

Multicast Routing and Addressing

Kaarle Ritvanen
Helsinki University of Technology
Department of Computer Science and Engineering
Kaarle.Ritvanen@hut.fi

Abstract

Multicast routing is necessary when multicast datagrams are to be sent to groups spanning several networks. The most popular protocol is PIM-SM because it scales well also in large networks. PIM-SM is based on shared distribution trees but also allows switching to source-rooted trees to improve performance. Multicast address allocation and assignment require some mechanism too. It can be done manually but there are also some protocols, such as MADCAP and MASC, which can be used to automate those tasks. This paper describes some of the techniques used in multicast routing and address allocation.

KEYWORDS: multicast, routing, address allocation, PIM

1 Introduction

Usually datagrams sent over the Internet have a single destination address which it will be delivered to. However, in certain cases it would be convenient if the datagrams could be sent to multiple receivers.

For instance, consider videoconferencing software that allows conferences among more than two participants. The same video stream is sent to every other participant by the currently speaking participant in the conference. Of course, that host could send multiple datagrams having the same content but different destination addresses, but this approach leads to inefficient use of bandwidth. It would be more efficient to duplicate the datagrams as late as possible on their path to the receivers. This technique is called multicasting.

The Internet Protocol (IP) incorporates a multicasting mechanism. In IPv4, a set of addresses, namely block $224.0.0.0/4$, is reserved to be used to identify multicasting groups [1]. In a similar fashion, also IPv6 address block $FF00::/8$ is reserved for this purpose [2].

1.1 Host Perspective

Since the multicast groups are dynamic, hosts must somehow be able to inform the routers about their memberships in such groups. Therefore there is the Internet Group Management Protocol (IGMP) which is used for this purpose with IPv4 [3]. In addition to allowing hosts to update their group membership information, it provides the routers a way of verifying the validity of the memberships. [4]

For IPv6, there is also an IGMP equivalent, which is called the Multicast Listener Discovery (MLD) protocol. MLD de-

fines a few extension messages for the IPv6 version of the Internet Control Message Protocol rather than defining a completely new protocol to be run over IPv6 although the mechanisms used by IGMP and MLD are essentially the same. [5]

Nevertheless, IGMP and MLD only define how hosts and routers share membership information. It does not define how the routers arrange the appropriate datagrams to arrive to their network. There are separate protocols for achieving that and they are discussed in more detail in Sections 2 and 3.

1.2 Multicast vs. Broadcast

Many link-layer protocols, such as the one used with Ethernet, support broadcasting. Broadcasting means that the packet is sent to every node within the link or subnet. Ethernet address $FF:FF:FF:FF:FF:FF$ is reserved for this purpose. Broadcasting is typically used when the address of the target host is not known, for example, when submitting Dynamic Host Configuration Protocol (DHCP) or Address Resolution Protocol queries. Broadcasting can be viewed as a special case of multicasting since it involves sending datagrams to a certain set of nodes.¹ [4]

However, hardware broadcast differs significantly from IP multicasting. It is a simple mechanism that does not require using any routing protocols. Unlike unicast and broadcast addresses, IP multicast addresses determine sets of nodes, the members of which can belong to arbitrarily many subnets, which makes multicasting an essentially more difficult issue than broadcasting. [4]

2 Multicast Routing Protocols

Many multicast routing protocols have been developed but there is no single all-purpose approach to this subject [4]. Common multicast routing protocols include Protocol Independent Multicast and Distance Vector Multicast Routing Protocol [6].

¹Broadcasting is considered a special case of multicasting in the Ethernet addressing scheme. A half of the Ethernet addresses are reserved for multicasting. Actually, IP multicasting utilizes a small fraction of them when it takes place in Ethernet or other multicast-capable network. If the least significant bit of the first octet of an Ethernet address is set, it is a multicast address. In this sense, the Ethernet broadcast address is only a certain, fixed multicast address.

2.1 Distance Vector Multicast Routing Protocol

Distance Vector Multicast Routing Protocol (DVMRP) was one of the first multicast routing protocols. However, it was still used in the global Internet to a large extent until the end of the 1990s. It was the protocol used in MBONE, a virtual multicasting backbone network.

DVMRP defines some extension message types to IGMP. Those messages are exchanged between DVMRP-capable routers in order to find out where the members of a certain group are located. DVMRP also defines a way to tunnel multicast traffic through unicast-only networks.

Originally, DVMRP was derived from another distance vector routing protocol, namely the Routing Information Protocol (RIP), which is used in conventional routing. The DVMRP constructs source-rooted *forwarding trees* for every pair formed of source address and group address. Each router determines its place in the tree so that its distance to the source is as small as possible. On receiving a multicast datagram, the router forwards it to every (virtual) interface except the one leading to the source along the tree.

