BGP convergence analysis

Diplomarbeit

Angefertigt nach einem Thema von
Prof. Anja Feldmann, Ph.D.
am Fachbereich Informatik
der Universität des Saarlandes

Betreuer: Olaf Maennel

von

Sara Bürkle

Saarbrücken, 16. Juni 2003

Hiermit erkläre ich an Eides statt, dass ich diese Arbeit selbstständig verfasst und keine anderen als die im Literaturverzeichnis angegebenen Quellen benutzt habe.

Sara Bürkle
Saarbrücken, 16. Juni 2003

Contents

List of Figures

List of Tables

Chapter 1

Introduction

During the past few years, the Internet has become a mission-critical component of our communication infrastructure. Most communication on the Internet is based on data transfer over connections between pairs of hosts. For this purpose, data is divided into small data packets, and each of those packets crosses the Internet independently from the rest.

In most instances, the data is not transferred on one direct physical connection between sender and receiver. Instead, the packets travel over intermediaries through this huge network that we call the Internet. The intermediaries are termed *routers*. This is because they have to decide for each data packet they receive which neighbouring router they forward it to en route to its final destination. In this way, routers determine the route a data packet takes through the Internet.

Obviously, routers have to gain knowledge about possible destinations. They have to know where the destination is located, whether it is reachable from their location, and by which path. So whenever a topology change occurs, the router should be informed as soon as possible in order to make a sound forwarding decision. As long as the involved routers negotiate the new routes but try to route data on invalid or contradictory paths, data may never reach its destination. This time of unsettled path selection after a topology change is the *convergence time*. Its duration is of critical interest to the stability of the Internet.

As the global Internet encompasses a huge number of computers, routing is a complex task. A part of this task is accomplished by the Border Gateway Protocol (*BGP*), used by hosts from different networks to exchange reachabil-

ity information.

BGP usually shows a convergence time in the order of minutes [25] and reacts quickly and dynamically to topology changes. Unfortunately, convergence can be delayed by unknown factors: Recent work by the research community [5, 10, 12, 13, 17, 20, 21, 22, 27, 30, 38] has shown that the dynamics of the protocol are poorly understood.

In the mid-90s, flapping routes, i. e. routes that kept changing to and fro, accounted for much of the instability observed in the global BGP system. It can be shown that routes may oscillate notoriously with uncoordinated configuration or for technical reasons [27]. Route-flap damping [41] was introduced into BGP to suppress nonessential BGP traffic dealing with those oscillating routes.

Today, route-flap damping is suspected to convey delayed convergence as a negative side-effect of suppressing unnecessary traffic. Damping in fact seems to overregulate the traffic, but it is not yet understood when and why this happens [26, 4].

On the whole, BGP is difficult to analyze because of its complexity, because of the size of today's Internet and because of its distributed character. To simplify BGP analysis, it was proposed that so-called *BGP Beacons* [3, 36] be installed. BGP beacons are routers that announce special unused network addresses in a previously defined and publicly documented way. 13 BGP beacons were set up during the summer of 2002 in distinct areas of the Internet. Since that time, the traces they leave in BGP traffic through their well-defined activity can be observed.

The goal of this thesis is to find patterns in observed BGP beacons reachability information, to characterize them, and to transfer these patterns to the analysis of global BGP data. This is done with statistical analysis of those parts of archived BGP traffic that refer to the reachability of the BGP beacons' prefixes.

The analysis of four months of BGP beacon traffic clearly reveals effects of ill-applied route-flap damping as well as instances of long convergence times. Most of the traffic, however, shows the usual convergence times of a few minutes. It is thus possible to divide the available beacon data into three classes with different statistical properties. This breakdown of the conver-

gence processes into distinct types helps to identify factors for convergence time. These factors, together with further insight gained by the BGP beacon study, help interpret patterns in convergence processes in global BGP data. In this way, dependencies between different timers and topology become much clearer, and more aspects of the big picture in BGP convergence behaviour are revealed.

Reader's Guide

The remainder of this thesis is organized as follows:

Chapter 2 gives an overview of routing in the Internet, especially the inter-domain routing protocol BGP (Border Gateway Protocol).

Chapter 3 describes our Internet routing data collection and analysis architecture and introduces metrics and terminology used in this thesis.

Chapter 4 analyzes BGP beacon traffic. BGP beacon traffic is a sample part of BGP traffic where the schedule of announcements and withdrawals of the networks is well-defined. After describing its characteristics, this chapter presents a classification of convergence behaviour that is deduced from four months of observed traffic.

Chapter 5 transfers the classification given in Chapter 4 onto global BGP traffic. In the global data, the convergence factors identified in the beacon analysis reappear, and characteristics of global convergence behaviour can be inferred by comparing the two studies. Additionally, the chapter estimates the impact of route-flap damping in the Internet.

Chapter 6 summarizes my findings, describes areas for future work and provides some concluding remarks.

Chapter 2

Background

The Internet is divided into a large number of different regions under autonomous administrative control, called *autonomous systems (ASes)*. Routing through the Internet depends on routing between ASes (*Exterior Gateway Protocols* [14]) and on routing inside the ASes (*Interior Gateway Protocols* [28]). Each AS usually employs special-purpose computers, or *routers*, concerned primarily with routing.

For Exterior Gateway Protocols, the *Border Gateway Protocol*, **BGP** [40] is the de facto standard. At the boundary of each autonomous system, so-called *border router*s from different ASes exchange routing information using BGP [16]. They interchange reachability information to destination IP address blocks, called *prefixes*. A prefix consists of two parts: the IP address[1] and the address mask indicating the size of the network.

Prefixes may each represent a single network or an aggregation of several network addresses. I therefore use the terms prefix and network synonymously. I will give an example for prefix aggregation: One AS provides Internet connectivity to four customers. Customer A has registered for the use of prefix a, namely `192.168.0.0/24`, customer B's network has prefix b, or `192.168.1.0/24`, and customers C and D accordingly preside over prefixes c and d, `192.168.2.0/24` and `192.168.3.0/24` respectively. When all customers A to D are reachable, the AS can aggregate prefixes a, b, c and d into one larger network or shorter prefix, prefix `192.168.0.0/22`. This aggregated prefix will be announced by the AS to the outside world. It incorporates four

[1] currently still mainly IPv4 addresses

more-specific prefixes, the prefixes of its four customers.

2.1 BGP basics

Today, BGP is the only Exterior Gateway Protocol used between ASes. It is an *incremental* protocol that sends update information only upon changes in the network topology.

BGP is a variant of the class of distance-vector protocols, where neighbouring routers exchange link cost information to possible destinations. BGP itself is a so-called path-vector protocol: Instead of distributing cost information it propagates full path information to each destination to avoid cycles.

If two ASes want to exchange traffic, they typically establish a BGP session between the incident routers; these routers are called (BGP) *peer*s. A BGP session is established between a pair of routers and uses TCP [39] as its underlying reliable transport mechanism. In a BGP peering session, the peers can send four kinds of messages:

OPEN and NOTIFICATION messages are used to open up a peering session or to tear it down because of errors, respectively.

KEEPALIVEs are used between peers to make sure that the connection still exists during periods of inactivity.

UPDATE messages carry network reachability information. An update either advertises a prefix or withdraws a previously announced prefix.

If a peer advertises a prefix, this can be seen as a commitment from the sending peer to reach the specified destination. By withdrawing a prefix, the sending peer indicates that it can no longer reach the destination. Peers in fact only exchange *best routes* to prefixes. This information is updated every time that the best route changes.

A route advertisement for a particular prefix includes a list of ASes that constitute the AS path to this prefix. A packet sent to the network labelled by this prefix would travel along this path. The route advertisement further contains much more information in the shape of attributes.

An example announcement update in human-readable format may look like this:

```
TIME: 01/01/03 00:05:34
TYPE: BGP4MP/MESSAGE/Update
FROM: 202.12.28.190 AS4777
TO: 193.0.0.1 AS12654
ORIGIN: IGP
ASPATH: 4777 2500 2500 2500 7660 22388 7570
NEXT_HOP: 202.12.28.190
ANNOUNCE
  192.207.156.0/24
```

The above update carries the information that peer `202.12.28.190` sent an announcement for prefix `192.207.156.0/24` on Jan 1, 2003 at 00:05:34 GMT. The AS path (`ASPATH`) by which the respective prefix is reachable is:

```
4777 2500 2500 2500 7660 22388 7570
```

In other words, prefix `192.207.156.0/24` at that time is reachable via an AS path consisting of a sequence of autonomous systems, starting with the peer AS `4777`. We say that the prefix is *originated* by AS `7570` and *transiting* ASes `22388`, `7660`, `2500`, and `4777`. Note that BGP updates travel from the originating AS through the transiting ASes to the observed peer, but the data packets that need to be routed take the other direction, they travel the AS path as given in the updates, read from left to right.

The BGP attributes provide further information about the prefix. Among other things, the `ORIGIN` attribute indicates where the path information came from: `IGP` (Interior Gateway Protocol) means that the network layer reachability information was introduced into BGP by the IGP of the originating AS. This can be the case when new prefixes are introduced inside an AS. Other possible values are `EGP` (the route is learned via an Exterior Gateway Protocol) and `INCOMPLETE` (the information is learned by some other means, usually manual configuration).

The `NEXT_HOP` attribute tells the router to which IP address it should forward packets. In our example, the `NEXT_HOP`-IP is `202.12.28.190`, the IP of the peering router itself. This is not necessarily the case since BGP allows third-party next hops.

There are other attributes available in BGP, e. g. the *Local Preference* at-

tribute, the *Multiple Exit Discriminator* (MED) and *Community* attributes. The local preference attribute is used within an AS to implement local policies for the best exit point. It is a valuable tool for ISPs to influence their costs for outgoing traffic. With the Multiple Exit Discriminator, an AS can indicate the best entry point to its neighbouring AS in case of multiple connections. The community attributes can be used to "colour" routes and to organize them into classes. Once coloured, the routers can control the distribution and acceptance of routes with a particular colour.[2]

Neighbouring routers exchange updates for prefixes whenever their best route changes. A router thus receives best routes from all of its peers for a number of prefixes. When looking at the updates the router receives, there may be updates from several peers about the reachability of one prefix. The inter-dependencies of updates from different peers or concerning different prefixes are far from being understood. Some of those updates may even be independent. Therefore, for statistical analysis, I group updates by peer and by prefix.

I will explain the reasons for this in more detail: On the prefix side, even if two prefixes represent two customers of the same ISP (i. e. same originating AS), one customer may become unreachable because of a router failure independently from the other customer. On the peer side, two example peers may be located at two very different locations in the world and in the BGP graph. Therefore, their knowledge of the reachability of prefix p may differ. They also receive and send updates at different times even if their AS paths to prefix p are similar or identical.

The independence of updates from different peers about one prefix is illustrated in Figure 2.1 on the next page where one prefix p is depicted by the black rectangle at the top. Prefix p can be reached via three of the router's peers, represented by the double circles below. Figure 2.1 on the next page presents the multitude of possible paths from each of these peers to prefix p that may be announced to them from some neighbour at some point in time. Some of those paths may get mentioned to the router in an update, if they were best route.

[2] I will use spelling for Canadian English in this text, e. g. colour and neighbour, but analyze and visualization.

