SAM Research Group
Internet-Draft
Intended Status: Informational
Expires: August 18, 2008

J. Buford
Avaya Labs

February 25, 2008

Hybrid Overlay Multicast Framework
draft-irtf-sam-hybrid-overlay-framework-02

Status of this Memo

Abstract

We describe an experimental framework for constructing SAM sessions using hybrid combinations of Application Layer Multicast, native multicast, and multicast tunnels. We leverage AMT relay and gateway elements for interoperation between native regions and ALM regions. The framework allows different overlay algorithms and different ALM control algorithms to be used.

Conventions used in this document

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 [1].

Table of Contents

1. Introduction

The concept of scalable adaptive multicast [BUF2007] includes both scaling properties and adaptability properties. Scalability is intended to cover:

- large group size
- large numbers of small groups
- rate of group membership change
- admission control for QoS
- use with network layer QoS mechanisms
- varying degrees of reliability
- trees connect nodes over global internet

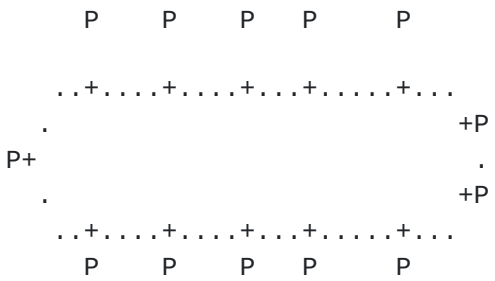
Adaptability includes

- use of different control mechanisms for different multicast trees depending on initial application parameters or application class
- changing multicast tree structure depending on changes in application requirements, network conditions, and membership
- use of different control mechanisms and tree structure in different regions of network depending on native multicast support, network characteristics, and node behavior

In this document we describe an experimental framework for constructing SAM sessions using hybrid combinations of Application Layer Multicast, native multicast, and multicast tunnels.

2. Definitions

2.1. Overlay Network



Overlay network - An application layer virtual or logical network in which end points are addressable and that provides connectivity, routing, and messaging between end points. Overlay networks are frequently used as a substrate for deploying new network services, or for providing a routing topology not available from the underlying physical network. Many peer-to-peer systems are overlay networks that run on top of the Internet.

In the above figure, "P" indicates overlay peers, and peers are connected in a logical address space. The links shown in the figure represent predecessor/successor links. Depending on the overlay routing model, additional or different links may be present.

2.2. Overlay Multicast

Overlay Multicast (OM): Hosts participating in a multicast session form an overlay network and utilize unicast connections among pairs of hosts for data dissemination. The hosts in overlay multicast exclusively handle group management, routing, and tree construction, without any support from Internet routers. This is also commonly known as Application Layer Multicast (ALM) or End System Multicast (ESM).

We call systems which use proxies connected in an overlay multicast backbone "proxied overlay multicast" or POM.

2.3. Peer

Peer: an autonomous end system that is connected to the physical network and participates in and contributes resources to overlay construction, routing and maintenance. Some peers may also perform additional roles such as connection relays, super nodes, NAT traversal, and data storage.

2.4. Multi-Destination Routing

Multi-Destination Routing (MDR): A type of multicast routing in which group member's addresses are explicitly listed in each packet transmitted from the sender [AGU1984]. XCAST [RFC5058] is an experimental MDR protocol. A hybrid host group and MDR design is described in [HE2005].

3. Assumptions

3.1. Overlay

Peers connect in a large-scale overlay, which may be used for a variety of peer-to-peer applications in addition to multicast sessions.

Peers may assume additional roles in the overlay beyond participation in the overlay and in multicast trees.

We assume a single structured overlay routing algorithm is used. Any of a variety of multi-hop, one-hop, or variable-hop overlay algorithms could be used.

