Revisiting IP Multicast

Juhoon Kim
(kimjuhoon@gmail.com)
Student Number: 312575

Seminar ”Internet Routing” ,
Technische Universitaet Berlin

WS 2006/2007 (January 17, 2008)

Abstract

This paper explains the history, the idea as well as usages and problems of traditional IP multicast. Furthermore, in this paper I discuss the very recent idea of IP multicast, FRM (Free Riding Multicast), which has been suggested in 2006.

1 Introduction

Since the 1990s Internet technology has been rapidly and widely developing. At the beginning of the Internet era there was not much demand for Internet speed, because most of the information on the Internet consisted of texts and small images. At that time it seemed as if no faster or better ways were required. Nevertheless, there were already attempts to find out more efficient methods to route same datagrams to multiple receivers through the network. One of the solutions became IP multicast and it is until today still actively discussed.

Because of the growth on the demand for the trasfer of the multimedia data such as real-time broadcasts of stock tickers, audio/video streams and software/data updates in these days, it is quite important to reapproach IP multicast. However, it is still not commonly implemented on the Internet, because the schemes of the IP multicast have often been complex and difficult to understand. As a result, in this paper I will focus on the traditional IP multicast first, then I will introduce the paper ”Revisiting IP Multicast”, which was published in 2006 by Sylvia Ratnasamy, Andrey Ermolinskiy and Scott Shenker.

2 IP Multicast

In the middle of the 1980s, Stanford university student Steve Deering and his advisor required a mechanism that made multicasted data flow through the network. The mechanism was supposed to be used on the network-distributed operating system ”Vsystem”, which Deering was developing at that time. His doctoral dissertation (”Multicast Routing in a Datagram Network” - December, 1991) was based on this idea and this paper later became the premier IP-Multicasting IETF document - RFC 1112. IP multicast is a datagram routing scheme to deliver a same datagram to large amount of receivers through the network. The basic mechanisms of Deering’s IP multicast were group membership advertisement (IGMP) and packet routing (DVMRP). [5] The first implementation of IP multicast was in BSD 4.4 in 1993.

2.1 What is IP multicast

Traditionally, there were two different ways to send datagrams to other hosts over the network, one is called ”unicast” and the other one is called ”broadcast”. Unicast is the sending of datagrams from one host to the other host, which is used by most of TCP/IP Internet connections.

As already mentioned in the introduction, there is much more bandwidth requirement to transfer video and audio streams, therefore unicast is not the right solution for such information transfer.

In contrast to unicast, broadcast means sending of datagrams from one host to all hosts over the network. The problem of broadcast, the possible waste of bandwidth, explains why multicast is required.

The IP multicast is a bandwidth-conserving technology that uses network traffic efficiently by simultaneously sending the datagram from the sender to a large number of hosts. The difference between broadcast and multicast is, that multicast only sends packets to the group, whose participants are interested in the information. This is done by the membership announcement and a special routing algorithm.

2.2 Class D IP addresses

”The Internet Assigned Numbers Authority (IANA) controls the assignment of IP multicast addresses. It has assigned the old class D address space to be used for IP multicast” [3]. This means that packets to a multicast group are sent to the class D addresses. The class D is in the range of 224.0.0.1 to 239.255.255.255.

Table 2.2.1 Well-known IP multicast addresses [3]

224.0.0.1	All systems on this subnet
224.0.0.2	All routers on this subnet
224.0.0.5	OSPF routers
224.0.0.6	OSPF designated routers
224.0.0.12	DHCP server/relay agent

2.3 IGMP (Internet Group Membership Protocol)

In the definition of the IP multicast, hosts can join and leave as well as create the multicast group at any time, thus multicast routers must renew its routing table by querying to end-hosts. For those purposes IP multicast requires the implementation of the IGMP. The IGMP is the communication protocol of the host and the multicast router. Its task is to manage the membership of multicast groups. The query is the message for checking if there is any host, which wants to join the group. The host answers with a report message, in case it wants to join the group. The group can be left by not replying to the qeury message for a certain time. Hosts use IGMP to report their host group memberships to close multicast routers. IGMP is defined by RFC-1112, RFC-2236 and RFC-3376. The latest version of IGMP is 3.

