Operating System

An Overview of Inter-Domain Multicast Routing

White Paper

Abstract

Multicast routing is used to distribute data to multiple recipients. Until recently, the multicast routing protocols have concentrated on multicasting within a single domain. There is, however, also a demand for multicasting on a global scale, across the Internet. To accomplish this, a new set of protocols is under development. This paper discusses these protocols. It first explains the current inter-domain solution, which involves MBGP and MSDP, and then discusses several proposed long-term solutions, focusing on BGMP and MASC, but also including Express and Simple Multicast. This paper is intended as an introduction to the subject of inter-domain multicasting, to be used by IT managers who want an overview of the subject before reading the relevant RFCs.

Microsoft Corporation • One Microsoft Way • Redmond, WA 98052-6399 • USA

0400

Contents

Introduction

While intra-domain multicast routing is fairly well established, with Protocol Independent Multicast-Sparse Mode (PIM-SM) accepted as the *de facto* multicast routing protocol, inter-domain multicast routing presents another set of issues. The goal is no longer to deploy IP-based multicast services from within a service provider's network. Instead, the focus is now on how services can be shared and distributed between providers. A number of issues must be resolved for this to occur. They include the following:

- Handling differences in topology and/or policy for unicast and multicast services.
- Avoiding third-party dependencies, which means that service providers don't have to rely on rendezvous points (RP) that lie in another service provider's domain. (This assumes, of course, that PIM-SM is the intra-domain multicast routing protocol.)
- Placing an RP where it is convenient rather than being forced to place it at an interconnection point.
- Establishing mechanisms for multicast address allocation.

This paper introduces some of the solutions that are emerging to make inter-domain multicast routing a viable technology for delivering content over the Internet. It first discusses the inter-domain routing protocol, MultiProtocol Border Gateway Protocol (MGBP), and the source-discovery protocol, Multicast Source Discover Protocol (MSDP). Together with PIM-SM as the intra-domain multicast routing protocol, these form a solution that is already being deployed with some success.

The paper then discusses a longer-term solution that is currently being investigated. This solution uses a hierarchical addressing scheme called the Multicast Address-Set Claim (MASC) protocol and the Border Gateway Multicast Protocol (BGMP). These two protocols, used in conjunction with MGBP and intra-domain routing protocols, may be how inter-domain multicast routing will be performed in the future. Two alternative solutions, Express and Simple Multicast, are also discussed.

The paper is intended for IT managers who would like a broad overview of inter-domain multicasting before they begin reading the appropriate RFCs (Request for Comments). It assumes a thorough knowledge of networking in general and intra-domain multicasting in particular, with an emphasis on the PIM-SM routing protocol. There is a white paper available on PIM-SM that explains this protocol and some general concepts, such as Reverse Path Forwarding (RPF), shared trees, and shortest path trees. It is located at:

http://www.microsoft.com/windows2000/library/howitworks/communications/trafficmgmt/pimsm2.asp

An understanding of autonomous systems (AS) and the Border Gateway Protocol (BGP), used for unicast inter-domain routing, is also suggested, but a brief overview of both follows.

Autonomous Systems

The classical definition of an AS, also called a *domain*, is a set of routers under a single technical administration. An AS has

- An interior gateway protocol (IGP), such as Open Shortest Path First (OSPF) and a set of metrics for routing packets within the AS.
- An exterior gateway protocol (EGP), such as BGP to route packets to other ASs.

In reality, it has become common for a single AS to use several IGPs and, sometimes, several sets of metrics. However, even if multiple IGPs and metrics are used, when viewed from the perspective of another AS, an AS seems to have a single coherent interior routing plan and presents a consistent picture of which destinations are reachable through it.

An AS number (written as AS*N*) is a 16-bit integer assigned by InterNIC. It is used by an exterior gateway protocol (EGP) such as the Border Gateway Protocol (BGP) to implement policy routing and to avoid top-level routing loops. Policy routing means that the type of routing information that crosses the border between two ASs can be controlled.

There are three types of AS:

- A stub AS is connected to only one other AS. For routing purposes, it could be regarded as a simple extension of the other AS. In fact, most networks with a single Internet connection don't have a unique AS number assigned, and their network addresses are treated as part of the parent AS.
- A transit AS has connections to more than one other AS and can be used as a conduit for traffic (*transit traffic*) between other ASs. Most large service providers are transit ASs. Two examples of transit ASs are UUNET and MCI.
- A multi-homed AS has connections to more than one other AS, but does not allow transit traffic to pass, although its interior hosts may route traffic through multiple ASs. This is the typical configuration for a large corporate network that does not want to pass traffic for others.

Interior routing policies and protocols must be established within each AS, enabling it to route packets internally. Exterior routing, with a routing protocol such as BGP, is used to interconnect the various ASs, with their independent interior routing protocols, into a single coherent Internet.

BGP-4 Overview

The Border Gateway Protocol, version 4 (BGP-4) is the current unicast exterior routing protocol used for the global Internet. It is a *path-vector* protocol, which is essentially a distance-vector protocol with a twist. Instead of using distance (or hops) to a destination as its primary metric, BGP uses a path to a destination. This path is a sequence of AS numbers, or, to put it another way, it is a route to a certain set of Classless Inter-Domain Routing (CIDR) prefixes. (CIDR is discussed in greater detail later in this paper.) Paths are tagged with various path attributes. Two of the most important are AS_PATH and NEXT_HOP, which are discussed below.

Two routers that are configured to speak BGP to each other so that they can exchange routing information are called *BGP peers*, and together they form a *BGP session*. For both routers to be sure that none of the information one has sent the other has been lost, the connection must be reliable. To accomplish this, BGP uses TCP as its transport protocol.

Once a connection is established, BGP peers exchange complete copies of their routing tables, which can be quite large. However, after this initial transfer, the routers only exchange revisions to those tables, which makes extended BGP sessions more efficient than shorter ones. While the session is ongoing, update messages are sent from one router to another each time one of the routers knows about a new BGP route or needs to withdraw a previous announcement.

The AS_PATH Attribute

One of BGP-4's functions is to detect (and avoid) loops at the AS level, using the AS_PATH attribute. A simplified view of AS_PATH is that it is the list of ASs that a packet traverses to reach its destination. Routers detect loops by checking for their own AS number in AS_PATHs received from neighboring ASs. This is shown in Figure 1, below:

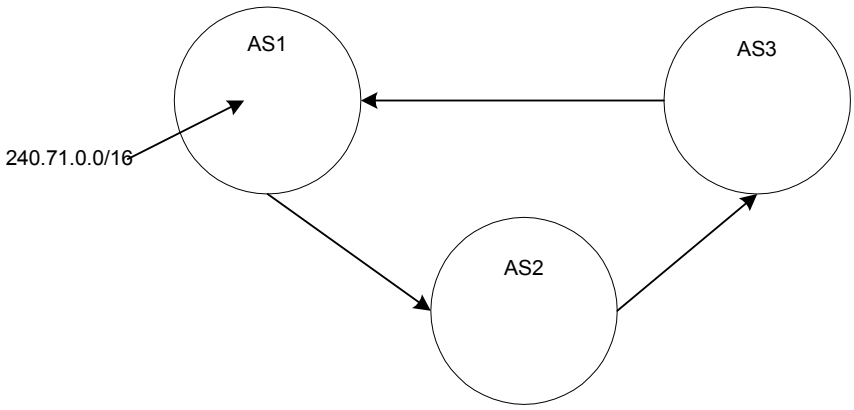

Figure 1. Preventing loops with AS_PATH

In this example, the prefix 240.71.0.0/16 is injected into the AS1 BGP routing table. AS1 advertises this route to AS2, making sure that AS1 appears in the AS_PATH attribute. This way, AS2 knows that AS1 was the first AS to inject the prefix into BGP. Next, AS2 advertises the route to AS3, appending its own AS number to AS_PATH. At this point, AS3 knows that the prefix originated with AS1 and was passed to AS2, and finally to AS3. If AS3 passes the advertisement to AS1, AS1 may decide not to accept it because it sees its own number in the AS_PATH attribute.

The syntax of the AS_PATH attribute is made more complex because of its support for path aggregation, which is when multiple paths are collapsed into one. For a more complete description, consult the relevant RFC.

The NEXT_HOP Attribute

Each time a BGP path advertisement crosses an AS boundary, the NEXT_HOP attribute changes to the IP address of the boundary router. However, when a BGP path advertisement passes through BGP speakers (BGP-enabled routers) within the same AS, the NEXT_HOP attribute is unchanged. This means the NEXT_HOP attribute is always the IP address of the first router in the next AS, even though this router may actually be several hops away. As a consequence, any interior routing protocols, such as OSPF (Open Shortest Path First), must include routing table entries that extend one hop beyond the AS boundary.

E-BGP and I-BGP

BGP routers communicate with neighboring BGP speakers in other ASs, and they communicate with other BGP speakers within their own AS. The common acronym for communication with neighbors in an exterior AS is E-BGP, while communication within an AS (the interior AS) is I-BGP. To prevent loops, an I-BGP speaker cannot pass routing information it learns from one I-BGP speaker on to another I-BGP speaker. This implies that all I-BGP speakers must be *fully meshed*, which means that each BGP speaker in the AS is in direct communication with every other BGP speaker. (This doesn't necessarily mean that the routers are physically connected, but that there is a TCP connection between them.) When a BGP routing update is received from a neighboring AS, it must be relayed directly to all the BGP speakers in the AS. There cannot, for example, be a BGP path from one router, through another, to a third within the same AS.

