2007-10

Multicast traffic aggregation in MPLS-based VPN networks

Martínez Yelmo, Isaías

IEEE Communications Society

IEEE Communications Magazine, vol. 45, issue 10, October 2007, pp. 78-85
http://hdl.handle.net/10016/2983

Multicast Traffic Aggregation in MPLS-Based VPN Networks

Isaias Martinez-Yelmo, David Larrabeiti, and Ignacio Soto, Universidad Carlos III de Madrid
Piotr Pacyna, AGH University of Science and Technology

ABSTRACT

This article gives an overview of the current practical approaches under study for a scalable implementation of multicast in layer 2 and 3 VPNs over an IP-MPLS multiservice network. These proposals are based on a well-known technique: the aggregation of traffic into shared trees to manage the forwarding state vs. bandwidth saving trade-off. This sort of traffic engineering mechanism requires methods to estimate the resources needed to set up a multicast shared tree for a set of VPNs. The methodology proposed in this article consists of studying the effect of aggregation obtained by random shared tree allocation on a reference model of a representative network scenario.

INTRODUCTION

Multiprotocol label switching (MPLS) has revolutionized service provider (SP) networks in recent years as the technology that enables effective multiservice exploitation of a packet-switched network. It satisfies all the connection-oriented carrier-grade requirements operators only used to find in more expensive technologies such as asynchronous transfer mode. The main short-term service driver for this technological move is, on one hand, the ability to transport regular TCP/IP traffic and IP telephony, layer 3 virtual private networks (L3VPNs), and layer 2 VPNs (L2VPNs) in a scalable and isolated way. On the other hand, MPLS makes it possible to amortize previous investments thanks to convergence and compatibility with connection-oriented frame relay and ATM networks, optical networks (through generalized MPLS, GMPLS), and any layer 2 technology. Consequently, there is a dual full service and network convergence motivation in this evolution that pushes the SPs to install MPLS in their networks.

One of the most important challenges in the evolution of MPLS multiservice backbones is multicast service. Although implementation of IP multicast service has evolved substantially since its inception in the early '80s, the existence of less scalable yet safer ubiquitously supported unicast-based alternatives, such as application layer multicasting, has hindered the global deployment and availability of this service to all Internet users. In fact, it seems that not all SPs are willing to make available a service that is not easy to manage, giving as an excuse lack of demand, even though a few SPs have shown that the service is viable by means of proper measures against denial-of-service attacks and traffic control. This situation may change in the next years as the transmission costs derived by peer-casting traffic are growing.

Irrespective of whether the multicast service is made available to end users or not, most multiservice SPs have deployed it in a controlled way in order to take advantage of its efficiency in the delivery of high-speed multipoint streams. An example of this are triple-play providers [1] that deliver TV channels over IP multicast to their asymmetric digital subscriber line (ADSL) set-top boxes. Usually, the last hop is delivered over IP unicast from a multimedia relay at the SP point of presence (PoP). This service can be delivered by IP/MPLS over a single point-to-multipoint (P2MP) label switched path (LSP) from a content delivery root to the relays, usually with caching capabilities for video on demand (VoD), or down to the subscribers.

But this is not the only service worth the cost of multicast deployment and management by SPs; the MPLS-based VPN service is getting momentum, and also the trend to hold high-quality IP multipoint videoconferencing, to broadcast corporate TV news channels or to perform fast bulk file/disk replication over the company's PC fleet. All such applications can take advantage of IP multicast service if available.

However, the challenge of multicast delivery is more complex for the MPLS-based VPN SP for scalability reasons, since the backbone must support an overlay of isolated virtually private P2MP LSP trees. These trees will likely be different for each VPN, and the trees may have dynamic structure. Therefore, LSP sharing is not straightforward.

Furthermore, even if the IP multicast service is not required, Ethernet multicast/broadcast emulation is still needed in the case of the multi-

 0163-6804/07/$20.00 © 2007 IEEE

site L2VPN. This is the context where P2MP LSPs may save bandwidth for the SP at the cost of a significant increase of forwarding state in core routers. As we shall review, experts have surrendered to the evidence that only an intelligent aggregation of multiple VPNs into the same multicast/broadcast tree can yield important bandwidth savings at a reasonable cost. How this partition and assignment of VPNs to trees should be made is an open research issue, given the diversity of topologies, traffic, and sites of different VPNs and backbone networks. On the other hand, high-rate flows may justify the setup of group-membership-aware multicast trees to avoid traffic in nodes not leading to group receivers.

The rest of this article is organized as follows. We describe how to build P2MP trees in an MPLS network suitable for arbitrary aggregation of VPN trees. We present the problem of how to bundle and share multipoint LSPs in a scalable way and the techniques being developed in the Internet Engineering Task Force (IETF) for this purpose. We explore the trade-off of state vs. bandwidth in the particular multicast VPN context. We draw a few practical conclusions and suggest directions for future work.

Signaling Point-to-Multipoint MPLS LSPs

A fundamental functionality required to take advantage of multicast in the network core is the ability to set up and use P2MP label-based forwarding entries. There are two protocols defined by the IETF to build LSPs in MPLS networks: Resource Reservation Protocol with Traffic Engineering (RSVP-TE) and Label Distribution Protocol (LDP). Both can be extended to support P2MP LSPs [2–4].

RSVP-TE builds the P2MP trees from the root to the leaves, whereas LDP builds the trees from the leaves to the root. In the case of IP multicast trees, LDP is intended to build the LSP following the IP multicast routing protocol. However, since all the solutions developed for scalable VPN multicast services are based on traffic engineered multi-VPN tree sharing, RSVP-TE is a more suitable tree setup protocol for this purpose. In fact, RSVP-TE indeed allows a tree to be constructed from a root router to a given set of leaf routers — in our case, the set of provider edge (PE) routers serving the sites of all the VPNs that have been selected to share the tree. This makes the LSP tree setup more versatile but also more complex as the signaling has to deal with explicit subtree descriptions.

