1 Multi-path BGP: motivations and solutions

F. Valera[1], I. van Beijnum[2], A. García-Martínez[1], M. Bagnulo[1]

[1]Universidad Carlos III de Madrid, [2]IMDEA Networks

Although there are many reasons towards the adoption of a multi-path routing paradigm in the Internet, nowadays the required multi-path support is far from universal. It is mostly limited to some domains that rely on IGP features to improve load distribution in their internal infrastructure or some multi-homed parties that base their load balance on traffic engineering. This chapter explains the motivations for a multi-path routing Internet scheme, commenting the existing alternatives and detailing two new proposals. Part of this work has been done within the framework of the Trilogy[1] research and development project, whose main objectives are also commented in the chapter.

1.1 Introduction

Multi-path routing techniques enable routers to be aware of the different possible paths towards a particular destination so that they can make use of them according to certain restrictions. Since several next hops for the same destination prefix will be installed in the forwarding table, all of them can be used at the same time. Although multi-path routing has a lot of interesting properties that will be reviewed in section 1.3, it is important to remark, that in the current Internet, the required multi-path routing support is far from universal. It is mostly limited to some domains that deploy multi-path routing capabilities relying on IGP (Intra-domain Gateway Protocol) features to improve the load distribution in their internal infrastructure and normally only allowing the usage of multiple paths if they all have the same cost.

However, multi-path routing would also present important advantages in the inter-domain routing environment.

In the Internet, for example, the routing system and the congestion control mechanisms which are two of its main building blocks, work in a completely

[1] Trilogy: Architecting the Future (2008-2010). ICT-2007-216372 (http://trilogy-project.org). The different research partners of this project are: British Telecom, Deutsche Telekom, NEC Europe, Nokia, Roke Manor Research Limited, Athens University of Economics and Business, University Carlos III of Madrid, University College London, Universit Catholique de Louvain and Stanford University

independent manner. That is, the route selection process is performed based on some metrics and policies that are not dynamically related to the actual load of the different available routes. On the other hand, when there is congestion in some parts of the network, the only possible reaction is to reduce the offered load. Current flow control mechanisms cannot react to congestion by rerouting excess traffic through alternative links because typically these alternatives are not known. Clearly, coupling routing and more specifically multi-path routing, and congestion control has significant potential benefits, since it would, for instance, enable to spread the traffic through multiple routes based on the utilization of the links.

This kind of coupling and interactions between multi-path and other techniques will be explained in section 1.2 since they constitute one of the main objectives of the Trilogy project, that is described in this section.

Despite the fact that multi-path alternatives for the inter-domain routing are not available yet in the Internet, some of the existing proposals are described in section 1.4. Finally, this chapter introduces two additional solutions in sections 1.4.4.2 and 1.4.4.3. These two solutions are some of the proposals being considered in the Trilogy project to provide non-equal cost multi-path routing at the inter-domain level. The goal of both mechanisms is to enable inter-domain multi-path routing in an incrementally deployable fashion that would result in increased path diversity in the Internet. Unlike the rest of the existing alternatives these new proposals imply minimum changes to the routers and to BGP (Border Gateway Protocol) semantics, are interoperable with current BGP routers, and have as one of their most important objectives an easier adoption of multi-path inter-domain solutions so their advantages can be realized earlier.

1.2 Trilogy project

1.2.1 Objectives

Trilogy is a research and development project funded by the European Commission by means of its Seventh Framework Programme. The main objective of the project is to propose a control architecture for the new Internet that can adapt in a scalable, dynamic, autonomous and robust manner to local operations and business requirements.

There are two main motivations for this objective. The first one is the traditional limited interaction that has always existed between congestion control, routing mechanisms, and business demands. This separation can be considered as the direct cause of many of the problems which are leading to a proliferation of disperse control mechanisms, fragmentation of the network into private environments, and growing scalability issues. Re-architecting these mechanisms into a more coherent whole is essential if these problems are to be tackled.

The second motivation comes from the observation of the success of current Internet. More than from its transparency and self-configuration, it comes from the fact that it is architected for change. The Internet seamlessly supports evolution in applications use and adapts to configuration changes; deficiencies have arisen where it is unable to accommodate new types of business relationships. To make the Internet richer and more capable will require more sophistication in its control architecture, but without imposing a single organizational model.

1.2.2 Trilogy technologies

Past attempts to provide joint congestion control and routing have proven that the objective of the Trilogy project is a challenging task. In the late 80's, a routing protocol that used the delay as the metric for calculating the shortest paths was tried in the ARPANET [16]. While this routing protocol behaved well under mild load conditions, it resulted in severe instabilities when the load was high [16]. Since that experience, it is clear that the fundamental challenge to overcome when trying to couple routing to congestion information is stability. Recent theoretical results [15, 10] have shown that it is indeed possible to achieve stability in such systems. The Trilogy project relies in these recent results in order build a stable joint multi-path routing and congestion control architecture. One key difference between Trilogy's architecture and the previous ARPANET experience is that Trilogy embeds multi-path routing capabilities. Intuitively stability is easier to achieve in a multi-path routing scenario where the load split ratio varies based on the congestion in the different paths than in a single-path routing approach, where all the traffic towards a given destination is shifted to an alternative path when congestion arises in the currently used path. So, multipath routing capabilities are one of the fundamental components for Trilogy's architecture. In addition, the distribution of traffic among the multiple routes is performed dynamically based on the congestion level of the different paths, as opposed to current multi-path routing schemes. Normal equal cost multipath practice is to perform round-robin distribution of flows among the multiple routes. It is possible to distribute the flows across the multiple routes in a way that optimizes the traffic distribution for a given traffic matrix [7, 26].

