How can Multi-Topology Routing be used for Intradomain Traffic Engineering?

Amund Kvalbein and Olav Lysne
Simula Research Laboratory, Oslo, Norway

ABSTRACT

Multi-Topology routing allows each router in a network to maintain several valid routes to each destination. This increases the possibilities to spread traffic towards a destination over multiple paths with connectionless routing protocols like OSPF or IS-IS. In this paper, we report early ideas on how this can be utilized as a Traffic Engineering tool. We look at both offline and online approaches, and argue that a Multi-Topology based solution has advantages over previous solutions in both paradigms.

Categories and Subject Descriptors

C.2.2 [**Computer Systems Organization**]: Network Protocols

General Terms

Algorithms

Keywords

Traffic Engineering, Multi-Topology Routing, Intradomain

1. INTRODUCTION

Most current intradomain Traffic Engineering (TE) methods are *offline* in the sense that a central entity computes a good routing based on some estimate of the demands, and this routing is not changed in response to short term traffic dynamics. In traditional link state protocols like OSPF or IS-IS, this is done by using heuristic methods to find a set of link weights that distributes the traffic well in the network [1, 2].

In networks relying on MPLS-TE [3], multi-commodity flow optimization techniques can be used to set up LSPs with the goal of optimizing some TE objective [4]. However, the use of MPLS in a network introduces the extra complexity of calculating, setting up and maintaining LSPs between each source and destination. In this paper, we focus only on destination-based routing, as opposed to the flow-based routing used in MPLS. Instead of setting up explicit paths between each source and destination, we rely on Multi-Topology routing to expose more of the underlying path diversity while staying in the destination-based routing paradigm.

The main problem with traditional offline TE methods based on link weight tuning is that they rely heavily on the available estimate of the traffic demands. It is hard to deduce a traffic matrix that accurately describes the demands over long time spans, and at the same time captures short term variations [5, 6]. Common TE methods try to optimize the routing for demand matrices that are averages over several days, or even months. However, network traffic is known to be highly dynamic and non-stationary, due to both natural variations in user demands and phenomena like flash crowds, BGP rerouting and failures. Recent inventions like overlay routing [7] and intelligent route control for multi-homed networks [8] make traffic patterns even harder to predict.

Offline TE methods based on link weight optimization are not well positioned to handle dynamic traffic changes. With a changed traffic matrix, we would like to run the optimization heuristic again, and install the new optimized link weights in the network to maintain the desired load balancing properties. However, changing link weights in an operational network is problematic, since such changes will lead to a period of routing instability as the routing protocol converges on the new topology [9]. Adjusting IGP link weights may also change the egress routers that are chosen in the BGP route-selection process, causing additional unwanted traffic shifts [10].

Several proposals have been made to mitigate the effects of traffic demand changes. Some schemes try to find a link weight setting that performs well also in the presence of link failures [11, 12, 13]. Others propose *oblivious* routing, that offer performance within certain boundaries under all possible traffic conditions [14, 15]. While such proposals can give improved performance in many corner-case scenarios, they must to a varying degree pay for this by decreased performance in the normal case.

The shortcomings of offline traffic engineering tools have led to proposals for *online* mechanisms [16, 17]. These methods rely on having several available paths between each source and destination. They use measurement techniques to monitor the quality of each path, and split traffic between them. By dynamically updating the split ratios, these methods can respond rapidly to traffic dynamics, without requiring any information about the demand matrix.

INM'07, August 27–31, 2007, Kyoto, Japan.

In this paper, we propose a new method for IGP traffic engineering with connectionless routing protocols that avoids the problems associated with link weight changes. Our method is based on Multi-Topology (MT) routing, which is currently being defined by the IETF for both OSPF [18] and IS-IS [19]. MT routing allows the routers to maintain several independent logical topologies, so that different types of traffic can be routed independently through the network. We advocate extending the MT mechanisms so that it can be efficiently used to spread traffic over several available paths.