Let us briefly describe P2MP trees set up in an MPLS network using RSVP-TE as proposed in RFC 4875 [4]. If a P2MP tree needs to be configured using RSVP-TE, it has to be explicitly defined from the root by means of a *Path* message. This *Path* message contains a *Session* object with a tree identifier for each P2MP (P2MP-ID) tree and the explicit routes for the branches of each tree. Thus, the specification of a tree in RSVP-TE is based on a set of secondary route objects (SROs), one per leaf, that describe the branches stemming from the primary path defined by an explicit route object (ERO). Thus, each branch will have its own identifier (S2L-ID) that is associated with its own ERO, as explained before. The purpose of the route objects is the following:

- ERO: This object defines the main branch of a tree or subtree in a P2MP MPLS tree when used in an RSVP-TE *Path* message. It contains the source routed LSP path from a source node to the leaf, and the distance from the router sending the object to the leaf. An ERO object is represented by *ERO* [*R*1, *R*2, …, *Rn*], where *Rn* is the *n*th router in the path.
- Secondary ERO (SERO): This object signals the secondary branches of a P2MP tree when used in an RSVP-TE *Path* message. Since each node is aware of building a tree with a set of EROs, there is no need to include the whole path from the root to the target leaf in each SERO object. This implicit sharing of subpaths to nodes gives a certain level of compression. A SERO object is described by *SERO* [*R*1, *R*2, …, *Rn*].

Apart from the ERO and SERO there are also a record route object (RRO) and a secondary RRO (SRRO), which are used for route recording purposes [4]. An example of the signaling procedure with the use of ERO and SERO objects is shown in Fig. 1. Router A is the root of the P2MP tree and has to send a *Path* message with objects that specify the branches to the three leaves C, E, and F. One path is chosen as primary (B-D-F in Fig. 1), and the others stem from this to build the tree. Once the P2MP tree has been signaled from the root to the leaves, the tree is traversed back hop by hop from the leaves to the root by *Resv* messages, as depicted in the figure.

Branching nodes just add forwarding entries (output interface and label) for the same P2MP-ID. It is foreseen that multipoint LSPs can also take advantage of multipoint/broadcast capability at the link layer, if available.

In the example, multicast MPLS frames could be sent by router D and shared by routers E and F if both have configured the same label for the tree. This could be achieved by allowing upstream routers to assign labels as in GMPLS [5]. This point is being standardized [6, 7].

Finally, it should be recalled that all RSVP-TE signaling must be periodically refreshed according to RSVP's softstate design in order to keep the tree up. A more detailed explanation of this process can be found in [4].

Multicast and MPLS-Based VPNs

One of the important applications contexts of P2MP LSPs is the delivery of the MPLS-based VPN service. There are two main VPN services being designed at the IETF supported by an IP MPLS networks: L3VPN supplying a routed service [8] and L2VPN that, when it emulates the full multicast/broadcast capability characteristic of the broadcast multiple access segment, is called virtual private LAN service (VPLS) [9].

The way they work is conceptually similar. Both try to keep the state in core routers bounded by tunneling multiple VPNs on shared LSPs,

A fundamental functionality required to take advantage of multicast in the network core is the ability to set up and use point-to-multipoint label-based forwarding entries. There are two protocols defined by the IETF to build LSPs in MPLS networks: RSVP-TE and LDP.

