Chapter 3

Multicast Trees

Multicasting is the transmission of an IP datagram to a "host group", a set of zero or more hosts identified by a single IP destination address. A multicast tree consists of multiple groups and nodes. A node must be able to join and leave the group at any time. It can also be a member of one or more groups at the same time. The details about multicast address specification are detailed in Section 2.5.7 in this thesis.

Multicast Listener Discovery (MLD) by Deering, S. et al. (1999) is a protocol used to manage groups and locate new members in IPv6. Unique trees are formed based on the source and group address to route the multicast traffic.

3.1 Multicast Listener Discovery (MLD)

According to Deering, S. et al. (1999), MLD is used to enable each IPv6 router to discover the presence of multicast address on its directly attached links, and to discover specifically which multicast addresses are of interest to these neighboring nodes. This information is then provided to whichever multicast routing protocol used by the router, in order to ensure that multicast packets are delivered to all links where there are interested receivers.

MLD is used to manage membership information in IPv6 multicast, performing similar jobs by Internet Group Management Protocol, Version 2 (IGMPv2) (Fenner,

W., 1997) in IPv4. The development of MLD is still ongoing. Extensions to the current MLD specification such as Multicast Listener Discovery Version 2 (MLDv2) for IPv6 by Vida, Rolland et al. (2002) is released as an Internet-Draft in the Internet Engineering Task Force (IETF).

According to Haberman, B. and Thaler, D. (2001), the existing MLD messages can be extended to support the host-to-router membership exchanges for anycast groups. MLD is incorporated into this thesis for anycast group membership management.

MLD is a sub-protocol of Internet Control Message Protocol for IPv6 (ICMPv6), proposed by Conta, A. and Deering, S. (1995). MLD messages type are a subset of the set of ICMPv6 messages and MLD messages are identified in IPv6 packets by a preceding Next Header value of 58. All MLD messages are sent with a link-local IPv6 Source Address, an IPv6 Hop Limit of 1, and an IPv6 Router Alert option (Partridge, C. and Jackson, A., 1999) in a Hop-by-Hop Options header.

According to Deering, S. et al. (1999), there are three types of MLD messages:

Multicast Listener Query (Type = decimal 130)

There are two subtypes of Multicast Listener Query messages:

- General Query – used to learn which multicast addresses have listeners on an attached link.
- Multicast-Address-Specific Query – used to learn if a particular multicast address has any listeners on an attached link.

Multicast Listener Report (Type = decimal 131)

- When a node starts listening to a multicast address on an interface, it immediately transmits an unsolicited Report for that address on that interface.
- Response when receive MLD query messages, either General Query or Multicast-Address-Specific.

Multicast Listener Done (Type = decimal 132)

When a node ceased to listen to a multicast address, it should sent a single Done message to the link-scope all-routers multicast address (FF02::2), carrying in its Multicast Address field the address to which it is ceasing to listen.

3.2 Multicast Distribution Trees

Multicast protocols can use two different approaches to build a distribution tree:

- Source-based Tree
- Shared Tree.

There are many articles on the multicast distribution trees and multicast routing protocols, such as articles from Cisco (2001a), Liljebladh, Mikael and Ralitza Gateva (1999), Nortel (1999) and Youn, C. H. (2000), Fall, Kevin (1999).

3.2.1 Source-based Tree

The simplest form of a multicast distribution tree is a source tree with its root at the source and branches forming a spanning tree through the network to the receivers. Because this tree uses the shortest path through the network, it is also referred to as a

shortest path tree (SPT). A router only accepts broadcast messages on the interface, which is on the shortest path to the source to prevent loops and duplicate messages. This is called reverse-path forwarding (RPF). Branches without subscribers will be pruned off from the distribution tree. The prune message can only be sent from a leaf network upstream towards the source to disconnect from the distribution tree. Upstream routers will start to send multicast flows (MLD Listener Query messages) to all leaf networks periodically so they can rejoin the distribution tree if they have sent a prune message. An unsolicited Report message is sent from a leaf network upstream to rejoin a distribution tree again without waiting for the messages to be broadcasted out again.

