Stardust.com, Inc.
1901 S. Bascom Ave, Suite 333
Campbell, California 95008
Phone: 408-879-8080
Fax: 408-879-8081

www.stardust.com

A Survey of the History of Internet Multicast

PLATINUM SPONSORS

GOLD SPONSORS

PUBLICATION SPONSORS

A Survey of the History of Internet Multicast

Abstract

Multicast communication---the one-to-many or many-to-many delivery of data---has become a hot topic. It is of interest in the research community, standards groups, and to network service providers. Even though much work has been done on multicast, there are still issues that have not been completely resolved. One result is that protocols are still evolving and standards are not yet complete. From a deployment perspective, the lack of standards has slowed progress, but efforts to deploy multicast as an experimental service are gaining momentum. The question now is how long it will be before multicast becomes a true Internet service. The goal of this paper is to describe the past, present, and future of multicast. Starting with the Multicast Backbone (MBone), we describe how the emphasis has been on developing and refining intra-domain multicast routing protocols. Starting in the middle to late 1990s, particular emphasis has been placed on developing inter-domain multicast routing protocols. We provide a functional overview of the currently deployed solution. The future of multicast may hinge on several research efforts that are working to make the provision of multicast less complex by fundamentally changing the multicast model. We briefly survey these efforts. Finally, attempts

are being made to deploy native multicast routing in both Internet2 networks and the commodity Internet. We examine how multicast is being deployed in these networks.

Introduction

Multicast communication---the one-to-many or many-to-many delivery of data---has become a hot topic. It is the focus of intense study in the research community. It has become a highly desired feature of many vendors' network products. It is growing into a true deployment challenge for Internet engineers. It is evolving into a highly touted service being offered by some Internet Service Providers (ISPs). And finally, it is starting to be used by a number of companies offering large-scale Internet applications and services. From almost all perspectives, multicast is developing into an important Internet service.

For all the potential multicast has, and for all the advocacy multicast has received, there are still some questions and concerns. First, by Internet standards, multicast is an old concept, yet by most measures, deployment has been very slow. To put deployment in perspective, compare multicast to the World Wide Web (WWW) and the HyperText Transfer Protocol (HTTP). IP multicast was first introduced in Steve Deering's Ph.D. dissertation in 1988 and tested on a wide scale during an “audiocast” at the 1992 Internet Engineering Task Force (IETF) meeting in San Diego[1]. The first WWW browser was written in 1990, and in 1993 there were about 100 sites on the WWW. So while multicast and the WWW are roughly the same age, multicast is in its infancy[2] while the WWW's success, influence, and use seem totally pervasive. Second, IP multicast is one of the first services to be deployed which requires additional “intelligence” in the network. Multicast requires a non-trivial amount of state and complexity in both core and edge routers. These requirements are at odds with the long-standing Internet belief that intelligence should be pushed to the edges of the network. While many in the Internet community realize that the new generation of network services will put demands on the network, the difficulty is in deploying and managing these services in an infrastructure that has a lengthy history of only offering best-effort, unicast service.

Given these concerns, the image of multicast may seem somewhat tarnished. Is multicast then more trouble than its efficiency gains and economies of scale are worth? This question is especially relevant if multicast is to be used in commercial applications offered by companies who hope to make money. The challenges are to define elegant protocols, to support an infrastructure on top of which new applications can be developed, to continue to investigate new ways of increasing efficiency and reducing complexity. Doing multicast “the right way” is a noble endeavor and an appropriate long-term research topic, but the demand for working multicast has created an environment in which even short-term functional solutions are very attractive.

In this paper, we describe the past, present, and future of multicast. The history of multicast will help the reader understand how multicast has evolved into its current state. Topics include a description of the Multicast Backbone (MBone) and an overview of the common intra-domain multicast routing protocols. More recently, multicast evolution has been primarily focused in the area of inter-domain protocol development. Multicast in the present can be characterized as an effort to deploy multicast on a wide scale using a triumvirate of inter-dependent routing protocols. Deployment efforts have been carried out in the two Internet2 backbone networks---the very high speed Backbone Network Service (vBNS) and Abilene---as well as in the commodity Internet (so designated in order to distinguish it from Internet2 networks). The future of multicast is rooted in the continued development, evaluation, and standardization of new protocols. However, unlike current efforts, which are focused primarily on routing, future efforts are likely to include other issues such as address allocation, management, and billing[3].

The remainder of this paper is organized as follows. Section 2 describes the early evolution of multicast, in particular the development of intra-domain multicast. The focus of Section 3 is on inter-domain multicast, including the best current practices and several of the efforts to define the next generation of protocols. Section 4 details inter-domain deployment efforts in the commodity Internet and in Internet2 networks. Section 5 is the conclusion of the paper.

Multicast in the Beginning: The Flat MBone

The early history of multicast took place from the first Internet-wide experiments in 1992, to the middle of 1997. During this time standardization and deployment in multicast focused on a single flat topology. The "flat" characteristic is in contrast to the Internet topology, which is based on a hierarchical routing structure. The initial multicast protocol research and standardization efforts were aimed at developing routing protocols for this flat topology. Beginning in 1997, when the multicast community realized the need for a hierarchical multicast infrastructure and inter-domain routing, the existing protocols were categorized as intra-domain protocols and work began on standardizing an inter-domain solution. In this section, we describe the standard IP multicast model and the evolution and characterization of intra-domain multicast protocols.

The First IP Multicast Model

Stephen Deering was the first to describe the standard multicast model for IP networks[4]. This model describes how end systems are to send and receive multicast packets. The model includes both an explicit set of requirements and several implicit requirements. An understanding of this model can give insight on why multicast is characterized as it is. The model is as follows[5]:

- **IP-Style Semantics:** A source can send multicast packets at any time, with no need to register, announce or schedule transmission. IP multicast is based on UDP (not TCP), so packets are delivered using a best-effort policy without reliability or congestion control.
- **Open Groups:** Sources only need to know a multicast address. They do not need to know group membership, and they do not need to be a member of the multicast group to which they are sending. A group can have any number of sources.
- **Dynamic Groups:** Multicast group members can join or leave a multicast group at will. There is no need to register, synchronize, or negotiate with a centralized group management entity.

This IP multicast model is an end-system specification and does not discuss requirements for how the network should perform routing. The model also does not specify any mechanisms for providing quality of service, security, or address allocation.

The Multicast Backbone

Interest in building a multicast-capable Internet, motivated by Deering's work[4], achieved critical mass in the late 1980s. The culmination of this work was the creation of multicast in the Internet[6] and the creation of the
Multicast Backbone (MBone)[7,8]. In March 1992, the MBone carried audio from the Internet Engineering Task Force (IETF)[1] in San Diego to 20 sites. While the conferencing software itself represented a considerable accomplishment, the most significant achievement was the deployment of a virtual multicast network.

The multicast routing function was provided by workstations running a daemon process called *mrouted* (pronounced m-route-d), which received unicast-encapsulated multicast packets and forwarded them over an appropriate set of outgoing interfaces. Connectivity among these machines was provided using point-to-point, IP-encapsulated *tunnels*. Each tunnel connected two endpoints via one logical link, but a tunnel could cross several Internet routers. Once a packet is received, it can be sent to other tunnel endpoints or broadcast on a local interface. Routing decisions were made using the Distance Vector Multicast Routing Protocol (DVMRP)[9]. An example of connectivity provided via a virtual topology is shown in Figure 1. In this earliest phase of the MBone, all tunnels were terminated on workstations, and the MBone topology was such that sometimes multiple tunnels ran over a common physical link. Multicast routing in the early MBone was actually a controlled form of flooding. The first versions of *mrouted* did not implement pruning. All recent versions incorporate pruning and it is now fully deployed.

