Hierarchical Distance-Vector Multicast Routing for the MBone *

Ajit S. Thyagarajan
Department of Electrical Engineering,
University of Delaware,
Newark, Delaware 19716
ajit@ee.udel.edu

Stephen E. Deering
Palo Alto Research Center,
Xerox Corporation,
Palo Alto, California 94304
deering@parc.xerox.com

Abstract: The exponential growth of the Multicast Backbone (MBone) has resulted in increased routing overhead and processing costs. In this paper we propose a two-level hierarchical routing model as a solution to this problem. This approach involves partitioning the MBone into non-overlapping regions using DVMRP as the inter-region routing protocol; intra-region routing may be accomplished by any of a number of existing multicast protocols. Our design is flexible enough to accommodate additional levels of hierarchy, and protocols other than DVMRP at the higher levels. The unique feature of this approach is the independence of the higher level routing protocol from the subnet addresses, which allows for easy incremental deployment with small changes to existing intra-region protocols.

1 Introduction

The Internet Multicast Backbone or *MBone* [Cas94] is an interconnected set of routers and subnets that provides IP multicast delivery in the Internet [RFC1112]. The MBone was established in early 1992 [Cas92], initially interconnecting 40 subnets (mostly Ethernets) in four different countries. In the three years since then, the MBone has grown exponentially to reach more than 1500 subnets in 25 countries (see Figure 1). This growth has been driven by the availability of popular new multicast applications, particularly those providing real-time audio and video teleconferencing capabilities to Internet hosts. With new multicast applications and multicast-based services appearing regularly, with multicast support becoming available on more types of host and router platforms, with sufficient bandwidth for typical MBone applications becoming more plentiful, and with the population of potential IP multicast users increasing as the Internet itself grows, it appears likely that the MBone will continue to grow at an ever increasing rate.

In order to decide where to forward IP multicast packets, the MBone routers run a *multicast routing protocol*. The protocol that is used in most of the MBone routers today

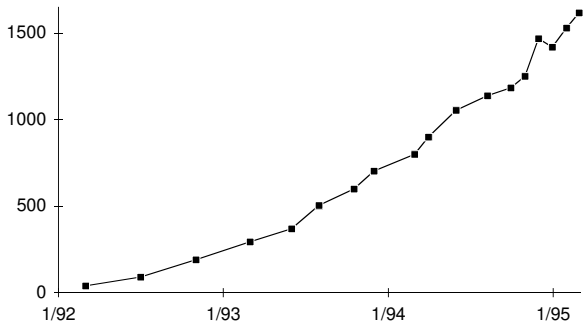

Figure 1: Growth of the MBone—Number of Subnets

is DVMRP, the Distance-Vector Multicast Routing Protocol [Dee91]. DVMRP routers treat the MBone topology as a single, "flat" routing domain. That means that each DVMRP router maintains a separate entry in its routing table for every subnet in the MBone, exchanges periodic routing messages with each of its neighbors identifying every subnet, and is susceptible to additional processing costs whenever there is a topological change anywhere in the MBone. As the number of subnets grows exponentially, those routing costs grow exponentially. If nothing is done about it, the memory and/or processing resources of the current DVMRP routers will soon be exceeded and MBone routing will collapse.

This problem is well-known in the realm of unicast routing, as is its solution: hierarchical routing [Kle77]. In hierarchical routing, a topology is logically partitioned into a number of separate routing domains, each running its own instance of a routing protocol; another routing protocol (or another instance of the same protocol) is used for routing from domain to domain. The routing protocol in each domain maintains detailed topological information only for its own domain, not for other domains, while the inter-domain protocol maintains information only about the interconnection of domains, not about their internal topologies. By applying such hierarchical partitioning recursively, (i.e., partitioning domains into subdomains, subdomains into subsubdomains, and so on), it is possible to provide routing in an arbitrarily large topology with only a bounded amount of topological information maintained at any one router. An example of this is unicast routing in the Internet, which started as flat routing among "networks" [RFC791], evolved into a two-level hierarchy of "networks" and "subnets" [RFC950], and

* This research was supported in part by the Advanced Research Projects Agency under contracts DABT63-94-C-0073 and NAG 2-638, and National Science Foundation grant NCR-93-01002. The views expressed here do not reflect the position or policy of the U.S. government.

is currently in the process of adding one or more additional levels of hierarchy through the mechanism of Classless Inter-Domain Routing (CIDR) [RFC1519].

