Handout #30
C. Noronha
Winter 98-99

EE384A: Network Protocols and Standards
Homework #5 - Solutions
IP Multicast

1. Internet Group Management Protocol (IGMP)

IGMP is a protocol that allows end stations to indicate their IP multicast group membership to routers. This protocol is only used to communicate group membership to neighboring routers. The protocol does not specify how routers propagate this information throughout the network. That is accomplished through DVMRP, MOSPF, or some other multicast routing protocol.

Membership Query Messages

On each network there is one router which is the Querier. This is the router with the lowest IP address. It is the responsibility of the Querier to periodically transmit Membership Query Messages, which cause the end stations to respond with their group membership information.
There are three types of query messages:

1.) General Query
This message is transmitted to the multicast address 244.0.0.1. It is used to learn the complete multicast reception state of all multicast receivers on a network. The General Query message is defined to have the Group Address field of the message equal to 0.

2.) Group-Specific Query
This message is used to determine if there are any end stations on a network that are interested in receiving traffic for a specific multicast address. This is indicated by setting the Group Address field equal to the appropriate multicast address.
The packet is transmitted with a destination address equal to the multicast address that the router is querying.

3.) Group-and-Source Specific Query
This message is used by routers to determine if there are any end-stations on a network interested in receiving traffic transmitted by the specified source to a specific multicast address. In addition to setting the Group Address field equal to the multicast address of interest, the router also lists the appropriate source addresses.
Like the Group-Specific Query, these messages are transmitted to the multicast address the router is querying.

Membership Report Messages

When end systems or other multicast receivers receive a Membership Query message, they respond after a random delay with a Membership Report Message.
The Membership Report messages indicate the multicast groups the end system is interested in receiving. Once a Membership Report has been transmitted indicating membership in a certain multicast address, all other end systems that are still waiting to transmit a Membership Report for the same address will cancel their transmissions. Membership Report messages may also be transmitted by a multicast receiver any time their group membership changes (e.g., the station wants to receive a new multicast group).

Membership Report messages contain a set of Group Records, one for each multicast address the end system is interested in receiving. Each Group Record indicates:

1.) The multicast address the end system is interested in receiving

2.) A list of source addresses

3.) A flag that indicates if the end system wishes to receive multicast traffic only from the list of source addresses or if the end system wishes to receive traffic from all sources except the ones listed in the Group Record.

Leaving a Multicast Group

There are two methods an end system may use to leave a multicast group. It may transmit a Membership Report message indicating it wishes to receive traffic from a specific list of sources, but the list of sources is empty.
The end system may also transmit a specific Leave Group message. The first method is preferred. An end system transmits a leave indication only if it was the last station to transmit a Membership Report for that multicast address.

When a router receives a leave indication, the router will immediately transmit a Group-Specific Membership Query to determine if there are any other end systems on the network which are interested in receiving traffic for the multicast address. After a period of time if the router does not hear any responses, it will stop forwarding multicast traffic for the specified address on the network.

2. IP Multicast Routing

We can group together the five algorithms discussed into two categories: (I) finding a delivery tree based on the group and not based on the source, and (II) finding a delivery tree based on the location of the source.

CBT and PIM-SM fall in category I, while DVMRP, MOSPF, and PIM-DM can be grouped together as category II. The latter group builds trees based on the source and destinations (group members' locations) and are not based on rendezvous points.

Category I

CBT

In CBT and PIM-SM the concept of a core router or a rendezvous point is central to how routes are created. In CBT, a core router is designated for each group. CBT makes use of a shared tree rather than a per-sender tree.

