# *Quantifying i-BGP Convergence inside Large ISPs*

JongHan Park,Pei-chun Cheng, Shane Amante, Dorian Kim, Danny McPherson, Lixia Zhang

University of California, Los Angeles
Computer Science Department
Technical Report # 110009
May 16, 2011

# Quantifying i-BGP Convergence inside Large ISPs

Jong Han Park* jpark@cs.ucla.edu
Pei-chun Cheng* pccheng@cs.ucla.edu
Shane Amante† Shane.Amante@Level3.com
Dorian Kim‡ dorian@blackrose.org
Danny McPherson§ dmcpherson@verisign.com
Lixia Zhang* lixia@cs.ucla.edu

## ABSTRACT

BGP is the global routing protocol used to propagate reachability information both between and within autonomous systems. 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.

## Categories and Subject Descriptors

C.2 [**Computer Communication Networks**]: Network protocols, Network operations

## General Terms

Measurements, Performance

## Keywords

i-BGP, routing convergence

## 1. INTRODUCTION

BGP [27] 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 [13, 25, 34, 37], 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.* [14–16] showed the existence of slow BGP routing convergence. Subsequent measurement studies confirmed the wide existence of slow convergence [21] and proposed a variety of BGP modifications to speed up BGP routing convergence [3, 4, 6, 23, 26, 29, 31, 38].

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 paper, 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 paper 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 3 ˜Section 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

*Computer Science Department, UCLA.
†Level-3 Communications Inc.
‡NTT Communications Inc.
§Verisign Inc.

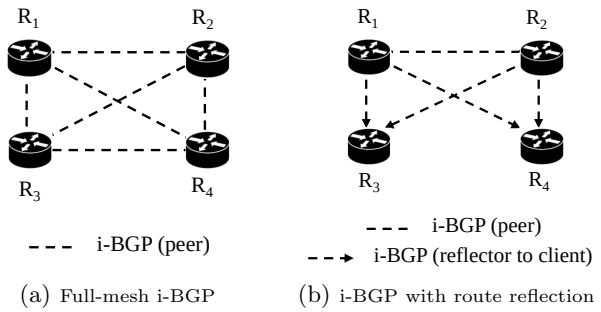
other hand, events that affect all routers (*i.e.,* AS-wide events) take much longer time to converge, caused mostly by delayed arrivals of external update messages at the studied ISPs.

- 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 5.3).
- 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 4.5). We make our implementation publicly available [22], which may be useful for the future research, as well as to the ISPs who wish to quantify their i-BGP convergence using i-BGP data collected over peering sessions.

This paper is organized as follows. In Section 2, we provide necessary background for this paper, including a briefing on different i-BGP architectures and their basic operations. Section 3 defines i-BGP convergence and describes a number of metrics which we use to characterize the i-BGP convergence. Section 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 5 presents our results on the i-BGP convergence characteristics and the impact of different i-BGP architectures on BGP convergence. Section 6 discusses the ramifications of our observations and discoveries. In Section 7, we briefly talk about related works, and finally in Section 8 we summarize our work and conclude.

# 2. INTERNAL BGP (I-BGP)

[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 1: Different i-BGP topologies**

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 in different ASes set up BGP sessions to exchange the reachability of the prefixes by sending BGP update messages. Such BGP sessions are called external BGP (e-BGP) sessions. BGP is also used between routers within the same AS to exchange BGP routing updates, and these sessions are called internal BGP (i-BGP) sessions.

## 2.1 The Full-Mesh I-BGP

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 [2] and AS confederations [32].

## 2.2 Route Reflection

### 2.2.1 *Basic Operation of Route Reflection*

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 1 illustrates the difference between intercon-

necting i-BGP routers via full-mesh and via RRs. Figure 1(a) shows an example of full-mesh i-BGP interconnections, where all i-BGP speakers are directly connected to each other. Figure 1(b) shows an example of route reflection deployment, where $R_1$ and $R_2$ serve as RRs and connect to i-BGP speakers $R_3$ and $R_4$, which are connected to both reflectors for redundancy. Since $R_3$ can learn $R_4$'s BGP reachability information from the RRs and vice versa, $R_3$ and $R_4$ do not need to interconnect.

