

Mboned
Internet-Draft
Intended status: Informational
Expires: January 7, 2017

M. Abrahamsson
T-Systems
T. Chown
Jisc
L. Giuliano
Juniper Networks, Inc.
July 6, 2016

Multicast Service Models
draft-acg-mboned-multicast-models-00

Abstract

The draft provides a high-level overview of multicast service and deployment models, principally the Any-Source Multicast (ASM) and Source-Specific Multicast (SSM) models, and aims to provoke discussion of applicability of the models to certain scenarios. This initial draft is by no means comprehensive. Comments on the initial content, and what further content would be appropriate, or indeed whether the draft is of value, are welcomed.

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 7, 2017.

Copyright Notice

Copyright (c) 2016 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. Multicast service models	3
3. Multicast building blocks	4
3.1. Multicast addressing	4
3.2. Host signalling	4
3.3. Multicast snooping	4
4. ASM service model protocols	5
4.1. Protocol Independent Multicast, Dense Mode (PIM-DM)	5
4.2. Protocol Independent Multicast, Sparse Mode (PIM-SM)	5
4.2.1. Inter-domain PIM-SM, and MSDP	5
4.3. Bidirectional PIM (BIDIR-PIM)	6
4.4. IPv6 PIM-SM with Embedded RP	6
5. SSM service model protocols	6
5.1. Source Specific Multicast (PIM-SSM)	6
6. Discussion	7
6.1. ASM Deployment	7
6.2. SSM Deployment	7
6.3. Other considerations	8
6.3.1. Scalability, and multicast domains	9
6.3.2. Reliable multicast	9
6.3.3. Inter-domain multicast peering	9
6.3.4. Layer 2 multicast domains	9
6.3.5. Anything else?	9
7. Use case examples	10
8. Conclusions	10
9. Security Considerations	10
10. IANA Considerations	10
11. Acknowledgments	10
12. References	10
12.1. Normative References	10
12.2. Informative References	12
Authors' Addresses	13

1. Introduction

IP Multicast has been deployed in various forms, both within private networks and on the wider Internet. While a number of service models have been published individually, and in many cases revised over time, there is, we believe, no high-level guidance in the form of an Informational RFC documenting the models, their advantages and

disadvantages, and their appropriateness to certain scenarios. This document aims to fill that gap.

This initial version of the document is not complete. There are other topics that can be included. The aim of this initial version is to determine whether this work is deemed of value within the IETF mboned WG.

2. Multicast service models

The general IP multicast service model [RFC1112] is that senders send to a multicast IP address, receivers express an interest in traffic sent to a given multicast address, and that routers figure out how to deliver traffic from the senders to the receivers.

The benefit of IP multicast is that it enables delivery of content such that any multicast packet sent from a source to a given multicast group address appears once and only once on any path between a sender and an interested receiver that has joined that multicast group. A reserved range of addresses (for either IPv4 or IPv6) is used for multicast group communication.

Two high-level flavours of this service model have evolved over time. In Any-Source Multicast (ASM), any number of sources may transmit multicast packets, and those sources may come and go over the course of a multicast session without being known a priori. In ASM, receivers express interest in a given multicast group address. In Source-Specific Multicast (SSM) the specific source(s) that may send traffic to the group are known in advance. In SSM, receivers express interest in a given multicast address and specific source(s).

Senders transmit multicast packets without knowing where receivers are, or how many there are. Receivers are able to signal to on-link routers their desire to receive multicast content sent to a given multicast group, and in the case of SSM from specific sender IP addresses. They may discover the group (and sender IP) information in a number of different ways. They may also signal their desire to no longer receive multicast traffic for a given group (and sender IP).

Multicast routing protocols are used to establish the multicast forwarding paths (tree) between a sender and a set of receivers. Each router would typically maintain multicast forwarding state for a given group (and potentially sender IP), such that it knows which interfaces to forward (and where necessary replicate) multicast packets to.

Multicast packet forwarding is generally not considered a reliable service. It is typically unidirectional, but a bidirectional multicast delivery mechanism also exists.

3. Multicast building blocks

In this section we describe general multicast building blocks that are applicable to both ASM and SSM deployment.

3.1. Multicast addressing

IANA has reserved specific ranges of IPv4 and IPv6 address space for multicast addressing.

Guidelines for IPv4 multicast address assignments can be found in [RFC5771]. IPv4 has no explicit multicast address format; a specific portion of the overall IPv4 address space is reserved for multicast use (224.0.0.0/4).

Guidelines for IPv6 multicast address assignments can be found in [RFC2375] and [RFC3307]. The IPv6 multicast address format is described in [RFC4291]. An IPv6 multicast group address will lie within ff00::/8.

3.2. Host signalling

A host wishing to signal interest in receiving (or no longer receiving) multicast to a given multicast group (and potentially from a specific sender IP) may do so by sending a packet using one of the protocols described below on an appropriate interface.

For IPv4, a host may use Internet Group Management Protocol Version 2 (IGMPv2) [RFC2236] to signal interest in a given group. IGMPv3 [RFC3376] has the added capability of specifying interest in receiving multicast packets from specific sources.

For IPv6, a host may use Multicast Listener Discovery Protocol (MLD) [RFC2710] to signal interest in a given group. MLDv2 [RFC3810] has the added capability of specifying interest in receiving multicast packets from specific sources.

Further guidance on IGMPv3 and MLDv2 is given in [RFC4604].

3.3. Multicast snooping

Is this appropriate in this document? There is discussion in [RFC4541].

4. ASM service model protocols

4.1. Protocol Independent Multicast, Dense Mode (PIM-DM)

PIM-DM is detailed in [RFC3973]. It operates by flooding multicast messages to all routers within the network in which it is configured. This ensures multicast data packets reach all interested receivers behind edge routers. Prune messages are used by routers to tell upstream routers to (temporarily) stop forwarding multicast for groups for which they have no known receivers.

PIM-DM remains an Experimental protocol since its publication in 2005.

4.2. Protocol Independent Multicast, Sparse Mode (PIM-SM)

The most recent revision of PIM-SM is detailed in [RFC7761]. PIM-SM is, as the name suggests, well-suited to scenarios where the subnets with receivers are sparsely distributed throughout the network. PIM-SM supports any number of senders for a given multicast group, which do not need to be known in advance, and which may come and go through the session. PIM-SM does not use a flooding phase, making it more scalable and efficient than PIM-DM, but this means PIM-SM needs a mechanism to construct the multicast forwarding tree (and associated forwarding tables in the routers) without flooding the network.

To achieve this, PIM-SM introduces the concept of a Rendezvous Point (RP) for a PIM domain. All routers in a PIM-SM domain are then configured to use specific RP(s). Such configuration may be performed by a variety of methods, including Anycast-RP [RFC4610].

A sending host's Designated Router encapsulates multicast packets to the RP, and a receiving host's Designated Router can forward PIM JOIN messages to the RP, in so doing forming what is known as the Rendezvous Point Tree (RPT). Optimisation of the tree may then happen once the receiving host's router is aware of the sender's IP, and a source-specific JOIN message may be sent towards it, in so doing forming the Shortest Path Tree (SPT). Unnecessary RPT paths are removed after the SPT is established.

4.2.1. Inter-domain PIM-SM, and MSDP

PIM-SM can in principle operate over any network in which the cooperating routers are configured with RPs. But in general, PIM-SM for a given domain will use an RP configured for that domain. There is thus a challenge in enabling PIM-SM to work between multiple domains, i.e. to allow an RP in one domain to learn the existence of a source in another domain, such that a receiver's router in one

domain can know to forward a PIM JOIN towards a source's Designated Router in another domain. The solution to this problem is to use an inter-RP signalling protocol known as Multicast Source Discovery Protocol (MSDP). [RFC3618].

Deployment scenarios for MSDP are given in [RFC4611]. MSDP remains an Experimental protocol since its publication in 2003. MSDP was not replicated for IPv6.

4.3. Bidirectional PIM (BIDIR-PIM)

BIDIR-PIM is detailed in [RFC5015]. In contrast to PIM-SM, it can establish bi-directional multicast forwarding trees between multicast sources and receivers.

Add more...

4.4. IPv6 PIM-SM with Embedded RP

Within a single PIM domain, PIM-SM for IPv6 works largely the same as it does for IPv4. However, the size of the IPv6 address (128 bits) allows a different mechanism for multicast routers to determine the RP for a given multicast group address. Embedded-RP [RFC3956] specifies a method to embed the unicast RP IP address in an IPv6 multicast group address, allowing routers supporting the protocol to determine the RP for the group without any prior configuration.

Embedded-RP allows PIM-SM operation across any network in which there is an end-to-end path of routers supporting the protocol. By embedding the RP address in this way, multicast for a given group can operate inter-domain without the need for an explicit source discovery protocol (i.e. without MSDP for IPv6). It would be desirable that the RP would be located close to the sender(s) in the group.

5. SSM service model protocols

5.1. Source Specific Multicast (PIM-SSM)

PIM-SSM is detailed in [RFC4607]. In contrast to PIM-SM, PIM-SSM benefits from assuming that source(s) are known about in advance, i.e. the source IP address is known (by some out of band mechanism), and thus the receiver's router can send a PIM JOIN directly towards the sender, without needing to use an RP.

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 IPv6, the address prefix FF3x::/32 is reserved for source-specific multicast use.

6. Discussion

In this section we discuss the applicability of the ASM and SSM models described above, and their associated protocols, to a range of deployment scenarios. The context is framed in a campus / enterprise environment, but the draft could broaden its scope to other environments (thoughts?).

6.1. ASM Deployment

PIM-DM remains an Experimental protocol, that appears to be rarely used in campus or enterprise environments. Open question: what are the use cases for PIM-DM today?

In campus scenarios, PIM-SM is in common use. The configuration and management of an RP is not onerous. However, if interworking with external PIM domains in IPv4 multicast deployments is needed, MSDP is required to exchange information between domain RPs about sources. MSDP remains an Experimental protocol, and can be a complex and fragile protocol to administer and troubleshoot. MSDP is also specific to IPv4; it was not carried forward to IPv6.

PIM-SM is a general purpose protocol that can handle all use cases. In particular, it is well-suited to cases where one or more sources may come and go during a multicast session. For cases where a single, persistent source is used, PIM-SM has unnecessary complexity.

As stated above, MSDP was not taken forward to IPv6. Instead, IPv6 has Embedded-RP, which allows the RP address for a multicast group to be embedded in the group address, making RP discovery automatic, if all routers on the path between a receiver and a sender support the protocol. Embedded-RP is well-suited for lightweight ad-hoc deployments. However, it does rely on a single RP for an entire group. Embedded-RP was run successfully between European and US academic networks during the 6NET project in 2004/05. Its usage generally remains constrained to academic networks.

BIDIR-PIM is designed, as the name suggests, for bidirectional use cases.

6.2. SSM Deployment

As stated in RFC4607, SSM is particularly well-suited to dissemination-style applications with one or more senders whose identities are known (by some mechanism) before the application

begins. PIM-SSM is therefore very well-suited to applications such as IP TV.

Some benefits of PIM-SSM are presented in RFC 4607:

"Elimination of cross-delivery of traffic when two sources simultaneously use the same source-specific destination address;

Avoidance of the need for inter-host coordination when choosing source-specific addresses, as a consequence of the above;

Avoidance of many of the router protocols and algorithms that are needed to provide the ASM service model."

A significant benefit of SSM is its reduced complexity through eliminating network-based source discovery. This means no RPs, shared trees, SPT switchover, PIM registers, MSDP or data-driven state creation. It is really just a small subset of PIM-SM, plus IGMPv3. This makes it radically simpler to manage, troubleshoot and operate.

SSM is considered more secure in that it supports access control, i.e. you only get packets from the sources you explicitly ask for, as opposed to ASM where anyone can decide to send traffic to a PIM-SM group address.

It is often thought that ASM is required for multicast applications where there are multiple sources. However, RFC4607 also describes how SSM can be used instead of PIM-SM for multi-party applications:

"SSM can be used to build multi-source applications where all participants' identities are not known in advance, but the multi-source "rendezvous" functionality does not occur in the network layer in this case. Just like in an application that uses unicast as the underlying transport, this functionality can be implemented by the application or by an application-layer library."

A disadvantage of SSM is that it requires hosts using SSM and (edge) routers with SSM receivers to support the new(er) IGMPv3 and MLDv2 protocols. The slow delivery of support in some OSes has meant that adoption of SSM has also been slower than might have been expected, or hoped.

6.3. Other considerations

6.3.1. Scalability, and multicast domains

One of the challenges in wider-scale multicast deployment is its scalability, if it is expected that multicast-enabled routers are required to hold state for large numbers of multicast sources/groups.

In practice, the number of groups a given router needs to hold state for is limited by the propagation of the multicast messages for any given group, e.g. because only a specific connected set of routers are multicast-enabled, or because multicast scope borders have been configured between multicast-enabled routers for access control purposes. Further, protocol policy/filters are typically used to limit state, as well as access control.

IPv4 multicast has no explicit indication of scope boundaries within its multicast address format. The prefix 239.0.0.0/8 is reserved for private use within a network, as per [RFC2365], and is believed to be in common usage. Other scopes within this range are defined, e.g. Organizational Local Scope, but whether this is in common use is unclear.

In contrast, IPv6 has specific flag bits reserved to indicate the scope of an address, e.g. link (0x2), site (0x5), organisation (0x8) or global (0xe), as described in [RFC7346]. Such explicit scoping makes configuration of scope boundaries a simpler, cleaner process.

6.3.2. Reliable multicast

Do we want to go here, and if so which protocols should we mention? FLUTE [RFC6726] might be one example.

6.3.3. Inter-domain multicast peering

Interdomain peering best practices are documented in [I-D.ietf-mboned-interdomain-peering-bcp].

6.3.4. Layer 2 multicast domains

Open question - do we want to look at L2 models, e.g. as might be applied at an IXP?

6.3.5. Anything else?

Anything else to add here?

7. Use case examples

Aim to add 2-3 deployment examples here, if deemed useful. Perhaps one PIM-SM/MSDP/Anycast-RP, one Embedded-RP, one SSM?

8. Conclusions

Do we wish to make a very strong recommendation here for the SSM service model, and thus for PIM-SSM, even in multi-source applications?

Is this document Informational or BCP? Currently assumed Informational.

9. Security Considerations

Do we need general text on multicast security here, or not?