The proposed approach is based on the theoretical results presented in [15]. The basic idea is to define a MPTCP (Multi-Path Transmission Control Protocol) that is aware of the existence of multiple paths. MPTCP will then characterize the different paths based on their congestion level. That means that MPTCP will maintain a separate congestion window for each of the available paths and will increase and reduce the congestion window of each path based on the experienced congestion. An MPTCP connection is constituted by multiple subflows associated to the different paths available and each subflow has its own congestion control. By coupling the congestion window of the different subflows, additional benefits may be obtained like the resource pooling benefits, described in [28] (this occurs when the networks resources behave as though they

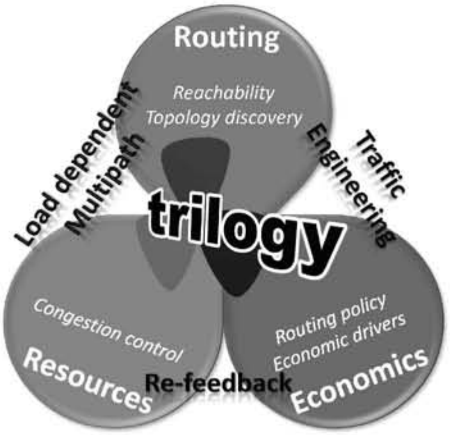

Figure 1.1 Three basic components of the Trilogy project

make up a single pooled resource and facilitates increasing reliability, flexibility and efficiency). While the coupling of the congestion windows of the different subflows of MPTCP allows users to move away from congested paths and leave space for flows that have more pressing needs due to the lack of path diversity toward their destination, Trilogy's architecture includes a third component that allows to provide accountability for the congestion caused in the network, a piece that is missing in the current Internet architecture, but deemed critical for the Next Generation Internet. This accountability component, called Re-ECN (Explicit Congestion Notification)[19] would allow users to be accountable for the congestion they generate.

These are the three main components of Trilogy's architecture, see Figure 1.1, and their interaction is detailed in [3].

The rest of the article will detail the multi-path routing component of the architecture, analyzing its most important motivations, different alternatives and particular proposals.

1.3 Multi-path routing

The adoption of a multi-path routing solution in the Internet will imply changes. Such changes imply costs that need to be assumed by the different business roles and in order to deploy an effective solution it is critical to have the right motivations for the affected parties. In particular, it is critical to have the right incentives, i.e. a scheme where the parties that have to pay for the costs also get some of the resulting benefits. In this section, some motivations to deploy a multi-path routing solution for the Internet are presented from the perspective of each of the stakeholders involved.

1.3.1 Higher network capacity

It is fairly intuitive to see that when multi-path routing is used it is possible to push more traffic through the network (and particularly when it is used in conjunction with congestion-dependent load-distribution). This is so basically because the traffic will flow through any path that has available capacity, filling unused resources, while moving away from congested resources. Using generalized cut constraints approach (see [17] and [14]), it is actually possible to model the capacity constraints for logical paths existing in a network and prove that the set of rates that a multi-path routing capable network that uses logical paths can accommodate is larger than set of input rates in the same network using uni-path routing directly over the physical paths. This basically means that the network provider can accommodate more traffic with its existing network, reducing its operation costs and becoming more competitive. From end users perspective, they will be able to push more traffic through their existing providers.

1.3.2 Scalable traffic engineering capabilities

The Internet global routing table contains over 300,000 entries and it is updated up to 1,000,000 times a day, according to recent statistics [12], resulting in the scalability challenges identified by the Internet community. There are multiple contributors to the global routing table, but about half of the routing table entries are more specific prefixes i.e. prefixes that are contained in less specific ones [18]. In addition, they exhibit a much less stable behavior than less specific prefixes, making them a major contributors to the BGP churn. Within those more specific prefixes, 40% can be associated with traffic engineering techniques [18] used by the ASes (Autonomous Systems) to change the normal BGP routing. Among the most compelling reasons for doing traffic engineering, we can identify avoiding congested paths. This basically means that ASes inject more specific prefixes to move a sub-set of traffic from a congested route towards a route with available capacity. In this case, more-specific prefixes act as a unit of traffic sinks that can be moved from one route to another when a path becomes congested. While this is a manual process in BGP, because of its own nature, these more specific prefix announcements tend to be more volatile than less specific prefixes announced to obtain real connectivity. Deploying a multi-path routing architecture would remove the need to use the injection of routes for more specific prefixes in BGP to move traffic away from congested links, especially when used in combination with congestion control techniques.

