Can Forwarding Loops Appear when Activating iBGP Multipath Load Sharing?

Simon Balon* and Guy Leduc

Research Unit in Networking
EECS Department- University of Liège (ULg)
Institut Montefiore, B28 - B-4000 Liège - Belgium
Simon.Balon@ulg.ac.be, Guy.Leduc@ulg.ac.be,

Abstract. We analyse the possible consequences of activating iBGP multipath load sharing in a given domain (or AS), which allows for load balancing over multiple exit routers. It has been stated that interdomain routing loops may appear in this case. We show that under reasonable assumptions (which reflect commercial relationships between ASes) such routing loops cannot appear. Furthermore we show that even if theses assumptions are not met, routing loops can only be transient.

Keywords: iBGP Multipath Load Sharing, Traffic Split, Traffic Engineering, Forwarding Loops

1 Introduction

Traffic Engineering in OSPF/ISIS networks consists in finding the best possible set of link weights ([5]). The routing scheme resulting from this link weight setting should reflect Traffic Engineering goals, i.e. good user performance and efficient use of network resources. Typically link weights optimizers use ECMP (Equal Cost Multi-Path) to split the traffic on multiple paths between one ingress node and one egress node. Using ECMP has multiple advantages. For example ECMP can be used to improve IP restoration ([9]). It is also a flexible routing technique and usually allows a good engineering of the network.

While it is considered valuable to split traffic on multiple paths inside a domain, splitting traffic on multiple interdomain paths is rarely envisaged. Indeed BGP typically chooses one (and only one) path among its multiple available ones. Although in an AS some destination prefixes are reachable via only one egress point, it is frequent that most of the prefixes (typically provider prefixes) are reachable via multiple BGP-equivalent routes (for example if the AS has multiple links connecting its providers). Using classical BGP one of these routes is chosen via the Hot-Potato criterion or a tie-break at a later stage of the BGP decision process. But it is also possible to configure BGP to allow the network operator to split traffic amongst multiple BGP-equivalent routes. This could move

* S. Balon is a Research Fellow of the Belgian National Fund for the Scientific Research (F.N.R.S).

the horizon of traffic engineering possibilities back, allowing an optimizer to take these traffic splits into account to better engineer the network, even allowing it to engineer the interdomain links ([2]).

But the situation is not as beautiful as it seems. Indeed splitting traffic amongst multiple available BGP-equivalent routes which may have a different AS-level paths can cause problems as explained in [8]. In that paper the authors state that forwarding loops could appear and they propose a solution. In this paper we show that contrary to what can be thought at first glance and under reasonable assumptions, forwarding loops should not appear in any case. These assumptions are based on the BGP router configurations that typically reflect commercial relationships.

The paper is organized as follows. In section 2 we introduce BGP basics and how iBGP multipath load sharing works. We also briefly describe why forwarding loops could appear with iBGP multipath load sharing. Section 3 presents the BGP configuration we assume in this paper. These are natural BGP configurations that should be respected in all the ASes. We show in section 4 that if these assumptions hold, no forwarding loops can appear. In sections 5 and 6 we analyse what happens if the assumptions we made about BGP are not respected. Indeed even if these should be respected in all the ASes it is impossible to be sure of that. We show in section 5 that even in this case no forwarding loops can appear when activating iBGP multipath load sharing. These can only appear at a later stage if the BGP configuration of an AS is changed. We show in section 6 that even in this case forwarding loops are only transient. Section 7 concludes the paper.

2 Routing Principles, iBGP Multipath Load Sharing and Forwarding Loops

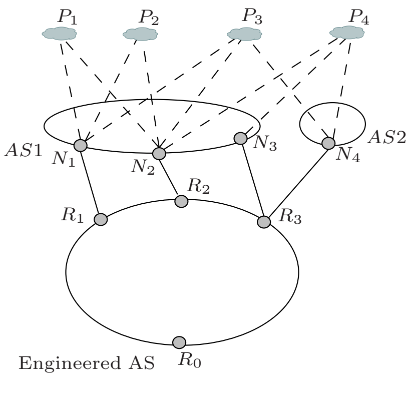

Fig. 1. Example Topology

We will explain the basic intradomain and interdomain routing principles on the example topology of figure 1. Routers R_0, R_1, R_2 and R_3 are part of the Engineered AS. This AS has two neighbouring ASes : AS_1 and AS_2. We consider four IP prefixes (P_1 to P_4) which are joinable through interdomain paths that are depicted by dashed lines. These are possible paths advertised by BGP.