The main idea in our contribution is to construct the set of logical topologies in such a way that any congested link can be avoided in at least one topology. Traffic is then spread over the topologies so that the load is well distributed over the available links. We discuss how this idea can give advantages over existing methods in both an offline and an online TE paradigm. We have previously proposed a solution for achieving fast recovery from component failures based on a similar approach [20], but we believe that the potential benefits of MT routing are even more significant in the context of traffic engineering. This paper mainly describes a set of ideas on how MT routing can be exploited for TE purposes. Some more evaluation results on some of the described methods are available in [21].

The rest of this paper is organized as follows. In Sec. 2, we introduce MT routing and describe the mechanisms that are needed to use it as a TE tool. In Sec. 3, we describe three different approaches for how MT routing can be used to control the traffic flow through the network, and describe some challenges for each approach. Finally, we summarize and conclude in Sec. 5.

2. MULTI-TOPOLOGY ROUTING

Multi-Topology routing allows the routers in an AS to maintain several parallel logical views of the network topology. The routers exchange topology-specific link state advertisements describing the properties of each link. Conceptually, the routers build a separate routing table for each topology. Data traffic is associated with a specific topology, and is routed according to the corresponding routing table.

2.1 Mapping traffic to topologies

An important question is how a router can decide which topology to forward an incoming packet in. The existing MT drafts describe how this can be done if the traffic in the different topologies belong to different address families like IPv4 vs IPv6 or unicast vs multicast. However, to efficiently use MT routing as a TE tool, we need a more generic way to associate a data packet with a topology, where all types of traffic can be routed in all topologies. We propose two possible ways of achieving this:

Explicit packet marking. With this approach, packets are associated with a topology by using bits in the IP header as a topology identifier. This can be done by reserving one or more ToS/DSCP values for each topology. Other possible methods for explicit IP packet marking are described in [22].

Tunnelling. A private address space can be assigned to each topology in a non-overlapping fashion, and each egress router in the network is given an IP address in each topology. Then a data packet can be associated with a topology by tunnelling it to the topology-specific IP address of the egress node of the original destination address.

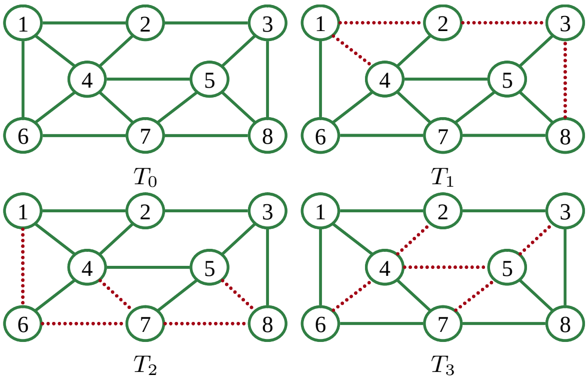

Figure 1: Example topologies. Dotted links are excluded from the routing process in the topology.

Both these methods will allow traffic to all destinations to be routed according to an arbitrary topology. This allows us to perform TE by deciding which traffic is routed according to the different topologies.

2.2 Building alternate topologies for traffic engineering

The next question is how the topologies should be constructed to efficiently facilitate TE. In our approach, we distinguish between the *original* topology where all links can be used for routing, and the *alternate* topologies, where some links are excluded from the routing process. We propose building a set of alternate topologies with two important properties:

1. For all links, there exists a topology where the link is not used for forwarding traffic.

2. All topologies are connected, so that in each topology, there is a valid routing path between each pair of nodes.

Creating the topologies in this way allows us to respond to congestion by moving traffic over to the alternate topology that avoids the congested link. It also gives us large flexibility, since traffic to any destination can be routed in any topology.

Figure 1 shows an example of how three alternate topologies can be built so that all links in the original topology are excluded from the routing process (dotted) in one of them.

We will now describe an algorithm for creating alternate topologies that satisfy the requirements above. The algorithm is taken from [23], where we proposed the use of multiple topologies for fast recovery from link failures. Here, we only describe the rules that control the topology creation.

