Policy Tree Multicast Routing: An Extension to Sparse Mode Source Tree Delivery

Horst Hodel

Computer Engineering and Networks Laboratory
Swiss Federal Institute of Technology
CH-8092 Zurich
hodel@ntb.ch

Abstract

Bandwidth-sensitive multicast delivery controlled by routing criteria pertinent to the actual traffic flow is very costly in terms of router state and control overhead and it scales poorly towards larger, wide-area networks. PIM-SM (Protocol-independent Multicast - Sparse Mode) has been introduced as a simple, flexible and scalable concept for Internet-wide multicasting. Yet, PIM's efficiency potential (like that of alternative wide-area multicast concepts) can only be fully exploited if it is based on Reverse Path Forwarding (RPF), a low-cost mechanism, which however does not select prescribed delivery paths if link parameters or routing policies are asymmetric. This paper presents an extension to protocols like PIM-SM, called Policy Tree Multicast Routing (PTMR). This concept leads to multicast delivery trees which, even under asymmetric conditions, readily comply with imposed macroscopic policies and moreover enable support of shortest path and QoS criteria. PTMR gives no consideration to how the policy-sensitive paths have been established; they may be imposed by the network itself but also by providers, recipients and even sources. PTMR amends the PIM-SM mechanism with a macroscopic control layer which marks (pegs) domain border routers on policy-sensitive forward paths. In a given domain, a pegged ingress border router is then joined by local group members as well as by its child pegs. This results in optimally fusing transit paths into local distribution trees. PTMR performance and policy control potential are restricted primarily: by the congestion of control messages at the multicast sources, introduced by source-originating tree construction; by the possible extent of policy-sensitive path aggregation; and by the intra-domain (transit) delivery conditions. PTMR is a single-layer protocol, appending PIM-SM with a policy dedicated delivery mode. Ist primary design target is to forward multicast traffic in accordance with any underlying multicast-relevant routing, including comprehensive policy routing. In comparison, Border Gateway Multicast Protocol (BGMP) is a proposal for an inter-domain MR protocol based on BGP-type routing, which focuses on fusing heterogenous multicast routing domains and which allows ASs to control multicast transit traffic.

1 Introduction

1.1 Multicast Routing Service

Many multicast applications call for network service with low end-to-end delay. Thus, multicast concepts which provide shortest delivery paths play a major role. In high-speed multicasting requiring simultaneous reception and in interactive conferencing, absolute delays and their minimizing are important. Constant delivery delay on the other hand is an issue for example when sending video or voice in real time.

However, in multimedia and image communication applications other criteria than minimum delay are increasingly gaining importance too, culminating in stringent QoS requirements [1][2]. Some even call for service options, e.g. in the inherent trade-off between fast and reliable delivery. On the IP level these requirements are met and guaranteed by excising network control and by resource reservation, backed by efficient routing which finds paths with sufficient resources to meet user requirements. Thus, Multicast Routing (MR) has to take into consideration a wide range of QoS requirements in a model where the network is characterized with multiple metrics such as bandwidth, delay and loss probability. In addition, more and more multicast applications establish Internet-wide associations and thus also require support by various providers.

In the Internet resources are not unlimited and traffic transit offers are increasingly discriminate. This leads to a growing demand for quality-of-service "guarantees" by users and their willingness to pay for those services, together with the need to protect network resources. Despite the Internet community's reluctance to invest in comprehensive policy routing, due to its complexity, it can be assumed that this issue, pioneered by such concepts as Inter-Domain Policy Routing (IDPR) [3] and Source Demand Routing Protocol (SDRP) [4], will have to be re-examined. IDPR is a policy routing protocol based on the node routing paradigm. SDRP is based on header routing and computes forwarding paths for special purposes on a per flow basis. Obviously, any concept devoted to policy routing will also have to encompass multicasting.

In summary, it must be anticipated that MR will be confronted with policy and QoS demands, which may not

only be expressed by network resource management and by providers, but also by recipients and even by sources, which want to individually select not only domain-specific paths but also QoS and delivery cost. Thus, source specific routing as well as QoS route selection, based on header routing and flow labels, and microscopic adherence to delivery prescriptions have to be coped with. In addition, MR based on underlying unicast routing should be able to assimilate this layer's policy routing mechanisms. In fact, since multimedia multicast adds considerably to the traffic heterogeneity on shared network resources, its introduction calls for improved policy control for other types of traffic as well.

Conventional IP multicasting is based on the group host model [5]. Thus, multicast support for varying service requirements can be realized through the use of different multicast groups. In addition, some (information distribution) applications even require source-specific delivery to given multicast groups [1]. The necessary address management and service registration can be provided by a Multicast Group Authority (MGA) [2]. However, the service-per-group model leads to separate multicast packet flows. Support for a single hierarchical flow from which routers (and receivers) extract relevant portions is not expected in the near term. [2].

1.2 Multicast Routing Concepts

When looking at cost, performance of MR concepts has to be judged primarily by their bandwidth consumption, but also by the required traffic overhead, router state dimension and processing expense. Efficient datagram multicasting based on the group host model requires two (usually more or less combined) functions: firstly, a source/group member handshake, i.e. sources and group members finding each other; secondly, aggregating, building and dynamically maintaining a delivery tree which is restricted to them. However, involving the multicast data source in the tree's deployment should be basically avoided, because of the resulting concentration of workload at the source and control traffic in its vicinity.

Different multicast concepts are characterized by their trade-off between the cost of excess data packets and that of exchanging and maintaining information for the source/group member handshake [5]. In the link state approach, the required route calculations are performed in parallel by each router, based on the topology image it is maintaining. Multicast delivery tree deployment based on such a concept is straightforward and moreover it does not require any source support. However, link state routing is not feasible for wide-area networks because of the resulting router table explosion. Augmenting it with multicast capability reduces its scalability even further, since group member location information has to be broadcast network-wide and maintained in all routers. [6].

On the other hand, when using distributed tree calculation (distance vector approach), multicast routers individually have to establish their functional position with respect to the delivery tree. Involving the source side in this task can be avoided by means of a special routing facility (Come-from Routing, CFR), which allows each router to identify the optimal incoming interface for multicast packets. Accordingly, it accepts packets from a given source exclusively on this interface, for itself and for further delivery. CFR assigns a different semantic to a route than conventional unicast forwarding algorithms in which a router chooses how it reaches others but doesn't choose or know how others reach itself. In particular, if a router B accepts a route from router A for source S, then conventionally B will forward packets **to** S, using A as the next hop. However, in the CFR case, the same route means that B will accept packets **from** source S, through the interface on which A is its neighbor. [7].

1.3 Reverse Path Forwarding

MR expense can be considerably reduced by applying Reverse Path Forwarding (RPF), a CFR concept utilized in variant forms in practical multicast protocols [5][8]. It is based on the simple idea that an actual delivery path to a node is the reverse of the path from this node to the source. Accordingly, a router's CFR interface is the one which it would use to send its own traffic to the multicast source. RPF-information can be established using a separate instance of a conventional routing protocol over the multicast topology. Installing a specific CFR protocol can be avoided altogether by having the routers consult their RIB (Routing Information Base) of the prevailing unicast routing fo RPF-information. If nothing else than the RPF information is looked up, multicasting will even stay independent of its type.

Multicast concepts which include their own routing protocol for multicast RIBs, which provide the RPF entries, lend themselves for enforcement of specific multicast policies and for optimal network resource allocation. But if the RPF multicasting is based on an existing unicast routing protocol, policy control of multicast delivery is curtailed because it has to be exercised via unicast policies imposed on the RPF paths. Another problem is that unicast and multicast policies may interfere with each other. (An example, taken from conventional shortest-path routing, is the RPF look-up based on unicast routing metrics which are tailored to locally enforce certain unicast policies or to establish routing priority in DMZs between network service providers.) It is even possible for multicast delivery on a given path to be blocked because the prevailing multicast policy is incompatible with the unicast policy on the reverse path (e.g. if a transit domain prohibits multicast traffic). A frequently practiced but very limited way to enforce separate multicast policy is to embed it in the multicast topology by means of tunnels, some of which may be needed for multicast connectivity anyhow.

Using RPF as approximation for genuine CFR leads to difficulties. Multicast delivery based on RPF information will only follow the paths prescribed by prevailing forward routing information if the environment is symmetric. This can be a serious drawback, because under asymmetric conditions minimum delivery delay cannot be guaranteed and

policy and QoS may not be obeyed. In the extreme case of asymmetric connectivity (due to unidirectional packet filters or extreme policies), forwarding on the chosen RPF path is not possible at all, thus blocking multicast packet delivery. Still, in terms of network resource allocation it can be advantageous if RPF selects a different path for multicast packets than the prevailing (unicast) forward path, e.g. for load spreading or for improved multicast path fusion.

1.4 Multicast Forwarding Algorithms

A multicast forwarding strategy with cost trade-off extremely on the excessive packet side consists in reaching group members by means of a global broadcast on a spanning tree. Efficiency can be considerably improved by subsequently pruning back the delivery tree to last-hop routers actually serving group members. However, in order to catch network changes, the broadcasting has to be periodically repeated. The required spanning tree is established by restricting the broadcast to CFR incoming interfaces. Broadcast-and-prune schemes are not efficient and exhibit poor scaling properties: They consume network resources for data packets traversing paths that do not lead to any recipients. Furthermore, all routers in the (global) network have to keep state for every active source/group pair. Broadcast-and-prune multicasting may be classified as a source-initiated, source-originating tree construction whereby the source/receiver handshake manifests itself in the pruning of non-member branches of the broadcast tree. (Figure 1).

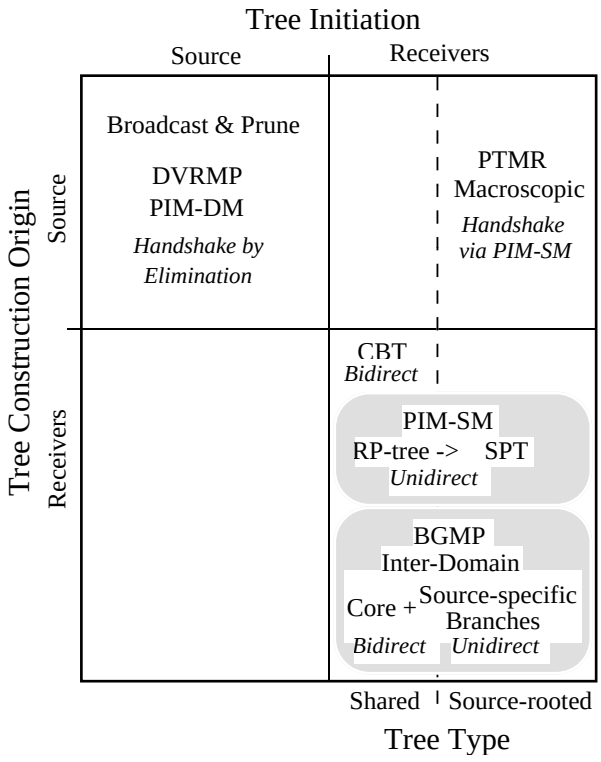

Figure 1: Multicast Tree Classification

Internet-wide multicast and sparse group member populations exacerbate the problems associated with Broadcast-and-prune concepts. Better suited in this case are schemes in which multicast data packets are only forwarded on router interfaces from which explicit Join messages initiated by last-hop routers have been received. This receiver-initiated, receiver-originating tree construction, which limits expansion of multicast transmissions precisely to the set of all recipients, is called Sparse Mode. In order to reduce the extreme source scaling factor inherent in receiver-initiated source/recipient handshake, a single delivery tree for a given group is constructed which is shared by all sources (shared tree architecture). Active group members join, based on the prevailing CFR, towards the delivery tree root. Multicast sources send their packets to the shared tree for delivery. In the so-called rendezvous concept, all sources send their multicast packets to the tree root (rendezvous point RP). The core-based version, on the other hand, is based on a bidirectional delivery tree which is defined by a neighbor data base: routers CFR-select their upstream neighbor towards the tree root (called core), which in turn identifies itself downstream. Bidirectional delivery allows immediate distribution of packets dispatched by tree-attached sources (e.g. source members). Packets from detached sources are IP-encapsulated and sent to the core. The core tree concept helps reduce dataflow and traffic concentration: multicast packets from tree-attached sources to recipients in their locality need not travel all the way up to the tree core and back down again.

