Internet Engineering Task Force
INTERNET DRAFT
Expires April 2000

C-Y Lee
L. Andersson
Nortel Networks
Ken Carlberg
SAIC
Bora Akyol
Pluris
October 1999

Engineering Paths for Multicast Traffic
<draft-leecy-multicast-te-01.txt>

Status of this memo

Abstract

This document describes a solution to engineer paths for IP multicast traffic in a network, by directing the control messages to setup multicast trees on engineered paths. This enables the network operator to have control over the topology of multicast trees.

This proposal partitions the multicast traffic engineering problem such that multicast routing protocols do not have to be modified to allocate resources for multicast traffic nor do resource allocation protocols such as RSVP or CR-LDP have to be able to setup forwarding states (in this case labels) like multicast routing protocols.

Resources are allocated on the same trip that paths are selected and setup. This prevent the problem of data being forwarded on branches of the tree where resources have not being allocated yet. An important aspect of this proposal is that it enables multicast paths

to be engineered in an aggregatable manner, allowing this solution to scale in the backbone.

1. Overview

In general, traffic is engineered to traverse certain paths so as to utilize resources in a network in a more optimal manner, while at the same time improving the level of service that can be offered.

In conventional IP routing, traffic may be engineered to use a path by configuring preferred links towards a destination with a lower metric. This method only allows traffic to be engineered based on the destination address. Since the forwarding is based on the destination address only, traffic cannot be engineered based on other attributes (which maybe useful for traffic engineering purposes) of the packet such as the source address of a packet or the requested service level. In contrast, MPLS abstracts the forwarding paradigm and allows traffic to be forwarded based on attributes (known as forwarding equivalence class (FEC) in MPLS) in addition to the destination address. This provides a versatile and convenient syntax for traffic engineering purposes.

This document describes a way to provide a basic traffic engineering mechanism for multicast. Traffic Engineering (TE) functionalities (in the MPLS entity) are used to decide where to forward the join control messages of multicast protocols, based on different traffic engineering requirements and to allocate resources. (Note that multicast data packets however are forwarded based on Layer 3 (L3) address information and are not label switched.)

Using this basic multicast traffic engineering mechanism, ISPs can define particular FECs for their network, resources required to receive traffic from certain root prefix, decrease fanouts at a node by limiting the number paths towards the node(prefix), allowing only certain paths to carry multicast traffic, experiment with heuristics to better engineer multicast trees, use a function to dynamically compute suitable paths based on current or predicted network resources. All these additional network or content provider specific functions to engineer traffic can be developed independently of the basic multicast traffic engineering scheme.

2.0 Motivation

The fundamental problem with doing multicast Traffic Engineering (TE) is the difficulty in doing it in a scalable manner. Multicast routes are very difficult (and some claim impossible) to aggregate. One can associate a label with a unicast route(prefix) and packets sent to that destination can be aggregated by associating them with the

label.

Since multicast routes are not aggregatable in general, associating a label with a multicast route implies per flow/group resource allocation. In essence, this kind of association will result in RSVP (or ATM) style resource allocation and is more applicable to per flow QOS than traffic engineering.

In contrast the approach taken in this proposal decouples traffic engineering from multicast route setup, thereby allowing the resources and paths for multicast data delivery to be independently allocated. What this implies is, resources and paths can be aggregated and engineered; and traffic can be statistically multiplexed, enabling network operators to provide differentiated services for multicast traffic in a scalable manner.

3.0 Scope

This draft described mechanisms which is applicable to multicast routing protocols such as PIM-SM, CBT, BGMP, Express or Simple Multicast, which will be called 'control driven' in this draft. The mechanisms to handle 'Data driven' or flood and prune protocols (eg DVMRP and PIM-DM) is FFS. This proposal assumes a multicast group/tree has a common service level requirement. It is envisaged that heterogeneous receivers requirement can be met by layer encoding data in different multicast groups or other variation of layer encoding.

It should be noted that the MPLS concepts of interest here are the FEC, ERO and resource allocation and path selection. Although the proposed scheme do not use label switching the solution is described in MPLS terms since the concepts of interest here have already been defined in MPLS.

