University of California
Los Angeles

Understanding the Impact of Internal BGP Route Reflection

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Computer Science

by

Jong Han Park

2011

The dissertation of Jong Han Park is approved.

Leonard Kleinrock

Mario Gerla

Ying Nian Wu

Lixia Zhang, Committee Chair

University of California, Los Angeles

2011

To my Parents, my Wife Younghee, and my Son Joonyoung

Table of Contents

List of Figures

List of Tables

Acknowledgments

First and foremost, it is my great pleasure to express my deepest gratitute to my dissertation advisor, Dr. Lixia Zhang, for her patience, support, and guidence throughout the research. I have to admit that often times were tough, but nonetheless I truly feel that she is the best advisor and that I was very fortunate to learn research under her supervision. She taught me research and how to become a better researcher, and I owe a profound debt of gratitude towards her. Second, I would like to thank my committee members: Dr. Leonard Kleinrock, Dr. Mario Gerla, and Dr. Ying Nian Wu for their valuable time, comments, and encouragements on the research.

I cannot thank enough to my two mentors, Dr. Mohit Lad and Dr. Ricardo Oliveira, for their guidence and valuable discussions in many different aspects including research and life. Their technical insights on different projects were always helpful, and I feel that I was very fortunate to spend the first 2 years of my Ph.D. study with them. Also, I would like to acknowledge my colleagues in Internet Research Laboratory at UCLA, Dr. Micheal Meisel, Pei-chun Cheng, and Dan Jen for their valuable discussions and comments on different research topics throughout my study. It is my pleasure to acknowledge and thank Jaeyoung Choi who visited our lab and worked together with me on BGP path diversity. I thank him for his hard work, discussions, and comments.

When our baby boy, Joonyoung, was born, my parents and parents-in-law visited for a few months and took turns taking care of him. Without their sacrifices, I could not have finished the degree in time.

There is one very special person in my life, my wife Dr. Younghee Kim, who always supports and believes in me. I thank her for her endless love, support,

and encouragement. This journey has been a very pleasant one because of her. Lastly, I would like to thank my 3-year-old boy Joonyoung for always being there with love and laughter.

Vita

1975	Born, Seoul, S. Korea
1999–2000	Assistant System Administrator, School of Medicine, University of California San Diego.
2000	B.S. (Computer Science), University of California, San Diego.
2000–2001	Software Engineer, 3GPP2 (CDMA) 1xRTT Software Development and Testing, Neopoint Inc., San Diego.
2001–2003	Software Engineer, Web Client Software Design and Development, Lims Technology, Daejun, S. Korea.
2003	Software Engineer, 3GPP (WCDMA) IP-based Node-B and RNC System Design, Corecess Inc., Seoul, S. Korea.
2003–2005	Senior Software Engineer, Massive Data Search Server Design and Development, Empas Inc., Seoul, S. Korea.
2005–2006	M.S. (Networked Systems), University of California, Irvine.
2006–2007	Research Staff, University of California, Irvine.
2008	Graduate Student Intern, Energy Efficient Networking, Intel Research, Oregon, U.S.A.
2007–present	Research Assistant, Computer Science Department, UCLA.
2007–present	Graduate Student Researcher, Computer Science Department, UCLA.

2008–present Teaching Fellow, Computer Science Department, UCLA.

Publications

Free-riding in BitTorrent Networks with the Large View Exploit Michael Sirivianos, Jong Han Park, Rex Chen, Xiaowei Yang, International Workshop on Peer to Peer Systems (IPTPS) 2007, Washington, U.S.A.

Dandelion: Cooperative Content Distribution with Robust Incentives Michael Sirivianos, Jong Han Park, Xiaowei Yang, Stanislaw Jarecki, USENIX Annual Technical Conference (ATC) 2007, Santa Clara, U.S.A.

A Study of Internet Routing Stability Using Link Weight Mohit Lad, Jong Han Park, Tiziana Refice, Lixia Zhang, UCLA CS Technical Report #080003

Investigating Occurrence of Duplicate Updates in BGP Announcements Jong Han Park, Dan Jen, Mohit Lad, Shane Amante, Danny McPherson, Lixia Zhang, Passive and Active Measurement Conference (PAM) 2010, Zurich, Switzerland. (best paper nominee)

Flap Damping with Assured Reachability Pei-chun Cheng, Jong Han Park, Keyur Patel, Lixia Zhang, Asian Internet Engineering Conference (AINTEC), November 2010, Bangkok, Thailand.

Understanding BGP Next-hop Diversity Jaeyoung Choi, Jong Han Park, Pei-chun Cheng, Dorian Kim, Lixia Zhang, 14th IEEE Global Internet Symposium (INFOCOM workshop), April 2011, Shanghai, China.

A Comparative Study of Architectural Impact on BGP Next-hop Diversity Jong Han Park, Pei-chun Cheng, Shane Amante, Dorian Kim, Danny McPherson, Lixia Zhang, UCLA CS Technical Report #100031, October 2010.

Quantifying i-BGP Convergence in Large ISPs Jong Han Park, Pei-chun Cheng, Shane Amante, Dorian Kim, Danny McPherson, Lixia Zhang, UCLA CS Technical Report #110009, May 2011.

Abstract of the Dissertation

Understanding the Impact of Internal BGP Route Reflection

by

Jong Han Park

Doctor of Philosophy in Computer Science

University of California, Los Angeles, 2011

Professor Lixia Zhang, Chair

The Internet has been growing rapidly in its size and become denser over time, and so have Internet Service Providers (ISPs). Today, the number of BGP routers inside a large ISP can be more than one thousand and spread across different geographical locations globally. To scale with the increasing number of routers, large ISPs have developed and used more scalable i-BGP architectures such as BGP route reflection without thorough analysis of their design. This lack of such analysis escalated interests and concerns on BGP performance inside a large ISP; BGP performance inside large ISPs is no longer simple to understand and can potentially have a noticeable impact on the end-to-end data plane performance.

As a first step to address these concerns, we perform measurement studies that define, quantify, and analyze two important BGP performance metrics inside large ISPs: path diversity and convergence delay. Our measurement analysis, based on BGP data collected from backbone production routers in two global-scale ISPs over a few years, shed lights on many interesting properties of existing BGP path diversity and convergence inside the ISPs, and may be useful in com-

prehensive and complete understanding of the end-to-end protocol performance and enhancements, more realistic simulations as well as designing the future global routing protocols.

CHAPTER 1

Introduction

The Internet consists of tens of thousands of independently administrated network domains called autonomous systems (ASes). Border Gateway Protocol (BGP) is used for both within a single AS and between different ASes to communicate reachability information. The original internal BGP (i-BGP) design used to distribute reachability information within an AS requires that all i-BGP routers are fully meshed and that a reachability learned from an i-BGP router is not forwarded to any other i-BGP router inside the full-mesh. This simple design to avoid routing loops does not scale as the Internet and Internet Service Providers (ISPs) grow rapidly in the size and become denser over time with the increasing number of i-BGP routers. Currently, the number of i-BGP routers in large ISPs is hundreds or even more than one thousand and spread across different geographical locations globally.

To solve this scalability problem in i-BGP, two solutions (AS confederations and route reflection) were proposed in 1996. Since then, both solutions have been widely adopted by large ISPs, yet, without a thorough analysis on the impact of the solutions. For example, both solutions create hierarhchies within the i-BGP topology and are known to have side effects. Some positive side effects include reduced networking provisioning cost, reduced memory usage for storing routing tables, and reduced number of update messages generated inside an ISP. However, these benefits come at a cost; there are also negative side effects on both routing

correctness and routing performance. The lack of thorough analysis escalated interests and concerns from the Internet community as the i-BGP topologies becomes icreasingly complex over time.

In this dissertation, we perform measurement studies focusing on route reflection (the domanant solution) that define, quantify, and analyze two important BGP performance metrics, namely BGP path diversity and convergence delay, inside large ISPs to understand the current state of the art as well as the impact of adopting route reflection on the two metrics. Our measurement analysis based on BGP data collected from backbone production routers in two global-scale ISPs over a few years, shed lights on many interesting properties of existing BGP path diversity and convergence inside the ISPs, and may be useful in comprehensive and complete understanding of the end-to-end protocol performance and enhancements, more realistic simulations as well as designing the future global routing protocols.

This dissertation is organized as follows. In Chapter 2, we provide an overview of the original full-mesh i-BGP design, the two solutions along with their known side effects that are particularly relevant for this dissertation. In the next two chapters (Chapter 3 and Chapter 4), we focus on the impact of route reflection, the dominant solution between the two solutions, and present our evaluation and analysis results on its impact on two important metrics of BGP performance: BGP path diversity and convergence delay inside an ISP respectively.

CHAPTER 2

Background

2.1 Routing in the Internet and Internal BGP

The Internet is made of tens of thousands of different networks called Autonomous Systems (ASes). Each AS represents a single administrative entity with its own unique AS number and IP address blocks called prefixes. Routers of different ASes set up BGP sessions in between to exchange BGP routing updates (inter-domain routing). Such BGP sessions are called e-BGP sessions. BGP sessions are also set up between routers within the same AS (intra-domain routing) to exchange BGP routing updates, and these sessions are called i-BGP sessions.

All routing protocols must have effective means to prevent routing loops. In e-BGP, routers detect any potential loops at inter-AS level by inspecting the AS_PATH attribute carried in all BGP messages. A router will drop a BGP message if the AS_PATH in the message already contains its own AS number. To avoid routing loops in i-BGP, the original design requires that all BGP routers in the same AS be directly connected to each other via pairwise i-BGP sessions. This full-mesh i-BGP connectivity allows each BGP router learns about reachability information directly from all other BGP routers in the same AS, eliminating the need to forward BGP updates learned from an i-BGP speaker to another i-BGP speaker, hence eliminating potential routing loops. However this full-mesh requirement leads to a total of $\frac{N*(N-1)}{2}$ i-BGP sessions in an AS, where N is the

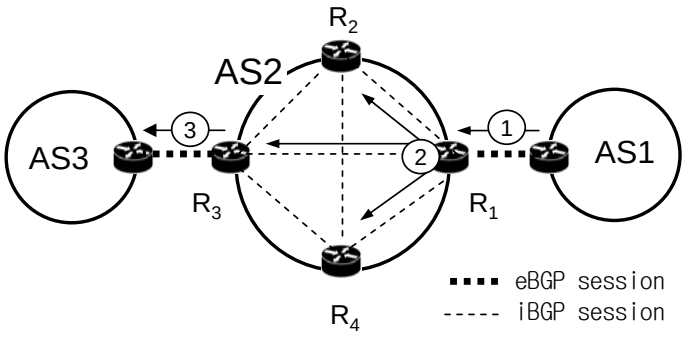


Figure 2.1: Inter-working of i-BGP and e-BGP in the Internet

number of BGP routers in the AS. Because i-BGP session management, such as such adding or removing an i-BGP, requires operator interventions, this full-mesh i-BGP connectivity also represents a high operational cost.

Figure 2.1 shows an example of the inter-working of e-BGP and i-BGP to propagate the reachability information to a destination network which is represented by an IP prefix. In Figure 2.1, AS2 maintains e-BGP sessions with AS1 and AS3 via its routers R_1 and R_3, respectively. Inside AS2, all BGP routers are inter-connected through i-BGP sessions. When AS1 announces a destination prefix d to AS2 over the e-BGP session with R_1, R_1 will propagate the information to all the other three routers in AS2 over its direct i-BGP sessions with them. If the routing policy permits, R_3 will further propagate d's reachability to its e-BGP neighbor, in this case the router in AS3. Note that within AS2, d's reachability message traverses only one i-BGP hop from R_1 to all the other routers. This process of propagating reachability information repeats until all ASes in the Internet learn how to reach prefix d.

The total number of i-BGP sessions in AS2 is $\frac{4*(4-1)}{2} = 6$ as shown in Figure 2.1. Following the same formula, the total number of i-BGP sessions for an AS

with 10, 100, or 1,000 routers would be 45, 4,950, or 499,500, respectively. Today, the number of BGP routers in a typical large AS can be several hundreds or even over a thousand, making the full-mesh i-BGP interconnections infeasible.

To alleviate this i-BGP scalability problem, the vendor and operator communities quickly proposed two solutions in 1996: route reflection and AS confederations. Both solutions have been deployed in operational networks, in certain cases AS confederation deployment is combined with route reflection. Overall, route reflection has a wider deployed base and is the focus of this chapter.

The objective of our comprehensive overview of route reflection are three-fold. First, we provide an overview of route reflection's operations and explain its pros and cons in detail (Section 2.1.3). Second, using the route reflection deployment in a large ISP as a case study, we 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 any protocol or implementation changes (Section 2.2). Finally in Section 2.3 we identify remaining issues in achieving the goals of both efficient routing information dissemination and system scalability.

2.1.1 Full-mesh I-BGP

BGP [RLH06] uses fully meshed internal BGP (i-BGP) sessions among all BGP routers in an autonomous system to disseminate BGP updates within one autonomous system. As a simple way of avoiding routing loops, the original i-BGP requires that all i-BGP routers within the same AS be connected in a full-mesh, and that reachability information learned from one i-BGP router must not be forwarded to any other i-BGP router. In this setting, the maximum number of i-BGP hops that an update can traverse is always 1. However, this full-mesh

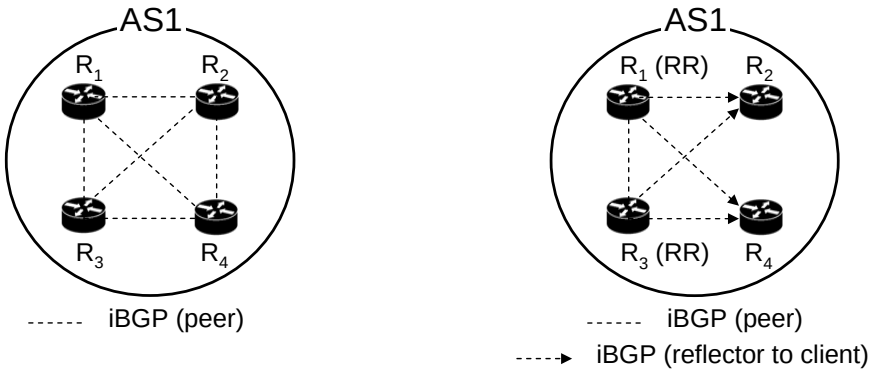
connection requirement results in the total number of i-BGP sessions growing as the square of the number of i-BGP routers inside the network. To mitigate this scalability problem, two alternative architectures have been proposed and widely used by large ISPs: route reflection [BCC06] and AS confederations [TMS07].

In a network with large numbers of BGP routers, this full-mesh requirement results in a large number of BGP sessions at each router. Furthermore, since BGP sessions are managed through manual configurations, this full-mesh requirement also leads to configuration changes at all routers whenever a router is added or removed.

2.1.2 AS Confederations

AS confederations [TMS07] take a divide-and-conquer approach to mitigate the i-BGP session scalability issue by grouping i-BGP routers together into sub-ASes, creating multiple sub-ASes within an AS. The smaller number of i-BGP routers in each sub-AS leads to a smaller number of i-BGP sessions within the sub-AS, making full-mesh connections feasible. The sub-ASes within the AS communicate with each other as they would in e-BGP. AS confederations prevent routing loops by introducing two new attributes: AS_CONFED_SET and AS_CONFED_SEQ, which are the counterparts of ORIGINATOR_ID and CLUSTER_LIST in route reflection.

Although the communication models of route reflection and AS confederation may look different, they both create hierarchies within i-BGP to solve the same scalability problem. The potential problems and additional delays explained earlier in route reflection also apply to AS confederations [Dub99, SD99].



(a) Full-mesh i-BGP

(b) i-BGP with route reflection

Figure 2.2: Different i-BGP topologies

2.1.3 Route Reflection

Route reflection [BCC06] was first developed in 1996 as one of the two proposed solutions to address the above mentioned BGP scalability problem; the other one is AS confederations [TMS07]. Between the two, route reflection has seen a larger deployed base. Route reflection deployment was rolled out more than 10 years ago, however it is generally known that the design did not go through thorough analysis studies. Only recently several studies appeared which analyze various impacts of route reflection on the overall routing system performance. The results from these studies show that route reflection can potentially decrease the network's robustness against failures [XWN03a, XWN03b], introduce delayed routing convergence [CCF05], reduce path diversity [UT06], adopt sub-optimal routes [VVK06], and even cause data forwarding loops [Dub99, DS99, GW02, SD99].

The simplest model of route reflection deployment is to select one BGP router in an AS to be the route reflector (RR), and have all the other routers in the AS set up i-BGP sessions with this RR. The RR receives BGP update messages from each i-BGP speaker and forwards (or reflects) them to all other i-BGP speakers.

Because the RR forwards updates among i-BGP speakers, it removes the need for i-BGP speakers to connect in a full-mesh. To avoid a single point of failure, ASes generally set up multiple RRs, which are interconnected in a full-mesh among themselves.

Figure 2.2 illustrates the difference between interconnecting i-BGP routers via full-mesh and via RRs. Figure 2.2(a) shows an example of full-mesh i-BGP interconnections, where all i-BGP speakers are directly connected to each other. Figure 2.2(b) shows an example of route reflection deployment, where R_1 and R_3 serve as RRs and connect to i-BGP speakers R_2 and R_4, which are connected to both reflectors for redundancy. Since R_2 can learn R_4's BGP reachability information from the RRs and vice versa, R_2 and R_4 do not need to interconnect. R_2 and R_4 are client routers of R_1 and R_3. A client is an i-BGP speaker that connects directly to an RR to learn the reachability information collected by other routers in the AS. In the view of R_2 and R_4, R_1 and R_3 are non-clients. Note that R_2 and R_4 require no special configurations; they are not aware of R_1 and R_3 being RRs. Only R_1 and R_3 require configuration changes. The relation between R_1 and R_3 is non-clients, and they can pass the reachability information learned from one i-BGP speaker to others in the same AS.

However an RR does not necessarily forward all the received reachability information to all i-BGP neighbors; the following rules apply depending on the type of i-BGP session from which the route is received:

1. the routes received from non-client i-BGP sessions are reflected only to clients;

2. the routes received from client i-BGP sessions are reflected to both clients and non-clients; and

3. the routes received from e-BGP sessions are reflected to both clients and non-clients.

For example in Figure 2.2(b), when R_2 receives a reachability information from its e-BGP neighbor (not shown in the figure), R_2 would forward the reachability information to all its i-BGP neighbors, namely R_1 and R_3 (Rule 3). Upon receiving this reachability information from their client, i.e., R_2, both R_1 and R_3 would further reflect this information to their clients, i.e., R_2 and R_4, and non-clients ,i.e., R_1 and R_3 to each other (Rule 2).

Because RRs forward reachability information learned from an i-BGP speaker to another i-BGP speaker, routing messages travel more than a single i-BGP hop, and it becomes possible to create loops. For example in Figure 2.2(b), an update message originated at R_2 can come back to R_2 through more than one RR (R_1 and R_3 in this case), forming a loop. To prevent such loops, two new attributes are added to BGP update messages: CLUSTER_LIST and ORIGINATOR_ID. An RR uses its router ID as the cluster ID. When forwarding a BGP update, if an RR finds its own cluster ID in the CLUSTER_LIST attribute of a received update, it discards the update; otherwise it prepends its cluster ID in the CLUSTER_LIST attribute before forwarding the update. In addition, the first router that injects a routing update 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 2.2(b), R_2 will discard all updates reflected back to itself after checking that ORIGINATOR_ID attribute contains its router ID.

AS confederations [TMS07], on the other hand, takes a divide-and-conquer approach to mitigate the i-BGP session scalability issue by grouping i-BGP routers together into sub-ASes, creating multiple sub-ASes within an AS. The smaller

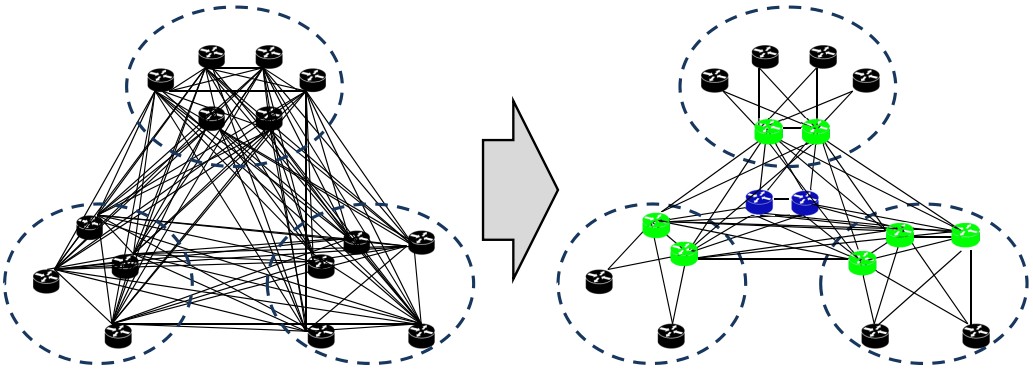


Figure 2.3: Before and after applying route reflection

number of i-BGP routers in each sub-AS leads to a smaller number of i-BGP sessions within the sub-AS, making full-mesh connections feasible. The sub-ASes within the AS communicate with each other as they would in e-BGP. Although the communication models of route reflection and AS confederation may look different, they inherit common properties of BGP routing, and potential problems in route reflection can also be replicated under AS confederations [SD99], possibly with a different degree of impact.

2.1.3.1 Benefits of Route Reflection

Reduced Number of i-BGP Sessions: Route reflection can effectively reduce the number of i-BGP sessions in an AS. A non-RR router only needs to establish a small number (typically two for redundancy) of i-BGP sessions with the RRs. Although an RR router generally has a larger number of BGP sessions, one can control this number through well-established engineering practices. Assuming a route reflection based AS with N i-BGP routers and K RRs, the total number of i-BGP sessions for a given RR can be computed as $\frac{K*(K-1)}{2} + C$, where K is the number of RRs in the network and C the number of client i-BGP routers con-

nected to the given RR. Typically K is a much smaller number than N, making the total number of i-BGP sessions for a given RR much smaller compared to that of full-mesh. For a given client, the number of i-BGP sessions is typically a constant (e.g., 2 for redundancy) regardless of the network size. Another advantage of route reflection is that it can be applied recursively to further reduce the total number of sessions. Figure 2.3 shows an example of a topology before and after route reflection is adopted. In this particular example, route reflection is applied twice recursively, creating a 3-level hierarchical route reflection topology.

Reduced Operational Cost: Creating, modifying, or removing BGP sessions require operator intervention. In the case of full-mesh i-BGP, any new router added to a network requires modifications to all the other routers' configurations. In the case of route reflection, adding or removing a client i-BGP router only requires configuration changes to the RRs the client connects to, with no impact on the other routers.

Reduced RIB-in Size: A BGP router R maintains three different types of routing tables: RIB-in, Loc-RIB, and RIB-out. An RIB-in contains unprocessed routing information that has been advertised to R by each of R's BGP neighbors. After examining the reachability information across all RIB-ins, the router decides a single best path for each destination D and stores this best path in Loc-RIB. R may or may not forward D's reachability information to its BGP neighbor routers depending on the routing policy, but because the routing policy to all the i-BGP neighbors is the same, R only needs one RIB-out to store reachability information to be propagated to all its neighbors. However the number of RIB-ins increases proportionally to R's number of BGP neighbors. If R has n neighbors each sending p prefixes, its total RIB-in size is in the order of $n \times p$. With full-mesh i-BGP sessions, n is the number of i-BGP neighbors in the full-mesh.

With route reflection, n for client i-BGP routers is the number of RRs the clients connect to and is typically a small number.

Reduced Number of BGP Updates: With a significant reduction in the number of its i-BGP neighbors, a client router naturally receives a significantly reduced number of updates. A route reflector R_r receives routing updates from all its neighbors, but since BGP only propagates the best path to each destination, R_r further propagates only those updates that change its best path selections. In sharp contrast to a full-mesh i-BGP setting where all BGP updates are propagated to all routers, RRs effectively shelter their client routers from a large percentage of incoming updates.

Assuming an AS has N BGP routers, if it uses full-mesh i-BGP connections, every i-BGP speaker processes roughly the same amount of updates coming from the $(N-1)$ sessions, putting a high processing demand on all the routers. If an AS adopts a simple route reflection topology with M RRs, only the RRs have $(M-1)+C$ i-BGP sessions (for a full-mesh connection among RRs and connections to C client routers); the rest client routers only need to connect to a few RRs. This differentiated processing load and memory requirements support a heterogeneous router environment where high-end routers with more capacity are used as reflectors and less capable routers can be used as clients, effectively extending their life time.

Incremental Deployability: Last but not the least, route reflection allows coexistence of RRs with conventional BGP routers that do not understand route reflection. A conventional BGP router B can be connected to RRs as a client, or a non-client (in which case B must also be connected to all other RRs). This allows a network to perform a gradual migration from the full-mesh i-BGP model to the route reflection model.

2.1.3.2 Caveats of Route Reflection

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

Robustness: With full-mesh i-BGP, a single router failure has limited impact on the rest of the network. That is, only the failed router is disconnected from the network and the rest routers in the network are not affected. In the case of route reflection, if a route reflector R_r fails, not only R_r itself is disconnected from the rest of the network, the client routers that used R_r to communicate with other routers also become disconnected and stop receiving routing updates. Furthermore, other routers can no longer get updates for the destinations connected to these client routers. To avoid such single point of failures, RRs are normally deployed in pairs, and each client router is usually connected to two or more RRs.

Prolonged Routing Convergence: An AS with route reflection can experience longer routing convergence compared to the full-mesh i-BGP interconnections [CCF05]. In the full-mesh i-BGP case, a BGP update travels only one i-BGP hop to reach all other i-BGP routers. However with route reflection, an update message may traverse more than one RR before reaching the final i-BGP router. Since each RR runs the best path selection process, there are both processing delay and transmission delay to cross a reflector. These additional delays in update propagation time can lead to a longer overall convergence delay.