BGP Policies

A policy is information that enables a BGP speaker to preferentially rank routes. Most protocols associate a path with an integer that represents the cost to reach the path's destination. Generally, the route with the lowest cost is the preferred route. BGP, on the other hand, does not use this method. Instead, it

allows complex policies to be developed for selecting the best route. The details of implementing these policies are largely undefined. The RFC says only that the computation should be based on "…preconfigured policy information. The exact nature of this policy information and the computation involved is a local matter." One way of implementing a simple policy is to use the AS_PATH attribute. Since it includes a list of ASs used to reach a destination, it can be used to implement a directive such as, "avoid all routes through AS*N*."

Multicast Border Gateway Protocol

BGP, as we stated, is a unicast exterior gateway protocol. A multicast counterpart must recognize that, for a variety of reasons, the service provider's multicast network need not be congruent with that of the unicast network. These differences may be because of topology and/or policy constraints. Some of the routers in the network may not be multicast-enabled, for example, or there may be portions of the network that, because of policy decisions, reject multicast traffic. Therefore, in order to construct distribution trees, an exterior multicast routing protocol must not only know the path back to the source; it must also be able to determine which parts of the infrastructure are available for multicasting. Additionally, for easier adoption, the protocol should be based on a familiar, existing unicast model. In short, a multicast exterior gateway protocol should fulfill these requirements:

- It should be able to distinguish between unicast and multicast topologies.
- It should use a familiar model for terminology, configuration, and operation.
- It should have a robust set of peering and policy controls.

Because BGP already satisfies the latter two requirements, the logical solution was to enhance BGP rather than to devise an entirely new protocol. In a general sense, these extensions, defined in RFC 2283, were added to enable BGP to carry routing information for network layer protocols other than IPv4 and not simply to enable multicast. Multicast is only one of the results of these additions. The acronym MBGP is often read as Multicast BGP, but the correct name is Multiprotocol BGP.

With MBGP, both unicast and multicast routes are carried in the same session but in different routing tables. Because MBGP is an enhanced version of BGP-4, all the familiar policy and configuration tools are available for multicast.

MBGP carries multicast routes by adding the Subsequent Address Family Identifier (SAFI) to either of two new path attributes. SAFI specifies whether the forwarding information is unicast, multicast, or unicast/multicast. The MP_REACH_NLRI attribute, which stands for MultiProtocol Reachable Network Layer Reachability Information, includes the set of reachable destinations, as well as the next hop information for forwarding to those destinations. The MP_UNREACH_NLRI attribute, which stands for MultiProtocol Unreachable Network Layer Reachability Information, is the set of unreachable destinations. These attributes are included in the BGP update messages, which are used to advertise single feasible routes to peers or to withdraw multiple unfeasible routes.

The upshot of all this is that, with MBGP, a router needs to know only its own internal topology and the path to each of the other ASs rather than the entire flat multicast topology. MBGP is backward compatible with BGP-4. Routers that don't understand the extended attributes simply ignore them.

Drawbacks to MBGP

Two drawbacks to MBGP are as follows:

- It increases the size of the routing tables.
- There is a potential for storing redundant information because multiple sets of routes for the same prefixes may be stored in the routers.

These issues, however, are rather trivial compared to the more important problem of convincing service providers to upgrade their network. The BGP routers that would use the extensions are probably critical production routers, already under a significant unicast load. Also, because it is a newer protocol, there is a higher probability that the software contains undocumented bugs.

Business pressures will probably be the impetus that drives ISPs to adopt MBGP. The convergence of Internet access and voice with broadcast video over cable has fueled the demand for full-service, single-bill providers, who supply both high-speed data services and some form of broadcast medium, such as video.

Multicast Source Discovery Protocol

MBGP resolves congruency issues, but it does not solve the problem of third-party dependencies or of flexibility in RP placement. Service providers do not want to rely on a competitor's RP for transmitting multicast traffic, but they must be able to get information from the multicast source to the receivers, regardless of where the source's RP is located. They also may need to place RPs someplace other than at a common interconnection point.

The Multicast Source Discovery Protocol (MSDP) was developed to solve these problems, at least in the short term. (We will discuss why this may only be a short term solution later in this paper.) MSDP uses inter-domain source trees rather than shared trees. This approach means that RPs require only the paths to the active sources outside of their own domains. There is no requirement to use a competitor's RP or for an ISP to locate an RP at a particular place. (MSDP only works with PIM-SM.)

With MSDP, the RP in one domain establishes peer relationships with RPs in other domains via a TCP connection. These relationships are used to exchange information about sources for particular groups. If a receiver in the local domain requests a source from a remote domain, a source tree is built the same way as it is in PIM-SM. Here is a summary of the steps used to determine that there are active sources in a different domain:

1. The source, located in some domain, originates traffic.
2. The source's designated PIM-SM router sends the data, encapsulated in a Register message, to the RP within the local domain.
3. The RP constructs a Source Active (SA) message and sends it to its MSDP peers. Each SA message contains the IP address of the data source, the address of the multicast group, and the IP address of the originating RP, which is the RP that is sending the SA. These messages are sent periodically, so long as the source is active.
4. Each peer RP with interested receivers sends a PIM-SM join message directly back to the source, not to the external RP. Once the reverse forwarding path has been established, and the RP is forwarding the data, group members have the option, according to PIM-SM conventions, of switching to a shortest-path tree.
5. Each peer floods the SA message away from the originating RP, using the AS_PATH parameter to perform a Reverse Path Forwarding (RPF) check back to the peer RP.

The following set of drawings illustrates this process. Figure 2, below, shows two domains, with the receivers located in AS1 and the source located in AS2:

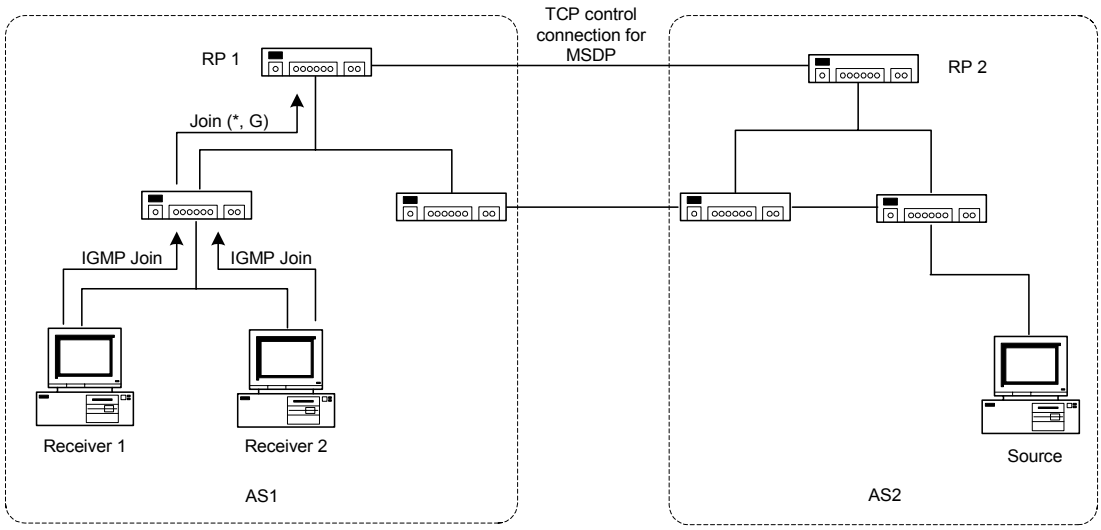

Figure 2. Receivers join a multicast group

The receivers send IGMP join messages to the leaf router, which then sends a PIM-SM join message to the domain's RP. This is the standard way for hosts to join a multicast group and for routers using the PIM-SM protocol to signal the RP to begin sending the multicast data. The oddity is that the source is not located within the same domain as the receivers. The question is: how does the RP in AS1 get the data from the source in AS2 once the source begins transmitting? Figure 3, below, shows what happens once the source becomes active:

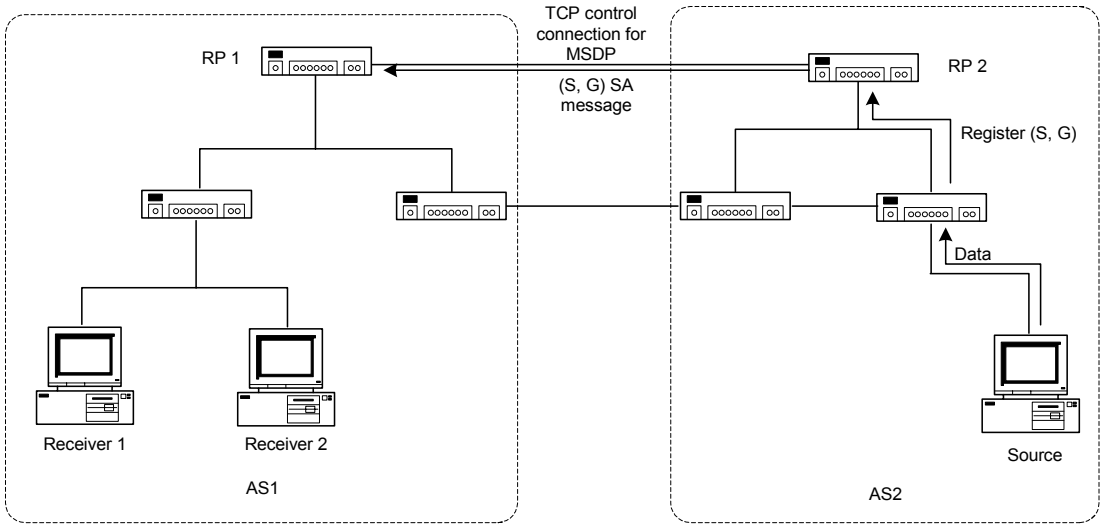

Figure 3. Multicast source begins transmitting data

Once the router in Domain 2 begins receiving data from the source, it sends a PIM-SM register message, with (S, G) state, to the RP, where S is the source address and G is the multicast group address. Again, this is simply using the

