Stardust.com, Inc.
1901 S. Bascom Ave, #333
Campbell, CA 95008 USA
Tel: 408-879-8080
Fax: 408-879-8081
www.ipmulticast.com
www.stardust.com

IP Multicast Initiative (IPMI)

The Evolution of Multicast: From the MBone to Inter-Domain Multicast to Internet2 Deployment

From the Stardust.com State-of-the-Art Series

Kevin C. Almeroth
Department of Computer Science
University of California, Santa Barbara

Abstract 2

Introduction 2

Evolution of Intra-Domain IP Multicast 4

Evolution of Inter-Domain IP Multicast 11

Inter-Domain Multicast Deployment 21

Conclusions 24

Acknowledgement 25

References 25

Abstract

Without a doubt, multicast communication, the one-to-many or many-to-many delivery of data has become a hot topic. From almost all perspectives, multicast has a very bright future. For all of the potential multicast has and for all of the advocacy multicast has received, there are still concerns. Two of these include the slow rate of deployment and the growing complexity of protocols. 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 was to develop and refine *intra-domain* multicast routing protocols. Starting in 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 which are looking to make the provision of multicast less complex by fundamentally changing the multicast model. We briefly survey these efforts. Finally, the Next Generation Internet (NGI) effort has targeted multicast as a critical service for deployment. We examine how multicast is being deployed in and between the networks that are part of the NGI.

Introduction

Without a doubt, 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 vendor 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 it is starting to be used by a number of companies offering large-scale Internet applications and services. From almost all perspectives, multicast has a very bright future.

For all of the potential that multicast has and for all of the advocacy multicast has received, there are still some 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 PhD dissertation in 1988 and tested on a wide scale during an “audiocast” at the 1992 Internet Engineering Task Force (IETF) meeting in San Diego [Audiocast]. 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 has very small deployment and use [Almeroth] compared to the WWW. Second, the standard IP multicast model and the way IP multicast is currently deployed requires a significant amount of state and complexity in routers. In fact,

multicast typically requires at least *per-group* state information and often even *per-source* information. This requirement is at odds with the traditional Internet IP service model and has become the bottleneck limiting the multicast scalability. Furthermore, as deployment continues to grow so does the complexity. The latest requirement being the need to deploy an inter-domain multicast solution. As the complexity grows some in the community have started to search for new ways of reducing the amount of complexity.

With these problems in mind, the image of multicast may seem somewhat tarnished. This is especially true if multicast is to be used as a money-making enterprise for commercial companies. But it is the challenge of defining elegant, protocols and service along with the financial prospects that makes providing scalable delivery of IP packets to thousands, millions, and even billions of devices (not necessarily people) an interesting and provocative research topic. In reality, the demand from industry for a production multicast service has created two goals. One goal is to make multicast work and the other is to develop elegant correct protocols. Doing multicast the right way is a noble endeavor and an appropriate long-term research topic, but the demand to make multicast work has created an environment where functional, even short-term, solutions are very attractive.

In this paper, we attempt to describe the past, present, and future of multicast. The evolution of multicast will help the reader understand some of the factors affecting the provision of multicast. The current state of multicast will help the reader understand how multicast is currently being deployed. And the future work on multicast protocols will help the reader understand how multicast will evolve. Starting with the beginning of the Multicast Backbone (MBone) we describe how the emphasis was placed on developing and refining intra-domain multicast routing protocols. Today there are a number of choices. Starting in middle to late 1990s, particular emphasis has been placed on developing inter-domain multicast routing protocols. We provide an operational overview of the set of inter-domain protocols. The latest developments in multicast are the creation of a number of new initiatives to study alternate models for providing multicast. We briefly describe these efforts. Finally, the Next Generation Internet (NGI), or Internet2 is currently deploying multicast. We briefly describe the peering relationships that connect the Internet2 backbones and various international high speed networks.

The remainder of this paper is organized as follows. First, we describe the early evolution of multicast, in particular the development of intra-domain multicast. Then we focus on inter-domain multicast including the best current practice and several of the efforts to define the next generation of protocols. Next we detail the deployment efforts in the Next Generation Internet,

particularly in the vBNS and Abilene backbone networks. The paper then concludes.

Evolution of Intra-Domain IP Multicast

From the first Internet-wide experiments in 1992 to the middle of 1997, research and deployment in multicast focused on a single flat topology. This topology is in contrast to the Internet topology, which is based on a hierarchical routing structure. The initial multicast protocol research and standardization efforts were targeted at developing routing protocols. Beginning in 1997, when the multicast community realized the need for inter-domain routing, the existing set of protocols were categorized as intra-domain protocols and work began on solving the issue of inter-domain multicast routing. In this section, we describe the standard IP Multicast model, and the evolution of intra-domain multicast protocols.

The standard IP Multicast model

Deering is responsible for describing the standard multicast model for IP networks [Deering Ph.D]. This model describes how end systems are to send and receive multicast packets. The model includes both an explicit set of requirements as well as some implicit requirements. An understanding of the model will help the reader understand part of the evolutionary path multicast has taken. The model is as follows [Deering Multicast]:

- **IP-style semantics**: A source can send UDP/IP packets at any time without a need to register or schedule transmission. Because packets are UDP they are delivered using best-effort.

- **Open groups:** Senders 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 they are sending to. A group can have any number of sources.

- **Dynamic groups:** Multicast group members can join or leave a multicast group without the need to register, synchronize or negotiate with any centralized group management entity.

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

Birth of the IP Multicast backbone

Interest in building a multicast-capable Internet began to achieve critical mass in the late 1980s culminated by Deering's work [Deering Ph.D]. This work lead to the creation of multicast communication in the Internet [RFC 1075, RFC 1112] and the creation of the Multicast Backbone (MBone) [Eriksson, MBone]. In 1992, the MBone carried its first worldwide event when 20 sites connected together to receive the March meeting of the Internet Engineering Task Force (IETF) [Audiocast]. This first audio conference allowed a few members from all over the world to hear what was being said at the San Diego IETF meeting. In addition to the conferencing software, 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 encapsulated multicast packets on an incoming interface and then forward packets over the appropriate set of outgoing interfaces. Connectivity between these machines was provided using point-to-point IP-encapsulated *tunnels.* Each tunnel connected two end-points via one logical link but typically crossed several Internet routers.

