Internet Draft
Document: draft-nickless-ipv4-mcast-bcp-01.txt
Expires: October 2001

B. Nickless
Argonne National Laboratory

April 2001

IPv4 Multicast Best Current Practice

Status of this Memo

Abstract

This document describes best current practices for IPv4 multicast deployment, both within and between PIM Domains and Autonomous Systems.

Table of Contents

Overview

Current best practice for IPv4 multicast service provision uses four different protocols: Internet Group Management Protcol, Protocol Independent Multicast (Sparse Mode), Border Gateway Protocol with multiprotocol extensions, and the Multicast Source Discovery Protocol. This document outlines how these protocols work together to provide end-to-end IPv4 multicast service. In addition, this document describes best current practices for configuring these protocols, individually and in combination.

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 [RFC2119].

Introduction and Terminology

IPv4 multicast [MCAST] is an internetwork service that allows IPv4 datagrams sent from a source to be delivered to more than one interested receiver. That is, a given source sends a packet the network with a destination address 224/4 CIDR [CIDR] range. The network transports this packet to all receivers (replicated where necessary) that have registered their interest in receiving these packets.

The letter S is used to represent the IPv4 address of a given source. The letter G is used to represent a given IPv4 group address (within the 224/4 CIDR range). A packet, or series of packets, sent by a sender with a given address S to a given group G is represented as (S,G). A set of packets sent to group G by multiple senders is represented as (*,G).

Any Source Multicast

Any Source Multicast (ASM) is the traditional IPv4 multicast [MCAST] model. IPv4 multicast sources send IPv4 datagrams to the network, with the destination address of each IPv4 datagram set to a specific ôgroupö address in the Class D address space (224/4). IPv4 multicast receivers register their interest in packets addressed to a group address, and the internetwork delivers packets from all sources in the internetwork to the interested receivers.

It is the responsibility of the internetwork to keep track of all the sources transmitting to a particular group, so that when a receiver wishes traffic sent to that group the network can forward traffic from all group sources.

IPv4 multicast receivers register their interest in packets sent to group addresses through the Internet Group Management Protocol Version 2 (IGMPv2) [IGMPV2]. IGMPv2 does not have any facility for receivers to specify which sources the receiver wants to receive from. That is, IGMPv2 only allows (*,G) registrations.

Source Specific Multicast

Source Specific Multicast (SSM) [SSM] is another IPv4 multicast model. IPv4 multicast sources send IPv4 datagrams to the network, with the destination address of each IPv4 datagram set to a specific ôgroupö address in the Class D address space (224/4). IPv4 multicast receivers register their interest in packets from a specific source that have been addressed to a group address, and the internetwork delivers packets from that source to the interested receivers.

It is the responsibility of each receiver to specify which sources, sending to which groups, the receiver wishes to receive datagrams from.

IPv4 multicast receivers register their interest in packets sent by specific sources to group addresses through the Internet Group Management Protocol Version 3 (IGMPv3) [IGMPV3]. That is, IGMPv3 supports (S,G) registrations.

Multiprotocol BGP

The topology of inter-domain IPv4 multicast forwarding is determined

by BGPv4 [BGPV4] policy, as is IPv4 unicast forwarding. BGP provides reachability information. Reachability information for IPv4 Unicast and IPv4 Multicast prefixes can be advertised separately. (See [MBGP] for details and the definition of Network Layer Reachability Information (NLRI) and Subsequent Address Family Information (SAFI).) The practical definition of reachability is different for IPv4 unicast (NLRI=unicast, SAFI=1) and IPv4 multicast (NLRI=Multicast, SAFI=2).

In current practice for BGP unicast advertisements (NLRI=Unicast, SAFI=1), reachability is interpreted to mean that IPv4 datagrams will be forwarded towards their destination host if sent to the NEXT_HOP address in the advertisement.

In the case of BGP multicast advertisements (NLRI=Multicast, SAFI=2), reachability is interpreted to mean two things:

First, IPv4 datagrams can be requested from sources within the advertised prefix range. Such requests are made to the advertised NEXT_HOP by means of the PIM Sparse Mode [PIM-SM] protocol, or (rarely) any other mutually agreed upon protocol that supports (S,G) requests.

Second, the MSDP [MSDP] speaker with the NEXT_HOP address will provide MSDP Source Active messages from PIM Rendevous Points within the advertised prefix range.

These two interpretations of BGP NLRI=Multicast flow from the original use of BGP to control Distance Vector Multicast Routing Protocol [DVMRP]. DVMRP is a ôdenseö routing protocol, which means traffic is flooded outwards from the sources to all possible receivers. In this situation, an IPv4 multicast router has to decide which incoming interface may accept IPv4 datagrams from a given source (to avoid forwarding loops). When the switch was made to use a ôsparseö forwarding model (requiring specific (S,G) requests for traffic to flow) both interpretations of BGP NLRI=Multicast became necessary for interoperability with the DVMRP-based model.

