

6man Working Group  
Internet-Draft  
Updates: 3306,3956,4607,4291 (if approved)  
Intended status: Standards Track  
Expires: November 24, 2013

M. Boucadair  
France Telecom  
S. Venaas  
Cisco  
May 23, 2013

Updates to the IPv6 Multicast Addressing Architecture  
draft-ietf-6man-multicast-addr-arch-update-01

Abstract

This document updates the IPv6 multicast addressing architecture by defining the 17-20 reserved bits as generic flag bits. The document provides also some clarifications related to the use of these flag bits.

This document updates RFC 3956, RFC 3306, RFC 4607 and RFC 4291.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

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 November 24, 2013.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. 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. Addressing Architecture Update . . . . .	2
3. Clarifications . . . . .	3
3.1. Flag Bits . . . . .	3
3.2. IANA Assigned SSM Block . . . . .	4
4. RFC Updates . . . . .	4
4.1. RFC3306 . . . . .	4
4.2. RFC3956 . . . . .	6
4.3. RFC4607 . . . . .	8
5. IANA Considerations . . . . .	9
6. Security Considerations . . . . .	9
7. Acknowledgements . . . . .	9
8. Normative References . . . . .	9
Authors' Addresses . . . . .	10

## 1. Introduction

This document updates the IPv6 multicast addressing architecture [RFC4291] by defining the 17-20 reserved bits as generic flag bits (Section 2). The document provides also some clarifications related to the use of these flag bits (Section 3.1) and also about IANA assigned SSM blocks (Section 3.2).

This document updates [RFC3956], [RFC3306], [RFC4607] and [RFC4291].

## 2. Addressing Architecture Update

Bits 17-20 of a multicast address are defined in [RFC3956] and [RFC3306] as reserved bits. This document defines these bits as generic flag bits so that they apply to any multicast address. Figure 1 and Figure 2 show the updated structure of the addressing architecture. The first diagram shows the update of the base IPv6 addressing architecture, and the second shows the update of so-called Embedded-RP.

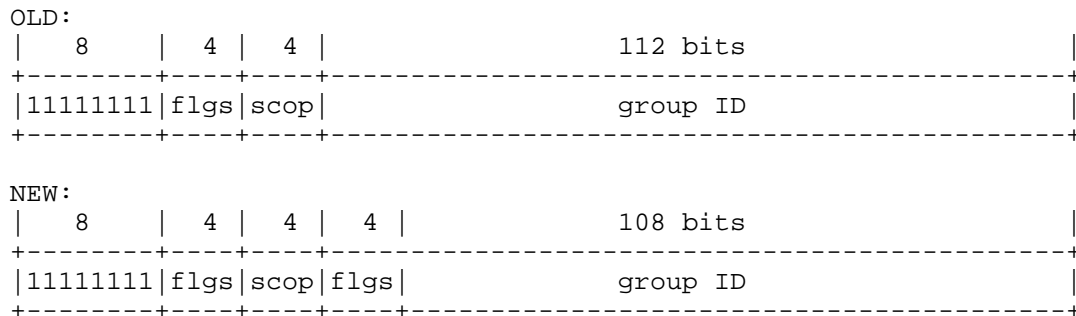


Figure 1: Updated IPv6 Multicast Addressing Architecture

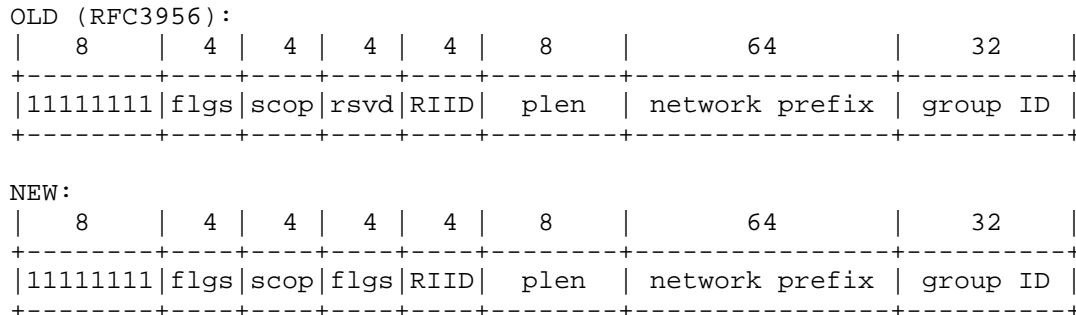


Figure 2: Embedded-RP with Updated IPv6 Multicast Address Arch.

Further specification documents may define a meaning for these flag bits. Defining the bits 17-20 as flags for all IPv6 multicast addresses allows addresses to be treated in a more uniform and generic way, and allows for these bits to be defined in the future for different purposes, irrespective of the specific type of multicast address.

### 3. Clarifications

#### 3.1. Flag Bits

Some implementations and specification documents do not treat the flag bits as separate bits but tend to use their combined value as a 4-bit integer. This practice is a hurdle for assigning a meaning to the remaining flag bits. Below are listed some examples for illustration purposes:

- o the reading of [RFC4607] may lead to conclude that ff3x::/32 is the only allowed SSM IPv6 prefix block.

- o [RFC3956] states only ff70::/12 applies to Embedded-RP. Particularly, implementations should not treat the fff0::/12 range as Embedded-RP.

To avoid such confusion and to unambiguously associate a meaning with the remaining flags, the following recommendation is made

Implementations MUST treat flag bits as separate bits.

### 3.2. IANA Assigned SSM Block

Another issue related to SSM is the IANA assigned SSM address block. Per [RFC4607], ff3x::4000:0001 through ff3x::7fff:fff is the block for IANA assignments (<http://www.iana.org/assignments/ipv6-multicast-addresses/ipv6-multicast-addresses.xml>). However, IANA assignments are permanent addresses and should not have the transient bit set. Quoting from [RFC4607]:

"T = 1 indicates a non-permanently-assigned ("transient") multicast address."

## 4. RFC Updates

### 4.1. RFC3306

This document changes Section 4 of [RFC3306] as follows:

OLD:

8	4	4	8	8	64	32	
-----	-----	-----	-----	-----	-----	-----	-----
11111111	flgs	scop	reserved	plen	network prefix	group ID	
-----	-----	-----	-----	-----	-----	-----	-----

flgs is a set of 4 flags:

+-+-+--+
0 0 P T
+-+-+--+

- o P = 0 indicates a multicast address that is not assigned based on the network prefix. This indicates a multicast address as defined in [ADDRARCH].
- o P = 1 indicates a multicast address that is assigned based on the network prefix.
- o If P = 1, T MUST be set to 1, otherwise the setting of the T bit is defined in Section 2.7 of [ADDRARCH].

The reserved field MUST be zero.

NEW:

	8		4		4		8		8		64		32	
+	-----	+	-----	+	-----	+	-----	+	-----	+	-----	+	-----	+
	11111111		flgs		scop		reserved		plen		network prefix		group ID	
+	-----	+	-----	+	-----	+	-----	+	-----	+	-----	+	-----	+

flgs is a set of 4 flags:                   +--+--+--+  
                                   |X|Y|P|T|  
                                   +--+--+--+

X and Y may each be set to 0 or 1.

- o P = 0 indicates a multicast address that is not assigned based on the network prefix. This indicates a multicast address as defined in [ADDRARCH].
- o P = 1 indicates a multicast address that is assigned based on the network prefix.
- o T is set according to the definition in Section 2.7 of [ADDRARCH]. Unicast-Prefix-based addresses would typically not be IANA assigned, so in most cases T would be set to 1.

This document changes Section 6 of [RFC3306] as follows:

OLD:

These settings create an SSM range of FF3x::/32 (where 'x' is any valid scope value). The source address field in the IPv6 header identifies the owner of the multicast address.

NEW:

T flag is set according to whether the addresses are assigned by IANA.

If the flag bits are to 0011, these settings create an SSM range of ff3x::/32 (where 'x' is any valid scope value). The source address field in the IPv6 header identifies the owner of the multicast address. ff3x::/32 is not the only allowed SSM prefix range. For example, ff2x::/32 would be IANA assigned SSM addresses.

#### 4.2. RFC3956

This document changes Section 2 of [RFC3956] as follows:

OLD:

As described in [RFC3306], the multicast address format is as follows:

	8		4		4		8		8		64		32	
+	-----	+	-----	+	-----	+	-----	+	-----	+	-----	+	-----	+
	11111111		flgs		scop		reserved		plen		network prefix		group ID	
+	-----	+	-----	+	-----	+	-----	+	-----	+	-----	+	-----	+

Where flgs are "0011". (The first two bits are as yet undefined, sent as zero and ignored on receipt.)

NEW:

As described in [RFC3306], the multicast address format is as follows:

	8		4		4		4		4		8		64		32	
+	-----	+	-----	+	-----	+	-----	+	-----	+	-----	+	-----	+	-----	+
	11111111		flgs		scop		flgs		rsvd		plen		network prefix		group ID	
+	-----	+	-----	+	-----	+	-----	+	-----	+	-----	+	-----	+	-----	+

flgs is a set of four flags:

+	-	+	-	+	-	+	-	+
	X		R		P		T	
+	-	+	-	+	-	+	-	+

X may be set to 0 or 1.

This document changes Section 3 of [RFC3956] as follows:

OLD:

	8		4		4		4		4		8		64		32	
+	-----	+	-----	+	-----	+	-----	+	-----	+	-----	+	-----	+	-----	+

```

|11111111|flgs|scop|rsvd|RIID|plen| network prefix | group ID |
+-----+-----+-----+-----+-----+-----+-----+
                                     +-+--+--+
flgs is a set of four flags:      |0|R|P|T|
                                     +--+--+--+

```

When the highest-order bit is 0, R = 1 indicates a multicast address that embeds the address on the RP. Then P MUST be set to 1, and consequently T MUST be set to 1, as specified in [RFC3306]. In effect, this implies the prefix FF70::/12. In this case, the last 4 bits of the previously reserved field are interpreted as embedding the RP interface ID, as specified in this memo.

The behavior is unspecified if P or T is not set to 1, as then the prefix would not be FF70::/12. Likewise, the encoding and the protocol mode used when the two high-order bits in "flgs" are set to 11 ("FFF0::/12") is intentionally unspecified until such time that the highest-order bit is defined. Without further IETF specification, implementations SHOULD NOT treat the FFF0::/12 range as Embedded-RP.

#### NEW:

```

| 8 | 4 | 4 | 4 | 4 | 8 | 64 | 32 |
+---+---+---+---+---+---+---+---+
|11111111|flgs|scop|flgs|RIID|plen| network prefix | group ID |
+---+---+---+---+---+---+---+---+
                                     +-+--+--+
flgs is a set of four flags:      |X|R|P|T|
                                     +--+--+--+

```

X may be set to 0 or 1.