In this paper, we propose the use of hierarchical routing for the MBone, initially using a two-level hierarchy of *regions* containing subnets. Intra-region multicast routing may be accomplished by any of a number of protocols, including DVMRP. Inter-region routing uses a modified version of DVMRP that computes multicast routes among regions rather than among subnets. Our design accommodates the eventual addition of more levels of hierarchy and the use of protocols other than DVMRP at any level.

In [Dee91], Deering describes an approach to hierarchical multicast routing based on hierarchy encoded in the (unicast) source addresses of multicast packets; it is a straightforward application of the traditional approach to hierarchical unicast routing, which also uses hierarchy encoded in addresses (though significant in *destination* addresses, rather than source addresses). While Deering's scheme would work in the MBone, in this paper we propose a novel alternative approach, employing region identifiers that are *not* encoded in addresses and using encapsulation for inter-region forwarding of packets. Our alternative is more amenable to incremental deployment and has lower routing overhead. Those benefits are particularly important for the MBone, which itself is being incrementally deployed (i.e., not all Internet routers support multicast routing yet), and in which many of the routers are general purpose hosts that are less able to cope with large routing overheads than are dedicated routers.

While the primary—and most urgent—motivation for deploying hierarchical routing in the MBone is reduction of the amount of topological information that MBone routers must store and exchange with other routers, important additional benefits derived from hierarchical routing are the following:

- The partitioning into regions allows multiple, different multicast routing protocols to be operated in different regions of the MBone, with a clean interface between them. The region boundaries isolate the damage that can be done by a misbehaving routing protocol, and allow new protocols to be tested and deployed on a region-by-region basis.

- The effects of low-level topology changes, such as link or router failures and recoveries, are isolated to only those routers in the same topological region as the affected components. Similarly, the effects of changes to the high-level interconnection topology of regions are limited only to the inter-region routers. Thus, the normal low degree of volatility within any region of the MBone does not sum to a high degree of volatility across the entire MBone, as is currently the case. This is particularly beneficial when using a distance-vector style routing protocol, such as DVMRP, which can incur long convergence times after some topology changes (due to the so-called "counting to infinity" problem [RFC1058]).

- Limits on the maximum diameter of the topology imposed by some routing protocols (including DVMRP, to ameliorate the counting-to-infinity problem) can be relaxed by using hierarchical routing. As different instances of routing protocols are used for different parts of the MBone, the limits apply only to those parts, not to the whole.

- Some new multicast routing protocols [Bal93, Dee94] do not have their own topology-determining component, but rather depend on the information provided by the Internet's unicast routing protocols. This approach works correctly only when the multicast and unicast topologies are identical which, as described in the next section, is not now the case for the MBone. Hierarchical routing enables the use of these new protocols within regions that satisfy the topological constraint, without waiting for it to be satisfied everywhere.

In the next section, we describe how the MBone is currently structured and what the consequences of that structure are for multicast routing. In section 3, we describe in detail our hierarchical DVMRP proposal. In section 4 we identify the costs of the proposed routing scheme, and in section 5 we discuss a number of related issues, before concluding in section 6

2 Current MBone Structure and Routing

Since IP multicast is a relatively recent addition to IP, very few of the installed Internet routers currently support multicast routing; most of the MBone routing and forwarding is being performed by general-purpose hosts, such as Unix workstations, running multicast routing software. An example is shown in Figure 2, where host H1 is serving as a multicast router between subnets (e.g., LANs) S1 and S2 because router R1 is not yet multicast-capable. H1 forwards IP multicast packets as required between S1 and S2, including those relayed to and from S3 by multicast-capable router R2. In this paper, we use the term *multicast router* (or just *router*, when the context is clear) to refer to either a multicast-capable Internet router or a host running multicast routing software.

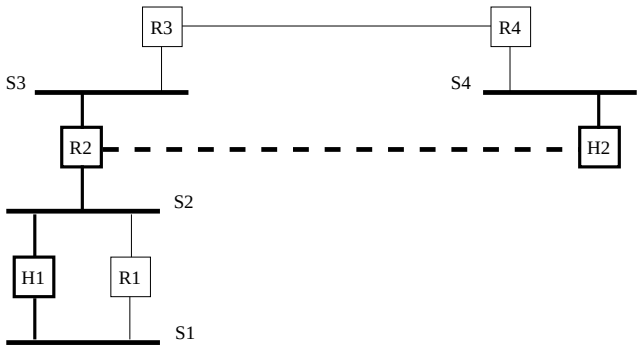

Figure 2: MBone Components

"Islands" of multicast routing capability, such as the S1–H1–S2–R2–S3 conglomerate in Figure 2, are connected to other islands via *tunnels* through non-multicast-capable parts of the Internet. A tunnel is a path between a pair of multicast routers over which a multicast packet is conveyed by encapsulating it inside a conventional unicast packet, addressed to the router at the far end of the tunnel. Tunnels are "installed" simply by manually configuring the router at each end with the IP address of the other end. The figure shows a tunnel configured between multicast routers R2 and

