Towards a systematic understanding of route reflection

Jong Han Park, Ricardo Oliveira, Shane Amante, Danny McPherson, Lixia Zhang

University of California, Los Angeles
Computer Science Department
Technical Report # 100006
Feb. 16, 2010

Towards a Systematic Understanding of Route Reflection

Jong Han Park * jpark@cs.ucla.edu
Ricardo Oliveira * rveloso@cs.ucla.edu
Shane Amante† Shane.Amante@level3.com
Danny McPherson‡ danny@arbor.net
Lixia Zhang* lixia@cs.ucla.edu

ABSTRACT

The original BGP design requires all BGP speakers within an autonomous network to be directly connected with each other to create a full mesh, and BGP update messages to be propagated to directly connected neighbors only. This requirement leads to BGP session scalability concerns in networks with large numbers of BGP routers. *Route reflection* was proposed by the operational community as a solution to address this scalability problem and has been widely deployed for a long time. However, measurement and analysis studies occurred only recently to understand its pros and cons. In this paper we provide an overview of route reflection, summarize the discoveries from published literature, and discuss the advantages and disadvantages of using route reflection as compared to using a fully connected iBGP mesh. We also use the route reflection deployment in a large ISP as a case study to show how one can use well engineered route reflector placement to overcome certain drawbacks as well as remaining issues for future study.

1. INTRODUCTION

Route reflection [1] was proposed in 1996 to address a BGP scalability problem and has been widely deployed since then. In the original BGP design [11], iBGP is the component to disseminate BGP updates within an autonomous system, and to avoid routing loop, all iBGP speakers are required to directly connect with each other, and BGP update messages are forwarded only to directly connected neighbors. In a large network with hundreds or even thousands of BGP routers, this full-mesh requirement results in a large number of BGP sessions on each router as well as a high operational cost whenever a router is added or removed, because all iBGP sessions are managed through manual configurations.

Route reflection is one of the two proposed solutions to address this scalability problem; the other one is *AS confederations* [13]. Between the two, route reflection has seen a larger deployed base. However before its deployment rolled out more than 10 years ago, the design did not go through thorough analysis studies. Only recently several studies appeared to analyze the full impacts of route reflection on the overall routing system performance. Work by others and ourselves show that route reflection can decrease the network's robustness to failures, introduce delayed routing convergence, reduce path diversity, lead to sub-optimal routes and even cause data forwarding loops.

In this paper we provide a comprehensive overview of BGP route reflection, including a discussion of its pros and cons as well as an outlook into its remaining issues. Our contributions are three-fold. First, in Section 2, we provide an overview of how route reflection operates and explain the pros and cons of route reflection in detail. Second, in Section 3, we provide a case study of route reflection deployment in a large ISP, which illustrates how one can use well engineered route reflector placement to overcome certain drawbacks in the route reflection deployment and further scale the routing system, without requiring any protocol or implementation changes. Third, we discuss interesting aspects observed from the ISP's route reflection deployment and identify remaining issues in achieving the goals of both efficient routing information dissemination and system scalability.

*University of California, Los Angeles.
†Level-3 Communications Inc.
‡Arbor Networks

2. ROUTE REFLECTION

In this section we first present a brief review of BGP basics, followed by an overview of route reflection; interested readers are referred to [1, 8] for more detailed descriptions on route reflection operations. We then provide an analysis of the pros and cons of the basic route reflection scheme. Generally speaking, the advantages of route reflection are well recognized, but its potential drawbacks did not receive much attention until lately. However as we will show in Section 3, some of the drawbacks can be mitigated through well engineered placements and configurations of route reflectors.

2.1 Routing in the Internet

The Internet is made of tens of thousands of different networks called Autonomous Systems (ASes) and BGP is the glue that connects them together. Routers in different ASes set up BGP sessions in between to exchange BGP routing updates (inter-domain routing). Such sessions are called eBGP sessions. BGP sessions are also used to exchange BGP routing updates between routers within the same AS (intra-domain routing), and these sessions are called iBGP sessions.

All routing protocols must have effective means to prevent routing loops. In eBGP, routers detect routing loops at inter-AS level by inspecting the *AS-path* attribute carried in BGP messages. A router will drop a BGP message if AS-path already contains its own AS number. To avoid routing loops in iBGP, the original design requires that all BGP routers in the same AS be directly connected to each other to create a full mesh, and that reachability information learned from any iBGP speaker must not be forwarded to any other iBGP speaker. This full-mesh requirement leads to a large number of BGP sessions on each router, and a high operational cost because operations such as creating, modifying, or removing iBGP sessions all require operator intervention.

