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

Table of Contents

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 throughout, 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 montoring 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 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 neccessitate 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 unneccessary 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 perspectve) are commonly used to improve the performance by reducing the amount of state maintained at the router, suppressing superfluous IGMP messages and improving responsivenss 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 is 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 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 faciliate 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 concantenation 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 characterstic 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 releated 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.smpte.org/st-2110>.

Authors' Addresses

Mike McBride
Futurewei

Email: michael.mcbride@futurewei.com

Olufemi Komolafe
Arista Networks

Email: femi@arista.com

Mboned
Internet-Draft
Intended status: Best Current Practice
Expires: September 10, 2020

M. Abrahamsson
T. Chown
Jisc
L. Giuliano
Juniper Networks, Inc.
T. Eckert
Futurewei Technologies Inc.

March 9, 2020

Deprecating ASM for Interdomain Multicast
draft-ietf-mboned-deprecate-interdomain-asm-07

Abstract

This document recommends deprecation of the use of Any-Source Multicast (ASM) for interdomain multicast. It recommends the use of Source-Specific Multicast (SSM) for interdomain multicast applications and recommends that hosts and routers in these deployments fully support SSM. The recommendations in this document do not preclude the continued use of ASM within a single organisation or domain and are especially easy to adopt in existing intradomain ASM/PIM-SM deployments.

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 September 10, 2020.

Copyright Notice

Table of Contents

1. Introduction

IP Multicast has been deployed in various forms, within private networks, the wider Internet, and federated networks such as national or regional research networks. While a number of service models have been published, and in many cases revised over time, there has been no strong recommendation made by the IETF on the appropriateness of those models to certain scenarios, even though vendors and federations have often made such recommendations.

This document addresses this gap by making a BCP-level recommendation to deprecate the use of Any-Source Multicast (ASM) for interdomain multicast, leaving Source-Specific Multicast (SSM) as the recommended interdomain mode of multicast. This document further recommends that all hosts and routers which support interdomain multicast applications fully support SSM.

This document does not make any statement on the use of ASM within a single domain or organisation, and therefore does not preclude its use. Indeed, there are application contexts for which ASM is currently still widely considered well-suited within a single domain.

The main issue in most cases with moving to SSM is application support. Many applications are initially deployed for intradomain use and are later deployed interdomain. Therefore, this document recommends applications support SSM, even when they are initially intended for intradomain use. As explained below, SSM applications are readily compatible with existing intradomain ASM deployments as SSM is merely a subset of ASM.

2. Background

2.1. Multicast service models

Any-Source Multicast (ASM) and Source-Specific Multicast (SSM) are the two multicast service models in use today. In ASM, as originally described in [RFC1112], receivers express interest in joining a multicast group address and routers use multicast routing protocols to deliver traffic from the sender(s) to the receivers. If there are multiple senders for a given group, traffic from all senders will be delivered to the receivers. Since receivers specify only the group address, the network, and therefore the multicast routing protocols, are responsible for source discovery.

In SSM, by contrast, receivers specify both group and source when expressing interest in joining a multicast stream. Source discovery in SSM is handled by some out-of-band mechanism in the application

layer, which drastically simplifies the network and how the multicast routing protocols operate.

IANA has reserved specific ranges of IPv4 and IPv6 address space for multicast addressing. Guidelines for IPv4 multicast address assignments can be found in [RFC5771], while guidelines for IPv6 multicast address assignments can be found in [RFC2375] and [RFC3307]. The IPv6 multicast address format is described in [RFC4291].

2.2. ASM routing protocols

2.2.1. PIM Sparse Mode (PIM-SM)

The most commonly deployed ASM routing protocol is Protocol Independent Multicast - Sparse Mode (PIM-SM), as detailed in [RFC7761]. PIM-SM, as the name suggests, was designed to be used in scenarios where the subnets with receivers are sparsely distributed throughout the network. Because receivers do not indicate sender addresses in ASM (but only group addresses), PIM-SM uses the concept of a Rendezvous Point (RP) as a 'meeting point' for sources and receivers, and all routers in a PIM-SM domain are configured to use specific RP(s), either explicitly or through dynamic RP discovery protocols.

To enable PIM-SM to work between multiple domains, an interdomain, inter-RP signalling protocol known as Multicast Source Discovery Protocol (MSDP) [RFC3618] is used to allow an RP in one domain to learn the existence of a source in another domain. Deployment scenarios for MSDP are given in [RFC4611]. MSDP floods information about all active sources for all multicast streams to all RPs in all the domains - even if there is no receiver for a given application in a domain. As a result of this key scalability and security issue, along with other deployment challenges with the protocol, MSDP was never extended to support IPv6 and remains an Experimental protocol.

At the time of writing, there is no IETF Proposed Standard level interdomain solution for IPv4 ASM multicast because MSDP was the de facto mechanism for the interdomain source discovery problem, and it is Experimental. Other protocol options were investigated at the same time but were never implemented or deployed and are now historic (e.g: [RFC3913]).

2.2.2. Embedded-RP

Due to the availability of more bits in an IPv6 address than in IPv4, an IPv6-specific mechanism was designed in support of interdomain ASM with PIM-SM leveraging those bits. Embedded-RP [RFC3956] allows

routers supporting the protocol to determine the RP for the group without any prior configuration or discovery protocols, simply by observing the unicast RP address that is embedded (included) in the IPv6 multicast group address. Embedded-RP allows PIM-SM operation across any IPv6 network in which there is an end-to-end path of routers supporting this mechanism, including interdomain deployment.

2.2.3. Bidir-RP

Bidir-PIM [RFC5015] is another protocol to support ASM. There is no standardized option to operate Bidir-PIM interdomain. It is deployed intradomain for applications where many sources send traffic to the same IP multicast groups because unlike PIM-SM it does not create per-source state. Bidir-PIM is one of the important reasons for this document to not deprecate intradomain ASM.

2.3. SSM Routing protocols

SSM is detailed in [RFC4607]. It mandates the use of PIM-SSM for routing of SSM. PIM-SSM is merely a subset of PIM-SM ([RFC7761]).

PIM-SSM expects the sender's source address(es) to be known in advance by receivers through some out-of-band mechanism (typically in the application layer), and thus the receiver's designated router can send a PIM JOIN directly towards the source without needing to use an RP.

IPv4 addresses in the 232/8 (232.0.0.0 to 232.255.255.255) range are designated as source-specific multicast (SSM) destination addresses and are reserved for use by source-specific applications and protocols. See [RFC4607]. For IPv6, the address prefix ff3x::/32 is reserved for source-specific multicast use.

3. Discussion

3.1. Observations on ASM and SSM deployments

In enterprise and campus scenarios, ASM in the form of PIM-SM is likely the most commonly deployed multicast protocol. The configuration and management of an RP (including RP redundancy) within a single domain is a well understood operational practice. However, if interworking with external PIM domains is needed in IPv4 multicast deployments, interdomain MSDP is required to exchange information about sources between domain RPs. Deployment experience has shown MSDP to be a complex and fragile protocol to manage and troubleshoot. Some of these issues include complex Reverse Path Forwarding (RPF) rules, state attack protection, and filtering of undesired sources.

PIM-SM is a general purpose protocol that can handle all use cases. In particular, it was designed for cases such as videoconferencing where multiple sources may come and go during a multicast session. But for cases where a single, persistent source for a group is used, and receivers can be configured to know of that source, PIM-SM has unnecessary complexity. Therefore, SSM removes the need for many of the most complex components of PIM-SM.

As explained above, MSDP was not extended to support IPv6. Instead, the proposed interdomain ASM solution for PIM-SM with IPv6 is Embedded-RP, which allows the RP address for a multicast group to be embedded in the group address, making RP discovery automatic for all routers on the path between a receiver and a sender. Embedded-RP can support lightweight ad-hoc deployments. However, it relies on a single RP for an entire group that could only be made resilient within one domain. While this approach solves the MSDP issues, it does not solve the problem of unauthorised sources sending traffic to ASM multicast groups; this security issue is one of biggest problems of interdomain multicast.

As stated in RFC 4607, SSM is particularly well-suited to dissemination-style applications with one or more senders whose identities are known (by some out-of-band mechanism) before the application starts running or applications that utilize some signaling to indicate the source address of the multicast stream (e.g., electronic programming guide in IPTV applications). PIM-SSM is therefore very well-suited to applications such as classic linear broadcast TV over IP.

SSM requires applications, host operating systems and the designated routers connected to receiving hosts to support Internet Group Management Protocol, Version 3 (IGMPv3) [RFC3376] and Multicast Listener Discovery, Version 2 (MLDv2) [RFC3810]. While support for IGMPv3 and MLDv2 has been commonplace in routing platforms for a long time, it has also now become widespread in common operating systems for several years (Windows, MacOS, Linux/Android) and is no longer an impediment to SSM deployment.

3.2. Advantages of SSM for interdomain multicast

This section describes the three key benefits that SSM with PIM-SSM has over ASM. These benefits also apply to intradomain deployment but are even more important in interdomain deployments. See [RFC4607] for more details.

3.2.1. Reduced network operations complexity

A significant benefit of SSM is the reduced complexity that comes through eliminating the network-based source discovery required in ASM with PIM-SM. Specifically, SSM eliminates the need for RPs, shared trees, Shortest Path Tree (SPT) switchovers, PIM registers, MSDP, dynamic RP discovery mechanisms (BSR/AutoRP) and data-driven state creation. SSM simply utilizes a small subset of PIM-SM, alongside the integration with IGMPv3/MLDv2, where the source address signaled from the receiver is immediately used to create (S,G) state. Eliminating network-based source discovery for interdomain multicast means the vast majority of the complexity of multicast goes away.

This reduced complexity makes SSM radically simpler to manage, troubleshoot and operate, particularly for backbone network operators. This is the main operator motivation for the recommendation to deprecate the use of ASM in interdomain scenarios.

Note that this discussion does not apply to Bidir-PIM, and there is (as mentioned above) no standardized interdomain solution for Bidir-PIM. In Bidir-PIM, traffic is forwarded to the RP instead of building state as in PIM-SM. This occurs even in the absence of receivers. Bidir-PIM therefore trades state complexity with unnecessary traffic (potentially a large amount).

3.2.2. No network-wide IP multicast group-address management

In ASM, IP multicast group addresses need to be assigned to applications and instances thereof, so that two simultaneously active application instances will not share the same group address and receive IP multicast traffic from each other.

In SSM, no such IP multicast group management is necessary. Instead, the IP multicast group address simply needs to be assigned locally on a source like a unicast transport protocol port number: the only coordination required is to ensure that different applications running on the same host don’t send to the same group address. This does not require any network operator involvement.

3.2.3. Intrinsic source-control security

SSM is implicitly secure against off-path unauthorized/undesired sources. Receivers only receive packets from the sources they explicitly specify in their IGMPv3/MLDv2 membership messages, as opposed to ASM where any host can send traffic to a group address and have it transmitted to all receivers. With PIM-SSM, traffic from sources not requested by any receiver will be discarded by the first-

hop router (FHR) of that source, minimizing source attacks against shared network bandwidth and receivers.

This benefit is particularly important in interdomain deployments because there are no standardized solutions for ASM control of sources and the most common intradomain operational practices such as Access Control Lists (ACL) on the sender's FHR are not feasible for interdomain deployments.

This topic is expanded upon in [RFC4609].

4. Recommendations

This section provides recommendations for a variety of stakeholders in SSM deployment, including vendors, operators and application developers, and also suggests further work that could be undertaken within the IETF.

4.1. Deprecating use of ASM for interdomain multicast

This document recommends that the use of ASM be deprecated for interdomain multicast, and thus implicitly, that hosts and routers that support such interdomain applications fully support SSM and its associated protocols. Best current practices for deploying interdomain multicast using SSM are documented in [RFC8313].

The recommendation applies to the use of ASM between domains where either MSDP (IPv4) or Embedded-RP (IPv6) is used.

An interdomain use of ASM multicast in the context of this document is one where PIM-SM with RPs/MSDP/Embedded-RP is run on routers operated by two or more separate administrative entities.

The focus of this document is deprecation of inter-domain ASM multicast, and while encouraging the use of SSM within domains, it leaves operators free to choose to use ASM within their own domains. A more inclusive interpretation of this recommendation is that it also extends to deprecating use of ASM in the case where PIM is operated in a single operator domain, but where user hosts or non-PIM network edge devices are under different operator control. A typical example of this case is a service provider offering IPTV (single operator domain for PIM) to subscribers operating an IGMP proxy home gateway and IGMPv3/MLDv2 hosts (computer, tablets, set-top boxes).

4.2. Including network support for IGMPv3/MLDv2

This document recommends that all hosts, router platforms and security appliances used for deploying multicast support the components of IGMPv3 [RFC3376] and MLDv2 [RFC3810] necessary to support SSM (i.e., explicitly sending source-specific reports). The updated IPv6 Node Requirements RFC [RFC8504] states that MLDv2 must be supported in all implementations. Such support is already widespread in common host and router platforms.

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

Multicast snooping is often used to limit the flooding of multicast traffic in a layer 2 network. With snooping, a L2 switch will monitor IGMP/MLD messages and only forward multicast traffic out on host ports that have interested receivers connected. Such snooping capability should therefore support IGMPv3 and MLDv2. There is further discussion in [RFC4541].

4.3. Building application support for SSM

The recommendation to use SSM for interdomain multicast means that applications should properly trigger the sending of IGMPv3/MLDv2 source-specific report messages. It should be noted, however, there is a wide range of applications today that only support ASM. In many cases this is due to application developers being unaware of the operational concerns of networks, and the implications of using ASM versus using SSM. This document serves to provide clear direction for application developers who might currently only consider using ASM to instead support SSM, which only requires relatively minor changes for many applications, particularly those with single sources.

It is often thought that ASM is required for multicast applications where there are multiple sources. However, RFC 4607 also describes how SSM can be used instead of PIM-SM for multi-party applications:

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

Some useful considerations for multicast applications can be found in [RFC3170].

4.4. Developing application guidance: SSM, ASM, service discovery

Applications with many-to-many communication patterns can create more (S,G) state than is feasible for networks to manage, whether the source discovery is done by ASM with PIM-SM or at the application level and SSM/PIM-SM. These applications are not best supported by either SSM/PIM-SSM or ASM/PIM-SM.

