BGP Beacons

Z. Morley Mao[*], Randy Bush[†], Timothy G. Griffin[‡], Matthew Roughan[§]

Draft: submitted for publication, please do not distribute or cite.

Abstract

The desire to better understand global BGP dynamics has motivated several studies using active measurement techniques, which inject announcements and withdrawals of prefixes from the global routing domain. From these one can measure quantities such as the BGP convergence time. Previously, the route injection infrastructure of such experiments has either been temporary in nature, or has use restricted to the experimenters. The routing research community would benefit from a permanent and public infrastructure for such active probes. We use the term *BGP Beacon* to refer to a publicly documented prefix having global visibility and a published schedule for announcements and withdrawals. A BGP Beacon is to be used for the ongoing study of BGP dynamics, and so should be supported with a long term commitment. We describe several BGP Beacons that have been set up at various points in the Internet. We then describe techniques for processing BGP updates when a BGP Beacon is observed from a BGP monitoring point such as Oregon's Route Views. Finally, we illustrate the use of BGP Beacons in the analysis of convergence delays, route flap damping, and update inter-arrival times.

1 What is a BGP Beacon?

The Border Gateway Protocol (BGP) [1, 2, 3] is central to the stability and robustness of the Internet. Passive monitoring of BGP updates has resulted in important insights into the dynamics of BGP [4, 5, 6]. Several public sources, such as Oregon's Route Views [7] and the RIPE Routing Service [8], provide BGP updates collected from a large number of points in the Internet. Passive measurements are not sufficient for all purposes and so active techniques have also been employed in the analysis of BGP dynamics [9, 10]. With the active approach, prefixes are announced and withdrawn from the global routing domain while quantities such as convergence time are measured. The main advantage of the active approach is the the *input* to the routing system is known (for example, an announcement or a withdrawal), which allows inferences to be made that would be difficult or impossible with purely passive measurements.

To date, the route injection infrastructure of such experiments has either been temporary in nature, or has its use restricted to the experimenters. Mounting such an infrastructure is often beyond the means of many interested in this area of research, and so we feel that the routing research community would benefit from a permanent and public infrastructure for such active routing probes. We use the term *BGP Beacon* to refer to a publicly documented prefix having global visibility and a published schedule for announcements and withdrawals. A BGP Beacon is to be used for the ongoing study of BGP dynamics, and so should be supported with a long term commitment. We describe two collection of BGP Beacons that have been set up at various points in the Internet. We then describe techniques for processing BGP updates when a BGP Beacon is observed from a BGP monitoring point such as Route Views or RIPE.

We illustrate the use of BGP Beacons with three case studies. Each study relies on the fact that we are monitoring updates that have been (indirectly) generated by a Beacon event. First, we revisit the convergence time issues studied in [9, 10]. Second, we investigate the potential that route flap damping [11] has for punishing "well behaved" routes. Simulation results in [12] have shown that route flap damping can punish "well behaved" as well as "misbehaving" routes. Here we use the BGP Beacons to validate those results in the global Internet. Even though the BGP Beacons we use have a fairly long cycle (two hours between each announce or withdraw event), we see that *even announcements* can trigger flap damping in as much as 10 percent of the time at some locations in the Internet. For out third study, we present an analysis of the interarrival times of updates generated by BGP beacons.

2 BGP Beacons

Currently, there are two groups of BGP Beacons that differ somewhat in implementation. There are four Beacons in the first group, called the *PSG Beacons* because the first was set up at `psg.com`, which are listed in Table 1. These Beacons were set up by Z. Morley Mao with the help of the Beacon hosts. A web site for these beacons [13] contains contains Beacon related scripts and Beacon data derived from Route Views. The Beacon *period* is the time between each event at the Beacon, where an event is either an announcement of the Beacon prefix or a withdrawal. Two hours was picked as an period for the first three Beacons since it seemed long enough for most route flap damping to expire.

In PSG beacons, two attributes of the announcements have been "hijacked" to serve as a timestamp and a sequence num-

[*] UC Berkeley
[†] IIJ
[‡] AT&T–Research
[§] AT&T–Research

Beacon	Prefix	Period	Upstream ASN(s)	Source ASN	Beacon ASN	Host ASN	Location	Host	Start date
1	`198.133.206.0/24`	2 hrs.	2914, 1239 (1)	3972	3972	3130	WA,US	Randy Bush	8/10/2002
2	`192.135.183.0/24`	2 hrs.	3701,2914	5637	5637	10876	OR,US	David Meyer	9/4/2002
3	`203.10.63.0/24`	2 hrs.	1221	1221	private	1221	Australia	Geoff Huston	9/25/2002
4	`198.32.7.0/24`	various	2914,8001	3944	private	3944	MD,US	Andrew Partan	10/24/2002

Table 1: The PSG Beacons

ber. The aggregator IP attribute, which is an IP address, is set to have the form `10.X.Y.X` where `0.X.Y.X` (in binary) represents the number of seconds since the start of the month (GMT). The aggregator ASN attribute, us a number that is incremented with each announcement and cycles through the values from 64,512 to 65,635. Note that values of both the Beacon timestamp and Beacon sequence numbers are within "private" spaces.

The second group of Beacons, called the RIPE Beacons [14], have been set up as a part of the RIPE Routing Information Services (RIS). Each BGP route monitor, located in nine different locations, is associated with a unique BGP Beacon prefix from `195.80.244.0/24` through `195.80.232.0/24`. Each RIPE Beacon has a period of 2 hours.

The implementation of these Beacons differs in the following ways. First, PSG Beacons currently have timestamps and sequence numbers, while the RIPE Beacons do not. Second, the PSG Beacons currently have what we call *anchor prefixes* (see Section 3) associated with them, which aid in the preprocessing of update data. Third, the PSG Beacons are not associated with BGP routing monitors, as are the RIPE Beacons.

For the rest of this study, we focus on PSG Beacons 1, 2, and 3, as Beacon 4's varying period may result in interaction between consecutive signals. We are not currently using the RIPE Beacons because of the lack of anchor prefixes needed for data cleaning.

2.1 Beacon software

The PSG Beacon daemon software is based on the open-source BGP software router written in perl available at `http://bgpd.sourceforge.net/`. The original software is purely passive and does not provide any functionality to advertise routes. We modified it to inject routing changes to an open BGP session triggered by user-defined interrupt signal. A cron job is set up to regularly send an interrupt to the Beacon daemon software, so that announce and withdraw updates are sent alternately. In the beginning, the updates are sent every 30 minutes. We quickly discovered that the prefix was not visible at all and this was due to the route flap damping mechanism [11] implemented by the upstream provider of the Beacon. Subsequently, the schedule is set to be 2 hours between consecutive updates to minimize the likelihood of route suppression.

2.2 Terminology for BGP update propagation

We use the term *input signal* to refer to any update generated at a BGP Beacon (either an announcement or a withdrawal). The network of BGP speaking routers can be thought of as a giant non-deterministic signal transducer [15], where each input signal *causes* various *output signals* to be generated at different locations in the Internet. Output signals generated by the Beacon input signals can vary considerably, depending on the Beacon, the monitor point, and the time observed. For example, here is a an output signal for an announcement from PSG Beacon 1 sent, as seen from one peer at Route-Views on January 11, 2003:

Time (GMT)	Type	AS Path
05:00:11	A	8121 19151 2914 1 3130 3927
05:00:39	A	8121 16631 174 1 3130 3927
05:01:08	A	8121 3491 1 3130 3927

This output signal contains three updates. The *signal duration* is the elapsed time from the first to the last update in the signal. For this example, the signal duration is 57 seconds. A signal containing only one update has a duration of 0 seconds.

2.3 Convergence time

A BGP monitor may be receiving BGP updates from more than one neighbor. For example, Route Views currently has over 20 neighbors. For any Beacon event, there will be some neighbor that sends an associated update first. The time between this first update and the last update in a signal collected at that point is called the *relative convergence time*. For example, if the first update received from any neighbor for the announcement event above was 05:00:06, then the example signal has a relative convergence time of 62 seconds. *End-to-end convergence time* is the time between the last update in a signal and the sending time of the input signal based on the Beacon timestamp in the aggregator field.

Ideally, we would like to know the end-to-end convergence times, but this requires clock synchronization. Both the Beacon machine and the monitoring sites should be NTP synchronized according to people who administer these machines. However, to our surprise, we discovered many instances where the receiving timestamp of a Beacon update is smaller than the sending timestamp of the update due to the problem with clock synchronization. As we do not have control over the machines at monitoring sites nor the machines that run the Beacon software, we sometimes use relative convergence times and signal durations to understand the convergence delays.