PIM-SM protocol. The MSDP protocol comes into play when RP 2 sends an SA message to its peer, RP 1, over the TCP control connection that exists between them. Figure 4, below, shows how RP 1 responds to this SA message:

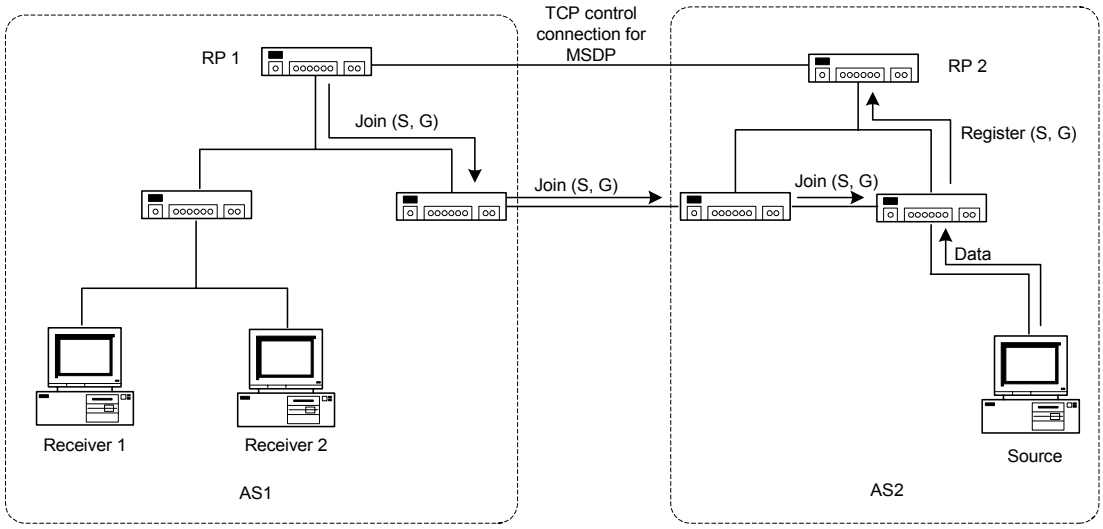

Figure 4. RP sends a join message to the inter-domain source

RP 1 has interested receivers, so it sends a join back to the source, effectively grafting that branch of the source tree onto the requesting domain's tree. Figure 5, below, shows that data now flows from the source in AS1 to the RP in AS2:

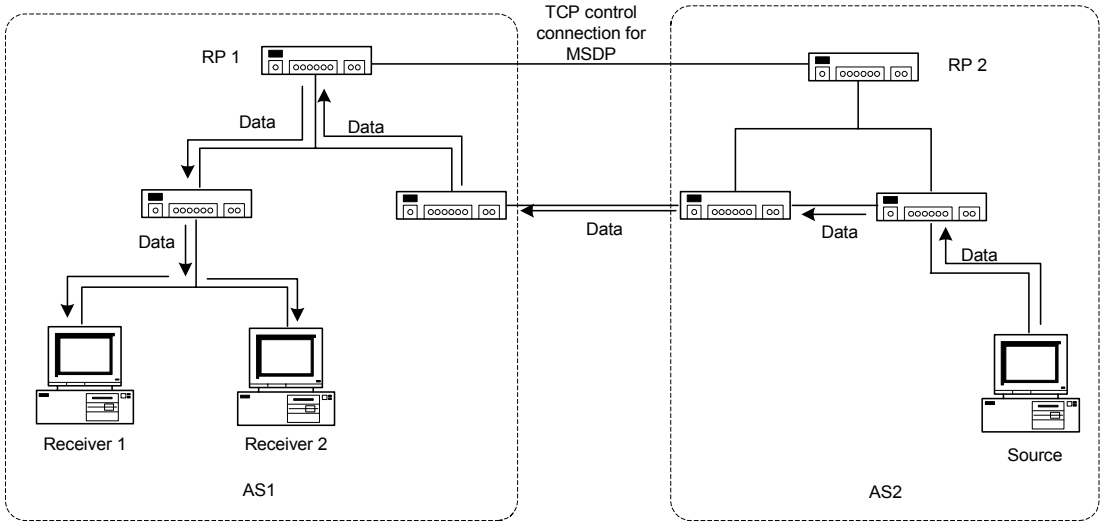

Figure 5. Data flows from source to RP

Once it starts receiving the data, RP 1 distributes the traffic to the receivers. Notice that RP 1 functions independently of RP 2 and that there was no need to locate RP1 at a point connecting AS1 and AS2.

Figure 6, below, shows how SA messages are flooded away from an originating RP toward the rest of the network:

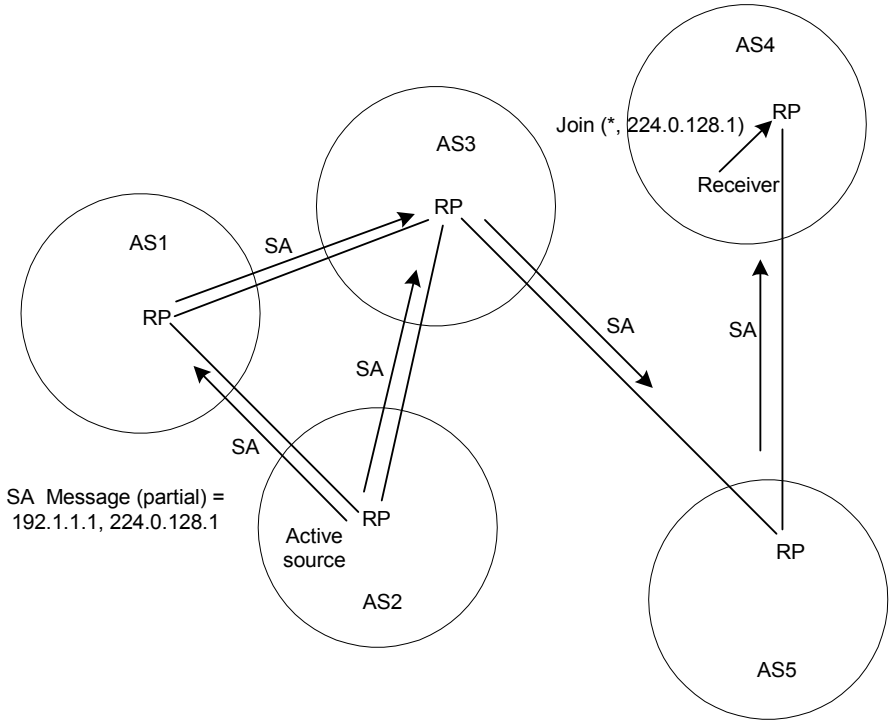

Figure 6. Flooding SA messages

This group of ASs has an active source in AS2 and an interested receiver in AS4. The RP in AS2 sends an SA message containing the source address and the multicast group address to each of its peers. They, in turn, send the message onward. Notice that the RP in AS3 will receive two SA messages. By examining the AS_PATH attribute, it can perform an RPF check and forward the message that arrived on the interface that is the shortest path back to the sender, while dropping the other message. In this case, the message from AS2 is forwarded while the message from AS1 is dropped. (The exception to this is a stub AS, which should be able to accept all SA messages because there is only one exit and that leads to the default peer.)

Problems with MSDP

There are three problems associated with MSDP. These are:

- Join latency
- Bursty sources
- Scalability

We will discuss each in turn.

Join Latency

Because some period of time elapses between SA messages, it's possible that a receiver that joins a group between messages will not receive some of the multicast data. For example, assume that an RP receives an SA message but has no interested receivers—it will discard the message. If a receiver then joins the group before the next SA message is received, it will suffer *join latency* proportional to the amount of time left before the next SA message is issued.

To solve this problem, MSDP routers can be configured to cache SA messages, in the hopes that when a new receiver joins the group, the source will still be active. If one MSDP peer caches SA messages, then other, non-caching, MSDP peers can take advantage of this cached information by sending an SA Request message to the caching MSDP peer. In other words, minimizing join latency means storing more state information--the more information that is cached, the shorter the latency period. Unfortunately, this is not a particularly desirable situation because, in inter-domain routing, large amounts of state may be required for an acceptably small latency time.

Bursty Sources

A similar problem exists with bursty sources. This is an important concern because burstiness is a characteristic of some commonly used multicast announcement tools such as SDR, which periodically sends session announcement protocol (SAP) packets. Typically only one, or, at most, a few, of these packets are sent at a time. We will use SAP packets as an example of why bursty sources can cause problems.

When an RP receives an SAP packet, it floods an SA message, and, once the reverse forwarding path is established, the receivers can send joins toward the source. There can be problems to this approach because no multicast forwarding state existed before the SAP packet was received. It takes some amount of time both to forward SA messages and for receivers to establish forwarding state. By the time all this happens, the initial SAP packets may have been dropped before the receivers can actually get the data.