The following example shows the inter-working between eBGP and iBGP. In Figure 1, AS2 maintains eBGP sessions with AS1 and AS3 using routers R_1 and R_3, respectively. Inside AS2, all BGP routers are inter-connected through iBGP sessions. When AS1 an-

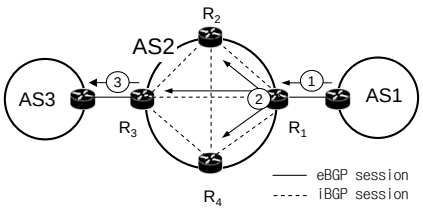
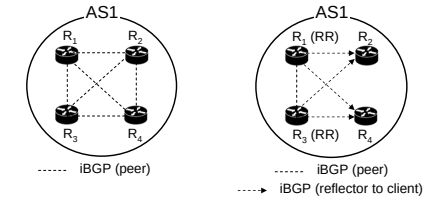


Figure 1: BGP in the Internet



(a) Full-mesh iBGP (b) iBGP with route reflection

Figure 2: Different iBGP topologies

nounces a destination reachability to AS2 over the eBGP session with R_1, R_1 will propagate the information to all the other three routers in AS2 over its direct iBGP sessions with them. R_3 further propagates this reachability information to its eBGP neighbor, in this case the router in AS3. Note that within AS2, the reachability message traverses only *one* iBGP hop from R_1 to all the other routers, thus the information does not loop back to the originating router R_1.

The total number of iBGP sessions in an AS with N BGP routers is N*(N-1)/2. For example, the total number of iBGP sessions in AS2 is 4*(4-1)/2 = 6 as shown in Figure 1. Similarly, the total number of sessions for an AS with 10, 100, or 1,000 routers would be 45, 4,950, or 499,500, respectively. Today, the number of routers in a typical large AS can be several hundreds or even over a thousand, making the full mesh iBGP interconnections infeasible.

To alleviate this scalability problem due to full-mesh iBGP interconnections, the vendor and operator communities proposed two solutions in 1996: route reflection and AS confederations. Both solutions have been deployed in operational networks, in some cases AS confederations are combined with route reflection. Overall, route reflection has a wider deployed base and is the focus of this paper.

2.2 Basic Operation of Route Reflection

The simplest model of route reflection operation is to select one BGP router to be *Route Reflector* (RR), and have all the other routers set up iBGP sessions with the route reflector. The route reflector receives BGP update messages from each iBGP speaker and forwards (or reflects) them to all other iBGP speakers. Because the route reflector can forward updates among iBGP speakers, iBGP speakers no longer need to connect in a full-mesh. To avoid single point of failure, in real deployment an AS sets up multiple route reflectors which are interconnected in a full mesh among themselves.

Figure 2 illustrates the difference between interconnecting iBGP routers via full mesh and via route reflectors. Figure 2(a) shows an example of full-mesh iBGP interconnections, where all iBGP speakers are directly connected to each other. Figure 2(b) shows an example of interconnection through route reflectors, where R_1 and R_3 serve as route reflectors and connect to iBGP speakers R_2 and R_4. R_2 and R_4 are connected to both reflectors for redundancy. Since R_2 can learn BGP reachability information that R_4 received from its eBGP neighbor and vice versa, R_2 and R_4 do not need to interconnect. We call R_2 and R_4 *client* of R_1 and R_3. A client is an iBGP speaker that connects directly to a route reflector to learn the reachability information collected by other routes in the AS. R_1 and R_3 also connect to each other, they are *non-clients* to each other.

A route reflector does not necessarily forward received reachability information to all iBGP neighbors; the following rules apply:

- the routes received from non-client iBGP sessions are reflected only to clients.
- the routes received from client iBGP sessions are reflected to both clients and non-clients.
- the routes received from eBGP sessions are reflected to both clients and non-clients.