Note that while MSDP is not strictly necessary for Autonomous Systems that only support Source Specific Multicast [SSM], MSDP depends on the latter interpretation of BGP NLRI=Multicast to avoid MSDP SA forwarding loops. There is a real danger of causing MSDP SA forwarding ôblack holesö unless MSDP peerings are set up at the same time as BGP NLRI=Multicast peerings.

MBGP also supports combined multicast and unicast advertisements (SAFI=3). Current practice is to interpret these advertisements to include all three meanings listed above: unicast forwarding, availability of traffic from multicast sources, and MSDP Source Active availability.

PIM Sparse Mode

The PIM Sparse Mode protocol [PIM-SM] is widely used to create forwarding state from IPv4 multicast sources to interested receivers.

The term ôPIM Sparse Mode domainö generally refers to the hosts and routers that share a PIM Sparse Mode Rendezvous Point.

In current practice, there is generally one PIM Sparse Mode domain per Autonomous System. Some Autonomous Systems choose to have multiple PIM Sparse Mode domains for scalability reasons.

Within a PIM Sparse Mode domain, the standard PIM Sparse Mode mechanisms are used to build shared forwarding trees and source specific trees from IPv4 multicast sources to interested receivers. IPv4 multicast sources are registered with the PIM Rendezvous Point (RP). Interested IPv4 multicast receivers make their group interest known through the Internet Group Management Protocol, and the associated PIM Designated Router (DR) sends PIM Join messages towards the RP to build the appropriate forwarding trees.

In the ASM model, PIM Sparse Mode Rendezvous Points have to co-operate in order to discover active sources and set up forwarding trees. MSDP is used to spread the knowledge of active sources within a multicast group. Source-specific (S,G) joins are used to set up forwarding from sources towards the interested receivers. No inter-PIM-domain shared forwarding tree is created.

In the SSM model, there is no need for PIM Sparse Mode Rendezvous Points because each receiver explicitly identifies the sources from which it desires traffic. Thus, the local PIM Designated Router that receives an IGMPv3 request for traffic can initiate the PIM-Sparse Mode source-specific (S,G) requests directly towards the source.

Internet Group Management Protocol

The Internet Group Management Protocol was designed to be used by hosts to notify the network that the hosts want to receive traffic on an IPv4 multicast group.

The IGMP design originally assumed a shared media network like Ethernet. When layer 2 switches became available, many vendors built in IGMP ôsnoopingö so as to avoid flooding IP multicast

traffic to all ports in a Virtual Local Area Network (VLAN). The best current practice for IPv4 multicast deployment in a switched Local Area Network context is to use IGMP snooping to avoid unnecessary IPv4 multicast flooding.

IGMPv2 [IGMPV2] supports the ASM model. IGMPv3 [IGMPV3] supports the ASM model as well as the SSM model.

Some wide area network access servers support IGMP and IPv4 Multicast over PPP connections. Host implementations also support the IGMP over PPP connections, even those that use dial-up modems. Such support contributes to the availability and utility of IPv4 multicast service, but only when configured by network operators.

Multicast Source Discovery Protocol

Current best practice is for Autonomous Systems to ask each other for traffic from specific sources transmitting to specific groups. It follows that inter-AS IP multicast forwarding trees are all source-specific. Thus, when a receiver registers an interest in datagrams addressed to a multicast group G (generally through an IGMPv2 (*,G) join) it is necessary for the associated PIM Sparse Mode Rendezvous Point (or other intra-AS protocol element, such as a Core Based Trees [CBT] Core Router) to arrange (S,G) joins towards each sender. Each inter-AS (S,G) join creates a branch of the forwarding tree towards the sender.

The Multicast Source Discovery Protocol [MSDP] is used to communicate the availability of sources between Autonomous Systems. MSDP-speaking PIM Sparse Mode Rendezvous Points (or other designated MSDP speakers with knowledge of all sources within an Autonomous System) flood knowledge of active sources to each other.

Model IPv4 Multicast-Capable BGPv4 Configuration

IPv4 multicast reachability is communicated between Autonomous Systems by BGPv4 prefix announcements. That is, prefixes are advertised with NLRI=Multicast (SAFI in {2,3}). As outlined above, the semantics of a BGPv4 advertisement of an IPv4 NLRI=Multicast prefix are currently interpreted to mean two things:

First, such an advertisement means that the router with the NEXT_HOP address of that advertisement will supply packets from any transmitting source S whose address matches the prefix advertised. In order to fulfill this expectation, any two BGPv4 speakers that communicate NLRI=Multicast advertisements must be able to ask each

other for (S,G) traffic. That is, they must have some protocol (most often PIM Sparse Mode) configured between them.