2.4 Beacon location terminology

As seen in Table 1, there are several AS numbers associated with a beacon: Source AS, Beacon AS, and Host AS. *Source AS* is the origin AS in the AS path of the updates associated with the beacon prefix. *Beacon* AS is the AS to which the Beacon daemon belongs. If EBGP session is used, where the BGP session is between two routers in different ASes, it can be either a private AS number or a special AS number for the purpose of Beacon experiments. In the case of IBGP session, the Beacon AS number is the same as the *Host AS* which is the AS that directly provides the network connectivity for the beacon prefix. In contrast, the upstream AS is the closest tier-1 ISP that provides connectivity to the Host AS. The Host AS itself can be the upstream AS in the case of beacon 3 hosted in Australia. The Source AS number is that of the Beacon AS unless it is private. In that case, the Source AS number is that of the Host AS. In general, there is no need for a special AS to create a Beacon.

2.5 Public monitoring points

There are several public monitoring points where BGP data are collected from several ISPs. Route Views [7] is one such monitoring point which peers with about 30 different networks and receive all the BGP updates. Since the Beacon prefixes are not aggregated and should be globally visible, they are visible in all the BGP feed available at the monitoring points.

3 Data cleaning and signal identification

For the Beacon analysis, it is important to clearly identify the output signals by associating the observed BGP updates with a single input signal. We describe in detail a novel methodology to achieve that goal. Not all updates observed related to the Beacon prefix are caused by our input signals. Some of them, for instance, may be caused by routing changes in an upstream AS from the observation point. Given a BGP feed at Route Views, for example, there are typically always some routing updates observed every second. Take a typical day, May 2 2003, there are about 70 updates observed from about 30 peers at Route Views. Such routing changes can be due to a variety of reasons such as routing policy change, link failure, congestion. BGP session requires keep-alive messages to be exchanged between the two neighbors. If timeout occurs due to congestion or link failure, the BGP session is reset. Upon session reestablishment, the entire routing table is exchanged resulting in a large number of updates observed. Such session

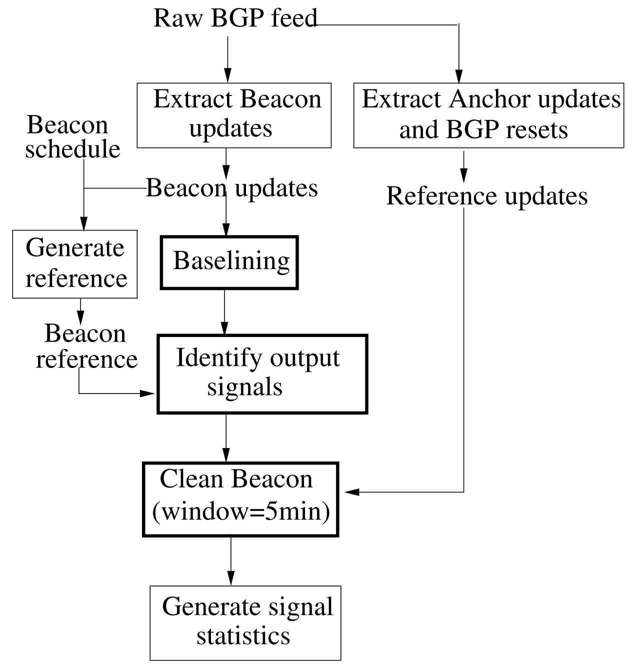


Figure 1: The process of cleaning Beacon data and identifying signals

resets can occur locally between the local router and the route monitoring software. In that case, such updates do not reflect actual routing changes that affect the forwarding plane of the data traffic. Work by Wang *et al.* [16] demonstrates the importance of filtering out updates due to local BGP session resets in analysis. Very different conclusions are drawn if such updates are not properly filtered.

We take a sequence of steps as shown in Figure 1 to clean the Beacon data to identify output signals and compute several signal statistics. This process consists broadly of three crucial steps: baselining, signal grouping, and noise filtering or cleaning.

3.1 Baselining

The goal of this step is to process the BGP data such that we can compare them from all peers in a sensible way. We first extract from the raw data the updates associated with the Beacon prefixes, *i.e.,* the Beacon updates. We found some peers at Route Views send out updates related to the BGP community and MED attribute changes. BGP policies can be set to prevent such attribute changes from being sent. These peers tend to send more updates and reveal more of the internals of the AS, as such attribute changes typically reflect the BGP dynamics within the last-hop AS. We also found that some peers at of Route Views send out consecutive updates that are identical. To do a reasonable comparison, during the baselining step, any update which is either identical to the previous update or differs only in its community or MED attribute values are eliminated. This reduces about 15% of the updates for all beacons based on Route Views data. 12 out of 23 peers at Route Views send such duplicate updates that differ only by these two attributes from previous updates. We verified that eliminating such updates have little or no effect on our analysis of interarrival update time (Section 6) and convergence delay (Section 7). However, the results of route flap damping (Section 5) will provide a lower bound, and the analysis of signal length can also provide an underestimate.

3.2 Signal identification

For the ease of analysis, we do additional processing to group updates together according to the input signals. To achieve that, we create *Beacon references* which identify the starting time of each output signal for the input signals injected. Any output signals received have timestamps greater than or equal to the beacon reference timestamps. Such timestamps can be easily generated if we have synchronized clocks, as the beacon schedule is recorded by the beacon daemon software. Similarly, we can generate such timestamps by using the Beacon timestamps in the BGP update aggregator field. However, as mentioned before, the clocks of the beacon machine and the monitoring machines are not well synchronized.

We thus resort to several algorithms to generate the Beacon references. If the monitoring sites receive the BGP feed directly from the AS that hosts the Beacon, the Host AS, then the timestamps of the messages from that AS can be used in the references. The BGP monitoring site, such as Route Views, usually establishes multihop BGP sessions with several different ISPs. The output signal coming from the Host AS almost always arrives first, as it typically travels through the fewest number of routers and smallest distance. Two of the Beacons (Beacon 1 and 3) fall into this category. Typically, the Host AS produces very clean output signals which consist of a single announcement given the input announcement signal and similarly a single withdrawal given the input withdrawal signal. It is important to point out that if there are no issues with time synchronization, the Beacon timestamps provide very accurate timestamps for the references that can be used for all monitoring sites. The time stamps based on the Host AS is specific to the monitoring site and requires that there is a BGP feed from the Host AS at the site. If other sites use such Beacon references generated from a different site, there may be offsets from the reference timestamps for the output signals, *i.e.,* some output signals may start earlier than specified due to clock differences.

If the BGP feed from the Host AS is not available at the monitoring sites, heuristics are used based on the Beacon schedule to determine the start of a new Beacon signal. The period of Beacon announcement is purposely set to be two hours for the first three Beacons. We therefore expect that most sites should already have converged on the final route long before the next input signal is injected. One simple heuristic is to look for large gaps between updates and use the Beacon schedule as a reference to identify the starting times of the output signals. It is easy to distinguish the start of a new announcement signal with the help of the sequence number in the aggregator field. The start of a withdrawal output signal is identified using the timing heuristic.

3.3 Noise filtering/cleaning

To differentiate the updates caused by our injected routing changes from updates caused by other effects, we propose the use of *anchor prefix* to detect such unexpected routing changes. An anchor prefix is a statically nailed down prefix belonging to the Beacon AS or the Host AS. Such a prefix can contain live hosts and thus does not experience routing changes caused by our experiments. It can also be an unused prefix purposed announced for the sake of reference with the Beacon prefix. Anchor prefixes serve as calibration points to identify unexpected routing changes. When there are no such routing changes, no routing updates associated with the anchor prefix can be observed. Anchor prefixes originate from the same AS as the Beacon, so that any routing change experienced by the anchor prefix must also be experienced by the Beacon prefix and is not caused by our purposely injected routing changes. In general, we found beacons hosted by larger ISPs, e.g., AS1221 in the case of Beacon 3 to be more stable as it has more redundant network connectivity. This is evidenced by the observation that Beacon 3's anchor prefix files are typically one third smaller than other beacons.

The Beacon signals are cleaned by deleting signals that can be affected by unexpected routing changes as experienced by anchor prefixes. For each anchor prefix update, we construct a window starting at W minutes before the update time and ends at W minutes after the update time. Any Beacon prefix signal that have an overlap with such a window is ignored, as they are likely to affected by external routing changes. Based on empirical observation, it takes on the order of minutes for a routing change to become globally visible. We set W to be 5 minutes in our analysis. We perform the same cleaning process using certain BGP STATE messages which indicate BGP session resets. Using Route Views data, we observe that the number of output signals remain fairly constant when $W >= 5$. The cleaning process deletes on average 2 to 3 percent of updates.

