A Route Deflection Approach to Minimize Routing Disruptions for Inter-AS Traffic Engineering

Kin-Hon Ho, George Pavlou, Stylianos Georgoulas and Mina Amin
Centre for Communication Systems Research, University of Surrey, Guildford, Surrey, GU2 7XH, UK
Email: { K.Ho, G.Pavlou, S.Georgoulas, M.Amin }@surrey.ac.uk

I. INTRODUCTION

A recent survey [1] has indicated an increasing usage of Border Gateway Protocol (BGP) route selection for inter-Autonomous System (AS) Traffic Engineering (TE) in response to changes in network conditions [2] such as traffic load and link capacities. The objective of inter-AS outbound TE is to control the flow of traffic exiting an AS, through optimal BGP route selection [3], so as to optimize inter-AS TE objectives. Common inter-AS TE objectives are, for example, to satisfy inter-AS link capacity constraints, to achieve inter-AS traffic load balancing, and/or to minimize peering cost. The most common technique to implement inter-AS TE is by configuring routing protocol policies or metrics such as BGP local preferences (*local-pref*) and Interior Gateway Protocol (IGP) link weights for hot-potato routing. In short, we call this BGP-based TE in this paper.

Unfortunately, it is known that changing inter-AS routes by re-configuring BGP policies may cause *routing disruptions* [4]. Routing disruption is defined as any transient or persistent perturbation of network performance caused by a routing change [4] which may result in long routing convergence, inbound traffic unpredictability and router processor overloading due to route re-computation. Hence, by considering these deficiencies of the BGP-based TE, an approach that not only achieves the inter-AS TE objectives but also minimizes routing disruptions is highly desirable. In this paper, we propose a simple inter-AS deflection routing approach, where a router makes a local traffic forwarding decision to divert traffic from the primary BGP route to the alternate one so as to optimize the inter-AS TE objectives. The merits of this route deflection approach are twofold: (1) minimizing routing disruptions and maintaining stable routing tables by keeping existing BGP routes intact; (2) achieving faster TE effects than the BGP-based TE that takes long time to re-converge onto the next best routes.

II. BGP-BASED TRAFFIC ENGINEERING AND ROUTING DISRUPTIONS

A. BGP-based Traffic Engineering

Consider the simple scenario of Figure 1(a) where transit AS AS-3 learns BGP routes to destination prefix *k* at egress routers *e1* and *e2* from AS-1 and AS-2 respectively. The value on each link within the AS represents the relevant IGP weight. For the purpose of load balancing or improving availability, we consider a common scenario in realistic transit ASes, where the two learnt routes to *k* have identical BGP route attributes such as *local-pref* and *AS-Path length*. Through the full-mesh internal BGP, *e1* and *e2* advertise their learnt BGP routes to other routers within the AS. When *i1* learns these routes from *e1* and *e2*, it selects the one learnt from *e1* as the best route because the lowest IGP cost path from *i1* to *e1* (*i1-e1* with total cost of 8) has smaller cost than that to *e2* (*i1-c1-i2-e2* with total cost of 9). This tie-break BGP route selection is also known as hot-potato routing.

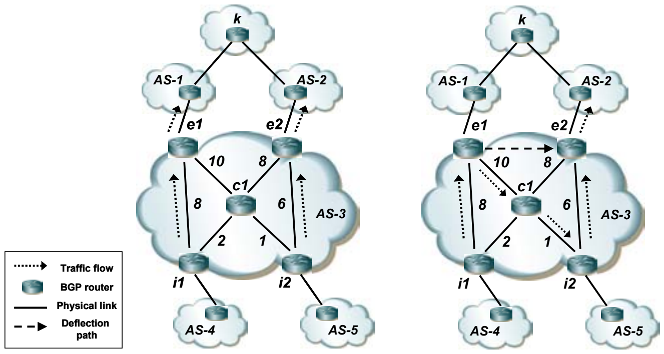

Figure 1. (a) A transit AS scenario (b) Inter-AS deflection routing

Most of the network overloading[1] is due to link failure or traffic upsurge. We assume that the inter-AS link connecting *e1* to AS-1 is overloaded. To reduce the overloading, AS-3 may perform BGP-based TE to alter the routing of traffic towards *k* from the overloaded link to another underloaded link, e.g. the link connecting *e2* to AS-2. To achieve this TE solution, the network operator may configure either of the following routing protocol settings:

- **BGP attribute**: set a higher *local-pref* value for the BGP route learnt at *e2* than the one learnt at *e1*. As a result, the traffic destined to *k* will only be routed through *e2* followed by AS-2.
- **IGP link weight**: change the IGP weight of link *c1-e2* from 8 to 5. As a result, *i1* selects the BGP route learnt from *e2* as the best route since the lowest IGP cost path from *i1* to *e2* (*i1-c1-e2* with the total cost of 7*)* has smaller cost than that to *e1* (*i1-e1* with the total cost of 8).