H2. R2 forwards a multicast packet to H2 by placing it inside a unicast packet addressed to H2, which normal Internet unicast routing will deliver to H2 via (multicast-ignorant) routers R3 and R4; the analogous process happens in the reverse direction for multicast packets forwarded from H2 to R2. The tunnel acts as a virtual point-to-point link between R2 and H2 (shown as a dashed line in the illustration) connecting the pair of MBone routers and, thus, the islands in which they reside. The set of multicast routers, the subnets to which they directly attach, and the interconnecting tunnels constitute what is known as the MBone.

The MBone may be thought of as a virtual network (or an overlay network) layered on top of the Internet. As more of the Internet routers are upgraded to support multicast routing, hosts will be retired from doing multicast routing, and tunnels are eliminated. The MBone will then become "less virtual" and eventually become one with the (unicast) Internet. Until that time, the topology of the MBone will differ from the topology of the Internet. In Figure 2, the unicast topology consists of all the boxes and solid lines, while the multicast topology is made up of only the thick-outlined boxes and the thick lines, including the dashed line.

Due to the difference in the MBone and the Internet topology, the multicast routers must run their own topology-discovery protocol, i.e., routing protocol, in order to decide where to forward multicast packets. In most of the current MBone, the protocol is the DVMRP. In some parts of the MBone, routers use different multicast routing protocols, in particular MOSPF [RFC1584] and PIM [Dee94]; these routers interoperate with DVMRP routers by implementing a subset of DVMRP—just enough to "fool" a DVMRP router into "doing the right thing".

MOSPF and PIM are able to do some information reduction on the routing data that they pass to DVMRP, in particular, replacing a set of subnet routes with a single network route, thus exploiting the hierarchy encoded in IP addresses. However, an examination of the current set of MBone routes reveals that doing such data reduction on all routes would result in a savings of only about 30%, rather than the orders-of-magnitude reduction necessary to survive the current growth rate. This modest reduction is due to the fact that, at this point in its deployment, very few networks (which correspond roughly to organizations or sites) have more than a single subnet on the MBone, i.e., most networks contain only one subnet, with respect to the MBone topology. It is for this reason that we propose a hierarchical multicast routing scheme based on address-independent region identifiers, as described in the next section.

3 Description of Hierarchical DVMRP

The MBone is currently organized as one flat "region", with most routers maintaining explicit routing information for every subnet in the network. We propose to organize the MBone into multiple, non-overlapping regions, each region being assigned a unique *region identifier*. Figure 3 shows such a network partitioned into four regions. Since *routing domains* are normally associated with unicast routing, this paper uses the word "region" to describe clusters of routers and subnets, each region encompassing one or more unicast routing domains. If the process of introducing hierarchy into the MBone were to be performed incrementally, there would be a small number of regions initially, each consisting of a large number of routing domains and as support for hierarchy in the MBone increases, these regions would shrink in size.

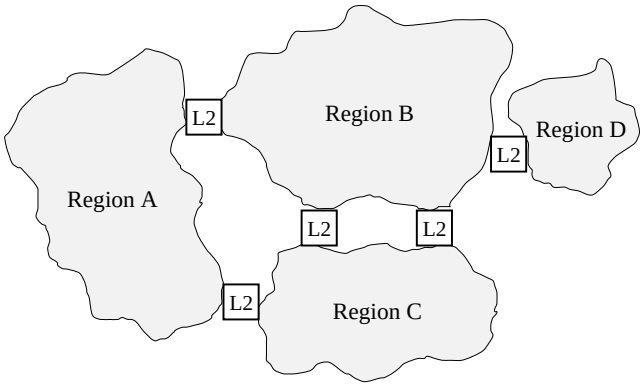

Figure 3: Regions interconnected by L2 routers

Each region consists of one or more "boundary routers" responsible for interconnecting different regions and forwarding multicast traffic between them. The multicast routers internal to a region run a "Level 1" (L1) multicast protocol for forwarding multicast traffic within the region and the boundary routers run a "Level 2" (L2) multicast protocol for forwarding inter-region traffic. Boundary routers include L1 functionality in order to participate in the L1 routing of each region to which they are attached as shown in Figure 4.

