A Framework for MPLS Path Setup in Uni-Directional Multicast Shared Trees

Ashraf Matrawy Chung-Horng Lung Ioannis Lambadaris
Department of Systems and Computer Engineering
Carleton University, Canada
{amatrawy,ioannis,chlung}@sce.carleton.ca

Abstract

Establishing multicast communications in MPLS-capable networks is an essential requirement for a wide-scale deployment of MPLS in the Internet. This paper outlines a framework for the setup of a MultiPoint-to-MultiPoint (MP2MP) Label Switched Path (LSP) for establishing uni-directional multicast shared trees. The presented framework is intended for multicast applications within a single autonomous domain and can be extended to cover inter-domain multicast sessions.

We propose the use of one (or more) control points in the network called Rendez-vous Points (RP) in a manner similar to PIM-SM shared trees. Senders of the multicast session have to register with the RP and establish unicast LSPs with the RP. Receivers who join the session have to send their join requests to the RP which acts as a root (and the sender) of a one-to-many tree by establishing a Point-to-MultiPoint (P2MP) LSP between the RP and the receiver. This architecture utilizes more than one RP to implement RP failure recovery, to provide load balancing within the domain, and to enable the extension of this framework to multiple domains by establishing LSPs between RPs in different domains. This architecture also has the advantage of using existing MPLS techniques and existing routing protocols and requires only the addition of more management capabilities at the RPs. The paper explains the framework in details and provides an example on how to set the LSP on a given topology. We also refer to some preliminary simulation results testing the scalability of the architecture in comparison with traditional multicast routing.

I. Introduction

Traditional routing protocols in IP networks have proved to be inadequate in meeting the Traffic Engineering (TE) requirements of the current Internet. TE is concerned with performance optimization of operational networks [1]. It aims at providing efficient and reliable network operations while optimally utilizing network resources. Current TE techniques and IP routing protocols have some shortcomings [1], [2] that stems from the fact that IP routing algorithms in general search for the shortest path without taking overall network utilization into account. This might result in overloading some links (the shortest) while keeping others (longer paths) underutilized.

MultiProtocol Label Switching (MPLS) is designed as a an alternative to traditional IPO routing that allows flexible traffic engineering. In MPLS, packet headers are examined at network's entry points (routers). Based on the packet's header and on traffic engineering settings, the packet is assigned a *label*. Once the packet passes the first router, intermediate routers will only examine the label and *switch* the packet based on this label. In addition to a more flexible traffic engineering and the faster switching at a sub-IP layer compared to routing at the IP layer, MPLS has the following advantages [3]:

1) Information that is not traveling with the packet, such as the incoming port at the ingress point, can be used to determine the packet's route.
2) Packet forwarding decisions can be made more complex without adding complexity to routers' architecture (except for those at the edge).
3) Since the first label assignment determines the packet's route in the network, the packet's entry point to the network maybe used to decide on the packet's route.
4) Explicit routing is easier done with MPLS than with traditional routing.
5) For QoS architectures, MPLS labels may be used to *infer* the packet's precedence.

Besides its TE role, MPLS is also useful in providing the Virtual Private Networks service and in enabling QoS support in IP networks.

Work on the MPLS architecture [4], [5], [6], [7] started with unicast forwarding. Since multicast communications are becoming and essential part of data networks architectures, some proposals [8], [9], [10], [11], [12] advocated providing