Because route reflectors forward reachability information learned from an iBGP speaker to another iBGP speaker, routing messages travel more than single iBGP hop and it is possible to create loops. For example in Figure 2(b), an update message originated at R_2 can come back to R_2 through more than one route reflector (R_1 and R_3 in this case), forming a loop. To avoid such loops, two new attributes are added to BGP update messages: *Cluster-list* and *Originator-ID*. A unique Cluster-ID is assigned to each route reflector. When a route reflector forwards a message, it prepends its Cluster-ID in the Cluster-list attribute. If a route reflector finds its own Cluster-ID in the Cluster-list attribute of a received update, it discards the update. In addition, every router is assigned a router ID, and the first router that injects a routing message into the network will record its router ID in Originator-ID attribute. If a router receives an update with an Originator-ID equal to its router ID, it discards the update. In Figure 2(b), R_2 will discard all updates reflected back to itself after checking that Originator-ID attribute contains its router ID.

2.3 Benefits of Route Reflection

Reduced number of iBGP sessions: Route reflection is an effective means to minimize the number of iBGP sessions in an AS. A non-RR router only needs to have a small number (typically two) of iBGP sessions with the route reflectors.

Reduced operational cost: Creating, modifying, or removing BGP sessions require human intervention. In case of full-mesh iBGP, any new router added to the network requires modifications in the configuration of all the other routers. On the other hand, in the case of route reflection, adding or removing a client iBGP router only requires configuration changes at the route reflectors the client connects to, with no impact on the rest of the routers.

Reduced RIB-in size: A BGP router keeps in memory one Routing Information Base for each neighbor to store all the received routes (RIB-in). For a router with n peers each sending p prefixes, its total RIB-in size is in the order of $n \times p$. With full-mesh iBGP sessions, n can be a very large number. With route reflection, n can be made much smaller.

Reduced number of BGP updates: With a significant reduction on the number of its iBGP neighbors, a client router naturally receives a significantly reduced number of updates. The route reflectors receive routing updates from all other routers, and since a BGP

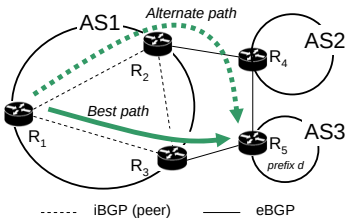


Figure 3: Packets can be dropped during path changes.

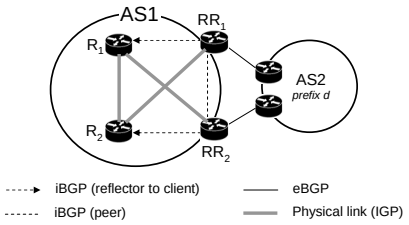


Figure 4: Route reflection with data forwarding loop

speaker only propagates the best path to its neighbors, only changes in the best path of a route reflector are propagated to its clients and non-clients. Therefore, route reflectors effectively filter out irrelevant incoming updates, in contrast to full-mesh iBGP where all BGP updates are propagated to all routers.

In a full-mesh every iBGP speaker has to process roughly the same amount of updates coming from same number of sessions, putting a high demand on capital expenses as the global routing table size continues to grow rapidly. The differentiated processing load and memory requirement for iBGP speakers in route reflection supports an heterogeneous router environment where high-end routers with more memory are used as reflectors and less capable routers can be used as clients, effectively extending their life time.

Incrementally deployable: Last but not the least, route reflection allows coexistence of route reflectors with conventional BGP routers that do not understand the concept of route reflection. A conventional BGP router B can be connected to route reflectors as a client, or a non-client (in which case B must also be connected to all other route reflectors to be part of a full RR mesh). This allows a network to perform an easy and gradual migration from the current full-mesh BGP model to the route reflection model.

2.4 Caveats of Route Reflection

Compared with the full mesh iBGP interconnections, although route reflection provides an effective alternative to address the iBGP scalability problem, it also brings several negative impacts on the overall routing system as listed below.

Robustness: With full-mesh iBGP, a single router failure has limited impact on the rest of the network. In case of route reflection, if a route reflector fails, not only all of its clients stop receiving routing updates, but also other routers can no longer get updates for the destinations connected to these client routers. To avoid such single point of failure, a client router is usually connected to two or more route reflectors.