If the SAP packets are sent after the original join state expires (the default value is three minutes), then the whole process must be repeated. In almost all cases, the first few packets of a transmission are lost. If the transmission is short as well as bursty, all the data may be lost. The original solution suggested in the MSDP protocol is that, to make sure the first few packets are delivered, SA messages carry data as well as control information. However, using a TCP control connection to carry data is considered a clumsy solution and discussions within the MSDP Working Group have resulted in the specification being modified to say that multicast data transmitted between peers should use UDP rather than TCP.

Scalability

If multicast becomes a highly successful method of transmitting data, the overhead associated with MSDP may become unacceptably large. If there are thousands of multicast groups, then the number of SA messages flooding the network could become too large to handle. When groups are dynamic, either because of bursty sources or frequent join/leaves, the management overhead can be significant. The consensus is that MSDP is an interim solution that will not suffice for the long-term. However, given that long-term solutions are not yet ready to be deployed, MSDP serves as a functional, acceptable compromise for the immediate future.

Long-Term Solutions

While the PIM-SM/MBGP/MSDP protocol suite is currently being used to implement inter-domain multicast routing, other, more long-term solutions are being investigated. Our focus will be on the most generally accepted proposal, which is made up of two complementary protocols. They are the Border Gateway Multicast Protocol (BGMP) and the Multicast Address-Set Claim (MASC) protocol. (The similarity in name and initialism between BGMP and MBGP is unfortunate and has caused endless confusion. The two are very different., as shown in the sections on BGMP and MBGP.) These protocols are used in conjunction with MBGP and some intra-domain multicast routing protocol.

BGMP creates bi-directional shared trees between domains for distributing multicast data. Incorporating concepts from PIM-SM and CBT (core-based trees) into its design, BGMP requires that each global multicast group be associated with a single RP. However, because it is an inter-domain protocol, BGMP expands the definition of an RP from a single router to an entire domain.

A key function, then, of BGMP is to decide where to place the RP. The solution lies in a strict multicast address allocation scheme that associates particular multicast addresses with particular domains. The BGMP architecture works in tandem with MASC, which is a hierarchical addressing protocol. Although BGMP is not dependent on MASC in particular, (another addressing protocol that does the same thing could be used) MASC is the currently accepted solution and the two are generally considered together. Because BGMP can really only be understood in the context of the underlying addressing scheme, we will first explain MASC, which is a part of an overreaching addressing architecture called Multicast Address Allocation (MALLOC), and then move on to BGMP.

Multicast Address Allocation

Multicast address allocation has traditionally been performed by session directories, specifically SD, which has been used since 1993, and SDR, which has been used since 1994. The basic model for both of them is the same. They perform the following tasks:

- Inform potential users that a session exists.
- Include the information required for joining the session.
- Inform the network of the multicast address that is being used so that other sessions will not also try to use it.

Although addresses are selected randomly, by advertising addresses that are already taken, collisions can be avoided. Collisions occur when a number of multicast groups use the same multicast address and traffic from each group is delivered to members of all groups.

Aside from preventing address collisions, the SD/SDR model incorporates some very desirable characteristics, such as the following:

- It is decentralized.
- It is robust.
- It requires no configuration.
- It is always available.

The difficulty is that, as with so many other multicast solutions, it will not scale. As the number of allocated addresses increases, the number of addresses that have been allocated, but are not yet known because of packet loss and propagation delay, becomes significant and the number of address collisions becomes unacceptable.

In general, an address allocation scheme that will work on a large scale must have, in addition to the characteristics already mentioned, the following attributes:

- It should be able to allocate addresses dynamically.
- It should support address aggregation.
- It should be hierarchical.

Dynamic Address Allocation

One of the underlying assumptions for allocating multicast addresses within a global Internet is that Class D addresses (multicast addresses) should be dynamically allocated by their administrative scope for a finite period of time. Dynamic allocation means that almost all applications would be assigned temporary addresses that could be reused. Only a few applications would be assigned permanent addresses. An example of such an application is the

Network Time Protocol, which uses the static address 224.0.1.1. In other words, static multicast addresses should only be assigned to applications that provide a part of the basic infrastructure.

Administratively scoped IP addresses have two basic characteristics:

- Packets addressed to administratively scoped multicast addresses do not go across administrative boundaries.
- Administratively scoped IP addresses are assigned locally. This means they do not need to be unique across administrative boundaries, thus permitting reuse of the multicast address space.

Routers located at the boundary of an administrative domain must support administrative scoping on each of their interfaces. These routers will not forward administratively scoped multicast packets outside their administrative domain.

Address Aggregation

The number of networks contained in the Internet is growing at an exponential rate. This means, for example, that since the BGP unicast routing tables basically have an equivalent number of entries, they too will grow at an exponential rate, if no alternative means of representation is used. The same reasoning that applies to unicast routing tables applies to multicast routing tables. For scalability, the number of advertised routes must be kept to a minimum. The CIDR system was implemented to improve the scaling factor of the Internet.

The CIDR System

Adopting CIDR has helped this situation because it enables *address aggregation*, which means that consecutive address prefixes can be combined into a single prefix. This reduces the number of routing table entries. For example, the address prefixes 128.8.0.0/16 and 128.9.00/16 can be aggregated to 128.8.0.0/15 because only their 16^{th} bits differ. .

When a border router, X, advertises this aggregate to a second border router, Y, then Y knows that, through X, it can reach the hosts in all the component address prefixes that make up the aggregate 128.8.0.0/15. There is no reason for X to advertise every single prefix. In other words, CIDR summarizes routing information so that the size of a router's table is reduced while its level of connectivity is maintained.

Here is another example. Figure 7, below, shows a service provider with a number of networks as its customers:

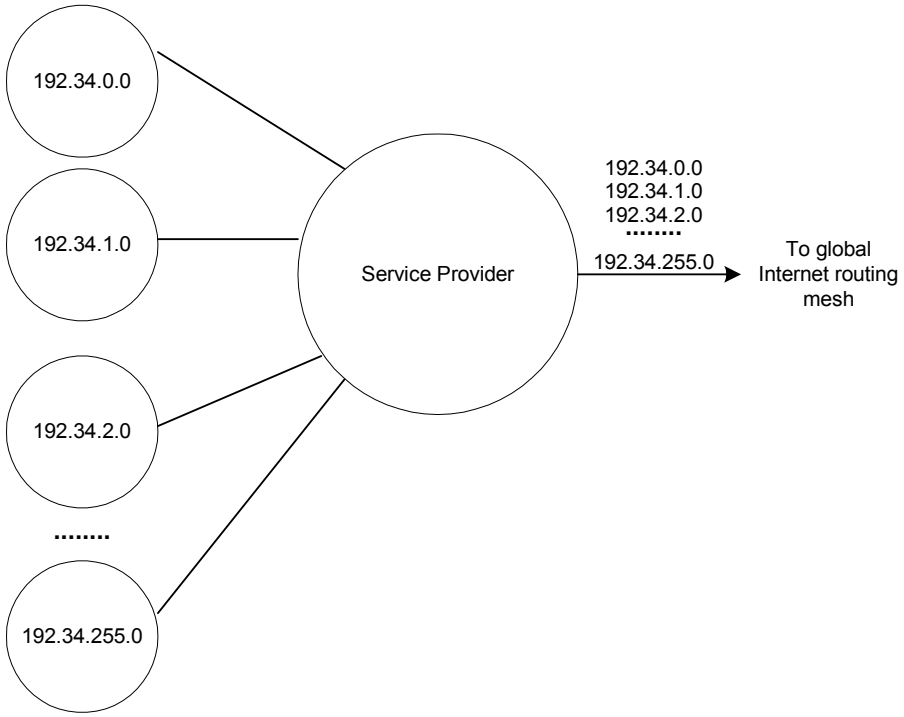

Figure 7. Example of inter-domain routing without CIDR

In this example, the service provider's customers are all Class C networks whose addresses start with 192.34. Despite this commonality, without CIDR, the service provider must announce each of these networks separately. The next example, shown in Figure 8, below, shows how CIDR can improve this situation:

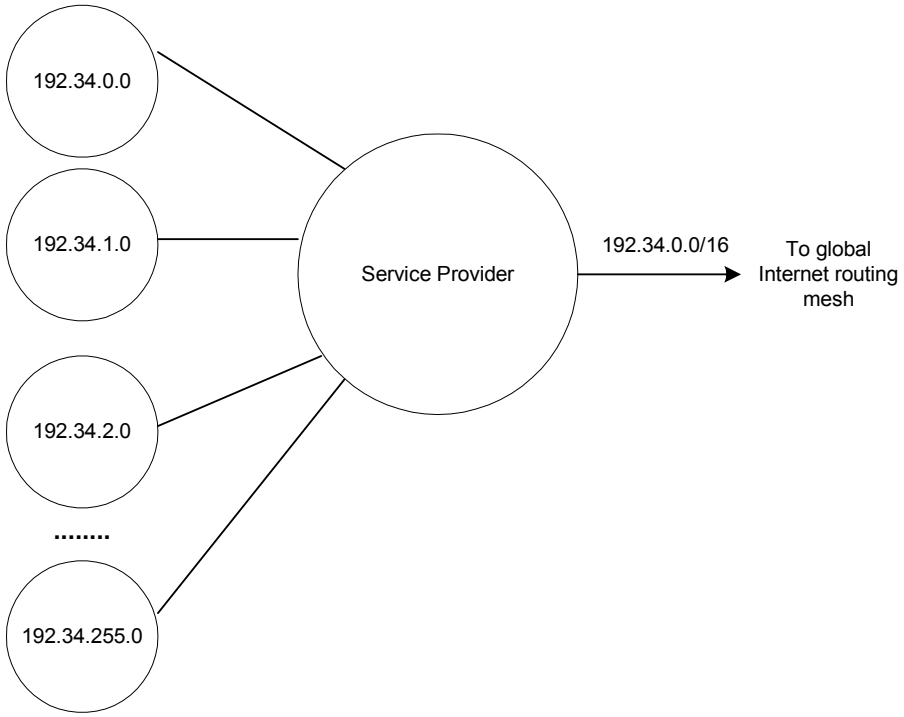

Figure 8. Example of inter-domain routing with CIDR

With CIDR, the service provider can aggregate every network address into a single advertisement. Note that the service provider itself doesn't see a reduction in the size of its routing table because it still must carry the specific routes needed to reach each customer. The reductions apply to other providers in the Internet. A single entry in their routing tables allows them to reach the service provider, which is all they need to know to reach the service provider's customers.

