# SEM: A New Small Group Multicast Routing Protocol

Ali Boudani, Bernard Cousin
IRISA/INRIA Rennes
Campus Universitaire de Beaulieu, Avenue du Général Leclerc
35042 Rennes, France
Tel: +33 2 9984 2537, Fax: +33 2 9984 2529
{aboudani, bcousin}@irisa.fr

***Abstract*—Recently several multicast mechanisms [3], known as small group multicast, were proposed that scale better with the number of multicast sessions than traditional multicast does. In this paper, we propose a new approach, Simple Explicit Multicast (SEM), which uses an efficient method to construct multicast trees and deliver multicast packets. SEM is original because it adopts the source-specific channel address allocation, reduces forwarding states in non branching node routers and implements data distribution using unicast trees.**

## I. Introduction

Data delivery to multiple destinations is a trade-off between bandwidth consumption (unicast transmission), state and signaling per multicast group (conventional multicast) and header processing per packet (small group multicast). A multicast routing protocol should be simple to implement, scalable, robust, use minimal network overhead, consume minimal memory resources, and inter-operate with other multicast routing protocols [4].

Multicast has become increasingly important with the emergence of network-based applications that consume a large amount of network bandwidth such as IP telephony, video conferencing, distributed interactive simulation (DIS) and software upgrading. Using the multicast services, a single transmission is needed for sending a packet to n destinations by sharing the link bandwidth, while n independent transmissions would be required using the unicast services. But, multicast suffers from the scalability problem. Indeed, a multicast router should keep forwarding state for every multicast tree passing through it. The number of forwarding states grows with the number of groups. Besides, each multicast tree need to be maintained which encounters limitations when the number of groups becomes large. Recently several multicast mechanisms, known as small group multicast, were proposed to scale better with the number of groups than traditional multicast does [3].

This document describes a new approach, Simple Explicit Multicast (SEM), which uses an efficient method to construct multicast trees and deliver multicast packets. In order to construct a multicast tree, the source encodes the list of destination addresses in a BRANCH message. This message has a role to discover routers acting as branching nodes in the multicast tree using the same mechanism used by Xcast [1]. Only branching node routers in the tree need to keep multicast forwarding states for a group. A special control plane is introduced to inform each branching node router about its next and previous hop branching node routers for a group. Instead, for multicast packets delivery, it uses recursive unicast trees, originally proposed in REUNITE [2]. Packets travel from a branching node router to another following the tree that has been construted by the BRANCH message. We propose that the source uses unicast encoding for multicast packets and sends them to its next hop branching node routers. Each branching node router acts as a source and packets travel from a branching node router to another.

The remainder of this paper is organized as follows. Section II presents some related works. Section III describes the SEM approach and some related issues are discussed. Section IV contains the approach analysis, simulation and evaluation for its forwarding state and messaging overhead. Section V is a summary followed by a list of references.

## II. Related Work

Some architectures aim to eliminate forwarding states at routers either completely by explicitly encoding the list of destinations in packets, instead of using a multicast address [1] or partially by using branching node routers in the multicast tree as in REUNITE [2] and HBH [7].

### A. Explicit Multicast

Explicit Multicast (Xcast) [1] is a newly proposed multicast scheme to support a very large number of small multicast groups and that by explicitly encoding the list of destinations in packets, instead of using a multicast address. Thus, the source encodes the list of destinations in the Xcast header, and then sends the packet to a router. Each router along the way parses the header, partitions the destinations based on each destination's next hop, and forwards a packet with an appropriate Xcast header to each of the next hops. An increased header processing per packet is cumbersome for high link speeds. Xcast+ [8] is an enhanced scheme for the support of receiver initiated join in explicit multicast which complements the existing Xcast. This is achieved by adding an IGMP join at receiver side and sending the join request through source-specific join message to the source and then by explicitly encoding the list of addresses of the multicast routers, instead of receiver addresses.