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 is possible to have routing loops. For example in Figure 1(b), an update message originated at $R_3$ can come back to $R_3$ through more than one RR ($R_1$ and $R_2$ in this case), forming a loop. To prevent such loops, two new attributes are added to BGP update messages: CLUSTER_LIST and ORIGINATOR_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. Thus in i-BGP with route reflection, one may find out the internal control path through which the update message traverse within the network by looking at CLUSTER_LIST attribute, just as one may find the ASes through which an e-BGP update has traversed by looking at AS_PATH attribute.

#### 2.2.2 Additional Delays caused by Route Reflection

In route reflection architecture, routing messages travel more than a single i-BGP hop. For example in Figure 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 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.

### 2.3 AS confederations

AS confederations [32] 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 [7, 30].

## 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.

### 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.

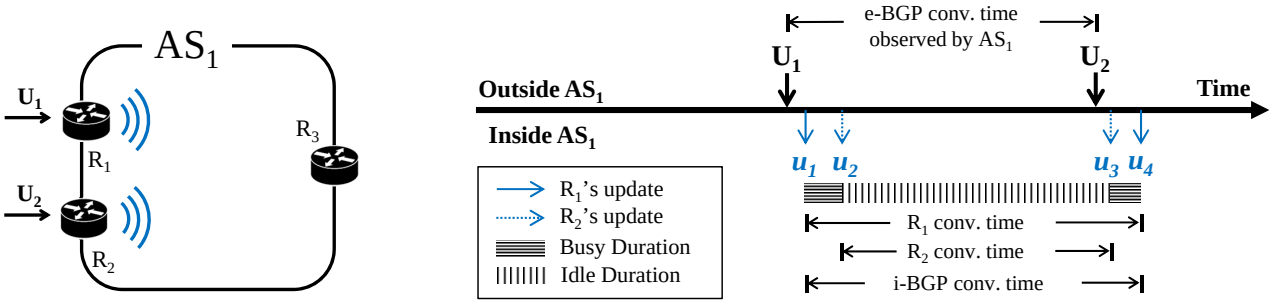
### 3.2 Metrics

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

#### 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 2 as an example to explain how we compute the convergence duration. In this example,




(a) External update arrivals

(b) i-BGP Updates Observed by Internal Routers

**Figure 2: i-BGP Convergence**

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 2(b) shows an example of busy and idle durations.

### 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 [35], 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 4.5, 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 paper.

### 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 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 paper, 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

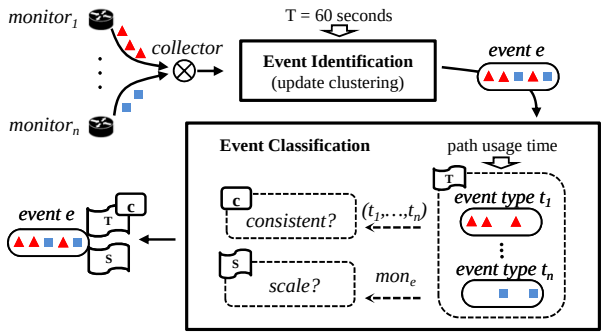


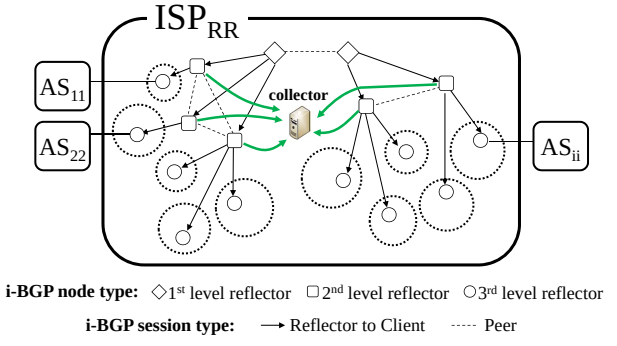

**Figure 3: High Level Data Processing**

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.

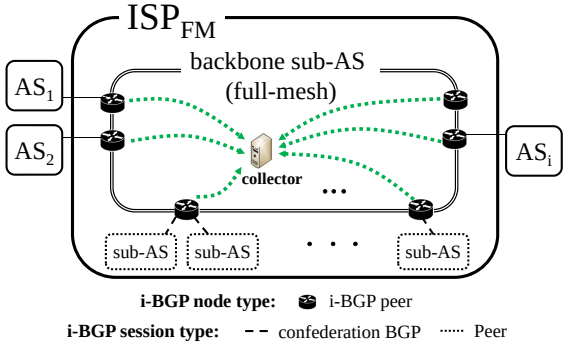
# 4. METHODOLOGY

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 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 [5,10,21,28,36] 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: Simplified i-BGP Topology of Two ISPs**

## 4.1 High Level Description of the two ISPs

### 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. To minimize the routing information propagation delay within the network, $ISP_{RR}$ does not use MRAI timer internally. Figure 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 $ISP_{RR}$.

