On the cost of using MPLS for interdomain traffic

Steve Uhlig and Olivier Bonaventure

Institut d'Informatique, University of Namur, Belgium
{obonaventure,suhlig}@info.fundp.ac.be

Abstract. Multi-Protocol Label Switching (MPLS) is currently heavily used inside autonomous systems for traffic engineering and VPN purposes. We study the cost of using MPLS to carry interdomain traffic by analyzing two one day traces from two different ISPs. Our study shows that a hybrid MPLS+IP solution can significantly reduce the number of LSPs and signalling operations by using MPLS for high bandwidth flows and pure IP for low bandwidth flows. However, the burstiness of the interdomain LSPs could be a problem.

1 Introduction

One of the basic assumptions of IP networks such as the Internet is that all IP packets are individually routed through the network based on the addresses contained in the packet header. This assumption is still valid today, but the complexity of per-packet routing coupled with the need to sustain the exponential growth of the Internet in terms of capacity and number of attached networks has led researchers to propose alternative solutions where per-packet routing on each intermediate hop is not always required. The first alternative solutions such as IP switching [NML98] and others tried to reduce the complexity of the packet forwarding operation by leveraging on the available ATM switches and establishing short-cut virtual circuits for IP flows. Later on, the IETF decided to standardize one IP switching solution under the name Multi-Protocol Label Switching (MPLS).

Although IP switching was initially proposed as a solution to increase the performance of core routers by using label-swapping instead of traditional IP routing in the core, this is not its only benefit. The main benefit of MPLS today is that it allows a complete decoupling between the routing and forwarding functions. With traditional IP routing, each router has to individually route and forward each packet. A consequence of this is that a packet usually follows the shortest path inside a single domain. With MPLS, IP packets are carried inside Label Switched Paths (LSPs). These LSPs are routed at LSP establishment time and the core routers forward the packets carried by the LSPs only based on their labels. This decoupling between forwarding and routing allows MPLS to efficiently support traffic engineering inside autonomous systems as well as transparent VPN services.

In this paper, we evaluate the cost of extending the utilization of MPLS across interdomain boundaries instead of restricting MPLS inside domains as is usually done today. The remainder of this paper is organized as follows. In section 2 we summarize the advantages of using MPLS across interdomain boundaries. To evaluate the cost of using MPLS in this environment, we collected traces from two different ISPs as described in section 3. We then analyze these traces in section 4 and show that a pure MPLS solution to carry all interdomain traffic would be too costly from a signalling point of view. We then analyze hybrid solutions where a fraction of the interdomain traffic is handled by MPLS while another fraction is handled by traditional IP hop-by-hop routing in sections 5 and 6.

2 Using MPLS at interdomain boundaries

Today, MPLS is considered as a key tool to be used inside (large) autonomous systems. This utilization of MPLS has been supported by a lot of research and development during the last few years. In contrast, the utilization of MPLS for interdomain traffic has not been studied in details. BGP has been modified to distribute MPLS labels and the RSVP-TE and CR-LDP signalling protocols support the establishment of interdomain LSPs. However, MPLS has not to our knowledge already been used to carry operational interdomain traffic.

In today's Internet, the behavior of interdomain traffic is mainly driven by several underlying assumptions of the BGP routing protocol. The first assumption is that once a border router announces an address prefix to a peer, this implies that this prefix is reachable through the border router. This reachability information implies that the border router is ready to accept any rate of IP packets towards the announced prefix. With BGP, the only way for a router to limit the amount of traffic towards a particular prefix is to avoid announcing this prefix to its peers. A second assumption of BGP is that all traffic is best-effort. This assumption was valid when BGP was designed, but will not remain valid in the near future with the deployment of applications such as Voice over IP or multimedia and streaming applications and the increasing needs to provide some QoS guarantees for "better than best-effort" traffic (e.g. Intranet, Extranet or traffic subject to specific service level agreements).

The utilization of MPLS for interdomain traffic could provide two advantages compared with traditional hop-by-hop IP routing. First, the utilization of MPLS will allow a complete decoupling between the routing and the forwarding functions. A border router could use a modified[1] version of BGP to announce a MPLS-reachability for external prefixes. This MPLS-reachability means that a peer could send traffic towards these announced prefixes provided that a LSP is first established to carry this traffic. At LSP establishment time, the border router will use connection admission control to decide whether the new LSP can be accepted inside its domain or not.

