Measuring the Shared Fate of IGP Engineering and Interdomain Traffic

Sharad Agarwal
Microsoft Research
sagarwal@microsoft.com

Antonio Nucci
Narus, Inc.
anucci@narus.com

Supratik Bhattacharyya
Sprint ATL
supratik@sprintlabs.com

Abstract

Typically, each Autonomous System (AS) tunes its local IS-IS or OSPF metrics without any coordination with other ASes. Such local optimizations can lead to sub-optimal end-to-end network performance, as suggested by the performance enhancements achieved by some overlay routing projects. We study the interaction of local IGP engineering in an ISP network with interdomain routing policies. Specifically, (a) how does hot-potato routing (the BGP policy of choosing the closest egress) influence the selection of IGP link metrics? and (b) how does traffic to neighboring ASes shift due to changes in the local AS's IGP link metrics?

In our measurement study, we find that the hot-potato routing policy interacts significantly with IGP engineering - ignoring this interaction resulted in metrics sub-optimal by as much as 20% *of link utilization. Further, the impact on neighboring ASes depends on peering locations and policies, and as much as* 25% *of traffic to a neighboring AS can shift the exit point. Such interdomain shifts can be detrimental to the performance of neighboring ASes. We rely on the actual measured network topology, IGP metrics, traffic matrix and delay bounds. Even though our results are specific to a single ISP, they show significant interaction between local IGP engineering and interdomain routing policies, and thus motivate further work on global network optimization and coordination among ISPs.*

1 Introduction

Network performance such as packet loss and delay typically changes over time for a flow. This is due to bursty traffic loads imposed by other applications sharing the Internet, and due to dynamic network conditions such as link failures. The network is performing sub-optimally when traffic on a path experiences congestion while other parallel paths remain lightly loaded. Network engineering is the task of maintaining optimal network performance by changing the paths that traffic flows on, which can include multiple hops inside a domain and across domains.

The Internet is made of many separate routing domains called Autonomous Systems (ASes), each of which runs an IGP (Interior Gateway Protocol) such as IS-IS [18] or OSPF [16]. The IGP handles routes to destinations within the AS, but does not calculate routes beyond the AS boundary. IGP engineering [7] (or traffic engineering (TE) or IGP optimization) is the tuning of local IS-IS or OSPF metrics to improve performance within the AS. Today, IGP engineering is an ad-hoc process where metric tuning is performed by each AS in isolation. That is, each AS optimizes paths within its local network for traffic traversing it without coordinating these changes with neighboring ASes. It is generally expected that there is sufficient isolation in intradomain routing changes between two ASes, by virtue of sufficient isolation between intradomain routing in an AS and interdomain routing and impact on traffic crossing AS boundaries.

Beyond the AS boundary, the choice of AS hops is determined by BGP [20] (Border Gateway Protocol). BGP engineering is a less developed and less understood process compared to IGP engineering. In addition to whether there is a physical link between two ASes over which routes and traffic can flow on, there are several BGP policies that determine which interdomain paths are exposed to a neighboring AS. Business peering policies can directly translate into which routes are exported to each AS [9, 22]. MEDs (multi-exit discriminators) [21] can be used to control traffic on multiple links between a customer AS and a provider AS. Selective announcement and AS-prepending are additional techniques [3]. After all these policies are applied, the remaining feasible paths can be subjected to the "hot-potato" routing policy. Hot-potato routing occurs when there are multiple egresses to reach a destination. BGP inside that AS will pick the egress point which is "closest" - i.e., has least IS-IS or OSPF cost from the traffic ingress point.

The first problem we consider is the case when IGP metrics change, resulting in hot-potato changes, which then causes traffic within an AS to shift egress points. Typically, IGP link metrics are selected to achieve a network-wide engineering objective such as minimizing maximum link utilization. One of the inputs to the selection process is a traffic matrix (TM) that represents the volume of traffic flowing

Proceedings of the 13th IEEE International Conference on Network Protocols (ICNP'05)
0-7695-2437-8/05 $20.00 © 2005 IEEE

between every ingress-egress pair within an AS [15]. When link metrics are set to new values, the IGP cost to reach various egress points from an ingress point may change. This may lead BGP to recompute the best egress point for certain destination prefixes, thereby leading to traffic shifts. In other words, the final flow of traffic in the network is different from what was considered during the IGP optimization stage. While this has been identified as a potential problem in prior work [5, 4], we are not aware of any prior work that has measured or quantified its impact using real data from an operational network. Certainly, this situation can be preempted by pro-actively considering hot-potato routing during the IGP metric selection process. However, that comes at a cost - it can increase the running time of algorithms that attempt the NP-hard IGP optimization problem. Therefore it is important to understand and quantify the extent to which this is a real issue.

The second problem we examine is how traffic to neighboring ASes shifts due to changes in the local ISP's IGP metrics. Local IGP tuning causes egress points to change for some traffic flows, thereby changing the ingress points for neighboring ASes. This can lead to sub-optimal network performance for those ASes. Their existing IGP metrics are no longer optimal because they were tuned for the previous TM. There may be high traffic load on some links. As a result, the IGP engineering of one AS has impacted other ASes. We are not aware of any prior work that has measured using operational network data how much traffic to neighboring ASes shifts between multiple peering points.

In this work, we use an existing IGP optimization tool and modify it to consider interdomain changes. We use real data gathered from the operational Sprint IP network to drive this tool and analyze potential impacts. This data includes the entire router level topology, existing IS-IS metrics, average link delays, link capacities, BGP routing tables and traffic flow information for the European side of the network. While the results presented are specific to a single ISP, they point to significant interaction between local IGP engineering and interdomain routing policies. Minimizing the maximum link utilization across the AS is the optimization goal of IGP engineering. We find that the hot-potato routing policy interacts significantly with it, to the extent that ignoring such shifts can result in metrics that are sub-optimal by as much as 20% of link utilization. This difference in maximum utilization is because compared to the current IGP metrics on the operational network, the new optimal metrics cause a significant amount of traffic to change egress points. Further, the impact on neighboring ASes highly depends on peering locations and policies, and as much as 25% of traffic to a neighboring AS can move to different interdomain links. Such interdomain shifts can be detrimental to the IGP performance of neighboring ASes. If these neighboring ASes in turn perform IGP engineering, it can result in shifts for this AS, potentially leading to a sub-optimal or unstable situation. Our findings show the need for further work on global network optimization and coordination of local optimization among ISPs.