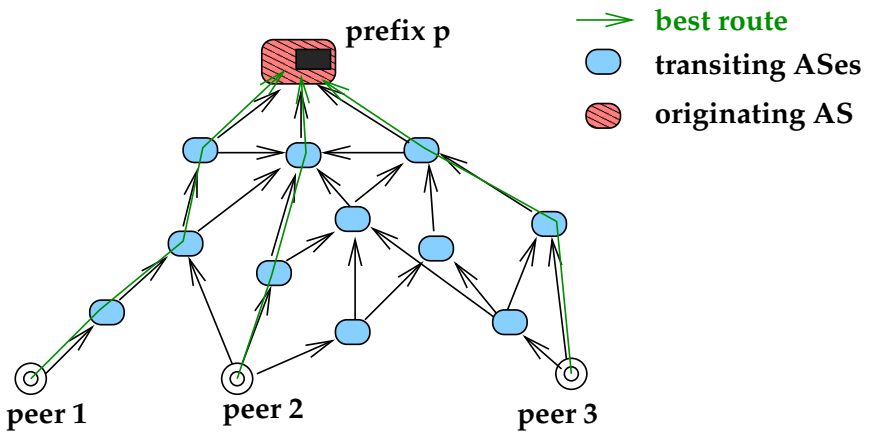

Figure 2.1: Illustration for independence of peer-prefix pairs

A snapshot of each peer's best route to prefix p is symbolized by the three green arrows leading from a peer through a number of transiting ASes (nodes) to the originating AS, the node containing prefix p. Of course the choice of the direction in which the arrow points is ambivalent, but packets sent from a peer are supposed to travel from the peer through the transiting ASes towards the originating AS.

Even if the router peers with two routers in the same AS both having the same best path, it is likely that the update timestamps of the updates will be different because each peer has its own location, configuration, processing load and timers. It is immediately obvious how to interpret time intervals between updates only for consecutive updates seen from one peer about one prefix, a so-called *peer-prefix pair*. That is why I am always grouping updates by peer and prefix before considering the statistical properties of those updates.

2.2 BGP route selection

The first quality of BGP is to provide reachability to the networks in the Internet. But as was seen above, BGP also provides an AS with the possibility to influence both incoming and outgoing traffic flows. In this way, BGP supports *policy* decisions in each AS, and this policy support is one of the most fundamental features of BGP to be deployed world wide: Routing in the Internet sends data packets over physical links. These links must be built, paid

and maintained, thus usage of a link is not for free. Neighbouring ASes sign contracts about link usage and fees (e. g. by traffic volume), and different links may have different costs, so every AS wants to control which links it uses for data transfer. This is done by policies that very often reflect business contracts.

Upon reception of a BGP advertisement, an AS applies its local policies to decide whether to accept a route. These decisions can be based on a variety of factors, such as the length of the AS path and local preferences for particular downstream providers. If the received route is in accordance with the local policies, it is added to the BGP routing table and stored on the router. As a router normally has more than one peer, such a BGP table is stored in the router for each BGP peering session. Only the best route is used to forward packets for a prefix. The peer though remembers all alternative routes to a prefix it has heard from some neighbour. In this way, an alternative route can be chosen immediately when the best route is no longer available.

The decision process for the best route is strongly influenced by the local policies at the AS. Those policies may change now and again, e. g. when a new business contract is signed. The policies thus reflect business concerns and autonomous routing decisions, and many ASes keep their policies confidential. Only at a few observation points, at the BGP collectors, are best-route announcements publicly available.

If the best route for a prefix changes, a second local policy is applied to decide whether or not to forward the advertisement to its peers. If it is forwarded, the router may adjust some of the attributes. It will at least add the unique AS number of its autonomous system to the AS path. Each AS on the path adds its number when the update is forwarded to a neighbouring AS. Some ASes add their AS number several times, like AS `2500` did in the example above in order to influence the preference of the route at other autonomous systems. In this way, BGP is also used to influence traffic flows into and out of the AS.

If a router receives a withdrawal, it deletes the entry in the BGP table. If the withdrawn route is currently the best route, the router looks into the routing tables for an alternative path to the prefix and either selects a new best route or marks the network as unreachable. If the prefix is now unreachable, the

router must send a withdraw notice to all peers concerned, i. e. to all peers that knew of the route through an earlier announcement.

On a lower level, when a data packet is received, the router has to forward it to the next physical neighbour en route to the final destination of the packet. The rest of the packet's path is of no immediate concern in this forwarding decision. This next hop to the neighbour may cross AS borders, but the path may lead through the router's own AS towards the best exit point first. The router therefore combines BGP's best route to a prefix with IGP information to determine the outgoing link for each prefix. The mapping of prefix and outgoing link is then stored in the *forwarding table*. Whenever the router receives a data packet that needs to be routed towards its destination, the router performs a longest prefix match for the destination address in the forwarding table to select the outgoing link for the packet.

Usually, routers send updates only upon changes of their best routes. A special case occurs when two routers initiate a BGP session: The two peers have to exchange their BGP routing tables. This is done by sending a BGP update for each prefix in the BGP routing tables. Depending on the number of prefixes in the routing table[3], this can create a large number of updates within a short time period. As routers often have to be rebooted when their configuration is changed, these updates may influence statistics concerning the number of updates. Depending on the analysis, these updates may have to be filtered from the observed BGP traffic. This filtering can be done rather easily for session resets on immediate neighbours, and there are heuristics proposed to deal with session resets on routers somewhere on the AS path [25].

To recapitulate the above: Whenever the best route to a prefix changes (be it because of a better route, a withdrawal or different attributes), the router sends the new best route to suitable peers (depending on the policy). If no announcement or withdrawal is sent for some specified time period, BGP sends a KEEPALIVE message to determine if the session is still alive. If the corresponding timer expires without reception of either updates or keepalives, the session is reset. In other words, the peering session is torn down and

[3] Nowadays, a "default-free" routing table contains about 120 000 prefixes. A router with "default route" just delegates the routing decision onto another router.

needs to be established again. If a router notices a session as being down, the corresponding peer's routing table has to be emptied and all best routes using this connection have to be deleted from the BGP table. In consequence, the appropriate updates, either announcements of alternative routes or withdrawals, have to be sent to other BGP peers.

2.3 BGP timers

Like most routing protocol specifications, BGP includes a minimum per-prefix advertisement timer to limit the BGP update rate. Without a limiting element, a peer would send updates to its neighbours every time any best route changed, even if the route to one prefix changed several times in a few seconds. In BGP, this timer is called the *Minimum Route-Advertisement Interval timer (MRAI timer)* [32]. The MRAI timer in current router products is usually implemented as a jittered per-peer timer and has a typical value of 30 seconds. [4]

A *jittered* timer is a timer that uses randomly varying values. The MRAI timer, for example, has a typical value of 30 seconds. But were exactly 30 seconds used at all routers, this could lead to self-synchronization of the routers. This would imply that all routers would send their updates at the same time and pace, every 30 seconds [7]. This is an undesired effect in BGP. It can be avoided by varying the timer interval randomly at each router. The timer is implemented in such a way that values between 25 and 30 seconds are normally observed.

The MRAI timer can cause multiple updates to be seen on one peering session. This procedure is illustrated in a simple example in Figure 2.2 on the next page: AS1 has added a prefix P and is therefore sending a BGP update to AS2 and AS3 for P with AS path: AS1. This update is received by AS2, added to the routing tables, and sent to AS4, since AS2 has not sent an update to AS4 within the last 30 seconds. AS4 receives the update, adds the prefix to its routing table and forwards the update to AS5. AS3 also receives the update and adds it to its routing table, but instead of sending the update immediately to AS4, AS3 has to wait until the Min-Route Advertisement Interval timer ex-

[4] The BGP specification recommends a 30 second timer interval for outgoing transit route advertisements.

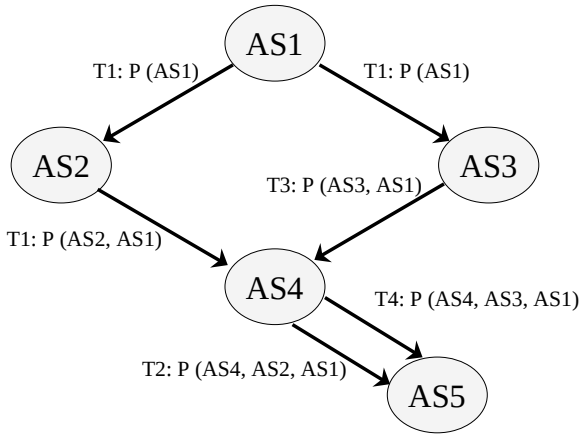

Figure 2.2: Example: BGP update propagation.

pires. Once AS4 receives the update from AS3 it realizes that this is a better path and reupdates its routing table entry for this prefix and sends another update for prefix *P* to AS5. In this rather simple example AS1 originated one update for prefix *P*, yet AS4 is sending two updates for the same prefix.

This is only one example how updates can be spawned. The MRAI timer is not even necessary for it: AS2 can forward the update more quickly than AS3 because the geographical distances between the routers are smaller, or because AS3 is larger and more AS-internal routers have to forward the information etc. It mainly depends on the topology and on the number of alternative paths that are available if one originated update will be observed as several updates at another peering session.

Early work on BGP quickly showed that, if left unconstrained, the peers tend to exchange a vast number of updates for flapping (oscillating) routes[5]. Route-flap damping [41] tries to avoid that an oscillating route pollutes the Internet with updates. The main idea is to remember the history of updates for a prefix and to "hold-down", or refuse to believe, updates that exceed certain parameters for expected behaviour. For example, a certain number of updates per hour can be too high. Other algorithms impose damping in a progressive fashion: the more flaps the longer the suppression times.

RIPE recommends parameters for route-flap damping [29], namely suppression periods of 60 minutes for /24-networks, periods of 30 – 45 minutes

[5]It is possible to distinguish between flaps and oscillations. I do not make this distinction.

for /22- to /23-networks, and 10 – 30 minutes for all others. Damping should not start before the 4th flap. The parameters recommended were not verified by simulations or tests. They are rather harsh, and the reason for this is probably that research had shown that uncoordinated BGP configurations as well as unstable links or interfaces may lead to persistent route oscillation [27],[13]. Indeed, route oscillation can be a serious problem to the BGP convergence process.

2.4 BGP convergence

BGP is an incremental protocol that sends updates only if something with regard to topology and/or reachability changes. One may suspect that if there is a change, it is followed by several updates sent between peers. Once all peers have settled on their best routes, the BGP world again becomes quiet. The time between the change and the time where all peers have chosen their new best route represents the concept of convergence time: A change causes BGP updates to be sent, and when all peers have settled on their new best route, the convergence process is finished.

This is the naive conception of what happens in BGP. However, due to the size of the Internet (currently about 15 000 ASes, 120 000 prefixes) and many other factors, there is constant noise in BGP traffic. It is thus necessary to look at the convergence process more closely.

Whenever the reachability of a network changes, there must be a cause for it. It is called the *instability event*. A partial list of the possible reasons is given here:

- failure or repair of a physical link
- reboot of a router (e. g. for software updates)
- policy changes in originating or transiting ASes
- addition/deletion of network prefixes.

The instability event can disrupt connections within an AS or between two ASes. When an AS-internal connection is involved, this triggers a convergence process within the AS, in the Interior Gateway Protocol [28]. Connecting IGP events to BGP, normally called IGP redistribution into BGP, is contrary to recommended operational practice. BGP normally should not be involved

in this instability event. The IGP shows a convergence time for its own convergence process though.

When the instability event affects connections between ASes, it causes an AS to withdraw a previously announced route or announce an alternative path. If we look at the effects of an instability seen in BGP, there is the time of the instability event, followed by the first BGP update that is sent by the affected BGP peer. This new reachability information propagates in some way through the net of border routers and eventually, there will be no new updates to be sent, and that is the point where the net is considered to be converged concerning this prefix. Seen from the point of view of one router, this router may receive a number of updates from every peer, compute its own best route and announce it to its peers. The whole procedure will be called a *convergence process* in BGP.

Note that there is a fundamental difference between the two concepts of *reachability* and *convergence*: When a new prefix p is added and a peer announces a route to it at time t_0, then starting from time t_0 the prefix p may be reachable. But as long as peers change their best route to p due to further information, data packets sent for communication purposes may get lost along the way. Hence the reachability of prefix p only is reliable once a stable route has been found. A peer may change its mind regarding the best route several times after the initial instability event. During this time, BGP has not yet converged on behalf of prefix p, although p may already be reachable.

