Per [RFC6514] and [RFC6515], this RT must be an IPv4-address-specific or IPv6-address-specific Extended Community (EC), whose "Global Administrator" field identifies the Upstream PE or the UMH. If the Global Administrator field identifies the Upstream PE, the "Local Administrator" field identifies a particular VRF in that PE.

The GTM procedures of this document require the use of this type of RT, in exactly the same situations where it is used in the MVPN specification. However, one adaptation is necessary: the "Local Administrator" field of this kind of RT MUST always be set to zero, thus implicitly identifying the global table, rather than identifying a VRF. We will refer to this kind of RT as a "PBR-identifying RT".

- The other kind of RT is the conventional RT first specified in [RFC4364]. It does not necessarily identify a particular router by address, but is used to constrain the distribution of VPN routes, and to ensure that a given VPN route is processed in the context of a given VRF if and only if the route is carrying an RT that has been configured as one of that VRF's "import RTs".

Whereas every VRF must be configured with at least one import RT, there is heretofore no requirement to configure any RTs for the global table of any router. As stated above, this document makes the use of PBR-identifying RTs mandatory for GTM. This document makes the use of non-PBR-identifying RTs OPTIONAL for GTM.

The procedures for the use of RTs in GTM are the following:

- If the global table of a particular PBR is NOT configured with any import RTs, then a received MCAST-VPN route is processed in the context of the global table only if it is carrying no RTs, or if it is carrying a PBR-identifying RT whose Global Administrator field identifies that PBR.
- The global table in each PBR MAY be configured with (a) a set of export RTs to be attached to MCAST-VPN routes that are originated to support GTM, and (b) with a set of import RTs for GTM.

If the global table of a given PBR has been so configured, the PBR will process a received MCAST-VPN route in the context of the global table if and only if the route carries an RT that is one of the global table's import RTs, or if the route carries a PBR-identifying RT whose global administrator field identifies the PBR.

If the global tables are configured with RTs, care must be taken to ensure that the RTs configured for the global table are distinct from any RTs used in support of MVPN (except in the case where it is actually intended to create an "extranet" [MVPN-extranet] in which some sources are reachable in global table context while others are reachable in VPN context.)

The "RT Constraint" procedures of [RFC4684] MAY be used to constrain the distribution of MCAST-VPN routes (or other routes) that carry RTs that have been configured as import RTs for GTM. (This includes the PBR-identifying RTs.)

In [RFC6513], the UMH-eligible routes (see section 5.1 of [RFC6513], "Eligible Routes for UMH Selection") are generally routes of SAFI 128 (Labeled VPN-IP routes) or 129 (VPN-IP multicast routes), and are required to carry RTs. These RTs determine which VRFs import which

such routes. However, for GTM, when the UMH-eligible routes may be routes of SAFI 1, 2, or 4, the routes are not required to carry RTs. This document does NOT specify any new rules for determine whether a SAFI 1, 2, or 4 route is to be imported into the global table of any PBR.

2.3. UMH-eligible Routes

[RFC6513] section 5.1 defines procedures by which a PE router determines the "C-root", the "Upstream Multicast Hop" (UMH), the "Upstream PE", and the "Upstream RD" of a given multicast flow. (In non-VPN multicast documents, the UMH of a multicast flow at a particular router is generally known as the "RPF neighbor" for that flow.) It also defines procedures for determining the "Source AS" of a particular flow. Note that in GTM, the "Upstream PE" is actually the "Upstream PBR".

The definition of the C-root of a flow is the same for GTM as for MVPN.

For MVPN, to determine the UMH, Upstream PE, Upstream RD, and Source AS of a flow, one looks up the C-root of the flow in a particular VRF, and finds the "UMH-eligible" routes (see section 5.1.1 of [RFC6513]) that "match" the C-root. From among these, one is chosen as the "selected UMH route".

For GTM, the C-root is of course looked up in the global table, rather than in a VRF. For MVPN, the UMH-eligible routes are routes of SAFI 128 or 129. For GTM, the UMH-eligible routes are routes of SAFI 1, SAFI 4, or SAFI 2. If the global table has imported routes of SAFI 2, then these are the UMH-eligible routes. Otherwise, routes of SAFI 1 or SAFI 4 are the UMH-eligible routes. For the purpose of UMH determination, if a SAFI 1 route and a SAFI 4 route contain the same IP prefix in their respective NLRI fields, then the two routes are considered by the BGP bestpath selection process to be comparable.

