Network Working Group J. Moy
Internet Draft Proteon, Inc.
Expiration date: January 1994 July 1993

# 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 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 forwarding extent 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 solely 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.

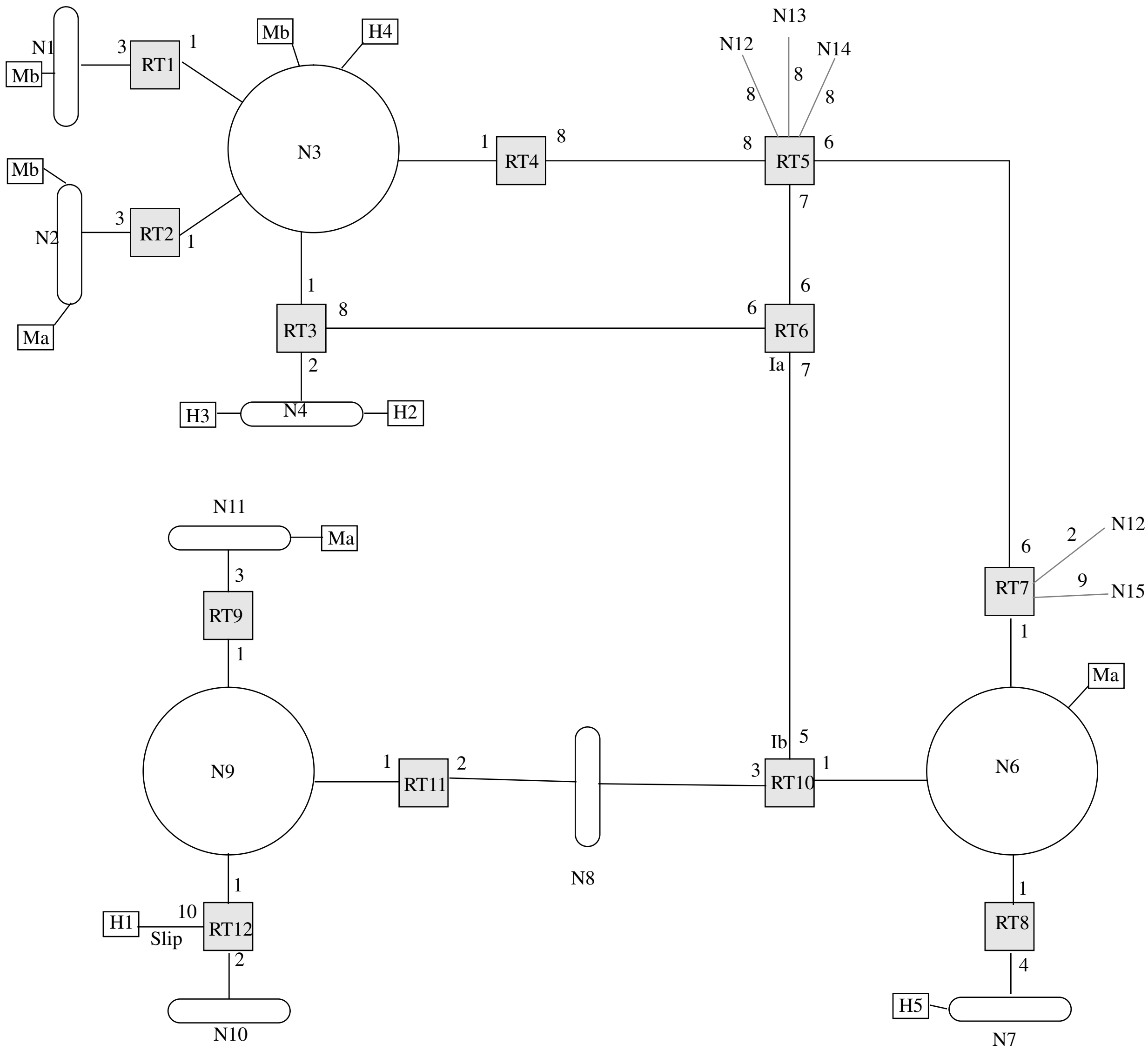



**Figure 1: A sample MOSPF configuration**

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 (Network 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 Network 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 (except for its IP TTL) 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.

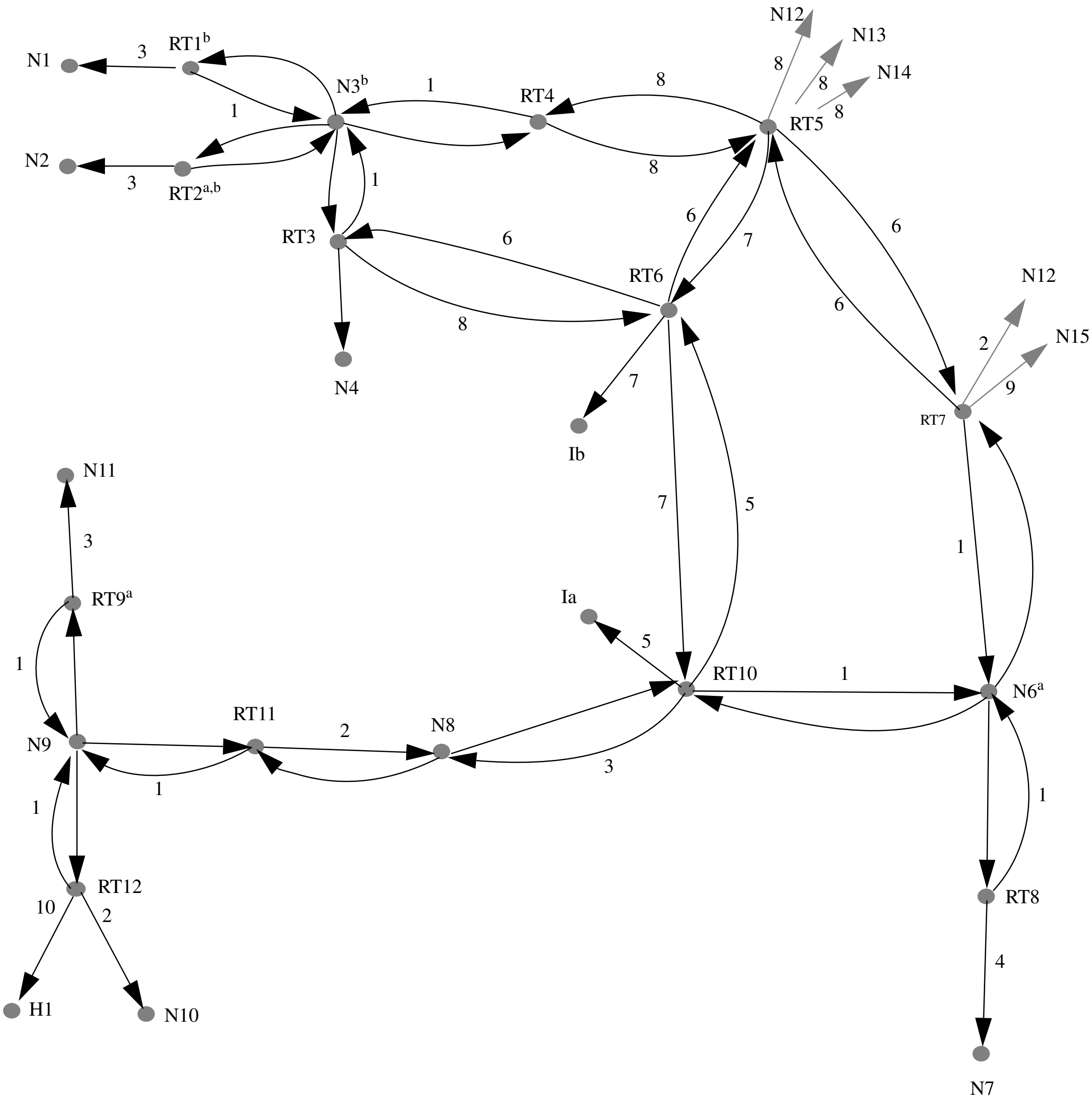



**Figure 2: The MOSPF database**

### 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 Management Protocol (IGMP; see [RFC 1112]). When a 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 router's **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 its 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.

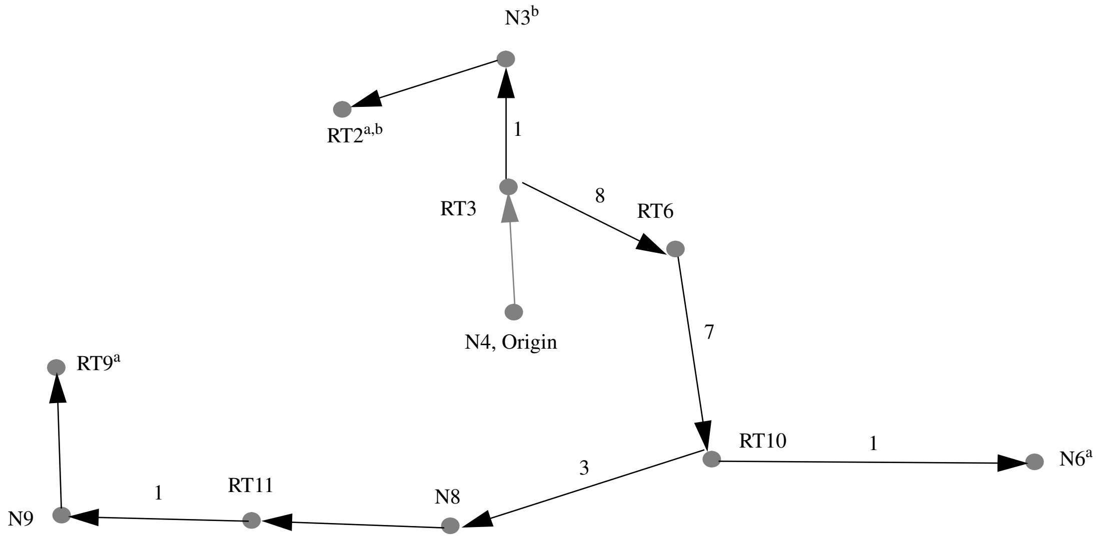



**Figure 3: Sample datagram's shortest-path tree, source N4, destination Group A**

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 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 RTX) position in the datagram's pruned shortest-path tree consists of 1) RTX's parent in the tree (this will be the forwarding cache entry's **upstream node)** and 2) the list of RTX'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. Router RT10's **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 forwarding extent 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 a 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 A. In that case, some routers will not even attempt to build a forwarding cache entry (e.g, router RT5) 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, RT7, 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), (N8:2) |
| RT11 | Net N8 | (N9:1) |
| 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, but not necessarily 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 single 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-link-LSAs).

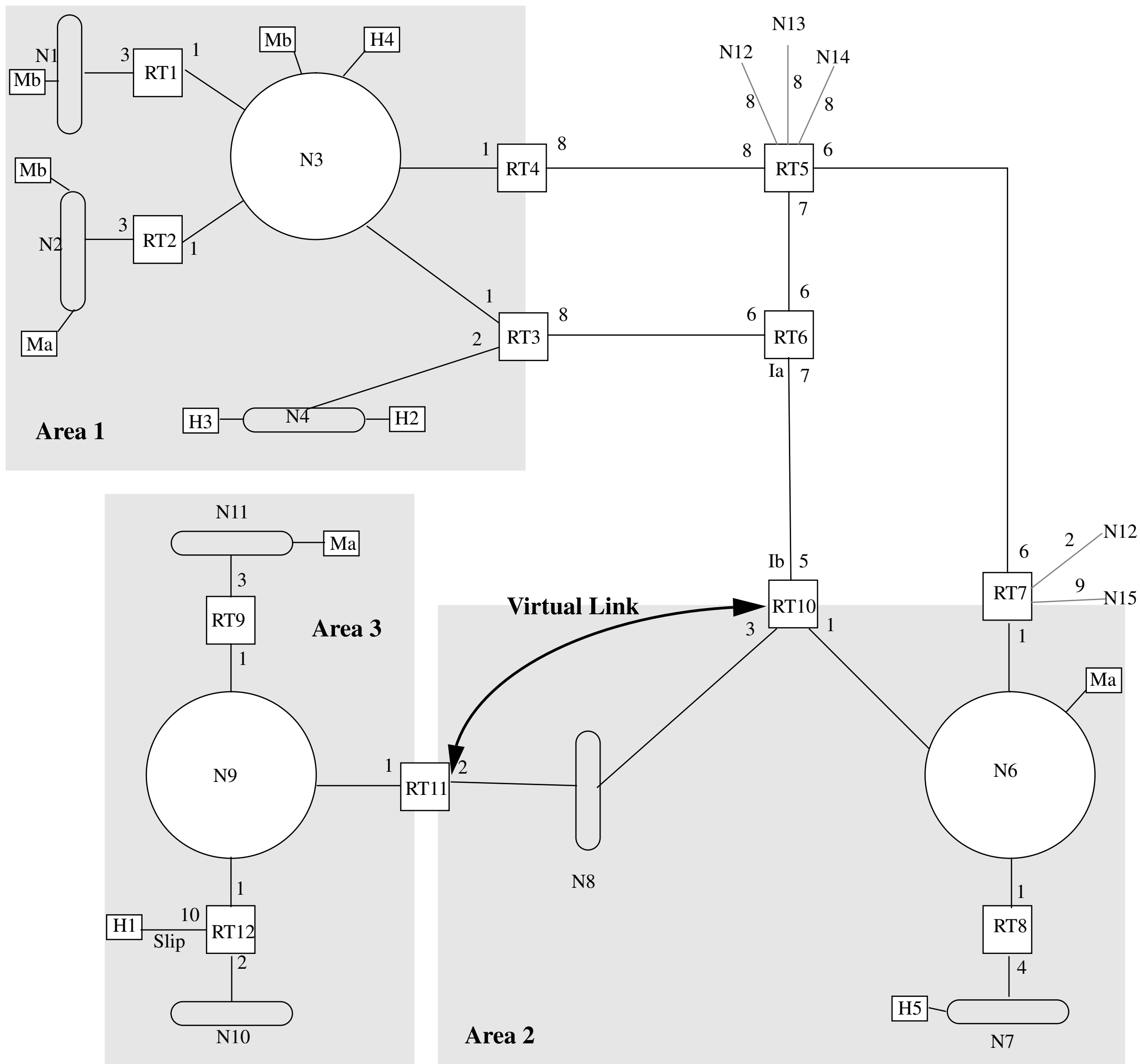



**Figure 4: A sample MOSPF area configuration**

## 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 AS external-link-LSA.

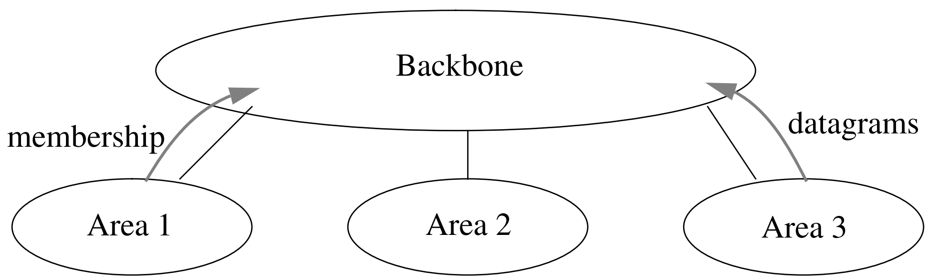



**Figure 5: Inter-area routing architecture**

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 AS external-link-LSAs. Note in Figure 7 that Router RT11 has condensed its routes to Networks N9-N11 and Host H1 into a single summary-link-LSA.