Prolonged routing convergence: An AS with route reflection can experience prolonged routing convergence compared to the full mesh iBGP interconnections. In the full-mesh iBGP case, BGP updates travel only one iBGP hop to reach all other iBGP speakers. However with route reflection, an update message can potentially traverse more than one route reflector before reaching the final iBGP speaker. Since each route reflector needs to run the best path selection process, there is both processing delay and transmission delay to cross a reflector. In Figure 2(a), if R_2 were to distribute an update message learned from an external peer, it will send the update through the direct iBGP sessions to R_1, R_3, and R_4. On the other hand, with route reflection in Figure 2(b), R_2 will first send the update to the route reflectors (R_1 and R_3). Upon receipt of the message, R_1 will determine the best route for the given destination among all available routes. If this update changes R_1's best path to the destination, R_1 will further distribute this message to R_3 and R_4. This extra iBGP hop through the route reflector adds to the delay before R_4 can receive the updates. As we will show in the next section, an AS may deploy a hierarchy of route reflectors to further scale the routing system, which in turn introduces additional delays in the routing update propagation time.

Besides the increased delay in routing message propagations, redundant route reflectors also introduce multiple parallel internal paths to propagate the reachability information to a given destination. For example, in Figure 2(b), R_2 can see three possible paths to reach a destination announced by R_4: 1) R_2-R_1-R_4, 2) R_2-R_3-R_4 and 3) R_2-R_1-R_3-R_4 . Thus when the destination becomes unreachable, R_2 will explore all the possible internal paths before converging to the unreachable state. Had all the routers been connected in a full mesh, R_2 would have only one path to reach it and the convergence could be faster.

This delayed convergence introduced by route reflection can worsen data plane performance. In [16], Wang *et al.* found one interesting full-mesh iBGP configuration, which can cause packet drops while there is a path fail-over event. We borrow Figure 3 from [16] to explain how route reflection may further lengthen the fail-over convergence time. In the converged state with the best route available, because a router does path poisoning on known, but unused routes, R_2 withdraws the path through R_2-R_4-R_5, and uses R_3-R_5 link to reach prefix d since this route has the shortest AS-path length; at this time only R_2 knows about this alternate path to reach prefix d. When the best route to prefix d through R_3 fails, R_1 can momentarily have a period of which there is no route available to reach prefix d if the withdraw message from R_3 is received first before the update sent by R_2 with the alternate route through R_4. During this period, R_1 will drop packets to this destination network until the update from R_2 arrives. Because route reflection adds further delay in delivering BGP update messages, it worsens this data plane performance degradation. In case of hierarchical RRs, update messages travel additional iBGP hops and the impact on data plane performance can also be increased.

Data forwarding loop: In an ideal route reflection picture, where a single route reflector connects to all client routers, route reflection should not introduce any data plane loops. However in real deployment, all client routers must connect to more than one route reflector to avoid single point of failure. This redundant connectivity to route reflectors can potentially introduce loops in data plane that are subtle and defeat intuitive inspection, as we show by the following example as reported in [3, 4, 7, 12, 15].

When a client router receives a data packet, it looks up the destination address and forwards the packet to the BGP nexthop address. Depending on the IGP connectivity, there can be multiple router hops between this client router and the BGP nexthop, as is the case in Figure 4. In Figure 4, RR_1 and RR_2 can each reach prefix d in AS2, and both announce this reachability to their clients R_1 and R_2, respectively. As far as the control plane (*i.e.* BGP

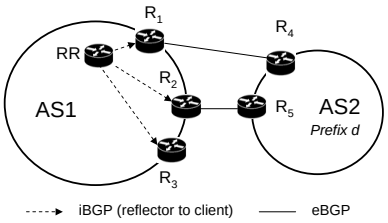


Figure 5: RR chooses *its* best route

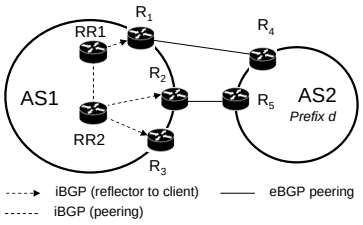


Figure 6: POP based route reflection

routing) is concerned, there is no routing loop. However when R_1 receives a data packet, it will try to send the packet to BGP nexthop RR_1 via R_2, expecting R_2 to further forward this packet to RR_1. However, R_2 believes that the BGP nexthop for destination d is RR_2 and sends the packet back to R_1, expecting that R_1 will forward the packet to RR_2. As a result of the inconsistencies between the control plane topology and physical connectivity, *i.e.* R_1 is connected to RR_1 on the control plane but connected to R_2 physically, and vise versa), packets heading to destination d would end up bouncing back and forth between R_1 and R_2.