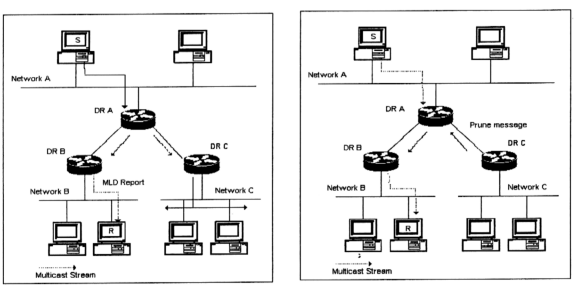

Figure 3.1 *The tree establishment processes for Source-based Tree*

The multicast stream to group G is split to both DR B and DR A (Figure 3.1, a). R is in network B and is a receiver of group G. Network C has no members of group G so DR C sends a prune message towards the source, DR A (Figure 3.1, b). DR A will send a General Query when the Query Interval timer expires. DR A will only send stream to DR B if DR B sends a Report message.

A source-based tree does not scale very well because it is not good to flood the entire Wide Area Network (WAN). Besides, a source-based tree will need much memory to maintain the different groups, prune states and all individual trees. However, a source-based tree is resilient to network failures since separate trees are maintained for each multicast recipient.

In IPv6 anycast, a tree is unnecessary for shortest-path routing. Details about shortest path routing for anycast will be discussed later. [Refer to Section 6.3]

3.2.2 Shared Tree

A shared tree, also called a core-based tree, has a special Rendezvous Point (RP). Unlike source-based tree, shared tree uses RP as a single common root placed at some chosen point in the network.

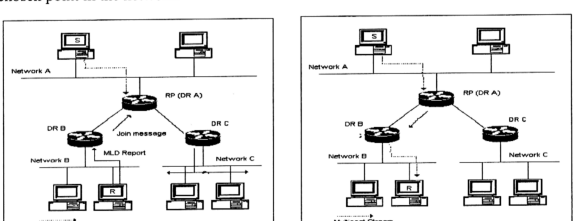

Figure 3.2 *Tree establishment processes for Shared Tree*

DR is the RP in this domain. S starts to send the multicast session for group G to the DR in the network A (DR A), which is the RP. R wants to be in group G and sends an MLD Report to the DR in the network (Figure 3.2, a). DR B will then look up the RP for the group G and send a join towards the RP. When the RP receives the join

messages, it starts to forward the multicast stream down on that interface (Figure 3.2, b).

A shared tree does not flood out messages as the source-based tree. So it scales better because it uses less network resources and has fewer overheads. An extra protocol is needed to distribute the information of the RPs. However, shared trees are more vulnerable to failures than a source-based tree if a failure occurs close to a RP or a RP goes down. Besides, the RP may become a bottleneck in a network if there are many sources that send much traffic. The shared tree approach is good in WANs with members that are far from each other and the network resources may not be so good.

3.3 Multicast Routing Protocols

There are different protocols for routing the multicast packets. All of them are either based on the source-based tree or the shared tree approach to form the distribution tree.

3.3.1 Distance Vector Multicast Routing Protocol (DVMRP)

Distance Vector Multicast Routing Protocol (DVMRP) (Waitzman, D. and Deering, S., 1998) is used today as the main routing protocol in the Multicast Backbone (Mbone). All LANs connected to the Mbone can use the other protocols if the border routers can interoperate with DVMRP. DVMRP is a distributed protocol that dynamically generates multicast delivery trees using RPF. It generates source-based tree by broadcasting and pruning.

There are some improvements in DVMRP. For example, the introduction of the Hierarchical DVMRP, that is an aggregation of subnets into shorter prefixes to reduce DVMRP routing table size. Automatic recognition of "stub trees" is introduced to reduce routing message overhead on slow links. Besides, DVMRP also provide support for host-specific pruning, tailor for IGMPv3 (Cain, Brad et al., 2002; Cisco, 2001b).

