# Making IP Traffic Engineering Robust to Intra- and Inter-AS Transient Link Failures*

Mina Amin, Kin-Hon Ho, Ning Wang, Michael Howarth and George Pavlou
Centre for Communication Systems Research, University of Surrey, UK
{M.Amin,K.Ho,N.Wang,M.Howarth,G.Pavlou}@surrey.ac.uk

**Abstract— Intra- and inter-AS transient link failures are common in operational IP networks. Robust intra- and inter-AS Traffic Engineering (TE) schemes have been proposed to optimize network performance against transient link failures. The existing literature has focused solely on either intra- or inter-AS link failure. They have, however, neglected the interactions between robust intra- and inter-AS TE, specifically the impact of intra-AS link failure on inter-AS TE and vice versa. As a result, the overall network performance may not be truly robust to link failures if the interactions are neglected. This paper proposes a joint robust TE approach that takes the interactions into account for achieving good network performance under both normal state and any single intra- or inter-AS link failure. We propose a two-phase heuristic to solve the problem and compare its performance with four alternative approaches that do not consider the interactions. Evaluation results reveal that our joint robust TE approach achieves higher robustness against intra- and inter-AS link failures than all the alternatives.**

## I. INTRODUCTION

THE Internet consists of thousands of Autonomous Systems (ASes), each of which runs an Interior Gateway Protocol (IGP) such as OSPF or IS-IS. For inter-domain traffic, the selection of the next-hop AS is determined by the Border Gateway Protocol (BGP). Today, traffic engineering is a technique that can be adopted by operators to optimize the performance of their operational IP networks. Engineering the traffic within an AS boundary based on IGP, called *intra-AS TE*, is effectively the tuning of the link weights [4,5,15,16], whereas selecting the best egress points for traffic to be sent to the next-hop ASes, called *inter-AS outbound TE*, is effectively the adjustment of BGP route attributes [1,6,7,11,21]. Recent studies in [2,3] have shown that both intra- and inter-AS link failures are part of the daily routines in large IP backbone networks, and most of these failures are common and transient. Over a 4-month period, 80% of inter-POP link failures lasted less than 10 minutes and 50% of them even lasted less than a minute [2]. In addition, more than 70% of these transient failures are single link failures. On the other hand, for 9452 eBGP peering link failures in 3 months in a transit ISP, 82% of them lasted for no more than 3 minutes [3].

When a link fails, traffic is diverted to alternative paths, thus increasing the load on these new serving paths and possibly leading to congestion. To avoid this, one might take a reactive approach of re-computing the IGP link weights and/or BGP route attributes after the failure. However, this may not be practical for two reasons. First, due to the transient nature of failures, there would be insufficient time for operators to re-compute the best post-failure TE configuration and implement it before the failed link is restored. Second, the new configuration will have to be advertised to every router in the network, and every router will have to re-compute the shortest path to every other router and to re-select its best egress point. This can lead to considerable instability, aggravating the situation already created by the link failure.

Although the reactive approach may not be appropriate or even feasible, transient link failures can be handled by computing the set of TE configurations in a proactive manner that is robust to all potential link failures. The goal of such a robust TE approach is to obtain a reasonably good network performance both under the normal state (i.e. absence of failures) and also under any potential link failure. Various kinds of robust TE approaches based on IGP link weight optimization [4,5] and BGP egress selection [6,7] have been proposed. These proposals, however, make their TE approaches robust either only to intra-AS or only to inter-AS transient link failures. They have neglected the interactions between robust intra- and inter-AS TE, specifically the impact of intra-AS link failures on robust inter-AS outbound TE and the impact of inter-AS link failures on robust intra-AS TE. As a result, the overall network performance may not be truly robust to link failures if these interactions are not considered.

In one scenario, if an inter-AS link (or egress point) fails, the inter-AS traffic is diverted from the failed egress point to other alternative egress points. This may cause a huge load increase not only at these new serving egress points but also at any link along the IGP paths between some ingress and the new egress points. In the other scenario where multiple egress routers have BGP routes that are equally good (i.e. they have the same local preference, AS path length, origin type, and multiple-exit-discriminator) for a routing prefix, each router in the AS directs the traffic to its closest egress point in terms of IGP distance. This is also known as Hot-Potato Routing (HPR). If an intra-AS link fails, the IGP distance between some ingress and egress points may change, causing thus some ingress points to divert the traffic to different egress points due to the HPR effect. These HPR changes are responsible for many of the large traffic shifts [13] in operational networks. Therefore, failure of an intra-AS link may shift a large proportion of traffic to other egress points and lead to a sudden load increase there. This may also result in excessive traffic to be sent to downstream ASes, violating the traffic exchange limits specified in their peering agreements.

Given the above interactions, we investigate the impact of both intra- and inter-AS transient link failures on robust TE. Accordingly, we propose a joint robust TE approach based on

* This work was undertaken in the context of FP6 Information Society Technologies AGAVE (IST-027609) project, which is partially funded by the Commission of the European Union.

978-1-4244-2066-7/08/$25.00 ©2008 IEEE

IGP link weight assignment for intra-AS and inter-AS outbound TE that is robust to all potential single intra- or inter-AS link failures. The goal is to find a set of IGP link weights that minimizes the intra- and inter-AS Maximum Link Utilization (MLU) under both the normal state and the worst case across all single link failure states while also taking HPR into account. We propose a two-phase heuristic algorithm to solve this problem and compare it with four IGP link weight optimization approaches in which two of them did not consider any link failure while the other two considered only intra-AS link failures. Nevertheless, all of them neglected the impact of both intra- and inter-AS link failures on the overall performance. Learning from our evaluations, we came to the following conclusions for the robust TE design: **1. Not only intra- but also inter-AS transient link failures should be considered.** The results reveal that the joint robust TE approach significantly improves both intra and inter-AS MLU, particularly under inter-AS link failures, in comparison to those IGP link weight optimization approaches that only consider intra-AS link failure. **2. The post-failure routing changes of hot-potato routing for inter-AS traffic should not be neglected when making changes to IGP link weights for TE.** We found that even if we make the TE approach robust to link failure, its performance may be offset by ignoring the effect of HPR which could change the originally optimized egress points for some inter-AS traffic flows and their IGP routes after link failures. This infers that not only inter-AS transient link failures but also the post-failure routing changes of HPR should be considered in a robust TE scheme.

In the next section, we further explain the TE and link failure interactions with an illustrative example. Section III presents the problem formulation of the joint robust TE approach. Then we detail our proposed two-phase heuristic in Section IV. In Section V, we review the four IGP link weight optimization approaches that will be used for our performance comparison. We present evaluation methodology and results in Sections VI and VII respectively. Section VIII provides a brief survey of related work. Finally, we conclude the paper in Section IX.

