Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 5, 2008

D. Thaler
M. Talwar
A. Aggarwal
Microsoft Corporation
L. Vicisano
Cisco Systems
T. Pusateri
!j

October 3, 2007

Automatic IP Multicast Without Explicit Tunnels (AMT) draft-ietf-mboned-auto-multicast-08

Status of this Memo

By submitting this Internet-Draft, each author represents that any applicable patent or other IPR claims of which he or she is aware have been or will be disclosed, and any of which he or she becomes aware will be disclosed, in accordance with Section 6 of BCP 79.

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

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

The list of current Internet-Drafts can be accessed at http://www.ietf.org/ietf/1id-abstracts.txt.

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

This Internet-Draft will expire on April 5, 2008.

Copyright Notice

Copyright (C) The IETF Trust (2007).

Abstract

Automatic Multicast Tunneling (AMT) allows multicast communication amongst isolated multicast-enabled sites or hosts, attached to a network which has no native multicast support. It also enables them to exchange multicast traffic with the native multicast infrastructure and does not require any manual configuration. AMT uses an encapsulation interface so that no changes to a host stack or applications are required, all protocols (not just UDP) are handled, and there is no additional overhead in core routers.

Table of Contents

1. Introduction

The primary goal of this document is to foster the deployment of native IP multicast by enabling a potentially large number of nodes to connect to the already present multicast infrastructure. Therefore, the techniques discussed here should be viewed as an interim solution to help in the various stages of the transition to a native multicast network.

To allow fast deployment, the solution presented here only requires small and concentrated changes to the network infrastructure, and no changes at all to user applications or to the socket API of end-nodes' operating systems. The protocol introduced in this specification can be deployed in a few strategically-placed network nodes and in user-installable software modules (pseudo device drivers and/or user-mode daemons) that reside underneath the socket API of end-nodes' operating systems. This mechanism is very similar to that used by "6to4" [RFC3056], [RFC3068] to get automatic IPv6 connectivity.

Effectively, AMT treats the unicast-only inter-network as a large non-broadcast multi-access (NBMA) link layer, over which we require the ability to multicast. To do this, multicast packets being sent to or from a site must be encapsulated in unicast packets. If the group has members in multiple sites, AMT encapsulation of the same multicast packet will take place multiple times by necessity.

2. Applicability

AMT is not a substitute for native multicast or a statically configured multicast tunnel for high traffic flow. Unicast replication is required to reach multiple receivers that are not part of the native multicast infrastructure. Unicast replication is also required by non-native sources to different parts of the native multicast infrastructure. However, this is no worse than regular unicast distribution of streams and in most cases much better.

The following problems are addressed:

1. Allowing isolated sites/hosts to receive the SSM flavor of multicast ([RFC4607]).

2. Allowing isolated non-NAT sites/hosts to transmit the SSM flavor of multicast.

3. Allowing isolated sites/hosts to receive general multicast (ASM [RFC1112]).

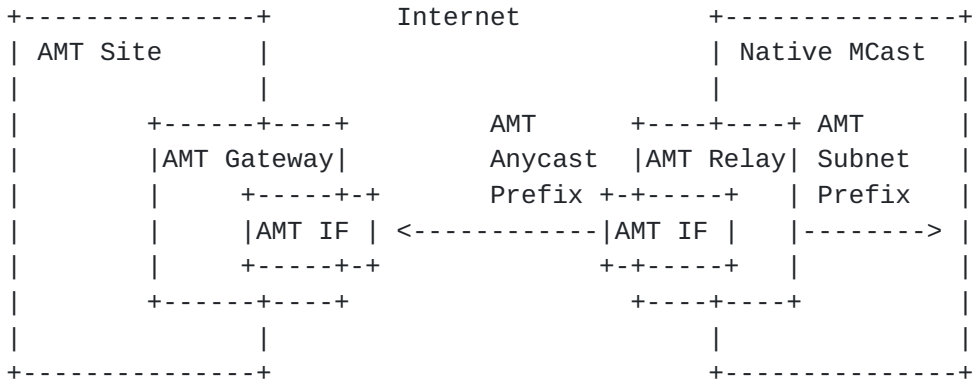
This document does not address allowing isolated sites/hosts to transmit general multicast. We expect that other solutions (e.g., Tunnel Brokers, a la [RFC3053]) will be used for sites that desire this capability.

Implementers should be aware that site administrators may have configured administratively scoped multicast boundaries and a remote gateway may provide a means to circumvent administrative boundaries. Therefore, implementations should allow for the configuration of such boundaries on relays and gateways and perform filtering as needed.

3. Requirements notation

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

4. Definitions



4.1. AMT Pseudo-Interface

AMT encapsulation of multicast packets inside unicast packets occurs at a point that is logically equivalent to an interface, with the link layer being the unicast-only network. This point is referred to as a pseudo-interface. Some implementations may treat it exactly like any other interface and others may treat it like a tunnel end-point.

4.2. AMT Gateway

A host, or a site gateway router, supporting an AMT Pseudo-Interface. It does not have native multicast connectivity to the native multicast backbone infrastructure. It is simply referred to in this document as a "gateway".

4.3. AMT Site

A multicast-enabled network not connected to the multicast backbone served by an AMT Gateway. It could also be a stand- alone AMT Gateway.

4.4. AMT Relay Router

A multicast router configured to support transit routing between AMT Sites and the native multicast backbone infrastructure. The relay router has one or more interfaces connected to the native multicast infrastructure, zero or more interfaces connected to the non-multicast capable inter-network, and an AMT pseudo-interface. It is simply referred to in this document as a "relay".

As with [RFC3056], we assume that normal multicast routers do not want to be tunnel endpoints (especially if this results in high fan out), and similarly that service providers do not want encapsulation

to arbitrary routers. Instead, we assume that special-purpose routers will be deployed that are suitable for serving as relays.

4.5. AMT Relay Anycast Prefix

A well-known address prefix used to advertise (into the unicast routing infrastructure) a route to an available AMT Relay Router. This could also be private (i.e., not well-known) for a private relay.

Prefixes for both IPv4 and IPv6 will be assigned in a future version of this draft.

4.6. AMT Relay Anycast Address

An anycast address which is used to reach the nearest AMT Relay Router.

This address corresponds to the setting the low-order octet of the AMT Relay Anycast Prefix to 1 (for both IPv4 and IPv6).