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

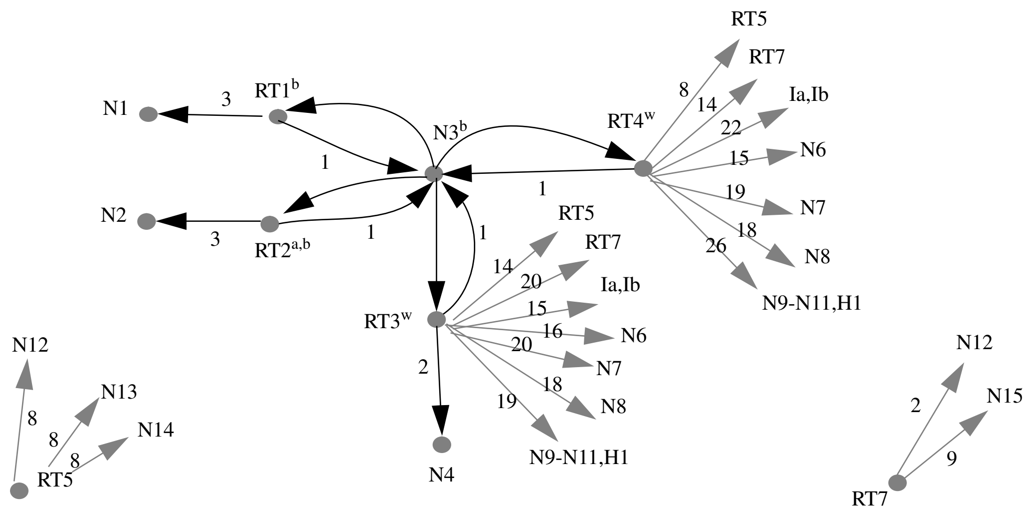



**Figure 6: Area 1's MOSPF database**

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).

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 multicast receivers** is introduced. A wild-card multicast 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

---

10. A router indicates that it is a wild-card multicast receiver 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.

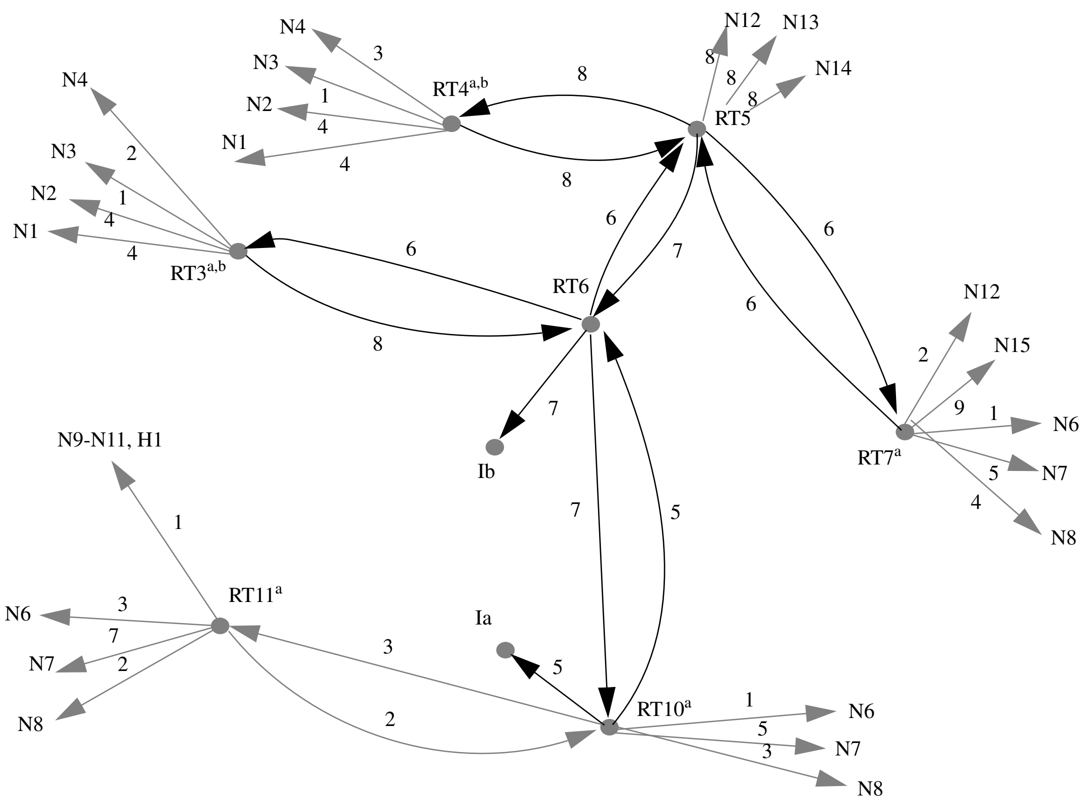



**Figure 7: The backbone's MOSPF database**

the backbone. Since at this point group membership information meets the multicast datagram traffic, delivery of the multicast datagrams becomes possible.

## 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 wildcard multicast receivers and OSPF summary-link-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.

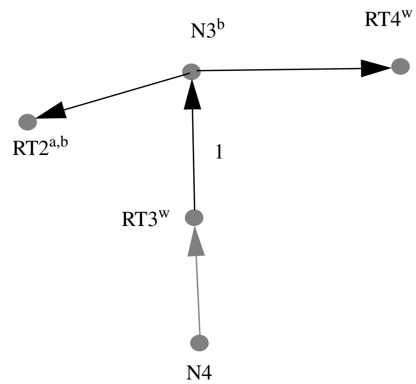



**Figure 8: Datagram's shortest-path tree: Area 1, source N4, destination Group A**

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 multicast receivers are pruned when producing the datagram shortest-path tree.

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 that receive 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 from there 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-link-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 multicast 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-link-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 short-

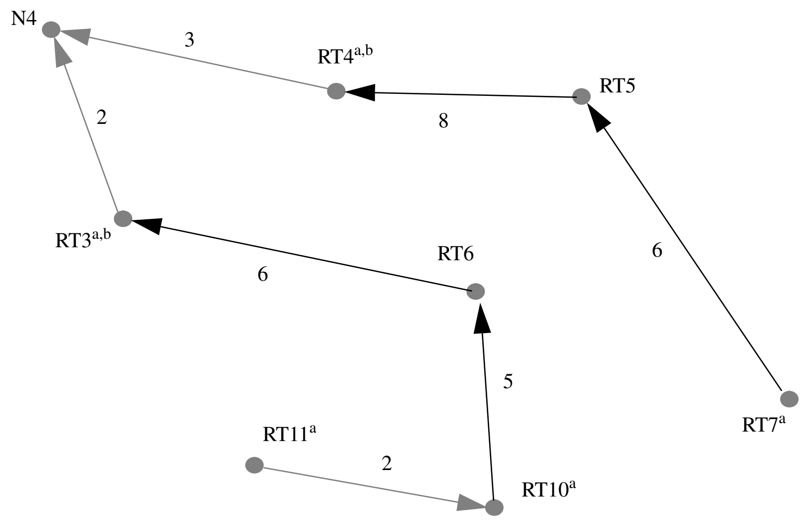



**Figure 9: Datagram shortest-path tree: Backbone, source N4, Destination Group A**

est-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.

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.

As an example, 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**[12].

As another example, consider 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 multicast receivers.

12. It is possible that through the operation of an inter-AS multicast routing protocol, Router RT7 knows that it does not have multicast 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.

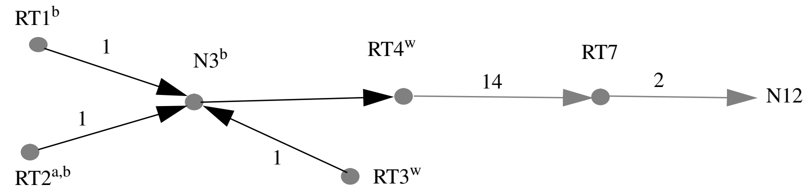



**Figure 10: Datagram shortest path tree: Area 1, source N12, Destination Group B**

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 multicast forwarder is not directly 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.

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 external links 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 must 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.

## 4.3 Inter-AS multicasting in a core Autonomous System

It may be the case that the MOSPF routing domain connects together many different Autonomous Systems, thereby serving as a "core Autonomous System" (e.g, the old NSFNet backbone). In this case, it could very well be that the majority of the MOSPF routers are also inter-AS multicast forwarders. Having each inter-AS multicast forwarder then declare itself a wild-card multicast receiver could very well waste considerable network bandwidth. However, as an alternative to declaring themselves wild-card multicast receivers, the inter-AS multicast routers could instead explicitly advertise all groups that they were interested in forwarding (to other "client" Autonomous Systems) in group-membership-LSAs. These advertised groups would have to be learned through an inter-AS multicast routing protocol (or possibly even statically configured).

This in essence allows the clients of the core Autonomous System to advertise their group membership into the core. However, since any client MOSPF domains will still have their inter-AS multicast forwarders configured as wild-card multicast receivers, this advertisement will be asymmetric: the core will not advertise its or others' group membership to the clients. The achieves the same inter-AS multicast routing architecture that MOSPF uses for inter-area multicast routing (see Figure 5).

# 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 expanding ring search 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

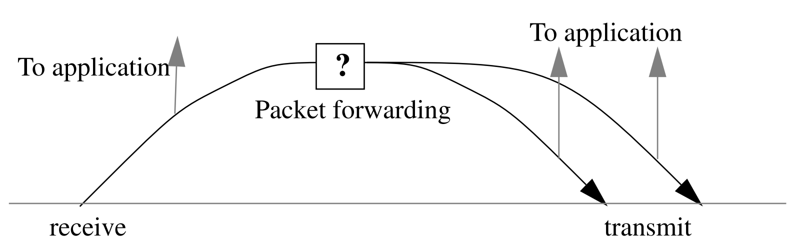



**Figure 11: RFC 1112 representation of internal group membership**

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

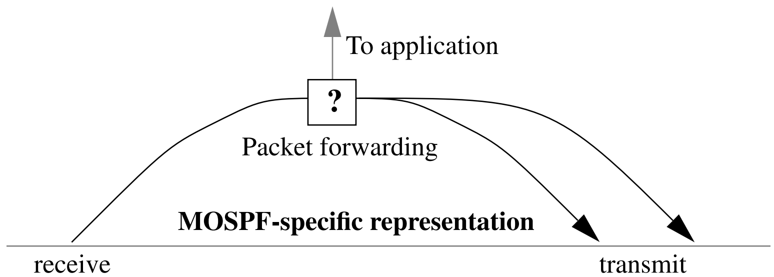



**Figure 12: MOSPF-specific representation of internal group membership**

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 links, and OSPF summary-link-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).

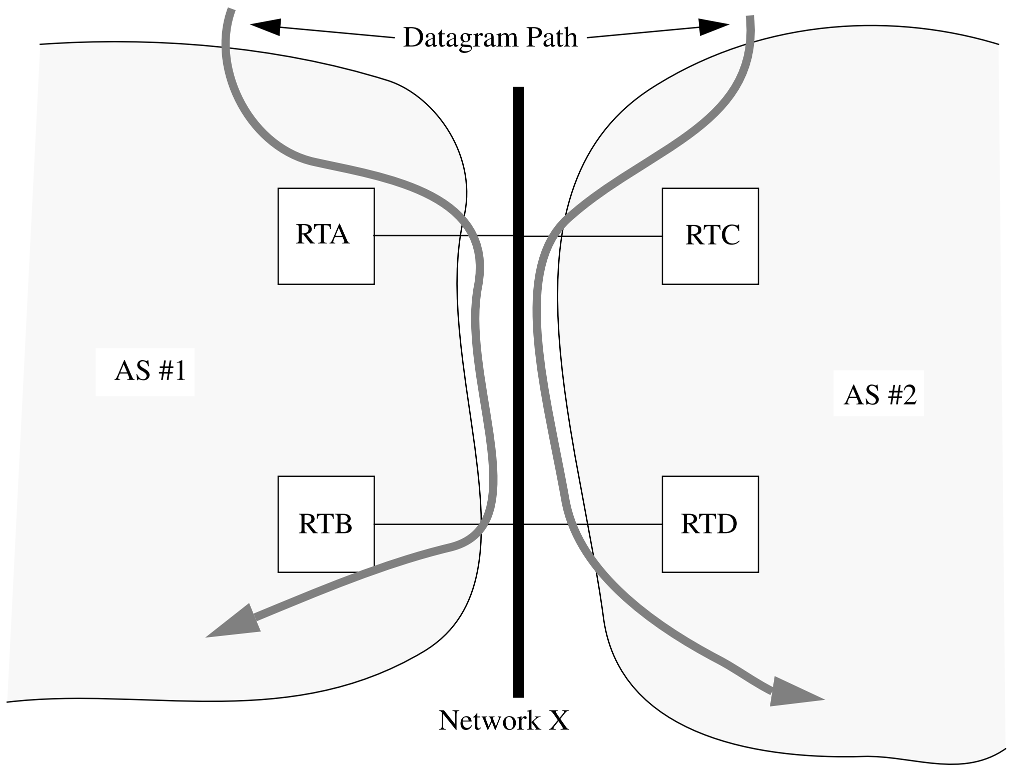
## 6.4 Networks on Autonomous System boundaries

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 (AS #1 in our example).

**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 to 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 6.1 above 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 multicast router to determine whether a datagram was received as a data-link multicast/broadcast or as a data-link unicast, for later use by the MOSPF forwarding mechanism. 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 non-backbone area router-LSAs (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-LSAs (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[13].

- **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.

---

13. 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.

## 8.3 Additions to the OSPF neighbor structure

The OSPF neighbor structure is defined in Section 10 of [OSPF]. In a MOSPF router, the following items are added to the OSPF neighbor structure:

- **Neighbor Options.** This field was already defined in the OSPF specification. However, in MOSPF there is a new option which indicates the neighbor's multicast capability. This new option is learned in the Database Exchange process through reception of the neighbor's Database Description packets, and determines whether group-membership-LSAs are flooded to the neighbor. See the items concerning flooding in Section 14 for a more detailed explanation.

## 8.4 The local group database

The local group database has already been introduced in Section 2.3.1. The current section attempts a more precise definition. The local group database tracks the group membership of the router's directly attached networks. Database entries are created and maintained by the IGMP protocol. Database entries can cause group-membership-LSAs to be originated, which in turn enable the pruning of datagram shortest-path trees. The local group database also dictates the router's responsibility for the delivery of multicast datagrams to directly attached group members.

Each entry in the local group database has three components: the multicast group, the attached network and the entry's age. A database entry is indexed by the first two components: multicast group and attached network. A database lookup function is assumed to exist, so that given a [multicast group, attached network] pair, the matching database entry (if any) can be discovered. A database entry for [Group A, Network N1] exists if and only if there are Group A members currently located on Network N1.

The three components of a local group database entry are defined as follows:

- **MulticastGroup**. The multicast group whose members are being tracked by this entry. Each multicast group is represented as a class D IP address. For the semantics of multicast group membership, see [RFC 1112].
- **AttachedNetwork.** Each database entry is concerned with the group members belonging to a single attached network. To get a complete picture of the local group membership (when for example building a group-membership-LSA), it may be necessary to consult multiple database entries, one for each attached network. Note that a router is only required to maintain entries for those attached networks on which the router has been elected Designated Router or Backup Designated Router (see Section 9).
- **Age.** Indicates the number of seconds since an IGMP Host Membership Report for multicast Group A has been seen on Network N1. If the age field hits Network N1's configured **IGMPTimeout** value, the local group database entry is removed (i.e., the entry has “aged out”). See Sections 9.2 and 9.3 for more information.