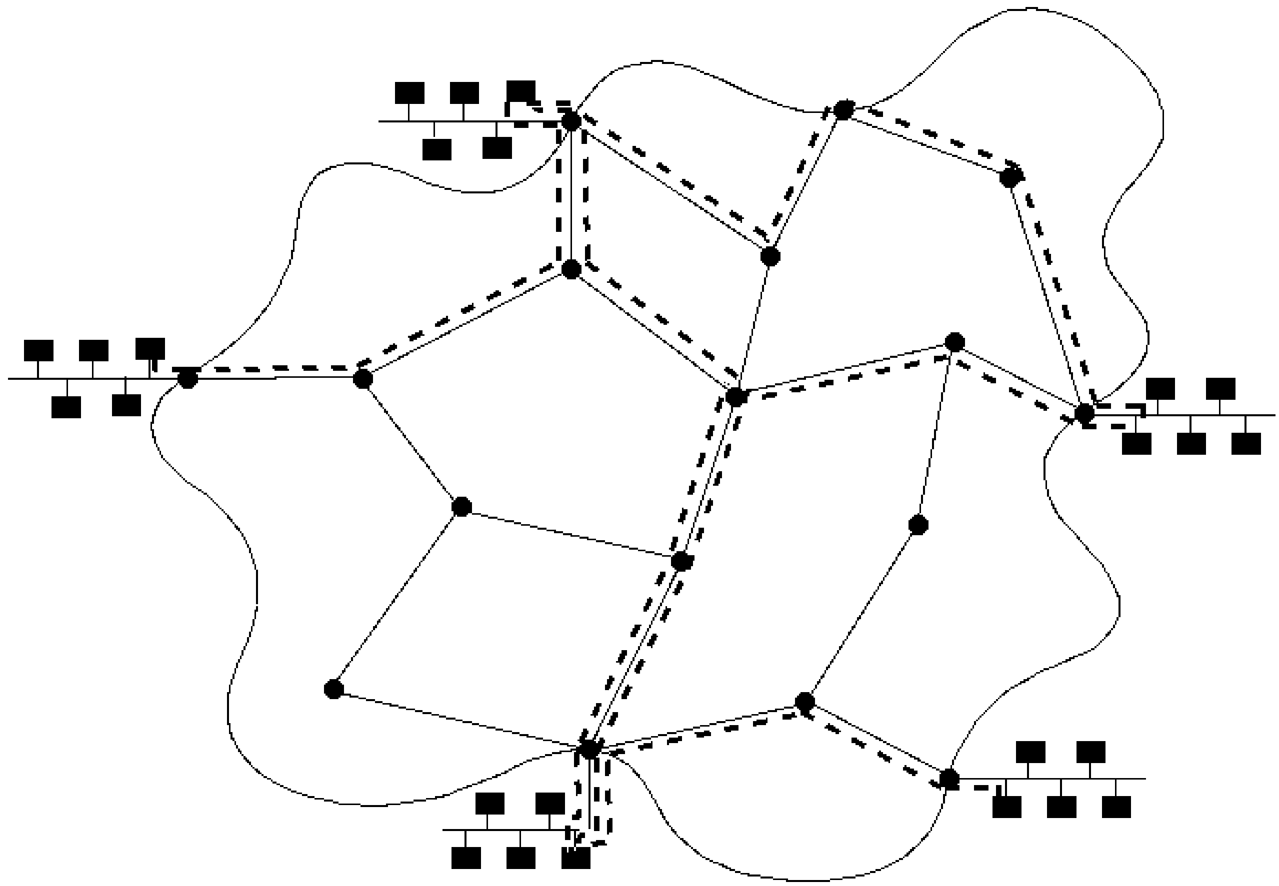

Figure 1: Generic tunnel-based topology representative of the early MBone.

The original multicast routing protocol, DVMRP, creates multicast trees using a technique known as *broadcast-and-prune*. Because of the way the tree is constructed by DVMRP, it is called a *reverse shortest path tree*. The steps to creating this type of tree are as follows:

1. A source broadcasts packets via its network card to the local network. The first-hop router receives the packet and sends it on all outgoing interfaces.

2. Each router receiving a packet performs a Reverse Path Forwarding (RPF) check. That is, each router checks to see if the incoming interface on which a multicast packet is received is the interface the router would use as an outgoing interface to reach the source. In this way, a router will choose to only receive packets on the one interface that it believes is the most efficient path back to the source. All packets received on the proper interface are forwarded on all outgoing interfaces. All others are discarded silently.

3. If a router has outgoing interfaces that are local networks, these routers are called *leaf routers*. A leaf router will check to see if it knows of any group members on its local interfaces. A router discovers the existence of group members by periodically issuing Internet Group Management Protocol (IGMP)[5,10,11] queries. If there are members, the leaf router forwards the multicast packet on the subnet. Otherwise, the leaf router

will send a *prune message* toward the source on the RPF interface, i.e., the interface the leaf router would use to forward packets to the source.

4. Prune packets are forwarded back toward the source, and routers along the way create prune state for the interface on which the prune message is received. If prune messages are received on all interfaces except the RPF interface, the router will send a prune message of its own toward the source.

In this way, reverse shortest path trees are created. These trees are constructed using the routing topology and can even be created for a virtual topology like the MBone. Broadcast-and-prune protocols are also known as *dense mode* protocols, because they are designed to perform better when the topology is densely populated with group members. Routers assume there are group members downstream, and so forward packets. Only when explicit prune messages are received does a router not forward multicast traffic. If a group is densely populated, routers are unlikely to ever need to prune. The key disadvantage of dense mode protocols is that state information must be kept for *each* source at *every* router in the network, regardless of whether downstream group members exist. If a group is not densely populated, significant state must be stored in the network and a significant amount of bandwidth may be wasted.

Progression of Intra-Domain Multicast

The MBone grew rapidly beginning in 1992 and continuing in the late 1990s. It was no longer a simple virtual network sitting on top of the Internet, but was rapidly becoming integrated into the Internet itself. In addition to simple DVMRP tunnels between workstations, the MBone began to have *native* multicast capability, i.e., routers are capable of handling multicast packets (see Figure 2). Furthermore, ongoing research has led to the development and deployment of two additional dense mode protocols. These are described below.

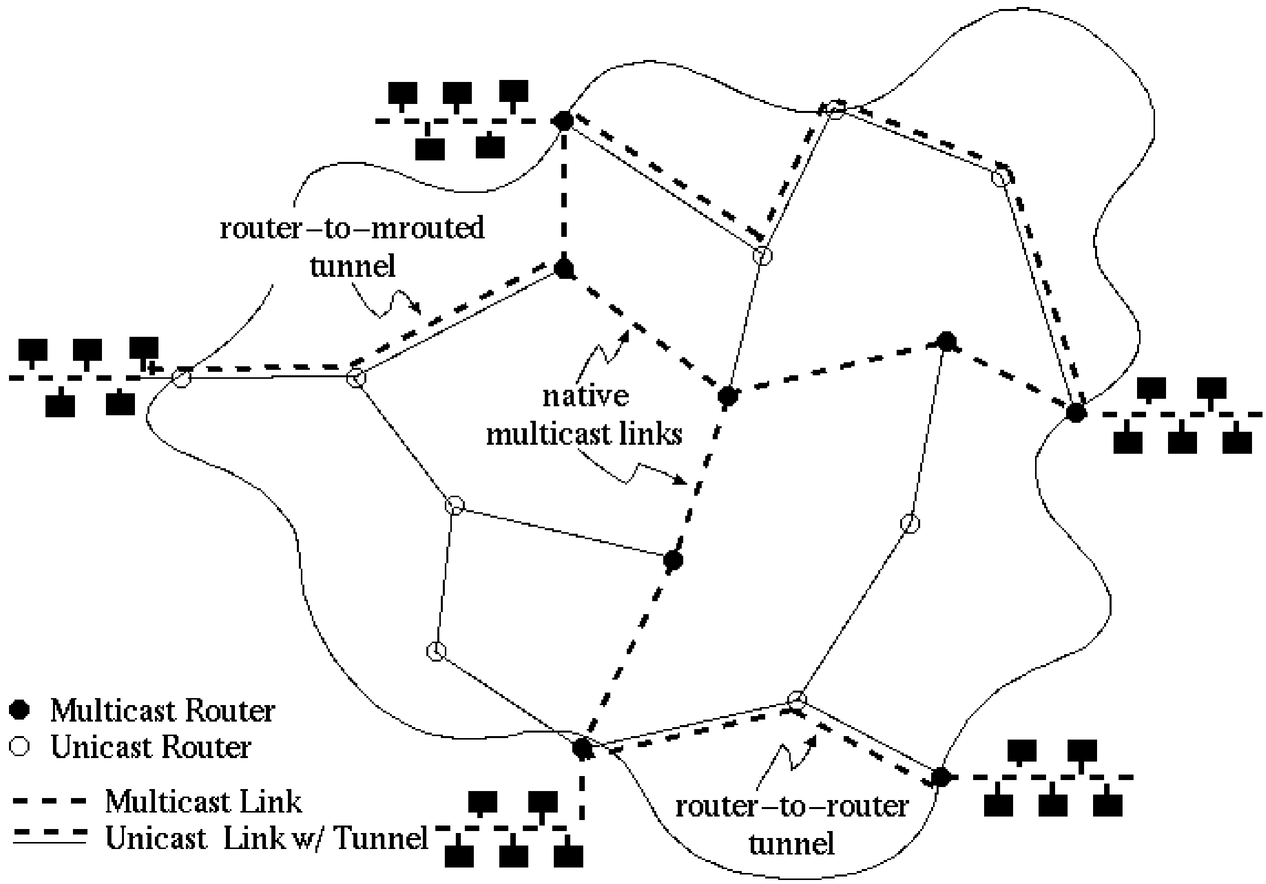

Figure 2: Example multicast topology with a combination of tunnels and native multicast links.