While shared trees require relatively low overhead, they have an inherent tendency towards traffic concentration. Thus, when sources send data simultaneously, latencies with considerable variance, late arrivals and packet drops may occur. Furthermore, it is hardly possible to support optimal routes to all recipients. Thus, shared trees do not provide minimum delay paths. [9]. Moreover, they exclude support of source-specific policies. On the other hand, in Distributed Simulation (DIS) [10], where large source member populations prevail, building individual source trees is not acceptable and thus shared trees are the appropriate concept. Futhermore, DIS application's strict requirements in terms of join latency can be better met by core-based concepts.

A large portion of MR expenses is caused by the source/group member handshake. This can be avoided if there is a single source involved, whose identity is known to the group members. In this case (Single-source Multicast [11]) receivers have to specify (S,G) pairs. Last-hop routers then join a unique source-specific delivery tree, utilising existing source joining resources. The single source delivery tree may be identified by an allocated range of group addresses.

1.5 Internet Multicast Routing Protocols

Well-known architectures for Internet MR are DVMRP [12] (routing protocol of the MBone multicast topology), CBT [13], MOSPF [14] and PIM [6].

DVMRP and one of the two alternative modes which PIM provides (PIM Dense Mode [15]) are Broadcast-and-prune

concepts. Current implementations of both DVMRP and PIM are based on RPF.

CBT (Core Based Trees) is a sparse mode concept which builds core-rooted (bidirectional) delivery trees, based on RPF information towards the core. An alternative group-shared tree concept is the PIM-SM (Sparse Mode) architecture which applies the rendezvous strategy [6][16] (See appendix). As a unique feature, it allows last-hop routers which see traffic from a given source to optionally join a separate source-rooted delivery tree (Shortest Path Tree, SPT). This mode not only provides optimal paths to all recipients and supports congestion-free delivery, but also allows the considering of source-specific delivery criteria. Thus, PIM-SM covers optimal delivery for heterogeneous applications.

In PIM-SM, a source's traffic is initially sent encapsulated to the relevant RP (register phase). If the traffic warrants it, and if it is possible, the RP then joins the source and establishes a source-specific multicast delivery branch. (Join latency is thus worse than that for source members in CBT.)

For establishing the RPF interfaces, DVMRP provides its own distance vector protocol which is run over the MBone topology. On the other hand, PIM and CBT obtain the RPF information from the prevailing unicast routing protocol.

MOSPF is the MR protocol based on OSPF [17]. From this protocol each router has a network topology image at its disposal, which for multicast is augmented with group member location information. MOSPF acts as a Broadcast-and-prune concept, but it performs its route calculations "in memory“. The forward link parameters provided by the link state data base allow the routers to calculate the actual forward routes, and thus link symmetry is no issue. In OSPF, maintaining the link state database is restricted to OSPF areas. Area-border routers inject into an area "summary link" information for extra-area destination paths, with a metric reflecting path length. This summarization of extra-area information reflects itself in the MR architecture: Group member populations in an area are summarized and represented by all of its area-border routers. The summary link information is then used to construct an RPF tree rooted in an extra-area multicast source. External routers (i.e. OSPF border routers) are treated as members of all groups. Multicast sources outside the OSPF domain (OSPF area cluster) are likewise approximated by summarized reverse paths, which are advertised within the domain by the external routers. As a consequence, the inter-domain routing protocol must be based on RPF in order to provide multicast packets at those entry points which are propagated within the domain.

Summing up, all current MR protocols, when applied in Internet-wide multicast, rely more or less on RPF information and hence perform inadequately in asymmetric situations.

1.6 Internet-wide Multicasting

In wide-area multicast, frugal bandwidth consumption is essential. Hence, optimal multicast traffic aggregation as well as low control overhead are critical issues. Furthermore, since wide-area networks are typically maintained by various Administrative Domains (ADs), heterogeneous environments have to be coped with. Also, individual AD policies need to be supported. They may be unsymmetrical and even source specific. Thus, specific inter-domain MR protocols which include a rich and well-developed policy model are called for [7].

PIM-SM and CBT achieve wide-area deployment with a single routing layer, i.e. they are essentially intra-domain protocols. More recent Internet-wide MR proposals address scalability as well as interoperability of various MR protocols with two-layer hierarchical MR models. This also allows a less aggressive reshaping of the distribution trees on the macroscopic layer. In two-layer concepts, CFR information for the two strata may be based on different criteria and may for example be derived from intra-domain and inter-domain unicast routing respectively. Within a (transit) domain, hierarchical MR leads to two distinct microscopic MR functions: Transit paths have to be connected and intra-domain delivery below the injection router has to be established, whereby special consideration has to be given to optimal fusion and contiguity of the delivery tree. A safe mechanism for avoiding the delivery of packet duplicates is having single injectors into a domain. However, this approach does not allow for load spreading. In order to adhere to criteria imposed on macroscopic multicast delivery, hierarchical MR frequently resorts to encapsulation across domains. While this approach drastically simplifies the interoperability with various intra-domain MR protocols, it prevents, within a domain, the fusing of transit paths and local traffic and thus may lead to a considerable increase in bandwidth consumption. Such transit tunneling occurs in Hierarchical DVMRP (HDVMRP) [18], a two-layer hierarchical model for the Mbone, employing DVMRP as an inter-region routing.

For wide-area MR to be efficient, any sort of global broadcast must be avoided. Thus, e.g., DVMRP and PIM-DM, which broadcast initial data packets, but also link state protocols, which broadcast membership information (e.g. MOSPF), do not qualify. For single-layer routing as well as for the macroscopic layer in a two-layer hierarchy, this leads to sparse mode concepts (e.g. CBT and PIM-SM). However, sparse mode architectures are based on shared trees, which may entail dependency on resources fielded by a third party domain. [7]. Bidirectional multicast delivery, as in CBT, exhibits the postulated low bandwidth consumption. In addition, it reduces third party dependency. (In wide-area deployment, its low source member join latency is less important). However, due to the bidirectional multicast packet dispersion, this concept fails with unidirectional policies and in other asymmetric situations and moreover cannot support source-specific policies. PIM-SM on the other hand offers optimal, source-specific trees, via the initial rendezvous mode for source/group member handshake. Furthermore, unidirectional shared trees have more aggressive loop prevention and share the same processing rules as source-specific entries which are inherently unidirectional. This for example considerably facilitates the switch-over procedure to a (unidirectional) source-rooted tree.

2 Forward Path Delivery Trees

2.1 Divergent Paths in Reverse Path Multicasting

Applying RPF is the key to lean, efficient and scalable MR. However, in an asymmetric environment divergent path situations may occur, i.e. the path based on RPF information may not coincide with the optimal one according to prevailing requirements for a given type of multicast traffic. If MR maintains a specific multicast RIB, divergent path may be caused by asymmetric conditions within the supporting routing environment. Otherwise, divergent path may originate in controlling multicast-specific delivery by routing information from a routing facility for a different traffic type (usually unicast routing), which may be based on a differing topology or on different criteria. Finally, this routing environment itself may already introduce asymmetries.

On the provider level, access, transit and route selection policies are generally not well concerted. Since they can be unidirectional, this may lead to asymmetries and thus divergent paths. Unpublished polices prevent a coordination altogether. For example, providers with overlapping domains tend to remit traffic as soon as possible to their neighbor in order to minimize their own delivery cost, resulting in an asymmetry when applied by both parties. Installing tunnels or applying source routing can cause asymmetric conditions too. Such measures may be taken for provider selection and resource management, but also in order to overcome policy model limitations of the prevailing inter-domain protocols. BGP [19] for example is based on the node routing paradigm. Hence, a domain which sends traffic to a neighboring domain for further delivery cannot intend this traffic to take a different route than that taken by traffic originating in this neighbor itself. Also, if BGP allows single route announcement only, it is faced with the regional network problem[1].

Asymmetries may also arise when routing boundaries are traversed: between regions with different routing protocols or, in hierarchical routing, between the two separate layers. Inter-domain unicast routing, for example, has a tendency to asymmetric paths because a router usually selects the closest border router for the outgoing traffic, independent of where the incoming traffic entered the domain.

In QoS routing, asymmetries for a specific service quality may be due to the given network resources. Asymmetries may also occur when selecting one of the offered service choices is not handled consistently. Similarly, if the RPF information is derived from unicast routing, the QoS path selected from the offered choice may not be the optimal path for the given multicast delivery requirements.

Even if multicast delivery is embedded in a homogeneous shortest-path routing environment, divergent paths may occur caused by asymmetric link parameters like delay, bandwidth, reliability and load factor. Path divergence may also be introduced if the unicast routing is manipulated by the configuring of differing routing metrics, weight factors or offsets. Furthermore, it may be evoked by load splitting or tie-breaking across multiple equal-length shortest paths. Lastly, asymmetries may be introduced by screening strategies for firewalling.

Generally, an asymmetry leads to multicast delivery on an alternative, unintended path. Even more severe is asymmetric connectivity, meaning that on the RPF path multicast traffic in the forward direction cannot flow at all. This can be brought about by the network topology (e.g. by tunnels or packet filters) and also on the routing side by policy constraints which are in effect unidirectional.

In PIM-SM for example, if RP mode delivery is blocked due to asymmetric connectivity, group members will never receive any traffic from sources using this RP-tree. When switching to an SPT, policy-sensitive delivery may still not be attained. Furthermore, missing forward connectivity will prevent traffic from the source coming down directly and thus the SPT switch-over procedure will not be completed. As a result, multicasting will remain on the RP-tree, even though a more desirable, policy-sensitive delivery might be possible. (Note however, that under adverse asymmetric delay conditions it is possible that staying on the RP-tree would lead for some recipients to a smaller delivery delay or to a better policy compliance!)

Current MR in the Internet is based on a unicast routing environment, applying intra-domain generic shortest-path criteria and on the inter-domain level path vectors. This environment is far from being symmetric. Traceroute experiments investigating routing asymmetries [20] showed that sequences of cities and ASs visited by routes in the two directions of a virtual path differ quite frequently. It was found that, overall, 50% of the paths includes an asymmetry in terms of cities. About two thirds of them were confined to a single hop. In terms of ASs, about 30% path asymmetries ocurred, mostly due to the addition of a single hop in one direction. Consequently, for the usually practised RPF multicasting this means that packets frequently do not use the same path as unicast traffic. However, since the extent of multicast applications in the Internet can still be overviewed, providers configure for the eliminating of inacceptable asymmetries (e.g. in conjunction with tunnels needed to connect multicast islands) and provide ad hoc patches as part of their traffic engineering (which focuses mainly on load spreading). As long as simple multicast connectivity or, at most, shortest path delivery inside ASs is expected, this performance level on the whole is acceptable. (In fact, present MR protocol specifications use RPF synonymously for CFR). However, the growing spread of

[1] Users gaining Internet access through a common (regional) network may be subscribed to different transit providers. These providers will announce to a given long-haul destination D only the paths to their own users. In the other direction, the regional network will have to pick for all their users a single path to D from the ones announced by the different transit networks. Thus an asymmetric situation may result which is usually resolved with tunnels.

multicasting and the new, exacting demands imposed on it increase the extent of asymmetric situations, Internet-wide as well as locally. Moreover, RPF is a weak policy mechanism. Hence, its indiscriminate application does not suffice anymore. Rather, immediate control of multicast traffic is called for, i.e. more elaborate mechanisms have to be applied in order to establish "true" forward path multicast delivery.

2.2 Genuine Come-from Routing

Asymmetry problems and thus divergent paths can basically be avoided by means of genuine Come-from Routing. This entails providing routers with specific multicast RIBs, containing actual CF-paths.

