CONSULTANCY RAPPORT FOR HOGESCHOOL OF AMSTERDAM

Multicast Configuration

ICT network

Thijs Hodiamont [*]
dabutch@os3.nl

Since the official approval of IPv4 multicast by the IETF in 1989 in the document RFC1112 "Host Extensions for IP Multicasting" it still remains a mystery technology to many network operators. While created in a time when bandwidth was sparse it is still very actual in the current world where bandwidth is virtually unlimited. Multicasts lowers the burden on the machines and networks and creates a transparent distribution of streams in video, audio or data formats. It is for this purpose that the Hogeschool of Amsterdam[1] wants a multicast configuration for the network in and between their locations. This article describes the theory of multicast and its routing protocols. The theory is applied to the network of the HvA and the choices that should be made there to create a multicast enabled environment.

Contents

1 Introduction

The growth of bandwidth at the workstation has increased the usage of video, audio and data streams across the internet. People listen to webradio, look at videostreams and download data. With the current unicast model this is not scalable for the future, multicast is the solution for all load and bandwidth problems that for example the webradio stations have.

If we where to keep using the unicast model to transmit data to our listeners this would implicate that for each listener a session has to be made. The processing load and amount of bandwidth this would generate might work for a couple of users but if one has hundreds or thousands of users this would very inefficiënt. The load and bandwidth grows linear with the amount of users listening.

Multicast introduces a way of reducing this load and bandwidth consumption that is ideally for all streaming providers and network operators. Less traffic means less costs as billing is done by the byte nowadays.

The fundaments of multicast will be discussed in section 2 followed by the various inter- and intradomain routing protocols in section 3. Finally all the theories are applied

[*] University of Amsterdam, studentnr 0346462

[1] Abbreviated with HvA in this document

to a case that results in a consultancy conclusion for the Hogeschool of Amsterdam in section 4.

2 Multicast fundamentals

Before we can understand how multicast operates it is mandatory to elaborate the most common methods of data delivery in the current Internet. This is followed by the real multicast basics and the differences this protocol has in routing with unicast.

2.1 Networking basics

The three main methods of data delivery are unicast, broadcast and multicast. The current internet and its routing is primarily based on delivering unicast data. However as we saw in section 1 this method is inefficient for certain forms of data. To illustrate these three methods of data delivery the example of a web radio station will be used.

2.1.1 Unicast

When using unicast it is mandatory for the server to create a one-to-one session for each listener. As seen in figure 1 this results in three sessions for the three listeners. This means three times the bandwidth for the first router and two times the bandwidth for the second router. It is clear that this does not scale with $listeners^n$ but despite this fact it is the most common use of webradio or streaming in general nowadays.

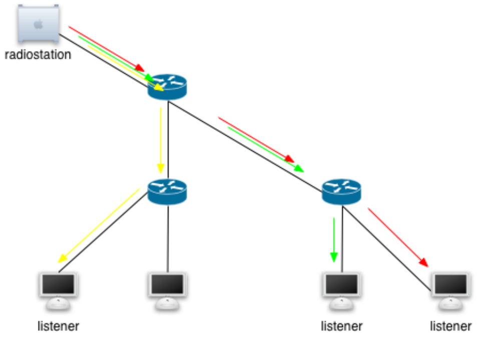

Figure 1: *Unicast data delivery*

2.1.2 Broadcast

Broadcast is the direct opposite of unicast. When using broadcast as data delivery only one stream is required and all hosts on the same network then receive this broadcast. This is a waste of network resources as just one or two clients are listening and as illustrated in figure 2 even if no hosts are listening the data is being received.

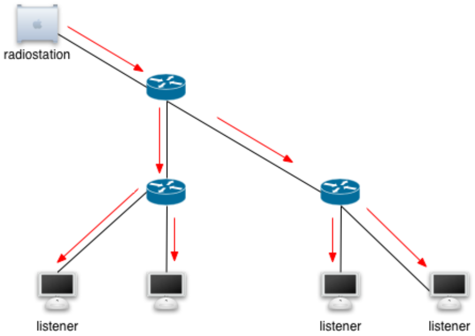

Figure 2: *Broadcast data delivery*

2.1.3 Multicast

Multicast has the best of both worlds. It provides one-to-many streams without using excessive bandwidth or resources. One stream is initiated and send to the first router in its path to the outside world. Using a separate protocol the router then knows which clients have subscribed for this service and distributes the service further into the network while multiplying the stream. Seen in figure 3 this results in just one packet of the stream on one network at all times.

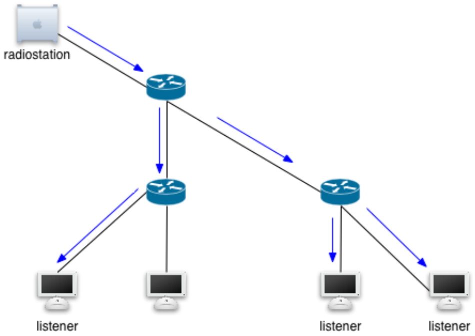

Figure 3: *Multicast data delivery*

2.2 Multicast basics

Multicast is often referred to as a one-to-many way of data delivery. A source is sending data to a multicast group address and a would-be receiver can subscribe himself to the same multicast address and start receiving the data without the source ever knowing it is sending data to this particular host.

2.2.1 Addressing

In normal IP unicast addressing a 48-bit MAC address is mapped to a 32-bit IP address. In multicast an IP address is always a **group** address and therefor it maps to its own MAC address to which all group members listen. The IP group address size is effectively 28-bit big as the first 4-bits in an multicast address are always set to `1110`. Multicast MAC addresses start with the 24-bit prefix `0x00:00:5e` . Given that the first byte of any ethernet address must be `01` to specify a multicast mac address, this implicates that the ethernet addresses corresponding to IP multicasting are in the range `0x01:00:5e:00:00:00` through `0x01:00:5e:7f:ff:ff`.This results in a 28-bit IP address that is mapped on a 23-bit MAC address and 5 bits of the IP address are lost. The lost 5 bits are the most significant bits of the 28 bits and this results in a 2^5:1 mapping of IP:MAC addresses. See figure 4 for an example.