MOSPF. Multicast Extensions to OSPF (MOSPF)[12] uses the Open Shortest Path First (OSPF)[13] protocol to provide multicast. Basically, MOSPF routers flood an OSPF area with information about group receivers. This allows all MOSPF routers in an area to have the same view of group membership. In the same way that each OSPF router independently constructs the unicast routing topology, each MOSPF router can construct the shortest path tree for each source and group. While group membership reports are flooded throughout the OSPF area, data is not. MOSPF is something of an oddity in terms of classification. It is considered a dense mode protocol because membership information is broadcast to each MOSPF router, but it is also considered an explicit join protocol because data is only sent to those receivers that specifically request it. The key to understanding MOSPF is to realize that it is heavily dependent on OSPF and its link state routing paradigm.

PIM-DM. Protocol Independent Multicast (PIM)[14] has been split into two protocols, a dense mode version called PIM-DM[15] and a sparse mode version called PIM-SM[16] PIM-DM is very similar to DVMRP; there are only two major differences. The first is that PIM (both dense mode and sparse mode) uses the unicast routing table to perform RPF checks. While DVMRP maintains its own routing table, PIM uses whatever unicast table is available. The name PIM is derived from the fact that the unicast table can be built using any unicast routing algorithm. PIM simply requires the unicast routing table to exist, and so is *independent* of the algorithm used to build it. The second difference between PIM-DM and DVMRP is that DVMRP tries to avoid sending unnecessary packets to neighbors who will then generate prune messages based on a failed RPF check. The set of outgoing interfaces built by a given DVMRP router will include only those downstream routers that use the given router to reach the source (successful RPF check). PIM-DM avoids this complexity, but the tradeoff is that packets are forwarded on all outgoing interfaces. Unnecessary packets are often forwarded to routers that must then generate prune messages because of the resulting RPF failure.

Intra-domain routing protocols evolved to address the disadvantages of dense mode protocols. A new class of protocols, called *sparse mode* protocols, was created. Instead of optimizing for dense groups, sparse mode protocols are designed to work more efficiently when there are only a few widely distributed group members. Instead of broadcasting traffic and triggering prune messages, receivers send explicit join messages. These join messages are sent to a router acting as a *core*. Sources are expected to send their data traffic to this same node. The use of a core as a "meeting place" for sources and receivers facilitates creation of the multicast tree. Two of the most popular sparse mode protocols are described below.

CBT. The Core Based Trees (CBT) protocol began its life as a research paper[17] and is now being standardized by the IETF[18]. CBT uses the basic sparse mode paradigm to create a single *shared tree* used by all sources. The tree is rooted at a core. All sources send their data to the core and all receivers send explicit join messages to the core. There are two differences between CBT and PIM-SM. First, CBT uses only a shared tree, and is not designed to use shortest path trees. Second, CBT uses *bi-directional* shared trees, but PIM-SM uses *unidirectional* shared trees. Bidirectional shared trees involve slightly more complexity, but are more efficient when packets traveling from a source to the core cross branches of the multicast tree. In this case, instead of only sending traffic "up" to the core, packets can also be sent "down" the tree. While CBT has significant technical merits and is on par technically with PIM-SM, few routing vendors provide support for CBT. The reason seems to be that the vendor community was only going to support one sparse mode protocol and the implicit selection was PIM-SM.

PIM-SM. PIM-SM[16] is much more widely used than CBT. It is similar to PIM-DM in that routing decisions are based on whatever underlying unicast routing table exists, but the tree construction mechanism is quite different. PIM-SM's tree construction algorithm is actually more similar to that used by CBT than to that used by PIM-DM. In the following description of sparse mode protocol operation, we use PIM-SM as our example.

1. A core, called a rendezvous point (RP) in PIM terminology, must be configured[19]. Different groups may use different routers for RPs, but a group can only have a single RP.

    - Information about which routers in the network are RPs, and the mappings of multicast groups to RPs, must be discovered by all routers.

    - RP discovery is done using a bootstrap protocol. However, because the RP discovery mechanism is not included in the PIM-SMv1 specification, each vendor implementation of PIM-SMv1 has its own RP discovery mechanism. For PIM-SMv2, the bootstrap protocol is included in the protocol specification.

- The basic function of the bootstrap protocol, in addition to RP discovery, is to provide robustness in case of RP failure. The bootstrap protocol includes mechanisms to select an alternate RP if the primary RP goes down.

2. Receivers send explicit *join messages* to the RP. Forwarding state is created in each router along the path from the receiver to the RP. A single shared tree, rooted at the RP, is formed for each group. As with other multicast protocols, the tree is a reverse shortest path tree---join messages follow a reverse path from receivers to the RP.

3. Each source sends multicast data packets, encapsulated in unicast packets, to the RP. When an RP receives one of these *register packets*, a number of actions are possible. First, if the RP has forwarding state for the group, i.e., there are receivers who have joined the group, the encapsulation is stripped off the packet and it is sent on the shared tree. However, if the RP does not have forwarding state for the group, it sends a *register-stop message* to the RP. This avoids wasting bandwidth between the source and the RP. Second, the RP may wish to send a join message toward the source. By establishing multicast forwarding state between the source and the RP, the RP can receive the source's traffic as multicast and avoid the overhead of encapsulation.

The basic goal of a sparse mode protocol is to use the RP as a “meeting place” for sources and receivers. Receivers explicitly join the shared tree, and sources register with the RP.

Sparse mode protocols offer a number of advantages over dense mode protocols. First, sparse mode protocols offer better scalability in terms of routing state. Only routers on the path between a source and a group member must keep state. Dense mode protocols require state in all routers in the network. Second, sparse mode protocols are more efficient because the use of explicit join messages means multicast traffic only flows across links that have been explicitly added to the tree.

Sparse mode protocols do have their few disadvantages. These are mostly related to the use of RPs. First, the RP can be a single point of failure. Second, the RP can become a hot spot for multicast traffic. Third, having traffic forwarded from a source to the RP and then to receivers means that non-optimal paths may exist in the multicast tree. The first problem is mostly solved with a bootstrap protocol that offers redundancy and fail over. The second and third problems are solved in CBT by using bidirectional trees. PIM-SM solves these problems by providing a mechanism to switch from a shared tree to a shortest path tree. This change occurs when a traffic rate threshold is violated. A leaf router will send a special message towards the source. Forwarding state is changed so traffic flows directly to the receiver from the source, instead of through the RP.

For clarity, it is worth summarizing the key multicast terminology. Multicast protocols use either a broadcast-and-prune or an *explicit join* mechanism. Broadcast-and-prune protocols are commonly called **dense mode protocols** and always use a **reverse shortest path tree** rooted at a source. Explicit join protocols, commonly called **sparse mode protocols**, can use either a reverse shortest path tree or a *shared tree*. A shared tree uses a **core** or a **rendezvous point** to bring sources and receivers together.

Weaknesses in the MBone

As the MBone has grown, it has suffered from an increasing number of problems, and these problems have been occurring with increasing frequency. The most important reason for this is the growing difficulty of managing a flat virtual topology. The same problems experienced with class-based unicast routing have manifested themselves in the MBone. As the MBone has grown, its size has become a problem, in terms of both routing state and susceptibility to misconfigurations. As a result, the multicast community has realized the need to deploy hierarchical, inter-domain routing.