R = 1 indicates a multicast address that embeds the address of the RP. P MUST be set to 1 according to [RFC3306], as this is a special case of unicast-prefix based addresses. This implies that for instance prefixes ff70::/12 and fff0::/12 are embedded RP prefixes, but all multicast addresses with the R-bit set to 1 MUST be treated as Embedded RP addresses. The behavior is unspecified if P is not set to 1. When the R-bit is set, the last 4 bits of the previously reserved field are interpreted as embedding the RP interface ID, as specified in this memo.

This document changes Section 4 of [RFC3956] as follows:

#### OLD:

It MUST be a multicast address with "flgs" set to 0111, that is, to be of the prefix FF70::/12,

NEW:

It MUST be a multicast address with R-bit set to 1.

It MUST have P-bit set to 1 when using the embedding in this document as it is a prefix-based address.

This document changes Section 7.1 of [RFC3956] as follows:

OLD:

To avoid loops and inconsistencies, for addresses in the range FF70::/12, the Embedded-RP mapping MUST be considered the longest possible match and higher priority than any other mechanism.

NEW:

To avoid loops and inconsistencies, for addresses with R-bit set to 1, the Embedded-RP mapping MUST be considered the longest possible match and higher priority than any other mechanism.

#### 4.3. RFC4607

This document changes the abstract of [RFC4607] as follows:

OLD:

IP version 4 (IPv4) addresses in the 232/8 (232.0.0.0 to 232.255.255.255) range are designated as source-specific multicast (SSM) destination addresses and are reserved for use by source-specific applications and protocols. For IP version 6 (IPv6), the address prefix FF3x::/32 is currently reserved for source-specific multicast use but others may be reserved in the future. This document defines an extension to the Internet network service that applies to datagrams sent to SSM addresses and defines the host and router requirements to support this extension.

NEW:

IP version 4 (IPv4) addresses in the 232/8 (232.0.0.0 to 232.255.255.255) range are designated as source-specific multicast (SSM) destination addresses and are reserved for use by source-specific applications and protocols. For IP version 6 (IPv6), the address prefix ff3x::/32 is currently reserved for source-specific multicast use but others may be reserved in the future. This



document defines an extension to the Internet network service that applies to datagrams sent to SSM addresses and defines the host and router requirements to support this extension.

This document changes Section 1 of [RFC4607] as follows:

OLD:

For IPv6, the address prefix FF3x::/32 is reserved for source-specific multicast use, where 'x' is any valid scope identifier, by [IPv6-UBM]. Using the terminology of [IPv6-UBM], all SSM addresses must have P=1, T=1, and plen=0. [IPv6-MALLOC] mandates that the network prefix field of an SSM address also be set to zero, hence all SSM addresses fall in the FF3x::/96 range. Future documents may allow a non-zero network prefix field if, for instance, a new IP- address-to-MAC-address mapping is defined. Thus, address allocation should occur within the FF3x::/96 range, but a system should treat all of FF3x::/32 as SSM addresses, to allow for compatibility with possible future uses of the network prefix field.

NEW:

For IPv6, all SSM addresses must have P=1 and plen=0 while T-bit is set according to whether the addresses are assigned by IANA [I-D.ietf-6man-multicast-addr-arch-update]. In particular, a system should treat all of ff3x::/32 and ff2x::/32 as SSM addresses, to allow for compatibility with possible future uses of the network prefix field. Other SSM prefixes can be defined in the future.

## 5. IANA Considerations

This document may require IANA updates. However, at this point it is not clear exactly what these updates may be.

## 6. Security Considerations

Security considerations discussed in [RFC3956], [RFC3306], [RFC4607] and [RFC4291] MUST be taken into account.

## 7. Acknowledgements

Many thanks to B. Haberman for the discussions prior to the publication of this document.

## 8. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC3306] Haberman, B. and D. Thaler, "Unicast-Prefix-based IPv6 Multicast Addresses", RFC 3306, August 2002.
- [RFC3956] Savola, P. and B. Haberman, "Embedding the Rendezvous Point (RP) Address in an IPv6 Multicast Address", RFC 3956, November 2004.
- [RFC4291] Hinden, R. and S. Deering, "IP Version 6 Addressing Architecture", RFC 4291, February 2006.
- [RFC4607] Holbrook, H. and B. Cain, "Source-Specific Multicast for IP", RFC 4607, August 2006.

#### Authors' Addresses

Mohamed Boucadair  
France Telecom  
Rennes 35000  
France

Email: mohamed.boucadair@orange.com

Stig Venaas  
Cisco  
USA

Email: stig@cisco.com

L3VPN  
Internet-Draft  
Intended status: Standards Track  
Expires: January 14, 2014

R. Kebler  
P. Kurapati  
Juniper Networks  
July 13, 2013

Multicast Traceroute for MVPNs  
draft-kebler-kurapati-l3vpn-mvpn-mtrace-00

Abstract

Mtrace is a tool used to troubleshoot issues in a network deploying Multicast service. When multicast is used within a VPN service offering, the base Mtrace specification does not detect the failures. This document specifies a method of using multicast traceroute in a network offering Multicast in VPN service.

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 January 14, 2014.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. 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.

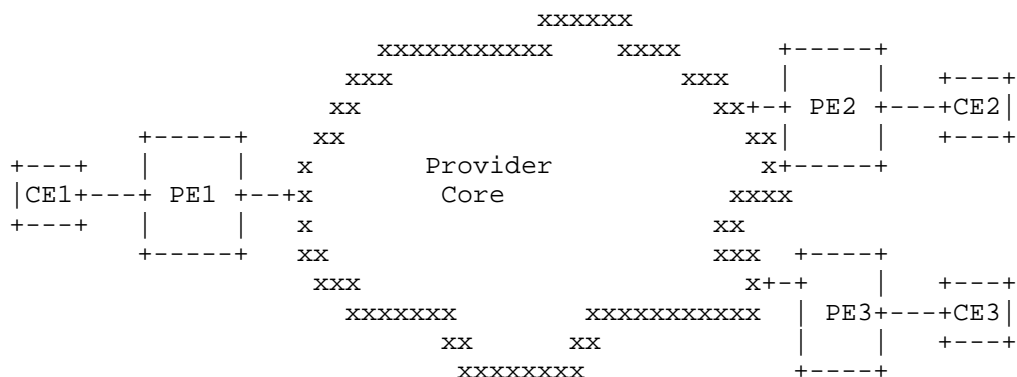
## Table of Contents

1. Introduction . . . . .	2
2. Overview . . . . .	3
3. Protocol Details . . . . .	4
3.1. Mtrace Query . . . . .	4
3.2. Mtrace Request . . . . .	4
3.2.1. Ingress PE Procedures . . . . .	6
3.3. Downstream Requests . . . . .	7
3.4. ASBR Behavior . . . . .	7
3.5. Virtual Hub and Spoke . . . . .	8
3.6. Inter-Area Provider Tunnels . . . . .	8
3.6.1. Egress PE . . . . .	8
3.6.2. ABR Behavior . . . . .	9
3.7. Mtrace MVPN Procedure . . . . .	9
4. Error Detection . . . . .	10
4.1. MVPN Error Codes . . . . .	11
5. Mtracev2 Extensions . . . . .	12
5.1. New Mtracev2 TLV Type . . . . .	12
5.2. MVPN Extended Query Block . . . . .	12
5.3. Leaf A-D Augmented Response Block . . . . .	13
5.4. PMSI Tunnel Attributes Augmented Response Block . . . . .	13
6. Mtrace2 Standard Response Block considerations . . . . .	13
7. IANA Considerations . . . . .	14
8. Security Considerations . . . . .	14
9. Acknowledgments . . . . .	14
10. Normative References . . . . .	14
Authors' Addresses . . . . .	15

## 1. Introduction

The current multicast traceroute [I-D.ietf-mboned-mtrace-v2] travels up the tree hop-by-hop towards the source. This verifies the basic multicast state back to the source, but is not sufficient to verify the MVPN state. The base Mtrace specification assumes that the routers in the path are directly connected through interfaces. In the case of Multicast traffic over VPN service, the PEs who are MVPN neighbors may be separated by several router hops. The path taken by the query can be completely different from the path taken through core by the actual multicast traffic. Consider a case in the below figure, where provider tunnel between PE2 (Source) and PE1 (Receiver) is not established correctly due to incorrect MVPN state on PE2. In the current form of Mtrace, the Query would result in a successful response since there is no error detection mechanism for MVPN state available currently. Even if one can infer from the statistics of the Mtrace Response that PE2 has an issue, the existing error codes are not sufficient to identify the root cause. Also, there could be a problem sending traffic over the provider tunnel from PE2 to PE1,

but the mtrace query will not even travel over this provider tunnel. Therefore, the mtrace successful response can be misleading. This draft ensures that the Response uses same provider-tunnel that the given C-S,C-G data would traverse and returns appropriate MVPN specific error codes which would help in identifying the root cause.



MVPN topology

## 2. Overview

As described in the Mtracev2 specification [I-D.ietf-mboned-mtrace-v2], a Querier initiates an Mtrace Query which is sent to the Last Hop Router. Last Hop Router converts this into a Request and sends it towards the First Hop Router. This draft introduces a new "Downstream Request" mechanism to allow the First Hop Router to send the mtrace request message back on the Provider tunnel to the Last hop router. The last hop router will then change it to Response and send it to the Querier who initiated the Query. If there is any error encountered by the Last hop router or First Hop router, a Response is directly unicasted to the querier with appropriate MVPN specific error codes added. Each hop in the path of Mtrace decrements the TTL value before sending the mtrace message.

Since the Mtrace is being extended for MVPNs, the Last Hop router and First Hop router SHOULD be a Provider Edge (PE) router so that the MVPN specific error codes can be contained within the provider space. The Request will be initiated by the egress PE and will travel upstream to the ingress PE. It is assumed that the Querier knows and can reach the egress PE. A Querier and egress PE can be the same router.

For Mtrace initiated by the CEs, the specification mentioned in Mtracev2 [I-D.ietf-mboned-mtrace-v2] SHOULD be followed. If a Mtrace message is received by the PE on CE facing interfaces containing MVPN specific extensions defined in this draft, it SHOULD be discarded.

### 3. Protocol Details

The protocol details that follow are described in terms of mtracev2. However, the same procedures can be achieved with mtracev1. The protocol extensions needed for mtracev2 are described in Section 5 and the protocol extensions for mtracev1 and described in section 6.