Version 3 of DVMRP employs an algorithm called Reverse Path Multicasting (RPM). A router indeed knows whether there are local group members beyond a certain interface. In contrast, it does not know whether there are any distant members, considering a multihomed network. RPM suggests that if a router does not know whether there are any members beyond a certain interface, the datagram should be forwarded there. If a router on a leaf network receives a datagram with a destination group that has no members in that network, it sends a *prune* request to the router from which the datagram originated. Then that router learns that forwarding packets with that address to that leaf network is not necessary. Similarly, networks adjacent to the leaf networks can send prune requests if they learn that neither they nor their leaf networks have any group members. Hence the information about group memberships (or actually the absence of them) flows up in the tree toward the source. A mechanism for canceling prunes is also provided. [7]

As the group memberships vary dynamically and as the routers do not have infinite amount of memory, it is necessary that the pruning information eventually expires. Therefore using DVMRP causes periodic flooding to networks without memberships [8]. DVMRP also suffers from the same scalability problems as other distance vector protocols. When the number of routers grows, propagation of the routing and membership information becomes slower. In addition, DVMRP has to keep track of a large amount of information which means that routers must have a lot of memory. Therefore it is not a suitable multicast routing protocol to be used in the global Internet. [4]

Moreover, the tunneling capability of DVMRP has caused problems. It is too easy to create virtual topologies that are inconsistent with the physical topology, which leads to inefficient routing.

2.2 Multicast Extensions to OSPF

Multicast Extensions to OSPF (MOSPF) is an attempt to create multicast routing protocol that scales better than

DVMRP. As its name suggests, MOSPF must be used in conjunction with the Open Shortest Path First (OSPF) routing protocol. MOSPF takes advantage of the link-state information maintained by OSPF. MOSPF defines an additional message type that is used to advertise the locations of multicast group members, similarly as link-states are advertised in OSPF. Using the link-state and group membership information, MOSPF routers are able to calculate pruned source-rooted shortest-path trees for multicast datagrams by using the Dijkstra's algorithm. MOSPF also defines a mechanism for inter-AS multicast forwarding. [9]

The biggest disadvantage of MOSPF is that every router must maintain membership information of every group. Therefore MOSPF also scales poorly if there are many multicast groups. When compared to DVMRP, MOSPF causes no useless data traffic. In contrast, it causes more routing information traffic because the membership databases have to be synchronized whenever someone joins or leaves a group. [4]

2.3 Protocol Independent Multicast

Protocol Independent Multicast (PIM) consists of two separate protocols, namely PIM Dense Mode (PIM-DM) and PIM Sparse Mode (PIM-SM) [10]. PIM-DM is intended to be used when the multicast group members are densely distributed across the network, whereas PIM-SM should be used when the distribution is wide and spans several networks. The operation of PIM-SM is described in detail in Section 3.

PIM-DM basically uses the same broadcast-and-prune strategy as DVMRP. The main difference between these two is that PIM-DM does not incorporate any topology discovery mechanism as DVMRP does, hence it is said to be protocol independent. Whereas DVMRP uses a distance vector protocol to build source-rooted trees, PIM-DM relies on the routing information the router has gathered by using a unicast routing protocol, such as RIP or OSPF. [11]

Often a protocol called Multiprotocol Extensions for BGP (MBGP) is used to gather the routing topology information for PIM. MBGP is an extension to the Border Gateway Protocol (BGP), and among other things it allows defining different topologies and policies for multicasting and unicasting. [12]

2.4 Deployment of Routing Protocols

It seems that DVMRP and PIM are the most used multicast routing protocols. The leading routing hardware vendor Cisco Systems supports both of them. For instance, Cisco Catalyst 3750 series devices support DVMRP and PIM [13]. The E-series routers of another large vendor, Juniper Networks, also support them [14].

DVMRP used to be the most popular protocol since it was used in MBONE but PIM has taken its place. In fact, nowadays MBONE is almost dead. DVMRP tunnels were effectively replaced by inter-domain MBGP by the end of 2000 [15]. Most WWW pages concerning MBONE have been last updated in the late 1990s.

Most router vendors support PIM, and some Internet Service Providers have begun using PIM in their backbones

and providing multicast services based on it. The problem with MOSPF is that it can be used only in OSPF networks. Typically, OSPF is not used on small or medium-sized networks. [16]

Perhaps even bigger problem in using MOSPF is that it runs Dijkstra's algorithm for each distribution tree and keeps track of a huge amount of state information. Therefore it is not suitable if there are many multicast groups to be handled. Cisco Systems has stated that they do not support MOSPF for this reason and encourage using PIM [17].

Neither PIM nor MOSPF specify how to gain information about multicast sources in other domains. This is one of the most difficult issues in multicast routing and one possible solution is discussed in Section 3.4. Anyway, since MOSPF scales poorly and is not superior to PIM in inter-domain routing, there is no good reason to use it.

3 Operation of PIM-SM

PIM-SM is a multicast routing protocol for situations where the group member distribution is sparse. Like PIM-DM, it is independent of unicast routing protocols. The major difference between PIM-SM and PIM-DM is that PIM-SM is a *demand-driven* protocol like MOSPF. Instead of broadcasting to all interfaces until a prune request is received, group memberships must explicitly be announced before datagrams are forwarded. Therefore, PIM-SM scales better to large networks than MOSPF because PIM-SM routers do not have to be aware of every group membership.

This section illustrates how PIM-SM works. The description of the protocol operation in Sections 3.1 and 3.2 is loosely based on [18].

3.1 Rendezvous Points and Shared Trees

PIM-SM uses the same shared distribution tree approach as the Core Based Trees (CBT) protocol [19]. Each PIM-SM domain has at least one router called Rendezvous Point (RP). If there are more RPs than just one, they are responsible for different multicast groups. So within a domain, exactly one RP is associated with each multicast address.

RPs correspond the core routers of CBT since they act as roots for shared distribution trees. Shared tree means that the same distribution tree is used for all source addresses. This implies that less state information needs to be maintained than with dense mode protocols which use separate trees for each source node.