Whenever a convergence process occurs in BGP, the convergence time can be defined in different ways: One convergence time is the time from the instability event to the stable best route on all routers in the Internet, i. e. to the time when the global BGP graph is converged. This *Internet-wide convergence* is a primarily theoretical concept, since no one has global knowledge of the entire Internet. Furthermore, the time it takes inside a router to make a decision can be measured as an *intra-router convergence time*. Since we usually look at BGP updates exchanged between two neighbouring routers and do not have any information about the time when an instability event happened, the *convergence time* in this thesis is measured as the time from the first update concerning the prefix to the last update. If the time of the instability event is known, it is possible to additionally estimate the time for

the total convergence process at an observation point: It is the time from this instability event to the stable best route on the observed router.

2.5 BGP analysis

So far it is impossible to point to the exact factors that lead to short or long convergence time. It seems clear however that the timers, MRAI and damping parameters, play a major role. Recent research from Z. Mao et al. [26] has shown that route-flap damping can delay convergence in cases where the original event is just a single flap. R. Bush et al. [4] suggest that damping should be considered harmful.

As BGP's dynamics in general are poorly understood, this intensified the endeavours to install the so-called BGP beacons. BGP beacons are unused prefixes with well-defined update schedule for research of BGP convergence in general and damping in particular. It is much easier to understand a series of updates seen at some peer if the original reachability change is known. So in fact, BGP beacons provide the first possibility to study the former of the two last-mentioned convergence times, namely the time starting with the instability event and ending with the update that leads to a stable state.

The talk of R. Bush et al. [4] at RIPE 43 convinced RIPE's RIS project to assist the beacon studies by adding the "Routing Beacon" functionality to their Remote Route Collectors (RRC) [34, 36][6]. With RIPE and other sponsors, it was possible to install 13 beacons in the course of summer 2002 [3].

BGP was designed in the mid 90s, for a much smaller Internet, but it survived several years of exponential growth with the help of IP address reorganization (CIDR) and other modifications. It is still an open research question how much more growth BGP will be able to sustain [19]. Apart from this, the research community has identified two main weaknesses concerning BGP, on the one hand security considerations, on the other hand the lack of understanding of BGP convergence properties [11, 8, 2]. The goal of this thesis is to improve the understanding of BGP convergence by way of analyzing BGP beacon traffic.

[6]The Routing Information Service, RIS, is actually offered by RIPE NCC, the RIPE Network Coordination Centre. The same is true for most services, which, for brevity, I associate to RIPE within this text.

Chapter 3

Definitions, Terminology and Methodology

In the last chapter, the background for understanding routing in the Internet, especially the routing protocol BGP, was given. This chapter builds upon this knowledge and explains the technical details used for BGP traffic analysis. I will provide information concerning where the BGP data analyzed comes from and what is done with it, as well as the terms and concepts needed for the statistical analysis in the following chapters.

3.1 BGP data collection architecture

As explained in Chapter 2, BGP reachability information is exchanged between pairs of routers, or BGP peers. They are said to be peering, in a peering session, and exchange reachability information as well as session information via TCP.

The basic idea for collecting and archiving BGP data is to use a pseudo-BGP speaker, the *BGP collector*, to only receive such reachability information from some willing peers while not sending them any reachability information. The collector logs all messages it receives from its peers archive them.

In September 1999, the first BGP archive was made publicly available. Most collectors available now have only been operating since about 2001. Before the collection of BGP data was started, BGP monitoring on a global scale was infeasible. BGP analysis has become more representative with the growing number of collectors: Data from more peers enables better analy-

sis, because it enables cross checks, gives redundancy in the case of collector failures and provides more different views on the global Internet.

One provider for BGP archives is RIPE (Réseaux IP Européens) [33], which offers nine different *Remote Route Collectors (RRC)*, RRC00 to RRC08, with two to twelve peers per collector. Another source of BGP data is Oregon's Route-views Project [35], which currently collects BGP traffic from 26 peers. In addition, our research group has access to the data from a local ISP, Saargate [37] with two peers.

The routing software used as BGP collector is Zebra [42]. The archived data does not only include the BGP traffic exchanged between the peers (updates) but also snapshots of the BGP tables. Updates and tables from RIPE and Routeviews are publicly available [34, 35].

3.2 BGP data analysis architecture

The tool `route_btoa` from the Multithreaded Routing Toolkit [18] is used to decode the BGP information into single prefix announcements and prefix withdrawals. The syntax of the withdrawal updates after being decoded by `route_btoa` is as follows:

```
Prot|Time|W|PeerIP|PeerAS|Prefix
```

An announcement update looks as follows (in one line):

```
Prot|Time|A|PeerIP|PeerAS|Prefix|ASpath|Origin|Nexthop|
    Local_Pref|MED|Community
```

These data fragments will be called *updates* throughout this thesis. They are not identical with the TCP data packets or BGP UPDATE messages that are exchanged between peers. These may each contain tens or hundreds of prefix announcements as well as a multitude of prefix withdrawals. Whenever the *number of updates* is mentioned, we refer to the total number of single prefix announcements or withdrawals.

Note that withdrawal and announcement follow the same pattern. The third field indicates the update type, either `W` for withdrawals or `A` for announcements. However, a previously announced AS path can be withdrawn

implicitly with the announcement of an alternative route. This is called an *implicit withdrawal*. This saves the sending of one update for that prefix. It should be noted that without knowledge of the actual state of the router, neither withdrawal nor implicit withdrawal reveal the previous best route to the prefix.

The use of the BGP attributes (local preference etc.) was described in Chapter 2. This thesis is confined to the analysis of the following parts of the updates:

- timestamp,
- prefix,
- peer IP,
- type of the update (announcement or withdrawal) and
- in the case of announcements: AS path.

The other information contained in an update will not be considered here.

3.3 Definitions

In the following chapters, several notions introduced here will be used extensively. All terms are linked in some way with *convergence time*. Convergence time is the time between an instability event and the time when the new routing state has become stable.

In practice, we consider the updates that one router is sending to another router. This is what a BGP collector gathers. At this specific observation point (from a peer of a BGP collector), some time after an instability event, there will be one or several updates in a row observed on behalf of the affected prefix. If the time between a pair of consecutive updates, their *interarrival time*, is "short", then these two updates are presumed to belong together, i. e. to be caused by the same instability event.

Obviously, "short" is a very relative term, and in fact the values chosen in BGP analysis to group updates vary considerably. Several consecutive update pairs with short interarrival times for the same peer form a sequence, called *update burst* [25], or simply *burst*. In an update burst, all pairs of consecutive updates with interarrival time smaller than some previously defined *timeout*

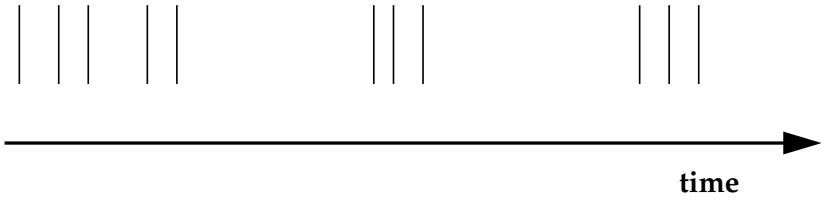

Figure 3.1: Example sequence of updates.

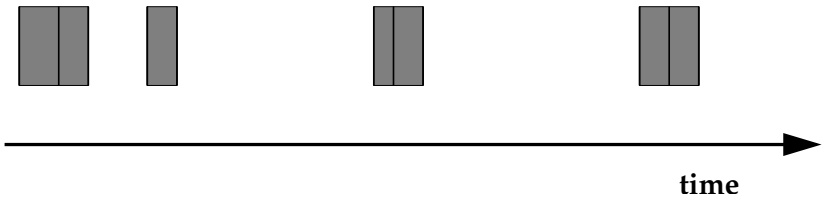

Figure 3.2: Updates grouped into bursts with small timeout value.

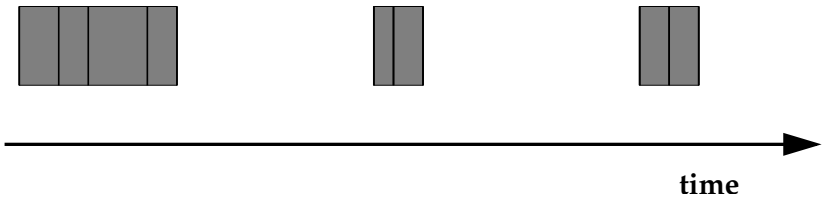

Figure 3.3: Updates grouped into bursts with medium sized timeout value.

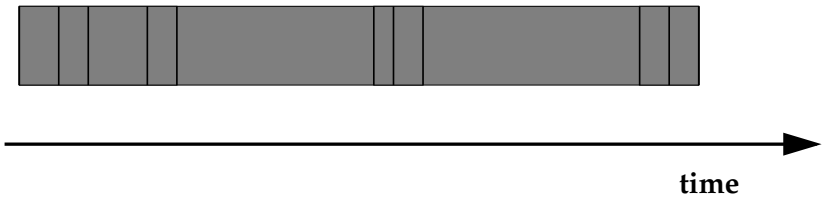

Figure 3.4: Updates grouped into bursts with large timeout value.

value are grouped together. This timeout value can be of the order of minutes or of an hour [23] depending on the purpose of the study.

Figures 3.1 to 3.4 illustrate how this works. The choice of the timeout value strongly influences the number of bursts and the number of updates in one burst. Figure 3.1 symbolizes the arrival of an update on the time line below as a short vertical segment. In the picture, 11 updates for some prefix are seen

at the peer. The interarrival time between the first and second update is only slightly larger than the one between the second and third. The gap between updates number 3 and 4 is a bit larger, such that the first 5 updates may be grouped into two different update bursts. This is the case for a timeout value larger than the interarrival time between the first two updates but smaller than the interarrival time between updates three and four. In Figure 3.2, the bursts generated with small timeout value are illustrated as shaded rectangles. The duration of a burst is the time difference between its first and its last update, i. e. the time represented by the length of the rectangle from the left to the right border.

Figures 3.3 and 3.4 visualize the update bursts that result from choosing a larger timeout values. The bursts are again represented as shaded rectangles.

A convergence process is expected to result in one update or a small number of updates, i. e. in one small update burst observed at a peer. Since route-flap damping may damp a route for 60 minutes, one may choose a timeout value larger than 3 600 seconds to collect all updates related to a single instability event in one burst. An update burst then presumably captures one convergence process, and the last update in the burst represents the *stable state* that was reached through the convergence process. In case of an announcement as the last update, I will also talk of a *stable path* instead of a stable state, because this path remains valid, i. e. in use as best path, at least as long as the timeout value.

However, one may want to use a smaller timeout to analyze interactions related to route-flap damping. With small timeouts, damping can be observed to separate one convergence process into two update bursts. Since the MRAI timer takes 30 seconds, it usually does not make sense to use a timeout value smaller than this value for analyzing BGP update bursts. But the choice of the timeout value depends on the object under investigation, and as of now, the research community has not come to terms as to which timeout value is appropriate for the analysis of a particular aspect of BGP.

The notion of an update burst is similar to the idea of a flow in TCP. Source and destination IP address and port number are used to identify a flow, but those IP addresses and ports may be reused together, so additionally a timeout value is used to discern different flows with the same source and destina-

tion IP and port.

Another important notion in this analysis is the *AS path length*: Only announcements contain an AS path, and the AS path length consequently is the number of AS numbers (ASNs) in the path to a prefix. For the AS path length, one can either use the number of ASNs in the path or the number of *unique* AS numbers in the path. Both versions can be justified: In some cases, one wants to know how many autonomous systems the data has to cross to reach a prefix: the number of unique AS numbers. However, best path decisions may rely on the total AS path length, and that is the point where the total number of ASNs is interesting. Both AS path length versions will be used here, and it will be made clear in the text of what kind the AS path length is.

