Network Working Group
Internet Draft
Expiration date: February 1993

J. Moy
Proteon, Inc.
September 1992

Multicast Extensions to OSPF

Status of this Memo

Abstract

This memo documents enhancements to the OSPF protocol enabling the routing of IP multicast datagrams. In this proposal, an IP multicast packet is routed based both on the packet's source and its multicast destination (commonly referred to as source/destination routing). As it is routed, the multicast packet follows a shortest path to each multicast destination. During packet forwarding, any commonality of paths is exploited; when multiple hosts belong to a single multicast group, a multicast packet will be replicated only when the paths to the separate hosts diverge.

OSPF, a link-state based routing protocol, provides a database describing the Autonomous System's topology. A new OSPF link state advertisement is added describing the location of multicast destinations. A multicast packet's path is then calculated by building a pruned shortest-path tree rooted at the packet's IP source. These trees are built on demand, and the results of the calculation are cached for use by subsequent packets.

The multicast extensions are built on top of OSPF version 2. The extensions have been implemented so that a multicast routing capability can be introduced piecemeal into an OSPF version 2 routing domain. Some of the OSPF version 2 routers may run the multicast extensions, while others may continue to be restricted to the forwarding of regular IP traffic (unicasts).

Please send comments to mospf@gated.cornell.edu.

1 Introduction

This memo documents enhancements to OSPF version 2 to support IP multicast routing. The enhancements have been added in a backward-compatible fashion; routers running the multicast additions will interoperate with non-multicast OSPF routers when forwarding regular (unicast) IP data traffic. The protocol resulting from the addition of the multicast enhancements to OSPF is herein referred to as the MOSPF protocol.

IP multicasting is an extension of LAN multicasting to a TCP/IP internet. Multicasting support for TCP/IP hosts has been specified in [RFC 1112]. In that document, multicast groups are represented by IP class D addresses. Individual TCP/IP hosts join (and leave) multicast groups through the Internet Group Management Protocol (IGMP, also specified in [RFC 1112]). A host need not be a member of a multicast group in order to send datagrams to the group. Multicast datagrams are to be delivered to each member of the multicast group with the same “best-effort” delivery accorded regular (unicast) IP data traffic.

MOSPF provides the ability to forward multicast datagrams from one IP network to another (i.e., through internet routers). MOSPF forwards a multicast datagram on the basis of both the datagram’s source and destination (this is sometimes called source/destination routing). The OSPF link state database provides a complete description of the Autonomous System’s topology. By adding a new type of link state advertisement, the **group-membership-LSA**, the location of all multicast group members is pinpointed in the database. The path of a multicast datagram can then be calculated by building a shortest-path tree rooted at the datagram’s source. All branches not containing multicast members are pruned from the tree. These pruned shortest-path trees are initially built when the first datagram is received (i.e., on demand). The results of the shortest path calculation are then cached for use by subsequent datagrams having the same source and destination.

OSPF allows an Autonomous System to be split into areas. However, when this is done complete knowledge of the Autonomous System’s topology is lost. When forwarding multicasts between areas, only incomplete shortest-path trees can be built. This may lead to some inefficiency in routing. An analogous situation exists when the source of the multicast datagram lies in another Autonomous System. In both cases (i.e., the source of the datagram belongs to a different OSPF area, or to a different Autonomous system) the neighborhood immediately surrounding the source is unknown. In these cases the source’s neighborhood is approximated by OSPF summary link advertisements or by OSPF AS external link advertisements respectively.

Routers running MOSPF can be intermixed with non-multicast OSPF routers. Both types of routers can interoperate when forwarding regular (unicast) IP data traffic. Obviously, the range of IP multicasts is limited by the number of MOSPF routers present in the Autonomous System (and their interconnection, if any). An ability to “tunnel” multicast datagrams through non-multicast routers is not provided. In MOSPF, just as in the base OSPF protocol, datagrams (multicast or unicast) are routed “as is” -- they are not further encapsulated or decapsulated as they transit the Autonomous System.

1.1 Terminology

This memo uses the terminology listed in section 1.2 of [OSPF]. For this reason, terms such as “Network”, “Autonomous System” and “link state advertisement” are assumed to be understood. In addition, the abbreviation LSA is used for “link state advertisement”. For example, router links advertisements are referred to as router-LSAs and the new link state advertisement describing the location of members of a multicast group is referred to as a group-membership-LSA.

[RFC 1112] discusses the data-link encapsulation of IP multicast datagrams. In contrast to the normal forwarding of IP unicast datagrams, on a broadcast network the mapping of an IP multicast destination to a data-link destination address is not done with the ARP protocol. Instead, static mappings have been defined from IP multicast destinations to data-link addresses. These mappings are dependent on network type; for some networks IP multicasts are algorithmically mapped to data-link multicast addresses, for other networks all IP multicast destinations are mapped onto the data-link broadcast address. This document loosely describes both of these possible mappings as **data-link multicast**.

The following terms are also used throughout this document:

- **Non-multicast router.** A router running OSPF version 2, but not the multicast extensions. These routers do not forward multicast datagrams, but can interoperate with MOSPF routers in the forwarding of unicast packets. Routers running the MOSPF protocol are referred to herein as either multicast-capable routers or MOSPF routers.
- **Non-broadcast networks**. A network supporting the attachment of more than two stations, but not supporting the delivery of a single physical datagram to multiple destinations (i.e., not supporting data-link multicast). [OSPF] describes these networks as non-broadcast, multi-access networks. An example of a non-broadcast network is an X.25 PDN.
- **Transit network.** A network having two or more OSPF routers attached. These networks can forward data traffic that is neither locally-originated nor locally-destined. In OSPF, with the exception of point-to-point networks and virtual links, the neighborhood of each transit network is described by a network links advertisement (network-LSA).
- **Stub network.** A network having only a single OSPF router attached. A network belonging to an OSPF system is either a transit or a stub network, but never both.

1.2 Acknowledgments

The multicast extensions to OSPF are based on Link-State Multicast Routing algorithm presented in [Deering]. In addition, the [Deering] paper contains a section on Hierarchical Multicast Routing (providing the ideas for MOSPF's inter-area multicasting scheme) and several Distance Vector (also called Bellman-Ford) multicast algorithms. One of these Distance Vector multicast algorithms, Truncated Reverse Path Broadcasting, has been implemented in the Internet (see [RFC 1075]).

The MOSPF protocol has been developed by the MOSPF Working Group of the Internet Engineering Task Force. Portions of this work have been supported by DARPA under NASA contract NAG 2-650.

2 Multicast routing in MOSPF

This section describes MOSPF's basic multicast routing algorithm. The basic algorithm, run inside a single OSPF area, covers the case where the source of the multicast datagram is inside the area itself. Within the area, the path of the datagram forms a tree rooted at the datagram source.

2.1 Routing characteristics

As a multicast datagram is forwarded along its shortest-path tree, the datagram is delivered to each member of the destination multicast group. In MOSPF, the forwarding of the multicast datagram has the following properties:

- The path taken by a multicast datagram depends both on the datagram's source and its multicast destination. Called source/destination routing, this is in contrast to most unicast datagram forwarding algorithms (like OSPF) that route based only on destination.
- The path taken between the datagram's source and any particular destination group member is the least cost path available. Cost is expressed in terms of the OSPF link-state metric. For example, if the OSPF metric represents delay, a minimum delay path is chosen. OSPF metrics are configurable. A metric is assigned to each outbound router interface, representing the cost of sending a packet on that interface. The cost of a path is the sum of its constituent (outbound) router interfaces[1].
- MOSPF takes advantage of any commonality of least cost paths to destination group members. However, when members of the multicast group are spread out over multiple networks, the multicast datagram must at times be replicated. This replication is performed as few times as possible (at the tree branches), taking maximum advantage of common path segments.
- For a given multicast datagram, all routers calculate an identical shortest-path tree. There is a single path between the datagram's source and any particular destination group member. This means that, unlike OSPF's treatment of regular (unicast) IP data traffic, there is no provision for equal-cost multipath.
- On each packet hop, MOSPF normally forwards IP multicast datagrams as data-link multicasts. There are two exceptions. First, on non-broadcast networks, since there are no data-link multicast/broadcast services the datagram must be forwarded to specific MOSPF neighbors (see Section 2.3.3). Second, a MOSPF router can be configured to forward IP multicasts on specific networks as data-link unicasts, in order to avoid datagram replication in certain anomalous situations (see Section 6.4).

While MOSPF optimizes the path to any given group member, it does not necessarily optimize the use of the internetwork as a whole. To do so, instead of calculating source-based shortest-path trees, something similar to a minimal spanning tree (containing only the group members) would need to be calculated. This type of minimal spanning tree is called a Steiner tree in the literature. For a comparison of shortest-path tree routing to routing using Steiner trees, see [Deering2] and [Bharath-Kumar].

2.2 Sample path of a multicast datagram

As an example of multicast datagram routing in MOSPF, consider the sample Autonomous System pictured in Figure 1. This figure has been taken from the OSPF specification (see [OSPF]). The larger rectangles represent routers, the smaller rectangles hosts. Oblongs and circles represent multi-access networks[2]. Lines joining routers are point-to-point serial connections. A cost has been assigned to each outbound router interface.

All routers in Figure 1 are assumed to be running MOSPF. A number of hosts have been added to the figure. The hosts labelled Ma have joined a particular multicast group (call it group A) via the IGMP protocol. These hosts are

1. Actually, OSPF allows a separate link cost to be configured for each TOS. MOSPF then potentially calculates separate paths for each TOS. For more details, see Section 6.2.

2. We also assume in this section that the pictured multi-access networks provide data-link multicast/broadcast services.