In Figure 2.2(a), if R_2 were to distribute an update message learned from an external peer, it will send the update through the direct i-BGP sessions to R_1, R_3, and R_4. On the other hand, with route reflection in Figure 2.2(b), R_2 will

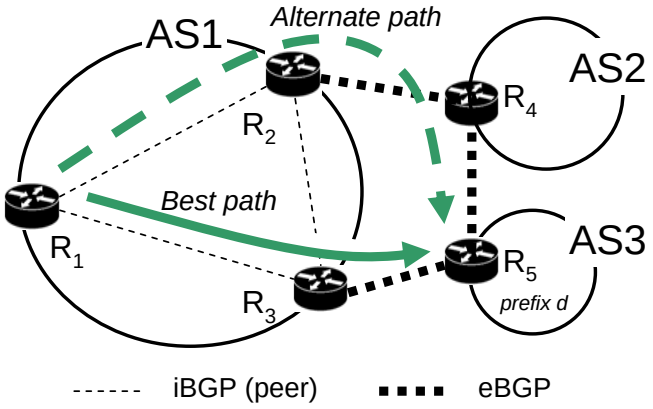


Figure 2.4: Packets can be dropped during path changes.

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 i-BGP hop through the RR adds to the delay before R_4 can receive the update. As we will show in the next section, an AS may deploy a hierarchy of RRs 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 paths to a given destination. For example, in Figure 2.2(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. We borrow Figure 2.4 from [WMJ06] by Wang et al. to illustrate this problem through an example. [WMJ06] shows that in a full-mesh i-BGP configuration, a path fail-over event can cause packet drops in the following way. Because a BGP router performs path poisoning[1] on known but less preferred routes, R_2 withdraws the path to destination d through R_2–R_4–R_5, and uses the path through R_3–R_5 to reach d since this path has the shortest AS_PATH length; at this time only R_2 knows about the alternate path through R_2–R_4–R_5 to reach prefix d. When the best route to prefix d through R_3 fails, R_1 can momentarily lose reachability to prefix d if the withdrawal 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 headed to d until the update from R_2 arrives.

Route reflection can worsen this data plane performance degradation due to prolonged routing convergence delay. Assume that AS1 in Figure 2.4 uses route reflection, and the number of i-BGP hops between R_1-R_2 is greater than one. When R_3-R_5 fails, R_3 will send the withdrawal message to R_1 and R_2, and R_2 in turn will send the update containing the reachability information of the alternative path to reach d as in the case of full-mesh i-BGP. However this update from R_2 to R_1 will have a longer delay simply because the update has to traverse more than one i-BGP hop. As a result, the time duration of d's reachability at R_1 will increase.

Data Forwarding Loop: In a simple route reflection configuration where a single RR connects to all client routers, there should be no data plane loops. However in real deployment, because all client routers must connect to multiple RRs to avoid single point of failure, this redundant connectivity to RRs can

[1]Path poisoning (or route poisoning) refers to a practice that a router explicitly withdraws an available path to prevent its neighbor routers from using the path, hence preventing potential routing loops.

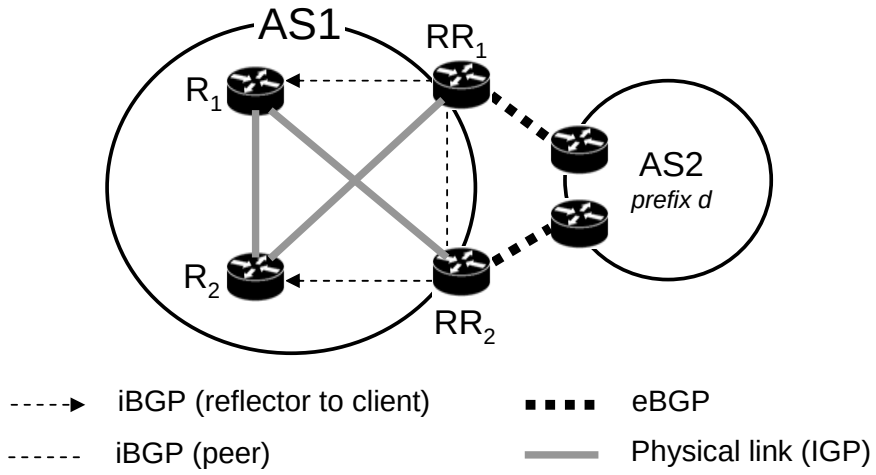


Figure 2.5: Route reflection with data forwarding loop

potentially introduce subtle data plane loops that defeat intuitive inspection, as we show by the following example borrowed from [GW02].

When a client router receives a data packet, it looks up the destination address and forwards the packet to the BGP next-hop router. Depending on the IGP connectivity, there can be multiple router hops between this client router and the BGP next-hop router, as is the case in Figure 2.5. In Figure 2.5, 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. As far as BGP 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 next-hop RR_1 via R_2, expecting R_2 to further forward this packet to RR_1. On the other hand, R_2 believes that the BGP next-hop 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: For a given BGP router, path diversity is a measure to quantify the number of different routes available to reach a given destination. A high path diversity for each destination prefix can increase the resiliency against failures and offer opportunities for traffic engineering [PTO08, UT06]. Since an RR only propagates its best route for a given destination, all the client routers of the same reflector use the same single best route to the destination as chosen by the RR. Figure 2.6 shows such an example: although both R_1 and R_2 are directly connected to AS2 to reach destination prefix d, if the reflector RR 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 until RR learns about the failure and switches to an alternative path to d, and then propagates the new path to all its clients. In contrast, full-mesh i-BGP interconnections not only would allow R_1 and R_2 to use their direct connection to AS2 to reach prefix d, but also allow R_3 to learn both paths and choose in between, and to be able to switch to the other path as soon as it learned about the failure from R_1 directly.

The above example shows that path diversity can potentially reduce routing convergence time because a router can switch to an alternative path immediately without waiting for BGP to converge in case of a failure. Intuitively one might believe that one could increase path diversity by increasing the number of RRs each client connects to. However this is not the case in the current practices. Although RRs are commonly deployed in pairs to avoid single point of failure, the pair of RRs are normally configured as pure replicas and always make the same routing decisions.

There have been several recent efforts to increase the path diversity in i-BGP to reduce the convergence time. [RFP11] by Raszuk et al. suggests to increase path diversity within an AS by modifying the best path selection in RRs, so that

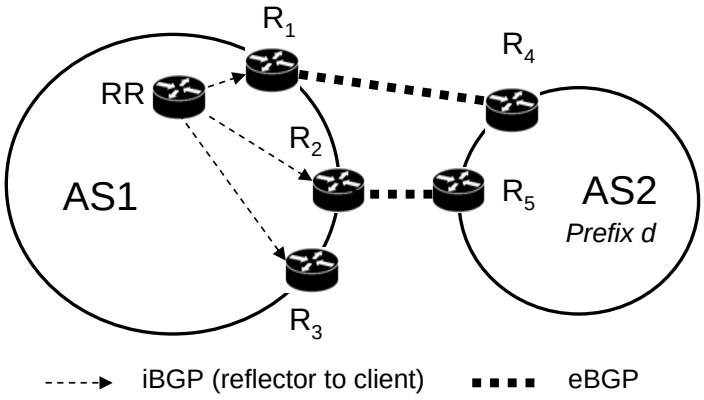


Figure 2.6: Route reflector chooses its best route

different RRs will advertise different paths to client routers. Another proposal is adding external-best option [MFC11] in BGP. By using best external option, a border BGP router can propagate more than one best external path to i-BGP neighbors inside an AS. This can increase the number of paths observed by i-BGP routers and decrease the number of hidden paths. Yet another proposal by Walton et al. [WRC10, SF09] suggests allowing any BGP router to propagate more than a single best path to increase the overall path diversity.

Note that BGP path diversity is an outcome of multiple impacting factors such as physical connectivity, BGP best path selection criteria (e.g., policies of different ISPs), and internal router topology and connectivity. Although conceptually the use of hierarchical route reflection system shall decrease the amount of alternative paths learned by a given router, the degree of reduction may differ depending on the above-mentioned network settings. More studies are necessary to understand the impact of hierarchical route reflection on path diversity.

Sub-Optimal Routes: An RR selects its best paths to reach the destination prefixes using its local routing information, and propagates these selected paths to its clients. It is most likely that not all the best paths chosen by the reflector would

be the best paths for each of all its clients. Therefore some client routers end up using sub-optimal paths to some destinations as reported in [BUM08, XWN03a]. For example in Figure 2.6, AS1 has two paths to reach prefix d in AS2, through R_1–R_4 and through R_2–R_5. Assuming that the link lengths in Figure 2.6 reflect the IGP distances of the routers, the route reflector RR would pass to R_1, R_2, and R_3 its own best path to prefix d in AS2, which is through R_1–R_4 (because RR itself is closer to R_1 than R_2). R_2 will still use its own best path through R_2–R_5 because of the BGP best path selection rule that prefers path learned from e-BGP over that learned from i-BGP. However, R_3 will use the path R_1–R_4, the only path learned from the RR. R_3's shortest path to prefix d should have been through R_2–R_5, had the AS1 used full-mesh i-BGP interconnections.

It is worth pointing out that, in a given network, the actual impact of the above drawbacks from route reflection heavily depends on the exact configuration and placement of RRs. Bates et al. suggest several approaches to minimize the negative impact of route reflection [BCC06], including placing an RR in the same POP (Point of Presence) with its clients, and making clients of the reflector in each POP fully meshed with each other for optimal routing within the POP.

In the next section, we perform a case study of route reflection deployment and see how well one can address some of the negative side effects by following the guidelines in [BCC06].

2.2 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_{RR} in the rest of this section). Our

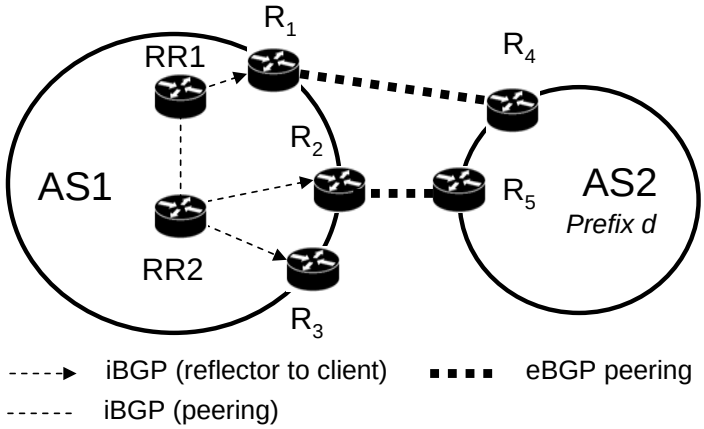


Figure 2.7: POP-based route reflection deployment

discussion focuses on two issues: (1) POP-based route reflector placement, as suggested by [BCC06], and consequent new issues in i-BGP scalability, and (2) hierarchical route reflection structure as a solution to address the scaling issues and the consequent impact on the overall routing system performance.

It is worth pointing out that the engineering techniques described in this section are not defined in the original route reflection specification [BC96]. Instead, they emerged to address operational constraints, and were later added to the updated specifications [BCC00, BCC06].

2.2.1 Circumventing the Drawbacks through RR Placement

In a network with route reflection, a client router can connect to any RR in the same network. However, as we discussed in Section 2.1.3.2, improperly configured client-reflector relations may lead to sub-optimal routing paths. Following the guidelines in [BCC06], ISP_{RR} configured a pair of RRs in each of its major POPs, so that client routers connect to the RRs residing in the same POP, making the logical i-BGP topology following the underlying geographic locations.

Given that a RR is located in the same POP with its clients, its best path selections should be the same as those made by its clients, at least at the granularity of the POP level. Thus some of the negative impacts from deploying route reflection mentioned in Section 2.1.3.2, 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 2.6 can be avoided by placing an RR in each POP. As shown in Figure 2.7, 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 RRs at every POP introduced its own scalability concerns. Large ISPs have routers at a large number of POPs, which may 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 initial problem of managing full-mesh i-BGP sessions among a large number of RRs in a global scale. ISP_{RR} circumvented the above issue by building a hierarchy of RRs.

2.2.2 Hierarchical Route Reflection

ISP_{RR} built a hierarchical route reflection structure by recursive application of route reflection. Since route reflection is an effective means to move i-BGP sessions away from full-mesh, one can apply the same idea again at the RR level, i.e., for a set of M POP level RRs that requires $\frac{M*(M-1)}{2}$ full-mesh i-BGP connections, one can simply set up a RR S to connect up the M RRs as its clients. As we already learned, for the overall routing system performance, this RR S should be placed as geographically close to all its clients as possible. However since the RRs are located at different POPs, no single location can satisfy this requirement. This problem can be alleviated to a large degree through the de-

ployment of multiple levels of route reflections. For example, although there is no location that is close to the POP level RRs in both east and west coast of the United States, one could have two higher level RRs, one on east coast and one on west coast, that are closer to the POP level RRs. To assure the propagation of global BGP routing reachability to all i-BGP routers, one only needs to create full-mesh i-BGP connections among all the top level RRs.

ISP_{RR} has several hundreds of i-BGP speakers distributed across two continents. It also has a heterogeneous set of routers with varying capabilities. To effectively control BGP routing information propagation in this large network and to control the routing scalability at individual routers, ISP_{RR} deploys route reflectors at each of its major POPs as described in [BCC06]; for small POPs which only have a small number of routers, they use the RRs located at the nearby major POPs. ISP_{RR} groups POPs into a few tens of regions, and sets 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 that between regions, the ISP has a pair of top layer RRs at the continent level which connect to the region level reflectors as clients.

Figure 2.8 depicts an example hierarchical route reflection system. 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 3^{rd} level round-shape RRs represent POPs. Consider a client router R_c in POP1 (not shown in Figure 2.8): under this hierarchical route reflection, R_c only needs to have i-BGP peering sessions with the two RRs in POP1. This reduced number of BGP neighbors leads to both a reduction in RIB-in size by more than an order of magnitude

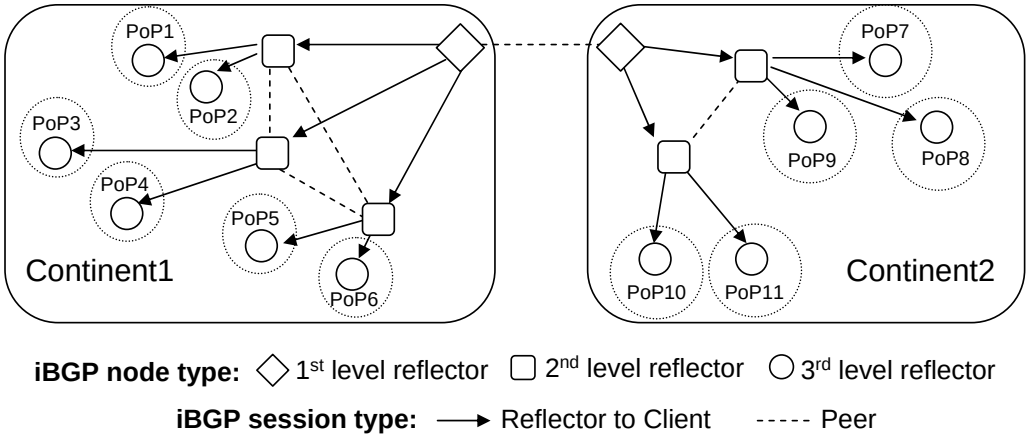


Figure 2.8: An example topology with hierarchical route reflection

compared to the full-mesh i-BGP connection, and a reduction in BGP update counts, since the updates only come from its i-BGP neighbors at POP1 instead from all the ISP_{RR}'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 higher level RRs get propagated through the RR hierarchy to reach client routers; those updates that do not affect the current best paths are filtered out by the RRs. However, these gains in RIB-in size and update count reductions do not come for free, as we explain next.

2.2.3 Impacts of Hierarchical Route Reflection

Increased Hop Distance and Paths: Under full-mesh i-BGP, any i-BGP speaker can reach any other i-BGP speaker with one i-BGP hop. Under a hierarchical route reflection, the distance for an update to travel from one i-BGP speaker to another is at least two hops (client-reflector-client), and in many cases longer. For example in Figure 2.8, the distance between a router R_c1 (e.g., a client of the RR in POP1) in Continent1 and another client router R_c2 in POP11

in Continent2 is 7 i-BGP hops. Due to various delays in propagating an update through each i-BGP hop, this increased hop count can represent a significantly prolonged BGP update propagation delay.

In addition to increased numbers of i-BGP hops, this hierarchical route reflection also leads to increased numbers of alternative paths that updates may travel through. With fully meshed i-BGP connections, each update goes over a single path from any i-BGP router to any other i-BGP router. Although Figure 2.8 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 in reality. Because RRs at each level are replicated, when R_c1 sends an update that affects the selection of path to destination d, two RRs in POP1 will each send the same update to the two regional RRs they are connected to; each of the two regional RRs will in turn send the received update to the two 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.

Additional Path Diversity Reduction: Multi-level hierarchical route reflection topology can also further reduce 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 ISP_{RR} in Continent1. The very top level RR in Continent1 will propagate only one (i.e., its best) route to the top level RR in Continent2. As a result, all the downstream i-BGP RRs and clients in Continent2 will only learn one route (i.e., the best route chosen by the top level RR in Continent1) to reach d, although there are in fact n routes

to reach d in Continent1.

A Mini-Internet within the Internet: We observe that in the topology shown in Figure 2.8, if one replaces reflector–client links as provider–customer links between ASes, and peer (conventional i-BGP session) links between RRs at the same level as peer–peer links between ASes, then this ISP_{RR}'s i-BGP topology remotely resembles that of the Internet's AS-level topology. As future work, it would be interesting to compare and contrast these two models in detail.

2.3 Summary and Future Directions

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

In the past, the number of BGP sessions that a router can handle was relatively small. Thanks to software and hardware technology advances, today's routers on the market are capable of handling thousands of i-BGP sessions [RFP11], removing one of the reasons for route reflection deployment. However the operational cost from configuring and maintaining full-mesh i-BGP sessions remains a strong motivation for deploying route reflections in a large network. Our study suggests that a number of open issues remain, and several potentials also exist, to make route reflection an effective solution towards future routing scalability. We identify the following items as our future work.

2.3.1 Separating Control Plane from Data Plane

As the Internet continues to grow in size, ISP_{RR} also grows rapidly over time and its overall topology becomes more complex. A recent trend in scaling and simplifying network management is to decouple a network's control plane from its data plane. In [FBR04], 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 [FBR04], the authors 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. For example, the top two layers of route reflectors in ISP_{RR} are configured to be responsible only for making an distributing routing information and decisions within its network, and they are not involved in data packet forwarding. The data forwarding is done by the client routers and RRs in the third layer, as well as by other non-i-BGP routers. Therefore, one could view ISP_{RR} as having a dedicated control plane solely for routing information propagation that is separated from data forwarding plane.

We also make three further observations. First, the recent effort in IETF SIDR Working Group [sid] 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. Second, the operational community is utilizing the RR redundancy to develop simple yet effective solutions to improve path diversity, as reported in [RFP11] and [WRC10]. Finally, a recently

proposed routing scalability solution, Virtual Aggregation [FXB10], 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.

2.3.2 Remaining Issues with Route Reflection

We sort the route reflection induced side effects identified in Section 2.1.3.2 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 [Jak10]; and designing effective route flap damping to prevent update flooding with minimized MRAI time value [CPP10].

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 design, 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 [RFP11, WRC10].

CHAPTER 3

BGP Next-hop Diversity and the Impact of Route Reflection

The Internet topological connectivity becomes denser over time. However the de facto routing protocol of the global Internet, BGP, lets each BGP router select and propagate only a single best path to each destination network. This simple design to prevent routing loops leads to a common concern that the rich connectivity is not fully utilized and the lack of alternative paths can reduce a network's robustness to failures as well as flexibility in traffic engineering, and can lead to slow adaptation to topological changes. Furthermore, many Internet service providers have replaced the full-mesh i-BGP connectivity model by route reflection to scale the i-BGP connections, which potentially can further reduce the number of alternative paths used by a network to reach external destinations.

In this chapter, we use i-BGP routing data collected from two global-scale large ISPs, ISP_{FM} and ISP_{RR} each with a different i-BGP architecture, over a 3-year time period to quantify and analyze BGP next-hop diversity for all external destinations to quantify actual BGP path diversity in the operational Internet and investigate how much impact route reflection deployment has on BGP path diversity reduction.

Our results show that both ISPs reach the majority of prefixes through multiple next-hop POPs. We use several case studies of prefixes with different diversity

degrees to study the major factors that influence the number of observed next-hops. Then, we take a step further and perform a comparative study by using i-BGP data collected from the two ISPs. Our results show that both ISPs have similar reduction (up to 42%) in the overall path diversity. Through simulations, we find that the first two topology-independent criteria in BGP best path selection, i.e., LOCAL_PREF and AS_PATH length, eliminated majority of the alternative paths, and the specifics of the i-BGP architectures and topologies have only a minor impact on the overall path reduction, which can further be mitigated through a carefully configured router topology.

3.1 Introduction to BGP Next-hop Diversity

Although a BGP router may learn multiple paths from its peers for a given destination, the BGP specification requires the router to select and propagate only one single best path. As the topological connectivity of the Internet grows denser over time [OPW10], it becomes increasingly desirable to fully utilize multiple available paths.

The number of routers also increased rapidly in large ISPs, currently reaching more than one thousand. To scale with the increasing number of routers that are distributed globally across different geographical locations, many ISPs have deployed route reflection: instead of connecting all i-BGP routers in a full mesh, an AS may use a hierarchy of route reflectors to pass reachability information around. Intuitively, an AS that deploys hierarchical RRs may only use a reduces number of alternative paths to reach external destinations, as compared to an AS with a full mesh i-BGP connectivity.

Currently in IETF, the operation community expresses avid desire to increase

the next-hop diversity, as it represents the opportunities in fast failure recovery, traffic engineering, and load balancing. As a result, several modifications to BGP have been proposed to allow BGP routers to propagate multiple paths for the same destination [RFP11, MFC11, WRC10, USF10]. Despite the promising effort on adding diversity, there has been little understanding on the more fundamental question: what is the existing next-hop diversity in the operation networks? Knowing and understanding the existing next-hop diversity is of significance as it can help us better understand the actual operational needs, and can shed light on important operational practices that influence the degree of next-hop diversity.

In this work, we define and measure the next-hop diversity as observed from all the backbone routers in two global-scale ISPs (referred to as ISP_{FM} and ISP_{RR} based on their internal *full-mesh* i-BGP and *route reflection* i-BGP connectivity, respectively) for all prefixes in the global routing table. Furthermore, to answer the question of whether hierarchical RR may have a negative impact of eliminating alternative paths, we perform a comparative study on BGP path diversity by comparing i-BGP routing data collected. Our findings can be summarized as follows:

- We show that the number of next-hop POPs and ASes for a given destination network varies widely in both ISPs. A significant number of prefixes have a high path diversity; more than 30% and 50% of all prefixes in ISP_{FM} and ISP_{RR} are reached via more than 10 next-hop POPs, mainly due to the topological connectivity between the origin AS and the two measured ISPs. However, we also observe that a considerable amount of prefixes are reached via a single next-hop POP; about 10% and 30% of all prefixes in ISP_{FM} and ISP_{RR} are reached via only 1 next-hop POP.

- We perform case studies to study the characteristics of prefixes with high,

moderate, and low diversity. We find that the topological location of the origin AS as seen by the measurement ISP is a major factor that influences next-hop diversity. More specifically, we discover that the prefixes with a very high next-hop diversity are mostly caused by the lack of geo-presence of ISP_{FM} and ISP_{RR} in some regions.

- Our simulations using the collected i-BGP data show that as much as 42% of alternative paths are eliminated in the studied ISPs, mainly by the first two topology-independent criteria in the BGP best path selection. The specifics of different i-BGP architectures have only a minor impact (less than 2.9%) on the number of alternative paths being used, and even this minor impact can be further mitigated by a well-engineered i-BGP placement and connectivity.

3.2 Background on BGP Next-hop Diversity

In this section, we provide a brief description of path diversity as used in this chapter. Then, we describe i-BGP hidden path phenomenon and explain how it hides alternative paths and reduces the number of overall visible paths.

3.2.1 Path Diversities in BGP

A BGP message reveals path diversities at two different levels: AS and next-hop level, which we refer as AS-path diversity and next-hop diversity respectively.

AS-path diversity: As briefly mentioned in Section 2.1, a BGP message carries AS_PATH attribute which records the AS-level path through which the message traveled to reach the receiving AS. Each AS paths represents an AS level path

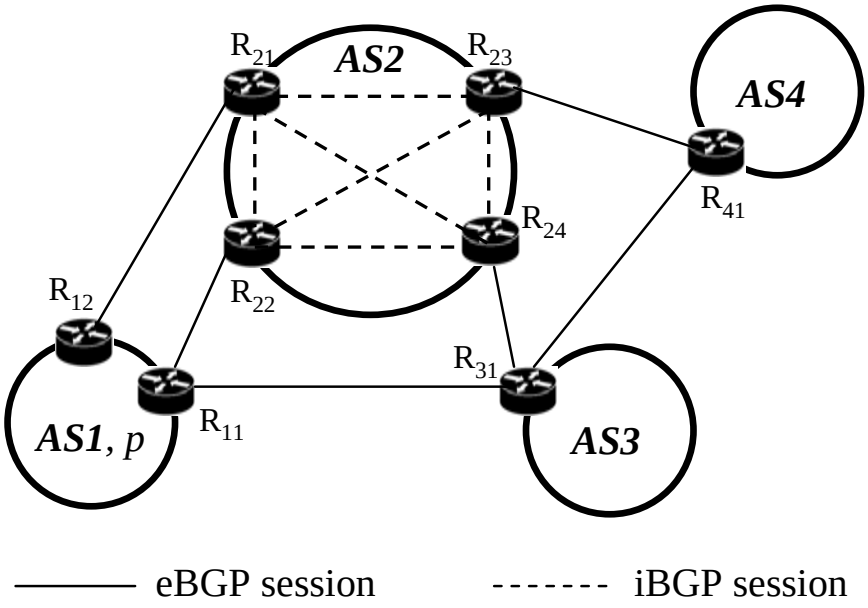


Figure 3.1: An example of BGP connectivity of an ISP

to reach such destination. For example in Figure 3.1, the announcement of the reachability to prefix p in AS1 will arrive at AS4 both through AS2 and AS3. Thus, AS4 learns two different paths (AS4–AS2–AS1 and AS4–AS3–AS1) to reach prefix p. Retaining multiple AS paths in AS4 could be helpful in case of a failure occurring outside of AS4. For example, if AS2 fails, AS4 will still be able to forward the data packets destined to prefix p to AS3, which will in turn forward them to AS1. However, as the receiving end, an operator has little control on the number of visible AS paths to reach a given destination. The alternative AS paths for a given destination may be hidden by the neighboring ASes due to various reasons such as policy, and the distributed nature of BGP routing protocol does not allow an operator to have much influence on the AS paths that are not propagated by the neighboring ASes.