### 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 [32] to scale with its network size. As in the case of $ISP_{RR}$,

$ISP_{FM}$ does not use MRAI timer inside its network. Figure 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 $ISP_{FM}$).

## 4.2 Data Collection and Preprocessing

In most of BGP data collection projects including Oregon RouteViews [33] and RIPE RIS [19], a collector (an i-BGP router) is used to set up BGP sessions with target routers (which we call *monitors* throughout the paper) and to passively record BGP data sent from the monitors in MRT [1] 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.

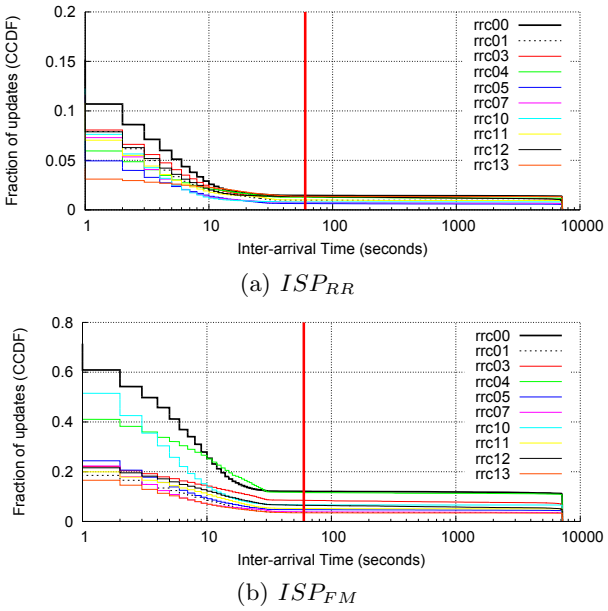
In $ISP_{RR}$, a collector is configured to maintain i-BGP *server-client* sessions to 18 route reflectors in the 2nd level and to passively record all i-BGP updates received during one year from May 2009 to April 2010. In $ISP_{FM}$, a collector is configured as one of the i-BGP peers in backbone sub-AS, 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.

## 4.3 Event Identification

A number of previous BGP data analytic studies [5, 10,21,28,36] 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.* [21] calculate the inter-arrival times of updates generated by BGP beacon prefixes [18], 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 [21] such that the inter-arrival times are calculated between two updates generated by all monitors inside the given AS.




(a) $ISP_{RR}$

(b) $ISP_{FM}$

**Figure 5: Inter-arrival Times of 10 Beacon Prefix Updates Observed Inside the two ISPs**

Figure 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 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 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.

### 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

$$S_e = \frac{mon_e}{mon_n} \quad (1)$$

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.

#### 4.4.2 Event Type

A number of previous works [15, 21] define different event types for a given routing event. To avoid confusion, we use the consistent definitions of event types. Table 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 [21] 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 paper, 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.

### 4.5 Geo-based Best Path Selection Inference

#### 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 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 router.

#### 4.5.2 Inferring the Best Path Selection in Peering Routers

| 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 1: Event Types**

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 geographical location of monitors instead, to approximate the IGP distance values. We confirmed with the operators in $ISP_{FM}$ that in general the IGP distance cost matches with the physical distance between the monitors.

# 5. RESULTS

In this section, we present our quantification results on the i-BGP convergence as defined in Section 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.

## 5.1 Number of Identified Events in Time

Figure 6 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.

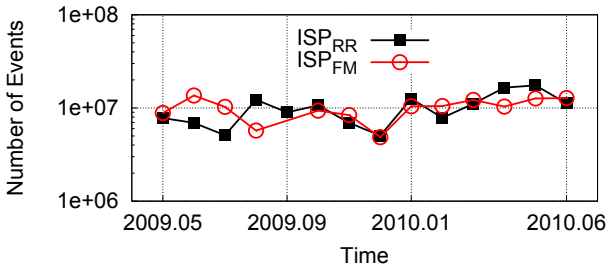



**Figure 6: Number of Identified Events from May 2009 to June 2010**

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 [12]. From both ISPs, we observe about 12% and 10% increase in the total number of prefixes during the 14 months.