If CFR is based on the distance vector (DV) approach, DVs have to convey the cost of CF routes. Routers generally do not know the CF cost of an incoming link; but their upstream neighbor does, since for him it is the conventional "go-to" cost it associates with the link in question. Thus, if a router adds to all cost entries in a CF distance vector which it dispatches on a link this link's cost, then the receiving router can use the result directly for maintaining its CF-RIB. In the link state approach, the link state database inherently contains link cost in both directions, but the CF paths have to be calculated and cached separately.

BGP-type path vector updates contain the sequence of ASs through which a network is reachable, or, when applying RPF, the AS path on which traffic is received from multicast sources in the advertised networks. Thus, by using separate, multicast-specific AS path updates, genuine CFR is possible, in the sense that ASs can control over which path multicast traffic from a given network is delivered.

The routing expense associated with genuine CFR may be enormous, particularly if distributed path synthesis (DV approach) has to be applied and when multiple metrics and QoS are involved. And, even if in Internet-wide multicast genuine CFR is consequently applied, routing problems at interfaces between heterogeneous domains and between hierarchical layers may persist.

Bidirectional delivery trees (e.g. in CBT) are built with CFR information towards the core. But since, given an asymmetry, multicast traffic flow can only comply in one direction with imposed criteria, it is basically irrelevant whether true CFR or (multicast-specific) RPF is applied. However, true CFR favors downward traffic flow, which generally tends to be heavier and furthermore is the delivery direction of multicast packets from detached sources.

Roughly, for the installing of genuine CFR, many unicast routing components more or less have to be duplicated. Still, there is saving potential. Since CF cost can be treated as an additional set of metrics (applying to multicast traffic), CFR may be incorporated in a conventional routing protocol which is designed to accommodate alternative metrics (e.g. routing protocols that support multiple TOS routes).

2.3 Source-originating Delivery Tree Deployment

In a distributed routing environment, multicast delivery trees are defined by (S,G) router states. In source-originating delivery tree deployment these states are set up by establishing individual paths to last-hop routers, according to imposed (forward) criteria for multicast traffic. They are then fused as much as possible. A source marks the prevailing multicast path to an individual last-hop router by sending, hop-by-hop, some sort of pilot packet towards it which is tagged with the given multicast group (and with the appropriate flow label). As necessary, new router states are installed or new outgoing interfaces added along the path. The fact that this process does not consider how the discovered path has been established and what route selection criteria were involved makes source-originating tree deployment independent of the underlying policy model as well as of the prevailing routing procedure. Particularly, enforcing source demand policies using header routing can be supported, whereby the marked routers provide a (loose) path set-up.

In order to dispatch the pilot packets, sources have to know the group member locations. However, for efficiency in a sparse environment, source/group member handshake needs to be initiated by the receivers. Accordingly, group members have to notify a source's first-hop router that they want to be included in the packet delivery. As a result, a first-hop router responsible for delivery tree deployment may be faced with a concentration of such requests, which may add up to an implosion when a source begins to be active. In addition, first-hop routers have to dispatch pilot packets at regular intervals in order to catch topology and routing changes. Besides, receiver initiation presupposes a mechanism by which group members learn about active sources, e.g. via a shared tree as in PIM-SM.

Besides imposing a heavy load on the first-hop router of multicast sources, source-originating delivery tree deployment requires large amounts of processing overhead and control traffic. Since it scales very poorly, this approach was, in connectionless networks, not considered to be practical. It may be assumed however, that in Internet-wide multicasting, by applying receiver-initiated, source-originating tree deployment to a macroscopic level only, complexity can be kept at a feasible size. A single router will then request the source's first-hop router to deploy a cluster-specific delivery path. Besides, since macroscopic routing may be less aggressive, a lower repetition rate for the periodically dispatched pilot packets is possible. Although control message congestion at the sources is now restricted, it still remains a pivotal issue. Macroscopic source-originating tree deployment is the basis of the presented "Policy Tree Multicast Routing".

3 Inter-Domain Multicast Routing

Inter-domain Multicast Routing provides scalability with regard to router memory and processing resources and it facilitates interoperability of MR protocols. Driven by the experience that Internet service providers are reluctant to deploy native multicast without being able to control multicast transit traffic, current IETF efforts focus on inter-domain MR which provides an AS with the flexibility and autonomy to control the receive path of multicast data traffic.

3.1 Framework

Multicast Domains
A Multicast Domain (MD) is a contiguous set of multicast routers defined by a (small) number of Multicast Border Routers (MBRs). No requirements are placed on the internal operation of an MD; any typical MR protocol may be employed (M-IGP). Specifically, an MD may exclusively connect immediately neighboring MBRs and thus does not need to have an MR protocol running (e.g. a common subnetwork, tunnels or an exchange).

Interoperability
MBRs are configured to forward packets between two or more independent MDs. The component-based model [21] provides a set of invariant rules, reducing the dimension of the interoperability problem. Accordingly, an MBR consists of routing components, one for every domain, each associated with a particular MR protocol. Each component may own more than one associated interface which runs the component's MR protocol. All components share a common multicast forwarding cache; for a consistent view they must be able to communicate with each other. Only the component owning an interface may change information about that interface in the forwarding cache. Additionally, each component typically keeps a separate RIB with any type of relevant entry. If a data packet is received on a given interface, the owning component decides whether it is to be accepted or dropped. But once accepted, it is processed according to the forwarding rules of all components. Interactions between components are described in terms of Alerts.

MDs may be organized into an arbitrary topology, with each pair of adjacent domains connected by one or more MBRs. For building a spanning MD delivery tree, an inter-domain routing protocol is required which calculates paths among domains. External routes need not be known inside domains. However, for a given traffic, single domain injectors have to be secured, and also joined.

For globally-scoped groups, domains have to provide their components with aggregate membership information, concerning themselves as well as other, internally reached domains. In cases where the M-IGP does not provide a mechanism for joining and pruning entire groups (as with DVMRP and PIM-DM), availability of an internal Domain-Wide Reporting (DWR) mechanism [22] is assumed.

In non-routing MDs, joining and pruning entire groups as well as individual sources within groups may be achieved by the Internet Group Management Protocol (IGMP) [23]. CBT, MOSPF and PIM-SM allow Joins and Prunes, but only the latter towards individual sources. Broadcast-and-prune protocols inherently are wildcard receivers for internally reached sources. But for externally reached sources, explicit membership has to be explicitly provided, by means of DWR. A simple heuristic to approximate DWRs is for MBRs to assume that when they are reached by traffic from internal sources at least one of them is also a group member. [21].

CFR Inter-domain Multicast Routing
The building of inter-domain multicast delivery trees can be based on CFR, with the routing information derived from an appropriate underlying inter-domain unicast routing protocol (EGP). This provides a mechanism for selecting and enabling MBRs for the inter-domain spanning tree forwarding of multicast packets, including single injection points. Multicast packets are then forwarded natively to internal group members, and/or egress MBRs, via intra-domain MR.

An EGP must be able to support a multicast RIB, i.e. a routing table used to calculate next hop MBRs towards potential multicast sources or roots of shared trees. An obvious EGP choice for supporting the demanded AS multicast policy framework is the Border Gateway Protocol (BGP) [19], because it already has much of the required functionality for building inter-domain multicast trees and beyond that it is widely deployed. BGP is path-vector based and allows ASs to impose access and transit policies. Path-vector based concepts with multicast-specific path updates (for so-called Come-from Paths) provide for genuine CFR, allowing ASs to express multicast specific policy independent of unicast policy. The recent multiprotocol extension [24] of BGP-4 provides a path attribute which allows to indicate whether path information is multicast specific or if unicast/multicast topologies and policies are congruent.

3.2 Border Gateway Multicast Protocol (BGMP)

Protocol Overview
BGMP [25] is a current proposal for inter-domain MR. It builds core-based bidirectional shared delivery trees among MDs. Each global multicast group has to be associated with a single root, i.e. with a root domain (G-root). Root domains are associated with unique ranges of IP multicast addresses (e.g. by means of MASC [26]). Selecting root domains is subject to administrative and performance input, i.e. considering ownership and avoiding poor locality.

BGMP offers optional source tree delivery (e.g. for improved M-IGP interoperation), but only in cases without overlap with the shared tree. This design choice avoids ambiguity and a complex tree switch-over procedure from a bidirectional shared tree to a source-rooted tree. BGMP trees are, in a sense, a hybrid between CBT and PIM-SM trees. Thus, path length and traffic concentration properties of shared tree delivery prevail. With the less rich connectivity of inter-domain MR, this is considered to be acceptable.

Each MBR has zero or more components associated with MDs running an M-IGP. In addition, each MBR with links

that do not fall inside an MD running an M-IGP, will have an inter-domain component that runs BGMP.

Typically, in BGMP MDs and ASs are aligned and the prevailing (appropriate) version of BGP is employed as the EGP. BGMP must be able to forward data and control packets to the next hop towards either a unicast source or a G-root. Thus, a fundamental requirement for any EGP is that it must be able to carry multicast prefixes. Again, the BGP multi-protocol extension satisfies this requirement. MBRs have peers inside an MD, with which they establish multicast connectivity by means of a prevailing M-IGP. Such internal peers are located within the same AS. MBRs speak "internal" BGP. In addition, MBRs have external peers in another AS, which are directly linked (across a non-routing MD) and to which an "external" BGP session is open. Routing MDs which are located in non-adjacent ASs may be connected with multicast tunnels. (Figure 2).

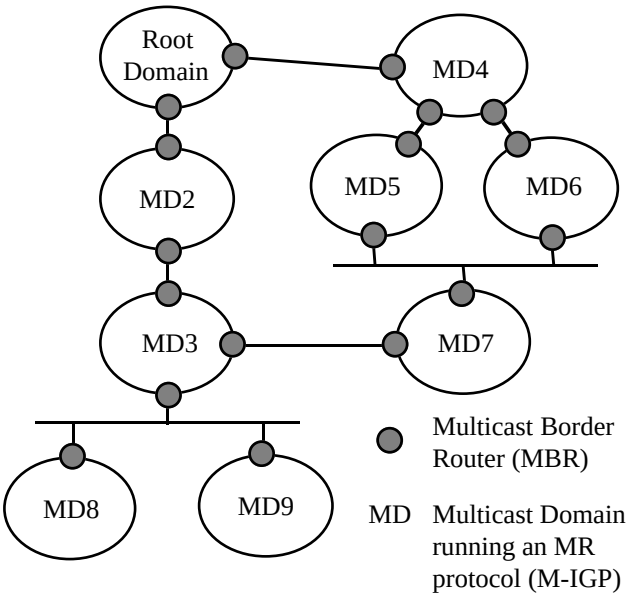

Figure 2: BGMP Inter-domain Topology

BGMP control is exercised by Join/prune messages, which are handled and interpreted only by other BGMP components. Joins and Prunes are sent over specific TCP connections between BGMP peers, as incremental updates. The TCP keep-alives serve as an explicit state refresh mechanism.

BGMP components maintain inter-domain tree state in response to: messages from EGP peers and notifications from M-IGP components on the same MBR (collectively called targets). The tree state table consists of (S-prefix, G-prefix) (including (*, G-prefix)) entries with a list of targets which have been explicitly joined.

BGMP generates inter-domain Joins in response to group membership inside a domain, learned by M-IGP, e.g. by means of DWRs. Then the next-hop MBR towards the G-root generates a (*,G) Join message, which is then forwarded MBR-by-MBR towards the G-root. Source-specific state triggers a source-specific Join.

When a given BGMP component receives a (*,G) Join Alert from another component, or a BGMP (*,G) Join message from an internal or external peer, it sends a BGMP Join message to the next-hop peer (towards G-root or S). If the next-hop MBR is an internal peer, i.e. the domain's injector, then at the given MBR a (*,G) or (S,G) Join Alert is sent to the M-IGP component. This Alert is handled according to native rules. It will transit the multicast traffic to the requesting MBR, fused with possible local delivery.