10. IANA Considerations

This document currently makes no request of IANA.

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

11. Acknowledgments

TBC if draft progresses...

12. References

12.1. Normative References

- [RFC1112] Deering, S., "Host extensions for IP multicasting", STD 5, RFC 1112, DOI 10.17487/RFC1112, August 1989, <<http://www.rfc-editor.org/info/rfc1112>>.
- [RFC2236] Fenner, W., "Internet Group Management Protocol, Version 2", RFC 2236, DOI 10.17487/RFC2236, November 1997, <<http://www.rfc-editor.org/info/rfc2236>>.
- [RFC2365] Meyer, D., "Administratively Scoped IP Multicast", BCP 23, RFC 2365, DOI 10.17487/RFC2365, July 1998, <<http://www.rfc-editor.org/info/rfc2365>>.
- [RFC2375] Hinden, R. and S. Deering, "IPv6 Multicast Address Assignments", RFC 2375, DOI 10.17487/RFC2375, July 1998, <<http://www.rfc-editor.org/info/rfc2375>>.

- [RFC2710] Deering, S., Fenner, W., and B. Haberman, "Multicast Listener Discovery (MLD) for IPv6", RFC 2710, DOI 10.17487/RFC2710, October 1999, <<http://www.rfc-editor.org/info/rfc2710>>.
- [RFC3307] Haberman, B., "Allocation Guidelines for IPv6 Multicast Addresses", RFC 3307, DOI 10.17487/RFC3307, August 2002, <<http://www.rfc-editor.org/info/rfc3307>>.
- [RFC3376] Cain, B., Deering, S., Kouvelas, I., Fenner, B., and A. Thyagarajan, "Internet Group Management Protocol, Version 3", RFC 3376, DOI 10.17487/RFC3376, October 2002, <<http://www.rfc-editor.org/info/rfc3376>>.
- [RFC3618] Fenner, B., Ed. and D. Meyer, Ed., "Multicast Source Discovery Protocol (MSDP)", RFC 3618, DOI 10.17487/RFC3618, October 2003, <<http://www.rfc-editor.org/info/rfc3618>>.
- [RFC3810] Vida, R., Ed. and L. Costa, Ed., "Multicast Listener Discovery Version 2 (MLDv2) for IPv6", RFC 3810, DOI 10.17487/RFC3810, June 2004, <<http://www.rfc-editor.org/info/rfc3810>>.
- [RFC3956] Savola, P. and B. Haberman, "Embedding the Rendezvous Point (RP) Address in an IPv6 Multicast Address", RFC 3956, DOI 10.17487/RFC3956, November 2004, <<http://www.rfc-editor.org/info/rfc3956>>.
- [RFC3973] Adams, A., Nicholas, J., and W. Siadak, "Protocol Independent Multicast - Dense Mode (PIM-DM): Protocol Specification (Revised)", RFC 3973, DOI 10.17487/RFC3973, January 2005, <<http://www.rfc-editor.org/info/rfc3973>>.
- [RFC4291] Hinden, R. and S. Deering, "IP Version 6 Addressing Architecture", RFC 4291, DOI 10.17487/RFC4291, February 2006, <<http://www.rfc-editor.org/info/rfc4291>>.
- [RFC4607] Holbrook, H. and B. Cain, "Source-Specific Multicast for IP", RFC 4607, DOI 10.17487/RFC4607, August 2006, <<http://www.rfc-editor.org/info/rfc4607>>.
- [RFC4610] Farinacci, D. and Y. Cai, "Anycast-RP Using Protocol Independent Multicast (PIM)", RFC 4610, DOI 10.17487/RFC4610, August 2006, <<http://www.rfc-editor.org/info/rfc4610>>.

- [RFC5015] Handley, M., Kouvelas, I., Speakman, T., and L. Vicisano, "Bidirectional Protocol Independent Multicast (BIDIR-PIM)", RFC 5015, DOI 10.17487/RFC5015, October 2007, <<http://www.rfc-editor.org/info/rfc5015>>.
- [RFC5771] Cotton, M., Vegoda, L., and D. Meyer, "IANA Guidelines for IPv4 Multicast Address Assignments", BCP 51, RFC 5771, DOI 10.17487/RFC5771, March 2010, <<http://www.rfc-editor.org/info/rfc5771>>.
- [RFC6726] Paila, T., Walsh, R., Luby, M., Roca, V., and R. Lehtonen, "FLUTE - File Delivery over Unidirectional Transport", RFC 6726, DOI 10.17487/RFC6726, November 2012, <<http://www.rfc-editor.org/info/rfc6726>>.
- [RFC7346] Droms, R., "IPv6 Multicast Address Scopes", RFC 7346, DOI 10.17487/RFC7346, August 2014, <<http://www.rfc-editor.org/info/rfc7346>>.
- [RFC7761] Fenner, B., Handley, M., Holbrook, H., Kouvelas, I., Parekh, R., Zhang, Z., and L. Zheng, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", STD 83, RFC 7761, DOI 10.17487/RFC7761, March 2016, <<http://www.rfc-editor.org/info/rfc7761>>.

12.2. Informative References

- [RFC4541] Christensen, M., Kimball, K., and F. Solensky, "Considerations for Internet Group Management Protocol (IGMP) and Multicast Listener Discovery (MLD) Snooping Switches", RFC 4541, DOI 10.17487/RFC4541, May 2006, <<http://www.rfc-editor.org/info/rfc4541>>.
- [RFC4604] Holbrook, H., Cain, B., and B. Haberman, "Using Internet Group Management Protocol Version 3 (IGMPv3) and Multicast Listener Discovery Protocol Version 2 (MLDv2) for Source-Specific Multicast", RFC 4604, DOI 10.17487/RFC4604, August 2006, <<http://www.rfc-editor.org/info/rfc4604>>.
- [RFC4611] McBride, M., Meylor, J., and D. Meyer, "Multicast Source Discovery Protocol (MSDP) Deployment Scenarios", BCP 121, RFC 4611, DOI 10.17487/RFC4611, August 2006, <<http://www.rfc-editor.org/info/rfc4611>>.

[I-D.ietf-mboned-interdomain-peering-bcp]

Tarapore, P., Sayko, R., Shepherd, G., Eckert, T., and R. Krishnan, "Use of Multicast Across Inter-Domain Peering Points", draft-ietf-mboned-interdomain-peering-bcp-03 (work in progress), May 2016.

Authors' Addresses

Mikael Abrahamsson
T-Systems
Stockholm
Sweden

Email: mikael.abrahamsson@t-systems.se

Tim Chown
Jisc
Lumen House, Library Avenue
Harwell Oxford, Didcot OX11 0SG
United Kingdom

Email: tim.chown@jisc.ac.uk

Lenny Giuliano
Juniper Networks, Inc.
2251 Corporate Park Drive
Hemdon, Virginia 20171
United States

Email: lenny@juniper.net

MBONED Working Group
Internet-Draft
Intended status: Best Current Practice
Expires: May 3, 2018

P. Tarapore, Ed.
R. Sayko
AT&T
G. Shepherd
Cisco
T. Eckert, Ed.
Huawei
R. Krishnan
SupportVectors
October 30, 2017

Use of Multicast Across Inter-Domain Peering Points
draft-ietf-mboned-interdomain-peering-bcp-14

Abstract

This document examines the use of Source Specific Multicast (SSM) across inter-domain peering points for a specified set of deployment scenarios. The objective is to describe the setup process for multicast-based delivery across administrative domains for these scenarios and document supporting functionality to enable this process.

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 <https://datatracker.ietf.org/drafts/current/>.

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

This Internet-Draft will expire on May 3, 2018.

Copyright Notice

Copyright (c) 2017 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

(<https://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. Overview of Inter-domain Multicast Application Transport . .	5
3. Inter-domain Peering Point Requirements for Multicast	6
3.1. Native Multicast	7
3.2. Peering Point Enabled with GRE Tunnel	8
3.3. Peering Point Enabled with an AMT - Both Domains Multicast Enabled	10
3.4. Peering Point Enabled with an AMT - AD-2 Not Multicast Enabled	12
3.5. AD-2 Not Multicast Enabled - Multiple AMT Tunnels Through AD-2	14
4. Functional Guidelines	16
4.1. Network Interconnection Transport Guidelines	16
4.1.1. Bandwidth Management	16
4.2. Routing Aspects and Related Guidelines	18
4.2.1. Native Multicast Routing Aspects	19
4.2.2. GRE Tunnel over Interconnecting Peering Point	19
4.2.3. Routing Aspects with AMT Tunnels	20
4.2.4. Public Peering Routing Aspects	22
4.3. Back Office Functions - Provisioning and Logging Guidelines	23
4.3.1. Provisioning Guidelines	24
4.3.2. Interdomain Authentication Guidelines	25
4.3.3. Log Management Guidelines	26
4.4. Operations - Service Performance and Monitoring Guidelines	27
4.5. Client Reliability Models/Service Assurance Guidelines .	29
4.6. Application Accounting Guidelines	29
5. Troubleshooting and Diagnostics	29
6. Security Considerations	30
6.1. DoS attacks (against state and bandwidth)	30
6.2. Content Security	32
6.3. Peering Encryption	34
6.4. Operational Aspects	34
7. Privacy Considerations	35
8. IANA Considerations	37
9. Acknowledgments	37
10. Change log [RFC Editor: Please remove]	37

11. References	39
11.1. Normative References	39
11.2. Informative References	40
Authors' Addresses	41

1. Introduction

Content and data from 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 content or other data offers significant savings of utilization of resources in any given administrative domain. End user demand for such content or other data is growing. Often, this requires transporting the content or other data across administrative domains via inter-domain peering points.

The objective of this Best Current Practices document is twofold:

- o Describe the technical process and establish guidelines for setting up multicast-based delivery of application content or other data across inter-domain peering points via a set of use cases.
- o Catalog all required information exchange between the administrative domains to support multicast-based delivery. This enables operators to initiate necessary processes to support inter-domain peering with multicast.

The scope and assumptions for this document are as follows:

- o Administrative Domain 1 (AD-1) sources content to one or more End Users (EUs) in one or more Administrative Domain 2 (AD-2). AD-1 and AD-2 want to use IP multicast to allow supporting large and growing EU populations with minimum amount of duplicated traffic to send across network links.
- o This document does not detail the case where EUs are originating content. To support that additional service, it is recommended to use some method (outside the scope of this document) by which the content from EUs is transmitted to the application in AD-1 that this document refers to as the multicast source and let it send out the traffic as IP multicast. From that point on, the descriptions in this document apply, except that they are not complete because they do not cover the transport or operational aspects of the leg from EU to AD-1.

- o This document does not detail the case where AD-1 and AD-2 are not directly connected to each other but only via one or more AD-3 (transit providers). The cases described in this document where tunnels are used between AD-1 and AD-2 can be applied to such scenarios, but SLA ("Service Level Agreement") control for example would be different. Other additional issues will likely exist as well in such scenarios. This is for further study.
- o For the purpose of this document, the term "peering point" refers to a network connection ("link") between two administrative network domains over which traffic is exchanged between them. This is also referred to as a Network-to-Network Interface (NNI). Unless otherwise noted, the peering point is assumed to be a private peering point, where the network connection is a physically or virtually isolated network connection solely between AD-1 and AD-2. The other case is that of a broadcast peering point which is a common option in public Internet Exchange Points (IXP). See Section 4.2.2 for more details about that option.
- o Administrative Domain 1 (AD-1) is enabled with native multicast. A peering point exists between AD-1 and AD-2.
- o It is understood that several protocols are available for this purpose including PIM-SM and Protocol Independent Multicast - Source Specific Multicast (PIM-SSM) [RFC7761], Internet Group Management Protocol (IGMP) [RFC3376], and Multicast Listener Discovery (MLD) [RFC3810].
- o As described in Section 2, the source IP address of the multicast stream in the originating AD (AD-1) is known. Under this condition, PIM-SSM use is beneficial as it allows the receiver's upstream router to directly send a JOIN message to the source without the need of invoking an intermediate Rendezvous Point (RP). Use of SSM also presents an improved threat mitigation profile against attack, as described in [RFC4609]. Hence, in the case of inter-domain peering, it is recommended to use only SSM protocols; the setup of inter-domain peering for ASM (Any-Source Multicast) is not in scope for this document.
- o The rest of the document assumes that PIM-SSM and BGP are used across the peering point plus AMT and/or GRE according to scenario. The use of other protocols is beyond the scope of this document.
- o An Automatic Multicast Tunnel (AMT) [RFC7450] is setup at the peering point if either the peering point or AD-2 is not multicast enabled. It is assumed that an AMT Relay will be available to a

client for multicast delivery. The selection of an optimal AMT relay by a client is out of scope for this document. Note that AMT use is necessary only when native multicast is unavailable in the peering point (Use Case 3.3) or in the downstream administrative domain (Use Cases 3.4, and 3.5).

- o The collection of billing data is assumed to be done at the application level and is not considered to be a networking issue. The settlements process for end user billing and/or inter-provider billing is out of scope for this document.
- o Inter-domain network connectivity troubleshooting is only considered within the context of a cooperative process between the two domains.

This document also attempts to identify ways by which the peering process can be improved. Development of new methods for improvement is beyond the scope of this document.

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) [RFC7450].
- o A service provider controls one or more application sources in AD-1 which will send multicast IP packets via one or more (S,G)s (multicast traffic flows, see Section 4.2.1 if you are unfamiliar with IP multicast). It is assumed that the service being provided is suitable for delivery via multicast (e.g. live video streaming of popular events, software downloads to many devices, etc.), and that the packet streams will be carried by a suitable multicast transport protocol.
- o An End User (EU) controls a device connected to AD-2, which runs an application client compatible with the service provider's application source.

- o The application client joins appropriate (S,G)s in order to receive the data necessary to provide the service to the EU. The mechanisms by which the application client learns the appropriate (S,G)s are an implementation detail of the application, and are out of scope for this document.

The assumption here is that AD-1 has ultimate responsibility for delivering the multicast based service on behalf of the content source(s). All relevant interactions between the two domains described in this document are based on this assumption.

Note that domain 2 may be an independent network domain (e.g.: Tier 1 network operator domain). Alternately, domain 2 could also be an Enterprise network domain operated by a single customer of AD-1. 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. Section 4 contains a comprehensive list of pertinent information that needs to be exchanged between the two domains in order to support functions to enable the application transport.