The remainder of the paper is organized as follows. In Section 2 we describe the problem further using an illustrative example and giving details of route selection. Section 3 has details on our methodology, including the data we collected from the operational network and the optimization tool. We present results on the scope of the problem, the impact on IGP engineering and the impact on neighboring ASes in Section 4. We describe related work in Section 5 and the paper concludes with Section 6.

2 Problem Description

Each Autonomous System (AS) on the Internet is a closed network of end hosts, routers and links, that runs its choice of intradomain routing protocol or IGP. It is typically a link state protocol such as IS-IS or OSPF where each link has a metric setting associated with it. The IGP determines how a network entity such as a router inside the AS reaches another router in the same AS. While there may be multiple possible paths between the ingress router and the egress router, typically the path with the lowest cost (sum of link metrics) is chosen. For example, consider Figure 1 where a destination is multi-homed to the ISP in two locations. Traffic entering the Miami PoP (Point-of-Presence) needs to go to the destination. BGP in the Miami routers has selected the San Francisco egress to reach the destination. Of the two paths, IS-IS picks the cheaper one with cost 45.

IGP engineering or traffic engineering (TE) is an important task in the network operation of most large ASes. It can be applied for a variety of goals. These include avoiding congestion in the network, maintaining SLA (service level agreement) delays and reducing the detrimental impact of link failures on congestion and delay. In IGP engineering, link metrics are changed to alter the relative costs of paths. In the above example, if the path with IS-IS cost 45 experiences heavy load, IGP engineering may increase its cost to 55, to balance the traffic load between the two paths from Miami to San Francisco. In reality, the optimization process considers the entire AS topology and all the flows traversing it and does a joint optimization. The traffic load is represented as a traffic matrix (TM) showing all the ingress and egress points and the demand of traffic between them.

The choice of which interdomain path to use rests with BGP, a path vector protocol. Assuming the destination is advertised from both the Seattle and San Francisco peering points, BGP has to apply several policy rules to determine which egress point traffic will go to. Shortest AS path lengths are preferred (rule 4 in Figure 2). If the destination advertised different MED values on the routes between the

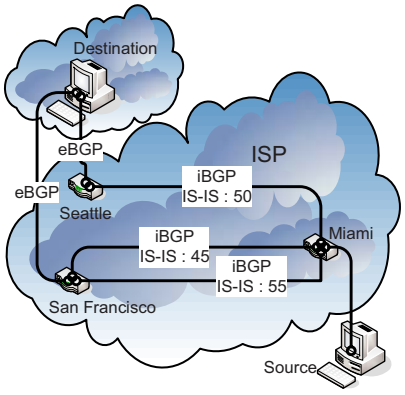

Figure 1. Example of Hot Potato Routing

1. Largest weight
2. Largest local preference
3. Local paths over remote paths
4. Shortest AS path
5. Lowest origin type
6. Lowest MED
7. eBGP over iBGP paths
8. Lowest IGP metric
9. Oldest path
10. Lowest router ID
11. Minimum cluster ID length
12. Lowest neighbor IP address

Figure 2. BGP Route Selection

two links, the route with the lowest MED value will be preferred. However, if the destination can be reached from both Seattle and San Francisco with the same AS path length, MED values, and there is no local policy preferring one over the other, the hot-potato policy will apply. This corresponds to rule 8 in Figure 2. BGP will chose the egress which has the lowest IGP cost from the ingress. So for traffic entering Miami, the cheapest exit to the destination is San Francisco, with an IS-IS cost of 45. In small ASes where multiple other ASes connect to it at the same point, this IGP cost will be practically indistinguishable between multiple BGP routes. However, for ISPs spanning a large geographic area, there can be a significant range in the IGP cost.

In this example, if due to IGP engineering the cheapest IS-IS cost from Miami to San Francisco goes to 55, then BGP will change the egress to Seattle for this destination. As a result, both the performance of the local AS and that of the destination AS will change. For the local AS, the TM will now be different because the egress point is different.

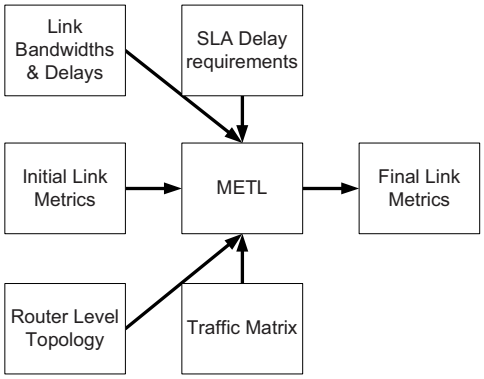

Figure 3. METL (igp MEtric assignment TooL)

If the IGP metric optimization process did not account for hot-potato routing, it will not expect the shift in egress. As a result, it may end up with more traffic load on the Miami to Seattle links than expected. Secondly, the destination AS will experience a different traffic load. Traffic that was ingressing its network in San Francisco, is now ingressing in Seattle. Its TM has changed and now it may have to re-do its IGP engineering.

It is important to measure the extent of this interaction using operational network scenarios to determine if further work is needed to solve global network optimization and coordination of local optimization among ISPs. To this end, we want to quantify two impacts. First, we want to identify by how much the maximum link utilization in the local AS changes as a result of hot-potato routing. IGP engineering optimizes metrics to reduce the maximum link utilization across the network because it makes the network more resilient to link failures and sudden traffic surges. If these new metrics cause hot-potato shifts that are not accounted for in the optimization process, the final link utilizations can be very different. The increase in maximum utilization is a measure of how important this problem is to the operation of an AS. Second, we want to find out how neighboring ASes are impacted. A large volume of traffic shift on the peering links to neighboring ASes can alter the internal link utilizations of those ASes.

