Internet Draft

Document:
draft-ietf-mboned-iesg-gap-analysis-00.txt

Expires:
January 2003

Editors: D. Meyer
Sprint

B. Nickless
Argonne National Laboratory

July 2002

Internet Multicast Gap Analysis from the MBONED Working Group for the IESG

1. Status of this Memo

This document is an Internet-Draft and is in full conformance with all provisions of Section 10 of RFC2026.

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

An overview of IP multicast as deployed in the Internet today, from the perspective of the MBONED working group. Existing infrastructure is examined critically. Suggestions for possible improvement of the overall architecture are presented for the IESG.

3. Table of Contents

1. Status of this Memo.....1
2. Abstract.....1
4. Overview and Background.....2
5. Conventions used in this document.....2

6. RFC 1112.....3
7. Source-Specific Multicast.....3
8. Host Extensions for IP Multicast.....3

9. Mapping of Multicast Group Addresses to Ethernet MAC Addresses..4
10. Local Subnet Receiver Interest Protocol (IGMP).....4
11. Collision-Sense Media Access Sender Model.....4
12. Multicast Gateways.....5
13. Dense Mode Internet Multicast Routing.....5
14. Reachability Protocol Independent Multicast Routing.....6
15. Sparse Mode Internet Multicast Routing.....7
16. Mixed Dense/Sparse Mode Internet Multicast Routing.....7
17. Bursty Sources vs. Sparse Mode Forwarding State Maintenance....8
18. Co-mingled Source Knowledge and Packet Forwarding.....8
19. Co-mingled IP and Ethernet Routing.....9
20. Inter-Domain IP Multicast Exchange Points.....9
21. IP Multicast Architectural Gaps.....11
22. Recommendations from MBONED to IESG.....11
23. Acknowledgements.....13
24. Security Considerations.....13
25. References.....14
26. EditorsÆ Addresses.....14

4. Overview and Background

At the IETF-54 meeting, the MBONED working group recommended that the MSDP working group publish their current work as an Informational RFC and shut down. Some participants in the MBONED and MSDP working groups believed that the recurring discussions about the operation of MSDP were proxy arguments about the IP Multicast service model, and how that model can be supported in the Internet. Participants came to rough consensus that the best place for these overall service model and deployment questions is the MBONED working group.

A two phase approach was adopted. The short-term objective is to document existing MSDP implementations and deployments. A longer-term objective for the MBONED working group is to perform a ôgap analysisö of the existing IP multicast service model, protocols, and deployment.

This document represents that ôgap analysis,ö and is intended as advice to IESG. The MBONED participants hope the IESG will consider this advice in the context of IESG guidance for further IP multicast protocol development and deployment work.

5. 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].

6. RFC 1112

The seminal specification for IP Multicast is RFC 1112. It describes five elements required for IP Multicast on a local subnet: extensions to host software, an IPv4 address range reserved for group addresses, a method for mapping multicast group addresses to Ethernet Media Access Control (MAC) addresses, a protocol to discover receivers interested in packets addressed to a group (Internet Group Management Protocol), and a collision-sense multiple access (CSMA) method for multicast packet senders.

This model was inspired by Ethernet. The IEEE 802.3 Ethernet specification includes a bit in the MAC address format that indicates the frame may be intended for multiple receivers. Ethernet interfaces can be programmed to receive these multicast MAC addresses and forward them to the host for processing. On a traditional 10Base5 Ethernet, any station can put a frame on the wire, with the only requirement being to sense collisions and retransmit if necessary.

RFC 1112 also suggests that gateways may exist for moving IP multicast datagrams to other subnets with interested receivers.

The RFC 1112 service model has since become known as Any Source Multicast (ASM). When a receiver registers interest in a group, it will be delivered datagrams from any source that transmits datagrams addressed to that group.

7. Source-Specific Multicast

Just as early Ethernet controllers were programmable to only receive frames with certain MAC addresses, RFC 1112 and IGMP Version 2 only allowed IP Multicast receivers to elect to receive datagrams addressed to specific group addresses. Receivers could not select the sources participating in a group from which they would receive.