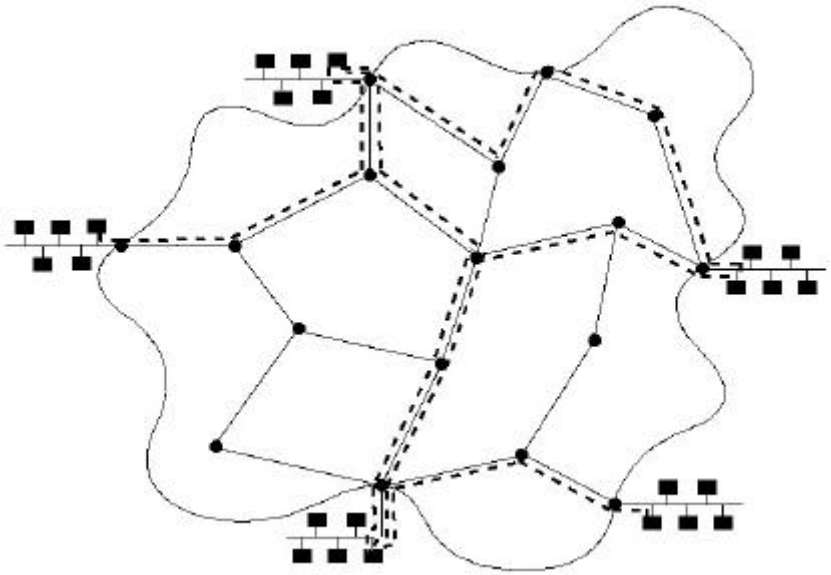

Figure 1: Example tunnel-based topology of the early MBone.

Once a multicast packet was received at a tunnel endpoint, it could be broadcast on a local network and received by other members. Routing between tunnel endpoints was provided using the Distance Vector Multicast Routing Protocol (DVMRP) [RFC 1075]. An example of connectivity

provided over this 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 instances of *mrouted* did not implement pruning. It was not until after several years of additional effort that pruning was successfully deployed. The technique used to create the multicast trees used to connect a source to a group of receivers became known as *broadcast-and-prune*. The resulting tree was called a *reverse shortest path tree* and it was rooted at the source. The steps to creating this tree are as follows:

1. The source broadcasts each packet on its local network. An attached router receivers the packet and send it on all outgoing interfaces.

2. Each router that receives a packet will perform 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 that the router would use as an outgoing interface to reach the source. In this way, a router will only receive packets on an 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 [1]

3. Eventually a packet will reach a router with some number of attached hosts. This *leaf router* will check to see if it knows of any group members on any of its attached subnets. A router discovers the existence of group members by periodically issuing Internet Group Management Protocol (IGMP) queries [IGMPv3, RFC 2236]. If there are group receivers the leaf router does nothing. If there are no group receivers the leaf router will send a *prune* towards the source on the RPF interface, i.e. on the interface the leaf router would use to forward packets to the source.

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

[1] In reality, the action for a packet that fails an RPF check depends on the protocol. What actually happens is that a prune is sent upstream on the interface to stop the delivery of these packets.

In this way, *reverse* trees are created. Furthermore, these trees can be constructed even on a virtual topology like the MBone. Broadcast-and-prune protocols are also known as *dense mode* protocols because they assume the topology is densely populated with group members and 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. 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 any 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.

Evolution of Intra-Domain IP Multicast

Since 1992, the MBone has grown tremendously. It no longer is a simple virtual network sitting on top of the Internet, but is rapidly becoming integrated into the Internet itself. In addition to simple DVMRP tunnels between workstations, the MBone now has native multicast capability provided in the routers themselves (see Figure 2). Other multicast routing protocols have been standardized and deployed.

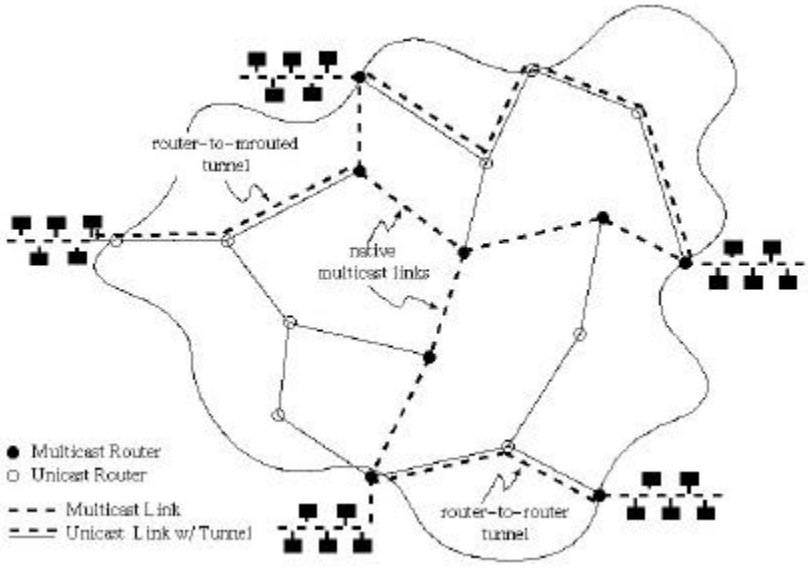

Figure 2: Example multicast topology with a combination of tunnels and native multicast links.

The key protocol advancement during this period was to address the limitations of dense mode protocols. A new class of protocols, called *sparse*

mode protocols was created. Instead of assuming many group members exist, sparse mode protocols assume that there are only a few group members and that they are distributed sparsely across the topology. Instead of broadcasting traffic and triggering prune messages, group members are expected to explicitly send join messages. Two sparse mode protocols in use today are described below.

MOSPF. Multicast Extensions to OSPF [MOSPF] uses the particular mechanisms of the Open Shortest Path First [OSPF] protocol to provide multicast. Basically, MOSPF routers flood information about group receivers throughout an OSPF area. 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 multicast group. While group membership reports are flooded throughout the OSPF area, data is not. This characteristic combined with the explicit join mechanism makes MOSPF a sparse mode protocol.