[RFC6513] defines procedures for determining which of the UMH-eligible routes that match a particular C-root is to become the "Selected UMH route". With one exception, these procedures are also applicable to GTM. The one exception is the following. Section 9.1.2 of [RFC6513] defines a particular method of choosing the Upstream PE, known as "Single Forwarder Selection" (SFS). This procedure MUST NOT be used for GTM (see section 2.3.4 for an explanation of why the SFS procedure cannot be applied to GTM).

In GTM, the "Upstream RD" of a multicast flow is always considered to

be zero, and is NOT determined from the Selected UMH route.

The MVPN specifications require that when BGP is used for distributing multicast routing information, the UMH-eligible routes MUST carry the VRF Route Import EC and the Source AS EC. To determine the Upstream PE and Source AS for a particular multicast flow, the Upstream PE and Source AS are determined, respectively, from the VRF Route Import EC and the Source AS EC of the Selected UMH route for that flow. These ECs are generally attached to the UMH-eligible routes by the PEs that originate the routes.

In GTM, there are certain situations in which it is allowable to omit the VRF Route Import EC and/or the Source AS EC from the UMH-eligible routes. The following sub-sections specify the various options for determining the Upstream PBR and the Source AS in GTM.

The procedures in sections 2.3.1 MUST be implemented. The procedures in sections 2.3.2 and 2.3.3 are OPTIONAL to implement. It should be noted that while the optional procedures may be useful in particular deployment scenarios, there is always the potential for interoperability problems when relying on OPTIONAL procedures.

2.3.1. Routes of SAFI 1, 2 or 4 with MVPN ECs

If the UMH-eligible routes have a SAFI of 1, 2 or 4, then they MAY carry the VRF Route Import EC and/or the Source AS EC. If the selected UMH route is a route of SAFI 1, 2 or 4 that carries the VRF Route Import EC, then the Upstream PBR is determined from that EC. Similarly, if the selected UMH route is a route of SAFI 1, 2, or 4 route that carries the Source AS EC, the Source AS is determined from that EC.

When the procedure of this section is used, a PBR that distributes a UMH-eligible route to other PBRs is responsible for ensuring that the VRF Route Import and Source AS ECs are attached to it.

If the selected UMH-eligible route has a SAFI of 1, 2 or 4, but is not carrying a VRF Route Import EC, then the Upstream PBR is determined as specified in section 2.3.2 or 2.3.3 below.

If the selected UMH-eligible route has a SAFI of 1, 2 or 4, but is not carrying a Source AS EC, then the Source AS is considered to be the local AS.

2.3.2. MVPN ECs on the Route to the Next Hop

Some service providers may consider it to be undesirable to have the PBRs put the VRF Route Import EC on all the UMH-eligible routes. Or there may be deployment scenarios in which the UMH-eligible routes are not advertised by the PBRs at all. The procedures described in this section provide an alternative that can be used under certain circumstances.

The procedures of this section are OPTIONAL.

In this alternative procedure, each PBR MUST originate a BGP route of SAFI 1, 2 or 4 to itself. This route MUST carry a VRF Route Import EC that identifies the PBR. The address that appears in the Global Administrator field of that EC MUST be the same address that appears in the NLRI and in the Next Hop field of that route. This route MUST also carry a Source AS EC identifying the AS of the PBR.

Whenever the PBR distributes a UMH-eligible route for which it sets itself as next hop, it MUST use this same IP address as the Next Hop of the UMH-eligible route that it used in the route discussed in the prior paragraph.

When the procedure of this section is used, then when a PBR is determining the Selected UMH Route for a given multicast flow, it may find that the Selected UMH Route has no VRF Route Import EC. In this case, the PBR will look up (in the global table) the route to the Next Hop of the Selected UMH route. If the route to the Next Hop has a VRF Route Import EC, that EC will be used to determine the Upstream PBR, just as if the EC had been attached to the Selected UMH Route.

If recursive route resolution is required in order to resolve the next hop, the Upstream PBR will be determined from the first route with a VRF Route Import EC that is encountered during the recursive route resolution process. (The recursive route resolution process itself is not modified by this document.)

The same procedure can be applied to find the Source AS, except that the Source AS EC is used instead of the VRF Route Import EC.

Note that this procedure is only applicable in scenarios where it is known that the Next Hop of the UMH-eligible routes is not be changed by any router that participates in the distribution of those routes; this procedure MUST NOT be used in any scenario where the next hop may be changed between the time one PBR distributes the route and another PBR receives it. The PBRs have no way of determining dynamically whether the procedure is applicable in a particular deployment; this must be made known to the PBRs by provisioning.