4.7. AMT Subnet Anycast Prefix

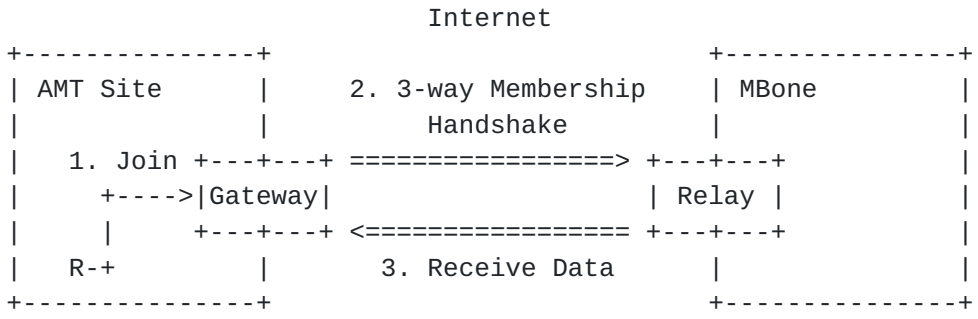
A well-known address prefix used to advertise (into the M-RIB of the native multicast-enabled infrastructure) a route to AMT Sites. This prefix will be used to enable sourcing SSM traffic from an AMT Gateway.

4.8. AMT Gateway Anycast Address

An anycast address in the AMT Subnet Anycast Prefix range, which is used by an AMT Gateway to enable sourcing SSM traffic from local applications.

5. Overview

5.1. Receiving Multicast in an AMT Site



Receiving Multicast in an AMT Site

AMT relays and gateways cooperate to transmit multicast traffic sourced within the native multicast infrastructure to AMT sites: relays receive the traffic natively and unicast-encapsulate it to gateways; gateways decapsulate the traffic and possibly forward it into the AMT site.

Each gateway has an AMT pseudo-interface that serves as a default multicast route. Requests to join a multicast session are sent to this interface and encapsulated to a particular relay reachable across the unicast-only infrastructure.

Each relay has an AMT pseudo-interface too. Multicast traffic sent on this interface is encapsulated to zero or more gateways that have joined to the relay. The AMT recipient-list is determined for each multicast session. This requires the relay to keep state for each gateway which has joined a particular group or (source, group) pair. Multicast packets from the native infrastructure behind the relay will be sent to each gateway which has requested them.

All multicast packets (data and control) are encapsulated in unicast packets. UDP encapsulation is used for all AMT control and data packets using the IANA reserved UDP port number for AMT.

Each relay, plus the set of all gateways using the relay, together are thought of as being on a separate logical NBMA link. This implies that the AMT recipient-list is a list of "link layer" addresses which are (IP address, UDP port) pairs.

Since the number of gateways using a relay can be quite large, and we expect that most sites will not want to receive most groups, an explicit-joining protocol is required for gateways to communicate group membership information to a relay. The two most likely

candidates are the IGMP/MLD protocol [RFC3376], [RFC3810], and the PIM-Sparse Mode protocol [RFC4601]. Since an AMT gateway may be a host, and hosts typically do not implement routing protocols, gateways will use IGMP/MLD as described in Section 7 below. This allows a host kernel (or a pseudo device driver) to easily implement AMT gateway behavior, and obviates the relay from the need to know whether a given gateway is a host or a router. From the relay's perspective, all gateways are indistinguishable from hosts on an NBMA leaf network.

5.1.1. Scalability Considerations

It is possible that millions of hosts will enable AMT gateway functionality and so an important design goal is not to create gateway state in each relay until the gateway joins a multicast group. But even the requirement that a relay keep group state per gateway that has joined a group introduces potential scalability concerns.

Scalability of AMT can be achieved by adding more relays, and using an appropriate relay discovery mechanism for gateways to discover relays. The solution we adopt is to assign addresses in anycast fashion to relays [RFC1546], [RFC4291]. However, simply sending periodic membership reports to an anycast address can cause duplicates. Specifically, if routing changes such that a different relay receives a periodic membership report, both the new and old relays will encapsulate data to the AMT site until the old relay's state times out. This is obviously undesirable. Instead, we use the anycast address merely to find the unicast address of a relay to which membership reports are sent.

Since adding another relay has the result of adding another independent NBMA link, this allows the gateways to be spread out among more relays so as to keep the number of gateways per relay at a reasonable level.

5.1.2. Spoofing Considerations

An attacker could affect the group state in the relay or gateway by spoofing the source address in the join or leave reports. This can be used to launch reflection or denial of service attacks on the target. Such attacks can be mitigated by using a three way handshake between the gateway and the relay for each multicast membership report or leave.

When a gateway or relay wants to send a membership report, it first sends an AMT Request with a request nonce in it. The receiving side (the respondent) can calculate a message authentication code (MAC)

based on (for example) the source IP address of the Request, the source UDP port, the request nonce, and a secret key known only to the respondent. The algorithm and the input used to calculate the MAC does not have to be standardized since the respondent generates and verifies the MAC and the originator simply echoes it.

An AMT Membership Query is sent back including the request nonce and the MAC to the originator of the Request. The originator then sends the IGMP/MLD Membership/Listener Report or Leave/Done (including the IP Header) along with the request nonce and the received MAC back to the respondent finalizing the 3-way handshake.

Upon reception, the respondent can recalculate the MAC based on the source IP address, the source UDP port, the request nonce, and the local secret. The IGMP/MLD message is only accepted if the received MAC matches the calculated MAC.

The local secret never has to be shared with the other side. It is only used to verify return routability of the originator.

Since the same Request Nonce and source IP address can be re-used, the receiver SHOULD change its secret key at least once per hour. However, AMT Membership updates received with the previous secret MUST be accepted for up to the IGMP/MLD Query Interval.

5.1.3. Protocol Sequence for a Gateway Joining SSM Receivers to a Relay

This description assumes the Gateway can be a host joining as a receiver or a network device acting as a Gateway when a directly connected host joins as a receiver.

- Receiver at AMT site sends IGMPv3/MLDv2 report joining (S1,G1).
- Gateway receives report. If it has no tunnel state with a Relay, it originates an AMT Relay Discovery message addressed to the Anycast Relay IP address. The AMT Relay Discovery message can be sent on demand if no relay is known at this time or at startup and be periodically refreshed.
- The closest Relay topologically receives the AMT Relay Discovery message and returns the nonce from the Discovery in an AMT Relay Advertisement message so the Gateway can learn of the Relay's unique IP address.
- When the Gateway receives the AMT Relay Advertisement message, it now has an address to use for all subsequent (S,G) entries it will join on behalf of attached receivers (or itself).

- If the gateway has a valid Response MAC from a previous AMT Query message, it can send an AMT Membership Update message as described below. Otherwise, the Gateway sends an AMT Request message to the Relay's unique IP address to begin the process of joining the (S,G). The gateway also SHOULD initialize a timer used to send periodic Requests to a random value from the interval [0, [Query Interval]] before sending the first periodic report, in order to prevent startup synchronization.

