

# Providing Multicast Communication in a Differentiated Services Network Using Limited Branching Techniques\*

Minaxi Gupta, Mostafa Ammar

*Networking and Telecommunication Group, College of Computing  
Georgia Institute of Technology, Atlanta, GA, U.S.A.*

## Abstract

The differentiated services (DS) paradigm has emerged as a scalable approach to provide QoS for unicast communication in the Internet. Utilizing the DS to provide QoS for multicast poses several challenges because of the multi-point aspect, dynamic group membership, and heterogeneous receiver resource requirements of multicast. Existing proposals to accomplish this goal either require extra per packet overhead or changing multicast routing tables. This paper proposes an alternate architecture called M-DS (multicast-DS), utilizing two inter-operable limited branching techniques, *edge-router branching* and *limited-core branching*, as appropriate within each DS domain. M-DS preserves the DS scalability, does not incur any per packet overhead due to extra headers in data packets, and routes packets using the IP multicast routing tables that are already set up in individual domains, albeit by enhancing the router functionality. We evaluate the performance (signaling overhead and extra bandwidth required) and show that it is practical to include it in the current DS framework.

## 1 Introduction

Multicast communication accomplishes one-to-many and many-to-many delivery of data in an Internet environment. It is scalable and efficient because it outperforms unicast even for a small number of receivers. A recent study [1] shows that even when there are 20-40 receivers, multicast can be 60-70% more efficient than unicast in the Internet. The differentiated Services (DS) architecture [2] is

a scalable method for implementing service differentiation for unicast communication in the Internet.

Dynamic join/leave of multiple potentially heterogeneous receivers poses unique challenges in providing support for multi-point multicast communication in a DS network. This is because when new receivers join the multicast group, branches may get added to the existing multicast tree without prior resource allocation and this can adversely affect the unicast and multicast traffic for which resources have previously been reserved.

This paper proposes a scalable architecture called M-DS, to provide QoS for multicast communication utilizing the DS framework. The architecture uses one of two limited branching techniques, namely, *edge-router branching* and *limited-core branching*. For each of the techniques, we focus on defining two signaling protocols; one for resource allocation on membership discovery and another for resource deallocation on membership termination on subnetworks. The signaling for resource allocation configures state in appropriate routers to be used during multicast data transmission. After this phase, data can begin to flow for multicast with QoS, as it would be in the case of DS for unicast. The signaling for resource deallocation resets the configuration changes made by the signaling for resource allocation. All the four signaling protocols have low message overhead. For both the techniques, flows are aggregated for scalability, just as in the DS framework for unicast. Also, both the techniques use the multicast state already set up in individual domains for routing packets and inter-operate with each other. Not changing multicast tables also facilitates the co-existence of IP multicast without the QoS requirements. These techniques have no per packet overhead during data transmission in terms

---

\*This work is supported by the AFOSR MURI grant F49620-00-1-0327, NSF grant ANI-9973115, and by a research grant from Bellsouth.

of extra headers in individual data packets. For details on the changes required in the routers to accommodate these techniques in a DS framework, refer to [3].

## 2 Related Work and Challenges

A DS domain is comprised of *boundary nodes* and *core nodes*. Boundary nodes interconnect the DS domain to other DS or non-DS capable domains while core nodes only connect to other core or boundary nodes within the same DS domain. Traffic enters a DS domain at an ingress node and leaves at an egress node.

The DS framework uses a six bit *DS field* from the IP header to define DS codepoints. All packets with the same codepoint that cross a link in a particular direction form a behavior aggregate. The DS boundary nodes at the customer egress set the appropriate codepoint in each packet in accordance with the customers' service level agreement (SLA) and the packet joins the correct behavior aggregate. From this point on, subsequent boundary or core nodes in various DS domains have no information about a particular customer's flow, they only deal with behavior aggregates. This contributes significantly to the scalability of the architecture.