B. Routing Disruptions

Configuring BGP policies or IGP link weights, however, is likely to cause routing disruptions. The problem with the BGP-based TE is that there are no mechanisms to alter the inter-AS routing other than re-configuring BGP policies, updating the BGP routing tables and then advertising new route updates within the AS and to upstream ASes. The routing disruptions, as a result of such a route change, are typically caused by the following reasons:

This work was undertaken in the context of FP6 Information Society Technologies AGAVE project (IST-027609), which is partially funded by the Commission of the European Union, and the UK EPSRC PAQMAN project (GR/S79992/01).

[1] In this paper, we assume inter-AS link overloading as the factor to initiate inter-AS TE. In fact, inter-AS TE can also be initiated for minimizing peering cost or other applicable objectives.

- route updating is slow which may take long time to converge due to the long convergence properties of BGP. During this time, service is disrupted.
- route re-computation increases the computation load on the router processor. Frequent route updating may affect the function of packet processing and forwarding.
- upstream ASes may change their best downstream routes so that traffic will no longer be routed through the advertising AS. This could have an unpredictable impact on the inbound traffic through the network.
- changing IGP link weights for inter-AS TE not only causes the problem of routing convergence, but also changes the routing of other traffic flows in the network. For instance, by changing the IGP link weight of *c1-e2* from 8 to 5, the shortest IGP path between *e1* and *e2* is changed from *e1-c1-i2-e2* to *e1-c1-e2*. This may cause some links to become overloaded.

In order to achieve inter-AS TE objectives while minimizing the routing disruptions, it is important to maintain stable BGP routing while providing alternate routes to forward the traffic around the overloaded links.

III. INTER-AS DEFLECTION ROUTING

A. Basic Operation

In this section, we propose a simple inter-AS deflection routing approach for achieving inter-AS TE while minimizing routing disruptions. Our inter-AS deflection routing is inspired by the previous proposal on intra-AS deflection routing [5]. The basic operation is that, when inter-AS TE is initiated to reduce overloading on an inter-AS link, instead of initiating update messages for re-computing a new route by re-configuring BGP policies, the incident egress router makes a local traffic forwarding decision to divert the traffic from the primary BGP route to the alternate one, by-passing the overloaded link.

Figure 1(b) illustrates the inter-AS deflection routing approach. In the BGP routing table of *e1*, AS-1 is the next hop of the inter-AS route to *k*. In order to reduce overloading on the inter-AS link connecting *e1* to AS-1, *e1* diverts the traffic for *k* from the primary BGP route to the alternate one through a new egress router *e2* (the deflection path is indicated by the dashed line in the figure) which then routes the traffic to *k*. It is emphasized that the alternate BGP route is only used for the purpose of inter-AS TE. As a result, the traffic received from *i1* for *k* will be first routed onto the path *i1-e1* and then immediately diverted onto the deflection path from *e1* to *e2* (e.g. the path *e1-c1-i2-e2*). In this paper, we call *e1* and *e2* as the *deflection router* and the *relay router* respectively.

With the inter-AS deflection routing, *e1* does not need to update the BGP routing table nor generate new route updates to effect inter-AS TE since no routing protocol policies/metrics (neither BGP nor IGP) will need to be re-configured. Therefore, the BGP routing table remains intact so as to minimize routing disruptions. In addition to this, since only the deflection router makes a local traffic forwarding decision, the TE effects can be achieved faster than BGP-based TE that relies on network-wide routing convergence.

B. Implementation

In this section, we present a potential implementation for inter-AS deflection routing by updating the Forwarding Information Base (FIB).

