

MBONED
Internet-Draft
Intended status: Informational
Expires: August 7, 2020

M. McBride
Futurewei
O. Komolafe
Arista Networks
February 4, 2020

Multicast in the Data Center Overview
draft-ietf-mboned-dc-deploy-09

Abstract

The volume and importance of one-to-many traffic patterns in data centers is likely to increase significantly in the future. Reasons for this increase are discussed and then attention is paid to the manner in which this traffic pattern may be judiciously handled in data centers. The intuitive solution of deploying conventional IP multicast within data centers is explored and evaluated. Thereafter, a number of emerging innovative approaches are described before a number of recommendations are made.

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 August 7, 2020.

Copyright Notice

Copyright (c) 2020 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
1.1. Requirements Language	3
2. Reasons for increasing one-to-many traffic patterns	3
2.1. Applications	3
2.2. Overlays	5
2.3. Protocols	6
2.4. Summary	6
3. Handling one-to-many traffic using conventional multicast	7
3.1. Layer 3 multicast	7
3.2. Layer 2 multicast	7
3.3. Example use cases	9
3.4. Advantages and disadvantages	9
4. Alternative options for handling one-to-many traffic	10
4.1. Minimizing traffic volumes	11
4.2. Head end replication	12
4.3. Programmable Forwarding Planes	12
4.4. BIER	13
4.5. Segment Routing	14
5. Conclusions	15
6. IANA Considerations	15
7. Security Considerations	15
8. Acknowledgements	15
9. References	15
9.1. Normative References	15
9.2. Informative References	16
Authors' Addresses	18

1. Introduction

The volume and importance of one-to-many traffic patterns in data centers will likely continue to increase. Reasons for this increase include the nature of the traffic generated by applications hosted in the data center, the need to handle broadcast, unknown unicast and multicast (BUM) traffic within the overlay technologies used to support multi-tenancy at scale, and the use of certain protocols that traditionally require one-to-many control message exchanges.

These trends, allied with the expectation that highly virtualized large-scale data centers must support communication between potentially thousands of participants, may lead to the natural assumption that IP multicast will be widely used in data centers,

specifically given the bandwidth savings it potentially offers. However, such an assumption would be wrong. In fact, there is widespread reluctance to enable conventional IP multicast in data centers for a number of reasons, mostly pertaining to concerns about its scalability and reliability.

This draft discusses some of the main drivers for the increasing volume and importance of one-to-many traffic patterns in data centers. Thereafter, the manner in which conventional IP multicast may be used to handle this traffic pattern is discussed and some of the associated challenges highlighted. Following this discussion, a number of alternative emerging approaches are introduced, before concluding by discussing key trends and making a number of recommendations.

1.1. Requirements Language

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

2. Reasons for increasing one-to-many traffic patterns

2.1. Applications

Key trends suggest that the nature of the applications likely to dominate future highly-virtualized multi-tenant data centers will produce large volumes of one-to-many traffic. For example, it is well-known that traffic flows in data centers have evolved from being predominantly North-South (e.g. client-server) to predominantly East-West (e.g. distributed computation). This change has led to the consensus that topologies such as the Leaf/Spine, that are easier to scale in the East-West direction, are better suited to the data center of the future. This increase in East-West traffic flows results from VMs often having to exchange numerous messages between themselves as part of executing a specific workload. For example, a computational workload could require data, or an executable, to be disseminated to workers distributed throughout the data center which may be subsequently polled for status updates. The emergence of such applications means there is likely to be an increase in one-to-many traffic flows with the increasing dominance of East-West traffic.

The TV broadcast industry is another potential future source of applications with one-to-many traffic patterns in data centers. The requirement for robustness, stability and predicability has meant the TV broadcast industry has traditionally used TV-specific protocols, infrastructure and technologies for transmitting video signals between end points such as cameras, monitors, mixers, graphics

devices and video servers. However, the growing cost and complexity of supporting this approach, especially as the bit rates of the video signals increase due to demand for formats such as 4K-UHD and 8K-UHD, means there is a consensus that the TV broadcast industry will transition from industry-specific transmission formats (e.g. SDI, HD-SDI) over TV-specific infrastructure to using IP-based infrastructure. The development of pertinent standards by the Society of Motion Picture and Television Engineers (SMPTE) [SMPTE2110], along with the increasing performance of IP routers, means this transition is gathering pace. A possible outcome of this transition will be the building of IP data centers in broadcast plants. Traffic flows in the broadcast industry are frequently one-to-many and so if IP data centers are deployed in broadcast plants, it is imperative that this traffic pattern is supported efficiently in that infrastructure. In fact, a pivotal consideration for broadcasters considering transitioning to IP is the manner in which these one-to-many traffic flows will be managed and monitored in a data center with an IP fabric.

One of the few success stories in using conventional IP multicast has been for disseminating market trading data. For example, IP multicast is commonly used today to deliver stock quotes from stock exchanges to financial service providers and then to the stock analysts or brokerages. It is essential that the network infrastructure delivers very low latency and high throughput, especially given the proliferation of automated and algorithmic trading which means stock analysts or brokerages may gain an edge on competitors simply by receiving an update a few milliseconds earlier. As would be expected, in such deployments reliability is critical. The network must be designed with no single point of failure and in such a way that it can respond in a deterministic manner to failure. Typically, redundant servers (in a primary/backup or live-live mode) send multicast streams into the network, with diverse paths being used across the network. The stock exchange generating the one-to-many traffic and stock analysts/brokerage that receive the traffic will typically have their own data centers. Therefore, the manner in which one-to-many traffic patterns are handled in these data centers are extremely important, especially given the requirements and constraints mentioned.

Another reason for the growing volume of one-to-many traffic patterns in modern data centers is the increasing adoption of streaming telemetry. This transition is motivated by the observation that traditional poll-based approaches for monitoring network devices are usually inadequate in modern data centers. These approaches typically suffer from poor scalability, extensibility and responsiveness. In contrast, in streaming telemetry, network devices in the data center stream highly-granular real-time updates to a

telemetry collector/database. This collector then collates, normalizes and encodes this data for convenient consumption by monitoring applications. The monitoring applications can subscribe to the notifications of interest, allowing them to gain insight into pertinent state and performance metrics. Thus, the traffic flows associated with streaming telemetry are typically many-to-one between the network devices and the telemetry collector and then one-to-many from the collector to the monitoring applications.

The use of publish and subscribe applications is growing within data centers, contributing to the rising volume of one-to-many traffic flows. Such applications are attractive as they provide a robust low-latency asynchronous messaging service, allowing senders to be decoupled from receivers. The usual approach is for a publisher to create and transmit a message to a specific topic. The publish and subscribe application will retain the message and ensure it is delivered to all subscribers to that topic. The flexibility in the number of publishers and subscribers to a specific topic means such applications cater for one-to-one, one-to-many and many-to-one traffic patterns.

2.2. Overlays

Another key contributor to the rise in one-to-many traffic patterns is the proposed architecture for supporting large-scale multi-tenancy in highly virtualized data centers [RFC8014]. In this architecture, a tenant's VMs are distributed across the data center and are connected by a virtual network known as the overlay network. A number of different technologies have been proposed for realizing the overlay network, including VXLAN [RFC7348], VXLAN-GPE [I-D.ietf-nvo3-vxlan-gpe], NVGRE [RFC7637] and GENEVE [I-D.ietf-nvo3-geneve]. The often fervent and arguably partisan debate about the relative merits of these overlay technologies belies the fact that, conceptually, it may be said that these overlays simply provide a means to encapsulate and tunnel Ethernet frames from the VMs over the data center IP fabric, thus emulating a Layer 2 segment between the VMs. Consequently, the VMs believe and behave as if they are connected to the tenant's other VMs by a conventional Layer 2 segment, regardless of their physical location within the data center.