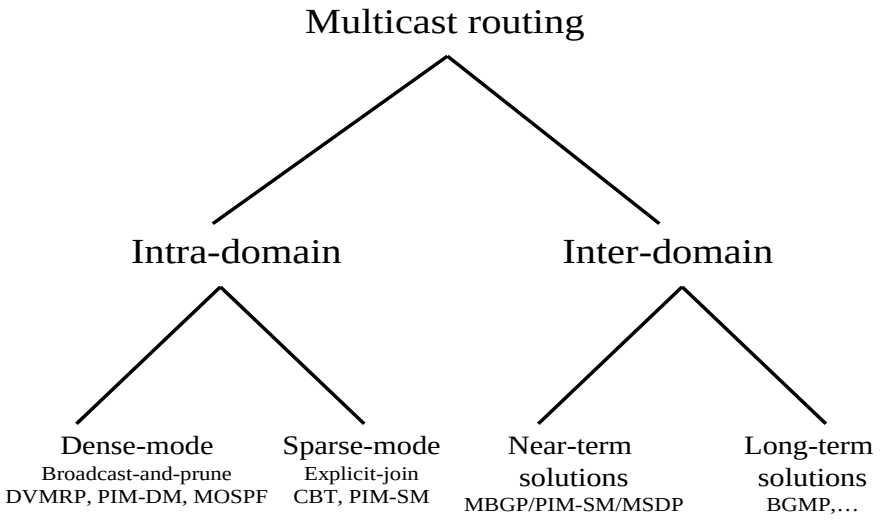

Fig. 1. A Classification of Multicast Routing Protocols

multicast support in MPLS networks. To provide multicast communications in MPLS networks, new elements have to be added to the MPLS architecture and some modifications/extensions will be required for the existing unicast MPLS architecture.

While MPLS offers flexibility and enhancements to unicast traffic, supporting multicast in an MPLS network is a nontrivial task. Some issues arise when an IP layer multicast tree (constructed using a multicast routing algorithm) is mapped to a layer 2 MPLS path. These issues include, the frequent flood/prune requests of multicast users, label consumption, big state tables, handling the co-existence of source and shared trees for the same multicast session, and implementing uni-directional and bi-directional shared trees. While in this paper we discuss the major issues, a detailed discussion on implementing IP multicast in MPLS networks can be found in [13].

In this paper, we present a framework for the setup of a MultiPoint-to-MultiPoint (MP2MP) Label Switched Path (LSP) for establishing uni-directional multicast shared trees. We use Rendez-vous Points (RP) in a manner similar to PIM-SM shared trees. This architecture utilizes more than one RP to implement RP failure recovery, to provide load balancing within the domain, and to enable the extension of this framework to multiple domains by establishing LSPs between RPs in different domains.

The rest of this paper is organized as follows. Section II presents an overview of multicast routing and discusses the need for different routing techniques to support different multicast applications. Section III defines the problems encountered in the area of multicast communications in MPLS networks. Related work on multicast Point-to-MultiPoint (P2MP) multicast trees is presented in Section IV. Our architecture is presented in Section V along with an example on how to setup MP2MP trees in Section VI. Section VII concludes the paper.

II. Multicast Routing

Understanding the difficulties in building a multicast tree in IP networks enables researchers to address the major problems when designing an architecture for multicast communications in MPLS networks. So this section briefly presents an overview of IP multicast routing algorithms [14]. Multicast routing started with a single flat routing space as opposed to the unicast hierarchical routing space. The first IP multicast on the Internet started in the early 1990s on the Multicast Backbone (MBone) [15]. MBone started with a few routers that can recognize forward multicast packets and that can run multicast routing. Because most routers on the Internet were not able to do that, unicast tunnels were created between the MBone routers creating a virtual topology on top of the Internet. Routing decisions were made using the Distance Vector Multicast Router Protocol (DVMRP) [16].

DVMRP uses the *flood-and-prune* technique. In this technique, messages from the source are sent to every router in the network. Routers will respond by sending a message back to the source on the interface the router finds the shortest back to the source. This method to construct a multicast tree is called a *Reverse Path Forwarding* tree. Leaf routers are responsible of keeping track of receivers that belong to a certain multicast session using the Internet Group Management Protocol (IGMP) [17]. When there are no receivers on a certain leaf sub-network, the corresponding router will send a *prune* message back to towards the source. The *flood-and-prune* techniques are suitable for networks that are densely populated where most routers are expected to be part of the multicast tree. They are usually referred to as Dense-mode techniques. Dense-mode protocols require the maintenance of a big amount of state information about all the routers in the network regardless of the existence of receivers.