Scalability. Large, flat networks are inherently unstable. Exacerbating this problem are organizational mechanisms that do not provide significant route aggregation. For these two reasons, the MBone has experienced substantial scalability problems. At its peak, the MBone had almost 10,000 routes. Unfortunately, most of these routes had long prefixes (between /28 and /32), which meant that very few hosts could be represented in each routing table entry. These scalability problems are not new. As the Internet has grown, unicast routing had to be fundamentally changed to enable continued growth and stability. The solutions---route aggregation and hierarchical routing---have proven successful, and the issue now is how to apply them to multicast.

Manageability. As the MBone has grown, it has become harder to manage. The MBone has no central management, and most tasks have been handled on a per-site basis. Most coordination takes place via the MBone mailing list. Because the MBone is a virtual topology and new sites can be connected anywhere, there should be a formal procedure for adding new sites. Because no such mechanism exists, the MBone has grown randomly, and there are many inefficiencies. Two types of inefficiency commonly observed are:

- **Virtual Topology Management:** The MBone is characterized as a set of multicast-capable islands connected by tunnels. The goal has always been to connect these islands in the most efficient manner, but over time sub-optimal tunnels have been created. Tunnels are often set up in very inefficient ways 1 for several examples). This behavior was observed very early in the history of the MBone, especially with regard to the MCI Backbone. To avoid the growing tangle of tunnels, engineers at MCI undertook the difficult task of enforcing a policy that tunnels through or into the MCI network would have to be terminated at designated border points. The goal was to resolve the observed problem of single physical links being crossed by several (up to 10) tunnels. The work of the MCI engineers set an example that helped keep the MBone reasonably efficient for a number of years.

- **Inter-Domain Policy Management:** Domain boundaries are another source of problems when trying to manage a flat topology. The model in today's Internet is to establish Autonomous System (AS) boundaries between Internet domains. ASes are commonly managed or owned by different organizations. Entities in one AS are typically not trusted by entities in another AS. As a result, exchange of routing information across AS boundaries is handled very carefully. Peering relationships among ASes are provisioned using the Border Gateway Protocol (BGP), which provides routing abstraction and policy control[20,21,22]. As a result of wide-scale use of BGP there is a commonly accepted procedure when two ASes wish to communicate. Because the MBone does not provide such an inter-domain protocol, it offers no protection across domain boundaries. When there is a single flat topology connected using tunnels, routing problems can easily spread throughout the topology.

Evolution of Inter-Domain Multicast

Inter-domain multicast has evolved out of the need to provide scalable, hierarchical, Internet-wide multicast. Protocols that provide the necessary functionality have been developed, but the technology is relatively immature. These protocols are being considered by the IETF, while simultaneously being evaluated through extensive deployment. The particular inter-domain solution in use is considered near-term, and is possibly only an interim solution. While the solution is functional, it lacks elegance and long-term scalability. As a result, additional work is underway to find long-term solutions. Some of these proposals are based on the standard IP multicast model. Others attempt to refine the service model in hopes of making the problem easier.

Near-term solution

The near-term solution for inter-domain multicast routing has three parts. The first is a straightforward extension of the inter-domain unicast route exchange protocol BGP. The second and third are additional protocols needed to build and interconnect trees across domain boundaries.

Multicast Routes in BGP

The first requirement follows from the need to make multicast routing hierarchical in the same manner as unicast routing. Route aggregation and abstraction, as well as hop-by-hop policy routing, are provided in unicast using the Border Gateway Protocol (BGP)[21]. BGP offers substantial abstraction and control among domains. Within a domain, a network administrator can run any routing protocol desired. Routing to hosts in an external domain is simply a matter of choosing the best external link.

BGP supports inter-domain routing by reliably exchanging network reachability information. This information is used to compute an end-to-end distance-vector-style path of AS numbers. Each AS advertises the set of routes it can reach and an associated cost. Each border router can then compute the set of ASes that should be traversed to reach any network. The use of a distance vector algorithm together with full path information allows BGP to overcome many of the limitations of traditional distance vector algorithms. Packets are still routed on a hop-by-hop basis, but less information is needed and better routing decisions can be made.

The functionality provided by BGP, and its well-understood paradigm for connecting ASes, are important catalysts for supporting inter-domain multicast. A version of BGP capable of carrying multicast routes would not only provide hierarchical routing and policy decisions, but would also allow a service provider to use different topologies for unicast and multicast traffic.

The mechanism by which BGP has been extended to carry multicast routes is called Multiprotocol Extensions to BGP4 (MBGP)[23][1]. MBGP is able to carry multiprotocol routes by adding the Subsequent Address Family Identifier (SAFI) to two BGP4 messages: MP_REACH_NLRI and MP_UNREACH_NLRI. Specifically for multicast, the SAFI field can specify unicast, multicast or unicast/multicast. With MBGP, instead of every router needing to know the entire flat multicast topology, each router only needs to know the topology of its own domain and the paths to reach each of the other domains. Figure 3 shows an example of several domains connected together by MBGP sessions. In one case, two domains are connected together using different connections for unicast and multicast.

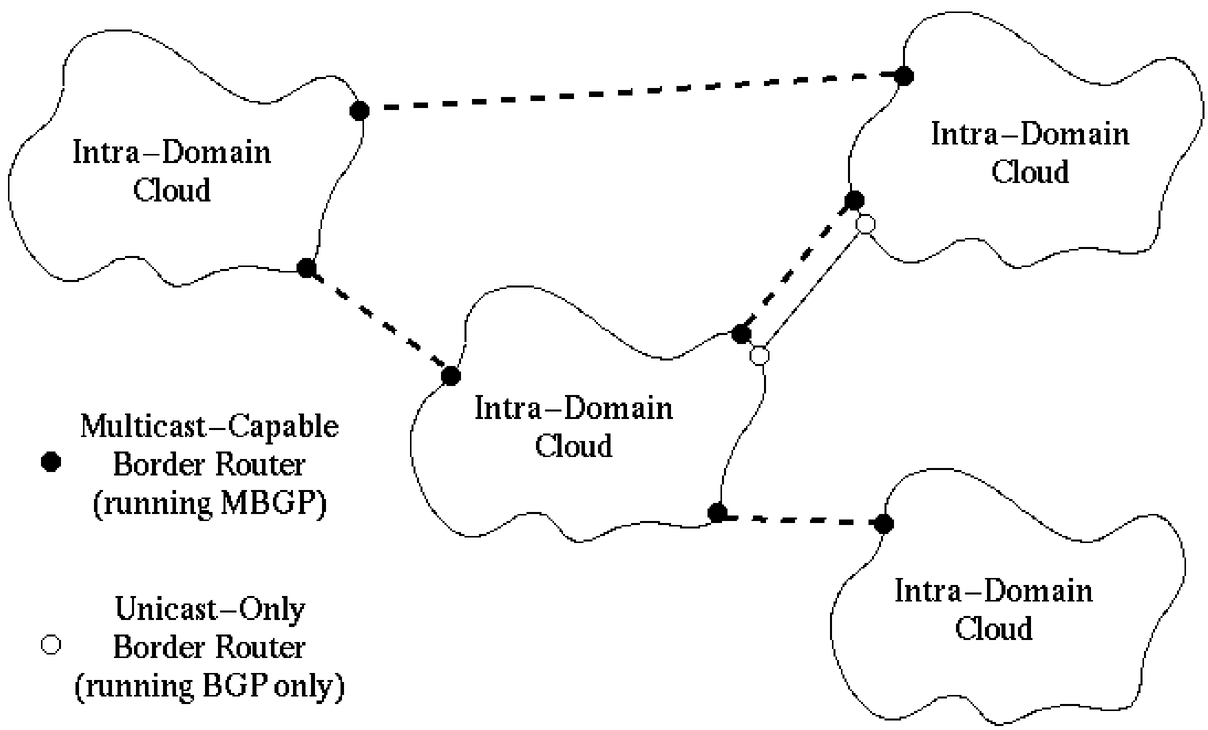

Figure 3: Inter-domain multicast topology running BGP and/or MBGP.