4.0 Approach

A control driven multicast routing protocol sends a 'join' message to graft a node to a multicast distribution tree, creating multicast routes in the process. Since the join messages are forwarded based on unicast routes, if the conventional routing table is used, the multicast routes setup will be based on conventional routes. To constrain multicast paths, the join message can be sent via paths, computed or statically configured.

This draft describes a scheme where multicast routing control messages (including join messages) are forwarded by the TE entity in a router on the constraint path.

To allow a router to process control messages, the control messages should contain the router alert option. The control message is identified at the egress router by its FEC. Based on the FEC, the MPLS entity can derive the path the control message should take and allocate resources as specified. A multicast routing protocol would setup the forwarding state on the ports/interface where the join is received. To enable the establishment of multicast forwarding state based on constraint (unicast) routes, multicast routing protocols which verify the Reverse Path Forwarding (RPF) must turn off this check or be able to obtain the 'constraint' RPF via a Constraint Based Routing (CBR) API. To prevent redundant data and loops, a loop avoidance scheme based on the concepts described in [MPLS-LOOP-AVOID] or [SM] can be used in the routing protocol. If there is a loop, the routing protocol should not create forwarding states for the group on the port where the join is received.

Other alternatives to send the join on the engineered path such as - extending CR-LDP/TE-RSVP to send and merge joins for the multicast tree associated with a label - changing the multicast routing protocol to send the join along the explicit route, either require multicast routing protocol functionalities to be present in MPLS or MPLS functionalities to be incorporated into multicast routing protocols. This proposal uses MPLS (label and explicit route object) to cause engineered paths to be selected but forward data using multicast routing. It does not require MPLS or multicast routing protocols to be merged, an exercise which tend to - result in redundant or the reinventing, of functionalities at L2/L3; increase the complexity of multicast traffic engineering while not providing any means of aggregating multicast traffic engineering.

The alternative approaches listed above require traffic to be engineered for each group/tree since multicast labels/routes are most likely to be not aggregatable. Each group must be assigned a different label as well. In contrast this proposal allows a network provider to aggregate the engineered path towards a root or root prefix (since resource allocation and path selection can be independent of the setup of forwarding states/routes). The root prefix could be a subnet or domain. Multicast traffic in the backbone network can then be, provisioned in a more scalable manner and statistically multiplexed on the (aggregated) engineered paths.

5.0 Procedure

5.1 At the Egress Router

At any egress router (a router where multicast data exits the network) the IP fields of interest in the control message (referred to as FEC here, for lack of a better term), the associated path

selection mechanisms are defined in a Traffic Configuration table. These FECs correlate to the control messages of routing protocols. (eg, destination = root prefix/target-node address, ToS=codepoint). Note that the message carrying this information traverses the network from egress to ingress. The path selection mechanisms can be based on, a static table or a Constraint Based Routing (CBR) table, or a path selection algorithm (dynamic). The resources required for the FEC can be statically configured at the egress router or obtain via other means as described in [MC_DS_PROV].

Figure 1 shows the passage of control messages in an egress router (dotted lines) and the interface between the various entities in the router (+++ lines)

When a join message arrives at the egress router the packet is processed by the appropriate multicast routing protocol, to setup multicast forwarding states. If there are already forwarding states, a join message is discarded, otherwise, the multicast routing protocol calls an API provided by the Multicast Traffic Engineering (MCTE) entity to get the next hop to the root.

The form of the API is represented in terms of the following:

```
get_MCTE_next_hop(Target-Node, Group);
```

Target-Node is a mandatory value. The value of Target-Node is in the form of an IP address. Group is not required for (a)-(c) and optional for (d) below. The return value is the next hop to the Target-Node.

The MCTE entity : a) obtains the route from conventional routing if no path or path selection mechanism is specified in the Traffic Configuration table, or b) obtains the manually configured explicit route in the Traffic Configuration table or c) obtains the explicit routes via a CBR process (Refer to [MPLS-TE] and [ISIS-TE]/[OSPF-TE] for details) or d) invokes the path selection algorithm, specified in the Traffic Configuration table. (Note: the routes in (a)-(c) are based on the network topology, whereas (d) may take into account the tree topology in the computation of routes)

The MCTE entity stores the route(s) obtained or computed for this FEC, and used these routes when it prepends a MCTE header in the control message later.