Instead, these applications are better served by routing protocols that do not create (S,G), such as Bidir-PIM. Unfortunately, today many applications use ASM solely for service discovery. One example is where clients send IP multicast packets to elicit unicast replies from server(s). Deploying any form of IP multicast solely in support of such service discovery is in general not recommended. Dedicated service discovery via DNS-SD [RFC6763] should be used for this instead.

This document describes best practices to explain when to use SSM in applications, e.g, when ASM without (S,G) state in the network is better, or when dedicated service-discovery mechanisms should be used, but specifying these practices is outside the scope of this document. Further work on this subject may be expected within the IETF.

4.5. Preferring SSM applications intradomain

If feasible, it is recommended for applications to use SSM even if they are initially only meant to be used in intradomain environments supporting ASM. Because PIM-SSM is a subset of PIM-SM, existing intradomain PIM-SM networks are automatically compatible with SSM applications. Thus, SSM applications can operate alongside existing ASM applications. SSM's benefits of simplified address management and significantly reduced operational complexity apply equally to intradomain use.

However, for some applications it may be prohibitively difficult to add support for source discovery, so intradomain ASM may still be appropriate.

4.6. Documenting an ASM/SSM protocol mapping mechanism

In the case of existing ASM applications that cannot readily be ported to SSM, it may be possible to use some form of protocol mapping, i.e., to have a mechanism to translate a (*,G) join or leave to a (S,G) join or leave for a specific source S. The general challenge in performing such mapping is determining where the configured source address, S, comes from.

There are existing vendor-specific mechanisms deployed that achieve this function, but none are documented in IETF documents. This may be a useful area for the IETF to work on as an interim transition mechanism. However, these mechanisms would introduce additional administrative burdens, along with the need for some form of address management, neither of which are required in SSM. Hence, this should not be considered a long-term solution.

4.7. Not filtering ASM addressing between domains

A key benefit of SSM is that the receiver specifies the source-group tuple when signaling interest in a multicast stream. Hence, the group address need not be globally unique, so there is no need for multicast address allocation as long the reserved SSM range is used.

Despite the deprecation of interdomain ASM, it is recommended that operators should not filter ASM group ranges at domain boundaries, as some form of ASM-SSM mappings may continue to be used for some time.

4.8. Not precluding Intradomain ASM

The use of ASM within a single multicast domain such as a campus or enterprise is still relatively common today. There are even global enterprise networks that have successfully been using PIM-SM for many years. The operators of such networks most often use Anycast-RP [RFC4610] or MSDP (with IPv4) for RP resilience, at the expense of the extra operational complexity. These existing practices are unaffected by this document.

In the past decade, some Bidir-PIM deployments have scaled interdomain ASM deployments beyond the capabilities of PIM-SM. This too is unaffected by this document, instead it is encouraged where necessary due to application requirements (see Section 4.4).

This document also does not preclude continued use of ASM with multiple PIM-SM domains inside organisations, such as with IPv4 MSDP or IPv6 Embedded-RP. This includes organizations that are federations and have appropriate, non-standardized mechanisms to deal with the interdomain ASM issues explained in Section 3.2.

4.9. Evolving PIM deployments for SSM

Existing PIM-SM deployments can usually be used to run SSM applications with little to no changes. In some widely available router implementations of PIM-SM, PIM-SSM is simply enabled by default in the designated SSM address spaces whenever PIM-SM is enabled. In other implementations, simple configuration options exist to enable it. This allows migration of ASM applications to

SSM/PIM-SSM solely through application-side development to handle source-signaling via IGMPv3/MLDv2 and using SSM addresses. No network actions are required for this transition; unchanged ASM applications can continue to co-exist without issues.

When running PIM-SM, IGMPv3/MLDv2 (S,G) membership reports may also result in the desired PIM-SSM (S,G) operations and bypass any RP procedures. This is not standardized but depends on implementation and may require additional configuration in available products. In general, it is recommended to always use SSM address space for SSM applications. For example, the interaction of IGMPv3/MLDv2 (S,G) membership reports and Bidir-PIM is undefined and may not result in forwarding of any traffic.

Note that these migration recommendations do not include considerations on when or how to evolve those intradomain applications best served by ASM/Bidir-PIM from PIM-SM to Bidir-PIM. This may also be important but is outside the scope of this document.

5. Future interdomain ASM work

Future work may attempt to overcome current limitations of ASM solutions, such as interdomain deployment solutions for Bidir-PIM, or source access control mechanisms for IPv6 PIM-SM with embedded-RP. Such work could modify or amend the recommendations of this document (like any future IETF standards/BCP work).

Nevertheless, it is very unlikely that any ASM solution, even with such future work, can ever provide the same intrinsic security and network and address management simplicity as SSM (see Section 3.2). Accordingly, this document recommends that future work for general-purpose interdomain IP multicast focus on SSM items listed in Section 4.

6. Security Considerations

This document adds no new security considerations. It instead removes security issues incurred by interdomain ASM with PIM-SM/MSDP such as infrastructure control plane attacks and application and bandwidth/congestion attacks from unauthorised sources sending to ASM multicast groups. RFC 4609 describes the additional security benefits of using SSM instead of ASM.

7. IANA Considerations

This document makes no request of IANA.

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

8. Acknowledgments

The authors would like to thank members of the IETF mboned WG for discussions on the content of this document, with specific thanks to the following people for their contributions to the document: Hitoshi Asaeda, Dale Carder, Jake Holland, Albert Manfredi, Mike McBride, Per Nihlen, Greg Shepherd, James Stevens, Stig Venaas, Nils Warnke, and Sandy Zhang.

9. Changelog

[RFC-Editor: Please remove this section.]

02 - Toerless: Attempt to document the issues brought up on the list and discussion by James Stevens re. use of Bidir-PIM intradomain and IGMP/MLD interop issues.

- NOTE: Text was not vetted by co-authors, so rev'ed just as discussion basis.

- more subsection to highlight content. Added more detailled discussion about downsides of ASM wrt. address management and intrinsic source-control in SSM. Added recommendation to work on guidance when apps are best suited for SSM vs. ASM/Bidir vs. service discovery. Added recommendation how to evolve from PIM-SM to SSM in existing deployments. Added section on possible future interdomain ASM work (and why not to focus on it).

01 - Lenny: cleanup of text version, removed redundancies.

00 - initial IETF WG version. See draft-acg-mboned-deprecate-interdomain-asm for work leading to this document.

10. References

10.1. Normative References

[RFC1112] Deering, S., "Host extensions for IP multicasting", STD 5, RFC 1112, DOI 10.17487/RFC1112, August 1989, <https://www.rfc-editor.org/info/rfc1112>.

[RFC3307] Haberman, B., "Allocation Guidelines for IPv6 Multicast Addresses", RFC 3307, DOI 10.17487/RFC3307, August 2002, <https://www.rfc-editor.org/info/rfc3307>.

[RFC3376] Cain, B., Deering, S., Kouvelas, I., Fenner, B., and A. Thyagarajan, "Internet Group Management Protocol, Version 3", RFC 3376, DOI 10.17487/RFC3376, October 2002, <https://www.rfc-editor.org/info/rfc3376>.

[RFC3810] Vida, R., Ed. and L. Costa, Ed., "Multicast Listener Discovery Version 2 (MLDv2) for IPv6", RFC 3810, DOI 10.17487/RFC3810, June 2004, <https://www.rfc-editor.org/info/rfc3810>.

[RFC3956] Savola, P. and B. Haberman, "Embedding the Rendezvous Point (RP) Address in an IPv6 Multicast Address", RFC 3956, DOI 10.17487/RFC3956, November 2004, <https://www.rfc-editor.org/info/rfc3956>.

[RFC4291] Hinden, R. and S. Deering, "IP Version 6 Addressing Architecture", RFC 4291, DOI 10.17487/RFC4291, February 2006, <https://www.rfc-editor.org/info/rfc4291>.

[RFC4607] Holbrook, H. and B. Cain, "Source-Specific Multicast for IP", RFC 4607, DOI 10.17487/RFC4607, August 2006, <https://www.rfc-editor.org/info/rfc4607>.

[RFC5771] Cotton, M., Vegoda, L., and D. Meyer, "IANA Guidelines for IPv4 Multicast Address Assignments", BCP 51, RFC 5771, DOI 10.17487/RFC5771, March 2010, <https://www.rfc-editor.org/info/rfc5771>.

[RFC7761] Fenner, B., Handley, M., Holbrook, H., Kouvelas, I., Parekh, R., Zhang, Z., and L. Zheng, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", STD 83, RFC 7761, DOI 10.17487/RFC7761, March 2016, <https://www.rfc-editor.org/info/rfc7761>.

[RFC8313] Tarapore, P., Ed., Sayko, R., Shepherd, G., Eckert, T., Ed., and R. Krishnan, "Use of Multicast across Inter-domain Peering Points", BCP 213, RFC 8313, DOI 10.17487/RFC8313, January 2018, <https://www.rfc-editor.org/info/rfc8313>.

10.2. Informative References

[RFC2375] Hinden, R. and S. Deering, "IPv6 Multicast Address Assignments", RFC 2375, DOI 10.17487/RFC2375, July 1998, <https://www.rfc-editor.org/info/rfc2375>.

[RFC3170] Quinn, B. and K. Almeroth, "IP Multicast Applications: Challenges and Solutions", RFC 3170, DOI 10.17487/RFC3170, September 2001, <https://www.rfc-editor.org/info/rfc3170>.

[RFC3618] Fenner, B., Ed. and D. Meyer, Ed., "Multicast Source Discovery Protocol (MSDP)", RFC 3618, DOI 10.17487/RFC3618, October 2003, <https://www.rfc-editor.org/info/rfc3618>.

[RFC3913] Thaler, D., "Border Gateway Multicast Protocol (BGMP): Protocol Specification", RFC 3913, DOI 10.17487/RFC3913, September 2004, <https://www.rfc-editor.org/info/rfc3913>.

[RFC4541] Christensen, M., Kimball, K., and F. Solensky, "Considerations for Internet Group Management Protocol (IGMP) and Multicast Listener Discovery (MLD) Snooping Switches", RFC 4541, DOI 10.17487/RFC4541, May 2006, <https://www.rfc-editor.org/info/rfc4541>.

[RFC4604] Holbrook, H., Cain, B., and B. Haberman, "Using Internet Group Management Protocol Version 3 (IGMPv3) and Multicast Listener Discovery Protocol Version 2 (MLDv2) for Source-Specific Multicast", RFC 4604, DOI 10.17487/RFC4604, August 2006, <https://www.rfc-editor.org/info/rfc4604>.

[RFC4609] Savola, P., Lehtonen, R., and D. Meyer, "Protocol Independent Multicast - Sparse Mode (PIM-SM) Multicast Routing Security Issues and Enhancements", RFC 4609, DOI 10.17487/RFC4609, October 2006, <https://www.rfc-editor.org/info/rfc4609>.

[RFC4610] Farinacci, D. and Y. Cai, "Anycast-RP Using Protocol Independent Multicast (PIM)", RFC 4610, DOI 10.17487/RFC4610, August 2006, <https://www.rfc-editor.org/info/rfc4610>.

[RFC4611] McBride, M., Meylor, J., and D. Meyer, "Multicast Source Discovery Protocol (MSDP) Deployment Scenarios", BCP 121, RFC 4611, DOI 10.17487/RFC4611, August 2006, <https://www.rfc-editor.org/info/rfc4611>.

[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>.

[RFC6763] Cheshire, S. and M. Krochmal, "DNS-Based Service Discovery", RFC 6763, DOI 10.17487/RFC6763, February 2013, <https://www.rfc-editor.org/info/rfc6763>.

[RFC8504] Chown, T., Loughney, J., and T. Winters, "IPv6 Node Requirements", BCP 220, RFC 8504, DOI 10.17487/RFC8504, January 2019, <https://www.rfc-editor.org/info/rfc8504>.

Authors' Addresses

Mikael Abrahamsson

Stockholm
Sweden

Email: swmike@swm.pp.se

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

Email: tim.chown@jisc.ac.uk

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

Email: lenny@juniper.net

Toerless Eckert
Futurewei Technologies Inc.
2330 Central Expy
Santa Clara 95050
USA

Email: tte+ietf@cs.fau.de

Internet Area
Internet-Draft
Intended status: Informational
Expires: April 29, 2021

C. Perkins
Blue Meadow Networks
M. McBride
Futurewei
D. Stanley
HPE
W. Kumari
Google
JC. Zuniga
SIGFOX

October 26, 2020

Multicast Considerations over IEEE 802 Wireless Media
draft-ietf-mboned-ieee802-mcast-problems-12

Abstract

Well-known issues with multicast have prevented the deployment of multicast in 802.11 (wifi) and other local-area wireless environments. This document describes the problems of known limitations with wireless (primarily 802.11) Layer-2 multicast. Also described are certain multicast enhancement features that have been specified by the IETF, and by IEEE 802, for wireless media, as well as some operational choices that can be taken to improve the performance of the network. Finally, some recommendations are provided about the usage and combination of these features and operational choices.

Table of Contents

1. Introduction

Well-known issues with multicast have prevented the deployment of multicast in 802.11 [dot11] and other local-area wireless environments, as described in [mc-props], [mc-prob-stmt]. Performance issues have been observed when multicast packet transmissions of IETF protocols are used over IEEE 802 wireless media. Even though enhancements for multicast transmissions have been designed at both IETF and IEEE 802, incompatibilities still exist between specifications, implementations and configuration choices.

Many IETF protocols depend on multicast/broadcast for delivery of control messages to multiple receivers. Multicast allows sending data to multiple interested recipients without the source needing to send duplicate data to each recipient. With broadcast traffic, data is sent to every device regardless of their interest in the data. Multicast is used for various purposes such as neighbor discovery, network flooding, address resolution, as well minimizing media occupancy for the transmission of data that is intended for multiple receivers. In addition to protocol use of broadcast/multicast for control messages, more applications, such as push to talk in hospitals, or video in enterprises, universities, and homes, are sending multicast IP to end user devices, which are increasingly using Wi-Fi for their connectivity.

IETF protocols typically rely on network protocol layering in order to reduce or eliminate any dependence of higher level protocols on the specific nature of the MAC layer protocols or the physical media. In the case of multicast transmissions, higher level protocols have traditionally been designed as if transmitting a packet to an IP address had the same cost in interference and network media access, regardless of whether the destination IP address is a unicast address or a multicast or broadcast address. This model was reasonable for networks where the physical medium was wired, like Ethernet. Unfortunately, for many wireless media, the costs to access the

medium can be quite different. Multicast over Wi-Fi has often been plagued by such poor performance that it is disallowed. Some enhancements have been designed in IETF protocols that are assumed to work primarily over wireless media. However, these enhancements are usually implemented in limited deployments and not widespread on most wireless networks.