Next-hop diversity: BGP announcement messages for a given prefix can be received from multiple AS neighbors (i.e., next-hop AS), potentially leading to

a high AS-path diversity. Furthermore, there can also be multiple routers (i.e., next-hop routers) to reach each of these neighboring ASes across different cities, which we refer as Point of Presence (i.e., next-hop PoP). For example in Figure 3.1, AS2 receives the reachability information on prefix p through both R_{21} and R_{22} from AS1, and BGP distinguishes these different paths to reach p in AS1 using an attribute named NEXT_HOP.

Maintaining visibility to multiple next-hop routers could be helpful in case of internal failures either on the paths to reach a particular next-hop router or the failure of the next-hop router itself. For example, when R_{12} fails in Figure 3.1, routers in AS2 can use R_{22}–R_{11} and will still be able to reach AS1. Between neighboring ASes, an operator is able to increase or reduce next-hop diversity. When higher next-hop diversity is desired, the operator could deploy more routers to peer with the neighboring ASes.

In general, a higher path diversity at both AS and next-hop level is desired for the purpose of robustness to internal and external failures, traffic engineering, and faster convergence. For example, when an AS (or a router) along the selected path fails and a BGP router has an alternative path in its routing table, the router can fail-over to the alternative path immediately without waiting for the convergence. In addition, a high degree of next-hop diversity offers operators flexibility to direct their traffic for better resource utilization (i.e., load balancing).

Note that ISP_{FM} does not use next-hop-self option. In contrast, ISP_{RR} uses next-hop-self option at the boundaries of its network. Due to such configuration difference, when we perform a comparative study of next-hop diversity between ISP_{FM} and ISP_{RR}, a direct comparison of next-hop diversity at the router level is not meaningful, and thus replaced next-hop router to next-hop POP (i.e., the city which the router is located) diversity.

3.2.2 BGP Best Path Selection

Regardless of different modes or architecture used, all BGP routers select only one best path for each destination prefix and propagate the selected path to neighbor routers. The best path selection considers the following criteria in the order listed [RLH06]: (1) highest LOCAL_PREF[1] , (2) shortest AS_PATH length, (3) lowest ORIGIN, (4) lowest MED, (5) prefer path learned from eBGP session over path learned from i-BGP session, (6) lowest IGP cost, and (7) lowest Router ID. The first 4 criteria examine BGP attributes whose values are independent from the router's location in the internal i-BGP topology, i.e., the preference of a path based on these 4 criteria would be the same regardless of the topological location of the router inside the AS. The last 3 criteria examine values that are topology-dependent and can result in different preference by different routers, depending on the topological location and connectivity of a given router inside the AS.

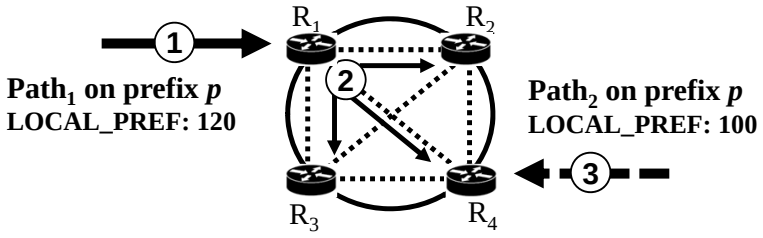
3.2.3 I-BGP Hidden Path Phenomenon

3.2.3.1 Hidden Paths at Border Routers

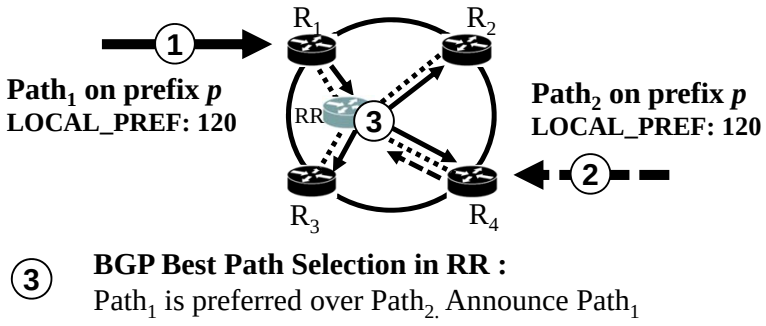
An i-BGP router does not announce the learned, but less preferred paths for a given destination. The less preferred paths are known only to the border router, and consequently i-BGP routers do not know the complete list of available paths to reach a given external destination.

Figure 3.2(a) shows an example of a less preferred path (due to lower LO-

[1]In BGP, LOCAL_PREF attribute value represents the policy of a given path by the local ISP. Typically, an ISP consistently assigns $\alpha > \beta$ as the LOCAL_PREF attribute value for a path via customer and peer respectively, such that the path via customer is preferred over that of peer. In both ISP_{FM} and ISP_{RR}, this is also true.



(a) Topology-independent hidden path



(b) Topology-dependent hidden path

Figure 3.2: Hidden path phenomenon in i-BGP

CAL_PREF attribute value in this case) hidden in a full-mesh i-BGP configuration. In this example, the less preferred path ($Path_2$) will not be announced[2] and known only by the border router (R_4) unless the current best path fails. When $Path_1$ fails, no router except R_4 can switch immediately to use $Path_2$, until R_4 announces $Path_2$ again. This inability to failover immediately to an alternatively path can have significant impact on the data plane performance [WMJ06].

[2]In the case that the less preferred path ($Path_2$) is announced first, the border router (R_4) would explicitly withdraw the path after learning about the more preferred path ($Path_1$).

3.2.3.2 Hidden Paths due to i-BGP Hierarchy

Depending on the i-BGP architecture and the internal router topology, the number of paths learned by a router to reach a destination can differ. Because only the best paths are further propagated from one side of sub-AS (or route reflector) boundary to the other side, the number of overall paths learned can be further reduced.

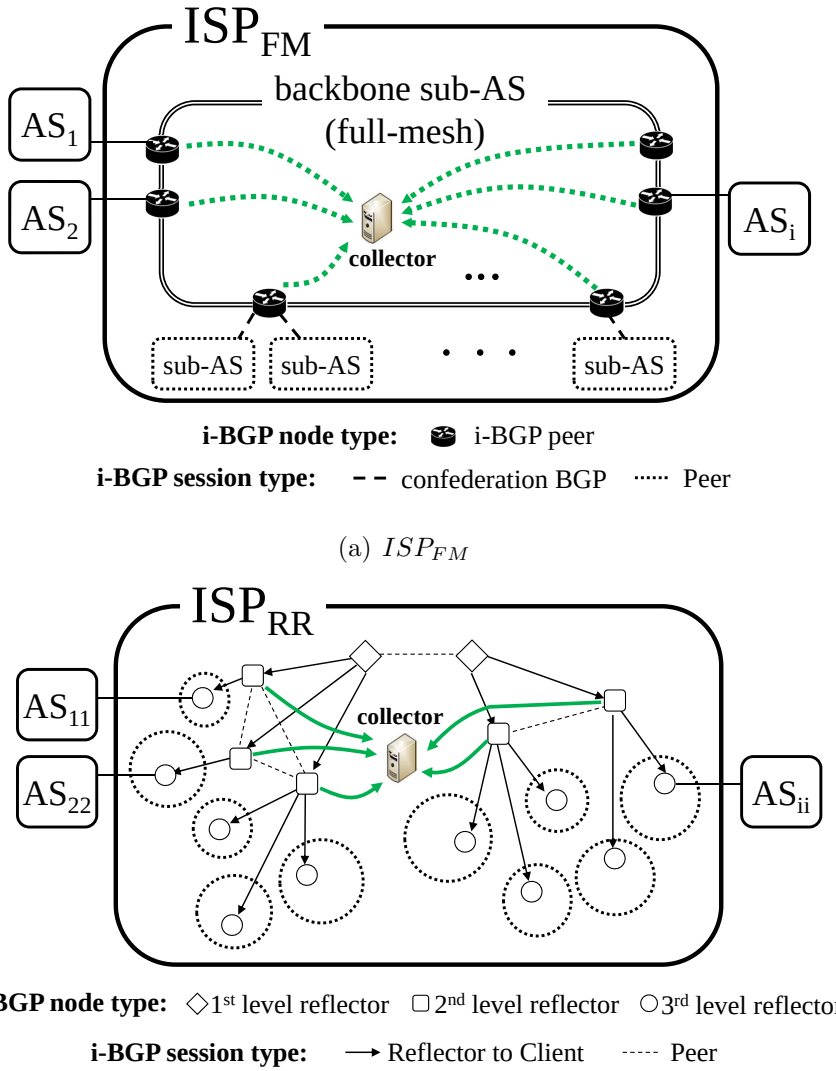
Figure 3.2(b) shows an example of an equally preferred path (at the BGP level) hidden in a route reflection i-BGP configuration. Although all equally preferred path are announced into the route reflector by the border routers, the route reflector chooses only one best path based on its topology-dependent BGP best path selection criteria and propagates only the selected path to its clients, preventing the clients from learning all BGP-level equally preferred paths.

3.3 Measuring BGP Next-hop Diversity

We used i-BGP data collected from ISP_{FM} and ISP_{RR}. In this section, we describe the high level network topology of the 2 ISPs, followed by data collection settings and how we measure the next-hop diversity.

3.3.1 A Brief Description of ISP_{FM}

ISP_{FM} is a global-scale large ISP which uses a single AS number globally in the Internet. It has several hundreds of i-BGP routers distributed across many countries in multiple continents, and uses AS confederations [TMS07] to scale with its network size. Figure 3.3(a) depicts a simplified topology of ISP_{FM} at a high level, where backbone sub-AS represents the backbone network of this ISP, consisting of more than one hundred i-BGP routers connected in a full-mesh



(a) ISP_{FM}

(b) ISP_{RR}

Figure 3.3: Simplified i-BGP topology of two ISPs

(hence referred to as ISP_{FM}).

ISP_{FM} deploys a BGP data collector which establishes an i-BGP peering session with each of the i-BGP routers in the backbone sub-AS to passively record all i-BGP updates received.

3.3.2 A Brief Description of ISP_{RR}

ISP_{RR} is another global-scale ISP which also uses one AS number globally in the Internet. It has several hundreds of i-BGP routers distributed across many countries in multiple continents. It deploys a hierarchical route reflection architecture by recursively applying route reflection. Figure 3.3(b) depicts a simplified hierarchical route reflection system built by ISP_{RR}. 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 3^{rd} level circle-shape RRs represent POPs. A collector (an i-BGP router) is configured as RR client to all route reflectors in the 2nd level route reflectors and passively record all i-BGP updates received.

The top 2 levels (1^{st} and 2^{nd}) of route reflectors in ISP_{RR} serve the sole purpose of distributing routing information to the rest of the network, i.e., they do not carry data traffic. We refer to these route reflectors as backbone routers in ISP_{RR}.

3.3.3 Quantifying Next-hop Diversity

From ISP_{FM} and ISP_{RR}, we gathered routing table snapshots (RIBs) from all backbone i-BGP routers. We first exclude 2 types of prefixes from this measurement study: internal prefixes and potential bogon prefixes with their length shorter than 8 or greater than 24. Then, from each RIB entry, we extracted NEXT_HOP and AS_PATH attributes to measure how many distinct next-hop POPs and ASes are visible collectively in the view of the backbone routers for a given destination.

Next-hop diversity can be measured at 3 different levels, namely next-hop

router, POP, and AS. Note that ISP_{FM} does not use next-hop-self option. In contrast, ISP_{RR} uses next-hop-self option at the boundaries of its network. Due to such configuration difference, a direct comparison of next-hop diversity at the router level is not meaningful, and thus omitted from our study.

3.4 BGP Next-hop Diversity in ISP_{FM}

3.4.1 Next-hop Diversity in ISP_{FM}

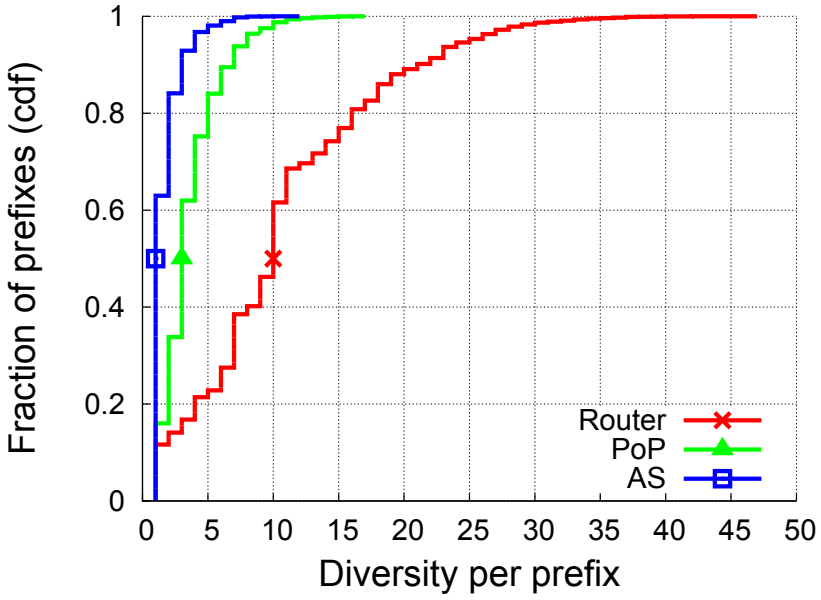


Figure 3.4: Distribution of next-hop diversity in ISP_{FM}

To quantify the next-hop diversity, we use the routing table snapshots taken from all backbone routers on July 1st, 2009. To ensure that the snapshots are representative, we also measured next-hop diversity using routing table snapshots taken at different times. In addition, we checked that the total number of prefixes in each snapshot and the set of unique neighbor ASes are roughly the same.

Figure 3.4 shows the cumulative distribution (CDF) of the number of next-hop ASes to reach a prefix in ISP_{FM}. For the total 276,712 prefixes, we observe that about 62% of all prefixes are reached via 1 neighbor, and almost all prefixes (about 96%) can be reached via less than or equal to 5 neighboring ASes.

Note that the number of next-hop ASes represents a gross diversity at the inter-domain routing level. For those prefixes that can only be reached through one neighbor, ISP_{FM} must wait for BGP to explore and settle down on the routes via other neighbors (if there is any) when the particular neighbor AS fails. The prolonged convergence delay in this case can potentially degrade the performance in the data plane [WMJ06]. However, such number of next-hop ASes only describe an abstract reachability at the logical AS level. In a typical operation settings, two ASes often set up peering sessions at different geographical locations using multiple BGP routers as explained earlier.

We further measure the number of available next-hop routers and their geographical locations (i.e., POPs) to reach a given destination. Figure 3.4 shows the distribution of the number of observed next-hop routers and POPs to reach each destination prefix. We observe that even though 18% of prefixes can still be reached via only one POP from one neighboring AS, the majority of the prefixes can be reached via 2 to 5 POPs. Furthermore, given that there often exist multiple routers in a given POP, the next-hop diversity is further amplified and varies widely from 1 up to 47. Most of the prefixes (88%) have more than 2 next-hops, and around 47% of all prefixes have their next-hop diversity between 6 and 12. We also observe that there exists a small fraction of prefixes (1.6%) with a very high next-hop diversity (>=30).

In Figure 3.4, we observe that prefixes with the same next-hop AS diversity can have different next-hop POP and router diversity. This indicates that the

amount of visible next-hop diversity can depend on not only the number of neighbor ASes but the number of peering routers with neighbor ASes through which ISP_{FM} reaches a given destination.

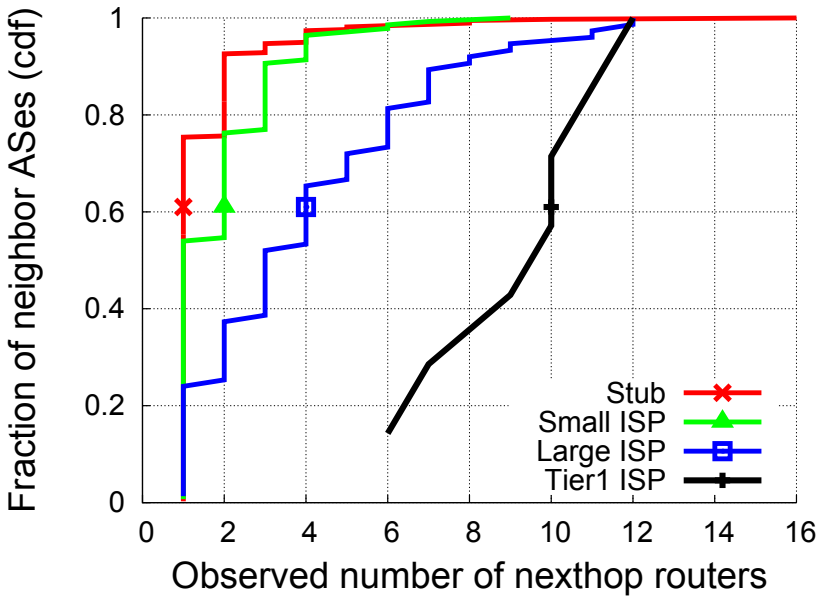


Figure 3.5: Observed connectivity of different neighbor types of ISP_{FM}

To show the number of peering routers differ across different neighbor ASes, we classified each neighbor AS as one of "Tier1", "Large ISP", "Small ISP", and "Stub" based on the classification found in [ZLM]. Then, we measured the number of peering routers for each of the types. Figure 3.5 shows that, in general larger neighbor ASes tend to have a higher number of routers peering with ISP_{FM}. This tendency is reflected in next-hop diversity.

For example, if two prefixes are reached via a Tier1 and a small ISP neighbor respectively, then based on Figure 3.5, the former prefix can have its next-hop diversity ranging from 6 to 12 while the diversity of the latter prefix can range from 1 to 9. Note that there exist few Stub ASes (e.g., UltraDNS, Amazon, Akamai, etc) whose number of peering routers is exceptionally high. This is due

to their specific business needs to provide global wide services, and connect to ISP_{FM} with many routers at different locations globally. The high number of peering routers enables these ASes to increase next-hop diversity in ISP_{FM} by simply announcing their prefixes over multiple peering routers. The existence of highly connected neighbors such as these large stub ASes, large ISPs, and Tier1 ISPs shown in Figure 3.5 suggests that, by utilizing available connectivity, there exist opportunities in ISP_{FM}'s current network to increase and exploit the existing path diversity.

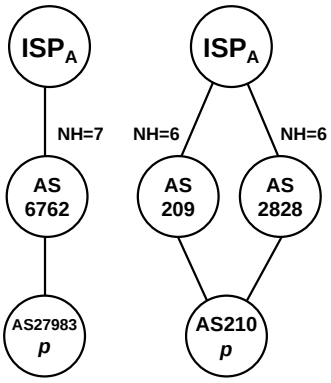
3.4.2 Case Studies: Prefixes with Low, Moderate, and High Next-hop Diversity

In this section, we take a closer look at representative cases of prefixes with the low, moderate, and high next-hop diversity to shed lights on the main factors that determine the amount of next-hop diversity for a given prefix.

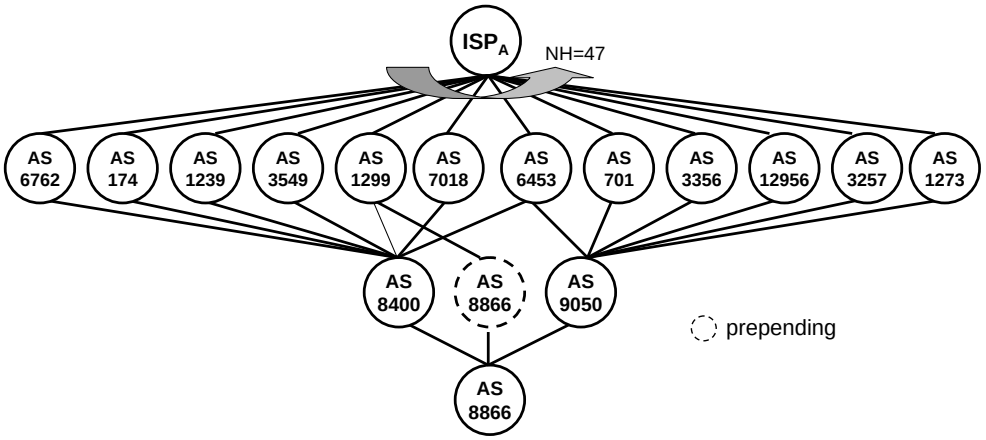
3.4.2.1 Prefixes with Low Diversity

In our study, prefixes with lowest next-hop diversity have announced by a neighboring AS through a single next-hop router like prefix 201.133.104.0/24 shown in Figure 3.6(a). There can be two reasons: (i) there is only one path to reach this prefix, and/or (ii) BGP's design choice to select and propagate only the best path to the neighbors prevents ISP_{FM} from being able to see other alternative paths. By further investigating the update messages, we found that the main reason is the latter; when the best (and the only visible) path fails, we could observe that oftentimes many alternative paths, which were hidden previously due to the BGP path selection, got exposed during i-BGP convergence process. As the result, the next-hop diversity for this prefix p would be 1 despite the fact that there do exist

(a) Case 1: Low Diversity



(b) Case 2: Moderate Diversity



(c) Case 3: High Diversity

Figure 3.6: Representative cases of prefixes with low, moderate, and high next-hop diversity

other paths.

We wondered if this observation is true for all prefixes identified to have the lowest diversity. Using historical AS level Internet topology available from

[ZLM], we verified that for all prefixes with next-hop diversity equal to one in our study, the prefixes do have multiple alternative next-hops. In other words, BGP's design choice to select and propagate only a single best path hides the alternative paths and prevented the prefixes in this class to have higher diversity, although alternative paths to reach these prefixes do exist.

The above examples show that how BGP path preference limits the next-hop diversity. However, as the path preference are configurable by design (i.e., via tunable parameters such as weight[3], LOCAL_PREF, etc.), a network operator may be able to adjust path preference to achieve higher next-hop diversity while respecting the network's routing policy.

3.4.2.2 Prefixes with Moderate Diversity

In Figure 3.6(b), we present two representative cases in moderate diversity. We classify prefixes whose next-hop POP diversity is between 5 and 14 as moderate, which is more than 50% of all prefixes in Figure 3.9(a). Prefix 190.103.225.0/24 announced by AS27983 is the first case. This prefix can be reached from ISP_{FM} through AS6762, a large ISP. The number of next-hops between ISP_{FM} and AS6762 were 7. Another representative case of a prefix with moderate next-hop diversity was prefix 204.113.217.0/24 announced by AS210. The AS and next-hop diversity are 2 and 12 respectively.

In both examples, the prefixes were reached through at least one neighbor AS which is a large ISP with at least 6 BGP peering sessions with ISP_{FM}. From these two cases, we can see that (i) the number of peering routers has an impact on the next-hop diversity; 190.103.225.0/24 announced by AS27983 has moderate diversity because its provider (AS6762) has 7 multiple peering routers

[3]supported by router vendors

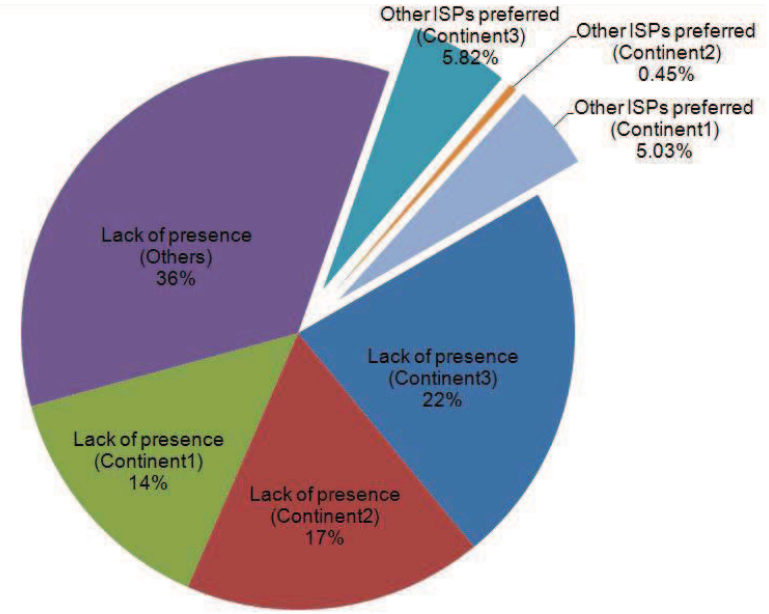


Figure 3.7: Geographical presence of ISP_{FM}

with ISP_{FM}. In addition, we also see that (ii) multi-homing helps increase path diversity; 204.113.217.0/24 announced by AS210 has 12 next-hop diversity in ISP_{FM} by multi-homing with AS209 and AS2828.