Each packet sent on the Internet follows a path which is defined by routing protocols. The exterior gateway protocol (EGP) defines the path at the network level. This path is called the AS path[1]. The EGP used in the Internet is BGP (Border Gateway Protocol). In each AS the path from each ingress router to each egress router is defined by the interior gateway protocol (IGP). The IGPs that are generally used in the Internet are OSPF and ISIS.

In an AS the path between ingress and egress routers are computed by a Shortest-Path algorithm based on the link weights. If ECMP is enabled, several equal shortest-paths can be used simultaneously to evenly split the traffic among them, by using a hash table that maps a hash of multiple fields in the packet header to one of these paths, so that all packets of a flow will follow the same path with limited packets reordering (see [4] for a performance analysis of hashing based schemes for Internet load balancing). Figure 2 shows an example of ECMP inside an AS. This figure assumes that there are two equal cost paths from R_0 to R_1.

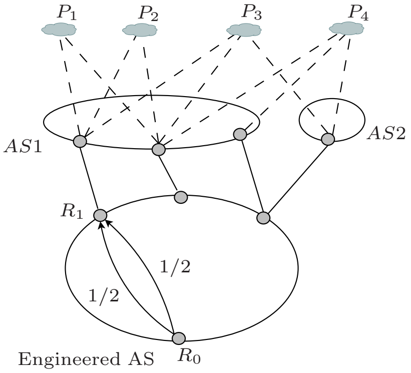

Fig. 2. Intradomain Equal Cost Multipath (ECMP)

BGP allows routers to exchange reachability information between neighboring ASes ([11]). Each AS is connected to several neighboring ASes by interdomain links. Depending on the connectivity of the network and on the destination of the packet, one or several neighboring ASes can be chosen to forward the packet to the destination. The choice of the BGP next-hop (i.e. the egress router in this AS or the border router in the next AS, that will relay the packet toward

[1] AS stands for Autonomous System. In the paper we use domain and AS interchangeably.

the destination) is based on the information exchanged with neighbors and on a local configuration implementing its routing policy.

There are two types of BGP sessions that are used to exchange routes between routers. eBGP sessions are used between routers in different ASes, while iBGP sessions are used between routers in the same AS. When a router receives a route on a iBGP or eBGP session, this route has to pass the input filter to be eligible in the BGP decision process that selects the best route(s) toward each destination prefix. The best route(s) selected by this process is(are) then announced on other BGP sessions after passing through an output filter.

The BGP route selection process, implementing routing policies, is made of several criteria ([3, 6]):

1) Prefer routes with the highest local preference which reflects the routing policies of the domain;
2) Prefer routes with the shortest AS-level Path;
3) Prefer routes with the lowest origin number, e.g., the routes originating from IGP are most reliable;
4) Prefer routes with the lowest MED (multiple-exit discriminator) type which is an attribute used to compare routes with the same next AS-hop;
5) Prefer eBGP-learned routes over iBGP-learned ones (referred to as the eBGP >iBGP criterion in the sequel);
6) Prefer the route with the lowest IGP distance to the egress point (i.e. the so-called hot-potato, or early exit, criterion);
7) If supported, apply load sharing between paths. Otherwise, apply a domain-dependent tie-breaking rule, e.g., select the one with the lowest egress ID.