We define a topology T as a set of nodes N and a set of links L. Given a network topology T_0, we build n additional topologies $T_1, \ldots, T_n$. We refer to T_0 as the original topology, and $T_1, \ldots, T_n$ as the alternate topologies. The alternate topologies are copies of the original topology T_0, with the difference that a subset of the links are removed in each alternate topology. Importantly, links are removed in such a way that each alternate topology $T_1, \ldots, T_n$ is still connected, and thus all nodes are still reachable in all topologies.

Input to our algorithm is the original topology T_0, and the number n of desired alternate topologies. The algorithm then iterates through all links and tries to remove each of them in one of the topologies T_i. A link can only be removed from a topology if doing so does not disconnect the topology. If a link cannot be removed in T_i, we try again in topology $T_{(imodn)+1}$ until all alternate topologies have been tried. For each link we want to remove, a new topology is chosen as the first T_i we try, so that the number of removed links is approximately equal over the different T_i.

We have previously shown that a surprisingly small number n of alternate topologies are needed by this algorithm [23]. Typically, less than five alternate topologies are needed to cover all links in current ISP networks.

3. TRAFFIC ENGINEERING USING MULTI-TOPOLOGY ROUTING

Given a set of topologies created by the algorithm described above, traffic engineering can be performed in several ways. In this section, we describe three fundamentally different possibilities, and discuss some of the advantages and challenges with each of them. Common for all three methods is that they try to avoid overload in the network by assigning traffic to different topologies in an intelligent manner.

3.1 Offline TE

MT routing can be used to perform offline TE based on the available estimate of the traffic demands. With this approach, each data packet is mapped to a topology at the ingress node. Once a data packet has been assigned to a topology, it will be routed shortest path according to that topology to the egress of the network.

Given an estimate of the traffic demands and the set of link weights used in the network, we can calculate the estimated load on each link. We show here a simple heuristic that assigns traffic to topologies on the granularity of ingress-egress flows, with the objective of minimizing the maximum link utilization. Alternatively, a more advanced method that splits traffic in the same ingress-egress flow on several topologies can be imagined.

The heuristic starts by routing all traffic in the original topology T_0. In this topology all links are available for routing, so we should route traffic here when possible. Our heuristic then identifies the most loaded link in the network and the ingress-egress flows that are routed over this link, before it moves one of these ingress-egress flows to the alternate topology where the link is not used for routing. The process is iterated until no further reduction in the maximum link utilization is achieved.

There are several possible ways to select the flow f that is moved to the alternate topology $t(l_{max})$. The easiest way would be to pick it randomly. However, it will often be beneficial to identify a flow of a certain size that should be moved. Selecting a large flow will give a larger reduction in the load on the congested link. By selecting larger flows, fewer iterations are needed in the algorithm.

Algorithm 1: Offline TE

```
repeat
    identify most loaded link $l_{max}$
    identify flows $\mathcal{F}$ that are routed over $l_{max}$
    select flow $f \in \mathcal{F}$
    move $f$ to topology $T(l_{max})$ where $l_{max}$ is not used
until no reduction in max utilization
```

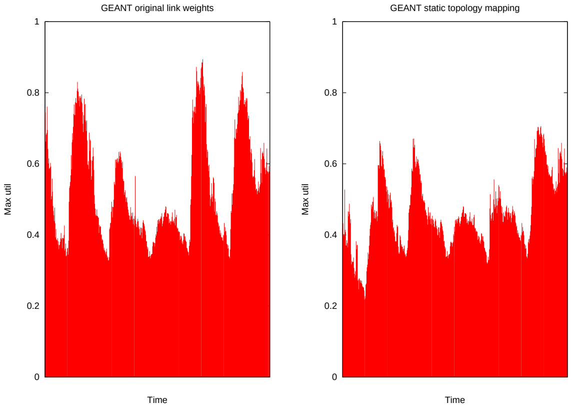

Figure 2: Maximum link utilization in the GEANT network

The main advantage of this MT approach compared to previous offline TE methods based on link weight tuning, is the increased ability to deal with traffic dynamics. MT routing makes it possible to adapt the routing to a changed traffic matrix without requiring a new IGP convergence. Instead, only the mapping from ingress-egress flow to topology needs to be changed. Changing this mapping will not trigger a routing re-convergence.

