Network Working Group S. Brim
Request for Comments: DRAFT Cornell University
Y. Rekhter
T.J. Watson Research Center, IBM Corp.
October 1992

IP Multicast Communications Using BGP

Status of this Memo

This document reflects the current status of recommendations for supporting inter-domain multicast packet forwarding using BGP. This RFC specifies an IAB standards track protocol for the Internet community, and requests discussion and suggestions for improvements. Please refer to the current edition of the "IAB Official Protocol Standards" for the standardization state and status of this protocol. Distribution of this document is unlimited.

This document is an Internet Draft. 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. Internet Drafts may be updated, replaced, or obsoleted by other documents at any time. It is not appropriate to use Internet Drafts as reference material or to cite them other than as a "working draft" or "work in progress".

Abstract

This document, a major revision of the previous version, reflects the current status of recommendations for supporting inter-domain multicast packet forwarding using BGP. Research is underway on other methods for inter-domain multicasting, but only what can be done today is considered here.

1 Introduction

Most communication in the Internet today is unicasting, where there is a single specific destination for every packet. On local area networks broadcasting is common, in which the the destination of a packet is every node on the network. Multicasting is like broadcasting in that it supports multiple recipients for a single packet, but packets are intended for a specific group. As examples, in a local area network environment multicasting is currently often used for communication between processors of loosely coupled systems, or for communication between routers or bridges. In these cases the sender wants to reach the members of a special group, but not every node on the network. Broadcasting is in fact a special case of multicasting in which the special group is all nodes.

Multicasting over wide areas, as opposed to just on local area networks, is an important capability the development of which has lagged far behind its need. Until recently there has only been one IP multicasting implementation which could be used on more than just a local area network [1], [9], but multicast packets had to be encapsulated in order to send them across autonomous system boundaries. Work is now in progress to develop support for wide-area multicasting using standard protocols and to make that support an integrated part of the Internet protocol suite. Part of that work is being done under the auspices of the IETF BGP Working Group, including this document which explores how multicasting can be supported between autonomous systems using the Border Gateway Protocol (BGP) [5], [7], [8].

The best introductory reference on multicast forwarding is by Deering [2]. It is highly recommended that this paper, plus some of its references, as well as the BGP RFCs be read before this one, since this one assumes a high degree of understanding of BGP and only summarizes some parts of Deering's presentation.

We speak in terms of multicast groups. In IP multicasting, multicast groups are known by their addresses, which are in the range from 224.0.0.0 to 239.255.255.255. Each group has a unique address. Each member of a group is known by one or more multicast addresses in addition to one or more unicast addresses. RFC 1112 [1] defines methods for mapping between IP multicast group addresses and the level 2 address spaces of ethernet, 802.3, all point-to-point protocols, and protocols with broadcast but no multicast capability (e.g. LocalTalk). Mappings are also defined for FDDI [4] and SMDS

[6].

The general issues in wide-area multicasting are:

- Discovering where packets should be sent. The destination address for a multicast packet refers to a multicast group, whose members and locations change over time. On a local area network all destinations hear the same packet, and there is no need for forwarding. When a multicast packet must be forwarded to the members of a group not on the same local area network, we need mechanisms by which the members can make themselves known to the routers and the routers can ensure that every member of a group receives at least one and preferably only one copy of a packet addressed to that group.

- Establishing efficient routing paths for multicast packets. A packet with multiple recipients must be replicated and sent on multiple links, but copies of the packet should travel over as few links as possible. We need routing protocols defined for use both within and between autonomous systems.

A major issue which has received increasing attention recently is how well inter-domain multicast routing can scale, given that we are now thinking in terms of an Internet which should be able to support a billion domains.

This document assumes that hosts will inform routers of their membership in multicast groups, probably via the Internet Group Management Protocol [1]. It attempts to explore three remaining problems -- communication between routers of where multicast packets should be sent, efficient propagation of packets between autonomous systems, and interactions between intra-autonomous system and inter-autonomous system routing protocols in the support of multicasting.