Consider the network of figure 1. Suppose that routes to P_1 are announced by N_1 to R_1 and N_2 to R_2 on eBGP sessions. Suppose that the routes announced by these two routers have the same attributes (i.e. local-preference, AS-path length, origin number and MED) after passing the input filters of routers R_1 and R_2 (this is very frequent in practice for routes that are received from the same neighboring AS[2]). Suppose also that these two routes are announced by R_1 and R_2 to R_0 on iBGP sessions. Usually the attributes are not changed when forwarding routes on iBGP sessions. So R_0 has two routes to reach P_1 and these two routes are equivalent w.r.t. criteria 1 to 4. Both are received on iBGP sessions so are also equivalent w.r.t. the 5th criterion. In this case R_0 will use its IGP distance to R_1 and R_2 to select the best route toward P_1. We say that this route is chosen using the hot-potato criterion by router R_0. Note that R_1 and R_2 will directly forward traffic toward this prefix on their interdomain link using the eBGP>iBGP criterion.

Now if R_1 and R_2 are at the same IGP distance from R_0, the 7th criterion will be used. By default only one next hop can be chosen and a tie-break selects the best route. But it is also possible to enable iBGP multipath load sharing [3,

[2] For the case study in [2] we have shown that 97.2% of the prefixes have multiple BGP-equivalent (w.r.t. criteria 1 to 4) egress points, which amounts to 35.6% of the traffic on average.

6] and balance the load on both paths. As for intradomain ECMP, a hash table is used to select the particular route of a packet. Figure 3 supposes that iBGP multipath is activated and that R_1 and R_2 are at the same distance from R_0. In this case the traffic going from R_0 to P_1 will be split evenly on both paths. Figure 4 presents the combined use of ECMP and iBGP multipath load sharing.

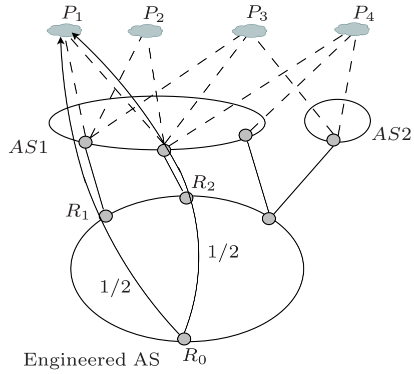

Fig. 3. iBGP multipath load sharing

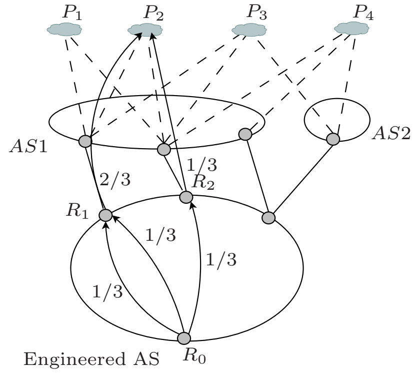

Fig. 4. ECMP + iBGP multipath load sharing

Note also that BGP ([11]) includes a loop prevention mechanism. When an AS receives a route whose ASPATH contains its AS number, it discards the route. This supposes that the ASPATH contains a full list of all the ASes along the path used to forward traffic toward this destination. If part of the ASPATH information is lost, this mechanism does not work anymore.

In [8] we can read that *Most of the current BGP implementations upon receiving multiple equal cost BGP routes from different peers can insert all of them* *(or a subset depending upon the local policies) in their forwarding table. This can be done to locally split the traffic across several paths. However, because BGP in its current state can only advertise one path to its peers, an implementation MUST choose from one of the best paths that it is using for the advertisement. This has implications for the BGP peers that receive such advertisements from ECMP capable BGP speakers. In the worst case it can lead to potential loops if the entire path information is not advertised to the peers.*

In [8] the authors present a first method to avoid forwarding loops using BGP AS_SET and AS_SEQUENCE. In next sections we analyse what happens if this method is not used and only one ASPATH is announced to other ASes. Contrary to what can be thought at first glance, we show that forwarding loops should not appear when using iBGP multipath on different ASPATH routes.

Definition 1 *A packet is trapped in a forwarding loop if there is a cycle of routers such that each router on the cycle forwards the packet to the next router on the cycle, leading the packet to be infinitely forwarded on the cycle.*

Of course forwarding loops should be avoided in practice. Note also that in IP networks, the time to live (TTL) field of the IP header will force routers to drop a packet which is trapped in a forwarding loop.

Definition 2 *A provider loop (for a particular destination prefix) is a cycle of ASes such that each AS on the cycle is the provider of the next AS.*

Note that a provider loop is also a customer loop if the cycle is analysed in the opposite direction.