Figure 2 shows the maximum link utilization at the granularity of 15 minutes in the European research network GEANT over a period of almost one week in 2005. The left plot shows the situation with the original link weights used in the GEANT network. These link weights are also used in the right plot, but here we have used Alg. 1 to spread the traffic over multiple topologies. We used 5 alternate topologies, in addition to the original topology. The flow-to-topology mapping is updated every hour based on the traffic demands experienced over the previous hour. The plots are based on traffic matrices calculated based on measurements from the GEANT network as described in [24].

3.2 Centralized online TE

Another option is to use MT routing in an online fashion, where the routing is dynamically adapted to the changing traffic conditions. The main advantage of this approach is that it does not require any knowledge of the traffic matrix. We first describe how this can be done in a centralized way, based on monitoring the link utilization of all links in the network.

The idea in this approach is again to start by routing all traffic in the original topology T_0, and to move traffic over to alternate topologies as needed. If the load on a link exceeds a selected threshold u_{max}, the network management system can instruct one or more ingress routers to change their flow-to-topology mapping, so that more traffic is routed in the

```
Algorithm 2: Centralized online TE
while true do
    if $u > u_{max}$ then
        move one flow $f \in \mathcal{F}$ from $T_0$ to $T_i$ where
        the link is not used
    end
    if $u < u_{min}$ then
        move one flow $f \in \mathcal{F}$ from $T_i$ back to $T_0$
    end
end
```

topology where the congested link is avoided. When the load on the link later falls below a lower threshold u_{min}, traffic can be moved back to the original topology where all links can be used for routing. Note that once a new mapping has been calculated by a central entity, each ingress router can independently apply the new mapping without the need for any global synchronization.

Alg. 2 shows the procedure that is performed by for each link by the management system at each sampling interval. This approach is simpler than the offline approach described above, since it requires no calculation of a global traffic matrix. Important challenges with this method is to determine u_{min}, u_{max} and the sampling period so that the system is stable yet responsive. It is important that the difference between u_{min} and u_{max} is large enough in order to avoid mapping flows back and forth between topologies.

An attractive aspect of MT-based TE, is that it can easily be combined with methods based on link weight tuning. For example, the link weights in the original topology T_0 can be set so that the chance of congestion is minimized under the expected long-term traffic conditions. Traffic is then only moved to alternate topologies in response to short term deviations from normal conditions. This is the reason why we have designed Alg. 2 so that traffic is moved back to T_0 when possible.

3.3 Distributed online TE

MT routing can also be used to perform traffic engineering in a distributed online fashion. With this approach, each ingress node in the network uses a lightweight probing mechanism to monitor the quality of the path to each egress in each topology. These measurements are then used to control the amount of traffic that is sent in each topology.

Previous methods for online traffic engineering like MATE [16] and TeXCP [17] depend on tools like MPLS-TE [3] for explicitly setting up several paths between each ingress and egress. With MT routing, several paths are available between an ingress and an egress. These paths are not necessarily disjoint, but all hops on the path can be avoided in at least one of the alternate topologies.

An important challenge when doing distributed online TE is to avoid instabilities. If the response to a changed traffic situation is too strong, the result may be an unstable system where flows are constantly moved from one topology to another.

Another open question is how the path quality in the different topologies can best be monitored. Previous online TE methods rely on using a separate probing agent for each path and for each ingress-egress pair. An interesting possibility with MT routing could be to allow packets to switch from one topology to another at intermediate routers. With this approach, each router would only have to monitor the status of its directly connected links. If one of these links are congested, traffic is moved to the topology where the link is avoided. Letting packets change topology in flight can potentially lead to severe packet reordering, with adverse consequences for TCP performance. It has been demonstrated that flow splitting can be performed at the granularity of bursts without severe reordering [25], at the cost of maintaining more state in the network. Packet reordering is an important issue that must be handled for in-flight topology switching to be a viable approach.