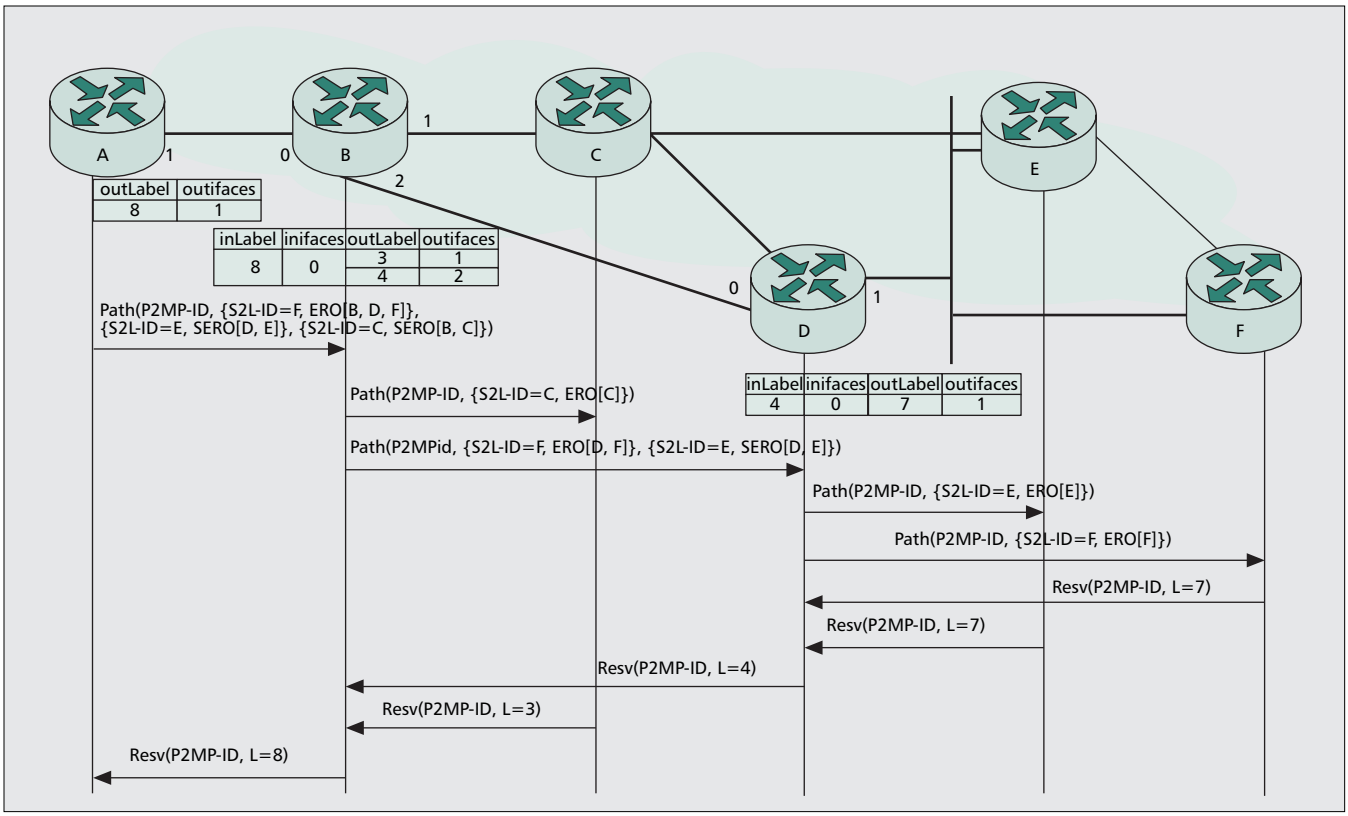

Figure 1. *A possible point-to-multipoint signaling sequence with RSVP-TE.*

and both try to cope with identical scalability problems using P2MP-based multicast. Consequently, the respective working groups have suggested parallel approaches, and the concepts explained here are valid for both types of VPNs.

BGP MPLS-Based VPNs

In a BGP MPLS-VPN the SP network is made up with a number of label switching routers (LSRs), identified as provider (P) for core LSRs and PE for routers interfacing with customer edge (CE) devices (a router or switch located at user premises) (Fig. 2).

If the MPLS network is configured for full mesh connectivity between PEs, a set of possibly merging LSPs are supposed to have been automatically established from any PE to any other PE in the network following the shortest paths with the help of LDP. A key idea of the MPLS concept is that the egress PE for a packet, or Ethernet frame in the case of L2VPN, is implicitly determined by the label the ingress PE sets on it, and the forwarding is based on that outer label all the way to the egress irrespective of its content. MPLS-based unicast VPN scalability is achieved by tunneling all VPN traffic into preexisting LSPs. Edge routers make use of these LSPs to traverse the core by means of label stacking, and reach the target egress PE irrespective of the VPN to which the traffic belongs. This way, the addition of a new VPN or a new VPN site does not imply an increase of forwarding state in the core, only at the involved edge routers. Thus, VPN-specific forwarding information is required only at PEs.

Furthermore, all VPN-specific information carried by the MPLS frame is tunneled (transported transparently through the network core) between the involved PEs. How this VPN-specific information is conveyed depends on the sort of VPN, layer 2 or 3. In the case of L3VPN it simply consists of another MPLS label pushed by the ingress PE before sending it over the generic carrier LSP; this label is previously agreed on by both endpoints of the LSP. In the case of L2VPN it requires additional encapsulation information to enable the multiplexing of different L2 technologies. Once the VPN and the origin of the incoming PDUs is properly identified at the egress PE, its forwarding must be based on its particular private forwarding table. To construct these private tables, routing (L3VPN) or bridging (L2VPN) information of a given VPN is distributed only to the PEs involved in the forwarding. One option to control this exchange is the Border Gateway Protocol (BGP). This is a more natural solution for L3VPN [8], as BGP was designed to deliver route information rather than labeled LAN connectivity information. For example, in RFC 4364, private routing information is exchanged by PEs, by means of i-BGP, completely isolated from other VPNs' routing schemes thanks to a Route-Distinguisher attribute, and its distribution and use is controlled by filtering on the Extended Community attribute.

Multicast Service for VPNs

The extension of the existing L3VPN and L2VPN architectures to multicast has been addressed by the respective IETF working groups in [9, 10], keeping a similar position on the management of multicast VPN traffic. The main issue is that, unlike in unicast VPNs, the implementation of

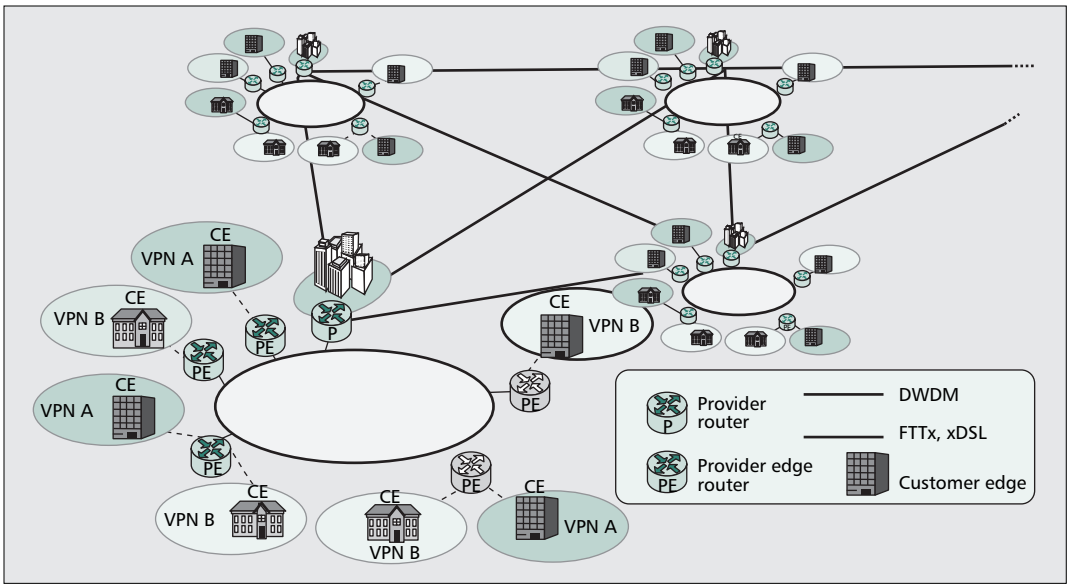

Figure 2. *Sample physical topology of a VPN SP network.*

An effort to take advantage of the P2MP functionality for the VPN service is underway at the IETF. The pragmatic approach to deal with the question is to design procedures that can be either driven by the routing protocols or used as multicast traffic engineering tools according to the existing TE databases.

optimal routing from a source to a set of destination PEs requires per-multicast source per-multicast group per-VPN state in the core (provider routers) to build the required specific distribution tree. That means adding yet another multiplying factor to the already complex IP multicast routing that may exceed the forwarding table and processing capacity of core routers.