Note that domain 2 may be an independent network domain (e.g., Tier 1 network operator domain). Alternately, domain 2 could also be an Enterprise network domain operated by a single customer.

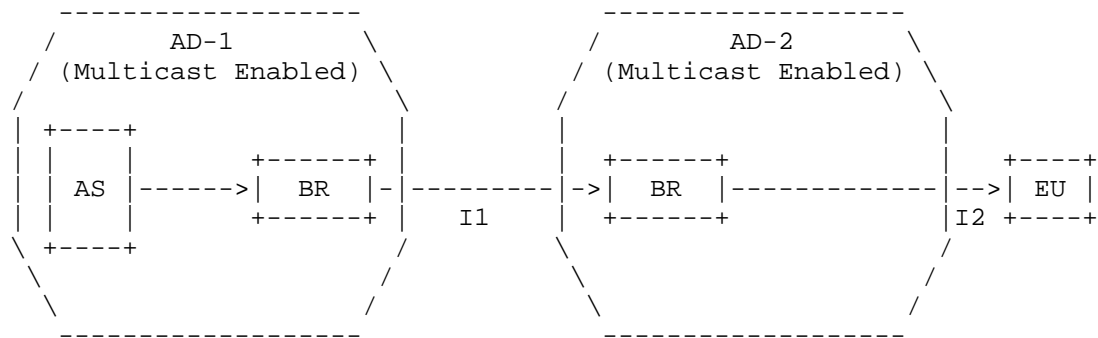
The Use Cases describing various architectural configurations for the multicast distribution along with associated requirements is described in Section 3. The peering point architecture and requirements may have some unique aspects associated with the Enterprise case. These unique aspects will also be described in Section 3. Section 4 contains a comprehensive list of pertinent information that needs to be exchanged between the two domains in order to support functions to enable the application transport.

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 five Use Cases for consideration in this document.

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 - see 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 (e.g., MBGP)
 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 an AMT enabled peering point.

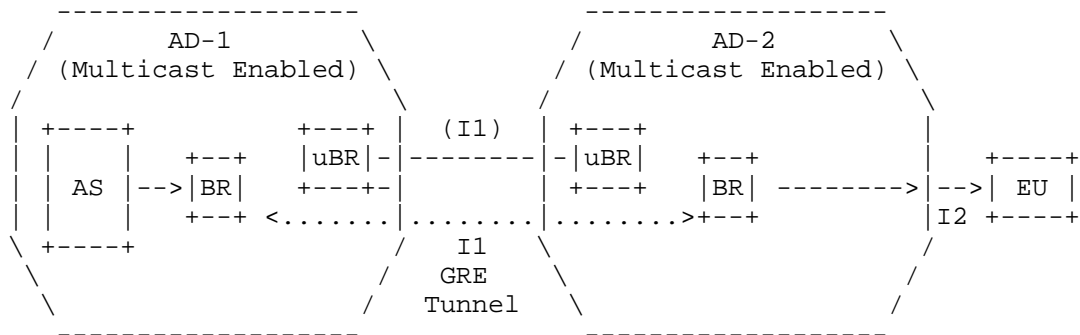
From the perspective of AD-1, the one disadvantage associated with native multicast into AD-2 instead of individual unicast to every EU in AD-2 is that it does not have the ability to count the number of End Users as well as the transmitted bytes delivered to them. This information is relevant from the 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:

- a. Dual homing for peering points between domains is recommended as a way to ensure reliability with full BGP table visibility.
- b. 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 the multicast traffic must support rate-adaption (see [BCP145]).
- c. 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. Another example is the use of a policy by which AD-1 delivers specified content to AD-2 only if such delivery has been accepted by contract.
- d. 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. See Figure 2.



AD = Administrative Domain (Independent Autonomous System)
 AS = Application (e.g., Content) Multicast Source
 uBR = unicast Border Router - not necessarily multicast enabled
 may be the same router as BR
 BR = Border Router - for multicast
 I1 = AD-1 and AD-2 Multicast Interconnection (e.g., MBGP)
 I2 = AD-2 and EU Multicast Connection

Figure 2: Content Distribution via GRE Tunnel

In this case, the interconnection I1 between AD-1 and AD-2 in Figure 2 is multicast enabled via a Generic Routing Encapsulation Tunnel (GRE) [RFC2784] between the two BR and encapsulating the multicast protocols across it.

Normally, this approach is chosen if the uBR physically connected to the peering link can or should not be enabled for IP multicast. This approach may also be beneficial if BR and uBR are the same device, but the peering link is a broadcast domain (IXP), see Figure 6.

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 an AMT enabled peering point.
- o Ability to support partial and/or incremental IP multicast deployments in AD-1 and/or AD-2: Only the path(s) between AS/BR (AD-1) and BR/EU (AD-2) need to be multicast enabled. The uBRs

may not support IP multicast or enabling it could be seen as operationally risky on that important edge node whereas dedicated BR nodes for IP multicast may be more acceptable at least initially. BR can also be located such that only parts of the domain may need to support native IP multicast (e.g.: only the core in AD-1 but not edge networks towards uBR).

- 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 The GRE must be established prior to stream starting.
- o The GRE tunnel is often left pinned up.

Architectural guidelines for this configuration include the following:

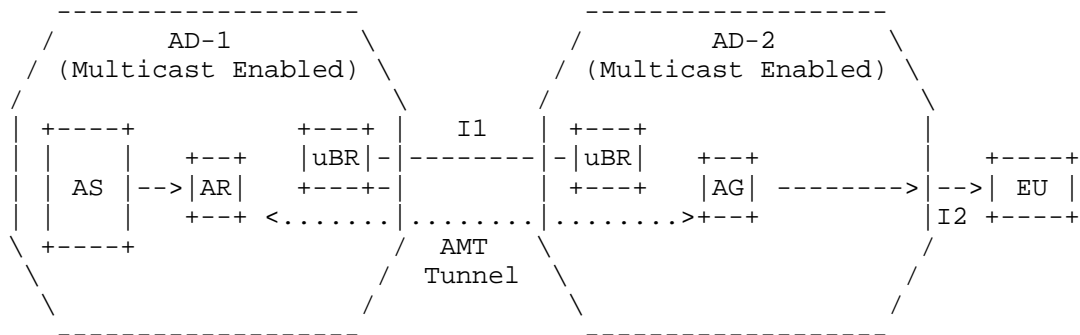
Guidelines (a) through (d) are the same as those described in Use Case 3.1. Two additional guidelines are as follows:

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



AD = Administrative Domain (Independent Autonomous System)
AS = Application (e.g., Content) Multicast Source
AR = AMT Relay
AG = AMT Gateway
uBR = unicast Border Router - not multicast enabled
otherwise AR=uBR (AD-1), uBR=AG (AD-2)
I1 = AMT Interconnection between AD-1 and AD-2
I2 = AD-2 and EU Multicast Connection

Figure 3: - 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 of bytes transmitted to all end users.
- 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 for the AG to automatically select the optimal AR.

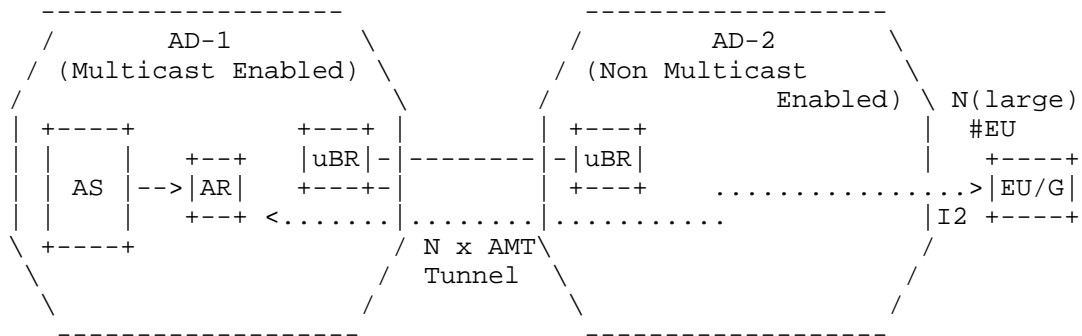
Architectural guidelines for this configuration are as follows:

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

- 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. Hence, the interconnection between AD-2 and the End User is also not multicast enabled. This Use Case is depicted in Figure 3.



AS = Application Multicast Source
uBR = unicast Border Router - not multicast enabled,
otherwise AR = uBR (in AD-1).
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 4: 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 Efficient use of bandwidth in AD-1 (The closer AR is to uBR, the more efficient).
- o Ability for AD-1 to introduce IP multicast based content delivery without any support by network devices in AD-2: Only application side in the EU device needs to perform AMT gateway library functionality to receive traffic from AMT relay.
- o Allows for AD-2 to "upgrade" to Use Case 3.5 (see below) at a later time without any change in AD-1 at that time.
- 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:

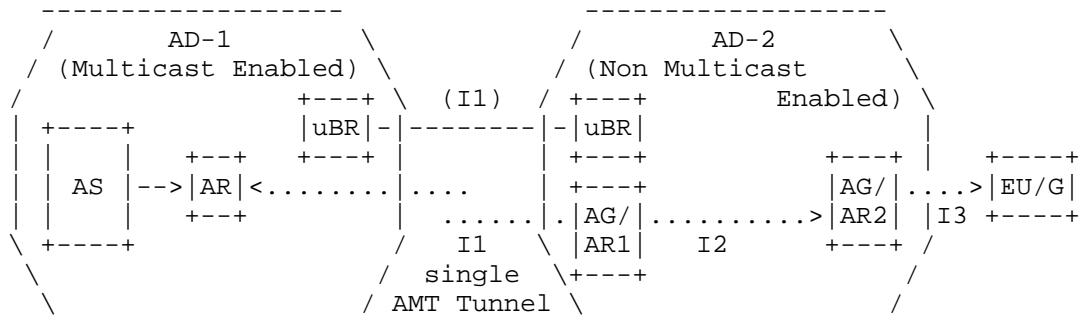
Guidelines (a) through (c) are the same as those described in Use Case 3.1.

- d. It is necessary that proper procedures are implemented such that the AMT Gateway at the End User device is able to find the correct AMT Relay for each (S,G) content stream. Standard mechanisms for that selection are still subject to ongoing work. This includes use of anycast gateway addresses, anycast DNS names, explicit configuration that is mapping (S,G) to a relay address or letting the application in the EU/G provide the relay address to the embedded AMT gateway function.

- 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:



```
uBR = unicast Border Router - not multicast enabled
otherwise AR=uBR (AD-1) or uBR=AGAR1 (AD-2)
```

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 5: AMT Tunnel Connecting AMT Relay and Relays

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 high amount of traffic due to the 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 in AD-2 AD-2. An AMT node comprises co-location of an AMT Gateway and an AMT Relay. No change is required by AD-1 compared to 3.4. This can be done whenever AD-2 seems fit (too much traffic across peering point).
- o One such node is at the AD-2 side of the peering point (node AGAR1 in above Figure).

- 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.
- o In the most simple option (not shown), AD-2 only deploys a single AGAR1 and lets EU/G build AMT tunnels directly to it. This setup already solves the problem of replicated traffic across the peering point. As soon as there is need to support more AMT tunnels to EU/G, then additional AGAR2 nodes can be deployed by AD-2.

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 necessary 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. Standard mechanisms for that selection are still subject to ongoing work. This includes use of anycast gateway addresses, anycast DNS names, or explicit configuration that is mapping (S,G) to a relay address. On the EU/G, this mapping information may come from the application.
- 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. Functional Guidelines

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 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 [RFC4760] and MBGP [RFC4760].
- o Bandwidth Allocation - If shared with other services, then there needs to be a determination of the share of bandwidth reserved for multicast delivery. See section 4.1.1 below for more details.
- o QoS Requirements - Delay and/or 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.

4.1.1. Bandwidth Management

Like IP unicast traffic, IP multicast traffic carried across non-controlled networks must comply to Congestion Control Principles as

described in [BCP41] and explained in detail for UDP IP multicast in [BCP145].

Non-controlled networks (such as the Internet) are those where there is no policy for managing bandwidth other than best effort with fair share of bandwidth under congestion. As a simplified rule of thumb, complying to congestion control principles means to reduce bandwidth under congestion in a way that is fair to competing competing (typically TCP) flow ("rate adaptive").

In many instances, multicast content delivery evolves from intra-domain deployments where it is handled as a controlled network service and of not complying to congestion control principles. It was given a reserved amount of bandwidth and admitted to the network so that congestion never occurs. Therefore the congestion control issue should be given specific attention when evolving to an interdomain peering deployment.

In the case where end-to-end IP multicast traffic passes across the network of two ADs (and their subsidiaries/customers), both ADs must agree on a consistent traffic management policy. If for example AD-1 sources non congestion aware IP multicast traffic and AD-2 carries it as best effort traffic across links shared with other Internet traffic and subject to congestion, this will not work: Under congestion, some amount of that traffic will be dropped, rendering the remaining packets often as undecodeable garbage clogging up the network in AD-2 and because this is not congestion aware, the loss does not reduce this rate. Competing traffic will not get their fair share under congestion, and EUs will be frustrated by extremely bad quality of both their IP multicast and other (e.g.: TCP) traffic. Note that this is not an IP multicast technology issue, but solely a transport/application layer issue: The problem would equally happen if AD-1 would send non-rate adaptive unicast traffic,, for example legacy IPTV video-on-demand traffic which typically is also non congestion aware. Because rate adaption in IP unicast video is commonplace today because of ABR (Adaptive Bitrate Video), it is very unlikely for this to happen though in reality with IP unicast.

While the rules for traffic management apply whether or not IP multicast is tunneled or not, the one feature that can make AMT tunnels more difficult is the unpredictability of bandwidth requirements across underlying links because of the way they can be used: With native IP multicast or GRE tunnels, the amount of bandwidth depends on the amount of content, not the number of EUs - and is therefore easier to plan for. AMT tunnels terminating in EU/G on the other hand scale with the number of EUs. In the vicinity of the AMT relay they can introduce very large amount of replicated traffic and it is not always feasible to provision enough bandwidth

for all possible EU to get the highest quality for all their content during peak utilization in such setups - unless the AMT relays are very close to the EU edge. Therefore it is also recommended to use IP multicast rate adaptation even inside controlled networks when using AMT tunnels directly to EU/G.