Some scenarios in which this procedure can be used are:

- all PBRs are in the same AS, or
- the UMH-eligible routes are distributed among the PBRs by a Route Reflector (that does not change the next hop), or
- the UMH-eligible routes are distributed from one AS to another through ASBRs that do not change the next hop.

If the procedures of this section are used in scenarios where they are not applicable, GTM will not function correctly.

2.3.3. Non-BGP Routes as the UMH-eligible Routes

In particular deployment scenarios, there may be specific procedures that can be used, in those particular scenarios, to determine the Upstream PBR for a given multicast flow.

Suppose the PBRs neither put the VRF Route Import EC on the UMH-eligible routes, nor do they distribute BGP routes to themselves. It may still be possible to determine the Upstream PBR for a given multicast flow, using specific knowledge about the deployment.

For example, suppose it is known that all the PBRs are in the same OSPF area. It may be possible to determine the Upstream PBR for a given multicast flow by looking at the link state database to see which router is attached to the flow's C-root.

As another example, suppose it is known that the set of PBRs is fully meshed via Traffic Engineering (TE) tunnels. When a PBR looks up, in its global table, the C-root of a particular multicast flow, it may find that the next hop interface is a particular TE tunnel. If it can determine the identify of the router at the other end of that TE tunnel, it can deduce that that router is the Upstream PBR for that flow.

This is not an exhaustive set of examples. Any procedure that correctly determines the Upstream PBR in a given deployment scenario MAY be used in that scenario.

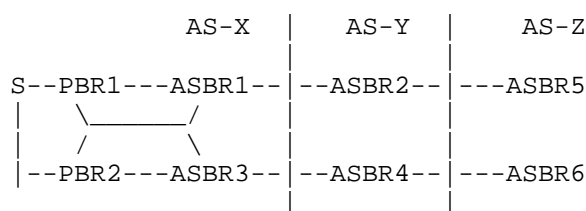
2.3.4. Why SFS Does Not Apply to GTM

To see why the SFS procedure cannot be applied to GTM, consider the following example scenario. Suppose some multicast source S is homed to both PBR1 and PBR2, and suppose that both PBRs export a route (of SAFI 1, 2, or 4) whose NLRI is a prefix matching the address of S. These two routes will be considered comparable by the BGP decision process. A route reflector receiving both routes may thus choose to redistribute just one of the routes to S, the one chosen by the bestpath algorithm. Different route reflectors may even choose different routes to redistribute (i.e., one route reflector may choose the route to S via PBR1 as the bestpath, while another chooses the route to S via PBR2 as the bestpath). As a result, some PBRs may receive only the route to S via PBR1 and some may receive only the route to S via PBR2. In that case, it is impossible to ensure that all PBRs will choose the same route to S.

The SFS procedure works in VPN context as along the following assumption holds: if S is homed to VRF-x in PE1 and to VRF-y in PE2, then VRF-x and VRF-y have been configured with different RDs. In VPN context, the route to S is of SAFI 128 or 129, and thus has an RD in its NLRI. So the route to S via PE1 will not have the same NLRI as the route to S via PE2. As a result, all PEs will see both routes, and the PEs can implement a procedure that ensures that they all pick the same route to S.

That is, the SFS procedure of [RFC6513] relies on the UMH-eligible routes being of SAFI 128 or 129, and relies on certain VRFs being configured with distinct RDs. Thus the procedure cannot be applied to GTM.

One might think that the SFS procedure could be applied to GTM as long as the procedures defined in [ADD-PATH] are applied to the UMH-eligible routes. Using the [ADD-PATH] procedures, the BGP speakers could advertise more than one path to a given prefix. Typically [ADD-PATH] is used to report the n best paths, for some small value of n. However, this is not sufficient to support SFS, as can be seen by examining the following scenario.



In AS-X, PBR1 reports to both ASBR1 and ASBR3 that it has a route to S. Similarly, PBR2 reports to both ASBR1 and ASBR3 that it has a route to S. Using [ADD-PATH], ASBR1 reports both routes to ASBR2, and ASBR3 reports both routes to ASBR4. Now AS-Y sees 4 paths to S. The AS-Z ASBRs will each see eight paths (four via ASBR2 and four via ASBR4). To avoid this explosion in the number of paths, a BGP speaker that uses [ADD-PATH] is usually considered to report only the n best paths. However, there is then no guarantee that the reported set of paths will contain at least one path via PBR1 and at least one path via PBR2. Without such a guarantee, the SFS procedure will not work.

