Static Internet Multicast

Masataka Ohta
Tokyo Institute of Technology,
Jon Crowcroft
University College London

Abstract

The current IP Multicast[1] model appears to achieve a level of simplicity by extending the IP unicast addressing model (historically the classful A,B, and C net numbers) from the mask and longest match schemes of CIDR[2], with a new classful address space, class D. The routing systems have been also built in a deceptively simple way in one of three manners - either broadcast and prune (DVMRP[3], Dense Mode PIM[17]), destination list based tree computation (MOSPF[4]) or single centered trees (current sparse mode PIM and CBT[16]). The multicast service creates the illusion of a spectrum that one can "tune in to", as an application writer. Due to this view, many have seen the multicast pilot service, the Mbone, as a worldwide Ethernet, where simple distributed algorithms can be used to allocate "wavelengths" and advertise them through "broadcast" on a channel (the session directory), associated with a spectrum.

These three pieces of the picture have tempted people to construct a distributed architecture for a number of next level services that cannot work at more than a modest scale, since they ignore the basic spirit of location independence for senders and receivers of IP packets, whether unicast or multicast. The problem is that many of these services are attempting to group activities at source, when it is only at join time that user grouping becomes apparent (if you like, multicast usage is a good example of very late binding). These services include Address Allocation and Session Creation, Advertisement and Discovery.

This memo proposes approaches to solve some current multicast problems rather statically with DNS[13] and URL based approach, and avoid the misguided pitfalls of trying to use address allocation to implement traffic aggregation for different sources or aggregation of multicast route policy control through control of such aggregated sources.

Note that a minor level of aggregation occurs in applications which source cumulative layered data (e.g. audio/video/game data[18] or layered multimedia conferencing tools[5] such as vic/rat and rlc[6]) - this memo is orthogonal to such an approach, which in any case ony results in a small constant factor reduction in state.

A lot of the IP multicast additional pieces of baggage are associated with the multimedia conferencing on Mbone - however, the commercial internet use of multicast includes many other applications - for these, SDR may not be the best directory model.

1 Introduction

Multicast and related applications have traditionally been developed in Routing and Transport areas. Naturally, designers have tried to solve many problems using techniques familiar to those working in the routing and transport areas, that is, with flooding or multicast.

Of course, global flooding or multicast do not scale very well, which means that scalable solutions that make use of these techniques only are often impossible in the world wide Internet.

An attempt to reduce the scalability requirement to localize multicast and flooding area through TTL or administrative scoping (intra-site, intra-provider multicast etc.) works only in a small scale experiment like Mbone. In the real Internet, senders and receivers of multicast communication, in general, may be using different providers and are distributed beyond AS boundaries.

As a result, there was a hope that address aggregation and unicast area topology report aggregation can solve the multicast scalability problems in the same way that they have bailed out the unicast Internet from problems with limitation of router memory and the capacity needed for route update reports:

- a) Unicast addresses refer to a location, however. Multicast addresses are logical addresses, and refer to sets of members who may be anywhere, and may be sent to by sources which are also in more than one of many places. This means that for unrelated multicast group (and we anticipate that, in general, we can expect relationship between groups only when the groups belongs to a single application and that there is far more group that is unrelated than there is group that is related), there is no meaningful allocation at session creation time of a mask/prefix style multicast address, either for destination group, or sources.

- b) To control the amount of state and routing control messages, the Internet has divided the routing systems into autonomous systems/regions, which can run their own routing, and need only report summarized information at the edge to another region. This serves two purposes in the Unicast world:

    1. Inter-domain routing protocols can be deployed that are different in different areas (this may be applied recursively).
    2. Summarization can be applied at "min-cut" points in the topology, and reachability information only needs to be exported/imported across borders.

Note that, autonomous system boundaries are merely for operational purpose of easy policy description. The boundary does not contribute to protocol issues to reduce the amount of routing information, which is accomplished with multi-layered OSPF[11] without BGP[12].

With multicast, while one could define inter-working boundaries and functions as the IDMR WG has, the principle goal of scaling the reports at a border cannot be achieved in a location independent manner (in the sense that without moving all the receivers to a particular region, there is no aggregation feasible).

As a result of this confusion, intra-domain multicast protocols, which are expected to operate within a single AS have been developed that scale poorly, even though there was no known inter domain multicast protocol which solves the scalability problem.