There is some confusion over exactly what functionality MBGP provides. To be clear, we offer the following example. If one domain advertises reachability for multicast, the message will say, "I have a path to sources on the networks listed in this message." It is important to understand that MBGP messages do not carry information about multicast groups, i.e., class D addresses are never carried in an MBGP message. Recall that multicast trees are constructed using a reverse path back to the source. Therefore, MBGP information is used when a join message is sent from an RP or receiver toward the source. This join message needs to know the best reverse path toward the source. MBGP provides this next-hop information between domains. If all unicast

[1] There is some ambiguity over terminology here. First, multiprotocol BGP4 is sometimes also referred to as BGP4+. Second, some think that MBGP stands for *Multicast* BGP. All three terms refer to the same protocol.

and multicast topologies were assumed to be the same, the reverse path join could simply follow the same next hop that any unicast traffic would follow. MBGP allows a network administrator to specify a different reverse path for the join to follow, and (subsequently) a different forward path when data is sent.

While MBGP is the first step toward providing inter-domain multicast, it alone is not a complete solution. MBGP is capable of determining the next hop to a host, but it is not capable of providing multicast tree construction functions. More specifically, what is the format of the join message? When should join messages be sent, and how often? Support for this functionality is not provided by MBGP; a true inter-domain multicast routing protocol is needed. Furthermore, conventional wisdom suggests that this protocol should not use the broadcast-and-prune method of tree construction. The near-term solution being advocated is to use PIM-SM, to establish a multicast tree between domains containing group members.

The Multicast Source Discovery Protocol

To summarize: various intra-domain routing protocols exist, there is a route exchange protocol to support multicast, and PIM-SM is to be used to connect receivers and sources across domain boundaries. But, there is still one function missing from the near-term solution. This function is needed when trying to connect sparse mode domains together. Given that PIM-SM is the only sparse mode protocol that has seen significant deployment, this function tends to be heavily influenced by PIM-SM. The problem is basically how to inform an RP in one domain that there are sources in other domains. The underlying assumption here is that a group can now have multiple RPs. However, the reality is that there is still only one RP per domain, but now multiple domains may be involved. The approach adopted is largely motivated by the perceived needs of the ISP community. In fact, the decision to have multiple RPs rather than a single root is what differentiates the near-term solution from other proposed solutions.

A problem arises when group members are spread over multiple domains. There is no mechanism to connect the various intra-domain multicast trees together. While traffic from all the sources for a particular group *within a particular domain* will reach the group's receivers, any sources outside the domain will remain disjoint. Why is this the case? Within a domain, receivers send join messages toward one RP and sources send register messages to the same RP. However, there is no way for an RP in one domain to find out about sources in other domains using different RPs. There is no mechanism for RPs to communicate with each other when one receives a source register message. This problem is summarized in Figure 4.

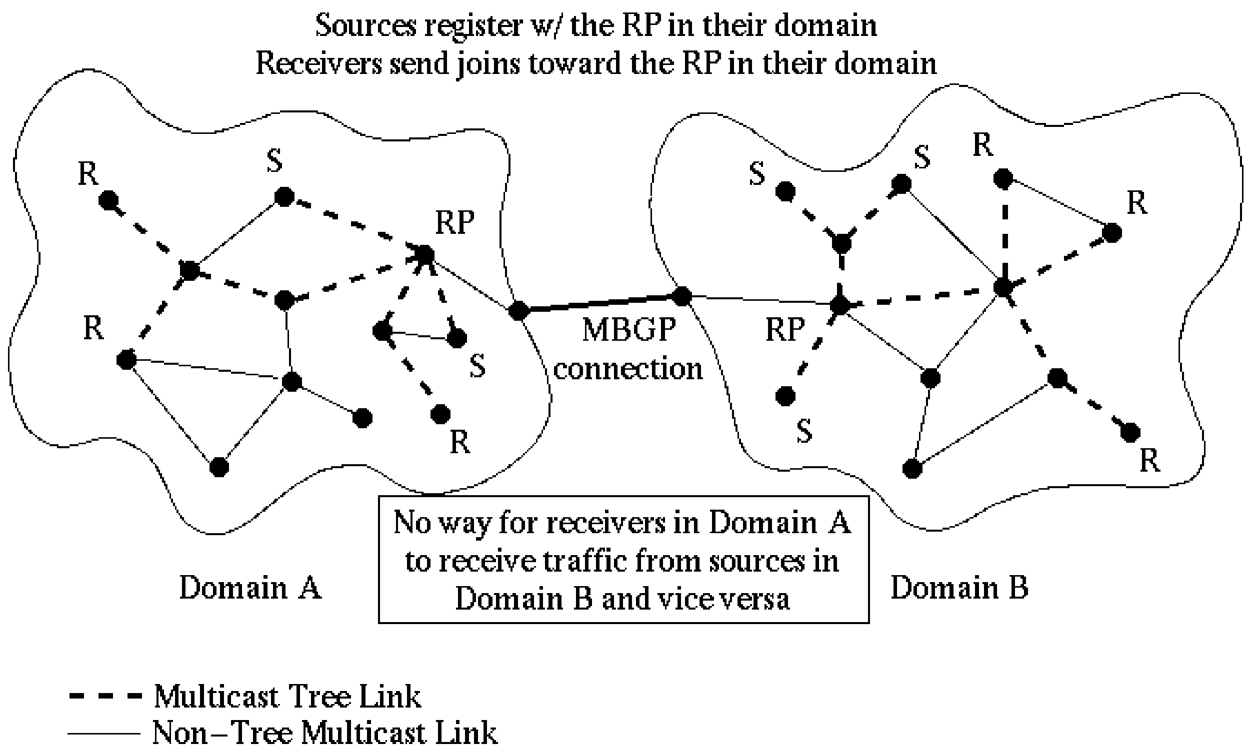

Figure 4: The problem of connecting sources and receivers across two sparse mode domains.

The decision to maintain a separate multicast tree and RP for each domain is driven by the need to reduce administrative dependencies between domains. Two potential problems are avoided in this way.

1. It is not necessary for two domains to co-administer a single sparse mode cloud. Relevant administrative functions include identifying candidate RPs and establishing the group-RP mapping.

2. It becomes possible to avoid multi-party dependencies, in which multicast delivery for sources and groups in one or more domains is dependent on another domain whose only function is to provide the RP. Dependencies can occur when all sources and receivers in the RP's domain leave or become inactive. The domain with the RP has no group members and yet is still providing the RP service. Depending on how multicast and inter-domain traffic billing is handled, this could be particularly undesirable.

The near-term solution adopted for this problem is a new protocol, appropriately named the Multicast Source Discovery Protocol (MSDP)[24]. This protocol works by having representatives in each domain announce to other domains the existence of active sources. MSDP is run in the same router as a domain's RP (or one of the RPs). MSDP's operation is similar to that of MBGP, in that MSDP sessions are configured between domains and TCP is used for reliable session message exchange. MSDP operation is described below, with each step shown in Figure 5.

1. When a new source for a group becomes active it will register with the domain's RP.

2. The MSDP peer in the domain will detect the existence of the new source and send a Source Active (SA) message to all directly-connected MSDP peers.

3. MSDP message flooding:

    - MSDP peers that receive an SA message will perform a *peer-RPF check*. The MSDP peer that received the SA message will check to see if the MSDP peer that sent the message is along the "correct" MSDP-peer path. These peer-RPF checks are necessary to prevent SA message looping.

    - If an MSDP peer receives an SA message on the correct interface, the message is forwarded to all MSDP peers except the one from which the message was received. This is called *peer-RPF flooding*.

4. Within a domain, an MSDP peer (also the RP) will check to see if it has state for any group members in the domain. If state does exist, the RP will send a PIM join message to the source address advertised in the SA message.

5. If data is contained in the message, the RP then forwards it on the multicast tree. Once group members receive data, they may choose to switch to a shortest path tree using PIM-SM conventions.

6. Steps 3-5 are repeated until all MSDP peers have received the SA message and all group members are receiving data from the source.