[1] The extensions to BGP required to support MPLS-reachability at interdomain boundaries are outside the scope of this paper. In this paper, we simply assume that some mechanism exists to announce routes reachable through MPLS.

A second advantage is that it would be easy to associate QoS guarantees to interdomain LSPs. These guarantees would be a first step towards the extension of traffic engineering across interdomain boundaries and could also be a way of providing end-to-end QoS by using guaranteed LSPs as a kind of virtual leased lines across domains.

3 Measurement environment

To better characterize the flows that cross interdomain boundaries, we considered traffic traces from two different ISPs. By basing our analysis on two different networks, we reduce the possibility of measurement biases that could have been caused by a particular network. The two ISPs had different types of customers and were both multi-homed.

3.1 The studied ISPs

The first ISP, WIN (`http://www.win.be`), was at the time of our measurements a new ISP offering mainly dialup access to home users in the southern part of Belgium. We call this ISP the "dialup" ISP in the remainder of this paper. When we performed our measurements, the dialup ISP was connected through E1 links to two different transit ISPs and at the Belgian national interconnection point, having peering agreement with about ten ISPs there.

The second ISP, Belnet(`http://www.belnet.be`), provides access to the commodity Internet as well as access to high speed European research networks to universities, government and research institutions in Belgium. We call this ISP the "research" ISP in the remainder of this paper. Its national network is based on a 34 Mbps backbone linking major Belgian universities. The research ISP differs from the dialup ISP in several aspects. First, the "customers" of the research ISP are mainly researchers or students with direct high speed connections to the 34 Mbps backbone, although some institutions also provide dialup service to their users. Second, the research ISP is connected to a few tens of external networks with high bandwidth links. It maintains high bandwidth peerings with two transit ISPs, the Dutch SURFNET network and is part of the TEN-155 European research network. In addition, the research ISP is present with high bandwidth links at the Belgian and Dutch national interconnection points with a total of about 40 peering agreements in operation.

3.2 Collection of traffic traces

To gather interdomain traffic traces, we relied on the `Netflow` [Cis99] measurement tool supported on the border routers of the two ISPs. Netflow provides a record at the layer-4 flow level. For a TCP connection, Netflow will record the timestamp of the connection establishment and connection release packets as well as the amount of traffic transmitted during the connection. For UDP flows, Netflow records the timestamp of the first UDP packet for a given flow,

the amount of traffic and relies on a timeout for the ending time of a UDP flow. The Netflow traces are less precise than the packet level traces used in many measurement papers [TMW97,NML98,FRC98,NEKN99] since Netflow does not provide information about the arrival time and the size of individual packets inside a layer 4 flow. However, `Netflow` allows us to gather day long traces corresponding to several physical links. Such a traffic capture would be difficult with per-link packet capture tools. The `Netflow` traces were collected at the border routers of the ISPs for unicast traffic and stored with a one-minute granularity. We recorded `Netflow` traces for the incoming traffic of the dialup ISP and incoming and outgoing traffic for the research ISP. Multicast traffic is not included in the traces we consider in this paper. The trace of the dialup ISP was collected in September 1999 while the trace for the research ISP was collected in December 1999.

The utilization of `Netflow` forces us to approximate the layer-4 flows as equivalent to fluid flows. More precisely, a flow transmitting M bytes between T_{start} and T_{stop} is modeled as a fluid flow transmitting $M/(T_{stop} - T_{start})$ bytes every second between T_{start} and T_{stop}. This approximation obviously leads to an incorrect estimation of the burstiness of the traffic and it can be expected that the utilization of `Netflow` underestimates the burstiness of interdomain flows.

3.3 Daily traffic evolution

The first noticeable difference between the two ISPs is the total amount of traffic carried by each ISP. The total amount of daily incoming traffic for the dialup ISP is about 37 GBytes. The research ISP received 390 GBytes during the studied day and transmitted 158 GBytes during the same period. The research ISP receives thus ten times more traffic than the dialup ISP. A closer look at the traffic of the research ISP shows that this traffic is mainly driven by TCP. For this ISP, 97.5 % of the incoming traffic in volume was composed of TCP packets. For the outgoing traffic, 95.8 % of the total volume was composed of TCP traffic. This prevalence of TCP is similar to the findings of previous studies [TMW97]. This implies that UDP-based multimedia applications do not seem to be yet an important source of unicast traffic, even in a high bandwidth network such as the research ISP.