The DS framework is specified with unicast in mind. Recently, there has been some work in the direction of utilizing the DS framework for providing service differentiation for multicast. Bless and Wehrle [4] propose to extend the multicast routing tables to include codepoints to provide QoS for multicast in the DS framework. Striegel and Manimaran [5] have proposed an encapsulation-based approach called *DSMCast* for providing multicast support in a DS domain. Their approach consists of adding a DSMCast header to each packet at the edge of the DS domain by the ingress router. Upon receiving such a packet, a core router will inspect the packet to determine which interfaces the packet should be replicated on based on the information contained in the DSMCast header. This solution keeps the core routers simple but incurs bandwidth overhead for every data packet, dependent on the number of receivers. Our approach, M-DS, is scalable both in terms of number of multicast groups, as well in terms of number of receivers.

Providing quality of service (QoS) for multicast communication using the DS framework poses unique challenges. Multicast sources generally do not know the identity of receivers. The source sends out one copy of the data and the IP layer multicast routers make duplicate copies where needed in the network to reach all receivers of the group. Also, group membership in multicast is dynamic and the receivers are heterogeneous. Out of these issues, heterogeneity is not a stumbling block because at the application layer, receivers with different resource requirements can join different multicast groups [6], so within a particular multicast group, the resource requirements are homogeneous. However, dynamic group membership makes it challenging to provide QoS for multicast communication.

The DS architecture for unicast can not be used as is to accomplish QoS for multicast without affecting other traffic adversely. The main reason for this is because scalability in the DS architecture for unicast is achieved by distinguishing between the functionality of core and boundary routers in each domain and by traffic aggregation. Except for the routers near the source, all routers along the way deal only with behavior aggregates in the DS architecture. In multicast, however, new members may join a multicast group dynamically and as a result, several routers in the core of the network may duplicate packets to reach the new receivers. Keeping the core routers simple to preserve the DS scalability would imply core routers assign the same codepoint to the duplicated packets as to the original packets. As a result, new branches can get added to the existing multicast tree without prior resource reservation. This problem is termed as *non-reservation subtree (NRS) problem* in [4]. NRS can potentially lead to violation of SLAs between the DS peers and hence compromises QoS for one or more classes of traffic.

## 3 Components And Assumptions

An example of the various entities of the M-DS architecture is shown in figure 1. It shows two DS domains, with the multicast source attached to domain *DS1* and receivers *R1* and *R2* on the same subnetwork attached to domain *DS2* through the designated router (DR). Each domain has boundary and core routers. The figure also shows the band-

width brokers  $BB1$ , and  $BB2$ , for domains  $DS1$  and  $DS2$  respectively. The DRs initiate the signaling with the BB of their domain upon each multicast group membership discovery and termination on their respective subnetworks. The bandwidth brokers (BB) handle resource allocation and deallocation requests in their domain by contacting their peers.

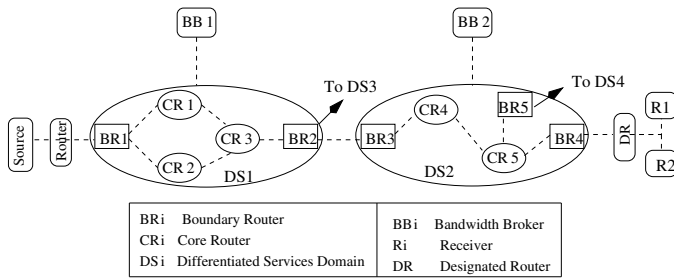


Figure 1: Components of the M-DS Architecture

The architecture makes the following assumptions about the infrastructure. First, each DS domain has static unicast SLAs for aggregates of flows with its peer domains for each of its ingress and egress router pairs. Admission requests for both unicast and multicast are handled dynamically by the BB of each domain by contacting the BB in the peer domain. These assumptions are consistent with those of the *QBone* BB work group [7]. Second, each BB is configured as a BGP-4 router and has TCP connections for communication with all its peer BBs and the DRs of its domains for communication during signaling. It also has access to unicast multicast routing information. Third, each BB has a data repository containing router configurations and policy information. It needs this to be able to make decisions to allocate and deallocate resources. Fourth, before sources start sending data, they register with the BB of their domain, which can then propagate this information to peer BBs. In this manner, the BBs know about the active multicast sources.<sup>1</sup> Fifth, we assume that all receivers know what QoS to ask for. The resource request can be carried in a manner similar to the one prescribed by the RSVP specification [10].