0-7803-7661-7/03/$17.00 

Whereas Xcast can support a very large number of small multicast groups, Xcast+ can support a very large number of medium size multicast groups. In all the newly proposed protocols the source knows the addresses of all the destinations before sending packets.The header processing time in every router grows with the number of the destination routers. The major difference between Xcast+ and SEM is that Xcast+ encodes the list of destinations in each packet while SEM uses this mechanism only with the BRANCH message. In both protocols the packet follows the unicast path between the source and all destinations. In SEM the packet will travel from a branching node router to another following the same unicast path. This seems a good solution in order to optimize the header processing time in every router.

### *B. REUNITE and HBH*

REUNITE [2] and HBH [7], use recursive unicast trees to implement multicast service. REUNITE does not use class D IP addresses. Instead, both group identification and data forwarding are based on unicast IP addresses. Only branching node routers for a group need to keep multicast forwarding state. All other non-branching node routers keep only multicast control state and simply forward data packets by unicast routing.

The HBH multicast routing protocol attempted to resolve some problems in REUNITE. First, HBH uses class D IP addresses for multicast sessions and not a unicast address as in REUNITE. Second, in REUNITE, when the first router that previously joined a group leaves the group, the tree maintenance become very complicated. Third, HBH attempted to resolve the asymmetric routing problem present in REUNITE. Finally, an HBH router keeps only the next hop router addresses and not the first router that join the session (multicast control table (MCT) and multicast forwarding table (MFT) has been modified).

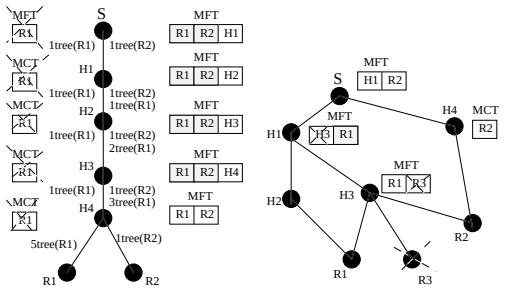
SEM (same as HBH) uses the unicast infrastructure to do packet forwarding with smaller routing tables, just as REUNITE does but uses (S, G) channels with class-D IP adresses to identifie multicast sessions. Using the IP multicast adressing model preserves compatibiliy with conventional multicast protocols. Since SEM uses the multicast addresses, The SEM control plane is compatible with the existing multicast protocols. SEM resolves also the asymmetric routing problem present in REUNITE since it uses the shortest path tree from the source to destinations. Besides, SEM eliminate all MCT and MFT entries in non branching node routers.

There are many similitaries between the SEM approach and the HBH approach but also many differences in the forwarding scheme and the control plane. As mentioned in HBH specifications, there is no formal definition of the interface between HBH and IP multicast and there are no details for forwarding protocol. Contrarily to other protocols, SEM can also provide statistics about the group members at any moment and ensures protection over denial of service attacks since the source is always aware about the unicast addresses of all destinations.

First, receivers in SEM use IGMP model and source specific join messages and the first join message reaches also the source itself but join messages in SEM are not periodic as in HBH. ALIVE messages between branching node routers are used to maintain the multicast tree since every branching node router knows its next and previous hop routers in the tree.

Second, the number of tree messages sent periodically in HBH is proportional to the number of destinations while we have only one tree message in SEM. Additionally, according to HBH specifications, when sending the tree message in order to construct the tree, this tree message passing by a router generates always a new tree message for all MFT entries including all previous receivers. In sparse mode networks, extra tree messages will inundate the network during the tree construction phase. If the number of receivers grows, the number of tree messages grows also.

Third, according to HBH specifications an MCT or an MFT exists in all routers between the source and the destination and this table is used to control and forward multicast packets. We found that there is no reduction at all in MFT sizes. Taking the network presented in Fig. 1(a), each router between the source and the destinations has an MFT. Unlike HBH, there is no need for MFT or MCT tables in non branching node routers. Packets in SEM, follow the unicast shortest path from the source to the destinations and these packets travel from branching node router to another while in HBH once the tree is constructed and marked entries expired, packets will follow an explicit path from the source to all destinations.