3 Methodology and Data

Our analysis methodology and data has several components. We use an IGP optimization tool, the network topology, traffic data to build a TM, BGP routing data and SLA constraints. We now describe each of these in detail.

3.1 IGP Optimization Tool

To analyze IGP engineering, we use a tool called METL (igp MEtric assignment TooL) that we developed. It takes

in the current state of the network, and outputs a set of final IGP metrics. The goal of METL is to output a new set of IGP metrics that meet the SLA constraints, while minimizing the maximum link utilization across the network. It is important to keep the worst link utilization as low as possible because traffic tends to be bursty and can grow over time. A description of the algorithm underlying the tool and detailed results of its performance can be found in our prior work [17]. Since prior work [7] has shown that optimal IGP metric assignment is NP-hard, METL uses heuristics to achieve a solution. METL provides results that are within 10% of tight theoretical lower bounds. It is not a goal of this paper to present METL in detail, and we believe our results are independent of the IGP optimization tool used. We use METL simply because we are familiar with it and because of the high quality of its solutions, as reported in our prior work.

As shown in Figure 3, it takes several inputs. It requires the router level topology of the network. This is a list of every link inside the AS that the IS-IS or OSPF protocol encompasses. Each link is annotated with the source router name and the sink router name. METL also requires the delay and capacity of each link. We obtain the topology and the link capacities from the router configurations of all the routers in the Sprint IP network. The link capacity is specified by the router line card type or a lower configured limit. We obtain the delay by performing extensive measurements of our network by exchanging ICMP messages between routers. The current IGP metric settings are also in the router configurations. The SLA (Service Level Agreement) constraints are also needed. This is a matrix that lists the delay guarantees promised to customers between different cities in the Sprint IP network [1].

3.2 Traffic and Routing Data

METL optimizes for lowest link utilization. The utilization is calculated by considering the shortest paths in the network [2] based on the IGP metrics and the input TM. The TM is a point-to-multipoint traffic matrix. Each row represents a flow, defined by the ingress router, traffic volume, and all the possible egress routers.

To build this TM, we use traffic data and BGP routing data. For traffic data, we use Cisco Netflow v5 ("Sampled Aggregated Netflow") to collect flow size information. Measuring flow data by capturing every packet at high

[1] An overview of the SLA for the Sprint IP network is at http://www.sprintworldwide.com/english/solutions/sla/

[2] Note that in the Sprint IP network, the routers are configured to use equal cost multipath routing (ECMP). Between a single source router and single exit router, if multiple IGP paths exist with equal IGP cost, then traffic is split equally along both paths. This equal traffic split is performed at each next-hop instead of in an end-to-end fashion. Since we wish to accurately model what occurs in the network, we account for ECMP in METL.

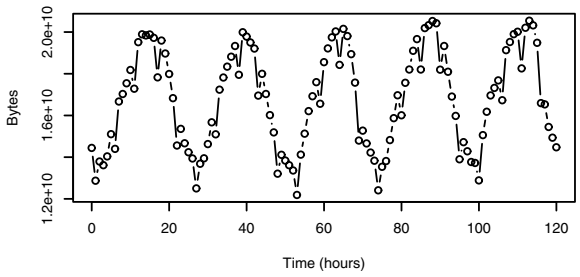

Figure 4. Total Traffic Entering Europe (1 hour bin)

packet rates can overwhelm a router's available processing power. Thus Cisco Netflow v5 uses periodic sampling to collect every 250th packet. We aggregate the measurements into one hour intervals. Due to hardware limitations of the operational routers, Cisco Netflow can only be enabled on a fraction of ingress routers. We have Netflow traffic information from all the ingress routers in the European portion of the Sprint network. Thus we can only build a partial TM. To work around this limitation we constrain our study of IGP optimization by allowing METL to only change the IGP metrics of European links. We believe it would unfairly skew our results to change US link metrics without the US ingress traffic. However, the tool still considers the full network topology and exit points to capture all the hot-potato dynamics. This is still a size-able problem - there are over 1300 links in the entire topology, 300 of which are in Europe. This approach of focusing on one continent of the network allows us to examine a scenario where we have complete network information and does not affect the validity of our findings for that portion of the network.

For the European Netflow data, we have 3 months of ingress traffic data at our disposal. We need to pick a time period out of these 3 months to build a TM. In Figure 4, we plot the total volume of ingress traffic in bytes over a typical 5 day period around 12 February 2004. As expected, there is a strong diurnal pattern. When optimizing IGP, network operation engineers typically consider peak time utilization of the network instead of off-peak time. Their objective is to bound the worst case performance of the network. Thus, we pick a 1 hour window during one of the daily peaks. We have considered multiple 1 hour peaks and the results are similar. Thus we present results from one representative period for conciseness.

This traffic data is now a point-to-point TM, where each flow entry gives the ingress Sprint router, the destination prefix and the number of bytes. We need to convert this into a point-to-multipoint TM, which lists the ingress Sprint router, all the candidate egress routers and the number of bytes. The candidate egress routers are all the ones that can sink the destination prefix and where the choice of egress router depends on the hot-potato IGP cost. For the same 1

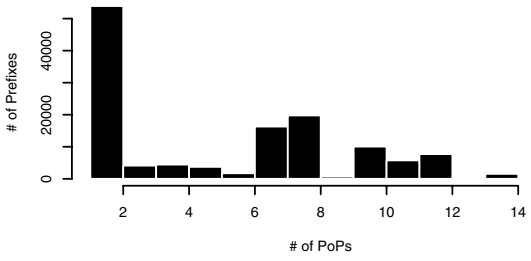

Figure 5. Distribution of Destination Prefixes by Exit PoPs; 12 Feb 2004

hour time period, we have BGP routing data from the Sprint IP network. We operate a software BGP router that connects to over 150 routers in the Sprint IP network. This collection router participates as a route reflector in the iBGP route reflector mesh [21] inside the Sprint IP network. Thus it sees all the candidate BGP routes that have reached the end of step 7 in Figure 2. We use a snapshot of the BGP routing table during the 1 hour time period, which lists all the candidate egress routers for every destination prefix. We correlate this with the traffic to obtain a point-to-multipoint TM.