2.4. Inclusive and Selective Tunnels

The MVPN specifications allow multicast flows to be carried on either Inclusive Tunnels or on Selective Tunnels. When a flow is sent on an Inclusive Tunnel of a particular VPN, it is sent to all PEs in that VPN. When sent on a Selective Tunnel of a particular VPN, it may be sent to only a subset of the PEs in that VPN.

This document allows the use of either Inclusive Tunnels or Selective Tunnels for GTM. However, any service provider electing to use Inclusive Tunnels for GTM should carefully consider whether sending a multicast flow to ALL its PBRs would result in problems of scale. There are potentially many more MBRs for GTM than PEs for a particular VPN. If the set of PBRs is large and growing, but most multicast flows do not need to go to all the PBRs, the exclusive use of Selective Tunnels may be a better option.

2.5. I-PMSI A-D Routes

2.5.1. Intra-AS I-PMSI A-D Routes

Per [MVPN-BGP], there are certain conditions under which is it NOT required for a PE router implementing MVPN to originate one or more Intra-AS I-PMSI A-D routes. These conditions apply as well to PBRs implementing GTM.

In addition, a PBR implementing GTM is NOT required to originate an Intra-AS I-PMSI A-D route if both of the following conditions hold:

- The PBR is not using Inclusive Tunnels for GTM, and

- The distribution of the C-multicast Shared Tree Join and C-multicast Source Tree Join routes is done in such a manner that the next hop of those routes does not change.

Please see also the sections on RD and RT usage.

2.5.2. Inter-AS I-PMSI A-D Routes

There are no GTM-specific procedures for the origination, distribution, and processing of these routes, other than those specified in the sections on RD and RT usage.

2.6. S-PMSI A-D Routes

There are no GTM-specific procedures for the origination, distribution, and processing of these routes, other than those specified in the sections on RD and RT usage.

2.7. Leaf A-D Routes

There are no GTM-specific procedures for the origination, distribution, and processing of these routes, other than those specified in the sections on RD and RT usage.

2.8. Source Active A-D Routes

Please see the sections on RD and RT usage for information applies to the origination and distribution of Source Active A-D routes. Additional procedures governing the use of Source Active A-D routes are given in the sub-sections of this section.

2.8.1. Finding the Originator of an SA A-D Route

To carry out the procedures specified in [RFC6514] (e.g., in Section 13.2 of that document), it is sometimes necessary for an egress PE to determine the ingress PE that originated a given Source Active A-D route. The procedure used in [RFC6514] to find the originator of a Source Active A-D route assumes that no two routes have the same RD unless they have been originated by the same PE. However, this assumption is not valid in GTM, because each Source Active A-D route used for GTM will have an RD of 0, and all the UMH-eligible routes also have an RD of 0. So GTM requires a different procedure for determining the originator of a Source Active A-D route.

In GTM, the procedure for determining the originating PE of a Source Active A-D route is the following:

- When a Source Active A-D route is originated, the originating PE MAY attach a VRF Route Import Extended Community to the route.
- When a Source Active A-D route is distributed by one BGP speaker to another, then
 - * if the Source Active A-D route does not carry the VRF Route Import EC, the BGP speaker distributing the route MUST NOT change the route's next hop field;
 - * if the Source Active A-D route does carry the VRF Route Import EC, the BGP speaker distributing the route MAY change the route's next hop field to itself.
- When an egress PE needs to determine the originator of a Source Active A-D route, then
 - * if the Source Active A-D route carries the VRF Route Import EC, the originating PE is the PE identified in the Global Administrator field of that EC;
 - * if the Source Active A-D route does not carry the VRF Route Import EC, the originating PE is the PE identified in the route's next hop field.

2.8.2. Optional Additional Constraints on Distribution

If some site has receivers for a particular ASM group G, then it is possible (by the procedures of [RFC6514]) that every PBR attached to a site with a source for group G will originate a Source Active A-D route whose NLRI identifies that source and group. These Source Active A-D routes may be distributed to every PBR. If only a relatively small number of PBRs are actually interested in traffic from group G, but there are many sources for group G, this could result in a large number of (S,G) Source Active A-D routes being installed in a large number of PBRs that have no need of them.