2.4 IP Multicast Routing Protocols

2.4.1 PIM (Protocol Independent Multicast)

PIM (Protocol Independent Multicast), like IGMP, is also a communication protocol, and it is used for routing multicast datagrams over the network, namely the IP multicast packet transmission. This transmission is going on between the local multicast router and the remote multicast router instead of between the host and the local router. Basically, there are four different kinds of PIM: PIM-SM, PIM-DM, Bidirectional PIM and PIM-SSM

- The first of this four is PIM-SM (Sparse Mode). PIM-SM works with shared trees which are unidirectional. These shared trees use a RP (Rendevous Point), where they are routed. PIM will forward the data down the shared tree or on the networks, no matter what size they are. PIM-SM is mostly implemented for wide-area usage.
- The second one is PIM-DM (Dense Mode). It works with the so called ”flood and prune mechanism” [3]. It builds shortest-path trees by flooding multicast traffic domain all over the network and that way branches of the tree, where no receivers are present, are pruned back. After that, the scaling properties of PIM-DM are usually poor.
- The third one is Bidirectional PIM. Whereas PIM-SM works in a unidirectional way, Bidirectional PIM builds shared bidirectional trees. Due to the fact that it never builds a shortest path tree, it may have longer end-to-end delays than PIM-SM. Nevertheless, one has to mention that it scales well because it needs no source-specific state.
- The last one is PIM-SSM (PIM Source Specific Multicast). PIM-SSM offers a safer and more scalable model for a limited account of applications. The reason is, that trees are built, which are routed in just one source.

2.4.2 MOSPF (Multicast Extension to OSPF)

OSPF (Open Shortest Path First) is a unicast link state routing protocol. MOSPF is the extension to OSPF routing protocol for IP multicast use. Over the MOSPF network each host can send IP multicast packets to the IP multicast group without any tunnels. MOSPF includes multicast information in the OSPF link state advertisement. From this information a MOSPF router can learn which groups are active. Basically, the packet of MOSPF is formed in the same way as for the OSPF Version 2. The difference from OSPF is one additional field in the hello packet, in the database description packets and in all kinds of link state advertisement packets. This additional field is ignored by all unicast routers and only processed by multicast routers. For the multicast extension to OSPF, one of the LSA (link state advertisement) packets is modified and a new LSA packet is added. The one, which is modified, is the router-LSA. The fifth bit was not used in OSPF, but in MOSPF, when it is set, the router is a wild-card multicast receiver.

Graph 2.4.2: the modified field in the router-LSA

1	2	3	4	5	6	7	8
—	—	—	—	W	V	E	B

- W (Wild-card): The router is a wild-card multicast receiver, when it is set.
- V (Virtual): The router is an end point of an active virtual link, when it is set.
- E (External): The router is an AS boundary router, when it is set.
- B (Border): The router is an area border router, when it is set.

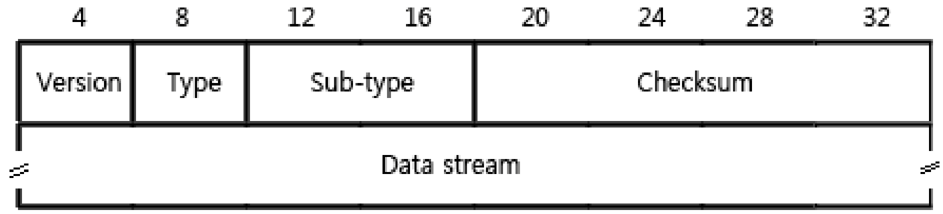
The other LSA, which is newly added, is the group-membership-LSA. The group-membership-LSA is used for the multicast group membership advertisement. This LSA is totally ignored by unicast routers.