#### 3.1. Mtrace Query

A Querier willing to perform a Mtrace on a MVPN issues a Mtrace Query. The format of the Query TLV is as specified in the Mtracev2 specification [I-D.ietf-mboned-mtrace-v2]. The (C-S,C-G) to be queried is populated in the source address and group address fields of the Mtrace2 Query block. A deployment may use wild card SPMSIs as defined in [RFC6625]. For example, a (C-\*,C-\*) wild card SPMSI or a (C-\*,ALL-PIM-ROUTERS) can be used to send messages like BSR across PEs as mentioned in section 5.3.4 MVPN specification [RFC6513]. A querier may be interested in knowing the health of such a SPMSI tunnel. In this case, the Multicast Address and Source Address fields of the Mtrace2 query can be filled with wild cards (all 1s) accordingly by the querier.

The Querier MUST add a MVPN Extended Query Block to include the RD of the C-S,C-G that it wishes to trace. When wild card SPMSIs are used, a PE could have subscribed to multiple upstream PEs for wild card SPMSIs. Hence, a query for a wild card SPMSI MUST also specify the upstream PE address that it is interested to query. The upstream PE address in the MVPN Extended Query Block MUST be filled only for wild card queries. For a regular (C-S,C-G) query, this field SHOULD be set to 0s by the querier and is ignored by the receivers.

This Query is sent to the Downstream PE (Last Hop Router) to initiate the mtrace towards the source. If a Querier does not receive a Response, it can retry sending Query messages with increasing TTL values to help diagnose where the Mtrace messages are being lost.

#### 3.2. Mtrace Request

The PE that receives the query will lookup the (C-S,C-G) using the RD of the query to distinguish the vrf. If the RD doesn't match any VRF, PE sends a response with error code set to BAD\_RD. The PE first checks the C-Mcast route that is matching (C-S,C-G) of the mtrace Query. It then finds the upstream multicast hop from the selected

C-Mcast route and unicasts the requests to the upstream multicast hop after decrementing the TTL. The Mtrace Request MUST have PMSI Tunnel Attributes Augmented Response Block populated with the PMSI attribute that the PE uses to receive the traffic for the given (C-S,C-G) traffic.

Upstream multicast hop can be same as upstream PE router in some cases, while it can be the ASBR or the BGP nexthop of the selected C-Mcast route in Inter-AS scenarios. The procedures for finding upstream multicast hop is discussed in detail under section 5.1 of MVPN specification [RFC6513].

When a wild card query is received, the PE will look for the upstream PE address in the MVPN Extended query block. The PE will then check if it has bound to the wild card SPMSI tunnel from the specified upstream PE. If it has, it will populate the Leaf A-D Augmented Response Block and PMSI Tunnel Attributes Augmented Response Block with the respective values. If the PE has not received any wild card SPMSI AD route from the specified upstream PE in the query, it should send a response with the error code set to NO\_WILD\_CARD\_SPMSI\_AD\_RCVD. If the PE has received wild card SPMSI AD route from the upstream PE, but has not responded with a LEAF-AD route, it should send a response with the error code set to NO\_WILD\_CARD\_SPMSI\_LEAF\_AD\_SENT.

For a non-wild-card query, the upstream PE address field in the MVPN Extended query block MUST be ignored by the PEs. It MUST follow the procedure to find the upstream multicast hop as discussed earlier.

If the route does not match any MVPN-TIB state, then the PE should send a Response to the Querier with the error code set to NO\_CMCAST\_STATE. If the PE cannot locate the upstream PE then it should send a response to the Querier with the NO\_UPSTREAM\_PE error code.

From the selected UMH route, the local PE extracts the ASN of the upstream PE (as carried in the Source AS Extended Community of the route), and the source-AS field of the mtrace Query is set to that AS.

If the local and the upstream PEs are in the same AS, then the RD in the mtrace Query is set to the RD of the VPN-IP route for the source/ RP.

Section 8 of MVPN specification [RFC6513] mentions two procedures (Segmented and Non-Segmented) for handling Inter-AS scenarios. If the local and the upstream PEs are in different ASes, and if segmented Inter-AS procedure is used, then the local PE finds in its

VRF an Inter-AS I-PMSI A-D route whose Source AS field carries the ASN of the upstream PE. The RD of the found Inter-AS I-PMSI A-D route is used as the RD of the mtrace Query. If Inter-AS I-PMSI A-D route is not found, a response with error code UNKNOWN\_INTER\_AS is sent.

To support non-segmented inter-AS tunnels, if the local and the upstream PEs are in different ASes, the local system finds in its VRF an Intra-AS I-PMSI A-D route from the upstream PE. The Originating Router's IP Address field of that route has the same value as the one carried in the VRF Route Import of the unicast route to the address carried in the Multicast Source field. The RD of the found Intra-AS I-PMSI A-D route is used as the RD in the mtrace Query. The Source AS field in the mtrace Query is set to value of the Originating Router's IP Address field of the found Intra-AS I-PMSI A-D route.

The PE receiving Mtrace Query will check for any errors. If any error is detected it will send the error back to the Querier. Otherwise, it will change the TLV value to be an Mtrace Request, and it will add a Mtrace2 Standard Response Block. It will also add a PMSI Tunnel Attributes Augmented Response Block with the attributes of the PMSI used to receive traffic for the S,G. If a Leaf-AD route was advertised to the upstream PE for this S,G then the PE will also include a Leaf-AD Augmented Response Block with the NLRI of the associated Leaf-AD route.

### 3.2.1. Ingress PE Procedures

The PE that receives the Request, will check the PMSI attributes of the sender of request to see if they match the values used to send traffic for the S,G. If the values do not match, then the PE uses the appropriate pmsi error code as specified in 'MVPN Error Codes' section and sends a mtrace Response back to the Querier. Also, if a Leaf A-D Augmented Response Block is included, the PE will validate that it has received this Leaf A-D route from the router that sent the Request. If not, then this PE should change the error code to BAD\_LEAF\_AD and send the Response to the Querier. If the PE expects that a Leaf A-D route is needed for the downstream PE to receive traffic, but did not receive one in the mtrace Request from the sending router, then it should use a NO\_LEAF\_AD\_RCVD error code for the mtrace Response. For a wild card SPMSI query, if the PE didn't receive LEAF AD route from the downstream PE, it should use NO\_LEAF\_AD\_RCVD error code.

When the upstream PE receives the Request, it will check for any errors. If there are errors detected, or if the TTL expired, then the PE will change the TLV code to be a Mtrace Response and unicast the response back to the Querier.



The ingress PE will also check it has local vrf connectivity for the source/RP. If it does not have any connectivity to the source/RP then it should use the base specification error code NO\_ROUTE and send an mtrace Response. Note that in a Virtual Hub and Spoke environment, it is possible for a PE to receive a mtrace Request and need to propagate it to another upstream PE. These procedures are outlined in the section "Virtual Hub and Spoke". If the PE does not expect to be receiving mtrace Responses from the mvpn core and have the route to the source located via another upstream PE, then it can use the base specification RPF\_IF error code.

If the PE that receives the Request is the ingress PE that has local vrf connectivity for the source, then it will add a Standard Response Block to the mtrace message. It will not include the additional PMSI Attributes Response Block. Then it will turn the Request into a Downstream Request by changing the value of the Type field of the TLV. It will send the mtrace message on the provider tunnel used to send the S,G data traffic.

### 3.3. Downstream Requests

When a router receives Mtrace Downstream Request, it will determine if it has added any of the Response Blocks for this mtrace message. If it does not locate its address in the list of Response Blocks, then it will silently discard this mtrace message. Otherwise, it will set the 'D' bit in its PMSI Tunnel Attributes Augmented Response Block to indicate that this message has been received on the PMSI tunnel.

If this router is the egress PE that provided the initial Response Block, then it will change the mtrace type to a Reply and sends the Reply to the Querier (the egress PE and the Querier may be the same router). Otherwise, this router must send the Downstream PE on the PMSI that it would normally send traffic for the S,G. Before sending the Downstream Request, the router must decrement the TTL and check for TTL expiry. If the TTL has expired, then this router must send the Response to the Querier with the appropriate code.

### 3.4. ASBR Behavior

When an ASBR receives a mtrace Request the ASBR finds an Inter-AS I-PMSI A-D route whose RD and Source AS matches the RD and Source AS carried in the mtrace Query. If no matching route is found and the ASBR is using segmented tunnels as described in MVPN specification [RFC6513], the ASBR sends an UNKNOWN\_INTER\_AS error code back to the Querier. If a matching route is found, the ASBR acts as a "first hop router" and modifies the Query type to DOWNSTREAM\_REQUEST. ASBR in this case MUST validate the PMSI attributes similar to the "first hop

router" and respond if there is any errors. ASBR MUST populate PMSI Tunnel Attributes Augmented Response Block with the Inter-AS provider tunnel information before sending the DOWNSTREAM\_REQUEST. Note that the mtrace request does not proceed upstream as it is assumed that performing a traceroute and exposing IP addresses across AS boundaries would not be desirable with Segmented Inter-AS Provider Tunnels.

To support non-segmented inter-AS tunnels as described in [RFC6513], instead of matching the RD and Source AS carried in the mtrace Query against the RD and Source AS of an Inter-AS I-PMSI A-D route, the ASBR should match it against the RD and the Originating Router's IP Address of the Intra- AS I-PMSI A-D routes. The Next Hop field of the MP\_REACH\_NLRI of the found Intra-AS I-PMSI A-D route is used as the destination for the mtrace Request.

### 3.5. Virtual Hub and Spoke

When a Virtual-Hub (V-HUB) as described in specification [I-D.ietf-l3vpn-virtual-hub] receives a mtrace Request the S,G may be reachable via one of its vrf interfaces. In this case, the V-HUB is an ingress PE and the procedure are defined in the Section "Ingress PE Procedures". Otherwise, the C-RP/C-S of the route is reachable via some other PE. This is the case where the received route was originated by a Virtual-spoke (V-spoke) that sees the V-HUB as the "upstream PE" for the given source, but the V-HUB sees another PE as the "upstream PE" for that source. In this case, the V-HUB should check the PMSI attributes sent in the mtrace Request against the Tunnel Attributes of the Provider Tunnel used to send traffic for the S,G from the upstream PE to the V-Spoke.

The V-HUB sends a mtrace Request to its upstream PE the same way as it would if it received a mtrace Query. V-HUB MUST add PMSI Tunnel Attributes Augmented Response Block of its own before sending the mtrace Request to the upstream PE. It may also add Leaf-AD Augmented Response Block if a Leaf-AD route was advertised upstream by the V-HUB. If the RD or Source-AS of the upstream PE is different, the V-HUB updates the MVPN Extended Query Block accordingly.