IEEE 802 wireless protocols have been designed with certain features to support multicast traffic. For instance, lower modulations are used to transmit multicast frames, so that these can be received by all stations in the cell, regardless of the distance or path attenuation from the base station or access point. However, these lower modulation transmissions occupy the medium longer; they hamper efficient transmission of traffic using higher order modulations to nearby stations. For these and other reasons, IEEE 802 working groups such as 802.11 have designed features to improve the performance of multicast transmissions at Layer 2 [ietf_802-11]. In addition to protocol design features, certain operational and configuration enhancements can ameliorate the network performance issues created by multicast traffic, as described in Section 5.

There seems to be general agreement that these problems will not be fixed anytime soon, primarily because it's expensive to do so and due to multicast being unreliable. Compared to unicast over Wi-Fi, multicast is often treated as somewhat of a second class citizen, even though there are many protocols using multicast. Something needs to be provided in order to make them more reliable. IPv6 neighbor discovery saturating the Wi-Fi link is only part of the problem. Wi-Fi traffic classes may help. This document is intended to help make the determination about what problems should be solved by the IETF and what problems should be solved by the IEEE (see Section 8).

This document details various problems caused by multicast transmission over wireless networks, including high packet error rates, no acknowledgements, and low data rate. It also explains some enhancements that have been designed at the IETF and IEEE 802.11 to ameliorate the effects of multicast traffic. Recommendations are also provided to implementors about how to use and combine these enhancements. Some advice about the operational choices that can be taken is also included. It is likely that this document will also be considered relevant to designers of future IEEE wireless specifications.

2. Terminology

This document uses the following definitions:

ACK
: The 802.11 layer 2 acknowledgement

AP
: IEEE 802.11 Access Point

basic rate
: The slowest rate of all the connected devices, at which multicast and broadcast traffic is generally transmitted

DTIM
: Delivery Traffic Indication Map (DTIM): An information element that advertises whether or not any associated stations have buffered multicast or broadcast frames

MCS
: Modulation and Coding Scheme

NOC
: Network Operations Center

PER
: Packet Error Rate

STA
: 802.11 station (e.g. handheld device)

TIM
: Traffic Indication Map (TIM): An information element that advertises whether or not any associated stations have buffered unicast frames

3. Identified multicast issues

3.1. Issues at Layer 2 and Below

In this section some of the issues related to the use of multicast transmissions over IEEE 802 wireless technologies are described.

3.1.1. Multicast reliability

Multicast traffic is typically much less reliable than unicast traffic. Since multicast makes point-to-multipoint communications, multiple acknowledgements would be needed to guarantee reception at all recipients. Since there are no ACKs for multicast packets, it is not possible for the Access Point (AP) to know whether or not a retransmission is needed. Even in the wired Internet, this characteristic often causes undesirably high error rates. This has contributed to the relatively slow uptake of multicast applications even though the protocols have long been available. The situation for wireless links is much worse, and is quite sensitive to the presence of background traffic. Consequently, there can be a high packet error rate (PER) due to lack of retransmission, and because the sender never backs off. It is not uncommon for there to be a packet loss rate of 5% or more, which is particularly troublesome for video and other environments where high data rates and high reliability are required.

3.1.2. Lower and Variable Data Rate

Multicast over wired differs from multicast over wireless because transmission over wired links often occurs at a fixed rate. Wi-Fi, on the other hand, has a transmission rate that varies depending upon the STA's proximity to the AP. The throughput of video flows, and the capacity of the broader Wi-Fi network, will change and will impact the ability for QoS solutions to effectively reserve bandwidth and provide admission control.

For wireless stations associated with an Access Point, the power necessary for good reception can vary from station to station. For unicast, the goal is to minimize power requirements while maximizing the data rate to the destination. For multicast, the goal is simply to maximize the number of receivers that will correctly receive the multicast packet; generally the Access Point has to use a much lower data rate at a power level high enough for even the farthest station to receive the packet, for example as briefly mentioned in section 2 of [RFC5757]. Consequently, the data rate of a video stream, for instance, would be constrained by the environmental considerations of the least reliable receiver associated with the Access Point.

Because more robust modulation and coding schemes (MCSs) have longer range but also lower data rate, multicast / broadcast traffic is generally transmitted at the slowest rate of all the connected devices. This is also known as the basic rate. The amount of additional interference depends on the specific wireless technology. In fact, backward compatibility and multi-stream implementations mean that the maximum unicast rates are currently up to a few Gbps, so

there can be more than 3 orders of magnitude difference in the transmission rate between multicast / broadcast versus optimal unicast forwarding. Some techiques employed to increase spectral efficiency, such as spatial multiplexing in MIMO systems, are not available with more than one intended receiver; it is not the case that backwards compatibility is the only factor responsible for lower multicast transmission rates.

Wired multicast also affects wireless LANs when the AP extends the wired segment; in that case, multicast / broadcast frames on the wired LAN side are copied to the Wireless Local Area Network (WLAN). Since broadcast messages are transmitted at the most robust MCS, many large frames are sent at a slow rate over the air.

3.1.3. Capacity and Impact on Interference

Transmissions at a lower rate require longer occupancy of the wireless medium and thus take away from the airtime of other communications and degrade the overall capacity. Furthermore, transmission at higher power, as is required to reach all multicast STAs associated to the AP, proportionately increases the area of interference.

3.1.4. Power-save Effects on Multicast

One of the characteristics of multicast transmission is that every station has to be configured to wake up to receive the multicast, even though the received packet may ultimately be discarded. This process can have a large effect on the power consumption by the multicast receiver station. For this reason there are workarounds, such as Directed Multicast Service (DMS) described in Section 4, to prevent unnecessarily waking up stations.

Multicast can work poorly with the power-save mechanisms defined in IEEE 802.11e, for the following reasons.

- Clients may be unable to stay in sleep mode due to multicast control packets frequently waking them up.
- Both unicast and multicast traffic can be delayed by power-saving mechanisms.
- A unicast packet is delayed until an STA wakes up and requests it. Unicast traffic may also be delayed to improve power save, efficiency and increase probability of aggregation.
- Multicast traffic is delayed in a wireless network if any of the STAs in that network are power savers. All STAs associated to the AP have to be awake at a known time to receive multicast traffic.
- Packets can also be discarded due to buffer limitations in the AP and non-AP STA.

3.2. Issues at Layer 3 and Above

This section identifies some representative IETF protocols, and describes possible negative effects due to performance degradation when using multicast transmissions for control messages. Common uses of multicast include:

- Control plane signaling
- Neighbor Discovery
- Address Resolution
- Service Discovery
- Applications (video delivery, stock data, etc.)
- On-demand routing
- Backbone construction
- Other L3 protocols (non-IP)

User Datagram Protocol (UDP) is the most common transport layer protocol for multicast applications. By itself, UDP is not reliable -- messages may be lost or delivered out of order.

3.2.1. IPv4 issues

The following list contains some representative discovery protocols, which utlize broadcast/multicast, that are used with IPv4.

- ARP [RFC5424]
- DHCP [RFC2131]
- mDNS [RFC6762]
- uPnP [RFC6970]

After initial configuration, ARP (described in more detail later) and DHCP occur much less commonly, but service discovery can occur at any time. Some widely-deployed service discovery protocols (e.g., for finding a printer) utilize mDNS (i.e., multicast) which is often the first service that operators drop. Even if multicast snooping [RFC4541] (which provides the benefit of conserving bandwidth on those segments of the network where no node has expressed interest in receiving packets addressed to the group address) is utilized, many devices can register at once and cause serious network degradation.

3.2.2. IPv6 issues

IPv6 makes extensive use of multicast, including the following:

- DHCPv6 [RFC8415]
- Protocol Independent Multicast (PIM) [RFC7761]
- IPv6 Neighbor Discovery Protocol (NDP) [RFC4861]
- multicast DNS (mDNS) [RFC6762]

- Router Discovery [RFC4286]

IPv6 NDP Neighbor Solicitation (NS) messages used in Duplicate Address Detection (DAD) and Address Lookup make use of Link-Scope multicast. In contrast to IPv4, an IPv6 node will typically use multiple addresses, and may change them often for privacy reasons. This intensifies the impact of multicast messages that are associated to the mobility of a node. Router advertisement (RA) messages are also periodically multicasted over the Link.

Neighbors may be considered lost if several consecutive Neighbor Discovery packets fail.

3.2.3. MLD issues

Multicast Listener Discovery (MLD) [RFC4541] is used to identify members of a multicast group that are connected to the ports of a switch. Forwarding multicast frames into a Wi-Fi-enabled area can use such switch support for hardware forwarding state information. However, since IPv6 makes heavy use of multicast, each STA with an IPv6 address will require state on the switch for several and possibly many multicast solicited-node addresses. Multicast addresses that do not have forwarding state installed (perhaps due to hardware memory limitations on the switch) cause frames to be flooded on all ports of the switch. Some switch vendors do not support MLD, for link-scope multicast, due to the increase it can cause in state.

3.2.4. Spurious Neighbor Discovery

On the Internet there is a "background radiation" of scanning traffic (people scanning for vulnerable machines) and backscatter (responses from spoofed traffic, etc). This means that routers very often receive packets destined for IPv4 addresses regardless of whether those IP addresses are in use. In the cases where the IP is assigned to a host, the router broadcasts an ARP request, gets back an ARP reply, and caches it; then traffic can be delivered to the host. When the IP address is not in use, the router broadcasts one (or more) ARP requests, and never gets a reply. This means that it does not populate the ARP cache, and the next time there is traffic for that IP address the router will rebroadcast the ARP requests.

The rate of these ARP requests is proportional to the size of the subnets, the rate of scanning and backscatter, and how long the router keeps state on non-responding ARPs. As it turns out, this rate is inversely proportional to how occupied the subnet is (valid ARPs end up in a cache, stopping the broadcasting; unused IPs never respond, and so cause more broadcasts). Depending on the address space in use, the time of day, how occupied the subnet is, and other

unknown factors, thousands of broadcasts per second have been observed. Around 2,000 broadcasts per second have been observed at the IETF NOC during face-to-face meetings.

With Neighbor Discovery for IPv6 [RFC2461], nodes accomplish address resolution by multicasting a Neighbor Solicitation that asks the target node to return its link-layer address. Neighbor Solicitation messages are multicast to the solicited-node multicast address of the target address. The target returns its link-layer address in a unicast Neighbor Advertisement message. A single request-response pair of packets is sufficient for both the initiator and the target to resolve each other's link-layer addresses; the initiator includes its link-layer address in the Neighbor Solicitation.

On a wired network, there is not a huge difference between unicast, multicast and broadcast traffic. Due to hardware filtering (see, e.g., [Deri-2010]), inadvertently flooded traffic (or excessive ethernet multicast) on wired networks can be quite a bit less costly, compared to wireless cases where sleeping devices have to wake up to process packets. Wired Ethernets tend to be switched networks, further reducing interference from multicast. There is effectively no collision / scheduling problem except at extremely high port utilizations.

This is not true in the wireless realm; wireless equipment is often unable to send high volumes of broadcast and multicast traffic, causing numerous broadcast and multicast packets to be dropped. Consequently, when a host connects it is often not able to complete DHCP, and IPv6 RAs get dropped, leading to users being unable to use the network.

4. Multicast protocol optimizations

This section lists some optimizations that have been specified in IEEE 802 and IETF that are aimed at reducing or eliminating the issues discussed in Section 3.

4.1. Proxy ARP in 802.11-2012

The AP knows the MAC address and IP address for all associated STAs. In this way, the AP acts as the central "manager" for all the 802.11 STAs in its basic service set (BSS). Proxy ARP is easy to implement at the AP, and offers the following advantages:

- Reduced broadcast traffic (transmitted at low MCS) on the wireless medium
- STA benefits from extended power save in sleep mode, as ARP requests for STA's IP address are handled instead by the AP.

- o ARP frames are kept off the wireless medium.
- o No changes are needed to STA implementation.

Here is the specification language as described in clause 10.23.13 of [dot11-proxyarp]:

> When the AP supports Proxy ARP "[...] the AP shall maintain a Hardware Address to Internet Address mapping for each associated station, and shall update the mapping when the Internet Address of the associated station changes. When the IPv4 address being resolved in the ARP request packet is used by a non-AP STA currently associated to the BSS, the proxy ARP service shall respond on behalf of the non-AP STA".

4.2. IPv6 Address Registration and Proxy Neighbor Discovery

As used in this section, a Low-Power Wireless Personal Area Network (6LoWPAN) denotes a low power lossy network (LLN) that supports 6LoWPAN Header Compression (HC) [RFC6282]. A 6TiSCH network [I-D.ietf-6tisch-architecture] is an example of a 6LowPAN. In order to control the use of IPv6 multicast over 6LoWPANs, the 6LoWPAN Neighbor Discovery (6LoWPAN ND) [RFC6775] standard defines an address registration mechanism that relies on a central registry to assess address uniqueness, as a substitute to the inefficient DAD mechanism found in the mainstream IPv6 Neighbor Discovery Protocol (NDP) [RFC4861][RFC4862].

The 6lo Working Group has specified an update [RFC8505] to RFC6775. Wireless devices can register their address to a Backbone Router [I-D.ietf-6lo-backbone-router], which proxies for the registered addresses with the IPv6 NDP running on a high speed aggregating backbone. The update also enables a proxy registration mechanism on behalf of the registered node, e.g. by a 6LoWPAN router to which the mobile node is attached.

The general idea behind the backbone router concept is that broadcast and multicast messaging should be tightly controlled in a variety of WLANs and Wireless Personal Area Networks (WPANs). Connectivity to a particular link that provides the subnet should be left to Layer-3. The model for the Backbone Router operation is represented in Figure 1.