3.1.1 Establishing Shared Trees

Each subnet of a PIM-SM domain has one Designated Router (DR) that connects the local hosts to multicast distribution trees. DRs obtain information about local memberships by IGMP or MLD as usual. DRs with local members in a multicast group periodically send PIM-SM Join/Prune messages toward² the RP of the group in question. Also the intermediate routers between those DRs and the RP periodically send messages of this type toward the RP.

²In this context, sending toward node N means that the packet is received and processed by the next router along the shortest route to N.

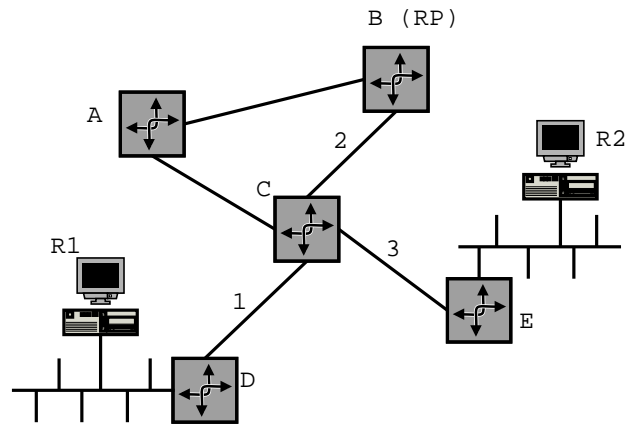


Figure 1: Joining to a group

Of course, Join/Prune messages must be sent immediately when some host joins or leaves a group so that it causes modifications to the distribution tree. Figure 1 shows an example of hosts R1 and R2 joining to a group. The following actions take place:

1. Host R1 informs router D that it wants to join group G. Router D adds group G to its routing table, outgoing interface set to the local area network interface and incoming interface set to the one leading to router C. A Join/Prune message for group G is sent to router C.
2. On receiving the message from router D, router C updates its routing table in similar fashion and sends a Join/Prune message to router B which is the RP for group G. As router B itself is the RP, it only updates its routing table without sending a further Join/Prune message.
3. Host R2 joins to group G. This makes router E send a Join/Prune to router C which in turn appends the interface leading to router E to the outgoing interface list for group G. However, no Join/Prune is sent to router B because router C already belongs to the shared forwarding tree of group G.

3.1.2 Sending Datagrams

Now suppose that host S connected to router A to send a datagram to group G, as depicted in Figure 2. This triggers the following actions:

1. Router A receives the datagram from host S, encapsulates it to a PIM-SM Register message and unicasts it to router B.
2. Router B decapsulates the datagram. It forwards the datagram to the interface leading to router C because it is the only interface in the outgoing interface list for group G.
3. Router C receives the datagram and looks up the outgoing interface list which contains the interfaces leading to routers D and E. The datagram is forwarded to them and they deliver it to hosts R1 and R2.

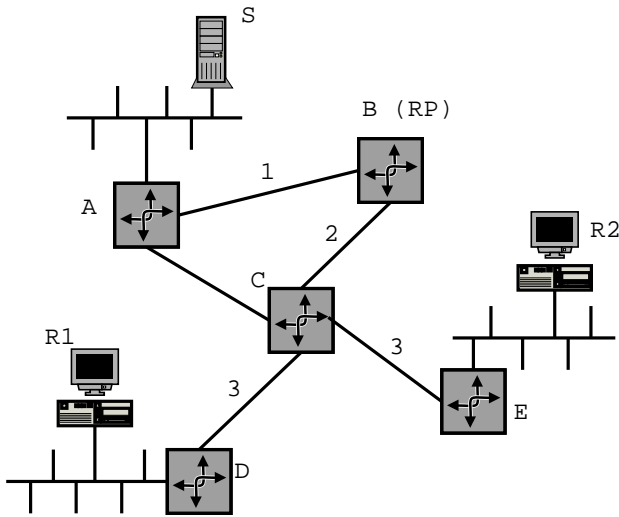


Figure 2: Sending to a group

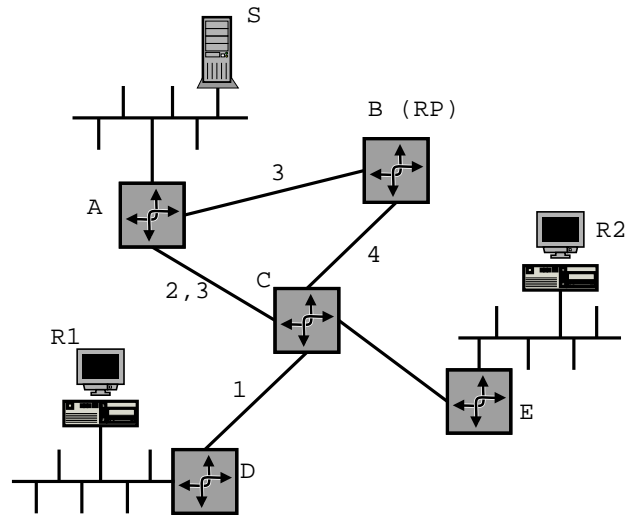


Figure 3: Switching to the shortest-path tree

After having received the Register message from the source, the RP (router B in this case) can send a source-specific Join/Prune message for group G and source S toward S so that the intermediate routers set up their routing tables with respect to group G and source S. The advantage is that multicast datagrams do not have to be encapsulated and unicast to the RP.