Naturally, in a Layer 2 segment, point to multi-point traffic can result from handling BUM (broadcast, unknown unicast and multicast) traffic. And, compounding this issue within data centers, since the tenant's VMs attached to the emulated segment may be dispersed throughout the data center, the BUM traffic may need to traverse the data center fabric.

Hence, regardless of the overlay technology used, due consideration must be given to handling BUM traffic, forcing the data center operator to pay attention to the manner in which one-to-many communication is handled within the data center. And this consideration is likely to become increasingly important with the anticipated rise in the number and importance of overlays. In fact, it may be asserted that the manner in which one-to-many communications arising from overlays is handled is pivotal to the performance and stability of the entire data center network.

2.3. Protocols

Conventionally, some key networking protocols used in data centers require one-to-many communications for control messages. Thus, the data center operator must pay due attention to how these control message exchanges are supported.

For example, ARP [RFC0826] and ND [RFC4861] use broadcast and multicast messages within IPv4 and IPv6 networks respectively to discover MAC address to IP address mappings. Furthermore, when these protocols are running within an overlay network, it is essential to ensure the messages are delivered to all the hosts on the emulated Layer 2 segment, regardless of physical location within the data center. The challenges associated with optimally delivering ARP and ND messages in data centers has attracted lots of attention [RFC6820].

Another example of a protocol that may necessitate having one-to-many traffic flows in the data center is IGMP [RFC2236], [RFC3376]. If the VMs attached to the Layer 2 segment wish to join a multicast group they must send IGMP reports in response to queries from the querier. As these devices could be located at different locations within the data center, there is the somewhat ironic prospect of IGMP itself leading to an increase in the volume of one-to-many communications in the data center.

2.4. Summary

Section 2.1, Section 2.2 and Section 2.3 have discussed how the trends in the types of applications, the overlay technologies used and some of the essential networking protocols results in an increase in the volume of one-to-many traffic patterns in modern highly-virtualized data centers. Section 3 explores how such traffic flows may be handled using conventional IP multicast.

3. Handling one-to-many traffic using conventional multicast

Faced with ever increasing volumes of one-to-many traffic flows, for the reasons presented in Section 2, it makes sense for a data center operator to explore if and how conventional IP multicast could be deployed within the data center. This section introduces the key protocols, discusses some example use cases where they are deployed in data centers and discusses some of the advantages and disadvantages of such deployments.

3.1. Layer 3 multicast

PIM is the most widely deployed multicast routing protocol and so, unsurprisingly, is the primary multicast routing protocol considered for use in the data center. There are three potential popular modes of PIM that may be used: PIM-SM [RFC4601], PIM-SSM [RFC4607] or PIM-BIDIR [RFC5015]. It may be said that these different modes of PIM tradeoff the optimality of the multicast forwarding tree for the amount of multicast forwarding state that must be maintained at routers. SSM provides the most efficient forwarding between sources and receivers and thus is most suitable for applications with one-to-many traffic patterns. State is built and maintained for each (S,G) flow. Thus, the amount of multicast forwarding state held by routers in the data center is proportional to the number of sources and groups. At the other end of the spectrum, BIDIR is the most efficient shared tree solution as one tree is built for all flows, therefore minimizing the amount of state. This state reduction is at the expense of optimal forwarding path between sources and receivers. This use of a shared tree makes BIDIR particularly well-suited for applications with many-to-many traffic patterns, given that the amount of state is uncorrelated to the number of sources. SSM and BIDIR are optimizations of PIM-SM. PIM-SM is the most widely deployed multicast routing protocol. PIM-SM can also be the most complex. PIM-SM relies upon a RP (Rendezvous Point) to set up the multicast tree and subsequently there is the option of switching to the SPT (shortest path tree), similar to SSM, or staying on the shared tree, similar to BIDIR.

3.2. Layer 2 multicast

With IPv4 unicast address resolution, the translation of an IP address to a MAC address is done dynamically by ARP. With multicast address resolution, the mapping from a multicast IPv4 address to a multicast MAC address is done by assigning the low-order 23 bits of the multicast IPv4 address to fill the low-order 23 bits of the multicast MAC address. Each IPv4 multicast address has 28 unique bits (the multicast address range is 224.0.0.0/12) therefore mapping a multicast IP address to a MAC address ignores 5 bits of the IP

address. Hence, groups of 32 multicast IP addresses are mapped to the same MAC address. And so a multicast MAC address cannot be uniquely mapped to a multicast IPv4 address. Therefore, IPv4 multicast addresses must be chosen judiciously in order to avoid unnecessary address aliasing. When sending IPv6 multicast packets on an Ethernet link, the corresponding destination MAC address is a direct mapping of the last 32 bits of the 128 bit IPv6 multicast address into the 48 bit MAC address. It is possible for more than one IPv6 multicast address to map to the same 48 bit MAC address.

The default behaviour of many hosts (and, in fact, routers) is to block multicast traffic. Consequently, when a host wishes to join an IPv4 multicast group, it sends an IGMP [RFC2236], [RFC3376] report to the router attached to the Layer 2 segment and also it instructs its data link layer to receive Ethernet frames that match the corresponding MAC address. The data link layer filters the frames, passing those with matching destination addresses to the IP module. Similarly, hosts simply hand the multicast packet for transmission to the data link layer which would add the Layer 2 encapsulation, using the MAC address derived in the manner previously discussed.

When this Ethernet frame with a multicast MAC address is received by a switch configured to forward multicast traffic, the default behaviour is to flood it to all the ports in the Layer 2 segment. Clearly there may not be a receiver for this multicast group present on each port and IGMP snooping is used to avoid sending the frame out of ports without receivers.

A switch running IGMP snooping listens to the IGMP messages exchanged between hosts and the router in order to identify which ports have active receivers for a specific multicast group, allowing the forwarding of multicast frames to be suitably constrained. Normally, the multicast router will generate IGMP queries to which the hosts send IGMP reports in response. However, number of optimizations in which a switch generates IGMP queries (and so appears to be the router from the hosts' perspective) and/or generates IGMP reports (and so appears to be hosts from the router's perspective) are commonly used to improve the performance by reducing the amount of state maintained at the router, suppressing superfluous IGMP messages and improving responsiveness when hosts join/leave the group.

Multicast Listener Discovery (MLD) [RFC 2710] [RFC 3810] is used by IPv6 routers for discovering multicast listeners on a directly attached link, performing a similar function to IGMP in IPv4 networks. MLDv1 [RFC 2710] is similar to IGMPv2 and MLDv2 [RFC 3810] [RFC 4604] similar to IGMPv3. However, in contrast to IGMP, MLD does not send its own distinct protocol messages. Rather, MLD is a subprotocol of ICMPv6 [RFC 4443] and so MLD messages are a subset of

ICMPv6 messages. MLD snooping works similarly to IGMP snooping, described earlier.