<sup>1</sup>This is essentially similar to how MSDP [8] requires the RPs [9] in each domain to advertise active sources.

## 4 Edge-Router Branching

We now describe the details of the edge-router branching technique (refer to [3] for a formal description of the technique). It uses existing unicast SLAs and exploits multicast scalability at domain granularity because it allows branching to occur only at the ingress and egress routers. If there are any branching points in the core of the domains, they are moved to the ingress of the domain. This keeps all the complexity confined to the edge of the network and hence the core nodes are kept scalable as in the case of DS for unicast.

### 4.1 Signaling Protocol for Admission

Signaling is initiated by the DRs of the domains upon a multicast group discovery on their respective subnetworks. To begin with, the DR sends a message to the BB of its group, giving it its own IP address, the QoS requirements and the multicast group address  $G$ . Subsequently, this BB may contact its peer BB and so on. Notice that for each subsequent receiver joining the same subnetwork served by this DR, no action needs to be taken as long as the QoS requirements and  $G$  are the same.

The BB extracts the QoS information and  $G$  and then finds out if the resources for this request have already been allocated in its domain. Because if they are, no new resources are to be allocated. The BB needs two pieces of information to make that decision.

1. Appropriate entries from the *current resource allocation table*. This table contains resource allocation information for each pair of ingress and egress routers in the domain along with the multicast groups they serve. It also contains the number of different subnetworks (identified by the IP addresses of the DRs) each ingress-egress router pair serves, directly or indirectly. Keeping the number of the subnetworks served corresponding to each ingress-egress router pairs helps the BB know when to deallocate the resources in its domain. Entries in this table are filled after making a decision that resources can be granted. To find out the appropriate entry, the BB first finds out the pair of ingress and egress routers in its domain in the path from the multicast source for group  $G$

to the DR. It does that using the routing information that it has access to as a BGP-4 router.

2. If the *branching point* for  $G$  lies in its domain. While the signaling protocol is in progress, multicast state is being set up in the domain using the IP multicast protocol in use in that domain. The core routers in every domain that determine that they are going to be the branching point for group  $G$  communicate this information to their BB by sending the corresponding multicast forwarding table entry. The BB sets a timer to get this information from the branching point router(s) in its domain. If the timer expires without the BB getting a reply, it assumes that no branching point exists in its domain.

Based on the above information, three cases arise for the ingress-egress router pair needed to satisfy this request, as outlined in figure 2. For examples of each of these cases, refer to [3].

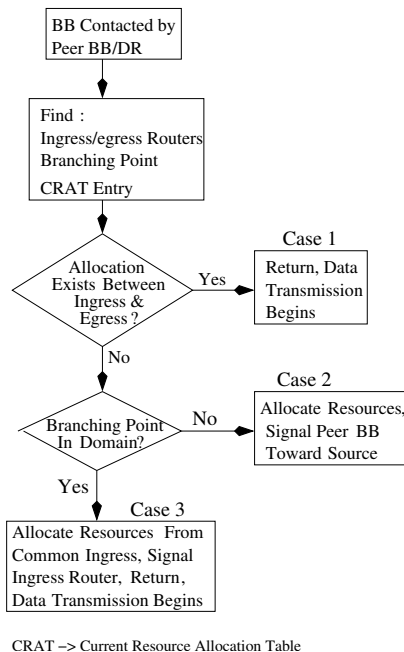


Figure 2: Signaling Steps in Edge-Router Branching Technique upon Membership Discovery

#### 4.1.1 Case 1: Allocation Exists

This case arises if the BB finds an entry in the current resource allocation table with the requested

QoS corresponding to the ingress and egress routers needed to serve the new request for group  $G$ . This implies that adequate resources have already been allocated. Nothing needs to be done other than updating the number of subnetworks served by this ingress-egress router pair in the current resource allocation table and the BB replies in the path to the DR affirmatively and signaling is terminated. The new receivers are grafted to the existing multicast tree and since multicast state is already set up, they can start getting data.