One simple workaround to this problem commonly used, yet inefficient, is replicating the frame at the ingress PE over all unicast LSPs leading to all other VPN sites. In the case of a VPLS VPN established with a reduced LSP connection topology, it is also possible to run an instance of the Spanning Tree Protocol (STP) per VPLS. In both cases there may be an important waste of bandwidth from not using the P2MP functionality of the network nodes. On the other hand, it is the only possible solution when there is no multipoint capability in the core network as can be the case with an optical GMPLS core.

An effort to take advantage of P2MP functionality for VPN service is underway at the IETF. The pragmatic approach to deal with the question is to design procedures that can be either driven by the routing protocols or used as multicast traffic engineering tools according to the existing TE databases, to make it possible for a VPN SP to flexibly trade off bandwidth and state. The IETF defines two types of aggregate multi-VPN trees to operate with in a more scalable way:

- **Aggregate inclusive tree** refers to a tree that carries all the multicast traffic (all the groups) of an aggregate of VPNs. This tree includes every PE that is a member of any of the VPNs in the aggregate; hence, the tree may uselessly deliver traffic to PEs not in the VPN by sending the packets there, or to PEs that do not have group members.
- **Aggregate selective tree** refers to a tree that can be used to carry traffic for a set of one or more multicast groups *G* belonging to a specified aggregate of VPNs. This tree should include only the PE leading to members of the group set *G*. In this case only PEs with members of *G* without sites of the originating VPN would receive unwanted traffic. This group-aware (selective) and membership-aware tree wastes less bandwidth than the inclusive one, but costs more in terms of forwarding state.

Therefore, aggregate inclusive trees are efficient in scenarios with multiple multicast groups (and also broadcast/unknown in VPLS), whereas selective trees are meant to carry just high-bandwidth multicast flows. An SP will usually partition the VPN set according to the best matched topologies and the number of aggregate inclusive trees it can manage (see next section) and manage the remainder with selective trees. The latter should often be source-rooted and can be automatically triggered by a bandwidth threshold (e.g., migration to the shortest path tree in PIM-SM).

A specific problem to be solved with this aggregate approach is demultiplexing of traffic belonging to different VPNs at the egress PEs. In unicast VPNs the question is solved by regular MPLS downstream binding. However, in multicast VPNs all downstream leaves should agree on a common label to identify the VPN traffic. Instead, [9, 10] propose to use upstream bindings originated by the ingress PE and advertised via iBGP. Since different ingress PEs may suggest the same label, it is necessary to keep a different label space for each tree. Hence, penultimate hop popping must be disabled, as the egress PE needs to identify first the tree that the MPLS frame has used according to the outer label and second the VPN instance to which the packet belongs by the inner label.

Figure 3 shows an example of MPLS forwarding setup for an L3VPN shared distribution tree.

Per-Source Trees vs. Shared Trees

Once decided on the set of VPNs that are going to share a distribution tree, there is also an option of sharing a mesh of source-rooted trees or a single shared tree.

In the former case, each PE must have configured the closest root — usually the PE itself