(a) Tree construction (b) R3 leaves the group

Fig. 1. HBH tree maintenance

Fourth, a receiver who leaves a group in HBH may affect the forwarding mechanism for the whole tree. Indeed, taking the Fig. 1(b) as an example, we have the final state of the tree where 3 receivers has joined the tree according to the HBH specifications (see [7] for more details). An R1 join message will reach the source and refresh the MFT entries (H1 is not stale because of this join message), so tree messages are to be sent to H1 and R2, also tree messages will be sent from H1 to H3 and R1. Tree message sent to R1 will force H3 to send a fusion (R1,R3) message to H1. When R3 leaves the group, then no join message will be sent to H3. In that case, R3 entry expires and MFT in H3 contains only R1 entry. H1 will continue receiving join messages from R1 and refresh the R1 entry but it will not receive any join message to refresh the H3 entry. When the entry timer expired, the entry will be destroyed. Since R1 entry in H1 is marked, R1 will never receive packets from S.

Finally, In the same example, when a tree message create an MFT entry in a router, a fusion message is sent to the previous

router and construct the distribution tree together with the tree message. During the tree construction where marked entries aren't expired yet, routers generate a lot of fusion messages. As an enhancement of HBH, only branching node routers should send fusion messages and only these routers create MFTs. In the other hand, it is not indicated in HBH if an MFT entry corresponds to a next hop router or to a destination. Indeed, when a fusion message marks the entries in an MFT router, only entries corresponding to leaf routers should be marked. Otherwise no data will be sent to destinations.

In conclusion, we can deduce that the tree discovery process is easier and simple in SEM than HBH. The presence of MCT and MFT in routers, the processing of tree and fusion messages and the huge number of these messages during the tree construction add some complexity to the HBH protocol that is simplified in SEM.

## III. SEM Protocol Description

In order to simplify address allocation in SEM, a group is identified by the channel (S, G) (called hereinafter session) where S is the source unicast address and G is a standard multicast address. A source creates and advertises a multicast session with a standard multicast address. In order to identify SEM session easily, compared with conventional multicast sessions, special multicast address range can be used. And thus, advertisement method using web pages will be useful.

Receivers send IGMP join (or leave) to the multicast router in their subnet in order to receive (or stop receiving) SEM packets from the source. This router sends source-specific join message (corresponding to (S, G) session) directly to the source. Intermediate routers don't need to keep the state information for the multicast session. Leave messages will be sent by the previous branching node router. Thus it is necessary for receivers to know the address of the source.

The source keeps track of the addresses of routers that sent source-specific join messages for the multicast session. The source encodes the list of router addresses in SEM header of a BRANCH message. The source then parses the header, partitions the destinations based on each destination's next hop, and send the BRANCH message to each of the next hops. The role of the BRANCH message is to discover routers acting as branching nodes in the multicast tree. We mean by branching node router, a router where packets arrive in an interface and should be forwarded to multiple interfaces (according to the next hop toward the destination routers). The SEM header contains also the previous hop branching router field (with initial value the source address S). The IP header will carry the protocol number PROTO_SEM. SEM packets are as follows:

[ IP header †Group address †Transport header †Payload ]

The IP header contains the source address and the destination address of the next hop branching node router.

Suppose that B, C, D, E, F and G want to receive packets distributed from S in Fig.2. This is accomplished as follows: B, C, D, E, F and G initiate IGMP join messages. When receiving the IGMP requests, R4, R8 and R9 each sends a source-specific join to S. S sends a BRANCH message with the list of multicast routers (R4, R8 and R9) in its SEM header to the first router, R1. IP header of the BRANCH message that S sends to R1 contains the source address S and the group address G and . SEM header contains the list of all destination leaf routers and the address of the previous hop branching node router (see Fig.2). Note that previous hop branching router initial value is the source address S itself.

