A Measurement Study on the Impact of Routing Events on End-to-End Internet Path Performance

Feng Wang
University of Mass., Amherst
fewang@ecs.umass.edu

Zhuoqing Morley Mao
University of Michigan
zmao@eecs.umich.edu

Jia Wang
AT&T Labs-Research
jiawang@research.att.com

Lixin Gao
University of Mass., Amherst
lgao@ecs.umass.edu

Randy Bush
Internet Initiative Japan
randy@psg.com

ABSTRACT

Extensive measurement studies have shown that end-to-end Internet path performance degradation is correlated with routing dynamics. However, the root cause of the correlation between routing dynamics and such performance degradation is poorly understood. In particular, how do routing changes result in degraded end-to-end path performance in the first place? How do factors such as topological properties, routing policies, and iBGP configurations affect the extent to which such routing events can cause performance degradation? Answers to these questions are critical for improving network performance.

In this paper, we conduct extensive measurement that involves both *controlled routing updates* through two tier-1 ISPs and active probes of a diverse set of end-to-end paths on the Internet. We find that routing changes contribute to end-to-end packet loss significantly. Specifically, we study failover events in which a link failure leads to a routing change and recovery events in which a link repair causes a routing change. In both cases, it is possible to experience data plane performance degradation in terms of increased long loss burst as well as forwarding loops. Furthermore, we find that common routing policies and iBGP configurations of ISPs can directly affect the end-to-end path performance during routing changes. Our work provides new insights into potential measures that network operators can undertake to enhance network performance.

Categories and Subject Descriptors

C.2.2 [**Computer-Communication Networks**]: Network Protocols—*Routing protocols*; C.4 [**Performance of Systems**]: Reliability, availability, and serviceability

General Terms

Measurement, Experimentation, Reliability

SIGCOMM'06, September 11–15, 2006, Pisa, Italy.

Keywords

Routing dynamics, BGP, active probing, failover event, recovery event, packet loss, packet reordering

1. INTRODUCTION

The deployment of interactive services such as VoIP and gaming on the Internet demands the network to maintain good end-to-end performance. Previous studies have shown that end-to-end performance on the Internet is unpredictable [20], and degraded end-to-end path performance is correlated with routing dynamics [15, 23, 6, 16, 19, 1]. Yet the causal relationship between routing changes and degraded data plane performance has not been established. In particular, very little is known about (1) how routing changes result in degraded end-to-end path performance in the first place, and (2) how factors such as topological properties, routing policies, and iBGP configurations affect the extent to which such routing events can cause performance degradation. The Internet is a system of immense scale. Routing events such as link failures or link repairs happen quite frequently as indicated by high volumes of routing updates [19, 17]. Answers to the above questions are critical for improving network performance and wide deployment of interactive services in the Internet.

So far, researchers have taken either the analytical approach or the measurement approach to understanding the impact of routing events on end-to-end performance. Neither can answer the above questions satisfactorily. In the analytical approach, artificial topology and routing policies are used [21, 26]. In the measurement approach, only correlation between routing dynamics and degraded end-to-end performance can be established [1]. In [15], similar to our experiment methodology, routing failures are artificially injected to understand their impact on end-to-end performance. However, what and how routing dynamics cause the degraded end-to-end performance has not been fully explored. The establishment of such a causal relationship can bring insight for the design of future interdomain routing protocols.

In this paper, we aim to study end-to-end performance under realistic topology and routing policies, while not limited by the black-box approach that most of measurement studies have taken. We control routing events by injecting well-designed routing updates at known times to emulate link failures and repairs. To understand the impact of routing events on the data plane performance, we select geographic and topologically diverse probing locations from the PlanetLab experiment testbed [22] to conduct active UDP based measurement probing while routing changes are in effect. In addition, we deploy frequent ping and traceroute to probe network routing states. This allows us to identify the root cause of intermittent loss of connectivity and degraded end-to-end path performance during routing changes. Our contributions are summarized as follows.

- To analyze the impact of routing events on end-to-end performance, we investigate several metrics to characterize the end-to-end loss, delay, and out-of-order packets. We find that while routing changes can lead to longer delay and out-of-order packets, routing changes impact end-to-end loss more significantly and can lead to loss bursts lasting as long as 20 seconds. Furthermore, our results show that one routing event can lead to *multiple* loss bursts. Our results have important implications for wide deployment of interactive applications such as VoIP.

- To understand the root cause for degraded end-to-end path performance during routing changes, we characterize the kinds of routing changes that can impact end-to-end path performance. We analyze the impact of topology, routing policies, and iBGP configurations on end-to-end path performance. Our results show that routing policies and iBGP configurations are the major causes of degraded performance observed.

- To demonstrate that our results are not limited by our measurement setup, we show that degraded end-to-end performance is experienced by a diverse set of hosts when there is a routing change. Further analysis shows that simply adding physical connectivity does not necessarily minimize the impact of routing changes on end-to-end path performance.

The paper is organized as follows. Section 2 provides the background to understand the causal relationship between routing events and data plane performance. We describe our measurement methodology in Section 3. We provide detailed data analysis in Sections 4 and 5. In Section 6, we argue that our measurement results are representative enough for common network topologies and configurations. Finally we discuss related work in Section 7 and conclude with Section 8.

2. BACKGROUND

In this section, we provide background to help illustrate the correlation between routing events and data plane performance. Border Gateway Protocol (BGP) [27] is the interdomain routing protocol that Autonomous Systems (ASes) use to exchange information on how to reach destination address blocks (or *prefixes*). BGP routers at the periphery of an AS learn how to reach external destinations through *eBGP sessions* with routers in other ASes. After applying local policies to the eBGP-learned routes, a BGP router selects a single best route and advertises it to other BGP routers within the same AS through *iBGP sessions*. In the simplest case, each router has an iBGP session with every other router (i.e., a *full-mesh* iBGP configuration). Large networks are often organized internally using *route reflectors* to make iBGP more scalable. Route reflectors usually connect to each other as a full mesh. Each route reflector and its clients (i.e., iBGP neighbors that are not route reflectors themselves) form a cluster. A route reflector reflects routes learned via one client to all other clients in the same cluster as well as other route reflectors. Similarly, it also reflects routes learned from other route reflectors to all its clients.

BGP is a path vector protocol. Each BGP advertisement usually includes the sequence of ASes for the path, along with other attributes such as the next-hop IP address. Before accepting an advertisement, the receiving router checks for the presence of its own AS number in the AS path to discard routes causing loops. By representing the path at the AS level, BGP hides the details of the topology and routing information inside each network.

BGP is a stateful protocol. Only routing changes are advertised. A router sends an advertisement of a new route for a prefix or a withdrawal when the route is no longer available. To limit the number of updates that a router has to process within a short time period, a rate-limiting timer, called the Minimum Route Advertisement Interval (MRAI) timer, determines the minimum amount of time that must elapse between routing updates to a particular destination [27] for the same neighbor. This is beneficial to reduce the number of updates explored, as a single routing change might trigger multiple transient routes during the *path exploration* or route convergence process before the final stable route is selected. Note that for a given router, the MRAI timer does not limit the rate of route selection, but only the rate of route advertisement. If new routes are selected multiple times while waiting for the expiration of MRAI, the latest selected route shall be advertised at the end of MRAI. Currently, the common default values of MRAI are 30 seconds for eBGP sessions and 5 seconds for iBGP sessions. To avoid long-lived black holes, RFC 1771 [27] specifies that the MRAI timer is only applied to BGP announcements, not to explicit withdrawals. However, router implementations might apply MRAI timer to both announcements and withdrawals. We show later that the MRAI values can impact the data plane performance.

BGP is a policy-based protocol. Each BGP router selects a single best route for each prefix by comparing the routes using their attributes. Rather than simply selecting the route with the shortest AS path, routers can apply complex policies to influence the best route selection for each prefix and to decide whether to propagate it to their neighbors. The policy configuration is usually guided by the commercial agreements between ASes, which determine AS relationships. In general, there are two dominant types of relationship: *provider-to-customer* and *peer-to-peer* [8]. In the former case, a customer pays the provider to be connected to the Internet. In peer-to-peer relationships, two ASes agree to exchange traffic on behalf of their respective customers free of charge. Note that network providers offer transit service only to its customers (i.e., a network provider only announces its own and its customer prefixes to its peer ASes). There are two commonly adopted routing policies: “prefer customer” and “no-valley”. Under the “prefer customer” routing policy, routes received from a network provider’s customers are always preferred over those received from its peers or any other routes. Under the “no-valley” routing policy, customers do not transit traffic from one provider to another, and peers do not transit traffic from one peer to another either. These rules directly match the commercial incentives among the networks.