#### 4.1.2 Case 2: No Branching Point

The second case arises if no resources have been allocated for  $G$  in this domain with the requested QoS. In this case, the BB does not find any entry with the requested QoS in the current resource allocation table corresponding to the ingress and egress routers needed to serve the new request. Also, no branching point router replies to the BB. The BB makes a decision to grant resources in the form of an SLA between the ingress-egress router pair based on policy information and SLA availability. Reference [3] provides details on the properties of the SLAs.

If the resources are granted, the BB updates the current resource allocation table for the latest bandwidth allocation corresponding to the ingress-egress router pair involved and sets the subnetwork count for the corresponding entry to one. The signaling does not terminate here for this case and the BB signals the upstream peer BB. Upon receipt of the message, the peer BB runs the same protocol as this BB.

#### 4.1.3 Case 3: Branching Needed

The third case arises if one or more core routers send the corresponding multicast forwarding table entry to the BB of their domain, informing that they would be the branching point for group  $G$ . This implies that the multicast tree passes through this domain but branching is required to graft the new receivers to the existing multicast tree. If the allocated resources are adequate for the new QoS request, new receivers can be grafted to the existing tree. The edge-router branching technique does not allow any branching point in the middle of any domain, hence to be able to graft new receivers to

group  $G$ , an additional SLA from the ingress router to the new egress is used if one is available.

Since all SLAs are unicast SLAs, to be able to use the new SLA, the BB moves all branching points from its core to the ingress of the domain and enhances the role of the ingress router to act as a branching point in addition to being an ingress router. This technique makes the core router functionality very scalable, just as in the DS for unicast communication but doing so amounts to introducing some changes in routing. The BB extracts the new routing information using the multicast forwarding table entry sent to it by the core branching point router(s) and conveys the appropriate routing information to the ingress router. The details on how the ingress router performs routing during data transmission and how core branching point router(s) use their existing multicast routing state are explained in [3]. There is no per packet bandwidth overhead during data transmission and some changes are required to be made to router functionality.

After moving the branching point to the ingress, the BB creates a new entry for this ingress-egress router pair in the current resource allocation table and sets the subnetwork count for the corresponding entry to one. To complete the signaling, the BB signals affirmatively in the path toward the DR and data transmission begins for the new receivers since multicast state is already set up.

At any point in the path to the source, if resources are denied, signaling returns from that BB all the way back to the DR, freeing all the resources and canceling all configuration changes and table updates. Most IP multicast routing protocols are based on soft state protocols and if data transmission does not begin, routing state vanishes automatically due to lack of a refresh. For additional state maintenance considerations, refer to [3].

## 4.2 Signaling Protocol for Departure

The protocol for the case when resources need to be deallocated for multicast group  $G$  is very similar to that of the case when they need to be allocated; only allocation of resources become deallocations. As in the case of allocation of resources, the DRs first contact the BB. The BB uses the routing information it has to look up the ingress-egress router pair serving this receiver in its domain. There

are two possible scenarios. First is that the subnetwork count corresponding to the ingress-egress router pair is one. This means this was the last receiver group served from this branch of multicast tree. Second is that the subnetwork count is greater than one, implying that there are more receivers on subnetworks in this domain or other domains that are being served by this ingress-egress router pair. In the first case, the resources can be deallocated and the corresponding entry from the current resource allocation table can be removed. Also, the changes made to the functionality of the ingress router of this domain are undone and peer BB in the path toward the source is signaled to carry out the leave protocol as well. In the second case, the subnetwork count is decremented by one because resources can not be deallocated yet and the adequate changes are made to the functionality of the ingress router. For the second case, the signaling can return from this point back to the DR confirming that the last receiver's leave from group  $G$  is complete.