Castro et al. [CAS2003] compared multi-hop overlays and found that tree-based construction in a single overlay out-performed using separate overlays for each multicast session. We use a single overlay rather than separate overlays per multicast sessions. We defer federated and hierarchical multi-overlay designs to later versions of this document.
Peers may be distributed throughout the network, in regions where native multicast (NM) is available as well as regions where it is not available.
An overlay multicast algorithm may leverage the overlay's mechanism for maintaining overlay state in the face of churn. For example, a peer may hold a number of DHT (Distributed Hash Table) entries. When the peer gracefully leaves the overlay, it transfers those entries to the nearest peer. When another peers joins which is closer to some of the entries than the current peer which holds those entries, than those entries are migrated. Overlay churn affects multicast trees as well; remedies include automatic migration of the tree state and automatic re-join operations for dislocated children nodes.

3.2. Overlay Multicast

The overlay supports concurrent multiple multicast trees. The limit on number of concurrent trees depends on peer and network resources

and is not an intrinsic property of the overlay. Some multicast trees will contain peers use ALM only, i.e., the peers do not have NM connectivity. Some multicast trees will contain peers with a combination of ALM and NM. Although the overlay could be used to form trees of NM-only peers, if such peers are all in the same region we expect native mechanisms to be used for such tree construction, and if such peers are in different regions we expect AMT to handle most cases of interest.

Peers are able to determine, through configuration or discovery:

- Can they connect to a NM router
- Is an AMT gateway accessible
- Can the peer support the AMT-GW functionality locally
- Is MDR supported in the region

3.3. NAT

Some peers in the overlay may be in a private address space and behind firewalls. We assume that mechanisms are available for the following, and that the mechanisms scale as the ratio of NATed peers to public address (public) peers grows, to a limit.

- Connectivity establishment between NATed peers and public peers
- Routing of overlay control messages to/from NATed and public peers.
- Routing of data messages over the topology of the tree

NAT traversal solutions developed elsewhere in IETF will be used, and new NAT traversal mechanisms are out of scope to this framework.

3.4. Regions

A region is a contiguous internetwork such that if native multicast is available, all routers and end systems can connect to native multicast groups available in that region.

A region may include end systems.

3.5. AMT

We use AMT [THA2007] to connect peers in ALM region with peers in NM region. AMT permits AMT-R and AMT-GW functionality to be embedded in

hosts or specially configured routers. We assume AMT-R and AMT-GW can be implemented in peers.

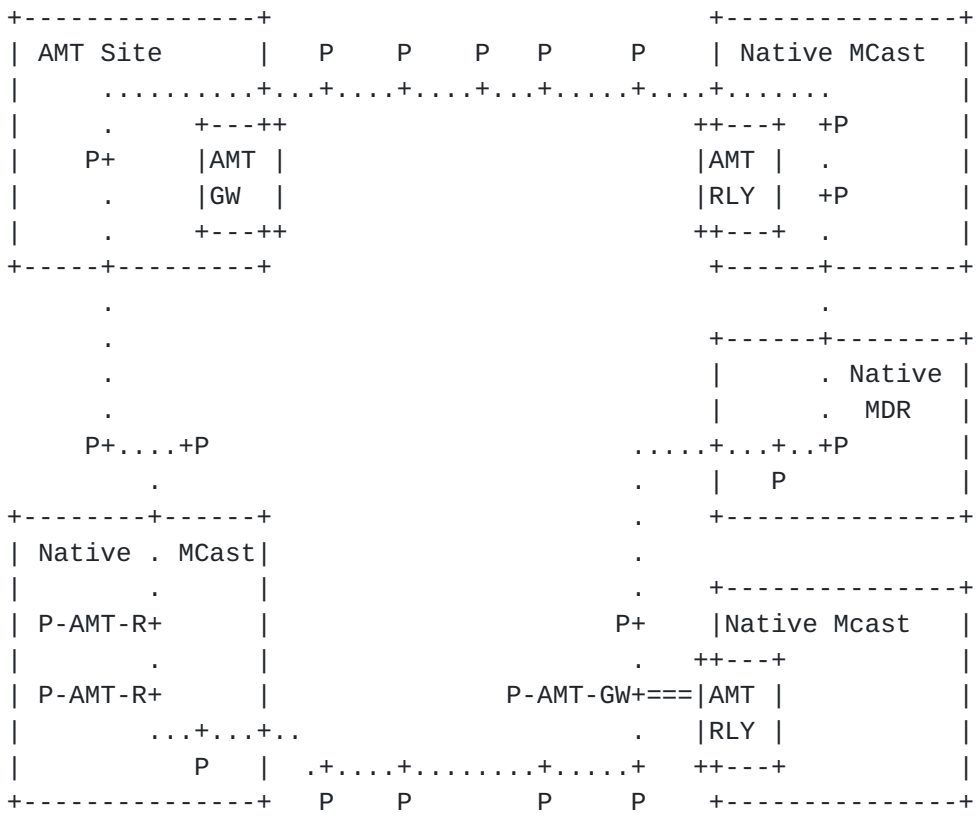
AMT has certain restrictions: 1) isolated sites/hosts can receive SSM, 2) isolated non-NAT sites/hosts can send SSM, 3) isolated sites/hosts can receive general multicast. AMT does not permit isolated sites/hosts to send general multicast.