Note that rate-adaptive IP multicast traffic in general does not mean that the sender is reducing the bitrate, but rather that the EUs that experience congestion are joining to a lower bitrate (S,G) stream of the content, similar to adaptive bitrate streaming over TCP. Migration from non rate-adaptive to rate adaptive bitrate in IP multicast does therefore also change the dynamic (S,G) join behavior in the network resulting in potentially higher performance requirement for IP multicast protocols (IGMP/PIM), especially on the last hops where dynamic changes occur (including AMT gateway/relays): In non rate-adaptive IP multicast, only "channel change" causes state change, in rate-adaptive also the congestion situation causes state change.

Even though not fully specified in this document, peerings that rely on GRE/AMT tunnels may be across one or more transit ADs instead of an exclusive (non-shared, L1/L2) path. Unless those transit ADs are explicitly contracted to provide other than "best effort" transit for the tunneled traffic, the IP multicast traffic tunneled must be rate adaptive to not violate BCP41 across those transit ADs.

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:

- o To be scalable,
- o To avoid or minimize new protocol development or modifications, and
- o To be robust enough to achieve high reliability and automatically adjust to changes and 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. However, 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 information is in the form of metadata and it contains instructions directing the EU client to launch an appropriate application if necessary, as well as 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 metadata 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 [RFC4604]. 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.
 - o Implement compatible multicast protocols to ensure proper multicast delivery across the peering points.

4.2.2. GRE Tunnel over Interconnecting Peering Point

If the interconnecting peering point is not multicast enabled and both AD's are multicast enabled, then a simple solution is to provision a GRE tunnel between the two AD's - 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 Multicast (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 met:

- 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

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 End-Point within the unicast-only network.

AMT Gateway (GW): The Gateway will reside on an End-Point - this could be any type of IP host such as a Personal Computer (PC), mobile phone, Set Top Box (STB) or appliances. The AMT Gateway receives join and leave requests from the Application via an Application Programming Interface (API). In this manner, the Gateway allows the End-Point 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 AD's. 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 AD's can coordinate and advertise special AMT Relay Anycast addresses with each other. Alternately, they may decide to

simply provision Relay addresses, though this would not be an optimal solution in terms of scalability.

Use Cases 3.4 and 3.5 describe more complicated AMT situations as AD-2 is not multicast enabled. For these cases, the End User device needs to be able to setup an AMT tunnel in the most optimal manner. There are many methods by which relay selection can be done including the use of DNS based queries and static lookup tables [RFC7450]. The choice of the method is implementation dependent and is up to the network operators. Comparison of various methods is out of scope for this document; it is for further study.

An illustrative example of a relay selection based on DNS queries and Anycast IP addresses process for Use Cases 3.4 and 3.5 is described here. 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. Note that this is strictly illustrative; the choice of the method is up to the network operators. The basic process is as follows:

- o Appropriate metadata is obtained by the EU client application. The metadata 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 metadata 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 from the metadata, 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. Details are outside the scope for this document. See [RFC4786].

forwarded by only one router onto the LAN, and PIM-SM/PIM-SSM detects requests for duplicate transmission and resolve it via the so-called "assert" protocol operation which results in only one BR forwarding the traffic. Assume this is AD-1a BR. AD-2b will then receive the multicast traffic unexpectedly from a provider with whom it does not have a mutual agreement for the traffic. Quality issues in EUs behind AD-2b caused by AD-1a will cause a lot of responsibility and troubleshooting issues.

In face of this technical issues, we describe the following options how IP multicast can be carried across broadcast peering point LANs:

1. IP multicast is tunneled across the LAN. Any of the GRE/AMT tunneling solutions mentioned in this document are applicable. This is the one case where specifically a GRE tunnel between the upstream BR (e.g.: AD-1a) and downstream BR (e.g.: AD-2a) is recommended as opposed to tunneling across uBRs which are not the actual BRs.
2. The LAN has only one upstream AD that is sourcing IP multicast and native IP multicast is used. This is an efficient way to distribute the same IP multicast content to multiple downstream ADs. Misbehaving downstream BRs can still disrupt the delivery of IP multicast from the upstream BR to other downstream BRs, therefore strict rules must be followed to prohibit that case. The downstream BRs must ensure that they will always consider only the upstream BR as a source for multicast traffic: e.g.: no BGP SAFI-2 peerings between the downstream ADs across the peering point LAN, so that only the upstream BR is the only possible next-hop reachable across this LAN. And routing policies configured to avoid fall back to the use of SAFI-1 (unicast) routes for IP multicast if unicast BGP peering is not limited in the same way.
3. The LAN has multiple upstreams, but they are federated and agree on a consistent policy for IP multicast traffic across the LAN. One policy is that each possible source is only announced by one upstream BR. Another policy is that sources are redundantly announced (problematic case mentioned in above example), but the upstream domains also provide mutual operational insight to help troubleshooting (outside the scope of this document).

4.3. Back Office Functions - Provisioning 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 AD's.

- o Functionality associated with logging, reporting, ordering, provisioning, maintenance, service assurance, settlement, etc.

4.3.1. Provisioning Guidelines

Resources for basic connectivity between AD's Providers need to be provisioned as follows:

- o Sufficient capacity must be provisioned to support multicast-based delivery across AD's.
- o Sufficient capacity must be provisioned for connectivity between all supporting back-offices of the AD's 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 receipt of the necessary metadata. Hence, this metadata must be provided to the EU regarding multicast URL - and unicast fallback if applicable. AD-2 must enable the delivery of this metadata to the EU and provision appropriate resources for this purpose.

Native multicast functionality is assumed to be available 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's to support this option.

4.3.2. Interdomain Authentication Guidelines

All interactions between pairs of AD's 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 credentials such as login passwords) for use by AD-1. Multiple accounts and multiple types or partitions of accounts can apply, e.g. customer accounts, security accounts, etc.

The reason to specifically mention the need for AD-1 to initiate interactions with AD-2 (and use some account for that), as opposed to the opposite direction is based on the recommended workflow initiated by customers (see Section 4.4): The customer contacts content source (part of AD-1), when AD-1 sees the need to propagate the issue, it will interact with AD-2 using the aforementioned guidelines.

4.3.3. Log Management Guidelines

Successful delivery (in terms of user experience) of applications or content via multicast between pairs of interconnecting AD's can be improved through the ability to exchange appropriate logs for various workflows - troubleshooting, accounting and billing, traffic and content transmission optimization, content and application development optimization and so on.

The basic model as explained in before is that the content source and on its behalf AD-1 take over primary responsibility for customer experience and the AD-2's support this. The application/content owner is the only participant who has and needs full insight into the application level and can map the customer application experience to the network traffic flows - which it then with the help of AD-2 or logs from AD-2 can analyze and interpret.

The main difference between unicast delivery and multicast delivery is that the content source can infer a lot more about downstream network problems from a unicasted stream than from a multicasted stream: The multicasted stream is not per-EU except after the last replication, which is in most cases not in AD-1. Logs from the application, including the receiver side at the EU, can provide insight, but can not help to fully isolate network problems because of the IP multicast per-application operational state built across AD-1 and AD-2 (aka: the (S,G) state and any other feature operational state such as DiffServ QoS).

See Section 7 for more discussions about the privacy considerations of the model described here.

Different type of logs are known to help support operations in AD-1 when provided by AD-2. This could be done as part of AD-1/AD-2 contracts. Note that except for implied multicast specific elements, the options listed here are not unique or novel for IP multicast, but they are more important for services novel to the operators than for operationally well established services (such as unicast). Therefore we detail them as follows:

- o Usage information logs at aggregate level.
- o Usage failure instances at an aggregate level.
- o Grouped or sequenced application access. performance, behavior and 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 AD's 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).

4.4. 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.

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. EUs 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 Network 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.

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 or 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.

4.6. Application Accounting Guidelines

Application level accounting needs to be handled differently in the application than in IP unicast because the source side does not directly deliver packets to individual receivers. Instead, this needs to be signalled back by the receiver to the source.

For network transport diagnostics, 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.

5. Troubleshooting and Diagnostics

Any service provider supporting multicast delivery of content should have the capability to collect diagnostics as part of multicast troubleshooting practices and resolve network issues accordingly. Issues may become apparent or identified either through network monitoring functions or by customer reported problems as described in section 4.4.

It is recommended that multicast diagnostics will be performed leveraging established operational practices such as those documented in [MDH-04]. However, given that inter-domain multicast creates a significant interdependence of proper networking functionality between providers there does exist a need for providers to be able to signal (or otherwise alert) each other if there are any issues noted by either one.

Service providers may also wish to allow limited read-only administrative access to their routers to their AD peers for troubleshooting. Of specific interest are access to active troubleshooting tools especially [Traceroute] and [I-D.ietf-mboned-mtrace-v2].

Another option is to include this functionality into the IP multicast receiver application on the EU device and allow for these diagnostics to be remotely used by support operations. Note though that AMT does not allow to pass traceroute or mtrace requests, therefore troubleshooting in the presence of AMT does not work as well end-to-end as it can with native (or even GRE encapsulated) IP multicast, especially wrt. to traceroute and mtrace. Instead, troubleshooting directly on the actual network devices is then more likely necessary.

The specifics of the notification and alerts are beyond the scope of this document, but general guidelines are similar to those described in section 4.4 (Service Performance and Monitoring). Some general communications issues are stated as follows.

- o Appropriate communications channels will be established between the customer service and operations groups from both AD's to facilitate information sharing related to diagnostic troubleshooting.
- o A default resolution period may be considered to resolve open issues. Alternately, mutually acceptable resolution periods could be established depending on the severity of the identified trouble.

6. Security Considerations

6.1. DoS attacks (against state and bandwidth)

Reliable operations of IP multicast requires some basic protection against DoS (Denial of Service) attacks.

SSM IP multicast is self protecting against attacks from illicit sources. Their traffic will not be forwarded beyond the first hop router because that would require (S,G) membership reports from

receiver. Traffic from sources will only be forwarded from the valid source because RPF ("Reverse Path Forwarding") is part of the protocols. One can say that [BCP38] style protection against spoofed source traffic is therefore built into PIM-SM/PIM-SSM.

Receivers can attack SSM IP multicast by originating such (S,G) membership reports. This can result in a DoS attack against state through the creation of a large number of (S,G) states that create high control plane load or even inhibit later creation of valid (S,G). In conjunction with collaborating illicit sources it can also result in illicit sources traffic being forwarded.

Today, these type of attacks are usually mitigated by explicitly defining the set of permissible (S,G) on e.g.: the last hop routers in replicating IP multicast to EUs; For example via (S,G) Access Control Lists applied to IGMP/MLD membership state creation. Each AD is expected to prohibit (S,G) state creation for invalid sources inside their own AD.

In the peering case, AD-2 is without further information not aware of the set of valid (S,G) from AD-1, so this set needs to be communicated via operational procedures from AD-1 to AD-2 to provide protection against this type of DoS attacks. Future work could signal this information in an automated way: BGP extensions, DNS Resource Records or backend automation between AD-1 and AD-2. Backend automation is the short term most viable solution because it does not require router software extensions like the other two. Observation of traffic flowing via (S,G) state could also be used to automate recognition of invalid (S,G) state created by receivers in the absence of explicit information from AD-1.

The second DoS attack through (S,G) membership reports is when receivers create too much valid (S,G) state to attack bandwidth available to other EU. Consider the uplink into a last-hop-router connecting to 100 EU. If one EU joins to more multicast content than what fits into this link, then this would impact also the quality of the same content for the other 99 EU. If traffic is not rate adaptive, the effects are even worse.

The mitigation is the same as what is often employed for unicast: Policing of per-EU total amount of traffic. Unlike unicast though, this can not be done anywhere along the path (e.g.: on an arbitrary bottleneck link), but it has to happen at the point of last replication to the different EU. Simple solutions such as limiting the maximum number of joined (S,G) per EU are readily available, solutions that consider bandwidth consumed exist as vendor specific feature in routers. Note that this is primarily a non-peering issue in AD-2, it only becomes a peering issue if the peering-link itself

is not big enough to carry all possible content from AD-1 or in case 3.4 where the AMT relay in AD-1 is that last replication point.

Limiting the amount of (S,G) state per EU is also a good first measure to prohibit too much undesired "empty" state to be built (state not carrying traffic), but it would not suffice in case of DDoS attack - viruses that impact a large number of EU devices.

6.2. Content Security

Content confidentiality, DRM (Digital Restrictions Management), authentication and authorization are optional based on the content delivered. For content that is "FTA" (Free To Air), the following considerations can be ignored and content can be sent unencrypted and without EU authentication and authorization. Note though that the mechanisms described here may also be desirable by the application source to better track users even if the content itself would not require it.

For interdomain content, there are at least two models for content confidentiality, DRM and end-user authentication and authorization:

In the classical (IP)TV model, responsibility is per-domain and content is and can be passed on unencrypted. AD-1 delivers content to AD-2, AD-2 can further process the content including features like ad-insertion and AD-2 is the sole point of contact regarding the contact for its EUs. In this document, we do not consider this case because it typically involves higher than network layer service aspects operated by AD-2 and this document focusses on the network layer AD-1/AD-2 peering case, but not the application layer peering case. Nevertheless, this model can be derived through additional work from what is describe here.

The other case is the one in which content confidentiality, DRM, end-user authentication and authorization are end-to-end: responsibilities of the multicast application source provider and receiver application. This is the model assumed here. It is also the model used in Internet OTT video delivery. We discuss the threads incurred in this model due to the use of IP multicast in AD-1/AD-2 and across the peering.

End-to-end encryption enables end-to-end EU authentication and authorization: The EU may be able to IGMP/MLD join and receive the content, but it can only decrypt it when it receives the decryption key from the content source in AD-1. The key is the authorization. Keeping that key to itself and prohibiting playout of the decrypted content to non-copy-protected interfaces are typical DRM features in that receiver application or EU device operating system.

End-to-end encryption is continuously attacked. Keys may be subject to brute force attack so that content can be decrypted potentially later, or keys are extracted from the EU application/device and shared with other unauthenticated receivers. One important class of content is where the value is in live consumption, such as sports or other event (concert) streaming. Extraction of keying material from compromised authenticated EU and sharing with unauthenticated EU is not sufficient. It is also necessary for those unauthenticated EUs to get a streaming copy of the content itself. In unicast streaming, they can not get such a copy from the content source (because they can not authenticate) and because of asymmetric bandwidths, it is often impossible to get the content from compromised EUs to large number of unauthenticated EUs. EUs behind classical 16 Mbps down, 1 Mbps up ADSL links are the best example. With increasing broadband access speeds unicast peer-to-peer copying of content becomes easier, but it likely will always be easily detectable by the ADs because of its traffic patterns and volume.