### 3.6. Inter-Area Provider Tunnels

#### 3.6.1. Egress PE

The egress PE does the same procedures as specified in Section "Mtrace Request" except it sends the Request upstream to the IP address determined from the Global Administrator field of the Inter-area P2MP Segmented Next-hop Extended Community as described in specification [I-D.ietf-mpls-seamless-mcast] . If the egress PE has

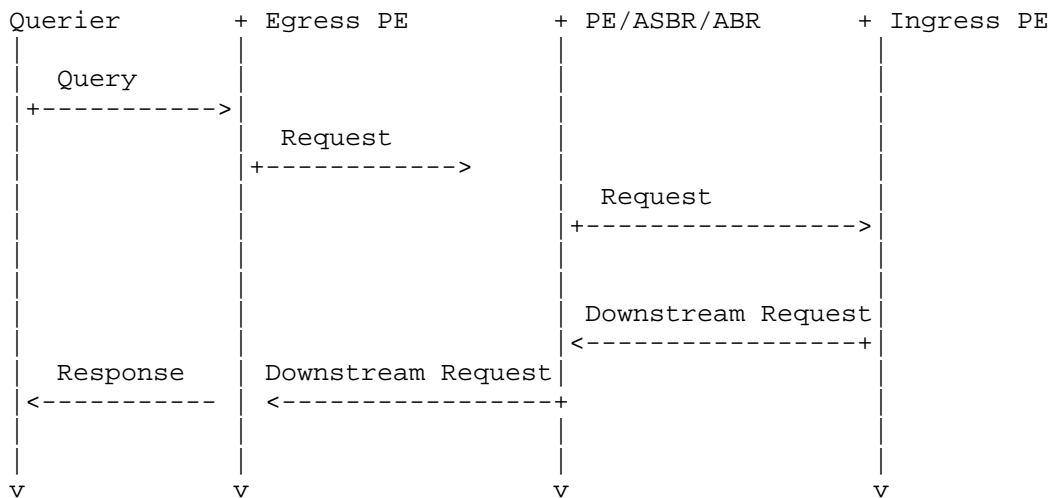
sent a Leaf-AD route then it must send a Leaf-AD Augmented Response Block with the NLRI of the Leaf A-D route.

### 3.6.2. ABR Behavior

ABR MUST find a S-PMSI or I-PMSI route whose NLRI has the same value as the Route Key field of the received mtrace Leaf-AD extended Query Block. If such a matching route is not found then a Response should be sent to the Querier with the NO\_LEAF\_AD\_RCVD. If the ABR has sent a Leaf-AD route then it must add a Leaf-AD Augmented Response Block with the values of Leaf A-D route NLRI. The upstream node's IP address is the IP address determined from the Global Administrator field of the Inter-area P2MP Segmented Next-hop Extended Community.

### 3.7. Mtrace MVPN Procedure

In this section, we will briefly discuss the Mtrace procedure taking a working and non-working network topology.



Mtrace MVPN Procedure

The above figure depicts the path of MTRACE in working condition. MTRACE request for MVPN can traverse multiple hops when a Virtual HUB is present or when segmented P2MP inter-area tunnels are used. If no error conditions are detected the downstream request will travel the same path as the regular multicast packet for the queried mroute would flow. The last hop router/egress router will convert it into a Response and send it back to querier

Let us consider a non-working case where Mtrace is expected to be used. Taking Virtual-HUB as an example, assume that there is a data-path issue between V-HUB and Egress Spoke. The below steps take place to determine the issue between V-HUB and egress Spoke

- 1 - Querier sends the Mtrace Query towards LHR (Egress PE-Spoke).
- 2 - Egress PE sends Request to V-HUB. V-HUB realises that the first hop router is a connected spoke and sends the request to Ingress Spoke PE.
- 3 - Ingress Spoke PE sends Downstream Request to V-HUB. The same is received by V-HUB. V-HUB sets the 'D' bit in its PMSI Tunnel Attributes Augmented Response Block.
- 4 - V-HUB sends Downstream request to ingress spoke. This is never received by the ingress spoke.
- 5 - The result of first 4 steps is that querier did not receive the response. This makes the querier fall back to TTL method.
- 6 - Querier reduces the TTL and the result will show that the hop from V-HUB to ingress spoke is missing thereby pointing the issue at the right place.

#### 4. Error Detection

All routers will check for normal multicast errors as defined in the Mtracev2 specification. In addition, they will check for errors specific to MVPNs and this specification.

All receiving routers will check the state of the Provider Tunnel used for forwarding traffic for the given S,G. The ability and manner to check if the Provider Tunnel is down depends on the Provider Tunnel type. If the Provider Tunnel is known to be down the PE will respond with a PTUNNEL\_DOWN error.

In some situations the router needs to send a Leaf AD route to the upstream PE. If the upstream expects a Leaf AD route, but did not receive one from the downstream PE, then the NO\_LEAF\_AD\_RCVD error will be sent.

The receiving router will check the values of the PMSI Tunnel attributes to see if they match the expected values for the PMSI. If an Inclusive-PMSI is used, then the router will verify that the values match those in the I-PMSI A-D route. If a Selective PMSI is used, then the Tunnel Attributes will be matched against the S-PMSI or Leaf A-D Route, depending on the Tunnel Type. If the values do not match, then a error code of the corresponding PMSI mismatch will be sent.

If a router receives a MVPN traceroute, but does not have the proper MVPN configuration, then it will respond with a UNEXPECTED\_MVPN error

#### 4.1. MVPN Error Codes

Value	Name	Description
-----	-----	-----
0x11	PTUNNEL_DOWN	The provide tunnel for this S,G is down.
0x12	NO_LEAF_AD_RCVD	The S-PMSI has not been joined by downstream neighbor
0x13	BAD_LEAF_AD	The Leaf A-D route does not match the expected values
0x14	BAD_RD	The RD is known to not exist on this PE
0x15	UNEXPECTED_MVPN	The MVPN traceroute message is unexpected
0x16	BAD_PMSI_ATTR_FLAG	Error matching the PMSI attribute flag
0x17	BAD_PMSI_ATTR_TYPE	Error matching the PMSI attribute type
0x18	BAD_PMSI_ATTR_LABEL	Error matching the PMSI attribute label
0x19	BAD_PMSI_ATTR_ID	Error matching the PMSI attribute tunnel identifier
0x1a	UNKNOWN_INTER_AS	Could not locate the Inter-AS provider tunnel segment.
0x1b	NO_UPSTREAM_PE	No valid upstream PE or route



0x1c NO\_CMCAST\_STATE No C-Mcast route for the requested query

0x1d NO\_WILD\_CARD\_SPMSI\_AD\_RCVD No Wild Card SPMSI AD is received from the upstream PE

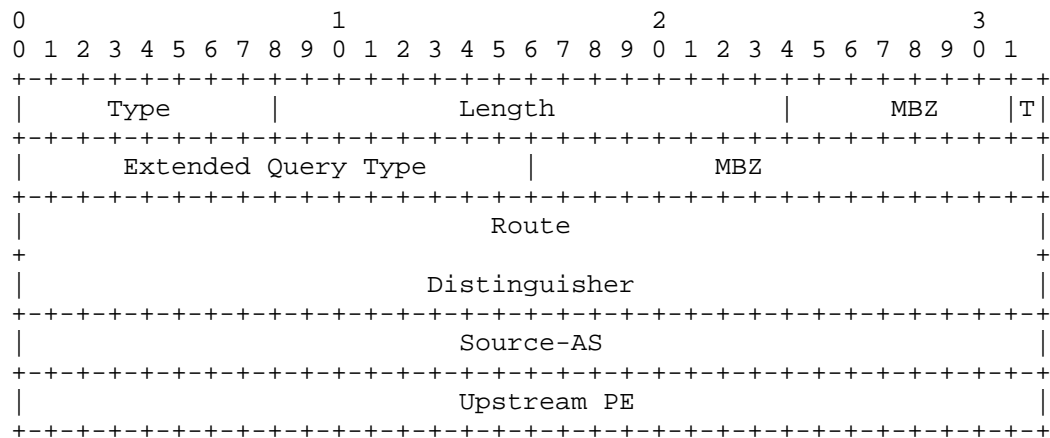
0x1e NO\_WILD\_CARD\_SPMSI\_LEAD\_AD\_SENT PE did not send LEAF-AD route for the wild card SPMSI

## 5. Mtracev2 Extensions

### 5.1. New Mtracev2 TLV Type

A new Mtracev2 TLV type will be created for the Mtrace2 Downstream Request.

### 5.2. MVPN Extended Query Block



#### MVPN Extended Query Block

Type: Mtrace2 Extended Query Block Type

Length: Length of the MVPN Extended Query Block

MBZ: Sent with all 0's, ignored on receipt

T bit: This bit should be 0

Extended Query Type: New type defined

MBZ: Sent with all 0's, ignored on receipt

Route Distinguisher: The RD of the S,G that should be traced

Source-AS: The Autonomous System Number (ASN) of the Source

Upstream PE: IP Address of the Upstream PE

### 5.3. Leaf A-D Augmented Response Block

```

0                               1                               2                               3
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|                                     Type                                     | Value .... |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

#### Leaf A-D Augmented Response Block

MBZ: Sent with all 0's, ignored on receipt

Type: New type defined

Value: The NLRI value of the associated Leaf A-D route

### 5.4. PMSI Tunnel Attributes Augmented Response Block

```

0                               1                               2                               3
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|                                     Type                                     |D|   MBZ   | Value.. |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

#### PMSI Tunnel Attributes Augmented Response Block

MBZ: Sent with all 0's, ignored on receipt

Type: New type defined

D: 'D' bit indicating that Downstream Request is received on PMSI

Value: The PMSI Tunnel Attribute as defined in RFC 6514

### 6. Mtrace2 Standard Response Block considerations



The PEs in the MVPN Mtrace add the Standard Response Block as defined in Mtrace2 [I-D.ietf-mboned-mtrace-v2]. For a PE, the incoming or outgoing interface can be a Tunnel. The First Hop Router (FHR) PE which is connected to the source SHOULD populate the incoming interface address with the respective interface connected to the CE. The outgoing interface address MAY be populated with 0 in this case. Other routers in the mtrace path MAY populate incoming and outgoing interface address fields as 0. 'Multicast Rtg Protocol' field MUST be populated with 0s by the Last Hop Router (LHR). First Hop Router (FHR) can populate this field with respective multicast routing protocol used towards its upstream CE. All the remaining fields of the Standard Response Block are populated as defined by the Mtrace2 [I-D.ietf-mboned-mtrace-v2] specification.