## II. Illustrative Example of Interactions

In Figures 1a-1e we illustrate how the aforementioned interactions, if not taken into account, can affect the robustness of the overall TE performance in terms of link failure. The performance metric we use is the intra- and inter-AS MLU under Normal State (NS) and some Failure States (FSs) where each FS corresponds to a single link failure. Link utilization is calculated as the total traffic load on the link divided by its bandwidth capacity. The intra-AS (or inter-AS) MLU under state *s* is the highest utilization among all the operational intra-AS (or inter-AS) links under that state.

The network in Figure 1 consists of three egress points (*j1*, *j2* and *j3*) with equal egress link capacity of 100 Mbps, two ingress points *i1* and *i2*, inter-AS traffic flows *t1=t_inter(i1,k1)*=40Mbps,*t2=t_inter(i1,k2)*=40Mbps,*t3=t_inter(i2,k3)*=20Mbps and remote destination prefixes *k1*, *k2* and *k3*, where *t_inter(i,k)* denotes the inter-AS traffic flow that enters the network from ingress point *i* and destined at prefix *k*. In this example, we assume that *k1* can be reached through all the egress points while *k2* can only be reached through *j2* and *k3* can be reached through *j1* and *j3* only. The network has several intra-AS links between ingress and egress points. The value on each link represents the IGP link weight. The capacity of bold links is 200Mbps while the capacity of the rest of the links is 100Mbps.

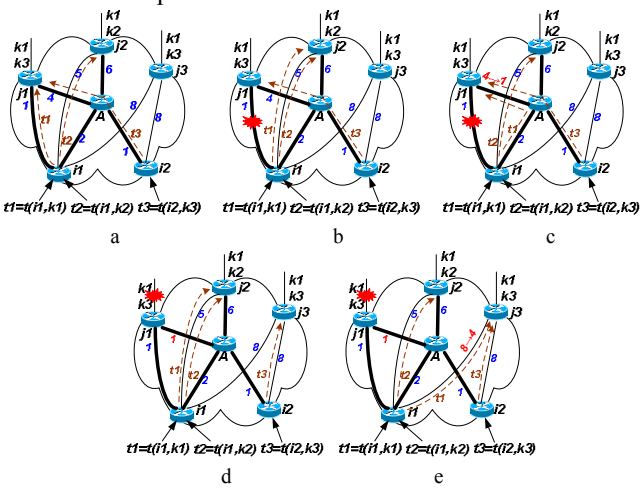


Figure 1. Traffic demand assignment under (a) NS, (b) *i1-j1* FS, (c) *i1-j1* FS with a changed IGP link weight, (d) *j1* FS, (e) *j1* FS with a changed link weight.

Note that throughout this paper we only consider the egress points that have "equally good" BGP routes towards each destination prefix. Therefore the egress point selection for the inter-AS traffic is determined by the IGP distance between individual ingress/egress pairs according to HPR. This scenario is inline with the fact that current ISPs often use HPR to control their inter-AS egress traffic [23].

Figure 1a shows the assignment of traffic flow *t1*, *t2* and *t3* to egress points *j1*, *j2* and *j1* respectively under NS. In this assignment, the inter- and intra-AS MLU would be on inter-AS link *j1* and intra-AS link *i1-j2* respectively and would be equal to ((40+20)/100,40/100)=(0.6,0.4).

Figure 1b shows the traffic flow assignment when intra-AS link *i1-j1* fails (i.e. *s={i1-j1}*). This failure disrupts the inter-AS traffic flow *t1* and shifts its egress point from *j1* to *j2* due to the HPR. The inter- and intra-AS MLU would then become ((40+40)/100,(40+40)/100)=(0.8,0.8) on inter-AS link *j2* and intra-AS link *i1-j2* respectively. Hence, the failure leads to an increase in the utilization of both intra- and inter-AS links.

However, this increased link utilization can be avoided if the IGP link weight of *A-j1* was set to 1. As shown in Figure 1c, when the intra-AS link *i1-j1* fails, the egress point of *t1* would not change and the inter and intra-AS MLU would be reduced to ((40+20)/100,40/100)=(0.6,0.4). Hence, an appropriate IGP link weight setting can avoid increase in the link utilization and change of egress points for the inter-AS traffic.

Figure 1d shows the traffic assignment when inter-AS link *j1* fails (i.e. *s={j1}*). This failure shifts *t1* and *t3* from *j1* to *j2* and *j3* respectively. The shifting of traffic increases both the inter- and intra-AS MLU, which would become ((40+40)/100, (40+40)/100)=(0.8,0.8). Note that, in this case, change of the egress point due to HPR and disruption of *t1* and *t3* are inevitable, since the egress point *j1* has no reachability to *k1* anymore. By comparing Figures 1c and 1d, we observe that, even though the overall network utilization under a failure of intra-AS link has been improved by an IGP link weight change, it remains poor when an inter-AS link fails.

Nevertheless, such poor overall network utilization would not happen if the IGP link weight of *i1-j3* was set to 4. As shown in Figure 1e, when the inter-AS link *j1* fails, the inter- and intra-AS MLU would become ((40+20)/100,40/100)=(0.6,0.4), which is identical to the results achieved under NS.

From this example, we can see that intra- and inter-AS link utilization can be improved with a set of appropriately configured link weights that takes into account both intra- and inter-AS transient link failures as well as the routing changes effects of HPR; that is the issue we investigate in this paper.

## III. JOINT ROBUST TE PROBLEM FORMULATION

### A. Inputs

*1)* ***Traffic Matrix (TM)*****:** this represents a matrix of traffic demand from each network point to each other over some time interval. In general, three types of traffic matrix can be identified in ISP networks. First of all, each element of the inter-AS traffic matrix, *t_inter(i,k)*, represents the total volume of inter-AS traffic from ingress point *i* towards destination prefix *k* that is reached through a downstream AS. Secondly, some traffic is destined locally within the network and we call this local traffic. Therefore, each element of this local traffic matrix, *t_loc(i,j)*, represents a volume of traffic from ingress point *i* destined to egress access point *j*. Finally, each element of the intra-AS traffic matrix, *t_intra(i,j)*, represents the total volume of intra-AS traffic from ingress point *i* destined to egress point *j*. Therefore, intra-AS traffic covers all the traffic that traverses the network including both the inter-AS traffic and local traffic. Thus, each element of the intra-AS traffic is the sum of local intra-AS and inter-AS traffic volume between each pair of ingress and egress nodes.