3.4.2.3 Prefixes with High Diversity

Our last case study explores prefixes with very high degree of next-hop diversity. Figure 3.6(c) shows a prefix 83.228.80.0/23 announced by AS8866, a regional ISP. AS8866 multi-homes with two providers (AS8400 and AS9050) which connect to many Tier1 and large ISPs. By becoming a customer of these two highly connected providers, prefix 83.228.80.0/23 in AS8866 inherently becomes visible through highly diverse paths from the perspective of ISP_{FM}.

In general, a common characteristic observed in prefixes with high degrees of next-hop diversity is that their origin ASes do not directly connect to ISP_{FM}.

From this observation, we hypothesized that the lack of geographical presence of ISP_{FM} can be a factor that determined the set of prefixes with high next-hop diversity. In the regions that ISP_{FM} does not provide connectivity, the origin ASes would connect to other ISPs when they wish to connect to the Internet. If these local ISPs happen to multi-home with many large ISPs except ISP_{FM}, then there will be many paths with equal AS_PATH length between the origin AS and ISP_{FM}, which leads to the very high next-hop diversity.

To verify our hypothesis, we checked the prefix origination point of prefixes with very high next-hop diversity against the POPs covered by ISP_{FM}. To find the location of prefix origination point, we used MaxMind GeoLite package [max] to map each prefix into a city. Then for these cities, we checked whether any POP of ISP_{FM} is present. Figure 3.7 verifies our hypothesis; in 89% of prefixes with very high next-hop diversity, ISP_{FM} did not have a presence.

This observation that some prefixes can have a very high diversity regardless of the ISP's intention can be an important input to the proposed BGP modifications [RFP11, MFC11, WRC10], which increase the diversity for all prefixes. Our results suggest that more intelligent approaches could be used to utilize the router resources more efficiently by increasing diversity selectively for interested prefixes, rather than over-provisioning these high diversity prefixes altogether.

3.4.3 BGP next-hop diversity changes in time

In this section, we seek to find out if there is a general trend of next-hop diversity changes over time. Due to the large amount of i-BGP routing data and the processing loads, we sampled the next-hop diversity of the first day of each month from July 2007 to July 2009. In addition, to better capture the next-hop diversity change in time, we only consider the prefixes that continuously exist over the

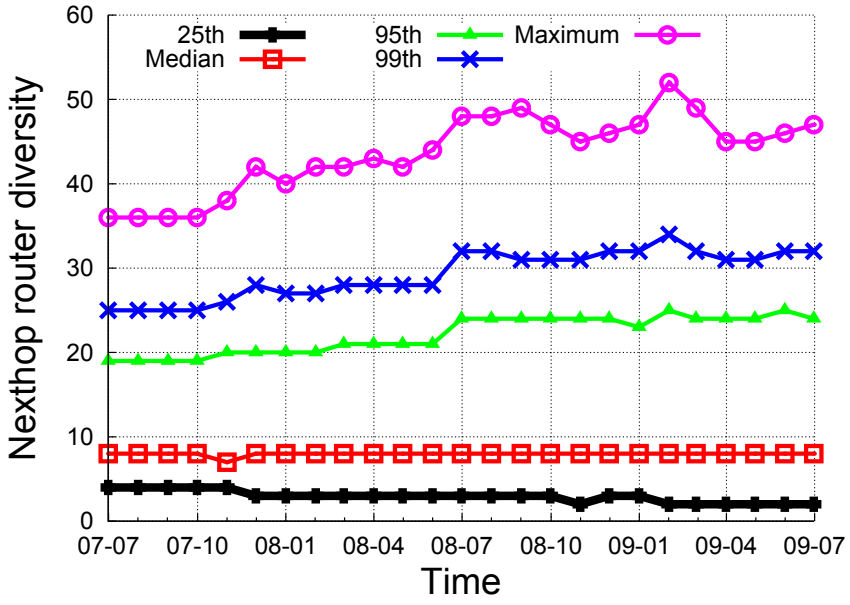
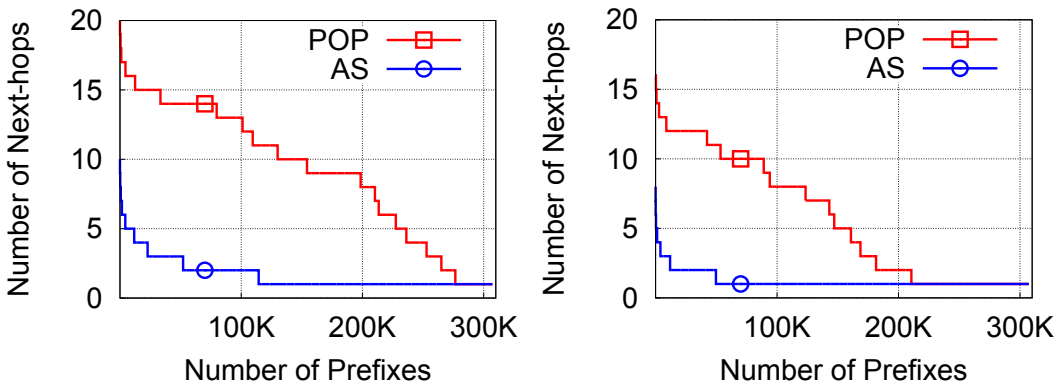


Figure 3.8: Next-hop Diversity Change in Time

entire two-year measurement period, which leaves us total 220,432 prefixes.

Figure 3.8 depicts next-hop router diversity changes at 25, 50, 95, 99 percentile, and maximum in next-hop router diversity distribution curves at different times. For example, on July 2007 (the leftmost data points), the median, 99%, and maximum next-hop diversity were 8, 25 and 36. Figure 3.8 shows that over the last two years, the median value stayed almost the same, though we checked that the individual prefix does shift its diversity to some extent. We do not observe any significant pattern of changes. As the individual prefix's diversity is determined by a complex interaction between the topological and geographical location of the origin AS, the inter-domain routing path from the origin to ISP_{FM}, the number of next-hop routers, and the BGP routing decisions, the path diversity changes in time with a seemingly unpredictable manner.

However, we also observed that the maximal next-hop diversity slowly increases in time, mainly due to the increased number of backbone routers inside



(a) ISP_{FM} (b) ISP_{RR}

Figure 3.9: Next-hop POP and AS diversity in ISP_{FM} and ISP_{RR}

ISP_{FM}. The maximum, 99 percentile, and 95 percentile next-hop router diversity gradually increased in time, and 25 percentile value decreased slightly. After further investigation, we found that the increasing trend in maximum, 99 percentile, and 95 percentile is mainly due to the increased number of peering routers between ISP_{FM} and its neighbors. Since July 2007, the number of backbone routers in ISP_{FM} gradually increased up to 19 additional routers by the end of July 2009. This also confirms to the findings of case studies we made previously in Section 3.4.2.

3.5 Comparing BGP Next-hop Diversity in ISP_{FM} and ISP_{RR}

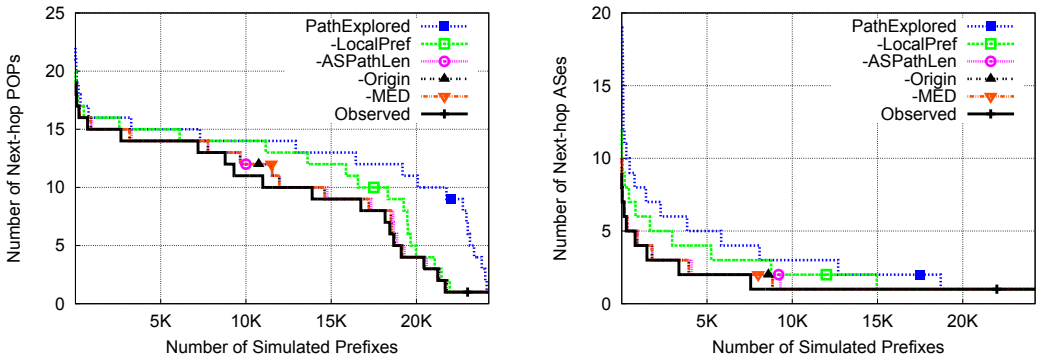
In this section, we measure next-hop diversity in ISP_{RR} and compare the results with the next-hop diversity in ISP_{FM}, based on our measurement results based on the routing table snapshots taken on June 3rd, 2010. To ensure that the snapshots are representative, we performed the same measurements using routing tables taken on each day during one week of June 3rd to 9th and on every 1st day

of each month from January to May in 2010. We verified that the distributions of next-hop POP and AS diversity are similar. In addition, we checked that the total number of prefix entries and the set of unique POPs and neighbor ASes are roughly the same.

Figure 3.9(b) shows the distributions of next-hop POP and AS diversity of the same 307,212 prefixes. The difference in the number of total prefix between the two ISPs mainly comes from the different announcements made by the neighboring ASes.

We make a number of common observations across the two ISPs. First, similar to what we observed in ISP_{FM}, a considerable and relatively larger number of prefixes can be reached via only one neighbor POP and AS; 34.02% and 84.42% of all prefixes have both their next-hop POP and AS diversity equal to 1. Second, as in the case of ISP_{FM}, overall next-hop POP diversity is relatively higher than next-hop AS diversity, indicating that ISP_{RR} also peers with its neighbor ASes in multiple POPs. Third, we observe a few groups of prefixes sharing the same degree of POP diversity (e.g., POP diversity equal to 12 and 8). Lastly, we find that the highest degree of diversity in ISP_{RR} is mostly related to how the origin ASes connect to ISP_{RR}. We identified the top 8,881 prefixes with the highest degree of next-hop diversity inside ISP_{RR}, announced by 1,336 unique origin ASes. Then, we used MaxMind GeoLite package [max] to map each prefix into a city. Finally for these mapped cities, we checked whether any POP of ISP_{RR} is present. We found that all 1,336 (100%) origin ASes that announced the prefixes with the highest degree of diversity do not directly connect to the two ISPs and that more than 91% of these origin ASes are located in regions that ISP_{RR} is physically absent.

Although both ISPs are classified as global-scale large ISPs, there is a notice-



(a) Next-hop POP diversity

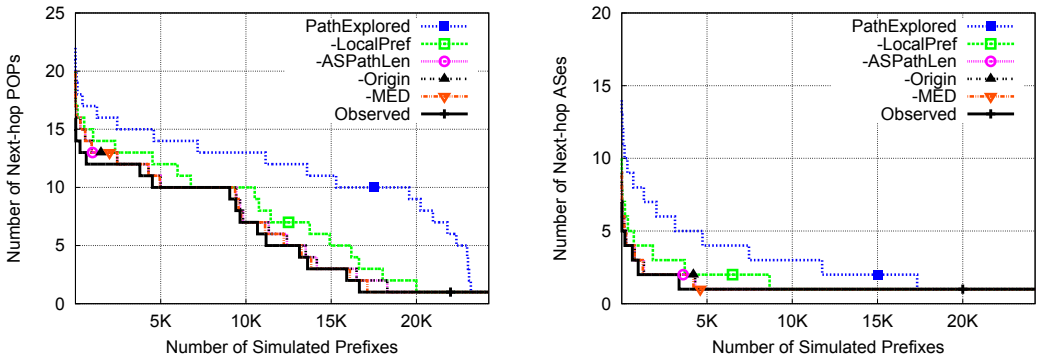
(b) Next-hop AS diversity

Figure 3.10: Next-hop diversity reduction in ISP_{FM}

able difference in next-hop diversity. First, we observe that the maximum number of next-hop POP and AS is different, potentially caused by the difference in their external connectivity. More importantly, we observe that the overall number of ISP_{RR}'s next-hop POPs and ASes to reach a given prefix is relatively lower, compared to ISP_{FM}. For example in ISP_{FM}, there are 10.17% and 62.76% of all prefixes with 1 next-hop POP and AS respectively. However in ISP_{RR}, we observe that relatively more prefixes (34.02% and 84.42%) have only 1 next-hop POP and AS respectively.

3.6 Investigating the Impact of Route Reflection on Next-hop Diversity Reduction

In this section, we further investigate different impacting factors on path diversity to explain the observed discrepancy by examining the i-BGP updates collected from the two ISPs for 6-month time period. More specifically, we focus on understanding the following 3 factors and their impact on the overall next-hop diversity: (1) *external connectivity*, (2) *topology-independent hidden path*, and (3)



(a) Next-hop POP diversity

(b) Next-hop AS diversity

Figure 3.11: Next-hop diversity reduction in ISP_{RR}

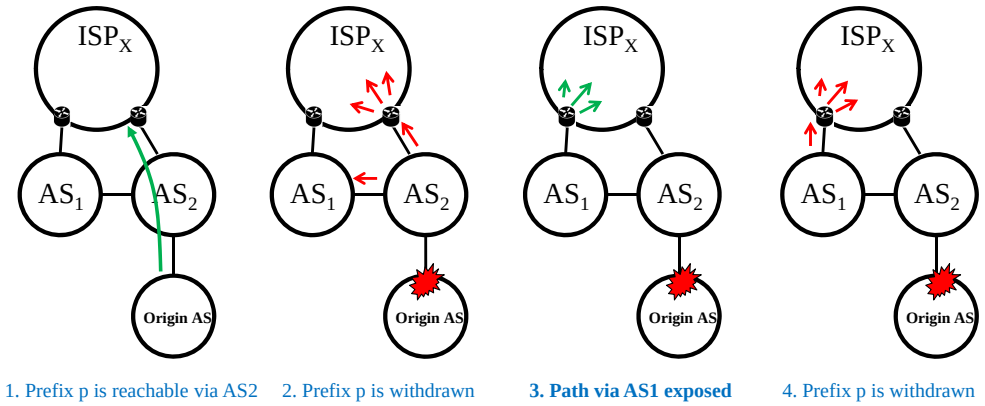


Figure 3.12: Inferring external connectivity

topology-dependent hidden path.

3.6.1 External Connectivity

As we have seen in the previous section, next-hop POP and AS diversity of a prefix can potentially be upper-bounded by the external BGP connectivity of the ISP with its neighbor ASes. The most straightforward approach to obtain the exact amount of external connectivity of an AS with its neighbor ASes is to examine the configurations of all border routers inside the AS. However, this requires access

to all border routers in each of the two measurement ISPs, which we did not have at the time of our measurement. Thus in this work, we estimate the amount of external connectivity by observing the routing dynamics in each of the two ISPs to examine how different (or similar) these 2 ISPs are in terms of the amount of their external connectivity. More specifically, we examine the i-BGP updates during a time period and estimate the external connectivity by recording the path exposed by the prefixes that have their paths explored. Figure 3.12 illustrates how we infer the external connectivity for a given ISP, ISP_x. In this figure, prefix p is initially reachable through AS2 (step 1). When the origin AS fails and the current best path is withdrawn (step 2), the previous hidden less preferred path would be explored before finally declaring that the prefix is not reachable (step 3 and 4).

One challenge in estimating the external connectivity by observing the routing dynamics is to determine the time duration. If the time duration is too long, the dynamics can include the permanent topology changes of the Internet [OPW10]. On the other hand, if the time duration is too short, we will not observe the prefixes which are inactive during the observation period, and the number of observed prefixes can be too small. To capture as many prefixes without including the permanent topology changes, we decided to look at multiple short time durations of one week that do not overlap over a longer period of time; to estimate the external connectivity, we use the i-BGP data collected over 6 months during the 1st week January, February, March, April, May, and June in 2010. Overall, we identified 88,236 prefixes announced by 12,727 origin ASes (38% of all ASes: 10 Tier-1s, 1,346 Transits, and 11,371 Stubs) which are approximately 1/3 of all prefixes and origin ASes in the global routing table. Additionally, we checked that the prefixes and their origin ASes cover various AS types, topological locations, and the overall next-hop diversity. Although we did not capture all prefixes and

ASes in the global routing table, our goal in this paper is to compare the relative difference of the external connectivity of the two ISPs, rather than precisely estimating all external connectivity for a given ISP. For this purpose, we believe that the total number of identified prefixes and origin ASes is sufficient.

In each of 6 independent simulations, the percentage in diversity reduction varies slightly. However, we essentially make the same observation across the multiple independent simulations, and the generality of our conclusion does not change. Therefore in this paper, we present the results on one week from June 3rd to June 9th in 2010 as the representative result for clarity. The number of identified prefixes during this week is 24,244 (about 7% of all prefixes), announced by 4,457 unique origin ASes (13.59% of all ASes: 5 Tier-1s, 648 Transits, and 3804 Stubs).

The blue lines (labeled *PathExplored* marked with filled square) in Figure 3.10 and 3.11 show the number of next-hop POPs and ASes based on the estimated external connectivity for the identified prefixes in ISP_{FM} and ISP_{RR}. The distributions of the estimated external connectivity between the 2 ISPs reveal that there is not a significant discrepancy, and therefore, we concluded that the external connectivity is not the dominating cause for the discrepancy observed in Figure 3.9.

3.6.2 Topology-independent Hidden Path

Given that the distribution of external connectivity of the 2 ISPs is similar, we measure the amount of topology-independent hidden path, which happens regardless of the i-BGP architecture or router topology, as described earlier in Section 3.2.3. To quantify the amount of next-hop diversity reduced by i-BGP hidden path phenomenon, we simulate the first 4 topology-independent criteria of

BGP best path selection algorithm and count how many external paths remain equally preferred by all routers inside the ISP after each of the criteria. The number of such remaining paths represents the path diversity after hidden path phenomenon caused by each of the first four BGP best path selection criteria.

3.6.2.1 ISP_{FM}

Figure 3.10 summarizes our simulation results for ISP_{FM}. In Figure 3.10(a) and Figure 3.10(b), each green (marked with a square), pink (marked with a circle), dotted black (marked with a triangle), orange (marked with an upside-down triangle) colored lines show the remaining next-hop POP and AS diversity respectively after each step of the first 4 best path selection criteria in ISP_{FM}. For example in Figure 3.10(a), our estimated external connectivity (i.e., blue line marked with a square) indicates that there are only 0.4% of prefixes initially with their next-hop POP diversity equal to 1. After considering the 1st criterion (LOCAL_PREF comparison), the green line (labeled -*LocalPref*) shows that more prefixes (7.36%) have the next-hop POP diversity equal to 1. This means, among multiple external paths to reach a given prefix, only *one* path stands out due to its higher LOCAL_PREF value, making other (less preferred) paths hidden inside the border routers.

Overall, the first 2 criteria contribute to most of the next-hop diversity reduction. After the 1st criterion (LOCAL_PREF comparison), about 10% of overall next-hop POP diversity is reduced in average. Then additional 12% next-hop POP diversity reduction happened after the 2nd criterion (AS_PATH length comparison).

3.6.2.2 ISP_{RR}

Figure 3.11 summarizes our simulation results for ISP_{RR}. As in the case of ISP_{FM}, the first 2 criteria of the best path selection are identified as the dominating factors that reduce next-hop diversity. However, the amount of reduction happened by each of the 2 criteria is quite different. In case of ISP_{RR}, the 1st criterion (LOCAL_PREF) had the most impact on next-hop diversity reduction (of about 29%), and is the main reason why the 2 ISPs have such discrepancy in the measured next-hop diversity in Figure 3.9. Our results reveal that although ISP_{RR} has a comparable amount of external connectivity compared to ISP_{FM}, relatively less number of paths are equally preferred after examining LOCAL_PREF attribute value and the subsequent topology-independent criteria.

3.6.3 Topology-dependent Hidden Path

The i-BGP hidden path phenomenon due to the first 4 topology-independent criteria of the best path selection happens regardless of the i-BGP topology. This implies that even in the full-mesh topology, the remaining next-hop diversity after the 4th criterion is the upper-bound, and that further reduction caused by the topology-dependent criteria represents the cost of moving away from the full-mesh topology.

Thus, we define the difference between measured diversity as seen by the backbone routers (i.e., black line labeled *BackBone*) and the diversity after the 4th criterion of best path selection (orange line labeled *-MED*) as the amount of diversity reduced due to topology and connectivity between the border routers and the backbone routers.

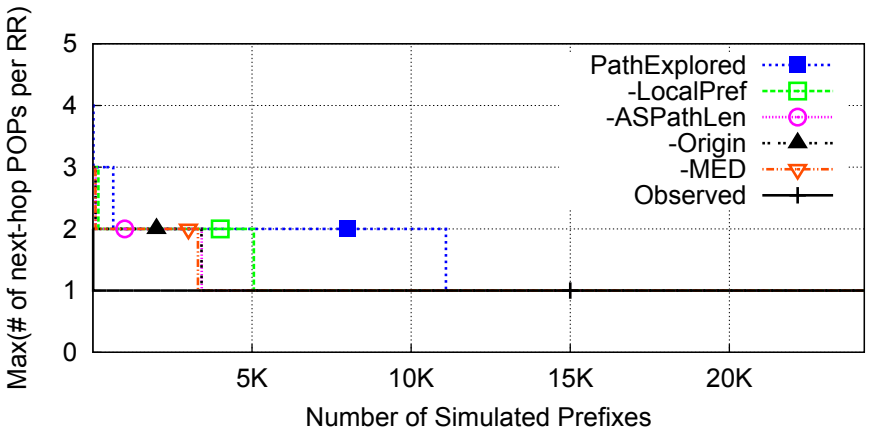


Figure 3.13: Maximum number of next-hop POPs per RR

3.6.3.1 The Impact of i-BGP Topology and Next-hop Diversity Reduction

In both Figure 3.10 and 3.11, we observe that the difference between the solid orange line (labeled *-MED* marked with an upside-down triangle) and the solid black line (labeled *BackBone* marked with a short vertical line) is relatively small. Overall, the average reduction due to the topology-dependent factors across all simulated prefixes is small; even with ISP_{RR}'s multi-level hierarchical route reflection architecture and its topology, there is only up to 2.9% reduction.

3.6.3.2 i-BGP Topology Design and Next-hop Diversity

In route reflection architecture, path diversity reduction happens essentially by deploying a relatively smaller number of route reflectors, compared to the available number of paths per route reflector. Given that ISP_{RR} has only a minor reduction in overall next-hop diversity, we further verify that in ISP_{RR} the number of route reflectors in the backbone routing infrastructure roughly match the number of available next-hop POPs. We first calculated the number of distinct

next-hop POPs observed by each route reflector before and after considering the first 4 BGP best path selection criteria, then chose the maximum number across all route reflectors. For example, if the number of observed next-hop POPs by two route reflectors are 2 and 5 respectively, the maximum number of next-hop POPs per route reflector (as shown in Figure 3.13) is max(2,5) = 5. If this number is equal to 1 for a given prefix, it implies that there is sufficient number of route reflectors in the network to preserve the observed next-hop POP diversity for that prefix.

Figure 3.13 summarizes our results. First, the number of maximum next-hop POP is 1 for the majority (more than 54%) of the prefixes. This indicates that the route reflectors are well-placed for these prefixes in terms of their next-hop POP diversity density. For the prefixes with the available next-hop POP greater than 1, there is a noticeable decrease in the maximum number of observed next-hop POPs per route reflector after considering the first 4 topology independent BGP best path selection criteria; for more than 32% of simulated prefixes, the number of next-hop POP decreased to 1. This result confirms again that in the current i-BGP design that reduces significant overall path diversity regardless of topology and that a more scalable i-BGP architectures can be used without much sacrifice in the overall path diversity reduction, when the i-BGP topology is carefully designed.

3.7 Discussions of Related Works

Prior works on path diversity fall into two classes: (i) quantifying existing path diversity and (ii) increasing path diversity.

Among prior works in the first class, Teixeira et al. [TMS03] measured the IP

level path diversity inside a ISP (Sprint)'s backbone network, and showed that Sprint has significant *IP level* path diversity among their POPs. In contrast, we measure the *BGP level* exiting point diversity. On the other hand, [HWJ06, NM01, VAZ07, VCS09] measured path diversity at the AS granularity. While our work measures path diversity from the perspective of a ISP, they mostly focused on multi-homing stub ASes because their common goal was to understand the impact of path diversity on data forwarding performance for a given multi-homing AS. Uhlig et al. [UT06] quantified path diversity in a ISP which configured its network with route reflection [MGW02]. Because they focused on the impact of route reflection on path diversity reduction, they used simulations to measure the path diversity inside an ISP using a small set of sampled prefixes. Our measurements confirm with their results in that many paths are not visible from the i-BGP routers because of the local routing policy. However, we measure next-hop diversity of production routers in a full-mesh network, and quantifies path diversity for all prefixes in the global routing table as well as the trend in overtime using the actual number of routers, which yield a more tangible and comprehensive understanding of path diversity and different impacting factors of path diversity in both general and corner cases.

The second class of prior works involve efforts to increase path diversity. Recently, the operator community starts to demand higher path diversity to accommodate the newly emerging applications [RFP11, WRC10, SF09]. This led to on-going efforts to increase path diversity by modifying the behavior of BGP. Walton et al. and Schrieck et al. [WRC10, SF09] propose a BGP capability, *Add-Path*, to distribute multiple paths for a given destination. This new extension increases the availability of additional paths, and can help reduce persistent route oscillations and route convergence within a network. While we deem it necessary to have a general way to exchange multiple paths between BGP routers, in this

chapter we showed that the majority of prefixes can be reached via more than one next-hop routers without changes to BGP. One interesting question is whether ISPs actually utilize the existing diversity before moving forward to increase it.

On the other hand, instead of modifying BGP, Raszuk et al. [RFP11] proposed to deploy multiple BGP route reflectors planes, and each additional plane incrementally increases the number of alternative paths. The key idea is to configure each reflector such that the N_{th} reflector could select and distribute the N_{th} best path. This technique echoes our observations in this chapter that changing BGP path preferences can greatly affect the diversity, but note that this technique might not be applicable to networks such as ISP_{FM}, that does not use route reflection to organize its network. We hope that our measurement results can serve as valuable input on these efforts to decide whether such mechanisms to increase path diversity are necessary.

3.8 Summary and Future Work

BGP has gone through many changes as it operates as the de-facto routing protocol in the Internet. Its original design required a BGP router to select and propagate only a single best path to its neighbors. This design choice is being reconsidered to increase path diversity. However, there has been little understanding on path diversity in the existing system, and the necessity and effectiveness of different proposals are not clear.