A second difference between the two ISPs is the daily evolution of the interdomain traffic. Figure 1 (left) shows that for the dialup ISP the peak hours are mainly during the evening while for the research ISP the peak hours are clearly the working hours (figure 1 (right)). For the dialup ISP, the links to the transit ISPs are congested during peak hours. For the research ISP, the links to the two transit ISPs are congested during peak hours, but not the links towards the interconnection points and the research networks with which the research ISP peers. For the research ISP, there is a clear asymmetry between the incoming and the outgoing traffic. The amount of incoming traffic is more than four times higher than the amount of outgoing traffic during peak hours. Outside peak hours, the amounts of incoming and outgoing traffic for the research ISP are similar. A similar asymmetry exists for the dialup ISP.

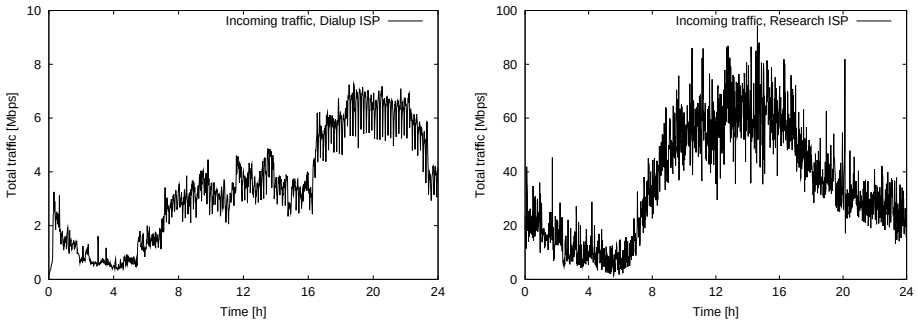

Fig. 1. Daily traffic evolution for dialup (left) and research (right) ISP

In the remainder of this paper, we focus our analysis on the incoming traffic since this is the predominant traffic for both ISPs.

4 Cost of a pure MPLS solution

Replacing traditional IP routing at border routers by MPLS switching would clearly have several implications on the performance of the border MPLSswitches. MPLS could in theory be used with a topology driven or a traffic driven LSP establishment technique. By considering one LSP per network prefix, a topology driven solution would require each autonomous system to maintain one LSP towards each of the about 70000 prefixes announced on the Internet. Such a pure topology-driven LSP establishment technique would imply the creation and the maintenance of 70000 LSPs by each autonomous system. This number of LSPs is clearly excessive.

To reduce the number of interdomain LSPs, we evaluate in this paper the possibility of using traffic-driven LSPs, i.e. LSPs that are dynamically established when there is some traffic towards some prefixes and released during idle periods. More precisely, we consider the very simple LSP establishment technique described in figure 2. In this section, we assume that *trigger* is equal to 1 byte, i.e. all IP traffic is switched.

To evaluate the cost of using MPLS for interdomain traffic, we have to consider not only the number of established LSPs, but also the number of signalling operations (i.e. LSP establishment and release). When considering interdomain traffic, the cost of using MPLS is not simply the CPU processing cost of the signalling messages by each intermediate router. We do not except that this cost would be the bottleneck. When an interdomain LSP is established, it will typically pass through several autonomous systems. When a border router will receive a LSP establishment request, it will have to verify whether the LSP can be accepted given the network policies, the current utilization of autonomous system links as well as authentication, billing and accounting issues. The handling of all these issues might be expensive.

```
After each one minute period and for each prefix p:
// Traffic(p) is traffic from prefix p during last minute
if (Traffic(p)≥ trigger)
{
    if (LSP(p) was established)
        LSP(p) is maintained;
    else
        Establish LSP(p);
}
else
{
    if (LSP(p) was established)
        LSP(p) is released;
}
```