*2)* ***Network Topology*****:** this contains information about the connectivity of intra-, inter-AS nodes and link capacity.

*3)* ***Reachability of Destination Prefixes*****:** this consists of the advertisements of destination prefixes received by each egress point. This reachability information can identify which destination prefix can be reached through which egress points and it may be obtained from the BGP routing information base (Adj-RIB-In) of each egress router.

### B. Problem Formulation

Given the inputs, the objective of the joint robust TE is to minimize the intra- and inter-AS Maximum Link Utilization (MLU) under NS and also to minimize the worst-case intra- and inter-AS MLU across all intra- and inter-AS FSs. Each intra-AS (or inter-AS) FS corresponds to the network with a specific intra-AS (or inter-AS) link failure. By all states we include the NS as well as all intra and inter-AS FSs. We denote intra-, inter-AS FSs and all states by $S^{Intra}$ , $S^{Inter}$ and $S^{All}$ respectively (i.e. $S^{Intra}=\{\forall l\in L\}$, $S^{Inter}=\{\forall j\in J\}$, $S^{All}=\{\varnothing\cup(\forall l\in L)\cup(\forall j\in J)\}$). As mentioned earlier, the intra-AS (or inter-AS) MLU under state *s* is defined as the highest utilization among all the operational intra-AS (or inter-AS) links under that state.

Also, the worst-case intra-AS (or inter-AS) MLU across all states is the highest utilization among the MLU of all intra-AS (or inter-AS) states.

To achieve our objective, the optimization problem is to compute a set of IGP link weights that by taking the HPR into account determines the routes between each pair of ingress and egress points as well as the egress points for inter-AS traffic. Prior to the problem formulation, we introduce the following notation: *I* and *J* are the set of ingress and egress points, *K* is the set of destination prefixes, *Out(k)* is the set of egress points that has routing reachability to prefix $k\in K$, and finally $W=(w_1,w_2,\ldots,w_l,\ldots,w_n)$ is a vector of IGP link weights where $w_l$ is the weight of link *l*.

We define $x^l_{(i,j)}(s,W)$ as a binary variable and its value is equal to 1 if intra-AS traffic flow *t_intra(i,j)* traverses intra-AS link *l* under state *s* with IGP link weight setting *W* and 0 otherwise. The worst-case intra-AS MLU across all states can be formulated as follows:

$$\underset{W}{Minimize}\, U^{intra}_{worst\_AllStates}=\underset{W}{Minimize}\,\underset{\forall s\in S^{All}}{Max}\, U^{intra}_{max}(s) \quad (1)$$

where

$$\forall s\in S^{All}: U^{intra}_{max}(s)=\underset{\forall l\neq s}{Max}(u^l_{intra}(s,W))=\underset{\forall l\neq s}{Max}\left(\frac{\sum_{\forall i\in I}\sum_{\forall j\in J} x^l_{i,j}(s,W).t\_intra(i,j)}{c^l_{intra}}\right) \quad (2)$$

$c^l_{intra}$ denotes the capacity of intra-AS link *l* and $u^l_{intra}(s,W)$ represents the utilization of *l* under state *s* with IGP link weight setting *W*. Note that the intra-AS MLU under NS ($U^{intra}_{max\_NS}$) can be calculated by (2) if state *s* represents only NS (i.e. $s=\varnothing$). If the failure states are limited to only intra-AS link failure (i.e. $s\in S^{Intra}$) then the expression in (1) represents the worst-case intra-AS MLU across only all intra-AS FSs (i.e. $U^{intra}_{worst\_IntraFSs}$). Similarly, if the failure states are limited to only inter-AS link failures (i.e. $s\in S^{Inter}$) then the expression in (1) represents the worst-case intra-AS MLU across only all inter-AS FSs (i.e. $U^{intra}_{worst\_InterFSs}$). In other words:

$$U^{intra}_{max\_NS}=U^{intra}_{max}(\varnothing) \quad (3)$$

$$U^{intra}_{worst\_IntraFSs}=\underset{\forall s\in S^{Intra}}{Max}\, U^{intra}_{max}(s) \quad (4)$$

$$U^{intra}_{worst\_InterFSs}=\underset{\forall s\in S^{Inter}}{Max}\, U^{intra}_{max}(s) \quad (5)$$

Clearly the worst-case intra-AS MLU under all FSs can be obtained as follows:

$$U^{intra}_{worst\_AllFSs}=\underset{\forall s\in S^{All}-\{\varnothing\}}{Max}(U^{intra}_{worst\_IntraFSs},U^{intra}_{worst\_InterFSs})=\underset{\forall s\in S^{All}-\{\varnothing\}}{Max}\, U^{intra}_{max}(s) \quad (6)$$

In a similar manner to the above robust intra-AS TE problem formulation, we define $y^j_{(i,k)}(s,W)$ as a binary variable and its value is equal to 1 if inter-AS traffic flow *t_inter(i,k)* is assigned to egress point *j* under state *s* with IGP link weight setting *W* and 0 otherwise. Hence, the worst-case inter-AS MLU across all states can be formulated as

$$\underset{W}{Minimize}\, U^{inter}_{worst\_AllStates}=\underset{W}{Minimize}\,\underset{\forall s\in S^{All}}{Max}\, U^{inter}_{max}(s) \quad (7)$$

where

$$\forall s\in S^{All}: U^{inter}_{max}(s)=\underset{\forall j\neq s}{Max}(u^j_{inter}(s,W))=\underset{\forall j\neq s}{Max}\left(\frac{\sum_{\forall i\in I}\sum_{\forall k\in K} y^j_{i,k}(s,W).t\_inter(i,k)}{c^j_{inter}}\right) \quad (8)$$

$c^j_{inter}$ denotes the capacity of inter-AS egress link *j* and $u^j_{inter}(s,W)$ represents the utilization of *j* under state *s* with IGP link weight *W*. Similar to (3) to (6) for the inter-AS utilization we have

$$U^{inter}_{max\_NS}=U^{inter}_{max}(\varnothing) \quad (9)$$

$$U^{inter}_{worst\_IntraFSs}=\underset{\forall s\in S^{Intra}}{Max}\, U^{inter}_{max}(s) \quad (10)$$

$$U^{inter}_{worst\_InterFSs}=\underset{\forall s\in S^{Inter}}{Max}\, U^{inter}_{max}(s) \quad (11)$$

$$U^{inter}_{worst\_AllFSs} = \underset{\forall s \in S^{All}-\{\emptyset\}}{Max}(U^{inter}_{worst\_IntraFSs}, U^{inter}_{worst\_InterFSs}) = \underset{\forall s \in S^{All}-\{\emptyset\}}{Max} U^{inter}_{max}(s) \quad (12)$$