Second, such an advertisement means that the router with the NEXT_HOP address of that advertisement will supply MSDP Source Active messages from any (e.g.) PIM Sparse Mode Rendezvous Point whose address matches the prefix advertised. To avoid MSDP ôblack holesö, Autonomous Systems with BGPv4 speakers that exchange NLRI=Multicast advertisements must also have appropriate MSDP peerings configured.

Model IPv4 Multicast Inter-domain PIM Sparse Mode Configuration

As outlined above, current practice is that each IPv4 BGPv4 NLRI=Multicast capable peering is capable of making (S,G) requests for traffic. Autonomous Systems predominantly use PIM Sparse Mode for this purpose. Whether PIM Sparse Mode is used or not, these peerings/adjacencies are configured in the following ways:

The minimum TTL Threshold for traffic crossing an Autonomous System peering is generally set to be 32. This value follows earlier practice [FAQ] that sets inter-institution TTL barriers at 16-32. It also provides a reasonable number of values both above and below the (maximum 255) barrier.

The PIM Sparse Mode Adjacency (or other inter-domain (S,G) request mechanism) should not make requests for traffic across the peering for sources in these groups:

224.0.1.39/32: CiscoÆs Rendezvous Point Announcement Protocol
224.0.1.40/32: CiscoÆs Rendezvous Point Discovery Protocol
239.0.0.0/8: Administratively Scoped IPv4 Group Addresses

The first two groups are used to determine where PIM Sparse Mode Rendezvous Points can be found within an Autonomous System. The latter group range is defined by RFC 2365 [RFC2365]. RFC 2365 has been generally interpreted to equate ôorganizationsö (see section 6.2) with Autonomous Systems. Some Autonomous Systems choose to interpret this differently.

Model PIM Sparse Mode Rendezvous Point Location

In order to participate in current-practice inter-Autonomous System IPv4 multicast routing, a PIM Sparse Mode Rendezvous Point (or other such MSDP-speaker) should have access to the full BGP NLRI=Multicast reachability table so as to arrange for (S,G) joins to the

appropriate external peer networks. This need arises when a (*,G) request comes in from a host. Access to the BGPv4 NLRI=Multicast reachability table is also important so that the (e.g.) PIM Sparse Mode Rendezvous Point will perform MSDP Reverse-Path-Forwarding (RPF) checks correctly.

PIM Sparse Mode Rendezvous Points are often located at the border router of an Autonomous System where the BGPv4 NLRI=Multicast reachability table is already maintained. If necessary, an MSDP Mesh Group can be created if there are multiple BGPv4 NLRI=Multicast speakers within an Autonomous System. (See Section 14.3 of [MSDP].)

The IPv4 address of each PIM Sparse Mode Rendezvous Point (or other such MSDP-speaker) must be chosen so that it is within an advertised BGPv4 NLRI=Multicast prefix. The MSDP RPF checks operate on the so-called ôRP-Addressö within the MSDP Source Active message, not the advertised source S. In the most widely deployed case, the RP-

Address is set by the MSDP-speaker to be the PIM Sparse Mode Rendezvous Point address.

Model MSDP Configuration Between Autonomous Systems

MSDP peerings are configured between Autonomous Systems. These peerings are statically defined. Thus, in practice, such MSDP-speaking (e.g.) PIM Sparse Mode Rendezvous Point(s) must be ôtied downö to known addresses and routers for the inter-AS peerings to operate correctly.

The so-called ôRP-addressö in MSDP Source Active messages must be addressed within prefixes announced by BGPv4 NLRI=Multicast advertisements. (Otherwise the RP-Address Reverse Path Forwarding checks done by peer MSDP-speaking Autonomous Systems will fail, and the MSDP Source Active messages will be discarded.) The most common RP-address in MSDP Source Active messages is the PIM Rendezvous Point IPv4 address.

In practice, MSDP speakers are configured to not advertise sources to external peers from the following groups. MSDP speakers are also configured to not accept source advertisements from external peers within the following groups:

224.0.1.2/32: SGI ôDogfightö game
224.0.1.3/32: RWHOD
224.0.1.22/32: SVRLOC
224.0.1.24/32: MICROSOFT-DS
224.0.1.35/32: SVRLOC-DA
224.0.1.39/32: CiscoÆs Rendezvous Point Announcement Protocol

```
224.0.1.40/32:     CiscoÆs Rendezvous Point Discovery Protocol
224.0.1.60/32:     HPÆs Device Discovery Protocol
224.0.2.2/32:      SunÆs Remote Procedure Call Protocol
229.55.150.208/32: Norton ôGhostö disk duplication software
232.0.0.0/8:       Source-Specific Multicast
239.0.0.0/8:       Administratively Scoped IPv4 Group Addresses
                   (with possible specific exceptions)
```

MSDP speakers are configured to not accept or advertise sources to or from external peers with Private Internet addresses [RFC1918].