3.3. Example use cases

A use case where PIM and IGMP are currently used in data centers is to support multicast in VXLAN deployments. In the original VXLAN specification [RFC7348], a data-driven flood and learn control plane was proposed, requiring the data center IP fabric to support multicast routing. A multicast group is associated with each virtual network, each uniquely identified by its VXLAN network identifiers (VNI). VXLAN tunnel endpoints (VTEPs), typically located in the hypervisor or ToR switch, with local VMs that belong to this VNI would join the multicast group and use it for the exchange of BUM traffic with the other VTEPs. Essentially, the VTEP would encapsulate any BUM traffic from attached VMs in an IP multicast packet, whose destination address is the associated multicast group address, and transmit the packet to the data center fabric. Thus, a multicast routing protocol (typically PIM) must be running in the fabric to maintain a multicast distribution tree per VNI.

Alternatively, rather than setting up a multicast distribution tree per VNI, a tree can be set up whenever hosts within the VNI wish to exchange multicast traffic. For example, whenever a VTEP receives an IGMP report from a locally connected host, it would translate this into a PIM join message which will be propagated into the IP fabric. In order to ensure this join message is sent to the IP fabric rather than over the VXLAN interface (since the VTEP will have a route back to the source of the multicast packet over the VXLAN interface and so would naturally attempt to send the join over this interface) a more specific route back to the source over the IP fabric must be configured. In this approach PIM must be configured on the SVIs associated with the VXLAN interface.

Another use case of PIM and IGMP in data centers is when IPTV servers use multicast to deliver content from the data center to end users. IPTV is typically a one to many application where the hosts are configured for IGMPv3, the switches are configured with IGMP snooping, and the routers are running PIM-SSM mode. Often redundant servers send multicast streams into the network and the network forwards the data across diverse paths.

3.4. Advantages and disadvantages

Arguably the biggest advantage of using PIM and IGMP to support one-to-many communication in data centers is that these protocols are relatively mature. Consequently, PIM is available in most routers and IGMP is supported by most hosts and routers. As such, no

specialized hardware or relatively immature software is involved in using these protocols in data centers. Furthermore, the maturity of these protocols means their behaviour and performance in operational networks is well-understood, with widely available best-practices and deployment guides for optimizing their performance. For these reasons, PIM and IGMP have been used successfully for supporting one-to-many traffic flows within modern data centers, as discussed earlier.

However, somewhat ironically, the relative disadvantages of PIM and IGMP usage in data centers also stem mostly from their maturity. Specifically, these protocols were standardized and implemented long before the highly-virtualized multi-tenant data centers of today existed. Consequently, PIM and IGMP are neither optimally placed to deal with the requirements of one-to-many communication in modern data centers nor to exploit idiosyncrasies of data centers. For example, there may be thousands of VMs participating in a multicast session, with some of these VMs migrating to servers within the data center, new VMs being continually spun up and wishing to join the sessions while all the time other VMs are leaving. In such a scenario, the churn in the PIM and IGMP state machines, the volume of control messages they would generate and the amount of state they would necessitate within routers, especially if they were deployed naively, would be untenable. Furthermore, PIM is a relatively complex protocol. As such, PIM can be challenging to debug even in significantly more benign deployments than those envisaged for future data centers, a fact that has evidently had a dissuasive effect on data center operators considering enabling it within the IP fabric.

4. Alternative options for handling one-to-many traffic

Section 2 has shown that there is likely to be an increasing amount one-to-many communications in data centers for multiple reasons. And Section 3 has discussed how conventional multicast may be used to handle this traffic, presenting some of the associated advantages and disadvantages. Unsurprisingly, as discussed in the remainder of Section 4, there are a number of alternative options of handling this traffic pattern in data centers. Critically, it should be noted that many of these techniques are not mutually-exclusive; in fact many deployments involve a combination of more than one of these techniques. Furthermore, as will be shown, introducing a centralized controller or a distributed control plane, typically makes these techniques more potent.

4.1. Minimizing traffic volumes

If handling one-to-many traffic flows in data centers is considered onerous, then arguably the most intuitive solution is to aim to minimize the volume of said traffic.

It was previously mentioned in Section 2 that the three main contributors to one-to-many traffic in data centers are applications, overlays and protocols. Typically the applications running on VMs are outside the control of the data center operator and thus, relatively speaking, little can be done about the volume of one-to-many traffic generated by applications. Luckily, there is more scope for attempting to reduce the volume of such traffic generated by overlays and protocols. (And often by protocols within overlays.) This reduction is possible by exploiting certain characteristics of data center networks such as a fixed and regular topology, single administrative control, consistent hardware and software, well-known overlay encapsulation endpoints and systematic IP address allocation.

A way of minimizing the amount of one-to-many traffic that traverses the data center fabric is to use a centralized controller. For example, whenever a new VM is instantiated, the hypervisor or encapsulation endpoint can notify a centralized controller of this new MAC address, the associated virtual network, IP address etc. The controller could subsequently distribute this information to every encapsulation endpoint. Consequently, when any endpoint receives an ARP request from a locally attached VM, it could simply consult its local copy of the information distributed by the controller and reply. Thus, the ARP request is suppressed and does not result in one-to-many traffic traversing the data center IP fabric.

Alternatively, the functionality supported by the controller can be realized by a distributed control plane. BGP-EVPN [RFC7432, RFC8365] is the most popular control plane used in data centers. Typically, the encapsulation endpoints will exchange pertinent information with each other by all peering with a BGP route reflector (RR). Thus, information such as local MAC addresses, MAC to IP address mapping, virtual networks identifiers, IP prefixes, and local IGMP group membership can be disseminated. Consequently, for example, ARP requests from local VMs can be suppressed by the encapsulation endpoint using the information learnt from the control plane about the MAC to IP mappings at remote peers. In a similar fashion, encapsulation endpoints can use information gleaned from the BGP-EVPN messages to proxy for both IGMP reports and queries for the attached VMs, thus obviating the need to transmit IGMP messages across the data center fabric.

4.2. Head end replication

A popular option for handling one-to-many traffic patterns in data centers is head end replication (HER). HER means the traffic is duplicated and sent to each end point individually using conventional IP unicast. Obvious disadvantages of HER include traffic duplication and the additional processing burden on the head end. Nevertheless, HER is especially attractive when overlays are in use as the replication can be carried out by the hypervisor or encapsulation end point. Consequently, the VMs and IP fabric are unmodified and unaware of how the traffic is delivered to the multiple end points. Additionally, it is possible to use a number of approaches for constructing and disseminating the list of which endpoints should receive what traffic and so on.

For example, the reluctance of data center operators to enable PIM within the data center fabric means VXLAN is often used with HER. Thus, BUM traffic from each VNI is replicated and sent using unicast to remote VTEPs with VMs in that VNI. The list of remote VTEPs to which the traffic should be sent may be configured manually on the VTEP. Alternatively, the VTEPs may transmit pertinent local state to a centralized controller which in turn sends each VTEP the list of remote VTEPs for each VNI. Lastly, HER also works well when a distributed control plane is used instead of the centralized controller. Again, BGP-EVPN may be used to distribute the information needed to facilitate HER to the VTEPs.

4.3. Programmable Forwarding Planes

As discussed in Section 2, one of the main functions of PIM is to build and maintain multicast distribution trees. Such a tree indicates the path a specific flow will take through the network. Thus, in routers traversed by the flow, the information from PIM is ultimately used to create a multicast forwarding entry for the specific flow and insert it into the multicast forwarding table. The multicast forwarding table will have entries for each multicast flow traversing the router, with the lookup key usually being a concatenation of the source and group addresses. Critically, each entry will contain information such as the legal input interface for the flow and a list of output interfaces to which matching packets should be replicated.