3.2 Shortest Path Trees

As it can be seen in Figure 2, the shared tree approach leads to a less efficient forwarding of datagrams than using shortest-path source-rooted trees. It would be more efficient if router A sent the datagrams directly to router C. This does not matter if the data rates are low. However, if some hosts are sending to a group very frequently, it may cause congestion near the RP. In addition, the transmission delay is increased because the datagrams are circulated via the RP.

To overcome this disadvantage, PIM-SM allows DRs to switch between shared trees and shortest path trees (SPT). Suppose that router D in Figure 3 notices that host S is sending to group G by a significant rate and decides to switch to SPT. This causes the following events to happen:

1. Router D sends a source-specific Join/Prune message to router C with the source address set to the address of S.
2. Router C sees from its unicast routing table that the shortest path to host S passes through router A. Therefore router C creates a new entry to its multicast routing table for group G and source address S and sets the incoming interface to the one leading to router A. However, this entry is not yet used. Router C sends router A a similar source-specific Join/Prune message it received from router D. Router A updates its routing table similarly. The outgoing interface of the new entry for source S is set so that the datagrams are sent also to router C.
3. Host S sends a datagram to group G. Router A now forwards it to both routers B and C.

4. As router C received a datagram originating from host S via router A, it knows that the shortest-path request has propagated all the way to the DR of S. (In this case, there was no intermediate routers between them, but this is not always the case.) Now router C can enable the new routing table entry and prune itself from the shared tree with respect to source address S and therefore it sends a Join/Prune message to router B with S inserted to the prune list. The original datagram is forwarded to routers D and E as usual.

After these steps multicast datagrams sent to group G by host S are delivered to router D using the shortest route. In fact, the route to router E was also shortened because the shortest route between host S and router D happened to cross with the shared-tree route to router E.

In this case, router B is the RP, so it does not send further prune requests although router C is its only child in the shared tree for group G. The RP cannot prune itself from the forwarding tree with respect to any source since it must know all the sources for the groups it manages. The reason for this becomes apparent in Section 3.4.

3.3 RP Selection

If there are several RPs within the domain, DRs must know which RP is associated with each multicast group address. One router in a domain acts as the Bootstrap Router (BSR) of the domain. That router is determined by an election procedure. All routers configured to act as RPs advertise themselves to the BSR. The advertisement messages include the multicast address blocks for which the router is willing to act as the RP. The BSR collects this information to a PIM-SM Bootstrap message which is then propagated hop-by-hop within the domain.

But what if several routers offer to act as the RP for some address blocks? The load is then balanced between them all. DRs use a certain hash algorithm to determine the correct RP for each group address.

3.4 Connecting PIM-SM Domains

PIM-SM builds forwarding trees within a single domain. A PIM-SM domain means a contiguous set of PIM-capable routers that are configured to operate within certain boundaries [10]. This implies that hosts can only receive multicast datagrams from the domain they belong to.

However, there are protocols that enable receiving datagrams also from other domains. Multicast Source Discovery Protocol (MSDP) is one such protocol. MSDP routers are connected to each other by TCP connections, thus forming a virtual topology. When an MSDP-capable RP receives a Register message from the DR of some new sender, it generates a Source-Active (SA) message containing the address of the source host, the group address and its own IP address. Then it sends the SA message to its MSDP peers (neighbors in the virtual topology) which in turn flood it to their other peers, in somewhat similar fashion DVMRP handles datagrams with non-pruned group addresses. The outcome is that every RP is aware of the sender unless the group is private, in which case no SA messages are sent. [20]

On the other hand, individual RPs know each active inter-domain multicast source (for non-private groups). Upon receiving an SA message for such group that has members in its domain, an RP sends a Join/Prune message for that group and that source address toward the source host like it does for intra-domain sources. After this the RP receives the datagrams sent by the source in the other domain and forwards them according to the normal PIM-SM convention. Of course, this means that every intermediate router must also support PIM-SM. [20]

SA messages are also cached for a while, just in case that a group gets members in that domain although it has not yet any. Active sources are also advertised periodically to ensure that datagrams are received from all inter-domain sources. [20]

Nevertheless, the SA flooding approach of MSDP is not itself very scalable. It works when the total number of RPs is low or there are not so many groups nor source hosts. Therefore someone might argue that why should we have different PIM-SM domains at all. Would it not be better to have only one domain within which the forwarding trees were built?

There are a few reasons for having separate domains. Firstly, domain boundaries impose a maximum for the distance between source hosts and RP. Since the RP of a certain group is determined by a hash function, it can happen that it is located on the opposite side of the domain. When the diameter of the domain is kept reasonably small, this is not a problem.

Another reason is security. The bootstrap protocol of PIM-SM described in Section 3.3 is prone to certain types of attacks. And messing up the system does not have to be intentional. Also misconfiguration of routers may make the entire domain unworkable. Moreover, if an ISP charges its customers per traffic basis, the customer would prefer having a separate PIM-SM domain in order to reduce the amount of traffic. If the RP of a group the customer uses resided outside the customer's network, the traffic would circulate via the ISP's network, thus imposing additional traffic costs. The location of the RP can be selected if the ISP refrains from advertising its own RPs for that group, however.