- The Relay responds to the AMT Request message by returning the nonce from the Request in a AMT Query message. The Query message contains an IGMP/MLD QUERY indicating how often the Gateway should repeat AMT Request messages so the (S,G) state can stay refreshed in the Relay. The Query message also includes an opaque security code which is generated locally (with no external coordination).

- When the Gateway receives the AMT Query message it responds by copying the security code from the AMT Query message into a AMT Membership Update message. The Update message contains (S1,G1) in an IGMPv3/MLDv2 formatted packet with an IP header. The nonce from the AMT Request is also included in the AMT Membership Update message.

- When the Relay receives the AMT Membership Update, it will add the tunnel to the Gateway in it's outgoing interface list for it's (S1,G1) entry stored in the multicast routing table. If the (S1,G1) entry was created do to this interaction, the multicast routing protocol running on the Relay will trigger a Join message towards source S1 to build a native multicast tree in the native multicast infrastructure.

- As packets are sent from the host S1, they will travel natively down the multicast tree associated with (S1,G1) in the native multicast infrastructure to the Relay. The Relay will replicate to all interfaces in it's outgoing interface list as well as the tunnel outgoing interface, which is encapsulated in a unicast AMT Multicast Data message.

- When the Gateway receives the AMT Multicast Data message, it will accept the packet since it was received over the pseudo-interface associated with the tunnel to the Relay it had attached to, and forward the packet to the outgoing interfaces joined by any attached receiver hosts (or deliver the packet to the application when the Gateway is the receiver).

- If later (S2,G2) is joined by a receiver, a 3-way handshake of Request/ Query/Update occurs for this entry. The Discovery/ Advertisement exchange is not required.

- To keep the state for (S1,G1) and (S2,G2) alive in the Relay, the Gateway will send periodic AMT Membership Updates. The Membership Update can be sent directly if the sender has a valid nonce from a previous Request. If not, an AMT Request messages should be sent to solicit a Query Message. When sending a periodic state refresh, all joined state in the Gateway is packed in the fewest number of AMT Membership Update messages.

- When the Gateway leaves all (S,G) entries, the Relay can free resources associated with the tunnel. It is assumed that when the Gateway would want to join an (S,G) again, it would start the Discovery/Advertisement tunnel establishment process over again.

This same procedure would be used for receivers who operate in Any-Source Multicast (ASM) mode.

5.2. Sourcing Multicast from an AMT site

Two cases are discussed below: multicast traffic sourced in an AMT site and received in the MBone, and multicast traffic sourced in an AMT site and received in another AMT site.

In both cases only SSM sources are supported. Furthermore this specification only deals with the source residing directly in the gateway. To enable a generic node in an AMT site to source multicast, additional coordination between the gateway and the source-node is required.

The gateway SHOULD allow for filtering link-local and site-local traffic.

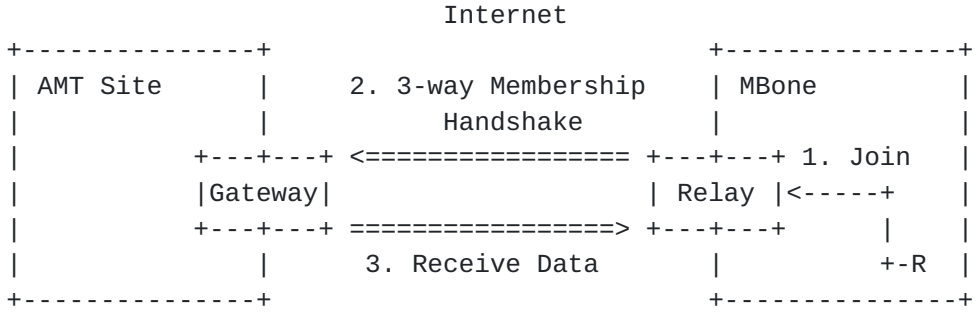
The general mechanism used to join towards AMT sources is based on the following:

1. Applications residing in the gateway use addresses in the AMT Subnet Anycast Prefix to send multicast, as a result of sourcing traffic on the AMT pseudo-interface.

2. The AMT Subnet Anycast Prefix is advertised for RPF reachability in the M-RIB by relays and gateways.

3. Relays or gateways that receive a join for a source/group pair use information encoded in the address pair to rebuild the address of the gateway (source) to which to encapsulate the join (see Section 7.2 for more details). The membership reports use the same three way handshake as outlined in Section 5.1.2

5.2.1. Supporting Site-MBone Multicast



Site-MBone Multicast

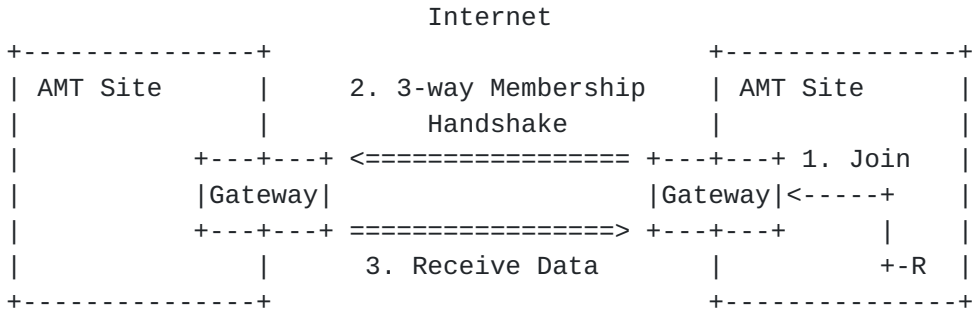
If a relay receives an explicit join from the native infrastructure, for a given (source, group) pair where the source address belongs to the AMT Subnet Anycast Prefix, then the relay will periodically (using the rules specified in Section 5.1.2) encapsulate membership updates for the group to the gateway. The gateway must keep state per relay from which membership reports have been sent, and forward multicast traffic from the site to all relays from which membership reports have been received. The choice of whether this state and replication is done at the link-layer (i.e., by the tunnel interface) or at the network-layer is implementation dependent.

If there are multiple relays present, this ensures that data from the AMT site is received via the closest relay to the receiver. This is necessary when the routers in the native multicast infrastructure employ Reverse-Path Forwarding (RPF) checks against the source address, such as occurs when PIM Sparse-Mode [RFC4601] is used by the multicast infrastructure.

The solution above will scale to an arbitrary number of relays, as long at the number of relays requiring multicast traffic from a given AMT site remains reasonable enough to not overly burden the site's gateway.

A source at or behind an AMT gateway requires the gateway to do the replication to one or more relays and receiving gateways. If this places too much of a burden on the sourcing gateway, the source should join the native multicast infrastructure through a permanent tunnel so that replication occurs within the native multicast infrastructure.