3 BGP Model Used

In this paper we consider the following common BGP configurations.

Assumption 1 *We consider import/export rules which state that ([10], [7], [1]) :*

- *an AS does not export to a provider or peer routes that it learnt from other providers and other peers;*
- *an AS can export to its customers any routes it knows of.*

This assumption (1) reflects that an AS does not want to provide transit services between its providers and peers.

Assumption 2 *We consider that routes learnt from customers should be preferred to routes learnt from either providers or peers, leaving ASes latitude to assign relative preferences among customer routes, and among peer and provider routes.*

This assumption (2) is the preference rule suggested in Guideline A of Gao and Rexford [7]. This is a logical assumption for commercial relationships. Indeed an AS earns money for the traffic it sends on its customer links while it does not earn money for the traffic it sends on its peer links and it pays for the traffic it sends on its provider links. So it should always prefer to send traffic to its customers than to its provider when it has the choice.

Our last assumption is the following (this is also assumed in [7]).

Assumption 3 *We assume that there is a hierarchical customer-provider relationship among ASes.*

This is equivalent to saying that there is no provider loop in the AS-level topology.

4 When Do Routers Use BGP Loop Prevention Mechanism?

The BGP loop prevention mechanism implemented in a BGP router consists in discarding routes whose ASPATH contains the AS number of the router[3] ([11]). When and how does this situation happen?

For this situation to happen, we have to be in the case of figure 5. AS_X receives a route for a destination prefix from AS_1. It announces this route to AS_2. Later AS_X receives back this route from AS_3 and discards the route because its AS number appears in the ASPATH.

We will demonstrate that this situation never happens if Assumptions 1, 2 and 3 are respected. We divide the problem into different cases, depending on the commercial relationship between AS_X and its neighbouring ASes for the particular destination prefix we consider. Note that applying this reasoning to each prefix known by AS_X allows us to generalize our result.

4.1 AS_1 is a provider or a peer of AS_X

AS_X has received the route from a provider or peer. So AS_X will export this route to AS_2 only if AS_2 is one of its customers (applying Assumption 1). Following the same reasoning the route is announced from AS_2 hop by hop to AS_3 and finally back to AS_X if all these links are provider to customer links. If it is not the case the route is stopped before coming back to AS_X. So AS_3 is a provider of AS_X and cycle A is a provider loop. This situation should not happen as we assumed in section 3 that there is a hierarchical customer-provider relationship among ASes (Assumption 3).

Now if cycle A is a provider loop (meaning that Assumption 3 is not respected), a forwarding loop could appear if AS_3 is preferred to AS_1 which are both providers. In this case BGP loop prevention mechanism will discard the route from AS_3 which could be chosen if this mechanism were not present.

[3] Note that this loop detection can also be performed on the sender-side. In this case a BGP router will not announce a route to a neighboring router if its AS number is in the ASPATH of this route.

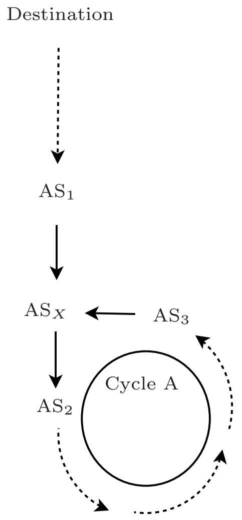

Fig. 5. AS topology

4.2 AS_1 is a customer of AS_X

As AS_1 is a customer, AS_X can announce the route on all its BGP sessions (Assumption 1). So AS_2 can be a customer, a peer or a provider of AS_X. We consider all these cases.

AS_2 is a customer of AS_X In this case, following the same kind of reasoning as in section 4.1, the route will come back to AS_X only if all the links from AS_2 to AS_3 and back to AS_X are provider to customer links. So this implies that cycle A is a provider loop (meaning that Assumption 3 is not respected).

The situation is a little bit different than in section 4.1, because anyway, if this situation happens, AS_X will always prefer the route from AS_1 which is a customer when compared to the route from AS_3 which is a provider (Assumption 2). In this case Assumptions 1 and 2 are sufficient to guarantee the absence of forwarding loops.