Tables 2 and 3 show the effect of cleaning on observed signals in terms of signal count, average signal duration, delay, and signal length. They demonstrate the importance of cleaning. For all four beacons, less than 5% of the signals have been deleted after cleaning. Overall, the average signal delay and signal duration have decreased for both announcement and withdrawal signals after cleaning. In some cases, the decrease is as much as 50% of the original value. However, the signal length remain mostly the same after cleaning. This means that cleaning removes those outliers with large inter-arrival time or long signal duration.

For each of the four PSG Beacons, we perform this sequence of steps and finally generate a set of signal statistics such as relative convergence, signal duration, end-to-end convergence time, number of updates.

Table 2: Effect of cleaning on observed announcement signals: signal count, average duration, delay, and length

Beacon	Before cleaning				After cleaning			
	count	avgDur (sec)	avgDelay (sec)	avgSigLen	count	avgDur (sec)	avgDelay (sec)	avgSigLen
1	33536	27.13	50.60	1.47	33318 (99.35%)	19.36	41.89	1.47
2	34522	9.13	29.56	1.20	33726 (97.69%)	6.75	25.21	1.17
3	32504	10.82	34.99	1.22	32188 (99.03%)	5.77	28.40	1.21
4	39044	41.95	63.66	1.52	37970 (97.25%)	22.79	43.16	1.46

Table 3: Effect of cleaning on observed withdrawal signals: signal count, average duration, delay, and length

Beacon	Before cleaning				After cleaning			
	count	avgDur (sec)	avgDelay (sec)	avgSigLen	count	avgDur (sec)	avgDelay (sec)	avgSigLen
1	33443	37.88	100	2.07	33261 (99.46%)	32.98	90.09	2.07
2	33860	45.24	109.23	2.19	33344 (98.48%)	42.94	94.38	2.19
3	32379	59.16	120.64	2.55	31182 (96.30%)	56.36	114.40	2.55
4	36633	96.33	139.63	3.43	35776 (97.66%)	75.65	115.90	3.41

4 BGP Implementation Impact: Cisco vs Juniper

On question of some interest is how much impact the different implementations of BGP have on the results. The BGP specification (RFC 1771) [1] defines the protocols to be used, but not how they should be implemented, and in some cases BGP implementations have contained bugs which meant they did not even match this specification. There are implementation differences between different router vendors, and even between models, and software versions from the same company. In addition, there are settable parameters which may impact behavior (*e.g.,* the MinRouteAdverTimer). As specified by the BGP RFC, MinRouteAdverTimer specifies the minimum amount of time a router needs to wait before second consecutive updates for the same router to the same neighbor. In an ideal world, these would have little impact on the operation of BGP, but studies have shown (*e.g.,* [17]) that at least some of these differences may have large impacts.

In this section we consider one example of the impact that differences in implementation may have on the behavior of BGP. Namely, we consider the difference between Cisco and Juniper implementations of BGP. The decision to compare these two (out of all the possibilities) is motivated by the fact that we know the make and model of two last hop routers (as seen by Route Views). In general this information is not publically available, but the last hop routers corresponding to Peer 147.28.255.1 and 147.28.255.2 are known to be Cisco and Juniper routers, respectively.

Table 4 presents a comparison between Juniper and Cisco routers (as seen from beacon 2). The table shows that the Juniper router sends about 25% more updates, has a similar update duration, and a substantially smaller average inter-arrival time for updates (around 60% of that for Cisco routers). The most startling difference, though, is in the number of short (< 26 second) inter-arrival times is much greater for the Juniper router.

The differences can be explained by the fact that, by default, the MinRouteAdverTimer is turned off in Juniper routers [18]. It is common practice for users to leave default settings alone, unless they have particular reason to do otherwise, and we know from discussion with the administrator of these routers that the defaults are used here. As shown in [17] having a small, or zero MinRouteAdverTimer can result in large numbers of additional updates (as seen here). Also shown in [9] was how the MinRouteAdverTimer spaces out the updates, thereby potentially increasing the convergence time – also as seen here.

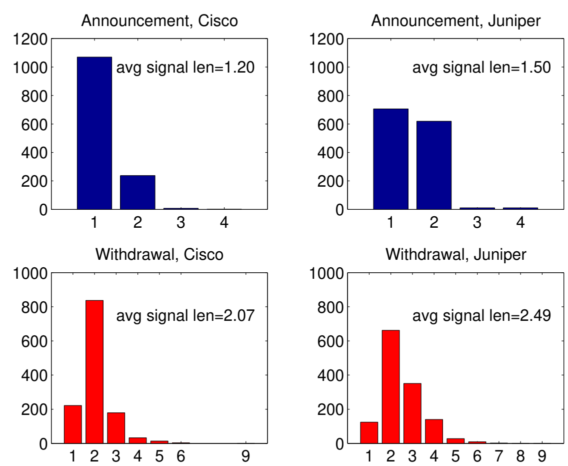


Figure 2: Beacon 2: Comparison of the numbers of updates for two known Cisco and Juniper routers from the same peer.

Given the obvious difference between the two router types

Peer	Type	signal length A	signal length W	duration A	duration W	inter-arrival A	inter-arrival W	% of short inter-arrivals A	% of short inter-arrivals W
147.28.255.1	Cisco	1.20	2.07	6.79	48.4	34.8	45.4	1.56	0.44
147.28.255.2	Juniper	1.50	2.49	7.13	44.3	14.2	29.6	12.76	4.37

Table 4: A comparison of the Cisco and Juniper Routers. The table shows average statistics (including the average signal length, or number of updates, the average duration, or convergence time, the average time between updates during a sequence of events, and the percentage of inter-arrival times less than 26 seconds), for announcement 'A', and withdrawal 'W' events, for the two known last hop routers.

above, the question naturally arises. Can we distinguish the other last hop routers as being of one type or the other? Figure 3 suggests that we can separate the two, though not perfectly. We could identify candidates from this figure, but needed to do a final verification by examining the update sequences in detail to confirm the findings. The routers that appear to be Juniper routers are marked on the plot.

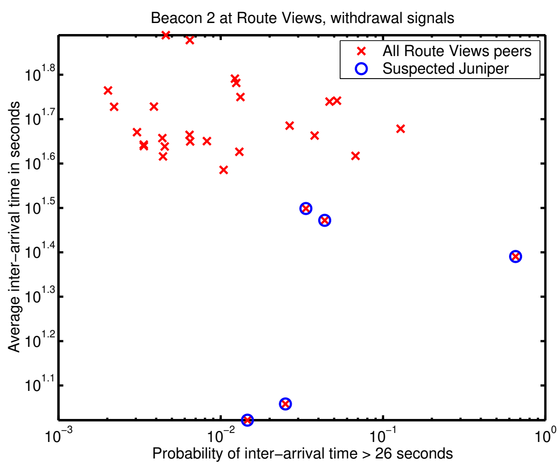


Figure 3: A scatter plot showing the routers classified as Juniper-like. Although there is not a completely clear distinction in the plot, detailed examination of the update sequences shows that the marked peer routers show similar characteristics to the known Juniper router.

The box plots in Figures 4 and 5 clearly demonstrate the difference in signal duration distribution as the signal length increases. For Juniper-like routers, there is no clear dependence between the signal duration and signal length. As the signal gets longer, the duration increases only slightly and keeps the values around 40 to 60 seconds. This is because Juniper's default rate-limiting algorithm allows updates to be sent in burst imposing little delay. In clear contrast, Cisco-like routers clearly show the 30 second rate-limiting behavior: there is a linear increase with slope of around 30 seconds in the duration as the signal gets longer. Furthermore, signals generated from Cisco-like peers have shorter signals compared to that of Juniper-like peers.