The form of the API provided by the path selection algorithm in (d) above is represented in terms of the following:

```
get_MCTE_route(Target-Node, Group, Type-of-Metric)
```

Target-Node is a mandatory value, and the rest are optional in their usage or applicability. The value of Target-Node is in the form of an IP address. The return value is a list of explicit route(s). (Note: currently, the above API assumes IPv4. A different API will be used for IPv6)

The other parameters of the API are optional. The Group represent an added level of granularity by which network administrators can base their traffic engineering decisions (e.g this allows per group/flow traffic engineering). (Note: currently, port values are not included due to the common practice of correlating session to group address).

Finally, the Type-of-Metric value correlates to different types of metrics used to distinguish one path from another. The default value is (1), which correlates to hop count. Other defined values consist of: (2) bandwidth, (4) delay, and (8) fan-out. In cases where the underlying algorithm (of get_MCTE_route) does not support metrics other than hop count, this field is ignored. The Type-of-Metric is specified with the path selection algorithm in the Traffic Configuration table.

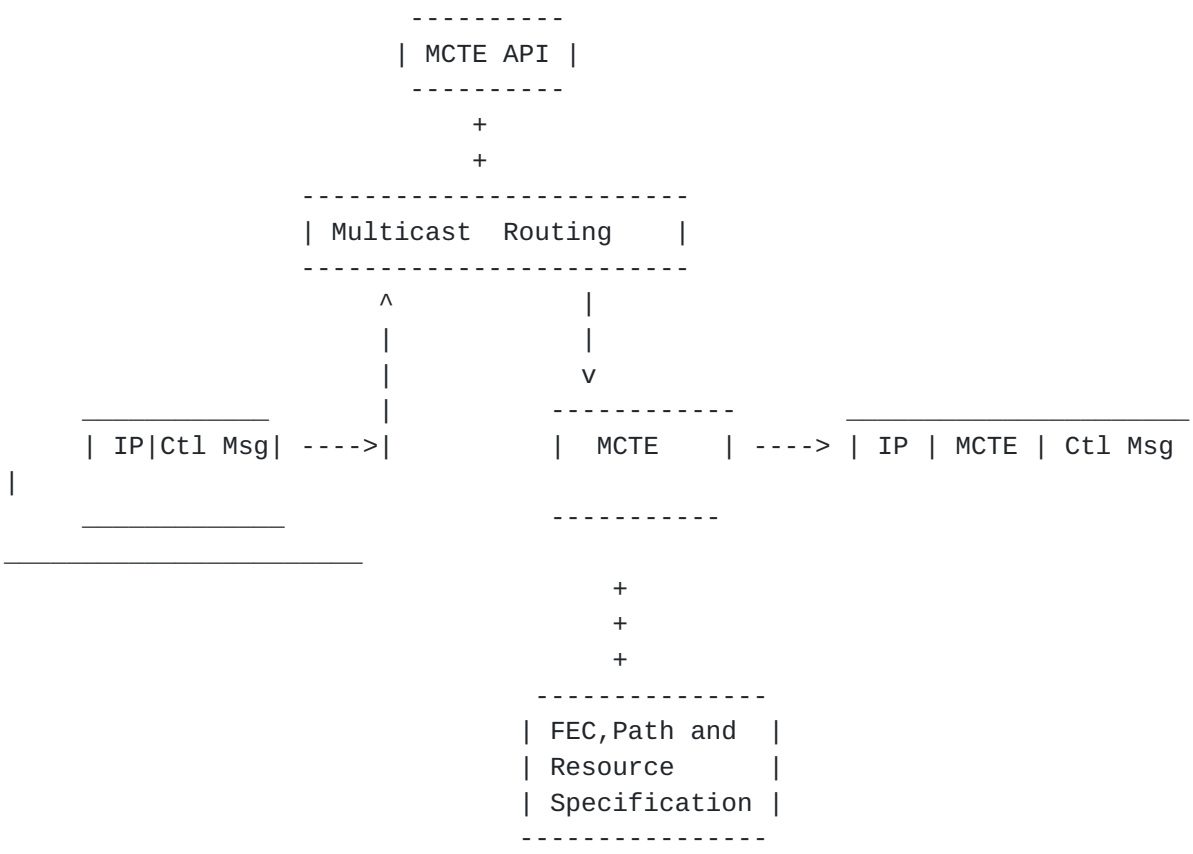


Fig. 1 At the egress (wrt data flow) router

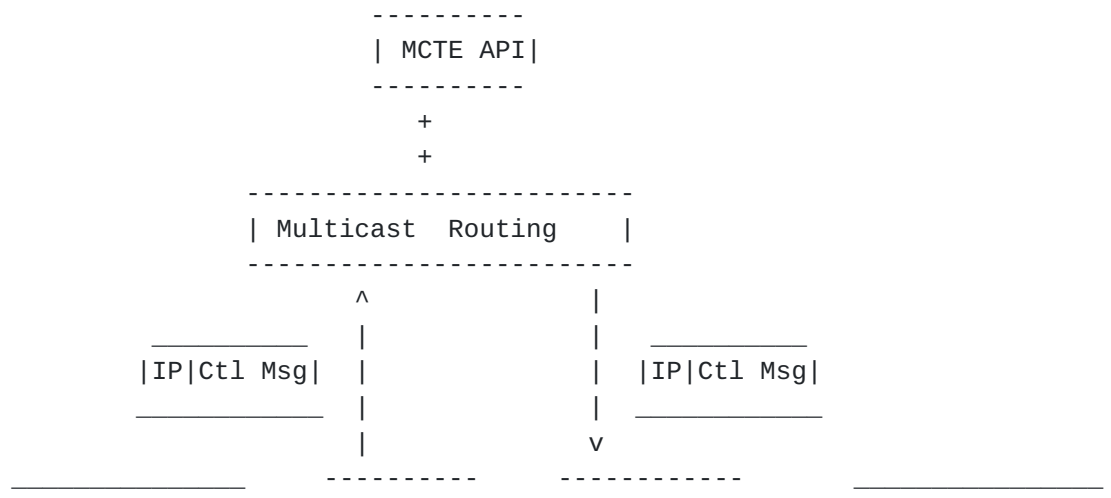