CBT uses "hard states", where messages are acknowledged and repeated after a time-out. In contrast, PIM-SM uses "soft-states", in which join messages are repeated at regular intervals, and in which the states can disappear if the information is not refreshed.

a) Formation of routes

- A core tree is defined for every group, composed of primary core routers and secondary core routers
- Join/leave information is based on IGMP. When a host on a network wants to join a group, the designated router sends an IGMP_join to a statically assigned "target" core. The join opens a path along the way to the core (or stops at the first on-tree router). An acknowledgment is sent back. The designated router is the router with the smallest address on the network.
- If the target core is not the primary core, it sends a join_request rejoin active to the primary core. The primary core sends back a join_ack primary rejoin ack.
- The router that forwards off a network need not be the DR for the network
- The core tree is built on demand. It grows when secondary core routers, lingered by join request, join the primary router. The primary core should be well known across the group.
- The tree is truncated in 2 cases :
  1) In case a router finds that the best route to a certain core is through one of its children, it sends a FLUSH.TREE message to all its children. These will stop using the parents and switch to the better path if they have local receivers.
  2) If a check for a group member presence finds no one in the router's subnet, the router sends a QUIT_REQUEST upstream to remove itself from the tree.
- The tree is maintained by a mechanism for routers to check the status of each others. It consists of a CBT_ECHO_REQUEST and CBT_ECHO_REPLY messages sent periodically.

b) Information exchanged between routers

- IGMP messages are used to determine group membership on a LAN and to determine the DR for it.
- JOIN_REQUEST, JOIN_ACK, QUIT_REQUEST with subcodes are used to connect to the CBT and to disconnect respectively. ACK packets are sent for JOIN_REQUEST as specified before.
- ECHO_REQUEST and REPLY messages are sent to determine reachability of routers on a network.

c) State information held in routers

- Routers have list of attached group members
- Routers have to know the addresses of the core routers and have specified target core routers for every group.
- Routers hold forwarding information by interface for every group
- A set of timers and flags used per group for every interface
- A list of children routers in the subnets

The underlying state information needed for this routing algorithm is saved as usual

PIM-SM

This protocol is based around the concept of Rendezvous Point (RP) for each multicast group. Routers that have members of the group attached have to contact the RP for join/prune actions.

PIM-SM may have one or more rendezvous points (RP). (There is a primary RP and alternative ones in case the primary RP is unreachable). These RPs are intentionally distinguished from CBT's cores, because traffic can be rerouted to bypass the rendezvous points for better efficiency.

The initiator of each multicast group selects a primary RP and a small set of alternative RPs, collectively called the RP-list. The routers must know this list.

a) Formation of routes

- When a router has a local host wanting to receive a group's multicast, the router has to contact the RP for that group if it is not already receiving the multicast. It sends a join message (*,group) to the RP. Along this route, routers add this entry to their forwarding tables. When the message reaches the RP, a path has been created and thus multicast starts flowing towards the router that sent the join message.

- When the traffic from a source to a host is large enough and there exists a shorter path than through the RP, the traffic switches to shortest path. Routers on the way change the entry to specific source and send prune messages to the appropriate interface when they receive the (S,G) message from the router with local members.

- When a router finds out that no receiver is present on its LAN anymore and if no route uses that LAN for the same group's multicast, it sends a prune message upstream. This message results in the truncation of unused branches of the new tree that does not include the leaving router.

- When new receivers come up, they cause the reopening of all sources multicast traffic that was closed when source specific routes were created.

b) Information exchanged between routers

- Routers periodically send join/prune messages to the RP exchanging the membership information
- DR routers are determined through examining of Hello messages, the router with the largest IP address on the network is the DR for the network.
- The DR collects group membership information through IGMP messages. It sends triggered join/prune and register messages towards the RP in addition to periodic ones.
- The RP information is collected by examining bootstrap information. Also, bootstrap messages are used to select a designated bootstrap router BSR. Candidate BSRs and candidate RPs are present in the system and an election mechanism is used to determine the domain BSR.
- The BSR periodically advertises the list of candidate RPs throughout the domain. Routers store this list and map groups to one of the CRPs (candidate RPs) that are members of the group using a given hashing function.
- Assert messages are exchanged when routers in a LAN or the multicast source find that they are sending duplicates. The assert mechanism allows the selection of the closest router to the RP on the LAN and other routers learn the result through examining the assert messages exchanged.
- The DR may loose an assert mechanism and the winning router takes the responsibility of sending (*,G) join message to the RP