It has been shown [14] that aggregation of multicast routing table entries, the number of which is a major scalability problem for IP multicast, is, in general, impossible.

The impossibility proof assumes nothing about QoS. That is, multicast QoS Flow state can be aggregated as good/bad as multicast best effort communication. RSVP may be extended to aggregate RSVP requests of strongly interrelated flows, for example, for streams with layered encoding, which may or may not share a single multicast address, latter case of which may result in a small constant factor of routing table entry reduction.

There may be a counter argument that a broadcast/prune in region (== big ether) and spare in other region for clumpy cast can overcome the problem. However, forwarding for "spare in other region" needs a routing table entry of its own. Moreover, even in the region, "broadcast/prune" scales worse than the theoretical lower bound of PIM-SM/CBT.

Thus, it is now necessary to thoroughly reconsider the architecture of multicast, Given a theoretical lower bound of multicast routing table entries, now is the time to find a multicast algorithm to achieve that lower bound. It is also meaningful to make the multicast architecture independent of unicast address hierarchy.

Fortunately, some problems can easily be solved for many common cases using techniques available in other areas without scalability problems.

Since the legacy multicast architecture was constructed carefully assuming routing table aggregation possible, it is necessary to change some of it to deploy new techniques. To solve hard scalability problems, it is necessary to recognize that all the details of all the protocols are tightly interrelated.

The multicast problems identified to be better solved in internet or application area in this memo are:

Multicast Address Allocation There was a proposal to allocate multicast address dynamically along the unicast address hierarchy. Such an allocation policy was expected to enhance the possibility of aggregation. However, as shown in the next section, it is impossible to aggregate multicast routing table. Then, while it is still possible to aggregate multicast address allocation, it is not meaningful.

However, it is meaningful to allocate multicast addresses statically through the DNS.

Multicast Core/RP Location CBT and PIM-SM were developed as intra-domain multicast protocols designed to be independent of the underlying unicast routing protocols. Naturally, they achieve the lower bound of spatial routing table size complexity. However, CBT and PIM-SM are not totally independent of unicast routing architecture, since they depends on flooding within an AS to locate the core or rendez-vous point. While this scales a little better than static assignment, it is still fairly bad. On the other hand, it is straight forward to use DNS to map from DNS multicast name to multicast address, core and RP. This solution may not be an option when dynamic multicast address assignment was a MUST and DNS dynamic update was not possible. However, this is now rectified since DNS update is being implemented now.

Multicast Session Announcement The announcement of multicast sessions can be performed over a special multicast channel. But the approach does not scale if the number of multicast channels increases. Of course, it is possible to introduce hierarchy of multicast session announcement channels. The real world complex structure makes the relationships between session announcement a complex network. Then, users join a session directory hierarchy by joining a group for some level, following the hierarchy, or following short-cut or following links, changing between several multicast groups to reach the final destination multicast for the session they seek. But as is proven, multicast costs routing table entries and associated protocol processing power of routers if a data of the multicast flows over the routers. So, it is desirable to constrain the number of multicast channels to be as small as possible.

If, instead, we use WWW[7] as EPG (Electric Program Guide) and embed SDP or SMIL information in RTP URLs, it can be used as multicast session announcement with arbitrary complex structure of hierarchy, short-cut or links with some caching, and we can use search techniques on this static data more easily.

Of course, neither DNS nor WWW scale automatically: they must continue to scale anyway and a lot of effort was already and will continue to be paid to make them better scale, more dynamic and more secure and their servers are becoming more capable (caching etc). DNS will be used for unicast name to address lookup forever and WWW will be the preferred way to retrieve information.

2 Meaningless Aggregation of Multicast Addresses

It is, in general, impossible to aggregate multicast routing table entries.

The minimum amount of state in each multicast router must be proportional to the number of multicast data flows which are running over it.

The locations of receivers are different, multicast application by multicast application. Multicast forwarding must be performed over a tree of receivers. The sources are different too. Thus, the tree is different multicast by multicast.

It is possible to aggregate multicast address allocation by making multicast location dependent with, say, a root domain. Then, it is possible to aggregate routing table entries to the root domain. For some type of central set of agencies (traditional broadcast TV/Radio) it might be possible to site their feeds at the same places in the Internet. But this is antithetical to the arbitrary growth allowed by random siting/evolution of content providers today, even in the Web. Sheer numbers preclude building unicast pipes from each source to a central set of sites.