1.3.3 Improved response to path changes

Logical paths that distribute load among multiple physical paths are more robust than each one of the physical paths, hence, using multiple logical paths would normally result in improved fault tolerance. However, it can be argued that current redundancy schemes manage to use alternative paths when the used path fails without needing to rely on multi-path. Nowadays, there are several mechanisms to provide fault tolerance in the Internet that would allow to switch to an alternative path in case the one actually used fails. Notably, BGP react to failures, and reroutes packets through alternate routes in case of failures but its convergence times may be measured in minutes and there are a certain amount of failures that are transparent to BGP because of aggregation. There are also other means to provide fault tolerance in the network, such as relying on the IGP, or in local restoration, and although some of them can have good response times, they are not able to deal with all the end-to-end failure modes, since they are not end-to-end mechanisms. On the other hand, end to end mechanisms for fault tolerance have been proposed, such as HIP (Host Identity Protocol) [20] or the REAP (REAchability Protocol) [6]. However, in all these cases, only one path is used simultaneously and because they are network layer protocols, it is challenging to identify failures in a transport layer agnostic way, resulting in response times that are measured in seconds [6]. The improved response to path changes that multi-path routing would allow is relevant to the end users, since they will obtain better resiliency, but it is also a motivation for the network operator, since the path change events would behave in a more congestion friendly manner.

1.3.4 Enhanced security

Logical paths that distribute load among multiple physical paths exhibit superior security characteristics than the physical paths. This is so due to a number of reasons. For instance, man-in-the-middle attacks are much harder to achieve, since the attacker needs to be located along the multiple paths, and a single interception point is unlikely to be enough. Same argument applies to sniffers along the path, resulting in enhanced privacy features. In addition, logical paths are more robust against denial-of-service attacks against any of the links involved in the paths, since attacking any link would simply imply that the traffic will move to alternative physical paths that compose the logical path. The result is that a multi-path routing based architecture results in improved security. This is a benefit for the end-user that would take advantage of the improved security features.

1.3.5 Improved market transparency

Consider the case where a site has multiple paths towards a destination through multiple transit providers. Consider now that it uses the different logical paths that include physical paths through its different transit providers. Since traffic will flow based on congestion pricing, at the end of the day, the client may be able to have detailed information about how much traffic has routed through each of its providers. Having more perfect information of the actual quality of the service purchased, allows clients to make more informed decisions about their providers, fostering competition and improving the market.

1.4 Multi-path BGP

In the Internet there are already some deployed alternatives in order to support the simultaneous usage of multiple paths to reach a certain destination. The best known solutions are the ones being used within the domain of a particular provider (intra-domain routing) since traffic can be conveniently controlled and directed while all the routing devices are under a single management entity and the multi-path solution is typically common throughout the domain. However these solutions are not directly applicable to the inter-domain routing framework since there are other important factors beyond the technical ones that must be considered, which are mainly related with policy and economic constraints.

This section provides an overview of the most relevant solutions proposed so far both for the intra-domain and the inter-domain environments, finally focusing on the motivations for other multi-path BGP alternatives and also exposing some of the problems that may arise when designing multi-path inter-domain protocols.

1.4.1 Intra-domain multi-path routing

One of the easiest frameworks to implement multi-path routing would be to use IP source routing, as far as the end systems are provided with enough topological information so as to calculate these multiple paths. However, apart from the security concerns on the use of source routing [4] and the lack of support for IP source routing in current routers, it also has some drawbacks strictly talking from a multi-path practical perspective like for example: scalability problems due to the provision of topology maps to the end systems, worse use of resources of the provider since traffic will typically be unbalanced in the network and some links may remain unused while others may become congested or less flexible routing scheme since IP traffic will normally flow following the same paths. One interesting feature that this scheme would enable is the usage of disjoint paths: since path selection is centrally done by the end systems it can be guaranteed that the selected paths do not partially overlap, improving resiliency that way.

Link state protocols like OSPF (Open Shortest Path First) [22] explicitly allow equal cost multi-path routing. When multiple paths to the same destination have the same cost, an OSPF router may distribute packets over the different paths. The Dijkstra algorithm makes sure each path is loop-free . A round robin schedule could be easily used for this, but there are protocols such as TCP that perform better if packets belonging to a certain flow follow the same path and for this, more complex techniques are often used (see [11] or [5]). Equal cost multi-path in general provides a better use of network resources than normal uni-path routing schemes and a better resilience and this is completely transparent to the end user. However, some times it may not provide enough path diversity so a more aggressive multi-path routing technique may be used. A well-known alternative for the intra-domain routing is the unequal cost multi-path used in EIGRP (Enhanced Interior Gateway Protocol) [1]. Unequal cost multi-path solutions imply using some other routes in addition to the shortest ones but these new routes do not guarantee loop freeness in the routing infrastructure. This is solved in most protocols using loop-free conditions like the ones defined in [27]. In essence, this comes down to a router only advertising routes to neighboring routers that have a higher cost than the routes that the router itself uses to reach a destination. See section 1.4.4.2 for further details.

OSPF is also capable of doing multi-topology routing. OSPF type-of-service routing (updated to be more general in [23]), overlays multiple logical topologies on top of a single physical topology. A single link may have different costs in different topologies. As such, the shortest paths will be different for different topologies. However, packets must be consistently forwarded using the same topology to avoid loops. This is different from other types of multi-path routing, where each link that a packet traverses brings the packet closer to its destination, in the sense that the cost for reaching the destination is smaller after each hop. This is also true in multi-topology routing, but only when a packet stays within the same topology, so multi-topology routing requires more complex IP forwarding function than regular hop-by-hop forwarding. If a packet is moved from one topology to another, it could face a higher cost towards its destination after traversing a link. A second topology change then creates a loop. This makes multi-topology routing appropriate for link state protocols where all routers have the same information, less suitable for distance vector protocols where each router only has a limited view of the network, and unsuitable for policy-based routing protocols such as BGP, where contractory policies may apply in different parts of the network.