## 8.5 The forwarding cache

The forwarding cache has already been defined in Section 2.3. The current section attempts a more precise definition. Each entry in the forwarding cache indicates how a multicast datagram having a particular [source network, destination multicast group, IP TOS] will be forwarded. A forwarding cache entry is built on demand from the local group database and the datagram's shortest-path tree. For more details, consult Sections 2.3.4 and 12.

Each entry in the forwarding cache has six components: the multicast datagram's source network, the destination multicast group, the IP TOS, the upstream node, the list of downstream interfaces and (possibly) a list of downstream neighbors. A forwarding cache entry is indexed by source network, destination multicast group and IP TOS. A lookup function is assumed to exist, so that given a multicast datagram with a particular [IP source, destination multicast group, IP TOS], a matching cache entry (if any) can be found.

The six components of a forwarding cache entry are defined as follows:

- **Source network**. The datagram's source network is described by a network/subnet/supernet number and its corresponding mask. The source network for a datagram is discovered via a routing table/database lookup of the datagram's IP source address, as described in Section 11.2.
- **Destination multicast group**. The destination group to which matching datagrams are being forwarded. For the semantics of multicast group membership, see [RFC 1112].
- **IP TOS.** The IP Type of Service specified by matching datagrams. Note that this means that the path of the multicast datagram depends on its TOS classification.
- **Upstream node.** The attached network/neighboring router from which the datagram must be received. If received from a different attached network/neighboring router, the matching datagram is dropped instead of forwarded. This prevents unwanted replication of multicast datagrams. It is possible that the upstream node is unspecified (i.e., set to NULL). In this case, matching datagrams will always be dropped, no matter where they are received from. It is also possible that the upstream node is specified as the placeholder EXTERNAL. This means that the datagram must be received on a non-MOSPF interface in order to be forwarded.
- **List of downstream interfaces.** These are the router interfaces that the matching datagram should be forwarded out of (assuming that the datagram was received from *upstream node*). Each interface is also listed with a **TTL** value. The TTL value is the minimum number of hops necessary to reach the closest (in terms of router hops) group member. This allows the router to drop datagrams that have no chance of reaching a destination group member.
- **List of downstream neighbors.** When the datagram is to be forwarded out a non-broadcast multi-access network, or if the interface's IPMulticastForwarding parameter is set to data-link unicast, the datagram must be forwarded separately to each downstream neighbor (see Sections 2.3.3 and 6.4). As done for downstream interfaces, each downstream neighbor is specified together with the smallest **TTL** that will actually reach a group member.

# 9 Interaction with the IGMP protocol

MOSPF uses the IGMP protocol (see [RFC 1112]) to monitor multicast group membership. In short, the Designated Router on a network periodically sends IGMP Host Membership Queries (see Section 9.1), which in turn elicit IGMP Host Membership Reports from the network's multicast group members. These Host Membership Reports are then recorded in the Designated Router's and Backup Designated Router's **local group databases** (see Section 9.2).

## 9.1 Sending IGMP Host Membership Queries

Only the network's Designated Router sends Host Membership Queries. This minimizes the amount of group membership information on the network, both in terms of queries and responses.

When a MOSPF router becomes Designated Router on a network, it checks to see that the network's **IPMulticast-Forwarding** parameter is set to data-link multicast (see Section B.2). If so, it starts the interface's **IGMP polling timer**. Then, whenever the timer fires (every **IGMPPollingInterval** seconds), the MOSPF router sends a Host Membership Query out the interface. The destination of the query is the IP address 224.0.0.1. For the format of the query, see [RFC 1112]. If/when the MOSPF router ceases to be the network's Designated Router, the IGMP polling timer is disabled and no more Hosts Membership Queries are sent.

Unusual behavior can result when multiple IP networks are assigned to a single physical network. MOSPF treats each such IP network separately, electing (possibly) a different Designated Router for each network. However, IGMP operates on a physical network basis only: when a Host Membership Query is sent, all group members on the physical network respond, regardless of their IP addresses. So unless the **IPMulticastForwarding** parameter is set to a value other than data-link multicast on all but one of the physical network's IP networks, excess multicast membership reporting will result.

## 9.2 Receiving IGMP Host Membership Reports

Received Host Membership Reports are processed by both the network's Designated Router and Backup Designated Router. It is the Designated Router's responsibility to distribute the network's group membership information throughout the routing domain, by originating group-membership-LSAs (see Section 10). The Backup Designated Router processes Reports so that it too has a complete picture of the network's group membership, enabling a quick cutover upon Designated Router failure.

An IGMP Host Membership Report concerns membership in a single IP multicast group (call it Group A). The Report is sent to the Group A address so that other group members may see the Report and avoid sending duplicates (see [RFC 1112] for details). When an IGMP Host Membership Report, sent on Network N[14], is received by a MOSPF router, the following steps are executed:

1. If the router is neither the Designated Router nor the Backup Designated Router on the network, the Report is discarded and processing stops.
2. If the Report concerns a multicast group in the range 224.0.0.1 - 224.0.0.255, the Report is discarded and processing stops. This range of multicast groups are for local use (single hop) only, and datagrams sent to these destinations are never forwarded by multicast routers.
3. Locate the entry for [Group A, Network N] in the **local group database**. If no such entry exists, create one. In any case, set the age of the entry to 0. Note that even if multiple hosts attached to Network N report membership in the same group, only a single local group database entry will be formed. See Section 8.4 for more details concerning the local group database.
4. If the router is the network's Designated Router, and a local group database entry was created in the previous step, it may be necessary to originate a new group-membership-LSA. See Section 10 for details.

14. Actually, when multiple IP networks have been assigned to the same physical network, the first thing that needs to be done is to associate an IP network with the received Host Membership Report. This is done in the same way that a receiving interface is associated with a received multicast datagram; see Section 11.1.

### 9.3 Aging local group database entries

Every local database entry has an age field. Suppose that there is a database entry for [Group A, Network N1]. The age field then indicates the length of time (in seconds) since the last Host Membership Report for Group A was received on Network N1. If the age of the entry reaches Network N1's configured IGMPTimeout value (see Section B.2), the entry is considered invalid and is removed from the database.

Note that when a router, after having been either Network N1's Designated Router or Backup Designated Router, but now being neither, will (after IGMPTimeout seconds) automatically age out all of its local group database entries associated with Network N1. For this reason, it is not necessary to purge local group database entries on OSPF interface state changes.

### 9.4 Receiving IGMP Host Membership Queries

If a MOSPF router has internal multicast applications, and if the applications have bound themselves to certain interfaces (using the RFC 1112 representation described in Section 5), then the MOSPF router responds to received Host Membership Queries by issuing Host Membership Reports. Identical to the operation of any IP host supporting multicast applications, the exact procedure for issuing these Host Membership Reports is specified in [RFC 1112]. Note that in this case, if the router has been elected Designated Router on a network, it must receive its own Host Membership Reports and Host Membership Queries.

If instead all of its applications have joined groups in an interface-independent fashion (using the MOSPF-specific representation described in Section 5), the MOSPF router does not respond to Host Membership Queries. Instead, the MOSPF router communicates this membership information by originating appropriate group-membership-LSAs (see Section 10.1).

## 10 Group-membership-LSAs

Group-membership-LSAs provide the means of distributing membership information throughout the MOSPF routing domain. Group-membership-LSAs are specific to a single OSPF area (see Section 3.1). Each group-membership-LSA concerns a single multicast group. Essentially, the group-membership-LSA lists those networks which are directly connected to the LSA's originator and which contain one or more group members. For more details on how the group-membership-LSA augments the OSPF link state database, see Section 2.3.1.

The creation of group-membership-LSAs is discussed in Section 10.1. The format of the group-membership-LSA is described in Section A.3. A router will originate a group membership-LSA for multicast group A when one or more of the following conditions hold:

1. The router is Designated Router on a network (call it Network X), the interface to Network X has its IPMulticast-Forwarding parameter set to data-link multicast (see Section B.2), and Network X contains one or more members of Group A.

2. The router is an inter-area multicast forwarder (see Section B.1), and one or more of the router's attached non-backbone areas contain Group A members. In this case, the router will originate a group-membership-LSA for Group A into the backbone. This is the way group membership is conveyed between areas (see Section 3.1).

3. The router itself has applications that are requesting membership in Group A, in an interface-independent fashion (see Section 5).

As for all other types of OSPF link state advertisements (e.g, router-LSAs, network-LSAs, etc.), group-membership-LSAs are aged as they are held in a router's link state database. To prevent valid advertisements from "aging out", a router must refresh its self-originated group-membership-LSAs every LSRefreshTime interval, by incrementing their LS sequence numbers and reissuing them. In addition, when an event occurs that would alter one of the router's self-originated group-membership-LSAs, a new instance of the LSA is issued with an updated (i.e., incremented by 1) LS sequence number. Note however that a router is not allowed to originate two new instances of the same advertisement within MinLSInterval seconds. For that reason, occasionally advertisement originations will need to be deferred. Also, an event may occur that makes it inappropriate for the router to continue to originate a particular LSA. In that case, the router flushes the advertisement from the routing domain by "premature aging". For more information concerning the maintenance of LSAs, see Sections 12, 12.4, 14 and 14.1 of [OSPF].

When one of the following events occurs, it may be necessary for a router to (re)issue one or more group-membership-LSAs:

1. One of the router's interfaces changes state. For example, the router may have become Designated Router on a particular network, causing the router to start advertising the network's group membership to the rest of the MOSPF system in group-membership-LSAs.

2. The router receives an IGMP Host Membership Report, causing a new local group database entry to be formed (see Section 9.2).

3. One of the router's local group database entries "ages out", because it is no longer being refreshed by received IGMP Host Membership Reports (see Section 9.3).

4. The router is an inter-area multicast forwarder, and the group membership of one of the router's attached non-backbone areas changes. This is detected by the reception of a new, or the flushing of an old, group-membership-LSA into/from the non-backbone area's link state database.

5. The group membership of one of the router's internal applications changes.

## 10.1 Constructing group-membership-LSAs

This section details how to build a group-membership-LSA. The format of a group-membership-LSA is described in Section A.3. Each group-membership-LSA concerns a single multicast group. The body of the advertisement is a list of the local transit nodes (the router itself and directly attached transit networks) that contain group members. Section 10 listed the conditions requiring the (re)origination of a group-membership-LSA. Note that if the router is an area border router, it may be necessary to originate a separate group-membership-LSA for each attached area.

The following defines the contents of a group-membership-LSA, as originated by Router X into Area A. It is assumed that the group-membership-LSA is to report membership in multicast group G:

- The advertisement fields that are not type-specific (**LS age**, **LS sequence number**, **LS checksum** and **length**) are set according to Section 12.1 of [OSPF].

- The **Options** field of a group-membership-LSA is not processed on receipt. However, for consistency, the Option field in these advertisements should have its MC-bit set, T-bit clear, and the E-bit should match the configuration of Area A (i.e., set if and only if Area A is not a stub area). The rest of the Options field is set to 0.

- The **Link State ID** is set to the group whose membership is being reported (Group G).
- The Advertising Router is set to the OSPF Router ID of the router originating the advertisement (Router X).
- The body of the advertisement is a list of local transit vertices that should be labelled with Group G membership (see Section 2.3.1). This list may include the advertising router itself, and any of the transit networks that are directly attached to said router. The following steps determine which of these transit vertices are actually included in the group-membership-LSA. Note that any particular vertex should be listed at most once, even though the following may indicate multiple reasons for a particular vertex to be listed. Also note that if no transit vertices are listed by the advertisement, the advertisement should not be (re)originated; if an instance of the advertisement already exists, it should then be flushed from the link state database using the premature aging procedure specified in Section 14.1 of [OSPF].

  a. Consider those entries in the local group database that describe Group G membership (see Section 8.4). Consider each such entry in turn. Each entry references one of Router X's attached networks (call it Network N). If either Network N does not belong to Area A, or if Router X is not Network N's Designated Router[15], Network N should not be added to the group-membership-LSA, and the next local group database entry should be examined. Otherwise, if N is a stub network (e.g., Router X is the only OSPF router attached to N), Router X adds itself to the advertisement by adding a vertex with Vertex type set to 1 (router) and Vertex ID set to Router X's OSPF Router ID. Otherwise, N is a transit network. In this case, Network N should be added to the advertisement by adding a vertex with Vertex type set to 2 (network) and Vertex ID set to the IP address of Network N's Designated Router (i.e., Router X's IP interface address on Network N).

  b. If Router X itself has applications requesting Group G membership on an interface-independent basis (see Section 5), it should add itself to the advertisement by adding a vertex with Vertex type set to 1 (router) and Vertex ID set to Router X's OSPF Router ID.

  c. If Router X is an inter-area multicast forwarder (see Section 3.1), Area A is the backbone area (Area ID 0.0.0.0), and at least one of Router X's attached non-backbone areas has Group G members (indicated by the presence of one or more advertisements in the areas' link state databases having Link State ID set to Group G[16]), then Router X should add itself to the advertisement by adding a vertex with Vertex type set to 1 (router) and Vertex ID set to Router X's OSPF Router ID.

Consider as an example the network configuration in Figure 4. Suppose that Router RT2 has been elected Designated Router for Network N3. Router RT2 would then originate (into Area 1) the following group-membership-LSA for Group B:

```
 ; RT2's group-membership-LSA for Group B
LS age = 0                    ;always true on origination
Options = (E-bit|MC-bit)
LS type = 6                   ;group-membership-LSA
Link State ID = Group B
Advertising Router = RT2's Router ID
        Vertex type = 1       ;RT2 itself (for stub N2)
        Vertex ID = RT2's Router ID
```

---

15. For this reason when a transit network has both MOSPF routers and non-multicast OSPF routers attached, care should be taken to ensure that a MOSPF router is elected Designated Router. This can be accomplished through proper setting of the routers' configured Router Priority.

16. Note that just because these advertisements exist in the link state database, it does not mean that the Group G members are reachable. Reachability does not enter into the building of the transit vertex list, in order to simplify the calculation. This is a trade-off. As a result, some multicast datagrams may be forwarded further than necessary, when the described Group G members actually are unreachable.

```
Vertex type = 2        ;Network N3 (since RT2 is DR)
Vertex ID = RT2's IP interface address on N3
```

### 10.2 Flooding group-membership-LSAs

When MOSPF routers and non-multicast OSPF routers are mixed together in a routing domain, the group-membership-LSAs are not flooded to the non-multicast routers[17]. As a general design principle, optional OSPF advertisements are only flooded to those routers that understand them.

A MOSPF router learns of its neighbor's multicast-capability at the beginning of the "Database Exchange Process" (see Section 10.6 of [OSPF], receiving Database Description packets from a neighbor in state Exstart). A neighbor is multicast-capable if and only if it sets the MC-bit in the Options field of its Database Description packets. Then, in the next step of the Database Exchange process, group-membership-LSAs are included in the **Database summary list** sent to the neighbor (see Sections 7.2 and 10.3 of [OSPF]) if and only if the neighbor is multicast-capable.

When flooding group-membership-LSAs to adjacent neighbors, a MOSPF router looks at the neighbor's multicast-capability. Group-membership-LSAs are only flooded to multicast-capable neighbors. To be more precise, in Section 13.3 of [OSPF], group-membership-LSAs are only placed on the **Link state retransmission lists** of multicast-capable neighbors[18]. Note however that when sending Link State Update packets as multicasts, a non-multicast neighbor may (inadvertently) receive group-membership-LSAs. The non-multicast router will then simply discard the LSA (see Section 13 of [OSPF], receiving LSAs having unknown LS types).