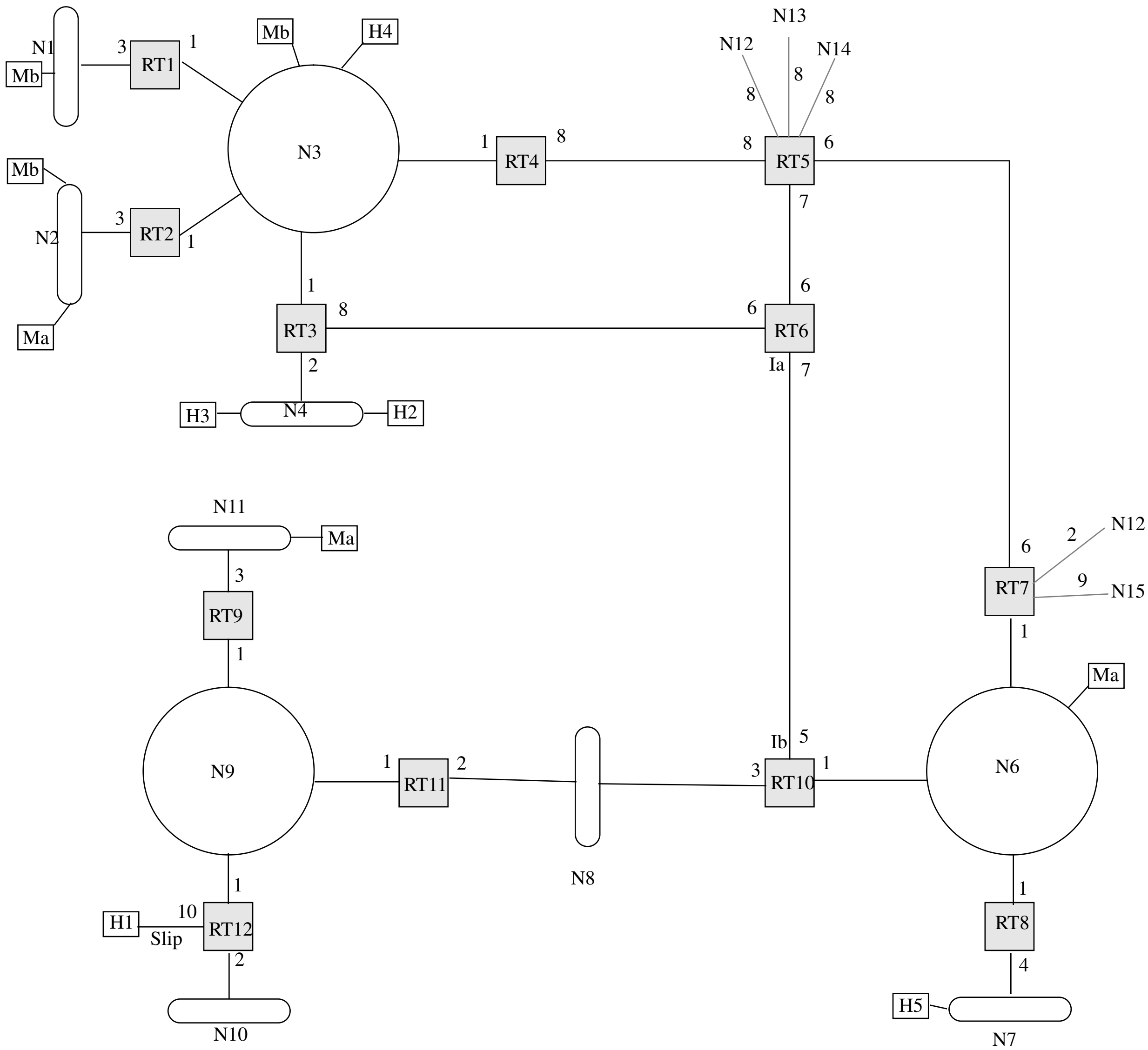


Figure 1: A sample MOSPF configuration

located on networks N2, N6 and N11. Similarly, using IGMP the hosts labelled Mb have joined a separate multicast group B; these hosts are located on networks N1, N2 and N3. Note that hosts can join multiple multicast groups; it is only for clarity of presentation that each host has joined at most one multicast group in this example. Also, hosts H2 through H5 have been added to the figure to serve as sources for multicast datagrams. Again, the datagrams' sources have been made separate from the group members only for clarity of presentation.

To illustrate the forwarding of a multicast datagram, suppose that host H2 (attached to network N4) sends a multicast datagram to multicast group B. This datagram originates as a data-link layer multicast on network N4. Router RT3, being a multicast router, has “opened up” its interface data-link multicast filters. It therefore receives the multicast datagram, and attempts to forward it to the members of multicast group B (located on networks N1, N2 and N3). This is accomplished by sending a single copy of the datagram onto network N3, again as a data-link multicast[3]. Upon receiving the multicast datagram from RT3, routers RT1 and RT2 will then multicast the datagram on their connected stub networks (N1 and N2 respectively). Note that, since the datagram is sent onto network N3 as a data-link multicast, router RT4 will also receive a copy. However, it will not forward the datagram, since it does not lie on a shortest path between the source (host H2) and any members of multicast group B.

Note that the path of the multicast datagram depends on the datagram’s source network. If the above multicast datagram was instead originated by host H3, the path taken would be identical, since hosts H2 and H3 lie on the same network (net N4). However, if the datagram was originated by host H4, its path would be different. In this case, when router RT3 receives the datagram, RT3 will drop the datagram instead of forwarding it (since RT3 is no longer on the shortest path to any member of group B).

Note that the path of the multicast datagram also depends on the destination multicast group. If host H2 sends a multicast to group A, the path taken is as follows. The datagram again starts as a multicast on network N4. Router RT3 receives it, and creates two copies. One is sent onto net N3 which is then forwarded onto network N2 by RT2. The other copy is sent to router RT10 (via RT6), where the datagram is again split, eventually to be delivered onto networks N6 and N11. Note that, although multiple copies of the datagram are produced, the datagram itself is not modified as it is forwarded. No encapsulation of the datagram is performed; the destination of the datagram is always listed as the multicast group A.

2.3 MOSPF forwarding mechanism

Each MOSPF router in the path of a multicast datagram bases its forwarding decision on the contents of a data cache. This cache is called the **forwarding cache**. There is a separate forwarding cache entry for each source/destination combination[4]. Each cache entry indicates, for multicast datagrams having matching source and destination, which neighboring node (i.e., router or network) the datagram must be received from (called the **upstream node**) and which interfaces the datagram should then be forwarded out of (called the **downstream interfaces**).

A forwarding cache entry is actually built from two component pieces. The first of these components is called the **local group database.** This database, built by the IGMP protocol, indicates the group membership of the router’s directly attached networks. The local group database enables the local delivery of multicast datagrams. The second component is the **datagram’s shortest path tree.** This tree, built on demand, is rooted at a multicast datagram’s source. The datagram’s shortest path tree enables the delivery of multicast datagrams to distant (i.e., not directly attached) group members.

---

3. Note that if N3 were a non-broadcast network, router RT3 would send separate copies of the datagram to routers RT1 and RT2. Since the IGMP protocol is not defined on non-broadcast networks, there could in this case be no group B member attached to network N3. However the multicast datagram would still be delivered to the group B members attached to networks N1 and N2.

4. Actually, in MOSPF there is a separate forwarding cache entry for each combination of source, destination and TOS. For a discussion of TOS-based multicast routing, see Section 6.2.

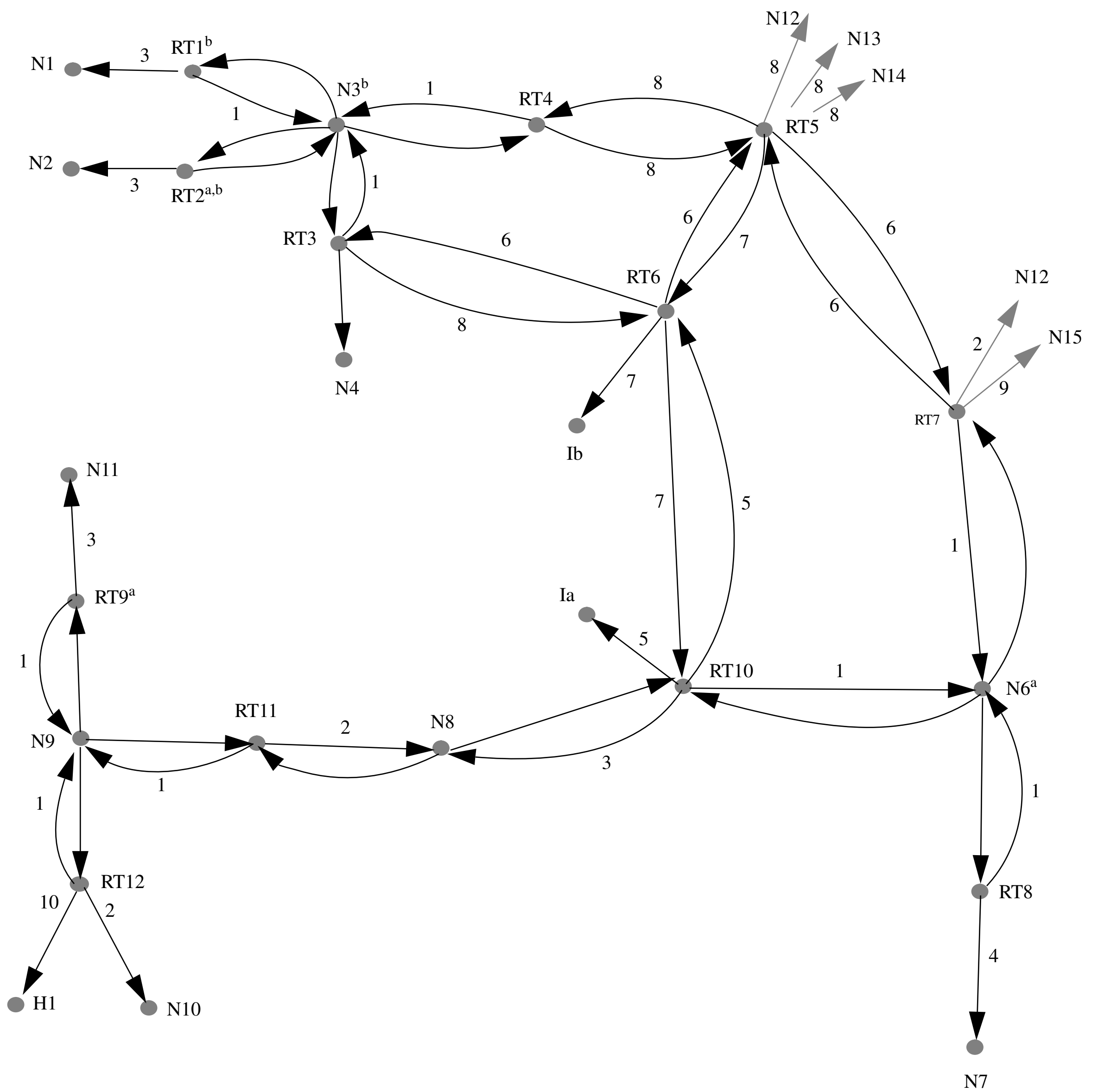


Figure 2: The MOSPF database

2.3.1 IGMP interface: the local group database

The **local group database** keeps track of the group membership of the router's directly attached networks. Each entry in the local group database is a *[group, attached network]* pair, which indicates that the *attached network* has one or more IP hosts belonging to the IP multicast destination *group*. This information is then used by the router when deciding which directly attached networks to forward a received IP multicast datagram onto, in order to complete delivery of the datagram to (local) group members.