3. EXPERIMENT METHODOLOGY

In this section, we describe our controlled Internet measurement and experiment methodology involving a BGP Beacon prefix from the Beacon routing experiment infrastructure [18]. During the period of a routing event, we actively probe a host in the Beacon prefix from a diverse set of hosts on the Internet. Our measurement methodology is also applicable to other studies correlating routing dynamics with data plane performance.

3.1 Controlled Routing Changes

Beacon prefixes [18] are a set of IP prefixes designed for experi-

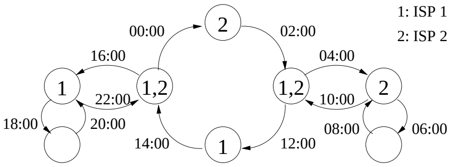


Figure 1: Time schedule (GMT) for BGP Beacon routing transitions.

Table 1: Classification of Beacon routing events

Beacon events	*BGP updates*	*Time schedule (GMT)*
Failover 1	Withdrawing route via $ISP1$	00:00, 04:00
Failover 2	Withdrawing route via $ISP2$	12:00, 16:00
Recovery 1	Restoring route via $ISP1$	02:00, 10:00
Recovery 2	Restoring route via $ISP2$	14:00, 22:00

mental purposes. There are no real users using addresses within the prefix. Their routing changes are well regulated: Beacon prefixes are announced or withdrawn every 2 hours with specific regular patterns. In our study, we use a multi-homed BGP Beacon, which has been active since September 2003. This Beacon has two tier-1 providers to which we refer as $ISP1$ and $ISP2$. Every two hours, the Beacon sends a route withdrawal or announcement to one or both providers according to the time schedule shown in Figure 1. Each circle denotes a state, indicating the providers offering connectivity to the Beacon. Each arrow represents a routing event and state transition, marked by the time that the routing event (either a route announcement or a route withdrawal) occurs.

There are 12 routing events every day. We focus on 8 routing events that keep the Beacon connected to the Internet; the other four serve the purpose of resetting the Beacon connectivity. These 8 Beacon events are classified into two categories. For a *failover beacon event*, we emulate a link failure scenario in which the Beacon originally announced routes through both providers but now withdraws the route through one of the two providers. That is, the Beacon changes from the state of using both providers to the state of using only a single provider. In a *recovery beacon event*, we emulate a link recovery or repair scenario in which the Beacon re-advertises a route previously withdrawn. That is, the Beacon changes from the state of using a single provider for connectivity to the state of using both providers. These two classes of routing changes emulate the control plane changes that a multi-homed site may experience in terms of losing and restoring a link to one or more of its providers; thus they represent real routing events on the Internet. Table 1 shows the classification of Beacon events and the time for the events.

3.2 Active Probing

The goal of active measurements is to capture the impact of routing changes on the end-to-end path performance of a diverse set of Internet paths. Knowing the time and the location of routing changes, we actively probe a host within the Beacon prefix (i.e., the Beacon host) from a set of geographically diverse sites from the PlanetLab infrastructure [22] using three probing methods: UDP packet probing, ping, and traceroute. Probing is performed every hour, thus both during injected routing events as well when there are no routing events for calibration purposes.

At every hour, every probing source sends a UDP packet stream marked by sequence numbers to the BGP Beacon host at $50msec$ interval. The probing starts 10 minutes before each hour and ends 10 minutes after that hour. (i.e., the probing duration is 20 minutes for each hour). Upon the arrival of each UDP packet, the Beacon host records the timestamp and sequence number of the UDP packet. In addition, ping and traceroute are also sent from the probe host towards the Beacon host, for measuring round-trip time (RTT) and IP-level path information during the same 20 minutes time period. Both ping and traceroute are run as soon as the previous ping or traceroute probe completes. Thus, their probing frequency is limited by the roundtrip delay and the probe response time from routers.

3.3 Data Plane Performance Metrics

In our study, we use the following metrics to measure the impact of routing events on end-to-end performance: loss, delay, and out-of-order packets. These metrics are selected as they are very basic and commonly used to capture data plane performance.

3.3.1 Packet Loss

We identify packet loss by observing gaps in sequence numbers of UDP probing packets. In our measurement, we use *bursty loss size*, which is defined as the maximum number of consecutive packets lost during a routing event.

3.3.2 Packet Delay

Ideally, we want to measure one-way delays from the probe host to the Beacon host to study the impact of routing changes on packet delays. However, such measurements are subject to limitations due to the clock skews on PlanetLab sites relative to the Beacon host. Instead, we measure roundtrip packet delays from the probe host to the Beacon host using ping probes.

3.3.3 Out-of-order Packets

Each PlanetLab probe site sends UDP packets with incrementing sequence numbers. However, packets may arrive out of order at the Beacon host. For example, if multiple paths are used for load balancing, packets can be reordered. Similarly, during routing changes, a packet sent out earlier may take a longer route compared to a later packet.

We identify out-of-order or reordered packets as follows. At the BGP Beacon host, we record the value of the next expected sequence number, which is the largest sequence number of the received packets incremented by 1. Each new in-order arriving packet has a sequence number greater than or equal to the value of the expected sequence number. The expected sequence number will be updated upon arrival of each in-order packet. A reordered packet occurs when the packet has a sequence number lower than the expected sequence number, which does not change upon the arrival of out-of-order packets. For example, for arriving packets with sequence numbers $\{1, 2, 4, 5, 3, 6\}$, packet 3 is out-of-order, as 3 is smaller than the expected sequence number of 6.

Let S_i be the sequence number of ith arriving packet. Thus, the expected sequence number S_{exp_i} is $max(S_j) + 1$, $\forall j \leq i$. Therefore, for a packet with sequence number S_i: if $S_i \geq S_{exp_i}$, it is in order. Otherwise, it is out of order.

We use two metrics to measure the degree of out-of-order delivery: *number of reordering* and *reordering offset*. The number of reordering is simply the number of packets that are considered out of order. The reordering offset measures for an out-of-order packet the difference between the actual arrival order and the expected arrival order. Using the above notation, reordering offset for S_i is $i - S_i$, assuming the initial sequence number is 1. For example, if arrived packets have sequence numbers $\{1, 2, 4, 5, 3, 6\}$, packet 3 is out-of-order, arriving as $5th$ packet, and its reordering offset is $5 - 3 = 2$. The reordering offset provides insights into the buffer size needed to restore proper order of received packets.

3.4 Identifying Routing Failures

We use a combination of active traceroute and ping measurements to identify whether packet loss bursts are caused by routing failures. Packet loss can be attributed to network congestion or routing dynamics. It has been shown that routing dynamics can lead to temporary route loss or forwarding loops [26, 1, 21]. We call such routing dynamics *routing failures*. An ideal method to identify whether a packet loss burst can be attributed to routing failures, is to correlate the loss burst with routing changes, including BGP and IGP routing information, from all routers involving in the burst. Unfortunately, identifying the root cause of packet loss requires obtaining such a large set of routing information from multiple ISPs and multiple routers, which is extremely difficult, if not impossible. Instead, we use ICMP response messages, as measured by traceroutes and pings to identify routing failures.

We derive loss bursts and correlate them with unreachable responses from traceroutes and pings. In particular, we correlate loss bursts with ICMP messages using the time window [-1 sec, 1 sec] since hosts in PlanetLab are time synchronized via NTP. When a router does not have a route entry for an incoming packet, it will send an ICMP network unreachable error message back to the source to indicate that the destination is unreachable if it is allowed to do so. Based on the ICMP response message, we can determine when and which router does not have a route entry to the Beacon host. Loss bursts that have corresponding unreachable ICMP messages are attributed to routing failures.

In addition, if a packet is trapped in forwarding loops, its TTL value will increase until the value reaches the maximum value at some router. The router will send a “TTL exceeded” message back to the source. We can observe forwarding loops from the traceroute data. In general, from traceroute and ping probes, we can determine whether a router loses its route to the Beacon host and whether there is a forwarding loop.