Normally, we do not know the instability event that caused an update burst. In these cases we can only measure the *(burst) duration* of a convergence process, i. e. the time between the first and the last update of the burst.

But there is a special case where the time of the instability is known because the instabilities are introduced on purpose: the BGP beacons. They are special unused prefixes created to study BGP dynamics.

There are 13 beacon prefixes at work at the time of this writing. They usually have a fixed schedule with four-hour periods. For example, at 00:00 GMT, a beacon prefix is announced, and two hours later, it is withdrawn and this cycle repeats. The exact schedule is publicly documented [3]. The update patterns observed at a peer following a *beacon event* are sometimes referred to as *echo*. There can be an echo of each beacon event at each peer.

In the special case of the BGP beacons, we know the exact time of the instability event that initiated the convergence process. Therefore, we can estimate the duration of the total convergence process from the instability event until a peer has reached a stable state. It is referred to as *beacon duration*. It is the time from the beacon event up to the last update following it.

Every publicly documented announcement or withdrawal of a beacon is called a *beacon event*, or simply *event*, whenever the meaning is clear from the context. The term *event* sometimes also refers to the two-hour time period starting at the beacon event and ending right before the following beacon event[1]. Events that announce a beacon prefix are called *announce events*, or

[1]The two hours following the beacon event are identified with it the way sovereign territory is

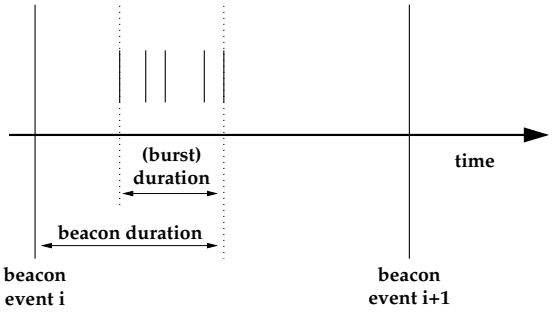

Figure 3.5: Illustration of duration and beacon duration.

sometimes *A-events* for short. The beacon events where the prefix is withdrawn are thus called *withdraw events* or *W-events*. Beacon events are a special case of an instability event because time and type are known.

Beacon duration and duration must not be confused. They refer to different values in observed BGP beacon traffic. As those two concepts will be very important, I included Figure 3.5 to illustrate the difference: During the beacon study, updates are assigned to a specified, fixed two-hour bucket that starts at one beacon event and ends at the next beacon event. The updates seen at one peer about one beacon prefix are symbolized again by short vertical segments. The beacon duration is the time from the beacon event to the last update seen at a peer, and the duration is the time from the first up to the last update of a burst. In part of the beacon study, I joined all updates observed during one beacon event into one update burst, regardless of the previous and following beacon event. The duration then is the time difference between the first and the last update seen at a peer. This implies that the duration is always less than or equal to the beacon duration.

The BGP beacons offer a wonderful opportunity: I studied the statistical properties of the BGP beacons' echos in four months of observed traffic. There are many conclusions that can be drawn from the observed BGP beacon traffic simply because there is more knowledge available about the instability event causing the updates. I will in the following chapter show what one can infer from the behaviour of the special prefixes that were created for research purposes.

identified with a country.

Chapter 4

BGP Beacons

In this chapter, I will analyze BGP beacon traffic to identify the factors leading to short or long convergence times. To this end, the observed beacon events are classified into three classes, each class representing a distinct convergence behaviour. Due to the special properties of the BGP beacons, one can draw new conclusions with regards to the global BGP convergence processes.

4.1 Beacon prefixes

In this chapter, I will consider updates for the BGP beacons only. There are 13 beacon prefixes at work at the time of this writing. They are composed of nine RIS beacons (each RRC collector announces one beacon prefix) and four privately sponsored beacons. The RIS beacon prefixes and their location are shown in Table 4.1 on the next page. They are announced ("network *a.b.c.d*") at 00:00h, 04:00h, 08:00h, 12:00h, 16:00h and 20:00h (GMT) with the corresponding withdrawals ("no network *a.b.c.d*") two hours later. All RIPE beacons have source AS `12654`, the RIPE ASN [33].

The addressing scheme of the four private beacons can be seen in Table 4.2 on the next page. While their schedule differs from the RIS beacons, they also use two-hour intervals. The exact schedule can be seen in Table 4.3 on page 27, where the schedule of the RIS beacons was also included.

To facilitate reference, I associate an ID with each beacon prefix, see the first row of Tables 4.1, 4.2, and 4.3. All RIS beacons are *R*-beacons in short, and Beacons 1 to 4 are the non-RIS BGP beacons (see Table 4.2). For the RIS

ID	Prefix:	RRC:	Location:
R_0	195.80.224.0/24	RRC00	RIPE NCC Amsterdam, NL
R_1	195.80.225.0/24	RRC01	LINX London, UK
R_2	195.80.226.0/24	RRC02	SFINX Paris, FR
R_3	195.80.227.0/24	RRC03	AMS-IX Amsterdam, NL
R_4	195.80.228.0/24	RRC04	CIXP Geneva, CH
R_5	195.80.229.0/24	RRC05	VIX Vienna, AT
R_6	195.80.230.0/24	RRC06	NSPIXP2 Otemachi, JP
R_7	195.80.231.0/24	RRC07	Netnod-IX Stockholm, SE
R_8	195.80.232.0/24	RRC08	MAE-WEST San Jose, CA, US

Table 4.1: RIS beacons: Addressing scheme and location.

ID	prefix	source AS	Beacon Host
1	198.133.206.0/24	3927	Randy Bush
2	192.135.183.0/24	5637	David Meyer
3	203.10.63.0/24	1221	Geoff Huston
4	198.32.7.0/24	3944	Andrew Partan

Table 4.2: Non-RIS beacons: Addressing scheme and hosts.

beacons, ID R_0 is associated to the beacon at RRC00 etc. up to beacon R_8 of RRC08.

Beacons R_0 through R_8 are timed by GMT, and the timestamps of the observed updates are captured in epoch seconds (also GMT)[1]. One can thus easily create two-hour bins for each beacon event. It starts at the time specified by the publicly documented schedule and ends one second before the next beacon event. As all time information is in GMT, there is no trouble in assigning all beacon updates to their respective two-hour bin.

However, beacons 1 to 4 are timed by local time, for example Pacific Time, which implies a shift of the schedule with respect to GMT in spring and autumn, when Pacific Standard Time changes to Pacific Daylight Savings Time and vice versa[2]. Fortunately, the timestamps in the BGP updates are in GMT

[1] Greenwich Mean Time is the same as Universal Time (UTC), only the definition is different. GMT orients on mean day lengths whereas the UTC operates on the fixed second length defined on atomic radiation. With leap seconds introduced into UTC, it is made sure that UTC is not further away from GMT than 0.9 seconds [15].

[2] Routeviews and Saargate collect their archive files under file names containing local timestamps. This is actually a bad idea as one hour of archive files gets overwritten in autumn, at the end of daylight savings time, when the clock is turned back one hour.

ID	prefix	Interval	First daily announcement	First daily withdrawal
1	198.133.206.0/24	2 hrs.	3:00 am GMT	1:00 am GMT
2	192.135.183.0/24	2 hrs.	3:00 am GMT	1:00 am GMT
3	203.10.63.0/24	2 hrs.	3:00 am GMT	1:00 am GMT
4	198.32.7.0/24	2 hrs. [3]	1:00 am GMT	3:00 am GMT
R_0	195.80.224.0/24	2 hrs.	0:00 am GMT	2:00 am GMT
R_1	195.80.225.0/24	2 hrs.	0:00 am GMT	2:00 am GMT
R_2	195.80.226.0/24	2 hrs.	0:00 am GMT	2:00 am GMT
R_3	195.80.227.0/24	2 hrs.	0:00 am GMT	2:00 am GMT
R_4	195.80.228.0/24	2 hrs.	0:00 am GMT	2:00 am GMT
R_5	195.80.229.0/24	2 hrs.	0:00 am GMT	2:00 am GMT
R_6	195.80.230.0/24	2 hrs.	0:00 am GMT	2:00 am GMT
R_7	195.80.231.0/24	2 hrs.	0:00 am GMT	2:00 am GMT
R_8	195.80.232.0/24	2 hrs.	0:00 am GMT	2:00 am GMT

Table 4.3: BGP beacons: Schedule.

epoch seconds anyway. Therefore, only one parameter, the withdraw and announce schedule, may shift in spring or autumn, and the update timestamps have to be interpreted relative to the beacon time.

To unify the time bins for the beacon events, I modified the timestamps in the non-RIS beacon updates such that the first daily announcement is at 0:00 GMT etc. In this way, all timestamps can be treated in the same way for binning purposes for all 13 beacons. The modifications necessary to shift the first daily announcement to 0:00 GMT are the following (all times in GMT):

- Beacon 1's update timestamps until October 27, 2002 at 10:00[4] are incremented by 3 600 seconds (1 hour). Afterwards, they are incremented by 10 800 seconds (3 hours).
- Beacon 2's update timestamps do not need to be changed after November 16, 2002. Those before were filtered out completely because of clock problems.
- Beacon 3's update timestamps are incremented by 3 600 seconds before October 26, 16:00, after this time they are incremented by 7 200 seconds.
- Beacon 4's update timestamps are decremented by 3 600 seconds until November 21, 14:00. After this time, the beacon follows a different, ex-

[3] Contrary to the information on the BGP beacon info web site [3], this period was changed to an experimental schedule in November 2002.

[4] Shift from Pacific Daylight Savings Time to Pacific Standard Time in the U.S.

perimental schedule. As the echoes are very distinct with the different schedule, the updates from beacon 4 after November 21 were filtered out for this study.

I can not guarantee not to confuse the announce event of 4 o'clock with the one of 8 o'clock with those heuristic timestamp modifications. But up to now, there were no time-of-day effects to be observed in BGP, so the only important thing is that announce events and withdraw events are not mixed. And by examining several events in a row (duration, last update type) from different observation points, it is generally easy to decipher what kind of beacon event was observed.

4.2 Data sets

My characterization and classification work is based on raw external BGP update traces from Ripe [33], Routeviews [35] and Saargate, a local ISP [37]. The results I present are based on the following raw data sets: 4 months of BGP updates, namely the trace that started on October 1, 2002 and ended on January 31, 2003. I used dumps of peering sessions gathered at RIPE's Remote Route Collectors (RRC00 to RRC08) [34], as well as dumps from Routeviews [35] and from Saargate [37]. (For the Saargate collector, there was no data available to us until October 25, at 8:31 GMT.) This chapter is based on only the updates for of the 13 BGP beacon prefixes.

Because of clock problems of collectors or beacons, the data from Routeviews until November 9, 2002 (0:00 GMT) was excluded from this work, as well as the data for Andrew Partan's Beacon (called Beacon 4) later than November 21, 14:00 and David Meyer's Beacon (Beacon 2) before November 16.

Applying this filter, the total number of updates amounts to 2 350 428 updates that represent the convergence process for 17 778 beacon events.[5] In theory, these roughly 18 000 events can be seen by every peer at every collector. In reality, roughly 60 % of the peer-beacon events do not reflect in the update trace. In total, there is data from 90 BGP peers in the traces.

For this analysis, close to 125 000 update trace files (chunks of 15 minutes)

[5] Nine RIPE beacons as well as Beacons 1 and 3 are available from October to January, which amounts to 1 476 events per beacon, Beacon 2 is available for 924 events, Beacon 4 for 618.

were read and filtered for the beacon updates. These update trace files contain a total of 1.7 billion updates, i. e. 0.14 percent of the updates are to the BGP beacon prefixes.