The **local group database** is built through the operation of the Internet Group Membership Protocol (IGMP; see [RFC 1112]). When an MOSPF router becomes Designated Router on an attached network (call the network N1), it starts sending periodic IGMP Host Membership Queries on the network. Hosts then respond with IGMP Host Membership Reports, one for each multicast group to which they belong. Upon receiving a Host Membership Report for a multicast group A, the router updates its **local group database** by adding/refreshing the entry [Group A, N1]. If at a later time Reports for group A cease to be heard on the network, the entry is then deleted from the local group database.

It is important to note that on any particular network, the sending of IGMP Host Membership queries and the listening to IGMP Host Membership Reports is performed solely by the Designated Router. A MOSPF router ignores Host Membership Reports received on those networks where the router has not been elected Designated Router.[5] This means that at most one router performs these IGMP functions on any particular network, and ensures that the network appears in the **local group database** of at most one router. This prevents multicast datagrams from being replicated as they are delivered to local group members. As a result, each router in the Autonomous System has a different local group database. This is in contrast to the MOSPF link state database, and the datagram shortest-path trees (see Section 2.3.2), all of which are identical in each router belonging to the Autonomous System.

The existence of local group members must be communicated to the rest of the routers in the Autonomous System. This ensures that a remotely-originated multicast datagram will be forwarded to the router for distribution to its local group members. This communication is accomplished through the creation of a **group-membership-LSA**. Like other link state advertisements, the group-membership-LSA is flooded throughout the Autonomous System. The router originates a separate group-membership-LSA for each multicast group having one or more entries in the **local group database**. The router's group-membership-LSA (say for group A) lists those local transit vertices (i.e., the router itself and/or any directly connected transit networks) that should not be pruned from group A's datagram shortest-path trees. The router lists itself in the group-membership-LSA for group A if either 1) one or more of the router's attached stub networks contain group A members or 2) the router itself is a member of group A. The router lists a directly connected transit network in the group-membership-LSA for group A if both 1) the router is Designated Router on the network and 2) the network contains one or more group A members.

Consider again the example pictured in Figure 1. If router RT3 has been elected Designated Router for network N3, then Table 1: lists the local group database for the routers RT1-RT4.

Router	*local group database*
RT1	[Group B, N1]
RT2	[Group A, N2], [Group B, N2]
RT3	[Group B, N3]
RT4	None

Table 1: Sample local group databases

5. The discussion in this section omits mention of the Backup Designated Router's role in the IGMP protocol. While the Backup Designated Router does not send IGMP Host Membership Queries, it does listen to IGMP Host Membership Reports, building "shadow" local group database entries in the process. These entries do not lead to group-membership-LSAs, nor do they influence delivery of multicast datagrams, but are merely maintained to ease the transition from Backup Designated Router to Designated Router, should the Designated Router fail. See Sections 2.3.4, 9 and 10 for details.

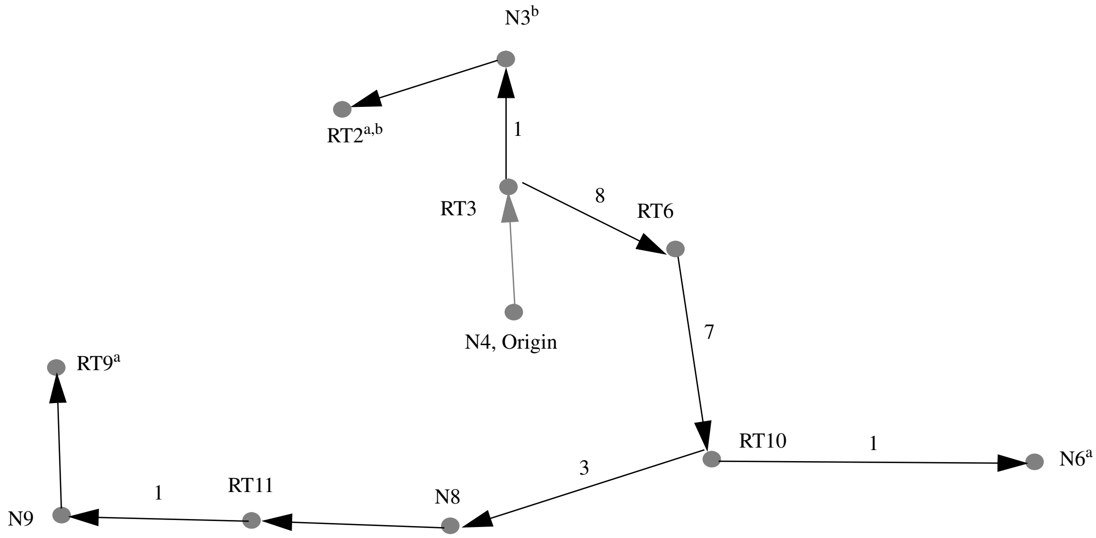


Figure 3: Sample datagram's shortest-path tree, source N4, destination Group A

In this case, each of the routers RT1, RT2 and RT3 will originate a group-membership-LSA for Group B. In addition, RT2 will also be originating a group-membership-LSA for Group A. RT1 and RT2's group-membership-LSAs will list solely the routers themselves (N1 and N2 are stub networks). RT3's group-membership-LSA will list the transit network N3.

Figure 2 displays the Autonomous System's link state database. A router/transit network is labelled with a multicast group if (and only if) it has been mentioned in a group-membership-LSA for the group (indicated by superscripts in Figure 2). When building the shortest-path tree for a particular multicast datagram, this labelling enables those branches without group members to be pruned from the tree. The process of building a multicast datagram's shortest path tree is discussed in Section 2.3.2.

Note that none of the hosts in Figure 1 belonging to multicast groups A and B show up explicitly in the link state database (see Figure 2). In fact, looking at the link state database you cannot even determine which stub networks contain multicast group members. The link state database simply indicates those routers/transit networks having attached group members. This is all that is necessary for successful forwarding of multicast datagrams.

2.3.2 A datagram's shortest-path tree

While the **local group database** facilitates the local delivery of multicast datagrams, the **datagram's shortest-path tree** describes the intermediate hops taken by a multicast datagram as it travels from its source to the individual multicast group members. As mentioned above, the datagram's shortest-path tree is a pruned shortest-path tree rooted at the datagram's source. Two datagrams having differing [source net, multicast destination] pairs may have, and in fact probably will have, different pruned shortest-path trees.

A **datagram's shortest path tree** is built "on demand"[6], i.e., when the first multicast datagram is received having a particular [source net, multicast destination] combination. To build the datagram's shortest-path tree, the following calculations are performed. First the datagram's source IP network is located in the link state database. Then using the router-LSAs and network-LSAs in the link state database, a shortest-path tree is built having the source network as root. To complete the process, the branches that do not contain routers/transit networks that have been labelled with the particular multicast destination (via a group-membership-LSA) are pruned from the tree.

As an example of the building of a datagram's shortest path tree, again consider the Autonomous System in Figure 1. The Autonomous System's link state database is pictured in Figure 2. When a router initially receives a multicast datagram sent by Host H2 to the multicast group A, the following steps are taken: Host H2 is first determined to be on network N4. Then the shortest path tree rooted at net N4 is calculated[7], pruning those branches that do not contain routers/transit networks that have been labelled with the multicast group A. This results in the pruned shortest-path tree pictured in Figure 3. Note that at this point the all the leaves of the tree are routers/transit networks labelled with multicast group A (routers RT2 and RT9 and transit network N6).

In order to forward the multicast datagram, each router must find its own position in the datagram's shortest path tree. The router's (call it router RT1) position in the datagram's pruned shortest-path tree consists of 1) RT1's parent in the tree (this will be the forwarding cache entry's **upstream node)** and 2) the list of RT1's interfaces that lead to downstream routers/transit networks that have been labelled with the datagram's destination (these will be added to the forwarding cache entry as **downstream interfaces**). Note that after calculating the datagram's shortest path tree, a router may find that it is itself not on the tree. This would be indicated by a forwarding cache entry having no upstream node or an empty list of downstream interfaces.

As an example of a router describing its position on the datagram's shortest-path tree, consider Router RT10 in Figure 3. The **upstream node** for the datagram is router RT6, and there are two **downstream interfaces**: one connecting to network N6 and the other connecting to network N8.

2.3.3 Support for Non-broadcast networks

When forwarding multicast datagrams over non-broadcast networks, the datagram cannot be sent as a link-level multicast (since neither link-level multicast nor broadcast are supported on these networks), but must instead be forwarded separately to specific neighbors. To facilitate this, forwarding cache entries can also contain **downstream neighbors** as well as **downstream interfaces**.

The IGMP protocol is not defined over non-broadcast networks. For this reason, there cannot be group members directly attached to non-broadcast networks, nor do non-broadcast networks ever appear in local group database entries.

As an example, suppose that network N3 in Figure 1 is an X.25 PDN. Consider Router RT3's forwarding cache entry for datagrams having source network N4 and multicast destination Group B. In place of having the interface to net-

---

6. One might imagine building all possible datagram shortest-path trees up front. However, this might be expensive, both in router CPU time and in router memory. It is hoped that building the datagram shortest-path trees on demand and caching the results will ease demands on router resources by spreading out the calculations over a longer period of time.

7. It is possible that, due to the existence of alternate paths, several different shortest-path trees are available. MOSPF depends on all routers constructing the exact same shortest path tree. For that reason, tie-breaking schemes have been implemented during tree construction to ensure that identical trees result. See Section 12 for more details.

work N3 appear as the downstream interface in the matching forwarding cache entry, the neighboring routers RT1 and RT2 would instead appear as separate downstream neighbors. In addition, in this case there could not be a group B member directly attached to network N3.

2.3.4 Details concerning forwarding cache entries

Each of the downstream interface/neighbors in the cache entry is labelled with a TTL value. This value indicates the number of hops a datagram forwarded out of the interface (or forwarded to the neighbor) would have to travel before encountering a router/transit network requesting the multicast destination. The reason that a hop count is associated with each downstream interface/neighbor is so that IP multicast's expanding ring search procedure can be more efficiently implemented. By expanding ring search is meant the following. Hosts can restrict the range of the IP multicast datagrams that they send by appropriate setting of the TTL value in the datagram's IP header. Then, for example, to search for the nearest server the host can send multicasts first with TTL set to 1, then 2, etc. By attaching a hop count to each downstream interface/neighbor in the forwarding cache, datagrams will not be forwarded unless they will ultimately reach a multicast destination before their TTL expires[8]. This avoids wasting network bandwidth during an expanding ring search.