3.5 Source-Specific Multicast

Source-Specific Multicast (SSM) defines new semantics for multicasting. It suggests that instead of using only the multicast group addresses, the channels³ are identified by the pair formed of the source and group address [21]. This change imposes the restriction that only one host can send to the channel. But in some cases, this makes perfect sense. Consider a multicast-based video-on-demand server that starts streaming periodically, for example every 15 minutes. During that period it collects a group of viewers and then uses multicast to deliver the datagrams. Another example could be a multicast-based TV or radio channel. The clients need to receive datagrams only from the streaming server. In fact, the clients do not even want to receive datagrams from anyone else. SSM actually provides very strict access control on who is allowed to send to a group [21].

As only one host is allowed to send to an SSM group, the complexity of routing is lower than in conventional multicasting. Only one source-rooted tree is needed for datagram delivery. There is no need for shared trees nor using source discovery protocol. [21]

Indeed, SSM requires that hosts are able to pass the source address to the router when joining to groups. IGMPv3 provides a mechanism for this [3]. For IPv6, MLDv2 adds support for SSM [22]. Of course, using SSM has implications on the application level too. Considering an SSM-based video-conferencing application, the participants must use some additional protocol to find out which hosts are sending to the group so they can join the appropriate channels.

The combination of SSM and PIM-SM is called PIM-SSM. SSM is used only with certain group addresses. Therefore PIM-SM routers know when to use PIM-SSM. In practice that means that the DRs send Join/Prune messages directly to the source instead of the RP. The result is that only source-rooted trees are built. Therefore it is not necessary to use MSDP with these addresses, either. [23]

4 Multicast Address Allocation and Assignment

Multicast routing protocols deal with the issue of getting the datagrams to each member of the group having a certain address. But how are these group addresses allocated to sites and assigned⁴ to nodes and applications? It is clear that a mechanism is needed for allocating and assigning these addresses to prevent collisions.

According to a document produced by the Multicast Address Allocation (MALLOC) Working Group of the Internet Engineering Task Force (IETF), there are three possible ways to allocate and assign multicast addresses: static, scope-relative and dynamic [24]. Although no difference between allocation and assignment is made in that document,

³In SSM terminology, groups (together with the source address) are called *channels*. Joining to a group is called *subscribing* a channel.

⁴Although used interchangeably by many multicast-related documents, address allocation and assignment have a slightly different meaning. Allocation refers to granting an administrative domain the right to use and assign certain addresses, whereas assignment means dividing the addresses among the nodes or applications.

this paper presents how this categorization applies to both allocation and assignment.

4.1 Address Allocation

This section describes different ways to allocate multicast addresses to organizational domains.

Static Statically allocated address blocks may be obtained from the Internet Assigned Numbers Authority (IANA). Although some organizations have managed to acquire their own multicast address blocks, IANA more willingly assigns static multicast addresses for applications rather than allocating them for organizations.

However, the IPv4 multicast address space is pretty small and IANA is constantly receiving address allocation requests. This problem was discussed in the meeting of IETF MBONE Deployment Working Group held on March 1, 2004. Previously made bad allocation choices were considered a problem because they make it more difficult to refuse requests by others. [25]

Scope-relative Address block $239.0.0.0/8$ has been defined to be the administratively scoped IPv4 multicast space. Addresses belonging to this block may be used as group addresses locally or within a single organization. Multicast datagrams sent to these addresses will not be sent across administrative boundaries, so these addresses do not need to be globally unique. Similarly, scoped multicast addresses have also been defined for IPv6 [2]. [26]

Dynamic In theory, it is possible that some protocol is used between administrative domains in order to prevent them from using the same multicast addresses. One such protocol is described in Section 5.2. However, using dynamic inter-domain allocation is rare.

Derived This allocation method is not listed in [24], but it is still important. Instead of directly allocating a static address block, Autonomous Systems (AS) can use so called GLOP addressing when it needs multicast group addresses. For each AS there are 256 addresses derived from its AS number that it is allowed to use for any purposes. These addresses belong to block $233.0.0.0/8$. The second octet of the address is the high-order octet of the 16-bit AS number and the third octet of the address is the low-order octet of the number. [27]

GLOP addressing is possible with IPv4, but there is quite a similar way to use IPv6 multicast addresses without explicit reservation. This mechanism is perhaps even better than GLOP addressing scheme because rather than being based on AS numbers, the addresses are derived from network prefixes. IPv6 addresses belonging to block $FF30::/12$ are reserved for this purpose. IPv6 unicast addresses contain a prefix, the length of which is usually at most 64 bits [28].⁵ The prefix and

its length are embedded to the address and the 32 least significant bits are reserved for the group ID. Including the prefix length information allows not only ASs to assign static addresses but also every other organization having control over any unicast prefix. [29]

4.2 Address Assignment

In addition to allocating multicast addresses, they also have to be assigned to nodes and applications within a domain. This section describes how that can be done.

Static Static addresses are needed by certain public protocols and they are assigned by IANA. For example, hosts running a Network Time Protocol (NTP) server can periodically send datagrams to group $224.0.1.1$. Thus, NTP clients are able to receive NTP clock synchronization messages even without knowing the unicast address of the server. They just have to join this well-known multicast group.

Assigning multicast addresses manually to private applications or nodes can also be viewed as static assignment. These addresses may be administratively scoped addresses, GLOP addresses or even addresses allocated by IANA for the organization. Anyway, they have to be in some way further assigned to nodes or applications.