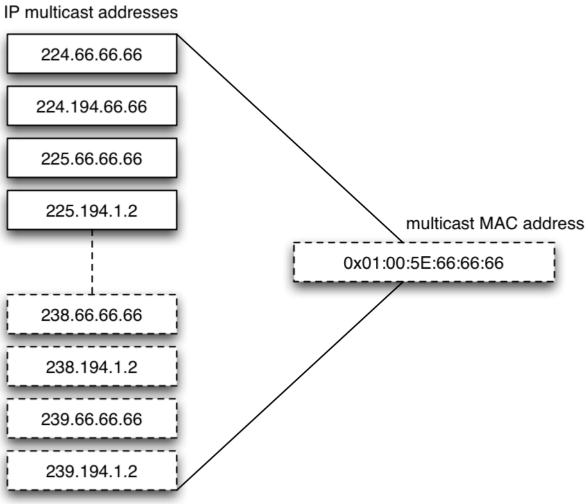

Figure 4: *IP:MAC address overlap*

This situation is inconvenient. A host subscribed to `224.194.66.66` will receive and process the headers of all packets with the given MAC destination. When it decapsulates the packet and sees the IP destination the packet is discarded, but already has used valuable system resources. This happens with traffic for all other groups that are displayed in figure 4. As a result, overlapping of these addresses is highly discouraged when building a network. Prevention can be done by tuning the multicast group addresses at application level.

intermezzo Why there is only a 23-bit MAC range available for multicast? The story goes that Steve Deering, the designer of multicast and graduate at that time, asked his advisor to buy a complete MAC address of 16 OUI's[2] for his research to map all 28 bits of IP to a unique MAC address. Unfortunately at that time, the IEEE charged $1000,- for each OUI and his advisor only agreed to buy one OUI with the remark that Steve could get half of that range. The other half would be used for other research projects.

Back to the theory. IP group addressing is done by using the special group of addresses located in the range `224.0.0.0/4` which goes from `224.0.0.0` to `239.255.255.255`. This range is controlled by the IANA[3] but certain ranges are reserved for specific usage.

- **224.0.0.0/24** link local multicast range
- **224.2.0.0/16** SAP/SDP[4] range
- **232.0.0.0/8** source specific multicast
- **233.0.0.0/8** AS-encoded[5], statically assigned GLOP range (RFC 3180)
- **239.0.0.0/8** administratively scoped multicast range (RFC 2365)

The most commonly used range is the link local range. Examples of predefined addresses are.

- **224.0.0.1/32** All systems on this subnet
- **224.0.0.2/32** All routers on this subnet
- **224.0.0.5/32** OSPF routers
- **224.0.0.6/32** OSPF designated routers