5.2.2. Supporting Site-Site Multicast



Site-Site Multicast

Since we require gateways to accept membership reports, as described above, it is also possible to support multicast among AMT sites, without requiring assistance from any relays.

When a gateway wants to join a given (source, group) pair, where the source address belongs to the AMT Subnet Anycast Prefix, then the gateway will periodically unicast encapsulate an IGMPv3/MLDv2 Report [RFC3376], [RFC3810] (including IP Header) directly to the site gateway for the source.

We note that this can result in a significant amount of state at a site gateway sourcing multicast to a large number of other AMT sites. However, it is expected that this is not unreasonable for two reasons. First, the gateway does not have native multicast connectivity, and as a result is likely doing unicast replication at present. The amount of state is thus the same as what such a site already deals with. Secondly, any site expecting to source traffic to a large number of sites could get a point-to-point tunnel to the native multicast infrastructure, and use that instead of AMT.

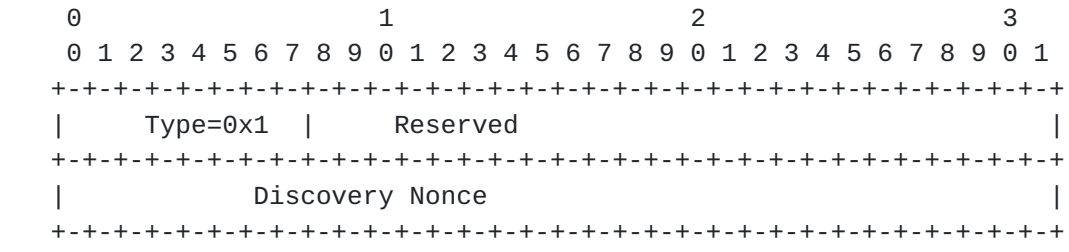
6. Message Formats

6.1. AMT Relay Discovery

The AMT Relay Discovery message is a UDP packet sent from the AMT gateway unicast address to the AMT relay anycast address to discover the unicast address of an AMT relay.

The UDP source port is uniquely selected by the local host operating system. The UDP destination port is the IANA reserved AMT port number. The UDP checksum MUST be valid in AMT control messages.

The payload of the UDP packet contains the following fields.



AMT Relay Discovery

6.1.1. Type

The type of the message.

6.1.2. Reserved

A 24-bit reserved field. Sent as 0, ignored on receipt.

6.1.3. Discovery Nonce

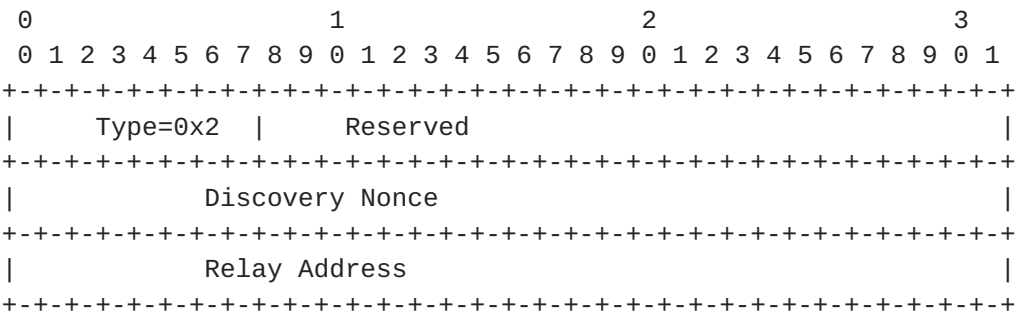
A 32-bit random value generated by the gateway and replayed by the relay.

6.2. AMT Relay Advertisement

The AMT Relay Advertisement message is a UDP packet sent from the AMT relay anycast address to the source of the discovery message.

The UDP source port is the IANA reserved AMT port number and the UDP destination port is the source port received in the Discovery message. The UDP checksum MUST be valid in AMT control messages.

The payload of the UDP packet contains the following fields.



AMT Relay Advertisement

6.2.1. Type

The type of the message.

6.2.2. Reserved

A 24-bit reserved field. Sent as 0, ignored on receipt.

6.2.3. Discovery Nonce

A 32-bit random value generated by the gateway and replayed by the relay.

6.2.4. Relay Address

The unicast IPv4 or IPv6 address of the AMT relay. The family can be determined by the length of the Advertisement.

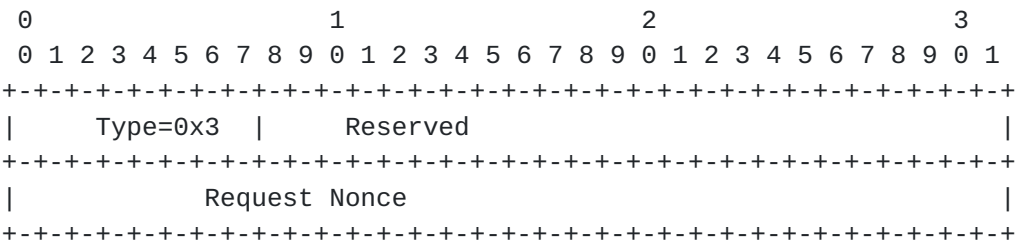
6.3. AMT Request

A Request packet is sent to begin a 3-way handshake for sending an IGMP/MLD Membership/Listener Report or Leave/Done. It can be sent from a gateway to a relay, from a gateway to another gateway, or from a relay to a gateway.

It is sent from the originator's unique unicast address to the respondents' unique unicast address.

The UDP source port is uniquely selected by the local host operating system. It can be different for each Request and different from the source port used in Discovery messages but does not have to be. The UDP destination port is the IANA reserved AMT port number. The UDP

checksum MUST be valid in AMT control messages.



AMT Relay Advertisement

6.3.1. Type

The type of the message.

6.3.2. Reserved

A 24-bit reserved field. Sent as 0, ignored on receipt.