As the number of modes in the MBone grew, more routers became able to support multicast and MBone started to have be a native multicast network. Sparse-mode multicast routing was proposed for multicast sessions where the number of routers expected to join a session is less than the total number of routers and when they are expected to be located in more than one domain. In Sparse-mode, receivers are expected to explicitly join the multicast session by sending a join-request.

Protocol Independent Multicast (PIM) [18], with its two modes PIM-DM and PIM-SM, evolved to be a widely used multicast routing algorithm. PIM-DM constructs multicast trees in a way similar to DVMRP. PIM-SM, on the other hand, is a complex protocol that is used to construct shared multicast trees where more than one source exist on the same tree. It uses a control point called Rendez-vous Point (RP). Each source has to register and send its packets to the RP. Receivers also have to send their join-requests to the RP which forwards the sources packets to the receivers.

All the protocols discussed so far are intra-domain protocols. When multicast started to be widely used, it became clear that hierarchical multicast routing is needed and inter-domain algorithms started to be developed. Due to the complexity of inter-domain multicast, a complete solution is not expected to be standardized by the IETF soon. Instead, a solution was adapted based on the protocols suite MBGP/PIM-SM/MSDP [14] which, despite the fact that it is functional, is not scalable. A classification diagram of multicast routing protocols is shown in Fig. 1.

III. Multicast Communications in MPLS Networks

As explained in Section II, the distribution of multicast traffic in IP networks is done using a distribution tree constructed by means of a multicast routing protocols. In order to distribute multicast traffic using MPLS networks, these multicast trees need to be mapped to layer 2 MPLS paths or Label Switched Paths (LSPs). This mapping procedure has to address some issues [13]. The most common of these issues summarized as follows.

- Label consumption in the multicast tree case will be high. In unicast, several unicast destinations can be aggregated to one LSP. It has not been studied yet how this can be done for multicast.
- Multicast trees usually experience frequent changes due to the join/leave requests of the users. Mapping these changes to layer 2 LSPs in the case of multicast MPLS should be efficient.
- Using shared trees is more efficient in the case of multicast MPLS than using multiple source trees. The drawback of shared trees is the requirement of using MP2MP LSPs. This requires the merging of some LSPs which may not supported by some Layer 2 technologies such as ATM.
- Multicast shared trees can be bi-directional in some cases. While they are more efficient in terms of resource utilization, they produce more merging points in the tree compared to uni-directional trees.

Mapping the multicast tree to layer 2 and the creation of LSPs are triggered by different events. Three types of events can trigger the creation of a multicast LSP [9], [13]:

- LSP Request: either by a multicast routing protocol or by resource reservation protocol.
- Topology changes: if the tree at layer 3 changes, the change should be mapped to layer 2.
- Traffic changes: in this case, LSPs are created only when traffic exists on this branch of the tree. This saves labels and is suitable for explicit-join multicast. This trigger is the one we adopt for our work.

In some cases, it can be beneficial to piggy-back label request on existing control messages instead of sending separate MPLS control messages. Two candidates for this are multicast routing messages [12] and RSVP-TE messages [19]. Using this mechanism in a certain network depends on a few factors such as which triggers are used in the network and whether multicast routing is available.

IV. P2MP LSP Establishment Approaches

This section provides a summary of the work that has been done so far in the area of IP multicast in MPLS networks. Some of this work is done by IETF-affiliated groups to develop standards while other work is a research effort concerned with the performance of multicast architectures in MPLS networks. While RFC3353 [13] is so far the main work in this area by the IETF, it does not specify any solution to the problem. Instead, it presents a framework for IP multicast deployment in an MPLS environment.