When a packet arrives at an MBR (internally or externally) and a matching entry is found in the BGMP tree state table, and furthermore the packet was received from the CFR target for S or for G respectively, then it is forwarded to all other targets in the target list, as required for bidirectional forwarding. If a target is an M-IGP component then the forwarding is subject to the rules of that M-IGP protocol. Data packets from sources in non-member domains may reach BGMP routers without matching tree state. These packets are sent natively to the next-hop EGP peer, according to the multicast RIB, until the inter-domain tree is reached. (Multicast sources must be located in a domain running BGMP at the borders!)

Finally, an MBR has to be able to forward external data for a matching tree state table entry to all members within the domain. This requirement has specific implications on the various M-IGPs. In Broadcast-and-prune protocols (like DVMRP, PIM-DM and MOSPF), i.e. those which build source-specific trees, the MBR receiving external packets is responsible for tunneling them to any MBR that is expected to inject them into the domain. In PIM-SM, if the next-hop MBR towards G-root is selected as RP, then RP-Joins/ Prunes transport membership information (for this as well as for internally reached MDs) to the appropriate router. Internally- as well as externally-sourced packets have to be brought to the RP for distribution. If the RP is the next-hop MBR, the fewest Registerings from other MBRs will have to be done (assuming that this MBR will receive the most packets from external sources).

Building source specific tree branches allows multi-homed domains to select as injector the MBR which is compatible with the M-IGP's source-specific CFR check. Thus, encapsulation from the injector to the "right" MBR can be avoided. Similarly, if a PIM-SM interior router sends (S,G) Joins, BGMP can initiate a source-specific tree branch. Thus, if not prevented by shared tree overlap, source-specific branches will allow to short-cut long paths on the bi-directional core based tree and thus to import packets of high data rate sources with lower overhead.

In CBT and MOSPF source-specific Joins cannot be supported. They can for example be avoided by not having the domain advertise source-specific multicast paths or by letting it "convert" (S,G) Joins into (*,G) Joins. In the latter case however, all data for group G will be pulled down to this domain, leading to bandwidth waste. Making CBT suited for BGMP MR domains is in progress. [27].

Comment
Satisfying the need for AS policy compliance and for multicasting across hetrogeneous MDs shifts architectural design priority away from group state minimization and delivery quality. BGMP obtains efficient policy support for ASs by aligning MDs with them and thus through the resources already provided by BGP. In order to provide genuine CFR on the AS level, a specific multicast RIB and multicast-specific path updates are necessary. In the worst case, i.e. if unicast/multicast topology and policies are incongruent, this almost doubles expenses in terms of router processing, memory space and control traffic overhead. Still, policy control potential is restricted to the path selection procedure and policy constraint support of BGP's underlying node routing paradigm and path vector concept. This implies for example that network-specific policies cannot be supported. Furthermore, considering CFR interfaces only is a policy mechanism which may not be discriminatory enough, even for AS policy. For example, traffic barriers imposed by AS policies may be by-passed if a source is covered by a prefix which is homed to more than one domain [7].

Due to the bidirectional multicast packet dispersion, BGMP is deficient in asymmetrical environments. Moreover, it cannot support source-specific delivery criteria. It is indeed possible for BGMP to comply with requests for source-specific delivery, but for the sake of reduced protocol complexity, only in restricted cases.

If local and inter-domain traffic is to be fused, providing Internet-wide multicast by interconnecting heterogeneous MR domains leads to considerable interoperability problems. They may require encapsulations and even protocol modifications. Some MR protocols, while having favorable properties in regional applications (or as a backbone), are less suited to supporting multicast transit traffic. All in all, some of the problems encountered in reaching the expected properties of next-generation inter-domain multicast routing require longer term engineering solutions.

It is argued [28] that instead of introducing a specific inter-domain protocol it would be better to create a unique multicast protocol, or to adapt an existing one, which could be run inside domains as well as across the Internet. Such a single-layer protocol obviously would have to be shared-tree based.

4 PTMR and its Architecture

4.1 Design rationale

The PTMR (Policy Tree Multicast Routing) model aims at an efficient solution for attaining policy-sensitive data packet delivery in Internet-wide multicast, across various domains, even under asymmetric conditions. PTMR's characteristic feature is the forwarding of multicast packets in accordance with any underlying multicast-relevant routing, including policy routing (supporting source-specific policies as well as shortest-path and QoS criteria).

The targeted source-specific policy control demands source-originating construction of delivery trees as well as unidirectional delivery on them. In wide-area multicasting, considering tree switch-over complexity, the source/receiver handshake has to be established via an unidirectional shared tree. This leaves, among the known concepts, PIM-SM as the only acceptable choice. Furthermore, source-originating tree construction is only feasible on macroscopic level, i.e. among MDs. Thus, if the prevailing policy control exceeds simple transit and route selection policies, MDs have to par-take in policy-dedicated delivery too, i.e. an M-IGP has to be applied which generates actual multicast-specific forward paths.

True forward path delivery is inherently provided by an M-IGP based on the link-state paradigm. But this concept's poor scalability restricts domain size considerably (comparable to OSPF areas). Furthermore, relying on link-state routing would mean relinquishing the notion of protocol-independent multicast. But above all, embedding a link-state M-IGP in a PIM-SM inter-domain concept leads to signifcant interoperability problems, especially with the supporting of source-specific Joins and Prunes. Unconditional tunneling of transit traffic on the other hand is highly undesirable because of its high bandwidth consumption.

This suggests the use of PIM-SM also as the M-IGP. For microscopic policy compliance this implies that domains have to be able to provide genuine CFR-joining, or that source-initiating tree construction would have to take place inside MDs as well, which would lead to additional interoperability problems.

PTMR in fact attacks the given problems with a single-layer MR architecture. On the macroscopic level it applies receiver-initiated, source-originating tree construction. Microscopic connectivity is based on CFR-joining. PTMR builds up on PIM-SM, a wide-area MR protocol which efficiently provides receivers with source location information. Since PIM-SM and PTMR are both being based on the join approach, this not only eases co-operation between the two concepts but also allows for synergy.

In a policy driven environment, PIM-SM based single-layer (as well as inter-domain) MR requires special administrative care for selecting a promiscuously accessable RP.

4.2 Definition of Terms (Glossary)

PIM Rendezvous Point (RP)
Routers that rendezvous receivers and senders for a group. Location of RPs is configured and their identity propagated. (*,G)-specific.

PIM RP-tree
Shared tree connecting group members to an RP. Recipients receive all packets from sources S for group G on this tree, except if their last-hop router has switched to the SPT. (S,G)-specific.

PIM SPT
PIM multicast tree for group members of group G, established by last-hop routers joining towards source S. (S,G)-specific.

PIM (S,G) router state
SPT router state for multicast packets from source S to group members G. Incoming interface is the CFR interface for S.

PIM Join and Prune messages
Messages which create or eliminate router state in order to create a PIM delivery tree, or modify it due to state, topology or membership changes. Joins/Prunes either apply to the RP-tree ((*,G)-specific) or the SPT ((S,G)-specific). They are periodically and event-driven sent hop-by-hop towards an RP or a source respectively.

PIM Designated Router (DR)
Elected router on a multi-access network (highest IP-address). DR's are for example responsible for sending PIM Join messages towards an RP or a source and for sending a source's initial multicast packets encapsulated to an RP.

Multicast Domain (MD)
A contiguous, convex set of multicast routers defined by a (small) number of Multicast Border Routers. PTMR MDs function as policy domains.

Multicast Border Router (MBR)
MBRs are responsible for interconnecting MDs and forwarding traffic between them. PTMR MBRs must be able to recognize their domain peers.

Last-hop Multicast Domain
A Last-hop MD is populated by one or several multicast group member clusters. Multicast sources forward packets to Last-hop MDs on a path which satisfies imposed path selection criteria. (S,G)-specific.

Policy Route
Macroscopic path from a source to a Last-hop MD, defined by a sequence of Peg-MBRs. It pegs out the path which the source has discovered for its multicast traffic according to prevailing policies. (S,G,Last-hop MD)-specific.

Peg Router
The MBRs through which a Mark Message enters transit MDs and the Last-hop MD are called Peg Routers. They peg out the Policy Route to the Last-hop MD. (S,G)-specific.

Last-hop Peg
The MBR through which a Mark message enters the Last-hop MD is designated as Last-hop Peg. Last-hop Pegs issue the Request messages. They also cache the First-hop Peg address. Modifications to the Policy Route may lead to a different Last-hop Peg. Initial Last-hop Peg for an MD is the MBR located on the PIM SPT. (S,G)-specific.

First-hop Peg
A PTMR router on the source's local network is established as First-hop Peg. Upon receiving a Request message, a First-hop Pegs issues a Mark message. (G)-specific.

Policy Tree (Inter-domain Tree)
Macroscopic multicast delivery tree formed by aggregated Policy Routes. Its nodes are Peg Routers, its leaves Last-hop Pegs. (S,G)-specific.

Peg Tree (Intra-domain Tree)
PTMR-tree segment with a Peg Router as root and a cluster of group members and/or Child Pegs as leaves.
(S,G)-specific.

PTMR-tree
Microscopically completed Policy Tree in form of concatenated Peg Trees. (S,G)-specific.

(S,G,Peg) router state
PTMR-tree router state for multicast packets from source S to group members G. Incoming interface is the CFR interface for the stated Peg Router, i.e. the root of the Peg Tree. For Peg Routers themselves this is the Parent-Peg.

Last-hop router
Routers which directly connect members of a group. Last-hop routers need to know the RP address and (in PTMR mode) the address of their current Last-hop Peg for every group. G-specific.

Request message
Sent from a Last-hop MD to the First-hop Peg. Request messages are sent event-driven by the initial Last-hop Peg and then periodically by the current one. Payload {S,G}.

Initially the Last-hop Peg has no cached First-hop Peg address. In this case it multicasts the Request message hop-by-hop upstream to the ALL-PTMR-ROUTERS group.
Payload {S,G, Targeted Neighbor}.

Mark message
Relayed by a source's First-hop Peg back to the requesting Last-hop Peg, in response to a Request message. Mark messages are routed hop-by-hop to the PTMR-neighbor on the prevailing policy-sensitive path for multicast data packets. They are addressed to the ALL-PTMR-ROUTERS group. Every contacted router forwards the Mark message on the appropriate interface towards the requesting Last-hop Peg. Contacted MBRs become Peg Routers. If they are not already on the required Policy Tree, they schedule themselves for joining it (with the subsequent Peg-join). The first Peg from the source to do this sets an Alter-flag in the Mark message. It will also be responsible for terminating the subsequent joining process. Payload {S, G, Targeted Neighbor, Last-hop Peg, Parent-Peg, Alter-flag}.

Announce message
Announce messages inform last-hop routers about a newly established Last-hop Peg, or confirm an established one. They are distributed by the Last-hop Peg at the end of an invoked Marking cycle. The Alter-flag status in the received Mark message is copied into the Announce message. If it is set, last-hop routers initialize a Peg-join. Announce messages are forwarded hop-by-hop on the local Peg Tree. They are addressed to the ALL-PTMR-ROUTERS group. Payload {S,G,Last-hop Peg, Alter-flag}.

Peg-join
Message for creating and modifying Peg Trees. Peg-joins either install a new (S,G,Peg) router state or change the Peg entry of an existing one. Peg-joins are sent hop-by-hop: by local routers towards their Last-hop Peg and by Peg Routers towards their parent. Payload {S,G,Parent-Peg}.

4.3 PTMR Concept

Policy Tree. The PTMR architecture is characterized by a structure called Policy Tree, which is the product of macroscopic receiver-initiated, source-originating tree construction. A policy tree is a multicast extension of the Policy Route, as first described in [29]: A macroscopic path from source to destination given by a sequence of domains which satisfies the policy requirements of the source and the involved domains and supports the requested service quality.

Macroscopic source-originating tree deployment guarantees compliance with prevailing macroscopic policies, independent of by whom and how they are imposed. If a domain wants to offer specific transit service quality, its internal CFR must be able to control the multicast traffic flow accordingly. This generally precludes the use of RPF.