However, it is still impossible to aggregate routing table entries to the receivers. The distribution pattern of receivers is unrelated to the location of the root domain. That is, a separate routing table entry is necessary for each multicast application.

A group of multicast receivers sharing a root domain may still have weak relationships in that most of them do not have any member in domains far from the root domain. Then, it is possible to share a default routing table entry, not to forward anything. But, such an entry is meaningless, because there is no data packet that will be forwarded for the entry and we still need unaggregated routing table entries for each multicast running over multicast routers.

Alternatively, it is possible to assign multicast addresses aggregated according to the statically or runtime detected distribution pattern of the receiver hosts, areas or domains. However, even with 32 receiver hosts, areas or domains, we need 32 bits for the aggregation prefix of the multicast addresses, which is too many for IPv4. Even IPv6[8] address space does not help a lot (96 receivers is not a great step forward!). Moreover, as the multicast membership changes dynamically, the multicast address itself must change dynamically.

That is, according to the current model of the Internet multicast, it is impossible to aggregate multicast routing table entries.

It is meaningless to try to aggregate multicast address assignment.

Still, it, of course, is meaningful and necessary to hierarchically delegate multicast address allocation.

3 The Difficulty of (Multicast) Address Assignment

Compared to the administrative effort for unicast address assignment by IANA, Internic, RIPE, APNIC and all the country NICs and development of the policy they used, it is trivially easy to develop a DHCP[9] protocol. The difficulty with DHCP was in the fact that the clients can not be reached by its IP address. It is even more trivial to develop a DHCP-like dynamic multicast address assignment protocol for clients unicast addresses which are already established. It could be as simple as a new option field of DHCP.

However, such a use of DHCP is meaningless, unless an administrator of the DHCP server has been delegated a block of unicast addresses and establishes a policy on how to assign them to clients. Similarly, we can argue that the DHCP-like mechanism for multicast is not a good solution.

Basically, multicast address assignment is not a protocol issue.

4 Recycling the Unicast Policy, Mechanism and Established Address Assignment for Multicast Policy, Mechanism and Address Assignment

If rather static allocation of multicast address is acceptable, it is possible to reuse the policy, mechanism, address assignment and protocol of unicast address assignment for multicast addresses..

For example, if we decide to use 225.0.0.0/8 for the static allocation, it is trivial to delegate the authority of multicast address 225.1.2.3 to an administrator of 3.2.1.in-addr.arpa, the administrator of 1.2.3.0/24.

We can simply define that the multicast DNS name should be looked up as:

```
3.2.192.225.in-addr.arpa.       CNAME mcast.3.2.192.in-addr.arpa.

mcast.3.2.192.in-addr.arpa.     PTR     bbc.com.

bbc.com.                        A       225.192.2.3
```

Then, if we construct applications that check the reverse mapping, unauthorized use of multicast addresses will be automatically rejected, which is what we are doing today with unicast addresses.

Note that the administrator of 3.2.192.in-addr.arpa is not the final person to be delegated the address but can further delegate the authority of mcast.3.2.192.in-addr.arpa. to someone else.

It should also be noted that, while the delegation uses the existing policy, mechanism, assignment and protocol, it does not mean that the multicast address must be used within the unicast routing domain of the unicast address block.

Just as MX servers or name servers can be located anywhere in the Internet regardless of the location of the hosts under the DNS domain they are serving, multicast channels can be used anywhere in the world.

The assingment policy automatically assure global uniqueness. But, it is still possible to have multicast addresses with local scopes, as long as they share globally unique well known DNS names, which is what we are using for intra-subnet multicast with IANA assigned well known names [15].

]sectionCore/RP location

The location of core of CBT or rendez-vous point of PIM-SM through DNS is straight forward as:

```
    bbc.com.    A       255.192.2.3
                RVP     london-station.bbc.com.

or

    bbc.com.    A       255.192.2.3
                CORE    london-station.bbc.com.
```

Again, just as MX servers or name servers can be located anywhere in the Internet regardless of the location of the hosts under the DNS domain they are serving, core or rendez-vous points can be located anywhere in the world.