Unlike with unicast it is not common to request a dedicated multicast range from IANA. Reason for this is the GLOP range. Each AS has a `/24` multicast address space in the `233.0.0.0/8` range. The way of calculating the range for an individual AS is as described in figure 5.

```
  0      7 0       7 0       7 0       7
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 |  233   |    16 bits AS    | local |
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 5: *GLOP address layout*

Consider this theory for AS1103[6]. Written in binary, left padded with `0`'s, we get `00000100 01001111`. Mapping the high order octet to the second octet of the address, and the low order octet to the third octet, we get `233.4.79.0/24`. This range is completely from SURFnet and for multicast use. This normally rules out the need for extra multicast space.
As with unicast IP there is also an administratively scoped range that should not be routed across domains. This range `223.0.0.0/8` can be compared with the ranges `10.0.0.0/8`, `172.16.0.0/16` and `192.168.0.0/16` used in IP unicast.

2.2.2 IGMP

The Internet Group Management Protocol is used by hosts to dynamically register themselves in a particular multicast group. Currently there exist three versions of IGMP.

[2] Organisational Unique Identifier - the high 24 bits of a MAC address that is assigned to "an organization" by the IEEE.

[3] Internet Assigned Numbers Authority

[4] Session Announcement Protocol / Session Description Protocol

[5] Autonomous System

[6] SURFnet

- **v1** described in RFC 1112
- **v2** described in RFC 2236
- **v3** described in RFC 3376

Normally a host performs two actions when joining a specific multicast group.

- the host starts listening on the layer 2 address that maps to the IP multicast group address
- the host informs the router of its interest in a particular group by sending a Host Membership Report message. This triggers the router to set up a path to receive the multicast data.

Routers periodically send out Host Membership queries to discover which hosts are listening to which groups on which networks that ensure that no unnecessary data is delivered.

The primary differences between version 1 and 2 is the way hosts are handled that leave a group. In version 1 a host just stops responding to the Host Membership queries. After a fixed amount of time the router assumes that there are no listeners on the LAN and the traffic forwarding is stopped. In version 2 an explicit leave was implemented in the form of a Leave Group message. When this message is send the router responds by sending a group specific query message to determine whether other hosts on that LAN are still interested. The latest version of IGMP, IGMPv3 adds support for exclude and include modes. Exclude mode enables a host to request multicast packets for a group from all sources except those specified. Include mode enables a host to request multicast packets for a group from only the sources that are listed.

Best common practice at time of writing is IGMPv2.

2.2.3 Reverse Path Forwarding

Unicast routing can be pretty straightforward. Forwarding is done by examining the IP destination of the packet, looking up the correct route and forwarding it on the correct interface. Unicast packets are routed from source to destination.

Multicast routing is slightly different. When routing multicast traffic a forwarding state is set up from receiver to the root of the distribution tree. Routers execute a reverse path forwarding check to determine which interface is closest to the root of the distribution tree. The RPF is then the incoming interface for a group. An example can be found in figure 6.

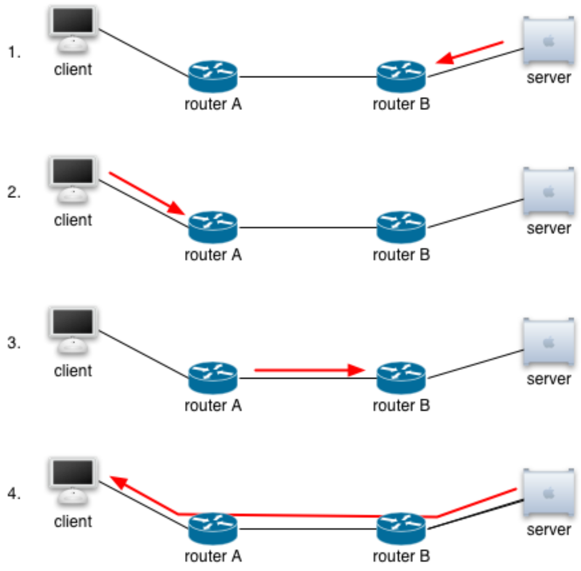

Figure 6: *RPF example*

Figure 6 shows an example how the reverse path forward check is performed.

1. Server sends data packets to multicast group address X. Router B creates a (Source,Group) state, adding its interface as the incoming interface for group X. Router B does not know of any interested parties and discards the packets.
2. Client announces to router A its interest in receiving of multicast packets for multicast group X via IGMP. This triggers router A to add its interface to the outgoing interface list for the X group.
3. Router A does a RPF lookup for servers address revealing that router B is the RPF neighbour for servers address. Router A proceeds to forward a (*,X) Join to its RPF neighbour for the RP address.
4. On receiving of the join message Router B forwards the data received from Server down the reverse path tree.

3 Multicast routing

In multicast routing there are three main areas to cover. The layer 2, interdomain routing and intradomain routing. Layer 2 does not really relate to routing but it is necessary to reduce the load on a switched network and therefor layer 2 is covered in this section. In intradomain routing there are protocols different available. These protocols are briefly looked at and this is completed with the interdomain routing protocols. Note that this article does not go deeply into all the specifications and details of each routing protocol. For more details on all protocols a must read is [1].

3.1 Layer 2

Normal behaviour of a layer 2 device in combination with multicast traffic is to forward it to all ports on the destination VLAN. Result of this is that all ports in this VLAN are cluttered with all available multicast traffic and unicast traffic might travel with great latency. To resolve this issue there are two protocols available that enable an opt-in method for multicast traffic on each port on a VLAN.

- Cisco Group Management Protocol (CCGP)
- IGMP snooping

The Cisco Group Management Protocol is a Cisco Systems proprietary protocol which has been used in Cisco environments for years. Now that computing power and bandwidth is more and more available CGMP has been mostly replaced by IGMP snooping which is the de facto standard on the Internet these days.

3.1.1 CGMP

CGMP is a Cisco Systems developed protocol that allows Cisco switches to learn about the existence of multicast clients in their LAN. CCMP is based on a client/server model where the router acts as the server and the switches as clients.
The procedure is rather simple. The designated router for a LAN receives an IGMP group join packet. The mac address for this host is registered in a table and a notification is send out on a known mac address. All switches listen on this mac address for changes in the multicast listener table. Each switch then processes this packet to see if it has to create entries in a forwarding table.

3.1.2 IGMP snooping

IGMP snooping is a layer 2 switch feature which is implemented by multiple switch vendors. There is no official standard for this technique and the probable reason for this is the layer violation that occurs when using this technique.
What occurs when using IGMP snooping is that a switch listens in into IGMP conversations between a router and a host. When the switch detects an IGMP group join on one of its ports it adds this port to the group of ports that are to receive multicast traffic for this group. On an IGMP group leave the port is removed from the group of ports that is to receive this multicast traffic. This procedure results in an optimisation in the network, multicast is only delivered to the interested listeners and the rest