In our data set, there are over $30,700$ entries in this TM. To reduce the processing time for our analysis, we consider a subset of this data. We consider the largest flows that contribute at least 80% of the total traffic entering Europe. As has been noted in prior work [2, 11], the elephant and mice phenomenon exists in our data set - a few destination prefixes sink the majority of traffic. By considering the largest subset of traffic, we reduce our processing time significantly. This reduces our point-to-multipoint TM to 530 entries. For the remainder of this paper, we treat this as the full input data set. We revisit this issue of running time in the conclusions.

3.3 Analysis

In this work, we want to study two issues. We want to know how badly IGP optimization suffers when considering a point-to-point TM versus considering hot-potato shifts using a point-to-multipoint TM. Secondly, we want to know how traffic to neighboring ASes changes as a result of local hot-potato shifts during IGP engineering.

For the first issue, we operate METL in two modes - TE and BGPTE. METL-TE ignores hot-potato routing; it assumes the egress points do not change when IGP metrics change. It uses the egress points that the input state of the network would use; in essence, it is using a point-to-point TM. We use the final output metrics from this mode to evaluate what the performance of the network would be without hot-potato shifts. We also evaluate how the network would actually perform with these new metrics after re-evaluating the final egress points. METL-BGPTE accounts for interdomain traffic shifts by re-calculating the egress point choice for every flow. It does this BGP route calculation in every iteration of the heuristic when evaluating a candidate set of IGP metrics. We evaluate the performance of these metrics compared to previous mode of operation.

To study the impact on other ASes, we consider how traffic on links to large neighbors changes between the original IGP metrics and the final optimized IGP metrics.

4 Results

Based on the METL tool, the full Sprint IP network topology of over 1300 links, actual network link delays, actual SLA constraints and the point-to-multipoint TM that we generate from actual network data, we now present results on how IGP engineering interacts with interdomain routes and traffic.

4.1 Scope of Problem

We begin by quantifying the extent of the hot-potato interaction problem, in terms of how many prefixes and how much traffic are vulnerable. Since the original deployment of BGP approximately 15 years ago, the number of ASes participating in it has grown to over $16,000$ today. If multiple paths exist to a destination, it can potentially exacerbate this issue. More specifically, the destination prefixes that are reachable from multiple PoPs or egress cities in the Sprint topology can experience hot-potato shifts.

In Figure 5, we plot the number of exit PoPs that prefixes in a typical routing table have. A typical iBGP routing table on a router inside the Sprint network will have about $150,000$ prefixes. Each prefix may have multiple routes, but after applying BGP route selection up to rule 8 in Figure 2, we are left with the candidates for hot-potato selection. Figure 5 considers only these remaining candidates for each prefix. We see that for about 40% of prefixes, there is only 1 egress point. That is, no matter how much IGP costs change inside the Sprint AS, those prefixes will always exit the same PoP. However, for the remaining 60% of prefixes, there are 2 or more candidate routes where the IGP cost can play the determining factor in route selection. We see that for a significant percentage of routes, over 6 different exit cities exist! After examining this data more carefully, we have found that the prefixes with over 6 PoPs are behind large ISPs or "Peer" ISPs. Typically, Peer ISPs filter out MED values and BGP communities that modify BGP local preferences when receiving routes from each other. This is in accordance with their business agreements which are different from the more common customer-provider relationship [9]. Thus destinations behind large ISPs are more susceptible to hot-potato routing because of this policy to

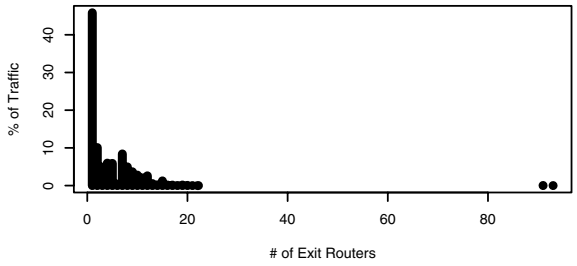

Figure 6. Distribution of European Traffic by Exit Routers; 12 Feb 2004

filter out certain BGP attributes. This leads us to consider the impact of IGP engineering on neighboring ASes, which we present later in this section.

In Figure 6, we plot the analog of Figure 5 for traffic. We show the distribution of traffic entering the European portion of the network by the number of possible exit routers. As in the previous figure, we only consider the BGP routes where the next selection rule is IGP cost. We see that for about 45% of traffic, there is only one network egress point. However, the remaining 55% of traffic is susceptible to hot-potato changes. Examining this data further, we find that the traffic with around 90 exit routers is actually network management traffic for the Sprint network.

IGP changes can occur for a variety of reasons, including link failures, the addition of network links, and IGP engineering. IGP changes in one part of the network can potentially cause BGP route selection to change across the network and thus cause large amounts of traffic to shift. Given the large number of prefixes and volume of traffic that are susceptible, it is now important to determine the extent of this interaction using operational network scenarios.

4.2 Impact of Hot-Potato Shifts on IGP Engineering

As we described in Section 2, IGP engineering that uses a point-to-point TM and ignores interdomain traffic shifts results in a simpler optimization algorithm with faster running time and less data collection constraints. While the resulting new IGP metrics may be optimal for the point-to-point TM, in reality they may cause BGP to recalculate egresses, causing them to be actually sub-optimal. We want to measure how high the final traffic link loads can be.