Reduced path diversity: Path diversity is a measure to quantify the number of different routes available to reach a given destination. Maintaining a high diversity for each destination prefix is desirable because it increases the resiliency of the network against failures and offers more opportunities for traffic engineering [9, 14]. Since a route reflector only propagates its best route for a given destination, all the client routers of the same reflector can only have one best route to the destination as chosen by the route reflector. Figure 5 shows such an example: although both R_1 and R_2 are directly connected to AS2 to reach destination prefix d, if the reflector chooses R_1 as the best path to d, then R_3 has to use that path as well. Furthermore, when the link between R_1 and R_4 fails, R_3 will have to wait for some time till RR learns about the failure, discovers an alternative path to d, and then propagates the new path to all its clients. In contrast, full-mesh iBGP interconnections would have allowed R_1 and R_2 to use their direct connection to AS2 to reach prefix d, R_3 would have learned both paths, and would have been able to switch to the other path as soon as it learned about the failure from R_1 directly.

It is perceivable that, for a client router, the number of routes to a given destination is upper-bounded by the number of the route reflectors it connects to. Thus to increase path diversity one could increase the number of route reflectors a client connects to. Route reflectors are already commonly deployed in pairs to avoid single point of failure. However in the current practices, this redundancy in route reflector connections does not help increase the path diversity – the pair of route reflectors are configured as pure replicas and always make the same routing decisions. In a recent proposal [10], Raszuk *et al.* have suggested to modify the best path selection in route reflectors, so that a client router can learn different paths from different reflectors as a way to increase path diversity within an AS.

Sub-optimal routes: A route reflector chooses *its own* best paths to all the destination prefixes, and then propagates these paths to all its clients. It is almost certain that not all these best paths chosen by the reflector would be the best paths for *each* of all its clients. As a result, some client routers end up using sub-optimal paths to some destinations as reported in [2, 17]. For example in Figure 5, AS1 has two routes to reach prefix d in AS2, through R_1-R_4 and through R_2-R_5. Assuming that the Figure 5 reflects the geographical distances of the routers, the route reflector would pass to R_1, R_2, and R_3 its own best path to prefix d in AS2, which would be through R_1-R_4 (because the reflector itself is closer to R_1 than R_2). In this case, R_2 will still use its best path through R_2-R_5 because of the BGP best path selection rule that prefers path learned from eBGP over iBGP. However, R_3 will use the path R_1-R_4, the only path learned from the route reflector. R_3's shortest path to prefix d would have been through R_2-R_5, had the AS1 used full-mesh iBGP interconnections.

It is worth pointing out that, in a given network, the impact of the above potential drawbacks from route reflection heavily depends on the exact configuration and placement of route reflectors. [1] suggested several approaches to minimize the negative impact of route reflection, including placing a route reflector in the same POP with its clients, and making clients of the reflector in each POP fully meshed for optimal routing within the POP.

3. CASE STUDY: ROUTE REFLECTION DEPLOYMENT IN A LARGE ISP

In this section, we take a closer look at route reflection by examining its deployment in a large ISP (which we will call ISP_x in the rest of this section). Our discussion focuses on two issues, (1) POP based RR placements and resulting new scalability challenges, and (2) hierarchical route reflection structure as a solution to address the RR scaling issues, and the consequent impact on the overall routing system performance.

3.1 Circumventing the Drawbacks through RR Placement

The definition of route reflection allows a client router to peer with any route reflector in the same network. However, as we discussed in Section 2.4, improperly configured client-reflector relations can lead to potentially negative impact on routing system performance. Following the guidelines in [1], ISP_x configured RR in each of its major POPs, so that client routers peer with the RR residing in the same POP, making the logical iBGP topology following the underlying geographic locations. To avoid single point of failure, ISP_x configured two RRs at each of its major POPs.

Given that a route reflector is located in the same POP with its clients, its best path selections should be the same as that made by its clients, at least at the granularity of the POP level. Thus some of the negative impact from deploying route reflection mentioned in Section 2.4, such as reduced path diversity and sub-optimal routing, should no longer exist at the POP level. For example, the sub-optimal route problem illustrated in Figure 5 can be avoided by placing an RR in each POP. As shown in Figure 6, if RR1 is placed in the same POP with R_1, and RR2 in the same POP with R_2 and R_3, then both R_2 and R_3 can use the path R_2-R_5 to reach prefix d.