Beacon 1's signal duration distribution is shown in Figures 6 and 7. Again we differentiate between Cisco-like and Juniper-like peers. Announcement signals have in general shorter durations compared to withdrawal signals. In addition, we again

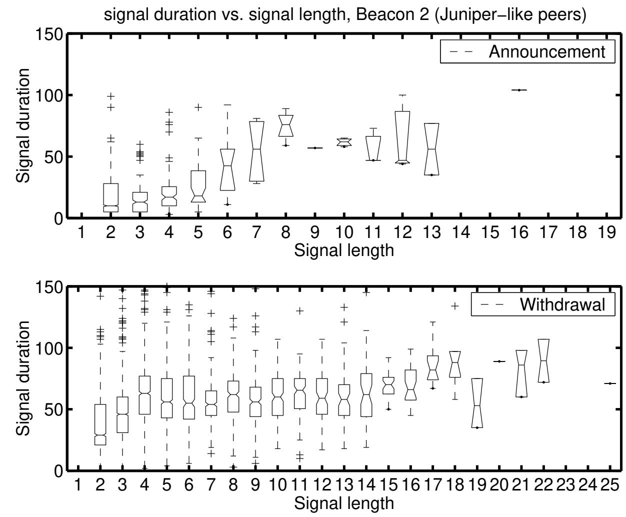


Figure 4: Beacon 2's signal duration distribution for each signal length for Juniper-like peers

see the signal duration for Cisco-like peers to be multiples of 30 seconds. The distribution for Juniper-like peers is much more spread out. Large number of signals converge within 30 seconds for both types of routers.

5 Route flap damping analysis

Route flap damping [11] is one of the two mechanisms in BGP aimed to achieve routing stability. Flap damping punishes unstable routes or routes that change frequently by suppressing them. It is designed to deal with routes that are unstable on a long time scale. In contrast, the other mechanism, the minimum route advertisement timer (minRouteAdverTimer) is designed to act on routes that are unstable on a short time scale. MinRouteAdverTimer specifies the minimum amount of time a router needs to wait before sending consecutive updates referring to the same prefix to the same neighbor. The hope of delaying the updates is for consecutive updates to be batched together to reduce update traffic. These two mechanisms can interact, as the timer determines the amount of updates that are propagated during the convergence process.

We now briefly describe how route flap damping works. The router keeps track of a penalty value for each route received from its neighbor belong to an external AS. Whenever the route changes, a penalty value is added. Both Juniper and Cisco

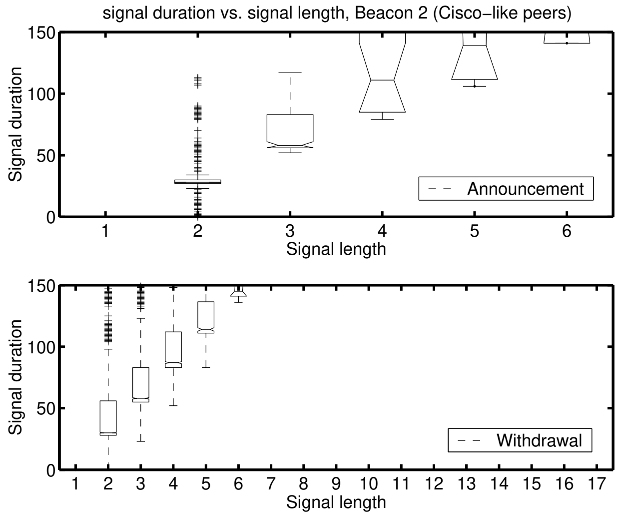
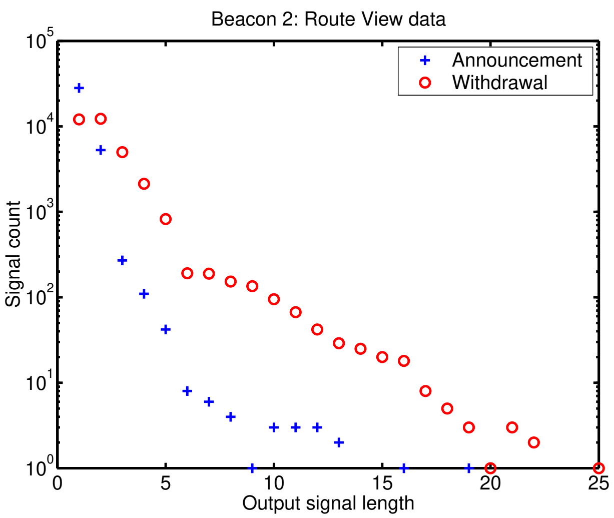


Figure 5: Beacon 2's signal duration distribution for each signal length for Cisco-like peers



routers have their own specific penalty increment as shown in Table 5. The penalty value decays exponentially over time and the associated configurable parameter is half life. It determines the amount of time it takes for the penalty to decrease to half of its original value. If the penalty ever exceeds the cutoff threshold, the route is considered suppressed, or no longer eligible as a usable route to forward traffic. There is limit on how long the route can be suppressed as defined by max suppress time. A withdrawal is sent out if the route was used in the forwarding table. Subsequently, if any other updates for this route are received, they will not be propagated. Once the penalty value decays below the reuse threshold, the route is considered usable again.

It has been shown [12] in simulations and commercial router testbed setting that in certain topologies, no matter how large the minRouteAdverTimer is, there are sufficient updates induced by a single route change to trigger route flap damping. This means that a single router reboot, which translates to a withdrawal message followed by an announcement message,

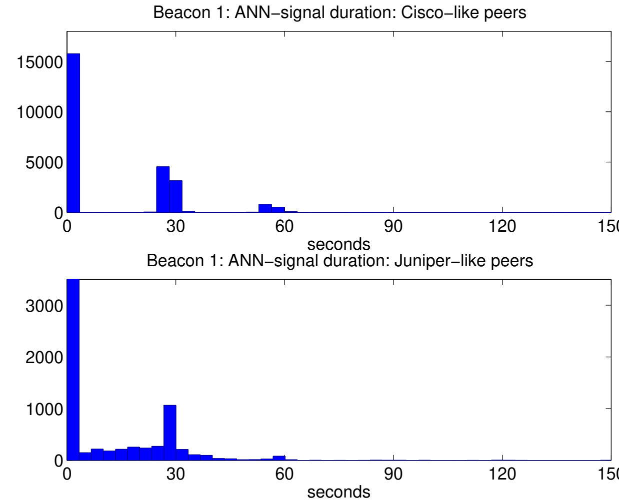


Figure 6: Comparing Beacon 1's announcement signal duration distribution for Cisco- like peers with that for Juniper-like peers

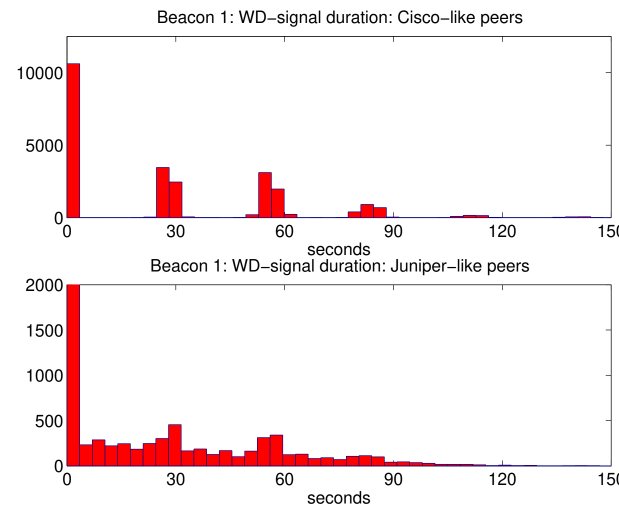


Figure 7: Comparing Beacon 1's withdrawal signal duration distribution for Cisco- like peers with that for Juniper-like peers

can cause the route to be suppressed somewhere on the Internet. As there is no feedback in the flap damping mechanism, it is difficult to determine which router suppresses the route. If the route suppressed is the only route to reach a destination prefix, then the destination becomes unreachable. It is thus very important to understand how likely this occurs in today's Internet.

It is very difficult to understand the extent at which route flap damping can suppress well-behaved or stable routes on today's Internet. The difficulty arises due to the complexity in inferring the root causes of BGP updates observed in passive measurements [15]. The beacon infrastructure provides a perfect medium for do such study, as routing changes are injected at known times and locations. Assuming the beacon prefix routes are not suppressed, we can simulate how likely a single routing change can cause the route to be suppressed. We implemented

Table 5: Default route flap damping parameter settings

RFD parameter	Cisco	Juniper
Withdrawal penalty	1000	1000
Readvertisement penalty	**0**	**1000**
Attributes change penalty	500	500
Cutoff threshold	**2000**	**3000**
Half-life (min)	15	15
Reuse threshold	750	750
Max suppress time (min)	60	60

the route flap damping algorithm using the Cisco and Juniper default parameters and calculated the percentage of signals that can trigger route suppression at the monitoring sites. Based on Route Views data, we observe about 5% of input signals are suppressed across all Route Views peers as shown in Figure 8. Some peers are much more likely to suppress the route than others because the large number of updates generated. The columns labeled with "peer max" in the Figure indicate the maximum percentage of suppressed signals on a per peer basis. Some peers suppress close to 45% of all signals received.