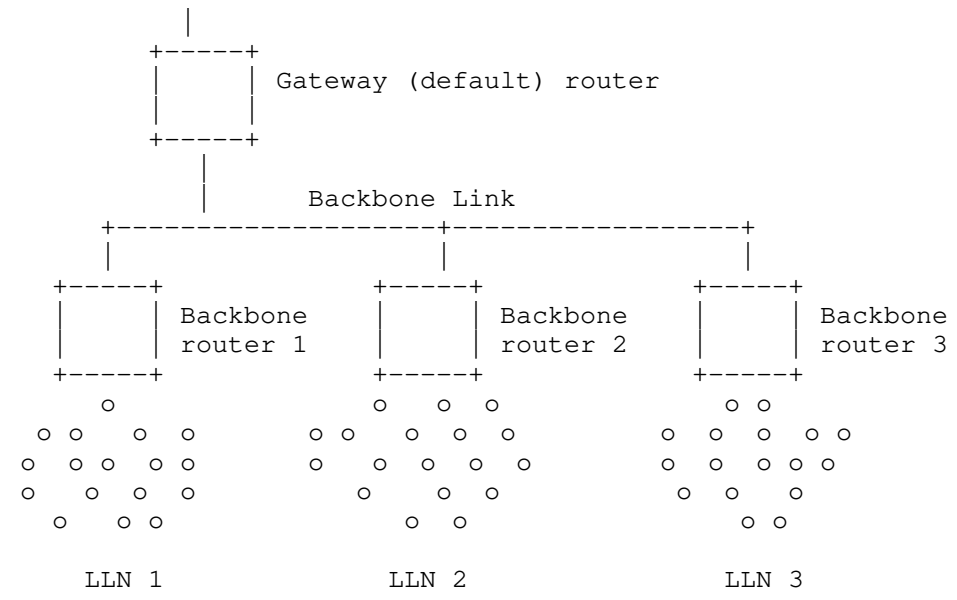


Figure 1: Backbone Link and Backbone Routers

LLN nodes can move freely from an LLN anchored at one IPv6 Backbone Router to an LLN anchored at another Backbone Router on the same backbone, keeping any of the IPv6 addresses they have configured. The Backbone Routers maintain a Binding Table of their Registered Nodes, which serves as a distributed database of all the LLN Nodes. An extension to the Neighbor Discovery Protocol is introduced to exchange Binding Table information across the Backbone Link as needed for the operation of IPv6 Neighbor Discovery.

RFC6775 and follow-on work [RFC8505] address the needs of LLNs, and similar techniques are likely to be valuable on any type of link where sleeping devices are attached, or where the use of broadcast and multicast operations should be limited.

4.3. Buffering to Improve Battery Life

Methods have been developed to help save battery life; for example, a device might not wake up when the AP receives a multicast packet. The AP acts on behalf of STAs in various ways. To enable use of the power-saving feature for STAs in its BSS, the AP buffers frames for delivery to the STA at the time when the STA is scheduled for reception. If an AP, for instance, expresses a DTIM (Delivery Traffic Indication Message) of 3 then the AP will send a multicast packet every 3 packets. In fact, when any single wireless STA associated with an access point has 802.11 power-save mode enabled,

the access point buffers all multicast frames and sends them only after the next DTIM beacon.

In practice, most AP's will send a multicast every 30 packets. For unicast the AP could send a TIM (Traffic Indication Message), but for multicast the AP sends a broadcast to everyone. DTIM does power management but STAs can choose whether or not to wake up and whether or not to drop the packet. Unfortunately, without proper administrative control, such STAs may be unable to determine why their multicast operations do not work.

4.4. Limiting multicast buffer hardware queue depth

The CAB (Content after Beacon) queue is used for beacon-triggered transmission of buffered multicast frames. If lots of multicast frames were buffered, and this queue fills up, it drowns out all regular traffic. To limit the damage that buffered traffic can do, some drivers limit the amount of queued multicast data to a fraction of the beacon_interval. An example of this is [CAB].

4.5. IPv6 support in 802.11-2012

IPv6 uses NDP instead of ARP. Every IPv6 node subscribes to a special multicast address for this purpose.

Here is the specification language from clause 10.23.13 of [dot11-proxyarp]:

> "When an IPv6 address is being resolved, the Proxy Neighbor Discovery service shall respond with a Neighbor Advertisement message [...] on behalf of an associated STA to an [ICMPv6] Neighbor Solicitation message [...]. When MAC address mappings change, the AP may send unsolicited Neighbor Advertisement Messages on behalf of a STA."

NDP may be used to request additional information

- Maximum Transmission Unit
- Router Solicitation
- Router Advertisement, etc.

NDP messages are sent as group addressed (broadcast) frames in 802.11. Using the proxy operation helps to keep NDP messages off the wireless medium.

4.6. Using Unicast Instead of Multicast

It is often possible to transmit multicast control and data messages by using unicast transmissions to each station individually.

4.6.1. Overview

In many situations, it's a good choice to use unicast instead of multicast over the Wi-Fi link. This avoids most of the problems specific to multicast over Wi-Fi, since the individual frames are then acknowledged and buffered for power save clients, in the way that unicast traffic normally operates.

This approach comes with the tradeoff of sometimes sending the same packet multiple times over the Wi-Fi link. However, in many cases, such as video into a residential home network, this can be a good tradeoff, since the Wi-Fi link may have enough capacity for the unicast traffic to be transmitted to each subscribed STA, even though multicast addressing may have been necessary for the upstream access network.

Several technologies exist that can be used to arrange unicast transport over the Wi-Fi link, outlined in the subsections below.

4.6.2. Layer 2 Conversion to Unicast

It is often possible to transmit multicast control and data messages by using unicast transmissions to each station individually.

Although there is not yet a standardized method of conversion, at least one widely available implementation exists in the Linux bridging code [bridge-mc-2-uc]. Other proprietary implementations are available from various vendors. In general, these implementations perform a straightforward mapping for groups or channels, discovered by IGMP or MLD snooping, to the corresponding unicast MAC addresses.

4.6.3. Directed Multicast Service (DMS)

There are situations where more is needed than simply converting multicast to unicast. For these purposes, DMS enables an STA to request that the AP transmit multicast group addressed frames destined to the requesting STAs as individually addressed frames [i.e., convert multicast to unicast]. Here are some characteristics of DMS:

- Requires 802.11n A-MSDUs

- o Individually addressed frames are acknowledged and are buffered for power save STAs
- o The requesting STA may specify traffic characteristics for DMS traffic
- o DMS was defined in IEEE Std 802.11v-2011
- o DMS requires changes to both AP and STA implementation.

DMS is not currently implemented in products. See [Tramarin2017] and [Oliva2013] for more information.

4.6.4. Automatic Multicast Tunneling (AMT)

AMT[RFC7450] provides a method to tunnel multicast IP packets inside unicast IP packets over network links that only support unicast. When an operating system or application running on an STA has an AMT gateway capability integrated, it's possible to use unicast to traverse the Wi-Fi link by deploying an AMT relay in the non-Wi-Fi portion of the network connected to the AP.

It is recommended that multicast-enabled networks deploying AMT relays for this purpose make the relays locally discoverable with the following methods, as described in [I-D.ietf-mboned-driad-amt-discovery]:

- o DNS-SD [RFC6763]
- o the well-known IP addresses from Section 7 of [RFC7450]

An AMT gateway that implements multiple standard discovery methods is more likely to discover the local multicast-capable network, instead of forming a connection to a non-local AMT relay further upstream.

4.7. GroupCast with Retries (GCR)

GCR (defined in [dot11aa]) provides greater reliability by using either unsolicited retries or a block acknowledgement mechanism. GCR increases probability of broadcast frame reception success, but still does not guarantee success.

For the block acknowledgement mechanism, the AP transmits each group addressed frame as conventional group addressed transmission. Retransmissions are group addressed, but hidden from non-11aa STAs. A directed block acknowledgement scheme is used to harvest reception status from receivers; retransmissions are based upon these responses.

GCR is suitable for all group sizes including medium to large groups. As the number of devices in the group increases, GCR can send block

acknowledgement requests to only a small subset of the group. GCR does require changes to both AP and STA implementations.

GCR may introduce unacceptable latency. After sending a group of data frames to the group, the AP has do the following:

- unicast a Block Ack Request (BAR) to a subset of members.
- wait for the corresponding Block Ack (BA).
- retransmit any missed frames.
- resume other operations that may have been delayed.

This latency may not be acceptable for some traffic.

There are ongoing extensions in 802.11 to improve GCR performance.

- BAR is sent using downlink MU-MIMO (note that downlink MU-MIMO is already specified in 802.11-REVmc 4.3).
- BA is sent using uplink MU-MIMO (which is a .11ax feature).
- Additional 802.11ax extensions are under consideration; see [mc-ack-mux]
- Latency may also be reduced by simultaneously receiving BA information from multiple STAs.

5. Operational optimizations

This section lists some operational optimizations that can be implemented when deploying wireless IEEE 802 networks to mitigate the issues discussed in Section 3.

5.1. Mitigating Problems from Spurious Neighbor Discovery

ARP Sponges

An ARP Sponge sits on a network and learns which IP addresses are actually in use. It also listen for ARP requests, and, if it sees an ARP for an IP address that it believes is not used, it will reply with its own MAC address. This means that the router now has an IP to MAC mapping, which it caches. If that IP is later assigned to an machine (e.g using DHCP), the ARP sponge will see this, and will stop replying for that address. Gratuitous ARPs (or the machine ARPing for its gateway) will replace the sponged address in the router ARP table. This technique is quite effective; but, unfortunately, the ARP sponge daemons were not really designed for this use (one of the most widely deployed arp sponges [arpsponge], was designed to deal with the disappearance of participants from an IXP) and so are not optimized for this purpose. One daemon is needed per subnet, the tuning is tricky (the scanning rate versus the

population rate versus retires, etc.) and sometimes buggy daemons have stopped, requiring a restart of the daemon and causing disruption.

Router mitigations

Some routers (often those based on Linux) implement a "negative ARP cache" daemon. Simply put, if the router does not see a reply to an ARP it can be configured to cache this information for some interval. Unfortunately, the core routers in use often do not support this. When a host connects to a network and gets an IP address, it will ARP for its default gateway (the router). The router will update its cache with the IP to host MAC mapping learned from the request (passive ARP learning).

Firewall unused space

The distribution of users on wireless networks / subnets may change in various use cases, such as conference venues (e.g SSIDs are renamed, some SSIDs lose favor, etc). This makes utilization for particular SSIDs difficult to predict ahead of time, but usage can be monitored as attendees use the different networks. Configuring multiple DHCP pools per subnet, and enabling them sequentially, can create a large subnet, from which only addresses in the lower portions are assigned. Therefore input IP access lists can be applied, which deny traffic to the upper, unused portions. Then the router does not attempt to forward packets to the unused portions of the subnets, and so does not ARP for it. This method has proven to be very effective, but is somewhat of a blunt axe, is fairly labor intensive, and requires coordination.

Disabling/filtering ARP requests

In general, the router does not need to ARP for hosts; when a host connects, the router can learn the IP to MAC mapping from the ARP request sent by that host. Consequently it should be possible to disable and / or filter ARP requests from the router. Unfortunately, ARP is a very low level / fundamental part of the IP stack, and is often offloaded from the normal control plane. While many routers can filter layer-2 traffic, this is usually implemented as an input filter and / or has limited ability to filter output broadcast traffic. This means that the simple "just disable ARP or filter it outbound" seems like a really simple (and obvious) solution, but implementations / architectural issues make this difficult or awkward in practice.

NAT

The broadcasts are overwhelmingly being caused by outside scanning / backscatter traffic. To NAT the entire (or a large portion) of the attendee networks would eliminate NAT translation entries for unused addresses, and so the router would never ARP for them. However, there are many reasons to avoid using NAT in such a blanket fashion.

Stateful firewalls

Another obvious solution would be to put a stateful firewall between the wireless network and the Internet. This firewall would block incoming traffic not associated with an outbound request. But this conflicts with the need and desire of some organizations to have the network as open as possible and to honor the end-to-end principle. An attendee on a meeting network should be an Internet host, and should be able to receive unsolicited requests. Unfortunately, keeping the network working and stable is the first priority and a stateful firewall may be required in order to achieve this.

5.2. Mitigating Spurious Service Discovery Messages

In networks that must support hundreds of STAs, operators have observed network degradation due to many devices simultaneously registering with mDNS. In a network with many clients, it is recommended to ensure that mDNS packets designed to discover services in smaller home networks be constrained to avoid disrupting other traffic.

6. Multicast Considerations for Other Wireless Media

Many of the causes of performance degradation described in earlier sections are also observable for wireless media other than 802.11.

For instance, problems with power save, excess media occupancy, and poor reliability will also affect 802.15.3 and 802.15.4. Unfortunately, 802.15 media specifications do not yet include mechanisms similar to those developed for 802.11. In fact, the design philosophy for 802.15 is oriented towards minimality, with the result that many such functions are relegated to operation within higher layer protocols. This leads to a patchwork of non-interoperable and vendor-specific solutions. See [uli] for some additional discussion, and a proposal for a task group to resolve similar issues, in which the multicast problems might be considered for mitigation.

Similar considerations hold for most other wireless media. A brief introduction is provided in [RFC5757] for the following:

- 802.16 WIMAX
- 3GPP/3GPP2
- DVB-H / DVB-IPDC
- TV Broadcast and Satellite Networks

7. Recommendations

This section provides some recommendations about the usage and combinations of the multicast enhancements described in Section 4 and Section 5.

Future protocol documents utilizing multicast signaling should be carefully scrutinized if the protocol is likely to be used over wireless media.

Proxy methods should be encouraged to conserve network bandwidth and power utilization by low-power devices. The device can use a unicast message to its proxy, and then the proxy can take care of any needed multicast operations.

Multicast signaling for wireless devices should be done in a way compatible with low duty-cycle operation.

8. On-going Discussion Items

This section suggests two discussion items for further resolution.

First, standards (and private) organizations should develop guidelines to help clarify when multicast packets should be sent wired rather than wireless. For example, 802.1ak [1] works on both ethernet and Wi-Fi and organizations could help decision making by developing guidelines for multicast over Wi-Fi including options for when traffic should be sent wired.

Second, reliable registration to Layer-2 multicast groups, and a reliable multicast operation at Layer-2, might provide a good multicast over wifi solution. There shouldn’t be a need to support 2^24 groups to get solicited node multicast working: it is possible to simply select a number of trailing bits that make sense for a given network size to limit the number of unwanted deliveries to reasonable levels. IEEE 802.1, 802.11, and 802.15 should be encouraged to revisit L2 multicast issues and provide workable solutions.