4. ALM Tree Operations

Peers use the overlay to support ALM operations such as:

- Create tree
- Join
- Leave
- Re-Form or optimize tree

There are a variety of algorithms for peers to form multicast trees in the overlay. We permit multiple such algorithms to be supported in the overlay, since different algorithms may be more suitable for certain application requirements, and since we wish to support experimentation. Therefore, overlay messaging corresponding to the set of overlay multicast operations must carry algorithm identification information.
For example, for small groups, the join point might be directly assigned by the rendezvous point, while for large trees the join request might be propagated down the tree with candidate parents forwarding their position directly to the new node.
In addition to these overlay level tree operations, some peers may implement additional operations to map tree operations to native multicast and/or AMT [THA2007] connections.



5. Hybrid Connectivity

In the above figure we show the hybrid architecture in six regions of the network. All peers are connected in an overlay, and the figure shows the predecessor/successor links between peers. The peers may have other connections in the overlay.

- No native multicast: Peers (P) in this region connect to the overlay

- Native multicast (NM) with a local AMT gateway (AMT GW). There are one or more peers (P) connected to the overlay in this region.

- Native multicast with a local AMT relay (AMT RLY). There are one or more peers (P) connected to the overlay in this region.

- Native multicast with one or more peers which emulate the AMT relay behavior (P-AMT-R) which also connect to the overlay. There may be other peers (P) which also connect to the overlay.

- Native MDR is a native multicast region using multi-destination routing, in which one or more peers reside in the region.

- Native multicast with no peers that connect to the overlay, but for which there is at least one peer in the unicast-only part of the network which can behave as an AMT-GW (P-AMT-GW) to connect to multicast sources through an AMT-R for that region. It may be feasible to also allow non-peer hosts in such a region to participate as receivers of overlay multicast; for this version, we prefer to require all hosts to join the overlay as peers.

6. Scenarios

6.1. ALM-Only Tree - Algorithm 1

Here is a simplistic algorithm for forming a multicast tree in the overlay. Its main advantage is use of the overlay routing mechanism for routing both control and data messages. The group creator doesn't have to be the root of the tree or even in the tree. It doesn't consider per node load, admission control, or alternative paths.

As stated earlier, multiple algorithms will co-exist in the overlay.

1. Peer which initiates multicast group:

```
groupID = create();  // allocate a unique groupId

                     // the root is the nearest peer in the overlay

// out of band advertisement/distribution of groupID, perhaps by
   publishing in DHT
```

2. Any joining peer:

```
// out of band discovery of groupID, perhaps by lookup in DHT

joinTree(groupID); // sends "join groupID" message
```

The overlay routes the join request using the overlay routing mechanism toward the peer with the nearest id to the groupID. This peer is the root. Peers on the path to the root join the tree as forwarding points.