Therefore, the problem of our joint robust TE can be formulated as follows:

$$\underset{W}{Minimize}(U^{intra}_{max\_NS}, U^{intra}_{worst\_AllFSs}, U^{inter}_{max\_NS}, U^{inter}_{worst\_AllFSs}) \quad (13)$$

subject to the following constraints:

$$\forall i,i' \in I, k \in K, s \in S, g \in Out(k): \sum_{j \in Out(k)/Q(i,g,k)} y^{j}_{i,k}(s,W) + \sum_{j' \in Q(i,g,k)} y^{j'}_{i',k}(s,W) \le 1 \quad (14)$$

$$\forall j \in J, i \in I, k \in K, s \in S \text{ if } y^{j}_{i,k}(s,W)=1 \text{ then } j \in Out(k) \quad (15)$$

$$\forall i \in I, k \in K, s \in S: \sum_{j \in Out(k)} y^{j}_{i,k}(s,W)=1 \quad (16)$$

$$\forall j \in J, i \in I, k \in K, s \in S: y^{j}_{i,k}(s,W) \in \{0,1\} \quad (17)$$

Constraint (14) is the proximity constraint [11], which ensures that the HPR is obeyed. In (14) $Q$ is a utility function that is used to specify, for a given ingress node $i$ and egress link $j$ and a given prefix $k$, the set of alternative egress links for $k$ that are closer than $j$. Thus $Q(i,j,k)$ is defined as the set of edge links where $Q(i,j,k)=\{g | g \in Out(k) \wedge d(i,g)<d(i,j)\}$. For more clarification refer to [11]. Constraint (15) ensures that if the traffic flow from ingress point $i$ destined to prefix $k$ is assigned to egress point $j$ under state $s$, then this prefix must be reachable through that egress point. Constraints (16) and (17) ensure that the traffic flow from ingress point $i$ to prefix $k$ is assigned to only one egress point that has routing reachability to this prefix under state $s$ (i.e. there is no traffic splitting).

According to (13), our joint robust TE is a complex quadruple-objective optimization problem. To simplify the problem, we first categorize these four objectives into two wider objectives at intra- and inter-AS levels. We therefore have the joint robust TE problem reduced to a bi-objective optimization problem as follows:

$$\underset{W}{Minimize}(U^{intra}_{max\_NS}, U^{intra}_{worst\_AllFSs}) \quad (18)$$

$$\underset{W}{Minimize}(U^{inter}_{max\_NS}, U^{inter}_{worst\_AllFSs}) \quad (19)$$

However, these two objectives may be in conflict: intra-AS resource utilization may only be improved at the expense of degradation in the utilization of inter-AS resources and vice versa. Consequently, we need to further simplify the problem in order to eliminate such conflict. We therefore resort to using the $\epsilon$-constraint method [12], in which the performance of an objective is optimized while the other one is constrained by not exceeding a tolerance value. Now the important question is which one of these objectives should be a constraint? There is anecdotal evidence [11] that inter-AS links are often bottleneck links in the Internet and significant amount of Internet traffic such as peer to peer traffic is routed across these links [20]. In addition, an inter-AS link is relatively more difficult to upgrade compared to an intra-AS link due to time-consuming and complicated negotiation between two ASes. It is also important to ensure that traffic exchange limits on peering agreements with downstream ASes are not violated. For these reasons, we place a constraint on the robust inter-AS TE objectives.

By placing a constraint on the utilization of inter-AS resources, the intra-AS resource utilization has to be optimized. However, this objective itself also consists of two conflicting objectives [4,5,16]: improving the worst-case intra-AS MLU under all FSs may lead to performance degradation in the intra-AS MLU under NS. To further simplify the problem, we adopt a weighted sum approach to transform these two intra-AS objectives into one. Therefore, the optimization problem of the joint robust TE can be formulated as follows:

$$\underset{W}{Minimize}(U^{intra}_{max\_NS}, U^{intra}_{worst\_AllFSs}) = \underset{W}{Minimize}((1-\alpha)U^{intra}_{max\_NS} + \alpha U^{intra}_{worst\_AllFSs}) \quad (20)$$

where $0 \le \alpha \le 1$, subject to the inter-AS utilization constraint:

$$U^{inter}_{worst\_AllStates} \le \varepsilon \quad (21)$$

where $0 < \varepsilon \le 1$. The constraint ensures that the inter-AS MLU across all states is less than $\varepsilon$. Since $U^{inter}_{worst\_AllStates}$ can be calculated as follows

$$U^{inter}_{worst\_AllStates} = \underset{\forall s \in S^{All}}{Max}(U^{inter}_{max\_NS}, U^{inter}_{worst\_IntraFSs}, U^{inter}_{worst\_InterFSs}) \quad (22)$$

the above constraint implies that

$$U^{inter}_{max\_NS} \le \varepsilon \quad (23)$$

$$U^{inter}_{worst\_IntraFSs} \le \varepsilon \quad (24)$$

$$U^{inter}_{worst\_InterFSs} \le \varepsilon \quad (25)$$

According to the above problem formulation, we aim to optimize the intra-AS MLU under NS and the worst-case MLU among all intra-AS FSs while respecting the inter-AS utilization constraint across all states. Since optimizing the intra-AS MLU for both NS and FSs has been proven to be NP-hard [4,5,16] and adding the inter-AS utilization constraint makes the problem even more complicated, we resort to heuristics to solve the problem efficiently.

## IV. Proposed Two-phase Heuristic

We propose a two-phase heuristic. The first phase consists of a local search algorithm to find an initial set of IGP link weights that satisfies the inter-AS utilization constraint (21). Based on this set of IGP link weights, in the second phase, we optimize the link weights towards intra-AS TE objective (20) while preserving the inter-AS utilization constraint.

### A. Phase I

The local search algorithm in phase 1 consists of three steps:

***Step 1***. *Initialization*: generate an initial solution ($W^{initial}$) by setting the weight of each link inversely proportional to its capacity. Run Dijkstra's SPF algorithm for $W^{initial}$ while taking into account HPR to determine the egress points for inter-AS traffic and the IGP routes between each pair of ingress and egress points. Calculate the initial worst-case inter-AS MLU under all states ($U^{inter\_initial}_{worst\_AllStates}$) using (22). Initialize the current solution ($W^{current} = W^{initial}$) and update the current performance metric ($U^{inter\_current}_{worst\_AllStates} = U^{inter\_initial}_{worst\_AllStates}$). If this value is less than the value of $\varepsilon$, then terminate the local search algorithm by returning the current IGP link weights as an input to the algorithm in phase II; otherwise proceed to steps 2 and 3.