AS_2 is a provider of AS_X In this case AS_2 has received the route from AS_X which is one of its customers and so it can announce it on all its BGP sessions (Assumption 1). Thus AS_3 can be a customer, a peer or a provider of AS_X. We consider all these cases.

a) AS_3 is a provider or a peer of AS_X In this case, AS_X will prefer the route coming from AS_1 (which is one of its customer) to the new route coming from AS_3 (which is a provider or a peer) (Assumption 2).

In this case Assumptions 1 and 2 are sufficient to guarantee the absence of forwarding loops. Note also that this (non-problematic) situation may happen without provider loop.

b) AS_3 is a customer of AS_X For AS_3 to announce the route to AS_X (which is its provider), it must have received this route from one of its customers (applying Assumption 1). By extending this reasoning we can deduce that the route has been propagated hop-by-hop on customer to provider links from AS_2 to AS_3. Otherwise the route would have been stopped between AS_2 and AS_3. In this case cycle A is also a provider loop and this should not happen (Assumption 3).

Note that if this situation happens (meaning that Assumption 3 is not respected), a forwarding loop could appear if AS_3 is preferred to AS_1 which are both customers (which respect Assumption 2). In this case BGP loop prevention mechanism will discard the route which could be chosen if this mechanism were not present.

AS_2 is a peer of AS_X In this case AS_2 will announce this route only to its customers (Assumption 1). So the route will be announced hop-by-hop on provider to customer links to AS_3 and then to AS_X (Assumption 1). AS_3 is a provider of AS_X and thus AS_X will prefer the route from AS_1 which is one of its customer to the route from AS_3 which is one of its provider (Assumption 2).

The conclusion is the same as in preceding paragraph labelled a).

4.3 Summary

Table 1 presents all the possible router configurations that result in AS_X receiving a route whose ASPATH contains its AS number. In all other router configurations it is not possible for AS_X to receive such a route.

Note that only two of these configurations could result in forwarding loops if BGP prevention mechanisms were not enabled. These two configurations are the lines marked with the label ”No if BGP prevention” in the ”Potential Forwarding loop” column (lines 1 and 5). Note that these two configurations imply that a provider loop is present in the network, which was supposed not to happen as stated in Assumption 3. Thus we can say that the BGP loop prevention mechanism is a kind of watchdog avoiding forwarding loops in misconfigured networks (i.e. networks which do not respect our Assumptions).

Anyway we cannot be 100 % sure that our assumptions are respected in the whole Internet. This is why the BGP loop detection check is still useful in today networks. In the next sections, we will analyse what happens if our assumptions are not respected and what is the impact of this point on the activation of iBGP multipath load sharing.

Line	AS_1	AS_2	AS_3	Provider loop	Potential Forwarding loop
1	Provider	Customer	Provider	YES	No if BGP prevention
2	Peer	Customer	Provider	YES	NO[4]
3	Customer	Customer	Provider	YES	NO
4		Provider	Provider or Peer	NO	NO
5			Customer	YES	No if BGP prevention
6		Peer	Provider	NO	NO

Table 1. All possible configurations (referring to fig. 5) leading AS_X to receive a route advertisement whose ASPATH contains its own AS number.