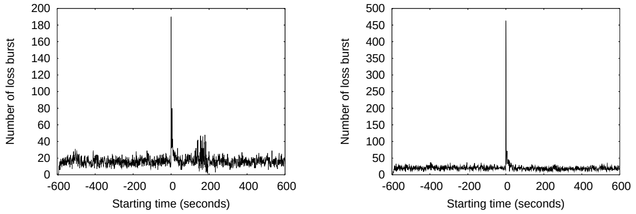
Since ICMP packets can be lost, disabled, or filtered by routers, it is possible that there is no corresponding ICMP message for some loss bursts even if those loss bursts might be caused by routing failures. As a result, we may underestimate the number of loss bursts due to routing failures. Therefore, the number of loss bursts caused by routing failures might be more than what can be identified by our methodology.

4. FAILOVER EVENTS

In this section, we characterize data plane performance during failover events. First, we observe that most packet loss bursts occur during failover events. Second, we present the extent to which packet loss is caused by routing failures. Finally, we show that routing failures can cause multiple loss bursts during one failover event. In addition, we characterize the locations that routing failures occur.

4.1 Data Plane Performance

We measure the performance (in terms of loss, delay, and packet reordering) based on UDP packet probes from 37 PlanetLab sites to the BGP Beacon during the entire month of July 2005. There are two kinds of failover events: (1) withdrawing the route advertised to $ISP1$ (denoted as “failover-1”) and (2) withdrawing the route advertised to $ISP2$ (denoted as “failover-2”). Each day, there are four failover events: two for each type. Among the 37 probing hosts, 14 hosts choose the path via $ISP1$ and 23 hosts choose the path via $ISP2$, when routes to both ISPs are announced. The withdrawal of the chosen route currently used by a host to reach the BGP Beacon forces the host to switch to the alternate, less preferred route (we refer to it as a *path change*).



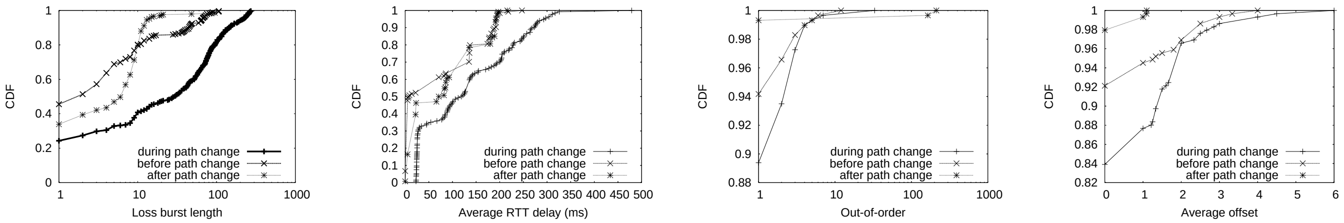
(a) Failover-1 (b) Failover-2

Figure 2: Number of loss bursts starting at each second.

At each probing host, UDP probing starts at 10 minutes before the injection of withdrawal messages and lasts for 20 minutes (i.e., till 10 minutes after the injection of withdrawal messages). To understand the packet loss around failover events, we measure the number of loss bursts starting at each second. Here, we consider consecutively lost packets as one loss burst. The time for the last received packet before the loss burst is the start time of the loss burst. Figure 2 shows the number of loss bursts over all probing hosts and failover events for the entire duration of our study. The x-axis represents the start time of a loss burst, where the start time is measured (in seconds) relative to the injection of withdrawal messages. We observe that the majority of loss bursts occur right after time 0, i.e., the time when a withdrawal message is advertised. The large number of loss bursts occurred during the time period [100 sec, 200 sec] in Figure 2(a) is most likely due to congestion because we observe no route changes in our traceroute measurements and no corresponding ICMP messages. After the failover event, traffic, including UDP probings, pings, and traceroutes, sent by probing hosts can cause congestion at some routers within $ISP2$ or the link between $ISP2$ and the Beacon. Note that there is no time synchronization problem because both the time for a loss burst occurring and the time for injecting a withdrawal message are measured by the clock on the BGP Beacon.

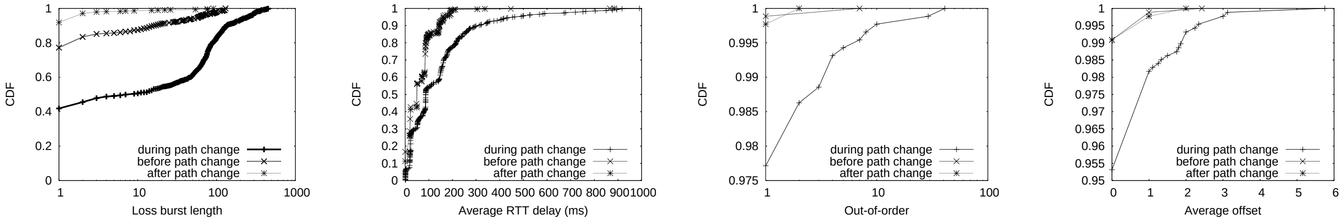
To understand the extent to which failover events can cause packet loss, we divide the time period that UDP packet probing is performed into three intervals: (1) *before path change*: the interval from the start time of UDP packet probing to the injection of withdrawal messages, (2) *during path change*: the interval from the injection of the withdrawal message to the time that path from the probing host to the Beacon is stabilized, and (3) *after path change*: the interval from the time the path from the probing host to the Beacon is stabilized till the end time of UDP packet probing. We use traceroute to estimate path change duration for each failover event, where we observe the IP-level path changing from the old stable path to the new stable path. The path change duration is measured by the time period between these two stable states. We measure the following four performance metrics during each of the three intervals of a failover event: (1) loss burst length (i.e., the number of consecutively lost packets in the loss burst), (2) round-trip delay, (3) number of reordered packets, and (4) offset of reordered packet.

Figure 3(a) shows distributions of loss burst length before, during, and after a path change for failover-1 events. The x-axis is shown in logscale. We find that the packet loss burst during path changes can have as many as 480 consecutive packets. Compared to the loss burst length during a path change, the packet loss burst length before and after a path change are quite short. Figures 3(b)-(d) show the cumulative distribution of the average round-trip delays, number of reordered packets, and the average reordering off-



(a) Loss burst length (b) Round-trip delay (c) Number of reordering (d) Reordering offset

Figure 3: Data plane performance during failover-1 events in which the route via $ISP1$ is withdrawn.



(a) Loss burst length (b) Round-trip delay (c) Number of reordering (d) Reordering offset

Figure 4: Data plane performance during failover-2 events in which the route via $ISP2$ is withdrawn.

set. We find that failover events have significant impact on packet round-trip delays. In the worst case, during path changes, packet round-trip delays can be more than 500msec. We observe that the number of reordered packets for most hosts during failover events is small. Only one PlanetLab host experiences more than 400 reordered packets after failover events, which is probably due to some anomalies along the path. However, the offset of reordered packets is larger during failover events than those before and after failover events. This indicates that path changes usually increase the degree of packet reordering and would require larger buffer sizes for real-time applications.

Figure 4 shows the performance characterization using the same metrics for failover-2 events (i.e., the route via $ISP2$ is withdrawn). Most observations we made for Figure 3 also hold here. These failover events have more impact on packet round-trip delays than the failover events when the route via $ISP1$ is withdrawn. In the worst case, the round-trip time could be 900 msec. More reordered packets are observed. Nevertheless, these reordered packets have smaller reordering offset on average. Because failover events have the most impact on loss burst length, we will focus on identifying the cause of the long packet loss bursts during path changes.

4.2 Root Causes of Loss Bursts

We correlate loss bursts with ICMP messages using the method described in Section 3.4. During the failover-1 events, 50% of loss bursts can be identified as caused by routing failures. During the failover-2 events, 52% of loss bursts are identified as caused by routing failures. To understand the extent to which routing failures affect packet loss, we focus on two kinds of routing failures: (1) loop-free routing failures and (2) forwarding loops.

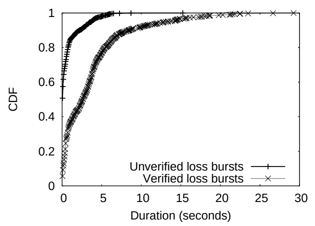
Table 2 shows the number of failover events, the number of loss bursts, and the amount of packet loss caused by routing failures. We verify that 23% of the loss bursts, corresponding to 76% of lost packets, are caused by routing failures, including both loop-free routing failures and forwarding loops. We are unable to verify the remaining 77% of loss bursts, which correspond to only 24% of packet loss. These loss bursts may be caused by either congestion or routing failures for which traceroute or ping is not sufficient (due to either insufficient probe frequency or lack of ICMP messages) for the verification.