Viewed in this way, there is nothing remarkable about the multicast forwarding state constructed in routers based on the information gleaned from PIM. And, in fact, it is perfectly feasible to build such state in the absence of PIM. Such prospects have been significantly enhanced with the increasing popularity and performance of network devices with programmable forwarding planes. These

devices are attractive for use in data centers since they are amenable to being programmed by a centralized controller. If such a controller has a global view of the sources and receivers for each multicast flow (which can be provided by the devices attached to the end hosts in the data center communicating with the controller), an accurate representation of data center topology (which is usually well-known), then it can readily compute the multicast forwarding state that must be installed at each router to ensure the one-to-many traffic flow is delivered properly to the correct receivers. All that is needed is an API to program the forwarding planes of all the network devices that need to handle the flow appropriately. Such APIs do in fact exist and so, unsurprisingly, handling one-to-many traffic flows using such an approach is attractive for data centers.

Being able to program the forwarding plane in this manner offers the enticing possibility of introducing novel algorithms and concepts for forwarding multicast traffic in data centers. These schemes typically aim to exploit the idiosyncracies of the data center network architecture to create ingenious, pithy and elegant encodings of the information needed to facilitate multicast forwarding. Depending on the scheme, this information may be carried in packet headers, stored in the multicast forwarding table in routers or a combination of both. The key characteristic is that the terseness of the forwarding information means the volume of forwarding state is significantly reduced. Additionally, the overhead associated with building and maintaining a multicast forwarding tree has been eliminated. The result of these reductions in the overhead associated with multicast forwarding is a significant and impressive increase in the effective number of multicast flows that can be supported within the data center.

[Shabaz19] is a good example of such an approach and also presents comprehensive discussion of other schemes in the discussion on related work. Although a number of promising schemes have been proposed, no consensus has yet emerged as to which approach is best, and in fact what "best" means. Even if a clear winner were to emerge, it faces significant challenges to gain the vendor and operator buy-in to ensure it is widely deployed in data centers.

4.4. BIER

As discussed in Section 3.4, PIM and IGMP face potential scalability challenges when deployed in data centers. These challenges are typically due to the requirement to build and maintain a distribution tree and the requirement to hold per-flow state in routers. Bit Index Explicit Replication (BIER) [RFC 8279] is a new multicast forwarding paradigm that avoids these two requirements.

When a multicast packet enters a BIER domain, the ingress router, known as the Bit-Forwarding Ingress Router (BFIR), adds a BIER header to the packet. This header contains a bit string in which each bit maps to an egress router, known as Bit-Forwarding Egress Router (BFER). If a bit is set, then the packet should be forwarded to the associated BFER. The routers within the BIER domain, Bit-Forwarding Routers (BFRs), use the BIER header in the packet and information in the Bit Index Forwarding Table (BIFT) to carry out simple bit-wise operations to determine how the packet should be replicated optimally so it reaches all the appropriate BFERs.

BIER is deemed to be attractive for facilitating one-to-many communications in data centers [I-D.ietf-bier-use-cases]. The BFIRs are the encapsulation endpoints in the deployment envisioned with overlay networks. So knowledge about the actual multicast groups does not reside in the data center fabric, improving the scalability compared to conventional IP multicast. Additionally, a centralized controller or a BGP-EVPN control plane may be used with BIER to ensure the BFIR have the required information. A challenge associated with using BIER is that it requires changes to the forwarding behaviour of the routers used in the data center IP fabric.

4.5. Segment Routing

Segment Routing (SR) [RFC8402] is a manifestation of the source routing paradigm, so called as the path a packet takes through a network is determined at the source. The source encodes this information in the packet header as a sequence of instructions. These instructions are followed by intermediate routers, ultimately resulting in the delivery of the packet to the desired destination. In SR, the instructions are known as segments and a number of different kinds of segments have been defined. Each segment has an identifier (SID) which is distributed throughout the network by newly defined extensions to standard routing protocols. Thus, using this information, sources are able to determine the exact sequence of segments to encode into the packet. The manner in which these instructions are encoded depends on the underlying data plane. Segment Routing can be applied to the MPLS and IPv6 data planes. In the former, the list of segments is represented by the label stack and in the latter it is represented as an IPv6 routing extension header. Advantages of segment routing include the reduction in the amount of forwarding state routers need to hold and the removal of the need to run a signaling protocol, thus improving the network scalability while reducing the operational complexity.

The advantages of segment routing and the ability to run it over an unmodified MPLS data plane means that one of its anticipated use

cases is in BGP-based large-scale data centers [RFC7938]. The exact manner in which multicast traffic will be handled in SR has not yet been standardized, with a number of different options being considered. For example, since with the MPLS data plane, segments are simply encoded as a label stack, then the protocols traditionally used to create point-to-multipoint LSPs could be reused to allow SR to support one-to-many traffic flows. Alternatively, a special SID may be defined for a multicast distribution tree, with a centralized controller being used to program routers appropriately to ensure the traffic is delivered to the desired destinations, while avoiding the costly process of building and maintaining a multicast distribution tree.

5. Conclusions

As the volume and importance of one-to-many traffic in data centers increases, conventional IP multicast is likely to become increasingly unattractive for deployment in data centers for a number of reasons, mostly pertaining its relatively poor scalability and inability to exploit characteristics of data center network architectures. Hence, even though IGMP/MLD is likely to remain the most popular manner in which end hosts signal interest in joining a multicast group, it is unlikely that this multicast traffic will be transported over the data center IP fabric using a multicast distribution tree built and maintained by PIM in the future. Rather, approaches which exploit idiosyncracies of data center network architectures are better placed to deliver one-to-many traffic in data centers, especially when judiciously combined with a centralized controller and/or a distributed control plane, particularly one based on BGP-EVPN.

6. IANA Considerations

This memo includes no request to IANA.

7. Security Considerations

No new security considerations result from this document

8. Acknowledgements

9. References

9.1. 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>>.

9.2. Informative References

- [I-D.ietf-bier-use-cases]
Kumar, N., Asati, R., Chen, M., Xu, X., Dolganow, A., Przygienda, T., Gulko, A., Robinson, D., Arya, V., and C. Bestler, "BIER Use Cases", draft-ietf-bier-use-cases-09 (work in progress), January 2019.
- [I-D.ietf-nvo3-geneve]
Gross, J., Ganga, I., and T. Sridhar, "Geneve: Generic Network Virtualization Encapsulation", draft-ietf-nvo3-geneve-13 (work in progress), March 2019.
- [I-D.ietf-nvo3-vxlan-gpe]
Maino, F., Kreeger, L., and U. Elzur, "Generic Protocol Extension for VXLAN", draft-ietf-nvo3-vxlan-gpe-07 (work in progress), April 2019.
- [RFC0826] Plummer, D., "An Ethernet Address Resolution Protocol: Or Converting Network Protocol Addresses to 48.bit Ethernet Address for Transmission on Ethernet Hardware", STD 37, RFC 826, DOI 10.17487/RFC0826, November 1982, <<https://www.rfc-editor.org/info/rfc826>>.
- [RFC2236] Fenner, W., "Internet Group Management Protocol, Version 2", RFC 2236, DOI 10.17487/RFC2236, November 1997, <<https://www.rfc-editor.org/info/rfc2236>>.
- [RFC2710] Deering, S., Fenner, W., and B. Haberman, "Multicast Listener Discovery (MLD) for IPv6", RFC 2710, DOI 10.17487/RFC2710, October 1999, <<https://www.rfc-editor.org/info/rfc2710>>.
- [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>>.
- [RFC4601] Fenner, B., Handley, M., Holbrook, H., and I. Kouvelas, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", RFC 4601, DOI 10.17487/RFC4601, August 2006, <<https://www.rfc-editor.org/info/rfc4601>>.
- [RFC4607] Holbrook, H. and B. Cain, "Source-Specific Multicast for IP", RFC 4607, DOI 10.17487/RFC4607, August 2006, <<https://www.rfc-editor.org/info/rfc4607>>.