PIM-SM. Protocol Independent Multicast [PIM] has been split into two protocols, a dense mode version called [PIM-DM] and a sparse mode version called [PIM-SM]. A similar protocol called Core Based Trees [CBT] was proposed in the research literature and standardized by the IETF [RFC 2189] but has not been significantly deployed. There are actually two versions of PIM-SM. In most cases the differences between the two versions are subtle. One of the not-so-subtle differences is mentioned in the second step of the PIM-SM operation described below:

1. One or more rendezvous points (RPs) are configured in the network.[2]

2. Information about which routers in the network are RPs and the mapping of multicast groups is exchanged using either an Auto-RP or a bootstrap protocol. The Auto-RP mechanism is a vendor specific implementation since there was no bootstrap protocol described in PIMv1. However, the bootstrap protocol is part of the PIMv2 standard. The key point to this discussion is that there are protocol mechanisms that inform routers in the network what the mapping from a group address to an RP IP address will be.

3. 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 tree is formed per group, called a *shared tree* and it is rooted at the RP.

[2] The decision on how many RPs to have and their placement in the network is a network planning issue and beyond the scope of this paper.

Like other multicast protocols, the tree is a reverse path tree because join messages follow a reverse path from receivers to the RP.

4. Each source sends multicast data packets encapsulated in unicast packets to the RP. In response to these *register* packets, the RP, if it has routing state for the particular group, will then send a join towards the source. When the first-hop router at the source receives the join message it stops encapsulating the source's packets and instead sends them as multicast packets.

Sparse mode protocols have 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 as opposed to dense mode protocols, which 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 leading to identified receivers. Sparse mode protocols do have a few disadvantages, and these are mostly related to the use of an RP. For example, the RP can be a single point of failure though the bootstrap mechanism provides some redundancy. Also, the RP can become a hot spot for multicast traffic and having traffic forwarded from a source to the RP and then to receivers mean non-optimal paths may exist in the multicast tree. To solve this problem, PIM-SM provides a mechanism to switch from a shared tree to a shortest path tree. This switch-over occurs when the traffic rate through a leaf router exceeds a specified threshold.

Finally, not only has progress been made in protocol development, but MBone growth has lead to increased multicast awareness by users, which has led to demand for new applications and better support for real-time data. Improvements have been made in transport layer protocols. For example, the Real-Time Protocol [RTP] assists loss- and delay-sensitive applications in adapting to the Internet's best-effort service model. With respect to applications, the MBone has seen an increasingly diverse set of media types. Originally, the MBone was considered a research effort and its evolution was overseen by members of the MBone community. Coordination of events was handled almost exclusively through the use of a global session directory tool, originally called *sd*, but now called *sdr*. As multicast deployment has continued, and as it has been integrated into the Internet as a native service, the informal use agreements and guidelines have faded. Even though sdr-based sessions remain at the core of Internet multicast events, their percentage of the total is shrinking. Other applications are being deployed that do not coordinate sessions through *sdr* or use RTP. This potpourri of tools has enriched the types applications available, but has stressed the ability to provide multicast according to the standard IP multicast model.

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 also called **dense mode protocols** and always use a **shortest path tree** rooted at a source. This class of protocols assumes hosts are receivers more often than not and require routers to explicitly prune unwanted traffic. Explicit join protocols also called **sparse mode protocols** can use either a shortest path tree or a **shared tree**. A sparse mode protocol assumes receivers do not necessarily want multicast traffic and so require explicit joins. A shared tree uses a **core** or a **rendezvous point** to bring sources and receivers together.

Problems with IP Multicast

As the MBone has grown it has been suffering from an increasing number of problems and with an increasing frequency. The most important cause can generally be described as 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 both in terms of routing state and susceptibility to more frequently occurring misconfigurations. The multicast community has realized the need to deploy hierarchical routing. A more detailed description of the problems experienced by the MBone include:

- **Scalability:** With a flat network, network routes must be known by every router. At its peak, the MBone had almost 10,000 routes. Unfortunately, most of these routes had long prefixes (between /28 and /32) which means that each routing table entry represented only a few hosts. There was almost no route aggregation, and engineers knew that as the network grew routing tables would become unmanageable. This situation was beginning to manifest itself and routing overhead and instability had made the MBone an unstable service. The need to aggregate routes and have a hierarchical topology were lessons already learned in unicast routing and the problems motivating these solutions were occurring in the MBone.

- **Manageability:** As the MBone has grown it has become harder to manage. No formal group is responsible for managing the topology. The task has been left to participants on the MBone mailing list. As the MBone has grown so have the inefficiencies. There are two types of inefficiencies that are commonly observed:

  - **Virtual topology (tunnel) management:** The MBone is characterized as a set of multicast-capable islands connected together by tunnels. The goal has always been to connect these islands together in the most

efficient manner, but over time, sub-optimal tunnels have been created. Tunnels are often set up in very inefficient ways (see Figure 1 for several examples). This behavior was observed very early in the MBone especially with regard to what was then the MCI Backbone. To avoid the growing tangle of tunnels, engineers at MCI undertook the difficult challenge 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 replace the observed cases when several (even up to 10) tunnels crossed a single physical link. The work by MCI engineers help set a precedent for other service providers 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. ASs 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 between ASs are provisioned using the Border Gateway Protocol (BGP), which provides routing abstraction and policy control [RFC 1771, Routing]. As a result of wide-scale use of BGP there is a commonly accepted procedure when two ASs wish to communicate. Because multicast does not provide such an inter-domain protocol there is no protection across domain boundaries. ASs must connect using tunnels and routing problems are propagated throughout the topology.

To summarize, the first problem is the complexity and instability of a large flat topology. The second problem is that there are no protocol mechanisms to build a hierarchical multicast routing topology. Solutions to these two problems were targeted as goals during 1997.

Evolution of Inter-Domain IP Multicast

Inter-domain multicast has evolved out of the need to provide Internet-wide multicast. The inter-domain solution that has been deployed today is near-term and 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. Proposals are based on both the standard IP multicast model as well as more radical proposals which attempt to re-define the service model and develop new protocols.