FIB is a condensation of the Routing Information Base. It is organized around destination prefixes, with each prefix associated with a next-hop address, outgoing interface, and so on. Figure 2 (left) shows the FIB of router *e1*. In the process of packet forwarding, the router uses the prefix as the key to perform a lookup operation, based on the longest prefix matching whereby a more specific prefix is preferred over a less specific one, in the FIB to produce the next-hop address and outgoing interface. Then the packets are forwarded to the corresponding outgoing interfaces. Since the objective of the inter-AS deflection routing is to divert traffic to alternate BGP routes, a straightforward implementation would be to alter the outgoing interfaces in the FIB.

prefix	next-hop	outgoing interface
k	AS-1	up

prefix	next-hop	outgoing interface
k	AS-1	right / exr-e1-e2

prefix	next-hop	outgoing interface	alternate interface
k	AS-1	up	right / exr-e1-e2

Figure 2. Updating FIB for inter-AS deflection routing

There are two ways to do that. The first way is to replace the default outgoing interface with a new interface to which the adjacent router can route the traffic to the designated alternate BGP route[2]. We illustrate this implementation using the example in Figure 1(b). The `right` interface of *e1* is associated with *c1* which has selected the BGP route learnt at *e2* as the best route to *k*. Hence, the `right` interface is eligible as the new outgoing interface to replace the default one, as illustrated in Figure 2 (middle). In fact, the new outgoing interface may also be an explicit route (e.g. `exr-e1-e2`) connecting *e1* to *e2*. The choice between using these two options will be discussed in the next section. Instead of replacing the default outgoing interface, the default can be preserved for use under normal situation, while an alternate outgoing interface entry is added for inter-AS deflection routing purposes. Figure 2 (right) shows such an extension to the FIB. Under normal situation, *e1* routes the traffic for *k* through the `up` interface to AS-1. If the inter-AS deflection routing is used, *e1* forwards the traffic using the alternate outgoing interface `right` to *c1* from which continues to forward the traffic to the designated alternate BGP route via the relay router *e2* along the path *c1-i2-e2*.

IV. ISSUES TO BE ADDRESSED

In this section, we discuss important issues that should be carefully addressed when using inter-AS deflection routing.

1) Routing Loop Avoidance

It is extremely important for inter-AS deflection routing not to create routing loops. A routing loop may be formed if the route through the deflection router to the prefix has been chosen by some intermediate nodes on the IGP path to the relay router as the best BGP route. We explain this scenario using the example in Figure 1(b).

[2] The designated alternate BGP route is selected for carrying the deflected traffic from the primary BGP route. It may be pre-computed or computed in an online manner based on some TE/optimization objectives. In the example of Figure 1(b), the route through *e2* is the designated alternate BGP route.

We assume that the IGP weight of link *c1-e1* is changed from 10 to 7. As a result of hot-potato routing and a route tie-break criterion using the lowest router ID, we assume that *c1* selects the route learnt from *e1* as the best route to *k*. However, a routing loop is formed since *e1* diverts the traffic towards *e2* through the `right` interface to *c1* while *c1* forwards the traffic back to *e1*. Such a routing loop can be avoided by either of the following approaches:

- ***Next-Hop routing***: carefully select the alternate BGP route so that the deflection router has at least one adjacent router selected for which the alternate BGP route is the best route to the prefix. Then, traffic is forwarded to the adjacent router through the corresponding outgoing interface.
- ***Explicit routing***: establish an explicit route between the deflection router and the relay router (i.e. between *e1* and *e2*) for the deflection path. The explicit routing approach assumes that, at the same time, the relay router has no deflection routing to the deflection router for the same prefix; otherwise, a routing loop will be formed.

There are some trade-offs between the two approaches. For the next-hop routing, the deflection router may not have any feasible adjacent router, simply because the adjacent routers have best routes to the prefix other than the designated alternate BGP route. In this case, an explicit route can always be established between the deflection router and the relay router. Regardless of whether the intermediate routers on the explicit route do not have the designated alternate BGP route in their routing table, they will still forward the traffic along the explicit route to the relay router. However, the explicit routing approach should be carefully implemented since the excessive use of explicit routes could cause a scalability problem in terms of the route states to be maintained.

2) *Sub-Optimal Intra-AS Resource Utilization*

It should be emphasized that the inter-AS deflection routing may result in sub-optimal intra-AS resource utilization since the deflection paths consume extra resources in the network. The resource utilization can be affected by:

- the location of the relay router: the closer the relay router to the deflection router, the shorter the deflection path.
- the selection of the deflection path: the path does not have to be the shortest to the relay router but could be a long one for achieving traffic engineering objectives.