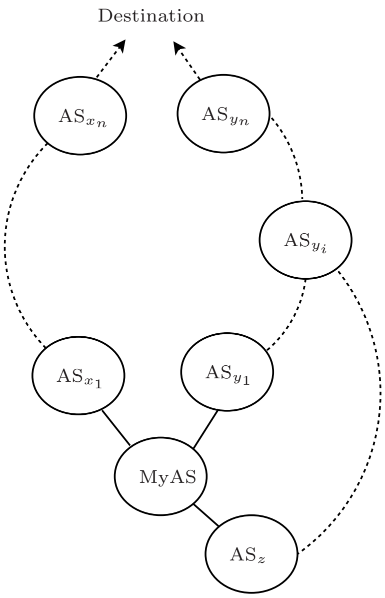

Fig. 6. iBGP mutipath AS topology

5 No Forwarding Loop When Activating iBGP Multipath Load Sharing

In this section we would like to analyse whether activating iBGP multipath load sharing can result in a forwarding loop or not. Indeed a BGP router which activates iBGP multipath on multiple routes will announce only one of these routes to its neighboring ASes. If later on, one AS on one route that has not been announced receives back this route, its BGP loop detection mechanism will be unable to detect the loop[5]. For such a situation to appear we have to be in the case of figure 5 in which one of the routers between AS_2 and AS_3 on cycle A enables iBGP multipath on at least two routes, one going to the destination via AS_X and another route in which AS_X is not present. Such a general topology is depicted on figure 6, where AS_{y_i} is the AS_X of figure 5, $AS_{y_{(i+1)}}$ is AS_1, cycle A is AS_{y_i} ... AS_{y_1} MyAS AS_z ... AS_{y_i} and MyAS is the AS on cycle A which enables iBGP multipath load sharing on multiple available routes : AS_{y_1} ... AS_{y_n} and AS_{x_1} ... AS_{x_n} which does not contain AS_{y_i}. We will show that even with such a topology no permanent forwarding loop can be installed. As this topology is built to reflect all the possible topologies that can lead to a permanent forwarding loop, this will imply that no forwarding loops can be created when using iBGP multipath load sharing. Note that optionally AS_{y_i} could be merged with AS_{y_1} and/or AS_z. Our reasoning can also be applied if iBGP multipath load sharing is used on more than one additional path to the destination in which AS_X is not present.

Suppose that at time $t = t_0$ iBGP multipath load sharing is not activated in the network and that MyAS has two BGP-equivalent routes w.r.t. criteria 1 to 6 whose ASPATH are AS_{x_1} ... AS_{x_n} and AS_{y_1} ... AS_{y_n}. One of the two available routes is chosen with some tie-break and this route is announced to AS_z. Suppose now that at time $t_1 > t_0$ we do activate iBGP multipath load sharing on these two routes and that we continue to announce the same route to AS_z. We will show that in this case no forwarding loop is created at time t_1. Indeed the route that was announced at time t_0 was either the route received from AS_{y_1} or the route received from AS_{x_1}. If it was the route received from AS_{y_1} no forwarding loop can be created because AS_{y_i} will see its AS number in the ASPATH received from AS_z. If it was the route received from AS_{x_1}, a forwarding loop cannot be created at time t_1. Indeed the route announced to AS_z is the same at time t_1 than at time t_0. So if AS_{y_i} prefers the route coming back from MyAS via AS_z to the route received from $AS_{y_{(i+1)}}$, it would already have chosen this route at time t_0 and the route with ASPATH AS_{y_1} ... AS_{y_n} would not have been available at MyAS.

[4] This is due to the fact that usually routes received from peers are preferred to routes received from providers even if this is not included in our assumptions. If we do not assume this preference rule, line 2 should just be merged with line 1.

[5] Of course this can only happen if at least one of our assumptions is not respected, as it has been shown in section 4.

6 Anyway Forwarding Loops Can Only Be Transient

Now suppose that in the preceding example, at time $t_2 > t_1$, the route selected by BGP at router AS$_{y_i}$ changes. There are two possibilities. Either both routes are used by activating iBGP multipath load sharing at router AS$_{y_i}$ or the route selected by BGP is now the route received back from MyAS via AS$_z$ instead of the route received from AS$_{y_{(i+1)}}$. We will analyse both cases separately.

6.1 Both routes are selected and used

We will see that this situation is impossible. Indeed this implies that at time t_2, AS$_{y_i}$ activates iBGP multipath load sharing and splits its traffic on its two available routes (the route received back from MyAS via AS$_z$ and the route received from AS$_{y_{(i+1)}}$). But iBGP multipath load sharing cannot select these two available routes as these do not have the same ASPATH length ($|AS_{y_i}\ ...\ AS_{y_n}| \leq |AS_{y_1}\ ...\ AS_{y_n}| = |AS_{x_1}\ ...\ AS_{x_n}| < |AS_{y_i}\ ...\ AS_z\ MyAS\ AS_{x_1}\ ...\ AS_{x_n}|$[6]). Indeed one condition for iBGP multipath load sharing to be activated on multiple routes is that these routes are equivalent w.r.t. BGP criteria 1 to 6, which implies equality of ASPATH lengths (via criterion 2).