## 11 Detailed description of multicast datagram forwarding

This section describes in detail the way MOSPF forwards a multicast datagram. The forwarding process has already been informally presented in Section 2.2. However, there are several obscure configuration options (e.g., the IPMulticastForwarding interface parameter) that have been presented elsewhere in this document, which may influence the forwarding process. This section gathers together all the influencing factors into a single algorithm.

It is assumed in the following that the datagram under consideration has actually be received on one of the router's interfaces. Locally generated datagrams (i.e., originated by one of the router's internal applications) are handled instead by the algorithm in Section 11.3.

Assume that the datagram's IP destination is Group G. The forwarding process then consists of the following steps:

1. Upon reception of the datagram, the MOSPF router notes the following parameters. These parameters are examined in later steps, to determine whether the datagram should be forwarded.
   a. The *receiving MOSPF interface* associated with the datagram. Based on the receiving physical interface, the receiving MOSPF interface is selected by the algorithm in Section 11.1.
   b. Whether the datagram was received as a link-level multicast/broadcast or as a link-level unicast. This information is used later in Step 7 to help determine whether the datagram should be forwarded.

---

17. Since the Designated Router controls flooding on the network, this is another reason to ensure that a MOSPF router is elected as Designated Router.

18. In other words, group-membership-LSAs will never be retransmitted to non-multicast routers.

2. A copy of the datagram should be passed to each internal application that has joined Group G on the receiving MOSPF interface (see Section 5).
3. If the datagram's IP source address matches the receiving MOSPF interface's IP address, the datagram should not be forwarded further, and should instead be discarded, completing the forwarding process. This keeps the router's own locally originated datagrams from being mistakenly replicated, in those cases where the receiving MOSPF interface receives its own multicast transmissions.
4. If Group G falls into the range 224.0.0.1 through 224.0.0.255 inclusive, the datagram should not be forwarded further. This range of addresses has been dedicated for use on a local network segment only.
5. Associate a source network (SourceNet) with the multicast datagram, as described in Section 11.2. If SourceNet cannot be determined (i.e., there is no available unicast route back to the datagram source), the datagram should not be forwarded further.
6. Look up the **forwarding cache** entry (see Section 8.5) matching the datagram's [SourceNet, Group G, TOS] combination. If the cache entry does not yet exist, one is built by the calculation in Section 12. In order for the datagram to be forwarded, the contents of the forwarding cache entry must be further verified against the received datagram's characteristics as follows:
    a. If the forwarding cache entry's **upstream node** is unspecified (i.e., NULL), then the datagram should not be forwarded further.
    b. Otherwise, suppose that the forwarding cache entry's upstream node is set to EXTERNAL. In this case, the datagram is forwarded further if and only if the receiving MOSPF interface is set to NULL (i.e., if and only if the datagram was received on a non-MOSPF interface).
    c. Otherwise, if the datagram's receiving MOSPF interface does not attach to the forwarding cache entry's upstream node, the datagram should not be forwarded further.
7. If the receiving MOSPF interface's IPMulticastForwarding parameter is set to **data-link unicast**, the datagram should be forwarded further only if it was received as a data-link unicast.
8. At this point the datagram is eligible for further forwarding. Before forwarding, the router checks to see whether it has any internal applications that have joined Group G on an interface-independent basis. If so, a copy of the datagram should be passed to each such requesting application process.
9. Examine each of the **downstream interfaces** listed in the forwarding cache entry. If the TTL in the datagram is greater than or equal to the TTL specified for the downstream interface, a copy of the datagram should be forwarded out the downstream interface. Before forwarding the datagram copy, the copy's TTL should be decremented by 1. On most interfaces, the datagram is forwarded as a data-link multicast/broadcast. The exact data-link encapsulation is dependent on the attached network's type:
    - On **ethernet** and **IEEE 802.3** networks, the datagram is forwarded as a data-link multicast. The destination data-link multicast address is selected as an algorithmic translation of the IP multicast destination. See [RFC 1112] for details.
    - On **FDDI** networks, the datagram is forwarded as a data-link multicast. The destination data-link multicast address is selected as an algorithmic translation of the IP multicast destination. See [RFC 1390] for details.
    - On **SMDS** networks, the datagram is forwarded using the same SMDS address that is used by IP broadcast datagrams. See [RFC 1209] for details.
    - On networks that support broadcast, but not multicast (e.g., the Experimental Ethernet), the datagram is forwarded as a data-link broadcast. See [RFC 1112] for details.

- On **point-to-point** networks, the datagram is forwarded in the same way that unicast datagrams are forwarded. See [RFC 1112] for details.

10. Examine each of the **downstream neighbors** listed in the forwarding cache entry. If the TTL in the datagram is greater than or equal to the TTL specified for the downstream neighbor, a copy of the datagram should be forwarded to the downstream neighbor (as a data-link unicast). Before forwarding the datagram copy, the copy's TTL should be decremented by 1.

ICMP error messages are never generated in response to received IP multicasts. In particular, ICMP destination unreachables and ICMP TTL expired messages are not generated by the above procedure if the router refuses to forward a multicast datagram.

## 11.1 Associating a MOSPF interface with a received datagram

A MOSPF interface must be associated with a received multicast datagram before it is forwarded (see Step 1a of Section 11), and with received IGMP Host Membership Reports before they are processed (see Section 9.2).

When there is only a single IP network assigned to the physical interface that received the datagram, the choice of receiving MOSPF interface is clear. When there are multiple logical IP networks attached to the receiving physical interface, the receiving MOSPF interface is selected as follows. Examine all of the MOSPF interfaces associated with the receiving physical interface. Discard those interfaces whose IPMulticastForwarding parameter has been set to disabled. The *receiving MOSPF interface* is then the remaining interface having the highest IP interface address (or NULL if there are no remaining interfaces)[19].

## 11.2 Locating the source network

MOSPF forwarding cache entries are indexed by the datagram's source IP network/subnet/supernet. For this reason, whenever an IP multicast datagram is received, the IP network belonging to the datagram's IP source address must be found. This is accomplished by the following algorithm:

Look up the OSPF TOS 0 routing table entry[20] corresponding to the datagram's IP source address, as described in Section 11.1 of [OSPF]. If this routing table entry describes an OSPF intra-area or inter-area route, the source network is set to be the network defined by the routing table entry's Destination ID and Address Mask (see Section 11 of [OSPF]). Otherwise (i.e., the routing table entry specifies an external route, or there is no matching routing table entry), the list of matching AS external-link-LSAs is examined. A matching AS external-link-LSA is one that describes a network which contains the datagram's IP source address. The list of matching AS external-link-LSAs is pruned in the following steps to determine the source network:

1. Those AS external-link-LSAs with MC-bit clear (see Section A.1), or with LS age set to MaxAge, or which have been originated by unreachable AS boundary routers are discarded.
2. AS external-link-LSAs specifying Type 1 external metrics are always preferred over those specifying Type 2 external metrics.

19. This last step will not be necessary if the configuration guidelines presented in Section 6.5 are followed.
20. The TOS 0 routing table entry is examined regardless of the TOS specified by the multicast datagram.

3. If there are still multiple AS external-link-LSAs remaining, those specifying the best matching (i.e., most specific) network are selected. The source network is then set to the network/subnet/supernet (possibly even the default route) described by the best matching AS external-link-LSAs. Note that AS external-link-LSAs specifying a cost of LSInfinity are eligible for this best match, as long as their MC-bit is set.[21]

It is possible that two different MOSPF routers may calculate the same multicast datagram's source network differently. For example, consider the network configuration shown in Figure 4. When calculating the source network for a datagram whose source is Network N10 and destination is Group Ma, Router RT11 would calculate the source network as Network N10 itself, while Router RT10 would calculate the source network as the aggregate of Networks N9-N11 and Host H1 (advertised in a single summary-link-LSA by Router RT11). However, despite the possibility of routers selecting different source networks, all routers will still agree on the datagram's shortest-path tree.

External sources are treated differently in the above calculation since it is likely that the Internet will have separate multicast and unicast topologies for some time to come. When the multicast and unicast topologies do merge, the MC-bit will be set on all AS external-link-LSAs and the above use of the LSInfinity metric (to indicate a route that is to be used for multicast traffic, but not unicast traffic), will no longer be necessary. At that time, the determination of source network for external sources will revert to the same simple routing table lookup that is used for internal sources.

As an example of the logic for external sources, suppose a multicast datagram is received having the IP source address 10.1.1.1. Suppose also that the three AS external-link-LSAs shown in Table 3: are in the router's OSPF database. The OSPF routing table lookup would yield the network 10.1.1.0 with a mask of 255.255.255.0, however the above calculation would choose a source network of 10.1.0.0 with a mask of 255.255.0.0, despite the fact that its matching LSA has a cost of LSInfinity.

| *Network* | *Mask* | *Cost* | *MC-bit* |
|---|---|---|---|
| 10.1.1.0 | 255.255.255.0 | Type 1: 10 | clear |
| 10.1.0.0 | 255.255.0.0 | Type 2: LSInfinity | set |
| 10.0.0.0 | 255.0.0.0 | Type 2: 1 | set |

**Table 3: Sample AS external-link-LSAs**

## 11.3 Forwarding locally originated multicasts

This section describes how a MOSPF router forwards a multicast datagram that has been originated by one of the router's own internal applications. The process begins with one of the router's internal applications formatting and addressing the datagram. Forwarding the locally originated multicast then consists of the following steps:

1. Find the router interface whose IP address matches the datagram's source address. Multicast the datagram out that interface, according to the Host extensions for IP multicasting specified in [RFC 1112].

2. If the router interface found in the previous step has been configured for MOSPF, and if its IPMulticastForwarding parameter is not equal to disabled, then set the receiving MOSPF interface to that interface. Otherwise, set the receiving MOSPF interface to NULL.

21. It is assumed that a MOSPF router that wants to stop advertising a route to an external destination will use the premature aging procedure specified in Section 14.1 of [OSPF], rather than setting the AS external-link-LSA's cost to LSInfinity.

3. Execute the MOSPF forwarding process described in Section 11, beginning with its Step 4.

The above algorithm amounts to the router always multicasting the datagram out the source interface, and the executing the basic forwarding algorithm (in Section 11) as if the datagram had actually been received on the source interface. In those cases where the router receives its own multicast transmissions, unwanted replication is prevented by Step 3 of Section 11. In fact, this specification has purposely presented the forwarding algorithm (both for received and for locally originated datagrams) so that the correct forwarding actions are taken independent of whether the router receives its own multicast transmissions.

# 12 Construction of forwarding cache entries

This section details the building of a MOSPF forwarding cache entry. A high level discussion of this construction has already been presented in Sections 2.3, 2.3.1, 2.3.2, 3.2, and 4.1. Forwarding cache entries are built on demand, when a multicast datagram is received and no matching forwarding cache entry is found (see Step 6 of Section 11). The parameters passed to the forwarding cache entry build process are: the datagram's source network (see Section 11.2) and its destination group address. These two parameters are called **SourceNet** and **Group G** in the following algorithm. The main steps in the build process are the following:

1. Allocate the forwarding cache entry. Initialize its Source network to SourceNet, its Destination multicast group to Group G and its IP TOS field to match the multicast datagram's TOS. Initialize its upstream node and list of downstream interfaces to NULL.

2. For each Area A to which the calculating router is attached:

   a. Calculate Area A's datagram shortest-path tree. This calculation is described in Section 12.2 below. In many ways it is similar to the calculation of OSPF's intra-area routes, described in Section 16.1 of [OSPF]. The main differences between the multicast datagram shortest-path tree calculation and OSPF's intra-area unicast calculation are listed in Section 12.2.9 below. As a product of each area's datagram shortest-path tree, the forwarding cache entry's list of outgoing interfaces is (possibly) updated.

      Area A's datagram shortest-path tree is dependent on the datagram's IP TOS. Section 12.2 describes the TOS 0 datagram shortest-path tree. The modifications necessary for non-zero TOS values are detailed in Section 12.2.8.

   b. Possibly set the forwarding cache entry's upstream node. Only one of the calculating router's attached areas will determine the forwarding cache entry's upstream node. This area is called the datagram's RootArea. The RootArea is initially set to NULL. After completing Area A's datagram shortest-path tree, the calculation in Section 12.2.7 will determine whether Area A is the datagram's RootArea.

3. Update the forwarding cache entry's list of outgoing interfaces, according to the contents of the local group database. This ensures multicast delivery to group members residing on the calculating router's directly attached networks. This process is described in Section 12.3.

These main steps are described in more detail below. The detailed description begins with an explanation of the major data structure used by the datagram shortest-path tree calculation: The Vertex data structure.

## 12.1 The Vertex data structure

A datagram shortest-path tree is built by the **Dijkstra** or **SPF** algorithm. The algorithm is stated herein using graph-oriented language: vertices and links. Vertices are the area's routers and transit networks, and links are the router interfaces and point-to-point lines that connect them. Each vertex has the following state information attached to it. Basically, this information indicates the current best path from the SourceNet to the vertex, and the position of the vertex relative to the calculating router. Note that a separate datagram shortest-path tree is built for each area, and that the vertices described below are also specific to a single area (called Area A).

- **Vertex type**. Set to 1 for routers, 2 for transit networks. Note that this coding matches the coding for vertices listed in the group-membership-LSA (see Section A.3).
- **Vertex ID**. A 32-bit identifier for the vertex. For routers, set to the router's OSPF Router ID. For transit networks, set the IP address of the network's Designated Router. Note that this coding matches the coding for vertices listed in the group-membership-LSA (see Section A.3).
- **LSA**. The link state advertisement describing the vertex' immediate neighborhood. Can be discovered by performing a database lookup in Area A's link state database (see Section 12.2 of [OSPF]), with LS type set to Vertex type and Link State ID set to Vertex ID.
- **Parent**. In the current best path from SourceNet to the vertex, the router/transit network immediately preceding the vertex. Note that the parent can change as better and better paths are found, up until the vertex is installed on the shortest-path tree.
- **IncomingLinkType**. This parameter is set to the type of link that led to Vertex's inclusion on the shortest-path tree. Listed in order of decreasing preference[22], the possible types are: ILVirtual (virtual links), ILDirect (vertex is directly attached to SourceNet), ILNormal (either router-to-router or router-to-network links), ILSummary (OSPF summary links), ILExternal (OSPF AS external links), or ILNone (the vertex is not on the shortest-path tree).
- **AssociatedInterface/Neighbor**. If the current best path from SourceNet to the vertex goes through the calculating router, this parameter indicates the calculating router's interface (or neighbor) which leads to the vertex.
- **Cost**. The cost, in terms of the OSPF link state metric, of the current best path from SourceNet to the vertex. Note that if the cost of the path is a combination of both external type 2 and internal OSPF metrics, that the vertex' cost parameter reflects both cost components. Remember that the type 2 cost component is always more significant than the type 1 component.
- **TTL**. If the current best path from SourceNet to vertex goes through the calculating router, TTL is set to the number of routers between the calculating router and the vertex. This includes the calculating router, but does not include the vertex itself.

## 12.2 The SPF calculation

This section details the construction of datagram shortest-path trees. Such a tree describes the path of a multicast datagram as it traverses an OSPF area. For a given datagram, each router in an OSPF area builds an identical tree. A router connected to multiple areas builds a separate datagram shortest-path tree for each area.