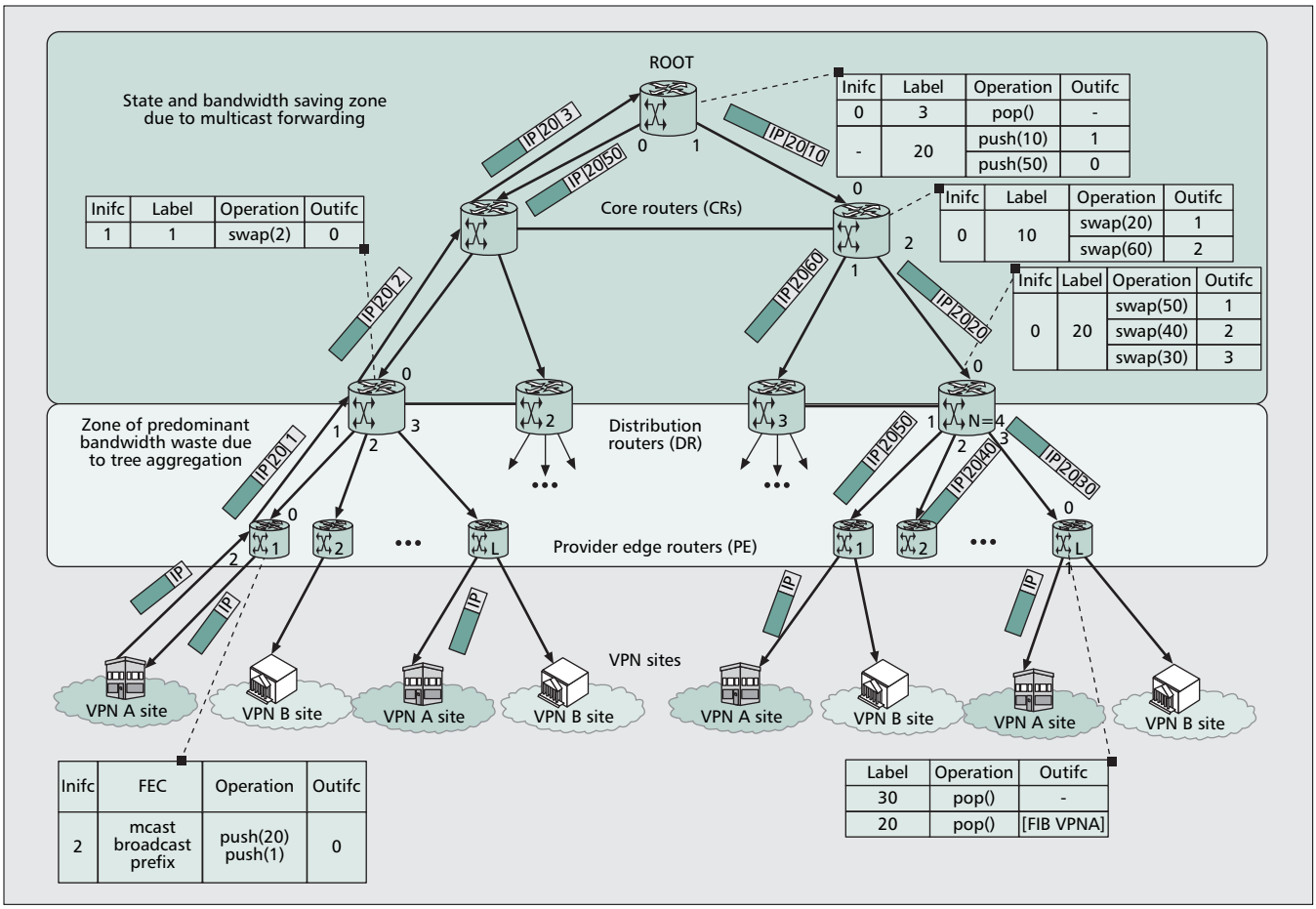

Figure 3. *Network topology with aggregated multicast tree example.*

— and the customer multicast/broadcast flows are sent to it for distribution.

In the latter case all the leaf PEs send their packets to the single root through a tunnel terminated at the root, and the root pops the outer label, checks the inner label previously assigned by the ingress PE for a given VPN group, and, according to it, swaps it to the label assigned to that VPN tree by the root. The forwarding entry also includes the second label to forward the traffic down the distribution tree shared with other VPNs (as in Fig. 3).

A choice between a single tree vs. a mesh of source-routed trees depends on available resources. The source-routed one provides more reliability and lower delay, at the cost of forwarding state and complexity; thus, a protected single shared tree is the most likely configuration. Furthermore, configuration of a shared tree is a task that needs human supervision and relies on a signaling protocol supporting P2MP LSP setup as already described for RSVP-TE. It would be desirable to have tools and protocols that automatically compute and build optimal aggregated trees. The promising way toward this is the enhancement of multicast routing protocols, since shared tree is a well-known method used by protocols such as CBT, PIM-SM, PIM-SSM, and PIM-BIDIR. However, creating TE methods to provide aggregation of VPNs with these protocols is an open issue nowadays.

Multicast Traffic Engineering

Although the computation of a multicast gain, in terms of bandwidth, as well as cost, in terms of per-node state savings, is a well-known problem in packet networks, only some recent works such as [11] study the VPN multiplying factor vs. the aggregation dividing factor.

Here we introduce a simple model for the estimation of this state-bandwidth gain caused by multicast VPN aggregation in an SP network, based on just a few network parameters. Without loss of generality, we shall constrain the analysis of the impact of aggregation by the least state consuming approach, aggregate inclusive shared trees. Here, all multicast traffic is delivered to all PEs that support sites belonging to the aggregated set of VPNs and share the P2MP tree. For this to happen, PEs must send traffic to the root as previously explained. Inclusive implies that the tree is not group-membership-aware, which implicitly means broadcast emulation down to a set of PEs. As already mentioned, our focus is on isolating the effect of aggregation of multiple VPN trees. Therefore, the method yields an estimate of bandwidth wasted during broadcast traffic delivery resulting from aggregation. The wasted bandwidth is caused by multicast or broadcast traffic of a VPN that reaches a PE with no site belonging to that VPN or a site with no group members there. Regardless of this, bandwidth is still saved in the core network.

We assume that intelligent aggregation has

been performed in the core (at the P level). This assumption is based on the observation of the physical setup of links and nodes of a typical VPN SP, as described in Fig. 2. SPs usually have one or two P nodes per metropolitan area network (MAN) and tens of PEs located in districts within the metropolitan area (often in local exchange premises). The idea is that, since wide area network (WAN) links are more expensive to maintain, the primary aggregation criterion should dictate to macroscopically bundle VPNs that are in the best match sets of cities. Then, in order to obtain conservative results and be less dependent on a concrete topology and VPN site distribution, we have assumed that no intelligent allocation of VPNs to aggregates is performed at PE level. In other words, VPN sites are randomly allocated to PEs, and VPNs are randomly allocated to one of the manually arranged aggregates according to a uniform distribution. This yields an overestimation that can be considered an upper bound for the required resources in a given scenario.

Although the actual topology of the network under study has an impact on the state-bandwidth trade-off and the resulting aggregation gain, a generic topology has to be studied for better presentation of the methodology first. Individual case studies can be elaborated for specific topologies with the use of the methodology presented in the article, and possibly with refinements or constraints specific for a particular setup. Hence, the proposed analysis of bandwidth consumption and forwarding state can be based on a sample topology shown in Fig. 3. Larger networks would expand to the right and left by increasing the degree of core nodes and replication. The method is universal, and can be adopted to study other topologies as well.

In Fig. 3 the unused redundant links, such as backup links between the PEs and the core network, are not depicted. The figure represents a single tree shared by several VPNs in which only VPNs A and B are shown. Hence, all the PEs send VPN traffic to the root via a P2P LSP that is replicated over the tree. Horizontal links represent direct level local connectivity which is used for analysis of the unicast traffic distribution. The network has a backbone of WAN core routers, which enhance the coverage by means of distribution routers (DRs) that provide service to a set of PE routers. According to a real MPLS network cited in [12], WAN connectivity and traffic distribution are usually performed by the same node, except in large metropolitan areas. Nevertheless, in order to keep the reference network homogeneous, DRs are present in all subtrees, as either WAN core routers or PEs. Finally, the CE routers or switches are served by PEs over point-to-point links.