## 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.

### 5.2.1 *Event Scale*

Figure 7 shows the distributions of event scale ($S_e$) of

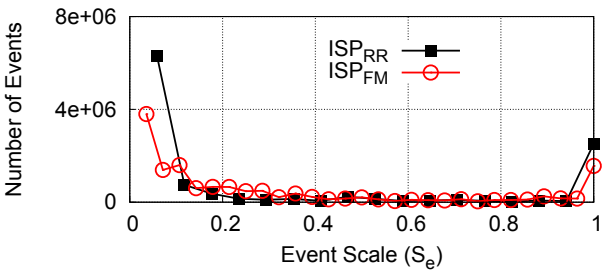



**Figure 7: Event Scale (June 2010)**

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 [17].

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 8 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 [8] that the BGP update churn observed from outside a large ISP can be due to uncorrelated and distributed local routing events across different locations.

### 5.2.2 *Local Events*

Table 2 shows the total number of local events identified from $ISP_{RR}$ and $ISP_{FM}$ during the month of June 2010. Figure 9 and Figure 10 summarize the characteristics of the local events inside $ISP_{RR}$ and $ISP_{FM}$ respectively, using the three metrics we introduced earlier in Section 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.

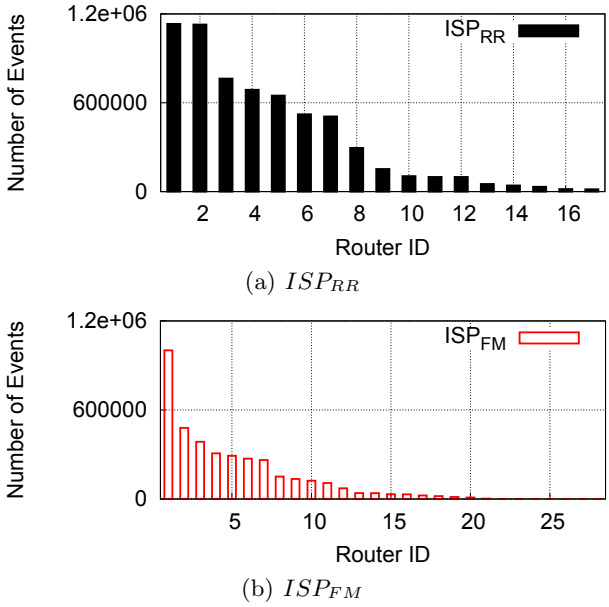



**Figure 8: Number of Local Events per Router**

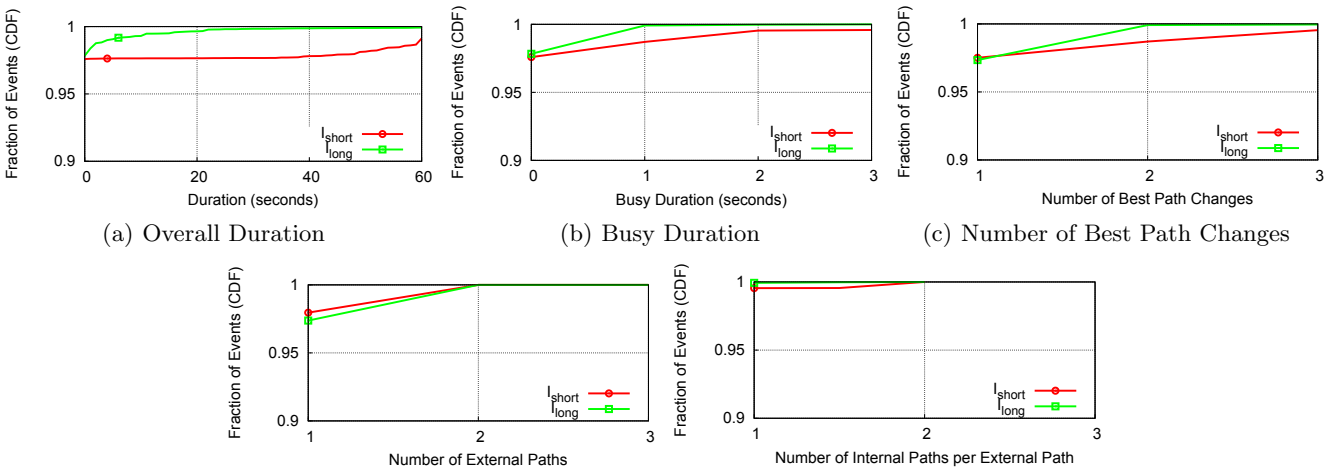
Because the fraction of such false positive local events is small enough (0.27% of overall events), we believe that the generality of our results is not be affected. In this section, we simply do not consider these local $I_{up}$ and $I_{down}$ events in our analysis.