As an example consider Router RT10's forwarding cache in Figure 3. Router RT10's cache entry has two **downstream interfaces**. The first, connecting to network N6, is labelled as having a group member one hop away (Network N6). The second, which connects to network N8, is labelled as having a group member two hops away (Router RT9).

Both the datagram shortest path tree and the local group database may contribute downstream interfaces to the forwarding cache entries. As an example, if a router has a local group database entry of [Group G, NX], then an forwarding cache entry for Group G, regardless of destination, will list the router interface to Network NX as a downstream interface. Such a downstream interface will always be labelled with a TTL of 1.

As an example of forwarding cache entries, again consider the Autonomous System pictured in Figure 1. Suppose Host H2 sends a multicast datagram to multicast group B. In that case, several routers will not even attempt to build a forwarding cache entry (routers RT5 and RT7) because they will never receive the multicast datagram in the first place. Other routers will receive the multicast datagram (since they are forwarded as link-level multicasts), but after building the pruned shortest path tree will notice that they themselves are not a part of the tree (routers RT1, RT4, RT8 and RT12). These latter routers will install an empty cache entry, indicating that they do not participate in the forwarding of the multicast datagram. A sample of the forwarding cache entries built by the other routers in the Autonomous System is pictured in Table 2:

Router	*Upstream node*	*Downstream interfaces (interface:hops)*
RT10	Router RT6	(N6:1), (N8:3)
RT11	Net N8	(N9:2)
RT3	Net N4	(N3:1), (RT6:3)
RT6	Router RT3	(RT10:2)

Table 2: Sample forwarding cache entries, for source N4 and destination Group A.

8. Note that the expanding ring search yields the nearest server in terms of hop count, not in terms of the OSPF metric.

Router	*Upstream node*	*Downstream interfaces (interface:hops)*
RT2	Net N3	(N2:1)

Table 2: Sample forwarding cache entries, for source N4 and destination Group A.

A MOSPF router must clear its entire **forwarding cache** when the Autonomous System's topology changes, because all the datagram shortest-path trees must be rebuilt. Likewise, when the location of a multicast group's membership changes (reflected by a change in group-membership-LSAs), all cache entries associated with the particular multicast destination group must be cleared. Other than these two cases, forwarding cache entries need not ever be deleted or otherwise modified; in particular, the forwarding cache entries do not have to be aged. However, forwarding cache entries can be freely deleted after some period of inactivity (i.e., garbage collected), if router memory resources are in short supply.

3 Inter-area multicasting

Up to this point this memo has discussed multicast forwarding when the entire Autonomous System is a single OSPF area. The logic for when the multicast datagram's source and its destination group members belong to the same OSPF area is the same. This section explains the behavior of the MOSPF protocol when the datagram's source and (at least some of) its destination group members belong to different OSPF areas. This situation is called **inter-area multicast**.

Inter-area multicast brings up the following issues, which are resolved in succeeding sections:

- Are the **group-membership-LSAs** specific to a specific area? And if they are, how is group membership information conveyed from one area to the next?
- How are the **datagram shortest-path trees** built in the inter-area case, since complete information concerning the topology of the datagram source's neighborhood is not available to routers in other areas?
- In an area border router, **multiple datagram shortest-path trees** are built, one for each attached area. How are these separate datagram shortest-path trees combined into a **single forwarding cache entry**?

It should be noted in the following that the basic protocol mechanisms in the inter-area case are the same as for the intra-area case. Forwarding of multicasts is still defined by the contents of the **forwarding cache**. The forwarding cache is still built from the same two components: the **local group database** and the **datagram shortest-path trees**. And while the calculation of the datagram shortest-path trees is different in the inter-area case (see Section 3.2), the local group database is built exactly the same as in the intra-area case (i.e., MOSPF's interface with IGMP remains unchanged in the presence of areas). Finally, the forwarding algorithm described in Section 11 is the same for both the intra-area and inter-area cases.

The following discussion uses the area configuration pictured in Figure 4 as an example. This figure, taken from the OSPF specification, shows an Autonomous System split into three areas (Area 1, Area 2 and Area 3). A single backbone area has been configured (everything outside of the shading). Since the backbone area must be contiguous, a single virtual link has been configured between the area border routers RT10 and RT11. Additionally, an area address range has been configured in Router RT11 so that Networks N9-N11 and Host H1 will be reported as a single route outside of Area 3 (via summary-LSAs).

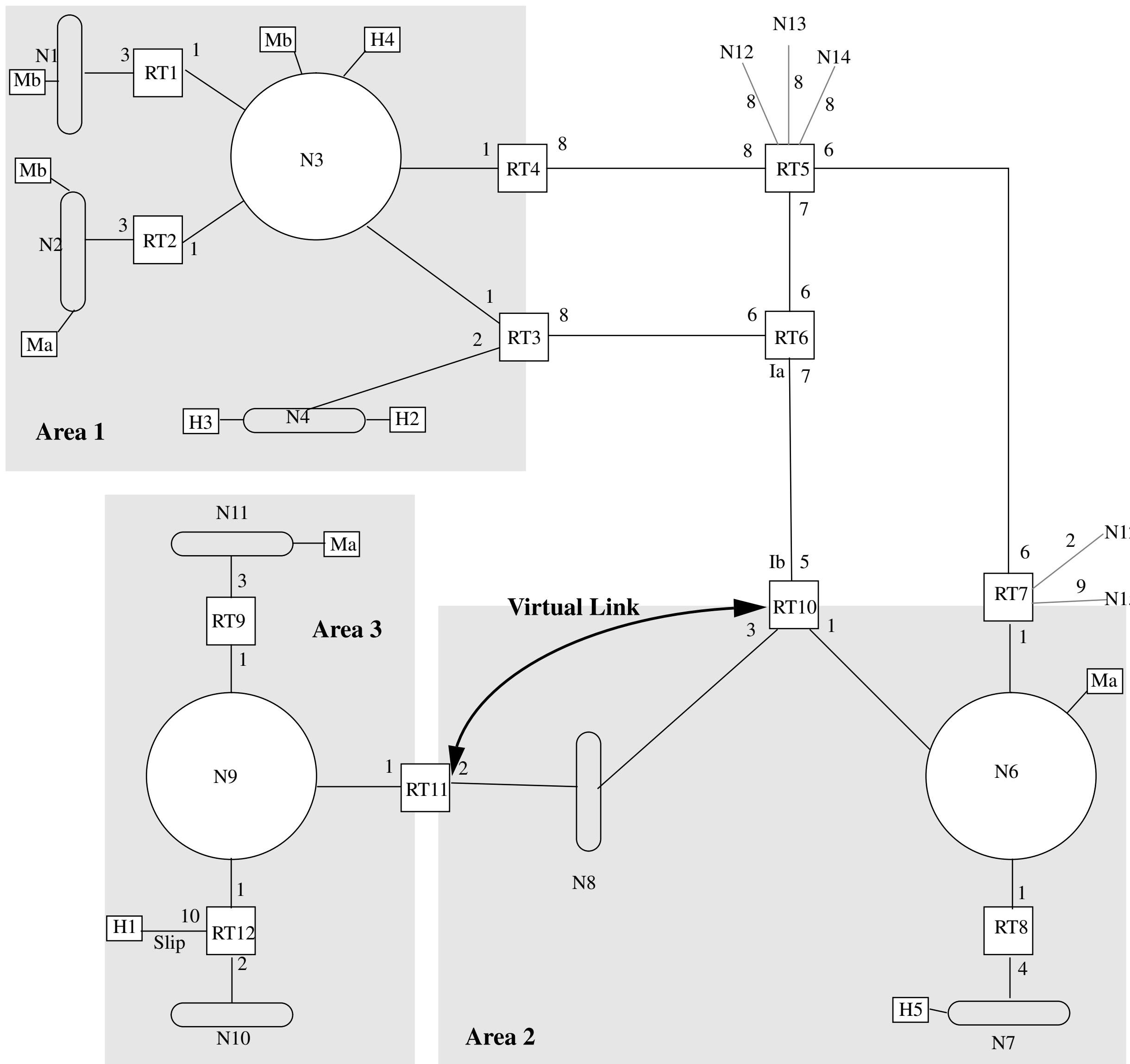


Figure 4: A sample MOSPF area configuration

3.1 Extent of group-membership-LSAs

Group-membership-LSAs are specific to a single OSPF area. This means that, just as with OSPF router-LSAs, network-LSAs and summary-link-LSAs, a group-membership-LSA is flooded throughout a single area only[9]. A router

9. This means that in MOSPF, just as in OSPF, the only kind of link state advertisement that can be flooded between areas is the external-link-LSA.

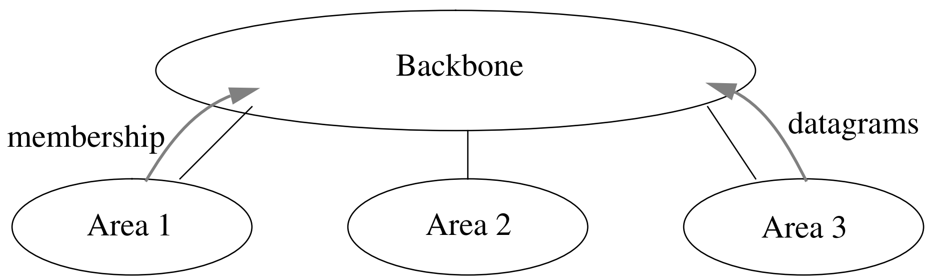


Figure 5: Inter-area routing architecture

attached to multiple areas (i.e., an area border router) may end up originating several group-membership-LSAs concerning a single multicast destination, one for each attached area. However, as we will see below, the contents of these group-membership-LSAs will vary depending on their associated areas.

Just as in OSPF, each MOSPF area has its own link state database. The MOSPF database is simply the OSPF link state database enhanced by the group-membership-LSAs. Consider again the area configuration pictured in Figure 4. The result of adding the group-membership-LSAs to the area databases yields the databases pictured in Figures 6 and 7. Figure 6 shows Area 1's MOSPF database. Figure 7 shows the backbone's MOSPF database. Superscripts indicate which transit vertices have been advertised as requesting particular multicast destinations. A superscript of "w" indicates that the router is advertising itself as a wild-card multicast receiver (see below). The dashed lines are OSPF summary-link-LSAs or external-link-LSAs.