22. This preference ordering is used in Step 5c of Section 12.2.

The datagram shortest-path tree is built by the **Dijkstra** or **SPF** algorithm, which is the same algorithm used to discover OSPF's intra-area unicast routes (see Section 16.1 of [OSPF]). The algorithm is stated herein and in [OSPF] using graph-oriented language: vertices and links. Vertices are the area's routers and transit networks, and links are the router interfaces and point-to-point lines that connect them. Basically, the algorithm manipulates two lists of vertices: the **candidate list** and the forming shortest-path tree. The candidate list consists of those vertices to which paths have been discovered, but for which the optimality of the discovered paths is yet unknown. At each cycle of the algorithm, the vertex closest to the tree's root, yet still remaining on the candidate list, is moved from the candidate list to the shortest-path tree. Then the neighbors of the just processed vertex are examined for possible addition to/modification of the candidate list. The algorithm terminates when the candidate list is empty.

The datagram shortest-path tree for Area A is constructed in the following steps. The datagram's SourceNet and its destination group G are inputs to the calculation (see Step 6 of Section 11). The datagram shortest-path tree also depends on the IP Type of service specified in the datagrams' IP Header. However, a discussion of TOS is deferred until Section 12.2.8; all calculations and costs in the current section concern TOS 0 only. Call the router performing the calculation Router RTX. At each step (and in the subordinate Sections 12.2.1 through 12.2.8) LSAs from Area A's link state database are examined. In all cases, any LSA having LS age equal to MaxAge is ignored. The main body of the calculation is in Steps 4 and 5, which are repeated until the candidate list becomes empty:

1. **Initialize the algorithm's data structures.** Clear the shortest-path tree. Initialize the state of each vertex in Area A (i.e., the area's routers and transit networks) to: Parent set to NULL, IncomingLinkType set to ILNone and AssociatedInterface/Neighbor set to NULL.

2. **Initialize the candidate list.** One or more vertices are initially placed on the candidate list, depending on the location of SourceNet with respect to Area A and Router RTX. This breaks down into the following cases (which are named for later reference):

   - **Case SourceIntraArea:** SourceNet belongs to Area A. In this case, the candidate list is initialized as in Section 12.2.1.

   - **Case SourceInterArea1:** SourceNet belongs to an OSPF area that is not directly attached to Router RTX. In this case, the candidate list is initialized as in Section 12.2.2.

   - **Case SourceInterArea2:** SourceNet does not belong to Area A, but it still belongs to an OSPF area that is directly attached to Router RTX. In this case, the candidate list is initialized as in Section 12.2.3.

   - **Case SourceExternal:** SourceNet is external to the OSPF routing domain, and Area A is not an OSPF stub area. In this case, the candidate list is initialized as in Section 12.2.4.

   - **Case SourceStubExternal:** SourceNet is external to the OSPF routing domain, and Area A is an OSPF stub area. In this case, the candidate list is initialized as in Section 12.2.5.

   Two different routers in Area A may select different initialization cases above. For example, consider the network configuration shown in Figure 4. When calculating the Area 3 datagram shortest-path tree for a datagram whose source is Network N7 (e.g., from Host H5) and destination is Group Ma, Router RT11 would initialize the candidate list using Case SourceInterArea2 while Router RT9 would use Case SourceInterArea1. Likewise, if Area 3 were configured as an OSPF stub area and the datagram source was the external network N12, Router RT11 would use Case SourceStubExternal while Router RT9 would use Case SourceInterArea1! However, despite the possibility of routers selecting different cases, all routers in an area will still initialize the candidate list (and in fact, run the rest of the SPF calculation) identically.

3. **If the candidate list is empty, the algorithm terminates**.

4. **Move the closest candidate vertex to the shortest-path tree.** Select the vertex on the candidate list that is closest to SourceNet (i.e., has the smallest Cost value). If there are multiple possibilities, select transit networks over routers. If there are still multiple possibilities remaining, select the vertex having the highest Vertex ID. Call the chosen vertex Vertex V. Remove Vertex V from the candidate list, and install it on the shortest-path tree.

   Next, determine whether Vertex V has been labelled with the Destination multicast Group G. If so, it may cause the forwarding cache entry's list of outgoing interfaces/neighbors to be updated. See Section 12.2.6 for details.

5. **Examine Vertex V's neighbors for possible inclusion in the candidate list.** Consider Vertex V's LSA. Each link in the LSA describes a connection to a neighboring router/network. If the link connects to a stub network, examine the next link in the LSA. Otherwise, the link (Link L) connects to a neighboring transit node. Call this node Vertex W. Perform the following steps on Vertex W:

   a. If W is already on the shortest-path tree, or if W's LSA does not contain a link back to vertex V, or if W's LSA has LS age of MaxAge, or if W is not multicast-capable (indicated by the MC-bit in the LSA's Options field), examine the next link in V's LSA.

   b. Otherwise determine the cost to associate with the link from V to W. If SourceNet belongs to Area A (Case SourceIntraArea in Step 2), use the cost listed for Link L in V's LSA. Otherwise, use the link's reverse cost: Examine W's LSA, and find the cost listed for the link connecting back to V. Actually, when V and W are both routers, there may be multiple links between them. In this case, use the smallest cost listed in W's LSA for any of the links connecting back to V and having the same Type (as specified in the Router-LSA; must be either: point-to-point connection or virtual link) as Link L[23].

   c. Calculate the cost from SourceNet to W, when using Link L. It is the sum of the cost of SourceNet to V (i.e., V's Cost parameter) plus the link cost calculated in Step 5b. Let this sum be Cost C. If W is not yet on the candidate list, install W on the candidate list, modifying its parameters as specified below (Step 5d). Otherwise, W is on the candidate list already. In this case, if:

      - C is less than W's current Cost, update W's parameters on the candidate list as specified below (Step 5d).
      - C is equal to W's current Cost, then the following tiebreakers are invoked. The type of Link L is compared to W's current **IncomingLinkType,** and whichever link has the preferred type is chosen (the preference order of link types is listed in Section 12.1's definition of **IncomingLinkType**). If the link types are the same, then a link whose Parent is a transit network is preferred over one whose Parent is a router. If the links are still equivalent, the link whose Parent has the higher Vertex ID is chosen. Whenever Link L is chosen, W's parameters are modified as below (Step 5d). Whenever the previously discovered link is chosen, the next link in V's LSA is examined instead.
      - C is greater than W's current Cost, examine the next link in V's LSA.

   d. At this point, a better candidate path has been found to Vertex W, using Link L. Modify Vertex W's parameters accordingly. W's **Parent** is set to Vertex V. W's **IncomingLinkType** is set to ILVirtual if Link L is a virtual link, otherwise IncomingLinkType is set to ILNormal. W's **Cost** parameter is set to C. W's TTL and AssociatedInterface/Neighbor parameters are set according to one of the following cases:

      - Vertex V is the calculating router itself. In this case, W's **TTL** parameter is set to 1. If Link L is a virtual link, W's **AssociatedInterface/Neighbor** is set to NULL. Otherwise, W's AssociatedInterface/Neighbor is set to the non-virtual interface connecting the calculating router to W which has the smallest cost value. Note that, in the reverse cost (inter-area and inter-AS multicast) cases, this may not be the interface corre-

23. No attempt is made to match the links' two halves. See Step 5d.

sponding to Link L. However, since W is only concerned with the node it is receiving the datagram from (the **upstream node**; see Section 11), and not with the particular interface the datagram is received on, the calculating router is free to pick the sending interface when there are multiple connecting links.

- Vertex V is upstream of the calculating router (i.e., V's AssociatedInterface/Neighbor is equal to NULL). In this case, Vertex W's TTL parameter is set to 0, and its **AssociatedInterface/Neighbor** is set to NULL.
- V is a transit network, and is directly downstream from the calculating router (i.e., V's AssociatedInterface/Neighbor is non-NULL and V's TTL is set to 1). W is then one of the calculating router's neighbors. In this case, W's **TTL** parameter is also set to 1. If network V has been configured for **data-link unicasting** (see Section B.2) or if V is a non-broadcast network, **W's AssociatedInterface/Neighbor** is set to W itself (a neighbor of the calculating router). Otherwise, W's AssociatedInterface/Neighbor is set to the calculating router's interface to Network V.
- Vertex V is downstream from the calculating router (i.e., V's AssociatedInterface/Neighbor is non-NULL), and either a) V is a router or b) V's TTL parameter is greater than 1. In these cases, W's **AssociatedInterface/Neighbor** parameter is copied directly from V. If V is a router, W's **TTL** parameter is set to V's TTL parameter incremented by one. If V is a transit network, W's TTL parameter is set directly to V's TTL parameter.

6. **If the candidate list is non-empty, go to Step 4.** Otherwise, the algorithm terminates.

After the datagram shortest-path tree for Area A is complete, the calculating router (RTX) must decide whether Area A, out of all of RTX's attached areas, determines the forwarding cache entry's **upstream node**. This determination is described in Section 12.2.7.

Examples of the above SPF calculation, with particular emphasis on the tiebreaking rules, are given in Appendix C.

### 12.2.1 Candidate list Initialization: Case SourceIntraArea

In this case, SourceNet belongs to Area A. The candidate list is then initialized as follows. Start with the LSA listed as Link State Origin in the matching OSPF routing table entry. If this LSA is not multicast-capable (i.e, its Options field has the MC-bit clear) the candidate list should be set to NULL. Otherwise, the vertex identified by the LSA is installed on the candidate list, setting its vertex parameters as follows: IncomingLinkType set to ILDirect, Cost set to 0, Parent to NULL and AssociatedInterface/Neighbor to NULL.

As a consequence of this initialization, note that if SourceNet is a stub network, then the datagram shortest-path tree will not actually be rooted at the datagram source, but will instead be rooted at the MOSPF router that attaches the stub network to the rest of the MOSPF system. For example, consider the network configuration shown in Figure 4. When calculating the Area 2 datagram shortest-path tree for a datagram whose source is Network N7 (e.g., from Host H5) and destination is Group Ma, Router RT11 (and all other routers attached to Area 2) will begin with the candidate list set to Router RT8. As another example, the datagram shortest-path tree pictured in Figure 3 is really rooted at Router RT3 instead of Network N4.

### 12.2.2 Candidate list Initialization: Case SourceInterArea1

In this case, SourceNet belongs to an OSPF area that is not directly attached to the calculating router (RTX). The candidate list is then initialized as follows. Examine the Area A summary-link-LSAs advertising SourceNet. For each such summary-link-LSA: if both a) the MC-bit is set in the LSA's Options field and b) the advertised cost is not equal to LSInfinity, then the vertex representing the LSA's advertising area border router is added to the candidate list. An added vertex' state is initialized as: IncomingLinkType set to ILSummary, Cost to whatever is advertised in the LSA, Parent to NULL and AssociatedInterface/Neighbor to NULL.

For example, consider the network configuration shown in Figure 4. When calculating the Area 1 datagram shortest-path tree for a datagram whose source is Network N7 (e.g., from Host H5) and destination is Group Ma, Router RT2 would initialize the candidate list to contain the two area border routers RT3 (with a cost of 20) and RT4 (with a cost of 19). See Figure 6 for more details.

### 12.2.3 Candidate list Initialization: Case SourceInterArea2

In this case, SourceNet belongs to an OSPF area other than Area A, but one that is still directly attached to the calculating router (RTX). The candidate list is then initialized in the following two steps:

1. Find the Area A summary-link-LSA that best matches SourceNet, excluding those summary-link-LSAs specifying cost LSInfinity or having unreachable Advertising Routers[24]. A matching summary-link-LSA is one that advertises a range of addresses containing SourceNet; the best matching is as usual the most specific match. Let SourceRange be the network described by the best matching summary-link-LSA.
2. Similar to the logic in the SourceInterArea1 case, examine all the Area A summary-link-LSAs which advertise SourceRange. For each such summary-link-LSA: if both a) the MC-bit is set in the LSA's Options field, b) the advertised cost is not equal to LSInfinity and c) the Advertising Router is reachable, then the vertex representing the LSA's Advertising Router is added to the candidate list. An added vertex' state is initialized as: IncomingLinkType set to ILSummary, Cost to whatever is advertised in the LSA, Parent to NULL and AssociatedInterface/Neighbor to NULL.

The reason why SourceRange is used, instead of simply using SourceNet (as was done in case SourceInterArea1), is that routing information may have been collapsed at area boundaries. In order for Area A's area border routers and its internal routers to construct the same Area A datagram shortest-path tree, they must both start at SourceRange — Area A's internal routers know nothing about SourceNet. Note that SourceRange is not discovered simply by looking at the calculating router's configured set of area address ranges, in order to avoid dependence on the configured area address ranges being synchronized across all area border routers.

For example, consider the network configuration shown in Figure 4. When calculating the Area 2 datagram shortest-path tree for a datagram whose source is Network N11 and destination is Group Ma, Router RT11 would calculate SourceRange to be the collection: Networks N9-N11 and Host H1. It would then initialize the candidate list to contain itself (RT11) only, with an associated Cost of 1 (since RT11 is advertising Networks N9-N11 and Host H1 in a summary-link-LSA with a cost of 1).

24. However, a summary-link-LSA is eligible for matching even if the MC-bit in its Options field is clear.

### 12.2.4 Candidate list Initialization: Case SourceExternal

In this case, SourceNet is external to the OSPF routing domain, and Area A is not an OSPF stub area. The candidate list is then initialized as follows. Note that an attempt may be made to add a Vertex W to the candidate list when W already belongs to the candidate list. When this happens, W's vertex parameters are updated if the Cost parameter it would be added with is better[25] (closer to SourceNet) than its previous value. When the costs are the same, W's parameters are still modified if the IncomingLinkType it would be added with is better (see IncomingLinkType's definition in Section 12.1) than its previous value.

For each AS external-link-LSA advertising SourceNet, the following steps are performed:

- If the AS external-link-LSA's MC-bit is clear or if its advertising router is not reachable, then the AS external-link-LSA is not used. AS external-link-LSAs having their MC-bit set and advertising a cost of LSInfinity can be used; these LSAs describe paths that can be used for multicast, but not unicast, data traffic (see Section 11.2).
- If the AS external-link-LSA's Forwarding address field is 0.0.0.0, the following vertices are added to the candidate list. If the Advertising AS boundary router (call it ASBR) belongs to Area A, the vertex representing the AS boundary router is added to the candidate list using parameters: IncomingLinkType set to ILExternal, Cost to whatever is advertised in the LSA, Parent to NULL and AssociatedInterface/Neighbor to NULL. Then, regardless of whether ASBR belongs to Area A, all Area A area border routers that are advertising reachable multicast-capable (MC-bit set) type 4 summary-link-LSAs for ASBR are added to the candidate list. Each such area border router is added with the parameters: IncomingLinkType set to ILSummary, Cost to the sum of whatever is advertised in the type 4 summary-link-LSA plus the value in the original AS external-link-LSA, Parent to NULL and AssociatedInterface/Neighbor to NULL.
- If the AS external-link-LSA's Forwarding address field is non-zero, the Forwarding address is looked up in the OSPF routing table. Then processing breaks into one of the following cases:
  - The Forwarding address is not usable. In this case, nothing is added to the candidate list. The Forwarding address is not usable if either it has no matching routing table entry, or if the matching routing table entry is neither of type intra-area nor of type inter-area.
  - The Forwarding address belongs to Area A[26]: the Forwarding address' matching routing table entry has Path-type of intra-area and its Associated area is Area A. In this case, the vertex represented by the matching routing table entry's Link State Origin field is added to the candidate list (assuming that the vertex is multicast-capable). The vertex is added with the parameters: IncomingLinkType set to ILExternal, Cost to whatever was advertised in the original AS external-link-LSA, Parent to NULL and AssociatedInterface/Neighbor to NULL.
  - The Forwarding address belongs to an area that is not attached to Router RTX[27]: the Forwarding address' matching routing table entry has Path-type of inter-area. Call the network represented by the matching routing table entry ForwardNet. For each reachable multicast-capable summary-link-LSA (in Area A) advertising ForwardNet, add the LSA's advertising area border router to the candidate list using parameters: IncomingLink-Type set to ILSummary, Cost to the sum of whatever is advertised in the summary-link-LSA plus the value in the original AS external-link-LSA, Parent to NULL and AssociatedInterface/Neighbor to NULL.