9. Security Considerations

This document does not introduce or modify any security mechanisms. Multicast is made more secure in a variety of ways. [RFC4601], for instance, mandates the use of IPsec to ensure authentication of the link-local messages in the Protocol Independent Multicast - Sparse Mode (PIM-SM) routing protocol. [RFC5796]specifies mechanisms to authenticate the PIM-SM link-local messages using the IP security (IPsec) Encapsulating Security Payload (ESP) or (optionally) the Authentication Header (AH).

As noted in [group_key], the unreliable nature of multicast transmission over wireless media can cause subtle problems with multicast group key management and updates. When WPA (TKIP) or WPA2 (AES-CCMP) encryption is in use, AP to client (From DS) multicasts have to be encrypted with a separate encryption key that is known to all of the clients (this is called the Group Key). Quoting further from that website, "... most clients are able to get connected and surf the web, check email, etc. even when From DS multicasts are broken. So a lot of people don’t realize they have multicast problems on their network..."

10. IANA Considerations

This document does not request any IANA actions.

11. Acknowledgements

This document has benefitted from discussions with the following people, in alphabetical order: Mikael Abrahamsson, Bill Atwood, Stuart Cheshire, Donald Eastlake, Toerless Eckert, Jake Holland, Joel Jaeggli, Jan Komissar, David Lamparter, Morten Pedersen, Pascal Thubert, Jeffrey (Zhaohui) Zhang

12. References

12.1. Informative References

[arpsponge]
Wessel, M. and N. Sijm, "Effects of IPv4 and IPv6 address resolution on AMS-IX and the ARP Sponge", July 2009, <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.182.4692>.

[bridge-mc-2-uc]
Fietkau, F., "bridge: multicast to unicast", Jan 2017, <https://github.com/torvalds/linux/commit/6db6f0eae6052b70885562e1733896647ec1d807>.

[CAB] Fietkau, F., "Limit multicast buffer hardware queue depth", 2013, <https://patchwork.kernel.org/patch/2687951/>.

[Deri-2010] Deri, L. and J. Gasparakis, "10 Gbit Hardware Packet Filtering Using Commodity Network Adapters", RIPE 61, 2010, <http://ripe61.ripe.net/presentations/138-Deri_RIPE_61.pdf>.

[dot11] "IEEE 802 Wireless", "802.11-2016 - IEEE Standard for Information technology--Telecommunications and information exchange between systems Local and metropolitan area networks--Specific requirements - Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specification (includes 802.11v amendment)", March 2016, <http://standards.ieee.org/findstds/standard/802.11-2016.html>.

[dot11-proxyarp] Hiertz, G., Mestanov, F., and B. Hart, "Proxy ARP in 802.11ax", September 2015, <https://mentor.ieee.org/802.11/dcn/15/11-15-1015-01-00ax-proxy-arp-in-802-11ax.pptx>.

[dot11aa] "IEEE 802 Wireless", "Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications Amendment 2: MAC Enhancements for Robust Audio Video Streaming", March 2012, <https://standards.ieee.org/standard/802_11aa-2012.html>.

[group_key] Spiff, "Why do some WiFi routers block multicast packets going from wired to wireless?", Jan 2017, <https://superuser.com/questions/730288/why-do-some-wifi-routers-block-multicast-packets-going-from-wired-to-wireless>.

[I-D.ietf-6lo-backbone-router] Thubert, P., Perkins, C., and E. Levy-Abegnoli, "IPv6 Backbone Router", draft-ietf-6lo-backbone-router-20 (work in progress), March 2020.

[I-D.ietf-6tisch-architecture] Thubert, P., "An Architecture for IPv6 over the TSCH mode of IEEE 802.15.4", draft-ietf-6tisch-architecture-29 (work in progress), August 2020.

[I-D.ietf-mboned-driad-amt-discovery]
Holland, J., "DNS Reverse IP AMT (Automatic Multicast Tunneling) Discovery", draft-ietf-mboned-driad-amt-discovery-13 (work in progress), December 2019.

[ietf_802-11]
Stanley, D., "IEEE 802.11 multicast capabilities", Nov 2015, <https://mentor.ieee.org/802.11/dcn/15/11-15-1261-03-0arc-multicast-performance-optimization-features-overview-for-ietf-nov-2015.ppt>.

[mc-ack-mux]
Tanaka, Y., Sakai, E., Morioka, Y., Mori, M., Hiertz, G., and S. Coffey, "Multiplexing of Acknowledgements for Multicast Transmission", July 2015, <https://mentor.ieee.org/802.11/dcn/15/11-15-0800-00-00ax-multiplexing-of-acknowledgements-for-multicast-transmission.pptx>.

[mc-prob-stmt]
Abrahamsson, M. and A. Stephens, "Multicast on 802.11", March 2015, <https://www.iab.org/wp-content/IAB-uploads/2013/01/multicast-problem-statement.pptx>.

[mc-props]
Stephens, A., "IEEE 802.11 multicast properties", March 2015, <https://mentor.ieee.org/802.11/dcn/15/11-15-1161-02-0arc-802-11-multicast-properties.ppt>.

[Oliva2013]
de la Oliva, A., Serrano, P., Salvador, P., and A. Banchs, "Performance evaluation of the IEEE 802.11aa multicast mechanisms for video streaming", 2013 IEEE 14th International Symposium on "A World of Wireless, Mobile and Multimedia Networks" (WoWMoM) pp. 1-9, June 2013.

[RFC2131] Droms, R., "Dynamic Host Configuration Protocol", RFC 2131, DOI 10.17487/RFC2131, March 1997, <https://www.rfc-editor.org/info/rfc2131>.

[RFC2461] Narten, T., Nordmark, E., and W. Simpson, "Neighbor Discovery for IP Version 6 (IPv6)", RFC 2461, DOI 10.17487/RFC2461, December 1998, <https://www.rfc-editor.org/info/rfc2461>.

[RFC4286] Haberman, B. and J. Martin, "Multicast Router Discovery", RFC 4286, DOI 10.17487/RFC4286, December 2005, <https://www.rfc-editor.org/info/rfc4286>.

[RFC4541] Christensen, M., Kimball, K., and F. Solensky, "Considerations for Internet Group Management Protocol (IGMP) and Multicast Listener Discovery (MLD) Snooping Switches", RFC 4541, DOI 10.17487/RFC4541, May 2006, <https://www.rfc-editor.org/info/rfc4541>.

[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>.

[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>.

[RFC4862] Thomson, S., Narten, T., and T. Jinmei, "IPv6 Stateless Address Autoconfiguration", RFC 4862, DOI 10.17487/RFC4862, September 2007, <https://www.rfc-editor.org/info/rfc4862>.

[RFC5424] Gerhards, R., "The Syslog Protocol", RFC 5424, DOI 10.17487/RFC5424, March 2009, <https://www.rfc-editor.org/info/rfc5424>.

[RFC5757] Schmidt, T., Waehlisch, M., and G. Fairhurst, "Multicast Mobility in Mobile IP Version 6 (MIPv6): Problem Statement and Brief Survey", RFC 5757, DOI 10.17487/RFC5757, February 2010, <https://www.rfc-editor.org/info/rfc5757>.

[RFC5796] Atwood, W., Islam, S., and M. Siami, "Authentication and Confidentiality in Protocol Independent Multicast Sparse Mode (PIM-SM) Link-Local Messages", RFC 5796, DOI 10.17487/RFC5796, March 2010, <https://www.rfc-editor.org/info/rfc5796>.

[RFC6282] Hui, J., Ed. and P. Thubert, "Compression Format for IPv6 Datagrams over IEEE 802.15.4-Based Networks", RFC 6282, DOI 10.17487/RFC6282, September 2011, <https://www.rfc-editor.org/info/rfc6282>.

[RFC6762] Cheshire, S. and M. Krochmal, "Multicast DNS", RFC 6762, DOI 10.17487/RFC6762, February 2013, <https://www.rfc-editor.org/info/rfc6762>.

[RFC6763] Cheshire, S. and M. Krochmal, "DNS-Based Service Discovery", RFC 6763, DOI 10.17487/RFC6763, February 2013, <https://www.rfc-editor.org/info/rfc6763>.

[RFC6775] Shelby, Z., Ed., Chakrabarti, S., Nordmark, E., and C. Bormann, "Neighbor Discovery Optimization for IPv6 over Low-Power Wireless Personal Area Networks (6LoWPANs)", RFC 6775, DOI 10.17487/RFC6775, November 2012, <https://www.rfc-editor.org/info/rfc6775>.

[RFC6970] Boucadair, M., Penno, R., and D. Wing, "Universal Plug and Play (UPnP) Internet Gateway Device - Port Control Protocol Interworking Function (IGD-PCP IWF)", RFC 6970, DOI 10.17487/RFC6970, July 2013, <https://www.rfc-editor.org/info/rfc6970>.

[RFC7450] Bumgardner, G., "Automatic Multicast Tunneling", RFC 7450, DOI 10.17487/RFC7450, February 2015, <https://www.rfc-editor.org/info/rfc7450>.

[RFC7761] Fenner, B., Handley, M., Holbrook, H., Kouvelas, I., Parekh, R., Zhang, Z., and L. Zheng, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", STD 83, RFC 7761, DOI 10.17487/RFC7761, March 2016, <https://www.rfc-editor.org/info/rfc7761>.

[RFC8415] Mrugalski, T., Siodelski, M., Volz, B., Yourtchenko, A., Richardson, M., Jiang, S., Lemon, T., and T. Winters, "Dynamic Host Configuration Protocol for IPv6 (DHCPv6)", RFC 8415, DOI 10.17487/RFC8415, November 2018, <https://www.rfc-editor.org/info/rfc8415>.

[RFC8505] Thubert, P., Ed., Nordmark, E., Chakrabarti, S., and C. Perkins, "Registration Extensions for IPv6 over Low-Power Wireless Personal Area Network (6LoWPAN) Neighbor Discovery", RFC 8505, DOI 10.17487/RFC8505, November 2018, <https://www.rfc-editor.org/info/rfc8505>.

[Tramarin2017] Tramarin, F., Vitturi, S., and M. Luvisotto, "IEEE 802.11n for Distributed Measurement Systems", 2017 IEEE International Instrumentation and Measurement Technology Conference (I2MTC) pp. 1-6, May 2017.

[uli] Kinney, P., "LLC Proposal for 802.15.4", Nov 2015, <https://mentor.ieee.org/802.15/dcn/15/15-15-0521-01-wng0-llc-proposal-for-802-15-4.pptx>.

12.2. URIs

[1] https://www.ieee802.org/1/pages/802.1ak.html

Appendix A. Changes in this draft between revisions 06 versus 07

This section lists the changes between revisions ...-06.txt and ...-07.txt of draft-ietf-mboned-ieee802-mcast-problems.

- Improved wording in section describing ARPsponge.
- Removed DRIAD as a discovery mechanism for multicast relays.
- Updated bibliographic citations, repaired broken URLs as needed.
- More editorial improvements and grammatical corrections.

Appendix B. Changes in this draft between revisions 05 versus 06

This section lists the changes between revisions ...-05.txt and ...-06.txt of draft-ietf-mboned-ieee802-mcast-problems.

- Included new text in Security Considerations to alert about problems regarding Group Key management caused by multicast unreliability and implementation bugs.
- Included DRIAD as a discovery mechanism for multicast relays.
- Corrected occurrences of "which" versus "that" and "amount" versus "number".
- Updated bibliographic citations, included URLs as needed.
- More editorial improvements and grammatical corrections.

Appendix C. Changes in this draft between revisions 04 versus 05

This section lists the changes between revisions ...-04.txt and ...-05.txt of draft-ietf-mboned-ieee802-mcast-problems.

- Incorporated text from Jake Holland for a new section about conversion of multicast to unicast and included AMT as an existing solution.
- Included some text about likely future multicast applications that will emphasize the need for attention to the technical matters collected in this document.
- Further modified text to be more generic instead of referring specifically to IETF conference situations.
- Modified text to be more generic instead of referring specifically to Bonjour.
- Added uPnP as a representative multicast protocol in IP networks.

- Referred to Linux bridging code for multicast to unicast.
- Updated bibliographic citations, included URLs as needed.
- More editorial improvements and grammatical corrections.

Appendix D. Changes in this draft between revisions 03 versus 04

This section lists the changes between revisions ...-03.txt and ...-04.txt of draft-ietf-mboned-ieee802-mcast-problems.

- Replaced "client" by "STA".
- Used terminology "Wi-Fi" throughout.
- Many editorial improvements and grammatical corrections.
- Modified text to be more generic instead of referring specifically to IETF conference situations.
- Cited [RFC5757] for introduction to other wireless media.
- Updated bibliographic citations.

Authors' Addresses

Charles E. Perkins
Blue Meadow Networks

Phone: +1-408-330-4586
Email: charliep@computer.org

Mike McBride
Futurewei Technologies Inc.
2330 Central Expressway
Santa Clara, CA 95055
USA

Email: michael.mcbride@futurewei.com

Dorothy Stanley
Hewlett Packard Enterprise
2000 North Naperville Rd.
Naperville, IL 60566
USA

Phone: +1 630 979 1572
Email: dstanley1389@gmail.com

Warren Kumari
Google
1600 Amphitheatre Parkway
Mountain View, CA 94043
USA

Email: warren@kumari.net

Juan Carlos Zuniga
SIGFOX
425 rue Jean Rostand
Labege 31670
France

Email: j.c.zuniga@ieee.org

Mboned
Internet-Draft
Intended status: Standards Track
Expires: July 17, 2019

J. Holland
Akamai Technologies, Inc.

January 13, 2019

DNS Reverse IP AMT Discovery
draft-jholland-mboned-driad-amt-discovery-03

Abstract

This document defines a new DNS resource record (RR) used to advertise addresses for Automatic Multicast Tunneling (AMT) relays capable of receiving multicast traffic from the owner of the RR. The new AMTRELAY RR makes possible a source-specific method for AMT gateways to discover appropriate AMT relays, in order to ingest traffic for source-specific multicast channels into multicast-capable receiving networks when no multicast connectivity is directly available between the sending and receiving networks.

Table of Contents

1. Introduction

This document defines DNS Reverse IP AMT Discovery (DRIAD), a mechanism for AMT gateways to discover AMT relays which are capable of forwarding multicast traffic from a known source IP address.

AMT (Automatic Multicast Tunneling) is defined in [RFC7450], and provides a method to transport multicast traffic over a unicast tunnel, in order to traverse non-multicast-capable network segments.