2 Reverse Path Forwarding

The only approach to wide-area routing of multicast packets that has been implemented so far uses "reverse path forwarding" and is described in RFC 1075 and in [2]. This approach would fit well in the BGP environment, offering low overhead and excellent interaction with IGPs. Also the implemented method is directly applicable to BGP already. However, it may not allow the level of administrative control of routing paths to which some network administrators have

become accustomed (see Section 4.1).

In every approach to forwarding multicast packets the problem faced by a particular router is to determine its position in the paths by which multicast packets from a particular source should be forwarded. A router needs to (1) determine whether to accept a particular multicast packet or to discard it, based on its originating source and immediate previous hop, and (2) once a packet has been accepted, decide which of its peers to forward it to, if any.

If a "source tree" defines the paths by which an end system sends unicast packets to all other end systems, then a "sink tree" defines how an end system is reached by unicast packets from all others. The goal in unicast routing is to make a destination reachable by packets from all sources; the goal in multicast routing is to ensure that a packet from a single source reaches multiple destinations. Obviously a set of paths that solves the first problem can be used to solve the second if we use it in the reverse direction. The basic reverse path forwarding approach uses the fact that the propagation of unicast IP routing information already causes the formation of a sink tree -- the graph of how unicast packets should flow to that IP entity from all others. Thus when multicast packets need to be routed from that network to multiple destinations, a broadcast tree with that source as the root has already been formed, and this approach simply arranges for the multicast packets to flow along certain branches of that tree, but in the opposite direction of the unicast packets.

This procedure establishes paths for efficient broadcast, but network bandwidth is still wasted by sending multicast packets along all branches of the sink tree even when there are no nodes on those branches interested in receiving them. Further mechanisms can be defined to dynamically ensure that multicast packets are only sent to those peers which are on paths leading to members of the destination multicast group, for example via the "prune" and "graft" messages described in RFC 1075. A prune message is sent to tell a BGP peer not to send it packets addressed from a particular source to a particular multicast group. A graft message is sent to cancel that directive.

Prune messages can be cached and timed out by the receiver, and repeated as necessary by the sender. A border router can maintain a table of which interfaces packets from a particular source to a particular target multicast group should and should not be forwarded on, depending on memory constraints and multicast activity in the Internet.

3 BGP and Reverse Path Forwarding

The BGP protocol itself does not have to be changed to support inter-domain multicasting, but implementation of IGMP "prune" and "graft" messages by the BGP speaker is required.

Functionally, in reverse path forwarding, if a border router which receives a multicast packet receives it on the link by which it would send a unicast packet to the originator of that multicast packet, then it will propagate that multicast packet to the other BGP peers which are using it to reach the originator and which have not sent a "prune" message for that {originator, group} combination. In all current implementations of the BGP protocol, a border router has an implicit confirmation of whether its external peers are using routes that it has offered to them through the "echo" inherent in the BGP update messages (as strongly encouraged in the BGP4 Internet Draft). This combined with prune messages can efficiently limit propagation of multicast packets to only those branches that want them.

However, without some extra features it is impossible for border routers to exchange prune and graft information across an autonomous system. A border router can use information obtained through examining LOC_PREF attributes and/or other means to detect if it is its own AS's exit point for sending unicast packets to a particular multicast source. If it is not, then the border router would never propagate multicast packets from that source into its AS or across its AS to others. However, if it is the AS's unicast exit point for a particular source, then without any way to gather further information it will have to forward multicast packets across its AS to all other AS border gateways, since (in reverse path forwarding) it has no way of knowing if there is a group member beyond one of the other border gateways or not.

There are two solutions to this problem which are reasonable. The first, which is recommended here, is to define new IGMP prune and graft messages. Prunes and grafts were originally meant to be messages from a router to one of its immediate neighbors, telling the neighbor whether it has recipients downstream from it with respect to a particular multicast {source, group} combination. To solve the above problem we can create new IGMP prune and graft messages which would be advisory -- these messages would be sent, for example, from one AS border router to another, telling it that the originator has no recipients downstream from it with respect to a particular multicast source that the AS is reaching through the recipient border

router.