One group of solutions proposed to the IETF is based on piggy-backing labels on multicast routing messages. In particular, some proposals assume that the join messages of the PIM-SM multicast routing protocol will be used to send labels upstream to the multicast source [10], [12]. This method was adopted in [20], [21] to build multicast tree for the PIM-SM (the source specific mode) where join messages propagate upstream till it reaches a router that is actually part of the multicast tree. At this point, label assignments take place and a special database that keeps track of mapping incoming labels to multiple outgoing labels. A simulation environment to study this solution was developed in [21] which is an extension to the simulators [22], [23].

The other group of mechanisms does not rely on multicast routing. Instead they rely on extending RSVP-TE to support the establishment of P2MP LSPs [24], [25]. The work in [24] specifies mechanisms for both sender-initiated and leaf-initiated signaling. It argues for the need to use RSVP-TE rather than conventional multicast routing protocols due to the lack of inter-domain multicast routing in some domains, the non-optimality of trees established by multicast routing, and the fact some multicast applications are not very dynamic so pre-established sender-initiated LSPs maybe suitable for these applications. In [25], similar work is based on RSVP-TE sender-initiated LSP setup with emphasis on optimizing packet replication and minimizing state in the network core.

V. MPLS MP2MP LSP Setup

This section presents our proposed architecture for MP2MP MPLS LSP setup to support multicast applications that require the establishment of uni-directional shared trees. This architecture requires the existence of at least one designated control point, called Rendez-vous Point (RP) in the network domain where the tree exists. An RP will serve as a meeting point where senders send packets to be distributed using a Source Specific Multicast (SSM) tree rooted at that RP.

A. Assumptions

These assumptions should hold true in the network domain where this architecture is applied.

- We are considering the case of multi-source multi-receiver multicast. Other methods are more applicable to single-source multicast [20].
- This work is concerned with multicast within an MPLS domain. If a branching node is located at the border with a non-MPLS receiver. Both layer 2 and layer 3 forwarding mechanisms should exist at this node.
- Hop-by-hop routing is assumed. Explicit routing is not considered at this stage of our work.
- The multicast application is dynamic, frequent join/leave requests are expected.
- An FEC (and an associate LSP) will be associated with each multicast tree, i.e. with the tree D-class address.
- Labels will be distributed starting from leaf LSRs (Label Switching Routers), i.e., label distribution will be Unsolicited-Downstream.
- Label distribution will be piggybacked with the join request messages.
- Any end-node can act as a sender and a receiver at the same time.

B. Information bases

In addition to the traditional MPLS mapping of incoming labels to output labels at each Label Switching Router (LSR), we propose the use of the following additional Information Bases (IB). A mapping of the multicast group address to the sources and their corresponding labels is the first IB we suggest. A similar entity, the LSG table (Label for Source and Group) is used in [21] for the single-source case. We modify this LSG to accommodate the multi-source case. In our architecture, this IB resides at the RP. An example of that IB is shown in Table I where we have two multicast groups $G1$ and $G2$. There are n sources subscribed to $G1$ and m sources subscribed to $G2$.

TABLE I
INFORMATION BASE MAPPING GROUPS TO SOURCES

Group and Label	Source	Incoming interface
G_1, L_{G_1}	S_{11}	$I_{S_{11}}$
	.	.
	S_{1i}	$I_{S_{1i}}$
	.	.
	S_{1n}	$I_{S_{1n}}$
G_2, L_{G_2}	S_{21}	$I_{S_{21}}$
	.	.
	S_{2i}	$I_{S_{2i}}$
	.	.
	S_{2m}	$I_{S_{2m}}$

TABLE II
INFORMATION BASE MAPPING INCOMING LABELS TO OUTGOING LABELS

(Group label)	(Outgoing interface,Label)
L_{G_1}	$(O_{R_{11}}, L_{R_{11}})$
	.
	$(O_{R_{1a}}, L_{R_{1a}})$
L_{G_2}	$(O_{R_{21}}, L_{R_{21}})$
	.
	$(O_{R_{2b}}, L_{R_{2b}})$

To save labels, all senders who request to register with a certain group will be asked by the RP to use the same label when they sent their traffic to the RP. This will save labels and will save space in the next IB as we will explain shortly. In this example, we call the label L_{G_1} for group G_1 and L_{G_2} for group G_2.

