MBONED Working Group
Internet Draft

Hugh LaMaster
Steve Shultz
NASA ARC/NREN
John Meylor
David Meyer
Cisco Systems

Category
draft-ietf-mboned-mix-01.txt

Informational
June, 1999

Multicast-Friendly Internet Exchange (MIX)

1. Status of this Memo

This document is an Internet-Draft and is in full conformance with all provisions of Section 10 of RFC 2026.

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.

2. Abstract

This document describes an architecture for a Multicast-friendly Internet eXchange (MIX), and the actual implementation at the NASA Ames Research Center Federal Internet eXchange (FIX-West, or FIX). The MIX has three objectives: native IP multicast routing, scalable interdomain policy-based route exchange, and to allow a variety of IGP protocols and topologies for intra-domain use. In support of these objectives, the MIX architecture defines the following components: a peer-peer routing protocol, a method for multicast forwarding, a method for exchanging information about active sources, and a medium which provides native multicast. This document describes the protocols and configurations necessary to provide a current, working multicast-friendly internet exchange, or MIX.

This memo is a product of the MBONE Deployment Working Group (MBONED) in the Operations and Management Area of the Internet Engineering Task Force. Submit comments to <mboned@ns.uoregon.edu> or the authors.

Acknowledgments

Thanks to the NASA HPCC program for supporting the NREN staff portion of this project; thanks to William P. Jones of the NASA ARC Gateway Facility for making the gateway facility available for housing this project.

3. Copyright Notice

4. Introduction

The MIX objective was to use current technology to implement a scalable, high-performance, efficient, native IP multicast architecture.

Past experience at ARC, NASA WANs, and at FIX-West, had shown that mrouted/DVMRP "Mbone" tunnels were an inefficient of routing multicast through an exchange point. Specifically, at FIX-West, the large number of tunnels often resulted in unicast traffic loads on the FIX FDDI that were 10 times the underlying multicast load. In addition, some WANs had multiple tunnels criss-crossing the same physical links, resulting in wasted WAN bandwidth. And, the separate workstation and router infrastructure for the "Mbone" tunnels created numerous problems. Maintenance of Unix system and tunnel

configurations was often ad hoc, because some of the network operators lacked the necessary expertise. And the hardware and software configuration and performance of the tunnel infrastructure was often out of step with the underlying router-based unicast structure. In addition, use of a single, shared, distance-vector IGP in the inter-domain space led to instability.

Therefore, it was desired to implement a new multicast internet exchange from the ground up, using current technology, and significantly improving performance, efficiency, and reliability.

Four elements were identified as being necessary for the MIX architecture in order to meet the objectives. These were to define a peer-peer routing protocol, a method for multicast forwarding, a method for exchanging information about distant sources and groups, and a non-switched broadcast medium.

NASA Ames Research Center hosts the Federal Internet eXchange (FIX-West, or, "the FIX") as well as hosting the Ames Internet eXchange (AIX), which is connected at high speed to the MAE-West, and, which also shares the same address space as the MAE-West. These facilities are co-located at the Ames Telecommunications Gateway Facility. It was felt that this would be an excellent location to test the viability of the native multicast technologies. The Multicast-friendly Internet eXchange (MIX) is co-located adjacent to the FIX for easy access from the existing FIX routers.

Choices were made for each element, and the MIX was implemented adjacent to the existing NASA ARC FIX gateway facility. At the time of writing, there are eight direct participants in the MIX, peering and exchanging routes and multicast traffic natively, and the performance and reliability have already far exceeded the tunneled infrastructure the MIX replaced.

5. Requirements and Technology

In order to meet the objectives for this multicast exchange, all peering partners had to agree mutually to standardize on the following four elements. These are:

- the protocol to be used for multicast route exchange
- the method for performing multicast forwarding
- the method for identifying active sources
- the physical medium for the multicast exchange

The elements chosen to implement the MIX were BGP4+ (also known as "MBGP") for routing and route exchange [BGP4+], PIM-SM for multicast

forwarding on the exchange, the MSDP protocol for information on sources and groups, and, FDDI for the multicast medium.

5.1. Routing

Two of the objectives of the MIX were to provide an EGP for scalable interdomain policy-based route exchange, and to allow a variety of IGP protocols and topologies for intra-domain use. As with unicast interdomain routing, BGP could be used as the EGP to exchange routes for multicast. However, the unicast and multicast routing paths and policies would have to be completely congruent. In practice, this is sometimes not the case. It is possible, however, to take advantage of the extensions in BGP4+ to deal with these policy and path incongruencies.

BGP4+ [BGP4+] describes extensions to (unicast) BGP that allow use of the existing BGP machinery to provide the necessary scalability, policy control, and route stability features and mechanisms to be applied to both unicast and multicast routes consistently.