Table 2: Overall packet loss caused by routing failures during failover events

Causes	*Failover events*	*Loss bursts*	*Lost packets*
Verified as routing failures	659 (56%)	846 (23%)	68343 (76%)
–Loop-free	451 (68%)	607 (71%)	37751 (55%)
–Forwarding loops	208 (32%)	239 (29%)	30592 (45%)
Unverified as routing failures	539 (44%)	2875 (77%)	21948 (24%)

As we will see later, more than half of the routing failures occur within $ISP1$. On the contrary, only a small portion of the routing failures occur within $ISP2$ upon withdrawal of the preferred route via $ISP2$. We continue to examine whether routing failures do occur within $ISP2$, which are not visible from ICMP messages. We use BGP updates collected from 12 routers within $ISP2$ to examine if those monitored routers experience routing failures. Among all the 724 failover events at those 12 backbone router ($2 \times 31 \times 12 = 724$), we observe 584 withdrawal messages from those monitored routers. That means that over 80% of all the failover events have routing failures. We also observe that the occurrence of withdrawal messages is right after the occurrence of failover events, and the withdrawal message is quickly replaced by an announcement. This means that during the failover events, routers within $ISP2$ indeed temporarily lose their routes to the Beacon. However, most of these transient routing failures are not visible as packet loss bursts in the data plane.

We measure the duration of a loss burst as the time interval between the latest received packet before the loss and the earliest one after the loss. Figure 5(a) shows the duration of loss bursts that can and cannot be verified as caused by routing failures. Again, we observe that the loss bursts that are verified as caused by routing fail-



(a) Loss burst verified vs.unverified as caused by routing failures

(b) Forwarding loops vs. loop-free routing failures

Figure 5: Duration for loss bursts.

ures last longer than those unverified loss bursts. Figure 5(b) further shows that loss bursts caused by forwarding loops last longer than those caused by loop-free routing failures.

4.3 How Routing Failures Occur

We use an example to illustrate the cause for packet loss during failover events. The PlanetLab host *planet02.csc.ncsu.edu* experiences packet loss during the routing failure occurred at 04:00:00 on July 30, 2005, where the host switches from the preferred path via $ISP1$ to the less preferred path via $ISP2$ after the withdrawal of the path via $ISP1$. Figure 6 shows the UDP packets received during the period from 04:00:00 to 04:00:30. The x-axis is the arrival order of each packet, and y-axis is the sequence number of the packet. We observe two major gaps (i.e., loss bursts) at 04:00:01 and 04:00:19. This can be explained by using the topology shown in Figure 7. The probing host *planet02.csc.ncsu.edu* is a customer of $ISP3$ which peers with both $ISP1$ and $ISP2$. Before the failover event, all routers except $e2$ have only one route via $e1$ to the Beacon. During the failover-1 event, the route via $ISP1$ is withdrawn by the Beacon. Routers $e1$ and $RR1$ within $ISP1$ will lose the route and explore the alternate path from router $e2$. During the path exploration, router $RR1$ cannot reach the Beacon indicated by the "Destination Host Unreachable" ping replies at time 04:00:01. After they obtain the path via peer $ISP2$, data traffic from the host to the Beacon will be forwarded via the peer link between $ISP1$ and $ISP2$. However, router $RR2$ in $ISP1$ cannot announce it to router $RR4$ in $ISP3$ because of the "no-valley" routing policy. So router $RR2$ will send a withdrawal to router $RR4$. As a result, router $RR4$ loses its route to the Beacon and is triggered to explore the alternate path. This is indicated by the "Destination Host Unreachable" ping message at time 04:00:19. In summary, the two long loss bursts shown in Figure 6 can be correlated with two unreachable ICMP responses, which indicate that these long loss bursts are caused by routing failures.

In the above example, during the round of path exploration, a router losing its routing entry is affected by the delay in obtaining the alternate route. If the router can obtain the alternate route without delay, the routing failure is not visible as packet loss bursts in the data plane. The latency in obtaining the alternate route is determined by the MRAI timer and the distance from the router that can provide the alternate route.

Based on BGP updates collected from these 12 backbone routers within $ISP2$, we identify the MRAI timer applied by $ISP2$. We observe many instances where there is little time difference between two consecutive announcements for the same prefix but with different BGP attributes. This observation implies that routers within $ISP2$ use a very small MRAI timer. This is verified by private communication with network operators of $ISP2$. The observation explains the fact that majority of large loss bursts do not occur within $ISP2$. On the other hand, most routers within $ISP1$ are Cisco routers, which have default 5 second MRAI timer. Thus, we suspect $ISP1$ does not use small MRAI timer so that routers within that ISP can experience routing failures. Our analysis here confirms the importance of MRAI timer setting on routing dynamics and subsequent impact on data plane performance.

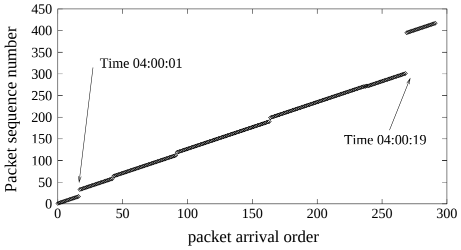


Figure 6: Long loss bursts experienced by "planet02.csc.ncsu.edu" during 04:00:00-04:00:30 on July 30, 2005.

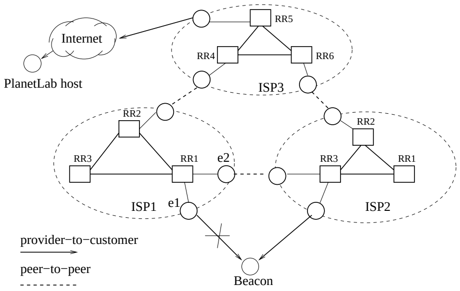


Figure 7: Topology of routers on the path from "planet02.csc.ncsu.edu" to the Beacon.

4.4 Multiple Loss Bursts Caused by Failover Events

As we have observed in the above example, a host can experience multiple loss bursts after the injection of a withdrawal message. This can be explained by the widely used "no-valley" routing policy. That is, when an AS obtains an alternate route from its peer, it cannot transit the route to another peer. So it will send a withdrawal message to the peer to invalidate the previous route. Thus, the withdrawal message can trigger the second loss burst within the peer.

In this section, we measure the number of loss bursts that a host experiences for each failover event. Figure 8 shows that, in over 75% of the cases, a hosts experiences fewer than two loss bursts as a result of a failover event, while a host can experience up to 6 loss bursts as a result of a single failover event. Figure 9 shows the percentage of packet loss in the first two loss bursts experienced by PlanetLab hosts for each failover event. We observe that the first two bursts contribute to the majority of packet loss. In the rest of the paper, we will focus on the first two loss bursts for failover events.

Among the first two loss bursts during failover-1 events, we can verify about 57% of the first loss bursts as caused by routing failures, and about 40% of the second loss bursts as caused by routing failures. The figures for the failover-2 events are 61% and 42%, respectively. In general, we observe that the number of the first loss bursts that can be verified as due to routing failures is larger than that of the second loss bursts.

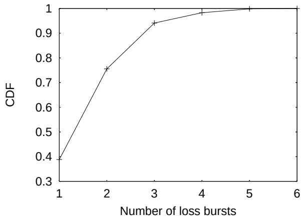
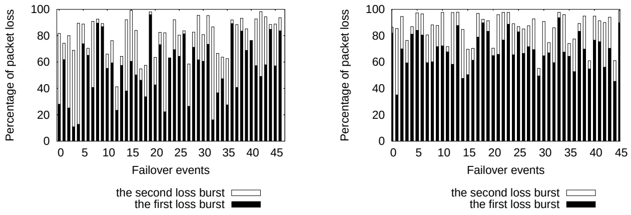


Figure 8: Multiple loss bursts may be experienced by end hosts.



(a) Failover-1

(b) Failover-2

Figure 9: Percentage of packet loss contributed by the first two loss bursts.

Table 3: Location of the first loss bursts caused by routing failures during failover events

Class	$ISP1$	$ISP2$	*Other tier-1*	*Non tier-1*
Failover-1	92%	0	5%	3%
Failover-2	0	9%	73%	18%

Table 4: Percentage of failover events involving routing failures for three different hosts.