Using i-BGP routing data collected from routers in the backbone routing infrastructure inside two global-scale ISPs, we show that there already exist opportunities in the existing network for the ISPs to utilize its diversity by showing that the majority of prefixes could be reached through multiple next-hop routers.

Also, our case studies reveal that the ISPs may further increase path diversity without any modification to BGP, by adjusting path preference values while respecting the network's policy. Furthermore, we find that a very small number of prefixes maintain a high degree of diversity, and in most cases, they happen specifically and regardless of the ISPs' intention, caused by the lack of geographical presence of ISP_{FM} and ISP_{RR} in the regions where origin ASes are located. We also find that in ISP_{FM}, the overall next-hop diversity have not changed much over the past two years, but the maximal next-hop diversity slowly increases in time, mainly due to the increased number of backbone routers.

Our measurement study based on the i-BGP data collected from two large ISPs quantifies the degree of path diversity in these 2 ISPs and reveals the most influential factors on BGP path diversity. Our results shows that although there is a significant overall path diversity reduction, the reduction caused by the specifics of i-BGP architecture inside ISP_{RR} is small. There are two main reasons. First, topology-independent criteria are high in the BGP best path selection decision making order and contribute significantly to the overall reduction. Second, a well-engineered i-BGP topology mitigates the topology-dependent reduction.

We discover that the overall alternative path reductions in the two large ISPs is mainly due to the topology-independent factors. However, there was a noticeable difference in the amount of reduction due to LOCAL_PREF attribute in BGP best path selection. We conjecture that this difference can be explained by the economic factors such as the access-circuit prices, transit prices, SLA's and peering policies which are affected by the different geographical regions that the two ISPs serve and leave the detailed analysis and verification as our future work.

In this work, we focused on understanding the static path diversity in different i-BGP architectures in the absence of failures. It remains as an open question

how different i-BGP architectures may impact BGP convergence in the presence of topological changes, which is the subject of our ongoing effort.

CHAPTER 4

Understanding BGP Convergence inside Large ISPs

In recent years there have been many measurement studies that use BGP updates between ASes to examine BGP routing dynamics across the global Internet. However, there has been virtually no measurement studies on BGP dynamics inside Internet service provider (ISP) networks. In this work, we use i-BGP data collected from two large ISPs during a 14-month period to define, quantify, and analyze i-BGP convergence. Our measurement results reveal interesting characteristics and performance issues of i-BGP convergence which have not been reported previously. More specifically, we quantify convergence delays of two different i-BGP architectures, namely full-mesh and hierarchical route-reflectors (HRR). We show that the delays due to HRR are insignificant in most cases, and can be further mitigated through carefully configured router topology.

4.1 Introduction to I-BGP Convergence

BGP [RLH06] is the global routing protocol used in the Internet to communicate reachability information between routers in different autonomous systems (ASes) as well as within a single AS. Because BGP dynamics have a direct impact on the data delivery performance [WMJ06, Zha04, PWM03, KKK07], in recent years

extensive measurement and analytic research efforts have been devoted to understanding BGP routing dynamics. As one of the seminal BGP measurement studies, Labovitz et al. [LMJ99, LAW01, LAA01] showed the existence of slow BGP routing convergence. Subsequent measurement studies confirmed the wide existence of slow convergence [OZP06] and proposed a variety of BGP modifications to speed up BGP routing convergence [ZAL04, PAM05, SMS06, BAS03, PZW02, CDZ05, DS04, SKM06].

As the Internet has grown in its size and connectivity density over time, so have the large ISPs. The rapid increase in both the number of routers in large ISPs and the complexity in their interconnections escalated interests and concerns on BGP routing dynamics *inside* a single autonomous system; such dynamics can have implications on the overall data packet delivery service and performance. However, most of the previous analytic studies focus on BGP dynamics at the inter-AS level, using a simplified model of the Internet represented as a graph where individual nodes represent ASes. The BGP dynamics inside each AS has largely remained as a missing puzzle for a comprehensive and complete understanding of the end-to-end routing performance.

In this chapter, we take a first step towards measuring BGP convergence inside large ISPs and the impact of different i-BGP architectures. Our measurement and analysis are based on i-BGP data collected during a 14-month period from two global-scale ISPs, each with a different i-BGP architecture. Our contributions and findings in this chapter can be summarized as follows.

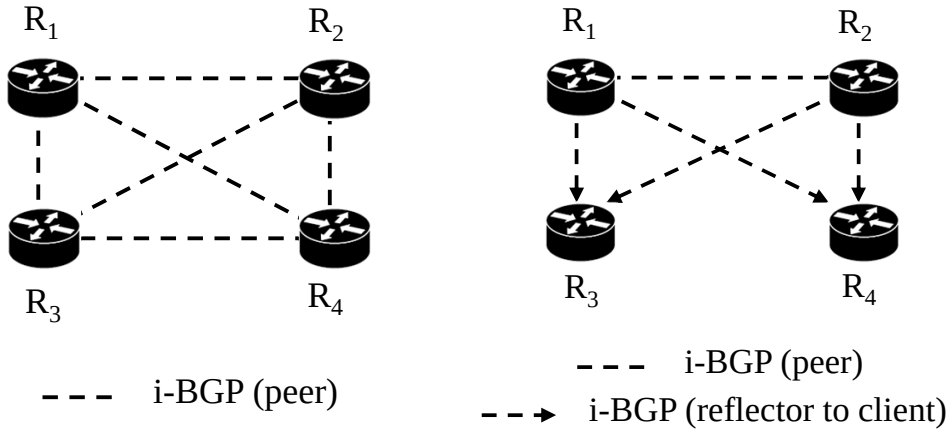
- We define, quantify, and characterize i-BGP convergence to provide the first quantitative assessment of i-BGP convergence of all prefixes in the global routing table from the view of two large ISPs (Section 4.3 ˜Section 4.5). We observe from both ISPs that the majority of routing dynamics inside

an ISP are either local (i.e., observed only at one particular POP) or AS-wide (i.e., observed in all POPs inside the AS) in their scale. Local events are mostly caused by local link failures and recoveries at different locations, which happen independently inside the studied ISPs and have a convergence duration with less than 1 second. On the other hand, events that affect all routers (i.e., AS-wide events) take much longer time to converge, caused mostly by delayed arrivals of external update messages at the studied ISPs.

- As a first step to understand the impact of increasingly complex i-BGP architectures and interconnections on i-BGP convergence, we perform several case studies to quantify additional delays caused by hierarchical route reflection architecture (HRR). Our results indicate that, although HRR introduces additional convergence delays, they are insignificant in most cases, and can be further mitigated by carefully engineered i-BGP topologies (Section 4.5.4).

- ISPs typically collect i-BGP data for monitoring and diagnosis purposes. Some ISPs collect i-BGP data by configuring a collector as a client to i-BGP routers, and others configure the collector as a peer with other i-BGP routers. In the latter case, the peering routers do not always send updates to the collector when their best path changes[1], making it difficult to examine the complete routing changes of individual peers. As part of our work in quantifying and characterizing i-BGP convergence, we introduce a geo-based BGP best selection inference that approximates the complete routing behavior of peering routers, using i-BGP data collected by a collector which is a member of i-BGP full-mesh (Section 4.4.5). We make our implementation publicly available [Par11], which may be useful for the future research,

[1]This is due to the design of i-BGP that an i-BGP router does not forward reachability information learned from other i-BGP routers.



(a) Full-mesh i-BGP (b) i-BGP with route reflection

Figure 4.1: Different i-BGP topologies

as well as to the ISPs who wish to quantify their i-BGP convergence using i-BGP data collected over peering sessions.

This chapter is organized as follows. In Section 4.2, we provide necessary background for this chapter, including a briefing on different i-BGP architectures and their basic operations. Section 4.3 defines i-BGP convergence and describes a number of metrics which we use to characterize the i-BGP convergence. Section 4.4 describes the data sets used in this study, how we process the collected data to identify events inside an ISP, and how we classify the identified events into different types. Section 4.5 presents our results on the i-BGP convergence characteristics and the impact of different i-BGP architectures on BGP convergence. Section 4.6 discusses the ramifications of our observations and discoveries. In Section 4.7, we briefly talk about related works, and finally in Section 4.8 we summarize our work and conclude.

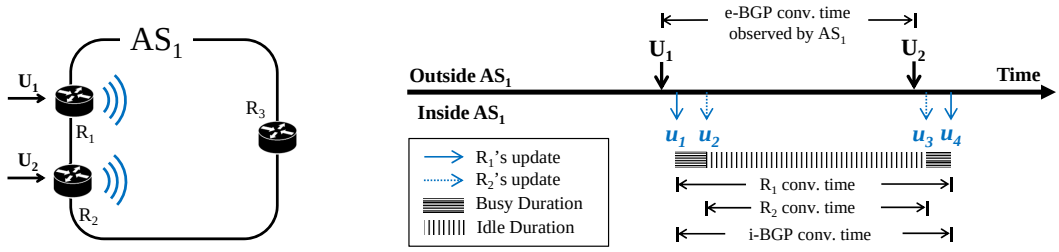
4.2 Additional Delays caused by Route Reflection

In route reflection architecture, routing messages travel more than a single i-BGP hop. For example in Figure 4.1(b), an update message originated at R_3 traverses more than one i-BGP hop (R_1 and R_2 in this case) to reach R_4. Thus, compared to a full-mesh configuration where R_2 would have communicated directly with R_4, route reflection introduces two additional delays in update propagation. First, the update has to go through a potentially longer physical path through either R_1 and R_2. Second, there is an additional processing delay at each BGP hop, such as BGP best path selection and routing loop detection.

Besides the increased delay caused by a longer physical path, creating hierarchies in an i-BGP topology also introduces multiple parallel paths to a given destination. For example, in Figure 4.1(b), R_3 can see three possible paths to reach a destination announced by R_4: (1) R_3–R_1–R_4, (2) R_3–R_2–R_4, and (3) R_3–R_1–R_2–R_4. Thus when the destination becomes unreachable, R_3 will explore all the possible internal paths before converging to the unreachable state. Had all the routers been connected in a full-mesh, R_3 would have only one path to reach it and the convergence could potentially be faster.

4.3 Defining I-BGP Convergence

In this section, we define i-BGP convergence and describe three metrics which we use to characterize the i-BGP convergence in detail.



(a) External update arrivals (b) i-BGP Updates Observed by Internal Routers

Figure 4.2: I-BGP convergence

4.3.1 I-BGP Convergence

We define i-BGP convergence as the process that all i-BGP routers, communicating over i-BGP sessions inside a single AS, as opposed to e-BGP convergence which considers the Internet-wide convergence, settle down to their best path after a routing information change to reach a given destination prefix. Different from the previous works which measure per-router view of convergence, we measure AS-wide convergence in the aggregated view of all i-BGP routers inside an AS.

4.3.2 Evaluating I-BGP Convergence: Metrics

We use three metrics to characterize i-BGP convergence in this chapter, namely (1) convergence duration, (2) number of updates, and (3) number of explored paths.

4.3.2.1 Convergence Duration

The convergence duration is the time that takes for routers to settle down to the next available best path after a routing information change and is directly related to the packet forwarding performance. In this work, we compute the convergence

duration for a given routing change as the relative time difference between the last update message and the first update message generated by all routers inside the AS for the given routing information change, and use it as one of our metrics to characterize the convergence.

We use Figure 4.2 as an example to explain how we compute the convergence duration. In this example, two external updates (U_1 and U_2) arrive at AS_1 through the border routers R_1 and R_2. Upon receiving these external updates, R_1 and R_2 further propagate this routing information inside AS_1 by sending the i-BGP update messages. The BGP routers inside AS_1 learn about the routing information change and decide whether they should change their best paths or not. In this particular example, the convergence duration for R_1 is time(u_4) - time(u_1), and the convergence duration for AS_1 is time(u_4) - time(u_1). This example is a special case, because the i-BGP convergence duration is equal to the router convergence duration of R_1. The reason for this is that both the first and the last update in this convergence event are generated by R_1.

Busy vs. Idle Durations: During a given i-BGP convergence, one or more external update messages may arrive at the receiving AS at different times, because external update messages are likely to traverse different physical path from the routing event origin to the receiving AS. When an external update message arrives, the received routing information will be distributed inside the AS in the form of i-BGP update messages, creating an i-BGP update burst (i.e., update churn). For a given i-BGP convergence, many external update messages may be received, and therefore, the convergence process can be considered as a series of i-BGP update bursts that happen upon each arrival of external updates. If the inter-arrival times of the external updates is longer than the duration of the update churn, there will be times in which the routers are idle in terms of the

number of updates for the given convergence. To examine the extent of this idleness during a convergence event, we divide the event duration into two types: (1) *busy duration*: the routers are busy creating the churn and settling down therefore have at least one update within a second, and (2) *idle duration*: the routers have already settled down and have no update within a second. Figure 4.2(b) shows an example of busy and idle durations.

4.3.2.2 Number of Best Path Changes

The number of best path changes of a given router is one of the dominant contributors on its processing load, and we use it as one of our metrics to characterize i-BGP convergence. In [WZP02], Wang et al. shows that an excessive amount of router load can lead to session resets, routing loops, and packet losses.

In the case that i-BGP update messages is collected using i-BGP server-client sessions, we can simply count the number of generated update messages to compute the number of best path changes made by a given router. However, if i-BGP update messages are collected by a collector which is a member of the i-BGP full-mesh, not all best path changes are visible from the collector's view, and has to be inferred. Later in Section 4.4.5, we describe a technique to infer the number of best path changes using i-BGP data collected over peering sessions. Note that the number of routers in the two studied ISPs differ and for comparison purposes, we compute the average number of best path changes per a router instead of the aggregated number in this chapter.

4.3.2.3 Number of Explored Internal and External Path

Every BGP update message contains reachability information, along with the path information on how to reach the destination. In e-BGP, the path typically

refers to the external path information recorded in NEXT_HOP and AS_PATH attributes, where NEXT_HOP is the next-hop router and AS_PATH is the AS-level path to reach the destination. i-BGP introduces internal (RR or Sub-AS) paths as briefly described in Section 4.2. To avoid ambiguity, we define *external path* as the external path information recorded in NEXT_HOP and AS_PATH attributes, and *internal path* as the internal (RR) path information recorded in CLUSTER_LIST (or AS_CONFED_SEQ) attribute. Note that throughout the chapter, when we say path without further specification, we mean the overall path (internal path + external path).

For a given i-BGP convergence, one may observe different number of internal and external paths explored. The number of external paths represents the number of external path learned by the AS to reach the destination from the egress point of the AS. On the other hand, the number of observed internal paths for a given external path represents the number of internal paths which an update message with the external path information traversed to reach the receiving router from the border router which initially injected the external update into the AS. Therefore, the number of internal paths for a given external path is the amount of i-BGP path explorations, happened for the given external update message, as it is injected and propagated to the routers inside the AS. It is worth mentioning that a more scalable i-BGP architecture, such as route reflection or AS confederations create a larger number of internal paths between i-BGP routers, and can potentially generate relatively more i-BGP updates compared to a full-mesh for a given external path. This potential update inflation caused by internal path exploration has been a concern of large ISPs which adopted a more scalable i-BGP architecture.

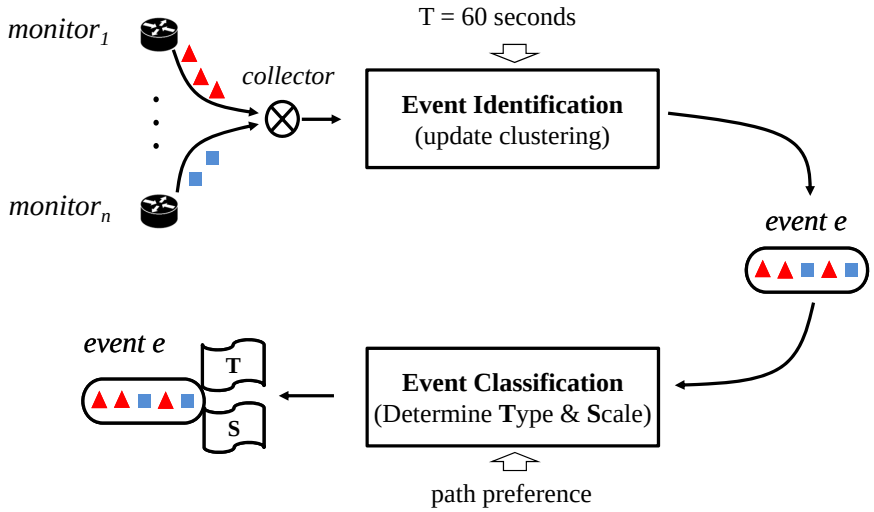
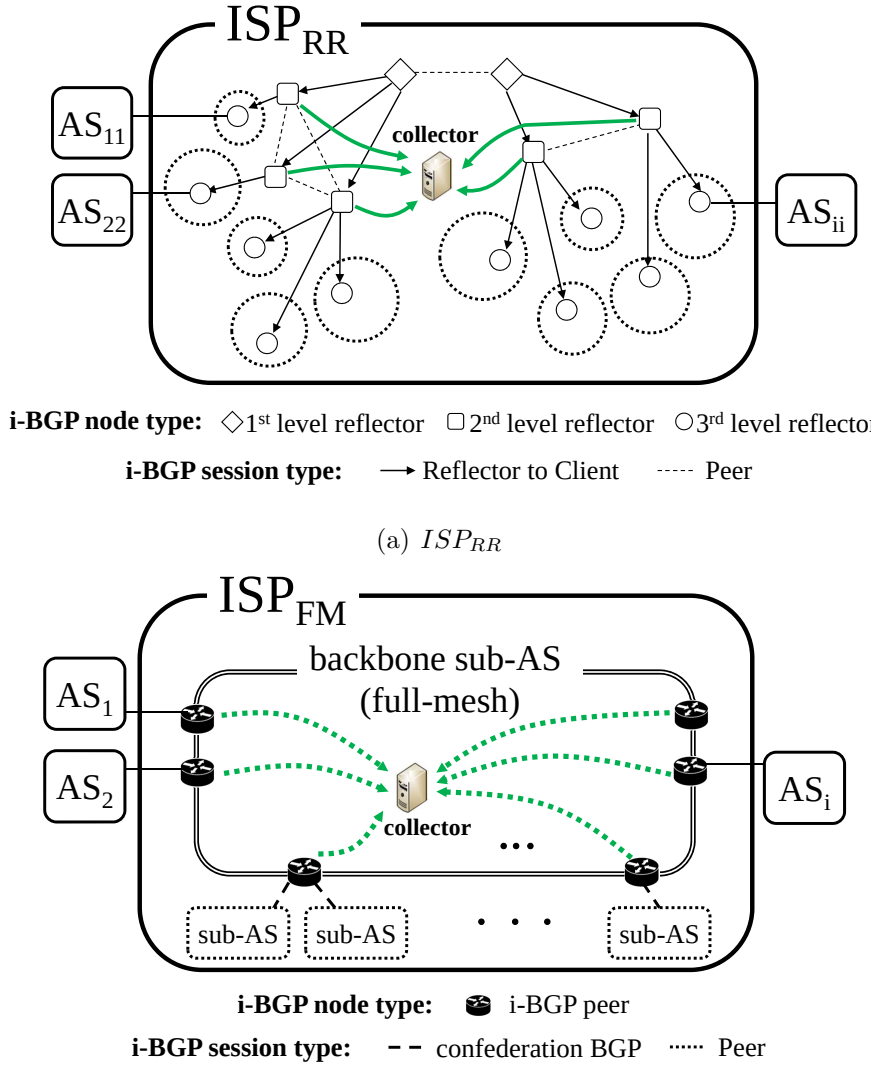


Figure 4.3: High level data processing

4.4 Measuring I-BGP Convergence

We used i-BGP data collected from two large ISPs, ISP_{RR} and ISP_{FM}, named after their i-BGP architecture in their backbone routing infrastructure. In this section, we describe the high level network topology of the two ISPs, and how we identify and classify routing events using the collected i-BGP data. Figure 4.3 depicts the high level view of our data collection and processing, which we explain in detail in this section.

Our methodology may be considered as a pastiche of previous approaches in that we fully utilize, whenever appropriate, the existing techniques, such as timer-based update clustering and inferring path preference based on path-usage time, proposed and validated in the previous works on e-BGP dynamics [RWX02, CGH03, FMM04, OZP06, WMR05] to avoid reinventing the wheel. At the end of this section, however, we describe a novel technique to infer the best path changes for a given router connected using i-BGP peering session, which may be helpful in the future research.



(a) ISP_{RR}

(b) ISP_{FM}

Figure 4.4: Simplified i-BGP topology of two ISPs

4.4.1 High Level Description of the two ISPs

4.4.1.1 ISP_{RR}

ISP_{RR} is a large ISP with several hundreds of i-BGP routers distributed across 22 countries in 2 different continents, and built a hierarchical route reflection

architecture by recursively applying route reflection as also described in the previous chapter. To minimize the routing information propagation delay within the network, ISP_{RR} does not use MRAI timer internally. Figure 4.4(a) depicts a simplified hierarchical route reflection system built by ISP_{RR}. The diamond-shape RRs at the top level represent continent level RRs; the square-shape RRs are at the 2^{nd} level of the hierarchy, each represents a regional RR, and the 3^{rd} level circle-shape RRs represent Points of Presence (POPs).

ISP_{RR} uses the top two levels of route reflectors for the sole purpose of distributing routing information to the rest of the network. We refer to this route reflector infrastructure in the upper two (1^{st} and 2^{nd}) levels of their route reflection hierarchy as backbone routers in ISP_{RR}. The collector connects to all 2nd level route reflectors to passively collect i-BGP updates. Note that ISP_{RR} uses server–client sessions when collecting i-BGP updates. In such configuration, the route reflectors send updates whenever their best paths change to the collector.

4.4.1.2 ISP_{FM}

ISP_{FM} is another large ISP with several hundreds of i-BGP routers distributed across 14 countries in 3 different continents, and uses AS confederations [TMS07] to scale with its network size as described in the previous chapter. As in the case of ISP_{RR}, ISP_{FM} does not use MRAI timer inside its network. Figure 4.4(b) shows a simplified topology of ISP_{FM} at a high level, where *backbone sub-AS* represents the backbone network of this ISP, consisting of more than one hundred i-BGP routers connected in a full-mesh (hence referred to as ISP_{FM}). In contrast to ISP_{RR}, ISP_{FM} configured the collector as a member of the full-mesh in its backbone routing infrastructure to passively collect i-BGP updates. In such configuration, the collector does not know all best path changes of the peers and

has to be inferred. Later in this chapter we explain in more detail why the best path changes are not visible from the collector and how we infer the best path changes in Section 4.4.5.

4.4.2 Data Collection and Preprocessing

In most of BGP data collection projects including Oregon RouteViews [Uni] and RIPE RIS [NCC], a collector (an i-BGP router) is used to set up BGP sessions with target routers (which we call *monitors* throughout the chapter) and to passively record BGP data sent from the monitors in MRT [mrt10] format. Similarly, we used a collector configured in both ISP_{RR} and ISP_{FM} to maintain i-BGP sessions with all monitors in the backbone routing infrastructure as shown in Figure 4.4.

In ISP_{RR}, a collector is configured to maintain i-BGP *server-client* sessions to 18 route reflectors in the 2nd level and to passively record all i-BGP updates received during one year from May 2009 to April 2010. In ISP_{FM}, a collector is configured as one of the i-BGP peers in backbone sub-AS, maintaining i-BGP *peering* sessions with 133 monitors in backbone sub-AS to passively record all i-BGP updates observed. Because of the peering session type of which a path learned from other i-BGP peering sessions is not forwarded, the collector has a limited view of best path changes in other peering monitors. Both ISP_{RR} and ISP_{FM} deployed more monitors in larger POPs. To avoid a potential bias towards large POPs with relative more monitors deployed, we select just one monitor for a given POP. The total number of selected monitors are 17 and 28 from ISP_{RR} and ISP_{FM} respectively.

BGP routers start their sessions by initially exchanging the whole routing table. To avoid identifying such table transfers as routing events, we identify the

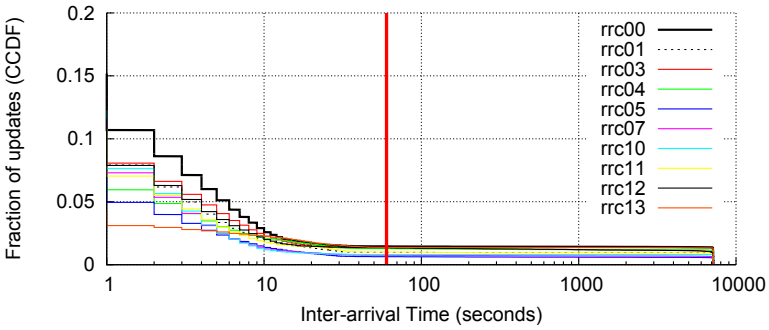
table transfers based on the BGP session state messages recorded together with the update messages by the collector and remove them out from our data. Additionally, we remove pure duplicate BGP update messages and update messages on internal prefixes and potential bogon prefixes that have prefix length smaller than 8 or greater than 24. The number of such prefixes is less than 5% of all prefixes.

4.4.3 Event Identification

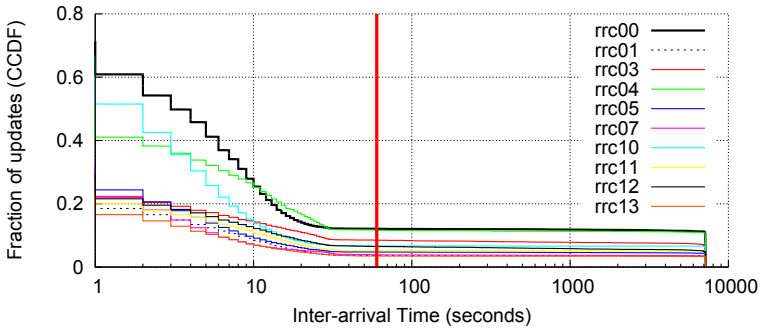
A number of previous BGP data analytic studies [RWX02, CGH03, FMM04, OZP06, WMR05] developed timer-based approaches to cluster routing updates into events. The intuition behind these approaches is that BGP updates often arrive in bursts. The two consecutive updates for a given prefix are assumed to be generated by the same routing event if they fall within a time interval threshold.