We ran METL-TE, which ignores hot-potato shifts. In Figure 7, the points marked by crosses show the expected utilization of each link in the European topology with these new metrics. We sorted the links by utilization, which is the ratio of traffic flowing on a link to the capacity of the link. For ease of presentation, we show the 50 most utilized links out of the 300 in Europe. For good network performance, we want the highest link load on any link to be as low as

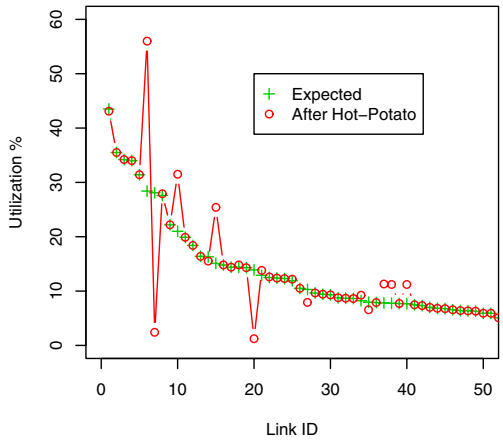

Figure 7. Link Utilization from METL-TE

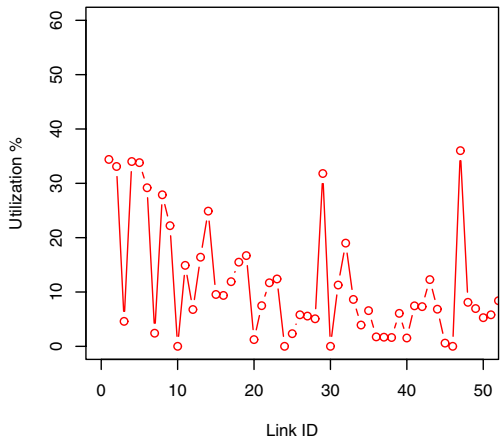

Figure 8. Link Utilization from METL-BGPTE

possible. That is, minimize the effect of bottleneck links. METL-TE has optimized the network and expects the final worst case link utilization to be 43.5%, which is on a high capacity link from router $R10$ to $R11$. This corresponds to the left most point on the graph. However, if we were to install these new metrics on the network, hot-potato shifts could result in different utilizations.

We re-evaluate these new IGP metrics while allowing BGP to change the exit points as needed, and show the resulting link utilizations by the circular points in Figure 7. We now see that the solution is actually significantly worse in reality - the highest utilization is 56.0%, on a low capacity edge link from router $R4$ to $EX2$. This link was expected to be only 28.0% utilized. 3.7% of the total volume of European traffic changed the exit point, which on individual links resulted in high utilization. Thus in this data set, ignoring hot-potato shifts in IGP engineering resulted in a set of IGP metrics that have 12.5% higher maximum link utilization.

Now we want to examine how much better IGP opti-

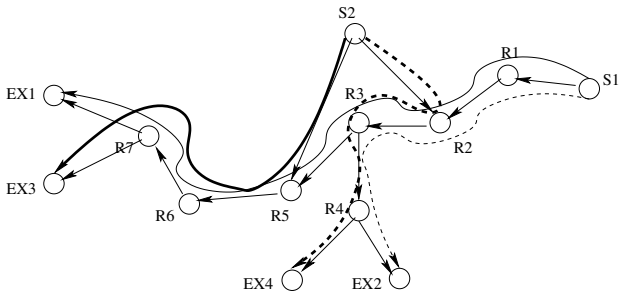

Figure 9. Example of Flow Shifts

mization can do if it uses a point-to-multipoint TM and recalculates BGP routes for every IGP metric change it considers. We ran METL-BGPTE using the same data set but allowing it to account for hot-potato shifts. The output set of IGP metrics cause the link utilizations shown in Figure 8. We see that now the maximum link load across the European links is 36.0%. The low capacity edge link from that suffered in the previous case now has a load of only 29.2%. Thus in this data set, modeling hot-potato shifts in IGP optimization resulted in a set of final IGP metrics that have 20% lower maximum link utilization than when ignoring BGP route recalculation.

By allowing the optimization heuristic to re-evaluate BGP routes in every iteration, we are allowing it to take advantage of hot-potato shifts to further reduce link utilization. To explain that, we present an example of traffic shift in Figure 9. All the circles depict routers in the Sprint network. $S1$ and $S2$ are ingress routers, $EX1$ to $EX4$ are egress routers and the remaining $R1$ to $R7$ are internal routers. Recall that in Figure 7, METL-TE expected the link $R4$ to $EX4$ to have only 28.0% utilization, but due to hot-potato shifts, it ended up with 56.0% utilization. This is because in the input traffic matrix, there are 11 flows going from $S1$ to $EX1$ and from $S2$ to $EX3$, as shown by the two solid lines in Figure 9. The original IGP metrics from the Sprint router configurations resulted in a hot-potato choice of sinking this traffic at $EX1$ and $EX3$ respectively. However, METL-TE reduced the link metric on $R3$ to $R4$. It continued to expect traffic to be sunk at $EX1$ and $EX3$. However, the destinations are multi-homed - one destination is behind both $EX3$ and $EX4$, and another is behind both $EX1$ and $EX2$. BGP re-evaluates its routes and picks $S1$ to $EX2$ and $S2$ to $EX4$ as the new routes. As a result, the link utilization of $R4$ to $EX4$ increases unexpectedly to 56.0%. When the IGP optimization was allowed to recalculate BGP routes every iteration in METL-BGPTE, it discovered a set of IGP metrics where BGP selected a path of $S2$ to $EX3$ and a path of $S1$ to $EX2$. This drove down the utilization of $R4$ to $EX4$. In essence, the heuristic discovered a solution where hot-potato induced shifts reduced the link utilization.