After the multicast forwarding states are setup, the control message is forwarded towards the root. If the control message matches a defined FEC, it is diverted to the MCTE entity. How the outgoing control message is diverted to the MCTE entity is implementation dependent. The MCTE entity calls an API provided by the MRP (Multicast Routing protocol)to find out whether the control message is a path setup (join), path teardown (leave) message or other maintenance message. If it is a path setup, resources specified in the Traffic Configuration table is allocated, if it is a path teardown message the resources are deallocated. If it is a maintenance control message, the control message is forwarded as is without any MCTE header and will be forwarded by the multicast routing protocol in intermediate routers as per normal.

If it is either a path setup or path teardown message, the MCTE entity prepends a MCTE header - containing the FEC, explicit routes (provided by the path selection mechanism) resources required (e.g Traffic Parameter, service level) and the protocol id of the control message. The IP protocol id is set to IPPROTO_MCTE.

The MCTE header is placed between the IP header and the control message. Resources as specified in the Traffic Configuration table are allocated/deallocated before the MCTE message is forwarded to the next hop returned by the path selection mechanism specified. To allow other routers to process this MCTE message (which includes the control message), the packet will be labeled as Router Alert.

5.2 At the Intermediate Routers

Figure 2 shows the passage of control messages in an intermediate router (dotted lines) and the interface between the various entities in the router (+++ lines)