## 7. IANA Considerations

New TLV Type for MTRACE\_MVPN\_QUERY, MTRACE\_MVPN\_REQUEST, MTRACE\_MVPN\_DOWNSTREAM\_REQUEST, MTRACE\_MVPN\_RESPONSE

## 8. Security Considerations

There are no security considerations for this design other than what is already in the mtracev2 specification.

## 9. Acknowledgments

The authors would like to thank Yakov Rekhter and Marco Rodrigues for their valuable review and feedback.

## 10. Normative References

[I-D.ietf-l3vpn-virtual-hub]

Jeng, H., Uttaro, J., Jalil, L., Decraene, B., Rekhter, Y., and R. Aggarwal, "Virtual Hub-and-Spoke in BGP/MPLS VPNs", draft-ietf-l3vpn-virtual-hub-08 (work in progress), July 2013.

[I-D.ietf-mboned-mtrace-v2]

Asaeda, H. and W. Lee, "Mtrace Version 2: Traceroute Facility for IP Multicast", draft-ietf-mboned-mtrace-v2-09 (work in progress), October 2012.

[I-D.ietf-mppls-seamless-mcast]

Rekhter, Y. and R. Aggarwal, "Inter-Area P2MP Segmented LSPs", draft-ietf-mppls-seamless-mcast-07 (work in progress), May 2013.

- [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.
- [RFC6625] Rosen, E., Rekhter, Y., Hendrickx, W., and R. Qiu, "Wildcards in Multicast VPN Auto-Discovery Routes", RFC 6625, May 2012.

Authors' Addresses

Robert Kebler  
Juniper Networks  
10 Technology Park Drive  
Westford, MA 01886  
USA

Email: rkebler@juniper.net

Pavan Kurapati  
Juniper Networks  
1194 N. Mathilda Ave  
Sunnyvale, CA 94089  
USA

Email: kurapati@juniper.net

Network Working Group  
Internet-Draft  
Expires: January 16, 2014

T. Morin, Ed.  
S. Litkowski  
Orange  
K. Patel  
Cisco Systems  
J. Zhang  
R. Kebler  
Juniper Networks  
July 15, 2013

Multicast state damping  
draft-morin-multicast-damping-00

Abstract

This document describes procedures to damp multicast routing state changes and prevent the churn due to the multicast dynamicity at the edge of a network. The procedures described in this document help avoid uncontrolled control plane load increase on the core routing infrastructure. New procedures are proposed inspired from BGP unicast route damping principles, but adapted to multicast. They cover multicast and multicast in VPNs contexts.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

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 January 16, 2014.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. 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 . . . . .	3
2. Terminology . . . . .	3
3. Overview . . . . .	3
4. Existing mechanisms . . . . .	4
4.1. Rate-limiting of multicast control traffic . . . . .	4
4.2. Existing PIM, IGMP and MLD timers . . . . .	4
4.3. BGP Route Damping . . . . .	5
5. Procedures for multicast state damping . . . . .	6
6. Procedures for multicast in IP VPNs . . . . .	8
6.1. Damping P-tunnel change events . . . . .	9
7. Procedures for Ethernet VPNs . . . . .	10
8. Operational considerations . . . . .	10
8.1. Enabling and configuring multicast damping . . . . .	10
8.2. Troubleshooting and monitoring . . . . .	11
8.3. Maximum values for exponential decay and thresholds parameters . . . . .	11
8.4. Default values . . . . .	11
9. IANA Considerations . . . . .	11
10. Security Considerations . . . . .	11
11. Acknowledgements . . . . .	12
12. References . . . . .	12
12.1. Normative References . . . . .	12
12.2. Informative References . . . . .	13
Authors' Addresses . . . . .	13

## 1. Introduction

When multicast receivers join and leave a said multicast group or channel at the edge of a network through multicast membership control protocols (IGMP, MLD), multicast routing protocols (e.g. PIM-SM, or mVPN) adjust multicast routing states accordingly to forward or prune multicast traffic to these receivers.

Mechanisms need to be put in place to ensure that the load put on the control plane of core routers remains under control regardless of the frequency at which multicast memberships changes are made by end hosts. By nature multicast memberships change based on the behavior of multicast applications running on end hosts, hence the frequency of membership changes can legitimately be much higher than the typical churn of unicast routing states.

This document describes procedures aimed at protecting the control plane of the core network infrastructure (more specifically edge routers, core routers and in the case of multicast in VPN contexts BGP Route Reflectors) while at the same time avoiding negative effects on the service provided, although at the expense of a minimal increase in average of bandwidth use in the network.

The base principle is described in Section 3. Existing mechanisms that could be relied upon are discussed in Section 4. Section 5 details the proposed procedures.

Sections 6 and 7 provide more specific details related to multicast in VPNs contexts.

Finally, Section 8 discusses operational considerations related to the proposed mechanism.

## 2. Terminology

TBC

## 3. Overview

The procedures described in this document allows the network operator to configure multicast routers so that they can delay the propagation of multicast state prune messages, when faced with a rate of multicast state dynamicity exceeding a certain configurable threshold. Assuming that the number of multicast states that can be created by a receiver is bounded, delaying the propagation of multicast state pruning results in setting up an upper bound to the

average frequency at which the router will send state updates to an upstream router.

From the point of view of a downstream router, this approach has no impact: the multicast routing states changes that it solicits to its upstream router will be honored without any additional delay. Indeed the propagation of joins is not impacted by the proposed defined procedures, and having the upstream router delay state prune propagation to its own upstream does not affect what traffic is sent to the downstream router. In particular, the amount of bandwidth used on the link downstream to a router applying this damping technique is not increased.

This approach increases the average bandwidth utilization on a link upstream to a router applying this technique: indeed, the bandwidth of a said multicast flow will be used for a longer time than if no damping was applied. That said, it is expected that this technique will allow to meet the goals of protecting the multicast routing infrastructure control plane without a significant average increase of bandwidth; for instance, damping events happening at a frequency higher than one event per X second, can be done without increasing the time during which a multicast flow is present on a link of more than X second.

To be practical, such a mechanism requires configurability, in particular, needs to offer means to control when damping is triggered and allow delaying Pruning for a longer period of time the more activity there is on a multicast state.

Note that the issues related to control plane load due to the dynamicity of multicast sources coming and going in the context of ASM multicast, are out of the scope of this document.

#### 4. Existing mechanisms

##### 4.1. Rate-limiting of multicast control traffic

[RFC4609] examines multicast security threats and among other things the risk described in Section 1. A mechanism relying on rate-limiting PIM messages is proposed in section 5.3.3 [RFC4609], but has the identified drawbacks of impacting the service delivered and having side-effects on legitimate users.

##### 4.2. Existing PIM, IGMP and MLD timers

In the context of PIM multicast routing protocols (), a mechanism exists that in some context may offer a form of de facto damping

mechanism for multicast states. Indeed, when active, the prune override mechanism consist in having a PIM upstream router delay for a certain time [prune override interval] before taking into account a PIM Prune message sent by a downstream neighbor. This mechanism has not been designed specifically for the purpose of damping multicast state, but as a means to allow PIM to operate on multi-access networks. See [RFC4601] section 4.3.3.

However, when active, this mechanism will prevent a downstream router to produce multicast routing protocol messages for a said multicast state that would result in the upstream router to send, to its own upstream, multicast routing protocol messages at a rate higher than  $1/[\text{prune override interval}]$ .

Similarly, the IGMP and MLD multicast membership control protocols can provide under the right conditions a similar behavior.

These mechanisms are not considered suitable to meet the goals spelled out in Section 1, the main reasons being that:

- o when enabled these mechanisms require additional bandwidth on the local link on which the effect of a Prune is delayed
- o to be active, they may require disabling features that may otherwise be required or useful; one typical example is explicit tracking for IGMP/MLD or PIM
- o on certain implementation, would require disabling behavior that cannot be turned off
- o do not provide a suitable level of configurability
- o do not provide a way to discriminate between multicast flows based on an averaged estimation of their recent past dynamicity

#### 4.3. BGP Route Damping

The procedures defined in [RFC2439] for BGP route flap damping are useful for operators who want to control the impact of unicast route churn on the routing infrastructure, and offer a standardized set of parameters to control damping.

These procedures are not directly relevant in a multicast context, for the following reasons:

- o they are not specified for multicast routing protocol in general

- o even in contexts where BGP routes are used to carry multicast routing states (e.g. [RFC6514]), these procedures do not allow to implement the principle described in this document, the main reason being that a damped route becomes suppressed, while the target behavior would be to keep advertising when damping is triggered on a multicast route

However, the set of parameters standardized to control the thresholds of the exponential decay mechanism can be relevantly reused. This is the approach proposed for the procedures described in this document (Section 5). Motivations for doing so is to help the network operator deploy this feature based on consistent configuration parameter, and obtain predictable results, without the drawbacks of exposed in Section 4.1 and Section 4.2.

## 5. Procedures for multicast state damping

This section describes procedures for multicast state damping satisfying the goals spelled out in Section 1. This section spells out procedures for (S,G) states in the PIM-SM protocol ([RFC4601] ; they apply unchanged for such states created based on multicast group management protocols (IGMP [RFC3376], MLD [RFC3810]) on downstream interfaces. How these procedures apply for any-source multicast (ASM) routing state will be covered in a further revision.

The following notions introduced in [RFC2439] are reused in these procedures:

figure-of-merit   \*\*a number reflecting the current estimation of past recent activity of an (S,G) multicast routing state, which evolves based on routing events related to this state and based an exponential decay algorithm ; the activation or inactivation of damping on the state is based on this number

cutoff-threshold parameter   value of the \*figure-of-merit\* over which damping is applied (configurable value)

reuse-threshold parameter   value of the \*figure-of-merit\* under which damping stops being applied (configurable value)

decay-half-life parameter   period of time used to control how fast is the exponential decay of the \*figure-of-merit\* (configurable value)

Additionally to these values a configurable "\*increment-factor\*" parameter is introduced, that controls by how much the figure-of-merit is incremented on multicast state update events.



Section 8.4 will propose default values for all these parameters.