---

25. Costs may have both a Type 2 and a Type 1 component; the Type 2 component is always most significant.
26. This case mirrors the SourceIntraArea candidate list initialization in Section 12.2.1.
27. This case mirrors the SourceInterArea1 candidate list initialization in Section 12.2.2.

- The Forwarding address belongs to another one of Router RTX's attached areas[28]: the Forwarding address' matching routing table entry has Path-type of intra-area and its associated Area is other than Area A. Call the network represented by the matching routing table entry ForwardNet. First find the Area A summary-link-LSA that best matches ForwardNet, excluding those summary-link-LSAs specifying cost LSInfinity or having unreachable Advertising Routers. Let ForwardRange be the network described by the best matching summary-link-LSA. Then, for each reachable multicast-capable summary-link-LSA (in Area A) advertising ForwardRange, add the LSA's advertising area border router to the candidate list using parameters: IncomingLinkType set to ILSummary, Cost to the sum of whatever is advertised in the summary-link-LSA plus the value in the original AS external-link-LSA, Parent to NULL and AssociatedInterface/Neighbor to NULL.

The above calculation can be restated as follows. Each of Area A's inter-area multicast forwarders and inter-AS multicast forwarders are examined. Those that have multicast-capable paths to SourceNet (represented as either a multicast-capable AS external link or the concatenation of a Type 4 summary link and a multicast-capable AS external link) are added to the candidate list as router vertices. (It is possible that, when considering a router that is both an inter-area multicast forwarder and an inter-AS multicast forwarder, two equal cost paths exist to SourceNet, one an AS external link and the other a concatenation of a Type 4 summary link and an AS external link. In this case, the concatenation of the Type 4 summary link and the AS external link is preferred). The added vertex' state is set as follows: IncomingLinkType set to ILSummary if the path is represented as a concatenation of a Type 4 summary link and an AS external link, IncomingLinkType set to ILExternal otherwise, Cost set to the cost of the shortest path from vertex to SourceNet, Parent set to NULL and AssociatedInterface/Neighbor set to NULL.

For example, consider the network configuration shown in Figure 4. When calculating the Area 2 datagram shortest-path tree for a datagram whose source is Network N14 and destination is Group Ma, the candidate list would be initialized to the two routers RT7 at a cost of 14 and RT10 at a cost of 19. This assumes that the external costs pictured in Figure 4 are external type 1s.

### 12.2.5 Candidate list Initialization: Case SourceStubExternal

In this case, SourceNet is external to the OSPF routing domain, and Area A is an OSPF stub area. The candidate list is then initialized similarly to case SourceInterArea1. The Area A summary-link-LSAs advertising DefaultDestination are examined. For each such summary-link-LSA having both its MC-bit set and its advertised cost not equal to LSInfinity, the vertex representing the LSA's advertising area border router is added to the candidate list. An added vertex' state is initialized as: IncomingLinkType set to ILSummary, Cost to whatever is advertised in the LSA, Parent to NULL and AssociatedInterface/Neighbor to NULL.

The most likely outcome of the above is that all of stub Area A's inter-area multicast forwarders will be installed on the candidate list, with appropriate costs.

### 12.2.6 Processing labelled vertices

When encountered during the SPF calculation, vertices labelled with the destination multicast group (Group G) may cause the forwarding cache entry's list of downstream interfaces/neighbors to be modified. A Vertex V in Area A is labelled with Group G if and only if at least one of the following holds:

---

28. This case mirrors the SourceInterArea2 candidate list initialization in Section 12.2.3.

1. V is a router, and its router-LSA indicates that it is a wild-card multicast receiver (i.e., bit W in its router-LSA is set). This may be true when V is an inter-area or inter-AS multicast forwarder.
2. V is listed in the body of a group membership-LSA. In particular, find the originator of Vertex V's LSA; call it Router Y. Then find the group-membership-LSA in Area A's link state database which has Link State ID = Group G and Advertising Router = Router Y (see Section A.3). If this group-membership-LSA exists, and if Vertex V is listed in the body of the LSA (see Sections 10 and A.3), then Vertex V is labelled with Group G.

When Vertex V is added to the shortest-path tree in Step 4 of Section 12.2, and if Vertex V is both downstream from the calculating router (i.e., Vertex V's **AssociatedInterface/Neighbor** is non-NULL) and labelled with Group G, then Vertex V's AssociatedInterface/Neighbor is added to the forwarding cache entry's **list of downstream interfaces/ neighbors**. In addition, Vertex V's TTL value is attached to the added downstream interface/neighbor. If the particular interface/neighbor had already been added to the list of downstream interfaces/neighbors, the list is simply modified by setting the downstream interface/neighbor's TTL value to the minimum of its existing TTL value and Vertex V's TTL value.

### 12.2.7 Merging datagram shortest-path trees

After the datagram shortest-path tree for Area A is complete, the calculating router (RTX) must decide whether Area A, out of all of its attached areas, determines the forwarding cache entry's **upstream node**. This is done by examining RTX's position on the Area A datagram shortest-path tree, which is in turn described by RTX's Area A Vertex data structure. If RTX's Vertex parameter IncomingLinkType is either ILNone (RTX is not on the tree), ILVirtual or ILSummary, then some area other than Area A will determine the upstream node. Otherwise, Area A might possibly determine the upstream node (i.e., may be selected the RootArea), depending on the following tiebreakers[29]:

- If RootArea has not been set, then set RootArea to Area A. Otherwise, compare the present RootArea to Area A in the following:
- Choose the area that is "nearest to the source". Nearest to the source depends on each area's candidate list initialization case, as it occurs in Step 2 of Section 12.2. The initialization cases, listed in order of decreasing preference (or nearest to farthest) are: SourceIntraArea, SourceInterArea1, SourceExternal and SourceStubExternal. Areas whose candidate list initialization falls into case SourceInterArea2 are never used as the RootArea. As an example, consider the network configuration shown in Figure 4. When calculating the datagram shortest-path tree for a datagram whose source is Network N7 (e.g., from Host H5) and destination is Group Ma, Router RT11 would set its RootArea to Area 2 (Case SourceIntraArea) instead of Area 3 (Case SourceInterArea2) or the backbone Area 0 (Case SourceInterArea).
- If there are still two equally good areas, and one of them is the backbone, set RootArea to the backbone (Area 0).
- If there are still two equally good areas, set RootArea to the area whose datagram shortest-path tree provides the shortest path from SourceNet to RTX. This is a comparison of RTX's Vertex parameter Cost in the two areas.
- If there are still two equally good areas, set RootArea to one with the highest OSPF Area ID.

If the above has set the RootArea to be Area A, the forwarding cache entry's upstream node must be set accordingly. This setting depends on the IncomingLinkType in RTX's Area A Vertex structure. If IncomingLinkType is equal to

29. Note that selecting the upstream node in this manner enforces the inter-area routing architecture outlined in Section 3.1. Namely, the multicast datagram is forwarded from the source area, over the backbone and then into the non-backbone areas. This is similar to the "hub and spoke" architecture for unicast forwarding described in Section 3.2 of [OSPF].

ILDirect, the upstream node is set to the appropriate directly-connected stub network. If equal to ILNormal, the upstream node is set to the Parent field in RTX's Area A Vertex structure. If equal to ILExternal, the upstream node is set to the placeholder EXTERNAL.

### 12.2.8 TOS considerations

The previous sections 12.2 through 12.2.7 described the construction of a TOS 0 (default TOS) datagram shortest-path tree. However, in a TOS-capable router, a separate tree may be built for each TOS. If a TOS-capable router receives a multicast datagram that specifies a non-zero TOS X, it first builds the TOS 0 datagram shortest-path tree. Then, if all the routers on the pruned tree are TOS-capable, a separate TOS X datagram shortest-path tree is calculated[30]. Otherwise, the TOS 0 tree is used for all datagrams, regardless of their specified TOS.

To determine whether there are any TOS-incapable routers on the pruned TOS 0 tree, the following additions are made to Section 12.2's tree calculation:

- A new piece of state information is added to each vertex: **TOS-capable path**. This indicates whether the present path from SourceNet to vertex, as represented on the datagram shortest-path tree, contains only TOS-capable routers.
- The **TOS-capable path** parameter is calculated when the vertex is first added to the candidate list and recalculated when/if the vertex' position on the candidate list is modified (see Section 12.2's Step 2 and Step 5d). The parameter is set to TRUE if both the vertex itself is TOS-capable and the vertex' parent has its TOS-capable path parameter set to TRUE; otherwise, TOS-capable path is set to FALSE.
- All routers on the TOS 0 datagram shortest-path tree are TOS-capable if and only if, whenever a vertex labelled with Group G is added to the shortest-path tree (Section 12.2.6), the value of the vertex' TOS-capable path parameter is TRUE.

The source of the multicast datagram is always located using a TOS 0 routing table lookup, regardless of the datagram's TOS classification (see Section 11.2). If the calculating router is not capable of TOS-based routing, it calculates only TOS 0 datagram shortest-path trees, and uses them to route datagrams independent of TOS value. Otherwise, when calculating the TOS X datagram shortest-path tree, the algorithm in Section 12.2 is used, with the modifications listed below.

- When calculating RangeNet and ForwardRange in Sections 12.2.3 and 12.2.4 respectively, only summary-link-LSAs having TOS 0 cost of LSInfinity are excluded (no change from the TOS 0 case). However, when adding vertices to the candidate list in Sections 12.2.2 through 12.2.5, the TOS X cost of the summary links and/or AS external links (and not the TOS 0 cost) are reflected in the added vertices' Cost parameter.
- In Step 5 of Section 12.2, the TOS X cost of Link L (in the appropriate direction) is used, not the TOS 0 cost.
- Non-TOS-routers are not added to the candidate list, and are thus excluded from the trees.

30. This procedure seems backwards. One would expect that the TOS X datagram tree would be built first. However, the SPF calculation must ensure that all routers participating in the forwarding of that datagram, both TOS-capable and non-TOS-capable, build the same tree. Since it is known that the non-TOS-capable routers will use the TOS 0 tree, the only safe way to use the TOS X tree is when you are guaranteed that the non-TOS-capable routers will decline to forward the datagram. This guarantee is clearly met when there are only TOS-capable routers on the TOS 0 datagram tree.

### 12.2.9 Comparison to the unicast SPF calculation

There are many similarities between the construction of a multicast datagram's shortest-path trees in Section 12.2 and OSPF's intra-area route calculation for unicast traffic (Section 16.1 of [OSPF]). Both have been described in terms of Dijkstra's algorithm. However, there are some differences. The major differences are listed below:

- In the multicast case, the datagram SPF calculation is rooted at the datagram's source. In the unicast case, each router is the root of its own unicast intra-area SPF calculation.
- In the multicast case, the datagram shortest-path tree is a true tree; i.e., between any two nodes on the tree there is one path. However, due to the provision for equal-cost multipath in [OSPF], the unicast SPF calculation may add additional links to the shortest-path tree.
- In order to avoid unwanted replication of multicast datagrams, MOSPF ensures that, for any given datagram, each router builds the exact same datagram shortest-path tree. This forces two differences from the unicast SPF calculation. First, it eliminates the possibility of equal-cost multipath. Secondly, when the MOSPF system contains multiple alternate paths, the algorithm must ensure that each MOSPF router deterministically chooses the same alternative. For this reason, tie-breaking mechanisms have been specified in Steps 2, 4 and 5b of Section 12.2.
- The calculation of datagram shortest path trees takes into account only those links that connect transit nodes (i.e, router to router or router to transit network links). The unicast SPF calculation in Section 16.1 of [OSPF] must additionally examine links to stub networks, although this is done after all the transit links are examined.
- While both the multicast and unicast trees select shortest paths on the basis of the OSPF metric, the datagram shortest-path trees also keep track of the TTL values between the root (datagram source) and all destinations (group members). This enables more efficient implementation of IP multicast's "expanding ring search" (see Section 2.3.4).
- In the multicast case, the algorithm is sometimes forced to use the link state cost for the reverse direction (i.e, the cost towards, instead of away from, the source). This is because the costs of OSPF summary-link-LSAs and AS external-link-LSAs, which sometime form the base of the multicast datagram shortest-path trees, are specified in the reverse direction (from the multicast perspective).
- There are potentially many more datagram shortest-path trees that need to be calculated (one for each source net, destination group and TOS combination), than the limited number of unicast SPF trees (one per each TOS). This is the main reason that the datagram shortest-path trees are calculated on demand; it is hoped that this will spread the cost of the SPF calculations over time[31].
- The way that the two algorithms handle TOS is different. In the multicast case, if a TOS-incapable node is encountered during the calculation of the TOS 0 datagram shortest-path tree, the TOS 0 datagram shortest-path tree is used instead of trying to build the TOS X tree (see Section 12.2.8). In the unicast case, the TOS X tree is always used, only falling back on the TOS 0 paths when a TOS X path does not exist.

## 12.3 Adding local database entries to the forwarding cache

After the datagram shortest-path trees have been built for each attached area, the forwarding cache has an upstream node and a list of downstream interfaces. In order to ensure the delivery of the multicast datagram to group members

---

31. Indeed, there will also be those cases where the router, not being on a particular datagram shortest-path tree, will never have to calculate the particular tree, since the router will not receive the datagram in the first place.

on directly attached networks, the local group database (Section 8.4) must then be scanned for possible addition to the list of downstream interfaces. All local group database entries having Group G as MulticastGroup are examined. Suppose [Group G, Network N] is one such entry. If the calculating router (RTX) is Network N's Designated Router, then RTX's Network N interface is added to the list of outgoing interfaces, with a TTL of 1. If the Network N interface was already present in the list of outgoing interfaces, its TTL is simply set to 1.

For example, consider the network configuration shown in Figure 4 when calculating the forwarding cache entry for a datagram whose source is Network N4 (e.g., from Host H2) and destination is Group Mb. After calculating the datagram shortest-path tree for Area 1, Router RT2 would have set it upstream node to Network N3 and its list of downstream interfaces to NULL. But then looking at its local group database, it would add its Network N2 interface with a TTL of 1 to its list of downstream interfaces.

# 13 Maintaining the forwarding cache

A MOSPF router may, for resource reasons, limit the size of its forwarding cache. At any time cache entries can be purged to make room for newer entries, since the purged entries can always be rebuilt when necessary. This memo does not specify an algorithm to select which entries to purge. However, care should be taken to ensure that any particular entry is not continually rebuilt and then purged again (i.e., thrashing should be avoided).

The building of the forwarding cache has been previously described in Section 12. There are events that force one or more forwarding cache entries to be deleted; these events are described below. Note that deleted cache entries will be rebuilt on an as-needed basis.