Section 4.1.5 of [RFC7450] explains that relay selection might need to depend on the source of the multicast traffic, since a relay must be able to receive multicast traffic from the desired source in order to forward it.

That section suggests DNS-based queries as a possible solution. DRIAD is a DNS-based solution, as suggested there. This solution also addresses the relay discovery issues in the "Disadvantages" lists in Section 3.3 of [RFC8313] and Section 3.4 of [RFC8313].

The goal for DRIAD is to enable multicast connectivity between separate multicast-enabled networks when neither the sending nor the receiving network is connected to a multicast-enabled backbone, without pre-configuring any peering arrangement between the networks.

1.1. Background

The reader is assumed to be familiar with the basic DNS concepts described in [RFC1034], [RFC1035], and the subsequent documents that update them, particularly [RFC2181].

The reader is also assumed to be familiar with the concepts and terminology regarding source-specific multicast as described in [RFC4607] and the use of IGMPv3 [RFC3376] and MLDv2 [RFC3810] for group management of source-specific multicast channels, as described in [RFC4604].

The reader should also be familiar with AMT, particularly the terminology listed in Section 3.2 of [RFC7450] and Section 3.3 of [RFC7450].

1.2. Terminology

1.2.1. Relays and Gateways

When reading this document, it’s especially helpful to recall that once an AMT tunnel is established, the relay receives native multicast traffic and sends unicast tunnel-encapsulated traffic to

the gateway, and the gateway receives the tunnel-encapsulated packets, decapsulates them, and forwards them as native multicast packets, as illustrated in Figure 1.

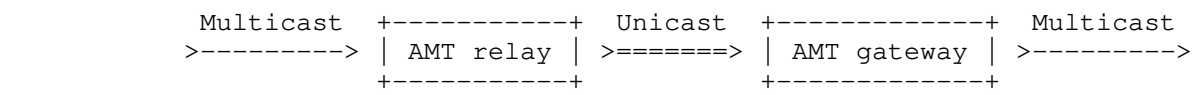


Figure 1: AMT Tunnel Illustration

1.2.2. Definitions

Term	Definition
(S,G)	A source-specific multicast channel, as described in [RFC4607]. A pair of IP addresses with a source host IP and destination group IP.
downstream	Further from the source of traffic.
FQDN	Fully Qualified Domain Name, as described in [RFC8499]
gateway	An AMT gateway, as described in [RFC7450]
relay	An AMT relay, as described in [RFC7450]
RPF	Reverse Path Forwarding, as described in [RFC5110]
RR	A DNS Resource Record, as described in [RFC1034]
RRType	A DNS Resource Record Type, as described in [RFC1034]
SSM	Source-specific multicast, as described in [RFC4607]
upstream	Closer to the source of traffic.

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119] and [RFC8174] when, and only when, they appear in all capitals, as shown here.

2. Relay Discovery Operation

2.1. Overview

The AMTRELAY resource record (RR) defined in this document is used to publish the IP address or domain name of an AMT relay that can receive, encapsulate, and forward multicast traffic from a particular sender.

The sender is the owner of the RR, and configures the RR so that it contains the address or domain name of an AMT relay that can receive multicast IP traffic from that sender.

This enables AMT gateways in remote networks to discover an AMT relay that is capable of forwarding traffic from the sender. This in turn enables those AMT gateways to receive the multicast traffic tunneled over a unicast AMT tunnel from those relays, and then to pass the multicast packets into networks or applications that are using the gateway to subscribe to traffic from that sender.

This mechanism only works for source-specific multicast (SSM) channels. The source address of the (S,G) is reversed and used as an index into one of the reverse mapping trees (in-addr.arpa for IPv4, as described in Section 3.5 of [RFC1035], or ip6.arpa for IPv6, as described in Section 2.5 of [RFC3596]).

Some detailed example use cases are provided in Section 3, and other applicable example topologies appear in Section 3.3 of [RFC8313], Section 3.4 of [RFC8313], and Section 3.5 of [RFC8313].

2.2. Signaling and Discovery

This section describes a typical example of the end-to-end process for signaling a receiver's join of a SSM channel that relies on an AMTRELAY RR.

The example in Figure 2 contains 2 multicast-enabled networks that are both connected to the internet with non-multicast-capable links, and which have no direct association with each other.

A content provider operates a sender, which is a source of multicast traffic inside a multicast-capable network.

An end user who is a customer of the content provider has a multicast-capable internet service provider, which operates a receiving network that uses an AMT gateway. The AMT gateway is DRIAD-capable.

The content provider provides the user with a receiving application that tries to subscribe to at least one (S,G). This receiving application could for example be a file transfer system using FLUTE [RFC6726] or a live video stream using RTP [RFC3550], or any other application that might subscribe to a SSM channel.

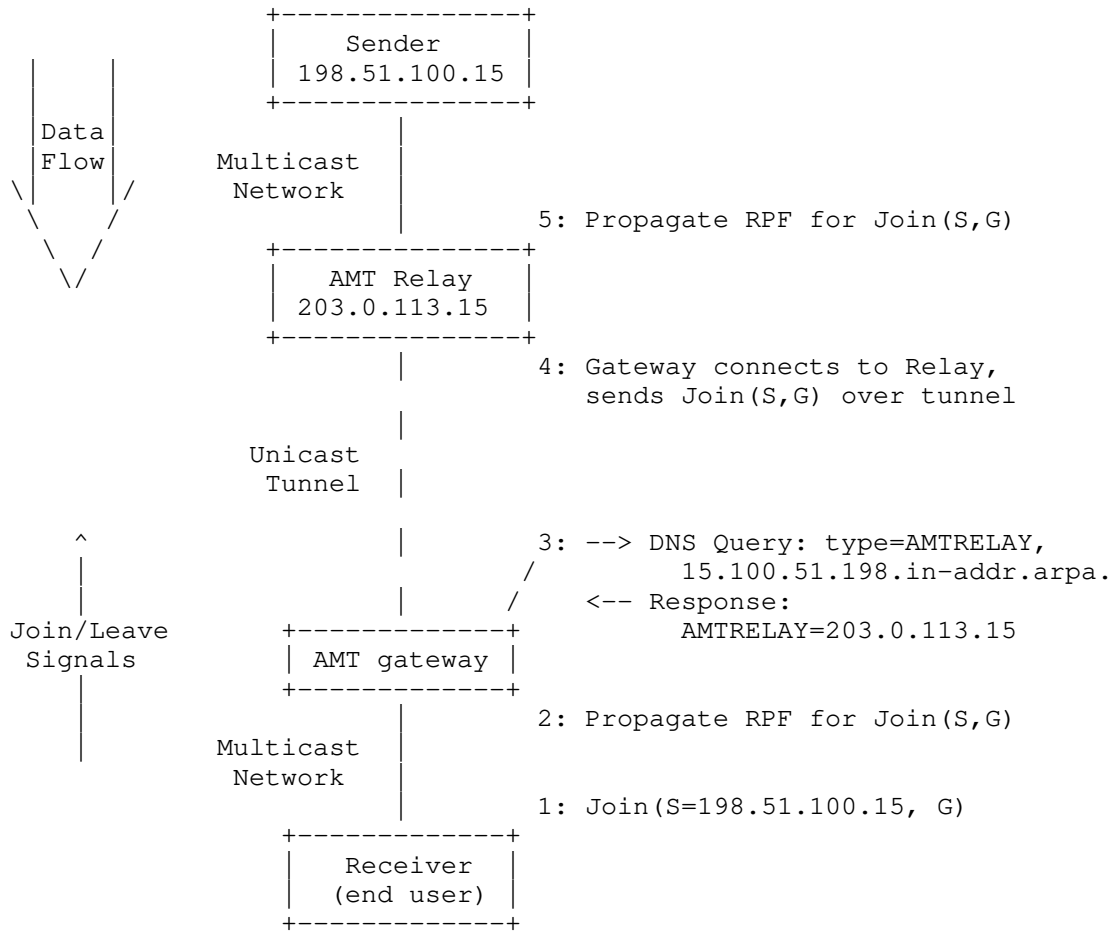


Figure 2: DRIAD Messaging

In this simple example, the sender IP is 198.51.100.15, and the relay IP is 203.0.113.15.

The content provider has previously configured the DNS zone that contains the domain name "15.100.51.198.in-addr.arpa.", which is the reverse lookup domain name for his sender. The zone file contains an

AMTRELAY RR with the Relay's IP address. (See Section 4.3 for details about the AMTRELAY RR format and semantics.)

The sequence of events depicted in Figure 2 is as follows:

1. The end user starts the app, which issues a join to the (S,G): (198.51.100.15, 232.252.0.2).

2. The join propagates with RPF through the multicast-enabled network with PIM [RFC7761] or another multicast routing mechanism, until the AMT gateway receives a signal to join the (S,G).

3. The AMT gateway performs a reverse DNS lookup for the AMTRELAY RRType, by sending an AMTRELAY RRType query for the FQDN "15.100.51.198.in-addr.arpa.", using the reverse IP domain name for the sender's source IP address (the S from the (S,G)), as described in Section 3.5 of [RFC1035].

   The DNS resolver for the AMT gateway uses ordinary DNS recursive resolution until it has the authoritative result that the content provider configured, which informs the AMT gateway that the relay address is 203.0.113.15.

4. The AMT gateway performs AMT handshakes with the AMT relay as described in Section 4 of [RFC7450], then forwards a Membership report to the relay indicating subscription to the (S,G).

5. The relay propagates the join through its network toward the sender, then forwards the appropriate AMT-encapsulated traffic to the gateway, which decapsulates and forwards it as native multicast through its downstream network to the end user.

2.3. Optimal Relay Selection

The reverse source IP DNS query of an AMTRELAY RR is a good way for a gateway to discover a relay that is known to the sender.

However, it is NOT necessarily a good way to discover the best relay for that gateway to use, because the RR IP will only provide information about relays known to the source.

If there is an upstream relay in a network that is more local to the gateway and able to receive and forward multicast traffic from the sender, that relay is better for the gateway to use, since more of the network path uses native multicast, allowing more chances for packet replication. But since that relay is not known to the sender,

it won't be advertised in the sender's reverse IP DNS record. An example network with this scenario is outlined in Section 3.1.2.

It's only appropriate for an AMT gateway to discover an AMT relay by querying an AMTRELAY RR owned by a sender when all of these conditions are met:

1. The gateway needs to propagate a join of an (S,G) over AMT, because in the gateway's network, no RPF next hop toward the source can propagate a native multicast join of the (S,G); and

2. The gateway is not already connected to a relay that forwards multicast traffic from the source of the (S,G); and

3. The gateway is not configured to use a particular IP address for AMT discovery, or a relay discovered with that IP is not able to forward traffic from the source of the (S,G); and

4. The gateway is not able to find an upstream AMT relay with DNS-SD [RFC6763], using "_amt._udp" as the Service section of the queries, or a relay discovered this way is not able to forward traffic from the source of the (S,G)

When the above conditions are met, the gateway has no path within its local network that can receive multicast traffic from the source IP of the (S,G).

In this situation, the best way to find a relay that can forward the required traffic is to use information that comes from the operator of the sender. When the sender has configured the AMTRELAY RR defined in this document, gateways can use the DRIAD mechanism defined in this document to discover the relay information provided by the sender.

2.4. DNS Configuration

Often an AMT gateway will only have access to the source and group IP addresses of the desired traffic, and will not know any other name for the source of the traffic. Because of this, typically the best way of looking up AMTRELAY RRs will be by using the source IP address as an index into one of the reverse mapping trees (in-addr.arpa for IPv4, as described in Section 3.5 of [RFC1035], or ip6.arpa for IPv6, as described in Section 2.5 of [RFC3596]).

Therefore, it is RECOMMENDED that AMTRELAY RRs be added to reverse IP zones as appropriate. AMTRELAY records MAY also appear in other zones, but the primary intended use case requires a reverse IP

mapping for the source from an (S,G) in order to be useful to most AMT gateways.

<TBD>

Please can a DNS expert review the following paragraph and perhaps help construct an equivalent and more clear explanation?

I borrowed the language from https://tools.ietf.org/html/rfc4025#section-1.2, but I'm not actually sure what "the fashion usual for PTR records" means, precisely...

PTR gives a domain name, and then we do what, an A/AAAA record lookup, and then a AMTRELAY lookup on the final name that has a valid A/AAAA after any CNAME/DNAME chain? - jake 2019-01-13

</TBD>

When the reverse IP mapping has no AMTRELAY RR but does have a PTR record, the lookup is done in the fashion usual for PTR records. The IP address' octets (IPv4) or nibbles (IPv6) are reversed and looked up with the appropriate suffix. Any CNAMEs or DNAMEs found MUST be followed, and finally the AMTRELAY RR is queried with the resulting domain name.

See Section 4 and Section 4.3 for a detailed explanation of the contents for a DNS Zone file.

3. Example Deployments

3.1. Example Receiving Networks

3.1.1. Tier 3 ISP

One example of a receiving network is an ISP that offers multicast ingest services to its subscribers, illustrated in Figure 3.

In the example network below, subscribers can join (S,G)s with MLDv2 or IGMPv3 as described in [RFC4604], and the AMT gateway in this ISP can receive and forward multicast traffic from one of the example sending networks in Section 3.2 by discovering the appropriate AMT relays with a DNS lookup for the AMTRELAY RR with the reverse IP of the source in the (S,G).