- Routers attached to hosts that want to send to a group send PIM register messages. The register messages are sent towards the RP in unicast. Register_stop are sent back from RP to indicate that it joined the (S,G) tree or the absence of downstream receivers

c) State information held in routers

- Routers hold a list of (*,G) route entries with incoming interfaces and out-going interfaces in addition to source specific (S,G) entries with the same information.
- Designated routers have a list of directly connected member with group information
- Routers have a list of RPs sent by the BSR of the domain. Of course, they have the address of the BSR.
- A set of flags associated with every entry in the routing table. Each router has a wc bit, an SPT bit and an RPT bit used in handling routes.
- A set of timers are needed for multiple actions performed by routers
- A list of neighbors is compiled based on Hello messages
- In addition, all the routing information necessary for the underlying routing protocol is saved

As shown above, there are many similarities between CBT and PIM-SM. The most striking difference is PIM-SM's option of moving from the RP-based tree to a shortest-path tree.

Category II

In category II, we have DVMRP, MOSPF, and PIM-DM.

Of the category II algorithms, DVMRP and PIM-DM are more similar to each other than to MOSPF since they employ some form of RPM. However, PIM stands out because it is protocol independent, unlike DVMRP and MOSPF.

DVMRP

DVMRP is an interior gateway protocol for IP multicasting. It implements the Reverse Path Multicasting (RPM) algorithm. (The main difference between RPM and RPF is that RPM allows pruning information to be propagated upstream.) It is based on RIP, which is used to determine the best path back to a source, and includes additional functions and information needed to implement multicasting.

a.) Formation of routes

- The first step in creating multicast routes: the DVMRP routers exchange Neighbor Probe messages to determine the routers closest to them.

- Once their neighbors are determined, the routers use a protocol similar to RIP to exchange Routing Reports and to maintain their multicast routing tables.

- When a multicast packet arrives the DVMRP router uses the multicast routing table to compute the shortest reverse path tree using the source of the packet as the root of the tree.

- If the packet arrived on the interface that corresponds to the shortest path to the source, the packet is forwarded on all outgoing links of the tree.

- To save unnecessary transmissions the routers may prune certain outgoing links. Pruning will occur in two situations:

  1.) If the outgoing interface is a leaf network (no Router Probe messages have been received on the interface), and none of the end stations on the leaf network have indicated, using IGMP, they wish to receive the multicast traffic

  2.) If all dependent downstream interfaces have indicated they do not wish to receive the multicast traffic by sending a DVMRP Prune messages. Dependent downstream interfaces are determined using a Poisonous Reverse technique described below.

- If a router in a pruned section of the multicast tree determines it needs to begin receiving multicast traffic, it can send a DVMRP Graft message upstream to add the link back to the tree.

- In the event there are multiple routers on the same network, the router with the lowest cost metric to the source and if there is a tie, the lowest IP address becomes the designated forwarder for the network. Only the designated forwarder should forward multicast packets on the network.
  Note: the designated forwarder is determined on a per source basis.

- If an internetwork contains a mixture of DVMRP-aware and DVMRP-unaware routers, IP tunneling may be used to transparently pass through DVMRP unaware portions of the network.

b.) Information exchanged between routers

- DVMRP Probe messages
  These messages are exchanged between routers to locate their neighbors and determine their capabilities.

- DVMRP Route Reports
  DVMRP routers periodically exchange route reports to maintain the multicast routing tables. These reports use the same procedure as RIP for route calculation. In addition, if a router depends on a certain interface to receive traffic from a specific network, the router will set the cost metric to infinity for the network on the dependent interface. This technique is known as Poisonous Reverse and it can be used to determine which routers are downstream and therefore may possibly be pruned.