Suppose an OSPF router has a local group database entry for *[Group Y, network X]*. The router then originates a group-membership-LSA for Group Y into the area containing network X. For example, in the area configuration pictured in Figure 4, Router RT1 originates a group-membership-LSA for Group B. This group-membership-LSA is flooded throughout Area 1, and no further. Likewise, assuming that Router RT3 has been elected Designated Router for network N3, RT3 originates a group-membership-LSA into Area 1 listing the transit network N3 as having group members. Note that in the link state database for Area 1 (Figure 6) both Router RT1 and network N3 have accordingly been labelled with Group B.

In OSPF, the area border routers forward routing information and data traffic between areas. In MOSPF. a subset of the area border routers, called the **inter-area multicast forwarders**, forward group membership information and multicast datagrams between areas. Whether a given OSPF area border router is also a MOSPF inter-area multicast forwarder is configuration dependent (see Section B.1). In Figure 4 we assume that all area border routers are also inter-area multicast forwarders.

In order to convey group membership information between areas, inter-area multicast forwarders "summarize" their attached areas' group membership to the backbone. This is very similar functionality to the summary-link-LSAs that are generated in the base OSPF protocol. An inter-area multicast forwarder calculates which groups have members in its attached non-backbone areas. Then, for each of these groups, the inter-area multicast forwarder injects a group-membership-LSA into the backbone area. For example, in Figure 4 there are two groups having members in Area 1: Group A and Group B. For that reason, both of Area 1's inter-area multicast forwarders (Routers RT3 and RT4) inject group-membership-LSAs for these two groups into the backbone. As a result both of these routers are labelled with Groups A and B in the backbone link state database (see Figure 7).

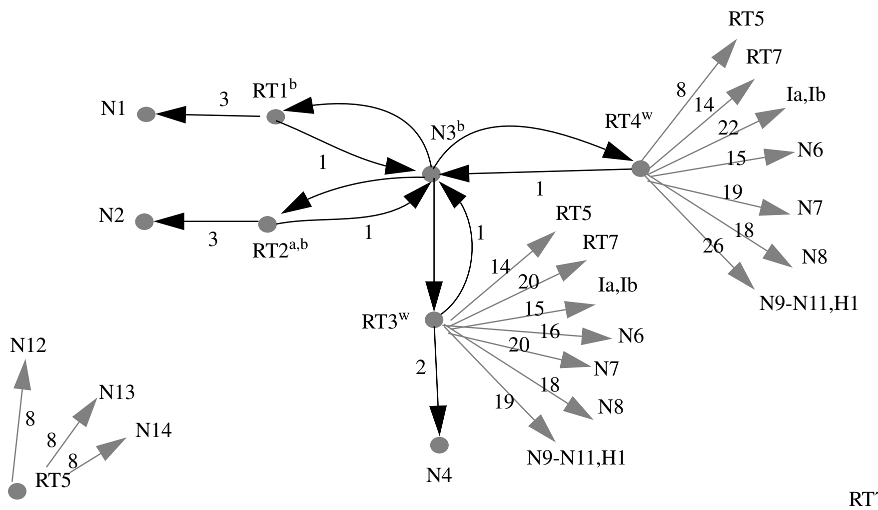


Figure 6: Area 1's MOSPF database

However, unlike the summarization of unicast destinations in the base OSPF protocol, the summarization of group membership in MOSPF is asymmetric. While a non-backbone area's group membership is summarized to the backbone, this information is not then readvertised into other non-backbone areas. Nor is the backbone's group membership summarized for the non-backbone areas. Going back to the example in Figure 4, while the presence of Area 3's group (Group A) is advertised to the backbone, this information is not then redistributed to Area 1. In other words, routers internal to Area 1 have no idea of Area 3's group membership.

At this point, if no extra functionality was added to MOSPF, multicast traffic originating in Area 1 destined for Multicast Group A would never be forwarded to those group A members in Area 3. To accomplish this, the notion of **wild-card receivers** is introduced. A wild-card receiver is a router to which all multicast traffic, regardless of multicast destination, should be forwarded. A router's wild-card multicast reception status is per-area. In non-backbone areas, all inter-area multicast forwarders[10] are wild-card multicast receivers. This ensures that all multicast traffic originating in a non-backbone area will be forwarded to its inter-area multicast forwarders, and hence to the backbone area. Since the backbone has complete knowledge of all areas' group membership, the datagram can then be forwarded to all group members. Note that in the backbone itself there is no need for wild-card multicast receivers[11]. As an example, note that Routers RT3 and RT4 are wild-card multicast receivers in Area 1 (see Figure 6), while there are none in the backbone (see Figure 7).

This yields the inter-area routing architecture pictured in Figure 5. All group membership is advertised by the non-backbone areas into the backbone. Likewise, all IP multicast traffic arising in the non-backbone areas is forwarded to the backbone. Since at this point group membership information meets the multicast datagram traffic, delivery of the datagrams becomes possible.

---

10. A router indicates that it is an inter-area multicast forwarder by setting the appropriate flag in its router-LSA. See Section 14.6 for details.

11. This is not quite true. As we shall see, any inter-AS multicast forwarders belonging to the backbone are designated as wild-card multicast receivers. See Section 4.

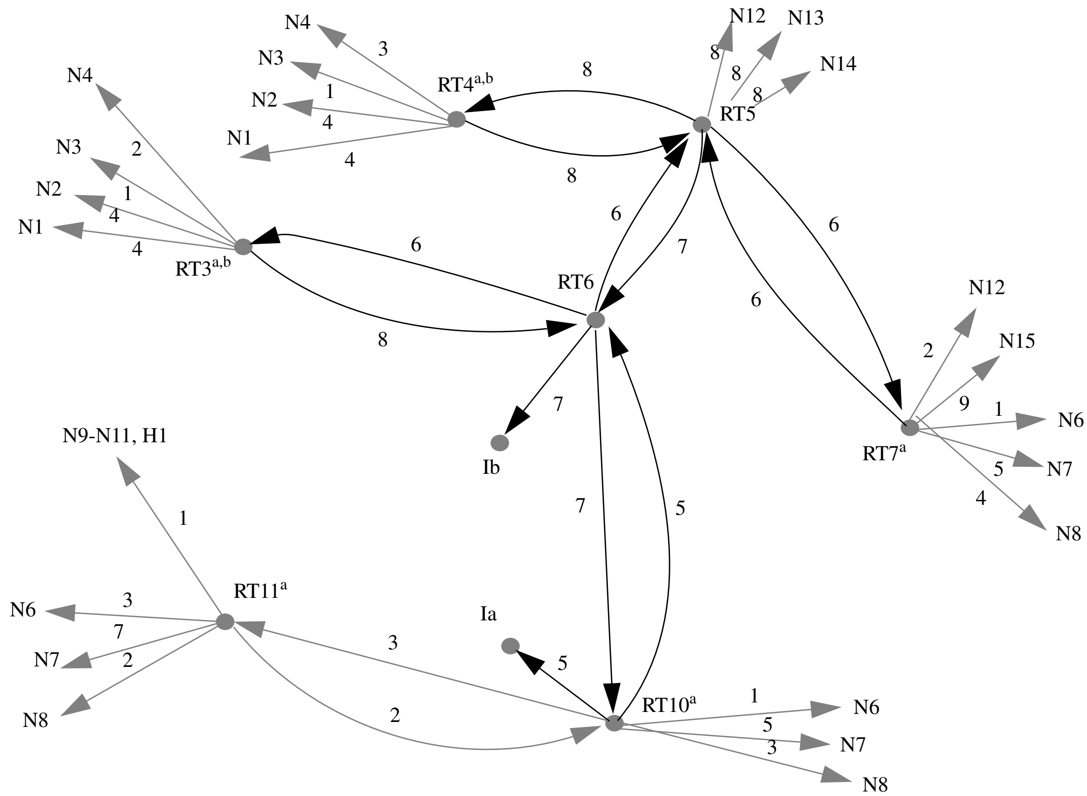


Figure 7: The backbone's MOSPF database

3.2 Building inter-area datagram shortest-path trees

When building datagram shortest-path trees in the presence of areas, it is often the case that the source of the datagram and (at least some of) the destination group members are in separate areas. Since detailed topological information concerning one OSPF area is not distributed to other OSPF areas (the flooding of router-LSAs, network-LSAs and group-membership-LSAs is restricted to a single OSPF area only), the building of complete datagram shortest-path trees is often impossible in the inter-area case. To compensate, approximations are made through the use of wild-card multicast receivers and OSPF summary-LSAs.

When it first receives a datagram for a particular [source net, destination group] pair, a router calculates a separate datagram shortest-path tree for each of the router's attached areas. Each datagram shortest-path tree is built solely from LSAs belonging to the particular area's link state database. Suppose that a router is calculating a datagram shortest-path tree for Area A. It is useful then to separate out two cases.

The first case, **Case 1: The source of the datagram belongs to Area A** has already been described in Section 2.3.2. However, in the presence of OSPF areas, during tree pruning care must be taken so that the branches leading to other areas remain, since it is unknown whether there are group members in these (remote) areas. For this reason, only those branches having no group members nor wild-card receivers are pruned when producing the datagram shortest-path tree.

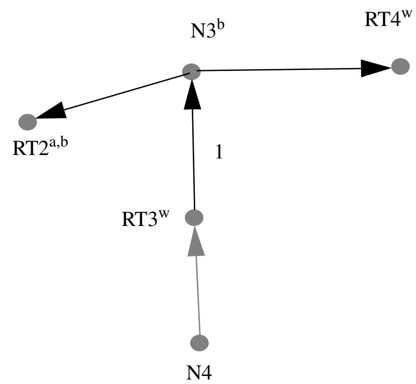


Figure 8: Datagram's shortest-path tree: Area 1, source N4, destination Group A

As an example, suppose in Figure 4 that Host H2 sends a multicast datagram to destination Group A. Then the datagram's shortest-path tree for Area 1, built identically by all routers in Area 1 receiving the datagram, is shown in Figure 8. Note that both inter-area multicast forwarders (RT3 and RT4) are on the datagram's shortest-path tree, ensuring the delivery of the datagram to the backbone and to Areas 2 and 3.

- **Case 2: The source of the datagram belongs to an area other than Area A.** In this case, when building the datagram shortest-path tree for Area A, the immediate neighborhood of the datagram's source is unknown. However, there are summary-LSAs in the Area A link state database indicating the cost of the paths between each of Area A's inter-area multicast forwarders and the datagram source. These summary links are used to approximate the neighborhood of the datagram's source; the tree begins with links directly connecting the source to each of the inter-area multicast forwarders. These links point in the reverse direction (towards instead of away from the datagram source) from the links considered in Case 1 above. **All additional links** added to the tree also **point in the reverse direction**. The final datagram shortest-path tree is then produced by, as before, pruning all branches having no group-members nor wild-card receivers.