Based on the preprocessed data from the four month period, with filters and modifications as described above, every update is assigned to one specific beacon event, of type either announcement or withdrawal. This is done using two-hour bins, e. g. every update with a timestamp between 2:00:00 and 3:59:59 is assigned to the 2 o'clock event etc. The underlying assumption for this binning is as follows: Instabilities in the beacon traffic only occur at the originating AS of the beacon prefix and only at the times specified by the beacon schedule. The characteristics of this data will be analyzed in the following sections.

4.3 Classification

As BGP and all its enhancements are designed by humans, the protocol itself is very well understood. But as mentioned in [4, 26], there is considerable doubt about the dynamics of the protocol in the global Internet, especially of BGP route-flap damping [41] and the recommended RIPE parameters for it [29].

It is known that BGP usually converges within 2 minutes [25], and that longer convergence times may occur, but it is not very well understood when and why the convergence is delayed. The community is eager to understand the dynamics and interactions of the protocol. The knowledge of the causes of BGP's dynamics could lead to improvement of the reliability of routing in the Internet. Additionally, understanding the interactions would enable us to enhance the convergence processes.

The idea of the BGP beacon analysis is to identify long convergence processes and to find the reasons for it. Long convergence processes imply a longer beacon duration[6], so at first, the beacon echoes will be separated according to their beacon duration.

It is easy to see that the beacon duration depends very much on the type of the beacon event. Announce events and withdraw events exhibit rather un-

[6]The beacon duration is the time from the original beacon update according to its schedule to the last update seen from a peer.

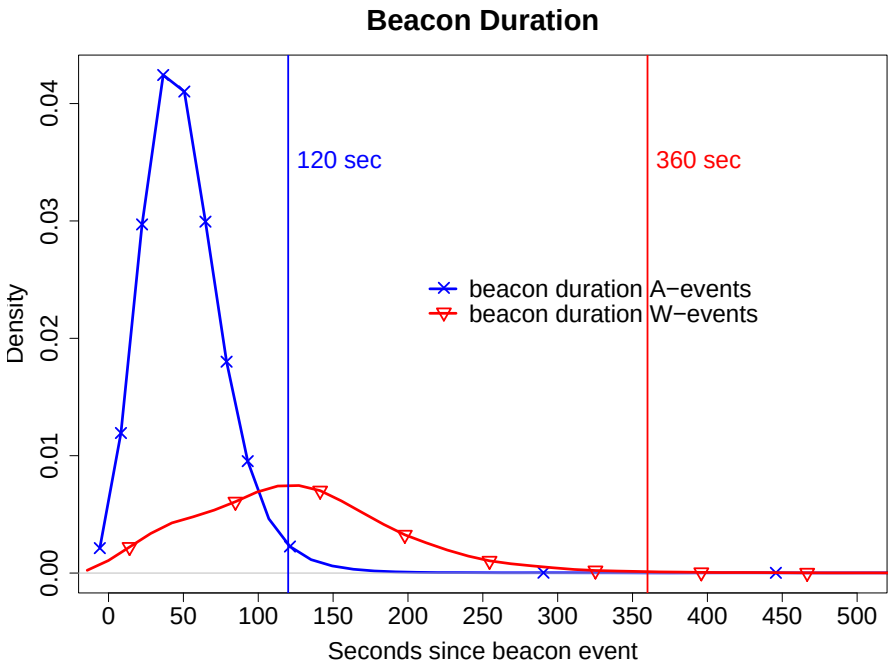

Figure 4.1: Beacon duration of all events.

equal beacon durations. Figure 4.1 shows the smoothed densities of the beacon durations of the announce events (in blue) versus the withdraw events (in red). The x axis plots the seconds since the beacon event, and it is cut off at 500 seconds after the beacon event for clarity. The red density may look as if there were values below 0, which is not possible by definition of the beacon duration. The reason for it is the smoothing of the density, which rounds sharp edges. (All densities in this thesis are smoothed, even if not mentioned explicitly.)

One can see from the blue density that for most of the announce events, the convergence process is finished after two minutes. The blue vertical line marks beacon duration 120 seconds. On the other hand, the convergence process for withdraw events (red density in Figure 4.1) is usually finished after six minutes (360 seconds, the value marked by the red vertical line). It is remarkable how different the events are in terms of beacon duration. While more than 92 % of the announce events have converged within 120 seconds after the beacon event, only 48.5 % of the withdraw events have reached a sta-

ble state at this time.

To further study the reasons of long convergence times, I wanted to separate the beacon events with longer convergence time from those with quick convergence time. Regarding only the beacon updates for October, the quantiles for 120 seconds beacon duration in announce events and 360 seconds in withdraw events were very similar, and I wanted to chose whole minutes for the separating value. So the values chosen for "quick" convergence are: **120 seconds** (2 minutes) for announce events and **360 seconds** (6 minutes) for withdraw events. The exact choice of the parameters is not critical for the characteristics of the classes. So far, each event ends up in either of two classes, one for quick convergence, and one for long convergence.

Last but not least, I expect from a successful convergence process that an instability in the shape of an announce (withdraw) event ends up in a stable announce (withdraw) state. The stable state is represented by the last update seen from a peer as there is no further update. When the stable state after an announce (withdraw) event is a withdrawal (announcement), it will be considered to be of the "wrong" type. There are beacon events with quick convergence (up to 2 or 6 minutes, respectively), but which end up in the wrong state. These beacon events are separated into an own class.

All in all, events are separated into three classes: the first class, called the **green class**, corresponds to the prevalent behaviour. The second one is the **red class**. It contains all events with long convergence times. The third is the **orange class**, which is a special case of the green class: the prefix converges quickly, but the last update is of the "wrong" type.

Note that not every beacon event is seen at every observation point. This does not only happen in failure cases, e. g. if the BGP speaker does not announce the prefix. It can also be due to aggregation or filter policies. Those invisible events are treated separately in Section 4.4.

The red class is the one of interest for studying BGP effects that are not fully understood. It is the class that contains all observed exceptions, e. g. due to failures or damping effects. However, it is important to consider the prevalent convergence characteristics for comparison. Therefore, the green class will be studied first. In addition, I discuss the implications and suggestions that can be drawn from the characteristics and special properties of

each class in this chapter. In this section, I will consider only the $\tilde{6}20\,000$ visible events. These are the basis for calculating percentages.

4.3.1 Green Events

As was shown in Figure 4.1, most of the announce events converge in less than two minutes, and most of the withdraw events converge in less than six minutes. Furthermore, the events mainly converge to the "right" state: The last update is a withdrawal for W-events or an announcement for A-events. As the green class is supposed to represent prevalent convergence behaviour, the following parameters are required for an event to belong to the green class:

- For an announce event to be classified as green, the event has to
  - converge in the first 120 seconds (2 minutes) and
  - the last update has to be an announcement.
- In case of a withdraw event, the event has to
  - converge in the first 360 seconds (6 minutes) and
  - the last update has to be of type withdrawal.

The events classified as green represent prevalent behaviour. 93.82 percent of the visible events (where at least one update appears) are classified as green.

As withdraw and announce events behave in quite distinct ways, I discuss their special characteristics in separate sections, one for announce and one for withdraw events. Overall, there are 617 299 visible events, which split almost evenly into A-events and W-events: 301 295 A-events and 316 004 W-events.

Green announce events

Of the 301 295 visible announce events, 272 605 (90.5 %) are classified as green, i. e. 90.5 % of the A-events lead to convergence within two minutes with an announcement as the last update. The median number of updates within green announce events is 1, but the maximum number of updates within one such event is 659 updates. Among the updates of green A-events, one can observe a maximum of 3 withdrawals. Only 1 615 green announce events carry one

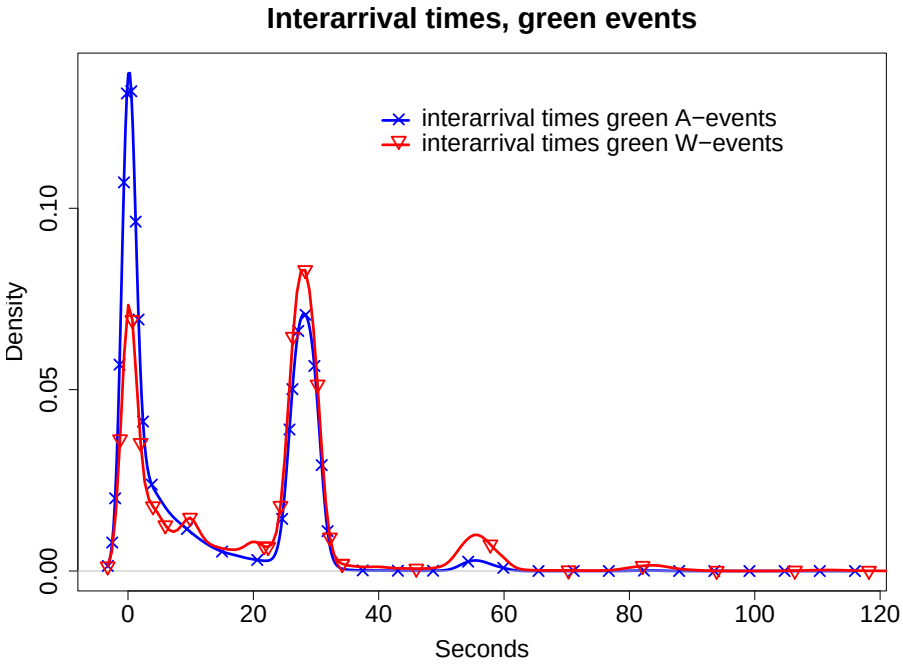

Figure 4.2: Interarrival times of green events.

or more withdrawals, that is less than 0.6 percent. 57 % of the green A-events consist of exactly one announcement.

The fact that a maximum number of 659 updates within one green announce event can be observed clearly shows that some routers do not abide by the MRAI timer of 30 seconds per peer-prefix pair. Indeed, the data even suggests that there are BGP implementations that either do not have an MRAI timer implementation or turn it off. This is surprising as the BGP specification contains the MRAI timer as a Must requirement.

Figure 4.2 shows the smoothed density of the interarrival times of both green A- and green W-events, the green A-events being drawn in blue and the green W-events in red. Since, by definition, green A-events have a beacon duration of up to 120 seconds, the interarrival times for these, are shown completely. The x axis is cut off at 120 seconds, so the density of the interarrival times of green W-events (red), is cut off. Note that I did not observe any significant peaks above 120 seconds.

Figure 4.3 on the next page shows two other density plots, the duration of the announce events (green) and the beacon duration (dark green) for the

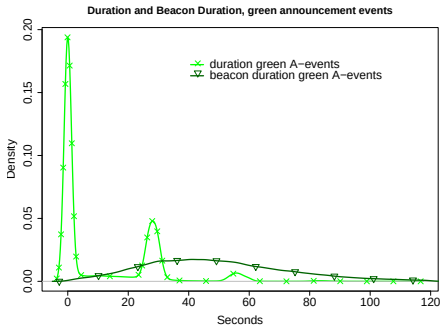

Figure 4.3: Duration and Beacon Duration: green A-event.

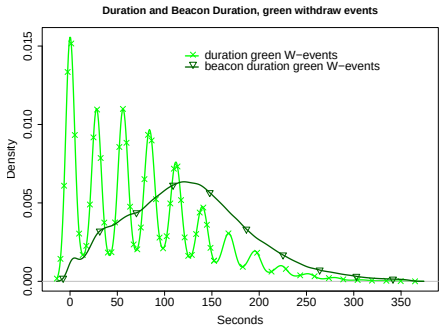

Figure 4.4: Duration and Beacon Duration: green W-event.

green class. The duration density as well as the interarrival times of the green A-events peak at multiples of the MRAI timer. This indicates that the MRAI timer is one of the primary components of the convergence time. Surprisingly, the beacon duration does not show any such peaks. This implies that the information is not delayed by 30 seconds at every hop, and that the total delay seen by a peer is evenly distributed over time.