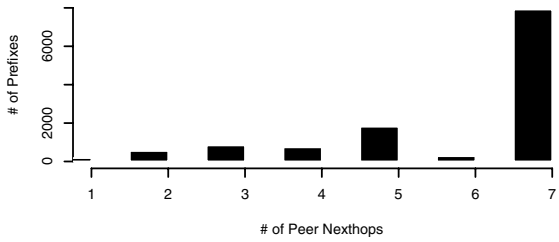

Figure 10. Prefix Distribution to Neighbor AS A

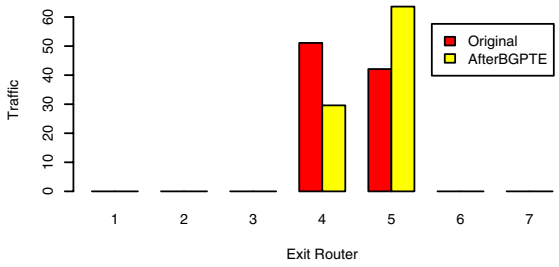

Figure 11. Traffic Distribution to Neighbor AS A

4.3 Impact of IGP Engineering on Neighbors

Unfortunately, the downside of using heuristics that increase local AS performance by inducing hot-potato traffic shifts is that they may affect traffic going to neighboring ASes. In the previous example, $EX1$ and $EX2$ are ingress points for a neighboring AS Y. Traffic has shifted between these two points due to the IGP engineering of the Sprint AS. As a result, the TM for AS Y is different, and its link utilizations will change. A significant enough change may require it to re-do its own IGP engineering.

We have calculated how the traffic shifts for various large neighbors of the Sprint network. For example, a large peer ISP A connects to the Sprint network in 7 locations. In Figure 10, we show for each prefix how many locations it is announced at. ISP A announces most prefixes at all 7 peering points. In Figure 11 we show how much of the measured European traffic exits at each of these 7 locations to ISP A. The dark bars show the distribution of traffic in the original network setting (i.e., with the IGP metrics from the router configurations). All 7 locations are in the US side of the topology, and routers 4 and 5 are on the East coast, closest to the European continent. Since these two exit points are cheapest in terms of IGP cost from Europe, they sink the majority of traffic. The light bars show how the traffic to these two points would change after we apply the new IGP metrics proposed by METL-TE, when allowing it to account for BGP exit point shifts. We see that now router 5 sinks more traffic than 4. In fact, about 22% of the traffic from Europe to this ISP changes its exit point. Due to privacy concerns, we cannot reveal the identity of the neighbor

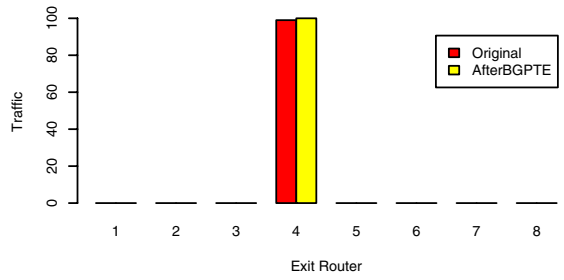

Figure 12. Traffic Distribution to Neighbor AS B

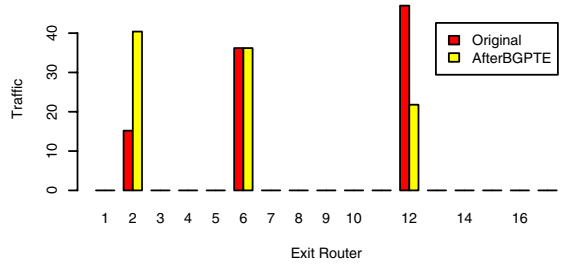

Figure 13. Traffic Distribution to Neighbor AS C

nor the volumes of traffic involved. Even though the BGP aware IGP optimization reduced link utilization by 20%, it came at the cost of changing traffic going to a neighboring AS by 22%. As we described in Figure 5, traffic to "peer" ISPs is particularly vulnerable. This shift in egress points can potentially have a detrimental impact on this neighbor.

However, the extent of such impact depends significantly on the locations that an AS connects to Sprint, and the BGP policies that govern those connections. Consider another ISP B that has the egress traffic distribution in Figure 12. Of the 8 peering locations, this ISP has 2 in Europe, 5 in US and 1 in Asia. In the extreme, IGP costs to reach a destination will mimic relative geographic distances due to the SLA delay constraints. So for our measured traffic entering Europe, any destinations in ISP B will naturally exit through one of the 2 European peering points. However, the Sprint router configurations have a BGP local preference for one of these peering points. As a result, no matter how much IGP metrics vary, hot-potato routing will not come into play because of this local BGP policy. A minor amount of traffic shifted into this exit router from another AS.

As a third example, consider the traffic distribution for a peer ISP C, shown in Figure 13. This large ISP has 17 peering locations, one of which is in Europe, and the rest are in the US. Two of the US locations are again on the East coast. Here, the BGP route advertisements from ISP C do not match as well as they did for ISP A in Figure 10. As a result, router 6 in Europe sinks some traffic, while routers 2 and 12 in the East coast of the US sink the rest. IGP metric changes due to local optimization in Europe shift 25% of

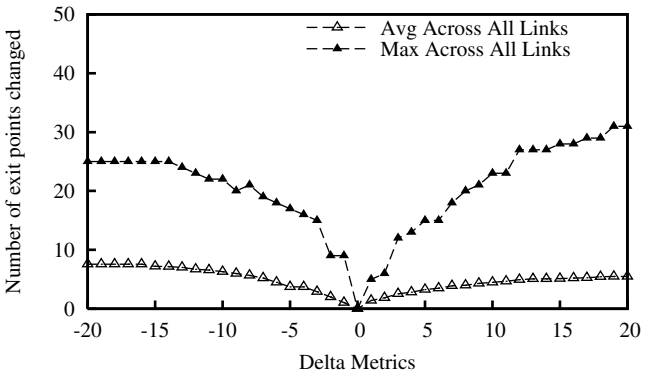

Figure 14. Flows Shifted With Individual Link Metric Perturbations

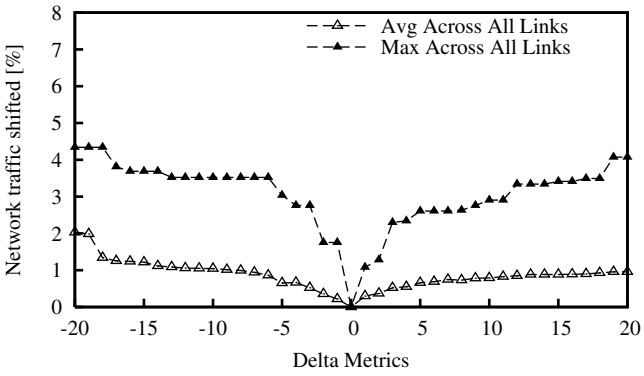

Figure 15. Network Traffic Shifted With Individual Link Metric Perturbations

the traffic to ISP C between the two US routers.