- When the internal topology of the MOSPF system changes, *all* forwarding cache entries must be deleted. This is because internal topology changes may invalidate the previously calculated datagram shortest-path trees. Since the multicast routing calculation depends on the result of the unicast routing calculations, the forwarding cache should be cleared *after* the unicast routing table is rebuilt. Internal topology changes are indicated when both a) a new instance of either a router-LSA or a network-LSA is received and b) the contents of the new advertisement (other than the LS age, LS sequence number and LS checksum fields) are different from the previous instance. This covers routers and links going up or down, routers that change from being multicast-incapable to being multicast-capable, etc.
- When a Type 3 summary-link-LSA (network summary) changes, those forwarding cache entries specifying datagram sources belonging to the range of addresses described by the updated summary-link-LSA must be deleted. See Sections 12.2.3 and 12.2.5.
- Suppose that the content of an AS external-link-LSA changes. If the AS external-link-LSA describes an external network N, then all forwarding cache entries specifying an external source network that is contained in N or that contains N (i.e., external sources that are a subset or a superset of N) must be deleted.
- When membership in a multicast group changes, all forwarding cache entries for the particular group must be deleted. Group membership changes are indicated when either a) the content of a group-membership-LSA changes or b) an entry in the local group database (see Section 8.4) changes.
- When the cost to an AS boundary router or to a forwarding address specified by one or more AS external-link-LSAs changes, all forwarding cache entries specifying an external network as datagram source must be deleted. In this case, potentially all inter-AS datagram shortest-path trees have been invalidated. The forwarding cache entries should be deleted after the new best cost to the AS boundary router/forwarding address has been calculated.

# 14 Other additions to the OSPF specification

MOSPF requires some modifications to the base OSPF protocol. All these modifications are backward-compatible. A router running MOSPF will still interoperate with an OSPF router when forwarding unicast traffic. Most of the modifications have been described earlier in this document. This section collects together those changes which have yet to be mentioned, organizing them by the affected Section of [OSPF].

## 14.1 The Designated Router

This functionality is described in Section 7.3 of [OSPF]. In OSPF, a network's Designated Router has two specialized roles. First, it originates the network's network-LSA. Second, it controls the flooding on the network, in that all of the routers on the network synchronize with the Designated Router (and the Backup Designated Router) only. For these reasons[32], when one or more of the network's routers are running MOSPF, the Designated Router should be running MOSPF also. This can be ensured by assigning all non-multicast routers the Router Priority of 0.

In MOSPF, the Designated Router also has the additional responsibility of monitoring the network's multicast group membership. This is done by periodically sending Host Membership Queries, and receiving Host Membership Reports in response (see Section 9). This is yet another reason why the Designated Router must be multicast-capable.

## 14.2 Sending Hello packets

This functionality is described in Section 9.5 of [OSPF]. A MOSPF router sets the MC-bit in the Options field of its Hello packets. This indicates that the router is multicast-capable; it does not necessarily indicate the state of the sending interface's **IPMulticastForwarding** parameter (see Section B.2). Setting the MC-bit in Hellos is done strictly for informational purposes. Neighbors receiving the router's Hello packets do not act on the state of the MC-bit. A neighbor's multicast-capability is learned instead during the Database Exchange Process (see Section 14.4).

## 14.3 The Neighbor state machine

This functionality is described in Section 10.3 of [OSPF]. When a neighbor enters state **Exchange**, the neighbor **Database summary list** is initialized (see the OSPF neighbor FSM entry for State: ExStart and Event: Negotiation-Done). This list describes of the portion of the router's link state database that needs to be synchronized with the neighbor. Group-membership-LSAs are included in the neighbor Database summary list if and only if the neighbor is multicast-capable. The neighbor's multicast capability is learned by examining the neighbor's Database Description packets (see Section 14.4).

## 14.4 Receiving Database Description packets

This functionality is described in Section 10.6 of [OSPF]. A neighbor's multicast-capability is learned through received Database Description packets. When the Database Description packet is received that transitions the neigh-

---

32. Group-membership-LSAs are not processed by non-multicast routers (see Section 10.2). Also, if the Designated Router was not running the multicast extensions, multicast datagrams would not be forwarded over the network because its network-LSA would have its MC-bit clear (see Step 5a in Section 12.2).

bor from ExStart to Exchange, the state of the MC-bit in the packet's Options field is examined. The neighbor is multicast-capable if and only if the MC-bit is set.

The neighbor's multicast capability controls whether group-membership-LSAs are summarized to the neighbor during the Database Exchange process (see Section 14.3), and whether group-membership-LSAs are flooded to the neighbor during the flooding process (see Section 10.2).

## 14.5 Sending Database Description packets

This functionality is described in Section 10.8 of [OSPF]. A MOSPF router sets the MC-bit in the Options field of its Database Description packets. This indicates to its adjacent neighbors that the router is multicast-capable; it does not necessarily indicate the state of the sending interface's **IPMulticastForwarding** parameter (see Section B.2).

When a router goes from being multicast-capable to multicast-incapable, or vice-versa, it must indicate this fact to its adjacent neighbors by restarting the Database Description process (i.e., rolling back the state of all adjacent neighbors to Exstart).

## 14.6 Originating Router-LSAs

This functionality is described in Section 12.4.1 of [OSPF]. A MOSPF router sets the MC-bit in the Options field of its router-LSA. This allows the router to be included in datagram shortest-path trees (see Step 5a of Section 12.2).

In addition, MOSPF has introduced a new flag in the router-LSA's **rtype** field: the **W-bit**. When the W-bit is set, the router is included on all datagram shortest-path trees, regardless of multicast group (see Section 12.2.6). Such a router is called a wild-card multicast receiver. The router sets the W-bit when it wishes to receive all multicast datagrams, regardless of destination. This will sometimes be true of inter-area multicast forwarders (see Section 3.1), and inter-AS multicast forwarders (see Section 4).

A router must originate a new instance of its router-LSA whenever an event occurs that would invalidate the LSA's current contents. In particular, if the router's multicast capability or its ability to function as either an inter-area or inter-AS multicast forwarder changes, its router-LSA must be reoriginated.

## 14.7 Originating Network-LSAs

This functionality is described in Section 12.4.2 of [OSPF]. In OSPF, a transit network's network-LSA is originated by the network's Designated Router. The Designated Router sets the MC-bit in the Options field of the network-LSA if and only if both a) the Designated Router is multicast-capable (i.e., running MOSPF) and b) the Designated Router's interface's **IPMulticastForwarding** parameter has been set to a value other than disabled (see Section B.2). When the network-LSA has the MC-bit set, the network can be included in datagram shortest-path trees (see Section 12.2.6).

It is intended that all routers attached to a common network agree on the network's IPMulticastForwarding capability. However, this agreement is not enforced. When there are disagreements, incorrect routing of multicast datagrams can result.

## 14.8 Originating Summary-link-LSAs

This functionality is described in Section 12.4.3 of [OSPF]. Inter-area multicast forwarders always set the MC-bit in the Options field of their summary-link-LSAs, regardless of whether the path described by the summary-link-LSA is actually multicast-capable. Indeed, it is possible that there is no multicast-capable path to the described destination. All other area border routers (ones that are not inter-area multicast forwarders) clear the MC-bit in the Options field of their summary-link-LSAs.

If its MC-bit is clear, the summary-link-LSA will not be used when initializing the candidate list in Sections 12.2.2, 12.2.3 and 12.2.5.

## 14.9 Originating AS external-link-LSAs

This functionality is described in Section 12.4.4 of [OSPF]. Unlike in summary-link-LSAs, an inter-AS multicast forwarder should clear the MC-bit in the Options field of one of its AS external-link-LSAs if it is known that there is no multicast-capable path from the described destination to the router itself. This knowledge may possibly be obtained, for example, from an inter-AS multicast routing algorithm (see Section 4). If the inter-AS multicast forwarder is unsure of whether a multicast-capable path exists between the described destination and the router itself, the MC-bit should be set in the AS external-link-LSA. All other AS boundary routers (ones that are not inter-AS multicast forwarders) clear the MC-bit in the Options field of their AS external-link-LSAs.

If its MC-bit is clear, the AS external-link-LSA will not be used when initializing the candidate list in Section 12.2.4.

When multicast connectivity to an external destination exists, but no unicast connectivity, an AS external-link-LSA can be originated having its MC-bit set and specifying a cost of LSInfinity. Such an AS external-link-LSA will still be used by the multicast routing calculation (see Section 12.2.4). As a result, when a MOSPF router wishes to stop advertising an AS external destination, it must use the premature aging procedure specified in Section 14.1 of [OSPF], rather than simply setting the AS external-link-LSA's cost to LSInfinity.

## 14.10 Next step in the flooding procedure

This functionality is described in Section 13.3 of [OSPF]. Group-membership-LSAs are specific to a OSPF single area, and are flooded to multicast-capable routers only. When flooding a group-membership-LSA, Section 13.3 of the OSPF specification is modified as follows: 1) The list of interfaces examined during flooding (called the **eligible interfaces** in Section 13.3 of [OSPF]) is the set of all interfaces attaching to Area A (the area that the group-membership-LSA is received from), just as for router-LSAs, network-LSAs and summary-link-LSAs. 2) When examining each interface, a group-membership-LSA is added to a neighbor's **link state retransmission list** if and only if both a) Step 1d of [OSPF]'s Section 13.3 is reached for the neighbor and b) the neighbor is multicast-capable. The neighbor's multicast capability is discovered during the Database Exchange process (see Section 14.4).

Note that, since on broadcast networks Link State Update packets are sent initially as multicasts, non-multicast routers may receive group-membership-LSAs. However, non-multicast routers will simply drop the group-membership-LSAs, for reasons of unrecognized LS type (see Step 2 of [OSPF]'s Section 13). Link State acknowledgments for group-membership-LSAs are not expected from non-multicast routers, and group-membership-LSAs will never be retransmitted to non-multicast routers, since the LSAs are not added to these routers' link state retransmission lists (see above paragraph).

For more information on flooding group-membership-LSAs, see Section 10.2.

## 14.11 Virtual links

This functionality is described in Section 15 of [OSPF]. When a MOSPF router (i.e., multicast-capable router) is both an area border router and an endpoint of a virtual link whose other endpoint is also multicast capable, the router must then also be an **inter-area multicast forwarder**. This is necessary to ensure that multicast datagrams will flow through the virtual link's transit area, from one endpoint to the other. When the backbone's datagram shortest-path tree is constructed in Section 12.1, it is assumed that virtual links are capable of forwarding multicast datagrams whenever both endpoints are multicast-capable.

# 15 References


[Bharath-Kumar] Bharath-Kumar, K. and Jaffe, J. M. **Routing to multiple destinations in Computer Networks.** IEEE Transactions on Communications, COM-31[3], March 1983.

[Deering] Deering, S. **Multicast Routing in Internetworks and Extended LANs**. SIGCOMM Summer 1988 Proceedings, August 1988.

[Deering2] Deering, S. **Multicast routing in a datagram internetwork.** Stanford Technical Report STAN-CS-92-1415, Department of Computer Science, Stanford University, December 1991.

[OSPF] Moy, J. **OSPF Version 2.** Internet Draft. March 1993.

[RFC 1075] Waitzman, D., Partridge, C. and Deering, S. **Distance Vector Multicast Routing Protocol.** RFC 1175, November 1988.

[RFC 1112] Deering, S.E.**Host extensions for IP multicasting.** RFC 1112, May 1988.

[RFC 1209] Piscitello, D.M. **Transmission of IP datagrams over the SMDS Service.** RFC 1209, March 1991.

[RFC 1340] Reynolds, J. and Postel, J. **Assigned Numbers.** RFC 1340, July 1992.

[RFC 1390] Katz, D. **Transmission of IP and ARP over FDDI Networks.** RFC 1390, January 1993.

# A Data Formats

This section documents the format of MOSPF protocol packets and link state advertisements (LSAs). All changes and additions made to the OSPF Version 2 data formats have been made in a backward-compatible manner. In other words, multicast routers running MOSPF can interoperate with (non-multicast) OSPF Version 2 routers when forwarding regular (unicast) IP data traffic.

The MOSPF packet formats are the same as for OSPF Version 2 (described in Appendix A of [OSPF]). One additional option has been added to the **Options** field that appears in OSPF Hello packets, Database Description packets and all link state advertisements. This new option indicates a router's/network's multicast capability, and is documented in Section A.1. The presence of this new option is ignored by all non-multicast routers.

To support MOSPF, one of OSPF's link state advertisements has been modified, and a new link state advertisement has been added. The format of the router-LSA has been modified (see Section A.2) to include a new flag indicating whether the router is a wild-card multicast receiver. A new link state advertisement, called the group-membership-LSA, has been added to pinpoint multicast group members in the link state database. This new advertisement is neither flooded nor processed by non-multicast routers. The group-membership-LSA is documented in Section A.3.

## A.1 The Options field

The OSPF Options field is present in OSPF Hello packets, Database Description packets and all link state advertisements. The Options field enables OSPF routers to support (or not support) optional capabilities, and to communicate their capability level to other OSPF routers. Through this mechanism routers of differing capabilities can be mixed within an OSPF routing domain.

When used in Hello packets, the Options field allows a router to reject a neighbor because of a capability mismatch. Alternatively, when capabilities are exchanged in Database Description packets a router can choose not to forward certain LSA types to a neighbor because of its reduced functionality. Lastly, listing capabilities in LSAs allows routers to route traffic around reduced functionality routers, by excluding them from parts of the routing table calculation.

Three capabilities are currently defined. For each capability, the effect of the capability's appearance (or lack of appearance) in Hello packets, Database Description packets and link state advertisements is specified below. For example, the ExternalRoutingCapability (below called the E-bit) has meaning only in OSPF Hello packets.