The second Information Base is the Label Information Base (LIB). It maps an incoming label to more than one label on different outgoing interface. An incoming label to the RP is unique per group, hence there will be as many entries per group in the LIB as the number of outgoing interfaces subscribed to this group. If we use a different label for every sender in a group, the LIB will grow linearly with the number of senders as we will have to repeat the outgoing interfaces with every sender's label. The LIB should exist on every branching node in the distribution tree rooted at the RP. However, the way this IB will be created at the RP is different from that of any other branching point due to the source merging process performed at the RP. In Table II, the mapping is shown between label L_{G_1} and a outgoing interfaces for group G_1. Similarly, label L_{G_2} is mapped to b outgoing interfaces for group G_2.

C. LSP establishment

The following are the details of the LSP establishment of this architecture.

- The selection of an RP is a network planning decision. In [26], some guidelines are provided on how to select RPs for PIM-SM.
- A backup RP should also be selected to support the recovery from a failure of the primary RP, to provide load-balancing capabilities, and to facilitate future extension of this architecture to support inter-domain multicast in MPLS networks. Both the primary and the backup RPs should work together to keep their Information Bases synchronized. The selection of a backup RP is beyond the scope of this paper. For the rest of this paper, when we refer to the RP, we mean the primary RP.
- Senders must register with the RP before they start to send traffic to the RP. Receivers, on the other hand, have to explicitly join the multicast session to start receiving the traffic from the RP.
- When a sender registers with the RP to send its traffic to a group G. This registration will involve the creation of a unicast LSP and label binding between the sender and the RP. The RP will assign a single label as the incoming label for this

group, L_G. A record for that relates this sender, incoming interface to the RP, the group incoming label L_G, and group G will be created in the LSG table described in Table I.

- If a source tries to register with an RP in a group G where no receiver has registered, a *source-register-error* should be sent to the source from the RP after a duration of *register-time-out* seconds, releasing the reserved labels.
- Similarly, if a receiver tries to join a group where no sender is registered, a *join-request-error* should be sent to that receiver from the RP after a duration of *join-time-out* seconds, releasing the reserved labels.
- A source-based multicast tree rooted at the RP will be built in the same way PIM-SM builds a multicast tree, i.e., based on explicit join requests from receivers.
- Requests to join group G from the end-receivers will travel upstream towards the RP with labels piggy-backed on them. A request travels upstream until it either reaches an LSR on the tree that receives packets from group G or reaches the RP.
  1) If the request reaches an LSR on the tree for group G, a branch from the tree starting at this LSR will be extended down towards the receiver. Label binding will be based on the label forwarded with the request message. An LSP will be setup from this LSR towards the receiver. An entry will be added to the LIB at this LSR. This is similar to the source-based tree creation proposed in [20].
  2) If the request reaches the RP, a record will be added to the LIB of the RP that maps the label for this group to the outgoing interface leading to this LSP (refer to Table II).
- If a receiver decides to leave the group, a *prune* message is sent towards the RP. The first LSR on the tree that receives this message, whether an LSR or the RP itself, will delete the record it has for the branch for this receiver and release the labels on the LSP downstream towards the receiver.

VI. Example

Fig. 2 illustrates the setup of an MP2MP LSP for a single multicast group G. In the figure, three unicast LSPs are established between senders S_1, S_2, and S_3. A source-based tree is established with its root at router RP. Labels from the three sources are mapped to those on the outgoing branches of RP. Label mapping at all LSR nodes will be done using unicast MPLS techniques except at RP.

Table III shows the mapping between the sources and group G at RP. The notation in the table is set so that I_{S_1} denotes the incoming interface at RP from S_1. Also note that all three sources will be asked by the Rp to use L_G as their labels when sending their traffic to the RP. This type of table only exists at RP and copy should be kept at the backup RP. Table IV shows the mapping of incoming labels to outgoing labels at RP. In the case other branching points exist in tree, tables similar to Table IV should be created at these points.