4.4 Impact of More Frequent Changes

Thus far, we have considered the case where a network operator chooses a set of IGP metrics for the entire network that optimizes utilization across all links. However, in practice, network-wide optimization is a rare occurrence. Typically an operator will apply a few, small link metric changes to react to sudden traffic surges or link additions or failures. She will incrementally adjust the metrics for a handful of links in a trial and error fashion without considering the global optimum. From an operational standpoint, it is important to understand the impact of such changes in light of BGP hot-potato shifts.

In Figure 14, we show the number of flows that shift by the change in metric for a link. For example, consider the points at +20 metric. For each link in the network, we calculate which BGP paths would change the egress if this

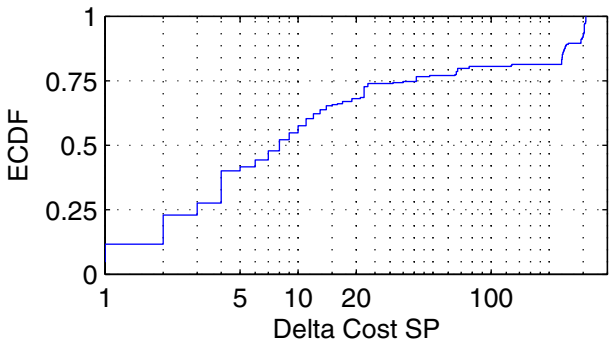

Figure 16. Cumulative Distribution of Cost Change Required to Cause Hot-Potato Shifts

link's metric was 20 higher, and then count the number of flows that go to these shifted egress points. We plot the number of flows shifted, averaged over all links as well as the maximum over all links. Thus if a random link is picked and its metric is increased by 20, it is likely that about 8 flows will change their egress point, but in the worst case it can be as many as 32. In Figure 15 we show how much traffic this corresponds to. The vertical axis is in percentage of the total volume of ingress traffic. This shows that as much as 4% of total network traffic can see hot-potato induced shifts if any particular link's metric is changed by $0 - 20$. These are relatively small metric changes - for the European network, the IGP link metrics are in the range of 1 to 60.

Hot-potato shifts can also be caused by link failures [23, 14]. A link failure can only increase the cost of paths between two points that used that link, since IS-IS or OSPF will now have to pick a more expensive path around it. Using our final IGP metric solution, we calculate for each flow by how much the cost to reach the chosen egress has to change before inducing a hot-potato shift to a different egress point. We plot the cumulative distribution across all flows that have multiple egress points in Figure 16. We see that for 50% of flows, an increase of only 8 will cause them to shift the exit point.

5 Related Work

We are not aware of any prior work that has measured and evaluated the interaction of hot-potato routing with link metric optimization using operational network data. Despite that, IGP engineering has been an active area of research. Fortz and Thorup [7] showed that finding optimal IGP link metrics is NP-hard with respect to several objectives and thus propose a local search heuristic. Other heuristics have also been considered including simulated annealing, Lagrangean relaxation, and genetic algorithms [19, 10, 25]. Fortz and Thorup [8] also consider assigning metrics in the context of traffic matrix changes and link failures. Our prior work [17] has incorporated Service Level Agreement constraints into the problem as well as handling short lived link failures.

A variety of other work has contributed to traffic matrix (TM) estimation which we do not cite here. However, the majority of this prior work considers a point-to-point TM model, where each traffic demand or flow has a single ingress point and a single egress point. In such a model, hot-potato routing shifts do not occur. Feldmann [5] has described an IGP optimization tool that considers a point-to-multipoint TM. The data needed by this tool as well as the manner in which it is to be processed is also described [4, 6]. However, there are no results showing how much traffic can be affected by IGP changes due to multiple BGP exit points. Results of how badly the optimization suffers without this complex analysis are also not presented. In this work, we provide these results based on data we have collected from an operational network.

While we are not aware of prior work that presents results on how hot-potato routing interacts with IGP engineering, we are aware of recent work that considers hot-potato routing in other scenarios. Hot-potato variations can occur during typical network operation [1, 24]. Teixeira et al [23] have examined how these variations occur when the IGP topology changes. While closely related, their focus is on network failure episodes, such as link failures, router failures and forwarding loops. We primarily consider interaction with IGP metric optimization. If we cannot do ad-hoc network management by assuming the logical separation of network boundaries holds, then coordination is required between intradomain TE across multiple ASes on the Internet. Multiple research challenges exist for achieving an Internet wide, scalable network management system. Recently, proposals [13, 12] to address these issues have appeared. We provide an analysis of existing network conditions to expose the extent of this problem.

6 Conclusions

This work is motivated by the current trend toward greater connectivity between ASes on the Internet. We need to revisit the isolation of IGP in one AS from the IGP of a neighboring AS. By using a real network topology, link delays, link capacities, delay constraints, routing tables and traffic matrix, we have evaluated how interdomain routing policies can interact with IGP engineering. We have found that in our data set, ignoring the hot-potato routing policy can result in IGP metrics that are sub-optimal by 20% of link utilization, which is a significant amount for typical network operation.