When IP multicast is being used without additional security, AD-2 is not aware which EU is authenticated for which content. Any unauthenticated EU in AD-2 could therefore get a copy of the encrypted content without suspicion by AD-2 or AD-1 and either live-decode it in the presence of compromised authenticated EU and key sharing, or later decrypt it in the presence of federated brute force key cracking.

To mitigate this issue, the last replication point that is creating (S,G) copies to EUs would need to permit those copies only after authentication of EUs. This would establish the same authenticated EU only copy deliver that is used in unicast.

Schemes for per EU IP multicast authentication/authorization (and in result non-delivery/copying of per-content IP multicast traffic) have been built in the past and are deployed in service providers for intradomain IPTV services, but no standard exist for this. For example, there is no standardized radius attribute for authenticating the IGMP/MLD filter set, but implementations of this exist. The authors are specifically also not aware of schemes where the same authentication credentials used to get the encryption key from the content source could also be used to authenticate and authorize the network layer IP multicast replication for the content. Such schemes are technically not difficult to build and would avoid creating and maintaining a separate network forwarding authentication/authorization scheme decoupled from the end-to-end authentication/authorization system of the application.

If delivery of such high value content in conjunction with the peering described here is desired, the short term recommendations are

for sources to clearly isolate the source and group addresses used for different content bundles, communicate those (S,G) patterns from AD-1 to the AD-2 and let AD-2 leverage existing per-EU authentication/ authorization mechanisms in network devices to establish filters for (S,G) sets to each EU.

6.3. Peering Encryption

Encryption at peering points for multicast delivery may be used per agreement between AD-1/AD-2.

In the case of a private peering link, IP multicast does not have attack vectors on a peering link different from those of IP unicast, but the content owner may have defined high bars against unauthenticated copying of even the end-to-end encrypted content, and in this case AD-1/AD-2 can agree on additional transport encryption across that peering link. In the case of a broadcast peering connection (e.g.: IXP), transport encryption is also the easiest way to prohibit unauthenticated copies by other ADs on the same peering point.

If peering is across a tunnel going across intermittent transit ADs (not discussed in detail in this document), then encryption of that tunnel traffic is recommended. It not only prohibits possible "leakage" of content, but also to protects the the information what content is being consumed in AD-2 (aggregated privacy protection).

See the following subsection for reasons why the peering point may also need to be encrypted for operational reasons.

6.4. Operational Aspects

Section 4.3.3 discusses exchange of log information, this section discussed exchange of (S,G) information and Section 7 discusses exchange of program information. All these operational pieces of data should by default be exchanged via authenticated and encrypted peer-to-peer communication protocols between AD-1 and AD-2 so that only the intended recipient in the peers AD have access to it. Even exposure of the least sensitive information to third parties opens up attack vectors. Putting for example valid (S,G) information into DNS (as opposed to passing it via secured channels from AD-1 to AD-2) to allow easier filtering of invalid (S,G) would also allow attackers to easier identify valid (S,G) and change their attack vector.

From the perspective of the ADs, security is most critical for the log information as it provides operational insight into the originating AD, but it also contains sensitive user data:

Sensitive user data exported from AD-2 to AD-1 as part of logs could be as much as the equivalent of 5-tuple unicast traffic flow accounting (but not more, e.g.: no application level information). As mentioned in Section 7, in unicast, AD-1 could capture these traffic statistics itself because this is all about AD-1 originated traffic flows to EU receivers in AD-2, and operationally passing it from AD-2 to AD-1 may be necessary when IP multicast is used because of the replication happening in AD-2.

Nevertheless, passing such traffic statistics inside AD-1 from a capturing router to a backend system is likely less subject to third party attacks than passing it interdomain from AD-2 to AD-1, so more diligence needs to be applied to secure it.

If any protocols used for the operational information exchange are not easily secured at transport layer or higher (because of the use of legacy products or protocols in the network), then AD-1 and AD-2 can also consider to ensure that all operational data exchange goes across the same peering point as the traffic and use network layer encryption of the peering point as discussed in before to protect it.

End-to-end authentication and authorization of EU may involve some kind of token authentication and is done at the application layer independently of the two AD's. 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. Implementation details are beyond the scope of this document.

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

7. Privacy Considerations

The described flow of information about content and the end-user described in this document aims to maintain privacy:

AD-1 is operating on behalf (or owns) the content source and is therefore part of the content-consumption relationship with the end-user. The privacy considerations between the EU and AD-1 are therefore in general (exception see below) the same as if no IP multicast was used, especially because for any privacy conscious content, end-to-end encryption can and should be used.

Interdomain multicast transport service related information is provided by the AD-2 operators to AD-1. AD-2 is not required to gain additional insight into the user behavior through this process that it would not already have without the service collaboration with AD-1 - unless AD-1 and AD-2 agree on it and get approval from the EU.

For example, if it is deemed beneficial for EU to directly get support from AD-2 then it would in general be necessary for AD-2 to be aware of the mapping between content and network (S,G) state so that AD-2 knows which (S,G) to troubleshoot when the EU complains about problems with a specific content. The degree to which this dissemination is done by AD-1 explicitly to meet privacy expectations of EUs is typically easy to assess by AD-1. Two simple examples:

For a sports content bundle, every EU will happily click on the "i approve that the content program information is shared with your service provider" button, to ensure best service reliability because service conscious AD-2 would likely also try to ensure that high value content, such as the (S,G) for SuperBowl like content would be the first to receive care in case of network issues.

If the content in question was one where the EU expected more privacy, the EU should prefer a content bundle that included this content in a large variety of other content, have all content end-to-end encrypted and the programming information not be shared with AD-2 to maximize privacy. Nevertheless, the privacy of the EU against AD-2 observing traffic would still be lower than in the equivalent setup using unicast, because in unicast, AD-2 could not correlate which EUs are watching the same content and use that to deduce the content. Note that even the setup in Section 3.4 where AD-2 is not involved in IP multicast at all does not provide privacy against this level of analysis by AD-2 because there is no transport layer encryption in AMT and therefore AD-2 can correlate by onpath traffic analysis who is consuming the same content from an AMT relay from both the (S,G) join messages in AMT and the identical content segments (that were replicated at the AMT relay).

In summary: Because only content to be consumed by multiple EUs is carried via IP multicast here, and all that content can be end-to-end encrypted, the only IP multicast specific privacy consideration is for AD-2 to know or reconstruct what content an EU is consuming. For content for which this is undesirable, some form of protections as explained above are possible, but ideally, the model of Section 3.4 could be used in conjunction with future work adding e.g.: dTLS [RFC6347] encryption between AMT relay and EU.

Note that IP multicast by nature would permit the EU privacy against the content source operator because unlike unicast, the content

source does not natively know which EU is consuming which content: In all cases where AD-2 provides replication, only AD-2 does know this directly. This document does not attempt to describe a model that does maintain such level of privacy against the content source but only against exposure to intermediate parties, in this case AD-2.

8. IANA Considerations

No considerations identified in this document.

9. Acknowledgments

The authors would like to thank the following individuals for their suggestions, comments, and corrections:

Mikael Abrahamsson

Hitoshi Asaeda

Dale Carder

Tim Chown

Leonard Giuliano

Jake Holland

Joel Jaeggli

Albert Manfredi

Stig Venaas

Henrik Levkowetz

10. Change log [RFC Editor: Please remove]

Please see discussion on mailing list for changes before -11.

-11: version in IESG review.

-12: XML'ified version of -11, committed solely to make rfcdiff easier. XML versions hosted on <https://www.github.com/toerless/peering-bcp>

-13:

- o IESG feedback. Complete details in:
<https://raw.githubusercontent.com/toerless/peering-bcp/master/11-iesg-review-reply.txt>
- o Ben Campbell: Location information about EU (End User) is Network Location information
- o Ben Campbell: Added explanation of assumption to introduction that traffic is sourced from AD-1 to (one or many) AD-2, mentioned that sourcing from EU is out of scope.
- o Introduction: moved up bullet points about exchanges and transit to clean up flow of assumptions.
- o Ben Campbell: Added picture for the GRE case, visualized tunnels in all pictures.
- o Ben Campbell: See 13-discus.txt on github for more details of changes for this review.
- o Alissia Cooper: Added more explanation for Log Management, explained privacy context.
- o Alissia Cooper: removed pre pre-RFC5378 disclaimer.
- o Alissia Cooper: removed mentioning of potential mutual compensation between domains if the other violates SLA.
- o Mirja Kuehlewind: created section 4.1.1 to discuss congestion control more detailed, adding reference to BCP145, removed stub CC paragraphs from section 3.1 (principle applies to every section 3.x, and did not want to duplicate text between 3.x and 4.x).
- o Mirja Kuehlewind: removed section 8 (conclusion). Text was not very good, not important to have conclusion, maybe bring back with better text if strong interest.
- o Introduced section about broadcast peering points because there were too many places already where references to that case existed (4.2.4).
- o Introduced section about privacy considerations because of comment by Ben Campbell and Alissia Cooper.
- o Rewrote security considerations and structured it into key aspects: DoS attacks, content protection, peering point encryption and operational aspects.

- o Kathleen Moriarty: Added operational aspects to security section (also for Alissia), e.g.: covering securing the exchange of operational data between ADs.
- o Spencer Dawkins: Various editorial fixes. Removed BCP38 text from section 3, superseded by explanation of PIM-SM RPF check to provide equivalent security to BCP38 in security section 7.1).
- o Eric Roscorla: (fixed from other reviews already).
- o Adam Roach: Fixed up text about MDH-04, added reference to RFC4786.

-13: Fix for Mirja's review on must for congestion control.

11. References

11.1. Normative References

- [RFC2784] Farinacci, D., Li, T., Hanks, S., Meyer, D., and P. Traina, "Generic Routing Encapsulation (GRE)", RFC 2784, DOI 10.17487/RFC2784, March 2000, <<https://www.rfc-editor.org/info/rfc2784>>.
- [RFC3376] Cain, B., Deering, S., Kouvelas, I., Fenner, B., and A. Thyagarajan, "Internet Group Management Protocol, Version 3", RFC 3376, DOI 10.17487/RFC3376, October 2002, <<https://www.rfc-editor.org/info/rfc3376>>.
- [RFC3810] Vida, R., Ed. and L. Costa, Ed., "Multicast Listener Discovery Version 2 (MLDv2) for IPv6", RFC 3810, DOI 10.17487/RFC3810, June 2004, <<https://www.rfc-editor.org/info/rfc3810>>.
- [RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, DOI 10.17487/RFC4760, January 2007, <<https://www.rfc-editor.org/info/rfc4760>>.
- [RFC4604] Holbrook, H., Cain, B., and B. Haberman, "Using Internet Group Management Protocol Version 3 (IGMPv3) and Multicast Listener Discovery Protocol Version 2 (MLDv2) for Source-Specific Multicast", RFC 4604, DOI 10.17487/RFC4604, August 2006, <<https://www.rfc-editor.org/info/rfc4604>>.

- [RFC4609] Savola, P., Lehtonen, R., and D. Meyer, "Protocol Independent Multicast - Sparse Mode (PIM-SM) Multicast Routing Security Issues and Enhancements", RFC 4609, DOI 10.17487/RFC4609, October 2006, <<https://www.rfc-editor.org/info/rfc4609>>.
- [RFC7450] Bumgardner, G., "Automatic Multicast Tunneling", RFC 7450, DOI 10.17487/RFC7450, February 2015, <<https://www.rfc-editor.org/info/rfc7450>>.
- [RFC7761] Fenner, B., Handley, M., Holbrook, H., Kouvelas, I., Parekh, R., Zhang, Z., and L. Zheng, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", STD 83, RFC 7761, DOI 10.17487/RFC7761, March 2016, <<https://www.rfc-editor.org/info/rfc7761>>.
- [BCP38] Ferguson, P. and D. Senie, "Network Ingress Filtering: Defeating Denial of Service Attacks which employ IP Source Address Spoofing", BCP 38, RFC 2827, DOI 10.17487/RFC2827, May 2000, <<https://www.rfc-editor.org/info/rfc2827>>.
- [BCP41] Floyd, S., "Congestion Control Principles", BCP 41, RFC 2914, DOI 10.17487/RFC2914, September 2000, <<https://www.rfc-editor.org/info/rfc2914>>.
- [BCP145] Eggert, L., Fairhurst, G., and G. Shepherd, "UDP Usage Guidelines", BCP 145, RFC 8085, DOI 10.17487/RFC8085, March 2017, <<https://www.rfc-editor.org/info/rfc8085>>.

11.2. Informative References

- [RFC4786] Abley, J. and K. Lindqvist, "Operation of Anycast Services", BCP 126, RFC 4786, DOI 10.17487/RFC4786, December 2006, <<https://www.rfc-editor.org/info/rfc4786>>.
- [RFC6347] Rescorla, E. and N. Modadugu, "Datagram Transport Layer Security Version 1.2", RFC 6347, DOI 10.17487/RFC6347, January 2012, <<https://www.rfc-editor.org/info/rfc6347>>.
- [INF_ATIS_10] "CDN Interconnection Use Cases and Requirements in a Multi-Party Federation Environment", ATIS Standard A-0200010, December 2012.
- [MDH-04] Thaler, D. and others, "Multicast Debugging Handbook", IETF I-D draft-ietf-mboned-mdh-04.txt, May 2000.

[Traceroute]
, <<http://traceroute.org/#source%20code>>.

[I-D.ietf-mboned-mtrace-v2]
Asaeda, H., Meyer, K., and W. Lee, "Mtrace Version 2:
Traceroute Facility for IP Multicast", draft-ietf-mboned-
mtrace-v2-20 (work in progress), October 2017.

Authors' Addresses

Percy S. Tarapore (editor)
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

Email: shep@cisco.com

Toerless Eckert (editor)
Futurewei Technologies Inc.