Later Ethernet controllers allowed more sophisticated filtering, including the capability of choosing from which senders the host

wished to accept frames. Similarly, Source-Specific Multicast (SSM) is an extension to the basic IP multicast model that allows receivers to select the source addresses from which to receive datagrams. IGMP Version 3 implements this service model extension for IPv4, and the Multicast Listener Discovery (MLD) protocol Version 2 implements this service model extension for IPv6.

8. Host Extensions for IP Multicast

Ultimately, user applications originate and accept IP multicast datagrams. RFC 1112 describes the extensions to various host software modules to support applications sending and receiving

datagrams. It also describes the operations a user application can perform to transmit and receive multicast datagrams.

9. Mapping of Multicast Group Addresses to Ethernet MAC Addresses

When preparing an IP datagram for transmission on an Ethernet, it is necessary for the host to specify the destination MAC address. (The source MAC address is typically constant, based on the IEEE-controlled address hard-wired into the Ethernet controller.) Before sending unicast datagrams, the Address Resolution Protocol (ARP) is generally used by a host to learn the destination MAC address for a given destination IPv4 address. RFC 2461 describes the equivalent Neighbor Discovery protocol procedure used for IPv6.

IPv4 multicast datagrams are not addressed to specific host addresses; instead, they are addressed to group addresses in the 224.0.0.0/4 range. Likewise, IPv6 multicast datagrams are addressed to group addresses in the FF00::/8 range.

RFC 1112 specifies a static 32:1 mapping from IPv4 multicast group addresses to Ethernet MAC addresses. One reason to define the 32:1 mapping was financial; to reserve enough Ethernet MAC addresses from the IEEE for a 1:1 mapping would have cost USD$16,000 in 1988. The 32:1 mapping reduced that cost to USD$1,000.

RFC 2464 (section 7) specifies a static 2^96:1 (that is, 79,228,162,514,264,337,593,543,950,336:1) mapping from IPv6 multicast group addresses to Ethernet MAC addresses. Note that it is impossible to determine the RFC 2375 scope directly from the Ethernet 802.3 MAC address, as the RFC 2464 mapping does not include the scope octet.

10. Local Subnet Receiver Interest Protocol (IGMP)

RFCs 1112 and 2236 define the Internet Group Management Protocol (IGMP) Version 2. IGMPv2 notifies the network of the interest of a host for datagrams addressed to a given group address. Again, this is analogous to a host notifying an Ethernet controller to accept frames with a given MAC address.

The IPv6 equivalent of IGMP is the Multicast Listener Discovery Protocol (MLD), originally defined in RFC 2710.

11. Collision-Sense Media Access Sender Model

Any host connected to a 10Base5 Ethernet can choose to transmit a frame at any time, subject only to the operation of the collision-sense media access (CSMA) protocol.

Given the static mapping of IPv4 group addresses to Ethernet MAC addresses, RFC 1112 specifies that an IPv4 datagram, addressed to a multicast group, can be transmitted by a host at any time.

Similarly, RFC 2464 implies that an IPv6 datagram addressed to a multicast group can be transmitted by a host at any time.

12. Multicast Gateways

RFC 1112 suggested that gateways might exist to pass multicast traffic between networks. Through the operation of the local subnet receiver interest protocol IGMP, these gateways can learn of the interest of receivers in multicast group datagrams.

As the technology for internetwork routing was unknown at the time of publication, RFC 1112 does not specify how that routing is to take place.

13. Dense Mode Internet Multicast Routing

Following the Ethernet broadcast model, the first scheme for routing IP multicast datagrams between Ethernets was for a multi-homed gateway to flood all multicast datagrams on each Ethernet to all other Ethernets. In other words, gateways become the IP multicast equivalent of Ethernet repeaters.