6.3.3. Request Nonce

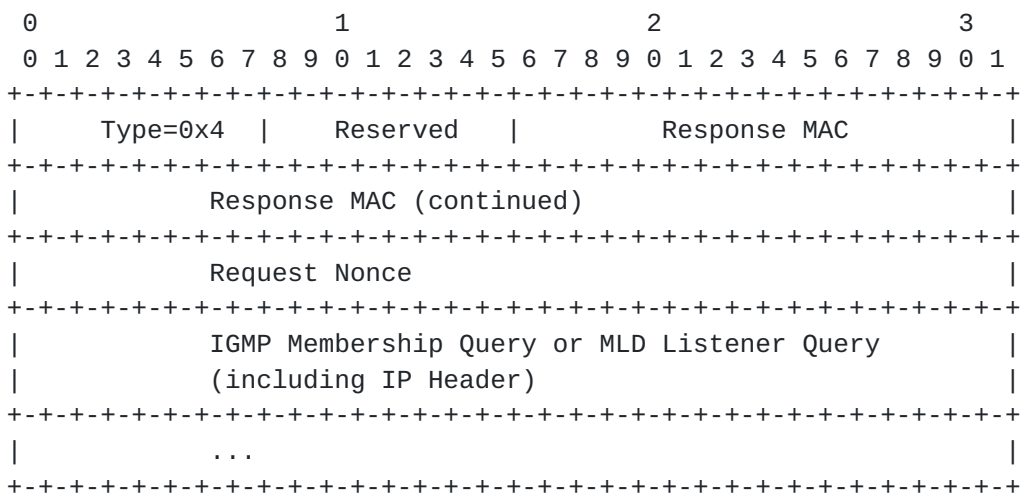
A 32-bit identifier used to distinguish this request.

6.4. AMT Membership Query

An AMT Membership Query packet is sent from the respondent back to the originator to solicit an AMT Membership Update while confirming the source of the original request. It contains a relay Message Authentication Code (MAC) that is a cryptographic hash of a private secret, the originators address, and the request nonce.

It is sent from the destination address received in the Request to the source address received in the Request which is the same address used in the Relay Advertisement.

The UDP source port is the IANA reserved AMT port number and the UDP destination port is the source port received in the Request message. The UDP checksum MUST be valid in AMT control messages.



AMT Membership Query

6.4.1. Type

The type of the message.

6.4.2. Reserved

A 8-bit reserved field. Sent as 0, ignored on receipt.

6.4.3. Response MAC

A 48-bit hash generated by the respondent and sent to the originator for inclusion in the AMT Membership Update. The algorithm used for this is chosen by the respondent but an algorithm such as HMAC-MD5-48 [RFC2104] SHOULD be used at a minimum.

6.4.4. Request Nonce

A 32-bit identifier used to distinguish this request echoed back to the originator.

6.4.5. IGMP/MLD Query (including IP Header)

The message contains either an IGMP Query or an MLD Multicast Listener Query. The IGMP or MLD version sent should default to IGMPv3 or MLDv2 unless explicitly configured to use IGMPv2 or MLDv1. The IGMP/MLD Query includes a full IP Header. The IP source address of the query would match the anycast address on the pseudo interface. The TTL of the outer header should be sufficient to reach the tunnel endpoint and not mimic the inner header TTL which is typically 1 for IGMP/MLD messages.

6.5. AMT Membership Update

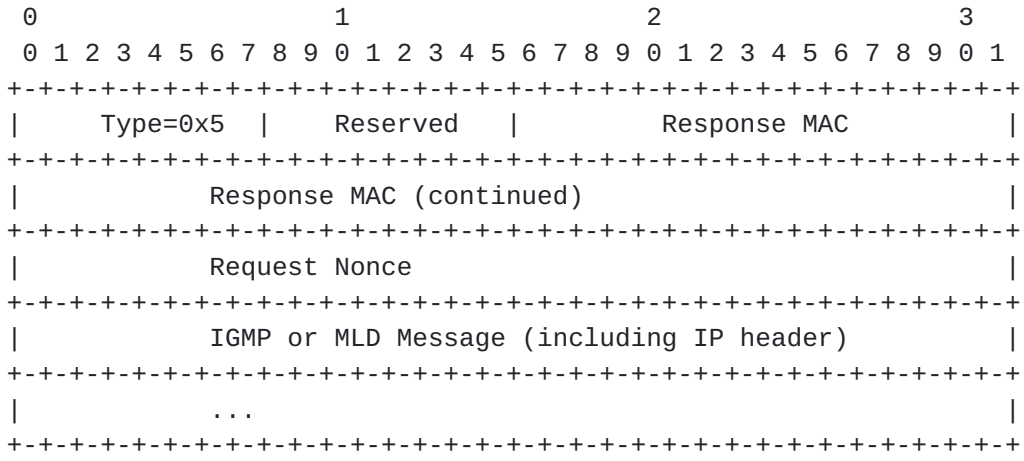
An AMT Membership Update is sent to report a membership after a valid Response MAC has been received. It contains the original IGMP/MLD Membership/Listener Report or Leave/Done received over the AMT pseudo-interface including the original IP header. It echoes the Response MAC received in the AMT Membership Query so the respondent can verify return routability to the originator.

It is sent from the destination address received in the Query to the source address received in the Query which should both be the same as the original Request.

The UDP source and destination port numbers should be the same ones sent in the original Request.

The relay is not required to use the IP source address of the IGMP Membership Report for any particular purpose.

The same Request Nonce and Response MAC can be used across multiple AMT Membership Update messages without having to send individual AMT Membership Query messages.



AMT Membership Update

6.5.1. Type

The type of the message.

6.5.2. Reserved

A 8-bit reserved field. Sent as 0, ignored on receipt.

6.5.3. Response MAC

The 48-bit MAC received in the Membership Query and echoed back in the Membership Update.

6.5.4. Request Nonce

A 32-bit identifier used to distinguish this request.

6.5.5. IGMP/MLD Message (including IP Header)

The message contains either an IGMP Membership Report, an IGMP Membership Leave, an MLD Multicast Listener Report, or an MLD Listener Done. The IGMP or MLD version sent should be in response the version of the query received in the AMT Membership Query. The IGMP/MLD Message includes a full IP Header.

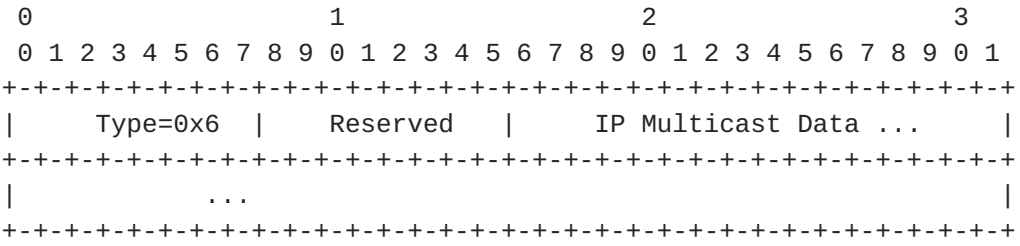
6.6. AMT IP Multicast Data

The AMT Data message is a UDP packet encapsulating the IP Multicast data requested by the originator based on a previous AMT Membership Update message.

It is sent from the unicast destination address of the Membership update to the source address of the Membership Update.

The UDP source and destination port numbers should be the same ones sent in the original Query. The UDP checksum SHOULD be 0 in the AMT IP Multicast Data message.