***Step 2***. *Neighborhood search*: a move is applied to transform the current solution into a neighbor solution. Perform a move by randomly picking up a link and increase or decrease its weight by a random value. Re-run Dijkstra's SPF algorithm for this new set of IGP link weights taking into account the HPR. Calculate the worst-case inter-AS MLU under all states ($U^{inter\_new}_{worst\_AllStates}$). If the new solution yields lower utilization than the current solution (i.e. $U^{inter\_new}_{worst\_AllStates} < U^{inter\_current}_{worst\_AllStates}$), accept the move by updating the current IGP link weights and performance

metric ($W^{current} = W^{new}$, $U^{inter\_current}_{worst\_AllStates} = U^{inter\_new}_{worst\_AllStates}$); otherwise repeat this step until such a solution is found.

***Step 3***. *Check stopping criterion*: repeat step 2 for the next iteration until the current worst-case inter-AS MLU under all states ($U^{inter\_current}_{worst\_AllStates}$) is less than the value of $\varepsilon$. However, if there is no significant improvement on $U^{inter\_current}_{worst\_AllStates}$ after a certain number of iterations, this means that the algorithm is unlikely to find solutions that satisfy the desired inter-AS utilization constraint, possibly due to high amount of traffic load. In this case, we have to increase the value of $\varepsilon$ by a step value denoted by $c$. In other words, $\varepsilon_{new} = \varepsilon + n \times c$, where $n$ is a positive integer value, acts as a coefficient for the step value. The increase in the value of $\varepsilon$ by coefficient $n$ continues until a solution that satisfies the constraint is found. Once the relaxed constraint is satisfied, terminate the local search algorithm by returning the current IGP link weights as an input to the intra-AS TE optimization in phase II.

### B. Phase II

Our algorithm in phase II follows the Tabu Search (TS) technique [14] with the following components:

*1) Neighborhood search:* we perform the following steps to identify the best move in the neighborhood:

***Step 1***. Identify two sets of intra-AS links – those whose utilizations are within a small percentage of the MLU (heavily utilized) and those whose utilizations are within a small percentage of the minimum link utilization (lightly utilized). Take the most utilized link in the first set into consideration.

***Step 2***. Increase the weight of the link by a random value in an attempt to remove the traffic from that link and reduce its load. Select a link randomly from the lightly utilized link set and decrease its weight by a random value in attempt to attract more traffic over this link from the highly utilized links.

***Step 3***. Run Dijkstra's SPF algorithm for the current IGP link weights with the HPR to re-calculate the egress points for the inter-AS traffic and the IGP routes for the intra-AS traffic. Then calculate objective function (20) and constraint (21).

***Step 4.*** Repeat step 3 until either a feasible solution that satisfies the constraint is found or the upper limit of repetition is reached.

***Step 5.*** Select the next most utilized intra-AS link and repeat steps 2 to 5 until all the links in the heavily utilized link set have been considered.

***Step 6.*** Among all feasible solutions, choose the one with the minimum intra-AS MLU and consider it as the current solution.

*2) Tabu list*: The tabu list memorizes the most recent moves, operating as a first-in-first-out queue. As suggested in [14], the size of the tabu list depends on the size and characteristics of the problem. In our problem, the tabu list consists of the links whose weights have been recently changed and the amount of increase/decrease applied to the corresponding link weight.

*3) Diversification*: The goal of diversification is to prevent the searching procedure from indefinitely exploring a region of the solution space that consists of only poor quality solutions. It is a modification of the neighborhood search and is applied when there is no obvious performance improvement after a certain number of iterations. For a diversification, several links are picked up from each of the lightly and heavily utilized link sets. The weights of the selected links from the former set are decreased while the weights of the selected links from the latter set are increased. Note that any solution produced by the diversification is acceptable if it is feasible.

*4) Stopping Criterion*: the search procedure stops if either the pre-defined maximum number of iterations is reached or there is no pre-defined performance improvement for objective function (20) after a certain number of consecutive diversifications.

## V. Alternative Approaches

We compare our joint robust TE with four alternative IGP link weight optimization approaches. The characteristics of these approaches are illustrated in Table1.

TABLE 1: VARIOUS IGP WEIGHT OPTIMIZATION APPROACHES

| Approach | TE for Normal State? | Robust to intra-AS link failure? | Consider HPR? | Robust to inter-AS link failure? |
|---|---|---|---|---|
| InvCap | No | No | No | No |
| Intra-As-Te | Yes | No | No | No |
| Intra-As-RobustTe | Yes | Yes | No | No |
| Intra-As-RobustBgpTe | Yes | Yes | Yes | No |
| Joint-RobustTe | Yes | Yes | Yes | Yes |

***1)*** **InvCap**: as often used by vendors, the IGP link weights are set inversely proportional to the link capacity.

***2)*** **Intra-As-Te**: the IGP link weights are optimized to achieve intra-AS load balancing only under NS. A notable work in this area is [15]. However, it aims to minimize a piece-wise linear cost function which is not easily comparable with our objective function (20). For ease of comparison, we consider the objective of this approach also to be minimizing the intra-AS MLU under NS:

$$\underset{W}{Minimize}\, U^{intra}_{max\_NS} \tag{26}$$

We adopt the Tabu Search heuristic proposed in [4] for this approach and modify its link weight optimization only for NS.

***3)*** **Intra-As-RobustTe**: the IGP link weights are optimized to achieve intra-AS load balancing under both NS and intra-AS FSs. The objective of this approach can be formulated as:

$$\underset{W}{Minimize}(U^{intra}_{max\_NS}, U^{intra}_{worst\_IntraFSs}) = \underset{W}{Minimize}((1-\beta)U^{intra}_{max\_NS} + \beta U^{intra}_{worst\_IntraFSs}) \tag{27}$$

where $0 \le \beta \le 1$. A notable work in this area is [16] with the consideration of an SLA constraint. We adopt their heuristic for this approach but without considering the SLA constraint. Note that since neither this approach nor **Intra-As-Te** account for the HPR effect, the egress points of inter-AS traffic are assumed to be fixed whenever IGP link weight is changed.

***4)*** **Intra-As-RobustBgpTe**: the link weights are optimized to achieve intra-AS load balancing under both NS and intra-AS FSs (the same as the problem formulation (27)) while taking into account the HPR. The closest related work to this approach is the METL-BGP TE tool [9]. However, they do not consider the impacts of inter-AS link failure on the overall network utilization. To implement this approach, we extend the heuristic in [16] by incorporating the HPR.