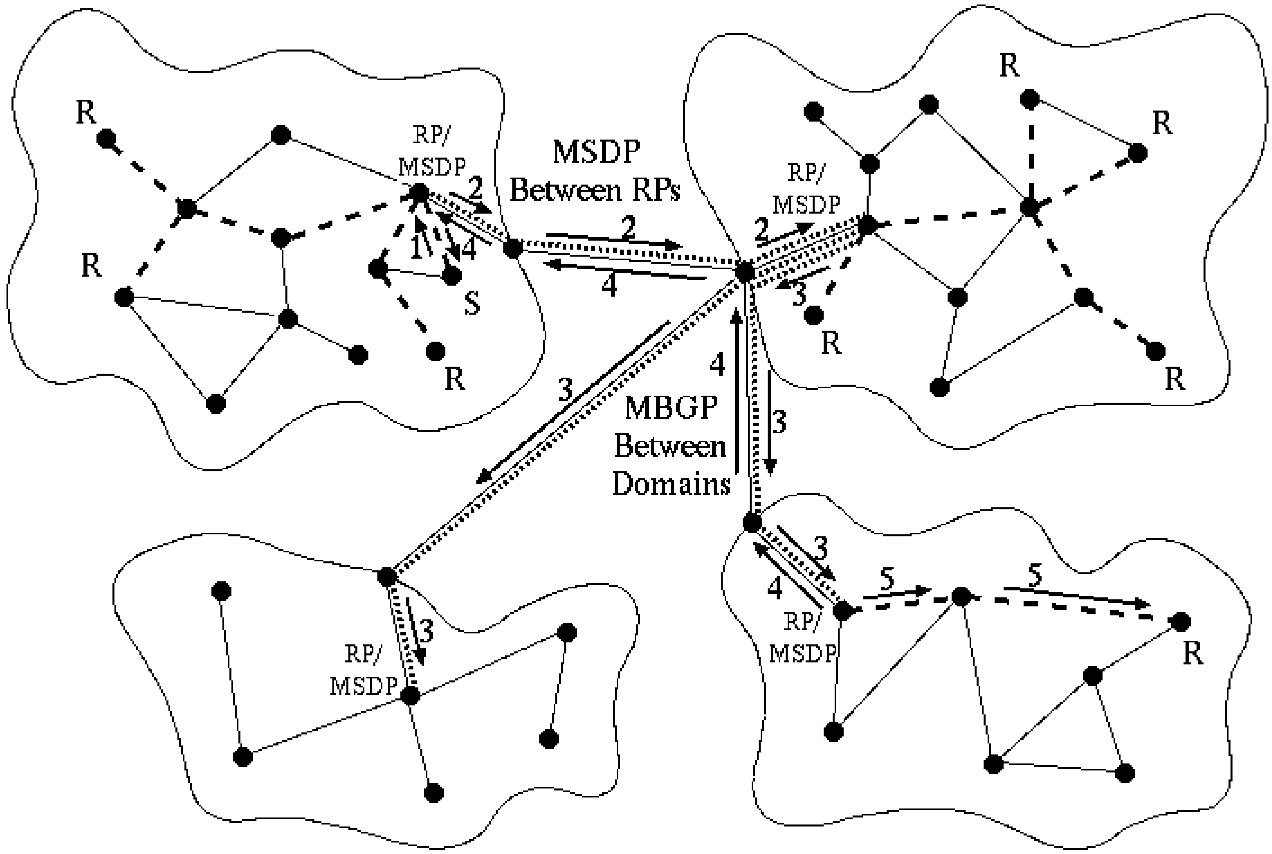

Figure 5: MSDP operation including flow of Source Active (SA) messages.

The short-term inter-domain solution just described is referred to with the abbreviations for the three relevant protocols: MBGP/PIM-SM/MSDP. However, while the given description is relatively complete, there are a number of details which are not discussed. And as with any system, most of the complexity is in the details. Furthermore, we have not yet discussed the limitations of the current solution in any detail. In particular, a qualitative assessment of the scalability, complexity, and overall quality of the protocols would be valuable.

The MBGP/PIM-SM/MSDP solution is relatively straightforward once a person understands all the abbreviations and understands the motivating factors that drove the design of the protocols. While some argue that the current set of protocols is not simple, it really is no more complex than many other Internet services, such as unicast routing. The key advantage of MBGP/PIM-SM/MSDP is that it is a functional solution largely built on existing protocols. Furthermore, it is already being deployed with a fair amount of success. The key disadvantage is that, as a long-term solution, the MBGP/PIM-SM/MSDP protocol suite may be susceptible to scalability problems. Further discussion of two particular problems follows.

Dynamic Groups with High Turnover. When multicast sources begin to transmit, the network is required to create routing state to control packet flow. We have already discussed how different types of multicast routing protocols accomplish this function. However, in the case of MSDP, information about the existence of sources must first be transmitted before routing state can be created. This extra complexity increases the overhead of managing groups. When groups are dynamic, due to either bursty sources or frequent group member join/leave events, the overhead of managing the group can be significant. A formidable task would be created for networks that must establish and remove information for thousands of sources and receivers scattered around the world. Two specific problems related to dynamic groups/sources are:

- *Join Latency*: Because MSDP SA messages are only sent periodically, there may be a significant delay between when new receivers join and when they hear the next SA message. To solve this problem, MSDP peers may be configured to cache SA messages. A non-caching MSDP peer can send an "SA-Request" message to an MSDP peer that does perform caching. This gives MSDP peers a mechanism to actively determine source, thereby reducing join latency. The tradeoff is the extra state and complexity of maintaining the cache.

- *Bursty Sources:* Some sources send short packet bursts separated by silent periods on the order of several minutes. One example is when a tool like *sdr* is used to periodically advertise a session. A problem potentially occurs when trying to establish a multicast tree for this kind of source. The problem begins when one or a few packets are sent to the RP. The RP will hear the packet and flood an SA message, and RPs in other domains will send join messages back to the source. However, because no multicast forwarding state existed when the packet was originally sent, and because it takes time to forward SA messages and have other RPs establish forwarding state, the original burst will not reach new receivers. Once state is established, all subsequent packets should reach these receivers. The problem occurs when the period of silence between packet bursts exceeds the forwarding state timeout value (typically 3 minutes). Because no packets are sent, the forwarding state is discarded. When another session announcement is sent, the same process of establishing state but losing the initial burst is repeated. In this way, no packets from bursty sources ever reach group members. The solution, specified in the MSDP protocol, is to have SA messages carry the first *n* data packets. This is not a particularly elegant solution, but it does solve the problem. The lack of elegance is making the protocol harder to standardize. Because data packets are delivered via SA messages, which are delivered over TCP connections, some in the multicast community wonder if this will have undesirable side effects or break assumptions of higher layer protocols. As a result, recent discussions in the MSDP working group have generated proposals that allow data to be carried in either GRE or UDP packets. The final decision on which data delivery options to support has not been made.

MSDP Scalability. The issue of scalability is an important one to consider for MSDP. Because of the way MSDP operates, if multicast becomes tremendously successful, the overhead of MSDP may become too large. The limitation occurs if multicast use grows to the point where there are thousands of multicast sources. The number of SA messages (plus data) being flooded around the network could become very large. The generally-agreed-upon conclusion is that MSDP is not a particularly scalable solution, and will likely be insufficient for the long term. But, given that long-term solutions are not ready to be deployed, MSDP is seen as an immediate solution to an immediate need.

Long-term proposals

While MBGP/PIM-SM/MSDP is a recognized near-term solution, there is still a need to continue researching and developing better solutions. Numerous such efforts are underway. These efforts can be broken down into two groups: efforts based on the standard IP multicast philosophy, and efforts which look to change this model in hopes of simplifying the problem. Two efforts, one in each of these areas are described next.

Long-Term Proposals

While MBGP/PIM-SM/MSDP is a recognized near-term solution, there is still a need to continue researching and developing better solutions. Numerous such efforts are underway. These efforts can be broken down into two groups: efforts based on the standard IP multicast philosophy, and efforts which look to change this model in hopes of simplifying the problem. Two efforts, one in each of these areas are described next.

Border Gateway Multicast Protocol

The Border Gateway Multicast Protocol (BGMP) was first proposed as a long-term solution for Internet-wide inter-domain multicast[25]. The key idea of BGMP is to construct bidirectional shared trees between domains using a single root. One of the functions of BGMP is then to decide in which particular domain to root the shared tree. BGMP relies on the belief that inter-domain dependencies can be avoided by using a strict address allocation scheme. Such an address allocation scheme allows domains to own specific addresses or specific ranges of addresses. The belief is that if a particular domain owns the address for a particular group, the domain will be significantly involved in the multicast service. Finally, this means dependency problems, even though there is a single root, should be highly unlikely. For example, a video-on-demand application will likely be rooted at the server; a video conference group will be rooted at the primary source or at a session coordinator. The belief is that no matter the type of session, one domain will always be the logical choice for the root domain.