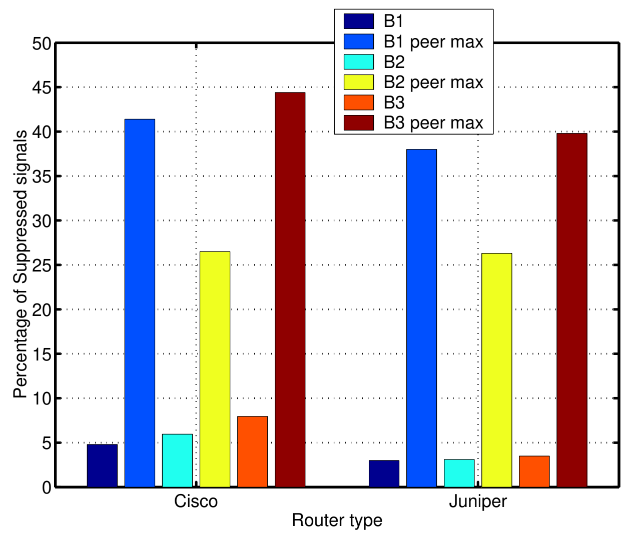


Figure 8: Overall percentage of suppressed signals due route flap damping for each beacon and on a per peer basis for Cisco and Juniper.

Figure 9 breaks down the suppressed signals between announcements and withdrawals. Since withdrawals typically generate more updates, naturally a much higher percentage of withdrawal signals are suppressed compared to announcement signals. In fact, at some peers, close to 90% of all withdrawal signals can trigger route suppression using Cisco's default setting. Overall, Cisco is more aggressive in suppressing routes than Juniper based on Route Views data for our three beacon prefixes for both announcement and withdrawal signals. In fact, this may not always be the case. Although Cisco's cutoff threshold is lower compared to Juniper's, it does not punish route readvertisement or an announcement that is preceded by

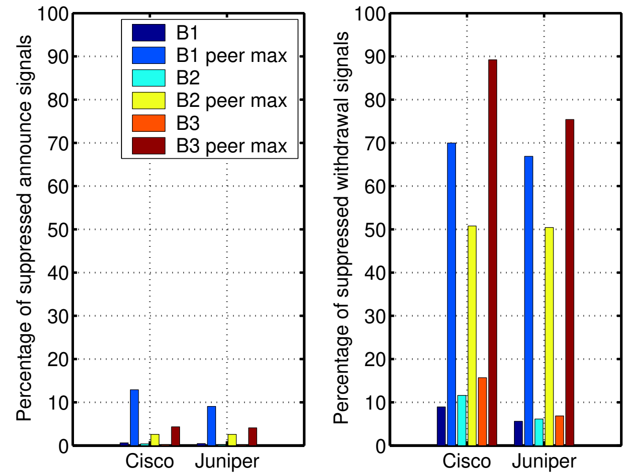


Figure 9: Percentage of suppressed beacon signals due to announcement and withdrawal

a withdrawal. Consequently, Cisco is more likely to suppress routes with the following update patterns: "AAAW" (A: announcement, W: withdrawal). There is no readvertisement in such patterns; therefore, Cisco's lower threshold will increase the probability of route suppression. Juniper is more likely to suppress routes with update patterns such as "AWAWA" where there are readvertisements. From our data, the former pattern is much more prevalent; therefore, Cisco is overall more aggressive in route suppression than Juniper.

Our previous analysis only provides a lower bound for the percentage of suppressed signals, as some routers in the network may already have suppressed the beacon prefix, resulting in possibly fewer updates observed at the monitoring site. Certain update patterns also indicate the presence of route suppression by some intermediate routers from a given monitoring point. For an announcement input signal, we sometime observe interarrival time between 1000 and 3600 seconds. After the long timeout, the update sequence always ends with an announcement. Such a long timeout is extremely unlikely to be caused by propagation delay, router processing delay, BGP path vector effects, or MinRouteAdverTimer values, even if they are accumulated along each router hop. Flap damping suppression duration is typically on the order of tens of minutes which match well with the timeout values we observe. In fact, based on the default Cisco parameter setting, the minimum suppression duration is about 30 minutes. For Juniper, the value is about 21 minutes.

If the data are cleaned properly and there are no missing updates, such timeout values provide a good indication that route suppression has occurred. The final announcement can be caused by the re-announcement of the route after its penalty has decayed below the reuse threshold. And the long break indicates the duration during which the route is suppressed. We show two such examples below (Tables 6, 7). In the first one, the last update before the long break is an announcement. In the second example, a withdrawal occurs before the long break. In both cases, the timeout value is about 40 minutes or 2400 seconds. We observe three transient routes from peer 216.18.31.102 in case 1. The final route goes through AS701, and it is very likely that the long break is due route flap damping occurring along the AS path "6539 701 1 3130 3927". ASes 6539, 701 or 1 may have suppressed the route originated by AS3927. AS3130 could not have suppressed the route, because the alternate path used still goes through it. As soon as this route is unsuppressed, it is chosen as the preferred route by AS6539, which apparently has at least three alternate routes to reach the destination AS3927 where the beacon prefix comes from. It very likely only suppresses the route learned from AS701; therefore, its connectivity to the beacon is not affected. In general, if the last update before the timeout is an announcement, it indicates that there is an alternate path available from the monitoring point. In the case of withdrawal, flap damping has affected all the available paths from the monitoring location.

The latter is exemplified by case 2 shown in Table 7. The route with AS path "11608 2914 3130 3927" appears to be preferred over the alternate route with AS path "11608 2914 1239 3130 3927". In this example, it is very likely that AS11608 suppresses the route received from AS2914. AS2914 is less likely to suppress the routes coming from both ASes 3130 and 1239, as this requires a router AS2914 to have received sufficient updates from both these neighbors. AS3130 is also unlikely to have suppressed the route from the origin AS 3927, as it is very close to the beacon source and gets fewer updates. In general, ASes farther away from the origin AS is more likely to experience more updates and thus more likely to suppress the route.

Table 6: Case 1: observation from peer 216.18.31.102 on Apr 3 2003 for Beacon 1:

Time (GMT)	Type	AS Path
23:00:17	A	6539 3561 1 3130 3927
23:00:44	A	6539 701 1 3130 3927
23:01:14	A	6539 3602 16914 852 1 3130 3927
23:42:46	A	6539 701 1 3130 3927

Table 7: Case 2: observation from peer 207.246.129.14 on Sep 17 2002 for Beacon 1:

Time (GMT)	Type	AS Path
22:38:45	A	11608 2914 1239 3130 3927
22:39:13	A	11608 2914 3130 3927
22:39:40	W	
23:24:26	A	11608 2914 3130 3927

In our data, we observed close to 1% such updates sequences indicative of route suppression in an intermediate router. We purposely separated consecutive updates with two hours, as the maximum suppress time is one hour. Therefore, for an announcement input signal, we should always expect to observe an ending announcement unless data are missing or anchor prefix is unstable. Similarly, withdrawal input signal should always result in an ending withdrawal in the output signals.

6 Inter-arrival time analysis

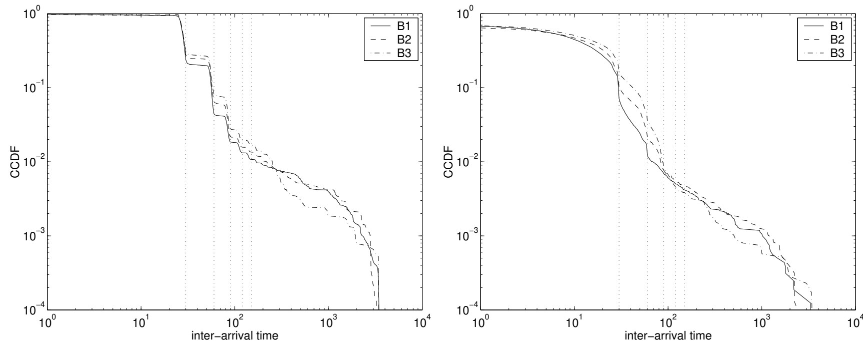
In this section we explore an aspect of BGP dynamics not considered in any previous work. A BGP update sequence has been typically considered (within this paper as elsewhere) to be a sequence of N updates, over some duration, but little attention has been given to the distribution of events within this time interval. However, within this sequence the updates are spaced according to an *inter-arrival* distribution. We provide a brief examination of some of the properties and ramifications of this distribution as an example of how BGP beacon data may used in the analysis of Internet routing dynamics.

