Network Working Group
Internet Draft

J. Moy
Proteon, Inc.
November 1991

Multicast Extensions to OSPF

Status of this Memo

This draft document will be submitted to the RFC editor as a protocol specification. Distribution of this memo is unlimited. Please send comments to mospf@comet.cit.cornell.edu.

Abstract

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

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

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

1 Introduction....4
1.1 Terminology....4
1.2 Acknowledgments....5
2 Multicast routing in MOSPF....5
2.1 Routing characteristics....6
2.2 Sample path of a multicast datagram....6
2.3 MOSPF forwarding mechanism....8
2.3.1 IGMP interface: the local group database....8
2.3.2 A datagram's shortest-path tree....11
2.3.3 Forwarding over Non-broadcast networks....12
2.3.4 Building a forwarding cache entry....13
3 Inter-area multicasting....14
3.1 Extent of group-membership-LSAs....16
3.2 Building inter-area datagram shortest-path trees....18
3.3 Area border routers: Merging datagram shortest-path trees....20
4 Inter-AS multicasting....21
4.1 Building inter-AS datagram shortest-path trees.....22
4.2 Use of the forwarding cache by inter-AS multicast forwarders....23
4.3 Stub area behavior....23
5 Optional capabilities....24
5.1 Mixing with non-multicast routers....24
5.2 TOS-based multicast....25
5.3 Assigning multiple IP networks to a physical network....25
5.4 Networks on Autonomous System boundaries....25
5.5 Recommended system configuration....27
6 Basic implementation requirements....28
7 Protocol data structures....28
7.1 Additions to the OSPF area structure....29
7.2 Additions to the OSPF interface structure....29
7.3 Additions to the OSPF neighbor structure....30
7.4 The local group database....30
7.5 The forwarding cache....30
8 Detailed description of multicast datagram forwarding....31
9 Interaction with the IGMP protocol....34
9.1 Sending IGMP Host Membership Queries....34
9.2 Receiving IGMP Host Membership Reports....34
9.3 Aging local group database entries....35
10 Group-membership-LSAs....35

10.1 Constructing group-membership-LSAs 36
10.2 Flooding group-membership-LSAs 37
11 Detailed construction of datagram shortest-path trees 38
11.1 The SPF calculation 39
11.2 TOS considerations 44
12 Maintaining the forwarding cache 45
13 Other additions to the OSPF specification 46
13.1 The Designated Router (Section 7.3 of [OSPF]) 46
13.2 Sending Hello packets (Section 9.5 of [OSPF]) 47
13.3 The Neighbor state machine (Section 10.3 of [OSPF]) 47
13.4 Receiving Database Description packets (Section 10.6 of [OSPF]) 47
13.5 Sending Database Description packets (Section 10.8 of [OSPF]) 47
13.6 Originating Router-LSAs (Section 12.4.1 of [OSPF]) 48
13.7 Originating Network-LSAs (Section 12.4.2 of [OSPF]) 48
13.8 Originating Summary-LSAs (Section 12.4.3 of [OSPF]) 48
13.9 Originating AS external-LSAs (Section 12.4.4 of [OSPF]) 48
13.10 Next step in the flooding procedure (Section 13.3 of [OSPF]) 49
13.11 Virtual links (Section 15 of [OSPF]) 49
14 References 50
A Data Formats 51
A.1 The options field 52
A.2 Router-LSA 53
A.3 Group-membership-LSA 55
B Configurable Constants 56
B.1 Global parameters 56
B.2 Router interface parameters 56

1 Introduction

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

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

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

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

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

1.1 Terminology

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

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

The following terms are also used throughout this document:

- **Non-multicast router.** A router running OSPF version 2, but not the multicast extensions. These routers do not forward multicast datagrams, but can interoperate with MOSPF routers in the forwarding of unicast packets.
- **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.

In MOSPF, multicast datagrams are routed on the basis of both their source and destination addresses. A multicast datagram is delivered to each member of the destination multicast group, along the shortest path between the source and the group member. Starting as a single packet, the multicast datagram is replicated as the paths to individual members diverge. Packet replication is performed only when necessary. When several group members share an initial path segment, a single datagram instance is forwarded over the common segment.

2.1 Routing characteristics

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

- The path taken by a multicast datagram depends both on the datagram's source and its multicast destination. Called source/destination routing, this is in contrast to most unicast datagram forwarding algorithms (like OSPF) that route based only on destination.
- The path taken between the datagram's source and any particular destination group member is the least cost path available. Cost is expressed in terms of the OSPF link-state metric. For example, if the OSPF metric represents delay, a minimum delay path is chosen. OSPF metrics are configurable. A metric is assigned to each outbound router interface, representing the cost of sending a packet on that interface. The cost of a path is the sum of its constituent (outbound) router interfaces[1].
- MOSPF takes advantage of any commonality of least cost paths to destination group members. However, when members of the multicast group are spread out over multiple networks, the multicast datagram must at times be replicated. This replication is performed as few times as possible (at the tree branches), taking maximum advantage of common path segments.
- For a given multicast datagram, all routers calculate an identical shortest-path tree. There is a single path between the datagram's source and any particular destination group member. This means that, unlike OSPF's treatment of regular (unicast) IP data traffic, there is no provision for equal-cost multipath.
- On each packet hop, MOSPF normally forwards IP multicast datagrams as data-link multicasts. There are two exceptions. First, on non-broadcast networks, since there are no data-link multicast/broadcast services the datagram must be forwarded to specific MOSPF neighbors (see Section 2.3.3). Second, a MOSPF router can be configured to forward IP multicasts on specific networks as data-link unicasts, in order to avoid datagram replication in certain anomalous situations (see Section 5.4).

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 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, H3 and H4 have been added to the figure to serve as sources for multicast datagrams. Again, the datagrams' sources have be made separate from the group members only for clarity of presentation.

---

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 5.2.

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

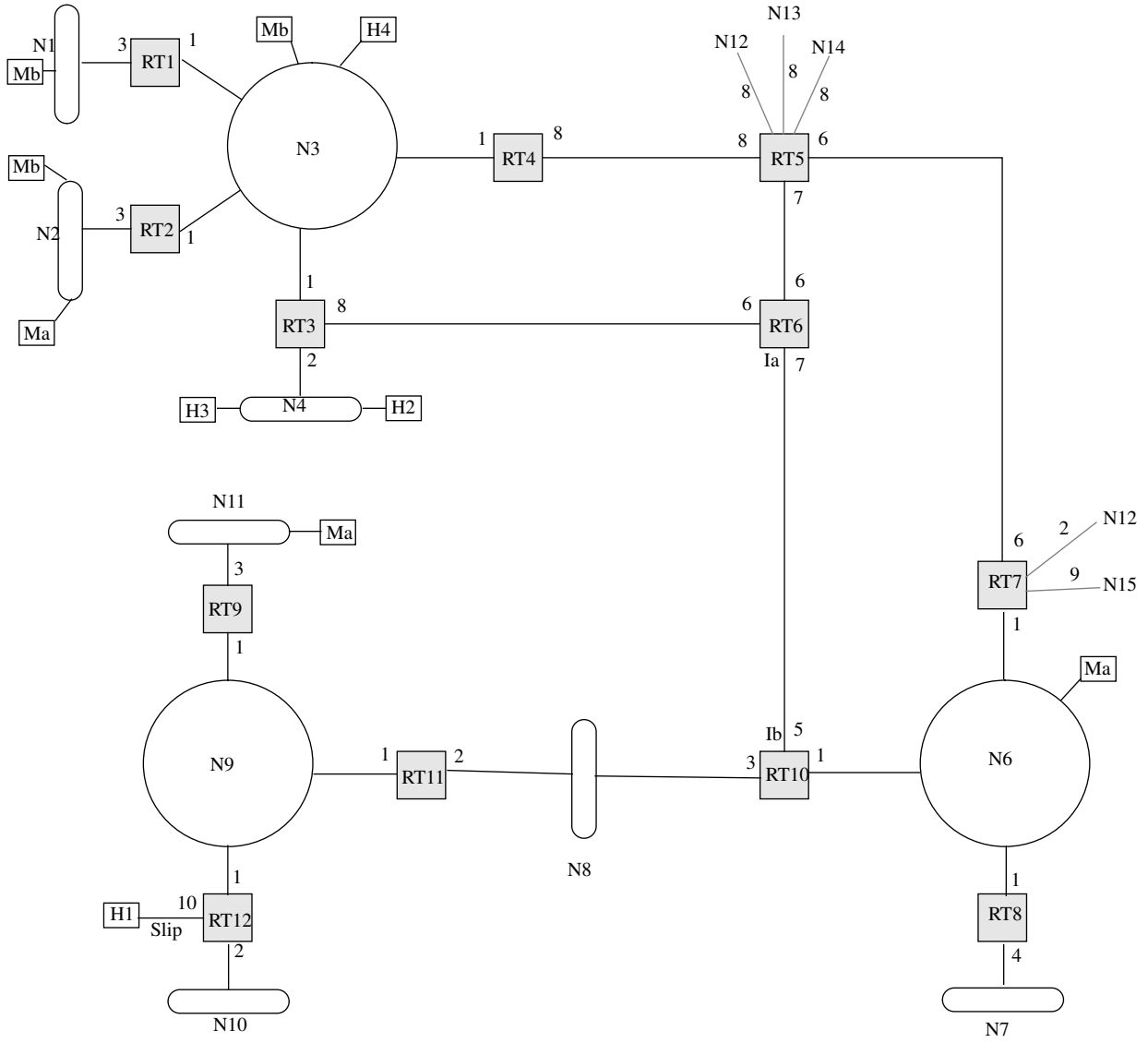


Figure 1: A sample MOSPF configuration

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 multi-

cast, router RT4 will also receive a copy. However, it will not forward the datagram, since it does not lie on a shortest path between the source (host H2) and any members of multicast group B.

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

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

2.3 MOSPF forwarding mechanism

Each MOSPF router in the path of a multicast datagram bases its forwarding decision on the contents of a data cache. This cache is called the **forwarding cache**. There is a separate forwarding cache entry for each source/destination combination[4]. Each cache entry indicates, for multicast datagrams having a particular source and destination, which 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 defines 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 locally attached) group members.

The following sections describe the **local group database** and the **datagram's shortest path tree** in greater detail. They also describe how these two components are combined to form MOSPF's **forwarding cache**.

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 IP multicast destination *group*. This information is then used by the router when deciding

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 5.2.