BGP4+ allows routes to be marked "unicast forwarding", "multicast forwarding", or "both unicast and multicast forwarding". In this way, BGP4+ supports different multicast and unicast forwarding paths and policies. This removes the dependency on unicast-only routing.

The ability of BGP4+ to support separate paths and policies for multicast is important for meeting the objectives of the exchange in various ways. It allows for a participant's multicast routing policy to be independent of its established unicast routing policy. This is important in order that the exchange can support providers migrating to BGP4+ as an IDMR. This is because it allows for the exchange of routes previously exchanged via DVMRP, even though those routes would not meet the existing unicast routing policy. It allows for different policy in the interim. For example, routes may be exchanged for BGP4+ multicast forwarding even though they would not be permitted under existing unicast routing policy. BGP4+ also provides for the possibility that even after full migration is complete, a separate multicast routing policy can be applied.

The exchange architecture imposes no requirements on the IGP or the multicast forwarding protocol or topology used internal to an AS.

5.2. Multicast Forwarding

The first requirement for the multicast forwarding protocol is that it be able to use routes exchanged via BGP4+. In addition, there is a requirement that the protocol only forward data upon explicit joins from participating peers. For these reasons, PIM-SM was selected.

The use of PIM on a shared LAN has certain consequences. It is necessary for all MIX participants to agree on certain configuration conventions affecting PIM forwarding on multi-access LANs. In particular, it is necessary to establish a standard protocol "metric preference" (also known as "distance" or process "precedence") to be used by all peers for the PIM Assert process, because the PIM Assert process [PIM-SM] uses the "metric preference" [PIM-SM] as a mechanism by which the multicast forwarder is chosen. If all parties are not following the convention, there may be black holes, in which a route appears to be valid, but traffic does not flow, or, there may be multicast loops, which can have deleterious consequences.

For the MIX, a standard set of metric preferences are applied to the BGP4+ routes as the convention for the PIM forwarding mechanism.

5.3. Active Sources

There are two current methods for distributing information about active sources to participating AS's. The AS's may be dense-mode regions, or, they may contain PIM-SM RP's. One method is to use dense-mode to flood data packets to dense-mode regions and to sparse-mode RPs co-located on the exchange. The second method is to use a protocol that allows each AS to share information about the sources contained within it without flooding data.

Recently, a new protocol, MSDP [MSDP] has been proposed that, when combined with PIM-SM, allows independent AS's to share information about distant sources and groups without flooding. Instead of flooding all data, only <S,G> information is flooded, and then, only to systems, such as PIM-SM RP's, which require the information. MSDP allows each AS to run its own sparse-mode region, independent of all other sparse-mode regions.

MSDP peering is established by all MIX participants. Most MIX-connected AS's are now running sparse-mode internally, or are actively migrating.

5.4. Medium

A primary objective for a multicast exchange is to provide support for native multicast among multiple peering partners.

There exist a number of unresolved issues regarding use of layer-2 switched media for multicast at interexchange points, and, until these issues are resolved, running native multicast on such media can be problematic.

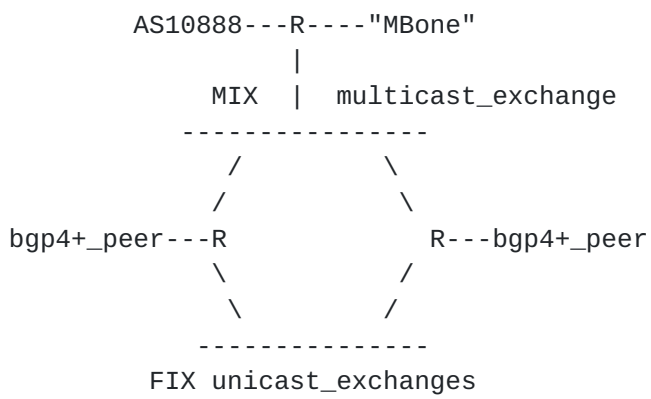
Fortunately, BGP4+ permits unicast and multicast to be carried on different media, permitting a multicast medium to be used independently of the unicast medium if necessary.

A FDDI concentrator was selected for the Ames MIX to provide the native multicast exchange medium. It was router-efficient, because it permitted the medium to do the multicast packet replication, with a single copy from a router being replicated to all neighbors. And FDDI was considered operationally convenient by most of the participants. Unicast traffic continues to be routed over the existing unicast exchange media.

Any medium which supports native multicast at layer 2 can be used efficiently. Other multicast exchanges are using switches with fast and gigabit ethernet ports and the switch configured with multicast as broadcast. Use of switching technology to handle layer 3 multicast requirements for inter-router communication is still unresolved. At some exchanges, native multicast is handled over ATM point-point VC's.