Hierarchical Organization

CIDR is only efficient when the address prefixes are assigned to ASs in a structured way. Networks can be aggregated and routing can be hierarchical only when the addressing is hierarchical, which means that the prefixes will nest, one within the other. The easiest way to ensure hierarchical addressing is to have sites obtain addresses from their providers rather than from a central repository. A decentralized approach is more efficient and scalable.

The Multicast Address Allocation Architecture

Any multicast addressing scheme must fit into the general solutions applied to unicast addressing. That is, it must be a hierarchical approach that allows a logical distribution of multicast address ranges. One proposal is called the

Multicast Address Allocation Architecture (MALLOC). This architecture uses a three-tiered approach comprised of these protocols:

- Multicast Address-Set Claim (MASC), which acts as a top-level address allocation protocol and operates between domains
- Address Allocation Protocol (AAP), which allocates addresses within a domain
- Multicast Address Dynamic Client Allocation Protocol (MADCAP), which is used by hosts to request addresses from a Multicast Address Allocation Server (MAAS)

Each of the protocols that make up MALLOC can be used independently from the other protocols.

This architecture is shown in Figure 9, below:

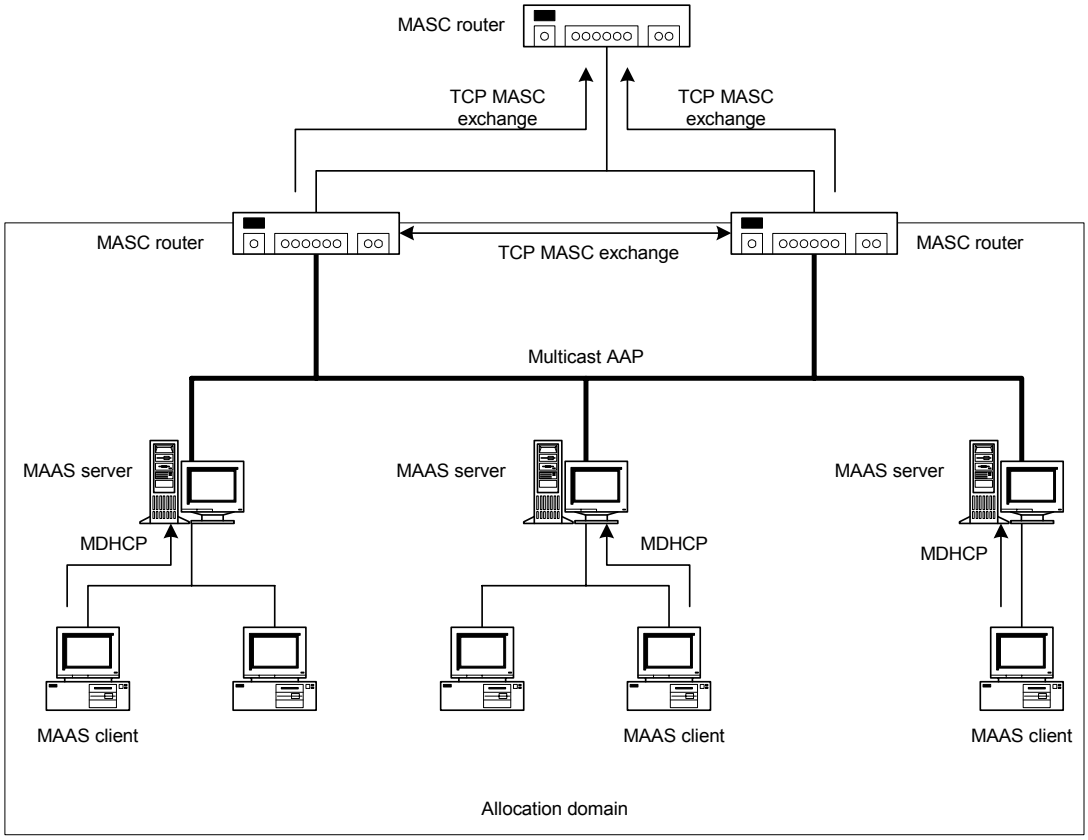

Figure 9. Multicast Address Allocation Architecture

This diagram shows a single allocation domain containing the three levels that make up MALLOC. An allocation domain is an administratively scoped multicast-capable region of the network and normally coincides with an AS.

The MASC protocol is the basis of the hierarchical address allocation architecture. It dynamically allocates multicast address ranges from which

groups within the domain get their specific multicast addresses. MASC nodes are typically routers.

The AAP protocol is an intra-domain protocol used for allocating multicast addresses within the domain. It is used by Multicast Address Allocation Servers (MAAS) to coordinate the allocation so duplicate addresses are not used. The MAAS servers themselves are not required as part of MALLOC, but not using them increases the chances of address collisions. This protocol works the same way as SDR, except that it only operates within a domain and it does not advertise sessions.

MADCAP is a simple request/response protocol that allows clients to dynamically request a multicast address from a server. Clients locate these servers either by being manually configured or by issuing a multicast request. Although MADCAP was originally based on DHCP (and was called MDHCP), they are now completely separate, with no dependencies between the two.

The sequence of steps used to allocate multicast addresses is shown in Figure10, below:

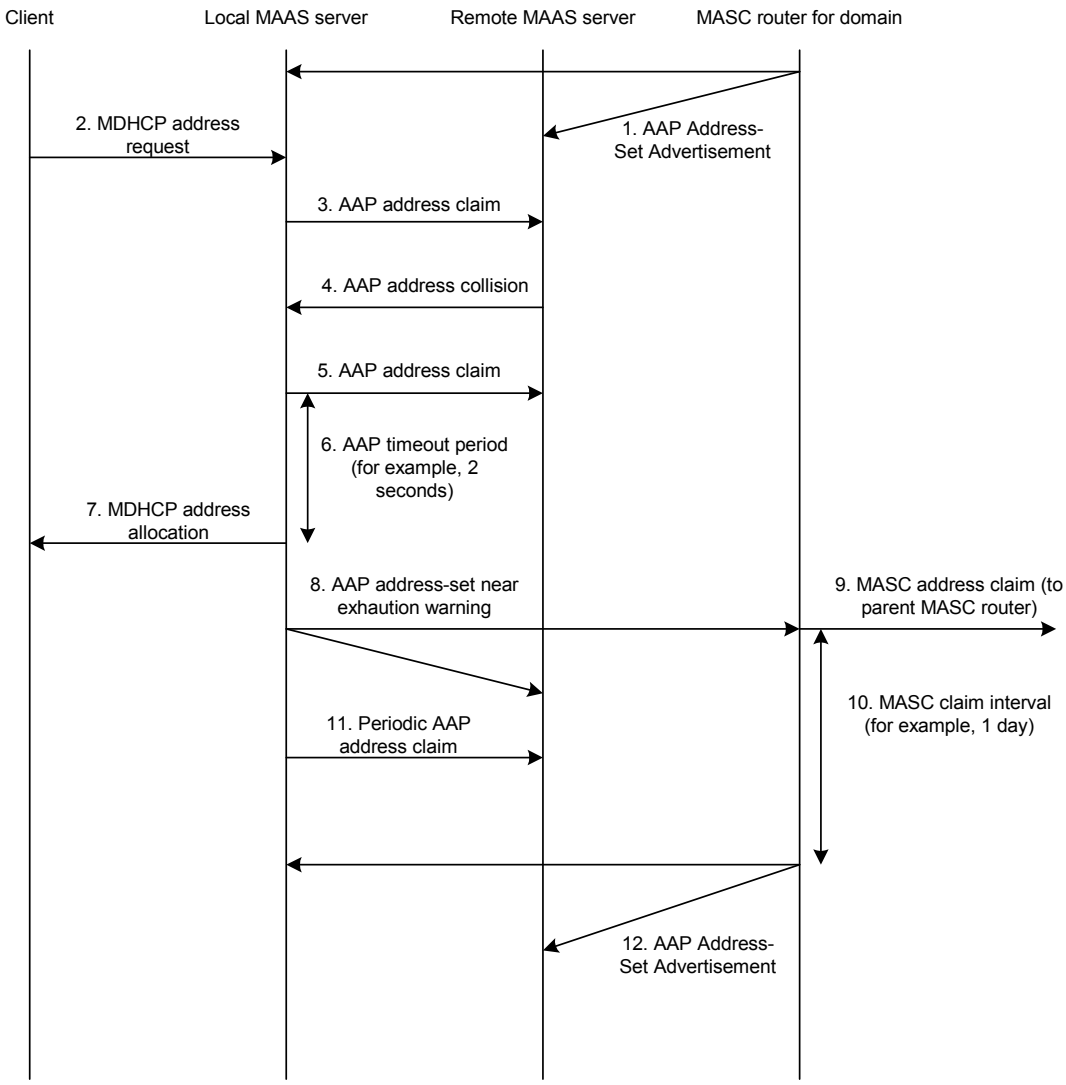

Figure 10. Multicast Address Allocation Process

This diagram shows the sequence of events that occurs when allocating multicast addresses. The MASC protocol procures ranges of addresses and distributes them to MAAS servers via the AAP protocol. Clients receive individual addresses from these ranges by communicating, via MADCAP, with a local MAAS server.