Email: tte+ietf@cs.fau.de

Ram Krishnan
SupportVectors

Email: ramkri123@gmail.com

MBONED WG
Internet-Draft
Intended status: Standards Track
Expires: April 29, 2018

M. McBride
C. Perkins
Huawei
October 26, 2017

Multicast Wifi Problem Statement
draft-mcbride-mboned-wifi-mcast-problem-statement-01

Abstract

There have been known issues with multicast, in an 802.11 environment, which have prevented the deployment of multicast in these wifi environments. IETF multicast experts have been meeting together to discuss these issues and provide IEEE updates. The mboned working group is chartered to receive regular reports on the current state of the deployment of multicast technology, create "practice and experience" documents that capture the experience of those who have deployed and are deploying various multicast technologies, and provide feedback to other relevant working groups. As such, this document will gather the problems of wifi multicast into one problem statement document so as to offer the community guidance on current limitations.

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 <https://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 29, 2018.

Copyright Notice

Copyright (c) 2017 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

(<https://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. Multicast over WiFi Problems	2
2.1. Low Reliability	3
2.2. Low Data Rate	4
2.3. High Interference	4
2.4. High Power Consumption	4
3. Common remedies to multicast over wifi problems	4
4. State of the Union	5
5. IANA Considerations	6
6. Security Considerations	6
7. Acknowledgments	6
8. Normative References	6
Authors' Addresses	6

1. Introduction

Multicast over wifi has been used to low levels of success, usually to a point of being so negative that multicast over wifi is not allowed. In addition to protocol use of broadcast/multicast for control messages, more applications, such as push to talk in hospitals, video in enterprises and lectures in Universities, are streaming over wifi. And many end devices are increasingly using wifi for their connectivity. One of the primary problems multicast over wifi faces is that link local 802.11 doesn't necessarily support multicast, it supports broadcast. To make make multicast over wifi work successfully we often need to modify the multicast to instead be sent as unicast in order for it to successfully transmit with useable quality. Multicast over wifi experiences high packet error rates, no acknowledgements, and low data rate. This draft reviews these problems found with multicast over wifi. While this is not a solutions draft, common workarounds to some of the problems will be listed, along with the impact of the workarounds.

2. Multicast over WiFi Problems

802.11 is a wireless broadcast medium which works well for unicast and has become ubiquitous in its use. With multicast, however, problems arise over wifi. There are no ACKs for multicast packets,

for instance, so there can be a high level of packet error rate (PER) due to lack of retransmission and because the sender never backs off. It is not uncommon for there to be a packet loss rate of 5% which is particularly troublesome for video and other environments where high data rates and high reliability are required. Multicast, over wifi, is typically sent on a low data rate which makes video negatively impacted. Wifi loses many more packets than wired due to collisions and signal loss. There are also problems because clients are unable to stay in sleep mode due to the multicast control packets continuing to unnecessarily wake up those clients which subsequently reduces energy savings. Video is becoming the dominant content for end device applications, with multicast being the most natural method for applications to transmit video. Unfortunately, multicast, even though it is a very natural choice for video, incurs a large penalty over wifi.

One big difference between multicast over wired versus multicast over wifi is that wired links are a fixed transmission rate. Wifi, on the other hand, has a transmission rate which varies over time depending upon the clients proximity to the AP. Throughput of video flows, and the capacity of the broader wifi network, will change and will impact the ability for QoS solutions to effectively reserve bandwidth and provide admission control.

The main problems associated with multicast over WiFi are as follows:

- o Low Reliability
- o Lower Data Rate
- o High interference
- o High Power Consumption

These points will be elaborated separately in the following subsections.

2.1. Low Reliability

Because of the lack of acknowledgement for packets from Access Point to the receivers, it is not possible for the Access Point to know whether or not a retransmission is needed. Even in the wired Internet, this characteristic commonly causes undesirably high error rates, contributing to the relatively slow uptake of multicast applications even though the protocols have been available for decades. The situation for wireless links is much worse, and is quite sensitive to the presence of background traffic.

2.2. Low Data Rate

For wireless stations associated with an Access Points, the necessary power for good reception can vary from station to station. For unicast, the goal is to minimize power requirements while maximizing the data rate to the destination. For multicast, the goal is simply to maximize the number of receivers that will correctly receive the multicast packet. For this purpose, generally the Access Point has to use a much lower data rate at a power level high enough for even the farthest station to receive the packet. Consequently, the data rate of a video stream, for instance, would be constrained by the environmental considerations of the least reliable receiver associated with the Access Point.

2.3. High Interference

As mentioned in the previous subsection, multicast transmission to the stations associated to an Access Point typically proceeds at a much higher power level than is required for unicast to many of the receivers. High power levels directly contribute to stronger interference. The interference due to multicast may extend to effects inhibiting packet reception at more distant stations that might even be associated with other Access Points. Moreover, the use of lower data rates implies that the physical medium will be occupied for a longer time to transmit a packet than would be required at high data rates. Thus, the level of interference due to multicast will be not only higher, but longer in duration.

Depending on the choice of 802.11 technology, and the configured choice for the base data rate for multicast transmission from the Access Point, the amount of additional interference can range from a factor of ten, to a factor thousands for 802.11ac.

2.4. High Power Consumption

One of the characteristics of multicast transmission is that every station has to be configured to wake up to receive the multicast, even though the received packet may ultimately be discarded. This process has a relatively large impact on the power consumption by the multicast receiver station.

3. Common remedies to multicast over wifi problems

One common solution to the multicast over wifi problem is to convert the multicast traffic into unicast. This is often referred to as multicast to unicast (MC2UC). Converting the packets to unicast is beneficial because unicast packets are acknowledged and retransmitted as needed to prevent as much loss. The Access Points (AP) is also

able to provide rate limiting as needed. The drawback with this approach is that the benefit of using multicast is defeated.

Using 802.11n helps provide a more reliable and higher level of signal-to-noise ratio in a wifi environment over which multicast (broadcast) packets can be sent. This can provide higher throughput and reliability but the broadcast limitations remain.

4. State of the Union

In discussing these issues over email and, most recently, in a side meeting at IETF 99, it is generally agreed that these problems will not be fixed anytime soon primarily because it's expensive to do so and multicast is unreliable. The problem of v6 neighbor discovery saturating the wifi link is only part of the problem. A big problem is that the 802.11 multicast channel is an afterthought and only given 100th of the bandwidth. Multicast is basically a second class citizen, to unicast, over wifi. Unicast may have allocated 10mb while Multicast will be allocated 1mb. There are many protocols using multicast and there needs to be something provided in order to make them more reliable. Wifi traffic classes may help. We need to determine what problem should be solved by the IETF and what problem should be solved by the IEEE.

Apple's Bonjour protocol, for instance, provides service discovery (for printing) that utilizes multicast. It's the first thing operators drop. Even if multicast snooping is utilized, everyone registers at once using Bonjour and the network has serious degradation. There is also a lot of work being developed to help save battery life such as the devices not waking up when receiving a multicast packet. If an AP, for instance, expresses a DTIM of 3 then it will send a multicast packet every 3 packets. But the reality is that most AP's will send a multicast every 30 packets. For unicast there's a TIM. But because multicast is going to everyone, the AP sends a broadcast to everyone. DTIM does power management but clients can choose to wake up or not and whether to drop the packet or not. Then they don't know why their Bonjour doesn't work. The IETF may just need to decide that broadcast is more expensive so multicast needs to be sent wired. 802.1ak works on ethernet and wifi. 802.1ak has been pulled into 802.1Q as of 802.1Q-2011. 802.1Q-2014 can be looked at here: <http://www.ieee802.org/1/pages/802.1Q-2014.html>. If we don't find a generic solution we need to establish guidelines for multicast over wifi within the mboned wg. A multicast over wifi IETF mailing list is formed (mcast-wifi@ietf.org) and more discussion to be had there. This draft will serve as the current state of affairs.

This is not a solutions draft, but to provide an idea going forward, a reliable registration to Layer-2 multicast groups and a reliable multicast operation at Layer-2 could provide a generic solution. There is no need to support 2^{24} groups to get solicited node multicast working: it is possible to simply select a number of trailing bits that make sense for a given network size to limit the amount of unwanted deliveries to reasonable levels. We need to encourage IEEE 802.1 and 802.11 to revisit L2 multicast issues. In particular, Wi-Fi provides a broadcast service, not a multicast one. In fact all frames are broadcast at the PHY level unless we beamform. What comes with unicast is the property of being much faster (2 orders of magnitude) and much more reliable (L2 ARQ).

5. IANA Considerations

None

6. Security Considerations

None

7. Acknowledgments

The following people have contributed information and discussion in the meetings and on the list which proved helpful for the development of the latest version this Internet Draft:

Dave Taht, Donald Eastlake, Pascal Thubert, Juan Carlos Zuniga, Mikael Abrahamsson, Diego Dujovne, David Schinazi, Stig Venaas, Stuart Cheshire, Lorenzo, Greg Shephard, Mark Hamilton

8. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.

Authors' Addresses

Mike McBride
Huawei
2330 Central Expressway
Santa Clara CA 95055
USA

Email: michael.mcbride@huawei.com

Charlie Perkins
Huawei
2330 Central Expressway
Santa Clara CA 95055
USA

Email: charlie.perkins@huawei.com

MBONED WG
Internet-Draft
Intended status: Standards Track
Expires: February 22, 2018

Zheng. Zhang
Cui. Wang
ZTE Corporation
Ying. Cheng
China Unicom
August 21, 2017

Multicast Model
draft-zhang-mboned-multicast-info-model-02

Abstract

This document intents to provide a general and all-round multicast model, which tries to stand at a high level to take full advantages of existed multicast protocol models to control the multicast network, and guides the deployment of multicast service. And also, there will define several possible RPCs about how to interact between multicast info model and multicast protocol models. This multicast information model is mainly used by the management tools run by the network operators in order to manage, monitor and debug the network resources used to deliver multicast service, as well as gathering some data from the network.

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 February 22, 2018.

Copyright Notice

Copyright (c) 2017 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. Design of the multicast model	3
3. UML Class Diagram for Multicast Info Model	4
4. Model Structure	5
5. Multicast Information Model	7
6. Notifications	17
7. Acknowledgements	17
8. Normative References	17
Authors' Addresses	18

1. Introduction

Currently, there are many multicast YANG models, such as PIM, MLD, and BIER and so on. But all these models are distributed in different working groups as separate files and focus on the protocol itself. Furthermore, they cannot describe a high-level multicast service required by network operators.

This document intents to provide a general and all-round multicast model, which tries to stand at a high level to take full advantages of these aforementioned models to control the multicast network, and guides the deployment of multicast service.

This multicast information model is mainly used by the management tools run by the network operators in order to manage, monitor and debug the network resources used to deliver multicast service, as well as gathering some data from the network.

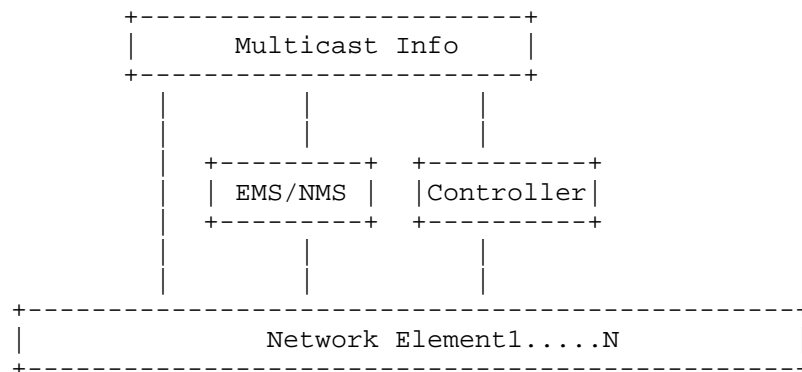


Figure 1: Example usage of Multicast Model

Detailedly, in figure 1, there is an example of usage of this multicast model. Network operators can input this model to a controller who is responsible to translate the information and invoke the corresponding protocol models into configurations to configure the network elements through NETCONF/RESTCONF/CLI. Or network operators can input this model to the EMS/NMS to manage the network elements or configure the network elements directly. On the other hand, when the network elements detect failure or some other changes, the network operators can collect these kind of notifications through this model to assist locating the exact failure and responding immediately. For example, when the network element suffers a failure of one MVPN neighbor, it can notify to the EMS/NMS or Controller or to other Multicast Model management tool directly to let the network operator take actions immediately.

Specifically, in section 3, it provides a human readability of the whole multicast network through UML class diagram, which frames different multicast components and correlates them in a readable fashion. Then, based on this UML class diagram, there is instantiated and detailed YANG model in Section 5.

In other words, this document does not define any specific protocol model, instead, it depends on many existed multicast protocol models and relates several multicast information together to fulfill multicast service.

2. Design of the multicast model

This model includes three layers: the multicast overlay, the transport layer and the multicast underlay information.

Multicast overlay defines the features of multicast flow, such as (vpnid, multicast source and multicast group) information, and (ingress-node, egress-nodes) nodes information. If the transport layer is BIER, there may define BIER information including (Subdomain, ingress-node BFR-id, egress-nodes BFR-id). In data center network, for fine-grained to gather the nodes belonging to the same virtual network, there may need VNI-related information to assist. If no (ingress-node, egress-nodes) information are defined directly, there may need overlay multicast signaling technology, such as MLD or MVPN, to collect these nodes information.

Multicast transport layer defines the type of transport technologies that can be used to forward multicast flow, including BIER forwarding type, MPLS forwarding type, or PIM forwarding type and so on. One or several transport technologies could be defined at the same time. As for the detailed parameters for each transport technology, this multicast information model can invoke the corresponding protocol model to define them.

Multicast underlay defines the type of underlay technologies, such as OSPF, ISIS, BGP, PIM or BABEL and so on. One or several underlay technologies could be defined at the same time. As for the specific parameters for each underlay technology, this multicast information model can depend the corresponding protocol model to configure them as well.

3. UML Class Diagram for Multicast Info Model

The following is a UML diagram for Multicast Info Model.