In the example, the network is supposed to have N DRs. Each DR supports L PEs. K VPNs are being served, with an average of M ($M < L$) sites per DR. Each site of every VPN is connected to a PE.

At this point it is necessary to decide how multicast forwarding entries are accounted for and how to measure used bandwidth for both a distribution based on unicast and multicast/broadcast.

Accounting for Forwarding State

The way multipoint forwarding entries have to be accounted for and compared to current routers' capacity is not straightforward because the available scalability studies provide figures for P2P TE LSPs [13], not for P2MP. It is not realistic to account for a multipoint forwarding entry n times with the same cost as a unicast one. It requires more switching resources and almost n times faster memory. In this work a conservative cost equivalence of one P2MP forwarding entry to n P2P entries is adopted.

As an example, according to [13] practical limits for TE P2P LSPs on a GSR Cisco-family router are 600 MPLS tunnel headends, up to 10,000 tunnel midpoints, with up to 5000 tailends per interface.

Therefore, the number of headends is the first direct limiting factor on the amount of shared trees that can originate from a router. However, as shown later, tunnel midpoint limits may also be reached at DRs for high-density VPNs. Although this technological limitation can be solved by adding more DRs to the network, it is important for proper network planning to understand the way midpoint forwarding state behaves.

In particular, state reduction due to VPN aggregation can reduce the cost of investment in the additional router.

Bandwidth Savings

The overall bandwidth consumption can be computed analytically and compared to the P2P case. Assuming that the multicast and broadcast traffic of all VPNs is the same and normalized to 1, the bandwidth consumed by unicast LSPs by the topology in Fig. 3 to support the multicast and broadcast traffic of the VPNs is given by

$$BW_{LSP} = K \cdot \left(N \cdot M \cdot 2 + M \sum_{i=1}^{\log_2(N)} 2^{i-1} \cdot (2i-1) \right). \quad (1)$$

Equation 1 reflects the cost of replication of a single packet originated at one of the leaves in the topology in Fig. 3 and sending it to all other leaves across the P2P VPNs. The first addend in Eq. 1 is the cost of distribution over the DR-PE links. The second addend is the cost of reaching DRs.

On the other hand, if a P2MP tree is used to distribute the multipoint traffic for each VPN, the total bandwidth consumption is

$$BW_{mLSP} = K \cdot \left(1 + \log_2(N) + \sum_{i=1}^{\log_2(N)} 2^i + M \cdot N \right). \quad (2)$$

The addend $1 + log_2 (N)$ is the cost of sending the packet from the PE to the root of the P2MP tree. The addend $\Sigma_{i=1}^{log2(N)} 2i$ is the cost of distributing the traffic from the root to the DRs. Finally, the addend $M * N$ is the cost of distributing the traffic along all the PEs that have a VPN site attached.

As VPN aggregation is performed, the only changing summand in Eq. 2 is the last one, which becomes $Z * N$, where $Z (Z \geq M)$ stands for the new average number of leaves that depend on a DR in a P2MP LSP after allocating uniformly the K VPN trees into W shared trees ($W \leq K$). The increase of bandwidth usage comes from the fact that when VPN A has no site in a PE where another VPN B sharing the tree has a site, A traffic over that DR-PE link is also considered

Although the actual topology of the network under study has an impact on the state-bandwidth tradeoff and on the resulting aggregation gain, a generic topology has to be studied for a better presentation of the methodology first.

