Paper title: The Evolution of Multicast: From the Mbone to Interdomain Multicast to Internet2 Deployment – Kevin C. Almeroth

Background and context

This paper was published in 2000 in a high impact (ISI impact score 3.068) IEEE bimonthly magazine publication. At the time of publication Multicast had been in existence for 12 years, but had not been widely adopted by the network community. It was implemented in a Multicast Backbone (Mbone) in 1992 which made use of IP-encapsulated tunnels to support multicast connectivity among machines. It evolved over the years to eventually use the native multicasting ability of routers and then became a hybrid of native multicast links and tunnels.

Multicast was also deployed in the two Internet 2 backbone networks spanning the US – the very-high speed Backbone Network Service (vBNS) and Abilene. Elsewhere multicast had also been implemented in the TEN-155 research network in Europe.

There was a flurry of Multicast protocols developed in those 12 years some of which were tested in Internet 2, MBONE, TEN-155 and others:

- Intra-domain for use in flat topologies (or tunneled flat topologies)
  - Deering's original Distance Vector Multicast Routing Protocol (DVMRP)
  - Multicast Extensions to OSPF (MOSPF)
  - Protocol Independent Multicast – Dense Mode (PIM-DM)
  - Protocol Independent Multicast – Sparse Mode (PIM-SM)
  - Core Based trees (CBT)

- Inter-domain for use when multicast spans a set of autonomous systems
  - Multiprotocol extensions to BGP (MBGP)
  - Multicast Source Discovery Protocol (MSDP)
  - Border Gateway Multicast Protocol (MGMP)

Up until this point, however, multicast was mainly an academic exercise within a single flat topology network where only a few experiments had been carried out such as streaming the audio of the 1992 IETF meeting to 20 sites on the Mbone and a “Rolling Stones” concert streamed over the Mbone in 1994.

The lack of Multicast growth on a wide-scale in the Internet was be attributed to ISP's not understanding how to charge for the service and worries about scaling issues in Internet routers that needed to keep a large amount of state information.

Moving to the present time, the recent surge of video on demand, Internet radio and IPTV has created some interest in using Multicast again to ensure more efficient delivery of these services. For example the BBC encourages Internet service providers to adopt multicast-addressable services in their networks by providing BBC Radio at higher quality than is available via their unicast-addressed services.

Overview

Multicast presents a very attractive solution to the problem of distributing content in a one-to-many fashion which is often required when, for example, a video is broadcast over a network to multiple users. However in order for Multicast to function effectively, it needs to break the foundational end-to-end argument that has been vigorously enforced by Internet Protagonists. This is because state

needs to be maintained in the network about which receivers are part of a multicast group address.

This article sought to describe the early evolution of intradomain multicast and a new set of next generation protocols for interdomain multicast.

Originally multicast had focussed on a flat topology which was not suitable for the hierarchical structure of the Internet. The Distance Vector Multicast Routing Protocol (DVMRP) was designed for these original flat topology networks. It created a multicast tree using a broadcast-and-prune mechanism. This allowed a reverse shortest path tree to be created by the source broadcasting packets and the leaf routers sending prune messages back to the source when no members exist. This mechanism was designed for topologies with densely populated group members where prune messages are rarely required and the state information stored for each source in every router is not considered unnecessary overhead.

Other dense mode protocols that were developed in the 90's were Multicast Extensions to OSPF (MOSPF) and Protocol Independent Multicast – Dense Mode (PIM-DM). MOSPF was essentially a link-state protocol like OSPF which flooded an OSPF area with information about group receivers allowing all MOSPF in the area to have the same view of group membership. This then allowed an MOSPF router to construct a shortest path tree for each source and group. PIM-DM makes use of whatever unicast routing table is available but results in some unnecessary packets being forwarded to routers which then need to generate prune messages.

When there are only a few widely distributed group members over a network topology, the dense mode protocols become inefficient and sparse mode protocols were required. Two protocols were designed for this type of network: Core Based Trees (CBT) and Protocol Independent Multicast – Dense Mode (PIM-SM). The key difference with sparse mode protocols was that receivers were expected to send explicit join messages to a router acting as a core and sources traffic would only reach receiver nodes that joined the multicast group.

CBT created a single shared tree which was rooted at the core router. It did not use shortest path trees but allowed bidirectional traffic on shared trees. PIM-SM also made use of whatever routing table existed but constructed the tree using the following mechanism: A Rendezvouz Point (RP) is configured for each multicast group and was discovered using a bootstrap protocol which included mechanisms for alternative RPs to be selected if the primary RP went down. Receivers sent explicit join messages via a reverse shortest path tree to the RP. PIM-SM is now widely adopted by routers whereas CBT is rarely used.

Sparse mode protocols had better scalability because less routing state needed to be stored and multicast traffic only flowed across links explicitly added to the tree. But the use of a core or RP made sparse mode protocols vulnerable to a single point of failure, traffic hot spots and non-optimal paths. These problems were solved somewhat in PIM-SM by the RP discovery process and the option to switch from a shared tree to a shortest path tree.