As a result of a protocol like BGMP, there is a need for a strict address allocation scheme. “Strict” means that ownership must be clearly defined and that there cannot be collisions. Therefore, the *sdr* mechanism of randomly choosing an address is not sufficient. Because of BGMP, as well as demands from ISPs and application writers, work is being conducted to develop the necessary address allocation schemes. Before discussing two of the proposals for address allocation, it is worthwhile to make two points. First, BGMP is relatively flexible, and can use any scheme as long as it provides strict address allocation. Second, independent of BGMP, there is a need for better address allocation. The sdr mechanism is not particularly scalable and is no longer sufficient even for the current MBGP/PIM-SM/MSDP solution. Proposals, usable in both the current model and with BGMP, are being considered by the IETF. They are described below:

> **Multicast Address-Set Claim**. The MASC protocol supports address allocation between domains\cite[25]. MASC includes mechanisms to guarantee that address collisions are immediately resolved. From a more abstract perspective, MASC provides the functionality required at the highest layer of a more general addressing

scheme called the Multicast Address Allocation Architecture (MAAA)[26]. MASC and its supporting protocols are specific instances of protocols that meet the requirements of the MAAA specification. In MAAA, there are three levels of address allocation: at the domain level, within a domain, and between hosts and the network. Work to develop protocols at each level is underway in the IETF. MASC would act as a top-level address allocation protocol and operate between domains; the multicast Address Allocation Protocol (AAP)[27] would allocate addresses within a domain; and the Multicast Address Dynamic Client Allocation Protocol (MADCAP)[28] would be used by hosts to request addresses from a Multicast Address Allocation Server (MAAS).

Glop Addressing. Another, much simpler, proposal is to statically allocate multicast addresses to each AS. A "glop" of addresses is assigned to each AS. The AS number is encoded as part of the address[29]. The first version of Glop is being evaluated with only part of the 224/4 address range. Only the 233/8 address range is being used. As a result, the first octet is static, the next two octets encode the AS number, and the final octet provides a range of addresses to be allocated. This proposal is gaining in popularity, but it has two limitations. First, because only 8 bits, or 256 addresses, are available to each AS, there is likely to be an insufficient number of addresses per AS. This problem could be solved by using more of the Class D address space, or by switching to IPv6 addressing. The second problem is that Glop does not specify a mechanism by which addresses are allocated within the domain. This problem could be solved by using a simple administrative procedure, by using a dynamic protocol like AAP/MADCAP, or by using an modified, intra-domain version of *sdr*.

Root Addressed Multicast Architecture

In response to the perceived complexity of MBGP/PIM-SM/MSDP and BGMP, and to the need to address additional multicast-related issues like security, billing, and management[3], some members of the multicast community are looking to make fundamental changes to the multicast model. One class of proposals that has received much attention recently is called the Root Addressed Multicast Architecture (RAMA)[30]. The premise for RAMA-style protocols is that most multicast applications are single-source or have an easily identifiable primary source. By making this source the root of the tree, the complexity of core placement in other multicast routing protocols can be eliminated. This tradeoff raises a number of important issues that are described at the end of this section. There are two primary RAMA-style protocols being discussed: Express Multicast[31] and Simple Multicast[32]. The key aspects of these two protocols are:

Express Multicast. Express is designed specifically as a single-source protocol. The root of the tree is placed at the source, and group members send join messages along the reverse path to the source. Express also provides mechanisms to efficiently collect information about subscribers. The protocol is specifically designed for subscriber-based systems that use logical channels. Representative applications include TV broadcasts, file distribution, and any single-source multimedia application. The key advantages of Express are that routing complexity can be reduced and that *closed groups* can be offered.

Simple Multicast. Simple Multicast and Express Multicast are similar, but Simple Multicast has the added flexibility of allowing multiple sources per group. A particular source must be chosen as the primary, and the tree is rooted at this node's first-hop router. Receivers send join messages to the source, and a bidirectional tree is constructed. Additional sources send packets to the primary source. Because the tree is bidirectional, as soon as packets reach a router in the tree they are forwarded both downstream to receivers and upstream to the core. The advantages and disadvantages of this proposal are being heavily debated, but the proposal's authors believe that it eliminates the address allocation problem and the need to place and locate RPs. Address allocation is done by using the core address and the multicast group address together to uniquely identify a group. By routing on this pair of addresses, each root/core/source can allocate, without collision, up to 2^{32} addresses.

The Express and Simple multicast proposals have received significant attention in both the research community and the IETF. There is another question in addition to that of the merits of these new protocols. If these protocols are standardized, will they be expected to replace all existing protocols, or will they work in parallel with the existing multicast infrastructure? If the RAMA-style protocols are expected to work in cooperation with existing protocols, there will be yet another set of protocols to deploy, evaluate, and interoperate with. This does not make the provision of Internet-wide multicast easier. If RAMA-style protocols are expected to replace the current set of protocols, the question becomes whether they have enough flexibility to support all types of multicast applications. The bottom line is that these new protocols are still proposals, and it is uncertain what their future will be.

Recent Multicast Deployment Efforts

As inter-domain multicast is standardized and deployed, the effort to transition from the MBone to native multicast is growing. Strong efforts are underway in both Internet2 and the commodity Internet.

Commodity Internet Deployment

Measuring the success of inter-domain deployment, either from a qualitative point of view or by taking a count of connected hosts, is a difficult problem. Published studies have so far only dealt with the MBone, although several studies that distinguish between the MBone and inter-domain multicast are currently underway (See http://imj.ucsb.edu/mantra/). It is beyond the scope of this paper to offer any quantitative results. However, it is possible to describe the plan, now being implemented, to transition from the MBone's flat virtual topology to a true inter-domain multicast infrastructure.

Now that inter-domain multicast routing is possible, the issue is how to deal with the MBone. While the rest of the Internet is working to deploy inter-domain multicast, the challenge is how to bring MBone users into the new infrastructure. The solution has been to make the MBone its own AS, called AS10888. All MBone tunnels and sites connected by tunnels are relegated to AS10888. Connectivity between AS10888 and other multicast-capable ASes is provided at the NASA Ames Multicast-friendly Internet eXchange (MIX)[33]. The NASA Ames MIX provides connectivity between the MBone (AS10888) and all other ASes that have deployed MBGP/PIM-

SM/MSDP. The deployment of inter-domain multicast can continue to grow while the flat routing topology that is the MBone is eliminated. Sites on the MBone will hopefully transition to native multicast by deploying whatever inter-domain solution is appropriate. When this occurs, these sites will no longer need their old MBone tunnels. Observational analysis suggests that this transition process is indeed occurring. Because of the differences in route aggregation between MBGP routes and MBone routes, it is difficult to quantify this assertion. However, the number of routes in the MBone has decreased dramatically, and the number of MBGP routes has increased dramatically.

Internet2 Deployment

For Internet2, the plan has always been to try and do multicast "the right way", to the extent possible given the currently available set of protocols. As a result, Internet2 multicast deployment is following guidelines set forth by the Internet2 Multicast Working Group. Briefly, these guidelines require all multicast deployed in Internet2 to be native and sparse mode. No tunnels are allowed, and all routers must support inter-domain multicast routing using MBGP/MSDP. To date, Internet2 has experienced a reasonable amount of success in deploying multicast. This success includes backbone deployment, connecting other high-speed networks, connecting member institutions, and running several high-bandwidth (on the order of 30 Mbps) multicast applications.

There are two Internet2 backbones in the United States. One is the vBNS[34,35] and the other is Abilene. The vBNS has been in existence since 1995, and from a very early stage has had basic dense mode capability. During the 1998 Internet2 Member Meeting in San Francisco, the inherent problems of dense mode protocols were painfully realized when tens of megabits of traffic were flooded across the network. As a result, vBNS engineers worked hard to transition the network to PIM-SM and MBGP/MSDP. As of mid-1999, the network had successfully deployed inter-domain multicast, and was in the process of establishing MBGP and MSDP peering relationships with other networks. Figure 6 shows the topology of the vBNS, including the existing MBGP and MSDP peering relationships. As vBNS engineers gain experience in using MBGP and MSDP, and as other network operators also gain experience, the rate at which new MBGP/MSDP peerings are added will increase. A number of additional networks, including several international high-speed networks, are planning to connect to the vBNS in the very near future.