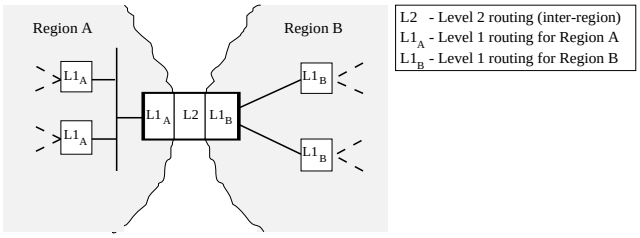

Figure 4: Logical representation of a boundary router.

This paper describes DVMRP as the L2 protocol, any of the other multicast protocols (including DVMRP) being the L1 protocol. The L2 routers exchange routing information using "region identifiers" instead of subnet addresses internal to the region. Packets generated from a region are tagged with the region's identifier, which is then placed in an encapsulation header for transit across regions. Upon receipt at the destination region, the encapsulation header is stripped off before final delivery to group members within the region.

We describe this approach in three phases for a multicast packet originating in Region A destined for group members in Regions A, C and D. Phase I describes routing within the source region, Phase II describes inter-region routing, and Phase III describes the final delivery process in the destination regions.

3.1 Phase I: Routing in Originating Region

We assume that a conventional multicast routing protocol, such as DVMRP or MOSPF, is operating within each region. A multicast packet that originates within a region is forwarded by the intra-region protocol to all subnets within the region that have members of the destination multicast group. In addition, the packet must be forwarded to all the boundary routers (L2 routers) attached to the region. The

method for accomplishing that is protocol-specific. For example, DVMRP uses a "broadcast and prune" approach, in which multicast packets are initially delivered to all routers in the region; those routers that are not on the path to a group member are subsequently "pruned off" the delivery tree. To ensure that DVMRP delivers to all boundary routers is simply a matter of suppressing the pruning messages from those routers. In the case of MOSPF, boundary routers may join a special "wildcard" group, which causes them to be included in the delivery tree of all multicast packets.

Figure 5(a) shows a multicast packet originating at S being forwarded to a member M of the destination group within Region A and to all the L2 routers attached to the region. The source and destination group addresses are used to forward the packet within the region. Figure 5(b) shows the packet header information used to route packets within the regions.

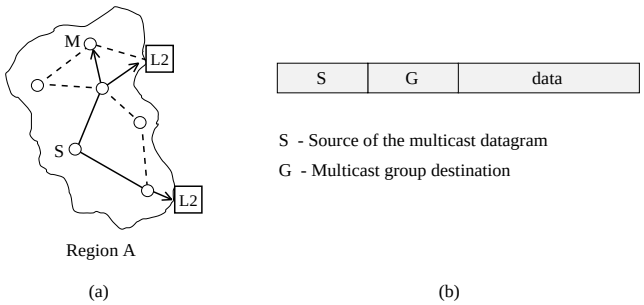

Figure 5: (a) Forwarding multicast packets within an originating region. The dotted lines show the connectivity within the region and the thick lines show the exact paths traversed to reach the destination members. (b) Packet header information.

A boundary router may decide, upon receiving a multicast packet originated within an attached region, that it need not be forwarded to other regions (according to the L2 routing protocol). The L1 component of the boundary router may take advantage of that knowledge to exclude itself from subsequent transmissions from the same source to the same multicast group. For example, in a region running DVMRP as its L1 protocol, a boundary router may send L1 DVMRP prune messages towards the source subnet in response to such L2 information.

3.2 Phase II: Routing between Regions

Each region is assigned a unique identifier which is configured into all of the L2 routers connected to that region. We call this identifier the Region-Id. With DVMRP as the Level 2 protocol, each region is the logical equivalent of a subnet and the Region-Ids are equivalent to the subnet addresses as used by L1 DVMRP routers. The L2 routers run a separate instance of the DVMRP routing algorithm with these Region-Ids as addresses.

Each L2 router also acquires information on all IP subnet addresses internal to its directly attached regions. It acquires this information by virtue of each boundary router having an L1 component as part of its implementation, as shown in Figure 4.

L2 routers join a reserved multicast group, ALL_BOUNDARY_ROUTERS (ABR) in all directly-attached regions. The ABR group is scope-limited to a single region, i.e. traffic to this particular group address does not cross region boundaries. An L2 router can distinguish packets from the different regions sent to the same destination ABR group address by the region the packet arrived from.