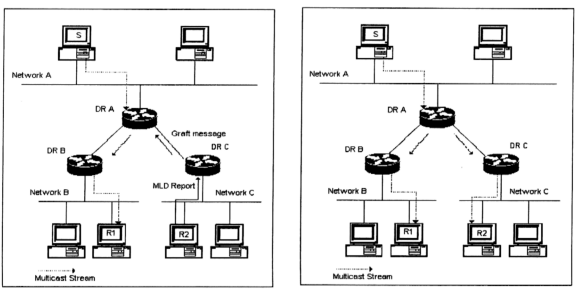

Figure 3.3 *Tree establishment processes for DVMRP*

DR C sent a prune message for the group G. R1 is the only group member until now. R2 decides to join and sends a MLD Report to DR C (left figure, Figure 3.3). DR C then sends a graft message towards the source because it has sent a prune message earlier. When DR A receives the graft message, it will remove the prune-state and start to send the multicast stream to DR C (right figure, Figure 3.3).

3.3.2 Multicast Open Shortest Path First (MOSPF)

Multicast Open Shortest Path First (MOSPF) (Moy, J., 1994) is a multicast enhancement of Open Shortest Path First (OSPF) Version 2 (Moy, J., 1998) to support multicast routing. The OSPF for IPv6, introduced by Coltun et al. (1999) is using the source-based tree approach. It is still a sparse mode protocol, as the multicast packets are not flooded out to paths where there are no receivers. OSPF is a protocol for routing unicast information and uses different costs for every path to find the path with the least cost. MOSPF has a two-level hierarchy because of the Autonomous System (AS) in OSPF. Routing information collected by the OSPF is used by MOSPF.

In MOSPF, multicast capable routers flag link state routing advertisements. Each router will indicate groups for which there are directly connected members. Link-state advertisements will be augmented with multicast group addresses to which local members have joined and link-state routing algorithm will be augmented to compute shortest-path distribution tree from any source to any set of destinations.

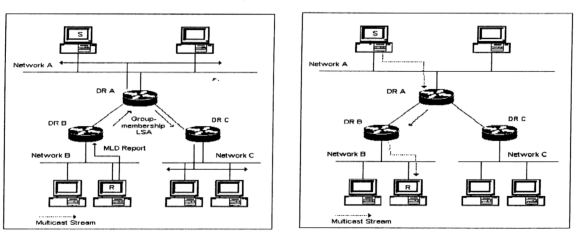

Figure 3.4 *The tree establishment processes for MOSPF*

Receiver R joins the group G by sending out an MLD Report to DR B. DR B will then flood the entire network area with a group-membership-LSA (left figure, Figure 3.4). The routers calculate the distribution tree and the multicast stream from S will reach receiver R.

3.3.3 Core Based Tree (CBT)

Core Based Tree (CBT) introduced by Ballardie, A. (1997) is based on the shared tree approach to form a distribution tree. A Rendezvous Point (RP), the core router is used as a meeting point for the sender and group receivers. A group will have only one RP. The upstream interface is the interface that leads towards the RP and not necessary the source. The downstream interface is likewise the interface leading from the RP and not necessary from the source.

In short, CBTs are bidirectional center-based shared trees routed at core where the receivers will send join messages to core and the senders will send data to core. There are no Shortest-Path Trees in CBTs. Although the hard state with acknowledged join from core or first on-tree router technique used by CBTs don't require source specific state, however it will increase the path lengths and may cause traffics concentration near the core. CBTs require explicit joining, where the routers send join messages towards the core. However, there is not much implementation of the CBTs currently.

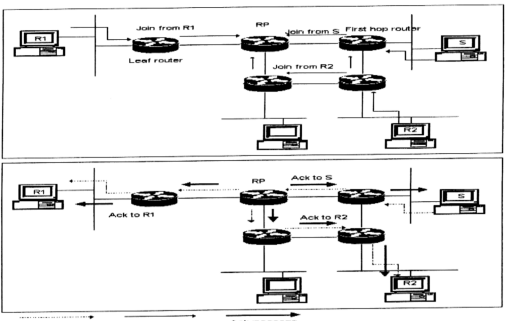

Figure 3.5 *Tree establishment processes for Core Based Tree*

The top figure in Figure 3.5 shows how all the routers join the same RP. ACKs are then sent back to the group members that joined and the multicast session can start (bottom figure, Figure 3.5). The paths used are independently on who's the sender. The path to R2 is not the optimal in this case when S is the sender.