4. INTERACTIONS WITH INTERDOMAIN ROUTING

BGP is the current de-facto standard for interdomain routing. When a BGP speaking router learns about several routes to a destination prefix, it will use a given route selection process to pick a single best path. If the first five tie breakers in this selection process are equal for two routes, the route with the shortest IGP path to the egress router will be preferred. This is known as hot-potato routing.

In our methods described above, we have used the commonly made assumption that traffic flows point-to-point from an ingress router to a single egress router. However, since there can be more than one valid egress for a given destination prefix, this is not always true. The hot-potato routing used in BGP implies that the IGP routing will also influence the egress point for a flow.

In a MT context, when we move traffic to an alternate topology, this might affect which egress router is the closer one with respect to IGP weights. For example, consider a situation where a destination prefix p is reachable through two egress routers R_1 and R_2. In the normal topology, an ingress router R_{in} may choose to send traffic towards R_1, since it is the closer with respect to IGP weights in T_0. If a link on the path from T_{in} to T_1 is congested, our method may respond to this by moving the traffic from T_{in} towards p to an alternate topology T_i, where the congested link is avoided. However, in this topology T_i some of the links are excluded from the routing, and hence R_2 might be the preferred egress router. As a consequence, when we take BGP hot-potato routing into account, our TE operations can result in some traffic being shifted from one egress router to another. Importantly, even if traffic is moved from one egress router to another, we still achieve our goal of moving traffic away from the congested link.

ISP policies are typically enforced by giving one route a higher local preference than another. The traffic shifting to an alternative egress router described here will only occur for prefixes where the IGP path length is used to determine the preferred egress router, and hence no policies will be violated.

An interesting extension of our method would be to take advantage of these traffic shifts to do traffic engineering on the edge links. This could be done by moving traffic to a different topology, so that a different egress router will be selected.

5. CONCLUSIONS

In this paper, we have advocated the use of Multi-Topology routing as an intradomain traffic engineering tool in connectionless IP networks. The main advantage of a MT-based approach over previous solutions based on link weight optimization, is the ability to change the routing in response to traffic dynamics without triggering an IGP re-convergence.

We have sketched three different approaches for MT traffic engineering. One uses the available estimate of the traffic demands to spread the traffic among the different topologies. The two other approaches do not rely on an estimate of the demands. Instead, they adapt the flow-to-topology mapping to the observed traffic in an online fashion. An attractive aspect of using MT routing for TE, is that it can be combined with other methods based on link weight tuning.

Many challenges remain before we have a complete TE method based on MT routing. Mechanisms for monitoring path characteristics must be defined, the stability of the methods must be proved, and solutions for reducing packet reordering must be devised. We believe all these problems can be solved, and are currently working on these and related issues.

6. REFERENCES

[1] B. Fortz and M. Thorup, "Internet traffic engineering by optimizing OSPF weights." in *Proceedings INFOCOM*, 2000, pp. 519–528.

[2] A. Sridharan, R. Guerin, and C. Diot, "Achieving near-optimal traffic engineering solutions for current OSPF/IS-IS networks," *IEEE/ACM Transactions on Networking*, vol. 13, no. 2, pp. 234–247, Apr. 2005.

[3] D. Awduche, L. Berger, D. Gan, T. Li, V. Srinivasan, and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP tunnels," RFC3209, dec 2001.

[4] D. Mitra and K. G. Ramakrishnan, "A case study of multiservice, multipriority traffic engineeringdesign for data networks," in *Proceedings GLOBECOM*, Rio de Janeiro, Brazil, Dec. 1999, pp. 1077–1083.

[5] A. Feldmann, A. Greenberg, C. Lund, N. Reingold, J. Rexford, and F. True, "Deriving traffic demands for operational IP networks: methodology and experience," *IEEE/ACM Transactions on Networking*, vol. 9, no. 3, pp. 265–280, June 2001.

[6] A. Soule, A. Lakhina, N. Taft, K. Papagiannaki, K. Salamatian, A. Nucci, M. Crovella, and C. Diot, "Traffic matrices: balancing measurements, inference and modeling," in *Proceedings SIGMETRICS*, Banff, Alberta, Canada, 2005, pp. 362–373.