Class	*Failover events*	*Events causing routing failure*
Customer of either $ISP1$ or $ISP2$	206	111 (53%)
Multihomed to $ISP1$ and $ISP2$	225	43 (19%)
Customer of other ISPs	1054	463 (43%)

One interesting observation is that when the second routing failures are triggered due to the "no valley" routing policy, traffic on data plane may traverse a path which violates the "no valley" routing policy. For example, in the previous example shown in Figure 7, we find that between the first large loss burst (at 04:00:01) and the second large loss burst (at 04:00:19), about 250 UDP packets are received by the Beacon. During this period of time, UDP packets traverse a path across three tier-1 ASes, "$ISP3$ $ISP1$ $ISP2$", which violates the "no-valley" routing policy. After the second loss burst, the IP-level path will change to traverse two tier-1 ASes, "$ISP3$ $ISP2$". The number of packets that traverse this "valley" path is determined by the delay of sending the withdrawal. In the worst case, if the MRAI timer are applied to withdrawal messages, the withdrawal message can be delayed as long as 30 seconds (the default MRAI timer applied on eBGP session).

4.5 Location of Routing Failures

After identifying routing failures causing packet loss during failover events, we further investigate a number of factors (e.g., interior network topologies, iBGP configurations, or MRAI timers) that can cause routing failures by analyzing the location of routing failures. From the source address of each ICMP message, we can identify which router loses its route entry if the ICMP message is unreachable, or which router is trapped in a forwarding loop if the ICMP message is an exceeded TTL or a forwarding loop is observed in traceroute data. We then derive the locations where routing failures occur according to the DNS name of the corresponding IP addresses. During each failover event, among all the first loss bursts caused by routing failures, we measure how many of them occur within $ISP1$, $ISP2$, or other ASes. As shown in Table 3, 92% of the first routing failures occur within $ISP1$ during failover-1 event, while the figure for $ISP2$ during failover-2 events is 9%. During failover-2 events, most of the first loss bursts occur within $ISP2$'s neighbors and those neighbors are tier-1 ASes. In addition, we find that about 55% and 96% of the second routing failures occur within other tier-1 ASes during failover-1 and failover-2 events, respectively. This means that routing failures occurring within $ISP1$ or $ISP2$ are propagated to their neighboring ASes.

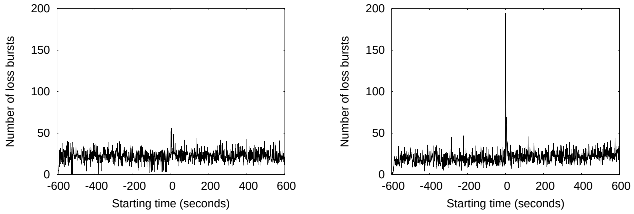
In addition, we evaluate the occurrence of routing failures from BGP updates, which are cascaded from $ISP1$ or $ISP2$ to other ASes. We first examine routing failures from BGP updates collected from 52 backbone routers within a tier-1 ISP. We observe that 134 withdrawal messages come from 4 monitored routers. We then use BGP updates from Oregon RouteView to examine routing failures occurring at other ASes. We observe 210 withdrawal messages from 7 ASes, which do not include $ISP1$ and $ISP2$. Those observations imply that routing failures during failover events indeed can be cascaded to other ASes.

We further classify PlanetLab hosts into three categories according to their connection to $ISP1$ and $ISP2$: (1) single-homed to either $ISP1$ or $ISP2$; (2) multi-homed to both $ISP1$ and $ISP2$; and (3) customer of other ISPs. In our measurement, the number of PlanetLab hosts in these three categories are 6, 6, and 25, respectively. Table 4 shows the number of failover events in which there is at least one loss burst caused by routing failures. We observe that every category of hosts experience packet loss caused by routing failures, and, as expected, multi-homed hosts experience less packet loss than the other two categories.

4.6 Methodology Evaluation

In this section, we evaluate our approach to correlating ICMP messages with loss bursts. We identify packet loss caused by routing failures by correlating loss with ICMP unreachable messages. The rationale is that if a destination network is unreachable from a router according to its routing table, an ICMP unreachable destination error message will be sent back to the source host. We assess the number of ICMP messages in the absence of routing change (i.e., at times other than the failover events). Recall that ping packets are sent to the Beacon when there is no Beacon event (10 minutes before and after 01:00, 03:00, 05:00, 09:00, 11:00, 13:00, 15:00, 17:00, 21:00, and 23:00). We observe a total of 3801 ICMP messages in our measurement during the period where there is no faildown events (during which the Beacon is completely withdrawn from both ISPs), only 0.6% of which are not caused by Beacon events. Thus, ICMP unreachable messages provide a good indication for routing failures.

Another issue that might introduce bias to our measurement is that some ISPs disable ICMP replies from their routers. We expect such policy to be typically uniformly applied to all the routers within a given ISP. In our measurement, we observe that ICMP messages come from 726 routers belonging to 68 ASes, and about 53% of those routers belong to 10 tier-1 ASes. In particular, 70 routers within $ISP1$ (i.e., 52% of $ISP1$'s routers visible in our measurement) generate ICMP messages. The corresponding figure for $ISP2$ is 24 routers (95% of $ISP2$'s routers observed). Given such a good coverage of ASes responding with ICMP messages and a high coverage from both $ISP1$ and $ISP2$, we conjecture that our



(a) Recovery-1

(b) Recovery-2

Figure 10: Number of loss bursts starting at each second during recovery events.

measurement is not significantly biased by ICMP blocking.

5. RECOVERY EVENTS

In this section, we investigate the end-to-end path performance during recovery events. In particular, we focus on those events which result in end-to-end path changes, i.e., the probe host chooses the restored provider when routes through both providers are available. Thus, the recovery event causes the probe host to switch to using the restored provider. We first present packet loss, delay, and packet reordering for all such events. Then we analyze the root causes of packet loss during recovery events.

5.1 Data Plane Performance

Similar to our analysis on failover events, We measure the performance (in terms of loss, delay, and packet reordering) based on UDP packet probes from 37 PlanetLab sites. There are two kinds of recovery events: (1) announcing the route to $ISP1$ (denoted as “recovery-1”) and (2) announcing the route to $ISP2$ (denoted as “recovery-2”). Among the 37 probing hosts, 12 hosts choose the restored path via $ISP1$ and 25 hosts choose the restored path via $ISP2$ when routes to both ISPs are available.

Figure 10 shows the number of loss bursts during recovery events across all probe hosts undergoing path changes for the entire duration of our study. Similar to our analysis on failover events, the x-axis represents the start time of a loss burst, where the start time is measured (in seconds) relative to the injection of announcement messages. During recovery-1 events in Figure 10(a), we do not observe a large number of loss bursts after time 0, i.e., the time when an announcement message is advertised. However, during recovery-2 events in Figure 10(b), we observe that loss bursts occur right after time 0, and can last for 100 seconds. Even though the magnitude of the loss burst peak may vary during two kinds of recovery events, more than half of end hosts (i.e., 29 PlanetLab hosts) in our measurement experience packet loss during recovery events.

To understand the extent to which recovery events can cause packet loss, we divide the period of time that UDP packet probing is performed into three intervals as we do for failover events: (1) *before path changes*, (2) *during path changes*, and (3) *after path changes*. Figure 11 (a) and Figure 12 (a) show the loss burst length during routing changes, compared with those before and after routing changes. We observe that the loss burst length during routing change does not show significant difference compared with those before or after routing change. Figure 11(a) also shows that there are more packets dropped after path change, which is most likely due to congestion because we observe no routing changes in our traceroute measurements and no corresponding ICMP messages. Recall that we have similar observations on the failover events shown in Figure 2(a). In addition, loss burst length can be as long as 180 packets and 140 packets for recovery-1 and recovery-2 events respectively. Such loss is most likely caused by routing failures. It is quite counter-intuitive that even during recovery events when both routes are available, packets would experience such long-lived loss bursts.

Table 5: Packet loss caused by routing changes during recovery events

Causes	*Recovery events*	*Loss bursts*	*Loss packets*
Verified as routing failures	41 (12%)	76 (4%)	1120 (26%)
–loop-free	17 (41%)	39 (51%)	480 (43%)
–forwarding loop	24 (59%)	37 (49%)	640 (57%)
Unverified as routing failures	290 (88%)	1714 (96%)	3266 (74%)

In addition to loss burst length during recovery events, we also measure packet round-trip delays and reordered packets. Figure 11 (b) shows performance using round-trip delay metric for the recovery events when the route via ISP1 is re-announced. We observe that the distribution of packet delays is similar compared to those for failover events, i.e., recovery events have impact on packet round-trip delays. However, from Figure 11 (c) and (d), the average reordering offset or the degree of reordering is smaller for recovery, about 2 compared to 6 for failover. The total amount of reordering is also significantly less. Interestingly we find reordering during recovery events is slightly smaller than that during normal time, indicating that recovery events do not contribute much to packet reordering.