1.4.2 Inter-domain multi-path routing

For the inter-domain environment there are also existing solutions providing limited multi-path routing. For instance, when there are parallel links between two eBGP (External BGP) neighbors, operators may configure a single BGP session between the two routers using addresses that are reachable over each of the links equally. This is normally done by assigning the address used for the BGP session (and thus, the NEXT_HOP address) to a loopback interface, and then having static routes that tell the router that this address is reachable over each of the parallel links. The BGP routes exchanged will now have a next hop address that is not considered directly reachable. Even though the BGP specification does not accommodate for this, implementations can typically be configured to allow it. They will then recursively resolve the BGP route's NEXT_HOP address, which will have multiple resolutions in the multi-path case. This will make the IP forwarding engine distribute packets over the different links without involvement from the BGP protocol. For iBGP (Internal BGP), the next hop address is not assumed to be directly reachable, so it is always resolved recursively. So in the case of iBGP, the use of multiple paths depends on the interior routing protocol or the configuration of static routes.

BGP is also capable of explicitly managing equal cost multi-path routing itself. This happens when a BGP router has multiple eBGP sessions, the router is configured to use multiple paths concurrently and the routes learned over different paths are considered sufficiently equal. The latter condition is implementation specific. In general, if the LOCAL_PREF , AS_PATH and MED are all equal, routes may be used concurrently. In this case, multiple BGP routes are installed in the routing table and packets are forwarded accordingly. Because all the relevant BGP attributes for the routes over different paths are the same, there is no impact to BGP loop detection or other BGP processing.

Apart from these existing solutions that are currently being applied, there are some other proposal that are worthwhile mentioning.

The source routing alternative is also possible for the inter-domain and similar comments would apply here than the ones made for the intra-domain (see [30] and [13]). In addition, one of the most important considerations now is that lack of flexibility for intermediate providers to apply their policies if packets come with a fixed path from the origin. In intra-domain routing this is not an issue since it is all related with a single provider but for the inter-domain routing, this is critical.

Some other solutions consist on overlays that run on top of the generic Internet routing mechanism. Additional paths are normally obtained tunneling packets between different nodes that belong to the overlay. The typical problems related to overlays are the additional complexity associated with the tunneling set up mechanisms and the overhead that the tunnels themselves introduce. One of these proposals is MIRO (Multi-path inter-domain ROuting, [29]) that reduces the overhead during the path selection phase by means of a co-operative path selection involving the different intermediate ASes (additional paths are selected on demand rather than disseminating them all every time). Another alternative is RON (Resilient Overlay Networks, [2]) that builds an overlay on top of the Internet routing layer and continuously probes and monitors the paths between the nodes of the overlay. Whenever a problem is detected, alternate paths are activated using the overlay.

Another recent solution is called path splicing [21] following the multi-topology idea and generating the different paths by running multiple protocol instances to create several trees towards the destination but without sharing many edges in common. While normal multi-topology schemes will just use different topologies for different packets (or flows) the idea here is to allow packets to switch between topologies at any intermediate hop, increasing the number of available paths for a given source-destination pair. The selection of the path is done by the end systems including certain bits in the the packets that select the forwarding table that must be used at each hop. This proposal claims for a higher reliability and faster recovery than normal multi-topology alternatives providing less overhead than overlay-based solutions.

1.4.3 Motivations for other solutions

Due to different reasons the previous proposals have still not been promoted into real alternatives. In this chapter two proposals are introduced based on the following motivations and assumptions for an early adoption:

- Change BGP semantics as little as possible.
- Change BGP routers as little as possible.
- Be interoperable with current BGP routers.
- Provide more path diversity that exists today.

In addition, it is worth noting that any solution should comply with the peering/transit Internet model based on economic considerations (see [8]). The rationale for this model is to realize that in most cases a site only carries traffic to or from a neighbor as a result of being paid for this (becoming a provider that serves a customer, or serving a paid peering), or because of an agreement exists in which both parties obtain similar benefit (peering). This results in the requirement to enforce two major restrictions:

- Egress route filtering restrictions: customer ASes should advertise its own prefixes and the prefixes of its customers, but they should never advertise prefixes received from other providers (for example, an AS should never advertise to its peers more than its own prefixes and those of its customers). In this way, a site does not offer itself to carry traffic for a destination belonging to a site for which it is not going to obtain direct profit.
- Preferences in route selection: routers should prefer customer links over peering links because sending and receiving traffic over customer links makes them earn money, and peering over provider links, because peering links at least does not cost them money. According to this, the multi-path route selection process can aggregate routes from many different customer links; or many peering links; or many provider links; but it can never mix links associated to different relationship types. Note that the administrator may even have specific preferences for routes received from neighbors with the same relationship with the site, because of economic reasons, traffic engineering, etcetera.