As noted above, the inter-arrival times for Juniper-like and Cisco-like routers are different, and so we shall consider these separately here. First, consider the Cisco-like cases. Figure 10 shows log-log plots of the CCDF of the inter-arrival times for events (including anouncements and withdrawals), for each of the three beacons. The x-axis being the time between updates (in seconds), and the y-axis, the probability that an interval exceeds this time. Note that in the results here, intervals are rounded up to the nearest second, so that all intervals less than 1 second will appear as one second.

Despite the difference in the two graphs, we see two regimes in both. This is most clear in the distributions for the Cisco-like routers. For each of the beacons we see a *body* region of step like decrease at slightly less than 30 second intervals (the vertical dashed line are drawn at exactly 30 second intervals). The 30 second intervals seems to match the typical value of the MinRouteAdverTimer described above. The second *tail* region seems to level out the distribution, followed by a sharp decrease, before the distribution is truncated at 3600 seconds by the data cleaning. The cut off between the two regions appears to be around 100 seconds. The Juniper-like routers show a similar division of the distribution, though the body part is less step like.

A natural hypothesis to make is that the two components of this distribution arise from different basic causes. We shall examine this hypothesis by trying to understand what two processes might produce these results. Out first approach is to do some distribution fits to the data, to gain an understanding of what we are seeing.

Figure 11 shows one such fit (on loglog and semilog axes), done by eye for the Cisco-like routers. The fit combines three components. Firstly, we model the step function taking the number of steps to be given by a geometric distribution, and with the length of the steps to be 28 seconds (plus a small Gaussian jitter). The second component is a small mass at 1 second, because of the discretization of the interval times, in particular



(a) Cisco-like last hop routers.

(b) Juniper-like last hop routers.

Figure 10: The inter-arrival time distribution for each of the three beacons as seen from Cisco-like and Juniper-like routers. The vertical dotted lines are drawn at 30 second intervals.

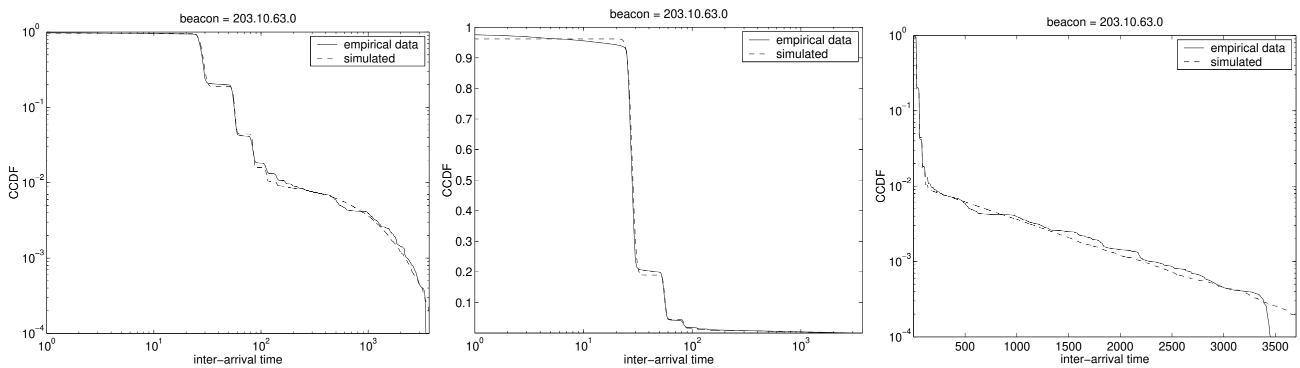


Figure 11: The inter-arrival time distribution for Cisco-like last hop routers, and beacon 1

all times below one second are rounded to one. The third component is a shifted exponential distribution, which matches the tail of the distribution. The figure shows the fitted curve as the dashed line. Note that the fit on the loglog plot is very visually satisfying, as it is on both of the semilog graphs.

Note that, in this fitting we are not seeking to gain a precise knowledge of the parameters involved. In fact, given the number of parameters we have to play with here, the data are not sufficient to achieve a precise parameterization (one can always fit a sufficiently complex curve to any data set). The aim is to gain an understanding of the processes that might be involved by seeing what type of distributions they generate. However, for the benefit of the reader we provide a precise definition of the distributions and parameters used to produce the fitted distribution.

The distribution is generated using:

$$X = \begin{cases} 28 * (1 + \text{Geom}(0.81)), & \text{with probability } 0.9524, \\ 1, & \text{with probability } 0.0381, \\ 90 + \text{Exp}(970), & \text{with probability } 0.0095, \end{cases}$$

where $\text{Exp}(970)$ refers to an exponentially distributed random variable with mean 970 (seconds), and $\text{Geom}(0.81)$ refers to a geometric distribution with parameter $p = 0.81$, and therefore mean $(1-p)/p = 0.2346$.

Rather than try to consider the exact nature of the distribution above, let us try to understand the implications of this form of distribution. This type of distribution could arise as a result of a random mixing between three different random variables:

- **geometric distribution:** The first part of the distribution is generated by a series of steps, each nearly 30 seconds apart. The action of the MinRouteAdverTimer would certainly explain the first step – this timer prevents a router from sending an announcement within some time of the last prior announcement of the prefix. The typical default (rarely changed in practise) of a Cisco router is 30 seconds, and the router adds jitter to this amount to prevent any possible synchronization effects. Hence, one would naturally expect a delay of around 30 seconds between announcements – hence the first step in the distribution. The second and further steps can then be explained as multiple MinRouteAdverTimer intervals. We can suggest a simple reason why one might see such gaps: Cisco's implementation of the MinRouteAdverTimer is not 'per prefix', but rather 'per peer'. That is, the router will send a series of announcements to a peer, and then wait for the MinRouteAdverTimer Hence, an announcement which arrives 'late' due to delays in prior transmission (through a series of AS's) can miss the next batch of transmissions, and be delayed for a step. This can happen multiple times as an announment traverses the network, and so we see multiple missed steps. The interesting thing is that this process can be so simply modelled by a geometric distribution, in which the probability of missing the next step does not depend on how many steps have already been missed.

- **mass at zero:** The discretization of timestamps to integer seconds results in discretization of the inter-event times – hence times between zero and one will tend to be lumped into one point at zero. Thus we need to include a probability mass at zero, which is indicative of the number of very short inter-arrival times.

- **shifted exponential:** This is the most puzzling part of the distribution, partly because we have the least data in this region (less than 1% of the distribution falls into the tail). The lack of a large data set, and the fact that these are truncated (by only using data sets for which the total time is less than one hour), means that one can model the tail almost as accurately using a power-law distribution. However, of the known BGP mechanisms, the most likely source of these delays is route flap damping, discussed above.

There is a natural test for the cause of the step like body of the distribution above. That is to consider the Juniper-like last hop routers, for whom we suspect the MinRouteAdverTimer is not used.

Figure 12 shows plots of the inter-arrival time distribution for the Juniper last hop routers (on loglog and semilog scales). This appears to be somewhat different from Figure 11. We can see much less evidence of a step-like decrease, and significantly more of the mass of the distribution appears before 30 seconds. Figure 12 also shows a fit to the distribution, though this time the fit has four components:

- a geometric distribution (step size ~ 30),

- mass at one (probability 0.1765);

- a convolution of a uniform (over $[0, 16)$) and exponential distribution (mean 10),

- a power-law tail (with $\alpha = 0.85$.).

In this case the much smaller (probability 0.0882), residual geometric component of the distribution can be easily explained by Cisco routers earlier in the path of the updates. The first part of the distribution, formed from the convolution of uniform and exponential distributions appears to be the fundamental difference between the two. It seems clear form these results the the MinRouteAdverTimer is responsible for the inter-arrivals times typically being multiples of approximately 30 seconds.

We display the power law fit to the tail for the Juniper-like routers, as in this case it appears to be a little better than the exponential tail, but note that it is not really possible to rigorously distinguish the two given the small amount of data in the tail, and the narrow range of scales across which it traverses. It will be interesting to study this as more data becomes available, to determine which model is better, as well as better determining the cause of this tail.

Finally, the above graphs lump two components together, the inter-arrivals for anouncements, and for withdrawals. It might be possible that the two components seen are actually derived from these two separate types of events. In Figure 13 we show the separate distributions for anouncements and withdrawals separated (and compared to the overall distribution). Note that the three curves all retain the same basic characteristics, though the curves for the announcement events both drop more sharply, and level off more than those for the withdrawals. The overall curve is more like the withdrawal curve, largely because withdrawal events generate more updates, and therefore more inter-arrival measurements.