Using this technique does have its effect on the hardware. Most switch vendors tell you their switch can do IGMP snooping at linespeed however enabling it on core switches is not a good idea. Enable IGMP snooping at host switches only. Another issue is the layer violation that occurs. To inspect a packet a layer 2 device has to look layer 3 packets. Despite the fact that there is no altering of the packets this does converts the switch into an entity that is between layer 2 and 3, However for now this is the most sensible solution as there is no alternative.

3.2 Intradomain

For intradomain routing here are several routing protocols available. The two most significant categories are the sparse and dense mode protocols. Both are explained in following subsections.

3.3 Sparse mode

Sparse mode protocols require an explicit join to initiate data delivery and therefor they are the opt-in method of multicast data delivery. This ensures that no unnecessary data is send across the network at the cost of some extra overhead opposed to dense mode protocols. In sparse protocols the root of the distribution tree is at a core node called a rendezvous point. When a host wants to join a group, its directly connected router (designated router) joins the distribution tree towards the RP. So traffic is received by the RP along the shortest path tree and forwarded to interested receivers across the domain via the shared tree or rendezvous point tree.
PIM sparse mode is the only example of a sparse protocol implemented by any major router vendor. Because of its sparse operation as well as protocol independence it is the de facto standard on the Internet today.

3.3.1 PIM-SM

PIM stands for Protocol Independent Multicast. At time of writing there are two versions available of PIM whereas the second version the current best practice is. PIMv1 messages are sent as IGMP messages where the type of the PIM message is distinguished by the IGMP code field. PIMv2 messages are sent using the IP protocol number 103. All routers connecting to a subnet must use the same PIM version, all PIM messages received with a different type of version are dropped. However for backwards compatibility some implementations automatically revert to version 1 of they receive a version 1 packet.

Group to RP mapping For PIM sparse mode to work properly it is mandatory for each router in a domain to know which rendezvous point is active for which multicast group. The definition of a PIM sparse mode domain is just that, a collection of routers that are physically connected and agree on the use of the same RP for all or a subset of group addresses in the `224.0.0.0/4` range. The three available options in choosing a RP are:

- static group-to-rendezvous point mapping
- Cisco Systems auto-RP
- PIM bootstrap router (BSR)

The following paragraphs will briefly describe each of these methods.

Static group-RP By far the most easy option to implement a RP point in the network. One machine is statically chosen as RP and all other routers in the network are manually configured to use this RP as their default RP. Obvious drawback is the fact that all routers have to be reconfigured every time this RP point is changed. Another drawback is the absence of any kind of redundancy regarding the RPs. However both these drawbacks can be addressed using an anycast configuration.

Auto group-RP Auto-RP is originally a Cisco Systems proprietary mechanism for dynamic group-to-RP mapping but nowadays auto-rp is supported by Juniper Networks as well.
Auto-RP elects an RP for PIM sparse mode using PIM dense mode as a flooding technique. Reason for this is the absence of an initial RP. These elected RPs start to announce themselves in a domain. All other routers that domain have joined the dense group and dynamically learn the address of the RP. Because of this reliance on the dense flooding method all routers must be configured in sparse-dense mode.
Each router in a PIM domain using auto-RP assumes one of the following roles:

- candidate RP
- mapping agent
- discovery-only

Every 60 seconds, a candidate RP sends an RP-announcement message detailing the group ranges for which it intends to server as RP. This message is send to the multicast group `224.0.1.39` .
The routers configured as mapping agent join the `224.0.1.39` group and listen for RP-announcement messages. Each mapping agent uses the following criteria to determine which RPs to announce as the active RP for each group:

- upon multiple RP announcement of the same group prefix and mask, accept the announcement only from the RP with the highest IP address
- reject a group prefix if it is already covered by a less-specific prefix advertised by the same RP
- accept all other announcements

After selecting an RP for each group range, the mapping agent announces RP-mapping messages to multicast group address `224.0.1.40` . Discovery only routers join this group and learn of the RP for each group range.
There is only one big drawback for this method and this is the dense mode character of the protocol. If no active RP for a multicast group can be found the routers will spread the group in dense mode through the network and this is not desirable.

Bootstrap group-RP PIM bootstrap was added to PIM version 2 as standardized way to provide dynamic group-to-RP mapping. Functionally seen is PIM bootstrap very similar to auto-RP. PIM bootstrap operates with one or more routers in the domain that are enabled to serve as candidate BSR's.
Each candidate BSR sends out messages on all of its interfaces. When neighbouring routers receive the message, they process the packet and forward a copy the packet out on all interfaces except for the interface on which the Bootstrap message was received.
If a candidate BSR receives a Bootstrap message with a BSR priority larger than its own, that routers stops announcing itself as candidate BSR. Eventually only one router in the domain will send out Bootstrap messages and are adapted by PIM routers.

Anycast RP In PIM sparse mode only one RP can be active for any single multicast group. Anycast RP is a clever mechanism that circumvents this limitation. Anycast means that multiple hosts, or in this case routers, share the same IP address. This address is then advertised by a routing protocol such as OSPF, IS-IS or BGP. Packets destined for the anycast address are then delivered to the next closest host with the anycast address.
With anycast RP, multiple routers are configured with the same IP address, typically on their loopback interface. This shared address is used in the RP-to-group mappings, which allows multicast groups to have multiple active RPs in a PIM domain. PIM sparse mode messages are sent towards the shared address, and they will reach the RP with the best routing metric from the originator of the message.
Thus anycast RP essentially forms multiple PIM sparse mode subdomains within the domain, with each subdomain consisting of one or more of the RPs and all of the PIM sparse mode routers. Because the domain is broken into subdomains, it is necessary to run MSDP between the RPs to exchange information about active sources between subdomains.
Anycast RP is mutually exclusive with the group-to-RP mapping mechanism so it can be used in conjunction with static RP, auto-RP or BSR. While auto-RP and BSR have their own methods of delivering load balancing and redundancy the best common practice is anycast static RP because of the simplicity.

3.4 Dense mode

Dense mode protocols are the opt-out protocols of multicast data delivery. These protocols assume that listeners are dense populated in the network and therefor flood the network with multicast data packets. In environments where listeners are dense populated, dense mode data delivery is more efficiënt than sparse mode data delivery due to the reduced overhead.