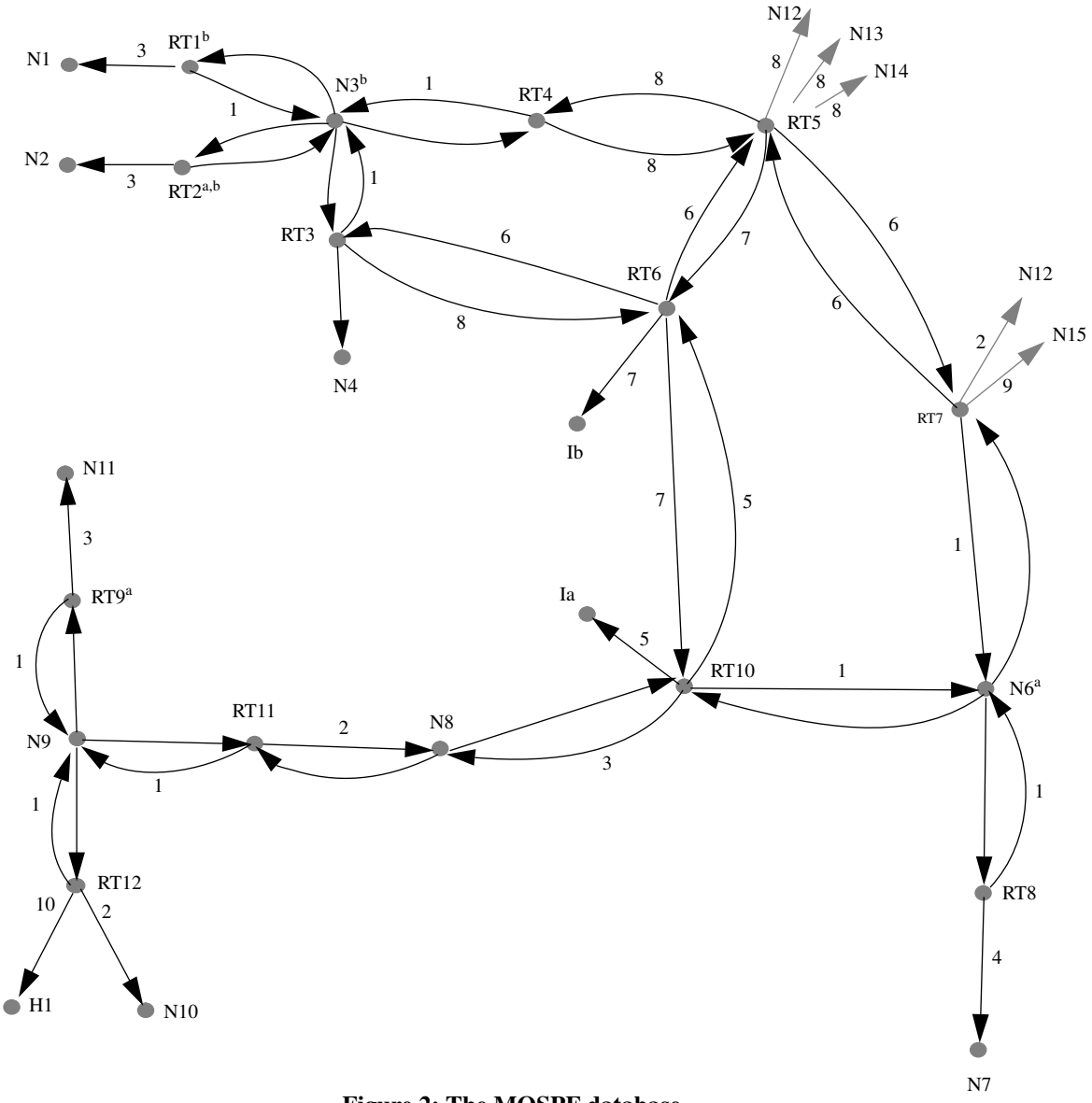


Figure 2: The MOSPF database

which directly attached networks to forward a received IP multicast datagram onto, in order to complete delivery of the datagram to (local) group members (see Section 2.3.4 for more details).

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

later time Reports for group A cease to be heard on the network, the entry is then deleted from the local group database.

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

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

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.

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.

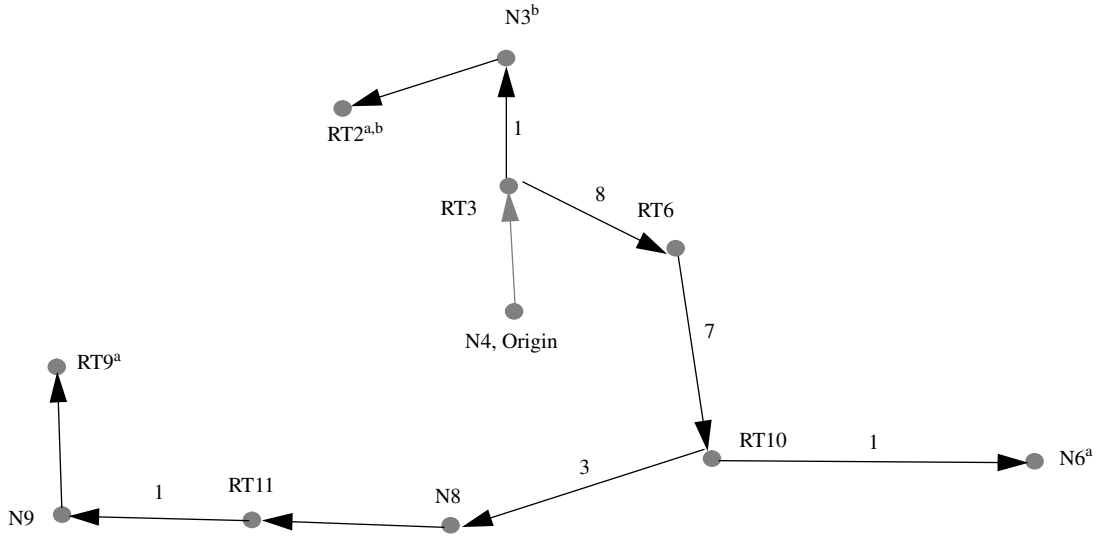


Figure 3: Sample datagram's shortest-path tree, source N4, destination Group A

Figure 2 displays the Autonomous System's link state database. A router/transit network is labelled with a multicast group if (and ony 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 the 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

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.

calculations are performed. First the datagram's source IP network is located in the link state database. Then from those router-LSAs and network-LSAs in the link state database, a shortest-path tree is built using the source network as root. To complete the process, the branches that do not contain routers/transit networks that have been labelled with the particular multicast destination (via a group-membership-LSA) are pruned from the tree.

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

In order to forward the multicast datagram, each router must find its own position in the datagram's shortest path tree. The router's (call it router RT1) position in the datagram's pruned shortest-path tree consists of 1) RT1's parent in the tree (this will be the forwarding cache entry's **upstream node)** and 2) the list of RT1's interfaces that lead to downstream routers/transit networks that have been labelled with the datagram's destination (these will be added to the forwarding cache entry as **downstream interfaces**). Each of the downstream interfaces is also labelled with the number of hops from RT1 to the nearest (in terms of hops) group member.

As an example of a router describing its position on the datagram's shortest-path tree, consider Router RT10 in Figure 3. The **upstream node** for the datagram is router RT6. Router RT10 then has two **downstream interfaces** for the datagram. The first, which connects 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).

The above discussion has not taken into account the effect of OSPF areas on datagram shortest-path trees. Building datagram shortest-path trees for datagrams that will be forwarded across OSPF area boundaries, and/or across Autonomous System boundaries, is more complicated. See Sections 3.2 and 4.1 for details.

2.3.3 Forwarding over 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**. When an downstream neighbor appears in a forwarding cache entry it has an attached hop count, just as the downstream interfaces do. In the process of forwarding a multicast datagram, one or more downstream neighbors may be discovered in the matching forwarding cache entry. The datagram is then forwarded separately (as a unicast) to each listed neighbor whose attached hop count is less than or equal to the TTL appearing in the datagram's IP header.

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.

---

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, a tie-breaking scheme has been implemented during tree construction to ensure that identical trees result. See Section 11 for more details.

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 network 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 would not be a group B member directly attached to network N3.

2.3.4 Building a forwarding cache entry

Recall that forwarding cache entries are built "on demand", whenever a datagram having a new [source net, multicast destination] combination is received. 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. To build a **forwarding cache** entry:

1. The **local group database** is scanned for all entries referencing the particular multicast group destination. Each such local group database entry lists an attached network having members of the particular multicast destination. group. As long as the router is Designated Router on the attached network, then the interface to the attached network is added to the forwarding cache entry as one of the **downstream interfaces**. The hop count (or TTL) associated with each of these added interfaces is set to 1.
2. The **datagram shortest path tree** entry for the [source network, multicast group destination] is built. The datagram's shortest path tree provides the **upstream node** for the cache entry and modifies/adds to list of **downstream interfaces/neighbors** to use.

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

Note that it is possible that in Step 2 above, after calculating the datagram's shortest path tree, a router finds that it itself is not on the tree. This would be indicated by a forwarding cache entry having no upstream node or an empty list of downstream interfaces/neighbors. This negative result is still cached however, and indicates that the router should not forward the multicast datagram.

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.

As an example of forwarding cache entries, again consider the Autonomous System pictured in Figure 1. Suppose Host H2 sends a multicast datagram to multicast group B. In that case, several routers will not even attempt to build a

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

forwarding cache entry (routers RT5 and RT7) because they will never receive the multicast datagram in the first place. Other routers will receive the multicast datagram (since they are forwarded as link-level multicasts), but after building the pruned shortest path tree will notice that they themselves are not a part of the tree (routers RT1, RT4, RT8 and RT12). These latter routers will install an empty cache entry, indicating that they do not participate in the forwarding of the multicast datagram. A sample of the forwarding cache entries built by the other routers in the Autonomous System is pictured in Table 2:

Router	*Upstream node*	*Downstream interfaces (interface:hops)*
RT10	Router RT6	(N6:1), (N8:3)
RT11	Net N8	(N9:2)
RT3	Net N4	(N3:1), (RT6:3)
RT6	Router RT3	(RT10:2)
RT2	Net N3	(N2:1)

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

In an OSPF area border router, the building of forwarding cache entries is more complicated, owing to the fact that a separate datagram shortest-path tree is built for each attached area. In this case, the multiple datagram shortest-path trees, along with the local group database, must be merged into a single forwarding cache entry. See Section 3.3 for details.

3 Inter-area multicasting

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

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

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

It should be noted in the following that the basic protocol mechanisms in the inter-area case are the same as for the intra-area case. Forwarding of multicasts is still defined by the contents of the **forwarding cache**. The forwarding cache is still built from the same two components: the **local group database** and the **datagram shortest-path trees**. And while the calculation of the datagram shortest-path trees is different in the inter-area case (see Sections 3.2 and 3.3), 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 mechanism described in Section 8 holds for both the intra-area and inter-area cases.