## VI. Evaluation Methodology

### A. Network Topology and Destination Prefixes

Our experiments were performed on two Point-of-Presence (POP) level topologies generated by BRITE [17]. The two POP level topologies have 50 nodes with 100 links and 100 nodes with 200 links. In each topology, all POP nodes are ingress points while only some of them, namely border POPs, are connected to adjacent provider ASes through inter-AS links and hence they can be both ingress and egress points. A similar network setup is also found in some ISP POP topologies provided by Rocketfuel [22]. We notice that the number of border POPs in these topologies is about half of the total POP nodes. Therefore, without loss of generality, we randomly select half of the POP nodes as border POP nodes each with only one inter-AS link. We also assume a homogenous environment in which the capacity of all the intra- and inter-AS links are OC-192 (9.6 Gbps) and OC-48 (2.5 Gbps) respectively.

For scalability and stability reasons, the joint robust TE can focus only on a small fraction of Internet destination prefixes, which are responsible for a large fraction of the Internet traffic [1]. In line with [11], we consider 1000 popular destination prefixes. In fact, each of them may not merely represent an individual prefix but also an aggregate of multiple destination prefixes that have the same set of candidate egress points [18]. This simplifies the problem by significantly reducing the number of prefixes to be considered. Nevertheless, the number of prefixes we consider could actually represent an even larger value of actual prefixes.

We assume that each border POP has reachability to all the considered destination prefixes. Therefore, during NS, the inter-AS traffic received at a border POP towards any destination prefix will exit the network through the same border POP without traversing the network. However, if the inter-AS link attached to this border POP fails, the inter-AS traffic will have to be routed within the network and then exit from another border POP.

### B. Traffic Matrices

We generate synthetic traffic matrices for our experiments. According to [18], inter-AS traffic volumes are top-heavy and follow the Weibull distribution with shape parameter 0.2-0.3. We therefore generate the inter-AS TM with this distribution using the shape parameter of 0.3. In addition, following the methodologies in [16], we generate local intra-AS TM using the Gravity Model (GM). In this model, the amount of incoming traffic at a POP is proportional to its size. Following the suggestions in [19], we randomly classify 40% of POPs as "small", 40% as "medium" and 20% as "big".

### C. Weighting parameter

By varying the weight parameters $\alpha$ and $\beta$ in objective functions (20) and (27) respectively and re-solving them, one can generate a trade-off curve between the two objectives of each function using the method of multi-objective programming [12]. If we solve the problem with $\alpha=0$ (or $\beta=0$), the problem is simply reduced to the intra-AS TE optimization for only NS. If $\alpha=1$ (or $\beta=1$), the problem completely ignores the performance under NS and only optimizes the worst-case intra-AS TE performance across all FSs. While a specific value of $\alpha$ (or $\beta$) allows us to achieve a balance between the two objectives, the most suitable value depends on the combination of network topology and traffic matrix.

### D. Constraint value and our heuristic parameters

For the local search algorithm, we start with $\varepsilon=0.1$ for the inter-AS utilization constraint in (21) (i.e. the load on each inter-AS link should not exceed 10% of its capacity). However, if no solution that satisfies the constraint can be found, we step up the value by $c=0.1$ to relax the constraint. In this case, it becomes $\varepsilon_{new} = \varepsilon + n \times c = 0.1 + 1 \times 0.1 = 0.2$. If the algorithm remains unable to find a feasible solution, this value is then gradually increased by $n \times c$ until such a solution is found. ISPs can set the constraint and step values based on their desired operational objectives.

According to our experiments we realized that by setting our heuristic parameters to the following values we can achieve sufficiently good results: In the local search, the constraint value is increased if the utilization improvement is less than 2% after 20 iterations. For tabu search, the size of tabu list is set to 20, the threshold of utilization improvement for diversification is set to 5% of the best visited solution after 20 iterations. The stopping criterion is satisfied if either the search procedure reaches 5 times the total number of considered destination prefixes or the utilization improvement is less than 5% of the best visited solution after 10 consecutive diversifications.

## VII. Evaluation Results

In this section we present our evaluation results. All the results presented in this paper are the average of 10 trials with independent network topologies and traffic matrices. Note that all the results with $MLU>1.0$ are not achievable. However they are illustrated for comparison purpose.

### A. Intra-AS MLU under NS

Figures 2a-2b show intra-AS MLU for the 50-POP and 100-POP topologies under NS. This metric refers to $U^{intra}_{max\_NS}$ in (3) or objective function (20). The x-axis represents the normalized intra-AS offered load, i.e. the total intra-AS traffic volume normalized by the total intra-AS capacity.

From both figures we observe that **InvCap** is the worst performer, which is expected since it does not perform link weight optimization for achieving load balancing. **Intra-As-Te** and **Intra-As-RobustTe** perform better than **InvCap** but worse than the other two. In fact, even though these approaches aim to minimize the intra-AS MLU under NS or FSs according to their objective functions (26) and (27) respectively, they do not take the effects of HPR into account in their IGP link weight optimization. As a result, the actual routing of traffic in the network can be different from what was produced from the optimization, which may result in sub-optimal performance. With the explicit consideration of HPR, the joint robust TE and **Intra-As-RobustBgpTe** approaches outperform the others. However, the joint robust TE approach performs slightly worse (about 9%-10% for 50-POP and 8%-10% for 100-POP) than **Intra-As-RobustBgpTe**. This is because it attempts to optimize the intra-AS MLU under the inter-AS utilization constraint,

whereas **INTRA-AS-ROBUSTBGPTE** does not consider it. Adding such constraint reduces the number of feasible candidate egress points and therefore leads to fewer available IGP routes that can be selected by the traffic. This may result in the situation where many traffic flows traverse the same link, thereby significantly increasing its utilization. Nevertheless, as will be shown in the following sections, the joint robust TE significantly improves the intra- and inter-AS MLU under FSs at this small cost of performance degradation under NS.

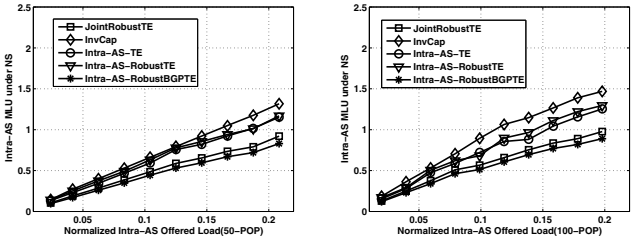


Figure 2a and 2b. Intra-AS MLU under NS

### B. *Intra-AS MLU under intra- and inter-AS FSs*

Figure 3 shows the worst-case intra-AS MLU across all intra- and inter-AS FSs. These metrics correspond to $U^{intra}_{worst\_IntraFSs}$ in (4) and $U^{intra}_{worst\_InterFSs}$ in (5) respectively.