6.2 The route received back from MyAS via AS$_z$ is now the best route

AS$_{y_i}$ has to change its BGP policies (i.e. its local pref values) for BGP to select the route received back from MyAS via AS$_z$ as best route instead of the route received from AS$_{y_{(i+1)}}$. Indeed the local prefs are the only way to force BGP to select a route whose ASPATH is longer (see BGP decision process in section 2). In this case a forwarding loop is created. But as AS$_{y_i}$ now has changed its route, it must withdraw the old route and advertise the new one to AS$_{y_{(i-1)}}$ and so hop by hop to MyAS. When MyAS receives the new route, it can detect the loop because its AS number appears in the ASPATH. So MyAS will stop using the route received from AS$_{y_1}$ and the forwarding loop is stopped. Note that at this time the router of MyAS which detects and stops the forwarding loop should alert the network operator that at least one of our assumptions is not respected somewhere. With such an alert the network operator could analyse the situation and look for the cause of the problem. Indeed this means that one of our 3 assumptions is not respected.

7 Conclusion

In this paper we have analysed how forwarding loops can appear in current BGP networks. We have shown that forwarding loops should not appear even if part of the ASPATH information is discarded, which can be the case when using iBGP

[6] $|ASPath|$ denotes the number of ASes of $ASPath$.

multipath load sharing for routes with different ASPATH. Indeed we have shown that BGP configurations reflecting commercial relationships ensure that no forwarding loops will appear. Anyway as it is not possible for a network operator to verify the good configuration of all the involved ASes, we have analysed what would happen in this case (i.e. if BGP configuration would not reflect commercial relationships). We have shown that even in this case, a forwarding loop cannot appear immediately after activating iBGP multipath load sharing. The forwarding loop could only appear if in addition to the aforementioned conditions, some ASes change their policies in a particular way. Moreover we have shown that even in this case, if a forwarding loop appears, it is only transient.

This leads us to conclude that activating iBGP multipath load sharing for routes with different ASPATH is not as dangerous as it may seem at first glance.

Acknowledgments

This work has been partially supported by the Walloon Region (TOTEM project). The authors would like to thank P. François and B. Quoitin (Université catholique de Louvain) for their comments on earlier drafts of this paper.

References

1. C. Alaettinoglu. Scalable Router Configuration for the Internet. In *Proceedings of the 1996 International Conference on Networking Protocols*, October 1996.
2. S. Balon and G. Leduc. Combined Intra- and Inter-domain Traffic Engineering using Hot-Potato Aware Link Weights Optimization. In *Submitted for publication*.
3. BGP Best path selection algorithm. http://www.cisco.com/warp/public/459/25.shtml.
4. Z. Cao, Z. Wang, and E. Zegura. Performance of Hashing-Based Schemes for Internet Load Balancing. In *Proceedings of INFOCOM*, 2000.
5. B. Fortz and M. Thorup. Internet Traffic Engineering by Optimizing OSPF Weights. In *Proceedings of INFOCOM*, pages 519–528, 2000.
6. Foundry enterprise configuration and management guide. http://www.foundrynet.com/services/documentation/ecmg/BGP4.html#17143.
7. L. Gao and J. Rexford. Stable Internet routing without global coordination. *IEEE/ACM Transactions on Networking*, pages 681–692, December 2001.
8. J. M. Halpern, M. Bhatia, and P. Jamka. Advertising Equal Cost Multipath Routes in BGP. *Internet Draft, Work In Progress*, February 2006.
9. G. Iannaccone, C.-N. Chuah, S. Bhattacharyya, and C. Diot. Feasibility of IP restoration in a tier 1 backbone. *IEEE Network*, 18(2), 2004.
10. J. L. Sobrinho. An Algebraic Theory of Dynamic Network Routing. *IEEE/ACM Transactions on Networking*, pages 1160–1173, October 2005.
11. J. Stewart. *BGP4 : Interdomain routing in the Internet*. Addison Wesley, 1999.