6. The NASA Ames Research Center Multicast-Friendly Internet Exchange

The Ames Multicast-friendly Internet eXchange, or MIX, began with the first beta-test trials in March 1998, and became operational, exchanging BGP4+ routes externally and using BGP4+ between multiple AS's, in May 1998. NREN implemented BGP4+ and internal BGP4+ and began trial external peerings in the same time frame, evolving from the first trials, to full deployment by October. As of June 1999, there were 10 AS's peering using PIM-SM, BGP4+, and MSDP to actively exchange multicast on the MIX FDDI. One of the AS's, AS10888, acts as a gateway between the DVMRP-based "Mbone" and the BGP4+ area. A router within AS10888 has been located on the MIX by NREN. The physical and logical topologies are as follows:



7. Topology, Architecture, and Special Considerations

BGP4+

-PIM Asserts and Metric preference

The PIM Assert mechanism requires that all routing protocols "compete" to see which router is allowed for forward onto the shared medium. To first order, the protocol metric preference is used to determine the forwarder. All MIX peers must coordinate routing protocol parameters so that one router does not inadvertantly win PIM asserts over a neighbor which has a functional path. This requires that BGP4+ routes have preference over other routes, such as BGP, OSPF, and DVMRP. In particular, it was necessary to standardize protocol metric preferences, and give BGP4+ routes the lowest, preferred, dynamic routing protocol metric preferences. For this reason, the standard set of BGP4+ metric preferences was chosen to be less than any other dynamic unicast routing protocol metric preferences. Any MIX routers which are using DVMRP must use a DVMRP metric preference higher than the BGP4+ metric preferences, rather than what many people have used previously as the DVMRP metric preference, of 0.

-Default

One transitional requirement is the necessity to have routes to "Mbone" sources, that is, sources within the global DVMRP routing region. Currently, the mechanism used is to have a single router in AS10888 on the MIX originate MBGP default to all external peers.

DVMRP routing

-DVMRP route redistribution

At present, all BGP4+ routes tagged with a particular community are redistributed at the MIX into DVMRP within AS10888. This is to provide DVMRP region users access to sources originating within AS's that are being routed via BGP4+ exclusively. Unless a particular community string is set, it is assumed that redistribution is not desired. In the reverse direction, instead of sending DVMRP routes into BGP4+, BGP4+ default is originated from the intermediary router.

In addition, local, stub-region DVMRP routes are redistributed into BGP4+ internally by several of the peers. As long as the regions remain stub regions, there is no danger, but, the possibility of a backdoor into the Mbone presents an ever-present threat of loops unless care is taken to redistribute only the routes which are known to be owned within the AS.

8. Conclusions and Recommendations

-Provide support for native multicast
-Use BGP4+ as a method of exchanging routes for inter-domain multicast
-Use PIM-SM with MSDP
-Concurrent use of BGP4+ and DVMRP for inter-domain routing is not recommended. It is strongly recommended to use BGP4+ exclusively for inter-domain route exchange.

9. Security Considerations

There are no security considerations unique to the multicast exchange.

10. References

[DVMRP] T. Pusateri, "Distance Vector Multicast Routing Protocol", <draft-ietf-idmr-dvmrp-v3-07.txt>, August 1998.

[BGP4+] T. Bates, R. Chandra, D. Katz, Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 2283, February 1998.

[BGP4+2] T. Bates, R. Chandra, D. Katz, Y. Rekhter, "Multiprotocol Extensions for BGP-4", Internet Draft, <draft-ietf-idr-bgp4-multiprotocol-v2-01.txt>, August 1998.

[PIM-SM] D. Estrin, D. Farinacci, A. Helmy, D. Thaler, S. Deering, M. Handley, V. Jacobson, C. Liu, P. Sharma, L. Wei, "Protocol Independent Multicast-Sparse Mode (PIM-SM): Protocol Specification", RFC 2362, June 1998.

[PIM-DM] S. Deering, D. Estrin, D. Farinacci, V. Jacobson, A. Helmy, D. Meyer, L. Wei, "Protocol Independent Multicast Version 2 Dense Mode Specification", Internet Draft, <draft-ietf-pim-v2-dm-01.txt>, November 1998.

[MSDP] D. Farinacci, Y. Rekhter, P. Lothberg, H. Kilmer, J. Hall, "Multicast Source Discovery Protocol (MSDP)", <draft-farinacci-msdp-00.txt>, June 1998.

11. Author's Address

Hugh LaMaster
Steve Shultz
NASA Ames Research Center
Mail Stop 233-21
Moffett Field, CA 94035-1000
email: hlamaster@arc.nasa.gov
shultz@arc.nasa.gov

David Meyer
John Meylor
Cisco Systems
San Jose, CA
email: dmm@cisco.com

jmeylor@cisco.com