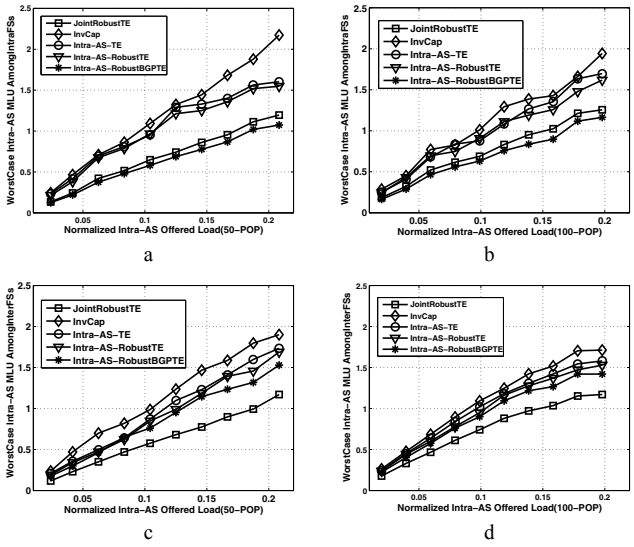


Figure 3. WorstCase intra-AS MLU across (a,b) intra-AS FSs, (c,d) inter-AS FSs for (50,100) POP

Figures 3a-3b show that **INVCAP** and **INTRA-AS-TE** appear to have the worst performance across all intra-AS FSs since they were not designed to be robust against intra-AS link failures. After these two approaches, **INTRA-AS-ROBUSTTE** has the worst performance due to the ignoral of HPR effects as we have explained in the previous section. This reveals that HPR is an essential consideration in the robust TE design. Therefore, **INTRA-AS-ROBUSTBGPTE** is the best performer in this case. Compared to it, the joint robust TE approach has slightly higher (about 7%-11% for 50-POP and 8%-13% for 100-POP) intra-AS MLU. This is because **INTRA-AS-ROBUSTBGPTE** optimizes only for *intra-AS FSs* whereas the optimization objective of the joint robust TE covers not only intra- but also inter-AS FSs. The two set of FSs may conflict with each other: reducing the intra-AS link utilization under intra-AS FSs may increase the utilization under inter-AS FSs. As a result, we may not be able to obtain the best intra-AS MLU in exchange for achieving a compromised solution for inter-AS FSs, and this is explained next. Figures 3c-3d show that the joint robust TE is the best performer regarding the worst-case intra-AS MLU across all inter-AS FSs (about 23%-33% for 50-POP and 17%-21% for 100-POP better than **INTRA-AS-ROBUSTBGPTE**, the second best approach). The reason is that it is the only TE approach that is designed to be robust against inter-AS link failures. Failure of inter-AS links can cause egress point changes and reroute the traffic through highly utilized parts of the network which overloads some intra-AS links. This explains why the four alternative approaches perform significantly worse than the joint robust TE approach under inter-AS link failures. In fact, in terms of the worst case intra-AS MLU across all FSs, i.e. $U^{intra}_{worst\_AllFSs}$ in (6), joint robust TE is about 11%-25% for 50-POP and 11%-20% for 100-POP better than **INTRA-AS-ROBUSTBGPTE**, the second best approach.

### C. *Inter-AS MLU under NS, intra- and inter-AS FSs*

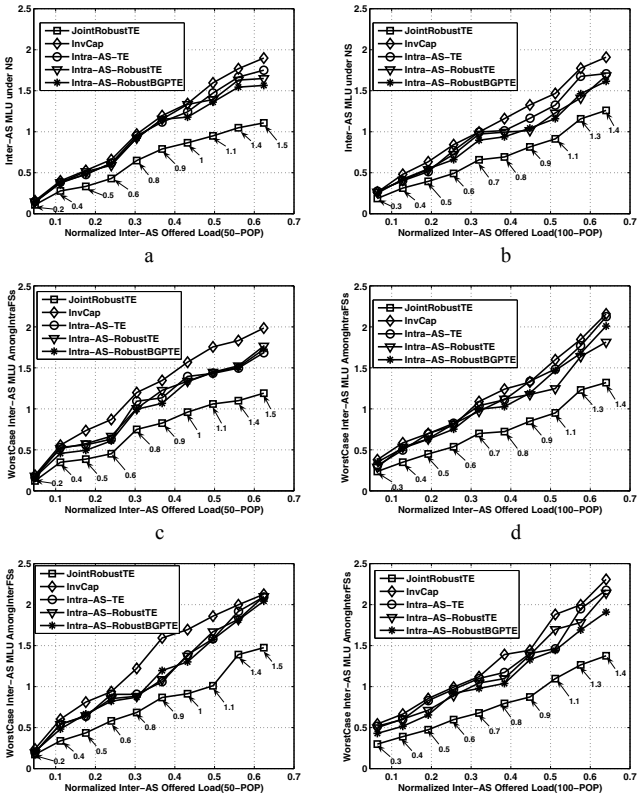


Figure 4. (a,b) Inter-AS MLU under NS, (c,d) WorstCase inter-AS MLU across intra-AS FSs and (e,f) WorstCase inter-AS MLU across inter-AS FSs for (50,100)

Figure 4 shows the inter-AS MLU. More specifically, Figures 4a-4b, 4c-4d and 4e-4f refer to the performance metrics $U^{inter}_{max\_NS}$ in (9), $U^{inter}_{worst\_IntraFSs}$ in (10) and $U^{inter}_{worst\_InterFSs}$ in (11) respectively. The x-axis represents the normalized inter-AS offered load, i.e. the total inter-AS traffic volume normalized by the total inter-AS link capacity. The values indicated by arrows are the inter-AS utilization constraint values (i.e. $\varepsilon$). A general observation of the figures is that if the TE approach considers neither inter-AS load balancing under NS nor impacts of link failure on the

utilization of inter-AS resources, like those four alternative approaches, a significant amount of traffic may be unpredictably assigned to some egress points and possibly cause severe congestion there. By comparing between Figures 4c-4d and 4e-4f, we found that intra- and inter-AS link failures equally contributed to the high utilization of inter-AS links. Hence, the robust TE approaches that neglect either intra- or inter-AS link failures may not make their performance truly robust. On the contrary, by considering both intra- and inter-AS link failures along with HPR, our joint robust TE approach improves all the performance metrics. In fact, in terms of the worst case inter-AS MLU across all FSs, i.e. $U^{inter}_{worst\_AllFSs}$ in (12), joint robust TE is about 17%-34% for 50-POP and 23%-35% for 100-POP better than **INTRA-AS-ROBUSTBGPTE**, the second best approach.