Ethernet bridges generally run the Spanning Tree Protocol (IEEE Specification 802.1d) to eliminate forwarding loops. Forwarding loops can also happen when there is more than one multi-homed gateway in an internetwork. The Distance Vector Multicast Routing Protocol (DVMRP) [RFC 1075] operates to eliminate forwarding loops. DVMRP, operating on a gateway, keeps track of the Reverse Path Forwarding (RPF) interface from which a multicast datagram with a given source address should arrive. If such multicast datagrams arrive on the appropriate RPF interfaces, they are replicated and flooded to all other interfaces by the gateway.

The effect of this procedure is to replicate all IP multicast datagrams transmitted by any host to all subnets. This limits the available bandwidth for all multicast traffic in the internetwork to that bandwidth available on the slowest link.

One optimization to this procedure is for gateways to use a receiver interest protocol such as IGMP to limit traffic flooded out an interface. Only if there are receivers interested in a group, on a network attached to a given interface, will the gateway flood datagrams addressed to that group out that interface. Of course, gateways are generally assumed by fellow gateways to be interested in all groups, so this optimization does not apply to networks with more than one gateway.

A second optimization further limits flooding. Consider a situation where a gateway has no interested receivers on all attached networks for a given group, yet receives an IP multicast datagram addressed to that group. DVMRP provides for gateways to send Prune messages out the appropriate RPF interface to notify fellow gateways that they have no interested receivers.

A third optimization provides for this limiting process to continue recursively. Once a gateway receives Prune messages from all other gateways on a network, and has no interested hosts, it stops forwarding messages out the attached interface. If all interfaces have such stoppages for a given group, it can generate its own Prune towards out the appropriate RPF interface to notify upstream gateways to stop sending datagrams addressed to a given group.

Through repeated application of this procedure, the distribution of multicast datagrams is limited only to the networks that have attached interested receivers, and to intermediate networks between sources and interested receivers. Multicast datagrams are distributed down a tree rooted at the source. Vertexes of the tree

are the gateways, and the edges of the tree are the networks connecting gateways.

This general strategy is known as ôdense modeö multicast routing.

As the number of multicast sources and receivers increase, the core of the multicast-enabled internetwork becomes more and more heavily loaded. Fewer opportunities for pruning occur.

As dense mode routing was experimentally deployed, a meta-stable failure mode was discovered. A gateway (or its attached network) can be overwhelmed with multicast traffic. Even though the gateway may have no interested receivers, it can fail to generate the required number of Prune messages. Unfortunately this failure mode can spread, because upstream gateways (closer to the sources) assume that packet replication and transmission is required, adding to their own load. Eventually the whole multicast internet collapses under the weight of un-Pruned traffic.

14. Reachability Protocol Independent Multicast Routing

In addition to controlling whether forwarding occurs (based on receiver interest), DVMRP maintains the topology of the forwarding trees from source to receivers. As its name implies, a distance-vector procedure similar to RIP is used.

Experience has shown that a distance-vector reachability protocol does not scale for large internets. Link-state protocols such as IS-IS and OSPF are generally used within administrative domains, and the Border Gateway Protocol is generally used between administrative domains. Convergence speed, policy flexibility, and other considerations motivate this diversity of reachability protocol use.

The Protocol Independent Multicast (PIM) multicast routing protocol takes its name from the fact that it can take its reachability information from any underlying reachability protocol. PIM concentrates on maintaining and controlling the multicast forwarding tree along the topology provided by whatever underlying reachability protocol(s) is/are used.

15. Sparse Mode Internet Multicast Routing

One way to eliminate the dense mode meta-stable failure mode is by adjusting the inter-gateway forwarding procedure to require

downstream gateways to explicitly request datagrams for a given group based on receiver interest.

In this regime, the forwarding tree is built from the interested receivers towards the source. Datagrams from the source are distributed back down the tree to the interested receivers.

Through IGMPv3 (or any similar SSM-style receiver interest discovery protocol) the receivers provide both pieces of information necessary for the internetwork to create and maintain the source-rooted forwarding tree: the IP addresses of the source and multicast group.