As a result of the peering/transit model, paths in the Internet can start going “up” from the originating site to a provider, and up to another provider, many times until it reaches a peering relationship, and then descend to a customer of this site, descend again, many times until it reaches the destination. Since it is impossible to find paths in which descending from a site to a customer and ascending again, or paths in which a peering link is followed by an ascending turn to a provider, it is said that Internet is “valley-free” [8] as a result of the application of the peering/transit model.

The “valley-free” model suggests that a loop in the advertising process (i.e. a route advertised to a site that already contains in the AS_PATH the AS number of that site) can only occur for a route received by a provider. This is because a customer or peer of a site S, cannot advertise a route that has been previously advertised by S, according to the restrictions stated above. The valley-free condition also assures that a route containing S that is received from a provider P1(S) was advertised by S to another provider. Since S only announces to its providers its own prefixes or customer prefixes, the prefixes received by any provider, whose selection would result in a loop, are its own prefix or customer prefixes. Note that these routes would never be selected because either the destination is already in the site, or because it always prefers customer links to provider links. Consequently, although there is a specific mechanism in BGP for detecting loops in the routes, the application of the peer/transit model by itself would be enough to assure that loops never occur. Of course, loop prevention mechanisms must exist in order to cope with routing instabilities, configuration errors, etcetera. However, we can extend this reasoning to the multi-path case to state that, if any multi-path BGP strategy complies with the peering/transit model, as requested before, the aggregation of routes with equal condition (just customer routes; if not, just peering routes; and if not just provider-received routes) will not result in route discarding due to loop prevention in the steady state for well-configured networks. However, any multi-path BGP mechanism must provide loop prevention to cope with transient conditions and configuration errors.

In this chapter we present two proposals that share some mechanisms, such as part of the route selection approach, and differ in others, such as the loop prevention mechanism.

1.4.4 LP-BGP and MpASS

1.4.4.1 Route selection and propagation

Because a router running BGP tends to receive multiple paths to the same destination from different neighboring routers, the modifications to allow for the use of multiple paths can be limited to each individual router and modifications to the BGP protocol are unnecessary. The selection process for multi-path BGP should take as a starting point the rules for uni-path BGP, deactivating the rules that are used for tie-breaking among similar rules to allow the selection of multiple routes instead of just a single one. Note that the more rules that are deactivated, the larger number of routes with the same preference can be selected for multi-path forwarding. However, only routes that are equivalent for the administrator must be selected, resulting this preference from economic reasons, traffic engineering considerations, or in general any policy that the administrator wants to enforce. So a modified multi-path router first applies normal BGP policy criteria and then selects a subset of the received paths for concurrent use. The attributes and rules through which relevant preferences of the administrator are enforced, in the order in which they are applied, are:

- Discard routes with lowest LOCAL_PREF . This rule enforces any specific wish of the administrator, and is the rule used to assure that only routes received from customers are selected; or if no routes from customers exist, only routes received from peers; or if none of the previous exist, routes received from providers.
- Discard routes with highest MED. This rule is used to fulfill the wishes of the customers in order to implement "cold potato" routing so that customers costs in terms of transit cost are reduced.
- Discard lowest ORIGIN. This rule is used in some cases as a traffic-engineering tool. If not, the impact of its application is low, since almost all routes should have equal ORIGIN attribute.
- Discard iBGP routes if eBGP routes exist. It is used to deploy hot-potato routing, which may be relevant to reduce internal transit costs. In addition, it also eliminates internal loops in route propagation. When applied, routers receiving a route from a external neighbor uses only external neighbors, so internal loops never occur. Routers not receiving a route from an external neighbor selects the router inside the AS that will send the packet out of the AS.
- Discard routes with highest cost to NEXT_HOP. This is also used to enforce hot-potato routing. However, some relaxation on this rule can be introduced, provided that prevention of loops in intra-domain forwarding is achieved by means such as some kind of tunneling like MPLS.

The rest of the rules (selecting route received from router with minimum loopback address, etcetera) are provided to ensure uniqueness in the result, so they can be removed for multi-path routing.

Therefore, a modified router first applies normal BGP policy criteria and then selects a subset of the received paths for concurrent use. Note that multiple paths mainly come from the possibility of ignoring AS_PATH length (although some conditions on this length could be established for accepting a route), and from accepting routes with different NEXT_HOP distances.

1.4.4.2 LP-BGP: Loop-freeness in multi-path BGP through propagating the longest path

In this particular proposal, after obtaining the different paths that will be installed in the forwarding table for the same destination prefix, the path with longest AS_PATH length to upstream ASes will be disseminated to neighboring routers where allowed by policy. Although disseminating a path that has a larger number of ASes in its AS_PATH seems counterintuitive, it has the property of allowing the router to use all paths with a smaller or equal AS_PATH length without risking loops (see Figure 1.2).

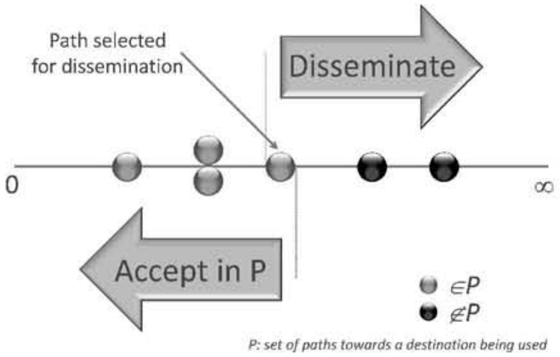

Figure 1.2 Multi-path selection in LP-BGP

However, this change has the implication that there is no longer a one-to-one relationship between the paths that packets follow through the network and the path that is advertised in BGP. The resulting obfuscation of the network's topology as seen by observers at the edge can either be considered harmful, for those who want to study networks or apply policy based on the presence of certain intermediate domains, or useful, for those intent on hiding the inner workings of their network.