Fig. 2. Simple LSP establishment technique

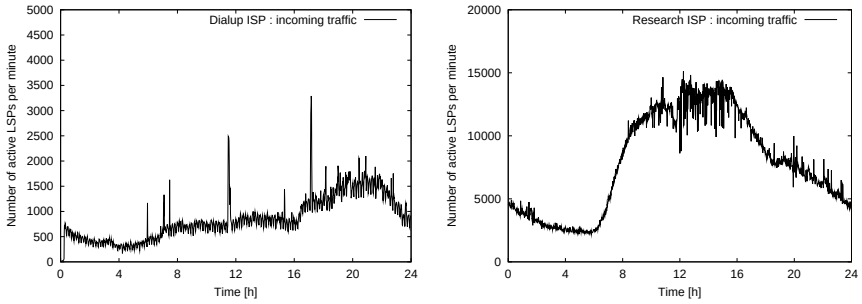

Fig. 3. Number of active LSPs for dialup (left) and research (right) ISP

Figure 3 compares the total number of LSPs that the border router of the ISP needs to maintain for the dialup (left) and the research ISP (right). In this figure, we only consider the incoming traffic as mentioned previously. This figure shows two interesting results. First, as expected, the number of LSPs follows the daily evolution of the traffic. Both ISPs need to maintain a larger number of LSPs during peak hours than during the night. Second, the research ISP with about ten times more traffic than the dialup ISP needs to maintain about ten times more interdomain LSPs than the dialup ISP. This means that with more capacity the research ISP communicates with a larger number of network prefixes than the dialup ISP rather than receiving more traffic from the same number of network prefixes. While the absolute number of LSPs stays in the range of 1000-2000 for the dialup ISP, the research ISP would require more than 10000 simultaneous LSPs during peak hours in order to switch every packet on a LSP. This number, given the cost of establishing interdomain LSPs, might be too high.

The second performance factor to be considered is the number of signalling operations. Figure 4 shows the number of per-minute signalling operations (LSP

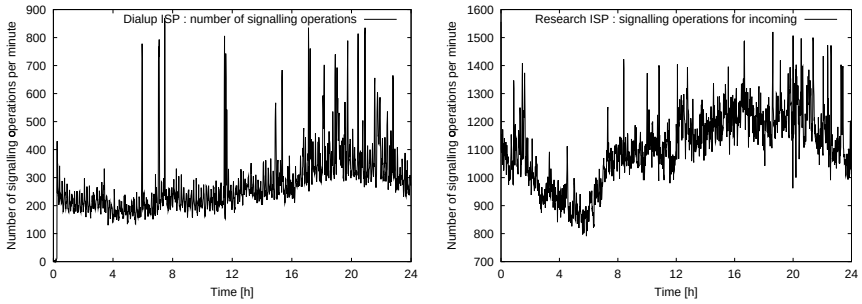

Fig. 4. Signalling overhead for dialup (left) and research (right) ISP

establishment and LSP release) for the dialup (left) and the research (right) ISP. For the dialup ISP, several hundreds of LSPs need to be established or released every minute. This is an important number compared to the 1000-2000 LSPs that are maintained by this ISP. For the research ISP, about 1000 signalling operations need to be performed every minute. This implies that during peak hours, 10 % of the LSPs are modified during each one minute interval. For both ISPs, the number of signalling operations would probably preclude the deployment of a pure MPLS solution to carry all interdomain traffic.

5 Reducing the number of LSPs

The previous section has shown the high cost of a pure MPLS solution to handle interdomain traffic. A pure MPLS solution is probably too costly from the signalling point of view, even for the relatively small ISPs that we considered in this study. To allow the efficient utilization of MPLS for interdomain traffic, we clearly need to reduce the number of interdomain LSPs as well as the number of signalling operations. This could be done in two different ways.

A first way would be to aggregate traffic from several network prefixes inside a single LSP. This would require a close cooperation with the routing protocol to determine which network prefixes can be aggregated inside each LSP. A potential solution was proposed in [PHS00]. Space limitations prohibit us to discuss this issue further in this paper.

A second way would be to utilize MPLS for high bandwidth flows and normal IP routing for low bandwidth flows. This would allow to benefit from MPLS capabilities to handle the higher bandwidth flows while avoiding the cost of LSPs for low bandwidth flows. This could be coupled with different types of routing for the two types of flows as proposed in [SRS99] where a different type of routing was developed for long-lived flows.

To evaluate whether the interdomain traffic of our two ISPs could be separated in two such classes, we analyzed the total amount of traffic received from each network prefix during the studied day. Figure 5 (left) shows that during this day, the dialup ISP received IP packets from slightly more than 15.000 different

network prefixes (out of about 70000 announced prefixes on the global Internet). However, it also shows that these prefixes are not equal. The most important prefix sends about 3.9% of the total traffic that enters the dialup ISP. The total traffic from the top 100 prefixes seen by the dialup ISP corresponds to 50% of the daily incoming traffic and 560 (resp. 1800) prefixes are required to capture 80% (resp. 95 %) of the daily incoming traffic.