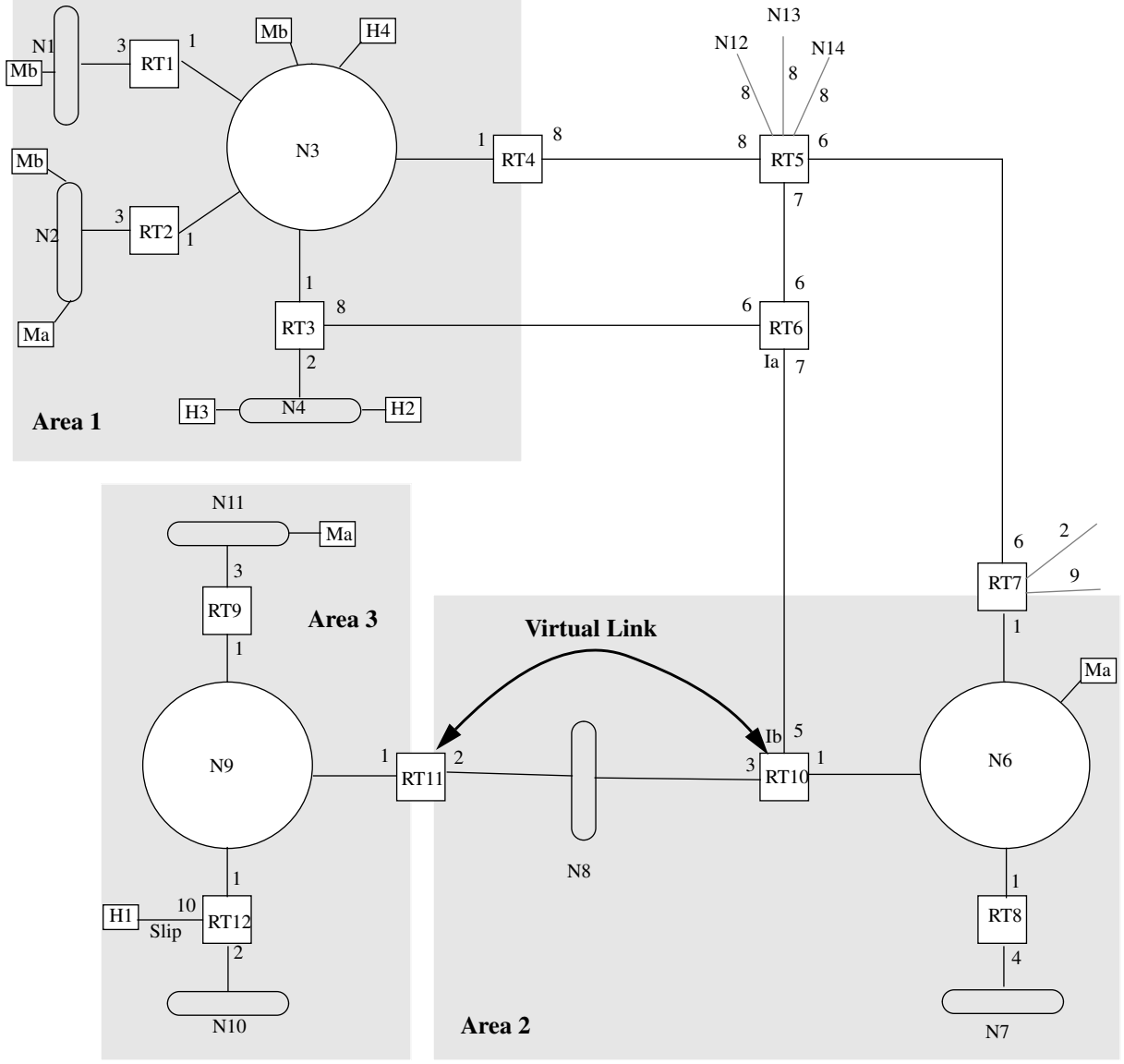


Figure 4: A sample MOSPF area configuration

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.

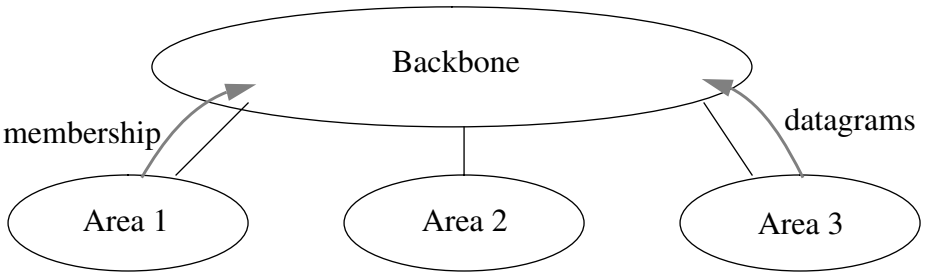


Figure 5: Inter-area routing architecture

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

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

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

In OSPF, the area border routers forward routing information and data traffic between areas. In MOSPF. a subset of the area border routers, called the **inter-area multicast forwarders**, forward group membership information and multicast datagrams between areas. Whether a given area border router is also an 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.

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

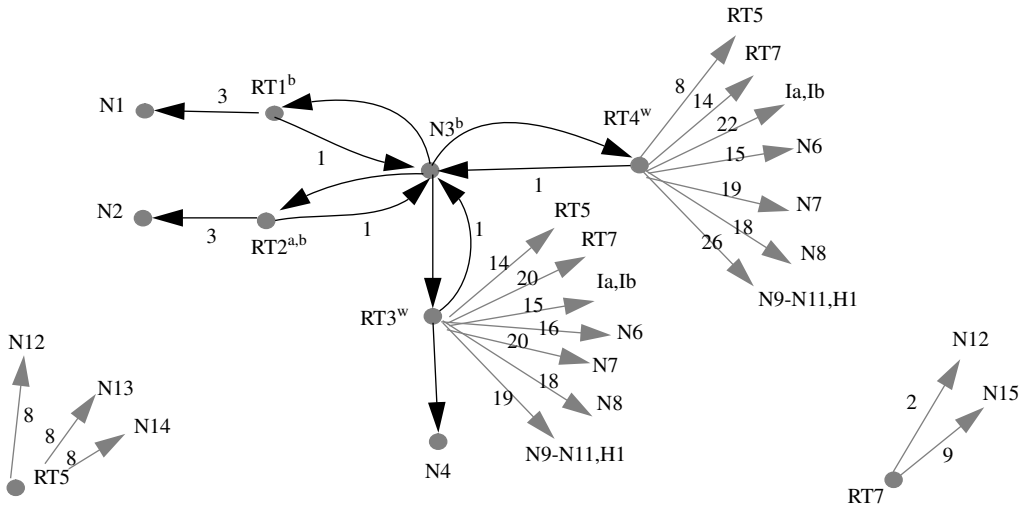


Figure 6: Area 1's MOSPF database

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 border router injects a group-membership-LSA into the backbone area. For example, in Figure 4 there are two groups having members in Area 1: Group A and Group B. For that reason, both of Area 1's inter-area multicast forwarders (Routers RT3 and RT4) inject group-membership-LSAs for these two groups into the backbone. As a result both of these routers are labelled with Groups A and B in the backbone link state database (see Figure 7).

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

At this point, if no extra functionality was added to MOSPF, multicast traffic originating in Area 1 destined for Multicast Group A would never be forwarded to those group A members in Area 3. To accomplish this, the notion of **wild-card receivers** is introduced. A wild-card receiver is a router to which all multicast traffic, regardless of multicast destination, should be forwarded. A router's wild-card multicast reception status is per-area. In non-backbone areas, all inter-area multicast forwarders[10] are wild-card multicast receivers. This ensures that all multicast traffic originating in a non-backbone area will be forwarded to the 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 exam-

10. A router indicates that it is an inter-area multicast forwarder by setting the appropriate flag in its router-LSA. See Section 13.6 for details.

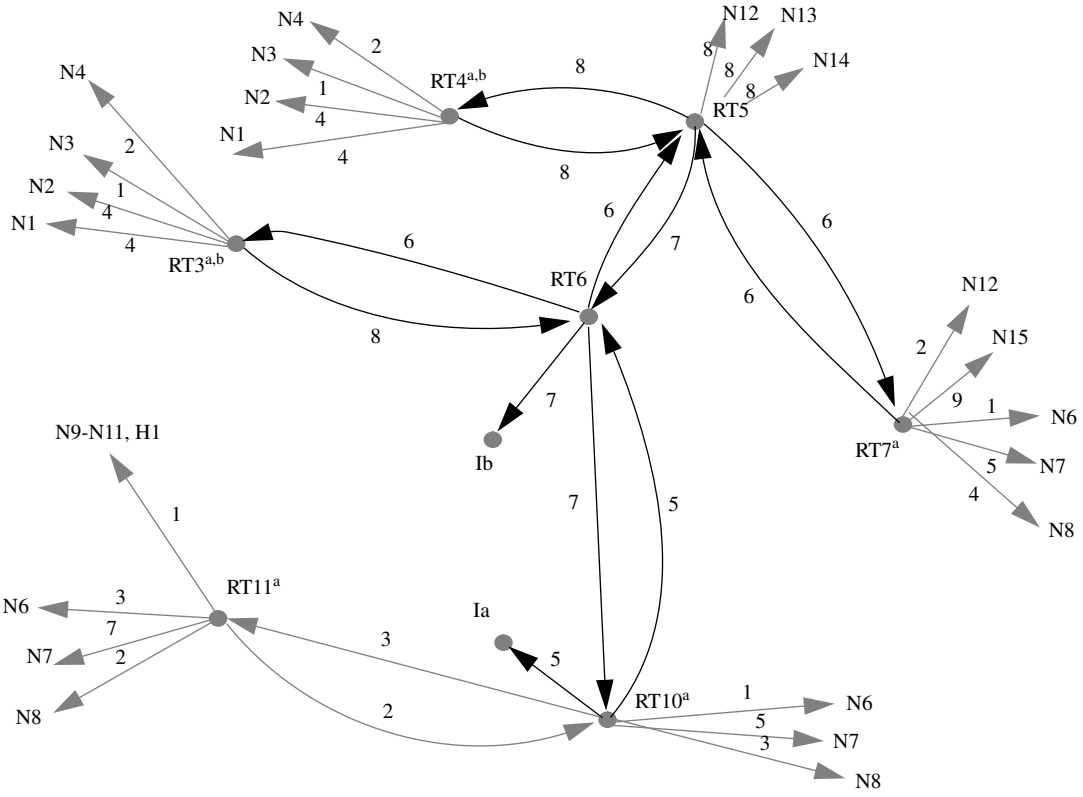


Figure 7: The backbone's MOSPF database

ple, 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 areas into the backbone. Likewise, all IP multicast traffic arising in the non-backbone areas is forwarded to the backbone. Since at this point group membership information meets the multicast datagram traffic, delivery of the datagrams becomes possible.