- DVMRP Prune
  If all downstream routers (routers which have sent a Poisonous Reverse Route Report) and all leaf networks are not interested in receiving multicast traffic, a router may propagate a DVMRP Prune message upstream to indicate it no longer is interested in receiving the multicast traffic.

- DVMRP Graft
  If a router has sent a Prune message and later wants to begin receiving multicast traffic for a specific group, then the router sends a DVMRP Graft message upstream.

c.) State information held in routers

- List of neighbor routers
- List of downstream routers for each possible source network
- Multicast routing tables
- Group membership information for each multicast group and each interface

PIM-DM

PIM-DM is intended for use in networks that have a high probability that any given network has group members (dense networks). The protocol is similar to DVMRP, but routers forward packets on all interfaces instead of only those interfaces in the shortest reverse path tree. PIM-DM also uses routing information from any existing routing protocol rather than using its own protocol, as in DVMRP or MOSPF. Thus, there is no need to exchange information for route formation.

Like DVMRP, PIM-DM employs RPM. However, it assumes that all downstream systems want to receive multicast datagrams unless pruning occurs. DVMRP builds a parent-child database and uses that to reduce duplicate packets. PIM-DM doesn't want to be dependent on the unicast routing protocol, so it uses an assert mechanism to resolve multiple forwarders. As in DVMRP, pruning state is maintained.

a.) Formation of routes

- Routers exchange PIM-Hello messages to determine their neighbor routers. If no PIM-Hello is received on a given interface, then that interface is considered a leaf network.
- The first time a packet for a given source/multicast address pair is received, the router broadcasts the packet on all outgoing links.
- If a router receives a multicast packet and none of the outgoing links are registered to receive that multicast address (either leaf networks with no IGMP registrations or non-leaf networks that are already pruned), the router sends a Prune message upstream. Once a network is pruned, the router will no longer forward packets on that network.
- Similar to the DVMRP-Graft messages, PIM-Join messages are used to remove the effects of a previous Prune message.
- If there are multiple routers with different paths to the source that are connected to the same network, PIM-Assert messages are used to elect one router as the "forwarder" for that network. The router with the lowest cost path to the source and the highest IP address if the path costs are the same is elected forwarder.

b.) Information exchanged between routers

- PIM-Hello messages
- PIM-Prune messages
- PIM-Join messages
- PIM-Assert messages

c.) State Information held in routers

- Routing information determined using a separate IP routing protocol
- Source/multicast address (S,G) pair entries for each port indicating if traffic should be forwarded on that port from the (S,G) pair.

MOSPF

Based on OSPF, network information available allows the calculation of source based shortest path tree at routers. Group membership information is collected through IGMP messages.

MOSPF is a protocol that makes use of detailed information about the entire network to minimize unnecessary datagram transmissions through intensive calculations at each router. Thus, unlike RPF-based methods, having the entire network map allows each router to calculate the shortest path tree using forward metrics. In a sense, MOSPF does the pruning ahead of time through exchanging group-membership link state advertisements.

a) Formation of routes

- Group membership information is collected by designated routers on each network and flooded in LSA throughout the AS.
- On demand, a shortest path tree based on the source is calculated. Branches that do not include receivers and are not on the path to other receivers are pruned. This tree is changed when topological changes occur.
- The calculation of the tree is different for the following 3 cases:

  1) Intra-area:
     When a group is contained entirely in one area or if no areas are configured, routers are able to calculate exact SPF tree in the area rooted at the specific source. This tree is pruned according to group membership information gathered by designated routers.
  2) Inter-area:
     If the source and some destinations are in different areas, the routers have less precise pictures of the overall graph. If the router performing these calculations is in the same area as the source, it should refrain from pruning branches that lead to wildcard multicast receivers, as they do not want to cut traffic for eventual receivers in other areas. If the router performing the calculations is in a different area than the source, it bases its tree on the source's area multicast forwarders' information.
  3) Inter-Autonomous System:
     Source and receivers are in different autonomous system. Routers in the source's AS must make sure that AS multicast forwards are not pruned out of multicast tree. Routers in the other AS base their trees on the AS multicast forwarder's information about external routers.