Group members receive source location information by way of the PIM RP-tree mode. Thus it is required that all group members are able to RPF-join the RP. The subsequent PIM-SM phase, if the source can be reached, provides source-specific multicast delivery until PTMR is operational. This allows spreading the process of marking individual multicast paths in order to reduce source congestion, the pivotal problem in source-originating tree construction. Since for reaching the policy-dedicated mode PIM SPT delivery is a transient phase only, its performance is less critical and thus it may be based for example on RPF.

The macroscopic PTMR topology is established by splitting up the Internet into Multicast Domains (MDs), surrounded by Multicast Border Routers (MBRs). MBRs must be able to recognize a peer MBR belonging to the same MD. MDs must be contiguous and convex, i.e. CFR must be secured and contained.

Marking Cycles. A Last-hop MD contains members of a given multicast group, which are summarized. On their behalf, an MBR induces, with a Request message, an active multicast source's First-hop router to establish a Policy Route to the requesting Last-hop MD, by discovering the policy-sensitive multicast packet path back to it. To this purpose, the DR returns a pilot message (Mark message) to the requesting MBR, which mimics multicast packets. The Policy Route is pegged out by the sequence of MBRs (Peg Routers) through which the Mark message enters transit MDs, ending with the Peg to the Last-hop MD (called Last-hop Peg). When the requesting MBR receives the requested Mark message, it multicasts an Announce message containing the address of the (new) Last-hop Peg to the local last-hop routers. Finally, with this information, the last-hop routers join towards their Last-hop Peg. (Figure 3).

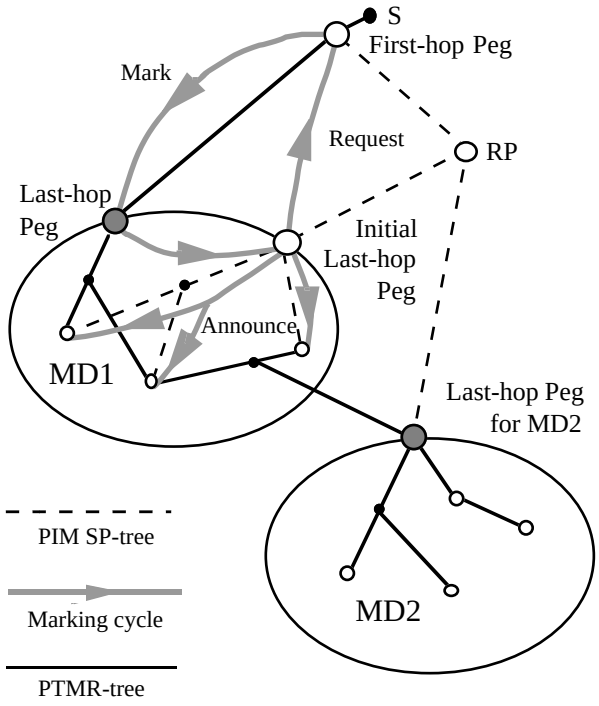

Figure 3: Construction of a PTMR-tree
(In MD 2 the Last-hop Peg and initial Last-hop Peg coincide!)

PTMR-tree. Aggregating (S,G) Policy Routes creates an (S,G) Policy Tree. This framework is microscopically completed into the actual multicast delivery tree (PTMR-tree). Last-hop routers in an MD join an intra-domain tree rooted in their Last-hop Peg (Peg Tree). Since this root is also the Parent-Peg for downstream Pegs (egress MBRs for the given MD!) on the same Policy Tree branch, Child Pegs join the Peg Tree as well. Thus, local distribution and transit paths are optimally fused. In other words: Microscopic multicast delivery is established by a concatenation of Peg Trees, each one rooted in a (Parent/Last-hop) Peg and with local group members and/or Child Pegs as its leaves. (Figure 4).

In generic PIM SPT delivery, all members of group G are joined directly towards the source S, i.e. all routers maintain (S,G) state with the CFR interface for S as the incoming interface. In order to impose policy on the delivery tree, PTMR extends this concept to a more general model. All router states are expressed in terms of (S,G,Peg). Local routers keep (S,G, Last-hop Peg) state, i.e. their incoming interface for (S,G) traffic is the CFR interface for their Last-hop Peg. Peg Routers, on the other hand, are connected to the Peg Tree of the upstream MD. Accordingly, their router state is (S,G,Parent-Peg), the first Peg being the source's First-hop Peg. Reversely expressed: A generic SPT is just a degenerated PTMR-tree, with all Pegs concentrated into the source, or: with all router states expressed by (S,G,Peg=S). (S,G,Peg) router states are installed and modified by a specific Peg-joining process: last-hop routers join their

Last-hop Peg, and from there on in succession each Peg joins its Parent-Peg.

Tree Maintenance. Policy on a PTMR-tree is imposed and maintained by the selection of Pegs only. Thus, the Policy Tree and the connectivity between its nodes can be maintained in two separate layers. Microscopically the tree is maintained like the PIM SPT, by periodic and event-driven Joins and Prunes based on state, topology and on group member information. On the Policy Tree level, the prevailing Policy Routes are periodically captured by Marking cycles and subsequently the Peg entries in the router states set accordingly by means of special Peg-joins. Policy changes are much less dynamic than connectivity changes and packet delivery is always secured on the already existing Policy Tree. Thus, periodic updating of the Policy Tree (by means of Marking cycles) may be exercised frugally, reducing the control message overhead and source congestion.

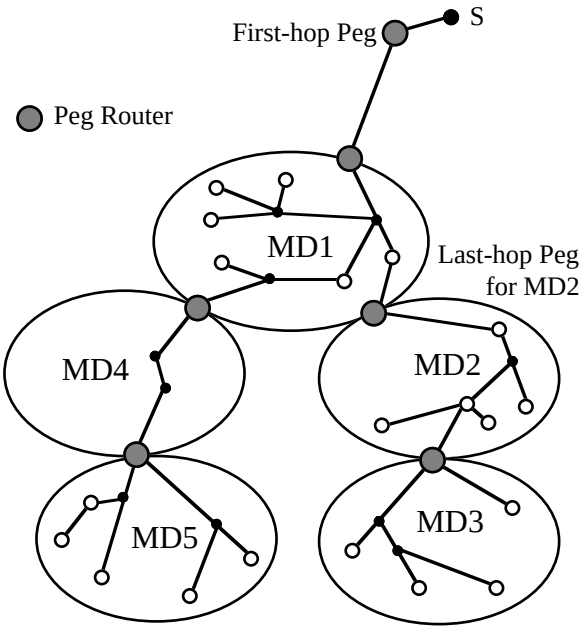

Figure 4: PTMR-tree

Initial PIM-SM Phase. In order to resolve the source congestion problem without considerably infringing on performance, the PTMR concept is based on the assumption that in Internet-wide MR symmetric conditions are prevalent. Thus, the PTMR concept gives first priority to efficient source-specific multicast packet delivery for this situation. On the other hand, asymmetric path environments are treated as an exception: PTMR establishes a priori a generic PIM SPT and devotes itself to asymmetries only afterwards, i.e. readiness of the PTMR-tree is delayed. If the existing delivery tree does not comply with prevailing policies, PTMR will modify it. For completely symmetric conditions, the PTMR-tree is isomorphic to the SPT, independent of the applied intra-domain CFR. Thus, if the SPT itself already adheres to the prevailing Policy Tree, PTMR will just add the Peg address to the routers' states, without further effect. Hence the initial PIM SPT strategy allows for time to reduce the request concentration at the source without performance loss in the most frequent case, i.e. policy-sensitive multicast under symmetric conditions.

PTMR steps in when last-hop routers on a regular PIM RP-tree have learned about an active source and decide to join it directly. The Join process follows the PIM procedure, but with an amendment: In order for MBRs to discover that they belong to a Last-hop MD, the Join procedure keeps track of tree branches ending within the domain (as opposed to those for transit). Once a Last-hop MBR on the SPT is established, it will request Policy Route deployment. If on the given SPT there is more than one MBR serving domain-local group members, then the number of individual Marking cycles increases accordingly. They may lead to deviating Policy Routes and thus to more than one Last-hop Peg. While this leads to additional delivery expense, the resulting load spreading may be desirable.

PTMR utilises PIM RP-mode for establishing the source/receiver mapping. However, if it is a priori known, as assumed in Single-source Multicasting [11], a unique single- source PTMR tree could be established, starting directly with the PTMR procedure.

5 PTMRT Performance Issues

5.1 Source Congestion and Delayed PTMR Readiness

In order to avoid a control message congestion at a source which becomes active, initial PTMR-tree construction is time-spread: Marking cycles triggered by an SPT switch-over are delayed by the initiating Last-hop Peg which holds back the Mark message for a random delay time within a given interval. PTMR-tree deployment starts when a last-hop router decides to move to an SPT, manifesting the need for an improved delivery service. In the symmetric case, policy-sensitive multicast packet delivery is attained with the completion of the relevant SPT branch. But in the presence of divergent paths, packet delivery has to be accepted temporarily on a path which may not comply with the prevailing policy requirements until the PTMR-tree branch is established[2].

[2] A temporary intermediate delivery mode could be basically avoided by having the protocol move to the PTMR-tree directly from the RP-tree. But policy-sensitive delivery under symmetric conditions would then also be affected by the request spreading. Another reason against this strategy is, that in a symmetric environment a PTMR-tree initiated by an MBR on the RP-path is generally not isomorphic to the SPT. This may result in non-optimal delivery paths.

If multicast traffic cannot flow on a prospective PIM source-specific path, due to unidirectional connectivity caused by policy or security measures, switching to the SPT is not completed and delivery remains on the RP-tree. But since PTMR postulates intact intra-domain connectivity, the SPT-join message will still reach and designate the Last-hop Peg, whereupon the PTMR-tree will be created or joined respectively.

Temporary delivery on the PIM SPT requires installing state but may in certain cases not bring better performance than staying on the RP-tree. However, this mode can be omitted by purposely interrupting its completion: a Last-hop MD can let the initial Last-hop Peg disrupt the SPT joining, and transit MDs can block SPT delivery by means of appropriate packet filters.

5.2 Policy Route Aggregation

On a PTMR-tree, multicast packets are distributed solely based on their group (destination) and source addresses. This allows, on a link, the aggregating of (S,G)-specific packets destined for group members in different Last-hop MDs. However, this concept only accommodates policies whose routing criteria for multicast packet distribution from a source to a given group can be mapped onto a single spanning tree.

(S,G)-specific Policy Routes may merge again at any Peg Router, leading to parallel paths which span one or several MDs, and thus to duplicate packet delivery. Occurrence of such parallel paths is manifested by unrelated Mark messages requesting a Peg Router to accept more than one Parent-Peg for the same (S,G) pair. Besides in such a contest situation, parallel paths also have to be dealt with in the course of modifying an existing Policy Route. PTMR resorts to a strategy which, in the contest case, amounts to an arbitration: A new span is favored over an existing one, if it has been in use for at least a given amount of time (state hold-down). In any case, the Marking cycle is carried through to the end. When subsequently completing the valid Policy Route, a defeated span will be pruned. In the contesting path case, a taken decision will persist at least until the next Mark message for the defeated Policy Route arrives, even if the winner in the meantime has lost group presence and therefore its Policy Route needs no support anymore[3].

Policies which lead to parallel Policy Routes should be avoided because they can cause insistence problems. If this is not possible on the administrative level, topological adjustments or different group address assignments for deviating policies may be made.

[3] In order to avoid this latency, Pegs at the merging point would have to cache the identity of all competing Last-hop MDs. Prune messages would need a Last-hop MD identifier and would have to be sent upstream beyond the point of group membership intervention, up to the merging node of the parallel paths.