For GTM, it is possible to constrain the distribution of (S,G) Source Active A-D routes to those PBRs that are interested in GTM traffic to group G. This can be done using the following OPTIONAL procedures:

- If a PBR originates a C-multicast Shared Tree Join whose NLRI contains (RD=0,*,G), then it dynamically creates an import RT for its global table, where the Global Administrator field of the RT contains the group address G, and the Local Administrator field contains zero. (Note that an IPv6-address-specific RT would need to be used if the group address is an IPv6 address.)
- When a PBR creates such an import RT, it uses "RT Constraint" [RFC4684] procedures to advertise its interest in routes that carry this RT.
- When a PBR originates a Source Active A-D route from its global table, it attaches the RT described above.
- When the C-multicast Shared Tree Join is withdrawn, so is the corresponding RT constrain route, and the corresponding RT is removed as an import RT of its global table.

These procedures enable a PBR to automatically filter all Source Active A-D routes that are about multicast groups in which the PBR has no interest.

This procedure does introduce the overhead of distributing additional "RT Constraint" routes, and therefore may not be cost-effective in all scenarios, especially if the number of sources per ASM group is small. This procedure may also result in increased join latency.

2.9. C-multicast Source/Shared Tree Joins

[RFC6514] section 11.1.3 has the following procedure for determining the IP-address-specific RT that is attached to a C-multicast route: (a) determine the upstream PE, RD, AS, (b) find the proper Inter-AS or Intra-AS I-PMSI A-D route based on (a), (c) find the next hop of that A-D route, (d) base the RT on that next hop.

However, for GTM, in environments where it is known a priori that that the next hop of the C-multicast Source/Shared Tree Joins does not change during the distribution of those routes, the proper procedure for creating the IP-address-specific RT is to just put the IP Address of the Upstream PBR in the Global Administrator field of the RT. In other scenarios, the procedure of the previous paragraph (as modified by this document's sections on "RD usage" and "RT usage") is applied by the PBRs.

3. Differences from other MVPN-like GTM Procedures

The document [SEAMLESS-MCAST] also defines a procedure for GTM that is based on the BGP procedures that were developed for MVPN.

However, the GTM procedures of [SEAMLESS-MCAST] are different than and are NOT interoperable with the procedures defined in this document.

The two sets of procedures can co-exist in the same network, as long as they are not applied to the same multicast flows or to the same ASM multicast group addresses.

Some of the major differences between the two sets of procedures are the following;

- The [SEAMLESS-MCAST] procedures for GTM do not use C-multicast Shared Tree Joins or C-multicast Source Tree Joins at all. The procedures of this document use these C-multicast routes for GTM, setting the RD field of the NLRI to zero.
- The [SEAMLESS-MCAST] procedures for GTM use Leaf A-D routes instead of C-multicast Shared/Source Tree Join routes. Leaf A-D routes used in that manner can be distinguished from Leaf A-D routes used as specified in [RFC6514] by means of the NLRI format; [SEAMLESS-MCAST] defines a new NLRI format for Leaf A-D routes. Whether a given Leaf A-D route is being used according to the [SEAMLESS-MCAST] procedures or not can be determined from its NLRI. (See [SEAMLESS-MCAST] section "Leaf A-D Route for Global Table Multicast".)
- The Leaf A-D routes used by the current document contain an NLRI that is in the format defined in [RFC6514], NOT in the format as defined in [SEAMLESS-MCAST]. The procedures assumed by this document for originating and processing Leaf A-D routes are as specified in [RFC6514], NOT as specified in [SEAMLESS-MCAST].
- The current document uses an RD value of zero in the NLRI in order to indicate that a particular route is "about" a Global Table Multicast, rather than a VPN multicast. No other semantics are inferred from the fact that RD is zero. [SEAMLESS-MCAST] uses two different RD values in its GTM procedures, with semantic differences that depend upon the RD values.
- In order for both sets of procedures to co-exist in the same network, the PBRs MUST be provisioned so that for any given IP group address in the global table, all egress PBRs use the same set of procedures for that group address (i.e., for group G,

either all egress PBRs use the GTM procedures of this document or all egress PBRs use the GTM procedures of [SEAMLESS-MCAST].

4. IANA Considerations

This document has no IANA considerations.

5. Security Considerations

The security considerations of this document are primarily the security considerations of the base protocols, as discussed in [RFC6514], [RFC4601], and [RFC5294].

This document makes use of a BGP SAFI (MCAST-VPN routes) that was originally designed for use in VPN contexts only. It also makes use of various BGP path attributes and extended communities (VRF Route Import Extended Community, Source AS Extended Community, Route Target Extended Community) that were originally intended for use in VPN contexts. If these routes and/or attributes leak out into "the wild", multicast data flows may be distributed in an unintended and/or unauthorized manner.