The MASC Protocol

Because it is the foundation of the MALLOC architecture, and because it is tied so closely to BGMP, we will examine the MASC protocol in some detail. Typically, MASC is run by one or more of a domain's border routers. Domains running MASC form a hierarchy parallel to the existing inter-domain structure. For example, a company's network has a regional network as its parent, and a

regional network has a backbone network as its parent. Backbone MASC domains that don't have a parent MASC domain are called *top-level* domains.

MASC dynamically distributes address ranges to domains using an approach called *listen and claim with collision detection*. In this approach, child domains listen to multicast address ranges selected by their parent domains, claim sub-ranges from the parents' ranges, wait a suitable amount of time to detect any collisions, and, if there are none, propagate these claims to their siblings. Top-level domains acquire their addresses by making claims from the global multicast address space, 224.0.0.0/4.

Address ranges are communicated to local MAASs using AAP and to other domains via MBGP. These address ranges, when injected into MBGP via a MASC speaker, are called *group routes*. The portion of the MBGP routing table that holds group routes is called the Group Routing Information Base (G-RIB).

As an example of how MASC works, consider the hierarchy of MASC domains shown below, in Figure 11:

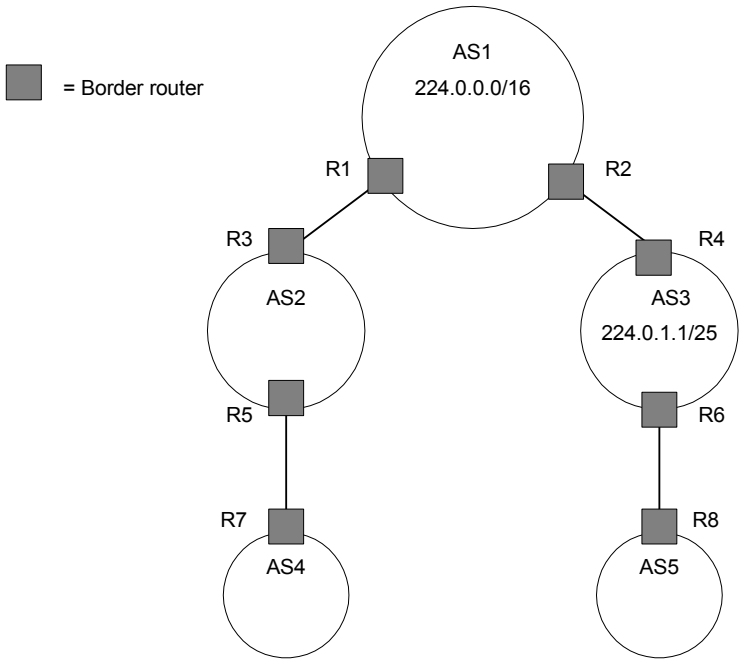

Figure 11. Example of MASC hierarchy

In this example, AS1 is a top-level domain. We can assume it is a backbone provider, while MASC domains AS2 and AS3, which have AS1 as their parent, are regional providers. AS2 and AS3 are siblings. The AS2 and AS3 domains have AS4 and AS5, respectively, as their children. The backbone provider, AS1, has already acquired, using MASC, the address range 224.0.0.0/16. Its child, AS3, has already acquired from it address range 224.0.1.1/25. We will show how AS2 acquires an address range.

MASC domain AS1 advertises its address range to its children. Child domain AS2 claims an address range (for example, 224.0.1.0/24) from its parent's address space by informing its parent of the claim. The parent, AS1, then propagates this claim information to its other children (in this case, AS3). If any of AS2's siblings are using the range, or a portion of the range, that AS2 has selected, they will send back collision announcements.

In this case, because AS3 is already using 224.0.1.1/25, it sends a collision announcement to AS2. When AS2 receives the collision announcement, it gives up its claim on that address space and makes a claim for a different range from its parent's space, for example, 224.0.128.0/24. It then listens for collision announcements, waiting for a long enough period of time to hear from all its siblings. Typical waiting periods can be one or two days. If no collision announcements occur, AS2 communicates the address range to its local MAASs and, via MBGP, to other domains as group routes.

The parent domain, AS1, keeps track of how much of its address space has been claimed. Once some threshold has been crossed, it claims more space, using the MASC protocol.

MASC and MBGP

As we said, once a MASC router successfully obtains an address range, it injects the range into its MBGP routing table as a group route. (It also sends the range to any other MASC routers in the domain and they, too, inject the range into their routing tables.) Using MBGP, the multicast address blocks are advertised throughout the Internet.

When one border router, X, advertises a group route to another border router, Y, it means that Y can use X as the next hop to forward multicast data packets (as well as control packets) towards the *root domain* for the address range represented by the group route. The root domain is the domain that initially injected the group route into the routing table.

For example, in Figure 10, the border router R3 advertises the group route corresponding to the address prefix 224.0.128.0/24 to border router R1 in AS1. Since all MBGP routers in a domain peer with each other to exchange routes received from external peers, the border router R2 learns of the group route received by R1. (If there are multiple paths to the root domain, one path is selected based on the attributes of the different group routes. This is standard MBGP behavior.)

The selected group route is stored by R1 in its G-RIB as (224.0.128.0/24, R3), which means that R3 is the next hop from R1 to reach the root domain for range 224.0.128.0/24. The other border router of AS1, which is R2, stores (224.0.128.0/24, R1) in its G-RIB because it uses R1 as the next hop to the root domain for that address range.

Address aggregation functions the same way as it would for unicast routes. Because the address range allocated to AS1, 224.0.0.0/16, encompasses AS2's address range, AS1's border routers need not propagate 224.0.128.0/24 to other domains. The border routers in other domains that need to reach the root domain for 224.0.128.0/24 can forward their packets using the group route corresponding to 224.0.0.0/16 that AS1 is already advertising. These packets will reach a border router in AS1 and that router can then use its more specific G-RIB entry for 224.0.128.0/24 to direct the packets to the root domain, AS2.

Objections to MASC

There are two main objections to MASC. They are as follows:

- It is too complex, making implementation difficult and costly.
- It cannot allocate the multicast address space fairly.

The first point is self-explanatory, but we will expand on the second point. Because multicast groups are much more dynamic than IP addresses, it is difficult to gauge how many addresses a domain will need. If routers underestimate, they must go back for extra blocks, which means the multicast address space becomes fragmented. Fragmentation violates the requirement that the address space be structured and logical. It also places an increasing burden on MBGP, requiring it to distribute more and more information as the number of blocks increases.

On the other hand, if routers overestimate, it is possible to run out of addresses. There are a total of 2^{28} multicast addresses, which at first glance seems like an enormous number. However, if every domain acquires a block of them before they are needed, as opposed to obtaining them on a one-by-one, as needed basis, it could be possible to exhaust the multicast address space.

In addition, there are a variety of proposals for using different parts of the multicast address space to encode information, such as the type of the group. One example is to allocate space to copyright-sensitive groups. This could be useful for keeping multicast data out of countries that don't acknowledge copyright law. Routers on the borders of copyright-violating domains would be configured to block multicast data from passing into those areas.

If copyright were the only policy to be encoded this way, it could multiply the number of blocks required by each domain by 2 because each domain might want to acquire a block from the copyright-sensitive range as well as a block from the copyright-insensitive range. Of course, it's unrealistic to say that this is the only policy that will ever be treated this way, and impossible to know how many policies will be accommodated by successively partitioning the address space. Using MASC would require each domain to acquire a block for each policy it might support. This again could mean running out of addresses.

Border Gateway Multicast Protocol

A domain's border routers run BGMP to construct an inter-domain shared tree for a multicast group. The same routers also run an intra-domain multicast routing protocol such as PIM-SM or Distance Vector Multicast Routing Protocol (DVMRP). These protocols are, as a class, called Multicast Interior Gateway Protocols (MIGP). Routers that run both an MIGP and BGMP are said to have an *MIGP component* and a *BGMP component*.

Shared trees always have a root RP, which, for BGMP, is some domain and not a single router. A key function of BGMP is to decide where this RP is located. As we said, the root domain is the domain where the initiator of the group is located. Therefore, to know where the root domain is, BGMP must know what domain owns the address range containing the group's multicast address. This is why BGMP relies on MASC, or an equivalent protocol. MASC associates multicast address ranges with specific domains. Once BGMP knows the root domain, it can construct a shared tree for distributing the multicast data.