Near-term solution

The near-term solution for inter-domain multicast routing has two parts. The first is a straightforward extension to the inter-domain unicast routing protocol called BGP. The second is an additional protocol whose functionality is necessitated by the situation when several sparse mode domains are connected together.

Carrying IP 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 is provided in unicast using the Border Gateway Protocol (BGP) [RFC 1771]. A network administrator for a domain can run any intra-domain routing protocol desired. To route packets to hosts external to the AS, the network need only know which external link is the best to reach the destination. The BGP mechanisms used to implement inter-domain routing are the reliable exchange of 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 that it can reach and an associated cost. This information is exchanged using TCP connections and each border router can then compute the set of ASs that should be traversed to reach any network. The use of a distance vector algorithm along with full path information addresses many of the limitations of traditional distance vector algorithms. Packets are still routed on a hop-by-hop basis but better routing decisions can be made with more knowledge of the AS topology.

The functionality provided by BGP and the well-understood paradigm for connecting ASs is the first step for providing inter-domain multicast routing. 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. Because some routing protocols rely on the underlying unicast routing tables, the creation of specific and separate state for multicast allows the two topologies to diverge. This is an important requirement as some network administrators wish to keep unicast and multicast traffic separate.

The mechanism by which BGP has been extended to carry multicast routes is through multiprotocol extensions to BGP4, also known as BGP4+[RFC 2283]. When using this protocol to carry multicast routes it is often referred to as “MBGP” which is assumed to stand for “Multicast Border Gateway Protocol”. However, while MBGP is not an official acronym it is acceptable terminology.

MBGP is able to carry multicast routes by adding the Subsequent Address Family Identifier (SAFI) to two BGP4 messages: MP_REACH_NLRI and MP_UNREACH_NLRI. SAFI can specify either unicast, multicast or unicast/multicast forwarding information. With MBGP, instead of every router needing to know the entire flat multicast topology, each router need only know its own internal topology and the path 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. This is an example of how MBGP can be used for unicast and multicast topologies that are not congruent.

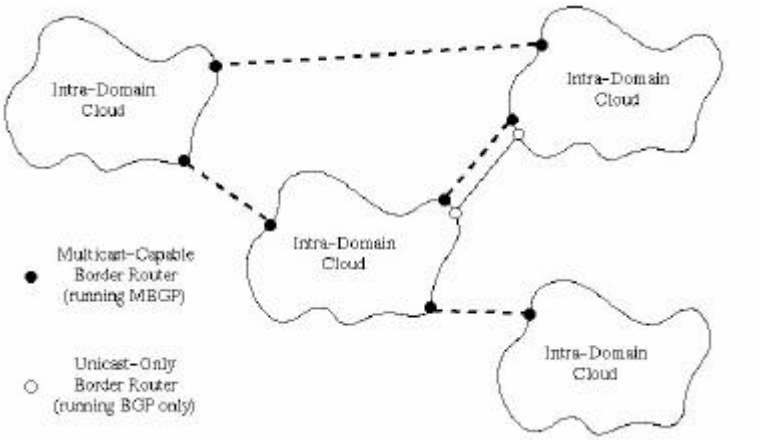

Figure 3: Example inter-domain multicast topology running BGP and/or MBGP.

While MBGP is the first step to providing inter-domain multicast, it alone is not a complete solution. While unicast BGP is capable of deciding the next hop for unicast packets, it is not capable of handling the multicast tree construction functions. As a result, there needs to be an inter-domain multicast routing protocol. 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; treat domains as nodes in a network; and establish a multicast tree among those domains containing group members.

The Multicast Source Discovery Protocol

There is still one more multicast function that is needed before the provision of inter-domain multicast can be considered complete. This function needs to be provided as a result of a problem that arises when connecting domains that are running sparse mode protocols. The problem is basically how to alert group

members in one domain that there are sources in other domains. Recall that the standard IP multicast model says sources do not need to be receivers and that sources do not need to know anything about a group's receivers. In a flat topology the problem of finding sources is accomplished in different ways depending on the type of routing protocol. The three basic mechanisms are:

- **Broadcast-and-prune:** In this case, senders broadcast packets to every single node in the network. Group members simply do nothing and multicast traffic from any and all sources will reach them.

- **Explicit-join:** In this case, all receivers must be pro-active and must send a join message to the group's RP. An RP in a specific domain knows about sources within the domain, but does not know about sources outside of the domain.

- **Domain-wide reports:** Protocols like MOSPF use an explicit join mechanism but instead of an RP, group membership information is carried throughout the topology. Every router will know what the multicast groups are, who the sources are, and who the receivers are. But again, MOSPF is an intra-domain protocol and information about sources is not transmitted across AS boundaries.

A particular problem arises when two sparse mode domains are connected together. While all the sources for a particular group *within a particular domain* will be known by the group's receivers, any sources outside of the domain will remain unknown. Why is this the case? The inter-domain solution using PIM-SM is designed to have its own RP. This characteristic ensures that “3rd party dependencies” do not occur. 3rd party dependencies can occur when there is only one RP per group. A problem occurs when all of a group's sources and receivers are located in one domain but the RP for the group is located in another domain. Two dependencies are thus created:

1. The service provider (AS) with the group sources and members must rely on a ``3rd party'' for service. This is potentially problematic as one service provider would be dependent on service from its competitors.

2. The service provider (AS) where the RP resides is carrying traffic for a group that it does not have sources or receivers. In most cases, without group members there is not likely to be little financial incentive to carry a competitor's traffic.

The adopted solution is to have an RP in each domain. An RP per domain leads to the problem of connecting the sources in one domain to receivers in other domains. This problem is summarized in Figure 4. The router closest to the group member sends joins towards the single RP that maps to the group address the receiver wishes to join[3]. Sources in domain A for this particular group register with the RP and traffic is forwarded using either the group-specific tree or the subsequently configured shortest path tree. The same activity for the same group takes place in domain B. The two domains cannot use the same RP since a 3rd party dependency might develop. In the end, there is no protocol mechanism which allows the necessary information to be exchanged (MBGP does not and cannot be used to provide this functionality).