The multi-path BGP modifications allow individual ASes to deploy multi-path BGP and gain its benefits without coordination with other ASes. Hence, as an individual BGP router locally balances traffic over multiple paths, changes to BGP semantics are unnecessary.

Under normal circumstances, the BGP AS_PATH attribute guarantees loop-freeness. Since the changes allow BGP to use multiple paths concurrently, but only a single path is disseminated to neighboring ASes, checking the AS_PATH for the occurrence of the local AS number is no longer sufficient to avoid loops. Instead, the the Vutukury/Garcia-Luna-Aceves LFI (Loop-free Invariant) [27] conditions are used to guarantee loop-freeness .

Intuitively, these conditions are very simple: because a router can only use paths that have a lower cost than the path that it disseminates to its neighbours (or, may only disseminate a path that has a higher cost than the paths that it uses), loops are impossible. A loop occurs when a router uses a path that it disseminated earlier, in which case the path that it uses must both have a higher and a lower cost than the path that it disseminates, situations that can obviously not exist at the same time. When the following two LFI conditions as formulated by Vutukury and Garcia-Luna-Aceves are satisfied, paths are loop-free:

$$FD_j^i \leq D_{ji}^k \quad k \in N^i$$
$$S_j^i = \{k | D_{jk}^i < FD_j^i \wedge k \in N^i\}$$

"where D^k_{ji} is the value of D^k_j reported to i by its neighbor k; and FD^i_j is the feasible distance of router i for destination j and is an estimate of D^i_j, in the sense that FD^i_j equals D^i_j in steady state but is allowed to differ from it temporarily during periods of network transitions." [27]. D^k_j is the distance or cost from router k to destination j. N_i is the set of neighbors for router i and S^i_j is the successor set that router i uses as next hop routers for destination j.

Our interpretation of the two LFI conditions as they relate to BGP is as follows:

$$cp(p_r) < cp_r(p_r)$$
$$P = \{p | cp(p) \leq cp(p_r) \wedge p \in \pi\}$$

Where P is the set of paths towards a destination that are under consideration for being used and π is the set of paths towards a destination disseminated to the local router by neighboring routers. p_r is the path selected for dissemination, $cp_r(x)$ the cost to reach a destination through path x that is reported to other routers and the cost $cp(x)$ is taken to mean the AS_PATH length of path x in the case of eBGP and the interior cost for iBGP. The interior cost is the cost to reach a destination as reported by the interior routing protocol that is in use.

Because the local AS is added to the AS_PATH when paths are disseminated to neighboring ASes, the smaller and strictly smaller requirements are swapped between the two conditions.

The BGP-4 specification [24] allows for the aggregation of multiple prefixes into a single one. In that case, the AS numbers in the AS_PATH are replaced with one or more AS_SETs, which contain the AS numbers in the original paths. Should the situation arise where a topology is not valley-free [8] and there is both a router that implements multi-path BGP as described in this chapter as well as, in a different AS, a router that performs aggregation through the use of AS_SETs, then routing loops may be possible. This is so because, depending on the implementation, a router creating an AS_SET could shorten the AS_PATH length and break the limitations imposed by the LFI conditions. To avoid these loops, P may either contain a single path with an AS_PATH that contains an AS_SET, or no paths with AS_PATHs that contain AS_SETs. Note that AS_SETs are rarely used today; a quick look through the Route Views project data reveals that less than 0.02% of all paths have one or more AS_SETs in their AS_PATH [25].

All paths that remain in the multi-path set after the previous steps and after applying policy are installed in the routing table and used for forwarding packets. The determination of traffic split ratios between the available paths is a topic for future work.

At this point, the path with the longest AS_PATH within P is selected for dissemination to BGP neighbors. As a result of the LFI conditions, multi-path-aware ASes will suppress looped paths with a multi-path-aware AS in the looped

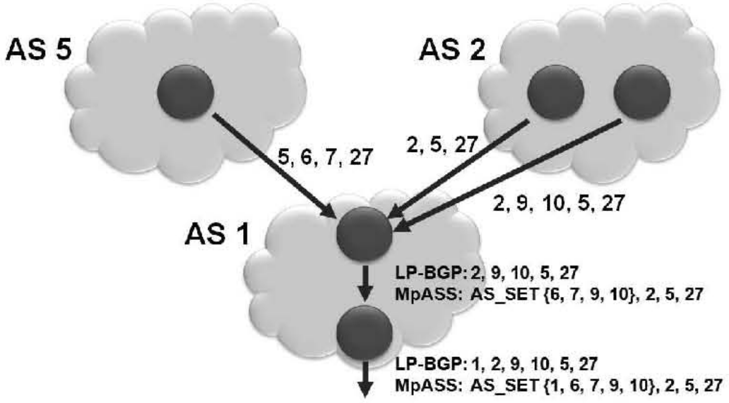
AS 5
AS 2
5, 6, 7, 27
2, 5, 27
2, 9, 10, 5, 27
AS 1
LP-BGP: 2, 9, 10, 5, 27
MpASS: AS_SET {6, 7, 9, 10}, 2, 5, 27
LP-BGP: 1, 2, 9, 10, 5, 27
MpASS: AS_SET {1, 6, 7, 9, 10}, 2, 5, 27