Dense mode protocols follow a *flood-and-prune* mechanism in which they flood the network with data to inform the routers of multicast sources. On arriving of this traffic on a router the data is forwarded to all inter-

faces except for the RPF interface. If there is no interest in the particular multicast source a *prune* message is send upstream and the traffic is stopped from the upstream router. Periodic reflooding is used to refresh state.

The major benefit of this protocol is simplicity. The great drawback however is the periodic bandwidth consumption if there are no listeners. There are only a few purposes for which this method is perfect and an example of this is Symantec Ghost.

3.4.1 DVMRP

DVMRP stands for Distance-Vector Multicast Routing Protocol. DVMRP was the first protocol to be deployed in the MBone[7]. It has standard dense mode flood-and-prune behaviour and implements a separate routing protocol on which RPF checks are performed. Note that as with all other distance vector router like RIP, there are limitations that include slow convergence and limited metrics. Nowadays most DVMRP implementations have been replaced by PIM sparse mode implementations.

3.4.2 PIM-DM

PIM dense mode has similar flood-and-prune behaviour as mentioned in DVMRP. The primary difference between the two protocols is that the latter introduces the concept of protocol independence. PIM can use the routing table populated by any underlying unicast protocol to perform RPF checks. This way PIM-DM can use the routing table that is filled by RIP, IGRP, OSPF, IS-IS, BGP and so on.

3.5 Sparse-dense mode

Sparse-dense mode is a PIM mode that is implemented by both Cisco Systems and Juniper Networks. This mode enables interfaces to operate in both sparse and dense mode on a per-group basis. Groups specified as dense groups are not mapped onto a RP and groups specified as sparse groups are. This mode is mainly used in networks that practice the auto-rp mechanism for PIM sparse mode. Drawback of this method is that if no RP for the specific group can be found failover is flooding the data through the network in dense mode.

3.6 Interdomain

Interdomain routing is not really about routing. It is merely about informing other domains about the active sources that are available in your own domain and vice versa. Multicast Source Discovery Protocol is the de facto standard for performing this tasks nowadays.

3.6.1 MSDP

In PIM-SM, the RP is configured to serve a range of multicast groups. The RP is responsible for knowing all of the active sources of all multicast groups in this range. There can only be one active RP for a given group and this presents interesting challenges when addressing redundancy, load balancing and interdomain connectivity. MSDP was developed to address these challenges.

MSDP introduced the ability for RPs to connect to other RPs and exchange out information about the active sources in their respective PIM-SM domains. With this capability each domain can have one or more RPs, enabling support for redundancy, load balancing and interdomain connectivity.

MSDP has been documented and standardized in RFC 3618 "Multicast Source Discovery Protocol (MSDP)" and both Cisco and Juniper have made implementations according to this RFC.

MSDP operates in a similar way as BGP by forming peer relationships with other MSDP routers via a TCP connection. MSDP peers within a domain facilitate redundancy and load balancing and MSDP peers between different domains allows for interdomain source discovery to occur.

A RP that is to participate in interdomain multicast routing must speak MSDP. However an MSDP speaker does not necessarily has to be a RP. Non-RP routers can be configured to speak MSDP with RPs to provide route reflection or MSDP transit traffic. A non-RP MSDP speaker does not originate any source information but simply relays source information from and to other domains.

Operation Upon receiving a PIM Register message the router generates an MSDP Source-Active message for the source-group pair and forwards the message to its configured MSDP peers. The SA message contains the source address, the group address and the address of the RP.

Upon receiving an SA message, a router checks to see if the message was received from its MSDP RPF-peer for the originator of the message. If the SA is received from a peer other than the RPF peer, the SA is ignored and discarded. On the other hand if the message was indeed received from its RPF peer the SA is forwarded to all other MSDP peers. This flooding guarantees that the SA message will be delivered throughout all peers but will not be looped back to the originator of the message.

If the MSDP speaking router is also a RP, additional processing of the MSDP SA message may be required. The RP determines if its domain has any interested members and if so the RP sends a PIM (S,G) Join message towards its RPF neighbor for the source to join the SPT.

SA caching Caching SA messages reduces join latency since the RP that receives a PIM join can quickly determine all the sources for the requested group by looking in its ow SA cache without have to ask other MSDP peers. This results in a faster delivery of multicast data and less overhead. Because of these advantages nearly all MSDP implementations have SA caching features. On Juniper Networks routers SA caching cannot be disabled, on Cisco Systems routers SA caching is a

[7] Global initiative to provide multicast connectivity. Deceased project.

configurable option.

Mesh group An MSDP mesh group can be configured for a group of MSDP peers that are competely meshed. MSDP mesh groups are able to reduce SA flooding by identifying the source of an SA message. If a message is received from a mesh group peer the message is sent to all nonmesh group peers and not to any other peer in the mesh group. If a message is received from a nonmesh group peer it is forwarded to all other mesh group peers. This way there are no SA storms between the meshed MSDP peers. Drawback in this method is that if no RP can be found the multicast group will be forwarded in dense mode.

4 HvA network

Now we know how multicast operates it is time to put the theory to good use. The HvA already has a small multicast configuration to receive streams from their upstream provider the national computer network for higher education and research in the Netherlands, SURFnet. In this chapter the requirements for an own multicast configuration will be set and some choices are made regarding routing protocols. With these choices it is possible to create a configuration for their networking equipment and implement it in the production network. The topology of the network is discussed after which choices regarding protocols and operating of these protocols can be made.

4.1 Topology

The Hogeschool of Amsterdam has several locations, all of which are connected to `border.hva.nl` through KPN metroconnect as seen in figure 7. Only the Leeuwenburg location has a redundant connection, all other locations have just one layer 3 point in their network.

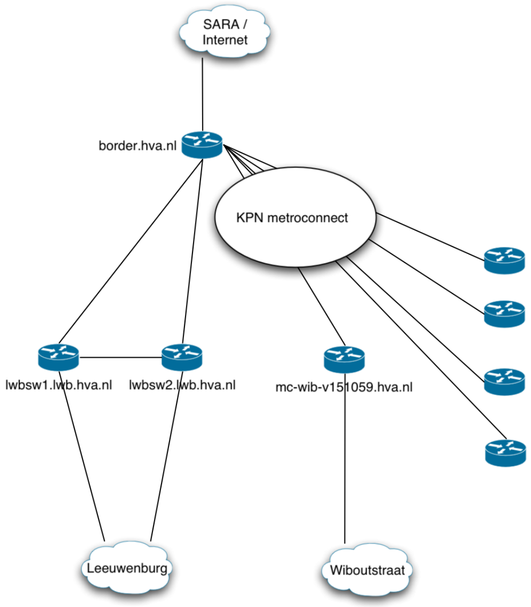

Figure 7: *current network topology HvA*

In the near future the HvA will take part in the GigaMAN[8] project in which they will have a ring topology as can be seen in figure 8. In this design all traffic will still go through `border.hva.nl` , but redundancy in the ring is created through the two Leeuwenburg routers.