[7] D. Andersen, H. Balakrishnan, R. Morris, and F. Kaashoek, "Resilient overlay networks," in *Proceedings SOSP*, 2001.

[8] A. Akella, B. Maggs, S. Seshan, A. Shaikh, and R. Sitaraman, "A measurement-based analysis of multihoming," in *Proceedings SIGCOMM*, 2003.

[9] A. Basu and J. G. Riecke, "Stability issues in OSPF routing," in *Proceedings of SIGCOMM*, San Diego, California, USA, Aug. 2001, pp. 225–236.

[10] R. Teixeira, A. Shaikh, T. Griffin, and G. M. Voelker, "Network sensitivity to hot-potato disruptions," in *Proceedings of SIGCOMM*, Portland, Oregon, USA, Aug. 2004, pp. 231–244.

[11] A. Nucci, B. Schroeder, S. Bhattacharyya, N. Taft, and C. Diot, "IGP Link Weight Assignment for Transient Link Failures," in *18th International Teletraffic Congress*, Berlin, Germany, Aug. 2003.

[12] B. Fortz and M. Thorup, "Robust optimization of OSPF/IS-IS weights," in *INOC*, Oct. 2003, pp. 225–230.

[13] A. Sridharan and R. Guerin, "Making IGP routing robust to link failures," in *Proceedings of Networking*, Waterloo, Canada, 2005.

[14] D. Applegate and E. Cohen, "Making routing robust to changing traffic demands: Algorithms and evaluation," *IEEE Transactions on Networking*, vol. 14, no. 6, pp. 1193–1206, Dec. 2006.

[15] H. Wang, H. Xie, L. Qiu, Y. R. Yang, Y. Zhang, and A. Greenberg, "COPE: traffic engineering in dynamic networks," in *Proceedings SIGCOMM*, 2006, pp. 99–110.

[16] A. Elwalid, C. Jin, S. H. Low, and I. Widjaja, "MATE: MPLS adaptive traffic engineering," in *Proceedings INFOCOM*, 2001, pp. 1300–1309.

[17] S. Kandula, D. Katabi, B. Davie, and A. Charny, "Walking the tightrope: responsive yet stable traffic engineering," in *Proceedings SIGCOMM*, 2005, pp. 253–264.

[18] P. Psenak, S. Mirtorabi, A. Roy, L. Nguen, and P. Pillay-Esnault, "MT-OSPF: Multi topology (MT) routing in OSPF," IETF Internet Draft (work in progress), Nov. 2006, draft-ietf-ospf-mt-07.txt.

[19] T. Przygienda, N. Shen, and N. Sheth, "M-ISIS: Multi topology (MT) routing in IS-IS," Internet Draft (work in progress), Oct. 2005, draft-ietf-isis-wg-multi-topology-11.txt.

[20] A. Kvalbein, A. F. Hansen, T. Čičić, S. Gjessing, and O. Lysne, "Fast IP network recovery using multiple routing configurations," in *Proceedings INFOCOM*, Apr. 2006.

[21] A. Kvalbein and O. Lysne, "Robust load balancing using multi-topology routing," Simula Research Laboratory, Tech. Rep. 2007-03, 2007.

[22] X. Yang and D. Wetherall, "Source selectable path diversity via routing deflections," in *Proceedings SIGCOMM*, sep 2006.

[23] A. Kvalbein, A. F. Hansen, T. Čičić, S. Gjessing, and O. Lysne, "Fast recovery from link failures using resilient routing layers," in *Proceedings 10th IEEE Symposium on Computers and Communications (ISCC)*, June 2005.

[24] S. Uhlig, B. Quoitin, J. Lepropre, and S. Balon, "Providing public intradomain traffic matrices to the research community," *ACM SIGCOMM Computer Communication Review*, vol. 36, no. 1, pp. 83–86, Jan. 2006.

[25] S. Kandula, D. Katabi, S. Sinha, and A. Berger, "Dynamic load balancing without packet reordering," *ACM SIGCOMM Computer Communication Review*, vol. 37, no. 2, pp. 51–62, apr 2007.