- [RFC4861] Narten, T., Nordmark, E., Simpson, W., and H. Soliman, "Neighbor Discovery for IP version 6 (IPv6)", RFC 4861, DOI 10.17487/RFC4861, September 2007, <<https://www.rfc-editor.org/info/rfc4861>>.
- [RFC5015] Handley, M., Kouvelas, I., Speakman, T., and L. Vicisano, "Bidirectional Protocol Independent Multicast (BIDIR-PIM)", RFC 5015, DOI 10.17487/RFC5015, October 2007, <<https://www.rfc-editor.org/info/rfc5015>>.
- [RFC6820] Narten, T., Karir, M., and I. Foo, "Address Resolution Problems in Large Data Center Networks", RFC 6820, DOI 10.17487/RFC6820, January 2013, <<https://www.rfc-editor.org/info/rfc6820>>.
- [RFC7348] Mahalingam, M., Dutt, D., Duda, K., Agarwal, P., Kreeger, L., Sridhar, T., Bursell, M., and C. Wright, "Virtual eXtensible Local Area Network (VXLAN): A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks", RFC 7348, DOI 10.17487/RFC7348, August 2014, <<https://www.rfc-editor.org/info/rfc7348>>.
- [RFC7432] Sajassi, A., Ed., Aggarwal, R., Bitar, N., Isaac, A., Uttaro, J., Drake, J., and W. Henderickx, "BGP MPLS-Based Ethernet VPN", RFC 7432, DOI 10.17487/RFC7432, February 2015, <<https://www.rfc-editor.org/info/rfc7432>>.
- [RFC7637] Garg, P., Ed. and Y. Wang, Ed., "NVGRE: Network Virtualization Using Generic Routing Encapsulation", RFC 7637, DOI 10.17487/RFC7637, September 2015, <<https://www.rfc-editor.org/info/rfc7637>>.
- [RFC7938] Lapukhov, P., Premji, A., and J. Mitchell, Ed., "Use of BGP for Routing in Large-Scale Data Centers", RFC 7938, DOI 10.17487/RFC7938, August 2016, <<https://www.rfc-editor.org/info/rfc7938>>.
- [RFC8014] Black, D., Hudson, J., Kreeger, L., Lasserre, M., and T. Narten, "An Architecture for Data-Center Network Virtualization over Layer 3 (NVO3)", RFC 8014, DOI 10.17487/RFC8014, December 2016, <<https://www.rfc-editor.org/info/rfc8014>>.
- [RFC8279] Wijnands, IJ., Ed., Rosen, E., Ed., Dolganow, A., Przygienda, T., and S. Aldrin, "Multicast Using Bit Index Explicit Replication (BIER)", RFC 8279, DOI 10.17487/RFC8279, November 2017, <<https://www.rfc-editor.org/info/rfc8279>>.

- [RFC8365] Sajassi, A., Ed., Drake, J., Ed., Bitar, N., Shekhar, R., Uttaro, J., and W. Henderickx, "A Network Virtualization Overlay Solution Using Ethernet VPN (EVPN)", RFC 8365, DOI 10.17487/RFC8365, March 2018, <<https://www.rfc-editor.org/info/rfc8365>>.
- [RFC8402] Filsfils, C., Ed., Previdi, S., Ed., Ginsberg, L., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing Architecture", RFC 8402, DOI 10.17487/RFC8402, July 2018, <<https://www.rfc-editor.org/info/rfc8402>>.
- [Shabaz19] Shabaz, M., Suresh, L., Rexford, J., Feamster, N., Rottenstreich, O., and M. Hira, "Elmo: Source Routed Multicast for Public Clouds", ACM SIGCOMM 2019 Conference (SIGCOMM '19) ACM, DOI 10.1145/3341302.3342066, August 2019.
- [SMPTE2110] "SMPTE2110 Standards Suite", <<http://www.smpste.org/st-2110>>.

Authors' Addresses

Mike McBride
Futurewei

Email: michael.mcbride@futurewei.com

Olufemi Komolafe
Arista Networks

Email: femi@arista.com

L3VPN
Internet-Draft
Intended status: Standards Track
Expires: January 5, 2015

R. Kebler
P. Kurapati
Juniper Networks
S. Asif
AT&T LABS
July 4, 2014

Multicast Traceroute for MVPNs
draft-kebler-kurapati-l3vpn-mvpn-mtrace-01

Abstract

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

Status of This Memo

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

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

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

This Internet-Draft will expire on January 5, 2015.

Copyright Notice

Copyright (c) 2014 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of

the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

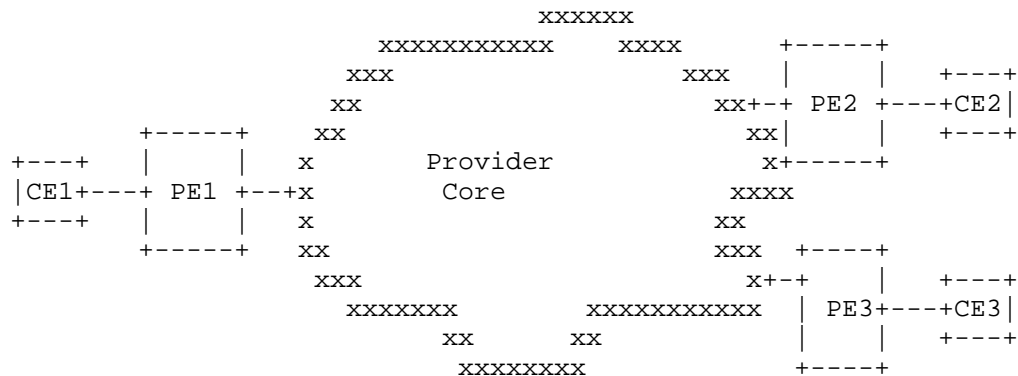
Table of Contents

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

1. Introduction

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

the Mtrace Response that PE2 has an issue, the existing error codes are not sufficient to identify the root cause. Also, there could be a problem sending traffic over the provider tunnel from PE2 to PE1, but the mtrace query will not even travel over this provider tunnel. Therefore, the mtrace successful response can be misleading. This draft ensures that the Response uses same provider-tunnel that the given C-S,C-G data would traverse and returns appropriate MVPN specific error codes which would help in identifying the root cause.



MVPN topology

2. Overview

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

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

can reach the egress PE. A Querier and egress PE can be the same router.

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

3. Protocol Details

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

3.1. Mtrace Query

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

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

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

3.2. Mtrace Request

The PE that receives the query will lookup the (C-S,C-G) using the RD of the query to distinguish the vrf. If the RD doesn't match any VRF, PE sends a response with error code set to BAD_RD. The PE first checks the C-Mcast route that is matching (C-S,C-G) of the mtrace Query. It then finds the upstream multicast hop from the selected C-Mcast route and unicasts the requests to the upstream multicast hop after decrementing the TTL. The Mtrace Request MUST have PMSI Tunnel Attributes Augmented Response Block populated with the PMSI attribute that the PE uses to receive the traffic for the given (C-S,C-G) traffic.