The L2 routing messages are similar in all respects to DVMRP L1 routing messages. An L2 routing message injected into a region has the source field set to the address of the boundary router as it is known in the region and the destination field to the ABR group. Since all L1 routers have complete address information on the region they belong to, these routing messages are forwarded to all other L2 routers connected to the region using the same rules for forwarding packets originating within the region.

Upon the arrival of a multicast packet from one its connected regions, an L2 router performs the following operations.

1. It checks if the source of the packet is one of the subnets in the region from which the packet arrived. If this check fails, the packet is ignored.

2. The packet is tagged with the Region-Id representing the region from which the packet originated.

3. For each region across which the L2 protocol decides to forward the packet, a copy of the packet is sent in the encapsulated form shown in Figure 6. The inner IP header contains the original source address and destination group address (S_i and G_i). The Tag field carries the originating Region-Id. The outer, encapsulation header carries the address of the sending boundary router as its source (S_0), and the ABR group address as its destination (G_o).

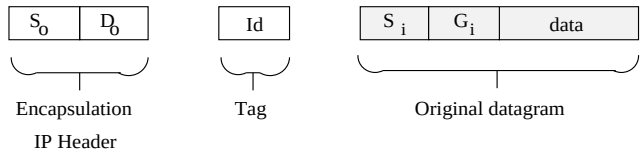

Figure 6: Encapsulation header attached to packets for forwarding between regions.

The resulting encapsulated packet is conveyed by the L1 routing from the sending boundary router to all other boundary routers of the region. Figure 7 shows such forwarding from boundary-router R1 across Region B, and from boundary-router R2 across Region C.

Upon the receipt of an encapsulated packet of this type, an L2 router performs the following operations.

1. The encapsulation header is stripped off and a reverse-path forwarding check is performed using the Region-Id to determine if the encapsulated packet arrived on the shortest path towards the source region, i.e. the region from which the packet originated. If the check fails, the packet is ignored.

2. The downstream regions are determined by using the Region-Id and the group address of the original packet as the key for the routing table lookup. For each downstream region across which that the packet is forwarded, the encapsulation process described above is repeated, retaining the Region-Id as the tag.

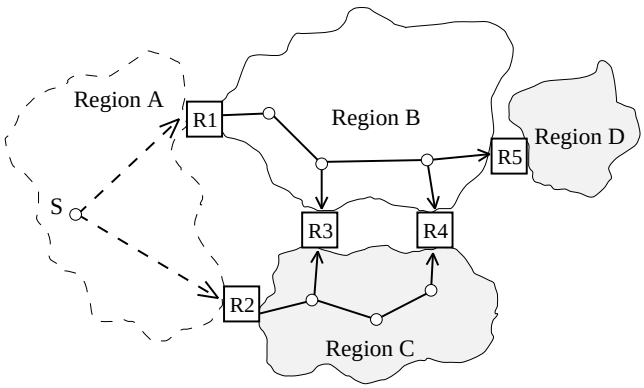

Figure 7: Routing of packets between Regions. Region A is the source. Regions C and D have members of the destination group. Region B is a transit region

3. If any of the connected regions have members of the destination multicast group, a copy of the original packet is injected into the region as described below in Phase 3 of the routing.

Membership information within a region is obtained by having the L1 routers within the region send *Region Membership Reports* (RMRs) to all of the region's boundary routers (via the ABR group), reporting which groups have members in the region. Periodically, and also whenever the group membership status changes on a subnet (that is, whenever the first member of a new group appears on the subnet, or the last member of a group departs from the subnet), at least one of the L1 routers attached to that subnet sends an RMR message. This is analogous to the operation of the IGMP protocol [RFC1112] by which hosts report their group memberships to their neighboring routers, but operating one level up in the hierarchy.

L2 routers that do not have any members of a particular group in any of their downstream regions send DVMRP prune messages to the upstream L2 router. If a member appears on any of the downstream regions, a DVMRP graft message is sent to the upstream router requesting restoration of delivery of traffic for that particular group.

3.3 Phase III: Routing in Destination Regions

Each L2 router advertises a *default route* with a certain preconfigured metric into each connected region. The *default route* is the route to be used in the absence of a specific routing entry for the source subnet of a packet. Each L1 router upon receipt of these default routing updates, determines the interface corresponding to the nearest L2 router using the default metric advertised by each L2 router. This process results in a set of logical sub-trees rooted at each L2 router, each subtree reaching a subset of the L1 routers within the region. Figure 8 shows the subtrees formed at each of the L2 routers R2, R3 and R4 attached to Region C.