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-

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.

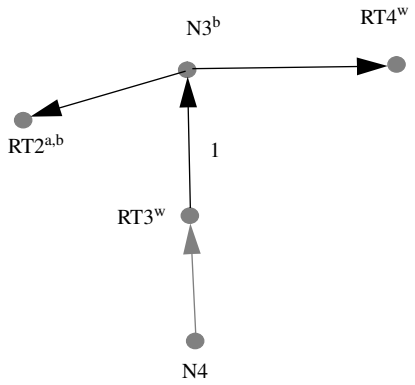


Figure 8: Datagram's shortest-path tree: Area 1, source N4, destination Group A

path trees is often impossible in the inter-area case. To compensate, approximations are made through the use of wild-card multicast receivers and OSPF summary-LSAs.

When it first receives a datagram for a particular [source net, destination group] pair, a router calculates a separate datagram shortest-path tree for each attached area. Each datagram shortest-path tree is built solely from LSAs belonging to the particular area's link state database. Those routers attached to multiple areas must calculate multiple trees and then merge them into a single forwarding cache entry; that is the subject of Section 3.3. Suppose that a router is calculating a datagram shortest-path tree for Area A. It is useful then to separate out two cases.

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

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

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

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

The reverse costs used for the entire tree in Case 2 are forced because summary-LSAs only specify the cost towards the datagram source. In the presence of asymmetric link costs, this may lead less efficient routes when forwarding multicasts between areas. (In fact, in those rare cases where there is asymmetric reachability within the OSPF routing domain, this scheme may even lead to non-delivery of multicast datagrams).

3.3 Area border routers: Merging datagram shortest-path trees

As mentioned above, for a given multicast datagram an area border router will end up computing multiple datagram shortest-path trees, one for each attached area. These datagram shortest-path trees must then be combined in a single **forwarding cache entry**.

Section 2.3.4 explains how a forwarding cache entry is built in the absence of OSPF areas. The router's place in the datagram shortest-path tree determines the **upstream node** and at least some of the **downstream interfaces/neighbors** found in the cache entry. In the inter-area case, the router examines its place in each of the datagram shortest-path trees. Only one of the trees will determine the forwarding cache entry's upstream node. This tree is selected in the following manner[12]:

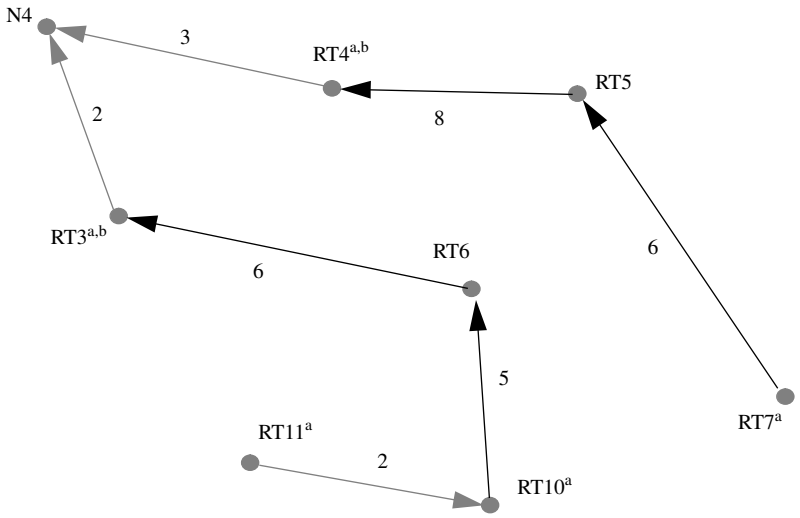


Figure 9: Datagram shortest-path tree: Backbone, source N4, Destination Group A

12. 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].

1. If the router is attached to the area which contains the datagram's source, then the datagram shortest-path tree for said area determines the upstream node.
2. Otherwise, examine the backbone's datagram shortest-path tree. If the router is on this tree, and the link pointing from the router back to the root is neither a virtual link nor a summary link, then the backbone's datagram shortest-path tree determines the upstream node.
3. Otherwise, **exclude** those trees where the link pointing from the router back to the root is either a **virtual link** or a **summary link**. Among those trees remaining, choose the one where the cost from the router to the root is smallest. In the case of ties, choose the tree associated with the area having the largest Area ID.

While only one of the datagram shortest-path trees determines the upstream node, all of the datagram shortest-paths trees contribute to the forwarding cache entry's list of downstream interfaces. The sole exception is that virtual links are never added as downstream interfaces.

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 defined:

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. When an inter-AS multicast forwarder receives a multicast datagram, this specification does not define how the router decides whether the previous datagram hop was a router in another Autonomous System or a MOSPF neighbor. This decision is assumed to be made through the operation of the inter-AS multicast routing protocol, and affects the inter-AS multicast forwarder's forwarding decision for the datagram.
3. MOSPF assumes that the inter-AS routing protocol is forwarding multicast datagrams in an RPF (reverse path forwarding; see [Deering] for an explanation of this terminology) fashion. In other words, it is assumed that a multicast datagram whose source (call it X) lies outside the MOSPF domain will enter the MOSPF domain at those points that are advertising (into OSPF) the best routes back to X. MOSPF calculates the path of the datagram through the MOSPF domain based on this assumption.

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

The presence of inter-AS multicast forwarders is reflected in the contents of the link state database[13]. 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**[14].

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

4.1 Building inter-AS datagram shortest-path trees.

When multicast datagrams are to be forwarded between Autonomous Systems, the datagram shortest-path tree is built as follows. Remember that the router builds a separate tree for each area to which it is attached. These trees are then merged into a single forwarding cache entry as specified in Section 3.3. 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 Autonomous System) having been described in Section 3.2. The only modification to these cases is that inter-AS multicast forwarders, as well as group members and inter-area multicast forwarders, must remain on the pruned trees. The new case is as follows:

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

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

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

---

13. As with inter-area multicast forwarders, inter-AS multicast forwarders indicate that status by setting a flag in their router-LSA. See Section 13.6 for details.

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

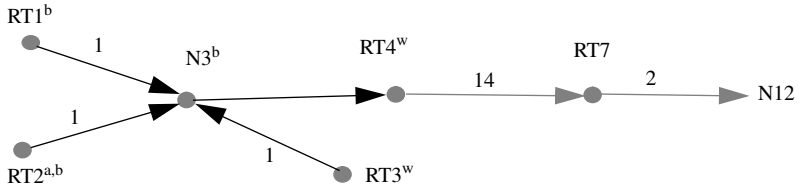


Figure 10: Datagram shortest path tree: Area 1, source N12, Destination Group B

In case 3 (datagram source in another AS) the requirement that all tree links point in the reverse direction (towards the source) accommodates the fact that summary links and AS externals already point in the reverse direction. This also leads to the requirement that the inter-AS multicast routing protocol operate in a reverse path forwarding fashion (see condition 3 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.

4.2 Use of the forwarding cache by inter-AS multicast forwarders

Inter-AS multicast forwarders build their forwarding cache entries using the very same procedures that have been specified in Sections 2.3.4 and 3.3. However, in an inter-AS multicast forwarder (call it Router X) a particular forwarding cache entry's **upstream node** may end up being an external to the AS. This would mean that, in order for Router X to forward the multicast datagram to other MOSPF routers, the pervious datagram hop would have to have been a router belonging to a different Autonomous System.

As an example, suppose that a host on network N12 sends a datagram to Destination Group B. Router RT7, an inter-AS multicast forwarder, will end up calculating an external router (via its external link to network N12) as the upstream node for the datagram. So RT7 must receive the datagram from a router in another AS before forwarding the datagram into the MOSPF system. It is assumed that RT7 can tell that the previous hop was in another AS through the operation of its inter-AS multicast routing protocol.

4.3 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]) hat are originated by the stub area's intra-area multicast forwarders.

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

5 Optional 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 Section B.

5.1 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 host 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 when 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 the needed multicast connectivity exists.
- When forwarding multicast datagrams between areas, mixing of MOSPF routers and non-multicast routers in the source area may cause unexpected loss of multicast connectivity. This is because in the inter-area routing of multicast datagrams the neighborhood of the datagram's source is approximated by OSPF summary link advertisements, and OSPF summary LSAs do not carry indications/guarantees of the summarized path's multicast routing capability.

5.2 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 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 non-zero TOS traffic. Using this logic, all routers in essence continue to utilize identical datagram shortest-path trees. See Section 11.2 for more details.

5.3 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 links 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 8 for details). As a result, unless this situation is dealt with 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 run IGMP (and therefore will not appear in local group database or forwarding cache entries either). As a result, the possibility of unwanted datagram replication is eliminated. The actual disabling of multicast forwarding on a subnet is done through appropriate setting of the **IP multicast forwarding** parameter on all router interfaces connecting to the subnet (see Section B.2).

5.4 Networks on Autonomous System boundaries

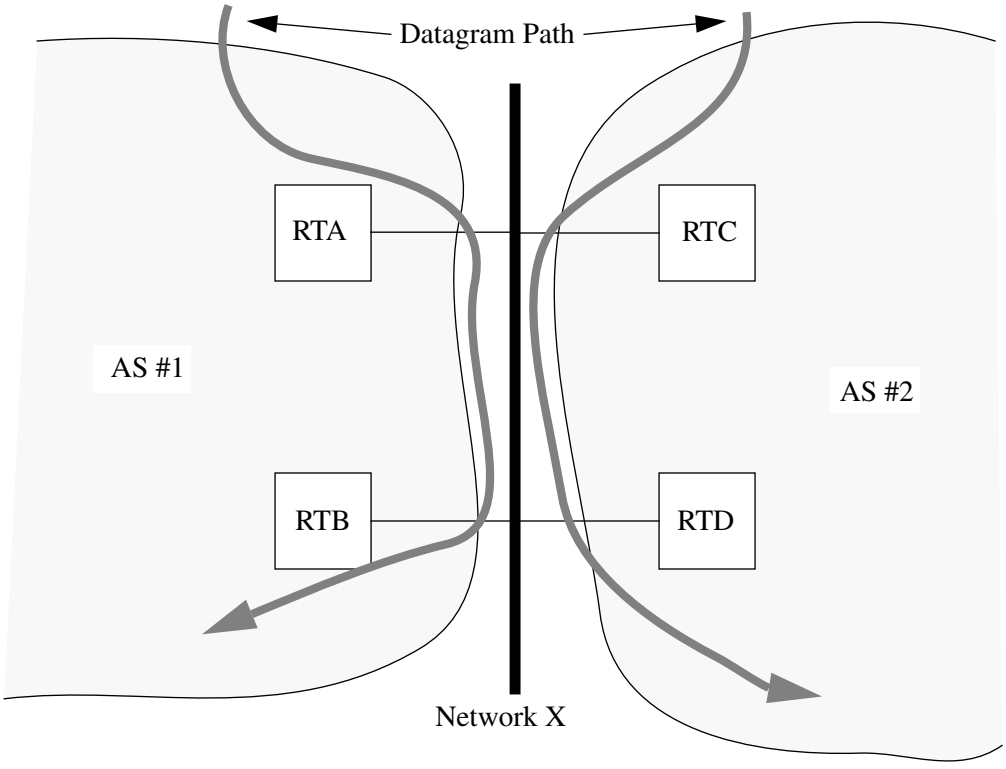
Another complication can arise on IP networks/subnets that lie on the boundary of a MOSPF Autonomous System. Similar to the unicast situation where these networks may be running multiple IGPs (Interior Gateway Protocols), these network may also be running multiple multicast routing protocols. It may 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 11. 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 11, 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 **data-link unicasting** parameter on those router interfaces attaching to the network (see Section B.2). As an example, in Figure 11 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-

Figure 11: Networks on AS boundaries



link unicast, but would reject any data-link multicast forwarded by Router A. This would eliminate replication of multicast datagrams downstream from Network X.