From Table 2 and Figure 9 and 10, 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 9(c) and 10(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.

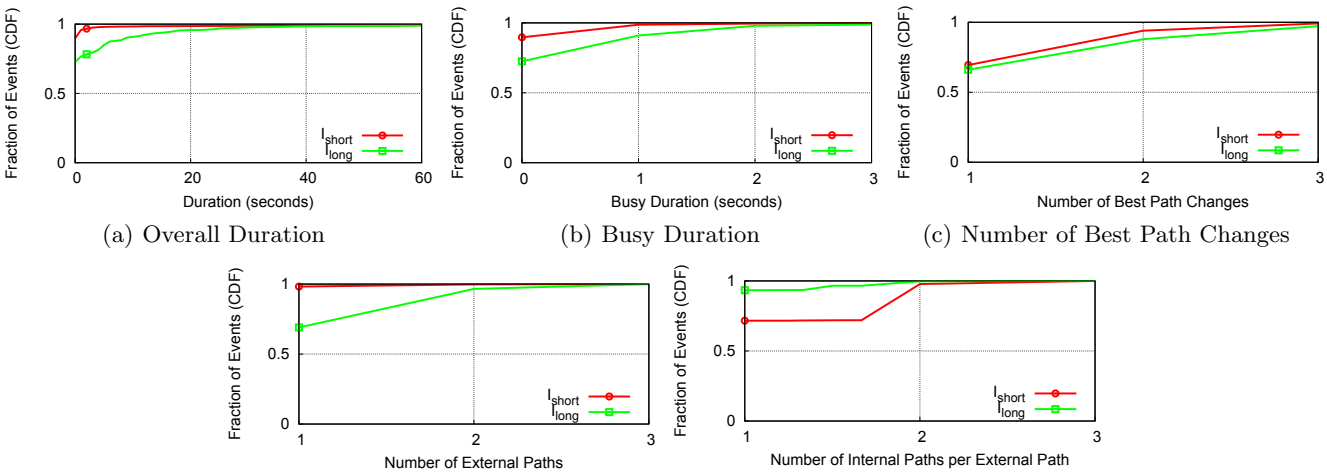
Third, we observe that a small fraction of local events have their convergence duration greater than a few seconds. These long durations (*e.g.,* the top 2.4% of $I_{short}$ events in $ISP_{RR}$) with can mostly be explained either by the inaccuracy of the timer-based event clustering technique which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




(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 9: Local Events Convergence in $ISP_{RR}$ During June 2010**




(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 10: Local Events Convergence in $ISP_{FM}$ During June 2010**

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

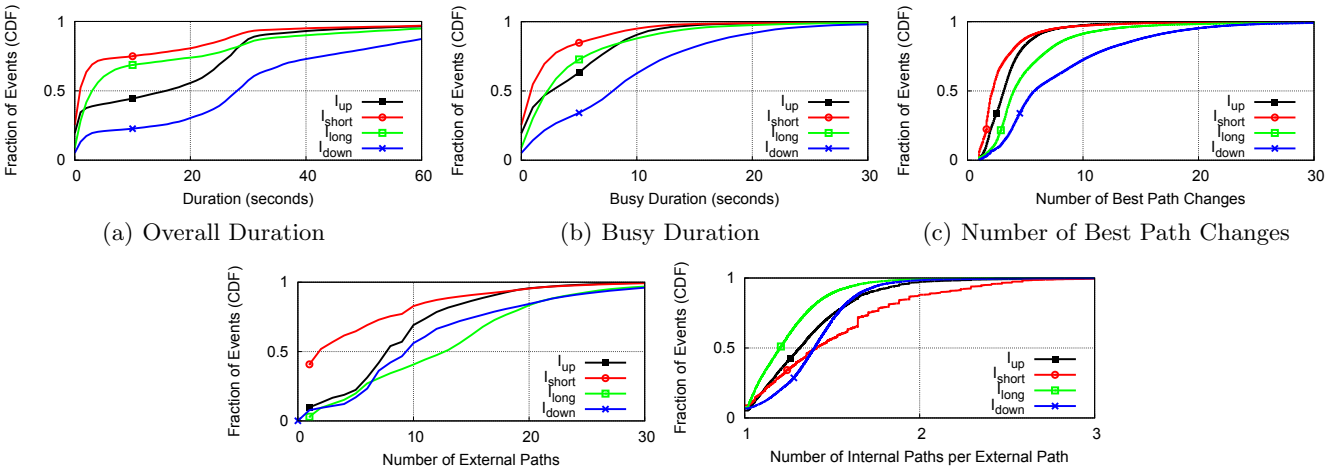
**Table 2: Number of Local Events in $ISP_{RR}$ and $ISP_{FM}$ During June 2010**