  As an example, suppose again that Host H2 in Figure 4 sends a multicast datagram to destination Group A. The datagram's shortest-path tree for the backbone is shown in Figure 9. The neighborhood around the source (network N4) has been approximated by the summary links advertised by routers RT3 and RT4. Note that all links in Figure 9's datagram shortest-path tree have arrows pointing in the reverse direction, towards network N4 instead of away from it.

The reverse costs used for the entire tree in Case 2 are forced because summary-LSAs only specify the cost towards the datagram source. In the presence of asymmetric link costs, this may lead to less efficient routes when forwarding multicasts between areas

Those routers attached to multiple areas must calculate multiple trees and then merge them into a single forwarding cache entry. As shown in Section 2.3.2, when connected to a single area the router's position on the datagram shortest-path tree determines (in large part) its forwarding cache entry. When attached to multiple areas, and hence calculating multiple datagram shortest-path trees, each tree contributes to the forwarding cache entry's list of downstream interfaces/neighbors. However, only one of the areas' datagram shortest-path trees will determine the forwarding cache entry's upstream node. When one of the attached areas contains the datagram source, that area will determine the upstream node. Otherwise, the tiebreaking rules of Section 12.2.7 are invoked.

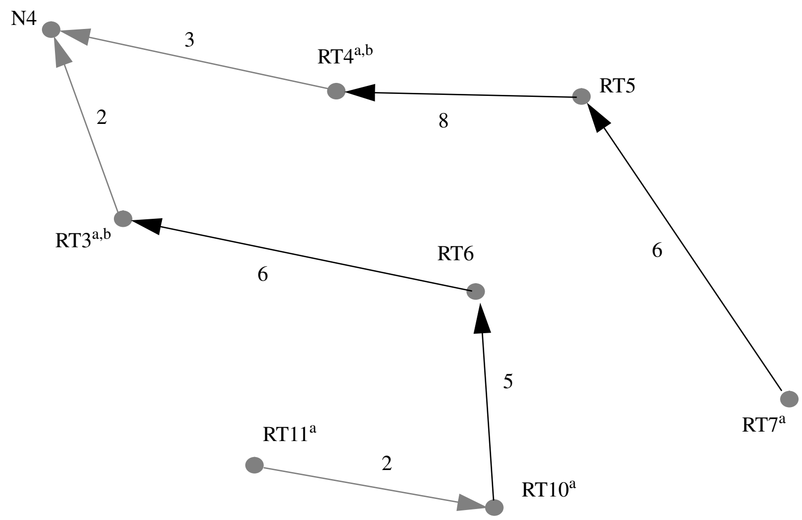


Figure 9: Datagram shortest-path tree: Backbone, source N4, Destination Group A

Consider again the example of Host H2 in Figure 4 sending a multicast datagram to destination Group A. Router RT3 will calculate two datagram shortest-path trees, one for Area 1 and one for the backbone. Since the source of the datagram (Host H2) belongs to Area 1, the Area 1 datagram shortest-path tree determines RT3's upstream node: Network N4. Router RT3 calculates two downstream interfaces for the datagram: the interface to network N3 (which comes from Area 1's datagram shortest-path tree) and the serial line to router RT6 (which comes from the backbone's datagram shortest-path tree). As for Router RT10, it calculates two trees, determining its upstream node from the backbone tree and its two downstream interfaces from the Area 2 tree. Finally, Router RT11 calculates three trees, determining its upstream node from the Area 2 tree and its downstream interface from the Area 3 tree.

4 Inter-AS multicasting

This section explains how MOSPF deals with the forwarding of multicast datagrams between Autonomous Systems. Certain AS boundary routers in a MOSPF system will be configured as **inter-AS multicast forwarders.** It is assumed that these routers will also be running an inter-AS multicast routing protocol. This specification does not dictate the operation of such an inter-AS multicast routing protocol. However, the following interactions between MOSPF and the inter-AS routing protocol are assumed:

1. MOSPF guarantees that the inter-AS multicast forwarders will receive all multicast datagrams; but it is up to each router so designated to determine whether the datagram should be forwarded to other Autonomous Systems. This determination will probably be made via the inter-AS routing protocol.

2. MOSPF assumes that the inter-AS routing protocol is forwarding multicast datagrams in an RPF (reverse path forwarding; see [Deering] for an explanation of this terminology) fashion. In other words, it is assumed that a multicast datagram whose source (call it X) lies outside the MOSPF domain will enter the MOSPF domain at those points that are advertising (into OSPF) the best routes back to X. MOSPF calculates the path of the datagram through the MOSPF domain based on this assumption.

MOSPF designates an inter-AS multicast forwarder as a **wild-card multicast receiver** in all of its attached areas. As in the inter-area case, this ensures that the routers remain on all pruned shortest-path trees and thereby receive all multicast datagrams, regardless of destination.

The presence of inter-AS multicast forwarders is reflected in the contents of the link state database[12]. Suppose that in Figure 1, both RT5 and RT7 were configured as inter-AS multicast forwarders. Then the link state database would look like the one pictured in Figure 2, with the addition of a) wild-card status for RT5 and RT7 (they would appear with superscripts of "w") and b) the external links originated by RT5 and RT7 being labelled as **multicast-capable**[13].

Consider instead the area configuration in Figure 4. Again suppose RT5 and RT7 are configured as inter-AS multicast forwarders. Then in Area 1's link state database (Figure 6), the external links originated by RT5 and RT7 would again be labelled as multicast-capable. However, note that in Area 1's database RT5 and RT7 are not labelled as wild-card multicast receivers. This is unnecessary; since Area 1's inter-area multicast forwarders (RT3 and RT4) are wild-cards, all multicast datagrams will be forwarded to the backbone. And in the backbone's link state database (Figure 7), RT5 and RT7 will be labelled as wild-cards.

4.1 Building inter-AS datagram shortest-path trees.

When multicast datagrams are to be forwarded between Autonomous Systems, the datagram shortest-path tree is built as follows. Remember that the router builds a separate tree for each area to which it is attached; these trees are then merged into a single forwarding cache entry. Suppose that the router is building the tree for Area A. We break up the tree building into three cases. This first two cases have already been described earlier in this memo: Case 1 (the source of the datagram belongs to Area A) having been described in Section 2.3.2 and Case 2 (the source of the datagram belongs to another OSPF area) having been described in Section 3.2. The only modification to these cases is that inter-AS multicast forwarders, as well as group members and inter-area multicast forwarders, must remain on the pruned trees. The new case is as follows:

- **Case 3: The source of the datagram belongs to another Autonomous System.** The immediate neighborhood of the source is then unknown. In this case the multicast-capable AS external links are used to approximate the neighborhood of the source; the tree begins with links directly attaching the source to one or more inter-AS multicast forwarders. The approximating AS external links point in the reverse direction (i.e., towards the source), just as with the approximating summary links in Case 2. Also, as in Case 2, **all links** included in the tree must **point in the reverse direction.** The final datagram shortest-path tree is then produced (as always) by pruning those branches having no group members nor wild-card receivers.

  As an example, suppose that a host on Network N12 (see Figure 4) originates a multicast datagram for Destination Group B. Assume that all external costs pictured are OSPF external type 1 metrics. Then any routers in Area 1 receiving the datagram would build the datagram shortest-path tree pictured in Figure 10. Note that all links in the tree point in the reverse direction, towards the source. The tree indicates that the routers expect the datagram to enter the Autonomous System at router RT7, and then to enter the area at router RT4.

  Note that in those cases where the "best" inter-AS boundary router is not attached to the area, the neighborhood of the source is actually approximated by the concatenation of a summary link and a multicast-capable AS external link. This is in fact the case in Figure 10.

---

12. As with inter-area multicast forwarders, inter-AS multicast forwarders indicate that status by setting a flag in their router-LSA. See Section 14.6 for details.

13. It is possible that through the operation of an inter-AS multicast routing protocol, router RT7 knows that it does not have connectivity to network N15 (even though it has unicast connectivity). In this case, RT7 would not advertise the external link to N15 as being multicast capable.

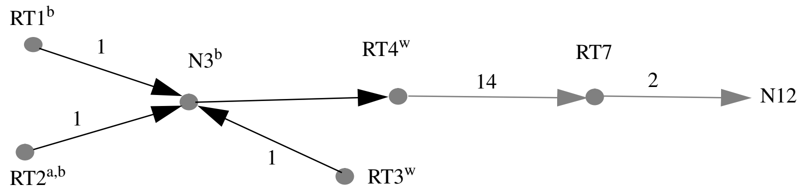


Figure 10: Datagram shortest path tree: Area 1, source N12, Destination Group B

In Case 3 (datagram source in another AS) the requirement that all tree links point in the reverse direction (towards the source) accommodates the fact that summary links and AS externals already point in the reverse direction. This also leads to the requirement that the inter-AS multicast routing protocol operate in a reverse path forwarding fashion (see condition 2 of Section 4). Note that Reverse path forwarding can lead to sub-optimal routing when costs are configured asymmetrically. And it can even lead to non-delivery of multicast datagrams in the case of asymmetric reachability.

Inter-AS multicast forwarders may end up calculating a forwarding cache entry's **upstream node** as being external to the AS. As an example, Router RT7 in Figure 10 will end up calculating an external router (via its external link to network N12) as the upstream node for the datagram. This means that RT7 will receive the datagram from a router in another AS before injecting the datagram into the MOSPF system.

4.2 Stub area behavior

AS external links are not imported into stub areas. Suppose that the source of a particular datagram lies outside of the Autonomous System, and that the datagram is forwarded into a stub area. In the stub area's datagram shortest-path tree the neighborhood of the datagram's source cannot be approximated by AS external links. Instead the neighborhood of the source is approximated by the **default summary links** (see Section 3.6 of [OSPF]) that are originated by the stub area's intra-area multicast forwarders.

Except for this small change to the construction of a stub area's datagram shortest-path trees, all other MOSPF algorithms (e.g., merging with other areas' datagram shortest-path trees to form the forwarding cache) function the same for stub areas as they do for non-stub areas.

5 Modelling internal group membership

A MOSPF router may itself contain multicast applications. A typical example of this is a UNIX workstation that doubles as a multicast router. This section concerns two alternative ways of representing the group membership of the MOSPF router's internal applications. Both representations have advantages. For maximum flexibility, the MOSPF forwarding algorithm (see Section 11) has been specified so that either representation can be used in a MOSPF router (and in fact, both representations can be used at once, depending on the application).