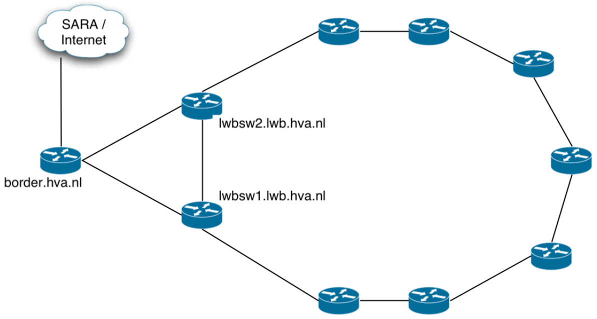

Figure 8: *GigaMAN network topology HvA*

All routers are Cisco based and have multicast enabled IOS versions. As SURFnet already talks PIM-SM and MSDP with their peers it is mandatory for the HvA that they do so also. If another protocol is chosen it is not possible to distribute their multicast sessions beyond the borderrouter of the HvA. This severely decreases the amount of options that are available.

4.2 Multicast

As said it is compulsory for the HvA to use PIM sparse mode and MSDP to ensure world wide distribution. If looked past this fact for just a moment it is possible to elaborate the choice for PIM sparse mode that would have been made anyway.
The choice between the categories of protocols is not difficult. Dense mode operates in opt-out mode and sparse mode in opt-in mode. While network resources are rather unlimited these days an opt-out method for multicast is not desirable for several reasons.

- Flooding is done periodically to keep state
- Flooding is done by all sources in an uncontrolled manner
- Flooding costs bandwidth for all hosts connected

An analogy with unsolicited e-mail is not hard to make. The sparse mode protocols might have some more overhead and abit more latency on the initial setup of the paths but it is much more efficient with bandwidth and has a more network friendly approach of data delivery. Managing a network with PIM sparse mode is easier as it is possible to choose a RP where all the data is concentrated. We can choose a convenient point in the network that can handle the amount of traffic and create redundancy in the network. Therefor a sparse mode network is always more preferable in big network environments.

[8] http://gigaman.gigaport.nl/

4.3 Intradomain

For intradomain routing PIM sparse mode will be used. This is the most suitable solution for the HvA as multicast traffic will be mostly streams. The occasional Symantec Ghost restore session will put some stress on the network but enabling sparse-dense mode is not worth the risks. The RPs will be placed on both of the Leeuwenburg routers, in what configuration is discussed in section 4.3.1.

4.3.1 Redundancy

The current star network has no options for redundancy. The border router is single point of failure for all links and therefor creating redundancy on the two Leeuwenburg routers has no function. However in the future ring network these two routers play an important role in the connection between the ring and SURFnet and are therefor the ideal points of placing rendezvous points on. In this way there is no unnecessary traffic on the links to SURFnet.
Another option which creates ultimate redundancy in case of multiple link failures is rendezvous points on all border routers. However in case of multiple link failures the functioning of multicast is the least of the worries of the network engineers. Having thought about redundancy there are two real options to implement.

1. two rendezvous points at both Leeuwenburg routers
2. rendezvous points on all routers of ring

The first option creates redundancy in the ring. Upon failure of on of the Leeuwenburg routers there is still a another rendezvous point available in the network. With option 2 the number of rendezvous points is extended to the number of routers at all locations. As this latter option is only an extension of option 1 this will not be recommend for now.
Now that the choice for two or more rendezvous is made, a mechanism of choosing the current active rendezvous point for a multicast group is necessary. As described in section 3.3.1 there are three ways of choosing this rendezvous point:

- static group-to-rendezvous point mapping
- Cisco systems auto-rendezvous point
- PIM bootstrap router (BSR)

Because of the reasonable simplicity in the network the choice for group to RP mapping is a static anycast RP on the both Leeuwenburg routers. This is the most simple solution for redundancy in the network. The two other mechanisms require more administration and do not provide any extra redundancy or load balancing features. Therefor a more simple design is preferred.
To implement failover between the two anycast RPs a MSDP meshing session between the two routers has to be enabled. A third session with the border router is necessary to let it function as a route collector. The two Leeuwenburg routers set up a session with the border router, which then sets up a session with the SURFnet routers. In this way the RPs are transparant to SURFnet, otherwise SURFnet has to set up a session with both RPs which is inefficiënt. In a picture this looks like figure 9

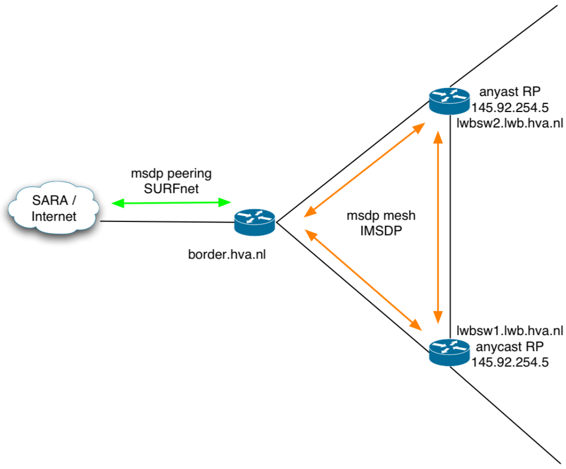

Figure 9: *HvA multicast configuration*

This method of implementing multicast is actually documented in an RFC. Namely RFC 3446 "Anycast Rendevous Point (RP) mechanism using Protocol Independent Multicast (PIM) and Multicast Source Discovery Protocol (MSDP)". In appendix B this rfc is added.

4.4 Interdomain

The interdomain part of the HvA network is solemnly to SURFnet. As they already use MSDP for their configurations it is mandatory that the HvA does so also. With each new configuration of an service to the outside world new ACLs have to be set but it is up to the network administrators to decide which. During the time of writing the configuration with SURFnet was not yet available so it is not possible to make any statements about that. No information about the ranges that the HvA could use for its streamings is available either.

4.5 Configuration

During the making of the decisions a multicast configuration based on the actual HvA network was implemented in a lab situation. This proof of concept configuration can be found in appendix A. A stage later in the project this configuration was actually used to make a proof of production in the production network to enable multicast routing within the HvA network.

5 Conclusion

This document has briefly elaborated the operating of protocols concerning multicast. Each having its own advantages and disadvantages the most simple solution was chosen. While simplicity is always a desired goal in network design, it is even more valuable when building and operating multicast networks because of the character of the protocol. With the given tips and the configuration in appendix A it should be very easy for the HvA to setup their production multicast network with a standardized configuration that should work on routing equipment of all vendors. Work that still remains is to determine which ACLs to place and which applications to use for multicast streaming. No real problems occured during the project and all went pretty smooth despite the fact that i was in a one man group.

List of Figures

References

[1] Wright Edwards, Giuliano. *Interdomain Multicast Routing*. Addison-Wesley, 2002. ISBN: 0-201-74612-3. 3

[2] W. Richard Stevens. *TCP/IP Illustrated, The Protocols*, volume 1. Addison-Wesley, 1994. ISBN: 0-201-63346-9.

[3] Frahim Froom, Sivasubramanian. *Building Cisco Multilayer Switched Networks (BCMSN)*. Cisco Press, 2004. ISBN: 1-58705-150-8.

A HvA Cisco configuration

The multicast configuration that is loaded on the three layer 3 devices in the HvA network.

A.1 border.hva.nl

The border router is only slaving the MSDP sessions from the two anycast RPs and therefor it only has a MSDP peering with both routers and with the SURFnet MSPD peer. PIM sparse mode is enabled on all relevant interfaces.