The payload of the UDP packet contains the following fields.



AMT IP Multicast Data

6.6.1. Type

The type of the message.

6.6.2. Reserved

An 8-bit reserved field. Sent as 0, ignored on receipt.

6.6.3. IP Multicast Data

The original IP Multicast data packet that is being replicated by the relay to the gateways including the original IP header.

7. AMT Gateway Details

This section details the behavior of an AMT Gateway, which may be a router serving an AMT site, or the site may consist of a single host, serving as its own gateway.

7.1. At Startup Time

At startup time, the AMT gateway will bring up an AMT pseudo-interface to be used for encapsulation. The gateway needs to discover an AMT Relay to send Membership Requests. It can send an AMT Relay Discovery at startup time or wait until it has a group membership to report. The AMT Relay Discovery message is sent to the AMT Relay Anycast Address. A unicast address (which is treated as a link-layer address to the encapsulation interface) is received in the AMT Relay Advertisement message. The discovery process SHOULD be done periodically (e.g., once a day) to re-resolve the unicast address of a close relay. To prevent startup synchronization, the timer SHOULD use at least 10 percent jitter.

If the gateway is serving as a local router, it SHOULD also function as an IGMP/MLD Proxy, as described in [RFC4605], with its IGMP/MLD host-mode interface being the AMT pseudo-interface. This enables it to translate group memberships on its downstream interfaces into IGMP/MLD Reports. Hosts receiving multicast packets through an AMT gateway acting as a proxy should ensure that their M-RIB accepts multicast packets from the AMT gateway for the sources it is joining.

7.2. Gateway Group and Source Addresses

To support sourcing traffic to SSM groups by a gateway with a global unicast address, the AMT Subnet Anycast Prefix is treated as the subnet prefix of the AMT pseudo-interface, and an anycast address is added on the interface. This anycast address is formed by concatenating the AMT Subnet Anycast Prefix followed by the high bits of the gateway's global unicast address.

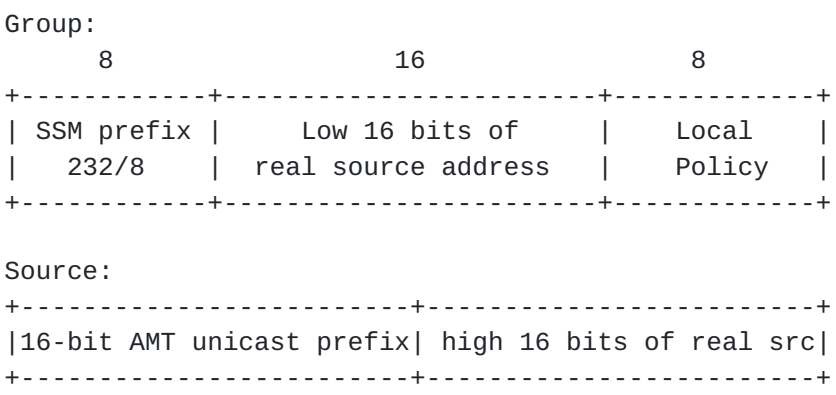
The remaining bits of its global unicast address are appended to the SSM prefix to create the group address and any spare bits may be allocated using local policy.

If a gateway wants to source multicast traffic, it must select the gateway source address and SSM group address in such a way that the AMT relay can have enough information to reconstruct the gateway's unicast address when it receives an SSM join for the source.

Note that multiple gateways might end up with the same anycast address assigned to their pseudo-interfaces.

7.2.1. IPv4

For example, if IANA assigns the IPv4 prefix x.y/16 as the AMT Subnet Anycast Prefix, and the gateway has global unicast address a.b.c.d, then the AMT Gateway's Anycast Source Address will be x.y.a.b. Since the IPv4 SSM group range is 232/8, it MUST allocate IPv4 SSM groups in the range 232.c.d/24.

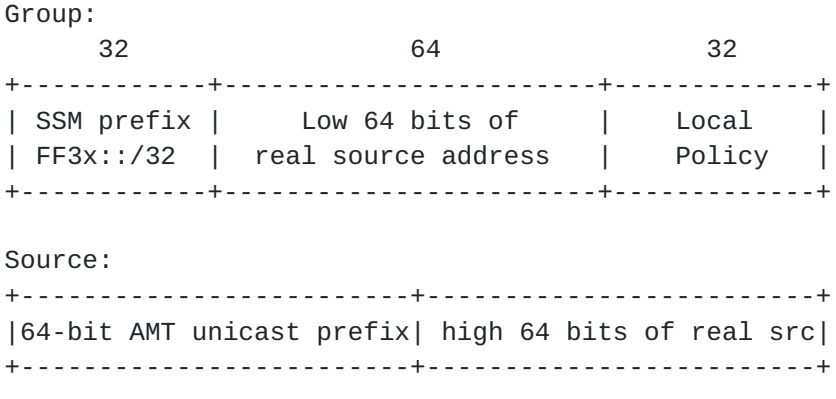


IPv4 format

This allows for 2^8 (256) IPv4 group addresses for use by each AMT gateway.

7.2.2. IPv6

Similarly for IPv6, this is illustrated in the following figure.



IPv6 format

This allows for 2^32 (over 4 billion) IPv6 group addresses for use by each AMT gateway.

7.3. Joining Groups with MBone Sources

The IGMP/MLD protocol usually operates by having the Querier multicast an IGMP/MLD Query message on the link. This behavior does not work on NBMA links which do not support multicast. Since the set of gateways is typically unknown to the relay (and potentially quite large), unicasting the queries is also impractical. The following behavior is used instead.

Applications residing in a gateway should join groups on the AMT pseudo-interface, causing IGMP/MLD Membership/Listener Reports to be sent over that interface. When UDP encapsulating the membership reports (and in fact any other messages, unless specified otherwise in this document), the destination address in the outer IP header is the relay's unicast address. Robustness is provided by the underlying IGMP/MLD protocol messages sent on the AMT pseudo-interface. In other words, the gateway does not need to retransmit IGMP/MLD Membership/Listener Reports and Leave/Done messages received on the pseudo-interface since IGMP/MLD will already do this. The gateway simply needs to encapsulate each IGMP/MLD Membership/Listener Report and Leave/Done message it receives.

However, since periodic IGMP/MLD Membership/Listener Reports are sent in response to IGMP/MLD Queries, a mechanism to trigger periodic Membership/Listener Reports and Leave/Done messages is necessary. The gateway should use a timer to trigger periodic AMT Membership Updates.