in [9].

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 duration 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.

### 5.2.3 AS-wide Events

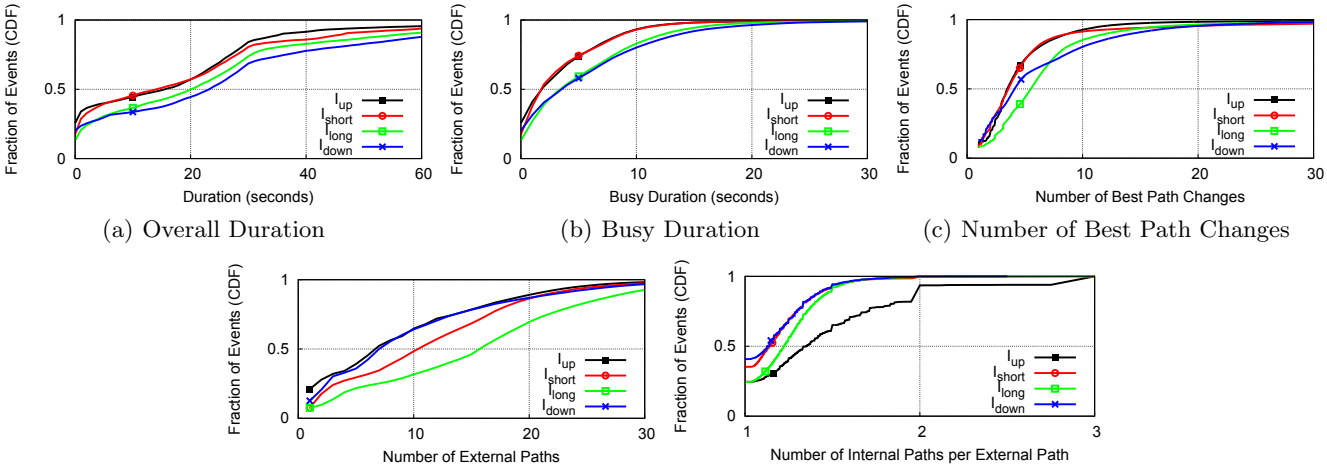
Table 3 shows the total number of identified AS-wide events from $ISP_{RR}$ and $ISP_{FM}$ during the same one month of June 2010. Figure 11 and Figure 12 summarize the characteristics of AS-wide i-BGP convergence using the three metrics we defined. As in the case of local events, we observe that the 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 investi-




(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 11: AS-wide Events Convergence in $ISP_{RR}$ During June 2010**




(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 12: AS-wide Events Convergence in $ISP_{FM}$ During June 2010**

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

**Table 3: Number of AS-wide Events in $ISP_{RR}$ and $ISP_{FM}$ During June 2010**

gation 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 11(a) and Figure 12(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 11(b) and Figure 12(b) are relatively lower in general compared to the overall durations shown in Figure 11(a) and Figure 12(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 duration in both $I_{short}$ and $I_{long}$. We verified that this indeed is the main cause for the observed shorter duration of $ISP_{RR}$.

## 5.3 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 impact of the various topologies on BGP convergence inside the network? As a first step to answer this question, 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.* [9], and therefore in this work, we focus mostly on the first two factors.

### 5.3.1 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 1(b) when $R_4$ withdraws a route to previously reachable prefix $p$ from its route reflectors $R_1$ and $R_2$, a number of update messages will be forwarded to $R_3$ through different control paths, as explained in Section 2. Until all withdrawal messages are received, $R_3$ 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 summarizes our results using i-BGP updates collected from $ISP_{RR}$ 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 13 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.