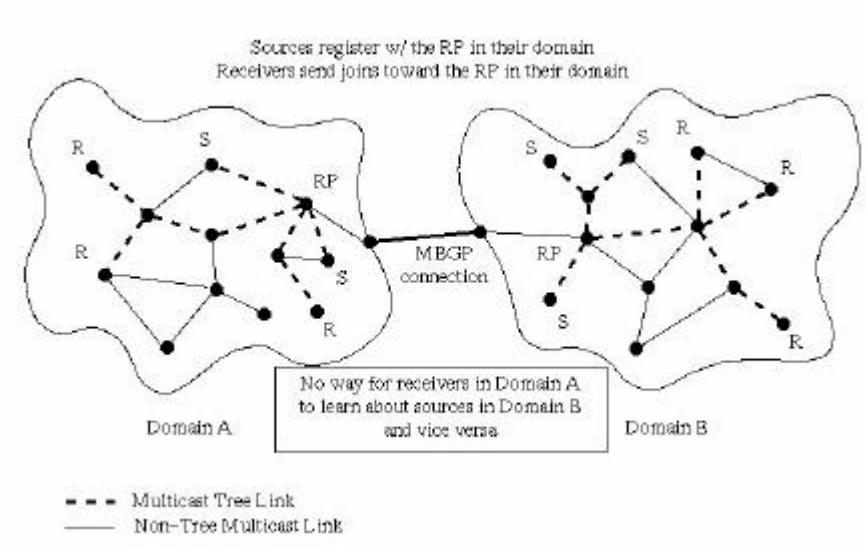

Figure 4: The problem of connecting groups spread across multiple sparse mode domains.

The solution to this problem is a new protocol, appropriately named the Multicast Source Discovery Protocol [MSDP]. The 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 MBGP in that MSDP sessions are configured between domains and TCP is used for reliable session message exchange. The basic MSDP mechanism is described as follows with each numbered step shown in Figure 5.

[3] We chose the term "closest router" in lieu of the more ambiguous terms of either "last-hop router" from the point of view of the source sending traffic or "first-hop router" from the point of view of the receiver sending join messages.

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 will send a Source Active (SA) message to all directly-connected MSDP peers.

3. MSDP message flooding:

   - MSDP peers who receive an SA message will perform a *peer-RPF check*. Depending on whether the two MSDP peers are in the same domain or not, a slightly different check is performed. But basically, the MSDP peer who received the SA message will check to see if the MSDP peer who sent the message is along the ``correct'' path. These peer-RPF checks are necessary to prevent looping of SA messages.

   - If the 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, which is 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 towards the source address advertised in the SA message. Once the reverse forwarding path has been established and the RP is forwarding data, the group members may then switch to a shortest path tree using PIM-SM conventions.

5. Repeat Steps 3 and 4 until all MSDP peers have received the SA message and all group members are receiving data from the source.

Thus ends the description of the design of short-term inter-domain multicast routing. However, while the description is relatively complete there are a number of details which were not discussed. And as with any system, the majority of the complexity is in the details. Furthermore, we have not yet discussed in any detail the limitations of the current solution. In particular, a qualitative assessment of the scalability, complexity, and overall quality of the protocols would be valuable.

The PIM-SM/MBGP/MSDP solution is relatively straightforward once a person understands all the acronyms and understands the motivating factors that drove the design of the protocols. So while some argue that the current set of protocols is not simple, it really is no more complex than solutions to

many of the Internet-wide problems. The key advantage of PIM-SM/MBGP/MSDP is it is a functional solution that is already being deployed with a fair amount of success. The key disadvantage is that as a long-term solution the PIM-SM/MBGP/MSDP protocol suite has poor scalability and does not perform well with reasonably dynamic groups and bursty sources. Each of these two issues is described more fully below.

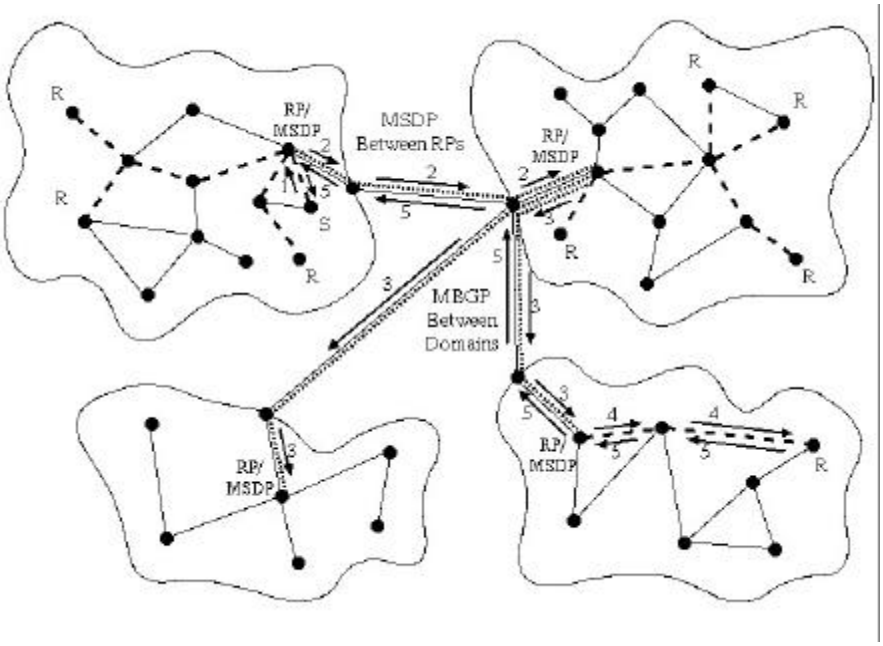

Figure 5: MSDP operation including flow of Source Active (SA) messages.

MSDP and dynamic groups. When multicast sources begin to transmit, the network is required to create some type of 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--either because of bursty sources or frequent group member join/leave events--the overhead of managing the group can be significant. A formidable task is created for a network that must establish and remove state information for thousands of receivers and transmitters scattered around the world.