Extreme values such as 659 updates for one prefix within 120 seconds from one peer should be avoidable. When peers in principle abide by the MRAI timer, the maximum number of updates seen in 2 minutes should lie below 10 updates. In fact, very large numbers of updates are rare (e. g. 24 instances with a number of updates above 200), therefore it seems okay to neglect them.

The main part of the A-events converge within 2 minutes. Considering the complexity of today's Internet and its distributed character, a convergence time of two minutes is quite acceptable.

Green withdraw events

Now let us examine the same statistics for green withdraw events. There are 316 004 visible withdraw events, of which 304 848 converged within 6 minutes with a withdrawal update and are thus classified as green events. While 90.5 % of the A-events are classified as green, I classified 96.5 % of the W-events as green. This may seem to point at a certain weakness of choosing our parameters at 120 and 360 seconds. But it is not primarily the choice of the parameters. Figure 4.5 on the next page shows the cumulative distribution function

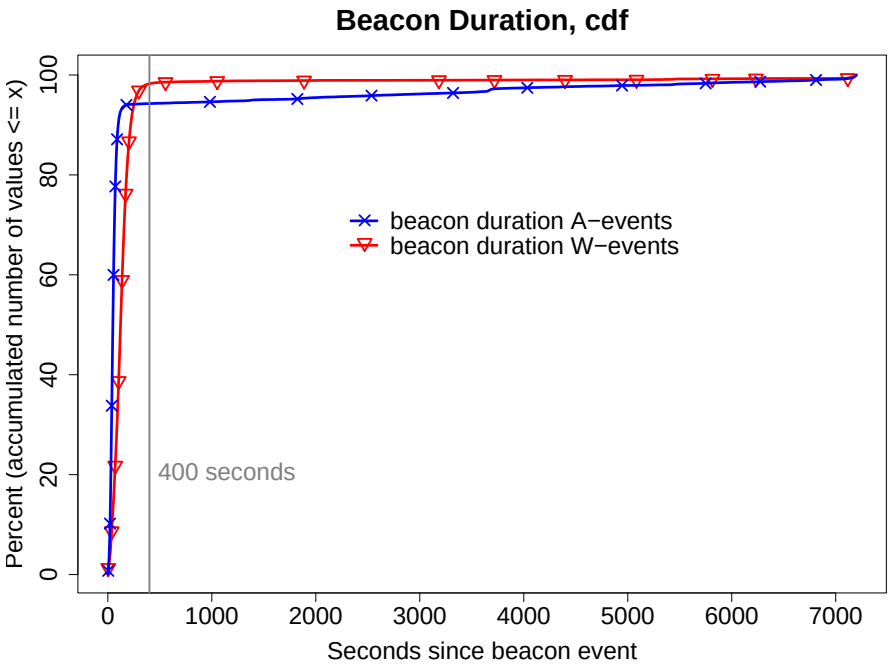

Figure 4.5: CDF: Beacon duration of all events.

(*cdf* in short) of the beacon duration of both A-events (in blue) and W-events (in red). The cdf for each beacon duration on the x axis in seconds shows the cumulated percentage of events that had a beacon duration of up to this value, relative to the total number of events. The two cdfs meet at beacon duration 252 seconds with a cdf percentage of 94.2 %. For smaller values, there are more A-events than W-events that converge within the specified limit, but for longer beacon durations, there are many more A-events than W-events with that beacon duration. This is something Figure 4.1 on page 30 did not reveal.

A slight rise of the blue cdf can be observed at about 3 500 seconds, but in total, the announce events with longer beacon duration are spread over the length of the time period. In contrast, the red cdf (W-events) rises sharply up to 400 seconds. A vertical grey line was included for this value. Not many values are larger than 400 seconds beacon duration for W-events. The values at which the sharp rises occur can be more easily gleaned from the densities shown in Figure 4.1. The characteristics of events with long beacon duration

are examined in Section 4.3.2.

Figure 4.5 shows that an announcement instability event will usually converge within two minutes, whereas the withdraw information needs double or triple this time. 90 % of the W-events have a beacon duration of up to only 220 seconds, which is "only" twice the convergence time of 90 % of the A-events. In general, distance vector protocols[7] show the behaviour that "good news travels fast, bad news travels slowly" [16]. This characterization confirms this for BGP, and the factor between the two modes in BGP is 2 to 3.

The median number of updates for these events is 3 (triple of the median of A-events), and the maximum number of updates within one event is 825 updates. The updates in the green W-events contain up to 22 withdrawals. 69 207 green withdraw events (22.7 %) carry no announcement, i. e. one fifth of the withdraw events that converge within 6 minutes do not use implicit withdrawals[8].

Of the total number of 1 410 131 updates in the 304 848 green W-events, 432 852 are withdrawals. This indicates that within quickly converging withdraw events, the withdrawals account for roughly one third of the updates (in the four months of beacon traffic 30.7 %). This is a larger percentage than in the overall beacon traffic: In the beacon updates, where 50 % of the initiating instability events are withdraw events, the withdrawals account for only 21.8 % of the beacon traffic.[9]

Figure 4.4 on page 34 plots the density of duration (green line) and beacon duration times (dark green) for the W-events of the green class. The MRAI timer is clearly recognizable as peaks in the duration density. This is even more pronounced than in Figure 4.3 for the green announce events.

Figure 4.2 plots the density of interarrival times of updates within green withdraw events in red. The density is cut off at 120 seconds so as to plot it with the density of the interarrival times of green announce events. There are only very few interarrival times greater than 120 seconds: 1.8 % of all interarrival time values are larger than 120 seconds.

The densities of interarrival times for both green events are alike. In other

[7] BGP more appropriately is called a path vector protocol because it propagates path information instead of cost information, but it belongs to the class of distance vector protocols.

[8] An announcement of an alternative route can implicitly withdraw the old route.

[9] From 2 350 428 beacon updates, 511 497 are withdrawals, i. e. 1 update in 5 is a withdrawal.

words, the interarrival times provide no good distinction for the event type. There are more interarrival times about 0 seconds in green A-events than in green W-events. Indeed, this is the most significant peak in the density for A-events. Due to the MRAI timer, there should be no interarrival times below approximately 25 seconds. Instead, both densities in Figure 4.2 show a significant peak for interarrival times of 0 seconds. Let us recall the green A-event that contained 659 updates. Such a large number of updates within 120 seconds implies short interarrival times. This indicates that there are router implementations that don't provide their BGP implementation with a default MRAI timer of 30 seconds. However, it is dubious if this implementation issue accounts for all 0 second interarrival times.

Summary of green events

Overall, most of the green events converge in a reasonable time. 6 minutes for withdraw events cannot be called a *very* quick convergence time, but there is no obvious way to enhance the convergence process for green events. Furthermore, the withdrawal of a prefix is bad news. Even if such information is known earlier, the prefix is still unreachable. Quick convergence thus would not help with regards to the underlying reachability problem.

The green class demonstrates clearly that there are 2 factors influencing the convergence time in the normal case of quick convergence:

- The MRAI timer
- and network or router delay.

The MRAI timer can be observed in the interarrival times of several updates that one peer sends in a row. If the number of updates per peer could be reduced, this would result in much lower convergence time: Each MRAI timer cycle saved at a peer lowers the convergence time by 30 seconds. It is also possible to change the value of the MRAI timer, but it is not clear how to define a globally optimal value [12]. Additionally, this may lead to a much larger number of updates per convergence process. Since the route-flap damping parameters currently deployed are adapted to the number of updates expected with MRAI timer of 30 seconds, this could result in much more damping of non-flapping routes.

Regarding the other factor, it would be a big step towards understanding routing in the Internet if causes and rules for the additional delay could be found. Link delays are usually only in the order of milliseconds, but delays seen in BGP are in the order of seconds. This may point to router processing overhead, protocol timers and IGP interactions as the possible reasons. One way to examine the impact of those influences is router testing, but this is beyond the scope of this thesis.

The green class identifies factors in quick BGP convergence. It remains to examine if different factors can be found for long convergence times in the red class.

4.3.2 Red Events

Red events contain all events with long convergence time. More precisely, red events are the events where the convergence process takes more than two minutes after an announce event, or where the convergence process takes more than six minutes after a withdraw event.

Since one beacon event consists of one update at the origin, one would naively expect the convergence process to take in the order of milliseconds as this is the delay observed in global data communications. The green class shows that convergence in BGP is in the order of seconds and minutes. To the newcomer, this is a surprise. One would thus expect few events to fall into the red class with even longer convergence times. Reasons for long convergence times can be failures of any kind, failure and repair at the beacon, or somewhere on the chosen best path, or at the collector.

It was shown that damping can be observed in beacon echoes [4], and damping would lead to red events. I thus expect damping effects to be observed in the update trace, but I also expect those to be of minor impact since session resets were not filtered from the update trace. Furthermore, route-flap damping is meant to damp only flapping routes. Therefore, one expects to observe damping effects in beacon traffic only as an exception from the rule.

To analyze the red events, I will present the basic statistical characteristics of this class, durations and beacon duration, interarrival times and number of updates. Those statistics do not provide us with an intuition of when the

updates show up within the event. Are the updates spread over the whole two-hour interval or do they appear in bursts? If they show up in bursts, we want to locate those bursts within the red events and to analyze the interarrival times between bursts.

28 820 red events were observed in four months. They account for 4.38 % of the events. 22 450 are red announce events, and 6 370 are red withdraw events. This difference is rather surprising: It is a known fact that path exploration in case of unreachable prefixes can lead to delays, "good news travels fast, bad news travels slowly" [16]. Path exploration can explain longer convergence times in withdraw events. It is probably one reason why observed beacon durations in W-events are mainly up to 6 minutes whereas those in A-events are up to 2 minutes. If the red events consisted mainly of failures, session resets and router reboots, they would distribute rather evenly over A- and W-events. Path exploration could possibly explain a large number of red W-events. Instead, there are more red A-events than red W-events. The high percentage of A-events in the red events indicates either a poorly chosen limit for the green events or a surprisingly large number of damping effects to be observed. However, the cdfs in Figure 4.5 show clearly that the choice of 2 minutes as the limit for quick convergence did not cause the large number of A-events to be classified as red. It will be thus my primary interest to identify route-flap damping and its impact on the red class.

The median number of updates within a red A-event is 3, the maximum is 1 575. This large number of updates again points to peers that do not abide by the MRAI timer: With interarrival times of 30 seconds, a maximum number of 240 updates can be expected in a two-hour interval. 15 774 (70.3 %) of the red A-events do not contain a withdrawal, 562 (8.8 %) of the red W-events do not contain an announcement. The median number of updates within a red W-event is 5, and the maximum is 241.

Figure 4.6 on the next page shows the smoothed density of the duration and beacon duration of the red announce events, and Figure 4.7 shows the same densities for the red withdraw events. In both plots, the duration density is shown in red and the beacon duration density in violet. The updates observed during the two-hour interval are treated as one update burst, isolated from all updates in the surrounding beacon events. Recall that the du-

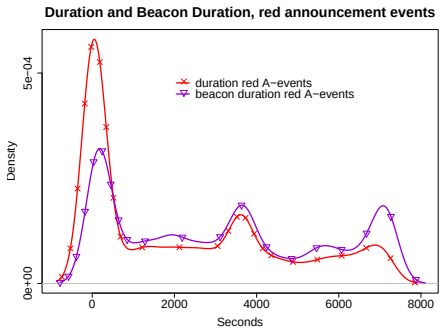

Figure 4.6: Duration and Beacon Duration: red A-events.