The L1 routers then construct suitable delivery trees for distribution of the multicast traffic within the Region. For example, DVMRP L1 routers may send prunes towards the nearest L2 routers if there are no downstream group members. Figure 8 shows the sub-tree rooted at R3 being pruned off due to the absence of any members in any of its downstream links.

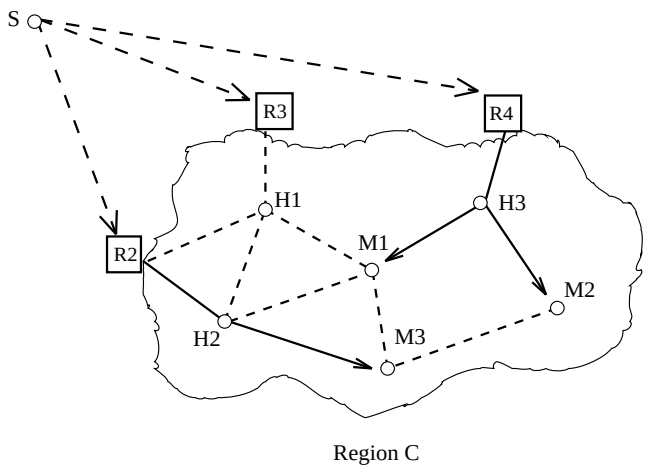

Figure 8: Routing packets in the destination Region C.

4 Protocol Evaluation

In the previous section, we described a scheme for achieving hierarchical DVMRP routing in the MBone. While this approach reduces the routing overheads, there are certain drawbacks that need to be carefully addressed.

4.1 Packet duplication over internal links

One of the problems in using this encapsulation technique is that regions which have multiple boundary routers, and hosts that are receivers of multicast traffic from external sources could have the same packet traverse a link in the encapsulated form as well as the original form. Boundary routers use the ABR group as the destination address in the encapsulated packet to communicate with other boundary routers. An L1 router in this region could potentially forward these encapsulated packets to other boundary routers in the region over the same links that it receives the decapsulated packets. For example, Figure 8 shows the delivery path R4-H3-M1 for external packets to reach member M1. However, router R4 received the same packets in the encapsulated form from router R2 over the links R2-H1-M1-H3-R4 as shown in Figure 7. Hence, the same packet traversed the links M1-H3 and H3-R4 first in the encapsulated form and then with the original header. This problem arises whenever there are members within a region having multiple boundary routers. Two possible ways by which this can be avoided are:

- Router decapsulation: The L1 routers could act like a "pseudo-L2" router by decapsulating transit packets and checking if there are any group members corresponding to each packet's destination multicast address in any of its attached subnets. The additional overhead using this method lies in examining each transit encapsulated packet.

- Host decapsulation: The encapsulated packets (for forwarding between L2 routers) have the ABR group as the destination address in order to prevent the L1 routers from having to partake in the decapsulation process. If the encapsulation header were to have the destination multicast group instead of the ABR group, these packets could be forwarded like regular multicast packets, with either the hosts or the "leaf" L1 router

stripping the encapsulation header and processing the original packet.

4.2 Effect on Fair Queuing and Resource Reservation Algorithms

Fair queuing [Dem89] and resource reservation [Zha93] algorithms associate the source and destination addresses of a packet, and possibly other fields, such as transport-level port numbers, to bandwidth and other constraints (e.g. type of traffic - unicast, multicast etc). These algorithms are implemented by identifying those fields in each packet header and making a queueing decision based on the constraints specified. Packets forwarded at level 2 of the hierarchy have an encapsulation header attached whose source and destination address are different from the original packet. This could cause existing algorithms to use the wrong source and destination addresses in the decision-making process.

We believe that a suitable alteration to the "lookup" code to examine both the encapsulated and the inner packet header will provide the requisite information required to make a queuing decision. The additional cost incurred will only be that of examining two headers to identify the exact fields required.

5 Other Issues

Before concluding, there are other related issues that arise as part of the hierarchical routing mechanism.

5.1 Enabling Multiple Levels of Hierarchy

The two-level hierarchical routing approach described can be extended to multiple levels using one of two approaches:

- Applying the encapsulating scheme recursively. It is possible to cluster several regions into "super-regions" inter-connected by L3 routers, and assigning identifiers to them. While this technique restricts the routing table size at each level to that of the addresses within each "super-region", it has the additional overhead of performing an additional level of encapsulation for communicating as well as the necessity to operate an RMR mechanism at each level.