A separate mechanism is necessary to prevent group members attached directly to Network X from receiving multiple datagram copies. By disabling the **IGMP capability** on certain router interfaces (see Section B.2), group membership on the boundary network will be monitored from only one AS, again preventing unwanted replication. For example, in Figure 11 both Router C and Router D could be configured to not send/receive IGMP Host Membership Queries/Reports on Network X. Any group member attached to Network X would then receive its multicast datagram from Router A only.

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.

5.5 Recommended system configuration

In order make MOSPF's selection of routes more predictable, it is recommended that all routers in any particular OSPF area have the same multicast and TOS capabilities.Keeping areas homogeneous ensures that IP multicast packets will follow relatively the same path as IP unicasts. In contrast, while heterogeneous areas will function, and will probably be necessary at least during the initial introduction of multicast routing, such areas may produce seemingly sub-optimal and unexpected routes. For example, see Section 5.1above for a detailed description of the possible pitfalls when mixing multicast and non-multicast routers.

As for the other options presented above, to achieve the most predictable results it is recommended that a router interface's **IP multicast forwarding** parameter only be disabled in those situations when multiple IP networks have been assigned to a single physical wire. And it is recommended that an interface's **IGMP capability** remain enabled and its **data-link unicasting** parameter remain disabled, unless there is the possibility for multiple multicast routing protocols to be running on the attached network.

6 Basic implementation requirements

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

- **Promiscuous multicast reception.** In a multicast router, it is necessary to receive all IP multicasts at the data-link level. On those interfaces where IP multicast datagrams are encapsulated by a wide range of data-link multicast destination addresses (e.g, ethernet and FDDI), this is most easily accomplished by disabling any hardware filtering of multicast destinations (i.e., by "opening up" the interface's multicast filter).
- **Data-link multicast/broadcast detection.** To avoid unwanted replication of multicast datagrams in certain exceptional conditions, it is necessary for the MOSPF forwarding mechanism to determine whether a packet was received as a data-link multicast/broadcast. See Section 5.4 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.

7 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 network 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 8 for more details.
- **Multicast routing capability.** Indicates whether the router will run 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 indicates its special functionality in its router-LSA (see Section 10), and also summarizes its (non-backbone) 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 13.11).
- **Inter-AS multicast forwarder.** Indicates whether the router will forward IP multicasts between Autonomous Systems. Such a router indicates its special functionality in its router-LSA (see Section 13.6). 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 13.9). 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.

7.1 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 non-backbone areas' membership by originating group-membership-LSAs into the backbone area. For more information, see Sections 3.1 and 10.

7.2 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 first three parameters, the *IP multicast forwarding* parameter, the *Data-link unicasting* parameter and the *IGMP monitoring* parameter are really descriptions of the attached network. As such, they should be configured identically on all router interfaces to a common network; otherwise incorrect routing of multicast datagrams may result. However, this synchronization of parameters is not enforced by the MOSPF protocol (e.g., these parameters are not reflected in the contents of a MOSPF router's Hello packets).

- **IP multicast forwarding.** This configurable parameter indicates whether IP multicasts should be forwarded over the attached network. The default value for this value is enabled. It should be disabled only in special circumstances, e.g., when multiple IP networks have been assigned to a single physical network. See Section 5.3 for more details.
- **Data-link Unicasting.** This configurable parameter indicates whether IP multicast datagrams should be forwarded on the attached network as data-link unicasts or as data-link multicasts. By default, they are forwarded as data-link multicasts. For an explanation of circumstances where forwarding as data-link unicasts is appropriate, see Section 5.4.
- **IGMP monitoring.** This configurable parameter indicates whether the router is allowed to monitor group membership on the attached network. By default, monitoring is enabled. See Section 5.4 for an explanation of when group monitoring should be disabled.
- **IGMP polling interval.** When the router is actively monitoring group membership on the attached network, it periodically send IGMP membership queries. *IGMP polling interval* is a configurable parameter indicating the number of seconds between IGMP membership queries. The router actively monitors group membership on the attached network when both a) IGMP is enabled on the associated interface and b) the router has been elected Designated Router for the network. See Section 9 for details.
- **IGMP timeout.** This configurable parameter indicates the length of time (in seconds) that a local group database entry associated with this interface will persist without another group 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 *IGMP polling interval.* See Section 9 for details.

7.3 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 13 for a more detailed explanation.

7.4 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 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:

- **Multicast group**. 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].
- **Attached network.** 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 for 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 **IGMP timeout** 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.

7.5 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 datagrams 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 11.

Each entry in the forwarding cache has six components: the source network (represented as a routing table entry), 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 routing table entry**. The OSPF routing table entry describing the source of matching datagrams (i.e, an IP network). The OSPF routing table is defined in Section 11 of [OSPF]. Note that there can be a separate routing table entry for varying values of TOS; regardless of the matching datagram's TOS, the routing table entry corresponding to the default TOS (TOS 0) is always chosen[15]. The routing table entry provides the following information: 1) the source network's mask, 2) the transit node closest to the source network and 3) an indication of whether the source is in a different OSPF area or in a different Autonomous System than the forwarding router. See Sections 8 and 11 for more details.
- **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 datagrams depends on their 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 from another Autonomous System in order to be forwarded (see Section 4.2).
- **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, the datagram must be forwarded separately to each downstream neighbor (see Section 2.3.3). As done for downstream interfaces, each downstream neighbor is specified together with the smallest **TTL** that will actually reach a group member.

8 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 (like the data-link unicasting option in Section 5.4) 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.

15. See Section 5.2 for an explanation of the affect of TOS on MOSPF routing.

Note that other processes in the router may handle the datagram before the MOSPF forwarding process is reached. For example, an inter-AS routing protocol may pre-process the datagram (see Step 1c below). Also note that the algorithm below covers the case where the router itself belongs to a multicast group, and hence may want to pass the datagram on to one or more application processes. However, unlike the paradigm presented in [RFC 1112], the router will pass the datagram to its application processes only if/when the datagram has been validated for forwarding[16].

The MOSPF forwarding process utilizes the MOSPF forwarding cache (see Section 7.5), and 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 OSPF interface* associated with the datagram. When there is only a single IP network assigned to the receiving physical interface, the choice of receiving OSPF interface is clear. However, when there are multiple IP networks assigned to the receiving physical network, the choice is made as follows[17]. Examine all of the OSPF interfaces associated with the receiving physical network. Discard those interfaces that have been disabled for IP multicast forwarding. Set the *receiving OSPF interface* to be the remaining interface having the highest IP interface address (or NULL if there are no remaining interfaces).

   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 4 to help determine whether the datagram should be forwarded.

   c. Whether the previous datagram hop was a router in another Autonomous System. This information will be used later in Step 3. Note: in the MOSPF forwarding process it is assumed that, if the previous datagram hop was in another AS, some other process in the router (e.g., the routing process associated with an inter-AS multicast routing protocol) has already decided that the datagram should be injected into the MOSPF system (see Section 4).

2. Look up the TOS 0 OSPF routing table entry that best matches the datagram's IP source. See Sections 11 and 11.1 of [OSPF] for a discussion of OSPF's routing table and its lookup algorithm. If no matching routing table entry exists the datagram is discarded, and the forwarding process is completed. Note that the routing table entry (if it exists) yields the datagram's source net and its mask, and is used in the next step to locate the matching forwarding cache entry. However, the exact path that would be used back to the source is unimportant. We call the matching routing table entry the *source routing table entry* in the remaining steps.

3. Look up the **forwarding cache** entry matching the datagram's [source routing table entry, IP multicast destination, TOS] combination. If the cache entry does not yet exist, one is built by calculating the datagram's shortest-path tree and merging this tree with the appropriate local group database entries (see Sections 2.3.4, 3.3 and 11.1). Section 7.5 describes the forwarding cache in detail.

   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:

16. In [RFC 1112], multicast group membership in a host is associated with a particular interface. However, group membership in a MOSPF router is associated with the router as a whole, not separately on each interface. Therefore, delaying datagram delivery to internal applications until the datagram has been cleared for forwarding eliminates unwanted datagram replication.

17. Note the contrast with base OSPF specification. The forwarding of unicast datagrams is based purely on destination and TOS, and so it is not necessary to label unicasts with their receiving OSPF interface. It is however necessary to label received OSPF protocol packets with their associated OSPF interface. But since these packets travel a single hop only, their receiving OSPF interface can be calculated from the protocol packets' IP source address.

a. If the forwarding cache entry's **upstream node** is unspecified (i.e., NULL), then the datagram should be discarded and the forwarding process is complete.

b. Otherwise, suppose that the previous datagram hop was a router in another Autonomous System (see Step 1c). If the forwarding cache entry's upstream node is EXTERNAL, then forward the datagram by proceeding to step 5. If the upstream node is other than external, the datagram is discarded and the forwarding process is completed.

c. Otherwise, if the datagram's receiving OSPF interface does not attach to the forwarding cache entry's specified upstream node (see Step 1a), the datagram is discarded and the forwarding process is complete.

4. If the receiving OSPF interface has been configured for **data-link unicasting**, and if the datagram was received as a data-link multicast/broadcast (see Step 1b), the datagram should be discarded and the forwarding process is complete.

5. At this point the datagram is eligible for forwarding. Before forwarding, the router checks to see whether it is itself is a member of the destination multicast group. If so, a copy of the datagram should be passed to each requesting (internal) application process.

6. 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.

7. Examine each of the **downstream interfaces** downstream. 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 group (i.e., class D address). 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 group (i.e., class D address). See [RFC 1188] 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.

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 when the router refuses to forward a multicast datagram.