However, to consider such hot-potato shifts, a point-to-

multipoint TM has to be employed for IGP engineering. This is challenging because we cannot rely on SNMP-based TM estimation techniques to obtain it. It requires Netflow information, which we could not even collect from the entire network due to different versions of deployed router operating systems. A second concern is the running time of IGP optimization. Hot-potato calculations in every iteration of the heuristic add significant complexity to the process. This has significantly increased the running time of METL, even when considering only 80% of the European ingress traffic. In the point-to-point TM mode, METL only takes 5 minutes to provide a solution on a dual 1.5 Ghz Intel processor Linux PC with 2 GB of memory. Using a point-to-multipoint TM for hot-potato shifts, it takes 55 minutes. Considering a smaller percentage of the traffic dramatically improves the running time since fewer flows are considered. When we consider 76% of the traffic, we obtain a solution that has a maximum link load of 41.5% after routing all 80% of the traffic. This is an increase from the 36% utilization from the 80% traffic solution, but the running time is only 10 minutes for this smaller TM.

A bigger concern is the impact on neighboring ASes. Such shifts in traffic to neighboring ASes can impact their network performance if large volumes of traffic shift. In our study, we found cases where as much as 25% of traffic to a neighboring ISP shifts the egress point due to local IGP engineering. We found several other scenarios where local routing policy and peering locations played an important part in determining if a neighbor can be affected by local AS changes. This is an important issue because it further increases the inter-dependence of performance between neighboring ASes. One AS improving its performance can reduce a neighbor's performance. If the neighbor then re-computes its IGP metrics, it can lead to instability. However, coordinating IGP engineering between competing ISPs without revealing private information is challenging.

Acknowledgements

Our work would not have been possible without the help of Sprintlink Engineering and Operations in collecting IS-IS, BGP, SNMP, and netflow traces from the network. In particular, we want to thank Ryan McDowell and Bjorn Carlsson.

References

[1] S. Agarwal, C. Chuah, S. Bhattacharyya, and C. Diot. The impact of BGP dynamics on intra-domain traffic. In *ACM SIGMETRICS*, June 2004.

[2] S. Bhattacharyya, C. Diot, J. Jetcheva, and N. Taft. Pop-level and access-link-level traffic dynamics in a tier-1 PoP. In *Internet Measurement Workshop*, 2001.

[3] N. Feamster, J. Borkenhagen, and J. Rexford. Guidelines for interdomain traffic engineering. In *ACM CCR*, 2003.

[4] N. Feamster and J. Rexford. Network-wide bgp route prediction for traffic engineering. In *SPIE ITCom*, 2002.

[5] A. Feldmann, A. Greenberg, C. Lund, N. Reingold, and J. Rexford. Netscope: Traffic engineering for IP networks. In *IEEE Network Magazine*, 1999.

[6] A. Feldmann, A. G. Greenberg, C. Lund, N. Reingold, J. Rexford, and F. True. Deriving traffic demands for operational IP networks: methodology and experience. In *ACM SIGCOMM*, pages 257–270, 2000.

[7] B. Fortz and M. Thorup. Internet traffic engineering by optimizing OSPF weights. In *IEEE INFOCOM*, March 2000.

[8] B. Fortz and M. Thorup. Optimizing OSPF/IS-IS weights in a changing world. In *IEEE JSAC Special Issue on Advances in Fundamentals of Network Engineering*, 2001.

[9] L. Gao. On inferring automonous system relationships in the Internet. In *IEEE INFOCOM*, November 2000.

[10] J. Harmatos. A heuristic algorithm for solving the static weight assignment optimisation problem in OSPF networks. In *Global Internet Conference*, November 2001.

[11] K.Papagiannaki, N. Taft, S. Bhattacharyya, P. Thiran, K. Salamatian, and C. Diot. A pragmatic definition of elephants in Internet backbone traffic. *Internet Measurement Workshop*, 2002.

[12] S. Machiraju and R. Katz. Verifying global invariants in multi-provider distributed systems. In *Hotnets Workshop*, 2004.

[13] R. Mahajan, D. Wetherall, and T. Anderson. Towards coordinated interdomain traffic engineering. In *Hotnets*, 2004.

[14] A. Markopoulou, G.Iannaccone, S.Bhattacharyya, C.N.Chuah, and C.Diot. Characterization of failures in an IP backbone network. In *IEEE INFOCOM*, 2004.

[15] A. Medina, N. Taft, K. Salamatian, S. Bhattacharyya, and C. Diot. Traffic Matrix Estimation: Existing Techniques and New Directions. In *ACM SIGCOMM*, August 2002.

[16] J. Moy. OSPF version 2. RFC 1583, IETF, March 1994.

[17] A. Nucci, B. Schroeder, S. Bhattacharyya, N. Taft, and C. Diot. IGP link weight assignment for transient link failures. In *International Teletraffic Congress*, August 2003.

[18] D. Oran. OSI IS-IS intra-domain routing protocol. RFC 1142, IETF, February 1990.

[19] M. Pioro, A. Szentesi, J. Harmatos, A. Juttner, P. Gajowniczek, and S. Kozdrowski. On OSPF related network optimisation problems. In *IFIP ATM IP*, July 2000.

[20] Y. Rekhter and T. Li. A Border Gateway Protocol 4 (BGP-4). RFC 1771, Network Working Group, March 1995.

[21] J. Stewart. *BGP4: Inter-Domain Routing in the Internet*. Addison-Wesley, 1998.

[22] L. Subramanian, S. Agarwal, J. Rexford, and R. H. Katz. Characterizing the Internet hierarchy from multiple vantage points. *IEEE INFOCOM*, 2002.

[23] R. Teixeira, T. Griffin, A. Shaikh, and G. Voelker. Network sensitivity to hot-potato disruptions. In *ACM SIGCOMM*, August 2004.

[24] R. Teixeira, A. Shaikh, T. Griffin, and J. Rexford. Dynamics of hot-potato routing in IP networks. In *ACM SIGMETRICS*, June 2004.

[25] Y. Wang, Z. Wang, and L. Zhang. Internet traffic engineering without full mesh overlaying. In *IEEE INFOCOM*, April 2001.