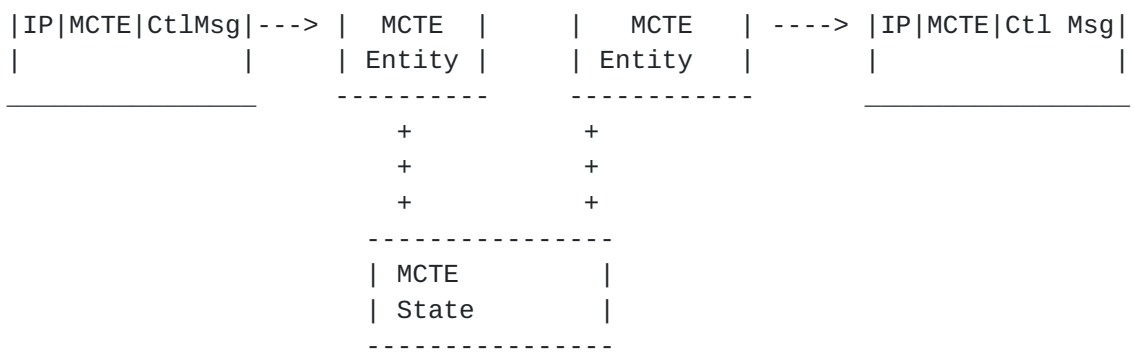


Fig. 2 At an intermediate router

When the next hop (or other intermediate nodes) receives the packet with Router Alert, it will be taken out of the forwarding path and directed to the MCTE entity since the IP protocol id is IPPROTO_MCTE.

The MCTE entity allocates/deallocates the resources requested by the MCTE message, creates a transient state for the MCTE message, called the MCTE state, for short. The appropriate mutlicast routing protocol (MRP), depending on the value of protocol id in the MCTE message, is then invoked. The exact mechanisms used in the router to accomplish this is implementation dependent.

The MRP creates the forwarding state for the group and forwards the join message towards the root. As in the egress router, the next hop towards the root is obtained from an MCTE API. Since the FEC for this control message matches the MCTE state created earlier, the join message is diverted to the MCTE entity. The MCTE entity placed the corresponding MCTE header on the control message and forwards the message to the next hop. The transient MCTE state is removed at this point.

Note that the FEC is only configured at the egress router (wrt to multicast data), intermediate routers are informed of the FEC information by previous hops. Similarly, the explicit or constraint route is only configured or computed at the egress router; the next hop and other intermediate nodes learn of the explicit routes via the explcit route list propagated from the egress router.

5.3 Loops

If the MPLS control message specifies looping explicit routes :

* then if the tree is uni-directional, only the join message will loop. Data will not loop since data flow is only in one direction

from root to members.

* then if the tree is bi-directional, the join message will loop, but because permanent states would not be established in this case, data will not be forwarded on the looping path.

However if there is a change in next hop towards the root at a node where there is already an existing forwarding state, then multicast routing protocols which uses bi-directional trees or a hybrid of uni-directional and bi-directional branches could invoke a loop avoidance procedure. One way to avoid loops in this case is (using splice message) described in [SM] and [MPLS-LOOP-AVOID].

6.0 Path Selection

This proposal allows different path selection algorithms to be used, depending on the FEC and path selection mechanism association. Paths can be configured, computed, discovered or obtain through other means.

A path selection mechanism will return the constraint routes given for e.g the group address, root of multicast tree and possibly other criteria. How the paths are selected are independent of this proposal, but a generic interface (API) between path selection algorithms and this multicast traffic engineering scheme is required and is specified in Section 5.1.

7.0 Applications

This section list some possible applications of this proposal.

a) A network operator may define an explicit route [Rx, Ry, Rz] towards a domain with prefix 10.0.0.0 for multicast traffic. Any member joining a group where the root address has the prefix 10.0.0.0 will have data delivered to it via the explicit route [Rz, Ry, Rx] (data is in the reverse direction of the join control message).

This explicit route may be a Loose Source Route, or a route calculated by an algorithm eg an Internal Gateway Protocol (IGP) which can provide constraint based routes.

It is worth noting that the explicit route can be the desired path from a root towards a member instead of the reverse path (from member towards the root).

b) Another variation of the above may define an additional field of interest in the FEC, the TOS. This will allow a network operator, to engineer paths or/and provision resources for traffic requiring

Expedited Forwarding [EF] or Assured Forwarding [AF]. (Refer to [MCPROV]).

c) To decrease fanout, egress routers (where multicast data traffic exits) can obtain the contraint routes towards the root of the tree and construct the tree along these paths instead. These routes can be statically configured or provided by an algorithm which takes into account fanout in route computation and this can be developed independently of the basic TE scheme described in this proposal.

d) Load Balancing - a load balancing algorithm can provide an alternative path that a control message can take depending on the service level requirement of the group and the current utilization of the equal cost paths.

e) Policy routing - Different paths may be defined for different groups.