MOSPF is defined by RFC-1584.

2.4.3 DVMRP (Distance Vector Multicast Routing Protocol)

DVMRP is the oldest routing protocol to transfer IP multicast packets over networks. DVMRP can run over various types of networks, including Ethernet local area networks (LANs). It can even run through routers that are not multicast-capable. The protocol sends multicast data in the form of unicast packets that are reassembled into multicast data at the destination. DVMRP got many ideas from RIP (Routing Information Protocol) and TRPB (Truncated Reverse Path Broadcasting) algorithm. The main difference from RIP is that RIP routes datagrams to a destination without a certain track, but DVMRP has a track for routing datagrams. DVMRP encapsulates packets in IP datagrams (IGMP). DVMRP packets consist of the header and the data stream.

Graph 2.4.3: The structure of DVMRP packets



- The current version is 1
- DVMRP type is 3
- Sub-type: Response (1), Request (2), Non-membership report (3), Non-membership cancellation (4).

The DVMRP is defined by RFC-1075.

2.5 MBone

Because most of the routers are not yet able to support IP multicast packet routing for the present, the virtual network called ”Multicast Backbone (MBone)”, which was introduced by Steve Deering and adopted by the IETF in March 1992, is built on the Internet. Until March 1997, there were more than 3,000 MBone servers on the Internet. The connections between MBone servers are called ”Tunnel”. The multicast packets are incapsulated while they are transfered through the tunnels, so that they just look like normal unicast datagrams. The multicast router, which receives the encapsulated IP multicast packet, modifies the destination ip address to the next MBone server and forwards it.

3 Free Riding Multicast

3.1 What is FRM (Free Riding Multicast)

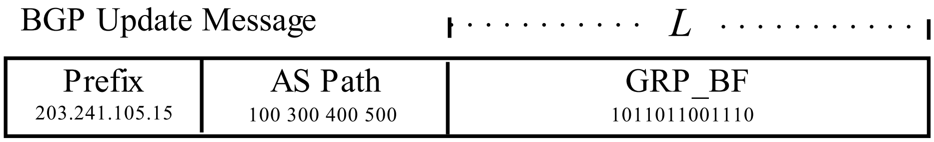
FRM is a new approach to implementary IP multicast. It was designed in 2006 by Sylvia Ratnasamy from Intel Research, Andrey Ermolinskiy from U.C. Berkeley and Scott Shenker from U.C. Berkeley and ICSI. They published their ideas in a paper with the title ”Revisiting IP Multicast”.
FRM is a new idea of IP multicast packet routing and forwarding, so that it simplifies the deployment, or design, of IP multicast, that is to say FRM makes IP multicast more practicable. In the paper the discussion focuses especially on the inter domain multicast.
The advantage of FRM is, that there is no requirement for new facility or protocol mechanisms, because FRM works well with already existing protocols. FRM uses the same BGP session, which is also used by the unicast routing system, so that one can avoid a lot of the multicast route computation, and the result of this is, that the implemenataion of IP multicast becomes much easier.
However, FRM does not only have advantages as written above, but also disadvantages such as more bandwidth consumption and overhead. The reason for this is, that FRM needs to attach more information for membership advertisement, therefore such disadvantages are

not avoidable. The question is, whether FRM is still more efficient than traditional IP multicast mechanisms although it has more bandwidth consumption and overhead. This will be discussed in section 3.4.