Oliveira et al. [OZP06] calculate the inter-arrival times of updates generated by BGP beacon prefixes [MBG03], announced from different topological locations in the Internet, and empirically determine the time threshold T. Because the root cause of each beacon event is known and the updates do not contain noise after preprocessing, we also use this approach to determine the time threshold. However, we make one slight modification: we cluster updates in the aggregated view of all monitors, as opposed to the view of a single monitor. Thus in our work, we modify the approach used in [OZP06] such that the inter-arrival times are calculated between two updates generated by all monitors inside the given AS.

Figure 4.5 shows the distribution of update inter-arrival times of the 10 beacon prefixes as observed from the 17 and 28 monitors inside $ISP_{RR}$and ISP_{FM} respectively. All the curves become flat before or at around 60 seconds (the



(a) ISP_{RR}



(b) ISP_{FM}

Figure 4.5: Inter-arrival Times of 10 Beacon Prefix Updates Observed Inside the two ISPs

vertical line on the figure). Based on this observation, we use T = 60 seconds as the inter-arrival time threshold when grouping updates into different events. Because the beacon prefixes are announced and withdrawn at a fixed interval of 7200 seconds, the tail drop of all the curves is at 7200 seconds as expected.

4.4.4 Event Classification

After we identify an event by clustering the update messages based on a time threshold, we classify each identified event by a different scale and type, based on the fraction of affected monitors inside the network and how the path changed after the event.

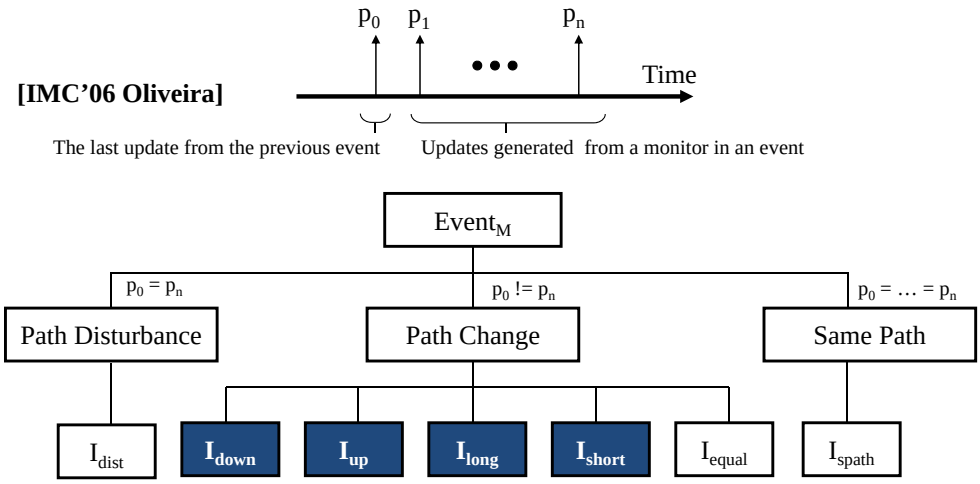


Figure 4.6: Event classification

4.4.4.1 Event Scale

After identifying an event, we determine the scale of the given event based on the fraction of monitors that are affected by this event. We define the scale S_e of a given event e as

$$S_e = \frac{mon_e}{mon_n} \tag{4.1}$$

where mon_e is the number of monitors affected (i.e., with at least one best path change) by the event e and mon_n is the total number of monitors. We define two special cases of event scale, namely *local* and *AS-wide*. In the case when $mon_e = 1$, we classify the event as a *local* event. In contrast, if $S_e = 1$, we classify the event as a *AS-wide* event.

Type	Descrption
I_{up}	A previously unreachable destination becomes reachable by the end of the event
I_{down}	A previously reachable destination becomes unreachable by the end of the event
I_{short}	The best path changes to a more preferred path by the end of the event (recovery)
I_{long}	The best path changes to a less preferred path by the end of the event (failover)
I_{spath}	One or more updates are generated and in all updates, the path does not change. These updates typically differ only in MED and COMMUNITY attributes,indicating that the internal BGP dynamics inside the monitors AS.
I_{pdist}	One or more updates are generated and in at least one update, the path is different. I_{pdist} events are likely to be resulted from multiple root causes, e.g., a transient failure which is followed quickly by a recovery, hence the name of the event type.
I_{equal}	The best path changes to anther path with equal preference

Table 4.1: Event types

4.4.4.2 Event Type

A number of previous works [OZP06, LAW01] define different event types for a given routing event. To avoid confusion, we use the consistent definitions of event types. Figure **??** shows different events types based on the paths changes before and after a given event. An event is divided into 3 different types at the highest level based on how the paths changed during the event. Table 4.1 lists the different event types along with a brief description.

I_{up} and I_{down} events are relatively easier to classify since one can identify them by looking at whether the prefix was reachable or not reachable in both the previous and current event. For events that involve path changes, it is necessary to compute and compare the preference of the path used before and after the event when classifying the event into one of I_{long}, I_{short}, and I_{equal}. This task can be challenging since many factors that determine the preference of a path, such as policy, is not visible from the observing monitor. In this work, we use usage-based path preference heuristic proposed in [OZP06] to infer the preference of a path. The basic intuition of the heuristic is that if a path is preferred over

another path for any reason, the observed usage time of the more preferred path will be greater than that of the less preferred one. The underlying assumptions are (1) both paths are available most of the time, and (2) the preference of the paths does not change during the measurement period. Note that, often, I_{spath} and I_{pdist} contain updates generated from more than one event (e.g., an active prefix with its reachability information changing very frequently), and the quantification results for the two events may not be very meaningful. Thus, we omit them from further analysis when we present our results.

Event Type Consistency: Given there are multiple monitors inside each ISP_{RR} and ISP_{FM}, it is possible that the event type identified by different monitors for a given routing event do not agree. For example in an event observed by two monitors, one monitor can identify the event type as I_{spath}, whereas the other monitor identifies the same event as I_{pdist}. In the case that the events types do not agree, we classify the event as inconsistent event. The inconsistent events are mainly caused by the limitation of timer-based update clustering approach, which cannot always cluster updates into events accurately. We observe that the overall fraction of inconsistent events ranges widely from as little as 2% up to as large as 10% of all events identified across different months. Our further investigation reveals that the inconsistencies caused by two factors: (1) the inaccuracy of the timer-based update clustering when two or more events are mistakenly clustered as one event and (2) by the inaccuracy of inferring the path preference purely based on the path usage-time without considering the path availability. In this chapter, we simply do not consider these inconsistent events and remove them from our further analysis for clarity. However, we believe that being able to accurately identify events and their types is important and leave this part as one of the future research directions.

4.4.5 Geo-based Best Path Selection Inference

4.4.5.1 Motivation

For the purpose of monitoring and diagnosis, ISPs often set up a collector to maintain i-BGP sessions with a set of monitors and passively collect i-BGP data. There are mainly two types of i-BGP sessions used: *server-client* and *peering* sessions.

A collector can be configured as a client of a route reflector and receive all best path changes of the route reflector (as in the case of ISP_{RR}). In this case, the amount of i-BGP data to be stored can be large. The other option is to deploy a collector as a member of i-BGP full-mesh (as in the case of ISP_{FM}). In the latter case, due to the i-BGP full-mesh update forwarding rule that prevents an i-BGP router from sending reachability information learned from other i-BGP routers to any other i-BGP routers in the full-mesh, the peering router does not send its best path changes if the path is learned from other i-BGP routers in the same full-mesh.

For example in Figure 4.1(a), assume that R_3 is the collector that maintains peering sessions with all other monitors in the full-mesh. Also assume that a prefix is initially reachable via two monitors, R_1 and R_2, and R_4 is using the path learned from R_2. However, when the path via R_2 fails, R_4 fails-over to the path learned from R_1. In this example, the best path changes to reach this prefix in R_4 is not visible by R_3 as in the original full-mesh i-BGP, because the paths learned by other i-BGP monitors are not forwarded and therefore not visible in the collected i-BGP data. Given only these partial information received by the collector, it can be challenging to understand the complete picture of each individual peer's routing behavior, including the best path changes made by the

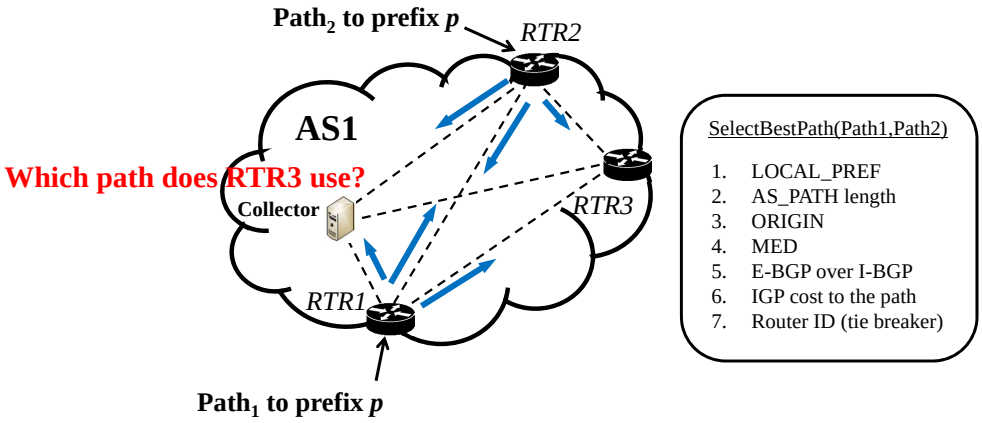


Figure 4.7: Geo-based best path selection inference

router.

4.4.5.2 Inferring the Best Path Selection in Peering Routers

The basic intuition behind the inference is that a monitor prefers the closest path in terms of IGP distance, when there are multiple equally preferred paths at the BGP level, as specified by the BGP best path selection algorithm. For every event, we store the following two pieces of information: (1) the list of announced (thus, equally preferred) paths before and after the event and (2) geographical locations of the monitors that announced each path in (1). If the nearest path for a given monitor r does not change after the event, then r is simply not affected by this event. On the other hand, if the nearest path changes after the event, then we assign r's new best path to one of the available nearest path.

Ideally, inferring the closest path using the actual IGP distance would yield the most accurate inference results. However, such IGP distance reveals a detailed data about the internal physical network topology of the ISP and was not available at the time of our measurements. Therefore in this work, we use geo-

graphical location of monitors instead, to approximate the IGP distance values. For example in case of i-BGP full-mesh as shown in Figure 4.7, assume that there are two paths to reach prefix p, announced by RTR1 and RTR3 respectively. The collector knows that there are two paths to reach prefix p but cannot find out, just by looking at the announcements, which path RTR2 will use as its best path because given the equally preferred paths, RTR2 will start considering the topology-dependent attributes which are only known to RTR2 and not available to the collector. So in our work in the case of ISP_{FM}, we infer the best path selection for each of the routers in the full-mesh based on geographical location of the router, since the topology dependent attribute tends to agree with the geographical distance. We confirmed with the operators in ISP_{FM} that in general the IGP distance cost matches with the physical distance between the monitors.

4.5 Quantification and Analysis Results

In this section, we present our quantification results on the i-BGP convergence as defined in Section 4.3. We first show the total number of identified events over 14-month period from May 2009 to June 2010. Then, we pick the most recent month (June 2010) to understand the convergence in more detail. Finally through several case studies, we study a number of additional convergence delays caused by more scalable i-BGP architectures such as hierarchical route reflection.

4.5.1 Number of Identified Events in Time

Figure 4.8 shows the number of identified events[2] from both ISP_{RR} and ISP_{FM} during the whole studied period. We make a number of interesting observations.

[2]The total number of events in ISP_{FM} during one month of September 2009 is omitted because the i-BGP data were not available during the month.

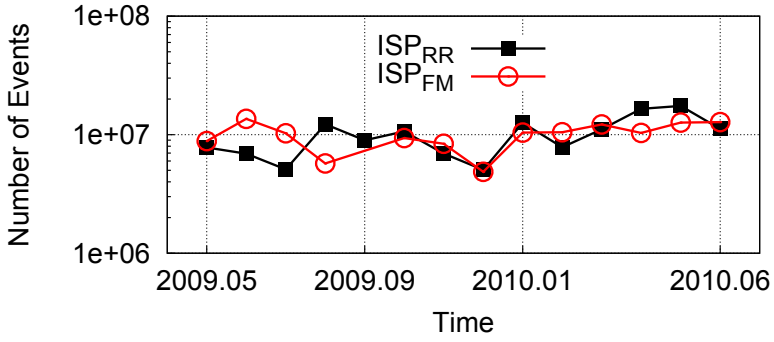


Figure 4.8: Number of identified events from May 2009 to June 2010

First, although the two ISPs have a very different i-BGP architectures, the number of overall events is comparable. Second, the number of events fluctuates in time inside both ISPs, and the fluctuation shows a similar pattern with the lowest number of events during the summer (July or August) and the winter (December). We further investigate what causes the total number of events to fluctuate widely in time and find that the number of events that affect the whole AS (i.e., AS-wide events) stays more or less the same throughout the 14-month measurement time period. However, the number of local routing events varies widely in time and is identified as the main cause behind the fluctuation observed. Lastly, although examining a longer period of time would be necessary to make a general statement about the trend, the number of overall events seems to be gradually increasing in time. The increasing number of overall events may be due to the fact that we define an event per prefix and that the number of prefixes in the global routing table increases in time [Hou]. From both ISPs, we observe about 12% and 10% increase in the total number of prefixes during the 14 months.

4.5.2 Characterizing i-BGP Convergence

To understand the characteristics of i-BGP dynamics and convergence in more detail, we choose the last month available from our dataset (June 2010) and present our results in terms of the metrics we introduced in Section 4.4.

4.5.2.1 Event Scale

Figure 4.9 shows the distributions of event scale (S_e) of all identified events from both ISP_{RR} and ISP_{FM}. We commonly observe from both ISPs that the majority of events are either local (i.e., involving only one monitor) or AS-wide (i.e., involving all monitors) and that the number of local events are a few times greater than the number of AS-wide events. This observation that i-BGP routing events have a small scale in most cases is consistent with e-BGP property that most e-BGP routing events are confined to a small scale [LPR08].

Given the majority of events are local in their scale, we further investigate the local events based on the monitor, which observes a given event to examine how the overall number of local events are contributed by different monitors. Figure 4.10 summarizes our results. A common observation across the two ISPs is that almost all monitors observe local events, contributing to the overall number of local events. However, some monitors observe more local events than others, and the contributing amount can be quite different amongst monitors. Although the two ISPs show a similar distribution, the geographical locations of the top 5 busiest routers with the most number of events do not overlap across the two ISPs and seem to be independent with each other. We observe that the high number of local route changes happens due to a set of local link failures and recoveries to another large neighboring AS. This confirms that the speculation made in [EKD10] that the BGP update churn observed from outside a large ISP

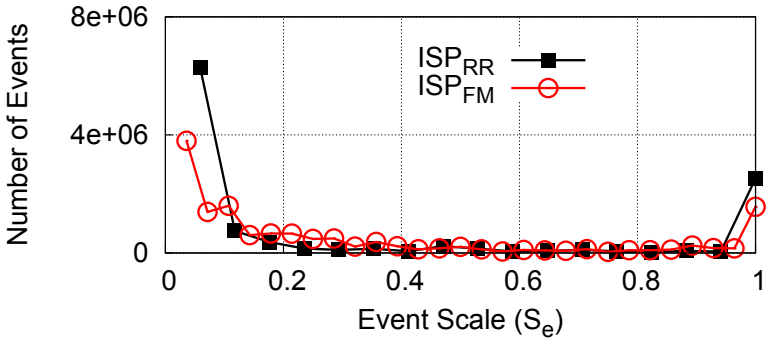
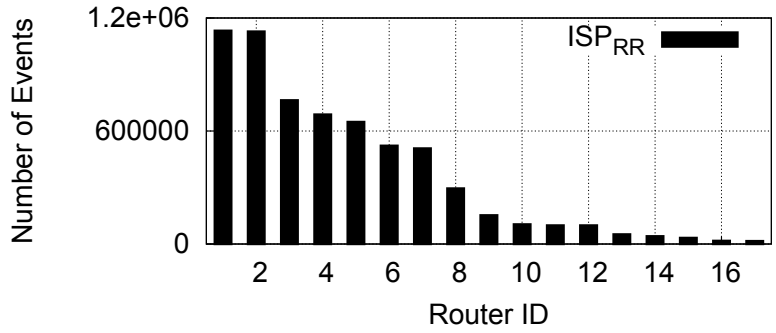
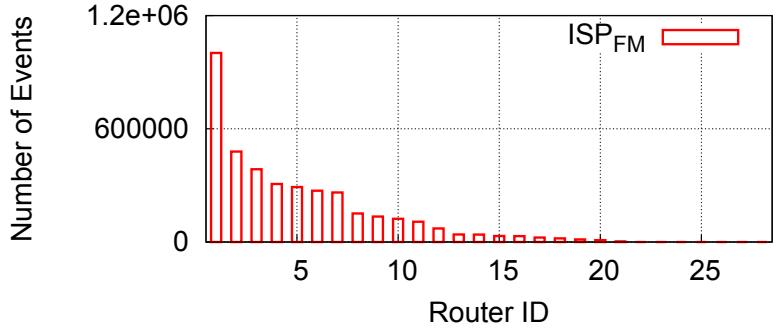


Figure 4.9: Event scale during June 2010



(a) ISP_{RR}

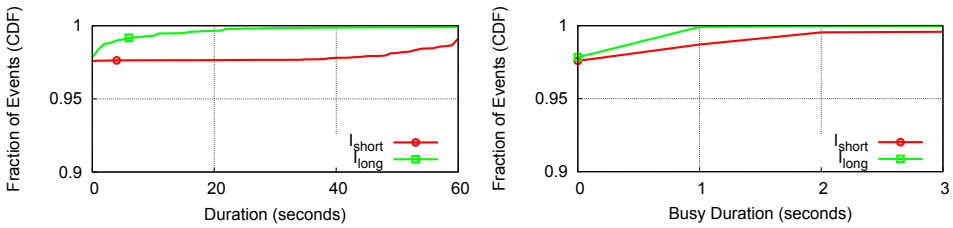


(b) ISP_{FM}

Figure 4.10: Number of local events per router

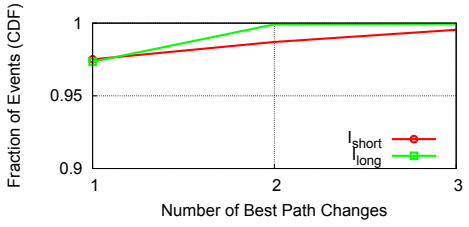
can be due to uncorrelated and distributed local routing events across different locations.

4.5.2.2 Local Events

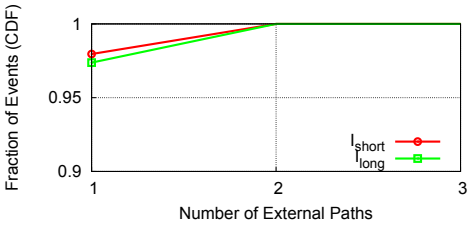


(a) Overall Duration

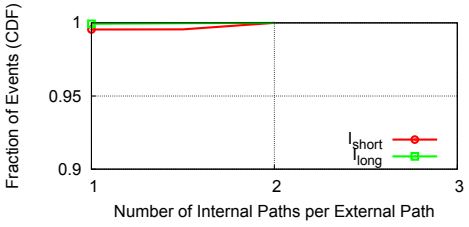
(b) Busy Duration



(c) Number of Best Path Changes



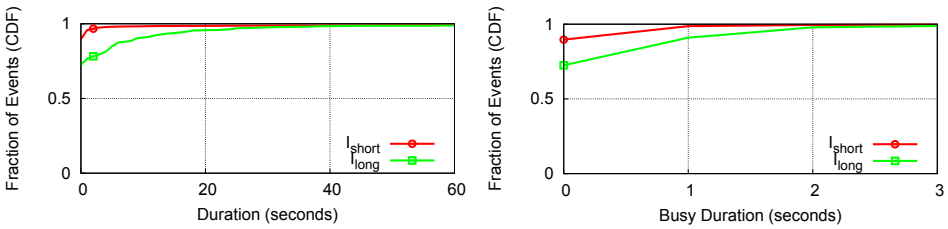
(d) Number of Explored E-Paths



(e) Number of Explored I-Paths/E-Path

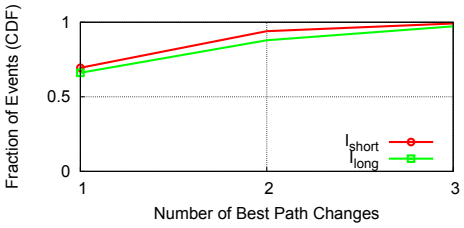
Figure 4.11: Local events convergence in ISP_{RR} during June 2010

Table 4.2 shows the total number of local events identified from ISP_{RR} and ISP_{FM} during the month of June 2010. Figure 4.11 and Figure 4.12 summarize the characteristics of the local events inside ISP_{RR} and ISP_{FM} respectively, using the three metrics we introduced earlier in Section 4.3. Ideally, I_{up} or I_{down} events should have AS-wide scale and should not be observed, as in the case in ISP_{FM}. In the case of ISP_{RR}, we checked that the identified local I_{up} and I_{down} are in fact AS-wide I_{up} or I_{down} events, but incorrectly broken into two separate events by the timer-based update clustering technique. Because the fraction of such false

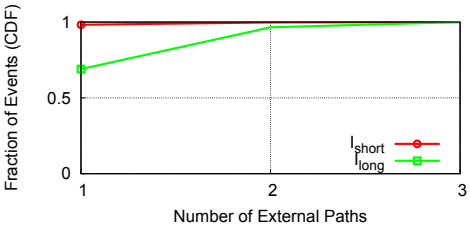


(a) Overall Duration

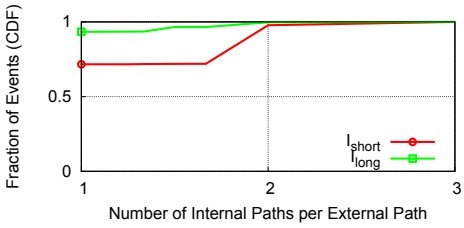
(b) Busy Duration



(c) Number of Best Path Changes



(d) Number of Explored E-Paths



(e) Number of Explored I-Paths/E-Path

Figure 4.12: Local events convergence in ISP_{FM} during June 2010

Types	ISP_{RR}	ISP_{FM}
I_{up}	23,627 (0.26%)	0 (0%)
I_{down}	23,732 (0.27%)	0 (0%)
I_{short}	1,265,395 (14.33%)	959,599 (7.51%)
I_{long}	1,199,760 (13.58%)	920,143 (7.20%)
I_{pdist}	126,268 (1.43%)	461,513 (3.61%)
I_{spath}	1,777,465 (20.12%)	1,148,943 (8.99%)

Table 4.2: Number of local events in ISP_{RR} and ISP_{FM} during June 2010

positive local events is small enough (0.27% of overall events), we believe that the generality of our results is not be affected. In this section, we simply do not consider these local I_{up} and I_{down} events in our analysis.

From Table 4.2 and Figure 4.11 and 4.12, we make a number of common observations on local I_{short} and I_{long} events from both ISPs.

First, the number of I_{short} events roughly matches with the number of I_{long} events, indicating that a failed link is eventually recovered within the one month time period. The overall convergence process of these two local events is quite simple; the majority of I_{short} and I_{long} events (more than 97% and 72% in ISP_{RR} and ISP_{FM} respectively) have convergence duration of less than one second and generate only one update message.

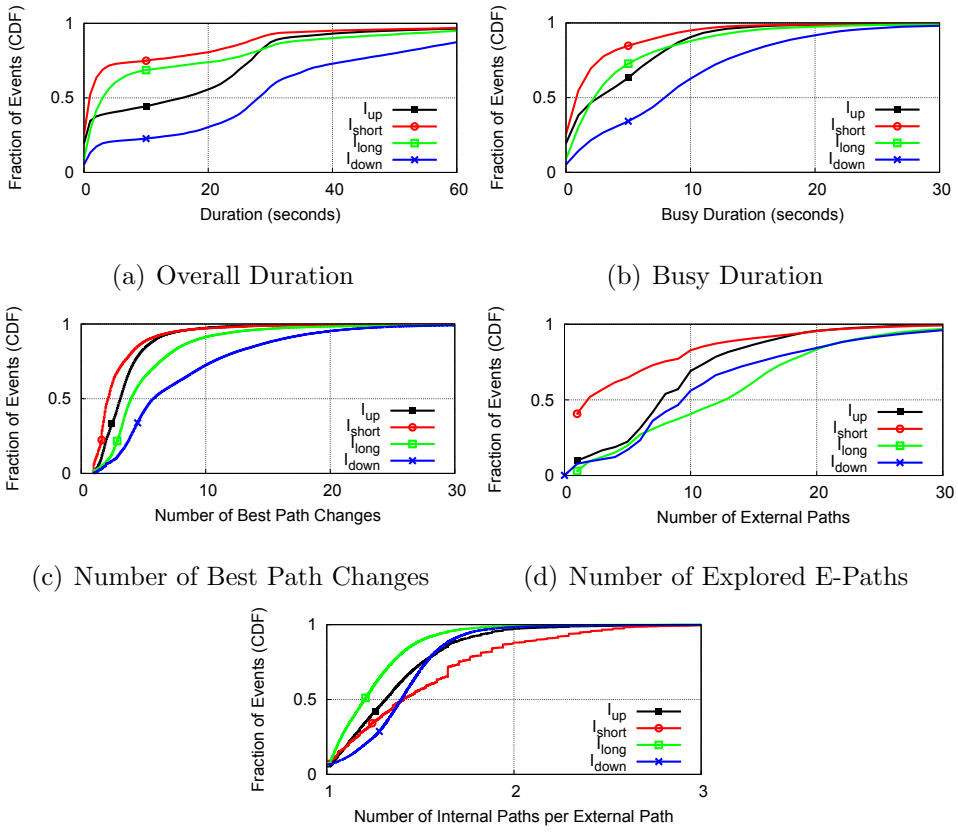
Second, when the local events have the duration with more than one second, the duration time is mostly determined by the idle time gaps between the update messages, and the duration can be large when the two (or more) update messages are separated with one or more large time gaps. In Figure 4.11(c) and 4.12(c), we observe that the number of update messages is less than 3 in almost all the cases, indicating that these relatively large durations are indeed caused by the large idle time gaps.