PTMR gives no consideration to how the discovered Policy Routes have been established and what route selection criteria are involved. Thus, if the network is furnished with a flow control concept, policy criteria with wider scope and higher granularity could be accommodated, Furthermore, load spreading to a given multicast group would be possible. Yet, the resulting larger number of policy trees would entail increasing complexity and overhead. Nevertheless, by using Last-hop Peg addresses as flow labels for example, the problem with Policy Routes merging again could be avoided.

5.3 Multicast Domains and Policy Compliance

PTMR MDs need not be aligned with ADs, routing domains and inter-domain (unicast) routing and its architecture allows arbitrary granularity of MDs. These actually function as policy domains. Existing ADs are an obvious choice. The border routers of the latter can then take on the required additional functionality.

In synergy with the PIM-SM substructure, completing a Policy Tree to a PTMR-tree is achieved by means of Peg Trees established with CFR-joining. Thus, microscopic coherence of transit paths is achieved by fusing them into local delivery trees. Due to this, transit service is equal to that for local delivery.

For PTMR-tree delivery to guarantee compliance with a given policy set, each transit domain has to complete its Policy Tree span according to the service it offered to the policy control entity. For non-specific transit, a domain may provide "best possible" service according to its own criteria. However, if a domain assumes responsibility for transit according to path-specific criteria, multicast packet delivery within a domain has to follow the microscopic path which was taken by the Mark message and led to the given Policy Route. Thus, its CFR must not produce divergent paths (disregarding cases where a divergent path offers equal service!). Local policy may decide to tunnel transit multicast data according to the required criteria, at the expense of bandwidth.

Compliance with prescribed intra-domain distribution may be secured by genuine CFR, but also by RPF if asymmetric situations can be avoided. Otherwise, reverting to RPF in individual domains decreases the possibility for the forwarding path to satisfy the conditions which led to the selection of a particular Policy Route. But PTMR operability still remains secured and, in the case of an additive metric, smooth degrading takes place. This allows for a flexible trade-off between (local) CFR cost and (global) PTMR performance.

The leanest PTMR solution results when all involved domains apply the RPF approximation. This strategy readily supports macroscopic inter-domain policies such as access, transit and route selection policies, even if asymmetric situations exist. On the other hand, guaranteeing full performance in terms of exacting policies

as well as shortest-path and QoS delivery requires applying genuine CFR in all domains, with the entailing cost.

An alternative approach to applying genuine CFR in order to overcome the deficient behavior of RPF in an asymmetric environment consists in controlling asymmetric situations. Since ADs are under a single administration, there exist various possibilities for avoiding asymmetries by means of administrative measures, topology modifications and traffic engineering. However, suppressing asymmetry may impede load sharing strategies.

On the other hand, the PTMR architecture permits arbitrary MD granularity. This allows confining MDs to regions in which symmetry can be guaranteed. For example, an AD could be subdivided into smaller MDs, particularly if it contains more than one routing region. The closer two consecutive Pegs in the Policy Route are, the shorter the RPF span will be. Thus, even if a diverging path results, it has, given an additive metric, less effect on delivery performance. Moreover, there is an increased possibility for the RPF path to satisfy the same route selection criteria as the discovered (policy-sensitive) path. In certain situations it might even be advantageous to strategically define MDs without expected group member population.

A multicast source which wants to enforce source demand routes, by means of header routing, includes the source route in the Mark message. PTMR inherently provides a microscopic (loose) path set-up, outlined by the Pegs. The spans between are filled according to domain-local CFR. Improved MD granularity allows for better adherence to a (strictly) given source route.

Finer MD granularity in order to improve policy compliance, however, has its price. The fact that the number of Last-hop MDs will increase not only results in a higher control traffic, router state and processing overhead, but also aggravates source congestion. On the other hand, designing larger MDs may lead to their having more than one Last-hop Peg, depending on the topology. Then, some of this overhead would incur anyway.

5.4 Shortest-path Delivery Trees

If the prevailing multicast policy calls for shortest-path routing, primarily for attaining minimum delay delivery, PTMR pegs out the Policy Route to a Last-hop Peg at shortest distance from the source. But, compared with the PIM SPT, the actual path length to local group members may increase, depending on their location within the Last-hop MD. However, in wide-area multicast the inter-domain segments are generally considerably longer than the intra-domain paths. In such cases, the possible delay increase for a specific group member is of minor influence. And when simultaneous reception is called for, the difference in delivery delay between members of the local group cluster due to varying path length still results within the Last-hop MD only.

In a symmetric environment, the PIM SPT immediately provides shortest paths to individual group members and PTMR will subsequently confirm them. But actual delivery occurs only on the shortest path if in all transit MDs symmetry prevails. If domains apply RPF, finer MD granularity may not only decrease the number of spans which are not shortest path but also their extent and thus their influence on the total path length. Still, under adverse asymmetric delay conditions within the transit MDs, it is possible that refraining from PTMR and staying on the SPT would lead to smaller transmission delays to individual group members. (This situation may also occur in PIM, when switching from RP-tree to SPT!)

6 PTMR Protocol Outline

6.1 Establishing Group Member Domains (Last-hop MDs)

Last-hop MDs, i.e. MDs populated with (S,G) group members, require at least one (S,G) Last-hop Peg which establishes a policy-sensitive path on their behalf. Last-hop Pegs have to be designated and released again if they do not serve group members any more. To this purpose, routers on a Peg Tree keep track of all (S,G) entries leading to domain-local branches, unlike those entries for transit to downstream MDs.

Last-hop routers tag the PIM (S,G) Join/Prune messages they generate as being domain-local. MBRs which receive domain-local Join/Prune messages remove all tags before sending them on. If a router receives a domain-local (S,G)-Join, the corresponding entry is labeled. If the label is newly added, then the Join message is sent further upstream (past routers which already have (S,G) state, but without label). An MBR receiving a tagged Join message is designated or confirmed as Last-hop Peg. Likewise, Prunes generated by last-hop routers have to proceed up to the point of domain-local intervention. Thus, if a Prune is contested only by branches not serving domain-local members, the entry's label has to be removed and the tagged prune message sent on. A Last-hop Peg Router receiving a domain-local Prune ceases this function, i.e. it is relieved of having to trigger Marking cycles.

An MBR which receives in the PIM RP-tree/SPT switch-over process an (S,G)-Join which is domain-local, it becomes initial Last-hop Peg. It is the task of the Marking cycles to commit this function subsequently to a peer MBR, according to the prevailing Policy Route.

6.2 First-hop Pegs

In response to a Request message, a Mark message has to be emitted at the source. In order not to involve sources in the routing process, a PTMR router on the source's local subnetwork has to serve as relay in the marking process and thus as First-hop Peg. Initially, Last-hop pegs do not have an entry for a source's First-hop Peg. In this case, the Request message is sent to the upstream RPF-neighbor towards the source, multicast hop-by-hop to the ALL-PTMR-ROUTERS group. The router to which the source is

directly attached is then established as First-hop Peg. (This is actually the first-hop router of the SPT - real or would-be!) With the first Mark message, Last-hop Pegs receive the address of the First-hop Peg (packet source address). They cache it, to be used as destination for their (unicast) Request messages. If the prevailing First-hop Peg becomes inactive, the source/First-hop Peg mapping has to be established anew via the RP mode.[4]

6.3 PTMR-tree Construction

Initial construction of an (S,G)-specific PTMR-tree proceeds from PIM-SM operation. Since the SPT is just a special form of PTMR-tree, initial construction of the latter is equivalent to its subsequent reshaping due to Policy Tree changes. Concretely, if a PIM SPT branch is being completed, it functions as the initial branch on the PTMR tree. PTMR-tree construction is again required when the Policy Tree is modified, either by adding a new branch or by altering an existing one due to policy changes. (When adding a new branch, Policy Route aggregation may require alterations to the existing Policy Tree too!)

Establishing a new (or an alternative) PTMR-tree branch starts with a Marking cycle, requested for an MD by its Last-hop Peg. It yields, in terms of Peg Routers, the prevailing Policy Route to the MD, i.e. every Peg has a record of its (S,G) Parent-Peg. If the chain of Pegs differs because the Policy Route changed, the (old) Last-hop Peg will then initiate an intra-domain Announce message multicast, by which last-hop routers will learn the address of the Last-hop Peg which they have to join. A new or modified Policy Route has to be microscopically connected and integrated into the existing PTMR-tree, by means of Peg Trees. To this purpose a special Peg-joining process is employed: last-hop routers have to join their Last-hop Peg and then the involved Pegs in succession join their parent.

A new or an alternative branch joining the (S,G) PTMR-tree may not only cross branches of existing Peg Trees for (S,G) traffic but also the SPT for the given (S,G) or even the RP-tree for G. Thus, in order not to lose or duplicate data packets, a procedure analog to the PIM RP-tree to SPT switch-over has to be applied. This means that new or differing states required for a PTMR-tree branch are initially scheduled only. During the Marking cycle, Peg Routers schedule their new Parent-Peg and subsequently, in the Peg-joining process, all involved routers schedule their (S,G,Peg) state. Effectively installing these states takes place in downstream succession when the scheduled routers see the first data packet coming down the new path. Routers which actually have to switch their incoming interface initiate an (S,G)-specific Prune on the abandoned path.

[4] Using a hop-by-hop procedure to find a PTMR router on the local subnetwork of a known source involves considerable processing and requires PTMR connectivity. Pertinant would be to apply an anycast or, in IPv4, a directed broadcast to the source's DR. But the latter function is poorly supported and problematic if subnetting is involved.

When an initial Last-hop Peg is activated, after a random delay it triggers a Marking cycle. From then on Marking cycles are periodically triggered by the active Last-hop Peg. A Marking cycle starts with the Last-hop Peg initiating a Request message for source S. For each received (S,G) Request message, source S's First-hop Peg starts a Mark message which is relayed back to the requesting Last-hop Peg. The Mark message is sent hop-by-hop to the PTMR-neighbor towards the requesting Last-hop Peg, on the prevailing policy-sensitive path for the given (S,G) traffic. (In an advanced concept, the routing decision could even be based on an attached flow label). Mark messages are multicast hop-by-hop to the ALL-PTMR-ROUTERS group. They contain the addresses of (S,G) as well as of the requesting Last-hop Peg and the targeted PTMR neighbor. Every MBR contacted on the path establishes itself as (S,G) Peg Router for its MD. Before forwarding the Mark message, it takes note of the conveyed Parent-Peg address and replaces it with its own address. If a contacted MBR is not already (S,G) Peg Router, i.e. if a new (or a different branch) for the Policy Tree is asked for, it schedules for joining the conveyed (S,G) Parent-Peg. On the other hand, if it is already in the correct (S,G,Peg) state it has to take no further action. Finally, if the contacted MBR is already (S,G) Peg but the Mark message conveys a different Parent-Peg address, this Peg is the merging point of two contesting parallel paths, caused either by conflicting Policy Routes or by a policy change. If the state's hold-down is expired, it schedules the new Parent-Peg; otherwise it retains its existing state. The first Peg which schedules a different Parent-Peg sets the Alter-flag in the Mark message and also makes a note of this.

The Mark message will finally reach the requesting Last-hop MD. If it is intercepted by one of the domain peers of the requesting Last-hop Peg, this router will become the new Last-hop Peg (i.e. the closest MBR for the given domain on the forward path from the source). It still forwards the Mark message towards the replaced Last-hop Peg. However, on this path the Parent-Peg Address field is not modified anymore.

Concluding the Marking cycle, the requesting Last-hop Peg distributes an Announce message on the prevailing Peg Tree (initially the intra-domain branches of the PIM SPT) to all of its last-hop routers, informing them about the address of the current Last-hop Peg. In addition, the status of the Alter-flag in the Mark message is copied into the Announce message. Announce messages are marked for the All-PTMR-ROUTERS group address and multicasted hop-by-hop on the Peg Tree. Others than last-hop routers ignore the content of the Announce message. Intermediate routers forward it, Child-Pegs however drop it.