3.2 Group Membership Advertisement

Hosts can join and leave a group at any time, therefore routers must be informed by hosts every certain amount of time if there is any change of group membership. According to the cisco documentation [3], this can be done by implementation of the Internet Group Memebership Protocol (IGMP, current version is 3).
The problem is that IGMP is only available between IP multicast supporting hosts and IP multicast supporting routers. However, the goal of FRM is, as stated above, to make implementation of IP multicast easier by using existing unicast channels (BGP). This way, one can say that FRM is extension to BGP.
The basic principle of FRM is that the information packet through BGP sessions carrys additional information of the group membership. It is actually more practical because the router can obtain routing information (the shortest path) and membership information at once.

Graph 3.2



The group information consists of the active group addresses and is denoted as GRP_BF. The length of this GRP_BF is L. The FRM encodes this group information with the Bloom filter algorithm. [1] It always gets false positive, so that it will never disregard the wish of joining the group. However, it always has more possiblities of sending and receiving information, even though these are not interesting for the hosts, because the more members are included in the group, the higher the possibility of false positive for the Bloom filter.
To solve this problem, the host must either just drop the information, which is not interesting, or inform the upstream ASes to stop sending information.

3.3 Forwarding

Existing intra-domain protocols should be used because of their complexity although FRM could also be extended to intra-domain scenario.
The multicast packets for the source and border routers are processed differently by FRM at the border router.

[1] The Bloom filter is an algorithm for finding out whether a given element is a member of the set or not. It was invented by Burton Bloom in 1970 and suggested as mechanism for the use of indexing and identifying web pages. This method is based on probability. It has two different states which are ”false positive” and ”false negative”. The element can be a member of the set if it is false positive, but not if it is false negative.

3.3.1 Forwarding on GRP_BF state at the Rs

Intra-domain multicast routing protocols[2] are used for delivering multicast packets between the source and the border router at the source's domain. The router at the source's domain seeks membership information from its BGP RIB[3] and constructs routing trees according to its AS_path.
This computation should be done by the processor of the Rs, because there is less forwarding load at that position and it is easier to eliminate overheads, moreover it is a very good position to cache or even to precompute the lookup results.

3.3.2 Forwarding on cached state at the Rs

The lookup results, which are stored at the position of the Rs, are indexed by their group addresses, therefore it is possible to seek group addresses directly by several exact-match methods such as CAMs[4] and direct memory data structures. The size of the caching data depends on the number of groups. The caching data can be stored in RAM at the Rs.

3.3.3 Forwarding on GRP_BF state at the Rt

Hence the Rs is not able to simply forward the packet to the other router, information on next hops is encoded in the ”shim” header[5]. For reducing the overhead by compressing, this information is also encoded by using the Bloom filter (denoted TREE_BF). TREE_BF, unlikely GRP_BF, has a fixed size of data length. Then FRM packs the TREE_BF as binary format and attaches it to the header.
This causes two problems. One is an inefficient use of bandwidth, which can be actually worse than traditional multicast methods because FRM needs a shim header for every single packet and this causes more transmission of information than with traditional methods. The other problem is the overhead from the cached results because now FRM includes now encoded information on the next hop list.
However, those payoffs are acceptable because the bandwidth consumption and overhead of the encoded tree information are quite scalable, to be accurate the Rt only needs to check whether its neighbor edges are in the encoded TREE_BF of the shim header, therefore the Rt only needs to know which neighbors it has and it is independent from any other groups.
For this task FRM uses TCAM[6] with the Bloom filter to find out whether the neighbor edge is in the shim header or not.
Because of this structure, which only depends on its neighbor edges, FRM does not need any wide-area protocol mechanisms.

[2] PIM-SM and CBT (Core Based Tree) are two well-known intra-domain multicast routing protocols

[3] Routing Information Base

[4] ”A method and apparatus for determining an exact match in a ternary CAM device.” (http://www.patentstorm.us/patents/6539455.html)

[5] Being the header format for MPLS (Multi-Protocol Label Switching), the shim header is inserted between the Layer 2 and Layer 3 headers. The format of the shim header is documented under RFC 3032.