Scope-relative IANA assigns scope-relative multicast addresses for certain types of services. These addresses are valid relative to the administratively scoped address blocks. For instance, $239.255.255.249$ refers to site-local DHCP servers and $239.251.255.249$ refers to organization-local DHCP servers. [26]

Dynamic When a normal application program needs a multicast group address, it could be allocated dynamically from a Multicast Address Allocation Server (MAAS) like unicast addresses are allocated from DHCP servers. This is convenient when the address is needed for only a limited time. The allocation can then be renewed if it turns out that the original lease time was too short. There are a few protocols for dynamically allocating multicast addresses. One of them is discussed in Section 5.1.

It is important to note that dynamic assignment can be used regardless of the way the addresses were allocated. They may be statically allocated or administratively scoped addresses, or even dynamically allocated addresses.

4.3 SSM Addresses

IANA has reserved IPv4 address block $232.0.0.0/8$ for SSM [21]. Similarly, some IPv6 addresses are intended to be used solely with SSM [29].

As described in Section 3.5, SSM channels are identified by the combination of the group address and the source address. This means that the group addresses do not have to

allocate unicast addresses only from block $2000::/3$, so this is not a problem.

⁵Actually, this is true only for such unicast addresses that do not belong to block $::/3$. Addresses outside this block are required to contain a 64-bit interface ID, thus leaving only 64 bits for the prefix. IANA currently

be globally unique. They only have to be unique among the applications used on the sending host. Therefore they can be used freely without further allocation by IANA. For the same reason, no MAAS is needed to coordinate address assignment between nodes.

5 Address Allocation Protocols

The MALLOC Working Group has suggested a three-layer allocation⁶ architecture. According to its suggestion, allocation of addresses takes place hierarchically, at three different layers. The motivation behind this modular architecture is that each layer can be replaced or updated independently. The layers are: [24]

1. Multicast clients requesting for single addresses from MAASs.
2. Intra-domain mechanism to ensure that MAASs do not allocate duplicate addresses. The MAASs could be manually configured or there might be a protocol for doing this. The Multicast Address Allocation Protocol is one such protocol [30].
3. Inter-domain allocation mechanism. If it is desirable to have fully dynamic allocation and assignment, there must be a protocol for making sure that the group addresses are unique across domain boundaries.

5.1 Multicast Address Dynamic Client Allocation Protocol

Multicast Address Dynamic Client Allocation Protocol (MADCAP) is a protocol that is used by applications to allocate multicast addresses from MAASs. MADCAP was originally based on DHCP and that is why there are many similarities between these two protocols. [31]

MADCAP has a server discovery mechanism which itself uses a statically allocated scope-relative multicast address. Clients can lease multicast addresses from the servers, renew their leases and release them. When allocating addresses, the clients pass the server the desired address family, scope and possibly some other options. [31]

It is important to note that MADCAP does not solve the problem how applications on different hosts can agree on a common group address. One of the hosts has to reserve the address which then must be communicated to the other hosts by an application level protocol, such as the Service Location Protocol (SLP) [32].

5.2 Multicast Address-Set Claim Protocol

Multicast Address-Set Claim (MASC) is an inter-domain allocation protocol. MAASs of different domains may use it to make sure that they are not assigning addresses from overlapping address blocks. In contrast to the traditional request-response protocols, MASC employs a claim-collide mechanism similar to the collision detection technique used in Eth-

ernet. The authors of the protocol argue that this makes the protocol more resilient to network failures. [33]

MASC nodes form a hierarchical virtual topology where the nodes are connected to each other by TCP connections. Each non-leaf node advertises free address prefixes it has reserved to its children. When a child node wants to allocate some addresses, it chooses a suitable portion from one of its parent's blocks⁷ and sends the parent a claim for that block. The claim is then sent to all the siblings of the originator. If none of them claims an overlapping set of addresses within a certain period (48 hours by default), the claimer is entitled to use the block. However, each prefix reservation has a finite validity period, after which it should be renewed if necessary. [34]

As the waiting period after the claim usually is relatively long, MASC servers must allocate a sufficient amount of blocks in advance. Unfortunately, predicting the future is a difficult task and therefore good algorithms are required. Robust algorithms are also useful when deciding the prefix to be claimed. It is better to pick a prefix that can be expanded as much as possible (by decrementing its length) because aggregatable address blocks require less space to store than completely random prefixes. [34]

5.2.1 Border Gateway Multicast Protocol

MASC can be used in connection with the Border Gateway Multicast Protocol (BGMP) [35]. BGMP is an inter-domain multicast routing protocol that builds shared trees between the domains, like CBT and PIM-SM do within a domain. The root of the tree for a certain group is located in the domain that has allocated the group address by MASC. [36]

BGMP was thought to be more scalable than MSDP on the same basis as PIM-SM scales better than DVMRP. Although IETF expected that it would replace MSDP in the long run, BGMP nowadays suffers from lack of interest of vendors and operators, and thereby it did not enter the standards track [37].

5.3 Deployment of Allocation Protocols

It seems that there are very few if any implementations of multicast address allocation protocols in addition to the reference implementations. These allocation protocols are hardly used by anyone. Anyway, Microsoft has added MADCAP support to its DHCP server although this does not imply that someone is really using it [38]. It seems that there are no MASC implementations besides the reference implementation written by Pavlin Radoslavov.