The above analysis provides us with one more insight that might not be immediately obvious. Regardless of the form of the tail of the distribution, it can be considered a 'heavy-tailed' distribution in the sense that there is a significant probability of an event several orders of magnitude larger than the typical events (thousands of seconds as opposed to around 30). This is important because it immediately explains our earlier finding that the convergence times are not well correlated with the number of updates seen.

The sum of a series of heavy-tailed random variables is well known to also have a heavy-tail, but a less well known result is the fact that the heavy-tail of the sum arises not from a sum of medium size events, but from single large events. The intuitive explanation for this effect is that "rare events happen in the most likely way". In this context, we may interpret this to mean that the longer convergence times are not typically the result of a long series of updates that take a long time to converge, but rather the result of a single long inter-arrival time between updates. The data seem to validate this intuition. If removing the heavy-tail from convergence times is considered important, then one must first concern oneself with the causes of the heavy-tail in the inter-arrival time (for instance flap damping), rather than trying to reduce the number of updates.

7 Convergence Redux

Given the insight we gained from our previous three studies, we now revisit the work by Labovitz *et al.* [9] conducted about three years ago. They analyzed BGP convergence behavior of four types of events: announcement (Tup), withdrawal (Tdown), fail over to a shorter route (Tshort), fail over to a longer route (Tlong). We only focus on the first two cases and leave it to future work to study the latter two cases. To study Tshort and Tlong, we need to modify our experiment setup to inject a withdrawal to one of the upstream ASes rather than

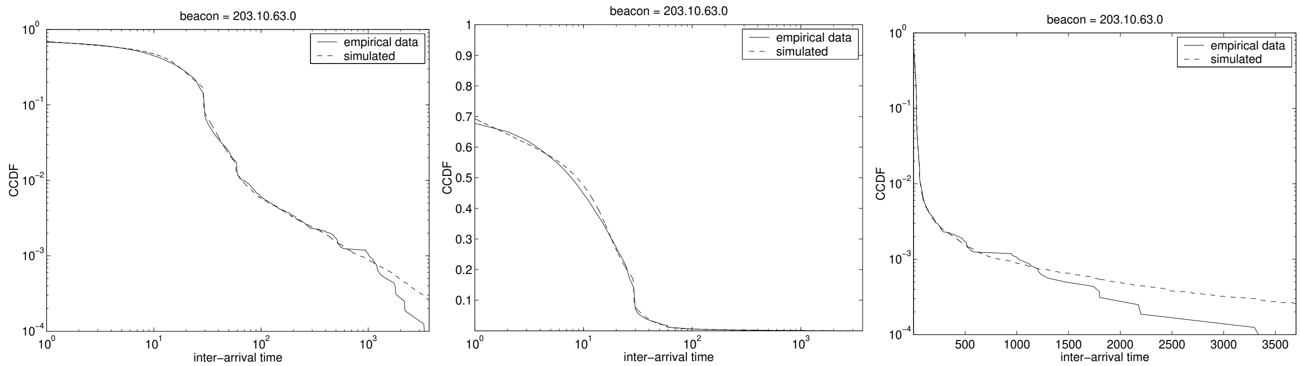


Figure 12: Empirical interarrival time distribution for Juniper-like routers comparing with simulated values

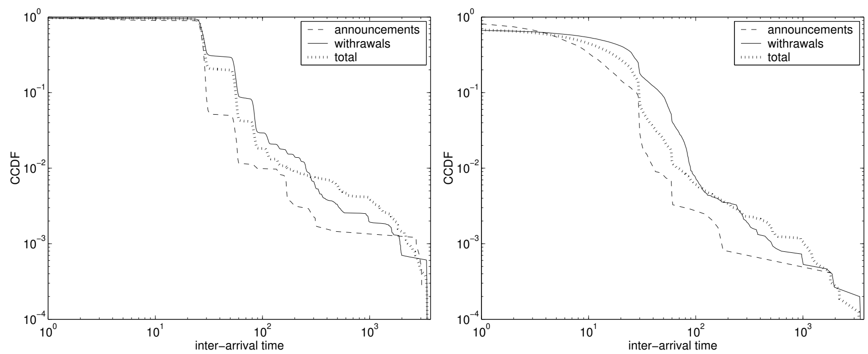


Figure 13: Interarrival time distribution for Juniper-like routers separated out by announcement and withdrawal signals

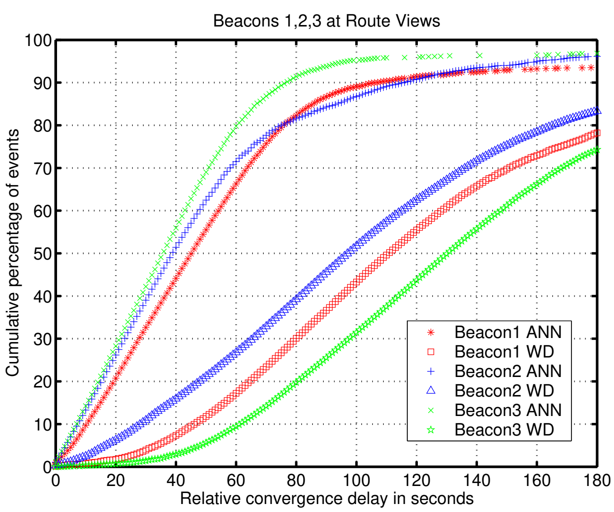


Figure 14: Cumulative distribution of relative convergence times for all three Beacons for both announcement and withdrawal signals

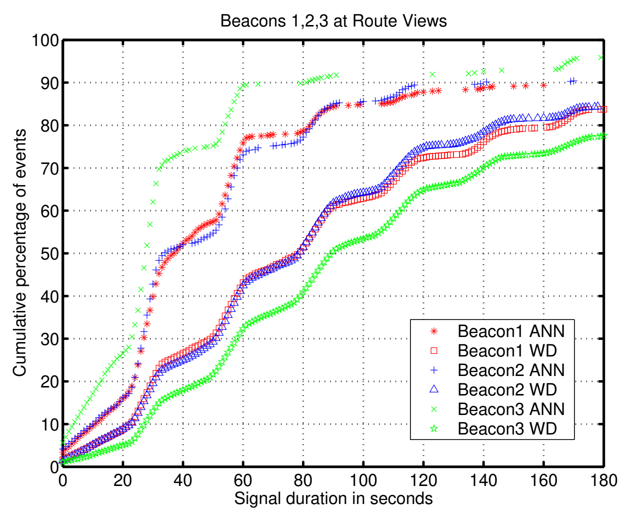


Figure 15: Cumulative distribution of signal duration for all three Beacons for both announcement and withdrawal signals

to both ASes in the case that the beacon is multihomed. As pointed out by Labovitz, the observed behavior of Tshort is very similar to Tup and that of Tlong is quite similar to Tdown.

Figures 14 and 15 present a cumulative distributions of the relative convergence times and signal durations, for PSG Beacons 1, 2 and 3. These results are entirely consistent with the results of [9], showing that these characteristics appear not to have changed significantly in the last few years. Our analysis is consistent also with a preliminary study of the RIPE Beacons,

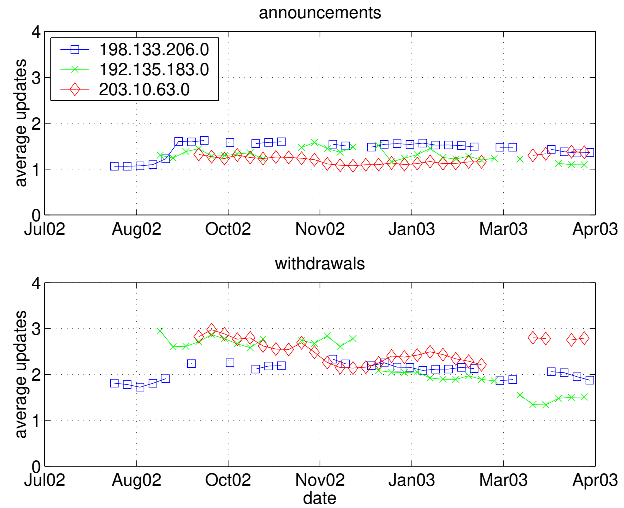


Figure 16: Variation in average signal length over time (Beacons 1,2,3).