16. Mixed Dense/Sparse Mode Internet Multicast Routing

A straightforward sparse mode forwarding protocol alone cannot support the RFC 1112 Any Source Multicast service model. Although the receivers supply the group address for which theyÆre interested in receiving datagrams, the internetwork is responsible for identifying active sources so the source-rooted forwarding trees can be created.

The hybrid approach taken in RFC 2362 (PIM Sparse Mode Version 2) and MSDP is to flood the initial datagrams from any sender, typically under strict rate controls. When a gateway receives one of these flooded datagrams from a given sender, whose group address matches that of an attached interested receiver, the gateway grafts itself to the source-rooted forwarding tree for that sender.

So long as the source continues to transmit packets, the forwarding tree associated with the source is preserved. After a period of quiescence the forwarding tree is torn down.

The result is a dual-plane routing architecture. A dense-mode, rate-limited, flooding plane distributes datagrams from newly active sources. A sparse-mode, source-rooted tree based forwarding plane distributes and replicates datagrams from established sources.

17. Bursty Sources vs. Sparse Mode Forwarding State Maintenance

Prior to the development of the client-server-based World Wide Web, a session announcement protocol (SAP) was developed to allow interested parties to discover and participate in multilateral multimedia conference. Periodically a multicast datagram would be generated and sourced, providing the multicast group addresses,

media formats, and other such information needed for interested parties to join the conference.

As in any real-world application protocol, two factors required an engineering trade-off: the bandwidth consumed by the announcements vs. the frequency of announcements. Recall that early internetwork multicast routing used a dense mode approach; datagrams were flooded everywhere unless they were explicitly known to be unwanted. Given the propensity of the entire internetwork multicast infrastructure to collapse under load, great emphasis was placed on limiting the total bandwidth consumed by the announcements. Thus, the SAP announcement frequency was often measured in tens of minutes.

As sparse-mode multicast routing became more widely deployed, this tens-of-minutes frequency of SAP announcements became a problem. Each time an SAP announcement was sourced, the sparse-mode source-based distribution tree would be created towards interested receivers. But due to the low frequency of each independent announcement, the distribution tree would have been deemed quiescent and would be torn down.

The resulting ôbursty sourceö traffic would often follow only the dense-mode, rate-limited flooding routing plane. The sparse-mode, higher-performance forwarding plane would assume the source has gone quiescent long before the next ôburstö.

18. Co-mingled Source Knowledge and Packet Forwarding

ThereÆs a chicken-and-egg problem at the heart of internetwork multicast routing. On the one hand, experience has shown that a source-based distribution tree is the most efficient way to forward datagrams from a source to all interested listeners. On the other hand, such a source-based distribution tree cannot be created until the source is known, and RFC 1112 decrees that sources be able to transmit at any time without warning.

In other words, RFC 1112 defines an active source as a source that has placed a datagram on the wire. But by the time the datagram has been placed on the wire, itÆs too late to create a source-based distribution tree to all interested receivers.

There have been several approaches taken to resolve this problem.

The first was to not use source-based distribution trees at all. Unfortunately this approach resulted in an internetwork that would collapse under load.

The second approach has been to create dual routing planes: a dense-mode plane to forward the initial datagrams from a source, and as a side effect create a source-based distribution tree in the sparse-mode plane. Unfortunately these dual routing planes lead to a great deal of complexity.

The third approach has been SSM: put the burden of spreading the knowledge of active sources on the application rather than the network. This approach has two major drawbacks: first, it requires the replacement or upgrade of edge IEEE 802.x devices to support IGMPv3 snooping, along with a wholesale upgrade to host operating systems. Second, it requires applications to develop their own rendezvous mechanisms.

19. Co-mingled IP and Ethernet Routing

For primarily historical reasons, the IETF has pushed vendors of nominally IEEE 802.x compliant equipment to also become IPv4-aware enough to understand the IGMP Version 2 protocol.

IEEE has responded with the GARP/GMRP protocol suite, which are intended to allow 802.x hosts to control MAC-layer multicast replication and filtering. However, the IETF has continued work on IGMP and MLD while ignoring media-specific protocols like GARP/GMRP.