If the gateway is behind a firewall device, the firewall may require the gateway to periodically refresh the UDP state in the firewall at a shorter interval than the standard IGMP/MLD Query interval. AMT Requests can be sent periodically to solicit IGMP/MLD Queries. The interval at which the AMT Requests are sent should be configurable to ensure the firewall does not revert to blocking the UDP encapsulated IP Multicast data packets. When the AMT Query is received, it can be ignored unless it is time for a periodic AMT Membership Update.

The relay can use the Querier's Robustness Variable (QRV) defined in [RFC3376] and [RFC3810] to adjust the number of Membership/Listener Reports that are sent by the host joining the group.

7.4. Responding to Relay Changes

When a gateway determines that its current relay is unreachable (e.g., upon receipt of an ICMP Unreachable message [RFC0792] for the relay's unicast address), it may need to repeat relay address discovery. However, care should be taken not to abandon the current relay too quickly due to transient network conditions.

7.5. Joining SSM Groups with AMT Gateway Sources

An IGMPv3/MLDv2 Report for a given (source, group) pair MAY be encapsulated directly to the source, when the source address belongs to the AMT Subnet Anycast Prefix.

The "link-layer" address to use as the destination address in the outer IP header is obtained as follows. The source address in the inclusion list of the IGMPv3/MLDv2 report will be an AMT Gateway Anycast Address with the high bits of the address, and the remaining bits will be in the middle of the group address.

Section 7.2 describes this format to recover the gateway source address.

7.6. Receiving AMT Membership Updates by the Gateway

When an AMT Request is received by the gateway from another gateway or relay, it follows the same 3-way handshake procedure a relay would follow if it received the AMT Request. It generates a MAC and responds with an AMT Membership Query. When the AMT Membership Update is received, it verifies the MAC and then processes the IGMP/MLD Membership/Listener Report or Leave/Done.

At the gateway, the IGMP/MLD packet should be an IGMPv3/MLDv2 source specific (S,G) join or leave.

If S is not the AMT Gateway Anycast Address, the packet is silently discarded. If G does not contain the low bits of the global unicast address (as described above), the packet is also silently discarded.

The gateway adds the source address (from the outer IP header) and UDP port of the report to a membership list for G. Maintaining this membership list may be done in any implementation-dependent manner. For example, it might be maintained by the "link-layer" inside the AMT pseudo-interface, making it invisible to the normal IGMP/MLD module.

7.7. Sending data to SSM groups

When multicast packets are sent on the AMT pseudo-interface, they are encapsulated as follows. If the group address is not an SSM group, then the packet is silently discarded (this memo does not currently provide a way to send to non-SSM groups).

If the group address is an SSM group, then the packet is unicast encapsulated to each remote node from which the gateway has received an IGMPv3/MLDv2 report for the packet's (source, group) pair.

8. Relay Router Details

8.1. At Startup time

At startup time, the relay router will bring up an NBMA-style AMT pseudo-interface. It shall also add the AMT Relay Anycast Address on some interface.

The relay router shall then advertise the AMT Relay Anycast Prefix into the unicast-only Internet, as if it were a connection to an external network. When the advertisement is done using BGP, the AS path leading to the AMT Relay Anycast Prefix shall include the identifier of the local AS.

The relay router shall also enable IGMPv3/MLDv2 on the AMT pseudo-interface, except that it shall not multicast Queries (this might be done, for example, by having the AMT pseudo-device drop them, or by having the IGMP/MLD module not send them in the first place).

Finally, to support sourcing SSM traffic from AMT sites, the AMT Subnet Anycast Prefix is assigned to the AMT pseudo-interface, and the AMT Subnet Anycast Prefix is injected by the AMT Relay into the M-RIB of MBGP.

8.2. Receiving Relay Discovery messages sent to the Anycast Address

When a relay receives an AMT Relay Discovery message directed to the AMT Relay Anycast Address, it should respond with an AMT Relay Advertisement containing its unicast address. The source and destination addresses of the advertisement should be the same as the destination and source addresses of the discovery message respectively. Further, the nonce in the discovery message MUST be copied into the advertisement message.

8.3. Receiving Membership Updates from AMT Gateways

The relay operates passively, sending no periodic IGMP/MLD Queries but simply tracking membership information according to AMT Request/ Query/Membership Update tuples received. In addition, the relay must also do explicit membership tracking, as to which gateways on the AMT pseudo-interface have joined which groups. Once an AMT Membership Update has been successfully received, it updates the forwarding state for the appropriate group and source (if provided). When data arrives for that group, the traffic must be encapsulated to each gateway which has joined that group or (S,G).

The explicit membership tracking and unicast replication may be done in any implementation-specific manner. Some examples are:

1. The AMT pseudo-device driver might track the group information and perform the replication at the "link-layer", with no changes to a pre-existing IGMP/MLD module.

2. The IGMP/MLD module might have native support for explicit membership tracking, especially if it supports other NBMA-style interfaces.

If a relay wants to affect the rate at which the AMT Requests are originated from a gateway, it can tune the membership timeout by adjusting the Querier's Query Interval Code (QQIC) field in the IGMP/MLD Query contained within the AMT Membership Query message. The QQIC field is defined in [RFC3376] and [RFC3810]. However, since the gateway may need to send AMT Requests frequently enough to prevent firewall state from timing out, the relay may be limited in its ability to spread out Requests coming from a gateway by adjusting the QQIC field.

8.4. Receiving (S,G) Joins from the Native Side, for AMT Sources

The relay sends an IGMPv3/MLDv2 report to the AMT source as described above in Section 5.1.2

9. IANA Considerations

9.1. IPv4 and IPv6 Anycast Prefix Allocation

The IANA should allocate an IPv4 prefix and an IPv6 prefix dedicated to the public AMT Relays to advertise to the native multicast backbone. The prefix length should be determined by the IANA; the prefix should be large enough to guarantee advertisement in the default-free BGP networks.

9.1.1. IPv4

A prefix length of 16 will meet this requirement.

9.1.2. IPv6

A prefix length of 32 will meet this requirement. IANA has previously set aside the range 2001::/16 for allocating prefixes for this purpose.

9.2. IPv4 and IPv6 AMT Subnet Prefix Allocation

It should also be noted that this prefix length directly affects the number of groups available to be created by the AMT gateway: in the IPv4 case, a prefix length of 16 gives 256 groups, and a prefix length of 8 gives 65536 groups.

9.2.1. IPv4

As described above in Section 7.2.1 an IPv4 prefix with a length of 16 is requested for this purpose.

9.2.2. IPv6

As described above in Section 7.2.2 an IPv6 prefix with a length of 32 is requested.

9.3. UDP Port number

IANA has previously allocated UDP reserved port number 2268 for AMT encapsulation.

All allocations are a one time effort and there will be no need for any recurring assignment after this stage.