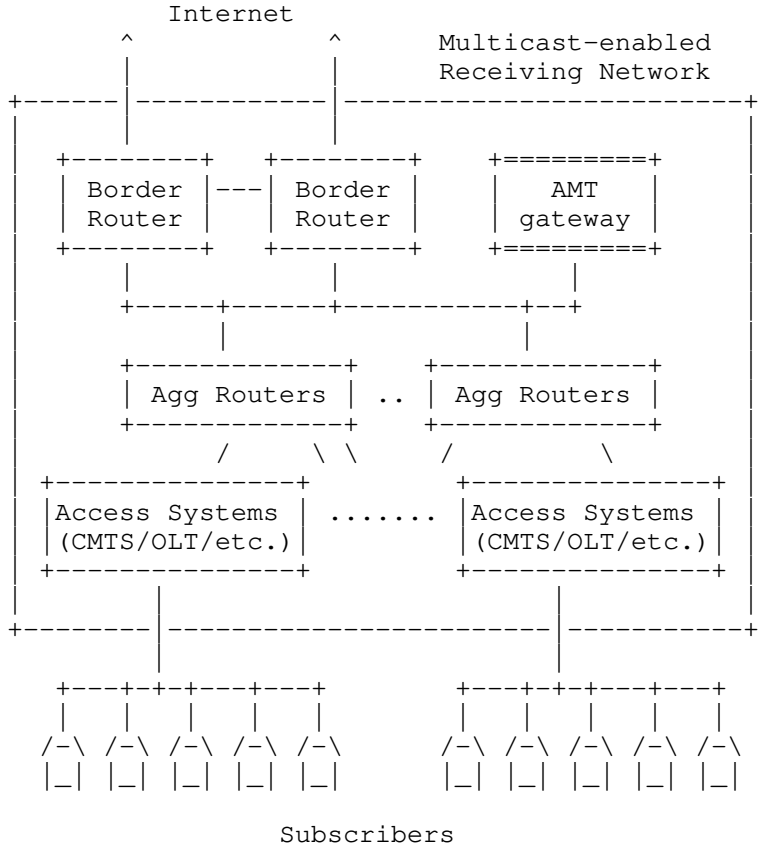


Figure 3: Receiving ISP Example

3.1.2. Small Office

Another example receiving network is a small branch office that regularly accesses some multicast content, illustrated in Figure 4.

This office has desktop devices that need to receive some multicast traffic, so an AMT gateway runs on a LAN with these devices, to pull traffic in through a non-multicast next-hop.

The office also hosts some mobile devices that have AMT gateway instances embedded inside apps, in order to receive multicast traffic over their non-multicast wireless LAN. (Note that the "Legacy Router" is a simplification that's meant to describe a variety of possible conditions- for example it could be a device providing a split-tunnel VPN as described in [RFC7359], deliberately excluding

multicast traffic for a VPN tunnel, rather than a device which is incapable of multicast forwarding.)

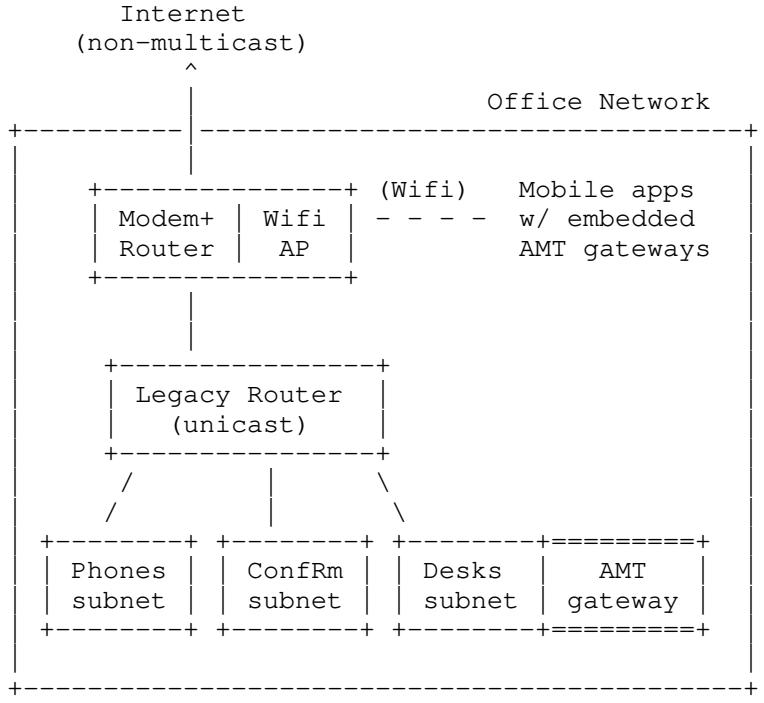


Figure 4: Small Office (no multicast up)

By adding an AMT relay to this office network as in Figure 5, it's possible to make use of multicast services from the example multicast-capable ISP in Section 3.1.1.

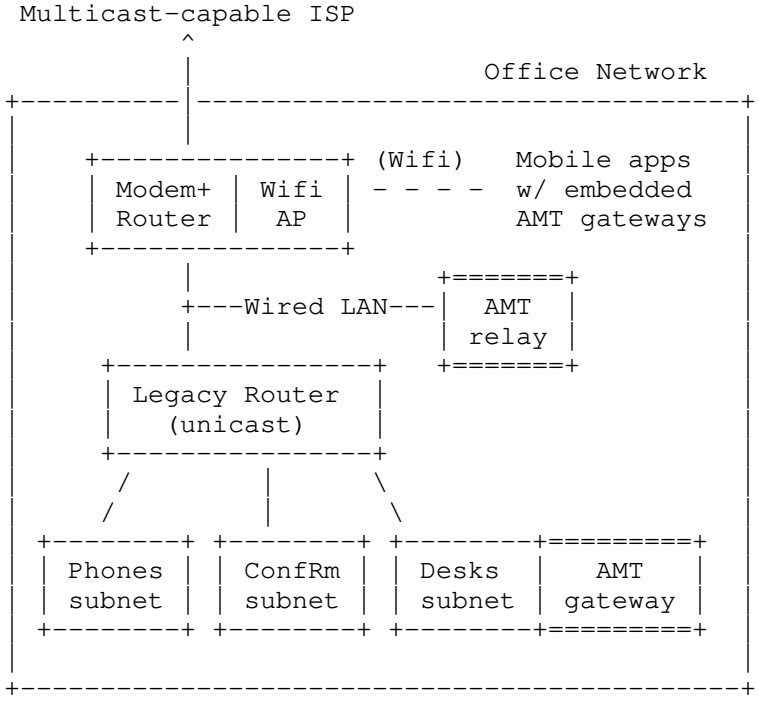


Figure 5: Small Office Example

When multicast-capable networks are chained like this, with a network like the one in Figure 5 receiving internet services from a multicast-capable network like the one in Figure 3, it's important for AMT gateways to reach the more local AMT relay, in order to avoid accidentally tunneling multicast traffic from a more distant AMT relay with unicast, and failing to utilize the multicast transport capabilities of the network in Figure 3.

For this reason, it's RECOMMENDED that AMT gateways by default perform service discovery using DNS Service Discovery (DNS-SD) [RFC6763] for _amt._udp.<domain> (with <domain> chosen as described in Section 11 of [RFC6763]) and use the AMT relays discovered that way in preference to AMT relays discoverable via the mechanism defined in this document (DRIAD).

It's also RECOMMENDED that when the well-known anycast IP addresses defined in Section 7 of [RFC7450] are suitable for discovering an AMT relay that can forward traffic from the source, that a DNS record with the AMTRELAY RRType be published for those IP addresses along with any other appropriate AMTRELAY RRs to indicate the best relative precedences for receiving the source traffic.

Accordingly, AMT gateways SHOULD by default discover the most-preferred relay first by DNS-SD, then by DRIAD as described in this document (in precedence order, as described in Section 4.2.1), then with the anycast addresses defined in Section 7 of [RFC7450] (namely: 192.52.193.1 and 2001:3::1) if those IPs weren't listed in the AMTRELAY RRs. This default behavior MAY be overridden by administrative configuration where other behavior is more appropriate for the gateway within its network.

The discovery and connection process for multiple relays MAY operate in parallel, but when forwarding multicast group membership reports with new joins from an AMT gateway, membership reports SHOULD be forwarded to the most-preferred relays first, falling back to less preferred relays only after failing to receive traffic for an appropriate timeout, and only after reporting a leave to any more-preferred connected relays that have failed to subscribe to the traffic.

It is RECOMMENDED that the default timeout be no less than 3 seconds, but the value MAY be overridden by administrative configuration, where known groups or channels need a different timeout for successful application performance.

3.2. Example Sending Networks

3.2.1. Sender-controlled Relays

When a sender network is also operating AMT relays to distribute multicast traffic, as in Figure 6, each address could appear as an AMTRELAY RR for the reverse IP of the sender, or one or more domain names could appear in AMTRELAY RRs, and the AMT relay addresses can be discovered by finding an A or AAAA record from those domain names.

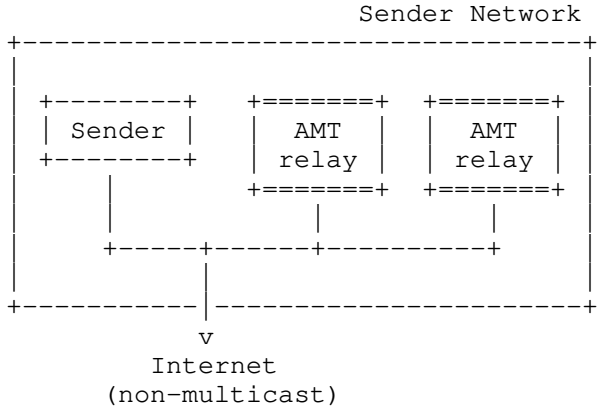


Figure 6: Small Office Example

3.2.2. Provider-controlled Relays

When an ISP offers a service to transmit outbound multicast traffic through a forwarding network, it might also offer AMT relays in order to reach receivers without multicast connectivity to the forwarding network, as in Figure 7. In this case it's RECOMMENDED that the ISP also provide a domain name for the AMT relays for use with the discovery process defined in this document.

When the sender wishes to use the relays provided by the ISP for forwarding multicast traffic, an AMTRELAY RR should be configured to use the domain name provided by the ISP, to allow for address reassignment of the relays without forcing the sender to reconfigure the corresponding AMTRELAY RRs.

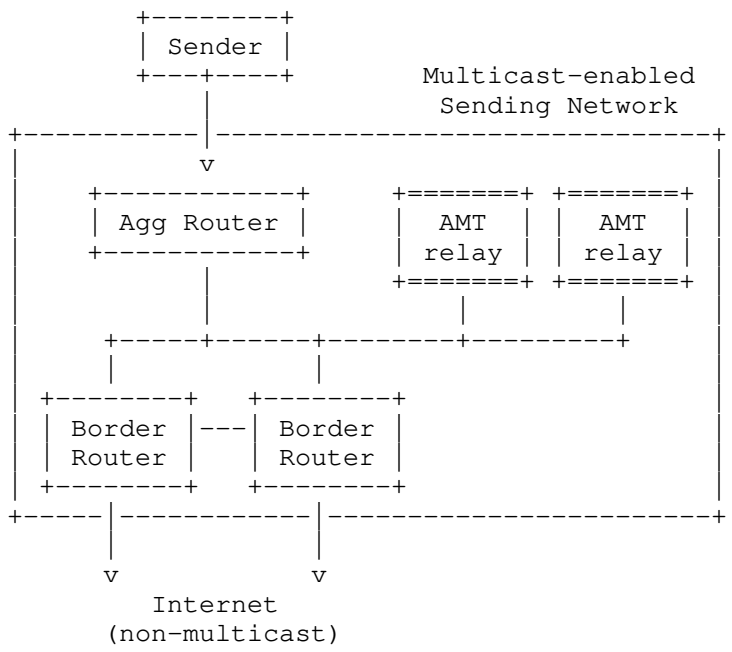


Figure 7: Sending ISP Example

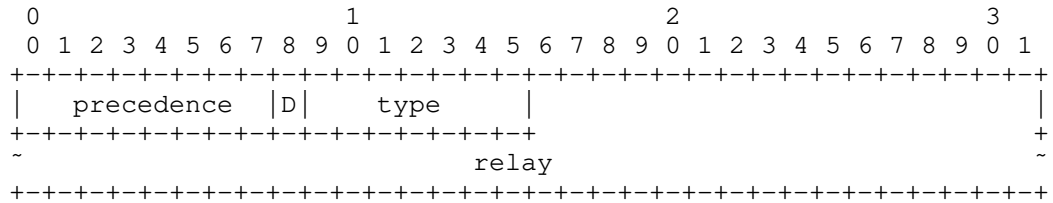
4. AMTRELAY Resource Record Definition

4.1. AMTRELAY RRType

The AMTRELAY RRType has the mnemonic AMTRELAY and type code TBD1 (decimal).

4.2. AMTRELAY RData Format

The AMTRELAY RData consists of a 8-bit precedence field, a 1-bit "Discovery Optional" field, a 7-bit type field, and a variable length relay field.



4.2.1. RData Format - Precedence

This is an 8-bit precedence for this record. It is interpreted in the same way as the PREFERENCE field described in Section 3.3.9 of [RFC1035].

Relays listed in AMTRELAY records with a lower value for precedence are to be attempted first.

Where there is a tie in precedence, the default choice of relay MUST be non-deterministic, to support load balancing. The AMT gateway operator MAY override this default choice with explicit configuration when it's necessary for administrative purposes.

For example, one network might prefer to tunnel IPv6 multicast traffic over IPv6 AMT and IPv4 multicast traffic over IPv4 AMT to avoid routeability problems in IPv6 from affecting IPv4 traffic and vice versa, while another network might prefer to tunnel both kinds of traffic over IPv6 to reduce the IPv4 space used by its AMT gateways. In this example scenario or other cases where there is an administrative preference that requires explicit configuration, a receiving network MAY make systematically different precedence choices among records with the same precedence value.

4.2.2. RData Format - Discovery Optional (D-bit)

The D bit is a "Discovery Optional" flag.

If the D bit is set to 0, a gateway using this RR MUST perform AMT relay discovery as described in Section 4.2.1.1 of [RFC7450], rather than directly sending an AMT request message to the relay.

That is, the gateway MUST receive an AMT relay advertisement message (Section 5.1.2 of [RFC7450]) for an address before sending an AMT request message (Section 5.1.3 of [RFC7450]) to that address. Before receiving the relay advertisement message, this record has only indicated that the address can be used for AMT relay discovery, not for a request message. This is necessary for devices that are not fully functional AMT relays, but rather load balancers or brokers, as mentioned in Section 4.2.1.1 of [RFC7450].

If the D bit is set to 1, the gateway MAY send an AMT request message directly to the discovered relay address without first sending an AMT discovery message.

This bit should be set according to advice from the AMT relay operator. The D bit MUST be set to zero when no information is available from the AMT relay operator about its suitability.

4.2.3. RData Format - Type

The type field indicates the format of the information that is stored in the relay field.

The following values are defined:

- type = 0: The relay field is empty (0 bytes).
- type = 1: The relay field contains a 4-octet IPv4 address.
- type = 2: The relay field contains a 16-octet IPv6 address.
- type = 3: The relay field contains a wire-encoded domain name. The wire-encoded format is self-describing, so the length is implicit. The domain name MUST NOT be compressed. (See Section 3.3 of [RFC1035] and Section 4 of [RFC3597].)