3.3.4 Protocol Independent Multicast Version 2 (PIM)

Protocol Independent Multicast Version 2 (PIM), a protocol proposed by IETF gets its name from the fact that it is IP routing protocol independent. PIM can leverage whichever unicast routing protocols that are used to populate the routing table including Enhanced Interior Gateway Routing Protocol (EIGRP) (Cisco, 2002a), OSPF, Border Gateway Protocol (BGP) (Cisco, 2002b) or static routes. PIM uses this unicast routing information to perform the multicast forwarding function; therefore it is IP protocol independent. PIM uses the unicast routing table to perform the Reverse Path Forwarding (RPF) check function and does not send and receive multicast

routing updates between routers like other routing protocols. There are two different approaches in PIM, PIM Dense Mode (PIM-DM) and PIM Sparse Mode (PIM-SM).

PIM Dense Mode (PIM-DM)

PIM-DM (Adams, Andres, 2002) is a protocol based on the assumption that the group members are densely distributed over the network. PIM-DM uses the source-based tree approach to form the distribution tree because of that assumption. PIM-DM initially floods multicast traffic throughout the network. Routers that do not have any downstream neighbors prune back the unwanted traffic and this process repeats every three minutes. The routers accumulate their state information by receiving the data stream through the flood and prune mechanism. These data streams contain the source and group information so that downstream routers can build up their multicast-forwarding table. PIM-DM can only support source trees – (S, G) entries and it cannot be used to build a shared distribution tree.

PIM Sparse Mode (PIM-SM)

PIM-SM (Estrin, D. et al., 1998) is based on the assumption that the group members are distributed over a wide area. A variant to the shared tree approach is used to form the distribution trees. When a multicast session begins the multicast flow will go through the RP. The shortest path can be used after the receiver receives the first packet by building a source-specific tree (see Figure 3.6). PIM-SM has the ability to interoperate with DVMRP. Discussions about PIM-SM are deeper than the other multicast routing protocol because PIM-SM is being implemented in this thesis and a PIM-SM extension model for anycast is defined later in Chapter 6.

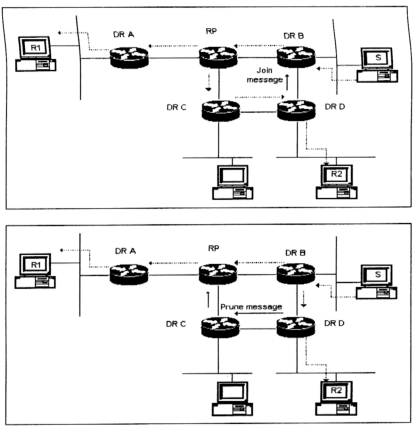

Figure 3.6 *Tree establishment processes for PIM-SM*

DR D decides to use the shortest path and sends a join message towards the source (top figure, Figure 3.6). As soon as DR D receives the first packet from S, DR D sends a prune message towards the RP.

A source that starts a new multicast session starts to send the packet as a multicast message to the first-hop router. The first-hop router will then encapsulate the multicast message with a Register message and send it as a unicast message to the RP for that group. If there are no members registered at the RP for that group, the first-hop router must continue to encapsulate the messages to the RP until it gets a Register-Stop message. A Register-Stop message is triggered when a group member

joins the group. The RP will send Register-Stop message periodically to the source as long as there is any member in the group.

A leaf router can switch to a shortest path tree by a request to the source after it has received at least one multicast packet from the RP. The leaf router will prune towards the RP when it receives the first datagram from the source and it will prevent the RP from being congested. However, "black holes" occur in the multicast flows if the shortest path is much shorter than the shared path – as if the packets traveling on the shared path have not reached the receiver before the first packet from the shortest path arrives.

A BootStrap Router (BSR) is responsible for collecting and sending out information regarding the RPs. It sends a BootStrap message out to all routers within the PIM domain periodically to spread the information of RPs. The mapping of groups and RPs is done through a hash function. A Candidate-RP (C-RP), a backup RP will take over the role of the RP when the ordinary RP is unavailable due to failures close to the RP or the RP crashes. A new BootStrap message with a new RP-set without the unavailable RP will be sent out by BSR to all routers. The routers will use a hash function to determine which C-RP should be used.

PIM-SM is said to be a good approach of the shared tree approach as it has the ability to switch to shortest path so that the RP is not so congested and the overhead of the protocol is kept down by using the router information sent by underlying routing protocol. However, big tables created by using one distribution tree for each sender in a group if there are many senders in each group.