TABLE III

IB mapping groups to sources at node RP

Group and Label	Source	Incoming interface
G, L_G	S_1	I_{S_1}
	S_2	I_{S_2}
	S_3	I_{S_3}

TABLE IV

IB mapping incoming labels to outgoing labels at node RP

Group label	(Outgoing interface,Label)
L_G	(O_{LSR_1}, L_{LSR_1})
	(O_{LSR_3}, L_{LSR_3})
	(O_{LSR_4}, L_{LSR_4})

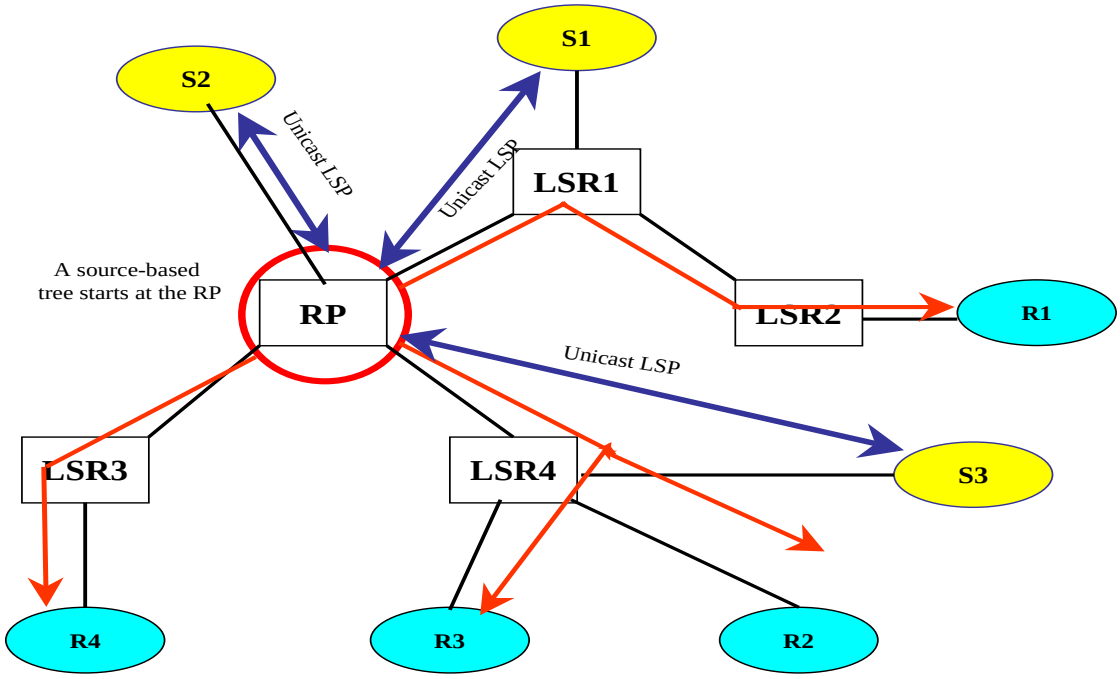

Fig. 2. Architecture for MP2MP LSP setup

VII. Conclusion and future work

In this paper, we have presented a framework for the creation of MP2MP MPLS LSP to support multicast communications that requires uni-directional shared trees. This framework does not require the use of multicast routing and uses most of MPLS unicast mechanisms. In the example we have shown in the paper, it seems that a bi-directional tree would be more efficient from some nodes, e.g. from S_3 to R_2 and R_3 but this is more complicated to establish and would result in more branching/merging points in the tree. We are in the process of creating a simulation environment, based on [22], [23], [21], to study the performance of this proposal. We evaluated the scalability of this architecture using simulation and reported the results in [27]. Other issues that are on our future research plan in this area are efficient load-balancing techniques, protection and recovery, and inter-domain multicast in MPLS networks.