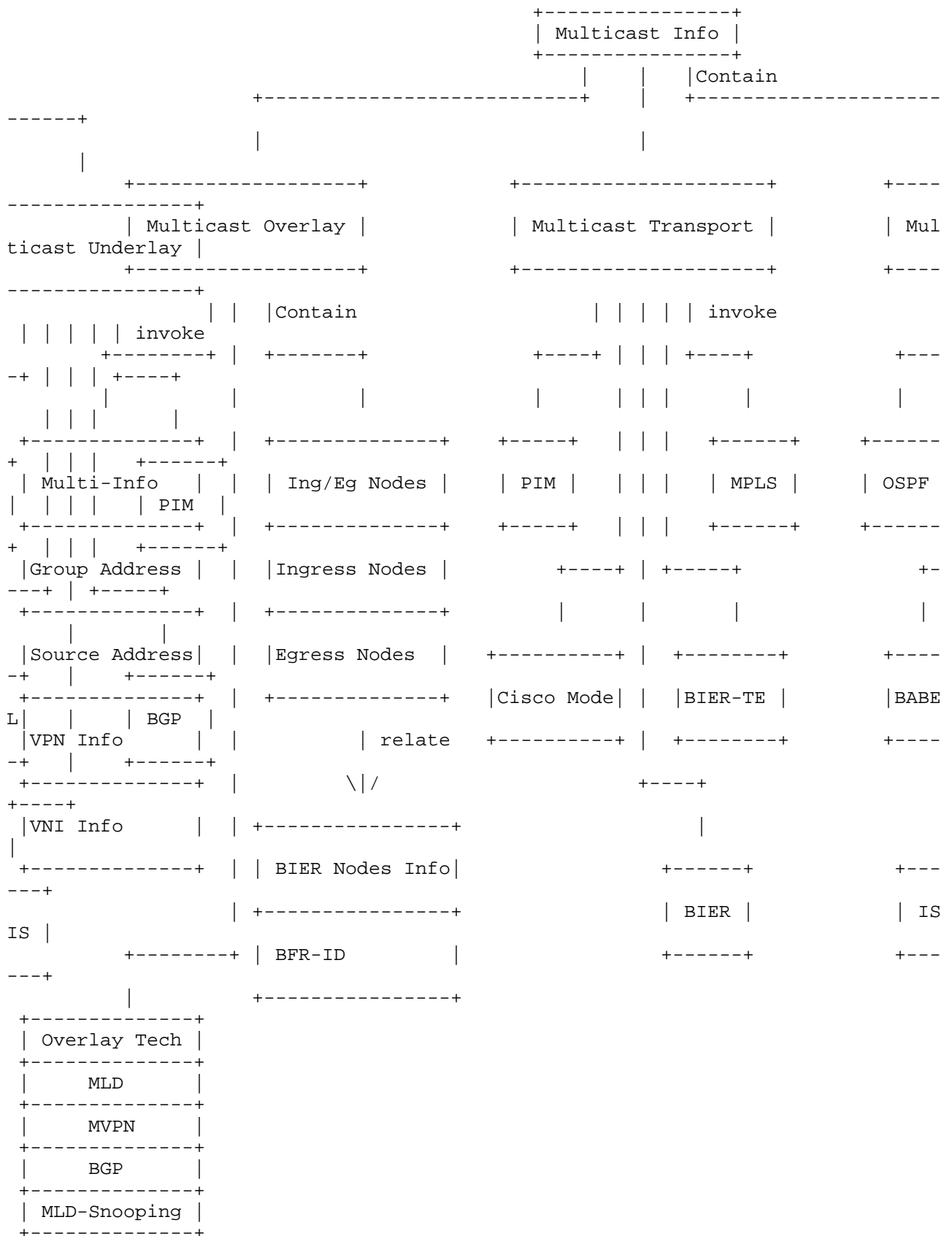


Figure 2: UML Class Diagram for Multicast Info Model

4. Model Structure

```
module: ietf-multicast-information
  +--rw multicast-information
    +--rw multicast-info* [vpn-id source-address source-wildcard group-address
s group-wildcard vni-type vni-value]
      +--rw vpn-id                uint32
      +--rw source-address         inet:ip-address
      +--rw source-wildcard        uint8
```

```

+--rw group-address          inet:ip-address
+--rw group-wildcard         uint8
+--rw vni-type               virtual-type
+--rw vni-value              uint32
+--rw multicast-overlay
|   +--rw nodes-information
|   |   +--rw ingress-node?   inet:ip-address
|   |   +--rw egress-nodes* [egress-node]
|   |   |   +--rw egress-node   inet:ip-address
|   +--rw bier-information
|   |   +--rw sub-domain?     sub-domain-id
|   |   +--rw ingress-node?   bfr-id
|   |   +--rw egress-nodes* [egress-node]
|   |   |   +--rw egress-node   bfr-id
|   +--rw overlay-technology
|   |   +--rw (overlay-tech-type)?
|   |   |   +--:(mld)
|   |   |   +--:(mvpn)
|   |   |   +--:(bgp)
|   |   |   +--:(mld-snooping)
+--rw multicast-transport
|   +--rw bier
|   |   +--rw sub-domain?     sub-domain-id
|   |   +--rw (encap-type)?
|   |   |   +--:(mpls)
|   |   |   +--:(non-mpls)
|   |   |   +--:(ipv6)
|   |   +--rw bitstringlength? uint16
|   |   +--rw set-identifier?   si
|   |   +--rw ecmp?            boolean
|   |   +--rw frr?            boolean
|   +--rw bier-te
|   |   +--rw sub-domain?     sub-domain-id
|   |   +--rw (encap-type)?
|   |   |   +--:(mpls)
|   |   |   +--:(non-mpls)
|   |   +--rw bitstringlength? uint16
|   |   +--rw set-identifier?   si
|   |   +--rw ecmp?            boolean
|   |   +--rw frr?            boolean
|   +--rw cisco-mode
|   |   +--rw p-group?        inet:ip-address
|   |   +--rw graceful-restart? boolean
|   |   +--rw bfd?            boolean
|   +--rw mpls
|   |   +--rw (mpls-tunnel-type)?
|   |   |   +--:(mldp)
|   |   |   |   +--rw mldp-tunnel-id?      uint32

```

```

| | | | +-rw mldp-frr? boolean
| | | | +-rw mldp-backup-tunnel? boolean
| | | | +---:(p2mp-te)
| | | | +-rw te-tunnel-id? uint32
| | | | +-rw te-frr? boolean
| | | | +-rw te-backup-tunnel? boolean
| | +-rw pim
| | +-rw graceful-restart? boolean
| | +-rw bfd? boolean
+-rw multicast-underlay
+-rw underlay-requirement? boolean
+-rw bgp
+-rw ospf
| +-rw topology-id? uint16
+-rw isis
| +-rw topology-id? uint16
+-rw babel
+-rw pim

```

5. Multicast Information Model

```

<CODE BEGINS> file "ietf-multicast-information.yang"
module ietf-multicast-information {

    namespace "urn:ietf:params:xml:ns:yang:ietf-multicast-information";

    prefix multicast-info;

    import ietf-inet-types {
        prefix "inet";
    }

    organization " IETF MBONED( MBONE Deployment ) Working Group";
    contact
        "WG List: <mailto:bier@ietf.org>
        WG Chair: Greg Shepherd
                <mailto:gjshep@gmail.com>
        WG Chair: Leonard Giuliano
                <mailto:lenny@juniper.net>

        Editor: Zheng Zhang
                <mailto:zhang.zheng@zte.com.cn>
        Editor: Cui Wang
                <mailto:wang.cuil@zte.com.cn>
        Editor: Ying Cheng
                <mailto:chengying10@chinaunicom.cn>
";

```

```
description
  "This module contains a collection of YANG definitions for
  managing multicast information.";

revision 2017-08-20 {
  description
    "Add BGP and MLD-snooping overlay and BIER-TE transport.";
  reference "https://tools.ietf.org/html/draft-zhang-mboned-multicast-info
-model";
}

revision 2016-12-08 {
  description
    "Initial version.";
  reference "https://tools.ietf.org/html/draft-zhang-mboned-multicast-info
-model";
}
/*feature*/
grouping general-multicast {
  description "The general multicast address information.";
  leaf source-address {
    type inet:ip-address;
    description "The address of multicast source. The value set to zero
      means that the receiver interests in all source that relevant to
      one group.";
  }
  leaf source-wildcard {
    type uint8;
    description "The wildcard information of source.";
  }
  leaf group-address {
    type inet:ip-address;
    description "The address of multicast group.";
  }
  leaf group-wildcard {
    type uint8;
    description "The wildcard information of group.";
  }
}

grouping m-addr {
  description "The vpn multicast information.";
  leaf vpn-id {
    type uint32;
    description "The vpn-id of the multicast flow.
      If there is global instance, the vpnid value should be zero.";
  }
  uses general-multicast;
}
```

```

typedef virtual-type {
    type enumeration {
        enum "vxlan" {
            description "The vxlan type.";
        }
        enum "virtual subnet" {
            description "The nvgre type";
        }
        enum "vni" {
            description "The geneve type";
        }
    }
    description "The collection of virtual network type.";
}

grouping multicast-nvo3 {
    description "The nvo3 multicast information.";
    leaf vni-type {
        type virtual-type;
        description "The type of virtual network identifier. Include the Vx
lan
            NVGRE and Geneve.";
    }
    leaf vni-value {
        type uint32;
        description "The value of Vxlan network identifier, virtual subnet I
D
            or virtual net identifier.";
    }
}

grouping multicast-feature {
    description
        "This group describe the different multicast information
        in various deployments.";
    uses m-addr;
    uses multicast-nvo3;
}

grouping ip-node {
    description "The IP information of multicast nodes.";
    leaf ingress-node {
        type inet:ip-address;
        description "The ingress node of multicast flow. Or the ingress
            node of MVPN and BIER. In MVPN, this is the address of ingress
            PE; in BIER, this is the BFR-prefix of ingress nodes.";
    }

    list egress-nodes {
        key "egress-node";
    }
}

```

```
        description "This ID information of one adjacency.";

        leaf egress-node {
            type inet:ip-address;
            description
                "The egress multicast nodes of multicast flow.
                Or the egress node of MVPN and BIER. In MVPN, this is the
                address of egress PE; in BIER, this is the BFR-prefix of
                ingress nodes.";
        }
    }
}
/* should import from BIER yang */
typedef bfr-id {
    type uint16;
    description "The BFR id of nodes.";
}

typedef si {
    type uint16;
    description
        "The type for set identifier";
}

typedef sub-domain-id {
    type uint16;
    description
        "The type for sub-domain-id";
}

typedef bit-string {
    type uint16;
    description
        "The bit mask of one bitstring.";
}

grouping bier-node {
    description "The BIER information of multicast nodes.";
    leaf sub-domain {
        type sub-domain-id;
        description "The sub-domain that this multicast flow belongs to.";
    }
    leaf ingress-node {
        type bfr-id;
        description "The ingress node of multicast flow. This is the
            BFR-id of ingress nodes.";
    }
    list egress-nodes {
```



```

        key "egress-node";
        description "This ID information of one adjacency.";

        leaf egress-node {
            type bfr-id;
            description
                "The egress multicast nodes of multicast flow.
                This is the BFR-id of egress nodes.";
        }
    }
}

grouping overlay-tech {
    description "The possible overlay technologies for multicast service.";
    choice overlay-tech-type {
        case mld {
            description "MLD technology is used for multicast overlay";
        }
        case mvpn {
            description "MVPN technology is used for multicast overlay";
        }
        case bgp {
            description "BGP technology is used for multicast overlay";
        }
        case mld-snooping {
            description "MLD snooping technology is used for multicast overl
ay";
        }
        description "The collection of multicast overlay technology";
    }
}

grouping multicast-overlay {
    description "The node information that connect the ingress multicast
    flow, and the nodes information that connect the egress multicast
    flow.";
    /*uses multicast-feature;*/
    container nodes-information {
        description "The ingress and egress nodes information.";
        uses ip-node;
    }
    container bier-information {
        description "The ingress and egress BIER nodes information.";
        uses bier-node;
    }
    container overlay-technology {
        description "The possible overlay technologies for multicast service
.";
        uses overlay-tech;
    }
}