Figure 12 shows the corresponding metrics for those events when the route via ISP2 is re-announced. Similar to Figure 11, we observe that packet delays are not different from that for failover events. The average offset for out-of-order packets is no more than three packets. From all the performance metrics, we find that recovery events have the most impact on loss burst length. Next, we identify the cause for packet loss during such routing changes.

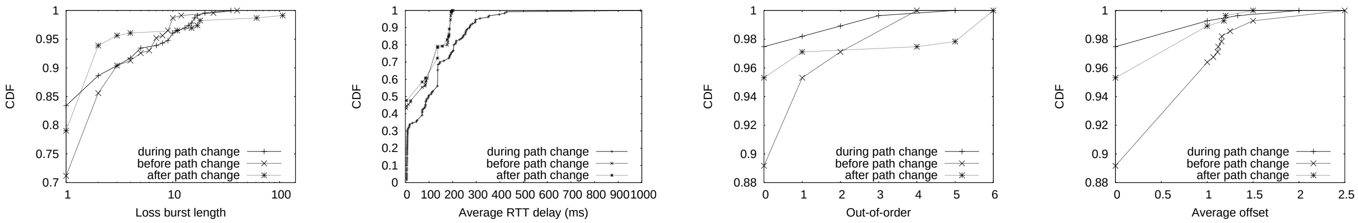
5.2 Root Causes of Loss Bursts

In general, we observe that during recovery events loss bursts are long. This motivates us to analyze whether some of packet loss bursts are caused by routing failures. This may appear to be unlikely since paths to both providers are available: the old route going through the less preferred provider is still usable while the routers switch to the more preferred, newly announced route. However, our measurement results show that routing failures indeed occur during recovery events.

Similar to our analysis on failover events, we correlate ICMP unreachable messages with loss bursts. From Table 5, we observe that 26% of packet loss is verified to be caused by routing failures. Note that the number of packet loss caused by routing failures might be more than what can be identified by our heuristic because ICMP messages may be filtered by some routers in the Internet.

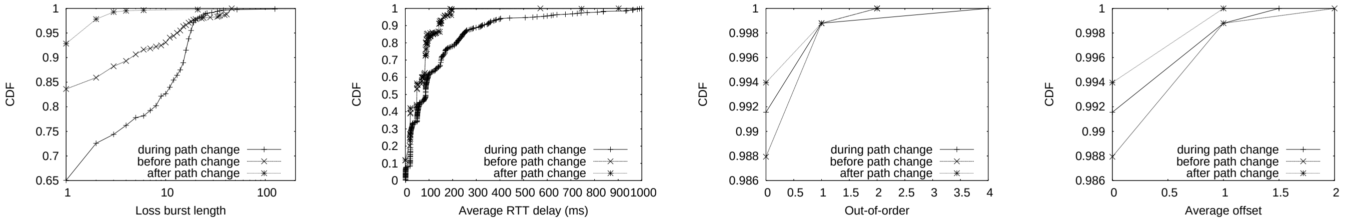
In addition, we evaluate routing failures from BGP updates, which are collected from 12 routers within $ISP2$. Over 724 recovery events, we observe 12 BGP withdrawals sent by those monitored routers. We also observe that there is little time difference between the withdrawal and the following announcement for the same prefix but with different BGP AS path. While the occurrence of routing failures during recovery event is rare, the routers within $ISP2$ indeed temporarily lose their routes to the Beacon.

As we do for failover events, we measure the duration of a loss burst as the time interval between the latest received packet before the loss and the earliest one after the loss. Figure 13(a) shows the cumulative distribution of the duration of loss bursts that are both



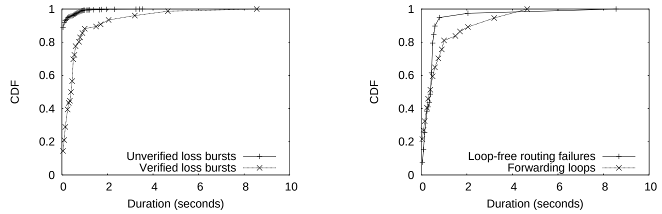
(a) Loss burst length (b) Round-trip delay (c) Number of out-of-order packet (d) Offset of Out-of-order

Figure 11: Performance during the recovery-1 events when the route via $ISP1$ is re-announced.



(a) Loss burst length (b)Round-trip delay (c) Number of out-of-order (d) Offset of Out-of-order

Figure 12: Performance during the recovery-2 events when the route via $ISP2$ is re-announced.



(a) Loss bursts verified vs. unverified as caused by routing failures (b) Forwarding loops vs. Loop-free routing failures

Figure 13: Duration of loss burst.

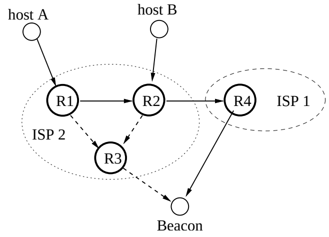


Figure 14: Topology for explaining packet loss burst during recovery.

verified and unverified to be caused by routing failures. We observe that verified loss bursts on average are longer than those unverified. In addition, during recovery events, more than 98% of routing failures last less than 5 seconds, while during failover events, about 80% of routing failures last less than 5 seconds as shown in Figure 5. This means that loss bursts caused by routing failures during recovery events last much shorter than those caused by routing failures during failover events. We also observe that unverified loss bursts last less than 4 seconds.

Figure 13(b) shows the duration of verified loss bursts that are caused by loop-free routing failures and forwarding loops. We observe that 57% of packet loss is due to forwarding loops, which is a little higher than that for failover events (47%). This implies that forwarding loops are also quite common during recovery events.

5.3 How Routing Failures Occur

Here, we illustrate an example to show how packets can be dropped due to path changes during recovery events. Figure 14 shows a simplified topology. In this figure, solid arrow lines indicate the path used before a recovery event, while dashed arrow lines represent the path used after the event. We analyze packet loss experienced by *vnl.cs.wust1.edu* (shown as PlanetLab host A in Figure 14) when a recovery event occurs. Before the event, the host reaches the Beacon via $ISP2$ followed by $R4$ in $ISP1$. After the event, the host reaches the Beacon directly via $ISP2$.

When the recovery event occurs, $R3$ receives the new path, and send it to both $R2$ and $R1$. Suppose that $R3$ waits for the expiration of MRAI timer to send the new route to $R1$. At the same time, suppose that $R3$ can send the new route to $R2$ because the MRAI timer has just expired. Note that the MRAI timer is maintained for each BGP session. As a result, $R2$ obtains the new route and switches to it. However, $R2$ cannot forward the new route to other iBGP routers so that it will send a withdrawal message to $R1$ to poison its previous route. Suppose there is no delay for the withdrawal, $R1$ loses its route entry to the Beacon until it obtains the new route from $R3$.

In this example, host A can experience packet loss during a recovery event, while host B may not. However, if there is no physical link between $R2$ and $R3$, the iBGP sessions between the two routers is via $R1$, host B still can experience packet loss because its packet to the Beacon is routed via $R1$, which can lose its route entry. The logical fully meshed iBGP sessions are widely deployed within large ISPs.

5.4 Multiple Loss Bursts Caused by Routing Failures

We find that multiple packet loss bursts can occur during recov-

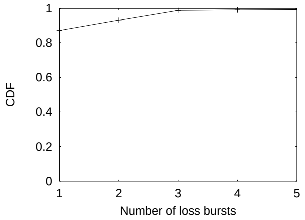


Figure 15: Multiple loss bursts may be experienced by end hosts during recovery events.

Table 6: Location of routing failures during recovery events

Class	$ISP1$	$ISP2$	*Other tier-1*	*Non tier-1*
Recovery-1	90%	0	0%	10%
Recovery-2	0	0%	59%	42%

ery events. As shown in Figure 15, we identify that some hosts can experience up to 5 loss bursts. Again, we focus on the first and second loss bursts, which contribute to the majority of packet loss. About 16% of the first loss bursts are identified as due to routing failures, while 8% of the second loss bursts are identified as caused by routing failures. This means that one recovery event can also cause *multiple routing failures*. According to our measurement, we find that more than half of the second routing failures are transient forwarding loops, and those forwarding loops last no more than 5 seconds. The reason is that when a route loses it routing entry during a recovery event, which is the first routing failure, it can propagate the withdrawal message to its neighboring ASes. When the neighbor receives the withdrawal message, it will explore available routes if there is none in its routing table. During the exploration, forwarding loops may occur.