## 5 Limited-Core Branching

The edge-router branching technique simplifies the core of the domain but introduces extra packet hops during data transmission. A good trade-off between making all core routers that are branching points complex versus incurring extra packet hops is to limit the number of core routers that would be allowed to be branching point routers. This is the motivation behind *limited core branching*.

For the edge-router branching technique, we assumed that SLAs in a domain existed only between ingress and egress router pairs, for the purposes of the limited-core branching technique, we assume that additional SLAs have been defined from certain special *core routers* to some egress points in addition to the usual unicast SLAs. These special core routers are the only ones that can be used as branching points in the domain. We do not address the issue of optimal placement strategy of special core routers, optimal number of such routers, and whether special core routers can be dynamically changed when receiver population changes. That remains an issue of future investigation.

The rest is a simple extension of three cases we described for edge-router branching. The manner in which the first two cases, namely when no re-

source allocations exist and when exact resource allocations exist and no branching is needed, are dealt with as before. The only difference is in the way the third case is handled. For every new membership discovery, if the BB finds out that one or more branching point(s) exist in its domain, before enhancing the functionality of the ingress router to duplicate packets to reach all downstream receivers, it first finds out if there is a special core router configured in its domain that is allowed to serve as a branching point. If there is, then for the path subsequent to that core router in that domain, that core's functionality is enhanced to be similar to that of the ingress router in edge-router branching. Only when the configured limit on the allowed branching point routers in the core of the domain is exhausted that the ingress router's functionality is enhanced to duplicate packets in addition to perform the DS specified functions.

Limited-core branching reaps the benefits of minimized packet hops during data transmission while keeping the technique scalable by limiting the number of special core routers allowed at the cost of a slight increase in complexity, which is controllable by individual domains. Any domain can opt to using limited-core branching independently of other domains and can configure as many special core routers as it chooses. This is because both the edge-router branching and limited-core branching techniques inter-operate.

## 6 Performance

The primary overheads of the M-DS architecture are in the form of signaling and increased packet hops for the multicast paths for some receivers. The main results of evaluating these overheads can be summarized as follows. First, the total signaling overhead per membership discovery and termination on sub-networks under both techniques is similar for all the topologies tested. It varies between 3.75 and 7 messages per receiver join (details in [3]) and is small compared to the average per second routing overhead of 23 messages/second per BGP-4 router [11]. Second, the simplicity of the edge-router branching technique comes at the cost of extra bandwidth consumption in terms of packet hops during data transmission. The effect is more pronounced when receivers are clustered together. Third, the mini-

mal controllable additional complexity of limited-core branching technique compared to edge-router branching saves the extra bandwidth consumption compared to edge-router branching by about 40%.

We used GT-ITM [12] to generate various topologies comprising of 744, 2646, and 6384 nodes each. Compared to the size of the Internet, these topologies seem small, but we believe that our simulations on these topologies produce results that prove the feasibility of deployment of both the techniques. The reason for this is the following. Signaling overhead depend on the number of domains in the topologies, not the number of nodes. In choosing the number of domains in the topologies, we used the results of a simple traceroute experiment we performed using random destinations across the globe. Our results indicated that most packets cross between 3 and 5 domains between source and destination. All our topologies have these properties. The details about the number of transit, stub, and total domains in these topologies are in [3].

We implemented a custom simulator in *C* to conduct the simulations. It implements Dijkstra's shortest path algorithm for multicast routing.