Another possibility would be to require the establishment of multicast "tunnels", as used by mrouted [9], between multicast-capable border routers. The tunnels would be used for sending encapsulated IGMP prunes and grafts between the border routers, bypassing the AS's internal routing. If tunnels are used, it would be best to have the multicast data packets carried in the tunnels as well -- one copy of a packet would be multicast into the AS if there were group members in the AS, and other copies would be encapsulated and sent directly to the other border routers that had not sent a prune for that {source, group} combination. The tunnels could be automatically created when the BGP connection is created.

Neither of these solutions would require changes to BGP, but both would couple multicast routing to knowledge of BGP routing information in the border routers. While the proposed solutions are similar to each other, the first one have an advantage of not requiring the establishment of multicast "tunnels", thus simplifying the operation of the protocol.

4 Potential Problems with Reverse Path Forwarding

4.1 Asymmetric routes

As long as the path by which one node reaches another is the exact reverse of how the other node reaches the first (symmetric routes), unicast and RPF-based multicast packets will flow along the same paths. However, the Internet supports, and frequently has, asymmetric routes between ASs. Network administrators currently set policies for how they want their networks to reach others, but, since in reverse path forwarding multicast packets flow according to how a node is reached, not according to how it reaches others, if routes are not symmetrical the behavior of the multicast packets will be controlled in the opposite way of what the network managers intended when they set up the controls for unicast traffic.

Discussions in IETF meetings suggest that while most network managers would not mind if multicast packets flowed from their ASs along the paths which others use to send unicast packets to them, there are some who would like to retain more control of how multicast packets flow through the Internet. There are ways to add source-based control, but they all add significant overhead either to protocol

traffic or to network administration. The cost to everyone seems to outweigh the benefit gained by a few. We can probably set up a mechanism similar to that in the "unified" routing scheme [3], where the majority of traffic is taken care of by simple, low-overhead routing, and for the small number of cases where it is necessary more complex routing can be used.

4.2 Incremental Implementation

One consideration is how easy it will be to get from the current Internet to one that mostly supports multicasting (getting to an Internet which fully supports multicasting is not a reasonable goal). Since in reverse path forwarding multicast routing depends directly on unicast routing, incremental implementation in the Internet might be awkward. There is no way to detect which routers support multicast routing, and thus no way to know if multicast packets can get between any two points, directly from the network itself. Tunnels may easily be set up, as described in RFC 1075, to reach between islands of multicast-supporting routers, but again with RPF there is no way of knowing when these (relatively inefficient) tunnels should be in place and when they are no longer necessary without frequent dialog between network operators.

Once again this is not an extreme difficulty, and network administrators are careful enough that they will probably be aware of their tunnel topology and their neighbors' activities, and able to control them effectively.

Another proposal, which has been called "multicast fireworks" because of the way multicast packets would "explode", essentially says that one should not require multicast forwarding to ever be completely deployed Internet-wide, that the Internet will be in a hybrid state, with some tunnels connecting multicast-capable ASs, for a very long time, perhaps forever.

4.3 Scaling

Many people have valid concerns about the capability of any multicast routing algorithm to scale to support 10^9 autonomous systems. In the case of reverse path forwarding, some people wonder about the involved in not propagating multicast group member locations in the first place, and essentially discovering them by sending data packets everywhere and using prune responses to clean up the forwarding tree