Figure 6: vBNS MBGP and MSDP peering topology.

The other Internet2 backbone is the Abilene network. Because Abilene is a newer network and has only recently (February 1999) become operational, the state of inter-domain multicast in Abilene is not nearly as advanced as it is in the vBNS. However, Abilene has been running PIM-SM since mid-1999, and has begun to establish its first set of inter-domain peering relationships. The challenge has been to climb the learning curve and establish multicast capability in the backbone. Now that the first MBGP/PIM-SM/MSDP peering relationships have been established, additional peerings are being added rapidly. The current topology is shown in Figure 7.

Figure 7: Abilene multicast map.

Conclusions

In this paper, we have presented a tutorial-style overview of multicast. We have covered the early development of intra-domain routing protocols, the evolution of the MBone, the needs and current solutions for inter-domain multicast, the set of next-generation protocols currently under investigation, and the current state of deployment in the Internet and Internet2. Whatever the future holds for multicast, it is likely to present major challenges for both research and deployment.

Acknowledgements

This paper would not have been possible without the expertise of many, people. It would be quite impossible for any list to do justice to them all. However, specific technical and qualitative suggestions were offered by Dave Meyer, Dave Thaler, Mostafa Ammar, Christophe Diot, Brian Levine, and Ben Chinowsky. Also, engineers from both the vBNS and Abilene contributed topologies for their respective networks. These individuals included Kevin Thompson, Steven Wallace, and Brent Sweeny.

References

[1] S. Casner and S. Deering, "First IETF Internet audiocast", *ACM Computer Communication Review*, pp. 92--97, July 1992.

[2] K. Almeroth, "A long-term analysis of growth and usage patterns in the Multicast Backbone (MBone)", in *IEEE Infocom*, (Tel Aviv, ISRAEL), March 2000.

[3] C. Diot, B. Lyles, B. Levine, and H. Kassem, "Requirements for the definition of new IP-multicast services", *IEEE Network*, January/February 2000.

[4] S. Deering, "Multicast routing in a datagram internetwork", Ph.D. Dissertation, 1991.

[5] S. Deering and D. Cheriton, "Multicast routing in datagram internetworks and extended LANs", *ACM Transactions on Computer Systems*, pp. 85--111, May 1990.

[6] S. Deering, "Host extensions for IP multicasting", Internet Engineering Task Force (IETF), RFC 1112, August 1989.

[7]S. Casner, "Frequently Asked Questions (FAQ) on the Multicast Backbone (MBone)", USC/ISI, December 1994. Available from ftp://ftp.isi.edu/mbone/faq.txt.

[8] H. Eriksson, "The multicast backbone", *Communications of the ACM*, vol. 8, pp. 54--60, 1994.

[9] D. Waitzman, C. Partridge, and S. Deering, "Distance vector multicast routing protocol (DVMRP)", Internet Engineering Task Force (IETF), RFC 1075, November 1988.

[10] W. Fenner, "Internet group management protocol, version 2", Internet Engineering Task Force (IETF), RFC 2236, November 1997.

[11] B. Cain, S. Deering, and A. Thyagarajan, "Internet group management protocol, version 3", Internet Engineering Task Force (IETF), draft-ietf-idmr-igmp-v3-*.txt, February 1999.

[12] J. Moy, "Multicast extensions to OSPF", Internet Engineering Task Force (IETF), RFC 1584, March 1994.

[13] J. Moy, "OSPF version 2", Internet Engineering Task Force (IETF), RFC 2178, April 1998.

[14] S. Deering, D. Estrin, D. Farinacci, V. Jacobson, G. Liu, and L. Wei, "PIM architecture for wide-area multicast routing", *IEEE/ACM Transactions on Networking*, pp. 153--162, Apr 1996.

[15] S. Deering, D. Estrin, D. Farinacci, V. Jacobson, A. Helmy, D. Meyer, and L. Wei, "Protocol independent multicast version 2 dense mode specification", Internet Engineering Task Force (IETF), draft-ietf-pim-v2-dm-*.txt, November 1998.

[16] D. Estrin, D. Farinacci, A. Helmy, D. Thaler, S. Deering, M. Handley, V. Jacobson, C. Liu, P. Sharma, and L. Wei, "Protocol independent multicast sparse-mode (PIM-SM): Protocol specification", Internet Engineering Task Force (IETF), RFC 2362, June 1998.

[17] T. Ballardie, P. Francis, and J. Crowcroft, "Core based trees (CBT): An architecture for scalable multicast routing", in *ACM Sigcom*, (San Francisco, California, USA), pp. 85--95, September 1995.

[18] A. Ballardie, "Core based trees (CBT version 2) multicast routing", Internet Engineering Task Force (IETF), RFC 2189, September 1997.

[19] B. Williamson, Developing IP Multicast Networks, Volume I (Fundamentals), Indianapolis, Indiana, USA: Cisco Press, 1999.

[20] C. Huitema, Routing in the Internet, Englewood Cliffs, New Jersey: Prentice Hall, Inc., 1995.

[21] Y. Rekhter and T. Li, "A border gateway protocol 4 (BGP-4)", Internet Engineering Task Force (IETF), RFC 1771, March 1995.

[22] P. Traina, "Experience with the BGP-4 protocol", Internet Engineering Task Force (IETF), RFC 1773, March 1995.

[23] T. Bates, R. Chandra, D. Katz, and Y. Rekhter, "Multiprotocol extensions for BGP-4", Internet Engineering Task Force (IETF), RFC 2283, February 1998.

[24] D. Farinacci, Y. Rekhter, P. Lothberg, H. Kilmer, and J. Hall, "Multicast source discovery protocol (MSDP)", Internet Engineering Task Force (IETF), draft-farinacci-msdp-*.txt, June 1998.

[25] S. Kumar, P. Radoslavov, D. Thaler, C. Alaettinoglu, D. Estrin, and M. Handley, "The MASC/BGMP architecture for inter-domain multicast routing", in *ACM Sigcomm*, (Vancouver, CANADA), August 1998.

[26] M. Handley, D. Thaler, and D. Estrin, "The internet multicast address allocation architecture", Internet Engineering Task Force (IETF), draft-ietf-malloc-arch-*.txt, December 1997.

[27] M. Handley, "Multicast address allocation protocol (AAP)", Internet Engineering Task Force (IETF), draft-ietf-malloc-aap-*.txt, August 1998.

[28] B. Patel, M. Shah, and S. Hanna, "Multicast address dynamic client allocation protocol (MADCAP)", Internet Engineering Task Force (IETF), draft-ietf-malloc-madcap-*.txt, February 1999.

[29] D. Meyer and P. Lothberg, "Static allocations in 233/8", Internet Engineering Task Force (IETF), draft-ietf-mboned-static-allocation-00.txt, May 1999.

[30] T. Ballardie, J. Crowcroft, C. Diot, C. Lee, R. Perlman, and Z. Wang, "On extending the standard IP multicast architecture", Technical Report, University College London Research Report RN/99/21, October 1999.

[31] H. Holbrook and D. Cheriton, "IP multicast channels: EXPRESS support for large-scale single-source applications", in *ACM Sigcomm*, (Cambridge, Massachusetts, USA), August 1999.

[32] R. Perlman, C.-Y. Lee, J. Crowcroft, Z. Wang, C. Diot, J. Thoo, and M. Green, "Simple multicast: A design for simple, low-overhead multicast", Internet Engineering Task Force (IETF), draft-perlman-simple-multicast-*.txt, February 1999.

[33] H. LaMaster, S. Shultz, J. Meylor, and D. Meyer, "Multicast-friendly internet exchange (MIX)", Internet Engineering Task Force (IETF), draft-ietf-mboned-mix-*.txt, June 1999.

[34] J. Jamison and R. Wilder, "vBNS: The internet fast lane for research and education", *IEEE Communications*, January 1997.

[35] J. Jamison, R. Nicklas, G. Miller, K. Thompson, R. Wilder, L. Cunningham, and C. Song, "vBNS: Not your father's internet", IEEE Spectrum, July 1999.