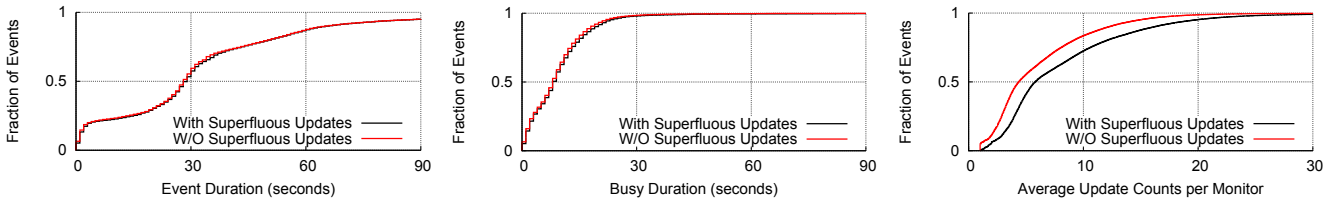
| Event Types | Duration (Busy) | Updates |
|---|---|---|
| $I_{up}$ | 0.29% (0.97%) | 2.72% |
| $I_{short}$ | 0.18% (0.65%) | 3.41% |
| $I_{long}$ | 0.34% (1.10%) | 12.79% |
| $I_{down}$ | 5.26% (7.21%) | 38.55% |

**Table 4: Summary of Average % Increase Caused by Superfluous Updates During June 2010**

### 5.3.2 Physical Path Stretch and Latency

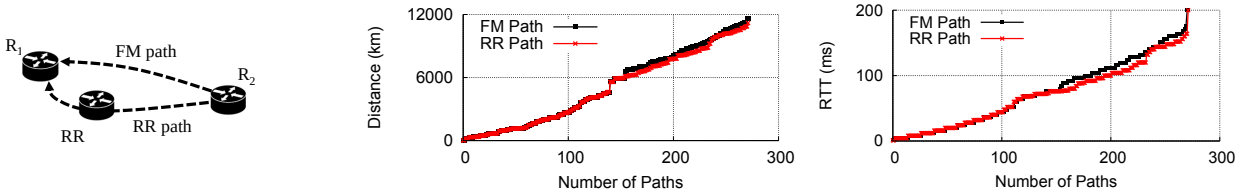
The alternative i-BGP architectures such as route reflection or AS confederations create a more scalable topology by essentially forming a hierarchical overlay topology on top of the existing full-mesh i-BGP topology. In these overlay i-BGP topologies, the update messages may travel only using the control 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 of the 17 route reflectors in the backbone 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 paper, we only present the representative result performed on April 26th, 2011 for clarity.

[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 13: AS-wide $I_{down}$ Convergence with and without Superfluous Updates During June 2010**




(a) Example of FM Path vs. RR Path (b) Path Length (c) Latency

**Figure 14: Full-Mesh vs. Route Reflection Path Length and Latency During June 2010**

Figure 14 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 $ISP_{RR}$. 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.

## 6. DISCUSSION

MRAI timer with the default value of 30 seconds in e-BGP is used between routers in different ASes to avoid overwhelming the neighboring router, by limiting the number of updates a router can send in a given time interval and has been identified as one of the most influential factors that leads to a slow BGP convergence in the Internet. In i-BGP, the default timer value suggested is 5 seconds. However, in both $ISP_{RR}$ and $ISP_{FM}$, MRAI timer is not used at all to minimize the convergence delay inside their networks. As a result, we observed that the convergence duration is very short (mostly under 1 second) for local events. On the other hand, AS-wide events are mostly caused by routing changes that happen outside the ISPs. As a result, the update messages often arrive in bursts with 30 seconds burst interval, mainly affected by the 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 bewteen making path changes during a given AS-wide event.

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.

## 7. RELATED WORKS

BGP convergence is closely related to data plane performance [25, 34, 37], 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* [14, 15, 18] 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 [21,28,36] 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 [5,10,21,28]. Lastly, the third type uses *simulations* to study BGP convergence and its proper-

ties [11, 20].

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.* [24] 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.

# 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 paper, 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 iBGP 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.

# 9. REFERENCES


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

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

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

[4] J. Chandrashekar, Z. Duan, Z.-L. Zhang, and J. Krasky. Limiting Path Exploration in BGP. In *Proceedings of IEEE INFOCOM*, 2005.