However, placing route reflectors at every POP also introduced its own limitations. Large ISPs tend to have routers at a large number of POPs, which may even be located in different continents. Route reflection requires that all RRs be connected in a full mesh,

putting a pair of RRs in every POP brings back the old issue of managing full mesh iBGP sessions among a large number of RRs in a global scale. In operational practice, the above issue is circumvented by creating a *hierarchy* of RRs.

3.2 Hierarchical Route Reflection

The basic idea behind a hierarchical route reflection structure is simple: Since route reflection is an effective means to move iBGP sessions away from full mesh, one can simply apply the same idea again at the RR level, *i.e.* for a set of N POP level RRs that require $N*(N-1)/2$ full mesh iBGP connections, one can simply set up a route reflector R to connect up the N RRs as its clients. However as we already learned, for the overall routing system performance, this route reflector R should be placed as geographically close to all its clients as possible. For a global scale ISP such as ISP_x, no single location can satisfy this requirement, hence multiple levels of RRs are needed. To assure the propagation of global BGP routing reachability to all iBGP routers, one only needs to create full mesh iBGP connections among all the RRs at the top level.

ISP_x has several hundreds of iBGP speakers distributed across multiple continents. It also has a heterogeneous router set. To effectively manage BGP routing information propagation in this large network and to control the routing scalability at individual routers. ISP_x deployed route reflectors at each of its major POPs as described in [1]; for small POPs which only have a small number of routers, they use the same RRs located at the nearest major POP. Because ISP_x has a large number of POPs to afford a full mesh connection of all POP level RRs, ISP_x grouped POPs into a few tens of regions, and set up a pair of RRs in each region that connect to the POP level RRs as their clients. Furthermore, since the geographical distance between continents is much further than those between regions, the ISP has a top layer of RRs at the continent level where the region level reflectors connect to as clients.

Figure 7 depicts an example hierarchical route reflection system to reflect the basic picture of the route reflection deployment in ISP_x. All RRs are deployed in pairs for necessary redundancy against single point of failure. To simplify the drawing, we omitted this detail. The diamond-shape RRs at the top level represent Continent level RRs; the square-shape RRs are at the 2^{nd} level of hierarchy, each represents a regional RR, and the $3rd$ level round-shape RRs represent POPs. Consider a client router, call it R_c, in POP1 (not shown in Figure 7): under this hierarchical route reflection, R_c only needs to have iBGP peers with 2 RRs and all other routers in POP1. This reduced number of peerings represents both a RIB-in size that is more than an order of magnitude smaller compared to the full mesh iBGP connection, and a BGP update sequence that only comes from its iBGP neighbors instead from all the ISP_x's BGP routers globally as would be the case with full mesh. Also, only those update messages that change the current best path chosen by RRs gets propagated through the RR hierarchy to reach all routers, and those that do not affect the current best path are filtered out by the RRs. However, such gains in RIB-in size and update message reduction do come with a cost, as we explain next.

3.3 Implications of Hierarchical Route Reflection

Under full-mesh iBGP, any iBGP speaker can reach any other iBGP speaker within one iBGP hop. Under a hierarchical route reflection the distance for an update to travel from one iBGP speaker to another is at least two hops (client-reflector-client), and in many cases longer. For example under the hierarchical route reflection shown in Figure 7, the distance between a router R_c1 (*e.g.* a client of the route reflector in POP1) in Continent1 and another client router R_c2 in POP11 in Continent2 is 7 iBGP hops. Due to various delays in propagating an update through each iBGP hop, this increase hop count can represent a significantly prolonged BGP update delay.

In addition to the difference in the number of iBPG hops, this hierarchical route reflection also presents a rather different picture in terms of the number of alternative paths that updates may travel through *on the control plane*. In full mesh iBGP connection each update has a *single* path to go from any router to any other router. Although Figure 7 seems also suggesting a single, albeit longer update propagation path between R_c1 in POP1 and R_c2 in POP11 due to the tree-like hierarchy of RRs, this is not the case because RRs at each level are replicated. When R_c1 sends an update that affects the selection of path to destination d, 2 RRs in POP1 will each send the same update to the 2 regional RRs they are connected to; each regional RR will in turn send each received updates to the 2 continental level RRs it connects to. Thus, one can see that in this 7-hop case there can be a large number of alternative paths that an update may go through from R_c1 to R_c2, which also contributes to prolonged routing convergence.