```

```
    }

/*transport*/

    grouping transport-bier {
        description "The BIER transport information.";
        leaf sub-domain {
            type sub-domain-id;
            description "The subdomain id that this multicast flow belongs to.";
        }
        choice encap-type {
            case mpls {
                description "The BIER forwarding depend on mpls.";
            }
            case non-mpls {
                description "The BIER forwarding depend on non-mpls.";
            }
            case ipv6 {
                description "The BIER forwarding depend on IPv6.";
            }
            description "The encapsulation type in BIER.";
        }
        leaf bitstringlength {
            type uint16;
            description "The bitstringlength used by BIER forwarding.";
        }
        leaf set-identifier {
            type si;
            description "The set identifier used by this multicast flow.";
        }
        leaf ecmp {
            type boolean;
            description "The capability of ECMP.";
        }
        leaf frr {
            type boolean;
            description "The capability of fast re-route.";
        }
    }

    grouping transport-bier-te {
        description "The BIER-TE transport information.";
        leaf sub-domain {
            type sub-domain-id;
            description "The subdomain id that this multicast flow belongs to.";
        }
        choice encap-type {
```

```
        case mpls {
            description "The BIER-TE forwarding depend on mpls.";
        }
        case non-mpls {
            description "The BIER-TE forwarding depend on non-mpls.";
        }
        description "The encapsulation type in BIER-TE.";
    }
    leaf bitstringlength {
        type uint16;
        description "The bitstringlength used by BIER-TE forwarding.";
    }
    leaf set-identifier {
        type si;
        description "The set identifier used by this multicast flow, especially in BIER TE.";
    }
    leaf ecmp {
        type boolean;
        description "The capability of ECMP.";
    }
    leaf frr {
        type boolean;
        description "The capability of fast re-route.";
    }
}

grouping transport-pim {
    description "The requirement information of pim transportation.";
    leaf graceful-restart {
        type boolean;
        description "If the graceful restart function should be supported.";
    }
    leaf bfd {
        type boolean;
        description "If the bfd function should be supported.";
    }
}

grouping mldp-tunnel-feature {
    description "The tunnel feature.";
    leaf mldp-tunnel-id {
        type uint32;
        description "The tunnel id that correspond this flow.";
    }
    leaf mldp-frr {
        type boolean;
        description "If the fast re-route function should be supported.";
    }
}
```

```
    leaf mldp-backup-tunnel {
        type boolean;
        description "If the backup tunnel function should be supported.";
    }
}

grouping p2mp-te-tunnel-feature {
    description "The tunnel feature.";
    leaf te-tunnel-id {
        type uint32;
        description "The tunnel id that correspond this flow.";
    }
    leaf te-frr {
        type boolean;
        description "If the fast re-route function should be supported.";
    }
    leaf te-backup-tunnel {
        type boolean;
        description "If the backup tunnel function should be supported.";
    }
}

/*typedef sub-domain-id {
    type uint16;
    description
        "The type for sub-domain-id";
}*/

grouping transport-mpls {
    description "The mpls transportation information.";
    choice mpls-tunnel-type {
        case mldp {
            uses mldp-tunnel-feature;
            description "The mldp tunnel.";
        }
        case p2mp-te {
            uses p2mp-te-tunnel-feature;
            description "The p2mp te tunnel.";
        }
    }
    description "The collection types of mpls tunnels";
}

grouping cisco-multicast {
    description "The Cisco MDT multicast information in RFC6037.";
    leaf p-group {
        type inet:ip-address;
        description "The address of p-group.";
    }
}
```

```
    }
  }

  grouping transport-cisco-mode {
    description "The transport information of Cisco mode, RFC6037.";
    uses cisco-multicast;
    uses transport-pim;
  }

  grouping multicast-transport {
    description "The transport information of multicast service.";
    container bier {
      uses transport-bier;
      description "The transport technology is BIER.";
    }
    container bier-te {
      uses transport-bier-te;
      description "The transport technology is BIER-TE.";
    }
    container cisco-mode {
      uses transport-cisco-mode;
      description "The transport technology is cisco-mode.";
    }
    container mpls {
      uses transport-mpls;
      description "The transport technology is mpls.";
    }
    container pim {
      uses transport-pim;
      description "The transport technology is PIM.";
    }
  }

  /*underlay*/
  grouping underlay-bgp {
    description "Underlay information of BGP.";
  }

  grouping underlay-ospf {
    description "Underlay information of OSPF.";
    leaf topology-id {
      type uint16;
      description "The topology id of ospf instance.";
    }
  }

  grouping underlay-isis {
    description "Underlay information of ISIS.";
```

```
    leaf topology-id {
        type uint16;
        description "The topology id of isis instance.";
    }
}

grouping underlay-babel {
    description "Underlay information of Babel.";
    /* If there are some necessary information should be defined? */
}

grouping underlay-pim {
    description "Underlay information of PIM.";
    /* If there are some necessary information should be defined? */
}

grouping multicast-underlay {
    description "The underlay information relevant multicast service.";
    leaf underlay-requirement {
        type boolean;
        description "If the underlay technology should be required.";
    }
    container bgp {
        uses underlay-bgp;
        description "The underlay technology is BGP.";
    }
    container ospf {
        uses underlay-ospf;
        description "The underlay technology is OSPF.";
    }
    container isis {
        uses underlay-isis;
        description "The underlay technology is ISIS.";
    }
    container babel {
        uses underlay-babel;
        description "The underlay technology is Babel.";
    }
    container pim {
        uses underlay-pim;
        description "The underlay technology is PIM.";
    }
}

container multicast-information {
    description "The model of multicast service. Include overlay, transport
and underlay.";

    list multicast-info{
```

```
        key "vpn-id source-address source-wildcard group-address group-wildc
ard vni-type vni-value";
        uses multicast-feature;
        description "The detail multicast information.";

        container multicast-overlay {
            description "The overlay information of multicast service.";
            uses multicast-overlay;
        }
        container multicast-transport {
            description "The transportation of multicast service.";
            uses multicast-transport;
        }
        container multicast-underlay {
            description "The underlay of multicast service.";
            uses multicast-underlay;
        }
    }
}
<CODE ENDS>
```

6. Notifications

TBD.

7. Acknowledgements

The authors would like to thank Stig Venaas, Jake Holland for their valuable comments and suggestions.

8. Normative References

[I-D.ietf-bier-architecture]

Wijnands, I., Rosen, E., Dolganow, A., Przygienda, T., and S. Aldrin, "Multicast using Bit Index Explicit Replication", draft-ietf-bier-architecture-07 (work in progress), June 2017.

[I-D.ietf-bier-bier-yang]

Chen, R., hu, f., Zhang, Z., dai.xianxian@zte.com.cn, d., and M. Sivakumar, "YANG Data Model for BIER Protocol", draft-ietf-bier-bier-yang-02 (work in progress), August 2017.

[I-D.ietf-pim-yang]

Liu, X., McAllister, P., Peter, A., Sivakumar, M., Liu, Y., and f. hu, "A YANG data model for Protocol-Independent Multicast (PIM)", draft-ietf-pim-yang-08 (work in progress), April 2017.

[RFC6020] Bjorklund, M., Ed., "YANG - A Data Modeling Language for the Network Configuration Protocol (NETCONF)", RFC 6020, DOI 10.17487/RFC6020, October 2010, <<https://www.rfc-editor.org/info/rfc6020>>.

[RFC6037] Rosen, E., Ed., Cai, Y., Ed., and IJ. Wijnands, "Cisco Systems' Solution for Multicast in BGP/MPLS IP VPNs", RFC 6037, DOI 10.17487/RFC6037, October 2010, <<https://www.rfc-editor.org/info/rfc6037>>.

[RFC6087] Bierman, A., "Guidelines for Authors and Reviewers of YANG Data Model Documents", RFC 6087, DOI 10.17487/RFC6087, January 2011, <<https://www.rfc-editor.org/info/rfc6087>>.

[RFC6513] Rosen, E., Ed. and R. Aggarwal, Ed., "Multicast in MPLS/BGP IP VPNs", RFC 6513, DOI 10.17487/RFC6513, February 2012, <<https://www.rfc-editor.org/info/rfc6513>>.

[RFC7223] Bjorklund, M., "A YANG Data Model for Interface Management", RFC 7223, DOI 10.17487/RFC7223, May 2014, <<https://www.rfc-editor.org/info/rfc7223>>.

Authors' Addresses

Zheng Zhang
ZTE Corporation
No. 50 Software Ave, Yuhuatai Distinct
Nanjing
China

Email: zhang.zheng@zte.com.cn

Cui(Linda) Wang
ZTE Corporation
No. 50 Software Ave, Yuhuatai Distinct
Nanjing
China

Email: lindawangjoy@gmail.com

Ying Cheng
China Unicom
Beijing
China

Email: chengying10@chinaunicom.cn

BESS
Internet-Draft
Updates: 6514 (if approved)
Intended status: Standards Track
Expires: July 22, 2018

Z. Zhang
L. Giuliano
Juniper Networks
January 18, 2018

MVPN and MSDP SA Interoperation
draft-zzhang-bess-mvpn-msdp-sa-interoperation-01

Abstract

This document specifies the procedures for interoperation between MVPN Source Active routes and customer MSDP Source Active routes, which is useful for MVPN provider networks offering services to customers with an existing MSDP infrastructure. Without the procedures described in this document, VPN-specific MSDP sessions are required among the PEs that are customer MSDP peers.

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 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 <https://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 July 22, 2018.

Copyright Notice

Copyright (c) 2018 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 (<https://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. Terminologies	2
2. Introduction	2
2.1. MVPN RPT-SPT Mode	4
3. Specification	4
4. IANA Considerations	5
5. Acknowledgements	5
6. References	5
6.1. Normative References	5
6.2. Informative References	6
Authors' Addresses	6

1. Terminologies

Familiarity with MVPN and MSDP protocols and procedures is assumed. Some terminologies are listed below for convenience.

- o ASM: Any source multicast.
- o SPT: Source-specific Shortest-path Tree.
- o C-S: A multicast source address, identifying a multicast source located at a VPN customer site.
- o C-G: A multicast group address used by a VPN customer.
- o C-RP: A multicast Rendezvous Point for a VPN customer.
- o EC: Extended Community.

2. Introduction

Section "14. Supporting PIM-SM without Inter-Site Shared C-Trees" of [RFC6514] specifies the procedures for MVPN PEs to discover (C-S,C-G) via MVPN Source Active A-D routes and then send (C-S,C-G) C-multicast routes towards the ingress PEs, to establish SPTs for customer ASM flows for which they have downstream receivers. (C-*,C-G)

C-multicast routes are not sent among the PEs so inter-site shared C-Trees are not used and the method is generally referred to as "spt-only" mode.

With this mode, the MVPN Source Active routes are functionally similar to MSDP Source-Active messages [RFC3618]. One or more of the PEs, say PE1, either act as a C-RP and learn of (C-S,C-G) via PIM Register messages, or have MSDP sessions with some MSDP peers and learn (C-S,C-G) via MSDP SA messages. In either case, PE1 will then originate MVPN SA routes for other PEs to learn the (C-S,C-G).

[RFC6514] only specifies that a PE receiving the MVPN SA routes, say PE2, will advertise (C-S,C-G) C-multicast routes if it has corresponding (C-*,C-G) state learnt from its CE. PE2 may also have MSDP sessions with other C-RPs at its site, but [RFC6514] does not specify that it advertise MSDP SA messages to those MSDP peers for the (C-S,C-G) that it learns via MVPN SA routes. PE2 would need to have an MSDP session with PE1 (that advertised the MVPN SA messages) to learn the sources via MSDP SA messages, for it to advertise the MSDP SA to its local peers. To make things worse, unless blocked by policy control, PE2 would in turn advertise MVPN SA routes because of those MSDP SA messages that it receives from PE1, which are redundant and unnecessary. Also notice that the PE1-PE2 MSDP session is VPN-specific, while the BGP sessions over which the MVPN routes are advertised are not.

If a PE does advertise MSDP SA messages based on received MVPN SA routes, the VPN-specific MSDP sessions are no longer needed. Additionally, this MVPN/MSDP SA interoperation has the following inherent benefits for a BGP based solution.

- o MSDP SA refreshes are replaced with BGP hard state.
- o Route Reflectors can be used instead of having peer-to-peer sessions.
- o VPN extranet mechanisms can be used to propagate (C-S,C-G) information across VPNs with flexible policy control.

While MSDP Source Active routes contain the source, group and RP address of a given multicast flow, MVPN Source Active routes only contain the source and group. MSDP requires the RP address information in order to perform peer-RPF. Therefore, this document describes how to convey the RP address information into the MVPN Source Active route using an Extended Community so this information can be shared with an existing MSDP infrastructure.

The procedures apply to Global Table Multicast (GTM) [RFC7716] as well.

2.1. MVPN RPT-SPT Mode

For comparison, another method of supporting customer ASM is generally referred to "rpt-spt" mode. Section "13. Switching from a Shared C-Tree to a Source C-Tree" of [RFC6514] specifies the MVPN SA procedures for that mode, but those SA routes are replacement for PIM-ASM assert and (s,g,rpt) prune mechanisms, not for source discovery purpose. MVPN/MSDP SA interoperation for the "rpt-spt" mode is outside of the scope of this document. In the rest of the document, the "spt-only" mode is assumed.

3. Specification

The MVPN PEs that act as customer RPs or have one or more MSDP sessions in a VPN (or the global table in case of GTM) are treated as an MSDP mesh group for that VPN (or the global table). In the rest of the document, it is referred to as the PE mesh group. It MUST not include other MSDP speakers, and is integrated into the rest of MSDP infrastructure for the VPN (or the global table) following normal MSDP rules and practices.

When an MVPN PE advertises an MVPN SA route following procedures in [RFC6514] for the "spt-only" mode, it SHOULD attach an "MVPN SA RP-address Extended Community". This is a Transitive IPv4-Address-Specific Extended Community. The Local Administrative field is set to zero and the Global Administrative field is set to an RP address determined as the following:

- o If the (C-S,C-G) is learnt as result of PIM Register mechanism, the local RP address for the C-G is used.
- o If the (C-S,C-G) is learnt as result of incoming MSDP SA messages, the RP address in the selected MSDP SA message is used.

In addition to procedures in [RFC6514], an MVPN PE may be provisioned to generate MSDP SA messages from received MVPN SA routes, with or without fine policy control. If a received MVPN SA route is to trigger MSDP SA message, it is treated as if a corresponding MSDP SA message was received from within the PE mesh group and normal MSDP procedure is followed (e.g. an MSDP SA message is advertised to other MSDP peers outside the PE mesh group). The (S,G) information comes from the (C-S,C-G) encoding in the MVPN SA NLRI and the RP address comes from the "MVPN SA RP-address EC" mentioned above. If the received MVPN SA route does not have the EC (this could be from a legacy PE that does not have the capability to attach the EC), the

local RP address for the C-G is used. In that case, it is possible that receiving PE's RP for the C-G is actually the MSDP peer to which the generated MSDP message is advertised, causing the peer to discard it due to RPF failure. To get around that problem the peer SHOULD use local policy to accept the MSDP SA message.

An MVPN PE MAY treat only the best MVPN SA route selected by BGP route selection process (instead of all MVPN SA routes) for a given (C-S,C-G) as a received MSDP SA message (and advertise corresponding MSDP message). In that case, if the selected best MVPN SA route does not have the "MVPN SA RP-address EC" but another route for the same (C-S, C-G) does, then the best route with the EC SHOULD be chosen. As a result, when/if the best MVPN SA route with the EC changes, a new MSDP SA message is advertised if the RP address determined according to the newly selected MVPN SA route is different from before. The previously advertised MSDP SA message with the older RP address will be timed out.

4. IANA Considerations

This document introduces a new Transitive IPv4 Address Specific Extended Community "MVPN SA RP-address Extended Community". An IANA request will be submitted for a subcode of 0x20 (pending approval and subject to change) in the Transitive IPv4-Address-Specific Extended Community Sub-Types registry.

5. Acknowledgements

The authors thank Eric Rosen and Vinod Kumar for their review, comments, questions and suggestions for this document. The authors also thank Yajun Liu for her review and comments.

6. References

6.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.
- [RFC3618] Fenner, B., Ed. and D. Meyer, Ed., "Multicast Source Discovery Protocol (MSDP)", RFC 3618, DOI 10.17487/RFC3618, October 2003, <<https://www.rfc-editor.org/info/rfc3618>>.

[RFC6514] Aggarwal, R., Rosen, E., Morin, T., and Y. Rekhter, "BGP Encodings and Procedures for Multicast in MPLS/BGP IP VPNs", RFC 6514, DOI 10.17487/RFC6514, February 2012, <<https://www.rfc-editor.org/info/rfc6514>>.

6.2. Informative References

[RFC7716] Zhang, J., Giuliano, L., Rosen, E., Ed., Subramanian, K., and D. Pacella, "Global Table Multicast with BGP Multicast VPN (BGP-MVPN) Procedures", RFC 7716, DOI 10.17487/RFC7716, December 2015, <<https://www.rfc-editor.org/info/rfc7716>>.

Authors' Addresses

Zhaohui Zhang
Juniper Networks

EMail: zzhang@juniper.net

Lenny Giuliano
Juniper Networks

EMail: lenny@juniper.net