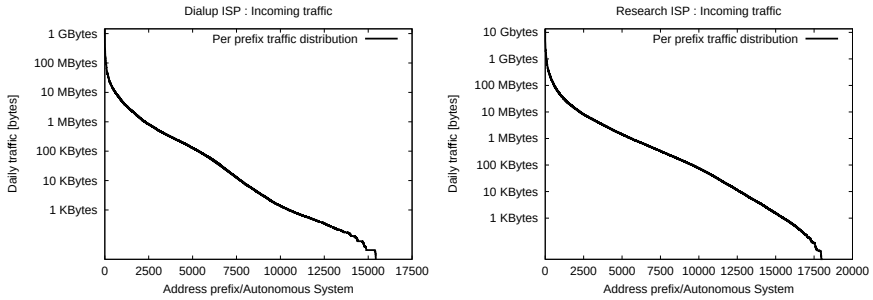

Fig. 5. Per prefix daily traffic distribution for dialup (left) and research (right) ISP

A similar trends exists for the research ISP as shown in figure 5 (right). The research ISP received IP packets from 18.000 different network prefixes during the studied day. For this ISP, the most important prefix sends 3.5% of the daily traffic. The total traffic from the top 100 prefixes seen by this ISP corresponds to 49.5 % of the daily traffic. Furthermore, the top 500 (resp. 1820) prefixes transmit 80 % (resp. 95 %) of the daily traffic towards the research ISP.

Based on this analysis, it seems possible to capture a large portion of the traffic by only considering the prefixes that transmit a large amount of data or the high bandwidth flows. The separation of the traffic into two different classes must be performed online at the border routers. For this, very simple techniques are required.

6 Cost of a hybrid MPLS+IP solution

As a simple mechanism to segregate the traffic between high bandwidth and low bandwidth flows, we consider the procedure described in figure 2 with a large trigger. This means that a LSP is maintained if we saw at least *trigger* bytes for a given prefix during the last minute. A LSP is released if we saw less than *trigger* bytes for a given prefix during the last minute. We assume that a LSP can be instantaneously established and thus seeing more than *trigger* bytes for a minute suffices to consider a LSP (dedicated to that prefix) as active (established or maintained) for the whole minute. This scheme is very simple, since only the last minute is considered in order to decide the action to perform on the LSP.

Other LSP establishment schemes are not considered in this paper due to space limitations.

By using this trigger-based mechanism, we use MPLS LSPs for high bandwidth flows and normal IP routing for low bandwidth flows. Figure 6 shows for both ISPs the amount of traffic captured by the high bandwidth flows as a function of the trigger and the number of signalling operations that are required to establish and release these high bandwidth LSPs. The captured traffic is expressed as a percentage of the total daily traffic of the ISP. The number of signalling operations is expressed as a percentage of the required signalling operations when all the traffic is switched by MPLS (i.e. $trigger$=1)

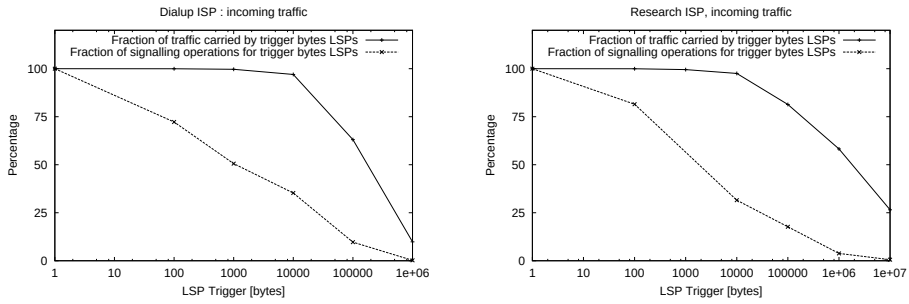

Fig. 6. Impact of LSP trigger for dialup (left) and research (right) ISP

Figure 6 (left) shows that for the dialup ISP if we only use MPLS for the prefixes that transmit at least 10 KBytes per minute, we still capture 97% of the total traffic while we reduce the number of signalling operations by a factor of 3. If we use MPLS for the prefixes that transmit at least 1 MBytes per minute, we only capture 10 % of the daily traffic. A similar situation holds for the research ISP. In this case, figure 6 (right) shows that if we used MPLS for prefixes that transmit at least 1 MByte per minute, then we still capture 58 % of the daily traffic and we reduce the number of signalling operations by a factor of 25 compared to a pure MPLS solution. Figure 6 shows clearly that using MPLS only for high bandwidth flows allows to reduce the number of signalling operations while still maintaining a good capture ratio.