5.5 Location of Routing Failures

We investigate the location of routing failures. Here, we present the results for both recovery events. According to the DNS name of the router from which a unreachable ICMP is sent, we identify the AS to which the router belongs. Table 6 shows the location for the first loss due to routing failures. We observe that for recovery-1 events, about 90% of routing failures occur within $ISP1$, while no routing failures are observed within $ISP2$ during recovery-2 events. Compared with routing failures during failover events, we find similar results indicating that a very small number of routing failures occurs within $ISP2$. On the other hand, routing failures do occur within $ISP2$'s neighbors, and the starting time of the first routing failures is just right after the time when the announcement is advertised. For the second loss burst due to routing failures, we find that all of them occur within $ISP1$'s and $ISP2$'s neighbors, and most of them are within $ISP2$'s neighbors.

From the example shown in Figure 14, we know that routing failures during recovery events depend on MRAI timer values. The new route is advertised to some routers without delay, while it is delayed for other routers due to the MRAI timer. As we mention in Section 4, we find that the MRAI timer applied in $ISP2$ is very small. This means that the new recovered route can be advertised to all routers within $ISP2$ with little delay. As a result, $ISP2$ does not experience any routing failures. Similarly, the new route can be advertised to its neighbors without delay. This explains the observation that the first loss bursts during recovery-1 events are right after the time when an announcement is sent, as shown in Figure 10(b). On the other hand, during recovery-1 events, most of the first loss bursts caused by routing failures occur within $ISP1$. Those routing failures can be explained by the delay due to the MRAI timer. Furthermore, the routes advertised by $ISP1$ or $ISP2$ can cause their neighbors to experience routing failures again.

6. REPRESENTATIVENESS OF THE BEACON EXPERIMENTS

In this section, we discuss the representativeness of the Beacon experiments. Beacon has one link to each of its two tier-1 providers. In general, an AS can have multiple upstream providers and/or connect to a single provider via multiple links. Therefore, the connectivity of destination prefixes in the Internet could be much more complex than the Beacon topology. We first characterize how destination prefixes are connected to upstream providers. We then investigate how the results or insight gained on the Beacon experiments can be applied in general. Finally, we discuss methods to avoid transient routing failures.

6.1 Characterizing Connectivity of Destination Prefixes

A prefix is *single-homed* if its origin AS advertises it to a single upstream provider. A prefix is *multi-homed* if it is advertised to multiple providers. On the other hand, a prefix can be advertised to a provider through multiple links. We classify destination prefixes into four categories according to the characteristics of their connectivity:

- Single-homed prefixes via a single upstream link.
- Single-homed prefixes via multiple upstream links.
- Multi-homed prefixes via a single upstream link.
- Multi-homed prefixes via multiple upstream links.

We characterize the connectivity of prefixes originated from a customer of a large tier-1 ISP. In particular, we examine how the ISP's customers advertise their prefixes to it. Within the ISP, we use a BGP monitor that has iBGP sessions to some of top-level backbone routers and to edge routers connecting to peer networks. A snapshot of BGP routing table from each router are collected on a daily basis. Thus, we are able to see all available routes for each destination prefix we examined. The results presented in this section are based on data collected on January 15, 2006.

Table 7 shows that over half of prefixes originated by customers of the tier-1 ISP are single-homed prefixes. In particular, 48% of prefixes are single-homed via a single upstream link. This is consistent with the observation on a different tier-1 ISP in a previous study [2]. We observe that 6% of prefixes are single-homed prefix via multiple upstream links, and the corresponding figures for multi-homed prefixes are 29% and 17%, respectively. Since single-homed prefixes via single upstream link do not have route redundancy, in the remainder of the paper, we will focus on single-homed prefixes via multiple upstream links and multi-homed prefixes. In the next two subsections, we show that the insight gained from the Beacon experiments can be applied to both multi-homed prefixes with a single upstream link and prefixes with multiple upstream links.

6.2 Routing Failures During Failover Events

Let's first look at an AS multihomed to a set of providers, which could be tier-1 or non-tier-1 ISPs. Each of those providers will learn more routes either from its peers or from its providers if it is a non-tier-1 ISP. Thus, multihoming to a set of providers can increase route diversity. However, multihoming only increases the number

Table 7: Connectivity of destination prefixes from a tier-1 ISP's customers.

Class	*A single upstream link*	*Multiple upstream links*
Single-homed prefixes	48%	6%
Multi-homed prefixes	29%	17%

of routes learns from peers or providers. Those routes are less preferable than that from customers when the "prefer-customer" routing policy is used. The higher preference of customer routes forces those routes with low preference invisible to other routers within the provider's network. Thus, during a failover event, a multi-homed prefix via a single upstream link will experience routing failure just like the Beacon prefix, even though it might have more providers than the Beacon.

On the other hand, connecting to a single provider at multiple locations will help it to avoid routing failures. The reason is that routes learned from the same customer typically have the same local preference so that they are visible to other routers in the provider's network. However, routing failures might still occur in this case. For example, BGP attributes, such as AS path length, BGP MED, IGP weight, or router ID, can lead to only one route available in a router. One possible method to avoid routing failures is to use the "hot potato" routing policy. The "hot potato" routing policy is applied to routes coming from the same AS, and each router selects the best route based on IGP distance. If router reflectors are used, for each route reflector, routes learned from its clients are always preferred over routes from others [13], the route reflector will see other available routes. However, applying hot potato policy to routes from the same customer may not be efficient because customers typically have limited connections within a geographic area.

6.3 Routing Failures During Recovery Events

In general, route diversity will be increased when a route is recovered. However, some routers may temporarily lose their routes to the destination during route recovery. Here, we use an example to show how route failures can occur in general during a recovery event. Figure 16 illustrates an AS with $k + m$ fully connected routers. Suppose that routers 1, ..., k use routes learned from other ASes to reach destination d. Here, the destination could be a single-homed prefix with multiple upstream links or a multi-homed prefix. Routers $k + 1$, ..., $k + m$ use routes from those k routers to reach d. Because of the fully meshed topology, all routers will have k available routes to the destination.

When a new route is advertised to router $k + m$, the new route may have the highest local preference. For example, routers 1, ..., k learn their routes from peers, and router $k + m$ learns the new route from a customer, or router $k + m$ has lowest router ID if all the routes come from customers. Thus, all routers will switch to the new route after they learn it from router $k + m$. Suppose that router $k + m$ propagate the new route to routers 1, ..., k without any delay, but postpones sending the new route to routers $k + 1$, ..., $k+m-1$ due to the MRAI timer. When routers 1, ..., k switch to the new route, they will send withdrawal messages to routers $k + 1$, ..., $k+m$ to poison their previously advertised route. If the withdrawal messages arrive at routers $k + 1$, ..., $k + m$ routers earlier than the new route does, routers $k + 1$, ..., $k + m$ will temporarily lose their routes to d.

As we have seen from the above example, both the location and preference of the recovered route can impact the occurrence of transient routing failures during recovery events. For example, if the recovered route is learned from router k, then no router will lose any route to d, and every router has k available routes during the recovery event. Or, if the recovered route does not have higher local preference, then only a subset of routers $1, 2, \ldots, i$ ($i < k$) switch to it.

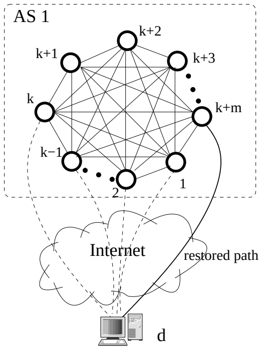


Figure 16: Routers may lose existing routes during recovery events.

6.4 Discussion

In summary, simply adding physical connectivity might not be sufficient in minimizing the impact of routing failures. Routing policies, iBGP configurations, MRAI timer values, and failure locations can have significant impact on the routing failures. We observe that the MRAI timer plays a crucial role during failover and recovery events. Because the MRAI timer is configured on a per BGP session basis and often shared across prefixes, the delay for the announcement is determined by BGP traffic of the session. During time periods of high BGP traffic volume, routing updates are most likely to be delayed. Of course, if the MRAI timer is small, the alternate path can be quickly obtained so that the probability of incurring routing failures is small. For example, in our measurement, we observe that $ISP2$ has very small MRAI timer values so that it seldom has routing failures during recovery event. Clearly, applying MRAI timers at a coarser granularity such as session based can save memory resources on routers, but it does have a negative effect on routing. Furthermore, the value of the timer applied to BGP updates can directly affect the failure duration. Our analysis implies that there is a need to reevaluate the mechanism to which MRAI timer is applied and the value of the timer. Another possible way to minimize the impact of routing failures during failover and recovery events is to store not only the best path but also the second best one at each router [24]. This can potentially reduce the chance that a router loses both routes at the same time.