References

[1] D. Awduche *et al.*, "Requirements for Traffic Engineering Over MPLS," *RFC 2702*.
[2] X. Xiao *et al.*, "Traffic Engineering with MPLS in the Internet," *IEEE Network*, March/April 2000.
[3] E. Rosen *et al.*, "MultiProtocol Label Switching Architecture," *RFC 3031*.
[4] "IETF MPLS Working Group," *http://www.ietf.org/html.charters/mpls-charter.html*.
[5] "MPLS Research Center," *http://www.mplsrc.com*.
[6] B. Davis and Y. Rekhter, *MPLS Technology and Applications*. Morgan Kaufmann, 2000.
[7] V. Alwayn, *Advanced MPLS Design and Implementation*. Cisco Press, 2002.
[8] A. Acharya and F. Griffoul, "IP Multicast Support in MPLS," *Proc. of IEEE ATM Workshop*, 1999.
[9] D. Ooms and W. Livens, "IP Multicast in MPLS Networks," *Proc. of High Performance Switching and Routing*, 2000.
[10] D. Ooms *et al.*, "MPLS for PIM-SM," *draft-ooms-mpls-pimsm-00.txt*, Work in progress.
[11] D. Farinacci, "Partitioning Label Space among Multicast Routers on a Common Subnet," *draft-farinacci-multicast-label-part-00.txt*, Work in progress.
[12] D. Farinacci *et al.*, "Using PIM to Distribute MPLS Labels for Multicast Routes," *draft-farinacci-mpls-multicast-02.txt*, Work in progress.
[13] D. Ooms *et al.*, "Overview of IP Multicast in a Multi-Protocol Label Switching (MPLS) Environment," *RFC 3353*.
[14] K. C. Almeroth, "The Evolution of Multicast: From the MBone to Inter-domain Multicast to Internet2 Deployment," *IEEE Network*, January/February 2000.

[15] H. Eriksson, “MBONE: The Multicast Backbone,” *Communications of The ACM*, August 1994.
[16] D. Waitzman *et al.*, “Distance Vector Routing Multicast Protocol (DVMRP),” *RFC 1075*, November 1988.
[17] W. Fenner, “Internet Group Management Protocol, Version 2,” *RFC 1112*.
[18] S. Deering *et al.*, “PIM Architecture for Wide-Area Multicast Routing,” *IEEE/ACM Trans. on Networking*, April 1996.
[19] D. Awduche *et al.*, “RSVP-TE : Extensions to RSVP for LSP Tunnels,” *RFC 3209*.
[20] A. Boudani and B. Cousin, “A New Approach to Construct Multicast Trees in MPLS Networks,” *Proc. of Seventh International Symposium on Computers and Communications*, 2002.
[21] A. Boudani *et al.*, “Multicast Routing Simulator over MPLS Networks,” *Proc. of the 36th Annual Simulation Symposium*, 2003.
[22] S. McCanne and S. Floyd, “ns Network Simulator,” *http://www.isi.edu/nsnam/ns/*.
[23] G. Ahn and W. Chun, “Design and Implementation of mpls network simulator supporting ldp and cr-ldp,” *Proc. of IEEE International Conference on Networks*, 2000.
[24] S. Yasukawa *et al.*, “Extended RSVP-TE for Multicast LSP Tunnels,” *draft-Yasukawa-mpls-rsvp-multicast-01.txt*, Work in progress.
[25] R. Aggarwal *et al.*, “Establishing Point to Multipoint MPLS TE LSPs,” *draft-raggarwa-mpls-p2mp-te-00.txt*, Work in progress.
[26] B. Williamson, *Developing IP Multicast Networks*. Cisco Press, 1999, vol. 1.
[27] A. Matrawy *et al.*, “ On the Scalability of using MPLS in Multicast Shared Trees,” *Submitted to Workshop on the Quest to Control Next Generation Transport Networks, at IEEE Globecom2004*

.