In order to go beyond the managed flat virtual topology of the MBone, an evolution to hierarchical interdomain routing was required similar to the evolution that occurred in the early days of the Internet. This solution was to move towards route aggregation, and hierarchical routing similar to that found in the interdomain Border Gateway Protocol (BGP)

Extending BGP to carry multicast routes between AS's would allow a service provider to use different topologies for unicast and multicast traffic. To this end a protocol called Multiprotocol Extrenstions to BGP (MBGP) was conceived which meant that instead of every router needing to know the entire flat multicast topology, it only needed to know the topology of its own domain and

the best interfaces to use to reach other domains. MBGP didn't provide a mechanism to construct multicast trees across domains and further work was required.

To address this issue the Multicast Source Discovery Protocol (MSDP) allowed representatives (the RP) in each domain to announce the existence of active sources to other domains by sending Source Active (SA) messages to all directly connected MSDP peers which in turn forwards to its connected MSDP peers. There were however two challenges to overcome: There would be join latency because of the delay between new receivers joining and hearing the next SA message which was sent periodically. Bursty sources would sometimes not reach new receivers because of the time between forwarding SA messages and other RP's establishing forwarding state. This occured when the time between packets bursts exceeds the forwarding state timeout value. The solution in the first case was to cache SA messages and in the second case to include some data packets in the SA message. Both of these were not very elegant solutions and led to some continued debate in the working MSDP group. MSDP's lack of scalability due to the large number and size of SA messages being flooded led people to believe that it was an intermediate solution and not a long term solution

A long term solution solution explored by the Border Gateway Multicast Protocol (BGMP) was to construct bidirectional shared trees between domains using a single root. The root would be placed in the domain that owned the address of the particular group. This has the advantage that the group will be rooted at the primary source and remove dependency problems. Addressing allocation became a new challenge in this sort of scheme and a strict address allocation scheme was required. Some of the proposed schemes were Multicast Address-Set Claim (MASC) which supported address allocation between domains. It supported the three layer address allocation scheme of the Multicast Address Allocation Architecture (MAAA) in which address allocation is broken into interdomain, intradomain and between hosts and network. GLOP is a simpler scheme which statically allocates multicast addresses to each AS where only 8 bits or 256 unique addresses are made available.

A number of other simpler protocols, which make the source the root of the tree were discussed: Root Addressed Multicast Architecture (RAMA), Express Multicast and Simple Multicast. Express Multicast could collect information about subscribers which made it ideal for applications like TV broadcasts and Simple Multicast allowed multiple sources per group. The author was uncertain about the future of these protocols and a recent scan of the status of Multicast shows that they didn't gain any widespread support.

The paper ended with a description of how interdomain multicast was to be deployed in the Internet. The solution was to make MBone its own AS running MBGP/PIN-SM/MSDP and then communicate with other AS's which also run these interdomain multicast protocols. MBone would also slowly remove its tunnels as more interdomain multicasting became active. There was evidence that this was already occuring by observing the decrease in the number of MBone routes and the increase in the number of MBGP routes. Internet 2 specified that all deployed multicast was not allowed to use tunnels and routers had to support interdomain multicast routing using MBGP or MSDP. Both the vBNS and Abilene Internet 2 networks were MBGP/PIM-SM/MSDP capable and had peering agreements with other networks although vBNS had progressed further with its multicast deployment.

Paper analysis

The paper gave a good birds-eye view of the current state of Multicast routing in 2000 in terms of standardization efforts, what was available, what was implemented in the US and what was possibility to come. One criticism is that the paper was very US-centric; it would have been good to see some mention of Multicast efforts in TEN-155 in Europe (Now Geant).

There was mention of ASes being very careful about what routing information they exchange but this is largely ignored in all future discussion about information that needs to be exchanged between ASes for interdomain multicasting.

In discussing Multicast from the point of view of an AS, the chicken and egg problem prevails. Why would ASes be interested in supporting extra Multicast routing traffic if users are not demanding multicast but then follows the adage, “they don't know what they've got until they've got it”. So perhaps the ball is in the court of the content providers to incentivize multicast as was done in the case of the BBC offering better quality streaming when multicast is used, which encourages users to force their ISP's to support multicast.

The discussion of CBT, which seemed like a technically sound protocol, was very brief. More could have been said about bidirectional shared trees and some analysis of when this could be better or worse than shortest path trees. A diagram comparing CBT and PIM-SM would also have made this clearer.

What would have helped in understanding multicast uptake in the Internet was some measure of the percentage of multicast data traffic collected at RPs. This is to distinguish between multicast routes being formed out of pure curiosity or as an academic exercise and actual usage.

There were a few layout problems in the paper, such as the Multiple Source Discovery Protocol section starts with a summary of the previous section. This title should have been moved to the paragraph above Figure 4.