MSDP-speakers are configured, wherever possible, to only advertise sources within prefixes that they are advertising as BGPv4 NLRI=Multicast (SAFI in {2,3}) announcements. That is, a non-transit Autonomous System would only advertise sources within the prefixes it advertises to its peers.

Based on recent events, MSDP peerings are configured with reasonable rate limits to dampen explosions of MSDP SA advertisements. These explosions can occur when malicious software generates packets addressed to many IPv4 multicast groups in a very short period of time. What ôappropriateö means for these rate limits will vary over time with the number of active IPv4 multicast sources in the

Internet. To determine an initial approximation for these rate limits, configure MSDP without rate limits initially, and then set the rate limits at some small multiple of the observed steady state rate. Another approach would be to set rate limits based on a small multiple of the current number of active sources in the Internet. The Mantra Project [MANTRA] maintains MSDP statistics, as well as other IPv4 multicast statistics.

Security Considerations

Autonomous Systems often configure router filters or firewall rules to discard mis-forwarded IPv4 datagrams. Such rules may explicitly list the IPv4 address ranges that are acceptable for incoming IPv4 datagrams. When IPv4 multicast is enabled, these rules need to be updated to disallow incoming IPv4 datagrams with addresses in the 239/8 CIDR range, and to allow incoming IPv4 datagrams with destination addresses in the 224/4 CIDR range.

PIM Sparse Mode Rendezvous Points are particularly vulnerable to Denial of Service attacks. As outlined above, it is important to put rate limits on MSDP peerings so as to protect your PIM Sparse Mode Rendezvous Points from explosions in the size of the cached MSDP Source Active table. Other denial of service attacks include

sending excessive Register-encapsulated packets towards the Rendezvous Point and flooding the Rendezvous Point with large numbers of IGMP joins.

Acknowledgements

Dino Farinacci created the (S,G) notation used throughout this document.

Marty Hoag, Simon Leinen, David Meyer, and Dave Thaler pointed out mistakes and made suggestions for improvement.

Marshall Eubanks described the vulnerability of PIM Sparse Mode Rendezvous Points to various denial of service attacks.

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.

References

[RFC2119] RFC 2119: Key Words for use in RFCs to Indicate Requirement Levels. S. Bradner. March 1997.

[MCAST] RFC 1112: Host extensions for IP multicasting. S.E. Deering. Aug-01-1989.

[CIDR] RFC 1519: Classless Inter-Domain Routing (CIDR): an Address Assignment and Aggregation Strategy. V. Fuller, T. Li, J. Yu, K. Varadhan. September 1993.

[IGMPV2] RFC 2232: Internet Group Management Protocol, Version 2. W. Fenner. November 1997.

[SSM] draft-holbrook-ssm-arch-02.txt: Source-Specific Multicast for IP. H. Holbrook, B. Cain. 1 March 2001.

[IGMPV3] draft-ietf-idmr-igmp-v3-07.txt: Internet Group Management Protocol, Version 3. B. Cain, S. Deering, B. Fenner, I Kouvelas, A. Thyagarajan. March 2001.

[BGPV4] RFC 1771: A Border Gateway Protocol 4 (BGP-4). Y. Rekhter, T. Li. March 1995.

[MBGP] RFC 2858: Multiprotocol Extensions for BGP-4. T. Bates, Y. Rekhter, R. Chandra, D. Katz. June 2000.

[PIM-SM] RFC 2117: Protocol Independent Multicast-Sparse Mode (PIM-SM): Protocol Specification. D. Estrin, D. Farinacci, A. Helmy, D. Thaler, S. Deering, M. Handley, V. Jacobson, C. Liu, P. Sharma, L. Wei. June 1997.

[MSDP] draft-ietf-msdp-spec-07.txt: Multicast Source Discovery Protocol (MSDP). D. Meyer (Editor). March 2001.

[DVMRP] RFC 1075: Distance Vector Multicast Routing Protocol. D. Waitzman, C. Partridge, S.E. Deering. November 1988.

[FAQ] http://netlab.gmu.edu/mbone_installation.htm

[RFC2365] RFC 2365: Administratively Scoped IP Multicast. D. Meyer. July 1998.

[RFC1918] RFC 1918: Address Allocation for Private Internets. Y. Rekhter, B. Moskowitz, D. Karrenberk, G. J. de Groot, E. Lear. February 1996.

[MANTRA] http://www.caida.org/tools/measurement/mantra

Author's Address

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