4.2.4. RData Format - Relay

The relay field is the address or domain name of the AMT relay. It is formatted according to the type field.

When the type field is 0, the length of the relay field is 0, and it indicates that no AMT relay should be used for multicast traffic from this source.

When the type field is 1, the length of the relay field is 4 octets, and a 32-bit IPv4 address is present. This is an IPv4 address as described in Section 3.4.1 of [RFC1035]. This is a 32-bit number in network byte order.

When the type field is 2, the length of the relay field is 16 octets, and a 128-bit IPv6 address is present. This is an IPv6 address as described in Section 2.2 of [RFC3596]. This is a 128-bit number in network byte order.

When the type field is 3, the relay field is a normal wire-encoded domain name, as described in Section 3.3 of [RFC1035]. Compression MUST NOT be used, for the reasons given in Section 4 of [RFC3597].

4.3. AMTRELAY Record Presentation Format

4.3.1. Representation of AMTRELAY RRs

AMTRELAY RRs may appear in a zone data master file. The precedence, D-bit, relay type, and relay fields are REQUIRED.

If the relay type field is 0, the relay field MUST be ".".

The presentation for the record is as follows:

```
IN AMTRELAY precedence D-bit type relay
```

4.3.2. Examples

In a DNS resolver that understands the AMTRELAY type, the zone might contain a set of entries like this:

```
$ORIGIN 100.51.198.in-addr.arpa.
10     IN AMTRELAY  10 0 1 203.0.113.15
10     IN AMTRELAY  10 0 2 2001:DB8::15
10     IN AMTRELAY 128 1 3 amtrelays.example.com.
```

This configuration advertises an IPv4 discovery address, an IPv6 discovery address, and a domain name for AMT relays which can receive traffic from the source 198.51.100.10. The IPv4 and IPv6 addresses are configured with a D-bit of 0 (meaning discovery is mandatory, as described in Section 4.2.2), and a precedence 10 (meaning they're preferred ahead of the last entry, which has precedence 128).

For zone files in resolvers that don't support the AMTRELAY RRType natively, it's possible to use the format for unknown RR types, as described in [RFC3597]. This approach would replace the AMTRELAY entries in the example above with the entries below:

[To be removed (TBD): replace 65280 with the IANA-assigned value TBD1, here and in Appendix B.]

```
10   IN TYPE65280  \# (
       6  ; length
       0a ; precedence=10
       01 ; D=0, relay type=1, an IPv4 address
       cb00710f ) ; 203.0.113.15
10   IN TYPE65280  \# (
       18 ; length
       0a ; precedence=10
       02 ; D=0, relay type=2, an IPv6 address
       20010db800000000000000000000000f ) ; 2001:db8::15
10   IN TYPE65280  \# (
       24 ; length
       80 ; precedence=128
       83 ; D=1, relay type=3, a wire-encoded domain name
       616d7472656c6179732e6578616d706c652e636f6d2e ) ; domain name
```

See Appendix B for more details.

5. IANA Considerations

This document updates the IANA Registry for DNS Resource Record Types by assigning type TBD1 to the AMTRELAY record.

This document creates a new registry named "AMTRELAY Resource Record Parameters", with a sub-registry for the "Relay Type Field". The initial values in the sub-registry are:

Value	Description
0	No relay is present.
1	A 4-byte IPv4 address is present
2	A 16-byte IPv6 address is present
3	A wire-encoded domain name is present
4-255	Unassigned

Values 0, 1, 2, and 3 are further explained in Section 4.2.3 and Section 4.2.4. Relay type numbers 4 through 255 can be assigned with a policy of Specification Required (as described in [RFC8126]).

6. Security Considerations

[TBD: these 3 are just the first few most obvious issues, with just sketches of the problem. Explain better, and look for trickier issues.]

6.1. Record-spoofing

If AMT is used to ingest multicast traffic, providing a false AMTRELAY record to a gateway using it for discovery can result in Denial of Service, or artificial multicast traffic from a source under an attacker’s control.

Therefore, it is important to ensure that the AMTRELAY record is authentic, with DNSSEC [RFC4033] or other operational safeguards that can provide assurance of the authenticity of the record contents.

6.2. Local Override

The local relays, while important for overall network performance, can’t be secured by DNSSEC.

6.3. Congestion

Multicast traffic, particularly interdomain multicast traffic, carries some congestion risks, as described in Section 4 of [RFC8085]. Network operators are advised to take precautions including monitoring of application traffic behavior, traffic authentication, and rate-limiting of multicast traffic, in order to ensure network health.

7. Acknowledgements

This specification was inspired by the previous work of Doug Nortz, Robert Sayko, David Segelstein, and Percy Tarapore, presented in the MBONED working group at IETF 93.

Thanks also to Jeff Goldsmith and Lenny Giuliano for helpful reviews and feedback.

8. References

8.1. Normative References

[RFC1034] Mockapetris, P., "Domain names - concepts and facilities", STD 13, RFC 1034, DOI 10.17487/RFC1034, November 1987, <https://www.rfc-editor.org/info/rfc1034>.

[RFC1035] Mockapetris, P., "Domain names - implementation and specification", STD 13, RFC 1035, DOI 10.17487/RFC1035, November 1987, <https://www.rfc-editor.org/info/rfc1035>.

[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>.

[RFC2181] Elz, R. and R. Bush, "Clarifications to the DNS Specification", RFC 2181, DOI 10.17487/RFC2181, July 1997, <https://www.rfc-editor.org/info/rfc2181>.

[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>.

[RFC3596] Thomson, S., Huitema, C., Ksinant, V., and M. Souissi, "DNS Extensions to Support IP Version 6", STD 88, RFC 3596, DOI 10.17487/RFC3596, October 2003, <https://www.rfc-editor.org/info/rfc3596>.

[RFC3597] Gustafsson, A., "Handling of Unknown DNS Resource Record (RR) Types", RFC 3597, DOI 10.17487/RFC3597, September 2003, <https://www.rfc-editor.org/info/rfc3597>.

[RFC3810] Vida, R., Ed. and L. Costa, Ed., "Multicast Listener Discovery Version 2 (MLDv2) for IPv6", RFC 3810, DOI 10.17487/RFC3810, June 2004, <https://www.rfc-editor.org/info/rfc3810>.

[RFC4604] Holbrook, H., Cain, B., and B. Haberman, "Using Internet Group Management Protocol Version 3 (IGMPv3) and Multicast Listener Discovery Protocol Version 2 (MLDv2) for Source-Specific Multicast", RFC 4604, DOI 10.17487/RFC4604, August 2006, <https://www.rfc-editor.org/info/rfc4604>.

[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>.

[RFC6763] Cheshire, S. and M. Krochmal, "DNS-Based Service Discovery", RFC 6763, DOI 10.17487/RFC6763, February 2013, <https://www.rfc-editor.org/info/rfc6763>.

[RFC7450] Bumgardner, G., "Automatic Multicast Tunneling", RFC 7450, DOI 10.17487/RFC7450, February 2015, <https://www.rfc-editor.org/info/rfc7450>.

[RFC8085] Eggert, L., Fairhurst, G., and G. Shepherd, "UDP Usage Guidelines", BCP 145, RFC 8085, DOI 10.17487/RFC8085, March 2017, <https://www.rfc-editor.org/info/rfc8085>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

8.2. Informative References

[RFC3550] Schulzrinne, H., Casner, S., Frederick, R., and V. Jacobson, "RTP: A Transport Protocol for Real-Time Applications", STD 64, RFC 3550, DOI 10.17487/RFC3550, July 2003, <https://www.rfc-editor.org/info/rfc3550>.

[RFC4025] Richardson, M., "A Method for Storing IPsec Keying Material in DNS", RFC 4025, DOI 10.17487/RFC4025, March 2005, <https://www.rfc-editor.org/info/rfc4025>.

[RFC4033] Arends, R., Austein, R., Larson, M., Massey, D., and S. Rose, "DNS Security Introduction and Requirements", RFC 4033, DOI 10.17487/RFC4033, March 2005, <https://www.rfc-editor.org/info/rfc4033>.

[RFC5110] Savola, P., "Overview of the Internet Multicast Routing Architecture", RFC 5110, DOI 10.17487/RFC5110, January 2008, <https://www.rfc-editor.org/info/rfc5110>.

[RFC5507] IAB, Faltstrom, P., Ed., Austein, R., Ed., and P. Koch, Ed., "Design Choices When Expanding the DNS", RFC 5507, DOI 10.17487/RFC5507, April 2009, <https://www.rfc-editor.org/info/rfc5507>.

[RFC6726] Paila, T., Walsh, R., Luby, M., Roca, V., and R. Lehtonen, "FLUTE - File Delivery over Unidirectional Transport", RFC 6726, DOI 10.17487/RFC6726, November 2012, <https://www.rfc-editor.org/info/rfc6726>.

[RFC6895] Eastlake 3rd, D., "Domain Name System (DNS) IANA Considerations", BCP 42, RFC 6895, DOI 10.17487/RFC6895, April 2013, <https://www.rfc-editor.org/info/rfc6895>.

[RFC7359] Gont, F., "Layer 3 Virtual Private Network (VPN) Tunnel Traffic Leakages in Dual-Stack Hosts/Networks", RFC 7359, DOI 10.17487/RFC7359, August 2014, <https://www.rfc-editor.org/info/rfc7359>.

[RFC7761] Fenner, B., Handley, M., Holbrook, H., Kouvelas, I., Parekh, R., Zhang, Z., and L. Zheng, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", STD 83, RFC 7761, DOI 10.17487/RFC7761, March 2016, <https://www.rfc-editor.org/info/rfc7761>.

[RFC8126] Cotton, M., Leiba, B., and T. Narten, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 8126, DOI 10.17487/RFC8126, June 2017, <https://www.rfc-editor.org/info/rfc8126>.

[RFC8313] Tarapore, P., Ed., Sayko, R., Shepherd, G., Eckert, T., Ed., and R. Krishnan, "Use of Multicast across Inter-domain Peering Points", BCP 213, RFC 8313, DOI 10.17487/RFC8313, January 2018, <https://www.rfc-editor.org/info/rfc8313>.

[RFC8499] Hoffman, P., Sullivan, A., and K. Fujiwara, "DNS Terminology", BCP 219, RFC 8499, DOI 10.17487/RFC8499, January 2019, <https://www.rfc-editor.org/info/rfc8499>.

Appendix A. New RRType Request Form

This is the template for requesting a new RRType recommended in Appendix A of [RFC6895].

A. Submission Date:

B.1 Submission Type:
[x] New RRTYPE [ ] Modification to RRTYPE
B.2 Kind of RR:
[x] Data RR [ ] Meta-RR

C. Contact Information for submitter (will be publicly posted):
Name: Jake Holland
Email Address: jakeholland.net@gmail.com
International telephone number: +1-626-486-3706
Other contact handles: jholland@akamai.com

D. Motivation for the new RRTYPE application.
It provides a bootstrap so AMT (RFC 7450) gateways can discover an AMT relay that can receive multicast traffic from a specific source, in order to signal multicast group membership and receive multicast traffic over a unicast tunnel using AMT.

E. Description of the proposed RR type.
This description can be provided in-line in the template, as an attachment, or with a publicly available URL.
Please see draft-jholland-mboned-driad-amt-discovery.

F. What existing RRTYPE or RRTYPEs come closest to filling that need and why are they unsatisfactory?
Some similar concepts appear in IPSECKEY, as described in Section 1.2 of [RFC4025]. The IPSECKEY RRType is unsatisfactory because it refers to IPSec Keys instead of to AMT relays, but the motivating considerations for using reverse IP and for providing a precedence are similar--an AMT gateway often has access to a source address for a multicast (S,G), but does not have access to a relay address that can receive multicast traffic from the source, without administrative configuration.

Defining a format for a TXT record could serve the need for AMT relay discovery semantics, but Section 5 of [RFC5507] provides a compelling argument for requesting a new RRType instead.

G. What mnemonic is requested for the new RRTYPE (optional)?
AMTRELAY

H. Does the requested RRTYPE make use of any existing IANA registry

or require the creation of a new IANA subregistry in DNS Parameters?
Yes, IANA is requested to create a subregistry named "AMT Relay Type Field" in a "AMTRELAY Resource Record Parameters" registry. The field values are defined in Section 4.2.3 and Section 4.2.4, and a summary table is given in Section 5.

I. Does the proposal require/expect any changes in DNS servers/resolvers that prevent the new type from being processed as an unknown RRTYPE (see RFC3597)?
No.

J. Comments:
It may be worth noting that the gateway type field from Section 2.3 of [RFC4025] and Section 2.5 of [RFC4025] is very similar to the Relay Type field in this request. I tentatively assume that trying to re-use that sub-registry is a worse idea than duplicating it, but I'll invite others to consider the question and voice an opinion, in case there is a different consensus.
https://www.ietf.org/assignments/ipseckey-rr-parameters/ipseckey-rr-parameters.xml

Appendix B. Unknown RRType construction

In a DNS resolver that understands the AMTRELAY type, the zone file might contain this line:

```
IN AMTRELAY 128 0 3 amtrelays.example.com.
```

In order to translate this example to appear as an unknown RRType as defined in [RFC3597], one could run the following program:

<CODE BEGINS>

```
$ cat translate.py
#!/usr/bin/python3
import sys
name=sys.argv[1]
print(len(name))
print(''.join('%02x'%ord(x) for x in name))

$ ./translate.py amtrelays.example.com.
22
616d7472656c6179732e6578616d706c652e636f6d2e
```

<CODE ENDS>

The length and the hex string for the domain name "amtrelays.example.com." are the outputs of this program, yielding a length of 22 and the above hex string.

22 is the length of the domain name, so to this we add 2 (1 for the precedence field and 1 for the combined D-bit and relay type fields) to get the full length of the RData.

This results in a zone file entry like this:

```
 IN TYPE65280  \# ( 24 ; length
        80 ; precedence = 128
        03 ; D-bit=0, relay type=3 (wire-encoded domain name)
        616d7472656c6179732e6578616d706c652e636f6d2e ) ; domain name
```

Author's Address

Jake Holland
Akamai Technologies, Inc.
150 Broadway
Cambridge, MA 02144
United States of America

Email: jakeholland.net@gmail.com