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).

Note that MOSPF routers themselves never respond to Host Membership Queries, and never send Host Membership Reports. Instead, when an application in a MOSPF router joins a multicast group, the router originates and appropriate group-membership-LSA (see Section 10.1). As a result, a MOSPF router's application group membership is not tied to a particular interface. In contrast, an OSPF router which is not running the multicast routing extensions (i.e., a non-multicast router) responds to received Host Membership Queries with Host Membership Reports, whenever it is itself a member of one or more multicast groups. In these routers, just as in normal internet hosts, group membership is tied to a particular network interface, as specified in [RFC 1112].

9.1 Sending IGMP Host Membership Queries

Only the network's Designated Router sends Host Membership Queries[18]. 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 **IP multicast forwarding** and **IGMP monitoring** parameters are enabled (see Section B.2). If so, it starts the interface's **IGMP polling timer**. Then, whenever the timer fires (every **IGMP polling interval** 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 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 **IP multicast forwarding** parameter (or the **IGMP monitoring** parameter) is disabled on all but one of the physical network's IP networks, excess multicast membership reporting will result. (Unwanted replication of multicast datagrams may also result; see Section 5.3).

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.

18. When DR is not multicast-router!

An IGMP Host Membership Report concerns membership in a single IP multicast group (call it Group A). The Report is also sent to the Group A address (see [RFC 1112] for details). When an IGMP Host Membership Report, sent on network N[19], 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 not 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 7.4 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.

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 IGMP timeout value, 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 IGMP timeout seconds) 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 interface state changes.

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), network X has been enabled for multicast forwarding and for IGMP monitoring (see Section B.2), and network X contains one or more members of group A.

19. 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 in Step 1a of Section 8.

2. The router is an inter-area multicast forwarder (see Section B.1), and one or more of the router's attached 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.

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, an 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 in 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 nodes) that contain group members. Section 10 listed the conditions requiring the 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 only if Area A is not a stub). 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.

  a. Consider those entries in the local group database that describe Group G membership (see Section 7.4). 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[20], no transit vertex should 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 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, 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 the attached non-backbone areas has group G members (indicated by the presence of one or more advertisements in their link state databases having Link State ID set to Group G[21]), 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.

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[22]. As a general design principle, optional OSPF advertisements are only flooded to those routers that understand them.

---

20. 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.

21. 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.

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

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. Such advertisements are only (reliably) 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[23]. Note however that when flooding 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 construction of datagram shortest-path trees

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. Each tree is "built on demand", i.e, when the first datagram having a particular source network, destination group and TOS combination is received. 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. These separate trees are then combined with the router's local group database to form a **forwarding cache entry** (see Sections 2.3.4 and 3.3). It is the forwarding cache that actually determines the forwarding of the multicast datagrams.

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 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 the candidate list. The algorithm terminates when the candidate list is empty.

However, there are several differences between the construction of datagram shortest-path trees and the OSPF intra-area route calculation (Section 16.1 of [OSPF]). These differences include:

- 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.

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

- In order to avoid unwanted replication of multicast datagrams, MOSPF ensures that, for any given datagram, each router build 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 3 and 4 of Section 11.1.
- 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, there are three variations in the building of datagram shortest-path trees, depending on whether the source is 1) in the same area as the SPF calculation, 2) in a different OSPF area or 3) outside of the MOSPF system altogether. These variations can be seen in Steps 2-4 of Section 11.1.
- 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 OSPF costs were specified in [OSPF] for the unicast case, and sometimes only incomplete information exists for the construction of multicast datagram shortest-path trees. See Steps 2 and 4 in Section 11.1.
- In the multicast case, not only is it necessary to calculate the interface(s) that the datagram should be forwarded out of, but it is also necessary to calculate the nodes that the datagram should be received from. (Another way of saying this is that the router is in the middle of the datagram shortest-path tree, but it is the root of the unicast SPF tree).
- 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[24].
- 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 a non-zero TOS (TOS X) datagram shortest-path tree, the TOS 0 datagram shortest tree is built instead (see Section 11.2). In the unicast case, a TOS-incapable node is added to the TOS X SPF tree using TOS 0 costs.

11.1 The SPF calculation

A router builds a separate datagram shortest-path tree for each attached OSPF area. Input to the calculation is the following set of parameters: the OSPF area through which the datagram is to travel (called **Area A** below), the multicast datagram's source network and matching source routing table entry (called **SourceNet** and **SourceRTE** below[25]; see Section 8 for their determination) and the datagram's multicast destination (called **Group G** below). Note that the

24. 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

25. The names SourceNet and SourceRTE re somewhat misleading, because information concerning the source may have been condensed well beyond the source's IP network. In fact, SourceRTE may well be the default routing table entry (i.e, the DefaultDestination 0.0.0.0).

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 11.2; all calculations and costs in the current section concern TOS 0 only.

The output of the calculation is essentially the router's position in the datagram shortest-path tree. This position, combined with the datagram shortest-path trees of the other attached areas and the router's local group database, provides the router's forwarding cache entry for the datagram (see Sections 2.3.4, 3.3 and 7.5). The router's position is described by the router/network from which the datagram should be received (Section 7.5's **upstream node**), and the interfaces/neighbors that the datagram should be forwarded out of/to (Section 7.5's **downstream interfaces** and **downstream neighbors**). Each downstream interface/neighbor is specified together with a TTL value, which indicates the number of router hops to the nearest Group D member.

As mentioned above, the SPF algorithm is stated in terms of vertices (routers and transit networks). At any particular time, the state of the algorithm consists of those vertices to whom optimal paths have been found (the partially-formed datagram shortest-path tree), and those vertices that are immediately eligible to be placed on the datagram shortest-path tree (these vertices form the candidate list). In addition, each vertex has the following set of state information attached:

- **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. First learned when the vertex is first added to the candidate list, the parent can change until the vertex is installed on the shortest-path tree (see Step 4 below).
- **Incoming link type**. This parameter is set to the type of link that led to Vertex's inclusion on the candidate list. Listed in order of decreasing preference, the possible types are: normal links (either router to router or router to network), virtual links, summary links or external links.
- **Associated interface/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. Calculated when the vertex is added to the candidate list (see Step 4 below). This may result in the list of downstream interfaces/neighbors being updated in Step 3.
- **Cost**. The cost, in terms of the OSPF link state metric, of the current best path from SourceNet to the vertex. As with the parent parameter, this parameter is initially when the vertex is first added to the candidate list (Step 4 below), and can change until the vertex is added to the shortest-path tree. 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 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 (if it is itself a router). See Step 4 below.

The datagram shortest-path tree is constructed in the following steps. The main body of the calculation is in Steps 3 and 4, which are repeated until the candidate list becomes empty. As the algorithm proceeds, the position of the calculating router (call it Router RTX) in the datagram shortest-path tree is discovered: the upstream node is determined when Router RTX adds itself to the SPF tree in Step 3, and the list of downstream interfaces/neighbors is determined in subsequent Step 3 iterations[26].

1. **Initialize the algorithm's data structures.** Clear the SPF tree and the candidate list. Set the upstream node to NULL, and clear the list of downstream interfaces/neighbors.

2. **Locate the source network, and add an initial set of vertices to the candidate list.** SourceRTE describes the best path(s) back to the datagram's source. OSPF routing table entries are described in Section 11 of [OSPF]. Examine SourceRTE's **Path-type** and **Area** fields. Depending on the setting of these two fields, the candidate list is initialized in one of the following ways:

   - **Case 1:** SourceRTE specifies a path-type of **intra-area**, with **Area equal to Area A**. In other words, the datagram's source belongs to Area A. In this case, SourceRTE contains a Link State Origin field (see Section 11 of [OSPF]). This field indicates the link state advertisement (router links or network links) that directly references SourceNet. The vertex described by this LSA (router or transit network) is then added to the candidate list[27]. The vertex' state is set as follows: Parent is set to NULL, associated neighbor/interface is set to NULL, Cost is set to 0 and TTL is set to 0.

   - **Case 2:** SourceRTE specifies a path-type of **inter-area**. In other words, the datagram's source belongs to an OSPF area that is not directly attached to the calculating router. In this case, examine the Area A summary-LSAs which have their Link State ID set to SourceNet. If the originator of such a summary-LSA is an inter-area multicast forwarder (which can be detected by examining the originator's router-LSA; see Section A.2 for details), the originator is added to the candidate list as a router vertex. The vertex' state is initialized as: Parent set to NULL, associated neighbor/interface set to NULL, cost set to that advertised in the summary-LSA and TTL set to 0.

   - **Case 3:** SourceRTE specifies a path type of **intra-area** with Area **not equal to Area A**. In other words, the datagram's source belongs to an OSPF area other than Area A, but which is still directly attached to the calculating router. Of all the destinations described by Area A's summary-LSAs, find the most specific destination that 1) still includes SourceNet and 2) is advertised by a reachable area border router. Call this Destination D[28]. Add each inter-area multicast forwarder that is originating a summary-LSA for Destination D to the candidate list. They are added as router vertices, with Parent set to NULL, incoming link type set to summary link, associated neighbor/interface set to NULL, cost set to that advertised in the summary-LSA and TTL set to 0.

   - **Case 4:** SourceRTE specifies a path type of **type 1 external** or **type 2 external,** and Area A is an OSPF **stub area.** This case reduces to Case 3 above, even though datagram's source is now in a different Autonomous System. When the Case 3 processing is performed, Destination D will end up being the DefaultDestination (0.0.0.0).

---

26. Note that the pruning of branches having no group members actually takes place as Router RTX locates its position on the tree. In other words, an downstream interface is only added if it actually leads to one or more Group D members.

27. 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. This makes the datagram shortest-path tree pictured in Figure 3 somewhat deceiving, as its root is really Router RT3 instead of Network N4.

28. The reason why Destination D is discovered by scanning the summary-LSAs, instead of simply using SourceNet (as was done in case 2), is that routing information may have been collapsed at the Area A boundary. Routers internal to Area A get indication of information collapsing simply by looking at the routing table; however, the area border routers need to look at the link state database. Additionally, Destination D is not discovered simply by looking at the router's configured set of area ranges, because we do not want to depend on the configured area ranges being synchronized across all of Area A's area border routers.

- **Case 5:** SourceRTE specifies a path type of **type 1 external** or **type 2 external**, and Area A is not an OSPF stub area. Examine each of Area A's inter-area multicast forwarders and inter-AS multicast forwarders. 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 summary link and the AS external link is preferred). The vertex' state is set as follows: Parent set to NULL, incoming link type set to summary link if the path is represented as a concatenation of a summary link and an AS external link, incoming link type set to external link otherwise, associated interface/neighbor set to NULL, Cost set to the cost of the shortest path from vertex to SourceNet, and TTL set to 0.