One specific problem MSDP must deal with is *join latency*. Because 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 problems, MSDP peers may be configured to cache SA messages in the

hopes that when a new receiver joins the source will still be active. If an MSDP peer caches SA messages, other MSDP peers can take advantage of the cached information. A non-caching MSDP peer can send an "SA-Request" message to an MSDP peer who does perform caching. Minimizing join latency basically translates into a tradeoff between storage of state information and latency. The more information that is cached the shorter the join latency. This is not a particularly desirable tradeoff as significant state may be required to provide sufficiently small latency.

Another problem is caused by "bursty sources". For example, consider the case of Session Announcement Protocol (SAP) packets. One or only a few SAP packets are sent periodically by the *sdr* tool. Someone advertising a session will send a packet, the RP will hear the packet, flood an SA message, and group receivers then send joins toward the source. Because no multicast forwarding state existed before the packet was originally sent and it takes a non-zero amount of time to forward SA messages and have receivers establish forwarding state, the original packet will be dropped. If the SAP packets are sent after the original join state times out (3 minutes), then the whole process must be re-done. Because tree establishment takes a non-zero amount of time the first few packets of a transmission are always lost. If the transmission is short and bursty, all packets may be lost. The solution specified in the MSDP protocol is that SA messages may carry data to ensure the first few packets are delivered. This is not a particularly elegant solution but it is functional. Furthermore, our description of MSDP says that MSDP peering sessions use TCP and that SA messages may multicast data via this TCP connection. Recent discussions in the MSDP Working Group have resulted in the MSDP specification being modified to say that multicast data transmitted between MSDP peers should use UDP instead of TCP.

MSDP scalability. The issue of scalability is an important characteristic to consider for MSDP. Based on 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 groups. The number of SA messages being flooded around the network, all being sent by TCP (or even UDP) and all containing some amount of multicast data, could be 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 was seen as an immediate solution to an immediate need.

Long-term proposals

While PIM-SM/MBGP/MSDP is a recognized near-term solution, there is still a need to develop a long-term solution. Numerous efforts are being undertaken in this direction. These efforts can basically be broken down into two groups: (1) efforts based on the standard IP multicast philosophy, and (2) efforts which look to change this model in hopes of simplifying the problem. Efforts 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 to Internet-wide, inter-domain multicast [MASC BGMP][4]. The key idea in BGMP is to construct bi-directional shared trees between domains and have only a single RP. One of the functions of BGMP is then to decide in which particular domain to root the shared tree. The solution offered by BGMP is unlikely to suffer from the 3rd party dependency problem because the protocol is dependent on a more strict address allocation scheme. In general, address allocation has become an important issue for commercial users of multicast [Diot]. BGMP anticipated this needed, and as a result, the BGMP architecture included its own address allocation scheme called the Multicast Address-Set Claim (MASC) protocol [MASC BGMP]. However, as we describe, BGMP is not dependent on MASC but rather on the requirement that multicast addresses be associated with a particular domain.

The two important address allocation issues are the requirement just described and the need to avoid group collisions. A collision occurs when two multicast groups use the same multicast address and traffic of each group is delivered to group members of both groups. The effect of group collisions can range from a simple annoyance to an effective denial-of-service attack. The current scheme is rather informal. A person can simply pick an address and use it, or *sdr* can be used. *sdr* also just picks a random address but tries to eliminate collision by choosing an address not currently in use by a session known to *sdr*. While *sdr* does a tolerable job of mostly avoiding collisions, it does not provide sufficient protection against malicious behavior.

The MASC is one possible solution and is part of a more general addressing scheme called the Multicast Address Allocation Architecture [Malloc]. There are three levels of address allocation: (1) at the domain level, (2) within a domain, and (3) between hosts and the network. MASC acts as a top-level

[4] BGMP should not be confused with MBGP. After reading this paper the differences should be obvious but the similarity in name and acronym has lead to constant confusion. Furthermore, BGMP has recently been renamed. It was previously known as Grand Unified Multicast (GUM).

address allocation protocol and operates between domains; the Address Allocation Protocol [AAP] allocates addresses within a domain; and the Multicast Address Dynamic Client Allocation Protocol [MADCAP] is used by hosts to request addresses from a Multicast Address Allocation Server (MAAS).

While MASC and supporting protocols provide the functionality needed by BGMP, it is not the only solution possible. Another proposal is to statically allocate multicast addresses. Each AS number receives a fixed number of addresses and the AS number is encoded as part of the address [Static]. This proposal, called GLOP addressing, is gaining popularity, but only solves part of the problem. Its limitation is that in the current proposal, which is only to evaluate the idea, only a 24 bits are allocated for GLOP addresses, 16 of which are used for the AS number. As a result, only 8 bits or 256 addresses are available per AS. This problem could be solved by using more of the Class D address space for GLOP addresses, or switching to IPv6 addressing.

Root Addressed Multicast Architecture

In response to the perceived complexity of PIM-SM/MBGP/MSDP and BGMP/MAAA and the need to address additional multicast-related issues like security, billing, and management [Diot], some members of the multicast community are looking to make fundamental changes in the multicast model as an alternate solution. One class of proposals being offered is called the Root Addressed Multicast Architecture (RAMA). Within this class, there are two important proposals: [Express Multicast] and [Simple Multicast][5]. The key aspects of these two protocols are:

- **Express Multicast:** Express is designed specifically as a single-source protocol. Express also offers 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, or any single-source multimedia application. The key advantages of Express are that the routing complexity can be reduced and *closed groups* can be offered.

- **Simple Multicast:** The distinction of Simple Multicast over traditional multicast protocols is that it performs routing based on 8 bytes of addressing, a 4-byte *core* address and a 4-byte Class D group address. The advantages and disadvantages of such a proposal are being heavily debated

[5] There is some confusion over the name "Simple Multicast". Because the name is a misnomer and not particularly descriptive, some call the Simple Multicast protocol by the more generic term "RAMA".

but the authors who propose the scheme believe it eliminates the address allocation problem and eliminates the need for a "bootstrap protocol" for locating shared tree RPs. The basic premise of Simple Multicast is that there is going to be a primary source. It is this source which becomes the root/core of the multicast tree. By routing on the destination *and* source address there can be 2^{32} addresses per root/core/source. This solves the address allocation problem. This also solves the RP bootstrap problem because the core of the tree is carried explicitly in each packet.