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

When a wild card query is received, the PE will look for the upstream PE address in the MVPN Extended query block. The PE will then check if it has bound to the wild card SPMSI tunnel from the specified upstream PE. If it has, it will populate the Leaf A-D Augmented Response Block and PMSI Tunnel Attributes Augmented Response Block with the respective values. If the PE has not received any wild card SPMSI AD route from the specified upstream PE in the query, it should send a response with the error code set to NO_WILD_CARD_SPMSI_AD_RCVD. If the PE has received wild card SPMSI AD route from the upstream PE, but has not responded with a LEAF-AD route, it should send a response with the error code set to NO_WILD_CARD_SPMSI_LEAF_AD_SENT.

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

If the route does not match any MVPN-TIB state, then the PE should send a Response to the Querier with the error code set to NO_CMCAST_STATE. If the PE cannot locate the upstream PE then it should send a response to the Querier with the NO_UPSTREAM_PE error code.

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

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

Section 8 of MVPN specification [RFC6513] mentions two procedures (Segmented and Non-Segmented) for handling Inter-AS scenarios. If the local and the upstream PEs are in different ASes, and if segmented Inter-AS procedure is used, then the local PE finds in its VRF an Inter-AS I-PMSI A-D route whose Source AS field carries the ASN of the upstream PE. The RD of the found Inter-AS I-PMSI A-D route is used as the RD of the mtrace Query. If Inter-AS I-PMSI A-D route is not found, a response with error code UNKNOWN_INTER_AS is sent.

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

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

3.2.1. Ingress PE Procedures

The PE that receives the Request, will check the PMSI attributes of the sender of request to see if they match the values used to send traffic for the S,G. If the values do not match, then the PE uses the appropriate pmsi error code as specified in 'MVPN Error Codes' section and sends a mtrace Response back to the Querier. Also, if a Leaf A-D Augmented Response Block is included, the PE will validate that it has received this Leaf A-D route from the router that sent the Request. If not, then this PE should change the error code to BAD_LEAF_AD and send the Response to the Querier. If the PE expects that a Leaf A-D route is needed for the downstream PE to receive traffic, but did not receive one in the mtrace Request from the sending router, then it should use a NO_LEAF_AD_RCVD error code for

the mtrace Response. For a wild card SPMSI query, if the PE didn't receive LEAF AD route from the downstream PE, it should use NO_LEAD_AD_RCVD error code.

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

The ingress PE will also check it has local vrf connectivity for the source/RP. If it does not have any connectivity to the source/RP then it should use the base specification error code NO_ROUTE and send an mtrace Response. Note that in a Virtual Hub and Spoke environment, it is possible for a PE to receive a mtrace Request and need to propagate it to another upstream PE. These procedures are outlined in the section "Virtual Hub and Spoke". If the PE does not expect to be receiving mtrace Responses from the mvpn core and have the route to the source located via another upstream PE, then it can use the base specification RPF_IF error code.

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

3.3. Downstream Requests

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

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

3.4. ASBR Behavior

When an ASBR receives a mtrace Request the ASBR finds an Inter-AS I-PMSI A-D route whose RD and Source AS matches the RD and Source AS carried in the mtrace Query. If no matching route is found and the ASBR is using segmented tunnels as described in MVPN specification [RFC6513], the ASBR sends an UNKNOWN_INTER_AS error code back to the Querier. If a matching route is found, the ASBR acts as a "first hop router" and modifies the Query type to DOWNSTREAM_REQUEST. ASBR in this case MUST validate the PMSI attributes similar to the "first hop router" and respond if there is any errors. ASBR MUST populate PMSI Tunnel Attributes Augmented Response Block with the Inter-AS provider tunnel information before sending the DOWNSTREAM_REQUEST. Note that the mtrace request does not proceed upstream as it is assumed that performing a traceroute and exposing IP addresses across AS boundaries would not be desirable with Segmented Inter-AS Provider Tunnels.

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

3.5. Virtual Hub and Spoke

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

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

3.6. Inter-Area Provider Tunnels

3.6.1. Egress PE

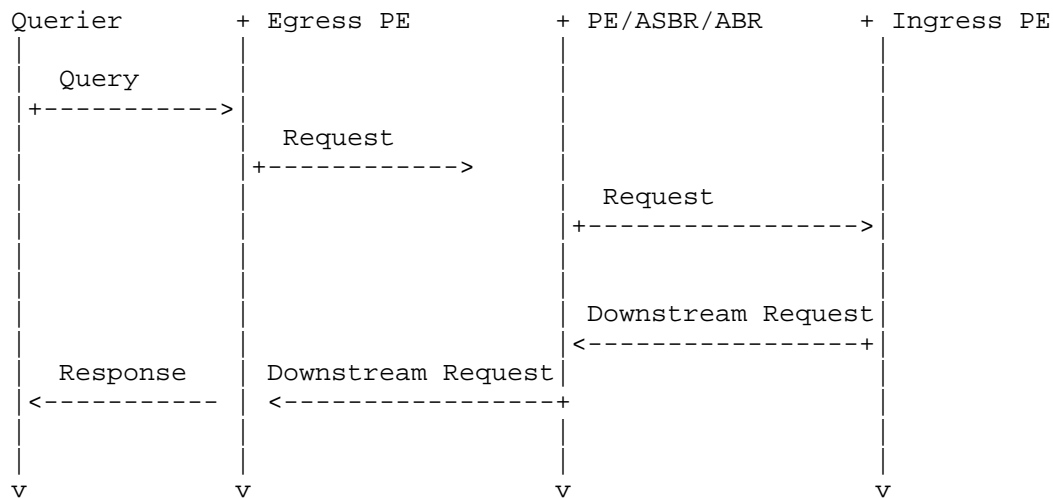
The egress PE does the same procedures as specified in Section "Mtrace Request" except it sends the Request upstream to the IP address determined from the Global Administrator field of the Inter-area P2MP Segmented Next-hop Extended Community as described in specification [I-D.ietf-mpls-seamless-mcast] . If the egress PE has sent a Leaf-AD route then it must send a Leaf-AD Augmented Response Block with the NLRI of the Leaf A-D route.

3.6.2. ABR Behavior

ABR MUST find a S-PMSI or I-PMSI route whose NLRI has the same value as the Route Key field of the received mtrace Leaf-AD extended Query Block. If such a matching route is not found then a Response should be sent to the Querier with the NO_LEAF_AD_RCVD. If the ABR has sent a Leaf-AD route then it must add a Leaf-AD Augmented Response Block with the values of Leaf A-D route NLRI. The upstream node's IP address is the IP address determined from the Global Administrator field of the Inter-area P2MP Segmented Next-hop Extended Community.

3.7. Mtrace MVPN Procedure

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



Mtrace MVPN Procedure

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

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

- 1 - Querier sends the Mtrace Query towards LHR (Egress PE-Spoke).
- 2 - Egress PE sends Request to V-HUB. V-HUB realises that the first hop router is a connected spoke and sends the request to Ingress Spoke PE.
- 3 - Ingress Spoke PE sends Downstream Request to V-HUB. The same is received by V-HUB. V-HUB sets the 'D' bit in its PMSI Tunnel Attributes Augmented Response Block.
- 4 - V-HUB sends Downstream request to ingress spoke. This is never received by the ingress spoke.