BGMP peers use TCP connections to exchange information. Peers send each other messages to exchange control information, such as which members have joined or left a group. The types of BGMP messages are as follows:

- Open, which is used to initialize a session with a BGMP peer.
- Update, which transfers Join/Prune and forwarder preference (FWDR_PREF) information between peers.
- Notification, which is used when an error is detected.
- Keepalive, which is used to keep the TCP connection from expiring.

Bidirectional Trees

Like CBT, BGMP constructs bi-directional shared trees, which means members can communicate with each other without going through the RP. In contrast, PIM-SM constructs unidirectional trees, which means that, for a host other than the RP to send data on the tree, the data must first be tunneled to the RP and then travel down the tree to the group members. Using bi-directional trees minimizes third-party dependencies and improves performance. This is shown in Figure 12, below:

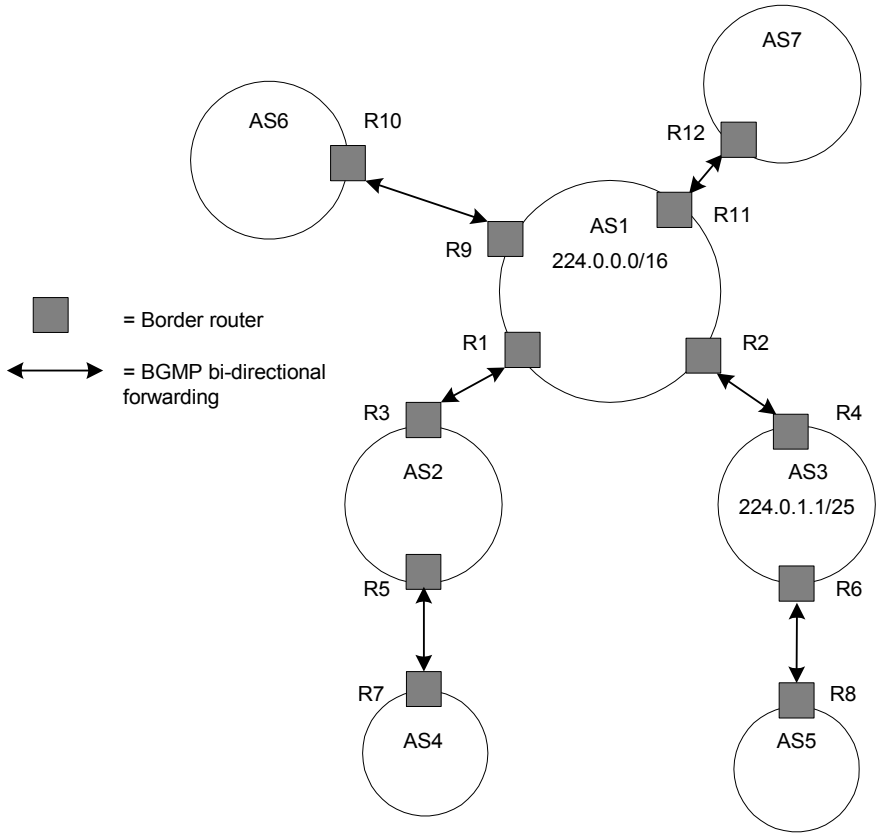

Figure 12. Example of bi-directional tree

Assume that AS2 is the root domain. In this example, group members located in AS3 and in AS6 can communicate with each other without having to rely on AS2. This eliminates a dependency on a third domain, and, because the path is shorter, improves performance. If the tree were unidirectional, data from AS3 , for example, would first have to go to AS2 (the root domain) and then, from AS2 to AS6.

Bi-directional Tree Construction

BGMP constructs shared trees using the group routes included in the G-RIB. As an example of how these trees are built, consider Figure 13, below:

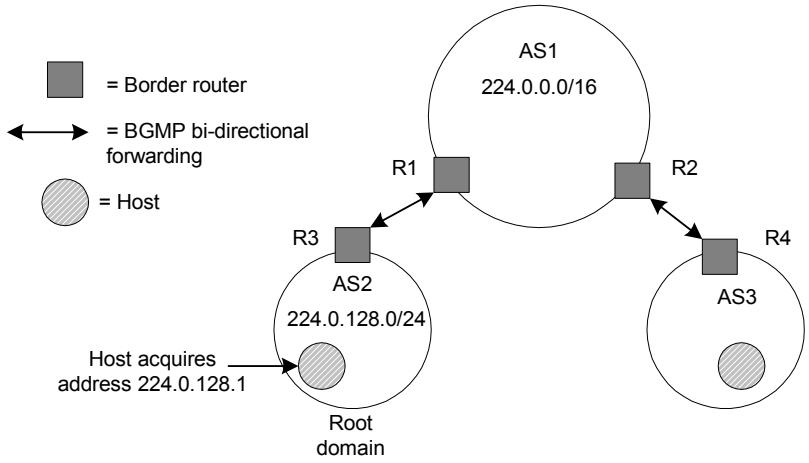

Figure 13. Example of building a shared tree

In this example, a host in AS2 acquires the multicast address 224.0.128.1 from the domain's assigned address range. This means AS2 is the root domain. When a host in AS3 joins this group, it sends an IGMP join message to its leaf router. The router generates a message compatible with the type of MIGP that it is running. For example, if it is running PIM-SM, it generates a Join message. If it is running DVMRP, and it has been pruned from the intra-domain source tree, it generates a Graft message. For simplicity, we will refer to join requests.

The BGMP component of AS3's *best exit router* (when there are multiple routers to choose from, this is the router selected by an MBGP algorithm for handling intra-domain traffic) receives this message. In this case, the router is R4, which is the only choice. It understands the message because it also has an MIGP component. The router must now find the root domain.

To do this, R4 looks up 224.0.128.1 in its G-RIB and finds (224.0.0.0/16, R2). It then creates a multicast forwarding entry that has a *parent target* and a list of *child targets*. The parent target is the BGMP peer that is the next hop toward the root domain. A child target is either the BGMP peer or the MIGP component from which the join request was received. Together, the parent target and the child targets form the *target list*. In the case of R4, the parent target is R2 and the only child target is its MIGP component. In this case, the target list could be written as (R2, MIGP in R4).

This multicast forwarding entry is also called a (*, G) entry. It means that if R4 receives any data packets that are intended for group G from any source, the packets will be forwarded to all the targets in the target list except the target from which the data packet came. Once a (*. G) entry is created, R4 sends a BGMP join message to the parent target, R2.

When R2 receives the join message from R4, it looks up 224.0.128.1 in its G-RIB and finds the entry (224.0.128.0/24, R1). This means that router R1 is the next hop to reach the root domain for address 224.0.128.1. Because R1 is an internal BGMP peer, R2 will transmit its join request to R1 using its MIGP

component. This means it creates a (*, G) entry with the MIGP component as the parent target and with R4 as the child target. The MIGP component of R2 performs whatever actions are necessary for a data packet to reach R1 from R2.

Once R1 receives the join request, it creates a (*, G) entry with its MIGP component as the child target and with R3 as the parent target. Since R3 is in the root domain for the group, it creates a (*, G) entry with its MIGP component as the parent target and with R1 as the child target. A join request is sent to the MIGP component, which joins group 224.0.128.1 using whatever procedure the MIGP component dictates.

Pruning a Bi-directional Tree

When a BGMP router or an MIGP component no longer leads to any group members, it removes itself from the child target list by sending a prune message to its parent target. Once the child target list is empty, the BGMP router removes the (*, G) entry and sends a prune message towards the root domain.

Source-specific Branches

BGMP can build source-specific branches of the shared-tree in cases where the shortest path to a source from a domain does not coincide with the bi-directional tree from a domain. Source-specific branches differ from source-rooted trees (or shortest-path trees) in that a source-specific branch stops when it reaches either a BGMP router on the bi-directional tree or the source domain.

Source-specific branches are useful for domains running MIGPs, such as DVMRP or PIM-DM, which build source-rooted trees within the domain. In these cases, if the border router that receives packets on the shared tree is not also on the shortest path to the source, it must encapsulate them in an MIGP wrapper and forward them to the border router that can inject them into the MIGP's domain. If the packets are not encapsulated and forwarded, they will fail the RPF check towards the source and be dropped.

Instead, if a a source-specific branch is built, packets can be brought directly into the domain via the router on the shortest path to the source. To accomplish this, once data is flowing, the router sends a source-specific join towards the relevant source. The join then propagates through the network, either until it reaches a branch of the bi-directional tree or the source domain.