Based on figure 6, a trigger between 10 and 100 KBytes (resp. 10 KBytes) for the research ISP (resp. dialup ISP) would be a reasonable compromise between the amount of traffic captured and the number of signalling operations. However, the number of signalling operations and the percentage of the captured traffic are not the only performance factors that we need to take into account. Another performance factor is the lifetime of the interdomain LSPs. Ideally, such a LSP should last for a long period of time so that the cost of establishing this LSP can be amortized over a long period. If the LSPs only last for a few minutes, then it would be difficult to dynamically establish high bandwidth interdomain LSPs.

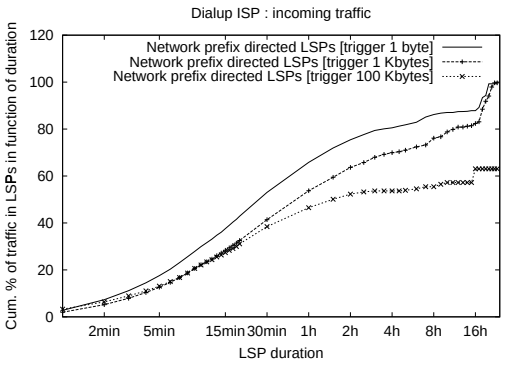

Fig. 7. LSP lifetime for dialup ISP

To evaluate the duration of these LSPs, we plot in figures 7 and 8 the cumulative percentage of the traffic that is carried by LSPs lasting at least x minutes. Figure 7 considers the cumulative amount of traffic that is carried by the LSPs as a function of their lifetime for the dialup ISP. This figure shows that if we consider a pure MPLS solution (trigger=1 byte), 17.5 % of the total traffic is carried by LSPs that remain established for up to five minutes. Thus, LSPs lasting more than five minutes capture more than 82.5 % of the total traffic. Five minutes is probably too short a duration for interdomain LSPs. If we now consider the LSPs that last for at least 30 minutes, they capture 47 % of the total traffic. However, when we utilize MPLS only for high bandwidth flows, the lifetime of the LSPs decreases. For example, if we consider the 100 KBytes LSPs, these LSPs only capture 63 % of the total traffic. Within the 100 KBytes LSPs, the LSPs that last for at least 30 minutes capture only 15 % of the daily traffic of the dialup ISP.

Figure 8 shows that the behavior of the research ISP is slightly different. If we consider a pure MPLS solution, then 18.3 % of the daily traffic is captured by LSPs that last up to five minutes. If we consider the LSPs that remain active for at least 30 consecutive minutes, these LSPs capture 65.5 % of the daily traffic. These values are better than for the dialup ISP. However, if we now consider the high bandwidth LSPs, we see an important decrease in the lifetime of these LSPs. If we consider the LSPs that transmit at least 10 KBytes per minute (a rather low bandwidth flow for the research ISP), they capture 97.5 % of the daily traffic. However, the 10 KBytes LSPs that last for at least 30 minutes only capture 38.2 % of the daily traffic. The situation is even worse when we consider the LSPs that carry at least 1 MByte per minute. All these LSPs carry 58 % of the daily traffic. However, among these high bandwidth LSPs, the LSPs that last only a single minute carry 36 % of the daily traffic of the research ISP. The high bandwidth LSPs that have a duration longer that five minutes carry only 3.6 % of the daily traffic and there are almost no LSP that remains active for more than 30 minutes.