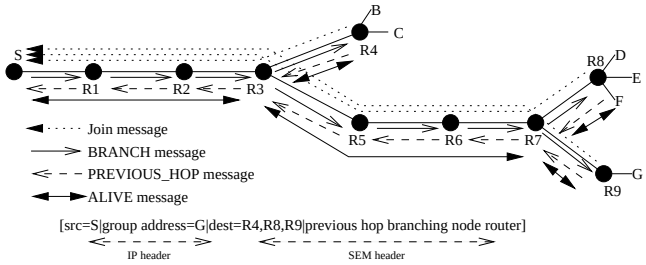



Fig. 2. Tree construction in SEM

When this BRANCH message arrives at a non branching node router for the (S, G), it is forwarded unchanged to the unique next hop router for all destinations. Otherwise, an entry is created at the branching node router. The entry contains the source address, the multicast address for the group and the list of unicast addresses of the next hop branching node routers (the list is initially empty). The branching node router replaces the previous hop branching router field in the BRANCH message with its own address before resending the BRANCH message.

The branching node router also sends a PREVIOUS_BRANCH message to the previous hop branching router (its address figure in previous hop branching router field in the BRANCH message). It is used to inform the previous hop branching router about the next hop branching router. The PREVIOUS_BRANCH message received by the previous hop branching node router updates the null list at the corresponding session entry (S, G) with the address of the next hop branching router (it can be extracted from IP header of the message). If an address to a next hop branching router already exists at the entry then the list should be simply updated and the new address should be added. For an optimization issue, the PREVIOUS_BRANCH message will not be generated if there is no changing in the corresponding entry. IP header of this message contains the router itself as the source and the previous branch router as the destination. The SEM header of the PREVIOUS_BRANCH message contains the source and the group addresses. At the end of this operation, we will obtain a path from the source to each destination router using the next hop branching node router addresses.

A BRANCH message is sent periodically by the source to ensure the maintenance of the tree. A timer is associated with the group entry at the source. If a new join or leave message arrives, then a new BRANCH message should be sent and the timer is set to zero. As a result of this tree maintenance, each branching node router contains an entry corresponding to each session. This entry contains the previous and the next hop branching node router for that router. When the source wants to send a packet to a group, the session entry is examined. The packet is forwarded directly to the next hop branching routers. The

packet is unicast to the next hop branching router with a payload containing G, the group multicast address. When the subsequent branching node routers receives the packet, the same operation is repeated. So if the router receiving the packet is not the next hop branching node router for that packet then it forwards the packet in unicast to the specific next hop branching node router. When the packet arrives at the router in the destination field, it will be then replicated and sent to each next hop branching node router. When the packet arrives to the leaf routers, then packet destination field should be replaced with the G address to ensure that it will be delivered through multicast to all receivers in the subnet.

ALIVE messages are used between branching node routers. When a router discovers that there are no more receivers for a group in its directly connected subnet, who wants to receive packets from the the source S, the router will stop sending ALIVE messages to previous branching node router. The previous branching node router eliminates the entry (it stops forwarding packets to the leaf router) and generates a source specific leave message (sent directly to the source). When receiving the leave message, the source eliminates the corresponding MFT entry and sends a new BRANCH message. Also, when a leaf router or a branching node router goes down, the previous branching node router will not receive ALIVE messages and eliminates the entry. It will send then a leave message toward the source who sends a new BRANCH message to rebuild the tree.

## IV. Protocol Evaluation

Our approach will be evaluated in terms of scalability (forwarding table size and control messages overhead) and efficiency (tree cost, delay and data processing).

We simulate SEM in NS (Network Simulator) [10] to validate the basic approach behavior and its effectiveness in state reduction and tree construction. The performance of SEM is compared to PIM, Xcast and HBH. PIM in our simulations refers to NS simulation of PIM-SM that constructs exclusively source specific trees. In addition to SEM we have simulated Xcast according to [1] and some of HBH mechanisms according to [7].

We present in our simulation two models generated using the GT-ITM generator [11]: each with flat graph of 100 nodes and all the links in the network are identical bidirectional links with 20Mbps bandwidth.. The topology of the first model is based on the first Waxman algorithm [12] and used as a dense mode network with 0.3 as the node degree distribution. The topology of the second model is based on a pure random algorithm in 5 domains and used as a sparse mode network. Four domains contain receivers and sources only, while the fifth domain is considered as the core domain. T sources and Nl receivers are randomly deployed in the network. A receiver join randomly the tree and there are no leave messages. Table I summarizes the parameters used in the simulation.