The Express and Simple multicast proposals have received significant attention both in the research community and within the IETF. The fate of these protocols is as yet unknown. One certainty, though, is that they will certainly affect how multicast service is provisioned in the Internet. The result will either be that these protocols will be standardized and deployed or their key features will be incorporated into existing protocols.

Inter-Domain Multicast Deployment

The successful deployment of multicast, or lack thereof, was one of the original motivations for developing inter-domain routing protocols. In this section we describe efforts to successfully deploy these protocols. Our description is divided into two parts, a discussion of the Internet2 architecture and a discussion of the commodity Internet, so named to distinguish it from Internet2.

Deployment in the commodity Internet

Measuring the success of inter-domain deployment, either from a qualitative point-of-view or a quantitative count of connected hosts, is a difficult problem. Published studies have so far only dealt with the MBone, though several studies are currently underway which distinguish between the MBone and inter-domain multicast. As such, it is beyond the scope of this paper to offer any quantitative results. However, it is possible to describe the plan underway 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 keep MBone users part of the new infrastructure. The solution has been to make the MBone its own AS, called AS10888. Connectivity between AS10888 and other multicast-capable ASs is provided at the NASA Ames Multicast Friendly Internet Exchange [MIX]. The NASA Ames MIX provides connectivity between the MBone (AS10888)

and 10 other ASs who peer using PIM-SM/MBGP/MSDP. The deployment of inter-domain multicast can continue to grow while the flat routing topology that is the MBone can be eliminated. Sites on the MBone will hopefully transition to native multicast by deploying whatever inter-domain solution is appropriate. At this point these sites will no longer need their old MBone tunnels. One challenge in this transition period is how to connect sparse mode backbones with dense mode stub networks. The long-term solution is simply to deploy sparse mode protocols throughout the network.

Deployment in Internet2

For Internet2, the plan has always been to try and do multicast "the right way" in so much as is possible given the currently available set of protocols.

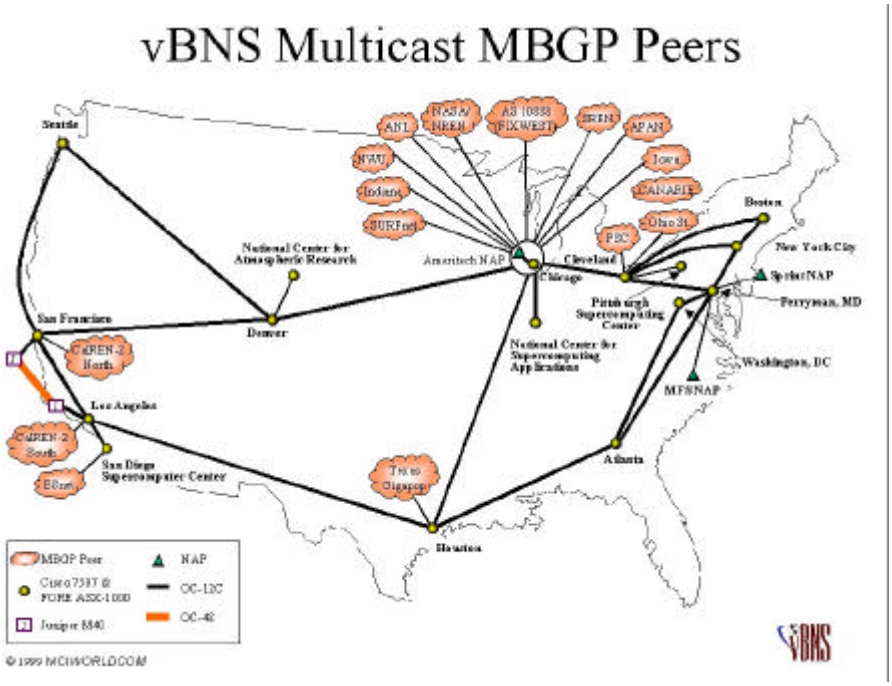

Figure 6: vBNS MBGP peering topology

As a result, the multicast deployment plan 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, and running high bandwidth multicast applications.

There are two Internet2 backbones in the United States, one is the very High Speed Backbone Network Service (vBNS) [vBNS 1997, vBNS 1999] and the other is Abilene. vBNS has been in existence since 1995, and from a very early stage, had basic dense mode capability. During the 1998 Internet2 Member Meeting in San Francisco, the inherent problems of dense mode protocols was painfully realized when tens of megabits of traffic was flooded across the network.

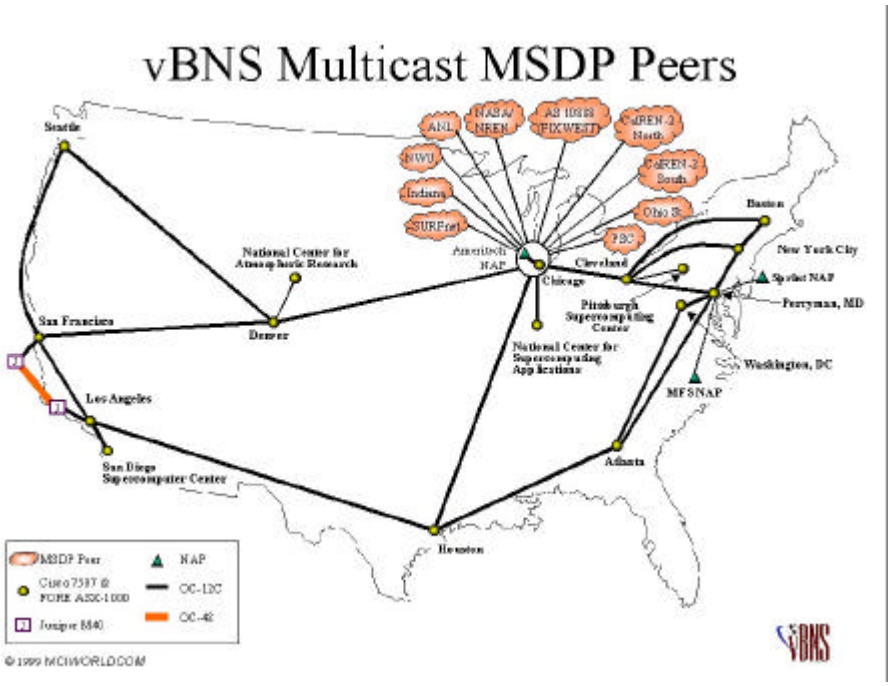

Figure 7: vBNS MSDP peering topology.