A set Alter-flag in a Mark message (and subsequently in the Announce messages which the last-hop routers receive) indicates that the Policy Route has changed and that there are Peg Routers scheduled to join a new or different Parent-Peg. In this case, a Peg-join has to be executed. In order to prepare the scheduled Policy Route for data flow, the intra-domain routers as well as the Pegs now have to be state scheduled. This is achieved by concatenating Peg-joins

towards the source, starting from the last-hop routers. Thus, if a last-hop router receives an Announce message with the Alter-flag set, it initiates a Peg-joining process. Peg-joins are sent hop-by-hop towards the Parent-Peg and either install a new (S,G,Peg) router state or change the Parent-Peg entry of an existing one. Every Peg which receives a Peg-join completes its state scheduling and sends the Join message on towards its scheduled or valid Parent-Peg. With the completion of the scheduling, the actual switch-over takes its course. Peg-joins can be distinguished from regular SPT-joins (which catch CFR changes towards the given Peg) by having a Peg address attached. (Figure 5).

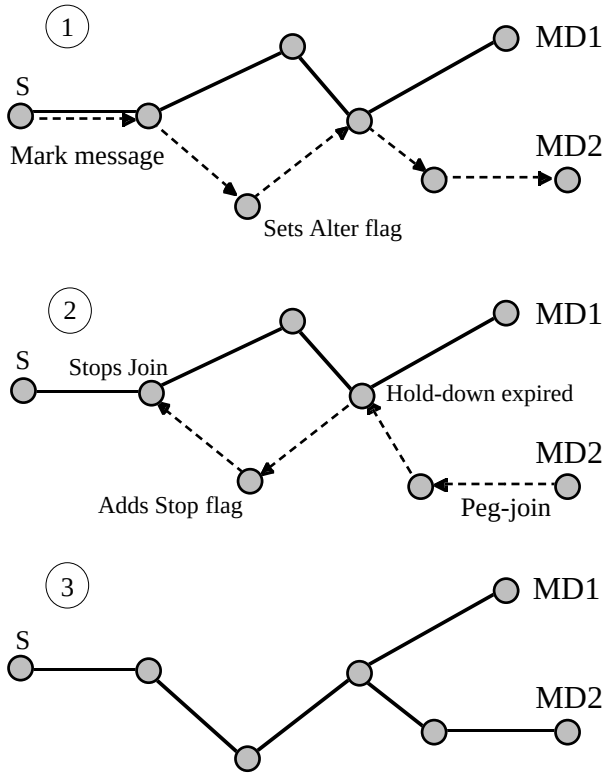

Figure 5: Peg Joining Process

In order to secure all required modifications to the Policy Route, Peg-joins have to proceed at least up to the last of the Pegs (on the way from the source!) whose Parent-Peg entry is already valid. Its downstream neighbor is the Peg which set the Alter-flag. When this router hands on the Peg-join, it appends it with a flag which instructs its Parent-Peg to terminate the joining process[5].

[5] Actually, the Peg-join process could stop when any first router on the existing PTMR-tree has been reached. But intra-domain routers cannot be aware of modifications in upstream MDs. With the chosen mechanism, if the Peg which notes down the Join stop is on a defeated (parallel) branch, then the Peg-joining will proceed up to the source!

If, in the course of an RPF-tree reshaping, a branch of an existing tree allowing (S,G) traffic is joined or crossed, all tree branches below the switching router are forced to follow. Thus, local last-hop routers and downstream MDs which are served by them will join the modified PTMR-tree immediately too. Otherwise, they will remain on their prevailing tree until they too receive a Mark message reflecting the current Policy Route. (Last-hop routers located on the RP-tree join only after having switched to the SPT).

6.4 Maintaining the PTMR-tree Mode

The Policy Tree for a given group is maintained according to prevailing policies, by means of periodic or event-driven Marking cycles. If for some reason the Marking cycle mechanism becomes dysfunctional, multicast packets may still be delivered, however on a frozen Policy Route. This problem is resolved thus: Last-hop routers which do not receive periodic Announce messages in due time anymore are made to switch back to the group's RP-tree.

Microscopic connectivity (i.e. the PTMR-tree) on the other hand is maintained by periodic and event-driven PIM Joins and Prunes based on state, topology and membership changes. This includes state changes due to modification of local delivery policies. If a source stops sending data packets or if their delivery on the present PTMR-tree gets disrupted, last-hop routers join the group's RP-tree again, in accordance with PIM's soft state refresh mechanism. If last-hop routers still see the source to be active, they will join towards it again.

A new last-hop router initially joins towards an RP for the given group G. If it decides to join the PIM SPT of a recognized active source and the new branch joins up the local Peg Tree for (S,G), then delivery on the relevant PTMR-tree is already established. With the next periodic Marking cycle, the new last-hop router will learn about its current Last-hop Peg; from then on it is able to join the Peg Tree directly. On the other hand, if the first MBR reached when joining the PIM SPT is not already an active (S,G) Last-hop Peg, it will become initial Last-hop Peg and hence triggers a new Marking cycle. Last-hop routers which do not have to serve local or downstream group members anymore prune themselves event-driven off the Peg Tree. If a Peg has no attached members of group G anymore, it prunes towards its Parent-Peg.

7 Summary

Policy-sensitive Multicast Delivery. For an acceptable overhead in terms of bandwidth consumption, control traffic, router state dimension and processing expense, wide-area multicasting needs to employ receiver-initiated, receiver-originating delivery tree construction, based on distributed route calculation. Such (sparse mode) concepts are inherently based on a joining mechanism, which in turn is dependent on Come-from Routing (CFR). If CFR is provided by a separate facility, it involves great expenditure. Thus, in currently deployed MR protocols, CFR multicasting is based on RPF information, which can be derived from the prevailing unicast routing. While this approximation entails a substantial cost reduction, it may lead in asymmetric environments to divergent paths. Such situations may be due to deriving CFR information from an extraneous routing facility or to asymmetric link and policy conditions, with the effect that shortest path delivery as well as (multicast-specific) policy control cannot be guaranteed.

Bidirectional multicast delivery trees allow optimal dispersion of multicast packets dispatched from sources located on the tree itself. But they cannot be applied in comprehensive policy routing which includes complying with shortest-path, QoS and other source-specific criteria and requires, in addition to "true" CFR, source-specific unidirectional multicast delivery.

Source-originating tree construction (for scalability in sparse mode it needs to be receiver-initiated as well) would lend itself to establishing policy-sensitive multicast delivery. Particularly, source demand policies could be enforced, even if source routes are employed. However, receiver-initiated source-originating tree construction requires a source to individually establish receiver-specific paths to known active receivers, which are then fused as much as possible into a delivery tree. This leads to a control message congestion at the sources. As a result, group member location and traffic aggregation mechanisms of source-originating tree construction exhibit poor scalability.

BGMP. Inter-domain MR models are currently emerging which not only cope with heterogeneous MD routing protocols and scalability issues but also accommodate policies imposed by provider domains. This shifts architectural design priority away from group state minimization and delivery quality. Inter-domain MR faces problems with the control plane interoperability of different (intra-domain) MR protocols, requiring multicast traffic encapsulations and even protocol modifications. Multicast Border Gateway Protocol (BGMP) is such an inter-domain MR proposal. It builds core-based inter-domain (bidirectional) delivery trees, by apply CFR based on BGP-type unicast routing. The underlying path vector concept provides genuine CFR on the domain level, but only if a specific multicast RIB and multicast-specific path updates are provided. In the general case, a separate multicast-specific (unicast) inter-domain routing facility has to be provided underneath. By aligning MDs and ASs, BGMP allows an efficient AS policy control because many of the required functions and resources are already provided by the presently deployed BGP. Still, if unicast/multicast topology and policies are not congruent, the expenses in terms of router processing, memory space and control traffic overhead increase accordingly.

With its bidirectional multicast packet dispersion, BGMP is deficient in asymmetrical environments. Moreover, it cannot support source-specific delivery criteria. It is indeed possible for BGMP to comply with requests for source-specific delivery, but for the sake of reduced protocol complexity, only in restricted cases.

BGPM's policy control potential is restricted to the path selection procedure and policy constraint support of BGP's underlying node routing paradigm and path vector concept. This implies for example that network-specific policies cannot be supported. Furthermore, to consider CFR interfaces only is a policy mechanism which may not be discriminatory enough, even for AS policy. For example, traffic barriers imposed by AS policies may be by-passed if a source is covered by a prefix which is homed to more than one domain.

PTMR. The PTMR concept whose architecture and protocol outline are presented in this paper aims at an efficient solution for attaining policy-sensitive data packet delivery in Internet-wide multicast, across various domains, even under asymmetric conditions. PTMR's characteristic feature is the forwarding of multicast packets in accordance with any underlying multicast-relevant routing, including comprehensive policy routing (supporting source-specific policies as well as shortest-path and QoS criteria). PTMR applies receiver-initiated, source-originating tree construction, but restricts it to a macroscopic layer (alleviating the source congestion problem inherent in this approach). Thus, policy-sensitive paths to receiver clusters (contained by MDs) are established, defined in terms of MDs. PTMR does not need to be aligned with ADs, routing domains and inter-domain (unicast) routing. MDs actually function as policy domains. The fact that PTMR is a single-layer concept, which is independent of the prevailing routing control(s), considerably alleviates the interoperability problem of MR domains and allows optimal fusion of transit traffic and local delivery. PTMR is based on PIM-SM which provides for the source/receiver handshake and for initial source-specific trees. Its affinity to PIM leads to synergy, on the protocol level as well as with individual mechanisms.

PTMR proceeds from PIM-SM operation; ist activation is optional. The source/receiver handshake is established by the PIM RP-mode. In order to reduce source congestion, establishing policy-dedicated delivery is randomly delayed. For completely symmetric conditions (assumed to be the most frequent case), a PTMR tree branch is isomorphic to that on the SPT and no state change occurs when switching from the SPT to the PTMR tree. Building a PTMR branch starts with individual group members switching to the SPT. But the PTMR procedure is only dependent on SPT construction for establishing the Last-hop Peg of the new Policy Tree branch. Thus its completion may be interrupted, either intentionally or due to blocked PIM SPT delivery. The given alternatives offer an optimal admini-

strative choice, based on the given expected session length, network environment conditions and delivery requirements:

- Remaining in PIM-SM mode:
  - RP-tree
    No tree switch-over and SPT state necessary
  - SPT
    Improved delivery quality
    Policy-compliant in a symmetric environment
- Activating PTMR:
  - Completing the SPT
    Low PTMR latency in a symmetric environment
    (Otherwise improved delivery quality)
    SPT state necessary
  - Not completing the SPT
    Increased PTMR latency in any case
    No intermediate SPT state necessary

If the source/receiver mapping is a priori known, as assumed in Single-source Multicasting, a unique single-source PTMR tree could be established, starting directly with the PTMR procedure.

PTMR-tree. The PTMR architecture is characterized by a structure called Policy Tree, which is the product of macroscopic source-originating tree construction. It is formed by the merging of Policy Routes, i.e. macroscopic paths from source to group member MDs (Last-hop MDs) given by a sequence of MDs which satisfies the policy requirements of the source and the involved domains and supports the requested service quality. Policy Routes are marked (pegged) with Peg Routers (domain ingress routers) according to prevailing forward routing criteria. A PTMR-tree, i.e. the microscopically completed Policy Tree, is established by last-hop routers as well as Child-Pegs CFR-joining towards their Parent-Peg. This results, within an MD, to a Peg-rooted delivery tree (called Peg Tree) which optimally aggregates all congruent paths, those for local delivery as well as those for transit. The fact that PTMR considers neither how a path which it pins to a Last-hop MD has as been established, nor what route selection criteria were involved, makes this concept independent of the underlying macroscopic policy model and the prevailing routing. Particularly, comprehensive policy routing can thus be supported, e.g. enforcing source-given policies using header routing, whereby the Pegs provide a (loose) path set-up.