On reception of updated multicast membership or routing information on a downstream interface I for a said (S,G) state, that results in a change of the state of the PIM downstream state machine (see section 4.5.3 of [RFC4601]), a router implementing these procedures MUST:

- o apply unchanged procedures for everything relating to what multicast traffic ends up traffic being sent on downstream interfaces, including interface I
- o increasing the *\*figure-of-merit\** for the (S,G) by the *\*increment-factor\** (updating the *\*figure-of-merit\** based on the decay algorithm must be done prior to this increment)
- o update the damping state for the (S,G) state: damping becomes active on the state if the recomputed *\*figure-of-merit\** is above the configured *\*cutoff-threshold\**
- o update the upstream state machine for (S,G) as per section 4.5.7 of [RFC4601], with the following change : if the state machine transitions to NotJoined state because of the reception of a PIM or IGMP/MLD message on a downstream interface (i.e. in the terminology of [RFC4601] *inheritedolist(S,G)* becomes NULL ), and if damping is active on the state, the router SHOULD NOT send the resulting Prune(S,G) message to its upstream neighbor ; this message MUST be sent when the damping state becomes, i.e. inactive when *\*figure-of-merit\** decays to a value below the configured *\*reuse-threshold\**

Same techniques as the ones described in [RFC2439] can be applied to determine when the figure-of-merit value is recomputed based on the exponential decay algorithm and the configured *\*decay-half-life\**. Given the specificity of multicast applications, it is REQUIRED for the implementation to let the operator configure the *\*decay-half-life\** in seconds, rather than in minutes. When the recomputation is done periodically, the period should be low enough to not significantly delay the inactivation of damping on a multicast state beyond what the operator wanted to configure (i.e. for a half-life of 10s, recomputing the *\*figure-of-merit\** each minute would result in a multicast state to remained damped for a time longer than what the parameters are supposed to command).

When a (S,G) state expires, its associated *\*figure-of-merit\** and damping state are removed as well.

These procedures do interact with PIM procedures related to refreshes

or expiration of multicast routing states. Indeed, PIM Prune messages triggered by the expiration of the (S,G) keep-alive timer, are not suppressed or delayed (see Section 8.3 for a discussion on why this specific aspect is not expected to impede the efficiency of damping procedures), and the reception of Join messages not causing transition of state on the downstream interface does not lead to incrementing the \*figure-of-merit\*.

Note that these procedures do not impact the PIM assert mechanism, in particular PIM Prune messages triggered by a change of the PIM assert winner on the upstream interface, are not suppressed or delayed.

Note also that no action is triggered based on the reception of PIM Prune messages (or corresponding IGMP/MLD messages) that relate to non-existing (S,G) state, in particular, no \*figure-of-merit\* or damping state is created in this case.

## 6. Procedures for multicast in IP VPNs

In VPN contexts, providing isolation between customers of a shared infrastructure is a core requirement resulting in even stringent expectations with regards to risks of denial of service attacks. Procedures for multicast support in IP VPNs are described in [RFC6513] and [RFC6514] and section 16 of [RFC6514] specifically spells out the need for damping the activity of C-multicast and Leaf Auto-discovery route.

The procedures described in Section 5 can be applied in the VRF PIM-SM implementation (in the "C-PIM instance"), with the corresponding action to suppressing the emission of a Prune(S,G) message being to not withdraw the C-multicast Source Tree Join (C-S,C-G) BGP route. Implementation of [RFC6513] relying on the use of PIM to carry C-multicast routing information MUST support this technique.

In the context of [RFC6514] where BGP is used to distribute C-multicast routing information, an additional option consists in applying damping at the level of the BGP implementation based on existing BGP damping mechanism, applied to C-multicast Source Tree Join routes and Shared Tree Join routes (and also Leaf A-D routes - see Section 6.1), and modified to provide the same effect of procedures described in Section 5 along the following guidelines:

- o not withdrawing (instead of not advertising) damped routes
- o providing means to configure the half-life in seconds if that option is not already available

- o using parameters for the exponential decay that are specific to multicast, based on default values and multicast specific configuration

Note that in a context where BGP Route Reflectors are used, it can be considered useful to also be able to apply damping on RRs. Additionally, for mVPN Inter-AS deployments, it can be needed to protect one AS from the dynamicity of multicast VPN routing events from other ASes. In that perspective, it is RECOMMENDED for implementations to support damping mVPN C-multicast routes directly into BGP, without relying on the PIM-SM state machine.

The choice to implement damping based on BGP routes or the procedures described in Section 5, is up to the implementor, but at least one of the two MUST be implemented; keeping in mind that in contexts where damping on RRs and ASBRs the BGP approach is RECOMMENDED.

Note well that damping SHOULD NOT be applied to BGP routes of the following sub-types: "Intra-AS I-PMSI A-D Route", "Inter-AS I-PMSI A-D Route", "S-PMSI A-D Route", and "Source Active A-D Route".

The following sub-sections describe additional procedures providing coverage against harmful effects of high multicast membership state dynamicity specific to mVPNs, and preserving the goals spelled out in Section 1.

#### 6.1. Damping P-tunnel change events

When selective P-tunnels are used (see section 7 of [RFC6513]), the effect of updating the upstream state machine for a said (C-S,C-G) state on a PE connected to multicast receivers, is not only to generate activity to propagate C-multicast routing information to the source connected PE, but also to possibly trigger changes related to the P-tunnels carrying (C-S,C-G) traffic. Protecting the provider network for an excessive amount of change in the state of P-tunnels is required, and this section details how it can be done.

A PE implementing these procedures for mVPN MUST damp Leaf A-D routes, in the same manner as it would for C-multicast routes (see Section 6).

A PE implementing these procedures for mVPN MUST damp the activity related to removing itself from a P-tunnel. Possible ways to do so depend on the type of P-tunnel, and local implementation details are left up to the implementor.

The following is proposed as example of how the above can be achieved.

- o For P-tunnels implemented with the PIM protocol, this consists in applying multicast state damping techniques describe in Section 5 to the P-PIM instance, at least for (S,G) states corresponding to P-tunnels.
- o For P-tunnels implemented with the mLDP protocol, this consists in applying damping techniques completely similar as the one described in Section 5, but generalized to apply to mLDP states
- o For root-initiated P-tunnel (P-tunnels implemented with the P2MP RSVP-TE, or relying on ingress replication), no particular action needs to be implemented to damp P-tunnels membership as soon as the activity of Leaf A-D route is damped
- o Another possibility is to base the decision to join or not join the P-tunnel to which a said (C-S,C-G) is bound, and to advertise or not advertise a Leaf A-D route related to (C-S,C-G), based on whether or not a C-multicast Source Tree Join route is being advertised for (C-S,C-G), rather than by relying on the state of the C-PIM Upstream state machine for (C-S,C-G)

## 7. Procedures for Ethernet VPNs

Specifications exists to support or optimize multicast and broadcast in the context of Ethernet VPNs ([I-D.ietf-l2vpn-vpls-mcast], [I-D.ietf-l2vpn-evpn]). The said specifications make use of S-PMSI and P-tunnels and for this reason, an implementation of these procedures MUST follow the procedures described in Section 6.1.

## 8. Operational considerations

### 8.1. Enabling and configuring multicast damping

In the context of flat multicast routing, it is proposed that enabling this multicast damping mechanism at the edge of a network providing a multicast service, for instance at receiver-facing routers or in ASBRs, will be sufficient to address the targeted issue. Additionally, these procedures can be enabled on core routers as well.

In the context of multicast VPNs, these procedures would be enabled on PE routers. Additionally in the case of C-multicast routing based on BGP extensions ([RFC6514]) these procedures can be enabled on ASBRs, and possibly Route Reflectors as well.

## 8.2. Troubleshooting and monitoring

Implementing the damping mechanisms described in this document should be complemented by appropriate tools to observe and troubleshoot damping activity.

More specifically it is RECOMMENDED to complement the existing interface providing information on multicast states with information on eventual damping of corresponding states (e.g. MRIB states). In the case of mVPN this applies also to information on P-tunnels damping, and when BGP is used for C-multicast routing propagation, to BGP C-multicast routes.

## 8.3. Maximum values for exponential decay and thresholds parameters

[TBC]

## 8.4. Default values

[TBC]

## 9. IANA Considerations

This document makes no request of IANA.

Note to RFC Editor: this section may be removed on publication as an RFC.

## 10. Security Considerations

The procedures defined in this document do not introduce additional security issues not already present in the contexts addressed, and actually aim at addressing some of the identified risks without introducing as much denial of service risk as some of the mechanisms already defined.

The protection provided relates to the control plane of the multicast routing protocols, including the components implementing the routing protocols and the components responsible for updating the multicast forwarding plane.

The procedures describe are meant to provide some level of protection for the router on which they are enabled by reducing the amount of routing state updates that it needs to send to its upstream neighbor or peers, but do not provide any reduction of the control plane load related to processing routing information from downstream neighbors.

Protecting routers from an increase in control plane load due to activity on downstream interfaces toward core routers (or in the context of BGP-based mVPN C-multicast routing, BGP peers) shall rely upon the activation of damping on corresponding downstream neighbors (or BGP peers) and/or at the edge of the network. Protecting routers from an increase in control plane load due to activity on customer-facing downstream interfaces or downstream interfaces to routers in another administrative domain, is out of the scope of this document and should rely upon already defined mechanisms (see [RFC4609]).

To be effective the procedures described here must be complemented by configuration limiting the number of multicast states that can be created on a multicast router through protocol interactions with multicast receivers, neighbor routers in adjacent ASes, or in multicast VPN contexts with multicast CEs. Note well that the two mechanism may interact: state for which Prune has been requested may still remain taken into account for some time if damping has been triggered and hence result in otherwise acceptable new state from being successfully created.

Additionally, it is worth noting that these procedures are not meant to protect against peaks of control plane load, but only address averaged load. For instance, assuming a set of multicast states submitted to the same Join/Prune events, damping can prevent more than a certain number of Join/Prune messages to be sent upstream in the period of time that elapses between the reception of Join/Prune messages triggering the activation of damping on these states and when damping becomes inactive after decay.

## 11. Acknowledgements

We would like to thank Bruno Decreane, Jeff Haas and Lenny Giuliano for discussions that helped shape this proposal. We would also like to thank Yakov Rekhter and Eric Rosen for their reviews and helpful comments. Thanks to Wim Henderickx for his comments and support of this proposal.

## 12. References

### 12.1. Normative References

[I-D.ietf-l2vpn-evpn]

Sajassi, A., Aggarwal, R., Henderickx, W., Balus, F., Isaac, A., and J. Uttaro, "BGP MPLS Based Ethernet VPN", draft-ietf-l2vpn-evpn-04 (work in progress), July 2013.

- [I-D.ietf-l2vpn-vpls-mcast]  
Aggarwal, R., Rekhter, Y., Kamite, Y., and L. Fang,  
"Multicast in VPLS", draft-ietf-l2vpn-vpls-mcast-14 (work  
in progress), July 2013.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate  
Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC2439] Villamizar, C., Chandra, R., and R. Govindan, "BGP Route  
Flap Damping", RFC 2439, November 1998.
- [RFC3376] Cain, B., Deering, S., Kouvelas, I., Fenner, B., and A.  
Thyagarajan, "Internet Group Management Protocol, Version  
3", RFC 3376, October 2002.
- [RFC3810] Vida, R. and L. Costa, "Multicast Listener Discovery  
Version 2 (MLDv2) for IPv6", RFC 3810, June 2004.
- [RFC4601] Fenner, B., Handley, M., Holbrook, H., and I. Kouvelas,  
"Protocol Independent Multicast - Sparse Mode (PIM-SM):  
Protocol Specification (Revised)", RFC 4601, August 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.