Arguably, this has marginalized IP multicast deployment, especially IPv6 multicast deployment. Only the very high-end IEEE 802.x devices have the sophistication to interpret IPv4/IGMP and IPv6/MLD datagrams.

20. Inter-Domain IP Multicast Exchange Points

Autonomous Systems often wish to exchange traffic. Exchange points have been developed to meet this demand. One popular type of exchange point is realized in an 802.x Ethernet switch. Each participating Autonomous System is provided an 802.x Ethernet port and an IP address on the exchange point network, to which the Autonomous System connects a router. Bilateral BGP sessions are then established between Autonomous Systems across the 802.x network fabric.

When an Autonomous System router wishes to deliver a unicast datagram to another Autonomous System router participating at such an exchange point, it follows this procedure:

- The datagramÆs IP Destination Address is compared to the Forwarding Information Base (FIB). The FIB returns a so-called ônext-hopö IP address. This next-hop address is generally assigned to another Autonomous SystemÆs router at the exchange

point.

- Through the operation of the Address Resolution Protocol (ARP), the 802.x Ethernet MAC address associated with the next-hop IP address is determined.

- An Ethernet frame is assembled with the destination MAC address set to the MAC address determined through ARP, and is transmitted to the Exchange Point 802.x Ethernet switch.

- The exchange point 802.x Ethernet switch examines the destination MAC address of the Ethernet frame. Based on that address, the exchange point switch delivers the Ethernet frame to the destination Autonomous SystemÆs router.

Consider an analogous procedure for multicast routing. The object is to graft the Autonomous SystemÆs router onto a source-rooted distribution tree across the exchange point. Here is one procedure that a downstream router can follow:

- The downstream router compares the source address to its Multicast Reachability Information Base (M-RIB). The M-RIB returns the IP address of an ôupstreamö router across the exchange point.

- Through the operation of the PIM Sparse Mode Protocol, the downstream router registers interest in that source and group addresses to the upstream router across the exchange point.

- Upon receipt of a matching datagram for the downstream router, the upstream router assembles an Ethernet frame and transmits it to the exchange point 802.x Ethernet switch. As per RFC 1112 or RFC 2464, the destination MAC address of the frame is statically derived from the destination group address of the datagram.

- The exchange point 802.x Ethernet switch examines the destination MAC address. As this MAC address is a multicast address, the 802.x Ethernet switch replicates this frame and sends it to all output ports.

This presents three problems:

First, the multicast traffic is needlessly replicated to all participants in the exchange point. In the unicast case above, the 802.x Ethernet exchange point switch could use the Ethernet destination MAC address to uniquely identify which port should receive a given frame. The static mapping of destination multicast

group address to Ethernet MAC addresses makes that determination impossible.

Second, the needlessly replicated multicast traffic can trigger the PIM Assert process, as per RFC 2362 Section 2.9. The PIM Assert process has been observed to override the policy decisions of downstream routers in exchange points.

Third, it is impossible for multicast traffic to pass through an exchange point more than once. Any given exchange point participant may not have a peering agreement with all other participants, requiring an intermediate hop through a transit Autonomous System participant. Due to the operation of ARP this is not a problem for unicast traffic, but due to the static mapping of multicast groups to (e.g.) Ethernet MAC addresses, this cannot work.

In summary, it is impossible to support IP multicast at an exchange point, when that exchange point is based on IEEE 802.3 Ethernet.

21. IP Multicast Architectural Gaps

In general, the IETF focus is on Internet protocols. IGMP snooping places the requirement of IPv4-awareness on IEEE-standardized 802.x Ethernet switches. The current drafts for IPv6 MLD seek to extend that requirement to include IPv6. The IESG would rightfully refuse to allow IETF working groups to impose such requirements on devices standardized by organizations outside the IETF (such as IEEE), but somehow has excepted the IP multicast work from this discipline.