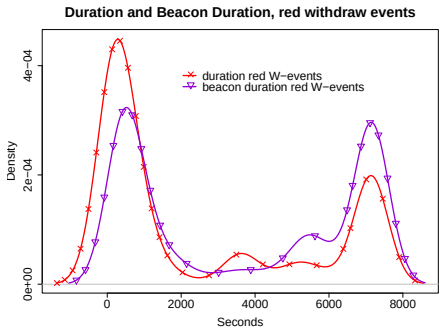

Figure 4.7: Duration and Beacon Duration: red W-events.

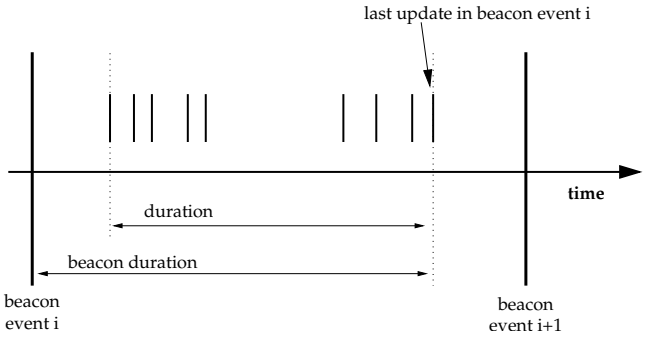

Figure 4.8: Illustration of duration and beacon duration in red events.

ration is the time from the first to the last update in the red event. The concept of duration and beacon duration in red events is illustrated in Figure 4.8.

In contrast to the green events (Figures 4.3 and 4.4 on page 34), duration and beacon duration of the red events have similar distributions. The resemblance is more pronounced for announce events and is in part due to the definition of the duration. Updates will be grouped into smaller update bursts later in this section.

The density for the red A-events shows a peak for small (beacon) durations below 500 seconds. This indicates that a significant portion of the red A-events may have failed the requirements for green events only by some small measure, convergence was delayed, but only in the order of minutes. This will be investigated further in Section 4.3.2 on page 49.

Figure 4.9 on the next page plots the density of the interarrival times of red events (separately for A- and W-events). Only values below 200 seconds

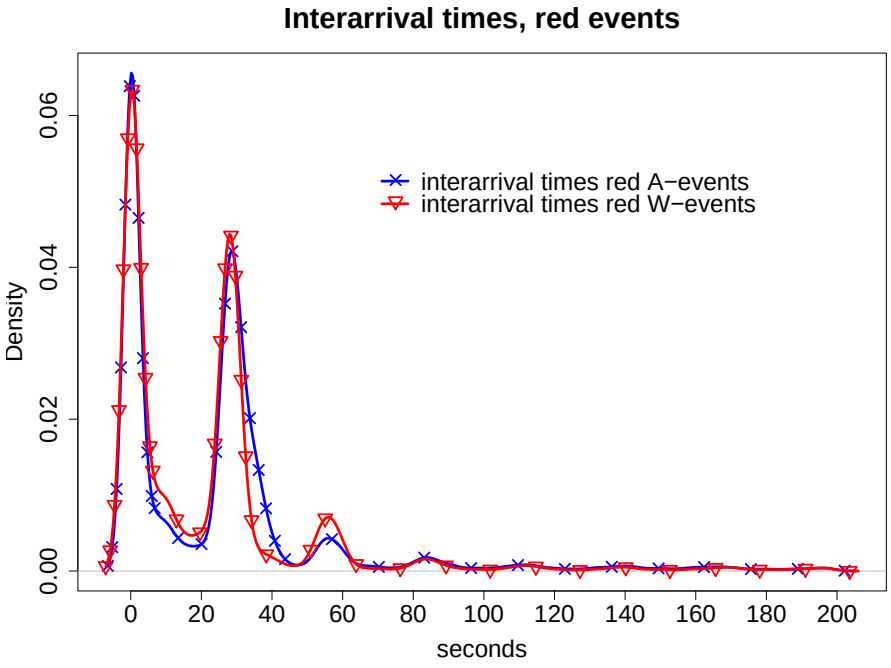

Figure 4.9: Interarrival times: red events.

are shown to reveal the details. This corresponds to 80 % (A-events) and 90 % (W-events) of the data, respectively. The impact of the MRAI timer is clearly visible with peaks of multiples of 30 seconds in both densities. There is no visible difference in the densities of interarrival times of red W-events and red A-events. The densities of interarrival times of green events in Figure 4.2 on page 33 did not show the same conformity: The density of green A-events was dominated by the peak around 0 seconds, the density of green W-events by the peak around 30 seconds.

The maximum interarrival time for red announce events is 7 197 seconds, and for red withdraw events 7 195 seconds. The median interarrival time of red A-events is 28, that of red W-events is 26. Both median values represent jittered MRAI timer values. The mean interarrival time is dominated by the large interarrival times. For red A-events it is 484.1, and for red W-events it is 365.5. Larger interarrival times are thus more common for A-events than for W-events.

The basic statistics presented here do not reveal much about the nature of red events. It is not possible to identify a reason for the large delays. There-

fore, I will in the following analyze the red class from another aspect.

Correlation with the number of updates in total

One may speculate that the convergence time depends on how much BGP traffic a router has to process. In a first attempt to find a reason for red events, I looked for a correlation of the fact that a red event is observed with the total number of BGP updates at a peer. Since we observe the updates on one peering session, it is possible to correlate the number of updates seen in different intervals with the colour of the classified beacon event in the same time interval on a per-peer basis. I compared the density of the number of updates seen in green events with the same density for red events. For both 15-minute as well as two-hour time bins, there is no difference to be seen in the two densities. With one exception: the density for the red events shows a peak at around 120 000 updates per time bin. This is the signature of a local session reset on the peer.

This implies that session resets may cause a beacon event to be classified as red. This is no surprise: After a session reset, the whole routing table has to be exchanged between the peers. This includes the beacon prefixes. If the session is reset after the first two (or six) minutes after the beacon event, this will of course lead to a red event. To get a feeling of the frequency of session resets, I consider the observed numbers: Session resets of immediate BGP peers appear in about 0.3 percent of the 15 minute bins on the RRC00 collector (12 peers) in October.

Session resets may cause an event to become a red event. But this is by far not the only possible reason[10], and besides it is feasible to filter out session resets for immediate BGP peers, but not for all session resets and router reboots of all peers on the AS path. Therefore, I did not filter out any session resets, as I could catch only an unknown portion of them. Session resets thus build up a part of the red events.

[10]The collectors have keepalives disabled. Therefore, the peering session recorded are not more susceptible to session resets than other usual peering sessions.

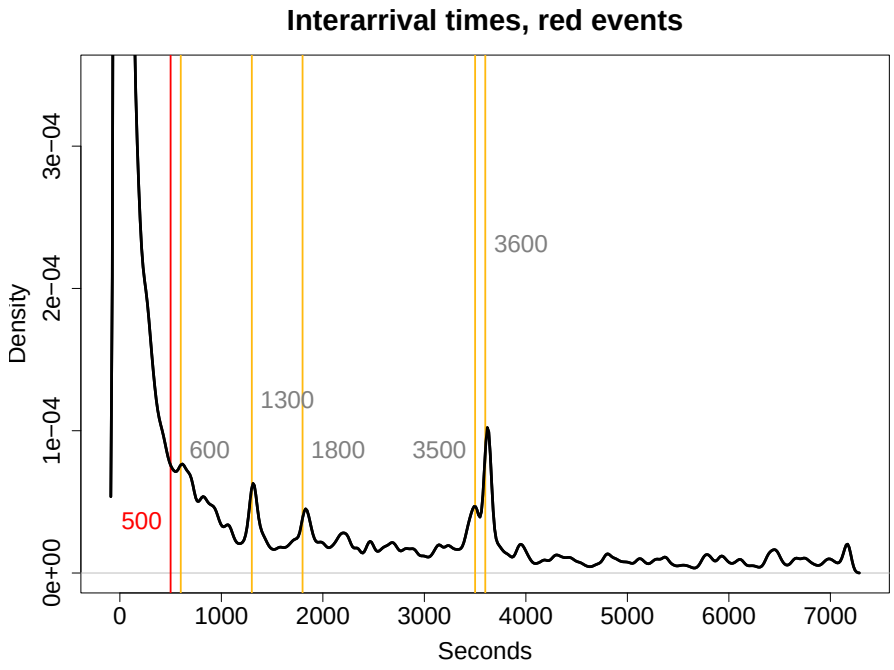

Figure 4.10: Interarrival times in red events.

Colour definition for bursts

The analysis so far still does not reveal any new information about causes for red events. Up to now, all updates in one event were considered as one entity, one large burst. This method does not reveal if updates show up in bursts or if they are spread over the two-hour interval. For the following analysis, updates in red events are grouped into "small update bursts". By grouping updates into small bursts, it is possible to discern details in the arrival process.

One first has to fix a timeout value by which the updates will be grouped together. There are indications for route-flap damping in red events, and I want to analyze this, so the timeout parameter should not group together the two small update bursts that result from damping. The minimum damping parameter that was recommended from RIPE [29] is 600 seconds. The timeout value should thus remain below 600 seconds.

In order to choose a value, I consider the interarrival times of updates within red events. Figure 4.9 on page 41 does not show enough detail, so Figure 4.10 shows the density of interarrival times within red events, zoomed

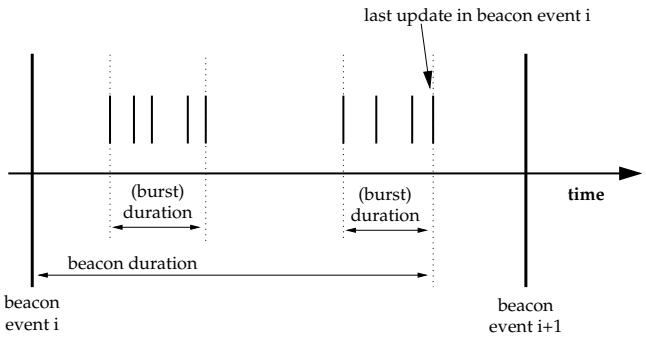

Figure 4.11: Several update bursts in one beacon event.

in on the y axis. The density has peaks at 600, 1 800 and 3 600 seconds (see shaded vertical lines marking these values), which represent minimum or maximum recommendations by RIPE [29] for route-flap damping of /24-networks.[11] The interarrival times thus reflect effects of route-flap damping. Furthermore, a peak is observed at 1 300 seconds. The RIPE recommendation does not directly provide us with an explanation for this peak.

Most updates have an interarrival time of up to 500 seconds. This value is marked by the red vertical line. Additionally, 500 seconds is below the minimum recommendation of RIPE [29] for all networks, therefore I chose 500 seconds as the timeout value to group the updates within red events into bursts. An illustration of this procedure can be seen in Figure 4.11. The updates in this event are grouped into two separate update bursts. The beacon duration still refers to the time difference between the beacon event and the last update burst, but several (burst) durations may be measured within one event. By grouping updates into bursts, each red event contains a list of bursts. To distinguish between different kinds of update bursts, I classified the bursts in a similar manner as the events themselves: by the burst duration as well as by the type of the last update in the burst that represents the momentary stable state.

A **burst** is assigned the colour **green** if

[11] RIPE recommends a value for max=min outage of 60 minutes for /24 or longer prefixes, and other values for shorter prefixes. If a specific damping implementation does not allow configuration of prefix-dependent parameters, they recommend to use the least aggressive set: max outage 30 minutes, min outage 10 minutes.

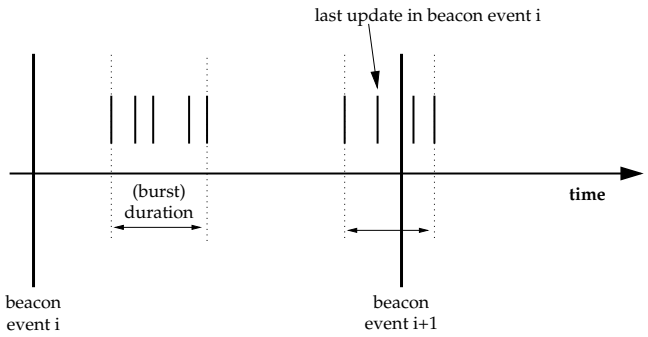

Figure 4.12: Illustration of a crossing burst.