### 6.1 Bandwidth Overhead

The graphs in figure 3 show the percentage extra packet hops (compared to the total hops using shortest path multicast routing) for various topologies for edge-router branching. The receivers are statically placed randomly and in clusters. The graph for clustered placement of receivers is not smooth because the extra hops are closely tied to the actual placement of receivers. We ran the tests on various topologies for any given number of nodes, with similar results. The overhead is less for random placement of receivers, about 10% extra packet hops compared to total hops for 140 receivers; compared to about 40% when the same number of receivers are clustered. This is expected because random placement of receivers does not lend itself to savings in packet hops because of lack of commonality in the path to the receivers. The overhead does not increase substantially beyond a certain percentage when more receivers are added to the clustered placement. This is evident from the fact that in going from 140 to 450 clustered receivers, the overhead only goes from 40% to 50%. This seems log-

ical because adding more receivers to the same domain after a point would not increase the overhead further.

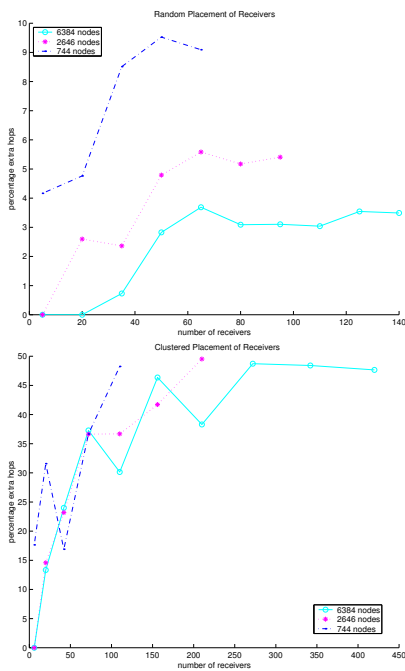


Figure 3: Percentage Extra Hops for Various Topologies for Edge-Router Branching

The graphs in figure 4 show the percentage extra packet hops for random and clustered placement of receivers for limited-core branching. The plots show three cases, when 1, 2, and 4 actual branching points are configured as special routers in each domain for the topology with 6384 nodes. For clustered placement of receivers, configuring 4 branching points as special routers in each domain brings down the percentage of extra hops to about 30% of actual packet hops under multicast routing, which is about 40% saving compared to edge-router branching technique. Also, note that configuring more branching points as special routers for random placement of receivers does not make a difference in the savings. This implies that decision about configuring branching point depends on the placement of receivers.

## 6.2 Signaling Overhead

To estimate the signaling overhead, the graphs in figure 5 show the number of signaling messages for when the receivers join dynamically. They show it

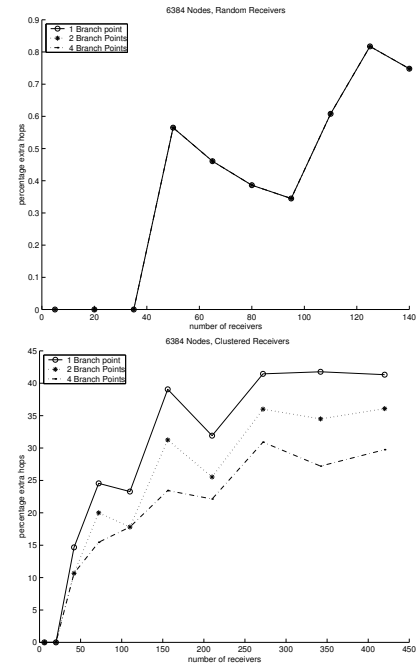


Figure 4: Percentage Extra Hops for Various Allowed Branching Points in Each Domain for Limited-Core Branching

for random and clustered placement of receivers for all three topologies for edge-router branching. The mean time between the arrival of receivers in the group for these plots is 60 minutes and the distribution of receivers' time in group is uniform. Since signaling is at the domain level, and all the simulation topologies have a similar number of domains, the signaling message overhead is almost independent of the number of nodes in the topology. As expected, the signaling overhead is more for random placement of receivers than for the clustered placement. The message overhead for 140 clustered placement receivers is about 50% less compared to similar number of randomly placed receivers. This is so because for clustered receivers, the signaling does not have to go all the way to the sender because of the presence of other receivers.