Third, we observe that a small fraction of local events have their convergence duration greater than a few seconds. These long durations (e.g., the top 2.4% of I_{short} events in ISP_{RR}) with can mostly be explained either by the inaccuracy of the timer-based event clustering technique which grouped updates generated by two or more independent events into one event, or can be attributed to the router processing delay as described in [FKM04].

One major difference between the two ISPs is that in ISP_{FM} there are relatively more events (about 25% of overall I_{short} and I_{long} events) with their dura-

tion spread out from 1 to 30 seconds. We find that this is mostly due to a failure and recovery of a link between ISP_{FM} and a neighbor AS at a particular POP during this specific month. Because ISP_{FM} does not use MRAI timer within its network, we suspect that the delay is due to the MRAI timer used in the routers between ISP_{FM} and the neighbor AS in their e-BGP session.

4.5.2.3 AS-wide Events



(a) Overall Duration

(b) Busy Duration

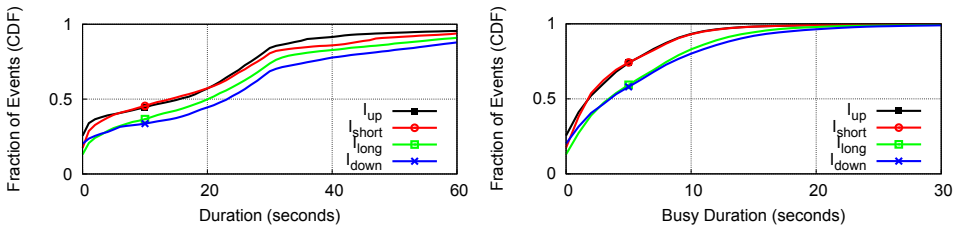
(c) Number of Best Path Changes

(d) Number of Explored E-Paths

(e) Number of Explored I-Paths/E-Path

Figure 4.13: AS-wide events convergence in ISP_{RR} during June 2010

Table 4.3 shows the total number of identified AS-wide events from ISP_{RR} and ISP_{FM} during the same one month of June 2010. Figure 4.13 and Fig-



(a) Overall Duration

(b) Busy Duration

(c) Number of Best Path Changes

(d) Number of Explored E-Paths

(e) Number of Explored I-Paths/E-Path

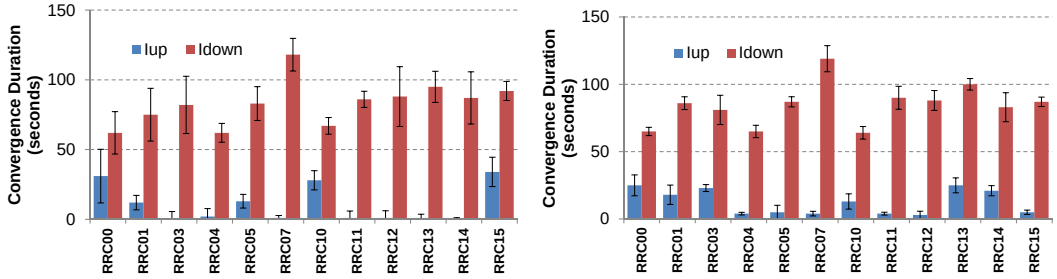
Figure 4.14: AS-wide events convergence in ISP_{FM} during June 2010

Types	ISP_{RR}	ISP_{FM}
I_{up}	222,501 (2.52%)	206,819 (1.62%)
I_{down}	220,105 (2.49%)	187,293 (1.47%)
I_{short}	367,172 (4.16%)	154,442 (1.21%)
I_{long}	375,808 (4.25%)	154,260 (1.21%)
I_{pdist}	1,174,469 (13.30%)	292,567 (2.29%)
I_{spath}	33,231 (0.37%)	257,563 (2.02%)

Table 4.3: Number of AS-wide events in ISP_{RR} and ISP_{FM} curing June 2010

ure 4.14 summarize the characteristics of AS-wide i-BGP convergence using the three metrics we defined. As in the case of local events, we observe that the number of I_{up} events roughly matches with the number of I_{down} events. Also, the number of I_{short} events matches with I_{long} events. We further make a number of common observations from both ISPs. First, there is a group of events with their duration less than 1 second. Our further investigation reveals that the convergence duration is closely related with the number of paths from the measurement ISP to reach a given prefix. If the number of paths to reach a given prefix is low (e.g., one path), the convergence duration is less than or near 1 second. Second, we observe in Figure 4.13(a) and Figure 4.14(a) that a large number of events have their convergence durations near the default e-BGP MRAI timer value (i.e., 30 seconds). Also, the busy durations shown in Figure 4.13(b) and Figure 4.14(b) are relatively lower in general compared to the overall durations shown in Figure 4.13(a) and Figure 4.14(a). These two observations indicate that AS-wide i-BGP convergence duration is affected heavily by the external update propagation delay due to the prevalent usage of MRAI timer outside the ISPs and that the routers are mostly idle during a given event.

There is one major difference between the two ISPs. In ISP_{RR}, I_{short} and I_{long} events have the shortest convergence duration, followed by I_{up}, and I_{down}. On the other hand in ISP_{FM}, I_{up} has the shortest convergence duration in general, followed by I_{short}, I_{long}, and I_{down}. This difference in the order of overall AS-wide convergence durations between different events, however, can be explained by the different connectivity to reach a particular destination, as briefly explained above. For example, assume ISP_1 has 1 best path (e.g., a large customer AS) to reach a set of prefixes. If this path becomes unstable and has many I_{long} and I_{short} events that affects the whole AS, the overall AS-wide convergence duration for I_{short} and I_{long} events can be biased towards having an overall short convergence



(a) ISP_{FM} (b) ISP_{RR}

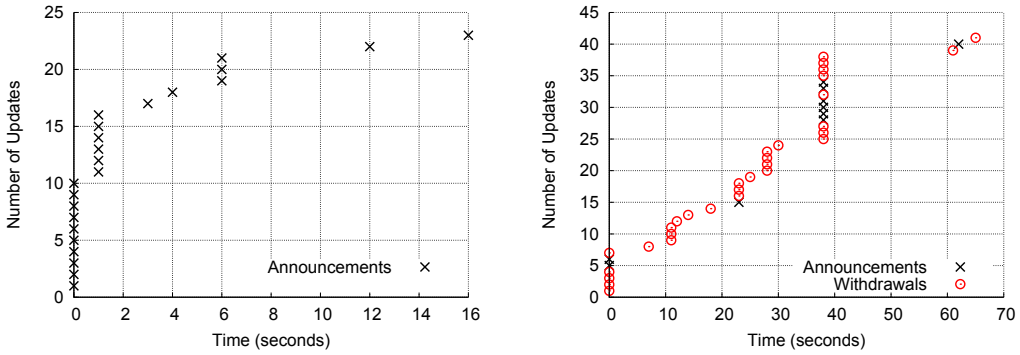
Figure 4.15: Convergence Duration of Beacon Prefixes During June 2010

duration in both I_{short} and I_{long}. We verified that this indeed is the main cause for the observed shorter duration of ISP_{RR}.

4.5.3 i-BGP Convergence of Beacon Prefixes in ISP_{FM} and ISP_{RR}

In this section, we perform case studies on i-BGP convergence duration using beacon prefix events as observed inside the two ISPs. The goal of the case studies is two-fold. First, we seek to have a concrete picture of the convergence inside the two ISPs and understand it in more detail by taking a close-up view. Second, by comparing the overall convergence durations of the beacon prefixes inside the two ISPs, we examine if there is any notable difference in the overall convergence delays, which can potentially be due to the hierarchical route reflection inside ISP_{RR} compared to ISP_{FM}.

Figure 4.15 shows the median convergence duration for I_{up} and I_{down} events during one month of June 2010 of a given beacon prefix from RIPE RIS along with 95% confidence intervals. We make the following two common observations across the two ISPs. First, the convergence duration of I_{up} events for a given beacon prefix ranges from 0 to 30 seconds in general and is always less than the convergence duration of I_{down} events for the same beacon prefix. Second, the



(a) I_{up} (b) I_{down}

Figure 4.16: Updates observed during I_{up} and I_{down} events of RRC00 beacon prefix inside ISP_{FM}

convergence duration of I_{down} events for a given beacon prefix is always greater than 60 seconds, mostly ranging from 60 to 90 seconds.

We use Figure 4.16 showing all i-BGP updates observed inside ISP_{FM} caused by the I_{up} and I_{down} events of RRC00 beacon prefix to explain (1) why the convergence duration for I_{up} events are shorter than that of I_{down} events and (2) the main impacting factors of convergence delays for the I_{up} and I_{down} events in this particular case of RRC00 beacon prefix. In case of I_{up} events, the most preferred paths arrive first at ISP_{FM} before other less preferred paths arrive (mostly due to the shorter physical and topological distance), and once after learning these most preferred paths, the routers in ISP_{FM} no longer make path changes. As a result, the convergence duration is short. The updates carrying the best paths, however, arrive at ISP_{FM} with different delays associated with different paths that the updates travel from the origin of the event to ISP_{FM}. For example, all updates within the first 2 seconds of the convergence are received from AS3257, and all the subsequent updates are received from another neighbor AS, AS6453.

On the other hand in case of I_{down} event, the routers explore paths in the descending order of the path preference as all available paths become withdrawn. Because the update propagation and path exploration within ISP_{FM} is quite fast and mostly under 1 second, we observe from Figure 4.16(b) that there are a few rounds of micro-convergence inside ISP_{FM}. That is, after the best path is withdrawn, routers momentarily settle down to the next best paths that are not yet withdrawn due to the external delays of the updates traveling from the origin of the events to ISP_{FM}. This process of micro-convergence repeats at the interval of 30 seconds (which is the default e-BGP MRAI timer value) until all paths to reach the destination are withdrawn.

Lastly, we make an interesting observation that the durations for I_{up} and I_{down} events are similar for a given beacon prefix across the two ISPs although their external connectivity to reach the beacon prefixes is different. We find that in some cases this is the result of having similar AS-level connectivity to the beacon prefixes from the two ISPs. For example, the beacon prefix from RRC00 (84.205.64.0/24) is reached from both ISPs through AS3257 and AS6453.

Besides the common observations mentioned above, we observe that the overall duration of both I_{up} and I_{down} events are slightly higher in ISP_{RR} for most beacon prefixes. There could be many reasons for this observed difference, and in the next section, we attempt to explain the differences by investigating whether the specifics of hierarchical route reflection deployment inside ISP_{RR} had any impact on this slightly higher convergence delay.

4.5.4 Impact of i-BGP Hierarchical Route Reflection on Convergence

The i-BGP topologies inside large ISPs have evolved over time by creating hierarchies and redundancies, with one question yet to be answered: what is the

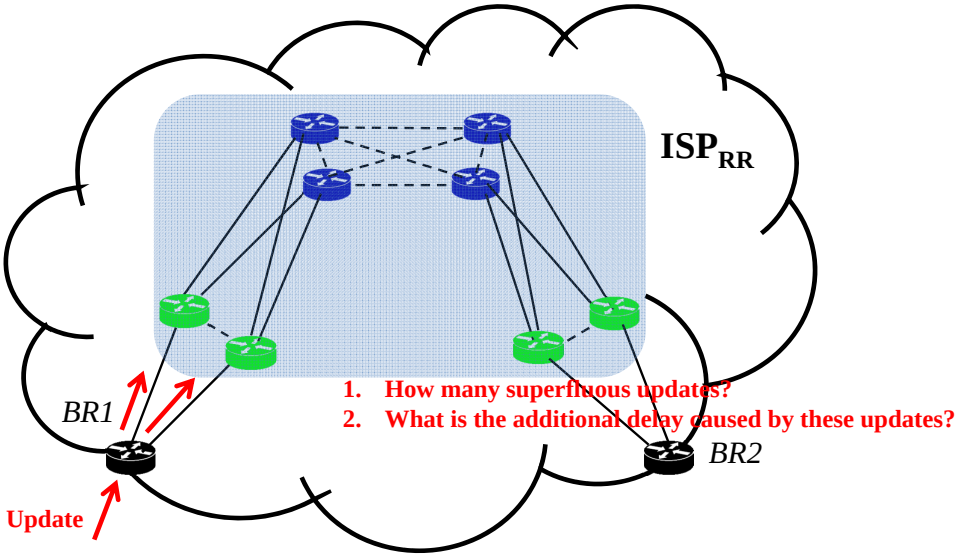


Figure 4.17: An example of superfluous updates in route reflection

impact of the various topologies on BGP convergence inside the network? As a first step to answer this question, we start by comparing the convergence time of the beacon prefixes in the two ISPs to examine if there is a noticeable difference in the overall convergence durations of the beacon prefix events. Then, we identify and study three most intuitive impacting factors that may cause an additional delay in i-BGP convergence, namely (1) superfluous i-BGP updates, (2) physical path stretch, and (3) BGP processing delay, to understand their impact on i-BGP convergence using i-BGP data collected from ISP_{RR}, which uses hierarchical route reflection architecture. Note that BGP processing delay has been studied in the past by Feldmann et al. [FKM04], and therefore in this work, we focus mostly on the first two factors.

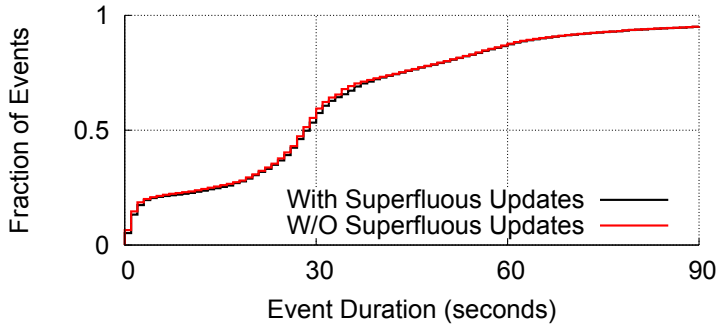
4.5.4.1 Superfluous i-BGP Updates Generated by Internal Path Exploration

Route reflectors are typically deployed in pairs to avoid single point of failure in route reflection.[3] As a result, a client typically connects to two or more redundant route reflectors and receive redundant routing information for any given event. For example in Figure 4.17 when one external update, which will change the best path to reach a given destination is received by the border router and the route reflectors in turn, the same reachability information will be duplicated through different route reflector paths within the route reflection topology. As a result, other border routers can receive multiple superfluous updates carrying the same reachability information. In the case of I_{down} events, until all superfluous update messages are received, other border routers would mistakenly believe that the prefix is still reachable.

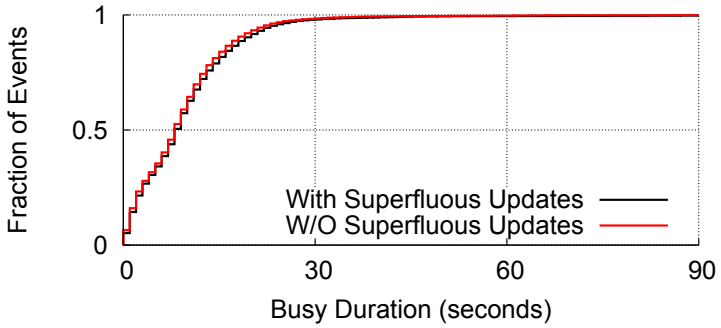
To quantify the extent of this additional delay due to creating redundant control paths, we first identify the update messages that are generated purely due to redundant control path by looking at the two additional BGP attributes that record the originator of the update and the control path used to forward the given update message from the originator to the receiving monitor (ORIGINATOR_ID and CLUSTER_LIST respectively). After identifying such superfluous updates which carry the same reachability information, we filter them out and re-apply our metrics (duration and number of best path changes) to check if there is a noticeable difference, compared to the results we have with the superfluous updates.

Table 4.4 summarizes our results using i-BGP updates collected from ISP_{RR}

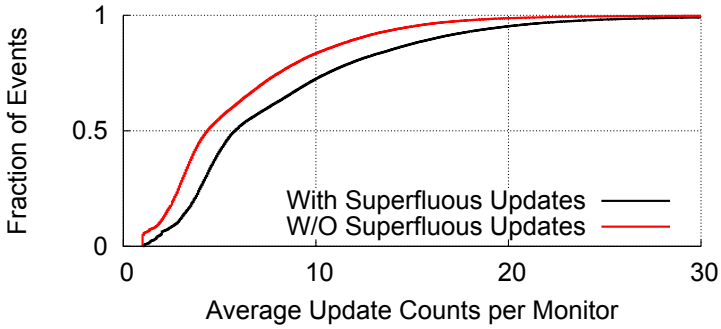
[3]Similarly, sub-ASes in AS confederations maintain more than one connection between each other for the same purpose.



(a) Overall Duration



(b) Busy Duration



(c) Number of Best Path Changes

Figure 4.18: AS-wide I_{down} convergence with and without superfluous updates during June 2010

during one month of June 2010. Across different types of AS-wide events, we observe that there is an increase, but the amount is not significant. Figure 4.18 shows the overall duration, busy duration, and the number of best path changes

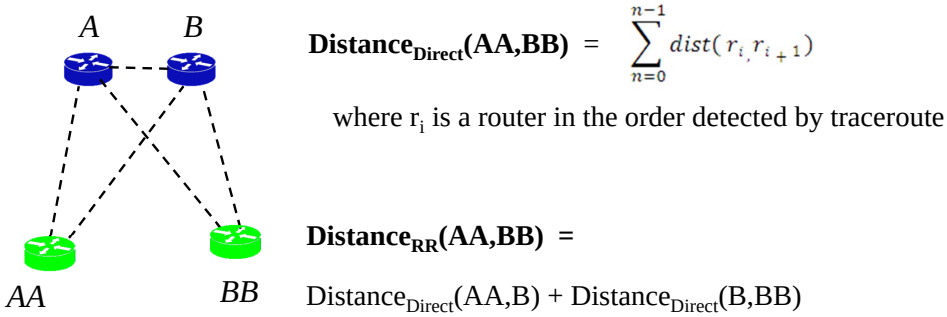
before and after removing the superfluous updates of I_{down} events (the worst case) in more detail. First, we observe that there is a slight difference on the overall duration and busy duration. The superfluous updates increased the overall duration and busy duration of I_{down} events by about 5% and 7% on average respectively. Additionally, we observe that there is a considerable increase in the number of best path changes made. Overall, we observe more than 38% increase in the number of best path changes on average due to the superfluous updates.

Event Types	**Duration (Busy)**	**Updates**
I_{up}	0.29% (0.97%)	2.72%
I_{short}	0.18% (0.65%)	3.41%
I_{long}	0.34% (1.10%)	12.79%
I_{down}	5.26% (7.21%)	38.55%

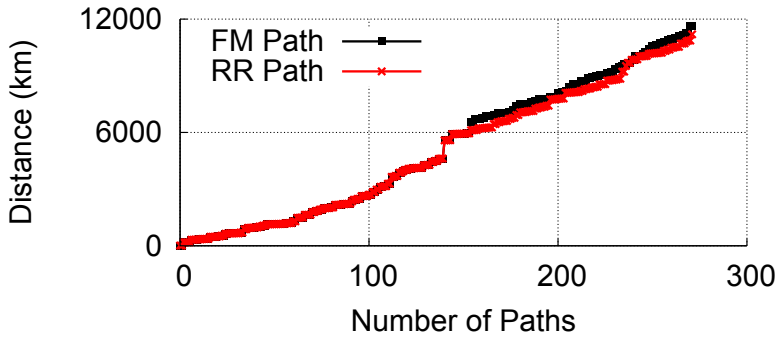
Table 4.4: Summary of average % increase caused by superfluous updates during June 2010

4.5.4.2 Physical Path Stretch and Latency

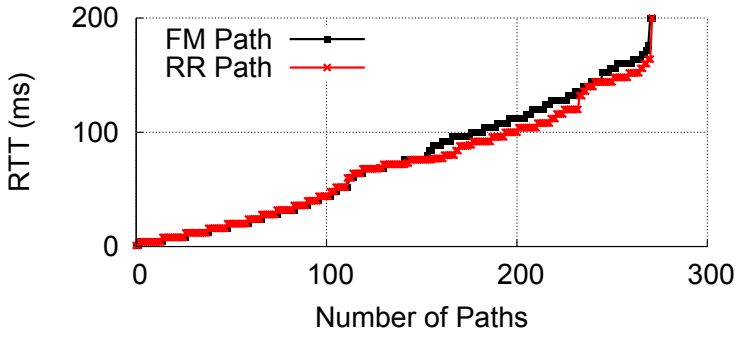
The alternative i-BGP architectures such as route reflection or AS confederations create a more scalable topology by essentially forming a hierarchical overlay topology on top of the existing full-mesh i-BGP topology. In these overlay i-BGP topologies, the update messages may travel only using the control paths that exist in the created overlay topology. As a result, an update message can often travel over a longer path, although there exists a shorter path. This can potentially delay the overall update propagation time. To measure the extent of this delay, we measure and compare the shortest physical path with the path in the route reflection topology by performing a traceroute and ping from each



(a) Example of direct path vs. RR path



(b) Path length



(c) Latency

Figure 4.19: Full-mesh vs. route reflection path length and latency during June 2010

of the 17 route reflectors in the backbone (i.e., the route reflectors in the top 2 levels) inside ISP_{RR}. There are 17x16 = 272 unidirectional paths in total. To calculate the physical distance from the obtained traceroute data, we first mapped

the router-level traceroute path to a POP-level path by examining the names of the routers, and finally calculated the distance by adding the POP-level distance from the source POP to the destination POP. We perform traceroute and ping at the same time, and across different times. In this chapter, we only present the representative result performed on April 26th, 2011 for clarity.

Figure 4.19 shows the distribution of physical path lengths in kilometers for the 272 paths, which would have been used in full-mesh i-BGP, compared with the route reflection paths as currently used by the route reflectors in the top 2 levels inside ISP_{RR}. Surprisingly, using the route reflection paths have slightly lower path length and latency in general in the case of ISP_{RR}. This indicates that (1) ISP_{RR}'s IGP metric is slightly different than the actual physical distance of the paths, and (2) by carefully designing the route reflection topology to align with the actual distance of the paths, one may avoid or even lower the overall latency.

4.6 Discussion

4.6.1 The Impact of MRAI Timer on i-BGP Convergence

MRAI timer is used in both i-BGP and e-BGP sessions to avoid overwhelming the neighboring router by limiting the number of updates during a given time interval. The default timer values are 5 seconds and 30 seconds for i-BGP and e-BGP sessions respectively. However for i-BGP sessions, a common practice is not to use MRAI timer in i-BGP sessions to expedite the convergence process, which is also the case in both ISP_{FM} and ISP_{RR}. As a result, we observed that the convergence duration is very short (mostly under 1 second) for majority of local events. Despite the observation that the convergence time observed for local

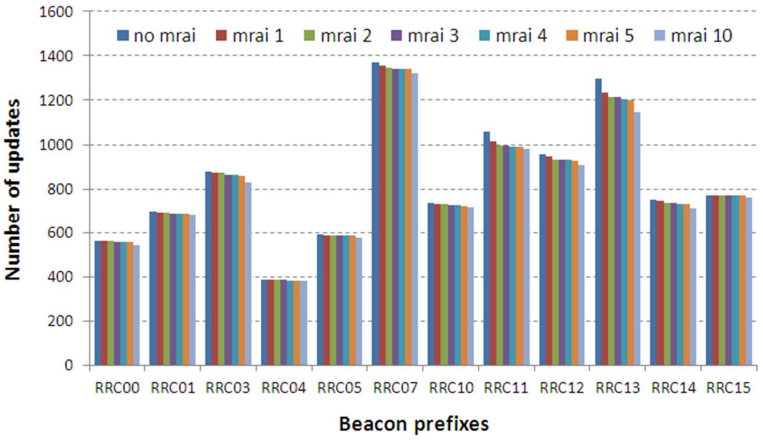
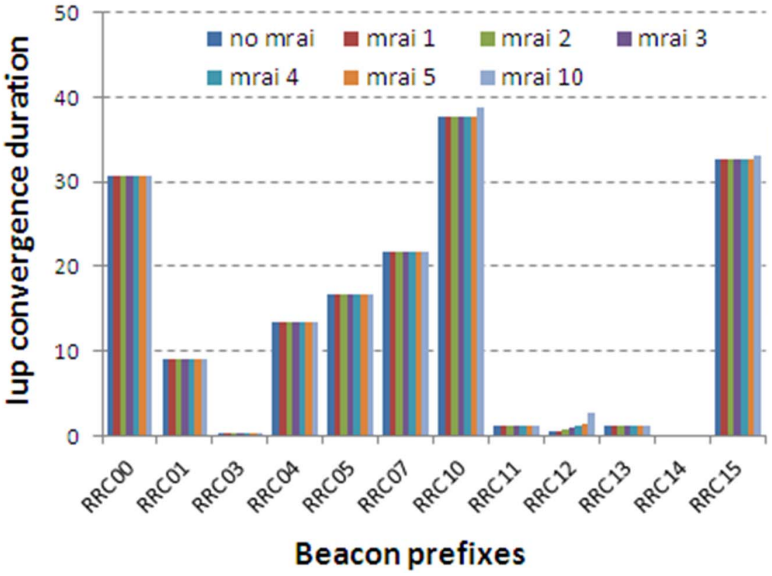


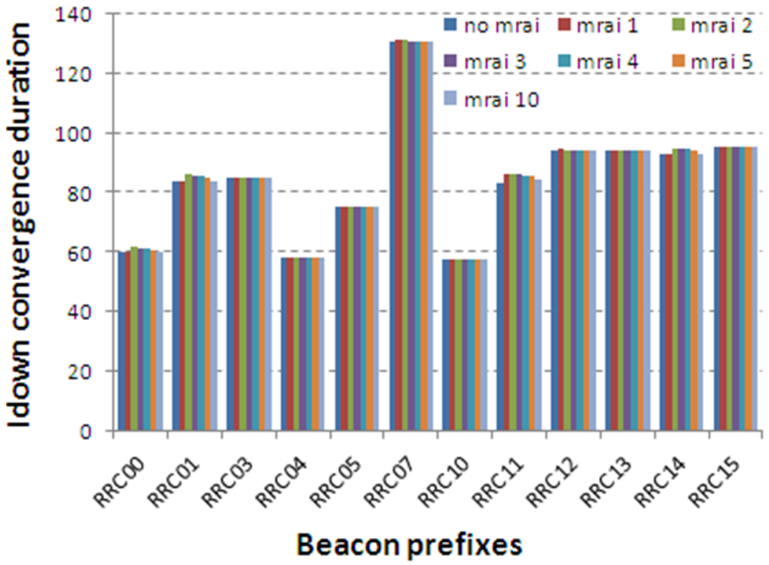
Figure 4.20: I-BGP MRAI timer values vs. update reduction in ISP_{FM}

events is quite fast, it is unclear how much impact the absence of MRAI timer had on reducing the overall convergence time (gain) or increasing the overall i-BGP update churn (loss) within the ISPs. As a first step to understand the tradeoffs of (not) using the MRAI timer in i-BGP sessions, we perform a set of simulations on beacon prefixes using the monitor-collector sessions in ISP_{FM} and ISP_{RR}. In practice, MRAI timer is implemented on per-peer basis. However in our simulations, we assume that the MRAI timer is applied on (peer,prefix) tuple and no WRATE for simplicity. In each simulation, we counted the total number of updates and measured convergence time for beacon prefix events as we vary the i-BGP MRAI timer values to 0, 1, 2, 3, 4, 5, 10 seconds. We repeated our simulations on different dates during June 2011. However for clarity, we present the results during one day of June 3rd 2010 for clarity.