```
!
ip multicast-routing
!
!
interface Loopback0
 description Unique IP address
 ip address 145.92.254.165 255.255.255.255
 ip pim sparse-mode
 ip sap listen
!
interface ifname
 ip pim sparse-mode
!
ip pim rp-address 145.92.254.161 override
ip msdp peer 145.92.254.163 connect-source Loopback0
ip msdp peer 145.92.254.164 connect-source Loopback0
ip msdp peer $ip_surfnet_msdp connect-source Loopback0
ip msdp mesh-group IMSDP 145.92.254.163
ip msdp mesh-group IMSDP 145.92.254.164
ip msdp cache-sa-state
ip msdp originator-id Loopback0
!
```

A.2 lwbsw1.lwb.hva.nl

The Leeuwenburg 1 router is one of the anycast RPs en therefor it has an anycast IP address configured on the second loopback interface. The first loopback device has a unique IP address to which all MSDP sessions with both `border.hva.nl` and `lwbsw2.lwb.hva.nl` point.

```
!
ip multicast-routing
!
interface Loopback0
 description Unique IP address
 ip address 145.92.254.164 255.255.255.255
 ip pim sparse-mode
 ip sap listen
!
interface Loopback1
 description Anycast RP address
 ip address 145.92.254.161 255.255.255.255
 ip pim sparse-mode
!
interface $ifname
 ip pim sparse-mode
!
ip pim rp-address 145.92.254.161 override
ip msdp peer 145.92.254.2 connect-source Loopback0
ip msdp peer 145.92.254.3 connect-source Loopback0
ip msdp mesh-group IMSDP 145.92.254.163
ip msdp mesh-group IMSDP 145.92.254.165
ip msdp cache-sa-state
ip msdp originator-id Loopback0
```