Formal Model. Formally, the PTMR model is a generalization of the PIM Sparse Mode architecture. In generic PIM SPT delivery, all members of group G are directly joined towards the source S, i.e. all routers maintain (S,G) state with the CFR interface for S as the incoming interface. In PTMR however, routers' incoming interface for S is the CFR interface for the Parent-Peg. Microscopically the PTMR-tree is maintained by periodic and event-driven Joins and Prunes based on CFR information and on group member presence. Policy on the other hand is imposed independently by the selection of Peg Routers. Since policy changes are much less dynamic than connectivity and membership changes, periodic updating of the Policy Tree may be exercised less aggressively, reducing control message overhead and source congestion.

Policy Compliance. For PTMR-tree delivery to guarantee compliance with a given policy set, multicast packet delivery within a domain has to follow the macroscopic path which was taken by the Mark message and led to the given Policy Route, i.e. each transit domain has to complete its Policy Tree span accordingly. Thus, its CFR must not produce divergent paths, i.e. genuine CFR has to be applied. This is very costly, but the RPF approximation cannot guarantee policy-sensitive multicast packet delivery. PTMR provides a flexible concept which allows MDs to contribute different service qualities or, reversely, prepare their transit delivery conditions according to what they offer to the global network. The minimum cost solution consists in all MDs resorting to RPF. In this case macroscopic access, transit and route selection policies are complied with. On the other hand, support for microscopic policies like shortest path and QoS can only be provided if no asymmetric conditions exist. This can be secured in an MD by either installing genuine CFR or by avoiding asymmetries through administrative measures. If the actual multicast transit provided by the MDs is inadequate, in the case of an additive metric, smooth degrading takes place.

The PTMR architecture allows arbitrary granularity of its MDs, i.e. policy domains. The obvious choice for MDs are the Administrative Domains. Finer MD granularity facilitates securing intra-domain symmetry in order to overcome the RPF deficiency. Also, in header routing, congruence between actual source route and the macroscopic (loose) set-up path can thus be improved. However, finer MD granularity increases PTMR overhead and the source congestion correspondingly.

PTMR Policy Tree construction adheres to requirements imposed specifically on multicast traffic. A Policy Tree is specific to a source and a given multicast group, but not more. Thus, only those policies can be accommodated whose routing criteria for (S,G)-specific multicast packet distribution can be mapped onto a single spanning tree. If the network is furnished with a flow control concept, policy criteria with wider scope and higher granularity could be accommodated. Yet the resulting larger number of policy trees would entail increasing complexity and overhead. On the other hand it would allow multicast traffic spreading.

Comparing PTMR and BGMP. PTMR is a single-layer wide-area multicast routing protocol based on arbitrary multicast policy domains. Being source-specific, it is burdened by the control expense of receiver-initiated source-originating tree construction and the control traffic congestion at the source. BGMP on the other hand is an inter-domain multicast routing protocol applied between multicast routing domains. Its main problem is inter-operability with intra-domain MR protocols. For building domain delivery trees, both protocols rely on underlying routing potentiality; BGMP consults a multicast-relevant RIB for CFR information and PTMR marks prevailing (policy-compliant) paths. If more than simple inter-domain

access and transit policies need to be supported, both concepts are dependent on appropriate intra-domain delivery. PTMR requires a priori extensive PTMR functionality deployment, expanded from PIM-SM technology, which brings substantial basic control expense. BGMP on the other hand allows smooth incremental transition from existing technology and resources, i.e. AS-aligned multicast inter-domain routing based on BGP (whose path vector mechanism allows the establishing of come-from paths) and with selected MR protocols for intra-domain routing. Thus, when only control of simple AS access and transit policies is required, BGMP's efficiency is optimal. However, the routing expense is dependent on the extent of divergent paths. If a richer policy model has to be supported, both concepts require a correspondingly elaborate underlying routing strata. For BGMP this leads to significant problems, e.g. with aligning underlying routing layers with multicast routing and with MD interoperability, and it would soon exceed the potential of BGP-type routing. For (unlimited) source-specific policy support, BGMP would also have to be appended by a tree switch-over procedure, which would be extremely complex if bidirectional shared tree delivery is retained. PTMR on the other hand is capable of accommodating any underlying routing protocol, including comprehensive policy routing, i.e. shortest-path, QoS and other source-specific delivery.

Appendix:
PIM-SM Protocol Synopsis

(Brief description of PIM-SM, emphasizing aspects relevant for understanding the PTMR extension.)

PIM-SM has been introduced as a simple, flexible and scalable architecture, designed to avoid excessive data traffic and routing state explosion in Internet-wide multicasting. It applies RPF-joining, with routing information derived from existing unicast routing. However, it is independent of its type (PIM: Protocol Independent Multicast!). [6][16].

PIM-SM is based on a group-shared delivery tree which forwards traffic from all active sources to a specific group, called (*,G) traffic. The root of the group-shared tree is called Rendezvous Point (RP). It is targeted by sources which become active. A last-hop router RPF-joins the group-shared tree (RP-tree) initially. Upon seeing traffic from source S and when the traffic type warrants it, it will switch to a shortest-path source tree (SPT), which is forwarding (S,G) traffic. According to the PIM-SM protocol, last-hop routers switching over to the SPT may wait until they have received a minimal number of data packets from the given source within some given interval. This avoids the overhead of SPT state when small numbers of packets are sent sporadically.

Recipients establish their branch of a multicast delivery tree by means of router states which are installed by sending a Join message hop-by-hop towards the source (or an PR). A router hands on a Join message only on its RPF interface, i.e. the interface it would use to send its own traffic to the multicast source (or an RP). In order to prune unwanted branches from the delivery tree, Prune messages are used. Prune messages are processed analog to Join messages, but they eliminate router state: If due to a received Prune a router eliminates an outgoing interface entry and it does not then have to serve members of the given group anymore, the Prune message is sent on upstream. (On an outgoing interface to a multi-access network, it must be made sure that no other branch contests the pruning!) Join/Prune messages are multicast hop-by-hop to the ALL-PIM-ROUTERS group. They contain the address of the targeted upstream neighbor.

If a source starts emitting multicast packets, the local elected router (Designated Router, DR) sends them IP-encapsulated to the RP (register phase). The RP then joins towards the sending source. A last-hop router first joins the RP-tree for the given group, by sending a Join message towards the RP. When switching to an SPT, it sends an SPT-join message directly towards the source. Routers on the way create an (S,G) state. The SPT-join may cross RP-tree branches on its way. Thus, in order to avoid loosing or duplicating packets during the transition phase, the switch-over is only scheduled, by marking the (S,G) entry with a flag. The effective (S,G) states are installed in downstream succession when the scheduled routers see the first data packet coming down the SPT path. The scheduling flag is then removed. Routers which actually have to switch their incoming interface to the SPT initiate a source-specific Prune on the abandoned RP-path. This is achieved by restricting the (*,G) shared forwarding state in the involved routers with a negative entry for source S. Last-hop routers on tree branches below the router which actually joins up the SPT are forced to this delivery mode too.

For maintaining delivery trees, PIM uses Joins for periodic and event-driven state refreshes. In steady state each router sends its parent periodic refreshes to capture state, topology and membership changes. If a router does not receive Join messages for an outgoing interface anymore, it eliminates the corresponding entry. On the other hand, if it does not receive any relevant data packets anymore, it will time-out the pertinent state. (This type of control is called soft state refreshing). When a last-hop router with attached active group members times out a source-specific state, it eliminates the corresponding negative entry towards the RP again. If unicast routing changes, routers send a Join on the new RPF incoming interface and a Prune on the previous one. Join/Prune messages contain aggregated information. A single message contains both a Join and a Prune list, each containing a set of source addresses tagged whether their joining/pruning applies to the RP-tree or the SPT. Routers with directly connected group members as well as sources have to know an RP for the multicast group. RPs are administratively designated and their identity is distributed.

Acknowledgments

This work was partially supported by Cisco Systems, San José and by Neu-Technikum Buchs (NTB), Switzerland. I would like to thank Cisco Systems for inviting me to spend a sabbatical with them and its IOS development team for their hospitality and support. Moreover, I would like to extend my thanks to the anonymous referees for comments on an earlier version of this paper. Last but not least, I am grateful to Mary Baker for the visitor privileges at Stanford University and to Paul Milsom.

References

[1] Z. Wang, J. Crowcraft. "Quality-of-Service Routing for Support of Multimedia Applications". IEEE Journal on Selected Areas in Communications, Vol. 14, No. 7. Sept. 1996.

[2] R. Braudes, S. Zabele. "Requirements for Multicast Protocols". RFC 1458, May 1993.

[3] M. Steenstrup. "An Architecture for Inter-Domain Policy Routing". RFC 1478, June 1993.

[4] D. Estrin, D. Zappala, T. Li, Y. Rekhter. "Source Demand Routing Protocol: Packet Format and Forwarding Specification".Work in progress, March 1995.

[5] S. E. Deering. "Multicast Routing in a Datagram Internetwork". Ph.D. Thesis, Stanford University, Dec. 1991.

[6] D. Estrin et all. "An Architecture for Wide-Area Multicast Routing". Proceedings of the ACM SIGCOMM, London, Aug. 1994.

[7] D.Meyer. "Some issues for an Inter-domain Multicast Routing Protocol". Work in progress, March 1997.

[8] Y. K. Dalal, R. M. Metcalf. "Reverse Path Forwarding of Broadcast Packets". Communications of the ACM, 21(12):1040-1048, Dec. 1978.

[9] L. Wei, D. Estrin. "The trade-offs of Multicast Trees and Algorithms". Proceedings of the 1994 Internat. Conference on Computer Communications and Networks, San Francisco, Sept. 1994.

[10] M. Myjak et all. "Scenarios and Appropriate Protocols for Distributed Interactive Simulation". Work in progress, March 1997.

[11] D. Cheriton, H. Holbrook. "Single-source Multicast". Work in progress.

[12] T. Pusateri. "Distance Vector Multicast Routing Protocol". Work in progress, Aug. 1997.

[13] A. Ballardie. "Core Based Trees (CBT) Multicast Routing Architecture". RFC 2201. Sept. 1997.

[14] J. Moy. "Multicast Extensions to OSPF". MOSPF: Analysis and Experience". RFC 1584, March 1994.

[15] D. Estrin et all. "Protocol Independent Multicast (PIM): Dense Mode Protocol Specification". Work in progress, Aug. 1997.

[16] D. Estrin et all. "Protocol Independent Multicast-Sparse Mode (PIM-SM): Protocol Specification". RFC 2117. June 1997.

[17] J. Moy. "OSPF Version 2". RFC 1247, July 1991.

[18] A. Thyagarajan, S. Deering. "Hierarchical Distance Vector Multicast Routing for the MBone". Proceedings of the ACM SIGCOMM, Cambridge, Mass., Oct. 1995.

[19] Y. Rekhter, T. Li, editors. "A Border Gateway Protocol (BGP-4)". RFC 1771, March 1995.

[20] V. Paxson "End-to-End Routing Behavior in the Internet". Proceedings of the ACM SIGCOMM, Stanford, CA., 26. Aug. 96

[21] D. Thaler. "Interoperability Rules for Multicast Routing Protocols". Work in progress, March 1997.

[22] W. Fenner. "Domain-Wide Reports". Work in progress.

[23] S. Deering, B. Cain, A. Thyagarajan. "Internet Group Management Protocol, Version 3". Work in progress. Dec. 1997

[24] T. Bates et all. "Multiprotocol Extensions for BGP-4". RFC 2283. Feb. 98

[25] D. Thaler, D. Estrin. D. Meyer "Border Gareway Multicast Protocol (BGMP): Protocol Specification". Work in progress, March 1998.

[26] D. Estrin et all. "The Multicast Address Set Claim (MASC) Protocol". Work in progress, Nov. 1997.

[27] A. Ballardie, B. Cain, Z. Zhang. " Core based Tree Multicast Border Router Specification". Work in progress, March 1998.

[28] M. Sola, M. Ohta, T. Maeno. Scalability of Internet Multicast Protocols". Submitted to INET '98 Conference. Geneva 23. July 1998.

[29] D. D. Clark. "Policy Routing in Internet Protocols". RFC 1102, May 1989.