The above cases have been described earlier in this memo. In particular, Section 2.3.2 describes intra-area datagram shortest-path trees (Case 1), Section 3.2 describes inter-area datagram shortest-path trees (Cases 2 and 3) and Section 4.1 describes inter-AS datagram shortest-path trees (Cases 4 and 5).

3. **Move 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 the transit network vertices first. 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. V is labeled with Group G if one of the following holds:

   - V is an inter-area multicast forwarder (determined by examining the router-type field in its router-LSA), and Area A is not the OSPF backbone area (Area ID 0.0.0.0).
   - V is an inter-AS multicast forwarder (determined by examining the router-type field in its router-LSA).
   - 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 12.2 of [OSPF]). 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.

   If Vertex V is the calculating router itself, set **upstream node** according to the following:

   - If Vertex V's **incoming link type** is virtual link, set upstream node to NULL.
   - Else, if Vertex V's **incoming link type** is summary link, set upstream node to NULL. The upstream node in the forwarding cache entry will be set via another area's datagram shortest-path tree. See Section 3.3 for details.
   - Else, if Vertex V's **incoming link type** is external link, set upstream node to EXTERNAL. This indicates that, in order to be forwarded, the datagram's previous hop must be a router belonging to another AS. See Section 4.2 for details.
   - Else, if Vertex V's **parent** is non-NULL, set upstream node to parent.
   - Else, SourceNet is directly connected to the calculating router. Set upstream node to SourceNet.

   Otherwise, if Vertex V is downstream from the calculating router (i.e., Vertex V's **associated neighbor/interface** is non-NULL), and if V is labelled with Group G, Vertex V's associated neighbor/interface is added to the **list of downstream interfaces/neighbors**, attaching Vertex V's TTL value to the added interface/neighbor. If the particular interface/neighbor had already been added to the list of downstream interfaces/neighbors, simply modify the list by setting the interface/neighbor's TTL value to the minimum of its existing TTL value and Vertex V's TTL value.

4. **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, examine the next link in V's LSA.

   b. Examine W's LSA. If W is not multicast-capable (indicated by the MC-bit in the LSA's Options field), it should not be added to any datagram shortest-path tree, and the next link in V's LSA should be considered instead.

   c. If W's LSA does not contain a link back to vertex V, the link under consideration should not be used for multicast traffic, and the next link in V's LSA should be considered instead.

   d. Determine the cost to associate with the link from V to W. If SourceNet belongs to Area A (Case 1 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[29].

   e. If the link cost calculated in Step 4d is LSInfinity, Link L should not be used for multicast datagrams, and the next link in V's LSA should be examined instead.

   f. Calculate the cost from SourceNet to W, when using Link L. It is the sum of V's cost plus the link cost calculated in Step 4d. 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. Otherwise, W is on the candidate list already. If:

      - C is less than W's current cost, update W's parameters on the candidate list as specified below (Step 4g).
      - 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 **incoming link type,** and whichever link has the preferred type is chosen (the preference order of link types is listed in this section's definition of **incoming link type**). If the link types are the same, then the link whose Parent has the higher Vertex ID is chosen. Whenever Link L is chosen, W's parameters are modified as below (Step 4g). 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.

   g. 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 **incoming link type** is set to virtual link if Link L is a virtual link, otherwise incoming link type is set to normal link. W's **Cost** parameter is set to C. If either Vertex V is the calculating router itself, or if Vertex V is downstream from the calculating router (i.e., Vertex V's **associated neighbor/interface** is non-NULL), then Vertex W's TTL and associated interface/neighbor parameters are set as follows:

      - Suppose 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 **associated neighbor/interface** is set to NULL. Otherwise, W's associated interface/neighbor is set to the non-virtual interface connecting V to W which has the smallest cost value. Note that, in the reverse cost (inter-area and inter-AS) cases, this may not be the interface corresponding to Link L.

29. No attempt is made to match the links' two halves. See Step 4g.

However, since W is only concerned with the node it is receiving the datagram from (the **upstream node**; see Section 8), and not with the particular interface the datagram is received on, V is free to pick the sending interface when there are multiple connecting links.

- Suppose Vertex V is downstream from the calculating router, and either a) V is a router or b) V's TTL parameter is greater than 1. In these cases, W's **associated interface/neighbor** parameter is copied 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.
- In the remaining case, V is a transit network, and is directly downstream from the calculating router (i.e., 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 associated interface/neighbor** is set to W itself (a neighbor). Otherwise, W's associated interface/neighbor is set to V's associated neighbor/interface.

5. **If the candidate list is non-empty, go to Step 3.** Otherwise, the algorithm terminates.

Note that, even though the **upstream node** and **list of downstream interfaces** produced above describe the calculating router's position on the pruned tree, the above calculation actually produces the entire (unpruned) shortest-path tree rooted at SourceNet. The pruned tree consists of those routers/networks which appear between SourceNet and vertices labelled with multicast Group G (see Step 3 above).

It is possible that the above calculation will fail to select an upstream node, or fail to select any downstream interfaces/neighbors. This does not necessarily mean that the router will not forward the multicast datagram, because the local group database, the datagram shortest-path trees of other areas and the group membership of internal applications must also be taken into account (see Sections 2.3.4 and 3.3). However, it is instructive to list two of the failure states. First, if on the datagram shortest-path tree the calculating Router is connected to SourceNet via a summary link (seconds and third bullets of Step 3), the upstream node is not set. In this case, the upstream node will be a router/network outside of Area A. Second, if on the datagram shortest-path tree the calculating Router is connected to SourceNet via a virtual link (see Step 3), the upstream node is again not set. In this case, the upstream node will be a router/network in the virtual link's transit area.

The above calculation reveals a limitation of the MOSPF interface parameter: **IP multicast forwarding**. Ideally, this parameter is set identically on all interfaces to a common network. However, the forwarding of multicast datagrams will be prevented on a transit network if and only if the network's IP multicast forwarding parameter is disabled in the network's Designated Router (see Step 4b above and Section 13.1). And the IP multicast forwarding parameter has no effect on point-to-point or stub networks.

11.2 TOS considerations

The previous section 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, 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. Otherwise, the TOS 0 tree is used for all datagrams, regardless of their specified TOS.

The reason for the above logic is so that separate trees can be used for separate TOS, even when TOS-capable and TOS-incapable routers are mixed in an area. The one thing that must be ensured is that there is no possibility of the TOS-incapable routers mistakenly forwarding the datagrams. This is why the TOS 0 datagram shortest-path tree is calculated first.

To determine whether there are any TOS-incapable routers on the pruned TOS 0 tree, the following additions are made to Section 11.1'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 (cases 1-5 of Section 11.1's Step 2, and the first bullet in Section 11.1's Step 4f), and recalculated when/if the vertex' position on the candidate list is modified (second bullet in Section 11.1's Step 4f). 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 pruned TOS 0 tree are TOS-capable if and only if, whenever a vertex labelled with Group G is added to the shortest-path tree (Step 3 in Section 11.1), the value of the vertex' TOS-capable path parameter is TRUE.

When calculating the TOS X datagram shortest-path tree, the algorithm in Section 11.1 is used, with the modifications listed below. In short: 1) The source of the multicast datagram is always located using the TOS 0 routing table entry, 2) When links are added to the candidate list, their TOS X cost are used and 3) If during the calculation a TOS-incapable link/router is encountered, that vertex is ignored. In more detail:

- 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.
- Regardless of the TOS X value, the routing table entry SourceRTE describes the TOS 0 path(s) back to the source. In other words, the routing table lookup for the datagram source is always performed with the default TOS value.
- In Case 3 of Step 2, regardless of TOS X's value, Destination D is determined by looking at the TOS 0 summary-LSAs of those area border routers that are reachable via TOS 0.
- When links are considered for addition to the candidate list (Cases 2-4 of step 2, and Step 4d), their TOS X cost is used. In particular, if any of the summary-LSAs considered for candidate list initialization in Cases 2-4 of Step 2 are TOS-incapable (determined by examining the T-bit in the summary-LSA's Option field), they are discarded. Also, in Case 5 of Step 2, only TOS X paths are considered; these are either TOS X AS external links, or the concatenation of a TOS X summary link and a TOS X external link.
- If the algorithm attempts to add a TOS-incapable router (determined by examining the T-bit in the router-LSA's option field) to the candidate list (Step 4g), Link L is ignored and the next link in the V's LSA is examined instead.

12 Maintaining the forwarding cache

A MOSPF router may, for resource reasons, limit the size of its forwarding cache. Old cache entries can be purged to make room for newer entries, since they can always be rebuilt if 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 built and then purged (i.e., thrashing should be avoided).

The building and maintenance of the forwarding cache has been previously described in Section 2.3.4. There are events that force one or more forwarding cache entries to be deleted. Each of these events (potentially) modify the datagram shortest-path tree calculation for all datagrams having given sources and/or destination groups:

- 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. 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) is different from the previous instance. Among other things, this covers routers and links going up or down, and also routers that change from being multicast-incapable to being multicast-capable.
- When a type 3 summary-LSA (network summary) changes, all forwarding cache entries must be examined. Those entries specifying datagram sources belonging to the range of addresses described by the updated summary-LSA must be deleted. See Cases 2 and 3 of Section 11.1's Step 2.
- When a type 4 summary-LSA (router summary) 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.
- When the content of an AS external-LSA changes, all forwarding cache entries specifying the named external network as datagram source 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 7.4) changes.

13 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].

13.1 The Designated Router (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[30], 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 DRPriority of 0.

30. 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 will not be forwarded over the network because its network-LSA will have its MC-bit clear (see Step 4a in Section 11.1).

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.

13.2 Sending Hello packets (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 that the Hello's sending interface has been enabled for **IP multicast forwarding** (see Section 5.3). 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 13.4).

13.3 The Neighbor state machine (Section 10.3 of [OSPF])

When a neighbor enters state **Exchange**, the neighbor **Database summary list** is initialized (see the neighbor FSM entry for State: ExStart and Event: NegDone). 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 13.4).

13.4 Receiving Database Description packets (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 neighbor 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 13.3), and whether group-membership-LSAs are flooded to the neighbor during the flooding process (see Section 10.2).

13.5 Sending Database Description packets (Section 10.8 of [OSPF])

A MOSPF router sets the MC-bit in the Options field of its Database Description packets. This indicates that the router is multicast-capable; it does not necessarily indicate that the Hello's sending interface has been enabled for **IP multicast forwarding** (see Section 5.3). Setting the MC-bit in Database Description packets indicates the router's multicast-capability to its adjacent neighbors.

Note that 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).