TABLE I
SUMMARY OF SIMULATION PARAMETERS

| | | |
|---|---|---|
| NT | 100 | number of nodes in the network |
| T | 10, 20, 30, 40, 50, 60 | percentage of sources in the network (number of trees) |
| Nl | 3, 6, 9, 12, 15, 18 | number of receivers for each source |

### A. Forwarding Table Size

We consider the parameter $\alpha$ of a distribution tree T to be the average number of multicast forwarding table entries per router for a tree:

$$\alpha(T) = \frac{Ne}{NT} \quad (1)$$

where *Ne* is sum of the total number of multicast forwarding table entries, i.e., the total number of (S, G) entries, on all the routers for distribution tree T, and *NT* is the number of routers on the tree. In a source specific distribution tree, every router contains one (S, G) forwarding table entry for the distribution tree, in which case *Ne = NT* and the value of the $\alpha$ parameter reaches its maximum 1.0 for source specific trees. The minimum $\alpha$ value for any particular tree is defined by the following equation:

$$\alpha_{min}(T) = \frac{Nb + Nl + Ns}{NT} \quad (2)$$

where *Nb* is the number of branching node routers on tree T, *Nl* is the number of leaf node routers on the tree, Ns is the number of sources of the tree which always 1, and *NT* is the total number of routers on tree T. The $\alpha$ parameter of a tree reaches its minimum when all uni-multicast routers on the tree are bypassed by dynamic tunnels.

We observed that in a multicast topology (constructed tree) resulting from a traceroute experiments from the IRISA (university of Rennes 1) to 5 sites in France, there are only 4 branching node routers out of 30 routers. We deduced that the $\alpha$ parameter value is smaller than 34% when using tunnels between branching node routers which implies that we can achieve over 66% reductions in multicast forwarding table size using our approach. The forwarding table size in all routers in the network using the pure random sparse mode model is shown in Fig. 3 and using the waxman model is shown in Fig. 4.

The horizontal axis is the percentage of sources that are active in the network, and the vertical axis is the overall forwarding table size in the network. The poly-lines labeled PIM-x and SEM-x show the overall forwarding table size for PIM-SM and SEM protocols respectively when the number of receivers per group is x.

The forwarding table size grows with the number of active groups and the number of receivers, as predicted in section IV-A. From Fig. 4 and Fig. 3 we can see that the relative state information reduction of SEM is roughly 40% and 80% respectively compared to PIM-SM. We deduce also that our protocol is more suitable for sparse mode networks and for groups with few members.

### B. Tree Cost and Control Overhead Analysis

Our approach has an advantage over conventional multicast protocols like PIM-SM and CBT since we don't force multicast packets to be sent all the way to the Rendez-Vous point and next