b) Information exchanged between routers

- Designated router is chosen for every network. It is responsible for sending periodic IGMP membership requests and collecting membership information and flood this information through the area using LSA
- Each area has multicast forwarder ABRs that summarize group membership to backbone in summary LSAs. This information is not re-advertised into other areas.
- Each area has wildcard multicast receivers (all multicast forwarders are wildcard multicast receivers). These receive all multicast traffic generated in the area and they allow multicast traffic to be exchanged between the areas they belong to.
- Selected ASBRs are multicast forwarders and provide summary information into other AS
- The usual OSPF LSAs are sent to provide underlying routing information.

c) State information held in routers

- Local group membership information is collected for the area by examining designated router's LSA.
- Entries for ABR that are multicast forwarders and wildcard receivers are added.
- OSPF border routers now include group membership information
- Routing cache is built based on tree and used later on for multicast forwarding.
- Multicast routers include group membership information

Example

Example with H1 as a source sending traffic to multicast group A:

Group A traffic - For DVMRP and PIM-DM, the first packet of the traffic will get flooded to all parts of the network except for leaf networks which do not have Group A members such as N10, N7, N4, and N1. When the packet reaches the routers of these leaf networks, such as RT12, RT8, RT3, and RT1, the routers will know through IGMP reports that there are no Group A members on that network and will not forward the packet onto the leaf network.

Of these routers, the ones that are leaf routers (RT8, RT1) will send prune messages to the router that is on the route to the source on the upstream interface.

These routers (RT3, RT10) will see the prune, and in DVMRP, the routers will know that there are other downstream routers (RT2) or systems directly attached to the interface which are still dependent on it for the (source,group) pair. Therefore, they will continue forwarding.

In PIM-DM, when these prune messages are sent, other dependent routers (RT2) who still wish to receive traffic will send PIM-Join messages to let the upstream router know that it still wishes to receive traffic for the (H1,A) pair. In the case of RT10, the router will know through IGMP that there is a member on N6, so it will not stop forwarding.

For the other router paths which the packet can take, such as the RT10-RT7-RT5-RT4 path:

In both PIM-DM and DVMRP, RT4 will see that RT3 has a shorter path to the source and is thus the designated router for N3. As a result since N3 is its only downstream interface, it will send a prune message back towards RT5. RT5 will choose RT6 as its upstream router because RT6 has equal metric back to the source as RT7, but has a lower IP address. RT5 will then send a prune message to RT6 to remove that branch from the multicast tree. RT7 will know that RT5 is not dependent on it for traffic and will thus send a prune message back to RT10. As before, however, RT10 will see through IGMP that there is member on the N6 interface that receives Group A traffic.

In MOSPF, the pruning process would not have to take place after packet transmission because with group membership LSAs the correct tree will already be known to the routers. (See RFC 1584 for an MOSPF example).

For the same scenario, we consider the core-based algorithms (CBT and PIM-SM). We consider that RT10 is a core router or rendezvous point (RP). The algorithms follow the procedures outlined above with respect to route formation. We are assuming that all the group members have already subscribed, so the trees are already formed. Host H1 starts sending to multicast group A. The datagram reaches RT12, which relays it on the broadcast network N9. RT9, because it is already on the tree, has set up a forwarding cache entry to pass the datagram to N11. RT11, being on the path to the core/RP, will forward the datagram toward RT10, which will in turn send it to all the routers that are on the tree.

Given a high enough volume of traffic or based on some other metric, hosts subscribed to Ma would be able to request to bypass RT10 and find a SP tree to H1 if a better one existed. Table entries consisting of (H1, Ma) would take precedence over (*, Ma) and choose the better path. One does not happen to exist for this example.