Figure 1.3 BGP propagation in LP-BGP and MpASS

part of the path, while regular BGP AS_PATH processing suppresses looped paths with no multi-path-aware ASes in the looped part of the path. To avoid loops for non-multi-path-aware iBGP routers, the selected path is also not disseminated over any BGP session through which the router learned a path that is in the multi-path set, and If the router previously disseminated a path over a session towards a neighboring router that supplied a path in the selected multi-path set P, it now sends a withdrawal for the multi-path destination.

1.4.4.3 MpASS: Multi-path BGP with AS_SETs

The main idea behind MpASS is to include in the AS_PATH all the AS numbers resulting from the union of the AS_PATH attributes of the routes aggregated so far. In particular, the AS_PATH is obtained by concatenating an AS_SEQUENCE structure containing the AS_PATH corresponding to the route that the BGP router would select from applying BGP uni-path selection rules, and an AS_SET structure that includes all the AS numbers of the rest of the routes, and the AS number of the site. This particular construction mechanism assures that all AS numbers are included and the length of the AS_PATH structure as defined for the AS_PATH length comparison rule [24], is equal to the length of the AS_PATH of the best route plus 1 (as it would occur for legacy uni-path BGP routers). In this way, when a legacy route applies the rule of discarding routes with larger AS_PATH length, this multi-path route is not penalized compared to the uni-path route that it would have generated.

Loop prevention is enforced by the check performed by regular uni-path BGP and it is not necessary to define any additional mechanism or particular condition, i.e. discarding routes that contain the AS number of the site of the router receiving the advertisement (see Figure 1.3). An additional characteristic is that the inclusion of all the AS numbers of the sites that may be traversed by a packet sent to the destination allows the application of policies based on the particular AS traversed when selecting a route. Legacy BGP routers receive a route that is indistinguishable to a regular BGP route, and if they select it, packets may benefit from the multiple available paths.

1.5 Conclusions and future work

Multi-path routing presents many advantages when compared with single-path routing: higher network capacity, scalable traffic engineering capabilities, improved response to path changes and better reliability, enhanced security, improved market transparency.

For the intra-domain routing environment there are different solutions that can be applied (and effectively are), and the fact of having the deployment constrained to a single routing domain particularly facilitates this task (only in the interior of a provider's network).

In the inter-domain routing framework, the situation is more complex because most of the different existing proposal imply important changes in the well established inter-domain communication technology based on BGP, linking different providers and each one with its own interests and requirements.

The European research and development project Trilogy considers multi-path routing as one of its main objectives. In the project, multi-path routing is considered together with congestion control mechanisms, and the different Internet economic drivers so as to try to improve the existing Internet communication mechanisms by means of providing a synergic solution based on the liaison of these three areas.

This chapter is focusing on one of these areas, the multipath routing, and we have presented two mechanisms for providing multiple routes at the inter-domain level that are being considered in the project. The mechanisms differ in the way routes are selected and how loop prevention is enforced. The first on, LP-BGP, has the potential to reduce the number of BGP updates propagated to neighboring routers, as updates for shorter paths do not influence path selection and are not propagated to neighboring routers. However, in longer paths there is more potential for failures, so the inclusion of long paths in the set of paths that a multi-path router uses, may expose it to more updates compared to the situation where only short paths are used. When propagating just the longest path BGP no longer matches the path followed by all packets. The second proposal (MpASS) allows the selection of routes with any AS_PATH length, since loop prevention relies on transporting the complete list of traversed AS numbers.

One difference among them is that LP-BGP may propagate a route with an AS_PATH larger than the best of the aggregated routes, so that the result of a multi-path aggregation may be a route less attractive to other BGP routers (presenting longer paths to customers may put service providers at a commercial disadvantage). Still, propagating the longest path has robust loop detection properties and operators may limit acceptable path lengths at their discretion, so the second disadvantage is relatively minor (they could require for instance all best routes to be equal length).

On the other hand, MpASS may suffer from excessive update frequency, since each time a new path is aggregated in a router, a new Update must be propagated to all other routers receiving this route, to ensure that loop prevention holds (note that in the uni-path case, BGP only propagates a route if the newly received improves the previous one, while in this case many routes may be gradually added to the forwarding route set). This problem can be relieved by setting a rate limit to the aggregation process.

As part of the future work we plan to do a deeper analysis of the stability properties of both protocols, i.e. routing convergence and convergence dynamics. Some intuition around the routing algebra theory developed by Griffin and Sobrinho [9] suggests the LP-BGP is stable and that MpASS is assured to be stable if only routes with equal AS_PATH length are aggregated, although more analysis is required to determine if the use of different lengths may lead to stable solutions.

Finally, an evaluation of the effect of applying these mechanisms in the real Internet is required in order to analyze the path diversity situation: is the current number of available paths too low, or too high? Is it enough to use equal length AS_PATH routes? What is the cost added to the already stressed inter-domain routing system? More work will continue in Trilogy, stay tuned.