5 - The result of first 4 steps is that querier did not receive the response. This makes the querier fall back to TTL method.

6 - Querier reduces the TTL and the result will show that the hop from V-HUB to ingress spoke is missing thereby pointing the issue at the right place.

4. Error Detection

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

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

In some situations the router needs to send a Leaf AD route to the upstream PE. If the upstream expects a Leaf AD route, but did not receive one from the downstream PE, then the NO_LEAF_AD_RCVD error will be sent.

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

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

4.1. MVPN Error Codes

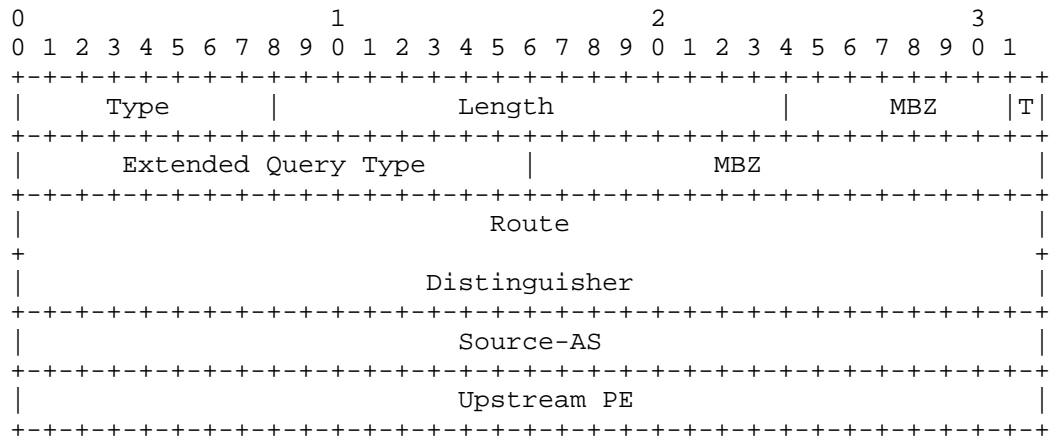
Value	Name	Description
-----	-----	-----
0x11	PTUNNEL_DOWN	The provide tunnel for this S,G is
down.		
0x12	NO_LEAF_AD_RCVD	The S-PMSI has not been joined by
		downstream neighbor
0x13	BAD_LEAF_AD	The Leaf A-D route does not match
		the expected values
0x14	BAD_RD	The RD is known to not exist on th
is PE		
0x15	UNEXPECTED_MVPN	The MVPN traceroute message is une
xpected		
0x16	BAD_PMSI_ATTR_FLAG	Error matching the PMSI attribute
flag		
0x17	BAD_PMSI_ATTR_TYPE	Error matching the PMSI attribute
type		
0x18	BAD_PMSI_ATTR_LABEL	Error matching the PMSI attribute
label		
0x19	BAD_PMSI_ATTR_ID	Error matching the PMSI attribute
tunnel		identifier
0x1a	UNKNOWN_INTER_AS	Could not locate the Inter-AS prov
ider		tunnel segment.
0x1b	NO_UPSTREAM_PE	No valid upstream PE or route
0x1c	NO_CMCAST_STATE	No C-Mcast route for the requested
query		
0x1d	NO_WILD_CARD_SPMSI_AD_RCVD	No Wild Card SPMSI SPMSI AD is rec
eived from the upstream PE		
0x1e	NO_WILD_CARD_SPMSI_LEAD_AD_SENT	PE did not send LEAF-AD route for
the wild card SPMSI		

5. Mtracev2 Extensions

5.1. New Mtracev2 TLV Type

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

5.2. MVPN Extended Query Block



MVPN Extended Query Block

Type: Mtrace2 Extended Query Block Type

Length: Length of the MVPN Extended Query Block

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

T bit: This bit should be 0

Extended Query Type: New type defined

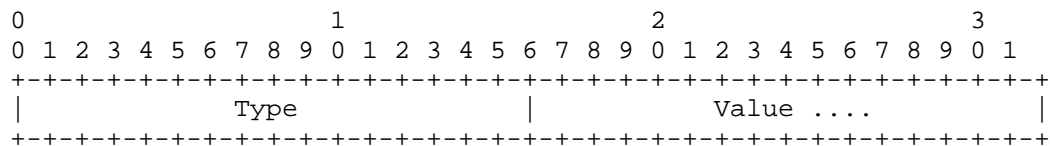
MBZ: Sent with all 0's, ignored on receipt

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

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

Upstream PE: IP Address of the Upstream PE

5.3. Leaf A-D Augmented Response Block



Leaf A-D Augmented Response Block

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

Type: New type defined

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

5.4. PMSI Tunnel Attributes Augmented Response Block

```

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

```

PMSI Tunnel Attributes Augmented Response Block

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

Type: New type defined

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

Value: The PMSI Tunnel Attribute as defined in RFC 6514

6. Mtrace2 Standard Response Block considerations

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

7. IANA Considerations

New TLV Type for MTRACE_MVPN_QUERY, MTRACE_MVPN_REQUEST,
MTRACE_MVPN_DOWNSTREAM_REQUEST, MTRACE_MVPN_RESPONSE

8. Security Considerations

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

9. Acknowledgments

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

10. Normative References

- [I-D.ietf-l3vpn-virtual-hub]
Jeng, H., Uttaro, J., Jalil, L., Decraene, B., Rekhter, Y., and R. Aggarwal, "Virtual Hub-and-Spoke in BGP/MPLS VPNs", draft-ietf-l3vpn-virtual-hub-08 (work in progress), July 2013.
- [I-D.ietf-mboned-mtrace-v2]
Asaeda, H. and W. Lee, "Mtrace Version 2: Traceroute Facility for IP Multicast", draft-ietf-mboned-mtrace-v2-10 (work in progress), July 2013.
- [I-D.ietf-mpls-seamless-mcast]
Rekhter, Y. and R. Aggarwal, "Inter-Area P2MP Segmented LSPs", draft-ietf-mpls-seamless-mcast-14 (work in progress), July 2014.
- [RFC6513] Rosen, E. and R. Aggarwal, "Multicast in MPLS/BGP IP VPNs", RFC 6513, February 2012.
- [RFC6514] Aggarwal, R., Rosen, E., Morin, T., and Y. Rekhter, "BGP Encodings and Procedures for Multicast in MPLS/BGP IP VPNs", RFC 6514, February 2012.
- [RFC6625] Rosen, E., Rekhter, Y., Hendrickx, W., and R. Qiu, "Wildcards in Multicast VPN Auto-Discovery Routes", RFC 6625, May 2012.