We illustrate the impact of the inter-AS deflection routing on resource utilization in Figure 3(a). We assume the total path hop count being the performance metric of resource utilization. Assume that routers *e1*, *e2* and *e3* have equivalent BGP routes to the same prefix. If *e1* needs to perform inter-AS deflection routing, it has two choices for the alternate BGP route: the route through *e2* or *e3*. If *e3* is selected, the traffic will be diverted from *e1* to *c4* followed by the path *c4-c5-c6-e3*, which takes four hops in total. In contrast, if *e2* is selected, *e1* diverts the traffic onto the path *e1-c1-e2* or *e1-c2-c3-e2*, which takes two and three hops respectively. Hence, selecting the route through *e2* will result in a shorter deflection path. Between the two candidate deflection paths to *e2*, *e1* may not need to choose the shortest one (i.e. *e1-c1-e2*) for traffic engineering purposes.

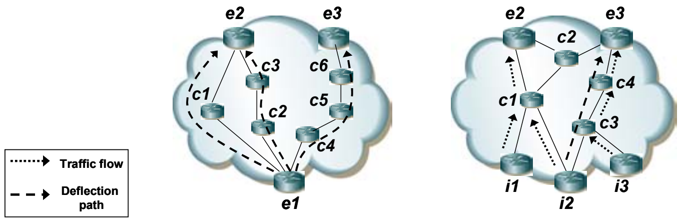

Figure 3. (a) Intra-AS resource utilization (b) Deflection granularity

On the other hand, if explicit routing is used and deflection path *e1-c2-c3-e2* is selected, *e1* may not need to establish an explicit route to *e2*. Instead, an explicit route from *e1* to some intermediate routers (e.g. *c2*) would be sufficient because *c2* will forward the traffic towards *e2* through *c3*. This improves scalability by minimizing the route states to be maintained. Some intelligent multi-objective algorithms may be devised to determine which paths should be used in order to optimize the traffic engineering objectives while improving scalability.

3) *Granularity of Deflection Routing*

Inter-AS deflection routing can be done in a coarse or fine-grained way. The coarse-grained way, as illustrated in Figure 1(b), is to divert the traffic to a prefix at the egress router, regardless of their ingress points. In order to perform the inter-AS deflection routing in a more fine-grained way based on (ingress, prefix) pair [3], the deflection can be done at the corresponding ingress point. However, due to their different locations, performing the deflection routing at ingress or egress points may lead to different intra-AS resource utilization. We illustrate the granularity of deflection routing in Figure 3(b).

We assume that the IGP weights of all links are unity. According to hot-potato routing, both *i1* and *i2* select *e2* as the best egress point while *i3* selects *e3*. If inter-AS deflection routing is to divert some traffic from *e2* to *e3*, then the traffic received from both *i1* and *i2* will be diverted onto the path *e2-c2-e3*. However, in order to achieve better network performance, the network operator may only want to divert the traffic from *i2* to *e3*. In this case, *i2* can itself divert the traffic using an explicit route to *e3* or a new outgoing interface to *c3* from which continues to forward the traffic to *e3*.

V. CONCLUSION

This paper proposes a simple inter-AS deflection routing approach to divert traffic from the primary BGP route to the alternate one so as to satisfy inter-AS traffic objectives. The approach minimizes routing disruptions and achieves faster TE effect than the BGP-based TE due to the localized traffic forwarding decision at the deflection router.

REFERENCES

[1] Y.R. Yang et al., "On Route Selection for InterAS Traffic Engineering," *IEEE Network Magazine*, November/December 2005, pp. 20-27.

[2] N. Feamster, J. Winick and J. Rexford, "A Model of BGP Routing for Network Engineering," Proc. *ACM SIGMETRICS*, June 2004.

[3] T.C. Bressoud, R. Rastogi and M.A. Smith, "Optimal Configuration for BGP Route Selection," Proc. *IEEE INFOCOM*, March/April 2003.

[4] R. Teixeira and J. Rexford, "Managing Routing Disruptions in Internet Service Provider Networks," *IEEE Comm. Magazine*, March 2006.

[5] Z. Wang and J. Crowcroft, "Shortest Path First with Emergency Exits," Proc. *ACM SIGCOMM*, September 1990, pp. 166-176.