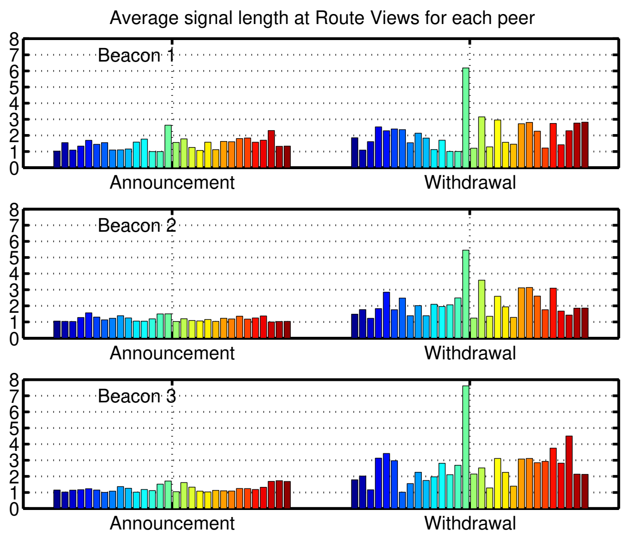


Figure 17: Average signal length for each peer

recently presented [19].

Figure 16 present the variations over time in signal length. These averages mask the wide variations seen among peers, which is presented in Figure 17. As in [9], we see more updates in signals associated with announcements and withdrawals. Figure 18 shows the variation in the relative convergence delay over time — there is large amount of variation but no clear trend.

Figure 19 illustrates an interesting time series of Beacon 1's duration during the course of our study. The gaps indicate the time periods during the beacon was down. In the month of August 2002, the Beacon was single-homed: with only one upstream provider AS2914. In September 2002, it became multihomed to AS2914 and AS1. There is little change for withdrawal signal duration; however, the duration doubled from around 30 seconds to 60 seconds. Upon further analysis, we

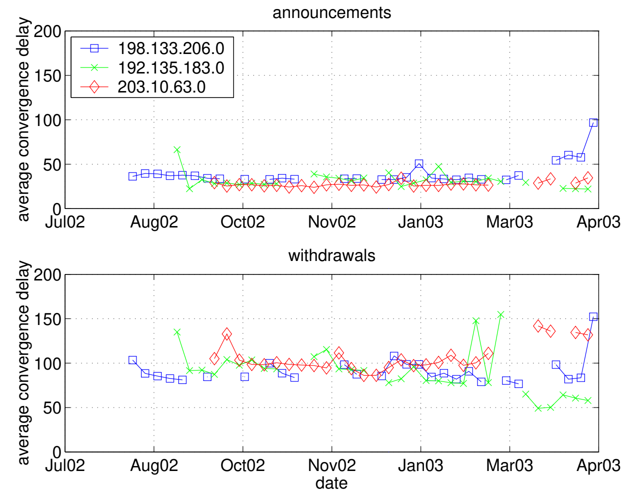


Figure 18: Variation in average relative convergence delay over time (Beacons 1,2,3).

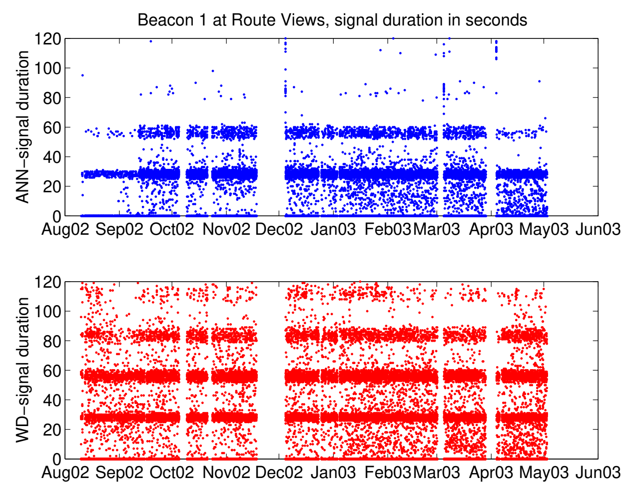


Figure 19: Beacon 1's signal duration variation over time, cutoff at 120 seconds. Max duration is 3525 seconds.

found that the average announcement signal length also doubled from around 1 to 2. After Beacon 1 is multihomed, the announcement signal is very likely to explore the alternate less preferred route first before settling on the final route. There is another interesting change occurred in April 2003, during which one of Beacon 1's upstream providers is changed from AS1 to AS1239. Apparently AS1239 seems to be better connected than AS1 and resulting in shorter announcement signal durations.

8 Conclusions and Open Problems

The paper describes a set of BGP Beacons that have been set up for public use, along with techniques for obtaining clean and useful data from these Beacons. Used in conjunction with public route monitors they provide a mechanism for performing controlled experiments with global Internet routing. We present several examples of how data from such experiments may be used to understand BGP routing dynamics.

Many interesting questions remain to be investigated, for instance, can we use this data to form a realistic and robust model of the dynamics of BGP? Such a model should, at the least, allow one to engineer BGP settings such as the MinRouteAdvertTimer, or route flap damping parameters to optimize global BGP convergence and stability.

To achieve such a model, we must gain a better understanding of some of the results discussed here – for instance, we need to better comprehend the processes that lead to the tail of the inter-arrival time. The results above suggest that the large inter-arrival times might be the best place to start improving BGP performance, but the data so far are not conclusive as to the cause of these gaps, though route flap damping is the likely culprit.

We have limited our attention to the PSG Beacons and Route Views monitoring data. We expect that using the RIPE Beacons will provide invaluable additional insights for this investigation, particularly because of the different connectivity and geographic locations of the RIPE Beacons. Furthermore, the on-going nature of the Beacons create a historical view of Internet routing dynamics that will allow measurement of the impact of changes to BGP.

Acknowledgments

We would like to thank all the PSG Beacon hosts for their cooperation. We also would like RIPE Information Service for providing Beacons as well monitors. Finally, we thank the people at Route Views for their continuous support and encouragement.

References

[1] Y. Rekhter and T. Li, "A Border Gateway Protocol," RFC 1771 (BGP version 4), March 1995.

[2] S. Halabi and D. McPherson, *Internet Routing Architectures*, Cisco Press, Indianapolis, Indiana, second edition, 2000.

[3] C. Huitema, *Routing in the Internet*, Prentice Hall, 2000.

[4] C. Labovitz, R. Malan, and F. Jahanian, "Internet Routing Stability," in *Proceedings of ACM SIGCOMM 1997*.

[5] C. Labovitz, R. Malan, and F. Jahanian, "Origins of Internet Routing Instability," in *Proceedings of INFOCOM 1999*.

[6] C. Labovitz, A. Ahuja, and F. Jahanian, "Experimental Study of Internet Stability and Wide-Area Network Failures," in *Proceedings of FTCS 1999*.

[7] "University of Oregon Route Views Archive Project," `www.routeviews.org`.

[8] Ripe NCC, "Routing Information Service Raw Data," .

[9] C. Labovitz, A. Ahuja, A. Bose, and F. Jahanian, "Delayed Internet Routing Convergence," in *Proceedings of ACM SIGCOMM 2000*.

[10] C. Labovitz, R. Wattenhofer, S. Venkatachary, and A. Ahuja, "The Impact of Internet Policy and Topology on Delayed Routing Convergence," in *Proceedings of INFOCOM 2001*.

[11] C. Villamizar, R. Chandra, and R. Govindan, "BGP Route Flap Damping," RFC 2439, 1998.

[12] Zhuoqing Morley Mao, Ramesh Govindan, George Varghese, and Randy H. Katz, "Route flap damping exacerbates internet routing convergence," in *Proceedings of ACM SIGCOMM 2002*.

[13] "PSG BGP Beacons," `http://www.psg.com/~zmao/BGPBeacon.html`.

[14] "RIPE BGP Beacons," `http://www.ripe.net/ris/beacon.html`.

[15] Tim Griffin, "What is the sound of one route flapping?," Network Modeling and Simulation Summer Workshop at Dartmouth, July 2002.

[16] Lan Wang, Xiaoliang Zhao, Dan Pei, Randy Bush, Daniel Massey, Allison Mankin, S. Felix Wu, and Lixia Zhang, "Observation and Analysis of BGP Behavior under Stress," in *Proceedings of Internet Measurement Workshop 2002*.

[17] Timothy G. Griffin and Brian J. Premore, "An Experimental Analysis of BGP Convergence Time," in *Proceedings of ICNP 2001*.

[18] Pedro Roque Marques), "Bgp route advertisement interval," Talk at RIPE 45 Meetings.

[19] Henk Uijterwaal, "Routing Beacons," `http://www.potaroo.net/iepg/november2002/`, November 2002.