The Internet is more complex than a simple CSMA-style Ethernet segment, where sources can transmit at any time. Experience clearly indicates that internetwork multicast datagram forwarding is most efficiently done by source-rooted distribution trees. Experience compels revisitation of the assumption that sources should be able to transmit at any time, yet receive the same level of service as that provided by a fully instantiated source-rooted distribution tree.

Registration of soon-to-be-active sources (along the lines of the unicast Address Resolution Protocol [ARP]) should be seriously considered.

Part of the registration of soon-to-be-active sources could include allocation of link-local media-specific multicast addresses, rather than relying solely on the static mappings defined in RFC 1112 and

RFC 2464.

The static mapping of IP multicast group addresses to media-specific multicast addresses (in particular, Ethernet) cannot operate properly at exchange points.

22. Recommendations from MBONED to IESG

The IESG should direct the MAGMA working group to develop a successor to IGMP/MLD.

- The successor should perform the receiver interest discovery functions of existing versions of IGMP/MLD, but in addition should support the registration of active sources.

- At least three modes of operation should be supported. As in IGMPv2/MLDv1, sources and receivers should be able to transmit to and receive from group addresses without respect to the identities of sources. In a second mode, analogous to IGMPv3/MLDv2, receivers should be able to select the sources from which they want to receive traffic for a particular group. A third mode should permit receivers to select the source address, group address, and upstream gateway from which to receive traffic.

- The successor should be media-agnostic. Media-specific multicast addresses should be treated as opaque handles. Examples of media-specific multicast addresses might include 802.x MAC addresses, ATM Forum NSAP point-to-multipoint addresses, etc.

- Adaptation layers for this successor protocol should be developed to use media-specific mechanisms for multicast transport and replication. For example, the IEEE 802.1p GARP/GMRP protocol should be used on Ethernet. ATM Forum UNI point-to-multipoint signaling should be used on ATM networks (c.f. RFC 2022).

- This successor protocol would provide for the dynamic assignment of media-specific addresses. As necessary, the media address assignment mechanism might control the creation and maintenance of media-specific intra-subnet distribution mechanisms, such as ATM point-to-multipoint switched virtual circuits. When in operation on an IEEE 802.3 Ethernet, this mechanism would supercede RFC 1112 Section 6.4 and RFC 2464 Section 7.

- The first application for this successor protocol would be at public internetwork exchange points. The third mode of operation (allowing receivers to select the source address, group address, and upstream gateway) would allow participants at exchange points to select their upstream neighbor towards a source based on explicit policy, rather than the vagaries of the PIM Assert mechanisms.

- This successor protocol may not be applicable for IP datagrams with TTL=1, so as to preserve semantics for link-local rendezvous (e.g. OSPF). Likewise, it may not be applicable for IPv6 RFC 2375 scopes 0, 1, and 2.

The IESG should encourage the development of a protocol to spread the knowledge of active sources to interested gateways. Given a successor to IGMP that supports the registration of active sources, this spreading of knowledge can happen independently of actual multicast datagram forwarding.

The IESG should discourage any further work on IGMP or MLD snooping, as implemented by otherwise nominally IEEE 802 compliant equipment.

Instead, the IESG should encourage the use of GARP/GMRP on IEEE 802 networks.

The IESG should guide protocols that use IP multicast to maintain a minimum frequency of datagram transmission, so as to preserve source-based forwarding trees.

23. Acknowledgements

This work was supported by the Mathematical, Information, and Computational Sciences Division subprogram of the Office of Advanced Scientific Computing Research, U.S. Department of Energy, under Contract W-31-109-Eng-38.

24. Security Considerations

Security considerations are not yet discussed in this draft memo.

25. References

Most of the references are called out in-line. This section will be completed more fully before final publication.

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

26. EditorsÆ Addresses

David Meyer Email: dmm@sprint.net

Bill Nickless
Argonne National Laboratory
9700 South Cass Avenue #221 Phone: +1 630 252 7390
Argonne, IL 60439 Email: nickless@mcs.anl.gov