As a result, vBNS engineers worked hard to transition the network to PIM-SM and to running MBGP/MSDP. As of mid-1999, the network has successfully deployed inter-domain multicast and is now in the process of establishing MBGP and MSDP peering relationships with other networks. Figures 6 and 7 show the topology of vBNS including the existing 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 on connecting to the vBNS in the very near future.

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 is not nearly as advanced as in vBNS.

However, as of mid-1999, Abilene is now running PIM-SM in the backbone and has begun to establish its first set of inter-domain peering relationships. The challenge has been to overcome the learning curve and establish multicast capability in the backbone. Now that the first PIM-SM/MBGP/MSDP peering relationships have been established, additional peerings will be added rapidly. The current peering topology is shown in Figure 8.

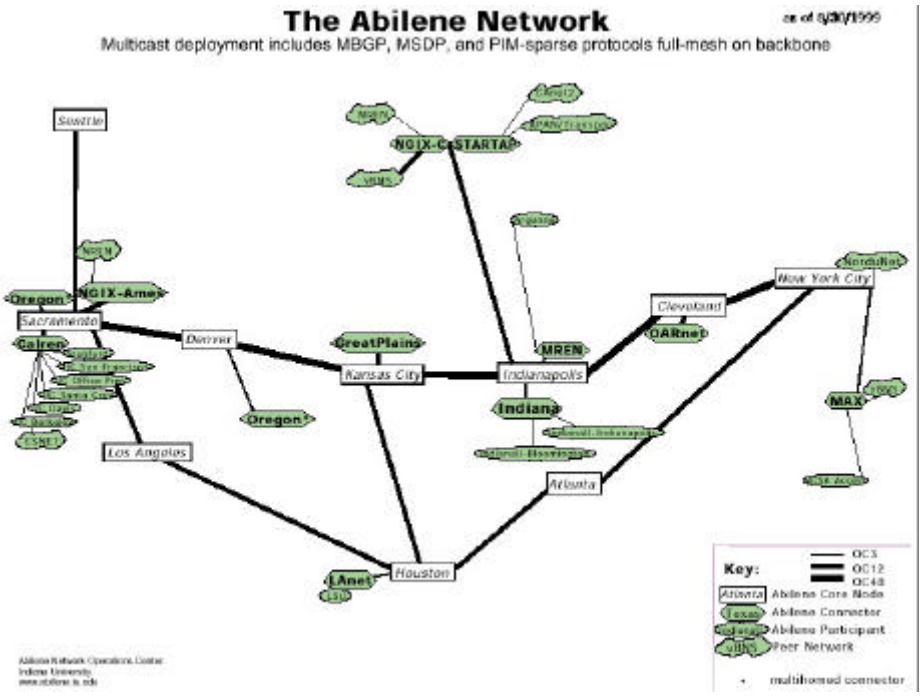

Figure 8: Abilene multicast map.

Conclusions

In this paper, we have presented a tutorial-style overview of the evolution of multicast, the current solution for inter-domain multicast, and a brief overview of the status of multicast deployment in the commodity Internet and Internet2. The evolution of multicast began with Deering's development of the multicast service model and the creation of the MBone. In the early years, significant progress was made on what became the intra-domain routing protocols. As the MBone grew, an exercise developed of finding the scalability bottleneck and fixing it. The most recent barrier was been the lack of inter-domain routing. Protocols have been developed to solve this problem and deployment is again progressing, both in the commodity Internet and Internet2. However, over the years the additional complexity required to fix new problems has increased. As a result, proposals for a new service model and the protocols to support it are being considered. Whatever the future for multicast, it has been

and will likely continue to hold interesting challenges for research as well as deployment.

Acknowledgement

This paper would not have been possible without the indirect expertise of many, many people. As a result it would be impossible to do justice through an actual list. However, specific technical and qualitative suggestions were offered by Dave Meyer, Dave Thaler, Christophe Diot, and Brian Levine. Also, engineers from both vBNS and Abilene drew their respective network topologies. These individuals included Kevin Thompson, Steven Wallace, and Brent Sweeny.

References

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

[Almeroth] K. Almeroth, “A long-term analysis of growth and usage patterns in the Multicast Backbone (MBone),” tech. rep., University of California, Santa Barbara, March 1999.

Available from http://www.cs.ucsb.edu/$\sim$almeroth/research.html\#mlisten.

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

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

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

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

[Diot] C. Diot, B. Lyles, B. Levine, and H. Kassem, “Requirements for the definition of new IP-multicast services,” tech. rep., Sprint ATL, June 1999.

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

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

[IGMPv3] 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.

[MADCAP] 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.

[Malloc] 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.

[MASC BGMP] 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.

[MBone] S. Casner, Frequently Asked Questions(FAQ) on the Multicast Backbone(MBone).

USC/ISI, December 1994.

Available from ftp://ftp.isi.edu/mbone/faq.txt.

[MIX] 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.

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

[MSDP] 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.

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

[PIM] 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, April 1996.

[PIM-DM] 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.

[PIM-SM] 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.

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

November 1988.

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

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

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

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

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

[Routing] C. Huitema, Routing in the Internet.

Englewood Cliffs, New Jersey: Prentice Hall, Inc., 1995.

[RTP] H. Schulzrinne, S. Casner, R. Frederick, and J.V., “RTP: A transport protocol for real-time applications.” Internet Engineering Task Force (IETF), RFC 1889, January 1996.

[Simple Multicast] T. Ballardie, R. Perlman, C. Lee, and J. Crowcroft, “Simple scalable internet multicast,” tech. rep., University College London, April 1999.

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

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

[vBNS 1999] 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.