[6] TCAM (Ternary Content Addressable Memory) is a method to find an exactly matching entry by a given key or to find the range the key is belonging to. TCAM has 0, 1 and ”don't care” matching states.

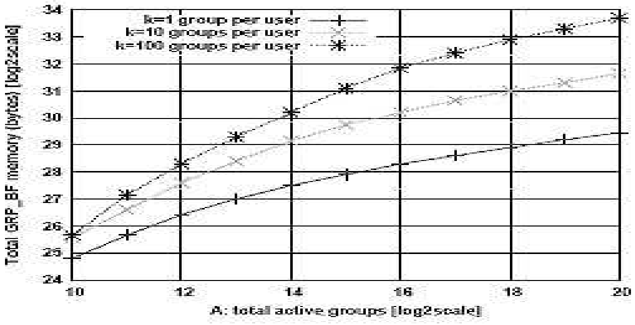
3.4 Overhead

As stated in prior sections FRM has several advantages. First, FRM decreases the complexity of IP multicast machanisms. Second, it makes ISP control membership advertisement easily and powerfully, because it uses existing BGP sessions. Third, FRM eases the configuration by the avoidance of the frequent selection of RPs.
However, FRM does not only have such advantages, it also has several disadvantages. As stated earlier two typical disadvantages are additional bandwidth costs and storage overhead.

3.4.1 Overhead due to the group membership advertisement

The packet within FRM includes encoded group mambership information, namely GRP_BF, this costs, of course, memory overhead. As we see in graph 3.5.1 [0], if we assume one million simultaneously active groups and ten groups per user, about 3GB of route processor memory is required [0]. According to the rapidly developing memory technology and to the current price of memory devices, the costs, as evaluated in the paper, are acceptable.

Graph 3.4.1 [0]



3.4.2 Bandwidth consumption due to the membership advertisement

There is, of course, extra bandwidth consumption for updating the group membership, however, that is quite manageable. In the paper [0] that was evaluated with the help of back-of-the-envelope calculations[7].
Ratnasamy, Ermolinskiy and Shenker use different calculations as examples to make the bandwidth consumption clear. If it is assumed that 5 hash functions are used by GRP_BFs and bit positions are represented as 24 bit values (in multiples of 256-bytes), the bandwidth consumption of a membership advertisement for each of all events to leave or join a group is approximately 15 bytes. [0]
If it is moreover assumed that there are 200000 prefixes in the BGP RIB and every prefix has at least one join/leave or a create event per second, the required bandwidth due to incoming GRP_BF update traffic at a border router is approximately 3MBps.[0] This is only a very small amount of the bandwidth capacity at a core BGP router.

[7] ”Back of the envelope (BotE) reasoning involves generating quantitative answers in situations where exact data and models are unavailable and where available data is often incomplete and/or inconsistent.” [4]

3.4.3 Bandwidth consumption due to FRM packet forwarding

Most of the bandwidth cost is caused by the per-packet shim header and too much transmission due to large subtrees, which cannot be encoded in a single shim header. The assumption in the paper [0] is 100 bytes of shim headers. However, according to the result of *total-tx*[8] evalution, which is one of two metrics for measuring the bandwidth consumption in the paper [0], the numbers of packets for the transmission are just slightly bigger than the ideal numbers of multicast. The paper also shows that FRM is much more efficient than the per-AS unicast method. Because due to most of the cases in the evalutation table (Chapter 6.1, table 2), the total number of packets of FRM is approximately half of the total number of packets of per-AS unicast. As it was expected one can see that the bigger the group is, the bigger the gap of the total number of packets between the ideal multicast and FRM gets.