Multiple-level hierarchical route reflection topology can also further worsen path diversity, because the total number of routes to a destination d is limited by the total number of the RRs at the highest level that d's reachability is propagated. As one approaches the top of the hierarchy, the number of RRs reduces. For example, assume that a prefix d originated at Continent1 can be reached through N egress points of this ISP in Continent1. The very top level route reflector in Continent1 will propagate only one (*i.e.* its best) route to the top level route reflector in Continent2. As a result, all the downstream iBGP route reflectors and clients in Continent2 will only learn one route (*i.e.* the best route from the top level route reflector in Continent1) to reach d, although there are in fact N routes to reach d in Continent1.

We would also like to make the following observation: the topology shown in Figure 7 remotely resembles that of AS-level Internet topology. If one replaces *reflector-client* link as *provider-customer* link and *peer* (conventional iBGP session) link as *peer-peer* link, AS-level Internet topology model can also be used to describe this Tier-1 ISP's iBGP topology, even though the connectivity first principles are different. As future work, it would be interesting to compare and contrast these two models in detail.

Another interesting observation is that the top two layers of route reflectors of ISP_x are configured to be responsible for distributing the routing information within its network *only*, that is they are not involved in data packet forwarding. The data forwarding is done by the client routers and RRs in the third layer, and by other non-iBGP speakers. Therefore, one may consider that ISP_x has a separate control plane solely for routing propagation, separated from data forwarding plane. In the following subsection, we provide a more detailed discussion about separating control plane from data plane.

4. SUMMARY AND FUTURE WORK

Two alternatives to full-mesh iBGP were proposed about a decade ago to address the iBGP scalability problem posed by the original full-mesh iBGP design. In this paper, we described the route reflection solution along with its advantages and disadvantages that have been identified over time. We looked into the route reflection deployment in a large ISP which provided a concrete example of what can be achieved through route reflection and what are the remaining issues.

In the past, the number of BGP sessions that a router can handle was relatively small. Due to software and hardware technology advances, today's routers on the market are capable of handling

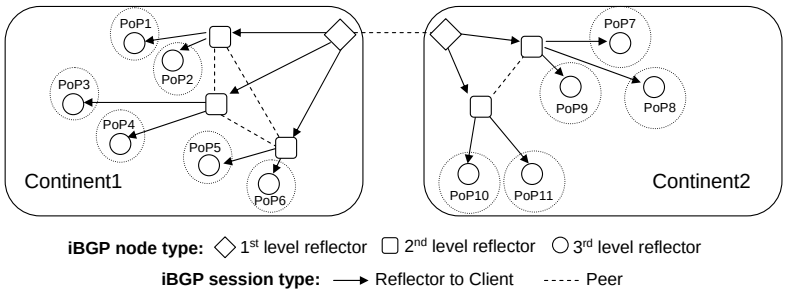


Figure 7: Simplified topology of the Tier-1 ISP using iBGP with hierarchical route reflection

thousands of iBGP sessions [10]. Although this may remove one of the reasons for deploying route reflection, the operational cost from configuring and maintaining large numbers of iBGP sessions remains a strong motivation for scaling the iBGP sessions in a large network. That is probably a main reason that route reflection has seen a wide adoption among large ISPs. However a number of open issues remain, and several potentials also exist, to make route reflection an effective solution towards future routing scalability. We identified the follow items as our future work.

4.1 Separating Control Plane from Data Plane

As the Internet continues to grow in size, ISP_x also grows rapidly over time and its overall topology become more complex to manage. A recent trend in scaling and simplifying network management is to decouple a network's control plane from its data plane. In [5], Feamster *et al.* argue for a (logically) centralized routing server (*i.e.* Routing Control Platform, RCP) to perform the routing decisions for all the routers in a network, effectively making the routers perform data forwarding functions only. However, there are major road blocks in implementing and deploying such a centralized control system. In [5], authors also recognize robustness, scalability, and routing correctness as major challenges in rolling out such a design.