- A "CIDR-type" approach in allocating region-ids. This technique merges topological information into the selection of region identifiers, thus enabling multiple levels of hierarchy by recursive aggregation of region-identifiers. This approach is more desirable since it does not have the additional overhead of adding another distinct level, the hierarchy being present in the region identifiers themselves. The reader is referred to [RFC1519] for further details on hierarchical aggregation of addresses.

5.2 Level 1 - Level 2 interoperability

We have so far described DVMRP as the inter-region (L2) multicast protocol, and any of a number of multicast protocols serving as the intra-region (L1) protocol. In order for smooth interoperability between the two levels of routing, a few changes have to be made to the various L1 protocols.

L1 routers now have to generate RMRs which carry group membership information on all attached subnets. The RMRs enable the L2 routers to determine the presence or absence of group members in each attached region. In addition, L1 routers must support the default route option as explained in Phase III of the protocol description.

Routing algorithms such as PIM and CBT [Bal93] which do not maintain explicit routing information of their own (rather, they rely on the unicast routing tables) have to be "fixed" at each interface to the L2 routers to provide "filtered" route updates, which contain only the addresses and prefixes that are local to the region. This is necessary since PIM L1 to DVMRP L2 routing updates can include entire unicast routing tables causing L2 routing to break.

5.3 Configuration parameters

Boundary routers must be configured with the value of two parameters for each region to which they directly attach: a Region-Id and a metric for the default route advertised into the region. It is possible to craft the default metrics advertised by each L2 router such that all the L1 routers within the region will forward traffic only from a single L2 router. This may be desirable in some regions if it is known that most of the external traffic will arrive through a single L2 router, the other L2 routers then serving as backup paths.

5.4 Avoiding multiple levels of encapsulation

Since multicasting is not yet fully deployed in the Internet, several multicast routers have *tunnels* between them. If two L2 routers were connected via tunnels, traffic exchanged between them has to be encapsulated twice, once due to the tunnel encapsulation, and the other due to Phase II forwarding described in Section 3. This dual-encapsulation process represents a significant overhead in forwarding each datagram. This can however be overcome using the address of the tunnel end-point as the destination address in the encapsulation header instead of the ABR group, the source region identifier being retained as a tag in the encapsulated packet. Since L2 forwarding is based only on the region identifier and the destination group of the original packet, the change in the destination field of the encapsulation header does not affect the forwarding mechanism at the destination L2 router.

5.5 Boundary through links vs. routers

In our discussion so far, we have assumed that the Regional boundaries fall across boundary routers as shown in Figure 4. It may be desirable in some circumstances to make the boundary fall between two routers instead, for example, if the organization on each side of a boundary wishes to configure and manage its own "half" of the boundary. This effect can be achieved by treating a "boundary link" as a separate, degenerate region. As long as there are no potential sources of multicast packets on the boundary link, there is no need for it to have a Region-Id, and its presence does not add to the size of the L2 routing table.

5.6 Boundary leakage issues

Misconfigured boundary routers and links across regions can lead to multiple regions being accidentally configured as one large region. Boundary routers can detect such anomalies from the periodic routing messages that they send into each directly attached region. If a routing message sent by an L2 router into one region appears from another region, this signifies either a misconfigured boundary router or a link causing traffic to leak through "backdoors". Boundary routers, on detection of such a condition, should stop forwarding data into those attached regions in order to prevent multicast routing loops.

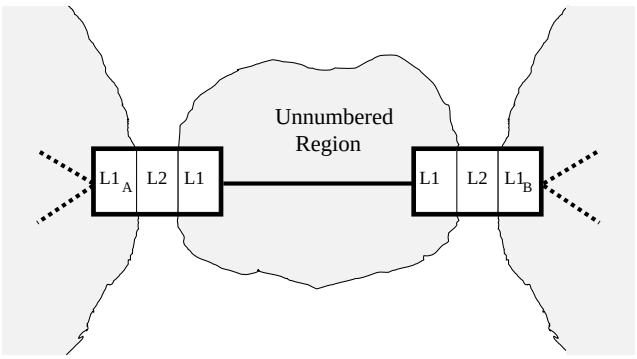

Figure 9: Decomposing a link between two L2 routers into an "unnumbered" region.