Internet providers often make extensive use of BGP communities (ie, adding, deleting, modifying communities throughout a network). As such, care should be taken to avoid deleting or modifying the VRF Route Import Extended Community and Source AS Extended Community. Incorrect manipulation of these ECs may result in multicast streams being lost or misrouted.

The procedures of this document require certain BGP routes to carry IP multicast group addresses. Generally such group addresses are only valid within a certain scope. If a BGP route containing a group address is distributed outside the boundaries where the group address is meaningful, unauthorized distribution of multicast data flows may occur.

6. Additional Contributors

Zhenbin Li
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China
Email: lizhenbin@huawei.com

Wei Meng
ZTE Corporation
No.50 Software Avenue, Yuhuatai District
Nanjing
China
Email: meng.wei2@zte.com.cn, vally.meng@gmail.com

Cui Wang
ZTE Corporation
No.50 Software Avenue, Yuhuatai District
Nanjing
China
Email: wang.cuil@zte.com.cn

Shunwan Zhuang
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China
Email: zhuangshunwan@huawei.com

7. Acknowledgments

The authors and contributors would like to thank Rahul Aggarwal, Huajin Jeng, Hui Ni, Yakov Rekhter, and Samir Saad for their contributions to this work.

8. Authors' Addresses

Lenny Giuliano
Juniper Networks
2251 Corporate Park Drive
Herndon, VA 20171
US
Email: lenny@juniper.net

Dante J. Pacella
Verizon
Verizon Communications
22001 Loudoun County Parkway
Ashburn, VA 20147
US
Email: dante.j.pacella@verizonbusiness.com

Eric C. Rosen
Cisco Systems, Inc.
1414 Massachusetts Avenue
Boxborough, MA, 01719
US
Email: erosen@cisco.com

Jason Schiller
Google
1818 Library Street
Suite 400
Reston, VA 20190
US
Email: jschiller@google.com

Karthik Subramanian
Cisco Systems, Inc.
170 Tasman Drive
San Jose, CA, 95134
US
Email: kartsubr@cisco.com

Jeffrey Zhang
Juniper Networks
10 Technology Park Dr.
Westford, MA 01886
US
Email: zzhang@juniper.net

9. References

9.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4364], Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks", RFC 4364, February 2006.
- [RFC6513] Rosen, E. and R. Aggarwal, "Multicast in MPLS/BGP IP VPNs", RFC 6513, February 2012.
- [RFC6514] Aggarwal, R., Rosen, E., Morin, T., and Y. Rekhter, "BGP Encodings and Procedures for Multicast in MPLS/BGP IP VPNs", RFC 6514, February 2012.
- [RFC6515] Aggarwal, R., and E. Rosen, "IPv4 and IPv6 Infrastructure Addresses in BGP Updates for Multicast VPN", RFC 6515, February 2012.

9.2. Informative References

- [ADD-PATH] "Advertisement of Multiple Paths in BGP", D. Walton, A. Retana, E. Chen, J. Scudder, draft-ietf-idr-add-paths-09.txt, October 2013.
- [RFC6368] Marques, P., Raszuk, R., Patel, K., Kumaki, K., and T. Yamagata, "Internal BGP as the Provider/Customer Edge Protocol for BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 6368, September 2011.
- [RFC4601] Fenner, B., Handley, M., Holbrook, H., and I. Kouvelas, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", RFC 4601, August 2006.
- [RFC4684] P. Marques, et. al., "Constrained Route Distribution for Border Gateway Protocol/MultiProtocol Label Switching (BGP/MPLS) Internet Protocol (IP) Virtual Private Networks (VPNs)", RFC 4684,

November 2006.

[RFC5294] Savola, P. and J. Lingard, "Host Threats to Protocol Independent Multicast (PIM)", RFC 5294, August 2008.

[MVPN-extranet] Rekhter, Y. and E. Rosen (editors), "Extranet Multicast in BGP/IP MPLS VPNs", draft-ietf-l3vpn-mvpn-extranet-04.txt, March 2014

[SEAMLESS-MCAST] Rekhter, Y., Aggarwal, R., Morin, T., Grosclaude, I., Leymann, N., and S. Saad, "Inter-Area P2MP Segmented LSPs", draft-ietf-mpls-seamless-mcast-09.txt, December 2013