The first representation is based on the paradigm presented in RFC 1112. In this case, an application joins a multicast group on one or more specific physical interfaces. The application then receives a multicast datagram if and only if it is received on one of the specified interfaces. If a datagram is received on multiple specified interfaces, the application receives multiple copies. Figure 11 shows this algorithm as it is implemented in (modified) BSD UNIX kernels. The figure shows the processing of a multicast datagram, starting with its reception on a particular interface. First copies of the datagram are given to those applications that have joined on the receiving interface. Then the forwarding decision (pictured as a box containing a question mark) is made, and the packet is (possibly) forwarded out certain interfaces. If these interfaces are not capable of receiving their own multicasts, a copy of the datagram must be internally looped back to appropriately joined applications.

The advantages to the RFC 1112 representation are as follows:

- It is the standard for the way an IP host joins multicast groups. It is simplest to use the same membership model for hosts and routers; most would consider an IP router to be a special case of an IP host anyway.
- It is the way group membership has been implemented in BSD UNIX. Existing multicast applications are written to join multicast groups on specific interfaces.
- The possibility of receiving multiple datagram copies may improve fault tolerance. If the datagram is dropped due to an error on the path to some interface, another interface may still receive a copy.
- The ability to specify a particular receiving interface may improve the accuracy of IP multicast's TTL scoping mechanism (see Section 2.3.4).
- Membership in the non-routable multicast groups (224.0.0.1 - 224.0.0.255) must be on a per-interface basis. An OSPF router always belongs to 224.0.0.5 (AllSPFRouters) on its OSPF interfaces, and may belong to 224.0.0.6 (AllDRouters) on one or more of its OSPF interfaces.

The second representation is MOSPF-specific. In this case, an application joins a multicast group on an interface-independent basis. In other words, group membership is associated with the router as a whole, not separately on each

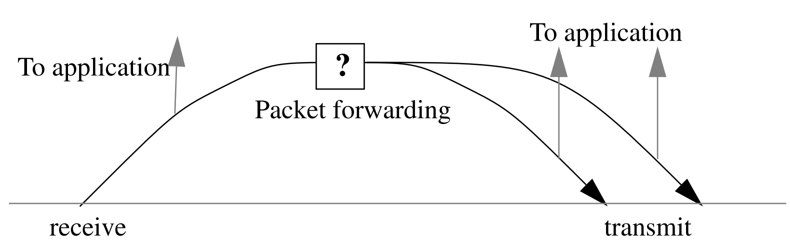


Figure 11: RFC 1112 representation of internal group membership

interface. The application then receives a copy of a multicast datagram if and only if the datagram would actually be forwarded by the MOSPF router. Figure 12 shows how this algorithm would be implemented. The datagram is

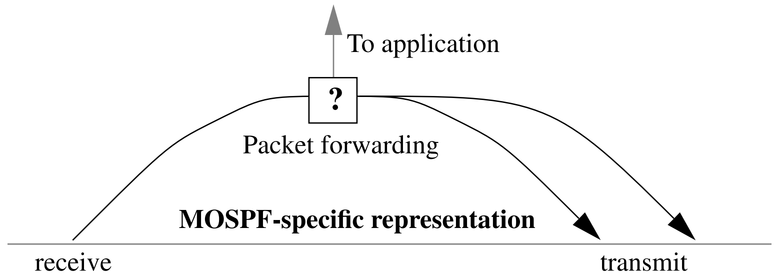


Figure 12: MOSPF-specific representation of internal group membership

received on a particular interface. If the datagram is validated for forwarding (i.e., the receiving interface connects to the matching forwarding cache entry's **upstream node**), a copy of the datagram is also given to appropriately joined applications. Note that this model of group membership is not as general as the RFC 1112 model, in that it can only be implemented in MOSPF routers and not in arbitrary IP hosts. However, it has the following advantages:

- The application does not need to have knowledge of the router interfaces. It does not need to know what kind or how many interfaces there are; this will be taken care of by the MOSPF protocol itself.
- As long as any interface is operational, the application will continue to receive multicast datagrams. This happens automatically, without the application modifying its group membership.
- The application receives only one copy of the datagram. Using the RFC1112 representation, whenever an application joins on more than one interface (which must be done if the application does not want to rely on a single interface), multiple datagram copies will be received during normal operation.

6 Additional capabilities

This section describes the MOSPF configuration options that allow routers of differing capabilities to be mixed together in the same routing domain. Note that these options handle special circumstances that may not be encountered in normal operation. Default values for the configuration settings are specified in Appendix B.

6.1 Mixing with non-multicast routers

MOSPF routers can be mixed freely with routers that are running only the base OSPF algorithm (called non-multicast routers in the following). This allows MOSPF to be deployed in a piecemeal fashion, thereby speeding deployment and allowing experimentation with multicast routing on a limited scale.

When a MOSPF router builds a datagram shortest-path tree, it omits all non-multicast routers. For example, in Figure 1, if router RT6 was not a multicast router, the datagram shortest-path tree in Figure 3 would be built with a more circuitous branch through router RT5, instead of through router RT6. In addition, non-multicast routers do not participate in the flooding of the new group-membership-LSAs. This adheres to the general principle that a router should not have to handle those link state advertisements whose format (or contents) the router does not understand.

Mixing MOSPF routers with non-multicast routers creates a number of potential problems. Certain mixings of MOSPF and non-multicast routers can cause multicast datagrams to take suboptimal paths, or in other cases can lead to the non-delivery of multicast datagrams. In addition, mixing MOSPF routers and non-multicast routers can cause the paths of multicast datagrams to diverge radically from the path of unicast datagrams. Such divergences can make routing problems harder to debug.

In particular, the following specific difficulties may arise when mixing MOSPF routers with non-multicast routers:

- Even though there is unicast connectivity to a destination, there may not be multicast connectivity. For example, if Router RT10 in Figure 1 becomes a non-multicast router, the group member connected to network N11 will no longer be able to receive multicasts sourced by Host H2. But the two hosts will be able to exchange unicasts (e.g., ICMP pings).
- When the Designated Router for a multi-access network is a non-multicast router, the network will not be used for forwarding multicast datagrams. For example, if in Figure 1 Router RT4 is Designated Router for network N3, and RT4 is non-multicast, network N3 will not be used to forward IP multicasts. This would mean that multicast datagrams originated by Hosts H2 and H3 would not be forwarded beyond their local network (N4), even though it seems that the needed multicast connectivity exists.
- When forwarding multicast datagrams between areas, mixing of MOSPF routers and non-multicast routers in the source area may cause unexpected loss of multicast connectivity. This is because in the inter-area routing of multicast datagrams the neighborhood of the datagram's source is approximated by OSPF summary link advertisements, and OSPF summary LSAs do not carry indications/guarantees of the summarized path's multicast routing capability.

6.2 TOS-based multicast

MOSPF allows a separate datagram shortest-path tree to be built for each IP Type of Service. This means that the path of a multicast datagram can vary depending on the datagram's TOS classification, as well as its source and destination.

For each router interface, OSPF allows a separate metric to be configured for each IP TOS. When building the shortest path tree for TOS X, the cost of a path is the sum of the component interfaces' TOS X metrics. Note that OSPF requires that a TOS 0 metric be specified for each interface. However, as a form of data compression, metrics need only be specified for non-zero TOS if they are different than the TOS 0 metric.

Additionally, OSPF routers can be configured to ignore TOS when forwarding packets. Such routers, called TOS-incapable, build only the TOS 0 portion of the routing table. TOS-incapable routers can be mixed freely with TOS-capable routers when forwarding unicast packets. The way this is handled for unicast packets is that the unicast is forwarded along the TOS 0 route whenever the TOS X route does not exist. However, MOSPF must treat this situation somewhat differently, since each router must build the exact same tree rooted at the datagram's source.

Like OSPF, MOSPF allows TOS-based routing to be optional. TOS-capable and TOS-incapable multicast routers can be mixed freely in the routing domain. TOS-incapable routers will only ever build TOS 0 datagram shortest-path trees. TOS-capable routers will first build TOS 0 datagram shortest-path trees. If these trees contain only TOS-capable routers, datagram shortest-path trees are then built separately for non-zero TOS values. Otherwise, the TOS 0 datagram shortest-path tree is used to forward all traffic, regardless of its TOS designation. Using this logic, all routers in essence continue to utilize identical datagram shortest-path trees. See Section 12.2.8 for more details.

6.3 Assigning multiple IP networks to a physical network

Assigning multiple IP networks/subnets to a single physical network causes some confusion in MOSPF. This is because the underlying OSPF protocol treats these IP networks/subnets as entirely separate entities, originating separate network-LSAs for each and forming separate adjacencies for each, while IGMP recognizes only the single underlying physical network. Adding to the problem is the fact that when a multicast datagram is received from such a multiply-addressed physical wire, there is no good way to choose the datagram's **upstream node** (which must be done in order to make the forwarding decision; see Section 11 for details). As a result, unless this situation is dealt with through configuration, unwanted replication of multicast datagrams may occur when they are forwarded over multiply-addressed wires.

As a remedy, MOSPF allows multicast forwarding to be disabled on certain IP networks/subnets. When multicast forwarding is disabled on the wire's "extra" subnets (i.e., all but one), the extra subnets will not appear in datagram shortest-path trees, nor will they appear in local group database or forwarding cache entries. As a result, the possibility of unwanted datagram replication is eliminated. The actual disabling of multicast forwarding on a subnet is done through setting the **IPMulticastForwarding** parameter to **disabled** on all router interfaces connecting to the subnet (see Section B.2).