13.6 Originating Router-LSAs (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 4b of Section 11.1).

In addition, MOSPF has introduced two new flags in the router-LSA's **rtype** field: **W1** and **W2**. The W1-bit is set if and only if the router has been configured as an inter-area multicast forwarder (see Section 3.1). The W2-bit is set if and only if the router has been configured as an inter-AS multicast forwarder (see Section 4). When either of these bits are set, the router is included on all datagram shortest-path trees, regardless of multicast group (see Step 3 of Section 11.1). Such a router is called a wild-card multicast receiver.

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 capability to function as either a inter-area or inter-AS multicast forwarder changes, its router-LSA must be reoriginated.

13.7 Originating Network-LSAs (Section 12.4.2 of [OSPF])

In OSPF, the network's network-LSA is originated by the 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 to the network has been enabled for **IP multicast forwarding** (see Section B.2). If the network-LSA has the MC-bit set, the network can be included in datagram shortest-path trees (see Step 4b of Section 11.1).

It is intended that all routers attached to a common network agree on the network's IP multicast forwarding capability. However, this agreement is not enforced. When there are disagreements, the Designated Router wins, since it originates the network-LSA.

13.8 Originating Summary-LSAs (Section 12.4.3 of [OSPF])

Inter-area multicast forwarders always set the MC-bit in the Options field of their summary-LSAs, regardless of whether the path described by the summary-LSA is actually multicast-capable. Indeed, it is possible that there is no multicast-capable path to the described destination. Note however that the setting of the MC-bit does not affect the building of MOSPF datagram shortest-path trees. Summary links are added to the candidate list in Section 11.1's Step 1 without ever looking at their MC-bits.

13.9 Originating AS external-LSAs (Section 12.4.4 of [OSPF])

Unlike in summary-LSAs, an inter-AS multicast forwarder should clear the MC-bit in the Options field of one of its AS external-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 its MC-bit is clear, the AS external-LSA will not be used when initializing the candidate list in Section 11.1's Step 1 Case 5.

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-LSA.

13.10 Next step in the flooding procedure (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-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 13.4).

Note that, because on broadcast networks Link State Updates 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.

13.11 Virtual links (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 11.1, it is assumed that virtual links are capable of forwarding multicast datagrams whenever both endpoints are multicast-capable.

14 References

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

[OSPF] Moy, J. **The OSPF Version 2.** RFC 1247. July 1991.

[RFC 1060] Reynolds, J. and Postel, J.**Assigned Numbers.** March 1990.

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

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

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 is documented in Section A.1. The presence of this new option is ignored by all non-multicast routers.

The format of the router-LSA has been updated to include two new flags indicating whether the router is either an inter-area multicast forwarder and/or an inter-AS multicast forwarder. Also an additional link state advertisement, the group-membership-LSA, has been added. This 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.

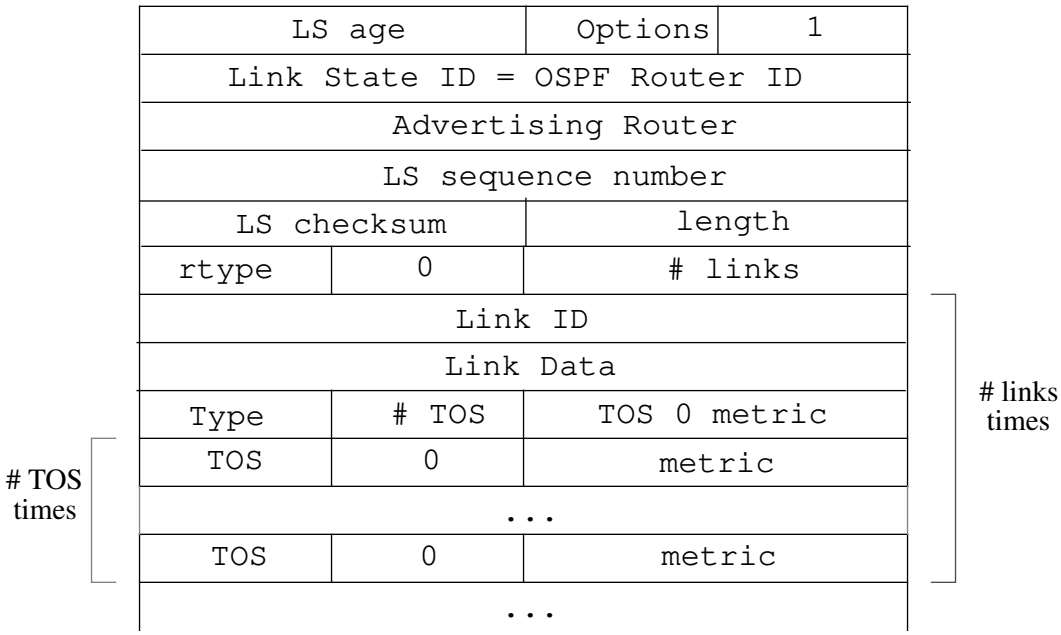
Three capabilities are currently defined. For each capability, the effect of the capability's appearance (or lack of appearance) in Hello packets, Database Description packets and link state advertisements is specified below. For example, the external routing capability (below called the E-bit) has meaning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

The OSPF options field

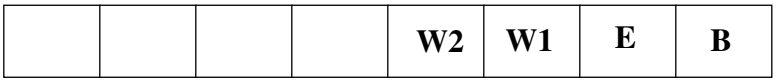
- **T-bit**. This describes the router's TOS capability. If the T-bit is reset, then the router supports only a single TOS (TOS 0). Such a router is also said to be incapable of TOS-routing. The absence of the T-bit in a router links advertisement causes the router to be skipped when building a non-zero 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**. Type 5 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 type 5 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 in order to tell adjacent neighbors whether the router will participate in the flooding of the new group-membership-LSAs.

A.2 Router-LSA

The OSPF router links advertisement indicates which of the router's interfaces are operational, how much they cost, and to what they connect. The contents of the router-LSA is 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. Two more flags have been added to the router-LSA for MOSPF. The first flag (labelled W1 below) indicates whether the router is also an inter-area multicast forwarder. The second flag (labelled W2 below) indicates whether the router is also an inter-AS multicast forwarder.



The router LSA



The rtype field

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

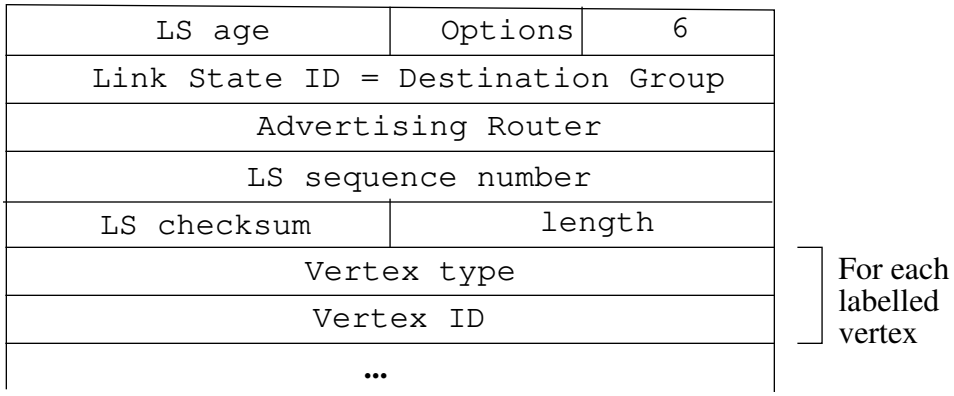
- **bit B**. When set, the router is an area border router (B is for border). These routers forward unicast data traffic between OSPF areas.
- **bit E.** When set, the router is an AS boundary router (E is for external). These routers forward unicast data traffic between Autonomous Systems.

- **bit W1**. When set, the router is an inter-area multicast forwarder. These routers forward multicast datagrams between OSPF areas. They function as wild-card multicast receivers to their attached non-backbone areas (hence the W).
- **bit W2**. When set, the router is an inter-AS multicast forwarder. These routers forward multicast datagrams between Autonomous Systems. They function as wild-card multicast receivers to all of their attached areas (hence the W).

A.3 Group-membership-LSA

Group-membership-LSAs are the type 6 link state advertisements. A router's group-membership-LSA indicates those directly attached networks which contain members of a particular multicast group. A router originates a multicast LSA for multicast group D when the following conditions are met for at least one 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. The group-membership-LSA specifies a list of transit graph vertices that should not be pruned when building any of group D's datagram shortest-path tree (see Section 2.3.2). The router includes itself in this list if it has one or more stub networks with group D members. The router also includes in this list all transit networks having group D members. See Section 10 for more information.

An area border router will also originate a group-membership-LSA for the backbone area when there are group D members in the router's attached non-backbone areas. Group-membership-LSAs are specific to a particular OSPF area. They are never flooded beyond their area of origination.



The group-membership-LSA

The multicast 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 multicast destination group address. There is no metric specified for the advertisement. Each transit vertex is specified by its Vertex type and Vertex ID (see Section 16.1 of [OSPF] 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 can be located in the link state database by setting the LS type = **Vertex type**, Link State ID = **Vertex ID** and Advertising Router = the Advertising Router mentioned in the group-membership-LSA's link state header.

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 capable of IP multicast routing. If the router is capable of multicast routing, 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 5.1) when forwarding regular (unicast) IP data traffic.
- **Inter-area multicast forwarder**. This parameter indicates whether the router forwards 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 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 13.11). 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 first three parameters, the **IP multicast forwarding** parameter, the **Data-link unicasting** parameter and the **IGMP monitoring** parameter are really descriptions of the attached network. As such, they should be configured identically on all router interfaces to a common network; otherwise incorrect routing of multicast datagrams may result.

- **IP multicast forwarding**. If this parameter is disabled, multicast datagrams will not be forwarded over the attached network. The router will then reset the **MC-bit** whenever it originates a network LSA for the network, causing the network to be excluded from all multicast datagram shortest-path trees. Also, multicast group membership will not be monitored for the network. This parameter is used when multiple IP networks have been assigned to the same physical network (see Section 5.3). This parameter can be disabled only on interfaces connecting to transit networks (see Section 11.1). By default, this parameter is enabled on all interfaces.
- **Data-link unicasting**. When this parameter is enabled, IP multicasts forwarded over the attached network will be forwarded as data-link unicasts instead of as data-link multicasts. This may be necessary on those network belonging to multiple Autonomous Systems (see Section 5.4). By default this parameter is disabled on all interfaces.

- **IGMP monitoring**. This parameter indicates whether the router is allowed to monitoring group membership on the attached network. It can be disabled in those cases where the administrator wants to be able to forward multicasts over the network, but still deny the network any directly attached group members (see Section 5.4). By default this parameter is enabled on all interfaces.
- **IGMP polling interval**. The number of seconds between IGMP membership queries sent out this interface. A router sends IGMP 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.
- **I GMP timeout**. If no IGMP membership reports have been heard on an attached network for a 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.