As mentioned in Section VI.D, we start with $\varepsilon = 0.1$. However the local search cannot find a feasible solution that satisfies the constraint until $\varepsilon$ is increased to 0.2 and 0.3 for the 50 and 100-POP topologies respectively. Note that, in practice, all the results with $\varepsilon > 1$ are undesirable due to egress point overload and potential packet losses. Nevertheless, even under this situation, the amount of overload is much smaller than the other alternative approaches.

### D. Overall Performance

At the cost of a small performance degradation of the intra-AS MLU under NS, the joint robust TE approach significantly outperforms the other alternatives in terms of the worst-case intra- and inter-AS MLU across all FSs.

For those alternative approaches, **INVCAP** performs the worst in all the performance metrics. Although **INTRA-AS-TE** and **INTRA-AS-ROBUSTTE** have considered optimization for NS and intra-AS FSs, they can only perform better than **INVCAP** due to the ignoring of both HPR effects and complete link failure scenarios. Clearly, **INTRA-AS-ROBUSTBGPTE** attempts to improve these deficiencies by incorporating the effects of HPR. However, it does not perform well compared to our joint robust TE approach due to the ignoring of inter-AS link failures and HPR impact on the overall network resource utilization. In summary, based on the improved performance of the joint robust TE approach, we suggest that for the robust TE design: (1) intra- and inter-AS transient link failures should be considered together, and (2) the routing changes of hot-potato routing under normal and post-failure states should not be neglected when making changes to IGP link weights.

## VIII. RELATED WORK

Prior intra-AS robust TE proposals [4,5,16] have computed a set of IGP link weights that is robust to failure of either any single link or critical intra-AS links. Moreover, some recent inter-AS robust TE work [6,7] have considered inter-AS link failure in their outbound TE optimization. Since all these methods have solely optimized either intra- or inter-AS TE objectives, their overall TE performance may be suboptimal or even very poor in case of any intra or inter-AS link failure. On the other hand, [8] has investigated the interactions between intra- and inter-AS TE and proposed a joint optimization. In addition, [9,10] have evaluated the behavior of HPR during the IGP link weight optimization. Nevertheless, none has investigated the impact of intra-AS link failure on inter-AS outbound TE as well as the impact of inter-AS link failure on intra-AS TE, which is the major difference between our work and the existing literature.

## IX. CONCLUSION

In this paper, we first investigated the interactions between intra (inter)-AS link failure and inter (intra)-AS TE and showed how this may be detrimental. To mitigate the interactions, we proposed a joint robust TE approach that optimizes intra-AS link utilization while preserving the inter-AS link utilization under both normal state as well as single intra- or inter-AS link failure states. By taking HPR into account, our joint robust TE approach optimizes the IGP link weights to achieve both intra-AS TE and inter-AS outbound TE under all the states. We solved the problem by a two-phase heuristic and compared its performance to four alternative approaches. Our evaluation results show that our approach achieves high robustness of TE performance against transient link failures. The other alternative approaches, however, do not satisfy all these objectives at the same time and hence their performance is less robust to link failures.

## REFERENCES


[1] N. Feamster et al., "Guidelines for Interdomain Traffic Engineering," ACM SIGCOMM Computer Communications Review, **33**(5), 2003.

[2] G. Iannaccone et al, "Analysis of Link Failures in a Large IP Backbone", Proc. ACM Internet Measurement Workshop (IMW), 2002.

[3] O. Bonaventure et al., "Achieving Sub-50 Milliseconds Recovery Upon BGP Peering Link Failures," Proc. ACM CONEXT, 2005.

[4] A. Nucci et al., "IGP Link Weight Assignment for Transit Link Failures," Proc. International Teletraffic Conference (ITC), 2003.

[5] A. Sridharan et al, "Making IGP Routing Robust to Link Failure," Proc. IFIP Networking, 2005.

[6] J. Qiu et al, "Robust Egress Interdomain Traffic Engineering," Proc. IEEE ICNP, 2006.

[7] M. Amin et al., "Making Outbound Route Selection Robust to Egress Point Failure," Proc. IFIP Networking, 2006.

[8] K.H. Ho et al., "Joint Optimization of Intra- and Inter-AS Traffic Engineering," Proc. IFIP/IEEE NOMS, 2006.

[9] S. Agarwal et al, "Measuring the Shared Fate of IGP Engineering and Interdomain Traffic," Proc. IEEE ICNP, 2005.

[10] S. Cerav-Erbas et al, "The Interaction of IGP Weight Optimization with BGP," Proc. ICISP, 2006.

[11] T.C. Bressound et al, "Optimal Configuration for BGP Route Selection," Proc. IEEE INFOCOM, 2003.

[12] V. Chankong et al, Multiobjective Decision Making-Theory and Methodology, Elsevier, New York, 1983.

[13] R.Teixeira et al, "Traffic Matrix Reloaded: Impact of Routing Changes," Proc. PAM, 2005.

[14] F. Golver, Tabu Search, Kluwer Academic Publisher, Norwell MA 1997.

[15] B. Fortz et al, "Internet Traffic Engineering by Optimizing OSPF Weights," Proc. IEEE INFOCOM, 2000.

[16] A. Nucci et al., "IGP Link Weight Assignment for Operational Tier-1 Backbones," IEEE/ACM Transactions on Networking, October 2007.

[17] The BRITE topology generator, http://www.cs.bu.edu/brite/

[18] A. Broido et al, Their Share: Diversity and Disparity in IP Traffic, Proc. PAM workshop, 2004.

[19] S. Bhattacharyya et al., "POP-level and Access-Link-Level Traffic Dynamics in a Tier-1 POP," Proc. ACM IWM, 2001.

[20] S. Saroiu et al., "An analysis of Internet Content Delivery System," Proc. USENIX OSDI, 2002.

[21] R. Teixeira et al., "TIE Breaking: Tunable Inter-domain Egress Selection," IEEE/ACM Transactions on Networking, October 2007.

[22] N. Spring et al., "Measuring ISP Topologies with Rocketfuel," Proc. ACM SIGCOMM, 2004.

[23] M. Caesar et al., "BGP Routing Policies in ISP Networks," IEEE Network, November 2005.