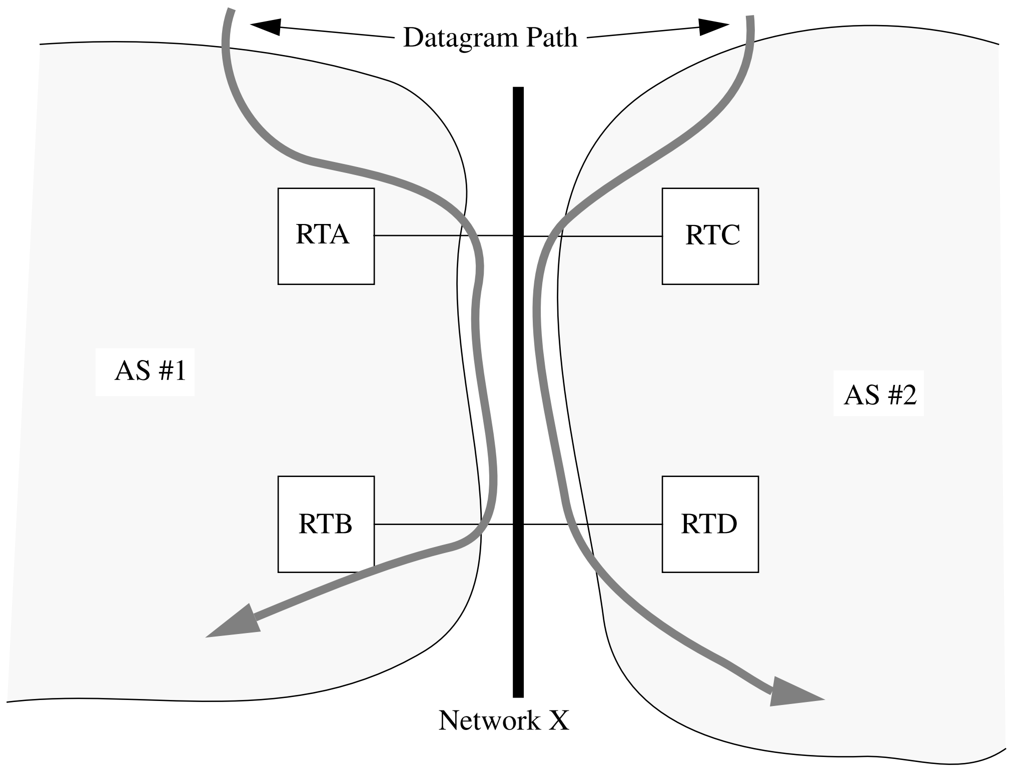
6.4 Networks on Autonomous System boundaries

Another complication can arise on IP networks/subnets that lie on the boundary of a MOSPF Autonomous System. Similar to the unicast situation where these networks may be running multiple IGPs (Interior Gateway Protocols), these networks may also be running multiple multicast routing protocols. It may then become impossible for a MOSPF router to determine whether a multicast datagram is being forwarded along the datagram shortest-path tree, or whether it has been inadvertently received from the other Autonomous System. Guessing wrong can lead to either unwanted replication or non-delivery of the multicast datagram. In addition, in order to prevent receiving duplicate multicast datagrams, group members on these boundary networks will probably want to declare their membership to one Autonomous System and not another.

For example, consider the two Autonomous Systems pictured in Figure 13. Network X is on the boundary of both ASes. One possible multicast datagram path is shown; the datagram originates in a third Autonomous System, and is then delivered to both AS #1 and AS #2 separately. The paths through the two Autonomous Systems may end up having certain boundary networks as common segments. In Figure 13, Network X is common to both paths. In this case, if both Autonomous Systems were running (separate copies of) MOSPF, the same datagram would appear twice on Network X as a data-link multicast. This would cause duplicate datagrams to be received by any group members on Network X or downstream from Network X.

MOSPF has two mechanisms to eliminate this replication of multicast datagrams. First, a system administrator can configure certain networks to forward multicast datagrams as data-link unicasts instead of data-link multicasts. This is done by setting the **IPMulticastForwarding** parameter to **data-link unicast** on those router interfaces attaching to the network (see Section B.2). As an example, in Figure 13 the routers in AS #2 could be configured so that Router C would send the multicast datagram out onto Network X as a data-link unicast addressed directly to Router D. Router D would accept this data-link unicast, but would reject any data-link multicast forwarded by Router A. This would eliminate replication of multicast datagrams downstream from Network X. In addition, if the IPMulticastForwarding parameter is set to data-link unicast on Network X, group membership will not be monitored on the network. This will prevent group members attached directly to Network X from receiving multiple datagram copies, since group membership on the boundary network will be monitored from only one AS.

Figure 13: Networks on AS boundaries



It should be noted that forwarding IP multicasts as data-link unicasts has some disadvantages when three or more MOSPF routers are attached to the network. First of all, it is more work for a router to send multiple unicasts than a single multicast. Second, the multiple unicasts consume more network bandwidth than a single multicast. And last, it increases the delay for some group members since multiple unicasts also take longer to send than a single multicast.

6.5 Recommended system configuration

In order make MOSPF's selection of routes more predictable, it is recommended that all routers in any particular OSPF area have the same multicast and TOS capabilities.Keeping areas homogeneous ensures that IP multicast packets will follow relatively the same path as IP unicasts. In contrast, while heterogeneous areas will function, and will probably be necessary at least during the initial introduction of multicast routing, such areas may produce seemingly sub-optimal and unexpected routes. For example, see Section 6.1above for a detailed description of the possible pitfalls when mixing multicast and non-multicast routers.

As for the other options presented above, to achieve the most predictable results it is recommended that a router interface's **IPMulticastForwarding** parameter be set to a value other than **data-link multicast** only when either a) multiple IP networks have been assigned to a single physical wire or b) multiple multicast routing protocols are running on the attached network.

7 Basic implementation requirements

An implementation of MOSPF requires the following pieces of system support. Note that this support is in addition to that required for the base OSPF implementation as outlined in Section 4.4 of [OSPF].

- **Promiscuous multicast reception.** In a multicast router, it is necessary to receive all IP multicasts at the data-link level. On those interfaces where IP multicast datagrams are encapsulated by a wide range of data-link multicast destination addresses (e.g, ethernet and FDDI), this is most easily accomplished by disabling any hardware filtering of multicast destinations (i.e., by "opening up" the interface's multicast filter).
- **Data-link multicast/broadcast detection.** To avoid unwanted replication of multicast datagrams in certain exceptional conditions, it is necessary for the MOSPF forwarding mechanism to determine whether a datagram was received as a data-link multicast/broadcast or as a data-link unicast. See Section 6.4 for more details.
- **An implementation of IGMP.** MOSPF uses the Internet Group Management Protocol (IGMP, documented in [RFC 1112]) to monitor multicast group membership. See Section 9 for details.

8 Protocol data structures

The MOSPF protocol is described herein in terms of its operation on various protocol data structures. These data structures are included for explanatory uses only, and are not intended to constrain a MOSPF implementation. Besides the data structures listed below, this specification will also reference the various data structures (e.g., OSPF interfaces and neighbors) defined in [OSPF].

In a MOSPF router, the following items are added to the list of global OSPF data structures described in Section 5 of [OSPF]:

- **Local group database.** This database describes the group membership on all attached networks for which the router is either Designated Router or Backup Designated Router. This in turn determines the group-membership-LSAs that the router will originate, and the local delivery of multicast datagrams (see Sections 2.3.1 and 10).
- **Forwarding cache.** Each entry in the forwarding cache describes the path of a multicast datagram having a particular [source net, multicast destination, TOS] combination. These cache entries are calculated when building the datagram shortest-path trees. See Sections 2.3.4 and 11 for more details.
- **Multicast routing capability.** Indicates whether the router is running the multicast extensions defined in this memo. A router running the multicast extensions must still run the base OSPF algorithm as set forth in [OSPF]. Such a router will continue to interoperate with non-multicast-capable OSPF routers when forwarding IP unicast traffic.
- **Inter-area multicast forwarder.** Indicates whether the router will forward IP multicasts from one OSPF area to another. Such a router declares itself a wild-card multicast receiver in its router-LSA (see Section 14.6), and also summarizes its attached areas' group membership to the backbone in group-membership-LSAs. When building inter-area datagram shortest-path trees, it is these routers that appear immediately adjacent to the datagram source at the root of the tree (see Section 3.2). Not all multicast-capable area border routers need be configured as inter-area multicast forwarders. However, whenever both ends of a virtual link are multicast-capable, they must both be configured as inter-area multicast forwarders (see Section 14.11).

- **Inter-AS multicast forwarder.** Indicates whether the router will forward IP multicasts between Autonomous Systems. Such a router declares itself a wild-card multicast receiver in its router-LSA (see Section 14.6). These routers are also assumed to be running some kind of inter-AS multicast protocol. They mark all external routes that they import into the OSPF domain as to whether they provide multicast connectivity (see Section 14.9). When building inter-AS multicast datagram trees, it is these routers that appear immediately adjacent to the datagram source at the root of the tree.

8.1 Additions to the OSPF area structure

The OSPF area data structure is described in Section 6 of [OSPF]. In a MOSPF router, the following item is added to the OSPF area structure:

- **List of group-membership-LSAs.** These link state advertisements describe the location of the area's multicast group members. Group-membership-LSAs are flooded throughout a single area only. Area border routers also summarize their attached areas' membership by originating group-membership-LSAs into the backbone area. For more information, see Sections 3.1 and 10.

8.2 Additions to the OSPF interface structure

The OSPF interface structure is described in Section 9 of [OSPF]. In a MOSPF router, the following items are added to the OSPF interface structure. Note that the IPMulticastForwarding parameter is really a description of the attached network. As such, it should be configured identically on all routers attached to a common network; otherwise incorrect routing of multicast datagrams may result[14].

- **IPMulticastForwarding.** This configurable parameter indicates whether IP multicasts should be forwarded over the attached network, and if so, how the forwarding should be done. The parameter can assume one of three possible values: **disabled**, **data-link multicast** and **data-link unicast**. When set to disabled, IP multicast datagrams will not be forwarded out the interface. When set to data-link multicast, IP multicast datagrams will be forwarded as data-link multicasts. When set to data-link unicast, IP multicast datagrams will be forwarded as data-link unicasts. The default value for this parameter is data-link multicast. The other two settings are for use in the special circumstances described in Sections 6.3 and 6.4. When set to disabled or to data-link unicast, IGMP group membership is not monitored on the attached network.
- **IGMPPollingInterval.** When the router is actively monitoring group membership on the attached network, it periodically sends IGMP Host Membership Queries. IGMPPollingInterval is a configurable parameter indicating the number of seconds between IGMP Host Membership Queries. The router actively monitors group membership on the attached network when both a) the interface's IPMulticastForwarding parameter is set to data-link multicast and b) the router has been elected Designated Router on the attached network. See Section 9 for details.
- **IGMPTimeout.** This configurable parameter indicates the length of time (in seconds) that a local group database entry associated with this interface will persist without another matching IGMP Host Membership Report being received. See Section 9 for details.
- **IGMP polling timer.** The firing of this interval timer causes an IGMP Host Membership Query to be sent out the interface. The length of this timer is the configurable parameter IGMPPollingInterval. See Section 9 for details.

14. Synchronization of the IPMulticastForwarding interface parameter is not enforced by the MOSPF protocol, since it is not included in the contents of a MOSPF router's Hello packets.