7. RELATED WORK

Similar to our work, a number of measurement studies have *correlated* routing instability and end-to-end performance [15, 23, 6, 16, 19, 1, 25]. Labovitz *et al.* studied BGP route instability, focusing on the stability of paths between Internet Service Providers and artificially injected routing failures to discover their effects on Internet path performance [15]. Markopoulou *et al.* has characterized failures that are correlated with IS-IS routing updates [19]. Feamster *et al.* studied the location and duration of end-to-end path failures and correlated such failures with BGP routing instability [6]. Teixeira *et al.* measured the effects of intradomain routing instability, but did not examine how this instability affects end-to-end performance [25]. Agarwal *et al.* correlated BGP routing changes with packet traces at aggregate level from a large backbone ISP and found that BGP routing instability usually has little impact on traffic shifts within a single AS [1]. Boutremans *et al.* used active and passive measurements to study the impact of network congestion, link failures, and IS-IS routing instability on voice over IP service on a tier-1 backbone network [3]. In contrast to the above existing work, our work focuses on how routing events such as link failures and repairs affect *end-to-end* Internet performance. Our work is partly motivated by the work by Paxson, who identified Internet failures, routing loops, and routing pathologies using end-to-end traceroutes and discovered that routing instability can disrupt end-to-end connectivity [20]. We take a step further by exploring the root causes in the form of topological properties for the data plane performance degradation due to routing changes.

There are also several studies on BGP routing instability [17, 16, 14, 12, 10, 11, 9, 5, 7]. For example, Labovitz *et al.* conducted a series of empirical studies on characterizing interdomain instability and the impact of policy and topology on convergence delays [17, 16]. Alternatively, Griffin *et al.* [14, 12] took theoretical approaches in explaining BGP dynamics observed empirically on the Internet. Gao *et al.* [10, 9] proposed guidelines and models for setting local routing policies in each AS to increase routing stability and reliability. These work is related to ours; however, they mainly focused on characterizing BGP dynamics and identifying root causes of such dynamics without further investigating in detail the impact of such dynamics on end-to-end performance.

8. CONCLUSIONS

Despite the fact that an increasing number of Internet applications, such as VoIP and gaming, rely on high availability of end-to-end paths, there is a lack of understanding of which and how routing events affect end-to-end path performance. In this work, we conduct extensive measurement involving both controlled routing updates through two tier-1 ISPs and active probes of a diverse set of end-to-end paths on the Internet to illuminate the impact of routing changes on data plane performance.

We find that during failover and recovery event, routers can experience routing failures. Based on our measurement, routing failures contribute to end-to-end packet loss significantly. During both failover events and recovery events, multiple loss bursts are likely observed and loss bursts can be significantly longer than those observed during recovery events. Multiple loss bursts can occur at different ASes. Furthermore, we show that common iBGP configuration and MRAI timer values play a major role in causing packet loss during routing events. Our study suggests that extending BGP to accommodate routing redundancy may eliminate majority of end-to-end path failures caused by routing events. The RCP architecture introduced in [4] is a potential candidate for providing redundancy within an AS.

Acknowledgements

We would like to thank the support from PlanetLab staff. We also thank Pei Dan and anonymous reviewers for their constructive comments. We are very grateful to Tim Griffin for his valuable comments on the experimental methodology and for initiating the Beacon project. The work is partially supported by NSF grants CNS-0325868, CNS-0430204, ANI-0208116, ANI-0085848, and the Alfred P. Sloan Fellowship.

9. REFERENCES

[1] Agarwal, S., Chuah, C.-N., Bhattacharyya, S., and Diot, C. The Impact of BGP Dynamics on Intra-Domain Traffic. In *Proceedings of ACM SIGMETRICS* (New York, NY, USA, June 2004).

[2] Agarwal, S., Nucci, A., and Bhattacharyya, S. Measuring the Shared Fate of IGP Engineering and Interdomain Traffic. In *ICNP* (2005), pp. 236–245.

[3] Boutremans, C., Iannaccone, G., and Diot, C. Impact of link failures on VoIP performance. In *Proceedings of ACM NOSSDAV* (May 2002).

[4] Caesar, M., Caldwell, D., Feamster, N., Rexford, J., Shaikh, A., and van der Merwe, J. Design and implementation of a Routing Control Platform. In *Proc. Networked Systems Design and Implementation* (2005).

[5] Chang, D. F., Govindan, R., and Heidemann, J. The temporal and topological characteristics of BGP path changes. In *Proceedings of IEEE ICNP* (November 2003).

[6] Feamster, N., Andersen, D., Balakrishnan, H., and Kaashoek, M. Measuring the Effects of Internet Path Faults on Reactive Routing. In *Proceedings of ACM SIGMETRICS* (San Diego, CA, June 2003).

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

[8] Gao, L. On Inferring Autonomous System Relationships in the Internet. *IEEE/ACM Transactions On Networking 9*, 6 (December 2001).

[9] Gao, L., Griffin, T., and Rexford, J. Inherently Safe Backup Routing with BGP. In *Proceedings of IEEE INFOCOM* (2001).

[10] Gao, L., and Rexford, J. A Stable Internet Routing without Global Coordination. *IEEE/ACM Transactions On Networking 9*, 6 (December 2001), 681–692.

[11] Griffin, T., and Wilfong, G. T. A Safe Path Vector Protocol. In *Proceedings of IEEE INFOCOM* (2000), pp. 490–499.

[12] Griffin, T. G., Shepherd, F. B., and Wilfong, G. The Stable Paths Problem and Interdoman Routing. *IEEE/ACM Transactions on Networking 10*, 2 (April 2002), 232–243.

[13] Griffin, T. G., and Wilfong, G. On the correctness of IBGP configuration. In *SIGCOMM '02: Proceedings of the 2002 conference on Applications, technologies, architectures, and protocols for computer communications* (2002), pp. 17–29.

[14] Labovitz, C., and Ahuja, A. The Impact of Internet Policy and Topology on Delayed Routing Convergence. In *Proceedings of IEEE INFOCOM* (Anchorage, Alaska, April 2001).

[15] Labovitz, C., Ahuja, A., Bose, A., and Jahanian, F. Delayed Internet routing convergence. *IEEE/ACM Transactions on Networking 9*, 3 (June 2001), 293–306.

[16] Labovitz, C., Ahuja, A., and Jahanian, F. Experimental Study of Internet Stability and Backbone Failures. In *Proceedings of FTCS* (1999), pp. 278–285.

[17] Labovitz, C., Malan, G. R., and Jahanian, F. Internet Routing Instability. *IEEE/ACM Transactions on Networking 6*, 5 (1998), 515–528.

[18] Mao, Z. M., Bush, R., Griffin, T., and Roughan, M. BGP Beacons. In *Proceedings of IMC* (2003).

[19] Markopoulou, A., Iannaccone, G., Bhattacharyya, S., Chuah, C., and Diot, C. Characterization of Failures in an IP Backbone, 2004.

[20] Paxson, V. End-to-end routing Behavior in the Internet. *IEEE/ACM Transactions on Network 5*, 5 (1997), 601–615.

[21] Pei, D., Zhao, X., Massey, D., and Zhang, L. A Study of BGP Path Vector Route Looping Behavior. In *ICDCS* (2004), pp. 720–729.

[22] PlanetLab. `http://www.planet-lab.org`.

[23] Roughan, M., Griffin, T., Mao, Z. M., Greenberg, A., and Freeman, B. Combining Routing and Traffic Data for Detection of IP Forwarding Anomalies. In *Proceedings of ACM SIGCOMM NeTs Workshop* (2004).

[24] Shand, M., and Bryant, S. IP Fast Reroute Framework. Internet Draft draft-ietf-rtgwg-ipfrr-framework-04.txt, October 2005.

[25] Teixeira, R., Shaikh, A., Griffin, T., and Rexford, J. Dynamics of hot-potato routing in IP networks, 2004.

[26] Wang, F., Gao, L., Wang, J., and Qiu, J. On Understanding of Transient Interdomain Routing Failures. In *Proceedings of IEEE ICNP* (2005).

[27] Y. Rekhter and T. Li. A border gateway protocol 4 (BGP-4). *RFC 1771* (1995).