3. Leave Tree:

   leaveTree(groupID) // removes this node from the tree

   Propagates a leave message to each child node and to the parent node. If the parent node is a forwarding node and this is its last child, then it propagates a leave message to its parent. A child node receiving a leave message from a parent sends a join message to the groupID.

4. Message forwarding:

   multicastMsg(groupID, msg);

- SSM tree - The creator of the tree is the source. It sends data messages to the tree root which are forwarded down the tree.

- ASM tree - A node sending a data message sends the message to its parent and its children. Each node receiving a data message from one edge forwards it to remaining tree edges it is connected to.

6.2. ALM tree with peer at AMT site (AMT-GW)

The joining peer connects to the tree using the ALM protocol, or, if the tree includes a peer in an NM region, then the peer can use the AMT GW to connect to the NM peer through the AMT relay. The peer can choose the delivery path based on latency and throughput.

If the peer is not a joining peer and is on the overlay path of a join request:

- If its next hop is a peer in an NM region with AMT-R, then it can select either overlay routed multicast messages or AMT delivered multicast messages.

- If its next hop is a peer outside of an NM region, then it could use either ALM only or use AMT delivery as an alternate path

6.3. ALM tree with NM peer using AMT-R

There are these cases:

- There is no peer in the tree which has an AMT-GW

  The NM peer uses ALM routing

- There is at least one peer in the tree which can function as P-AMT-GW

  The NM peer can join the tree using ALM routing and/or connecting to the P-AMT-GW.

- There is at least one peer in the tree which is in an AMT-GW region

  The NM peer can join the tree using ALM routing and/or connecting to the AMT-GW.

6.4. ALM tree with NM peer with P-AMT-R

Either the NM peer supports P-AMT-R or another peer in the multcast tree in the same region is P-AMT-R capable.

The three cases above apply here, replacing AMT-R with P-AMT-R.

6.5. Mixed Region Scenarios

In version 2 of this document we elaborate on:

- ALM tree topology vs NM topology and NM-ALM edges
- Single NM-ALM edge nodes vs multi NM peers from same region in the tree
- Initial tree membership is ALM vs initial tree membership is NM

For ALM tree topology vs NM topology, all peers belong to the overlay, but only P-ALM peers use overlay routing for multicast data transmission. As a default behavior, a P-NM peer should generally prefer to join the tree via an AMT-GW node. But there may be special cases (small trees, short multicast sessions, trees where most of the members are known to be P-ALM) in which the peer can override this to specify an ALM-only join. A P-NM peer may also accept P-ALM children which don't use the AMT tunnel path to participate in the multicast tree.

Consider 3 types of tree links: P-ALM to P-ALM, P-NM to P-NM and P-ALM to/from P-NM:

- P-ALM to P-ALM - This is a normal ALM tree path with management strictly in the overlay

- P-NM to P-NM - If the peers are in the same region, then the data path use native multicast capability in that region, and control occurs in ALM layer for ALM tree coordination and NM layer for native multicast purposes. If the peers are in different NM regions, then, if AMT gateways are available and configured to support an AMT tunnel between the regions, a tunnel is created using the AMT protocol (or already exists for this multicast group). The peers connect to their respective AMT gateways using the AMT procedure.

- P-ALM to/from P-NM - The connection can be either ALM or AMT tunnel depending on the context.

We expect two new functions are needed to build hybrid trees:

- joinViaAMTGateway(peer, AMT-GW, group_id) where 'Peer' is the peer requesting to join the ALM group identified by group_id, and AMT-GW is the ip address of the AMT gateway that the peer uses in its native multicast region. Request is transmitted to one or more parent peer candiates and/or rendezvous peers for the specified group id, according to the usual join protocol in this overlay. If the parent peer is a P-AMT-GW, then a tunnel is formed using the AMT protocol from the P-AMT-GW to the specified AMT-GW. If parent peer is a peer P-NM in native multicast region, then the tunnel is created between P-NM's AMT-GW and the specified AMT-GW, using the AMT protocol. If parent peer is a P-ALM, then the requested is propagated to other peers in the tree according to the join rules.