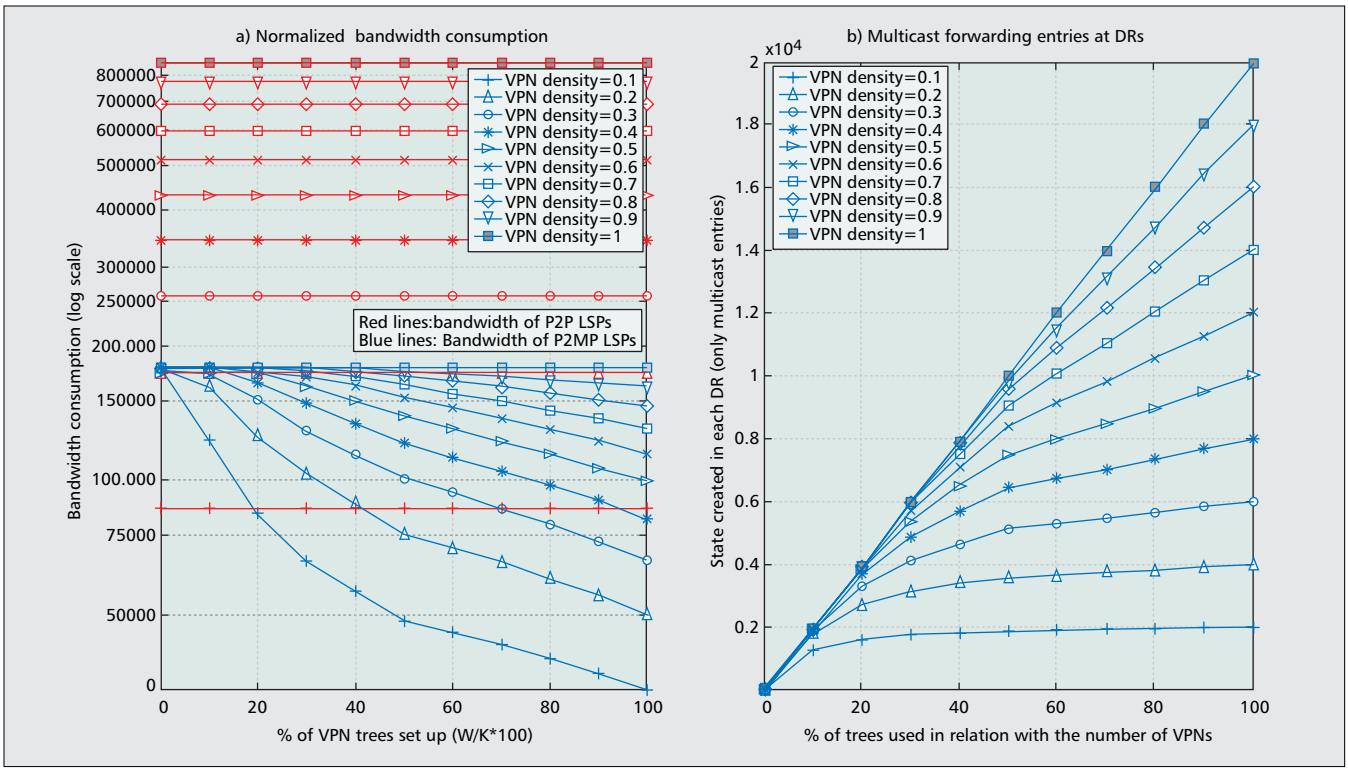

Figure 4. *Bandwidth and forwarding entries used in a multicast-enabled MPLS network.*

useless. In this situation DRs support $W * Z$ forwarding entries. It must be noted that $M \le Z \le L$.

Simulation Results

In order to estimate the benefits of deploying a shared tree, a simulation was run on a network supporting 1000 VPNs. The number of DR leaves was $N = 8$, and there were 20 PEs per DR. VPNs were allocated randomly to the pre-established shared trees, and a number of VPN sites were uniformly assigned to PEs. Since this latter variable is quite relevant, an additional parameter called *density of sites* was introduced to denote the fraction of PEs that have at least one site of a given VPN attached. It is assumed that all VPNs have the same density. The value of W ranged from 1 (a single shared tree for all the VPNs) to 1000 (a P2MP tree per VPN). On each iteration, the value of Z is calculated by randomly grouping different VPNs to a shared tree and obtaining all the PEs that belong to any of the grouped VPNs in a DR. Matlab 7.0 was used to perform this simple task.

The results are shown in Fig. 4. In Fig. 4a, bandwidth consumed by shared trees is plotted in blue for different tree sharing rates of the served VPNs, showing the values for different density of sites. It can be observed the way VPN aggregation increases bandwidth consumption with respect to the reference value defined when a single tree per VPN is used.

For the same number of shared trees, better performance is obtained when the density of sites is higher because the probability of being attached to the same PEs increases. Furthermore, if the density of sites grows, the benefit of including new shared trees is less significant because with a few trees the consumed bandwidth is near the optimal value.

The amount of bandwidth consumed by P2P LSPs for various densities of sites is shown in red in Fig. 4a. The intersection point between the bandwidth consumed by P2P LSPs (red lines) and by P2MP trees (blue lines) of the same density of sites gives the minimum percentage of trees that makes P2MP LSPs more effective than P2P LSPs. For instance, for a density of sites = 0.1, the amount of shared trees should be at least 20 percent of the number of VPNs. Thus, 100 percent means that there are no sharing trees (i.e., there is one tree per VPN).

The cost in terms in forwarding entries, estimated as previously discussed, at DRs is shown in Fig. 4b. The graphic shows that for high densities, the state space savings obtained due to aggregation is larger than at low densities. Furthermore, regarding absolute values, it can be seen that forwarding states at DRs should be checked carefully when designing multicast-supporting MPLS VPNs. Anyway, we recall that the aggregation of VPNs was performed randomly. A more intelligent aggregation of VPNs should provide better results in saved bandwidth and required forwarding entries. The graphic shows almost a linear growth of the cost for high densities, but also important savings in bandwidth even with fairly small densities. As a consequence, it can be concluded that the deployment of aggregated multicast in VPNs is effective in a wide range of conditions even if the level of aggregation imposed by the constrained nodes' memory is high.

Conclusions and Future Work

This article has reviewed the current practical approaches to a scalable implementation of multi-

cast VPN service over an IP-MPLS multiservice network using P2MP trees. These proposals are based on the aggregation of traffic into shared trees to manage the forwarding state vs. bandwidth saving trade-off when P2MP trees are used in MPLS.

The work presented also illustrates the nature of this problem and shows how MPLS traffic engineering LSP trees can be used to alleviate it. The amount of wasted bandwidth depends heavily on the specific scenario and on intelligent assignment of VPNs to shared trees for a target distribution of VPN sites over provider edges. The proposed methodology allows us to study the effect of aggregation by random shared tree allocation on a representative exemplary VPN network model that should maintain most characteristics of a production network. This analysis provides a practical upper bound on the cost of P2MP MPLS trees and implied bandwidth savings at different tree sharing rates for a target network size. This conservative estimate of resources allows us to quantify the effect of aggregation and understand its behavior at different densities in order to provide guidance for MPLS VPN network design.

Many issues remain open. Today the allocation of a VPN to a shared tree is performed by hand by setting the same root for VPN multicast trees that hold many common leaves (PEs). Intelligent engineering of multicast traffic is required to automatically perform this process, and the next few years will probably provide research results on methods and software tools.

Finally, adaptation to the implementation-specific features of future optical multipoint-capable optical switches and, after that, to the delivery of multipoint labeled optical burst switching capabilities is another area of future research. The fraction of multipoint-capable optical switches in a network and wavelength assignment constraints to construct light trees will drive the distribution of VPNs bundles to trees. This long-term issue is one of the research topics under study within the IST e-Photon/One project.

The fraction of multipoint-capable optical switches in a network and wavelength assignment constraints to construct light trees will drive the distribution of VPNs bundles to trees. This long-term issue is one of the research topics under study within the IST e-Photon/One project.