Figure 4.20 and 4.21 summarizes our simulation results for ISP_{FM}. We observe that the use of MRAI timer in ISP_{FM} does reduce the total number of updates and increase the convergence time in general. However, the reduction in the total number of updates and the the increase in the convergence time duration are not significant. We find that this is due to the i-BGP full-mesh design that



(a) I_{up} events



(b) I_{down} events

Figure 4.21: I-BGP MRAI timer values vs. convergence time increase in ISP_{FM}

the routers do not forward reachability information. In i-BGP full-mesh peering sessions, an i-BGP router does not send an update after receiving an update from another i-BGP router by design. Therefore for an i-BGP router to send an update

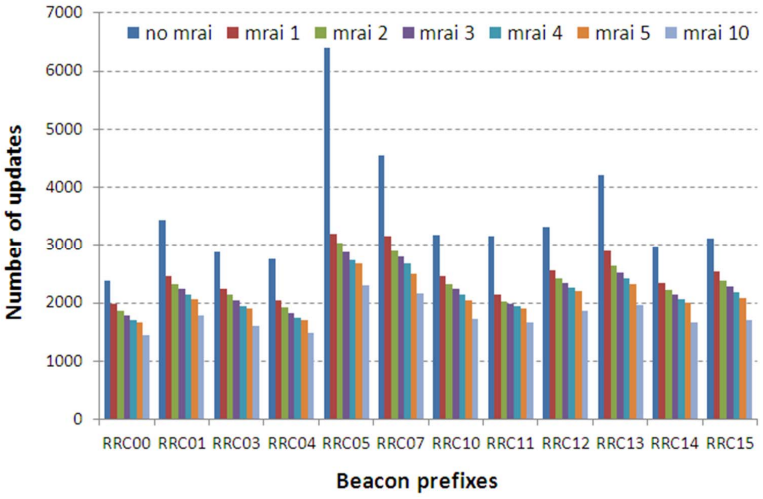
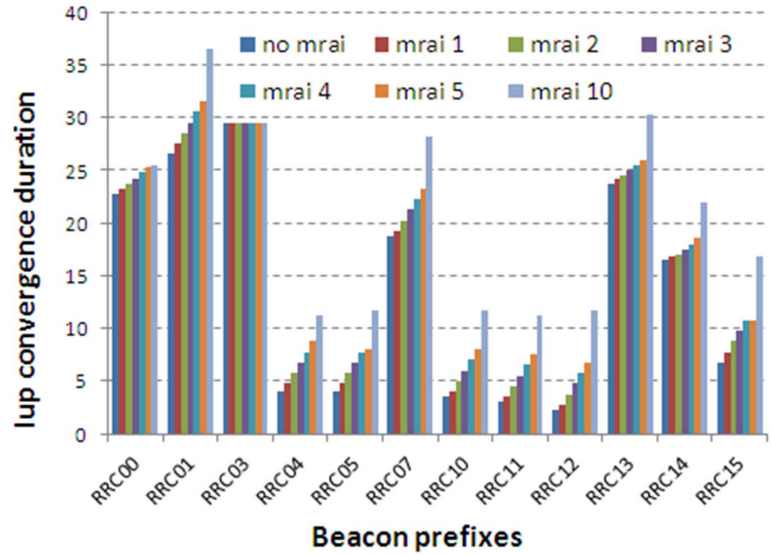


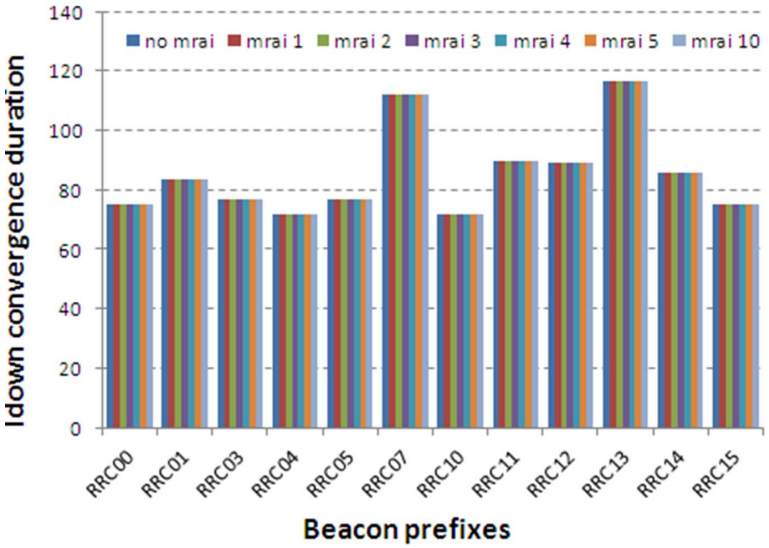
Figure 4.22: I-BGP MRAI timer values vs. update reduction in ISP_{RR}

to its neighbors, it must be the case that the router received an update message from an e-BGP neighbor. Because the arrival of external updates for a given router is often affected by the e-BGP MRAI timer of 30 seconds, i-BGP timer with up to 10 seconds in the simulation (which is much less than 30 seconds) did not have a major impact on neither the update reduction nor the convergence time increases.

Figure 4.22 and 4.23 summarizes our simulation results for ISP_{RR}. We observe that the use of MRAI timer in ISP_{RR} also reduces the total number of updates and increases the convergence time in general. However unlike the case of ISP_{FM}, both the reduction in the total number of updates and the increase in the convergence time duration are significant; there is up to 50% reduction (e.g., RRC05 beacon with MRAI timer = 1) and 30% increase (e.g., RRC15 beacon I_{up} event with MRAI timer = 10) in the total number of updates and convergence time respectively. Unlike the previous i-BGP full-mesh peering sessions, the time gab between the two consecutive updates from a given route reflector to the collector depends on the inter-arrival times of external updates *across all*



(a) I_{up} events



(b) I_{down} events

Figure 4.23: I-BGP MRAI timer values vs. convergence time increase in ISP_{RR}

border routers in ISP_{RR}, which can be much shorter than the inter-arrival times of updates for a given router.

In e-BGP on the other hand, MRAI timer with the default value of 30 seconds

is used between routers in different ASes and has been identified as one of the most influential factors that leads to a slow BGP convergence in the Internet [OZP06]. We observed that the AS-wide events are mostly caused by routing changes that happen outside the ISPs and that the update messages often arrive in bursts with 30 seconds burst interval, mainly affected by the prevalent usage of e-BGP MRAI timers in the path through which the update messages traverse to reach the given ISP. Because BGP convergence inside an ISP is much faster than the burst arrival rate (30 seconds), we observed that there is a large time gap between making path changes during a given AS-wide event.

4.6.2 The Impact of HRR on i-BGP Convergence

As speculated, multi-level hierarchical route reflection incurs more overhead. However, the overhead was noticeable in terms of the control plane load (i.e., number of updates), and in terms of convergence duration there was only a slight increase. Interestingly in terms of physical path stretch, i-BGP overlay paths in route reflection topology reduced the actual distance and latency. Therefore, there was not an additional delay in the studied ISP. However, this example shows that designing the topology carefully to follow the physical path is important in mitigating the potential additional delays.

4.7 Related Works

BGP convergence is closely related to data plane performance [WMJ06, Zha04, PWM03], and there have been extensive studies on BGP convergence and its properties.

There were mainly three types of work that measure BGP convergence. The

first type performed *active measurements* [LAA01,MBG03,LAW01] using a small set of prefixes in controlled environments. After injecting controlled BGP announcements, they showed that BGP converges slowly in the order of minutes and sometimes longer and further analyzed the root causes of the observed slow convergence. The second type is *passive measurement* studies [RWX02, WMR05, OZP06] using collected BGP data. Our work belongs to this type since we use passively collected i-BGP data from the measurement ISP to quantify and understand i-BGP convergence. These passive measurement studies share many similarities with our work because the source data format is the same. For example, we also use timer-based update clustering approach used in [RWX02, CGH03, FMM04, OZP06]. Lastly, the third type uses *simulations* to study BGP convergence and its properties [Gri01, NC02].

Most of the previous works, including the ones mentioned above, focus on BGP dynamics at the AS level (e.g., e-BGP convergence). We step down a level and take a detailed look inside a single node to shed lights on the BGP convergence properties within a single ISP. As one of the most closely related work, Pei et al. [PM06] collected i-BGP data for a set of small prefixes to study the convergence behavior of virtual private networks (VPN) within a single ISP. In this work, we study all prefixes in the global routing table as seen by the measurement ISP to study the i-BGP dynamics as well as the impact of i-BGP architecture on convergence delay.

4.8 Conclusions

Both inter-AS and intra-AS BGP measurement studies are required to achieve a comprehensive and complete understanding of the end-to-end routing performance. Unfortunately up to now most BGP measurement and analytical studies

have been limited to the BGP behaviors at inter-AS level, with virtually no measurement study on BGP dynamics *within* individual ASes. In this chapter, we conducted the first systematic measurement study to define, quantify, and analyze i-BGP convergence using i-BGP data collected from two large ISPs.

Our work provides a number of interesting characteristics of i-BGP convergence and performance quantification results. We discover that most routing events are either local or AS-wide in their scale. The local failures and recoveries involve different independent locations and routing convergence is quite fast; the majority of local events converge within 1 second. The duration of AS-wide events are mostly affected by the two factors: (1) the connectivity between the measured ISP and the destination prefix being affected, and (2) the external update propagation delays outside the measured ISP.

We take a step further to measure the overhead and performance differences between the full-mesh i-BGP architecture and the hierarchical route reflections (HRR). Our results show that, although HRR brings an increase in the routing update counts, this additional overhead is not significant in most cases, and can be mitigated through a carefully engineered i-BGP topology.

References

[BAS03] A. Bremler-Barr, Y. Afek, and S. Schwarz. "Improved BGP Convergence via Ghost Flushing." In *Proceedings of IEEE INFOCOM*, pp. 927–937, March 2003.

[BC96] T. Bates and R. Chandra. "BGP Route Reflection An alternative to full mesh IBGP." RFC 1966 (Experimental), June 1996. Obsoleted by RFC 4456, updated by RFC 2796.

[BCC00] T. Bates, R. Chandra, and E. Chen. "BGP Route Reflection - An Alternative to Full Mesh IBGP." RFC 2796 (Proposed Standard), April 2000. Obsoleted by RFC 4456.

[BCC06] T. Bates, E. Chen, and R. Chandra. "BGP Route Reflection: An Alternative to Full Mesh Internal BGP (IBGP)." RFC 4456 (Draft Standard), April 2006.

[BUM08] Marc Olivier Buob, Steve Uhlig, and Mickael Meulle. "Designing Optimal iBGP Route Reflection Topologies." *Networking*, April 2008.

[CCF05] Matthew Caesar, Donald Caldwell, Nick Feamster, Jennifer Rexford, Aman Shaikh, and Jacobus van der Merwe. "Design and Implementation of a Routing Control Platform." In *NSDI'05: Proceedings of the 2nd conference on Symposium on Networked Systems Design & Implementation*, pp. 15–28, Berkeley, CA, USA, 2005. USENIX Association.

[CDZ05] Jaideep Chandrashekar, Zhenhai Duan, Zhi-Li Zhang, and Jeff Krasky. "Limiting Path Exploration in BGP." In *Proceedings of IEEE INFOCOM*, 2005.

[CGH03] D. Chang, R. Govindan, and J. Hiedemann. "The Temporal and Topological Characterestics of BGP Path Changes." In *Proceedings of ICNP*, november 2003.

[CPP10] P-C. Cheng, Jong Han Park, Keyur Patel, and Lixia Zhang. "Route Flap Damping with Assured Reachability." In *AINTEC '10: Asian Internet Engineering Conference*, 2010.

[DS99] Rohit Dube and J. G. Scudder. "Route Reflection Considered Harmful." http://ietfreport.isoc.org/all-ids/draft-dube-route-reflection-harmful-00.txt, May 1999.

[DS04] Shivani Deshp and Biplab Sikdar. "On the Impact of Route Processing and MRAI Timers on BGP Convergence Times." In *Proceedings of Global Telecommunications Conference*, 2004.

[Dub99] R. Dube. "A Comparison of Scaling Techniques for BGP." *SIGCOMM Comput. Commun. Rev.*, October 1999.

[EKD10] Ahmed Elmokash, Amund Kvalbein, and Constantine Dovrolis. "BGP Churn Evolution: A Perspective from the Core." In *Proceedings of IEEE INFOCOM*, March 2010.

[FBR04] Nick Feamster, Hari Balakrishnan, Jennifer Rexford, Aman Shaikh, and Kobus 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.

[FKM04] Anja Feldmann, Hongwei Kong, Olaf Maennel, and Er Tudor. "Measuring BGP Pass-through Times." In *Passive and Active Measurement Workshop (PAM)*, pp. 267–277, 2004.

[FMM04] A. Feldmann, Olaf Maennel, Z. Morley Mao, A. Berger, and B. Maggs. "Locating Internet Routing Instabilities." In *Proceedings of ACM SIGCOMM*, September 2004.

[FXB10] P. Francis, X. Xu, H. Ballani, D. Jen, R. Raszuk, and L. Zhang. "FIB Suppression with Virtual Aggregation." http://tools.ietf.org/html/draft-ietf-grow-va-02, March 2010.

[Gri01] Timothy G. Griffin. "An Experimental Analysis of BGP Convergence Time." In *Proceedings of ICNP*, pp. 53–61, 2001.

[GW02] Timothy G. Griffin and Gordon Wilfong. "On the Correctness of i-BGP Configuration." *SIGCOMM Comput. Commun. Rev.*, **32**(4):17–29, 2002.

[Hou] Geoff Houston. "BGP Routing Table Analysis Reports." `http://bgp.potaroo.net/`. (Online).

[HWJ06] Junghee Han, David Watson, and Farnam Jahanian. "An Experimental Study of Internet Path Diversity." In *IEEE Transactions on Dependable and Secure Computing*, October 2006.

[Jak10] Paul Jakma. "Revisions to the BGP 'Minimum Route Advertisement Interval." http://tools.ietf.org/html/draft-ietf-idr-mrai-dep-02, February 2010.

[KKK07] Nate Kushman, Srikanth Kandula, and Dina Katabi. "Can You Hear Me Now?! It Must be BGP." In *SIGCOMM Comput. Commun. Rev.*, ,, March 2007.

[LAA01] Craig Labovitz, Abha Ahuja, Abhijit Abose, and Farnam Jahanian. "Delayed Internet Routing Convergence." *IEEE/ACM Transactions on Networking*, **9**(3):293 – 306, June 2001.

[LAW01] C. Labovitz, A. Ahuja, R. Wattenhofer, and S. Venkatachary. "The Impact of Internet Policy and Topology on Delayed Routing Convergence." In *Proceedings of IEEE INFOCOM '01*, April 2001.

[LMJ99] C. Labovitz, G. R. Malan, and F. Jahanian. "Origins of Internet Routing Instability." In *Proceedings of IEEE INFOCOM '99*, pp. 218–26, New York, NY, 1999.

[LPR08] Mohit Lad, Jong Han Park, Tiziana Refice, and Lixia Zhang. "A Study of Internet Routing Stability Using Link Weight." Technical Report UCLA/CSD-080003, University of California, Los Angeles, 2008.

[max] "MaxMind - GeoIP." `http://www.maxmind.com/app/ip-location`.

[MBG03] Z. M. Mao, R. Bush, T. Griffin, and M. Roughan. "BGP Beacons." 2003.

[MFC11] P. Marques, R. Fernando, E. Chen, and P. Mohapatra. "Advertisement of the Best External Route in BGP." `http://tools.ietf.org/html/draft-ietf-idr-best-external-03`, March 2011. (Internet Draft).

[MGW02] D. McPherson, V. Gill, D. Walton, and A. Retana. "Border Gateway Protocol (BGP) Persistent Route Oscillation Condition." RFC 3345 (Informational), August 2002.

[mrt10] "MRT Routing Information Export Format." `http://www.ietf.org/internet-drafts/draft-ietf-grow-mrt-13.txt`, September 2010. [Online].

[NC02] J. Nykvist and L. Carr-Motykova. "Simulating Convergence Properties of BGP." In *Proceedings of 11th International Conference on Computer Communications and Networks*, October 2002.

[NCC] RIPE NCC. "Routing Information Service." http://www.ris.ripe.net/.

[NM01] Krishna Nayak and Dan McKernan. "Measuring Provider Path Diversity from Traceroute Data: work in progress." In *CAIDA-ISMA Workshop*, December 2001.

[OPW10] Ricardo Oliveira, Dan Pei, Walter Willinger, Beichuan Zhang, and Lixia Zhang. "The (in)Completeness of the Observed Internet AS-level Structure." In *IEEE/ACM Transactions on Networking*, 2010.

[OZP06] Ricardo Oliveira, Beichuan Zhang, Dan Pei, Rafit Itzak-Ratzin, and Lixia Zhang. "Quantifying Path Exploration in the Internet." In *Proceedings of Internet Measurement Conference*, 2006.

[PAM05] Dan Pei, Matt Azuma, Dan Massey, and Lixia Zhang. "BGP-RCN: Improving BGP Convergence through Root Cause Notification." *Computer Networks*, June 2005.

[Par11] Jong Han Park. "BGP Best Path Change Inference Project." `http://sourceforge.net/projects/infer-bpc/`, May 2011.

[PM06] Dan Pei and Jacobus Van der Merwe. "BGP Convergence in Virtual Private Networks." In *Proceedings of the 6th ACM SIGCOMM Conference on Internet Measurement*, pp. 283–288, New York, NY, USA, 2006. ACM.

[PTO08] Cristel Pelsser, Tomonori Takeda, Eiji Oki, and Kohei Shiomoto. "Improving Route Diversity through the Design of iBGP Topologies." *IEEE International Conference on Communications*, May 2008.

[PWM03] Dan Pei, Lan Wang, Daniel Massey, S. Felix Wu, and Lixia Zhang. "A Study of Packet Delivery Performance during Routing Convergence." In *Proceedings of IEEE International Conference on Dependable Systems and Networks (DSN)*, 2003.

[PZW02] Dan Pei, Xiaoliang Zhao, Lan Wang, Dan Massey, Allison Mankin, S. Felix Wu, and Lixia Zhang. "Improving BGP Convergence Through Consistency Assertions." In *Proceedings of IEEE INFOCOM*, 2002.

[RFP11] R. Raszuk, R. Fernando, K. Patel, D. McPherson, and K. Kumaki. "Distribution of Diverse BGP Paths." `http://tools.ietf.org/html/draft-ietf-grow-diverse-bgp-path-dist-03`, January 2011. (Internet Draft).

[RLH06] Y. Rekhter, T. Li, and S. Hares. "A Border Gateway Protocol 4 (BGP-4)." RFC 4271 (Draft Standard), January 2006.

[RWX02] Jennifer Rexford, Jia Wang, Zhen Xiao, and Yin Zhang. "BGP Routing Stability of Popular Destinations." In *Proceedings of the ACM SIGCOMM Workshop on Internet Measurement*, 2002.

[SD99] John G. Scudder and Rohit Dube. "BGP Scaling Techniques Revisited." *SIGCOMM Comput. Commun. Rev.*, October 1999.

[SF09] V. V. den Schrieck and P. Francois. "Analysis of Paths Selection Modes for Add-Paths." http://tools.ietf.org/html/draft-vvds-add-paths-analysis-00, July 2009.

[sid] "IETF Secure Intra-Domain Routing Working Group." http://datatracker.ietf.org/wg/sidr/charter/.

[SKM06] A. Sahoo, K. Kant, and P. Mohapatra. "Speculative Route Invalidation to Improve BGP Convergence Delay under Large-Scale Failures." In *Proceedings of 11th International Conference on Computer Communications and Networks*, October 2006.

[SMS06] Wei Sun, Z. M. Mao, and K. G. Shin. "Differentiated BGP Update Processing for Improved Routing Convergence." In *Proceedings of ICNP*, pp. 280–290, November 2006.

[TMS03] Renata Teixeira, Keith Marzullo, Stefan Savage, and Geoffrey M. Voelker. "In search of path diversity in ISP networks." *ACM Sigcomm conference on Internet measurement*, 2003.

[TMS07] P. Traina, D. McPherson, and J. Scudder. "Autonomous System Confederations for BGP." RFC 5065 (Draft Standard), August 2007.

[Uni] University of Oregon. "Route Views Project." http://www.routeviews.org.

[USF10] J. Uttaro, V. V. den Schrieck, P. Francois, R. Fragassi, A. Simpson, and P. Mohapatra. "Best Practices for Advertisement of Multiple Paths in BGP.", November 2010.

[UT06] Steve Uhlig and Sebastien Tandel. "Quantifying the BGP Routes Diversity Inside a Tier-1 Network." *Networking*, **3976**, April 2006.

[VAZ07] Vijay Vasudevan, David G Andersen, and Hui Zhang. "Understanding the AS-level Path Disjointness Provided by Multi-homing.", 2007.

[VCS09] P. D. Arjona Villicana, C. C. Constantinou, and A. S. Stepanenko. "The Internet's Unexploited Path Diversity.", 2009.

[VVK06] Mythili Vutukuru, Paul Valiant, Swastik Kopparty, and Hari Balakrishnan. "How to Construct a Correct and Scalable i-BGP Configuration." In *IEEE Infocom*, April 2006.

[WMJ06] Feng Wang, Z. Morley Mao, Lixin Gao Jia Wang, , and Randy Bush. "A Measurement Study on the Impact of Routing Events on End-to-End Internet Path Performance." In *Proceedings of ACM SIGCOMM*, 2006.

[WMR05] Jian Wu, Z. Morley Mao, and Jennifer Rexford. "Finding a Needle in a Haystack: Pinpointing Significant BGP Routing Changes in an IP Network." In *Proceedings of 2nd symposium on Networked Systems Design and Implementation (NSDI)*, 2005.

[WRC10] D. Walton, A. Retana, E. Chen, and J Scudder. "Advertisement of Multiple Paths in BGP." `http://www.ietf.org/id/draft-ietf-idr-add-paths-04.txt`, August 2010. (Internet Draft).

[WZP02] Lan Wang, Xiaoliang Zhao, Dan Pei, Randy Bush, Daniel Massey, Allison Mankin, S. Felix Wu, and Lixia Zhang. "Observation and Analysis of BGP Behavior under Stress." In *Proceedings of the 2nd ACM SIGCOMM Workshop on Internet Measurment*, pp. 183–195, New York, NY, USA, 2002. ACM Press.

[XWN03a] Li Xiao, Jun Wang, and K. Nahrstedt. "Reliability-aware i-BGP Route Reflection Topology Design." In *ICNP '03: Proceedings of the 11th IEEE International Conference on Network Protocols*, 2003.

[XWN03b] Li Xiao, Jun Wang, and Klara Nahrstedt. "Optimizing i-BGP Route Reflection Network." *IEEE International Conference on Communications*, May 2003.

[ZAL04] Hongwei Zhang, Anish Arora, and Zhijun Liu. "A Stability-Oriented Approach to Improving BGP Convergence." In *In SRDS04*, pp. 90–99, 2004.

[Zha04] Beichuan Zhang. "Destination Reachability and BGP Convergence Time." In *Proceedings of IEEE Globecom, Global Internet and Next Generation Networks*, pp. 1383–1389, 2004.

[ZLM] Beichuan Zhang, Raymond Liu, Dan Massey, and Lixia Zhang. "Internet Topology Project." http://irl.cs.ucla.edu/topology/.