Subject index

References

[1] Albrightson B., Garcia-Luna-Aceves J., Boyle J. (1994). EIGRP–A fast routing protocol based on distance vectors. In *Proc. Networld/Interop 94, Las Vegas. Proceedings.* 136–147.

[2] Andersen D., Balakrishnan H., Kaashoek F., Morris R. (2001). Resilient overlay networks. In *ACM SOSP Conference. Proceedings.*

[3] Bagnulo M., Burness L., Eardley P., García-Martínez A., Valera F., Winter R. (2009). Joint Multi-path Routing and Accountable Congestion Control. In *ICT Mobile Summit. Proceedings.*

[4] Bellovin S. (1989). Security problems in the TCP/IP protocol suite. *ACM SIGCOMM Computer Communication Review 19,* 2, 32–48.

[5] Chim T., Yeung K., Lui K. (2005). Traffic distribution over equal-cost-multi-paths. *Computer Networks 49,* 4, 465–475.

[6] de la Oliva A., Bagnulo M., García-Martínez A., Soto I. (2007). Performance analysis of the reachability protocol for ipv6 multihoming. *Lecture Notes in Computer Science* **4712**, 443–454.

[7] Fortz B., Thorup M. (2000). Internet traffic engineering by optimizing OSPF weights. In *IEEE INFOCOM 2000. Proceedings.* Vol. 2. 519–528.

[8] Gao L., Rexford J. (2001). Stable Internet routing without global coordination. *IEEE/ACM Transactions on Networking 9,* 6, 681–692.

[9] Griffin T., Sobrinho J. (2005). Metarouting. *ACM SIGCOMM Computer Communication Review 35,* 4, 1–12.

[10] Han H., Shakkottai S., Hollot C., Srikant R., Towsley D. (2006). Overlay TCP for multi-path routing and congestion control. *IEEE/ACM Transactions on Networking 14,* 6, 1260–1271.

[11] Hopps C. (2000). Analysis of an Equal-Cost Multi-Path Algorithm. *RFC2992*.

[12] Huston G. (2009). Potaroo.net. *[Online]. Available: http://www.potaroo.net/*.

[13] Kaur H., Kalyanaraman S., Weiss A., Kanwar S., Gandhi A. (2003). BANANAS: An evolutionary framework for explicit and multipath routing in the Internet. *ACM SIGCOMM Computer Communication Review 33,* 4, 277–288.

[14] Kelly F. (1991). Loss networks. *The annals of applied probability 1,* 3, 319–378.

[15] Kelly F., Voice T. (2005). Stability of end-to-end algorithms for joint routing and rate control. *ACM SIGCOMM Computer Communication Review 35,* 2, 5–12.

[16] Khanna A., Zinky J. (1989). The revised ARPANET routing metric. *ACM SIGCOMM Computer Communication Review 19,* 4, 45–56.

[17] Laws C. (1992). Resource pooling in queueing networks with dynamic routing. *Advances in Applied Probability 24,* 3, 699–726.

[18] Meng X., Zhang B., Huston G., Lu S. (2005). IPv4 address allocation and the BGP routing table evolution. *ACM SIGCOMM Computer Communication Review 35,* 1, 71–80.

[19] Moncaster T., Briscoe B., Menth M. (2009). Baseline Encoding and Transport of Pre-Congestion Information. *IETF draft. draft-ietf-pcn-baseline-encoding-02*.

[20] Moskowitz R., Nikander P., Jokela P., Henderson T. (2008). Host Identity Protocol. *RFC5201*.

[21] Motiwala M., Elmore, M. Feamster N., Vempala S. (2008). Path Splicing. In *ACM INFOCOM. Proceedings.*

[22] Moy J. (1998). OSPF Version 2. *RFC2328*.

[23] Psenak P., Mirtorabi S., Roy A., Nguyen L., Pillay-Esnault P. (2007). Multi-Topology (MT) Routing in OSPF. *RFC4915*.

[24] Rekhter Y., Li T., Hares S. (2006). A Border Gateway Protocol 4 (BGP-4). *RFC4271*.

[25] Routeviews. (2009). University of Oregon Route Views Project. *[Online]. Available: http://routeviews.org/*.

[26] Sridharan A., Guerin R., Diot C. (2005). Achieving near-optimal traffic engineering solutions for current OSPF/IS-IS networks. *IEEE/ACM Transactions on Networking 13,* 2, 234–247.

[27] Vutukury S., Garcia-Luna-Aceves J. (1999). A simple approximation to minimum-delay routing. In *ACM SIGCOMM. Proceedings.* ACM New York, NY, USA, 227–238.

[28] Wischik D., Handley M., Bagnulo M. (2008). The resource pooling principle. *ACM SIGCOMM Computer Communication Review 38,* 5, 47–52.

[29] Xu W., Rexford J. (2006). MIRO: Multi-path interdomain routing. In *Proceedings of the 2006 conference on Applications, technologies, architectures, and protocols for computer communications.* Vol. 36. ACM New York, NY, USA, 171–182.

[30] Zhu D., Gritter M., Cheriton D. (2003). Feedback based routing. *ACM SIGCOMM Computer Communication Review 33,* 1, 71–76.