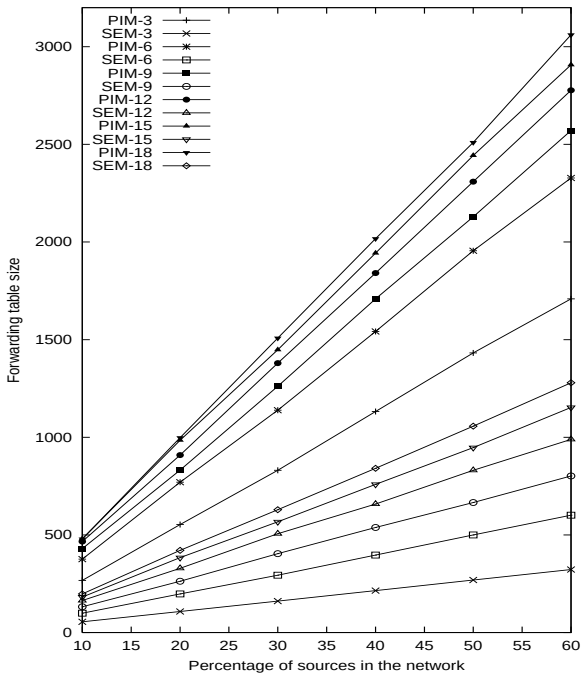



Fig. 3. Forwarding table size - pure random sparse mode model

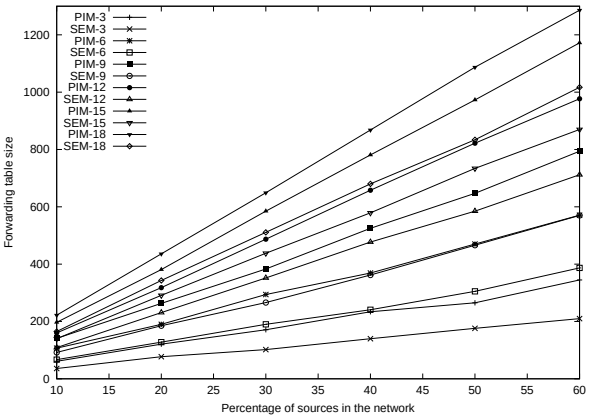



Fig. 4. Forwarding table size - Waxman model

to receivers. Packets follows only shortest paths between source and receivers. Besides there is no switching between shared tree and source specific tree. In HBH, as mentioned in Section II-B, periodic join messages will reach always the source and during the tree construction, tree messages and fusion messages are considered as extra overhead messages.

Otherwise, the control overhead of SEM can be measured using the total number of control packets sent per link or the total percentage of bandwidth spent on control traffic. In both PIM-SM and SEM, each distribution tree needs to be refreshed periodically. SEM uses BRANCH messages, PREVIOUS_HOP messages and ALIVE mesages to ensure the tree maintenance. First join message reaches always the source, while in PIM-SM it is intercepted by the nearest router that already joined the session. The number of control packets needed to refresh the states in PIM-SM and SEM would have been roughly the same, if there are no dynamic join and leaves since ALIVE messages between two branching node routers have the same impact as periodic join messages between routers in PIM-SM. The extra overhead in SEM is a result of periodic BRANCH messages and PREVIOUS_HOP messages. Currently the refresh period is fixed but in the future, we can make it adaptive to a number of factors including data rate of the flow, and the length between two branching node routers.

We used the simulation model represented in Fig. 5 when k, the number of routers between the source and the branching node router varies between 4 and 20 and n the number of receivers per group varies between 3 and 18.

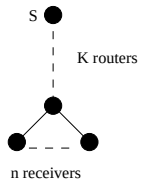



Fig. 5. Simulation model

Fig. 6 represents the number of the overall control packets for SEM and PIM-SM. The poly-lines labeled PIM-x and SEM-x show the number of control messages needed for PIM-SM and SEM protocols respectively when the number of receivers per group is x. We deduce from Fig. 6 that using our technique

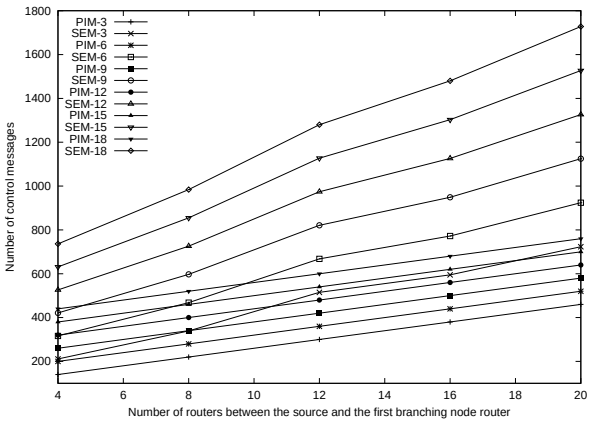



Fig. 6. Control messages vs. number of routers in PIM-SM and SEM

may doubles or triples in the worst cases the number of control packets needed for the tree maintenance comparing to PIM-SM.