For example, see Figure 14, below:

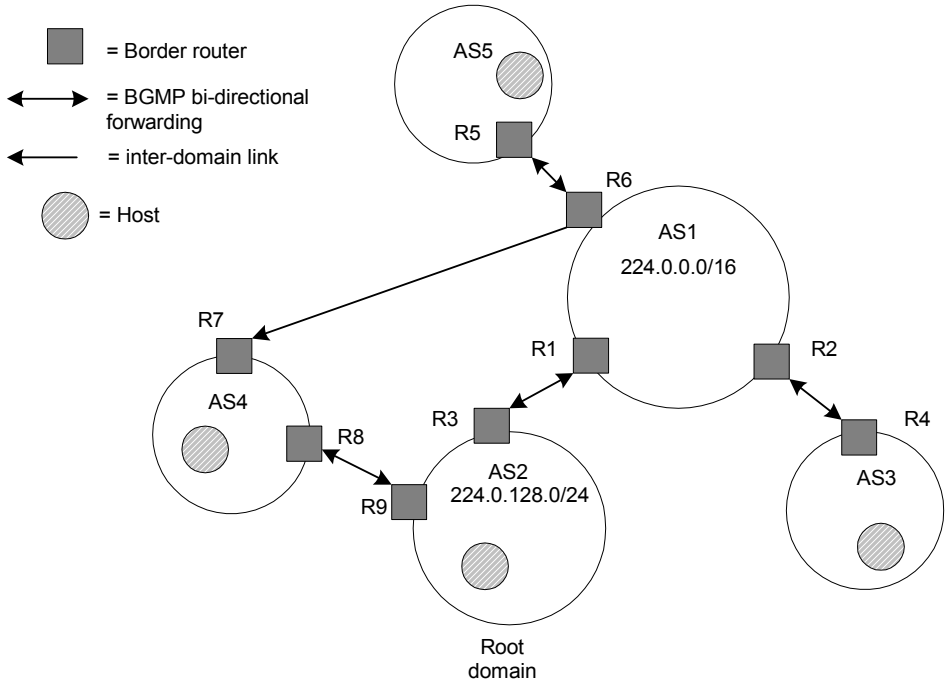

Figure 14. Example of source-specific branch

In this example, first assume there are no source-specific branches. Members of group 224.0.128.1 are located in domains AS2, AS3, AS4, and AS5. The source for the group is located in AS5. The root domain is AS2. There is a bi-directional tree set up, but AS4 also has an inter-domain link to AS1 via router R7. This means that the shortest path from AS4 to AS5 is through R7.

We will assume that AS4 uses DVMRP as its MIGP, which means that internal routers will only accept packets from a source if they receive that packet from their neighbor towards the source. Since only R8 is on the bi-directional shared tree, data from a source in AS5 will be received by R8, which must then encapsulate and forward the packets to R7 in order to avoid internal RPF check failures. The packets are then injected by R7 into the DVMRP domain so that members in A4 receive the multicast traffic.

To avoid this extra step, R7 can send a source-specific join toward the source in AS5. It must also create a multicast forwarding entry called an (S, G) entry. The parent target of this entry is the next hop toward the source (S), which is R6. The child target is R7's MIGP component. Data packets that arrive from R6 can, therefore, be accepted and forwarded to other targets listed in the (S, G) entry. The source-specific join propagates towards the source, similarly to how a shared-tree join propagates towards the root domain.

As the source-specific join is forwarded, it sets up (S, G) state in the intermediate border routers until it reaches a border router that is on the shared-tree for the group. In this example, R6 is on the shared tree. On

receiving the source-specific join, R6 creates an (S, G) entry and adds R7 to the child target list. The join is not propagated any further.

Subsequent data packets sent by S and received by R6 are forwarded to all the other targets in the (S, G) entry, including R7. Once it begins receiving data from R6, R7 sends a source-specific prune to R8, and starts dropping the encapsulated copies of S's data that is now flowing through R7. Since R8 has no other child targets for (S, G), it propagates the prune up the shared tree to R9 to stop receiving packets from S along the shared tree.

Alternatives to MASC/BGMP

The MALLOC/BGMP strategy offers a comprehensive solution to inter-domain multicast addressing. It maintains the original model for Internet multicasting, which is as follows:

- Receivers announce their interest.
- Senders simply send the data.
- Routers deliver the data from senders to receivers.

However, accomplishing this requires more complexity than was originally anticipated. In brief, the required pieces to the puzzle are the following:

- MASC for inter-domain address allocation.
- AAP for intra-domain address allocation.
- MADCAP for hosts to dynamically request an address.
- MBGP to distribute the G-RIB information.
- BGMP to use the G-RIB to build inter-domain multicast trees.
- MIGP (such as DVMRP, PIM, or CBT) for inter-domain multicast routing.

These protocols are still being investigated and tested. In response to the complexity of this approach, and to address some additional issues such as security and billing, some alternatives have been proposed. One class of suggestions is called Root Addressed Multicast Architecture (RAMA). We will mention two proposals that use this approach.

Express Multicast

Express is designed for situations where there is only a single source. One of its features is that it offers mechanisms for collecting information about subscribers. This is very important to service providers. It is difficult to attract advertising revenue without some knowledge of the customer base.

In essence, Express is the same as point-to-multipoint Asynchronous Transfer Mode (ATM). Express defines a multicast group as the pair (S, G). Multicast addresses only need to be unique with respect to the source, S. Only S is allowed to transmit and trees are unidirectional. (This may be a problem because some reliable multicast protocols require bi-directional trees so that receivers can multicast ACKs or NAKs, or find nearby repair nodes.) If there is a group with more than one sender, there are two choices. One is to create multiple groups, one per source. The other is to have senders other than S tunnel their packets to S and have S distribute the packets.

Simple Multicast

Simple Multicast is similar to Express. The main difference is that, while Simple Multicast also uses a single shared tree, it is bi-directional rather than unidirectional. The basic premise, once again, is that there is a primary source, which becomes the root of the shared tree. Along with this premise, Simple Multicast uses eight bytes of addressing: a four-byte core (or RP) address and a four-byte Class D group address.

This scheme means that routers don't have to figure out the RP address from the multicast address. Instead, the RP address is part of the identity of the group. The group is identified by the pair (C, G) rather than simply by the four-byte, Class D address, G. The address, G, does not need to be globally unique. It only needs to be unique with respect to C, because only the eight-byte quantity (C, G) needs to be unique. The proponents of Simple Multicast believe that this eliminates both the need for an address allocation scheme and for a bootstrap protocol to locate shared-tree RPs. The bootstrap protocol is unnecessary because the core of the tree is carried explicitly in each packet.

References

This section is a list of relevant RFCs and Internet drafts. It is listed alphabetically.

AAP

Handley, Mark, and Stephen R. Handley. *Multicast Address Allocation Protocol*.

http://search.ietf.org/internet-drafts/draft-ietf-malloc-aap-02.txt

BGP4

Rekhter, Yakov, and Tony Li. *A Border Gateway Protocol 4.*

http://www.ietf.org/internet-drafts/draft-ietf-idr-bgp4-09.txt

BGMP

Bates, Tony, Ravi Chandra, Dave Katz, and Yakov Rekhter. *Multiprotocol Extensions for BGP4*.

http://search.ietf.org/internet-drafts/draft-ietf-bgmp-spec-00.txt

CIDR

Fuller, Vince, Tony Li, Jessica Yu, and Kannan Varadhan. *Classless Inter-Domain Routing (CIDR): an Address Assignment and Aggregation Strategy*.

ftp://ftp.isi.edu/in-notes/rfc1519.txt

Express

Holbrook, Hugh, and Dave Cheriton. This is currently unpublished. There is an Express Web page at:

http://gregorio.stanford.edu/holbrook/express/

MADCAP

ftp://ftp.isi.edu/in-notes/rfc2730.txt

MALLOC

Thaler, David, Mark Handley, and Deborah Estrin. *The Internet Multicast Address Allocation*.

http://search.ietf.org/internet-drafts/draft-ietf-malloc-arch-04.txt

MASC

Estrin, Deborah, Ramesh Govindan, Mark Handley, Satish Kumar, Pavlin Radoslavov, and Dave Thaler. *The Multicast Address-Set Claim Protocol*.

http://search.ietf.org/internet-drafts/draft-ietf-malloc-masc-05.txt

MSDP

ftp://ftp.ietf.org/internet-drafts/draft-ietf-msdp-spec-05.txt

Simple Multicast

Perlman, Radia, Cheng-Yin Lee, Tony Ballardie, Jon Crowcroft, Zheng Wang, Thomas Maufer, Christophe Diot, and Joseph Thoo. *Simple Multicast: A Design for Simple, Low-Overhead Multicast*.

http://search.ietf.org/internet-drafts/draft-perlman-simple-multicast-03.txt

Summary

Although intra-domain multicasting is fairly well established, inter-domain multicasting presents a different set of issues. Inter-domain multicast routing requires the following:

- Reliable methods for receivers, which are distributed across the Internet, to access sources.
- A way to implement policies between domains to control the distribution of multicast data.
- No third-party dependencies.

Intra-domain routing protocols don't address these problems, so a new set of protocols is being developed. Several solutions are being proposed. Some of them are geared toward the short-term, while others are intended for the long-term.

In the short-term, the most widely accepted model uses these protocols:

- PIM-SM as the intra-domain routing protocol.
- MBGP to distinguish between unicast and multicast topologies.
- MSDP to connect multiple PIM-SM domains together.

Because of concerns about scaling, other solutions are being proposed for the long-term. The most accepted depends on linking multicast address allocation with multicast routing. An addressing architecture scheme called MALLOC links addresses with domains and distributes them, and is made up of three protocols:

- MASC for inter-domain address allocation.
- AAP for intra-domain address allocation.
- MADCAP for hosts to dynamically request an address.

In addition to address protocols, two routing protocols are also used. One is MBGP, which is used to distribute routing information across domains. The other is BGMP, which is used to build inter-domain shared trees.

Alternative methods are also being proposed. The two most popular assume a single source that is the root of a shared tree. These approaches are called Express and Simple Multicast.

For More Information

For the latest information on Windows 2000 Server, check out our Web site at http://www.microsoft.com/windows/server and the Windows NT Server Forum on MSN™ at http://computingcentral.msn.com/topics/windowsnt.