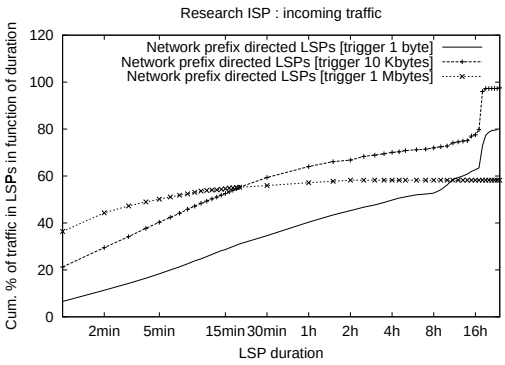

Fig. 8. LSP lifetime for research ISP

The difference between the evolution of the LSP lifetime with the bandwidth of the LSP for the two ISPs can probably be explained by two factors. The first factor is the congestion level. Most of the incoming traffic of the dialup ISP is received through its two heavily congested transit ISP links. On the other hand, the external links of the research ISP, especially those towards the research networks and the interconnection points, are only lightly congested. The second factor is the maximum bandwidth that a user can consume. A user of the dialup ISP is limited by its dialup modem while a user of the research ISP may easily receive traffic at several Mbps.

The burstiness of interdomain traffic of the research ISP implies that it would be difficult to utilize guaranteed bandwidth interdomain LSPs to optimize the traffic of the research ISP. A closer look at the behavior of these LSPs shows that it is difficult to predict the bandwidth that one LSP would need for the upcoming minute. For the research ISP, the solutions proposed in [DGG+99] are not applicable. Either the reserved bandwidth is much smaller than the traffic carried by the LSP or the reservation is much larger than the actual traffic. In both cases, the utilization of guaranteed bandwidth interdomain LSPs does not seem to be a good solution to perform interdomain traffic engineering for our research ISP. This is due to the current nature of the best-effort traffic and the large capacity of our research ISP. The situation would probably change with the deployment of differentiated services and the utilization of traffic conditioners such as shapers. QoS sensitive applications such as multimedia, streaming or voice over IP would behave differently from the best-effort applications we found in our two ISPs.

7 Conclusion

In this paper, we have analyzed the cost of using MPLS to carry interdomain traffic. Our analysis was carried out by studying full day traffic traces for two different ISP. We have shown that utilizing exclusively MPLS to handle all the

interdomain traffic would be too costly when considering the number of LSPs and the number of signalling operations that are required to establish and release dynamically such LSPs.

We have then shown that the cost of MPLS could be significantly reduced by using MPLS for high bandwidth flows and traditional hop-by-hop IP routing for low bandwidth flows. We have evaluated a simple trigger-based mechanism to distinguish between the two types of LSPs. The utilization of such a mechanism can significantly reduce the number of signalling operations and the number of LSPs. The optimal value for the trigger depends on the total bandwidth of the ISP. However, we have also shown that the burstiness of the interdomain LSPs could be a significant burden concerning the utilization of MPLS to perform interdomain traffic engineering with guaranteed bandwidth LSPs.

Acknowledgements

This paper would not have been written without the traffic traces provided by Belnet and WIN. We especially thanks Rudy Van Gaver, Damien Collart and Marc Roger for their help.

References

[Cis99] Cisco. NetFlow services and applications. White paper, available from http://www.cisco.com/warp/public/732/netflow, 1999.

[DGG+99] N. Duffield, P. Goyal, A. Greenberg, P. Mishra, K. Ramakrishnan, and J. van der Merwe. A flexible model for resource management in Virtual Private Networks. In *SIGCOMM1999*, pages 95–108, September 1999.

[FRC98] A. Feldmann, J. Rexford, and R. Caceres. Efficient policies for carrying web traffic over flow-switched networks. *IEEE/ACM Transactions On Networking*, 6(6):673–685, December 1998.

[NEKN99] K. Nagami, H. Esaki, Y. Katsube, and O. Nakamura. Flow aggregated, traffic driven label mapping in label-switching networks. *IEEE Journal on Selected Areas in Communications*, 17(6):1170–1177, June 1999.

[NML98] P. Newman, G. Minshall, and T. Lyon. IP switching - ATM under IP. *IEEE/ACM Transactions On Networking*, 6(2):117–129, April 1998.

[PHS00] P. Pan, E. Hahne, and H. Schulzrinne. BGRP: A framework for scalable resource reservation. Internet draft, draft-pan-bgrp-framework-00.txt, work in progress, January 2000.

[SRS99] A. Shaikh, J. Rexford, and K. Shin. Load-sensitive routing of long-lived IP flows. In *SIGCOMM1999*, pages 215–226, September 1999.

[TMW97] K. Thompson, G. Miller, and R. Wilder. Wide-area internet traffic patterns and characteristics. *IEEE Network Magazine*, 11(6), November/December 1997. also available from http://www.vbns.net/presentations/papers.