Another evaluation in the paper (Chapter 6.2) [0] is the comparison of the redundant transmission between per-AS unicast and FRM. This time *per-link-tx* [9] is used instead of *total-tx* for measuring the bandwidth consumption. 10 million users are assumed on the network. In this evaluation we see more than 90% of links require only one transmission per link. With FRM's tree-encoded forwarding, the ratio of links, which requires only one transmission per link, goes up to 99.5%. This means that only less than 0.5% of the links have redundant transmissions. In the worst case the number of transmissions is going up to 157, but that is still much less than the worst case of unicast (6950). Moreover, there are ways to optimize the worst case of FRM's redundant transmissions. With optimization even in the worst case the FRM link sees only two redundant transmissions. Compared to FRM, in unicast 6% of links see redundant transmissions and in the worst case the number of transmissions per link goes up to 6950.

3.4.4 Storage costs due to FRM packet forwarding

At the position of the router Rt, the number of forwarding entries depends on its neighbors. This means that it depends on the AS degree on its domain. If the aggregate is used for the efficiency of FRM forwarding, it adds additionally two hops more for the forwarding. As we see from the evaluation in the paper [0], the CDF[10] of forwarding entries per AS shows very small forwarding tables. More than 90% of links have less than 10 forwarding entries in both basic FRM and the FRM which is using aggregate links, and just less than 1% of all links have more than 100 forwarding entries. In the worst case they increase up to 2400 forwarding entries without aggregate links, however with aggregate links the forwarding entries in the worst case increase up to 14,071 due to two additional hops per link as already discussed above.

[8] *total-tx* is the total number of packets which are transmitted from the source to all receivers in the network

[9] *per-link-tx* is the number of transmissions per link used to multicast a single packet from the source to all receivers.

[10] CDF (cumulative distribution function) describes the probability distribution of a real-valued random variable

3.5 Optimizations of bandwidth consumption due to FRM packet forwarding

Although its efficiency is much higher than the one of unicast, the ways for optimizing FRM's worst case transmission should be discussed here. Ratnasamy, Ermolinskiy and Shenker suggested two different ways of optimization:

- no leaves: The ASes at the leaves do not have to be encoded into the shim header, because it is possible that the provider AS, which receives advertisement packets from the customer AS, detects the customer AS.
- aggregate links: If the edges from an AS are too inefficiently seperated, then the router replaces them with all related edges from AS A. This reduces transmissions due to large subtrees, which cannot be encoded in a single shim header.

4 Summary

There is no doubt about the requirements of more efficient Internet routing methods, therefore IP multicast was actively discussed and researched since the middle of the 90s. Currently, variety protocols are designed for IP multicast. However, because of its complexity the deployment of IP multicast is still not common.

FRM is a new, effective, efficient and simple approach to IP multicast. It was designed for making the inter-domain multicast protocol easier to realize but it is also extendable to the intra-domain protocol. FRM is basically restructuring of inter-domain multicast routing protocols. FRM is an extention to BGP, so that it uses existing unicast channels.

FRM encodes group information for the membership advertisement and the neighbor hops for the packet forwarding. This is done by the Bloom filter method and direct matching lookup methods such as CAM.

FRM pays storage costs and consumes more bandwidth in exchange for simplicity of protocol, ease of the configuration and ISP control over sources/subscribers. However, these overheads are, due to current storage technology and its efficiency, very acceptable and manageable.

References

[0] S. Ratnasamy, A. Ermolinskiy, S. Shenker: *Revisiting IP Multicast*;, ACM SIGCOMM 2006.

[1] Thomas Albert Maufer: *Deploying IP Multicast in the Enterprise*;, 1997.

[2] Iljitsch van Beijnum: *BGP Building Reliable Networks with the Border Gateway Protocol* ; O'Reilly 2002.

[3] CISCO Documentation: `http://www.cisco.com/univercd/cc/td/doc/cisintwk/ito_doc/ipmulti.htm`

[4] Praveen K. Paritosh and Kenneth D. Forbus: *Analysis of Strategic Knowledge in Back of the Envelope Reasoning*; AAAI 2005.

[5] IP-Multicasting Technology: `http://www.intelligraphics.com/articles/ipmulticasting1_article.html`