The IETF MBONE Working Group is aware of the fact that there currently is no workable solution for global dynamic allocation. Moreover, application developers and deployers lack knowledge of MADCAP and SLP, which could be used to allocate temporary addresses and propagate them to other hosts. [25]

The reason for the lack of public interest towards dynamic multicast address allocation is that the use of multicast is altogether relatively rare. If there were masses of software

⁶Again, term *allocation* is used to mean both address allocation and assignment. In this model, layer 3 corresponds to allocation, whereas assignment is performed at layers 1 and 2.

⁷A MASC node can have several parents, which improves the availability of address blocks to be claimed.

employing multicast, it would be very convenient to allocate and assign group addresses dynamically. But as multicast has been and still is mainly a subject of academic research whereas commercial applications are few, dynamic allocation is not worth the additional complexity it causes.

6 Conclusions

Multicast routing means delivering datagrams across network boundaries to an arbitrary, dynamic set of receivers in a reasonable and efficient way. Three multicast routing protocols were discussed, namely DVMRP, MOSPF and PIM.

6.1 Comparison of Routing Protocols

DVMRP is a distance vector protocol originally based on RIP. DVMRP uses a flood-and-prune strategy to build distribution trees. Nowadays DVMRP is seldom used. Even those multicast regions still hanging MBONE usually run PIM internally while DVMRP is used between them. DVMRP itself does not address any interoperability issues but the other protocols can provide it the necessary information. DVMRP also provides a way to tunnel multicast traffic over non-multicast networks. [39]

MOSPF defines some new message types for OSPF in order to enable multicasting. MOSPF builds one distribution tree for each pair of source and group addresses which makes its scalability poor. The tree construction is based on Dijkstra's algorithm and the OSPF link-state advertisements. In addition, every MOSPF router of a certain area must be aware of all group memberships inside that area. That is a serious drawback since it causes much signaling overhead. However, this information is used to completely eliminate unnecessary data traffic and to calculate shortest-path distribution trees.

PIM actually defines two protocols for two different node distribution schemes. Both of them are independent of the unicast routing protocol used. Except for that, PIM-DM is quite similar to DVMRP. PIM-SM tries to minimize the overhead caused by flooding. The branches are explicitly joined to and pruned from the distribution trees. Nevertheless, this increases signaling overhead because the forwarding tree has to be updated even when no datagrams are sent to the group. Another objective of PIM-SM is to keep the amount of state information small. PIM-SM supports both shared and shortest-path distribution trees.

PIM-SM is currently the best alternative for delivering datagrams to sparsely distributed set of receivers. PIM-DM, DVMRP and MOSPF consume too much resources and their use is limited to relatively small environments [39]. Table 1 summarizes the main properties of these protocols.

6.2 Address Allocation

Perhaps the currently most reasonable way to do multicasting is to use the administratively scoped addresses, GLOP addresses or SSM unless you manage to persuade IANA that you really should have your own multicast address or address block. There are some address allocation and assignment

protocols but they are not used in practice. The use of multicast is relatively rare, so there is no real need for dynamic address allocation and assignment, at least yet.

References

- [1] S. Deering. *Host Extensions for IP Multicasting*. RFC 1112, IETF Network Working Group, 1989.
- [2] R. Hinden, S. Deering. *Internet Protocol Version 6 (IPv6) Addressing Architecture*. RFC 3513, IETF Network Working Group, 2003.
- [3] B. Cain et al. *Internet Group Management Protocol, Version 3*. RFC 3376, IETF Network Working Group, 2002.
- [4] D. E. Comer. *Internetworking with TCP/IP, Volume 1: Principles, Protocols and Architectures*. 4th ed. Prentice Hall, Upper Saddle River, New Jersey, USA, 2000.
- [5] S. Deering et al. *Multicast Listener Discovery (MLD) for IPv6*. RFC 2710, IETF Network Working Group, 1999.
- [6] D. Waitzman et al. *Distance Vector Multicast Routing Protocol*. RFC 1075, IETF Network Working Group, 1988.
- [7] T. Pusateri. *DVMRP Version 3*. IETF Inter-Domain Multicast Routing Working Group, Internet Draft (draft-ietf-idmr-dvmrp-v3-11), 2003.
- [8] T. Ferrari. *Introduction to Multicast*. INFN National Center for Telematics and Informatics, 2000. [referenced on February 3, 2004]
<http://www.cnaf.infn.it/~ferrari/papers/myslides/mcast-intro/>
- [9] J. Moy. *Multicast Extensions to OSPF*. RFC 1584, IETF Network Working Group, 1994.
- [10] D. Estrin et al. *Protocol Independent Multicast — Sparse Mode (PIM-SM): Protocol Specification*. RFC 2362, IETF Network Working Group, 1998.
- [11] A. Adams et al. *Protocol Independent Multicast — Dense Mode (PIM-DM): Protocol Specification (Revised)*. IETF PIM Working Group, Internet Draft (draft-ietf-pim-dm-new-v2-04), 2003.
- [12] T. Bates et al. *Multiprotocol Extensions for BGP-4*. RFC 2858, IETF Network Working Group, 2000.
- [13] Anon. *Configuring IP Multicast Routing*. Cisco Systems, Inc. [referenced on February 6, 2004]
<http://www.cisco.com/univercd/cc/td/doc/product/lan/cat3750/12119ea1/3750scg/swmcast.pdf>
- [14] Anon. *IP Services — Multicast*. Juniper Networks, Inc. [referenced on February 6, 2004]
http://www.juniper.net/products/ip_infrastructure/e-series/ip_services_multicast.html