## 12.2. Informative References

- [RFC4609] Savola, P., Lehtonen, R., and D. Meyer, "Protocol  
Independent Multicast - Sparse Mode (PIM-SM) Multicast  
Routing Security Issues and Enhancements", RFC 4609,  
October 2006.

## Authors' Addresses

Thomas Morin (editor)  
Orange  
2, avenue Pierre Marzin  
Lannion 22307  
France

Email: thomas.morin@orange.com

Stephane Litkowski  
Orange  
France

Email: [stephane.litkowski@orange.com](mailto:stephane.litkowski@orange.com)

Keyur Patel  
Cisco Systems  
170 W. Tasman Drive  
San Jose, CA 95134  
USA

Email: [keyupate@cisco.com](mailto:keyupate@cisco.com)

Jeffrey (Zhaohui) Zhang  
Juniper Networks Inc.  
10 Technology Park Drive  
Westford, MA 01886  
USA

Email: [zzhang@juniper.net](mailto:zzhang@juniper.net)

Robert Kebler  
Juniper Networks Inc.  
10 Technology Park Drive  
Westford, MA 01886  
USA

Email: [rkebler@juniper.net](mailto:rkebler@juniper.net)





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.

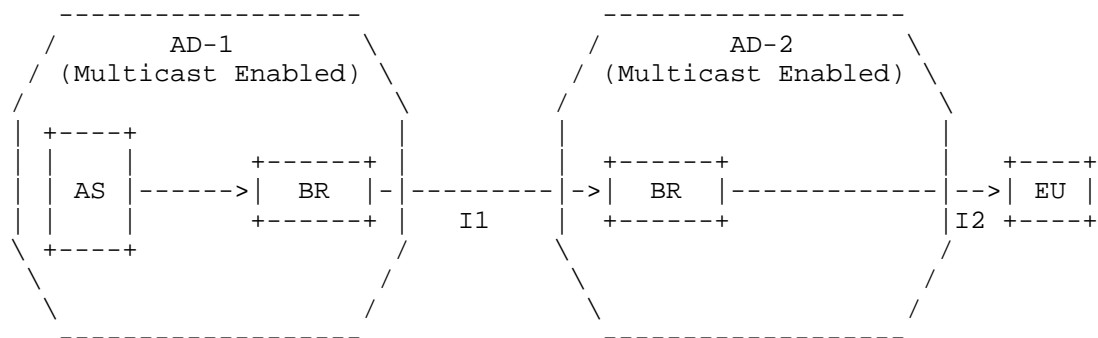
The Use Cases describing various architectural configurations for the multicast distribution along with associated requirements is described in section 3. Unique aspects related to the Enterprise network possibility will be described in this section. A comprehensive list of pertinent information that 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.

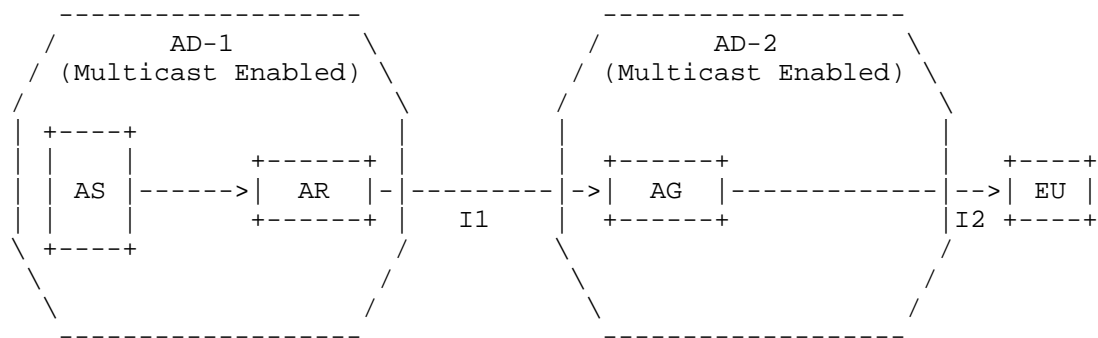
- o GRE tunnels are typically configured manually between peering points to support multicast delivery between domains.



- 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.

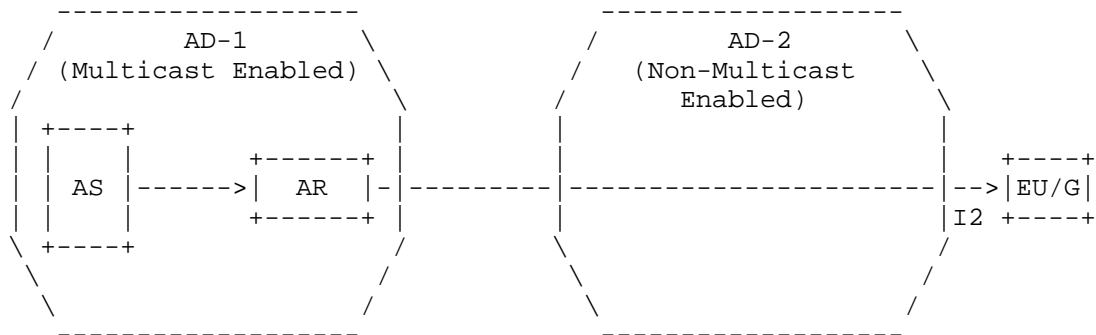
Architectural guidelines for this configuration are as follows:

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:

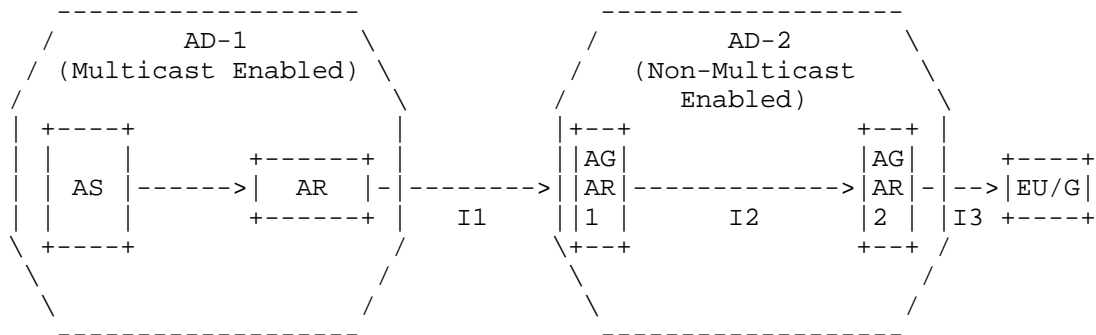
Guidelines (a) through (c) are the same as those described in Use Case 3.1.

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 a 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



mboned WG  
Internet-Draft  
Intended status: Standards Track  
Expires: December 31, 2014

C. Xie  
Q. Sun  
China Telecom  
C. Wang  
W. Meng  
ZTE Corporation  
B. Khasnabish  
ZTE USA, Inc  
June 29, 2014

IPv4-IPv6 Multicast Address Conversion  
draft-tsao-mboned-v4v6mcast-dynamic-conversion-02

Abstract

This draft describes a mechanism for stateless conversion of IPv4 multicast address to IPv6 multicast address and vice versa, using different rules. These rules can be used in both IPv4-IPv6 translation or encapsulation. This solution can be used in any scenarios describe in [RFC6144].

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 December 31, 2014.

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.

## Table of Contents

1. Introduction . . . . .	3
2. Convention and Terminology . . . . .	4
3. Architecture . . . . .	5
4. IPv4/IPv6 Multicast Address Conversion . . . . .	6
4.1. Rule Design . . . . .	6
4.2. IPv4 Multicast Address Suffix-embedded IPv6 Multicast Address . . . . .	7
4.3. Full IPv4 Multicast Address-embedded IPv6 Multicast Address . . . . .	7
5. Forwarding . . . . .	9
5.1. From IPv4 Multicast System to IPv6 Multicast System . . .	9
5.2. From IPv6 Multicast System to IPv6 Multicast System . . .	9
6. Backwards compatibility . . . . .	10
7. Security Considerations . . . . .	11
8. References . . . . .	12
8.1. Normative References . . . . .	12
8.2. Informative References . . . . .	12
Authors' Addresses . . . . .	13

## 1. Introduction

This draft describes a mechanism for stateless translation between IPv4 multicast address and IPv6 multicast address using different rules. These rules can be used in both IPv4-IPv6 translation or encapsulation. This solution can be used in any scenarios describe in [RFC6144].

The approach described in this draft is fully compatible with [I-D.ietf-mboned-64-multicast-address-format].



## 2. Convention and Terminology

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].

Rule\_IPv6\_M\_Prefix/Length:

Define an IPv6 Prefix assigned by a Service Provider for a IPv4/IPv6 Multicast Address Conversion rule.

Rule\_IPv4\_M\_Prefix/Length:

Define an IPv4 Prefix assigned by a Service Provider for a IPv4/IPv6 Multicast Address Conversion rule.

Rule\_IPv4\_Offset:

Define an offset where IPv4 Multicast Address should be embedded in the IPv6 Multicast Address.

Rule\_IPv4\_Type:

Defined whether an IPv4 Multicast Address Suffix or a full IPv4 Multicast Address is embedded in the IPv6 Multicast Address. Value 0 is default and means IPv4 Multicast Address Suffix is embedded in the IPv6 Multicast Address. Value 1 means a full IPv4 Multicast Address is embedded in the IPv6 Multicast Address.