We experimented with exponential distribution as well with the mean time in group ranging from 2 minutes to 2 hours, the results were almost the same. Since the protocols involve a similar amount of signaling overhead for both join and leave of each receiver, the corresponding plots for the overhead when receivers leave were identical. The signal-

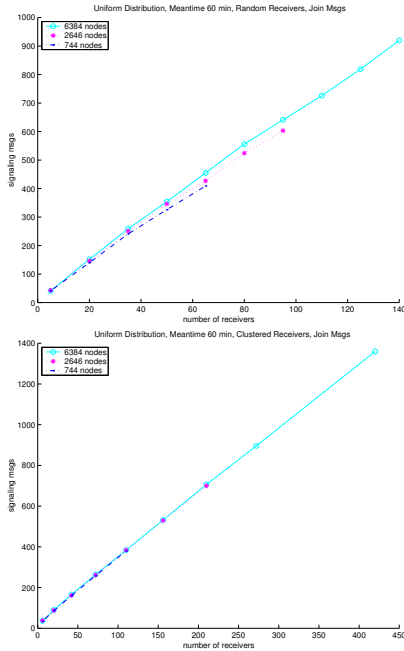


Figure 5: Signaling Messages for Receiver Join for Various Topologies

ing in the case of limited-core branching involves only as many additional messages as the number of branching points in each domain. Hence the signaling overhead for it is expected to be similar to that in the case of edge-router branching.

The above results on the signaling are conservative because we assume only a maximum of one receiver per subnetwork. In reality, because of the manner in which IGMP works, signaling would be carried out once for every membership initiation and termination on a subnetwork. For all other receivers on the same subnetwork, there is no signaling overhead.

## 7 Conclusions

This paper proposes M-DS, a scalable architecture that consists of two inter-operable techniques for providing support for multicast communication in a DS network. Both the techniques, *edge-router branching* and *limited-core branching* define two signaling protocols each to be run upon membership discovery and termination on subnetworks.

These techniques keep the core of the network simple and are scalable like the DS framework. They do not involve any per packet bandwidth over-

head during data transmission. This is because they do not introduce any extra headers in the data packets. The architecture allows each domain to run its individual IP multicast protocol. Both the techniques use the IP multicast routing state already set up in individual domains for forwarding packets during actual data transmission. Incorporating these techniques in the DS framework requires that the functionality of all the routers be enhanced. But for actual data transmission, only a few routers would need to use this enhanced functionality.

## References

- [1] K. Almeroth. Validating the Multicast Mystique. IEEE Infocom. Apr 2001.
- [2] S. Blake et. al. An Architecture for Differentiated Services. RFC 2475. Dec 1998.
- [3] M. Gupta and M. Ammar. Providing Multicast Communication in a Differentiated Services Network Using Limited Branching Techniques. Tech Report GIT-CC-02-27. May 2002.
- [4] R. Bless and K. Wehrle. IP Multicast in Differentiated Services Network. Internet Draft. Sept 1999.
- [5] A. Streigel and G. Manimaran. A Scalable Approach to Diffserv Multicasting. ICC 2001.
- [6] X. Li, M. Ammar, and S. Paul. Video Multicast over the Internet. IEEE Network Magazine. April 1999.
- [7] B. Teitelbaum P. Chimento. QBone Bandwidth Broker Architecture (work in progress). QBone Bandwidth Broker Work Group.
- [8] D. Meyer and B. Fenner. Multicast Source Discovery Protocol (MSDP). Internet Draft. May 2001.
- [9] D. Estrin et. al. Protocol Independent Multicast-Sparse Mode (PIM-SM): Protocol Specification. RFC 2362. June 1998.
- [10] R. Braden et. al. Resource ReSerVation Protocol (RSVP) – Version 1 Functional Specification. RFC 2205. September 1997.
- [11] Internet Performance Measurement and Analysis. <http://www.merit.edu/ipma.trends>.
- [12] K. Calvert, M. Doar, and E. Zegura. Modeling Internet Topology. IEEE Communication Magazine. July 1997.