after the fact. Under traditional RPF, every multicast group with global scope periodically sends at least one packet to every part of the world, regardless of whether there are group members there or not. Since it is not necessary for a node to be a member of a group in order to send messages to that group, the alternative of propagating membership information (instead of the using the "probe" data packets) would require propagating membership information for each group everywhere, to any node that might want to send to that group. Propagating knowledge of group membership would require at least one packet for each member-containing network to be sent to every leaf of the Internet, each time that member-containing network transitioned between having zero and at least one member. On the other hand using data packets and prune messages would require one packet to be sent to every constituent of the Internet for the entire group, as opposed to sending one for each member-containing network. More data packets would be sent periodically, but the frequency would depend on the times specified in the prune messages. It is expected that these times will be long and that routers will use graft messages as necessary. Since a graft message will only be sent if data packets for a particular group is desired, graft messages are only incrementally more traffic than the data itself will be and are not significant as overhead. Thus, independent of the topology of the Internet, it is always cheaper to use the prune/graft approach than it is to propagate membership information. Finally, prune messages need not apply to just the particular group and source for the packet that triggers them. It can be shown that if the source and group fields in the prune message are prefix-based, and prunes are sent which cover all unwanted groups and sources, essentially in anticipation of future data "probe" packets, that very few of these packets will ever be sent.

Since reverse path forwarding works with whatever address prefixes are in the route information base at any BGP node, and keeps only cached information about active multicast sources and groups, the amount of stored information required will continue to scale well as the Internet grows.

There is a draft document based on Tony Ballardie's ongoing thesis work, in which he and others propose "core-based trees", basically that members of a group form a tree based on a well-known set of "core" nodes, and that senders of packets to that group need know nothing about the membership; they should simply send their packets toward the core. When the packets hit the tree formed by the members they will begin following all branches of the tree from that point. This scheme seems to have great potential, in that it doesn't flood

the Internet with either membership notifications or "probe" data packets, and thus it should scale well. Policies can be applied to some degree and traffic will flow from a source toward the tree basically according to the source's preferences. However, there is a chance that it might have significant overhead in maintaining trees, since participants must be sure that a particular "core" node is functioning, and adapt rapidly if it is not. The detailed mechanisms of actually making the scheme work robustly are still being explored and a subject of future research.

5 Acknowledgments

The development of some of the ideas presented in this document was supported by the Defense Advanced Research Project Agency through grant NAG 2-593 from the NASA Ames Research Center. This work would not have been possible without the help of the IETF BGP Working Group, John Moy, and Steve Deering.

References

[1] S.Deering, "Host extensions for IP multicasting", RFC 1112, Network Information Center, Aug. 1989.

[2] S. Deering, "Multicast Routing in a Datagram Internetwork", PhD thesis, Electrical Engineering Dept., Stanford University, Dec. 1991.

[3] D.Estrin, Y.Rekhter, and S.Hotz, "A Unified Approach to Inter-Domain Routing", RFC 1322, Network Information Center, May 1991.

[4] D.Katz, "A Proposed Standard for the Transmission of IP Datagrams over FDDI Networks", RFC 1188, Network Information Center, Oct. 1990.

[5] K.Lougheed and Y.Rekhter, "A Border Gateway Protocol 3 (BGP-3)", RFC 1267, Network Information Center, Oct. 1991.

[6] D.Piscitello and J.Lawrence, "A Specification of the Transmission of IP Datagrams Over SMDS", RFC 1209, Network Information Center, Mar. 1991.

[7] Y.Rekhter and P.Gross, "Applications of the Border Gateway Protocol in the Internet", RFC 1268, Network Information Center, Oct. 1991.

[8] Y.Rekhter and T.Li, "A Border Gateway Protocol 4 (BGP-4)", Internet Draft, Network Information Center, June 1992.

[9] D.Waitzman, C.Partridge, and S.Deering, "Distance vector multicast routing protocol", RFC 1075, Network Information Center, Nov. 1988.

Security Considerations

Security issues are not discussed in this memo

Authors' Addresses

Scott W. Brim
Cornell Information Technologies
143 Caldwell Hall
Cornell University
Ithaca, NY 14853
USA

Phone: +1-607-255-5510
EMail: Scott_Brim@cornell.edu

Yakov Rekhter
T.J. Watson Research Center IBM Corporation
P.O. Box 218
Yorktown Heights, NY 10598

Phone: +1-914-945-3896
EMail: yakov@watson.ibm.com