### C. Data Processing and Delay

Having a medium group size, the Xcast header processing in every router for all packets travelling from the source to the destinations is considered to be very expensive for router ressources and increases the delay. Indeed, the header processing time for an Xcast packet is approximately proportional to the number of destinations and the header processing time for a simple unicast packet. Using the SEM protocol, only BRANCH

messages need extra header processing time. Comparing to Xcast, the packet header processing and thus delay in SEM are minimized. SEM supports larger number of members and consumes less of router ressources than Xcast.

## V. Conclusion and Future Works

In this paper, we presented SEM, a new small group multicast protocol. This protocol uses an efficient method to construct multicast trees and deliver multicast packets. Indeed, for multicast tree construction, a BRANCH message is sent to all destinations. The BRANCH message header contains all destination addresses and uses the same mechanism used by Xcast. When a BRANCH message discovers that a router is acting as a branching node, it creates a multicast state in that router. For multicast packets delivery, this protocol uses the branching node routers mechanism similar to that used in REUNITE and HBH.

SEM is a promising approach since it adopts the source-specific channel address allocation and implements data distribution using unicast trees. The application areas for SEM include conferencing, multi-player games and collaborative working.

As a result of analysis, while SEM has some control overheads compared to Xcast and Xcast+, its cost of packet header processing is minimized. Besides, while REUNITE has some advantages, it has a higher protocol complexity and larger number of control messages. SEM presents also many advantages over HBH protocol especially during the tree construction and state forwarding reduction in non branching node routers. We confirmed through simulation that SEM can significatly reduce multicast forwarding states and presents many advatages over other multicast protocols.

Our future work will focus on the latency problem, since join and leave messages take extra time comparing to other multicast protocols. A solution for this problem could be sending packets in xcast mode during the tree construction phase. Once the tree is constructed, packets will be send in SEM mode. We will also try to reduce the conrol overhead caused by the periodic BRANCH messages. We will study the incremental deployement, interoperability with other multicast protocols and the possibility of including QoS parameters inside SEM tree construction.

## References


[1] R. Boivie, N. Feldman, Y. Imai, W. Livens, D. Ooms, and O. Paridaens. Explicit multicast (Xcast) basic specification. IETF Internet draft, October 2000.

[2] I. Stoica, T. Eugene, and H. Zhang. REUNITE: A recursive unicast approach to multicast. In *INFOCOM (3)*, pages 1644–1653, 2000.

[3] D. Ooms. Taxonomy of xcast/sgm proposals. IETF Internet draft, July 2000.

[4] M. Ramalho. Intra- and Inter-domain multicast routing protocols: A survey and taxonomy. *IEEE Communications Surveys and Tutorials*, 3(1):2–25, First Quarter 2000.

[5] J. Tian and G. Neufeld. Forwarding state reduction for sparse mode multicast communication. In *INFOCOM (2)*, pages 711–719, March 1998.

[6] P. Radoslavov, D. Estrin, , and R. Govindan. Exploiting the bandwidth-memory tradeoff in multicast state aggregation. Technical report 99-697, University of Southern California, Dept. of CS, July 1999.

[7] L. HMK Costa, S. Fdida, and O. CMB Duarte. Hop-by-hop multicast routing protocol. In *ACM SIGCOMM'2001*, pages 249–259, August 2001.

[8] M. Shin, Y. Kim, J. Lee, and S. Kim. Explicit multicast extension supporting receiver initiated join. IETF Internet draft, February 2001.

[9] B. Cain, S. Deering, B. Fenner, I. Kouvelas, and A. Thyagarajan. Internet group management protocol, version 3. IETF Internet draft, January 2002.

[10] K. Fall. and K. Varadhan. The NS Manual. UC Berkeley, LBL, USC/ISI, and Xerox PARC, January 2001.

[11] E. Zegura, K. Calvert, and S. Bhattacharjee. How to model an internetwork. In *INFOCOM*, 1996.

[12] B. Waxman. Routing of multipoint connections. *IEEE Journal on Selected Areas in Communications*, 6(9):1617–1622, December 1988.