	MOSPF	DVMRP	PIM-DM	PIM-SM
Root of tree	Source	Source	Source	RP or source
Tree is built	Before sending	After sending datagram and receiving prune		On joining
Traffic overhead	Signaling	Excess data traffic due to flooding		Signaling
Neighbor discovery	OSPF link-state	Own mechanism	Unicast routing information	
Traffic tunneling	No	Yes	No	No

Table 1: Routing protocol properties

- [15] P. Rajvaidya, K. C. Almeroth. *Analysis of Routing Characteristics in the Multicast Infrastructure*. Proceedings of IEEE INFOCOM 2003, San Francisco, California, USA, March 30 – April 3, 2003.
- [16] R. B. Bellman. *The Push for IP Multicasting*. Business Communications Review, Vol. 27, No. 6, pp. 28–32. 1997.
- [17] Anon. *Cisco IOS Multicast Q&A*. Cisco Systems, Inc. 2003. [referenced on February 25, 2004]
http://www.cisco.com/warp/public/cc/pd/iosw/prodlit/mcast_qp.pdf
- [18] S. Deering et al. *The PIM Architecture for Wide-Area Multicast Routing*. IEEE/ACM Transactions on Networking, Vol. 4, No. 2, pp. 153–162. 1996.
- [19] A. Ballardie. *Core Based Trees (CBT) Multicast Routing Architecture*. RFC 2201, IETF Network Working Group, 1997.
- [20] B. Fenner (ed), D. Meyer (ed). *Multicast Source Discovery Protocol (MSDP)*. RFC 3618, IETF Network Working Group, 2003.
- [21] S. Bhattacharyya (ed). *An Overview of Source-Specific Multicast (SSM)*. RFC 3569, IETF Network Working Group, 2003.
- [22] R. Vida (ed), L. Costa (ed). *Multicast Listener Discovery Version 2 (MLDv2) for IPv6*. Internet Draft (draft-vida-ml-d-v2-08), 2003.
- [23] D. Meyer et al. *Source-Specific Protocol Independent Multicast in 232/8*. IETF MBONE Deployment Working Group, Internet Draft (draft-ietf-mboned-ssm232-07), 2004.
- [24] D. Thaler et al. *The Internet Multicast Address Allocation Architecture*. RFC 2908, IETF Network Working Group, 2000.
- [25] S. Leinen (ed). *MBONE Deployment Working Group Minutes*. 59th IETF Meeting, Seoul, South Korea, February 29 – March 4, 2004. [referenced on April 5, 2004]
<http://www.ietf.org/proceedings/04mar/minutes/mboned.htm>
- [26] D. Meyer. *Administratively Scoped IP Multicast*. RFC 2365, IETF Network Working Group, 1998.
- [27] D. Meyer, P. Lothberg. *GLOP Addressing in 233/8*. RFC 3180, IETF Network Working Group, 2001.
- [28] R. Hinden et al. *IPv6 Global Unicast Address Format*. RFC 3587, IETF Network Working Group, 2003.
- [29] B. Haberman, D. Thaler. *Unicast-Prefix-based IPv6 Multicast Addresses*. RFC 3306, IETF Network Working Group, 2002.
- [30] M. Handley, S. R. Hanna. *Multicast Address Allocation Protocol (AAP)*. IETF MALLOC Working Group, Internet Draft (draft-ietf-malloc-aap-04), 2000.
- [31] S. R. Hanna et al. *Multicast Address Dynamic Client Allocation Protocol (MADCAP)*. RFC 2730, IETF Network Working Group, 1999.
- [32] E. Guttman et al. *Service Location Protocol, Version 2*. RFC 2608, IETF Network Working Group, 1999.
- [33] P. I. Radoslavov et al. *The Multicast Address-Set Claim (MASC) Protocol*. RFC 2909, IETF Network Working Group, 2000.
- [34] P. I. Radoslavov et al. *A Claim–Collide Mechanism for Robust Distributed Resource Allocation*. USC Department of CS Technical Report 99-711. 1999.
- [35] D. Thaler. *Border Gateway Multicast Protocol (BGMP): Protocol Specification*. IETF BGMP Working Group, Internet Draft (draft-ietf-bgmp-spec-06), 2004.
- [36] S. Kumar et al. *The MASC/BGMP Architecture for Inter-Domain Multicast Routing*. Proceedings of ACM SIGCOMM '98, pp. 93–104. Vancouver, British Columbia, Canada, September 2–4, 1998.
- [37] B. Fenner. *Note for draft-ietf-bgmp-spec*. IETF BGMP Mailing List, 2004. [referenced on April 7, 2004]
<http://www1.ietf.org/mail-archive/working-groups/bgmp/current/msg00001.html>
- [38] Anon. *Multicast Address Allocation*. Microsoft Corporation. [referenced on February 10, 2004]
http://www.microsoft.com/technet/prodtechnol/windowsserver2003/proddocs/standard/sag_DHCP_und_MulticastAllocation.asp
- [39] J. Lin, R-S. Chang. *A Comparison of the Internet Multicast Routing Protocols*. Computer Communications, Vol. 22, No. 2, pp. 144–155. 1999.