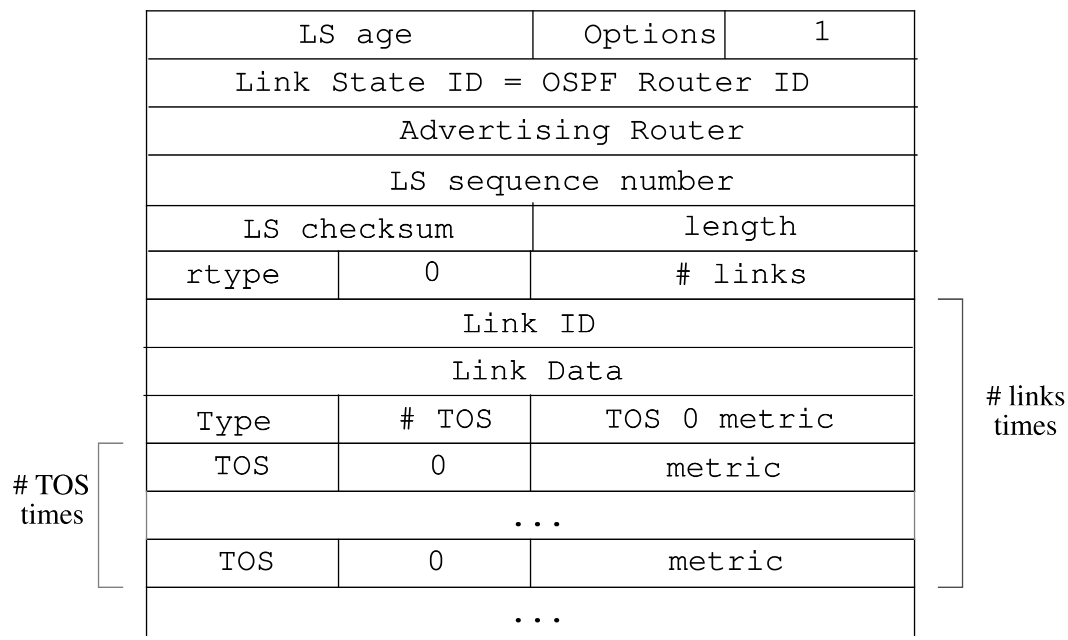
| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | | **MC** | **E** | **T** |

**The OSPF Options field**

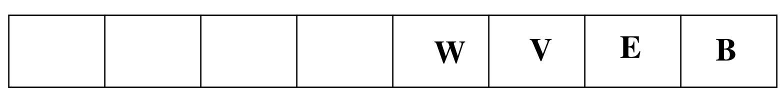
- **T-bit**. This describes the router's TOS capability. If the T-bit is reset, then the router supports only a single TOS (TOS 0). Such a router is also said to be incapable of TOS-routing. The absence of the T-bit in a router links advertisement causes the router to be skipped when building a non-zero TOS shortest-path tree. In other words, routers incapable of TOS routing will be avoided as much as possible when forwarding data traffic requesting a non-zero TOS. The absence of the T-bit in a summary link advertisement or an AS external link advertisement indicates that the advertisement is describing a TOS 0 route only (and not routes for non-zero TOS).
- **E-bit**. AS external link advertisements are not flooded into/through OSPF stub areas. The E-bit ensures that all members of a stub area agree on that area's configuration. The E-bit is meaningful only in OSPF Hello packets. When the E-bit is reset in the Hello packet sent out a particular interface, it means that the router will neither send nor receive AS external link state advertisements on that interface (in other words, the interface connects to a stub area). Two routers will not become neighbors unless they agree on the state of the E-bit.
- **MC-bit.** The MC-bit describes the multicast capability of the various pieces of the OSPF routing domain. When calculating the path of multicast datagrams, only those link state advertisements having their MC-bit set are used. In addition, a router uses the MC-bit in its Database Description packets to tell adjacent neighbors whether the router will participate in the flooding of the new group-membership-LSAs.

## A.2 Router-LSA

An OSPF router originates a router-LSA into each of its attached areas. The router-LSA describes the state and cost of the router's interfaces to the area. The contents of the router-LSA are described in detail in Section A.4.2 of [OSPF]. There are flags in the router-LSA that indicate whether the router is either a) an area border router or b) an AS boundary router or c) the endpoint of a virtual link. One more flag has been added to the router-LSA for MOSPF; it is called bit W below. This flag indicates whether the router wishes to receive all multicast datagrams regardless of destination (i.e., is a wild-card multicast receiver).




**The router LSA**




**The rtype field**

The following defines the flags found in the **rtype** field. Each flag classifies the router by function:

- **bit B**. When set, the router is an area border router (B is for border). These routers forward unicast data traffic between OSPF areas.
- **bit E.** When set, the router is an AS boundary router (E is for external). These routers forward unicast data traffic between Autonomous Systems.
- **bit V**. When set, the router is an endpoint of an active virtual link (V is for virtual) which uses the described area as its Transit area.

- **bit W**. When set, the router is a wild-card multicast receiver. These routers receive all multicast datagrams, regardless of destination. Inter-area multicast forwarders and inter-AS multicast forwarders are sometimes wild-card multicast receivers (see Sections 3 and 4).

## A.3 Group-membership-LSA

Group-membership-LSAs are the Type 6 link state advertisements. Group-membership-LSAs are specific to a particular OSPF area. They are never flooded beyond their area of origination. A router's group-membership-LSA for Area A indicates its directly attached networks which belong to Area A and contain members of a particular multicast group. A router originates a group-membership-LSA for multicast group D when the following conditions are met for at least one directly attached network: 1) the router has been elected Designated Router for the network and 2) at least one host on the network has joined Group D via the IGMP protocol.

A router may also originate a group-membership-LSA for Group D if the router itself has internal applications belonging to Group D. In addition, area border routers originate group-membership-LSAs into the backbone area when there are group members in the router's attached non-backbone areas. See Section 10 for more information concerning the origination of group-membership-LSAs.

| | | | |
|---|---|---|---|
| LS age | Options | 6 | |
| Link State ID = Destination Group | | | |
| Advertising Router | | | |
| LS sequence number | | | |
| LS checksum | length | | |
| Vertex type | | | For each labelled vertex |
| Vertex ID | | | |
| ... | | | |

**The group-membership-LSA**

The group-membership-LSA consists of the standard 20-byte link state header (see Section A.4.1 of [OSPF]) followed by a list of transit vertices to label with the multicast destination. The advertisement's **Link State ID** is set to the destination multicast group address. There is no metric associated with the advertisement. Each transit vertex is specified by its Vertex type and Vertex ID (see Section 12.1 for an explanation of this terminology):

- **Vertex type.** Set equal to 1 for a router, and 2 for a transit network. Note that the only router that may be included in the list is the Advertising Router itself.
- **Vertex ID.** For router vertices, this field indicates the router's OSPF Router ID. For transit network vertices, this field indicates the IP address of the network's Designated Router. Note that the link state advertisement associated with the transit vertex is the LSA whose LS type = **Vertex type**, Link State ID = **Vertex ID** and Advertising Router = the group-membership-LSA's Advertising Router.

# B Configurable Constants

This section documents the configurable parameters used by OSPF's multicast routing extensions. These parameters are in addition to the configurable constants used by the base OSPF protocol (documented in Appendix C of [OSPF]). An implementation of MOSPF must provide the ability to set these parameters, either through network management or some other means.

## B.1 Global parameters

The following parameters apply to the router as a whole.

- **Multicast capability**. An indication of whether the router is running MOSPF. If the router is running MOSPF, it will perform the algorithms as set forth in this specification. Otherwise, the router is still able to run the basic OSPF algorithm (as set forth in [OSPF]), and will be able to interoperate with multicast capable routers (see Section 6.1) when forwarding regular (unicast) IP data traffic.
- **Inter-area multicast forwarder**. This parameter indicates whether the router will forward multicast datagrams between OSPF areas. Such a router summarizes group membership information to the backbone, and acts as a wild-card multicast receiver in all its attached non-backbone areas (see Section 3.1). 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). By default, all multicast-capable area border routers are configured as inter-area multicast forwarders.
- **Inter-AS multicast forwarder**. This parameter indicates whether the router forwards multicast datagrams between Autonomous Systems. Such a router acts as a wild-card multicast receiver in all attached areas (see Section 4). It is also assumed that an inter-AS multicast forwarder runs some kind of inter-AS multicast routing algorithm.

## B.2 Router interface parameters

The following parameters can be configured separately for each of the router's OSPF interfaces. Remember that an OSPF interface is the connection between the router and one of its attached IP networks. 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.

- **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**. The number of seconds between IGMP Host Membership Queries sent out this interface. A multicast-capable router sends IGMP Host Membership Queries only when it has been elected Designated Router for the attached network. See [RFC 1112] for a discussion of this parameter's value.

- **IGMP timeout**. If no IGMP Host Membership Reports have been heard on an attached network for a particular multicast group A after this period of time, the entry *[Group A, attached network]* is deleted from the router's **local group database**. See Section 9 for more information.

# C Sample datagram shortest-path trees

In MOSPF, all routers must calculate exactly the same datagram shortest-path trees. In order to ensure this in internetworks having redundant links, a number of tie-breakers were defined in the MOSPF routing table calculation (see Steps 4 and 5c of Section 12.2, and Sections 12.2.4 and 12.2.7). This section illustrates the use of these tie-breakers on a sample topology.

Three different examples are given. All examples use the same physical topology and the same set of OSPF interface costs (see the left side of Figure 14). The source of the datagram is always Host H1 on the network at the top of the figure (192.9.1.0), and the destination group members are the two hosts labelled with Group Ma at the bottom of the figure. The first case shows an example of intra-area multicast, while the remaining two cases show the influence of OSPF areas on the path of a multicast datagram.

## C.1 An intra-area tree

The datagram shortest-path tree resulting from the intra-area case is shown on the right of Figure 14. The root of the tree is the source network (192.9.1.0), and the leaves are the two routers (RT4 and RT3) directly attached to the stub networks containing Group Ma members.

There are equal-cost paths available to both group members. For the group member on the left, the path could go either through network 10.1.0.0 or through network 10.2.0.0. By the tie-breaking rules, the path through 10.2.0.0 is chosen since it has the larger IP network number (see Step 5c of Section 12.2).

For the group member on the right, the path could go either over Network 10.2.0.0 or over the serial line connecting routers RT2 and RT3. The path over Network 10.2.0.0 is chosen after executing two tie-breaking rules. First, Network 10.2.0.0 is placed on the shortest-path tree before Router RT3 since networks are always chosen over routers (see Step 4 of Section 12.2). Then, given a choice of either Network 10.2.0.0 or Router RT2 for RT3's parent on the tree, Net 10.2.0.0 is again preferred since it is a network (see Step 5c of Section 12.2)

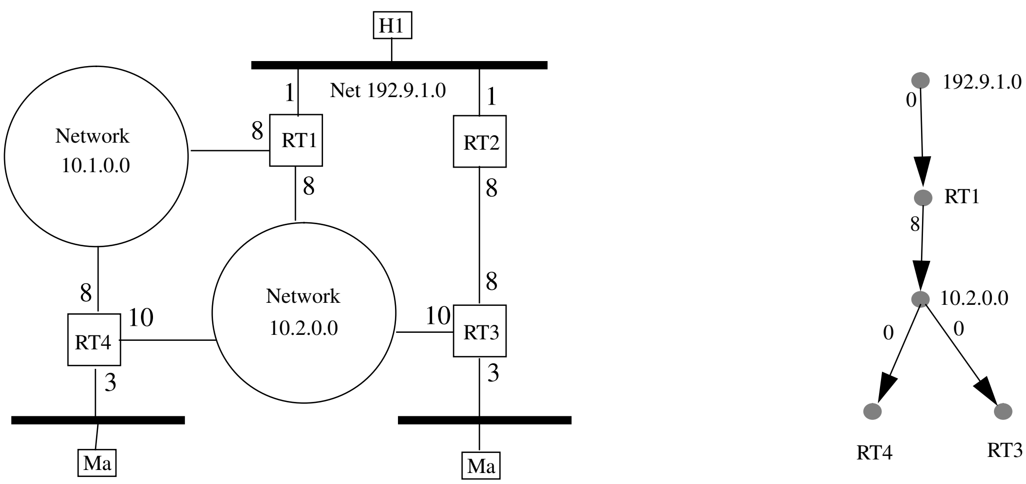



**Figure 14: An intra-area tree**

## C.2 The effect of areas

In Figure 15 below, the previous diagram has been modified by the inclusion of OSPF areas. The datagram source is now part of the OSPF backbone (Area 0), while the rest of the topology is in Area 1. In this case, since the datagram source and the group members belong to different areas, reverse costs are used when building the tree (see Step 5b of Section 12.2). This actually eliminates the equal cost paths from the diagram, and leads to the Area 1 datagram shortest-path tree on the right of Figure 15.

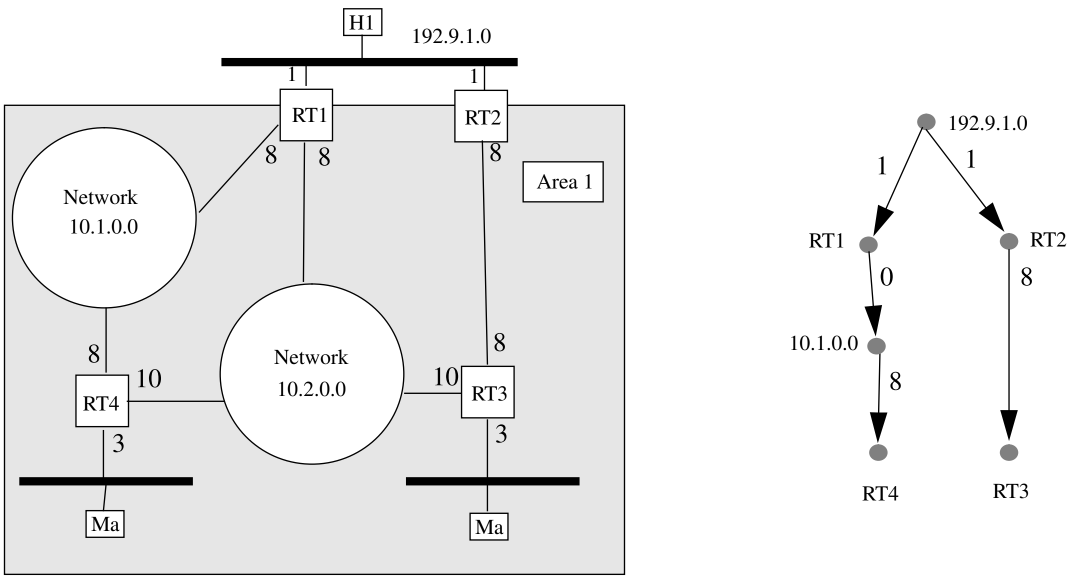



**Figure 15: The effect of areas**

## C.3 The effect of virtual links

In Figure 16 below, Network 10.1.0.0 has been configured as a separate area (Area 1), while everything else belongs to the OSPF backbone (Area 0). In addition, a virtual link has been configured through Area 1, enhancing the backbone connectivity. In this case, both the source and the group members belong to the same area, so forward costs are used. However, since virtual links are preferred over regular links (see Step 5c of Section 12.2), the backbone datagram shortest-path tree uses Network 10.1.0.0 instead of 10.2.0.0 on the path to the left group member. This leads to the tree on the right of Figure 16.

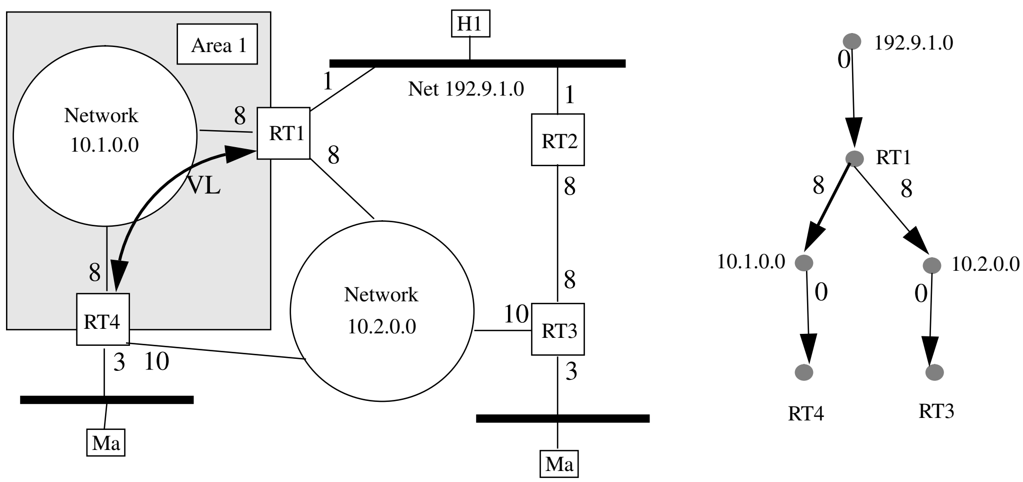



**Figure 16: The effect of virtual links**

## Security Considerations

Security issues are not discussed in this memo.

## Author's Address

John Moy
Proteon, Inc.
9 Technology Drive
Westborough, MA 01581

Phone: (508) 898-2800
Email: jmoy@proteon.com

This document expires in January 1994.