10. Security Considerations

The anycast technique introduces a risk that a rogue router or a rogue AS could introduce a bogus route to the AMT Relay Anycast prefix, and thus divert the traffic. Network managers have to guarantee the integrity of their routing to the AMT Relay Anycast prefix in much the same way that they guarantee the integrity of all other routes.

Within the native MBGP infrastructure, there is a risk that a rogue router or a rogue AS could inject a false route to the AMT Subnet Anycast Prefix, and thus divert joins and cause RPF failures of multicast traffic. As the AMT Subnet Anycast Prefix will be advertised by multiple entities, guaranteeing the integrity of this shared MBGP prefix is much more challenging than verifying the correctness of a regular unicast advertisement. To mitigate this threat, routing operators should configure the BGP sessions to filter out any more specific advertisements for the AMT Subnet Anycast Prefix.

Gateways and relays will accept and decapsulate multicast traffic from any source from which regular unicast traffic is accepted. If this is for any reason felt to be a security risk, then additional source address based packet filtering MUST be applied:

1. To prevent a rogue sender (that can't do traditional spoofing because of e.g. access lists deployed by its ISP) from making use of AMT to send packets to an SSM tree, a relay that receives an encapsulated multicast packet MUST discard the multicast packet if the IP source address in the outer header does not match the source address that would be extracted using the rules of Section 7.2.

2. A gateway MUST discard encapsulated multicast packets if the source address in the outer header is not the address to which the encapsulated join message was sent. An AMT Gateway that receives an encapsulated IGMPv3/MLDv2 (S,G)-Join MUST discard the message if the IP destination address in the outer header does not match the source address that would be extracted using the rules of Section 7.2.

11. Contributors

The following people provided significant contributions to earlier versions of this draft.

Dirk Ooms
OneSparrow
Belegstraat 13; 2018 Antwerp; Belgium
EMail: dirk@onesparrow.com

12. Acknowledgments

Most of the mechanisms described in this document are based on similar work done by the NGTrans WG for obtaining automatic IPv6 connectivity without explicit tunnels ("6to4"). Tony Ballardie provided helpful discussion that inspired this document.

In addition, extensive comments were received from Pekka Savola, Greg Shepherd, Dino Farinacci, Toerless Eckert, Marshall Eubanks, John Zwiebel, and Lenny Giuliano.

Juniper Networks was instrumental in funding several versions of this draft as well as an open source implementation.

13. References

13.1. Normative References

[RFC0792] Postel, J., "Internet Control Message Protocol", STD 5, RFC 792, September 1981.

[RFC3376] Cain, B., Deering, S., Kouvelas, I., Fenner, B., and A. Thyagarajan, "Internet Group Management Protocol, Version 3", RFC 3376, October 2002.

[RFC3810] Vida, R. and L. Costa, "Multicast Listener Discovery Version 2 (MLDv2) for IPv6", RFC 3810, June 2004.

[RFC4605] Fenner, B., He, H., Haberman, B., and H. Sandick, "Internet Group Management Protocol (IGMP) / Multicast Listener Discovery (MLD)-Based Multicast Forwarding ("IGMP/MLD Proxying")", RFC 4605, August 2006.

[RFC4607] Holbrook, H. and B. Cain, "Source-Specific Multicast for IP", RFC 4607, August 2006.

13.2. Informative References

[RFC1112] Deering, S., "Host extensions for IP multicasting", STD 5, RFC 1112, August 1989.

[RFC1546] Partridge, C., Mendez, T., and W. Milliken, "Host Anycasting Service", RFC 1546, November 1993.

[RFC2104] Krawczyk, H., Bellare, M., and R. Canetti, "HMAC: Keyed-Hashing for Message Authentication", RFC 2104, February 1997.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC3053] Durand, A., Fasano, P., Guardini, I., and D. Lento, "IPv6 Tunnel Broker", RFC 3053, January 2001.

[RFC3056] Carpenter, B. and K. Moore, "Connection of IPv6 Domains via IPv4 Clouds", RFC 3056, February 2001.

[RFC3068] Huitema, C., "An Anycast Prefix for 6to4 Relay Routers", RFC 3068, June 2001.

[RFC4291] Hinden, R. and S. Deering, "IP Version 6 Addressing Architecture", RFC 4291, February 2006.

[RFC4601] Fenner, B., Handley, M., Holbrook, H., and I. Kouvelas, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", RFC 4601, August 2006.

Authors' Addresses

Dave Thaler
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052-6399
USA

Phone: +1 425 703 8835
Email: dthaler@microsoft.com

Mohit Talwar
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052-6399
USA

Phone: +1 425 705 3131
Email: mohitt@microsoft.com

Amit Aggarwal
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052-6399
USA

Phone: +1 425 706 0593
Email: amitag@microsoft.com

Lorenzo Vicisano
Cisco Systems
170 West Tasman Dr.
San Jose, CA 95134
USA

Phone: +1 408 525 2530
Email: lorenzo@cisco.com

Tom Pusateri
!j
222 E. Jones Ave.
Wake Forest, NC 27587
USA

Email: pusateri@bangj.com

Full Copyright Statement

Copyright (C) The IETF Trust (2007).

This document is subject to the rights, licenses and restrictions contained in BCP 78, and except as set forth therein, the authors retain all their rights.

This document and the information contained herein are provided on an "AS IS" basis and THE CONTRIBUTOR, THE ORGANIZATION HE/SHE REPRESENTS OR IS SPONSORED BY (IF ANY), THE INTERNET SOCIETY, THE IETF TRUST AND THE INTERNET ENGINEERING TASK FORCE DISCLAIM ALL WARRANTIES, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO ANY WARRANTY THAT THE USE OF THE INFORMATION HEREIN WILL NOT INFRINGE ANY RIGHTS OR ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

Intellectual Property

The IETF takes no position regarding the validity or scope of any Intellectual Property Rights or other rights that might be claimed to pertain to the implementation or use of the technology described in this document or the extent to which any license under such rights might or might not be available; nor does it represent that it has made any independent effort to identify any such rights. Information on the procedures with respect to rights in RFC documents can be found in BCP 78 and BCP 79.

Copies of IPR disclosures made to the IETF Secretariat and any assurances of licenses to be made available, or the result of an attempt made to obtain a general license or permission for the use of such proprietary rights by implementers or users of this specification can be obtained from the IETF on-line IPR repository at http://www.ietf.org/ipr.

The IETF invites any interested party to bring to its attention any copyrights, patents or patent applications, or other proprietary rights that may cover technology that may be required to implement this standard. Please address the information to the IETF at ietf-ipr@ietf.org.

Acknowledgment

Funding for the RFC Editor function is provided by the IETF Administrative Support Activity (IASA).