6 Conclusions

We have described an approach to adding hierarchical multicast routing to the MBone, initially using a two-level hierarchy with DVMRP at the top level. The modifications to DVMRP to operate as an inter-region protocol, and the modifications of the various existing protocols (including DVMRP) to operate as inter-region protocols, are small and straightforward. Thus, we expect to be able to deploy this scheme quickly, in order to allow the MBone to keep growing at its current rate.

The use of address-independent region identifiers as the basis of the top-level routing enables significant reduction in routing table size, regardless of the "aggregatability" of the MBone subnet addresses. It also enables easy incremental deployment of the Level 2 routers—the region boundaries need not correspond one-to-one with the addressing boundaries used for unicast routing.

In addition to the reduction of routing table size, deployment of hierarchical multicast routing will reduce the degree of topological volatility that any router must handle and relax the current constraints on maximum MBone diameter. It will also provide a clean architecture for accommodating more than one multicast routing protocol at a time and enable the safe deployment of new multicast routing protocols that rely on information from unicast routing.

Our design extends to deeper hierarchies and the use of protocols other than DVMRP at the top level. It is too early to say whether or not the MBone will evolve in those directions. For example, if the new protocols that rely solely on unicast routing for topology discovery, such as PIM, prove to be successful, there may be no need to deepen the hierarchy; instead, the problem of routing table growth will revert to unicast routing.

A final observation is that, interestingly, the hierarchical routing scheme proposed here—using address-independent region identifiers and encapsulation—*cannot* be applied directly to the problem of *unicast* route scaling. That is because, as a unicast packet travels towards it destination, the forwarding decision is made on successively finer-grain information about the location of the destination. Even though the routers far from the destination route on only course-grain information (the high-order parts of the destination address), all of the location information must be present in the original packet for the routers near the destination to use. Current multicast routing algorithms work the opposite way: they make routing decisions based on where a packet *came from*—the root of a multicast delivery tree—and the further the packet travels, the coarser the information needed about its source. As the multicast packet crosses boundaries, the routers can encapsulate it with a header that contains less information than the original, as in the scheme described in this paper: for its first hops, forwarding is based on the origin subnet number; for its transit hops among regions, forwarding is based only on source region; and for its final hops, forwarding is based on a single default route for all distant sources.

7 Acknowledgements

We would like to thank Deborah Estrin, Dino Farinacci, and Van Jacobson for their considerable help in developing the approach to hierarchical multicast routing described in this paper.

References

[Bal93] Ballardie, A. J., P. F. Francis, and J. Crowcroft, "Core Based Trees", In *Proceedings of the ACM SIGCOMM*, San Francisco, 1993.

[Cas92] Casner, S. and S. Deering, "First IETF Internet Audiocast", *ACM SIGCOMM Computer Communications Review*, Vol.22, No.3, July 1992.

[Cas94] Casner, S., "Are you on the MBone?", *IEEE Multimedia*, Vol.1, No.2, Summer 1994.

[Dee91] Deering, S. E., *Multicast Routing in a Datagram Internetwork*, Ph.D. Thesis, Stanford University, December 1991.

[Dee94] Deering, S., D. Estrin, D. Farinacci, V. Jacobson, C. Liu, and L. Wei, "An Architecture for Wide-Area Multicast Routing", *Proceedings of the ACM SIGCOMM 1994*, August 1994.

[Dem89] Demers, A., S. Keshav, and S. Shenker, "Analysis and Simulation of a Fair Queuing Algorithm", *SIGCOMM '89*, 1989.

[Kle77] Kleinrock, L. and F. Kamoun, "Hierarchical Routing for large networks; performance evaluation and optimization". *Computer Networks*, Volume 1, pages 155-174, 1977.

[RFC791] Postel, J., "Internet Protocol",RFC 791, September 1981.

[RFC950] Mogul, J. and J. Postel, "Internet Standard Subnetting Procedure", RFC 950, August 1985.

[RFC1058] Hedrick, C., "Routing Information Protocol", RFC 1058, June 1988.

[RFC1112] Deering, S., "Host Extensions for IP Multicasting", RFC 1112, August 1989.

[RFC1519] Fuller, V., T. Li, J. Yu, and K. Varadhan, "Classless Inter-Domain Routing (CIDR): an Address Assignment and Aggregation Strategy", RFC 1519, September 1993.

[RFC1584] Moy, J., "Multicast Extensions to OSPF", RFC 1584, March 1994.

[Zha93] Zhang L., S. E. Deering, D. Estrin, S. Shenker, and D. Zappala, "RSVP: A Resource Reservation Protocol", *IEEE Network*, September, 1993