We observe from the operational practice that route reflection *can* be used as a simple, incrementally deployable means to steer a network towards separating its control plane from the data plane, as ISP_x has already started evolving its network towards that direction. For example, the top two layers of RRs shown in Figure 7 can be considered as its core routing infrastructure that are responsible for handling routing propagations and decisions only, and that are entirely separated from the data forwarding plane. Data packet forwarding is handled by the routers near the peripheral of this control hierarchy only, by routers at third RR layer or below.

We also make three further observations. First, the operational community is utilizing the RR redundancy to develop simple yet effective solutions to improve path diversity, as reported in [10]. Second, the recent effort in IETF SIDR Working Group to secure the global system requires new functionality and processing power at routers to verify all routing updates, a separate control plane can ease such new functional deployment. Finally, a recently proposed routing scalability solution, Virtual Aggregation [6], can also find an incrementally deployable path through route reflection. All signs indicate that we should pursue the use of route reflection as an effective and incrementally deployable vehicle towards scaling the global routing system through the separation of control and data planes.

4.2 Remaining Issues with Route Reflection

Of all the route reflection induced side effects identified in Section 2.4, we sort them into two categories. The first one concerns routing convergence. Route reflection deployment in a global-scale ISP desires a hierarchical structure, which can prolong routing propagation and worsen routing convergence. Efforts along the following directions are underway to address this issue: using redundant standby paths to assure data plane performance during routing convergence; minimizing MRAI timer to speed up routing propagation; and designing effective route flap damping to prevent update flooding with minimized MRAI time value.

The second category concerns how best to build and utilize redundant RRs that can address robustness, path diversity, and sub-optimal paths all at once. By definition, an RR plays a more important role than a client router, thus it requires redundancy against a single point of failure. Redundant RRs, in turn, can also be used to increase path diversity and reduce sub-optimal routing as suggested in [10].

5. REFERENCES

[1] T. Bates, E. Chen, and R. Chandra. RFC 4456: BGP Route Reflection: An Alternative to Full Mesh Internal BGP, April 2006.
[2] M. O. Buob, S. Uhlig, and M. Meulle. Designing Optimal iBGP Route Reflection Topologies. *Networking*, April 2008.
[3] R. Dube. A Comparison of Scaling Techniques for BGP, October 1999.
[4] R. Dube and J. G. Scudder. Route Reflection Considered Harmful, May 1999.
[5] N. Feamster, H. Balakrishnan, J. Rexford, A. Shaikh, and K. van der Merwe. The Case for Separating Routing from Routers. In *ACM SIGCOMM Workshop on Future Directions in Network Architecture (FDNA)*, Portland, OR, September 2004.
[6] P. Francis, H. Ballani, and T. Cao. A White Paper on Reducing FIB Size through Virtual Aggregation, June 2008.
[7] T. G. Griffin and G. Wilfong. On the Correctness of iBGP Configuration. *SIGCOMM Comput. Commun. Rev.*, 32(4):17–29, 2002.
[8] S. Halabi and D. McPherson. Internet Routing Architectures, 2nd ed., August 2000.
[9] C. Pelsser, T. Takeda, E. Oki, and K. Shiomoto. Improving Route Diversity through the Design of iBGP Topologies. *IEEE International Conference on Communications*, May 2008.
[10] R. Raszuk, K. Patel, I. Kouvelas, R. Fernando, and D. McPherson. Distribution of Diverse BGP Paths, July 2009.
[11] Y. Rekhter, T. Li, and S. Hares. RFC 4271: A Border Gateway Protocol 4 (BGP4), January 2006.
[12] J. G. Scudder and R. Dube. BGP Scaling Techniques Revisited, October 1999.
[13] P. Traina, D. McPherson, and J. Scudder. RFC 5065: Autonomous System Confederations for BGP, August 2007.
[14] S. Uhlig and S. Tandel. Quantifying the BGP Routes Diversity Inside a Tier-1 Network. *Networking*, 3976, April 2006.
[15] M. Vutukuru, P. Valiant, S. Kopparty, and H. Balakrishnan. How to Construct a Correct and Scalable iBGP Configuration. In *ACM/IEEE Infocom*, April 2006.
[16] F. Wang, Z. M. Mao, L. G. Jia Wang, , and R. Bush. A measurement study on the impact of routing events on end-to-end internet path performance. In *Proceedings of ACM SIGCOMM 2006*, 2006.
[17] L. Xiao, J. Wang, and K. Nahrstedt. Optimizing iBGP Route Reflection Network. *IEEE International Conference on Communications*, May 2003.