ACKNOWLEDGEMENTS

We would like to acknowledge the anonymous reviewers for their insightful comments. This work has been partially supported by the European Union under the IST e-Photon/One+ (FP6-IST-027497) project and by the CAPITAL project (Spanish MEC, TEC2004-05622-C04-03).

REFERENCES

[1] J. Sanchez, P. Manzanares, and J. Malgosa, "Spanish Telco Strategies Facing New Integrated Digital Transmission Advances," Global Commun. Newsletter, *IEEE Commun. Mag.*, vol. 44, no. 2, Feb. 2006.

[2] S. Yasukawa, "Signaling Requirements for Point-to-Multipoint Traffic-Engineered MPLS Label Switched Paths (LSPs)," RFC 4461, Apr. 2006; http://www.ietf.org/rfc/rfc4461.txt

[3] I. Minei *et al.*, "Label Distribution Protocol Extensions for Point-to-Multipoint and Multipoint-to-Multipoint Label Switched Paths," internet draft, draft-ietf- mpls-ldp-p2mp-02.txt, Dec. 2006, work in progress.

[4] R. Aggarwal, D. Papadimitriou, and S. Yasukawa, "Extensions to Resource Reservation Protocol — Traffic Engineering (RSVPTE) for Point-to-Multipoint TE Label Switched Paths (LSPs)," RFC 4875, May 2007; http://www.ietf.org/rfc/rfc4875.txt

[5] L. Berger, "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Resource ReserVation Protocol-Traffic Engineering (RSVPTE) Extensions," RFC 3473, Jan. 2003, updated by RFCs 4003, 4201, 4420; http://www.ietf.org/rfc/rfc3473.txt

[6] R. Aggarwal, Y. Rekhter, and E. Rosen, "MPLS Upstream Label Assignment and Context Specific Label Space," internet draft draft-ietf-mpls-upstream-label-02.txt, Mar. 2007, work in progress.

[7] R. Aggarwal and J. L. L. Roux, "MPLS Upstream Label Assignment for RSVP-TE," Internet draft draft-ietf-mpls-rsvp-upstream- 01.txt, Mar. 2007, work in progress.

[8] E. Rosen and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)," RFC 4364, Feb. 2006; http://www.ietf.org/rfc/rfc4364.txt

[9] R. Aggarwal, Y. Kamite, and L. Fang, "Multicast in VPLS," Internet draft draft-ietf-l2vpn-vpls-mcast-01.txt, Mar. 2007, work in progress.

[10] E. C. Rosen and R. Aggarwal, "Multicast in MPLS/BGP IP VPNs," Internet draft draft-ietf-l3vpn-2547bis-mcast-04.txt, Apr. 2007, iwork in progress.

[11] G. Apostolopoulos and I. Ciurea, "Reducing the Forwarding State Requirements of Point-to-Multipoint Trees Using MPLS Multicast," *ISCC 2005 Proc.*, June 2005, pp. 713–18.

[12] E. Osborne and A. Simha, *Traffic Engineering with MPLS, ser. Networking Technology*, Cisco Press, 2002; http://www.ciscopress.com/bookstore/product.asp?isbn=1587050315&rl=1

[13] Cisco, "MPLS Traffic Engineering (TE) Scalability Enhancements," 2003; http://www.cisco.com/univercd/cc/td/doc/product/software/ios120/120newft/120limit/120st/120st14/scalable.htm

BIOGRAPHIES

ISAIAS MARTINEZ-YELMO [S '04] (imyelmo@it.uc3m.es) received an M.Sc. in telecommunication engineering in 2003 from University Carlos III de Madrid (UC3M) and an M.Sc. in telematics in 2007 from University Carlos III de Madrid and University Politecnica de Catalua, both in Spain. He is a research and teaching assistant in telematics engineering and Ph.D. student in telematics at University Carlos III de Madrid since 2004. His research activities are focused on NGN networks and peer-to-peer overlay networks.

DAVID LARRABEITI [M '96] (dlarra@it.uc3m.es) is a full professor of switching and networking architectures at UC3M. He got his M.Sc. and Ph.D. in telecommunications engineering from University Politecnica de Madrid in 1991 and 1996, respectively. From 1998 to 2006 he was an associate professor at UC3M and led a number of international research projects. His research interests include the design of the future Internet infrastructure, ultra-broadband multimedia transport, and traffic engineering over IP-MPLS backbones. At UC3M he is responsible for the e-Photon/ONe+ network of excellence on Optical Networking.

IGNACIO SOTO (isotog@it.uc3m.es) received a telecommunication engineering degree in 1993 and a Ph.D. in telecommunications in 2000, both from the University of Vigo, Spain. He was a research and teaching assistant in telematics engineering at the University of Valladolid from 1993 to 1999. In 1999 he joined UC3M, where he has been an associate professor since 2001. His research activities focus on mobility support in packet networks and heterogeneous wireless access networks. He has been involved in international and national research projects related to these topics, including the EU IST Moby Dick and Daidalos projects. He has published several papers in technical books, magazines, and conferences, lately in the areas of efficient handover support in IP networks with wireless access, network mobility support, and security in mobility solutions. He has served as a Technical Program Committee member for INFOCOM.

PIOTR PACYNA (pacyna@kt.agh.edu.pl) received an M.Sc. degree in computer sciences in 1995 and a Ph.D. in telecommunications in 2005 from AGH University of Technology, Krakow, Poland. He has spent his sabbatical leaves at CNET France Telecom and UC3M. He has been working as a lecturer in the Department of Telecommunications of AGH University of Science and Technology. His research focuses on next-generation IP networks, mobility, and security. He has been active in several ACTS and IST research projects: BTI (1997–2000) and Moby Dick (2001–2003), and IST Integrated Projects Daidalos (2003–2006) and Daidalos II (2006–), and has authored several research papers.