### 3. Architecture

All of the scenarios that are describe in [RFC6144] can be easily illustrate using the diagram show in Figure 1 below:

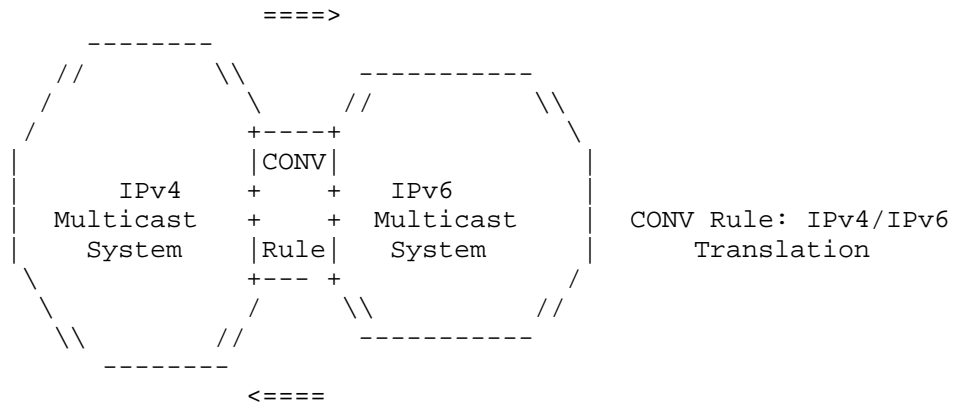


Figure 1: IPv4-IPv6 Address Conversion

As shown in this diagram(Fig.1), there is a conversion node between an IPv4 Multicast System and IPv6 Multicast System. Every conversion node must be provisioned with at least one rule defined in the document used for IPv4/IPv6 Multicast Address Conversion. There are also two arrows: an arrow from IPv4 Multicast system to IPv6 Multicast System means IPv4 Multicast system initiates the multicast flow. Another arrow from IPv6 Multicast system to IPv4 Multicast System means IPv6 Multicast system initiates the multicast flow. And this also means that the algorithmic described in this document support both IPv4-initiated communication and IPv6-initiated communication.

#### 4. IPv4/IPv6 Multicast Address Conversion

This section specifies the rule(s) for IPv4/IPv6 multicast address conversion.

##### 4.1. Rule Design

Every CONV node must be provisioned with at least one rule. When there are several rules for IPv4/IPv6 Conversion assigned for a CONV node, this node should choose the rule which is longest match prefix for the destination IP address in multicast flow.

Each rule includes the following:

Rule\_IPv6\_M\_Prefix (including prefix length)

Rule\_IPv4\_M\_Prefix (including prefix length, optional)

Rule\_IPv4\_Offset (optional)

Rule\_IPv4\_Type (optional)

Rule\_IPv6\_M\_Prefix/Length is according to section 2.7 of [ADDRARCH][RFC3513], or based on [RFC3306]. This parameter is mandatory.

Rule\_IPv4\_M\_Prefix/Length is in IPv4 multicast group address scope. By default, this parameter is empty, which means match any IPv4 group address in the destination address field in the receiving packet. This parameter is optional.

Rule\_IPv4\_Offset defines the offset where IPv4 multicast address is embedded in the IPv6 multicast address. By default, the value is 96, which means embedded the IPv4 multicast address in the last 32 bits of the IPv6 multicast address. This parameter is optional.

Rule\_IPv4\_Type defines two kinds of IPv6 Multicast Address format: one format is IPv4 Multicast Address Suffix is embedded in the IPv6 Multicast Address, and corresponding Rule\_IPv4\_Type value is 0; another format is Full IPv4 Multicast Address is embedded in the IPv6 Multicast Address, and corresponding Rule\_IPv4\_Type value is 1. By default, Rule\_IPv4\_Type value is 0. This parameter is optional.

When Rule\_IPv6\_M\_Prefix is SSM mode, the corresponding Rule\_IPv4\_M\_Prefix in the same rule should be SSM mode. When Rule\_IPv6\_M\_Prefix is ASM mode, the corresponding Rule\_IPv4\_M\_Prefix in the same rule should be ASM mode.

If Rule\_IPv6\_M\_Prefix is ASM mode but the corresponding Rule\_IPv4\_M\_Prefix is SSM mode, the CONV node should process this rule as invalid. Also, if Rule\_IPv6\_M\_Prefix is SSM mode but the corresponding Rule\_IPv4\_M\_Prefix is ASM mode, the CONV node should process this rule as invalid.

#### 4.2. IPv4 Multicast Address Suffix-embedded IPv6 Multicast Address

When Rule\_IPv4\_Type value is 0, the concentrated IPv6 Multicast Address format is as follow:

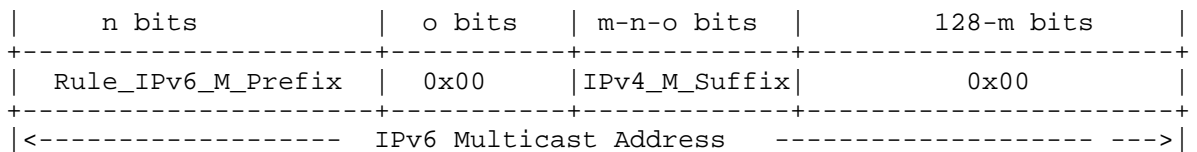


Figure 2: IPv6 Multicast Address Format for Rule\_IPv4\_Type=0

The IPv6 Multicast Address is created by combining the Rule\_IPv6\_M\_Prefix and IPv4\_M\_Suffix and all zeros. Where the IPv4\_M\_Suffix is embedded is dependent with the Rule\_IPv4\_Offset(m). From the above format, with the Rule\_IPv4\_Offset(m), can induce the embedded position of the IPv4\_M\_Suffix. Then can concentrate the IPv6 Multicast Address as above. The IPv4\_M\_Suffix illustrates as follow:

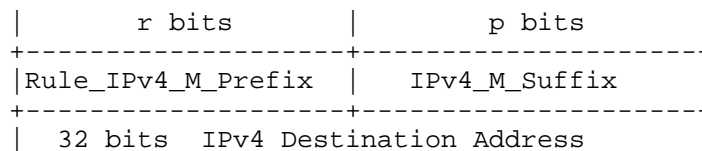


Figure 3

If Rule\_IPv4\_Offset value is 0, puts the IPv4\_M\_Suffix in the last (32-r) bits in the 128-bits IPv6 Multicast Address.

#### 4.3. Full IPv4 Multicast Address-embedded IPv6 Multicast Address

When Rule\_IPv4\_Type value is 1, the concentrated IPv6 Multicast Address format is as follow:

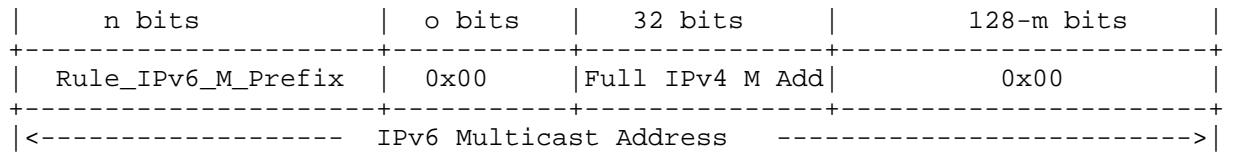


Figure 4: IPv6 Multicast Address Format for Rule\_IPv4\_Type=1

The IPv6 Multicast Address is created by combining the Rule\_IPv6\_M\_Prefix and Full IPv4 Destination Address and all zeros. Where the Full IPv4 Destination Address is embedded is dependent with the Rule\_IPv4\_Offset(m). From the above format, with the Rule\_IPv4\_Offset(m), can induce the embedded position of the Full IPv4 Destination Address. Then can concentrate the IPv6 Multicast Address as above. The Full IPv4 Destination Address is the destination IPv4 address in the multicast flow.

## 5. Forwarding

### 5.1. From IPv4 Multicast System to IPv6 Multicast System

When a CONV node receives IPv4 multicast flow from IPv4 Multicast System, the CONV node should check whether there is a Rule\_IPv4\_M\_Prefix longest match with the destination IPv4 multicast address. If there is no such rule which has a longest match prefix, the CONV node should drop these IPv4 multicast flow. If there is a rule which has a longest match prefix with the destination IPv4 multicast address, then do the IPv4-IPv6 conversion according to this rule. And then derive the IPv6 multicast address. The CONV node then checks the IPv6 multicast routing table, finds the outgoing interface and forwards the IPv6 multicast flow into the IPv6 Multicast System.

### 5.2. From IPv6 Multicast System to IPv6 Multicast System

When a CONV node receives IPv6 multicast flow from IPv6 Multicast System, the CONV node should check whether there is a Rule\_IPv6\_M\_Prefix longest match with the destination IPv6 multicast address. If there is no such rule which has a longest match prefix, the CONV node should drop these IPv6 multicast flow. If there is a rule which has a longest match prefix with the destination IPv6 multicast address, then do the IPv4-IPv6 conversion according to this rule. If the Rule\_IPv4\_Type value is 0, then derives the IPv4\_M\_Suffix from the destination IPv6 address at the Rule\_IPv4\_Offset, concentrates the Rule\_IPv4\_M\_Prefix with the IPv4\_M\_Suffix as the destination IPv4 multicast address. If the Rule\_IPv4\_Type value is 1, then derives the destination IPv4 address from the destination IPv6 address at the Rule\_IPv4\_Offset. The CONV node then checks the IPv4 multicast routing table, finds the outgoing interface and forwards the IPv4 multicast flow into the IPv4 Multicast System.

## 6. Backwards compatibility

This solution is fully compatible with the multicast address format in the "draft-ietf-mboned-64-multicast-address-format".

## 7. Security Considerations

To be added later on as-needed basis.



## 8. References

### 8.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC3306] Haberman, B. and D. Thaler, "Unicast-Prefix-based IPv6 Multicast Addresses", RFC 3306, August 2002.
- [RFC3513] Hinden, R. and S. Deering, "Internet Protocol Version 6 (IPv6) Addressing Architecture", RFC 3513, April 2003.
- [RFC6144] Baker, F., Li, X., Bao, C., and K. Yin, "Framework for IPv4/IPv6 Translation", RFC 6144, April 2011.

### 8.2. Informative References

- [I-D.ietf-mboned-64-multicast-address-format]  
Boucadair, M., Qin, J., Lee, Y., Venaas, S., Li, X., and M. Xu, "IPv6 Multicast Address With Embedded IPv4 Multicast Address",  
draft-ietf-mboned-64-multicast-address-format-05 (work in progress), April 2013.

Authors' Addresses

Chongfeng Xie  
China Telecom  
Room 502, No.118, Xizhimennei Street  
Beijing  
China

Email: xiechf01@gmail.com,xiechf@ctbri.com.cn

Qiong Sun  
China Telecom  
Beijing  
China

Email: bingxuere@gmail.com,sunqiong@ctbri.com.cn

Cui Wang  
ZTE Corporation  
No.50 Software Avenue, Yuhuatai District  
Nanjing  
China

Email: wang.cuil@zte.com.cn

Wei Meng  
ZTE Corporation  
No.50 Software Avenue, Yuhuatai District  
Nanjing  
China

Email: meng.wei2@zte.com.cn,vally.meng@gmail.com

Bhumip Khasnabish  
ZTE USA,Inc  
55 Madison Avenue, Suite 160  
Morristown, NJ 07960  
USA

Email: bhumip.khasnabish@zteusa.com,vumipl@gmail.com