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

[6] S. Deshp and B. Sikdar. On the Impact of Route Processing and MRAI Timers on BGP Convergence Times. In *Proceedings of Global Telecommunications Conference*, 2004.

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

[8] A. Elmokash, A. Kvalbein, and C. Dovrolis. BGP Churn Evolution: A Perspective from the Core. In *Proceedings of IEEE INFOCOM*, March 2010.

[9] A. Feldmann, H. Kong, O. Maennel, and E. Tudor. Measuring BGP Pass-through Times. In *Passive and Active Measurement Workshop (PAM)*, pages 267–277, 2004.

[10] A. Feldmann, O. Maennel, Z. M. Mao, A. Berger, and B. Maggs. Locating Internet Routing Instabilities. In *Proceedings of ACM SIGCOMM*, September 2004.

[11] T. G. Griffin. An Experimental Analysis of BGP Convergence Time. In *Proceedings of ICNP*, pages 53–61, 2001.

[12] G. Houston. BGP Routing Table Analysis Reports. `http://bgp.potaroo.net/`. (Online).

[13] N. Kushman, S. Kandula, and D. Katabi. Can You Hear Me Now?! It Must be BGP. In *SIGCOMM Comput. Commun. Rev.*, ,, March 2007.

[14] C. Labovitz, A. Ahuja, A. Abose, and F. Jahanian. Delayed Internet Routing Convergence. *IEEE/ACM Transactions on Networking*, 9(3):293 – 306, June 2001.

[15] 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.

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

[17] M. Lad, J. H. Park, T. Refice, and L. Zhang. A Study of Internet Routing Stability Using Link Weight. Technical Report UCLA/CSD-080003, University of California, Los Angeles, 2008.

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

[19] R. NCC. Routing Information Service. http://www.ris.ripe.net/.

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

[21] R. Oliveira, B. Zhang, D. Pei, R. Itzak-Ratzin, and L. Zhang. Quantifying Path Exploration in the Internet. In *Proceedings of Internet Measurement Conference*, 2006.

[22] J. H. Park. BGP Best Path Change Inference Project. `http://sourceforge.net/projects/infer-bpc/`, May 2011.

[23] D. Pei, M. Azuma, D. Massey, and L. Zhang. BGP-RCN: Improving BGP Convergence through Root Cause Notification. *Computer Networks*, June 2005.

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

[25] D. Pei, L. Wang, D. Massey, S. F. Wu, and L. Zhang. A Study of Packet Delivery Performance during Routing Convergence. In *Proceedings of IEEE International Conference on Dependable Systems and Networks (DSN)*, 2003.

[26] D. Pei, X. Zhao, L. Wang, D. Massey, A. Mankin, S. F. Wu, and L. Zhang. Improving BGP Convergence Through Consistency Assertions. In *Proceedings of IEEE INFOCOM*, 2002.

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

[28] J. Rexford, J. Wang, Z. Xiao, and Y. Zhang. BGP Routing Stability of Popular Destinations. In *Proceedings of the ACM SIGCOMM Workshop on Internet Measurement*, 2002.

[29] 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.

[30] J. G. Scudder and R. Dube. BGP Scaling Techniques Revisited. *SIGCOMM Comput. Commun. Rev.*, October 1999.

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

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

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

[34] F. Wang, Z. M. Mao, L. G. Jia Wang, , and R. Bush. A Measurement Study on the Impact of Routing Events on End-to-End Internet Path Performance. In *Proceedings of ACM SIGCOMM*, 2006.

[35] L. Wang, X. Zhao, D. Pei, R. Bush, D. Massey, A. Mankin, S. F. Wu, and L. Zhang. Observation and Analysis of BGP Behavior under Stress. In *Proceedings of the 2nd ACM SIGCOMM Workshop on Internet Measurment*, pages 183–195, New York, NY, USA, 2002. ACM Press.

[36] J. Wu, Z. M. Mao, and J. 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.

[37] B. Zhang. Destination Reachability and BGP Convergence Time. In *Proceedings of IEEE Globecom, Global Internet and Next Generation Networks*, pages 1383–1389, 2004.

[38] H. Zhang, A. Arora, and Z. Liu. A Stability-Oriented Approach to Improving BGP Convergence. In *In SRDS04*, pages 90–99, 2004.