CORE and RVP RRs have exactly the same syntax as PTR RR. Their query type values are *to be assigned by IANA* While the current CBT nor PIM-SM does not allow a single multicast group has multiple cores or rendez-vous points, future extension may. Thus, at the DNS level, a single node may have multiple CORE or RVP RRs. That is, the following DNS node is a valid node:

```
bbc.com.    A       255.192.2.3
            RVP     london-station.bbc.com.
            RVP     wales-station.bbc.com.
```

5 Session Announcement

The proposal is essentially to use a URL of RTP combined with SDP like:

```
rtp://london-station.bbc.com/?t=2873397496+2873404696&
m=audio+3456+RTP/AVP+0&m=video+2232+RTP/AVP+31
```

The URL contains all the necessary information to establish a session, including the domain name (or multicast address), port number(s), RTP payload type and optional QoS requirement.

Then, users surfing over WWW can actively search or randomly encounter some multicast or unicast RTP URL.

If the user clicks the label of the URL, the user will be queried whether he want to receive (should be default for multicast) or send data or both (should be default for unicast). He will also queried the source or destination of the data with appropriate default (his TV at the living room) and the multicast session begins, if necessary, with RSVP[10].

6 References

RFC ("Request for Comment") documents are Internet Protocol Specifications and can be retrieved from `ftp://doc.ic.ac.uk/`. Their status is defined in the current RFC that is called "Assigned Numbers".

Internet Drafts are Internet Work in progress documents, and can be retrieved from `ftp://ftp.isi.edu/in`

References

[1] D. R. Cheriton and S. E. Deering, "Host Groups: A Multicast Extension for Datagram Internetworks," ACM Computer Communication Review, vol. 15, Sept. 1985. Ninth Data Communications Symposium

[2] Hans Werner Braun, Peter S. Ford, and Yakov Rekhter, "CIDR and the evolution of the Internet Protocol," in Proceedings of the International Networking Conference (INET), (San Francisco, California), pp. BBA-1 - BBA-5, Internet Society, Aug. 1993. also SDSC Report GA-A21364 and in ConneXions September 1993.

[3] RFC 1075 Distance Vector Multicast Routing Protocol. D. Waitzman, C. Partridge, S.E. Deering. Nov-01-1988.

[4] RFC 1584 Multicast Extensions to OSPF. J. Moy. March 1994.

[5] Mark Handley, J. Crowcroft, Carsten Bormann, Jurg Ott, The Internet Multimedia Conferenceing Architecture Internet Draft draft-ietf-mmusic-confarch-01, also, to appear in Computer Networks and ISDN Systems, late 1998.

[6] L.Vicisano, L.Rizzo, J. Crowcroft Layered Congestion Avoidance for Relaible Multicast IEEE Infocom, San Francisco, April 1998.

[7] M. Handley, J.Crowcroft World Wide Web, Beneath the Surf UCL Press, 1994.

[8] RFC 2101 IPv4 Address Behaviour Today. B. Carpenter, J. Crowcroft, Y. Rekhter. February 1997.

[9] R. Droms, "Dynamic host configuration protocol," Request for Comments (Proposed Standard) 1541, Internet Engineering Task Force, Oct. 1993.

[10] Lixia Zhang, Stephen Deering, Deborah Estrin, Scott Shenker, and Daniel Zappala, "RSVP: a new resource ReSerVation protocol," IEEE Network, vol. 7, pp. 8-18, Sept. 1993.

[11] J. Moy, "OSPF version 2," Request for Comments (Draft Standard) 1583, Internet Engineering Task Force, Mar. 1994

[12] D. Katz, Y. Rekhter, T. Bates, and R. Chandra, "Multiprotocol extensions for BGP-4," Internet Draft, Internet Engineering Task Force, Jan. 1998. Work in progress.

[13] RFC 1591 Domain Name System Structure and Delegation. J. Postel. March 1994.

[14] [MANOLO] http://web.jet.es/sola/inet98.html

[15] [IANA] http://www.ietf.org

[16] [CBT] Tony Ballardie, Paul Francis, and Jon Crowcroft, "Core based trees (CBT)," in SIGCOMM Symposium on Communications Architectures and Protocols (Deepinder P. Sidhu, ed.), (San Francisco, California), pp. 85-95, ACM, Sept. 1993. also in em Computer Communication Review 23 (4), Oct. 1992.

[17] [PIM] Steve Deering, Deborah Estrin, D. Farinacci, Van Jacobson, C. G. Liu, and L. Wei, "An Architecture for Wide-Area Multicast Routing," in SIGCOMM Symposium on Communications Architectures and Protocols, (London, UK), pp. 126-135, Sept. 1994.

[18] Shareware Game from INRIA, France. http://www.inria.fr/rodeo/MiMaze/