```
!
```

A.3 lwbsw2.lwb.hva.nl

The Leeuwenburg 2 router is one of the anycast RPs en it has a similar configuration as the Leeuwenburg 1 router. An anycast IP address is configured on the second loopback interface. The first loopback device has a unique IP address to which all MSDP sessions with both `border.hva.nl` and `lwbsw1.lwb.hva.nl` point.

```
!
ip multicast-routing
!
interface Loopback0
 description Unique IP address
 ip address 145.92.254.163 255.255.255.255
 ip pim sparse-mode
 ip sap listen
!
interface Loopback1
 description Anycast RP address
 ip address 145.92.254.161 255.255.255.255
 ip pim sparse-mode
!
interface $ifname
 ip pim sparse-mode
!
ip pim rp-address 145.92.254.161 override
ip msdp peer 145.92.254.163 connect-source Loopback0
ip msdp peer 145.92.254.165 connect-source Loopback0
ip msdp mesh-group IMSDP 145.92.254.163
ip msdp mesh-group IMSDP 145.92.254.165
ip msdp cache-sa-state
ip msdp originator-id Loopback0
!
```

B RFC 3446

Network Working Group
Request for Comments: 3446
Category: Informational

D. Kim
Verio
D. Meyer
H. Kilmer
D. Farinacci
Procket Networks

January 2003

Anycast Rendevous Point (RP) mechanism using Protocol Independent Multicast (PIM) and Multicast Source Discovery Protocol (MSDP)

Status of this Memo

This memo provides information for the Internet community. It does not specify an Internet standard of any kind. Distribution of this memo is unlimited.

Copyright Notice

Abstract

This document describes a mechanism to allow for an arbitrary number of Rendevous Points (RPs) per group in a single shared-tree Protocol Independent Multicast-Sparse Mode (PIM-SM) domain.

1. Introduction

PIM-SM, as defined in RFC 2362, allows for only a single active RP per group, and as such the decision of optimal RP placement can become problematic for a multi-regional network deploying PIM-SM.

Anycast RP relaxes an important constraint in PIM-SM, namely, that there can be only one group to RP mapping can be active at any time. The single mapping property has several implications, including traffic concentration, lack of scalable register decapsulation (when using the shared tree), slow convergence when an active RP fails, possible sub-optimal forwarding of multicast packets, and distant RP dependencies. These properties of PIM-SM have been demonstrated in native continental or inter-continental scale multicast deployments. As a result, it is clear that ISP backbones require a mechanism that allows definition of multiple active RPs per group in a single PIM-SM domain. Further, any such mechanism should also address the issues addressed above.

The mechanism described here is intended to address the need for better fail-over (convergence time) and sharing of the register decapsulation load (again, when using the shared-tree) among RPs in a domain. It is primarily intended for applications within those networks using MBGP, Multicast Source Discovery Protocol [MSDP] and PIM-SM protocols, for native multicast deployment, although it is not limited to those protocols. In particular, Anycast RP is applicable in any PIM-SM network that also supports MSDP (MSDP is required so that the various RPs in the domain maintain a consistent view of the sources that are active). Note however, a domain deploying Anycast RP is not required to run MBGP. Finally, a general requirement of the Anycast RP scheme is that the anycast address MUST NOT be used as the RP address in the RP’s SA messages.

The keywords MUST, MUST NOT, MAY, OPTIONAL, REQUIRED, RECOMMENDED, SHALL, SHALL NOT, SHOULD, SHOULD NOT are to be interpreted as defined in BCP 14, RFC 2119 [RFC2119].

2. Problem Definition

The anycast RP solution provides a solution for both fast fail-over and shared-tree load balancing among any number of active RPs in a domain.

2.1. Traffic Concentration and Distributing Decapsulation Load Among RPs

While PIM-SM allows for multiple RPs to be defined for a given group, only one group to RP mapping can be active at a given time. A traditional deployment mechanism for balancing register decapsulation load between multiple RPs covering the multicast group space is to split up the 224.0.0.0/4 space between multiple defined RPs. This is an acceptable solution as long as multicast traffic remains low, but has problems as multicast traffic increases, especially because the network operator defining group space split between RPs does not always have a priori knowledge of traffic distribution between groups. This can be overcome via periodic reconfigurations, but operational considerations cause this type of solution to scale poorly.

2.2. Sub-optimal Forwarding of Multicast Packets

When a single RP serves a given multicast group, all joins to that group will be sent to that RP regardless of the topological distance between the RP and the sources and receivers. Initial data will be sent towards the RP also until configured the shortest path tree switch threshold is reached, or the data will always be sent towards the RP if the network is configured to always use the RP rooted shared tree. This holds true even if all the sources and the receivers are in any given single region, and RP is topologically distant from the sources and the receivers. This is an artifact of the dynamic nature of multicast group members, and of the fact that operators may not always have a priori knowledge of the topological placement of the group members.

Taken together, these effects can mean that (for example) although all the sources and receivers of a given group are in Europe, they are joining towards the RP in the USA and the data will be traversing a relatively expensive pipe(s) twice, once to get to RP, and back down the RP rooted tree again, creating inefficient use of expensive resources.

2.3. Distant RP Dependencies

As outlined above, a single active RP per group may cause local sources and receivers to become dependent on a topologically distant RP. In addition, when multiple RPs are configured, there can be considerable convergence delay involved in switching to the backup RP. This delay may exist independent of the toplogical location of the primary and backup RPs.

3. Solution

Given the problem set outlined above, a good solution would allow an operator to configure multiple RPs per group, and distribute those RPs in a topologically significant manner to the sources and receivers.

3.1. Mechanisms

All the RPs serving a given group or set of groups are configured with an identical anycast address, using a numbered interface on the RPs (frequently a logical interface such as a loopback is used). RPs then advertise group to RP mappings using this interface address. This will cause group members (senders) to join (register) towards the topologically closest RP. RPs MSDP peer with each other using an address unique to each RP. Since the anycast address is not a unique address (by definition), a router MUST NOT choose the anycast unicast address as the router ID, as this can prevent peerings and/or adjacencies from being established.

In summary then, the following steps are required:

3.1.1. Create the set of group-to-anycast-RP-address mappings

The first step is to create the set of group-to-anycast-RP-address mappings to be used in the domain. Each RP participating in an anycast RP set must be configured with a consistent set of group to RP address mappings. This mapping will be used by the non-RP routers in the domain.

3.1.2. Configure each RP for the group range with the anycast RP address

The next step is to configure each RP for the group range with the anycast RP address. If a dynamic mechanism, such as auto-RP or the PIMv2 bootstrap mechanism, is being used to advertise group to RP mappings, the anycast IP address should be used for the RP address.

3.1.3. Configure MSDP peerings between each of the anycast RPs in the set

Unlike the group to RP mapping advertisements, MSDP peerings must use an IP address that is unique to the endpoints; that is, the MSDP peering endpoints MUST use a unicast rather than anycast address. A general guideline is to follow the addressing of the BGP peerings, e.g., loopbacks for iBGP peering, physical interface addresses for eBGP peering. Note that the anycast address MUST NOT be used as the RP address in SA messages (as this would case the peer-RPF check to fail).

3.1.4. Configure the non-RP’s with the group-to-anycast-RP-address mappings

Finally, each non-RP router must learn the set of group to RP mappings. This could be done via static configuration, auto-RP, or by PIMv2 bootstrap mechanism.

3.1.5. Ensure that the anycast IP address is reachable by all routers in the domain

This is typically accomplished by causing each RP to inject the /32 into the domain’s IGP.

3.2. Interaction with MSDP Peer-RPF check

Each MSDP peer receives and forwards the message away from the RP address in a "peer-RPF flooding" fashion. The notion of peer-RPF flooding is with respect to forwarding SA messages [MSDP]. The BGP routing tables are examined to determine which peer is the next hop towards the originating RP of the SA message. Such a peer is called an "RPF peer". See [MSDP] for details of the Peer-RPF check.

3.3. State Implications

It should be noted that using MSDP in this way forces the creation of (S,G) state along the path from the receiver to the source. This state may not be present if a single RP was used and receivers were forced to stay on the shared tree.

4. Security considerations

Since the solution described here makes heavy use of anycast addressing, care must be taken to avoid spoofing. In particular unicast routing and PIM RPs must be protected.

4.1. Unicast Routing

Both internal and external unicast routing can be weakly protected with keyed MD5 [RFC1828], as implemented in an internal protocol such as OSPF [RFC2328] or in BGP [RFC2385]. More generally, IPSEC [RFC2401] could be used to provide protocol integrity for the unicast routing system.

4.1.1. Effects of Unicast Routing Instability

While not a security issue, it is worth noting that if unicast routing is unstable, then the actual RP that source or receiver is using will be subject to the same instability.

4.2. Multicast Protocol Integrity

The mechanisms described in [RFC2362] should be used to provide protocol message integrity protection and group-wise message origin authentication.

4.3. MSDP Peer Integrity

As is the the case for BGP, MSDP peers can be protected using keyed MD5 [RFC1828].

5. Acknowledgments

John Meylor, Bill Fenner, Dave Thaler and Tom Pusateri provided insightful comments on earlier versions for this idea.

This memo is a product of the MBONE Deployment Working Group (MBONED) in the Operations and Management Area of the Internet Engineering Task Force. Submit comments to <mboned@ns.uoregon.edu> or the authors.

6. References

[MSDP]   D. Meyer and B. Fenner, Editors, "Multicast Source Discovery Protocol (MSDP)", Work in Progress.

[RFC2401] Kent, S. and R. Atkinson, "Security Architecture for the Internet Protocol", RFC 2401, August 1995.

[RFC1828] Metzger, P. and W. Simpson, "IP Authentication using Keyed MD5", RFC 1828, August 1995.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC2362] Estrin, D., Farinacci, D., Helmy, A., Thaler, D., Deering, S., Handley, M., Jacobson, V., Liu, C., Sharma, P. and L. Wei, "Protocol Independent Multicast-Sparse Mode (PIM-SM): Protocol Specification", RFC 2362, June 1998.

[RFC2328] Moy, J., "OSPF Version 2", STD 54, RFC 2328, April 1998.

[RFC2385] Heffernan, A., "Protection of BGP Sessions via the TCP MD5 Signature Option", RFC 2385, August 1998.

[RFC2403] Madson, C. and R. Glenn, "The Use of HMAC-MD5-96 within ESP and AH", RFC 2403, November 1998.

7. Author’s Address

Dorian Kim
Verio, Inc.
EMail: dorian@blackrose.org

Hank Kilmer
EMail: hank@rem.com

Dino Farinacci
Procket Networks
EMail: dino@procket.com

David Meyer
EMail: dmm@maoz.com

8. Full Copyright Statement

Acknowledgement

Funding for the RFC Editor function is currently provided by the Internet Society.