8.0 Acknowledgments

The authors are grateful to Dirk Ooms and Yunzhou Li for reviewing this draft and their helpful suggestions to improve this proposal, Jamal Hadi-Salim for his technical advice and Jon Crowcroft for providing insightful comments.

References

[ARCH] E. Rosen, A. Viswanathan, R. Callon, "Multiprotocol Label Switching Architecture", Work in Progress, July 1998.

[MPLS-TE] Awduche, D. et al., "Requirements for Traffic Engineering over MPLS", Internet Draft, draft-ietf-mpls-traffic-eng-00.txt, October 1998.

[CRLDP] L. Andersson, A. Fredette, B. Jamoussi, R. Callon, P. Doolan, N. Feldman, E. Gray, J. Halpern, J. Heinanen T. E. Kilty, A. G. Malis, M. Girish, K. Sundell, P. Vaananen, T. Worster, L. Wu, R. Dantu, "Constraint-Based LSP Setup using LDP", Work in Progress, January, 1999.

[ISIS_TE] Smit, H. and T. Li, "ISIS Extensions for Traffic Engineering," draft-ietf-isis-traffic-00.txt, work in progress.

[OSPF-TE], D Katz, D Yeung, "Traffic Engineering Extensions to OSPF", draft-katz-yeung-ospf-traffic-00.txt

[TE-RSVP] D. Awduche, L. Berger, D-H. Gan, T. Li, G. Swallow, Vijay Srinivasan,

Internet Draft, draft-ietf-mpls-rsvp-lsp-tunnel-02.txt, September 1999

Multicast Routing with resource reservation,
Journal of High Speed Networks 7 (1998) 113-139,
B. Rajagopalan, R. Nair

CBT, Core Based Tree Multicast Routing,
Internet-Draft, March 1998, Ballardie, Cain, Zhang

PIM-SM, Protocol independent multicast-sparse mode Specification,
RFC-2117, June 1997
Estrin, Farinacci, Helmy, Thaler, Deering, Handley,
Jacobson, Liu, Sharma, and Wei.

BGMP, Border Gateway Multicast Protocol Specification,
Internet-Draft, March 1998, Thaler, Estrin, Meyers

Express, H. Holbrook, D. Cheriton
Sigcomm Paper

SM, Simple Multicast, Internet-Draft, March 1999,
draft-perlman-simple-multicast-02.txt, Perlman et al

YAM, K. Carlberg, J. Crowcroft
Hipparch 1998

[MPLS-LOOP-AVOID] "Avoiding Loops in MPLS", Internet Draft,
draft-leecy-mpls-loop-avoid-00.txt, June 1999
C-Y Lee, L. Andersson, Y. Ohba,

[CLARK] D. Clark and J. Wroclawski, "An Approach to Service
Allocation in the Internet", Internet Draft

[DSHEAD] K. Nichols and S. Blake, "Definition of the
Differentiated Services Field (DS Byte) in the IPv4 and IPv6
Headers", Internet Draft, May 1998.

[AF] J.Heinanen, F.Baker, W. Weiss, J. Wroclawski
Assured Forwarding PHB Group RFC2597, June 1999

[EF] V.Jacobson, K. Nichos, K. Poduri,
Expedited Forwarding Per Hop Behavior, RFC2598, June 1999

[MCPROV] C-Y Lee,
Provisioning Resources for Multicast Traffic in a
Differentiated Services Network, Internet Draft October 1999

Authors' Information

Cheng-Yin Lee
Nortel Networks
PO Box 3511, Station C
Ottawa, ON K1Y 4H7, Canada
leecy@nortelnetworks.com

Loa Andersson
Nortel Networks Inc
Kungsgatan 34, PO Box 1788
111 97 Stockholm
Sweden
Phone: +46 8 441 78 34
obile: +46 70 522 78 34
email: loa_andersson@nortelnetworks.com

Ken Carlberg
SAIC
S 1-2-8
1710 Goodridge Drive
McLean, VA. 22102
carlberg@time.saic.com

Bora Akyol
Pluris Terabit Network Systems
10445 Bandley Drive
Cupertino, CA 95014
USA
akyol@pluris.com
Phone: (408) 861-3302
Fax: (408) 863-0271
email: akyol@pluris.com