- leaveViaAMTGateway(peer, AMT-GW, group_id)where 'Peer' is the peer requesting to leave the ALM group identified by group_id, and AMT-GW is the ip address of the AMT gateway that the peer uses in its native multicast region. Request is transmitted the parent peer which is associated with the AMT-GW or provides that role. If the parent peer is a P-AMT-GW, then it removes the child from its AMT children list and may tear down the AMT tunnel P-AMT-GW to the specified AMT-GW if no other children are using it. If parent peer is a peer P-NM in native multicast region, then the tunnel is created between P-NM's AMT-GW and the specified AMT-GW, using the AMT protocol.

Regarding initial tree membership being either P-NM or P-ALM node(s),

we expect the general case should be that hybrid tree formation is supported transparently regardless.

7. Open Issues and Further Work

- AMT [THA2007] has some restrictions on connecting isolated sites/hosts as SSM/ASM sources and receivers. Further analysis is needed to insure that OM data path is consistent with these constraints and whether additional operating restrictions between the overlay and AMT need be specified.

- For NM regions with no AMT support, specifics of how peers self-select as P-AMT-GW and P-AMT-RLY, and what additional behavior if any is needed beyond that specified in [THA2007].

- We expect that the evolution of this document will lead to protocol specification related to the interopation points of the hybrid interfaces of the network.

8. Security Considerations

Overlays are vulnerable to DOS and collusion attacks. We are not solving overlay security issues.
For this version we assume centralized peer authentication model similar to what is proposed for P2P-SIP.

9. IANA Considerations

This document has no actions for IANA.

10. References

10.1. Normative References

[1] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 199

[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.

[RFC5058] R. Boivie, N. Feldman , Y. Imai , W. Livens , D. Ooms, "Explicit Multicast (Xcast) Concepts and Options", IETF RFC 5058. November 2007.

10.2. Informative References

[AGU1984] L. Aguilar, Datagram Routing for Internet Multicasting, Sigcomm 84, March 1984.

[BUF2007] J. Buford, S. Kadadi. SAM Problem Statement. Dec 2006. Internet Draft draft-irtf-sam-problem-statement-01.txt, work in progress.

[CAS2002] M. Castro, P. Druschel, A.-M. Kermarrec, An. Rowstron, Scribe: A large-scale and decentralized application-level multicast infrastructure IEEE Journal on Selected Areas in Communications, Vol.20, No.8. October 2002.

[CAS2003] M. Castro, M. Jones, A. Kermarrec, A. Rowstron, M. Theimer, H. Wang and A. Wolman, "An Evaluation of Scalable Application-level Multicast Built Using Peer-to-peer overlays," in Proceedings of IEEE INFOCOM 2003, April 2003.

[HE2005] Q. He, M. Ammar. Dynamic Host-Group/Multi-Destination Routing for Multicast Sessions. J. of Telecommunication Systems, vol. 28, pp. 409-433, 2005.

[MUR2006] E. Muramoto, Y. Imai, N. Kawaguchi. Requirements for Scalable Adaptive Multicast Framework in Non-GIG Networks. November 2006. Internet Draft draft-muramoto-irtf-sam-generic-require-01.txt, work in progress.

[THA2007] D. Thale, M. Talwar, A. Aggarwal, L. Vicisano, T. Pusateri. Automatic IP Multicast Without Explicit Tunnels (AMT). Internet Draft draft-ietf-mboned-auto-multicast-08, Work in progress. Oct. 2007.

Author's Address

John Buford
Avaya Labs
307 Middletown-Lincroft Road
Lincroft, NJ 07738
USA

Email: buford@samrg.org

Full Copyright Statement

Intellectual Property

Acknowledgment

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