Authors' Addresses

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

Email: rkebler@juniper.net

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

Email: kurapati@juniper.net

Saud Asif
AT&T LABS
200 S Laurel Ave.
Middletown, NJ 07748
USA

Email: sasif@att.com

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

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

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

Status of this Memo

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

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

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

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

Copyright Notice

Copyright (c) 2014 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

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

Abstract

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

Table of Contents

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

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

1. Introduction

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

The objective of this Best Current Practices document is twofold:

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

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

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

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

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

2. Overview of Inter-domain Multicast Application Transport

A multicast-based application delivery scenario is as follows:

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

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

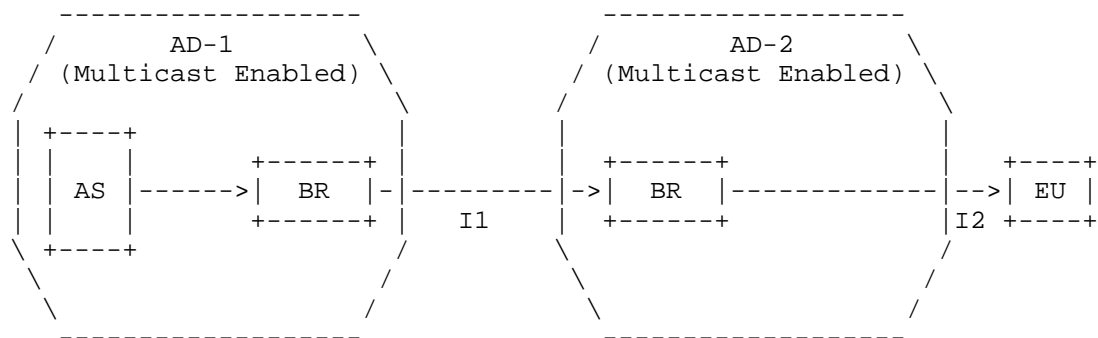
The Use Cases describing various architectural configurations for the multicast distribution along with associated requirements is described in section 3. Unique aspects related to the Enterprise network possibility will be described in this section. A comprehensive list of pertinent information that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

3. Inter-domain Peering Point Requirements for Multicast

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

3.1. Native Multicast

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



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

Figure 1 - Content Distribution via End to End Native Multicast

Advantages of this configuration are:

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

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

Architectural guidelines for this configuration are as follows:

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

3.2. Peering Point Enabled with GRE Tunnel

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

Advantages of this configuration:

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

Disadvantages of this configuration:

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

Architectural guidelines for this configuration include the following:

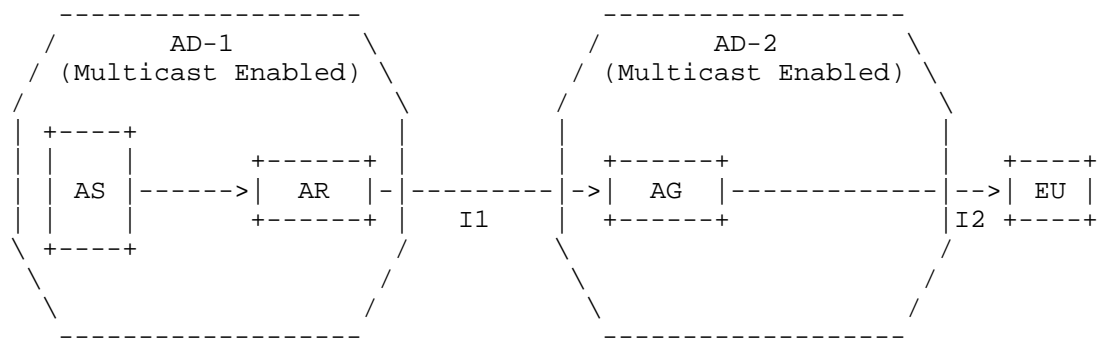
Guidelines (a) through (d) are the same as those described in Use Case 3.1.

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

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

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

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



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

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

Advantages of this configuration:

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

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

Disadvantages of this configuration:

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

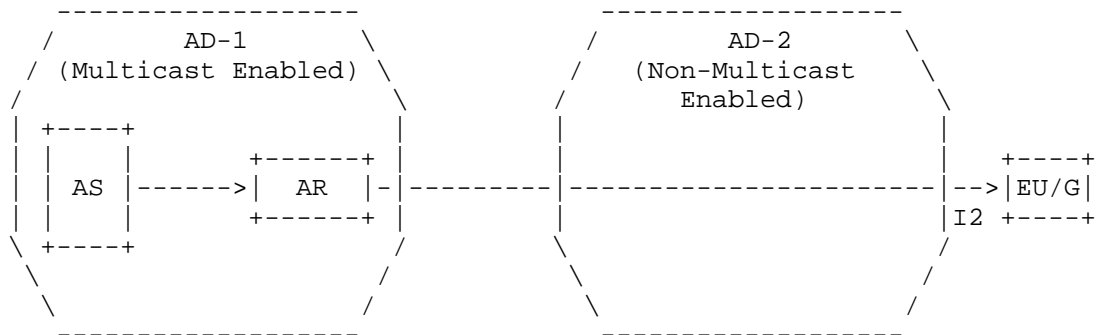
Architectural guidelines for this configuration are as follows:

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

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

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

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



AS = Application Multicast Source
 AR = AMT Relay
 EU/G = Gateway client embedded in EU device
 I2 = AMT Tunnel Connecting EU/G to AR in AD-1 through Non-Multicast Enabled AD-2.

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

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

Advantages of this configuration:

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

Disadvantages of this configuration:

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

Architectural guidelines for this configuration are as follows:

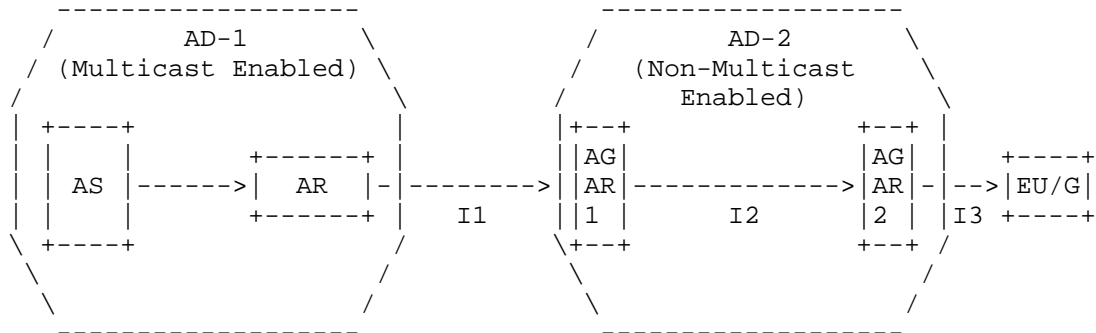
Guidelines (a) through (c) are the same as those described in Use Case 3.1.

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

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

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

This is a variation of Use Case 3.4 as follows:



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

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

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

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

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

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

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

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

Architectural guidelines for this configuration are as follows:

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

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

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

4. Supporting Functionality

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

4.1. Network Interconnection Transport and Security Guidelines

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

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

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

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

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

4.2. Routing Aspects and Related Guidelines

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

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

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

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

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

4.2.1 Native Multicast Routing Aspects

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

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

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

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

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

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

4.2.2 GRE Tunnel over Interconnecting Peering Point

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

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

4.2.3 Routing Aspects with AMT Tunnels

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

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

There are essentially two components to the AMT specification:

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

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

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

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

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

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

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

4.3. Back Office Functions - Billing and Logging Guidelines

Back Office refers to the following:

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

4.3.1 Provisioning Guidelines

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

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

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

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

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

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

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

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

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

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

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

4.3.2 Application Accounting Billing Guidelines

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

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

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

4.3.3 Log Management Guidelines

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

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

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

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

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

4.3.4 Settlement Guidelines

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

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

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

4.4. Operations - Service Performance and Monitoring Guidelines

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

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

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

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

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

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

4.5. Client Reliability Models/Service Assurance Guidelines

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

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

5. Security Considerations

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

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

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

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

6. IANA Considerations

7. Conclusions

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

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

8. References

8.1. Normative References

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

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

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

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

8.2. Informative References

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

9. Acknowledgments

Authors' Addresses

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

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

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

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

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