- it is inside of an announce event, and the burst has
  - a duration of at most 120 seconds (2 minutes) and
  - the last update is an announcement, or
- it is inside of a withdraw event, and the burst has
  - a duration of at most 360 seconds (6 minutes) and
  - the last update is a withdrawal.

With the help of the illustration in Figure 4.11, it becomes obvious that the duration of the first burst in a beacon event is less than or equal to the beacon duration. That is to say that, in green events, the duration of the bursts (with timeout 500 seconds) is usually[12] less than 120 seconds. To avoid too much confusion, I used the same parameters, 120 seconds and 360 seconds, for the green bursts as for the beacon duration in green events.

A **burst** is considered **orange**, if it obeys the timing limits of the green bursts but ends with the "wrong" update. These bursts would be classified green if I ignored the condition on the last update.

All other bursts are assigned the colour **red**.

As the colour for a burst is assigned only contingent upon the event type, there is no obvious assignment for a burst that starts in one event and ends in another event. Such a *crossing burst* is illustrated in Figure 4.12. Updates out of two consecutive beacon events are grouped together into one update burst because the interarrival times between the consecutive updates are below 500

[12] An exception to this rule are the crossing bursts that are explained three paragraphs below.

seconds. The first update in this burst belongs to beacon event i, but the last update belongs to event $i+1$. It is possible for a burst to span over more than two events. Since the crossing bursts do not fit into the classification, I ignored bursts that crossed event borders. Fortunately, these account for only 0.4 percent of all bursts. More specifically, all updates are grouped into update bursts using a timeout of 500 seconds, crossing bursts are deleted, the remaining bursts are assigned to their respective two-hour beacon event and classified accordingly. Ignoring the crossing bursts removes 2 151 of the 28 820 red events. However, all bursts will be considered again for another analysis in Chapter 5.

After assigning colours to the bursts, every red event consists of a list of coloured bursts, which I call the **(burst) history**. After the deletion of the crossing bursts, all green events consist of one green burst and all orange events consist of one orange burst. A red event can consist of one green burst if this burst does not start at the beginning of the beacon event. The colour of the event depends on the beacon duration, whereas the colour of the burst depends on the burst duration.

Burst history in red events

I consider 26 669 burst histories (93.7 % of the original red events) in this section. Table 4.4 on the next page summarizes the percentage and total number of the top-ten burst histories of red events. 27.8 % of the burst histories consist of one green burst, 13.0 % of one orange burst, and 8.5 % of one red burst. Together, almost half of the events considered here (49.3 %) consist of only one burst. Below, I will analyze red events consisting of exactly one burst in more detail.

The fact that half of the red events with one burst, or 27.9 % of all red events, contain one green burst, is rather intriguing. Red events with two green bursts may be caused by route-flap damping, as well as red events with the history orange-green. But one green burst seems to exclude a possible impact of route-flap damping. To give insight into those events, I will analyze the red events with one green burst, separately.

Since one important aspect for the analysis of route-flap damping is the interarrival time between small update bursts, I will examine red events con-

Percentage	Total number	History
27.80	7413	green
22.96	6124	green-green
12.97	3460	orange
8.52	2273	red
6.72	1793	green-orange
4.88	1301	green-red
2.99	797	orange-orange
2.92	780	green-green-green
2.70	721	orange-green
0.99	264	red-green

Table 4.4: Top ten of the burst histories in red events.

sisting of two or more green bursts more closely. They represent 26.4 % of all burst histories. (Note that not all of these are listed in Table 4.4, as they are not all in the top ten.) Burst histories list up to 7 green bursts in a row within one red event.

Red events consisting of only one burst

The red events consisting of one burst (see Table 4.4) sum up to 13 146 and can be divided into three groups: The ones consisting of a green (56.4 %) burst, those consisting of an orange (26.3 %) one and thos with a red burst (17.3 %). Figure 4.13 on the next page plots the density of the beacon duration of these single-burst red events, according to their classification. The green density represents the green bursts, the densities of red and orange bursts are coloured in their respective colour.

The densities show that most green bursts as well as most red bursts appear at the beginning of the two-hour time period of the beacon event. The green bursts have lower beacon durations than the red ones, i. e. the last update of the burst appears earlier. This is not surprising, since the red bursts also have a longer duration than the green ones by definition: The burst duration of a red burst in a W-event, e. g., is at least 361 seconds.

Almost all green bursts are finished 500 seconds after the beacon event, see the green vertical line. This indicates that some red events are similar to green events: A short update burst appears at the beginning of the two-hour beacon event. However, in red events, the delay is for some reason larger. The

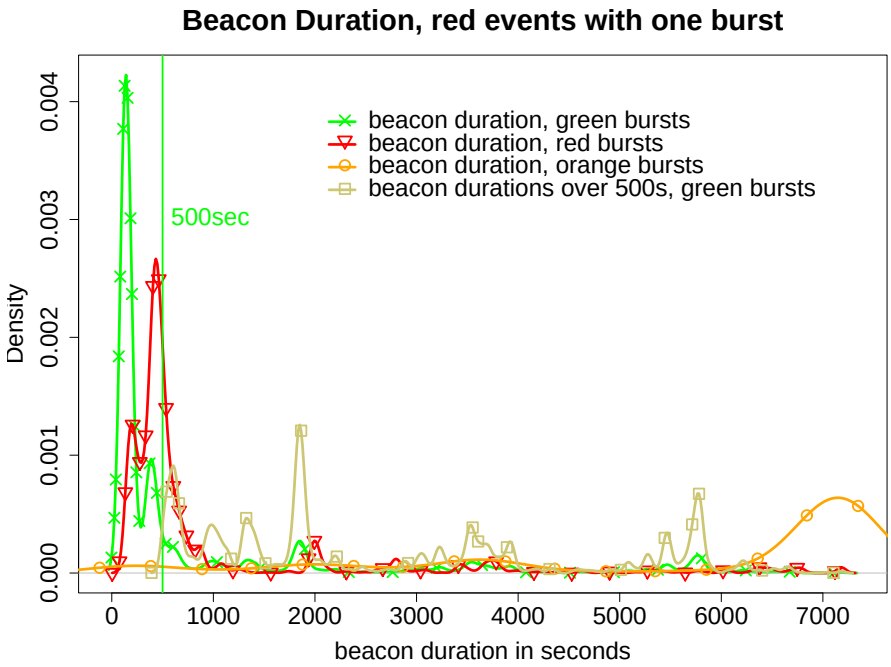

Figure 4.13: Beacon duration of single-burst red events.

red events that are similar to green events represent about 2 % of all beacon events. However, some of the green bursts are spread over the whole time frame and do not appear at the beginning of the beacon event. To show such details, the greyish green density in Figure 4.13 plots the density of the beacon duration for red events with one green burst and beacon duration of more than 500 seconds. Bursts seem to show up in waves with a periodicity of about 30 minutes. This may be an effect of route-flap damping. The green bursts will be analyzed more thoroughly in the next subsection.

The red density is very similar to the green one, except for the additional delay. An event is classified as red only if the beacon duration is above 120 (360) seconds. Additionally, for a burst in this red event to be classified as red, the duration of the burst needs to be above 120 (360) seconds. These two selection criteria add up to a beacon duration significantly higher than the one for red events with a green burst.

The orange density has its only peak above 7 000 seconds. Since the maximum beacon duration by definition is 7 199, this is rather suspicious. One

explanation is that these bursts belong to the next event. This can happen if a beacon fires too early or a collector records updates with wrong timestamps. Further analysis shows that RRC01 [34] is the main contributor to this peak in the density. RRC01 contributes 2 241 single orange bursts with beacon duration above 7 000 seconds. Note that there are only a total of 3 460 single-orange-burst red events. Probably, the clock on RRC01 was slightly off from November 23 to December 9 2002.

Some beacons are more involved than others: From the identified 2 241 orange bursts contributed by RRC01, 520 bursts (23.2 %) refer to beacon R_5. Beacons 2, R_0 and R_3 each contribute between 300 and 400 bursts (about 15 %). Presumably, these beacons are closer in terms of topology to the peers of RRC01. Note that it is possible to observe a green burst at the beginning of a beacon event together with an orange bursts near the end of this event. In this case, the clock problems of RRC01 may cause some of the red events with history green-orange. (Green-orange histories represent 6.7 % of the histories in red events.) Clock problems can cause green events to be classified as red, or vice versa. The beacon project heavily relies on synchronization of clocks on both beacons and collectors. Lack of synchronization can lead to the problems discussed above.

Summarising our results from the subsection: The single-burst red events represent 49.3 % of the red events. This corresponds to a fraction of 2.15 % of all beacon events. They contribute a significant fraction of the red events. A good part of them just show a larger delay in reaching a peer: 53.4 % of the single-burst red events have a beacon duration of up to 500 seconds. Another part is caused by faulty timestamp generation. Overall, many of the single-burst red events can be explained and may be considered harmless.

Red events consisting of only one green burst

This section analyzes in more detail a subclass of single-burst red events: the red events that consist of exactly one green burst. The single green bursts account for 56.4 % of the single-burst red events. These 7 413 events correspond to 27.9 % of the red events and 1.2 % of all beacon events. Why is it that green bursts are seen so late in the beacon event that the events are classified as red?

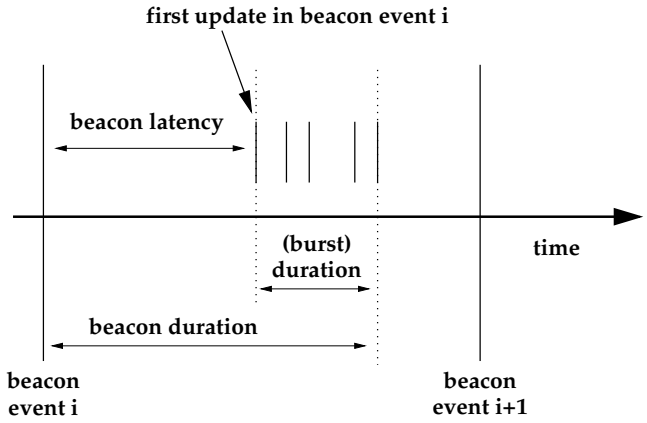

Figure 4.14: Illustration of beacon latency.

Figure 4.15 on the next page plots the duration of the green bursts against the beacon duration. In fact, it also shows the time to the first update in the beacon event, which I will call the *beacon latency*: It is the time from the beacon event to the first update, see Figure 4.14. It can be computed from the burst duration and the beacon duration, as it is the time difference between beacon duration and burst duration.

Green bursts in red events can be found both in announce events and in withdraw events. Green bursts in red announce events have a duration of up to two minutes, and their beacon duration is larger than two minutes. For green bursts in red withdraw events, the burst duration has to be less than six minutes, but the beacon duration is larger than six minutes. Both kinds of green bursts are represented in Figure 4.15. The x value represents the burst duration, and the y value shows the beacon duration in seconds.

The three horizontal green lines at the bottom of Figure 4.15 mark beacon durations of 120, 240, 360 and 480 seconds. It can be observed that many bursts have a beacon duration between 121 and 240 seconds and a duration of up to 120 seconds. The points in this rectangle represent green bursts in announce events[13]. This rectangle contains 58.0 % of the 7 478 single-green-burst red events. They can be interpreted in the following way: If the beacon latency had been lower, these announce events would have been classified as

[13] Withdraw events with beacon duration of up to 360 seconds are classified as either green or orange.

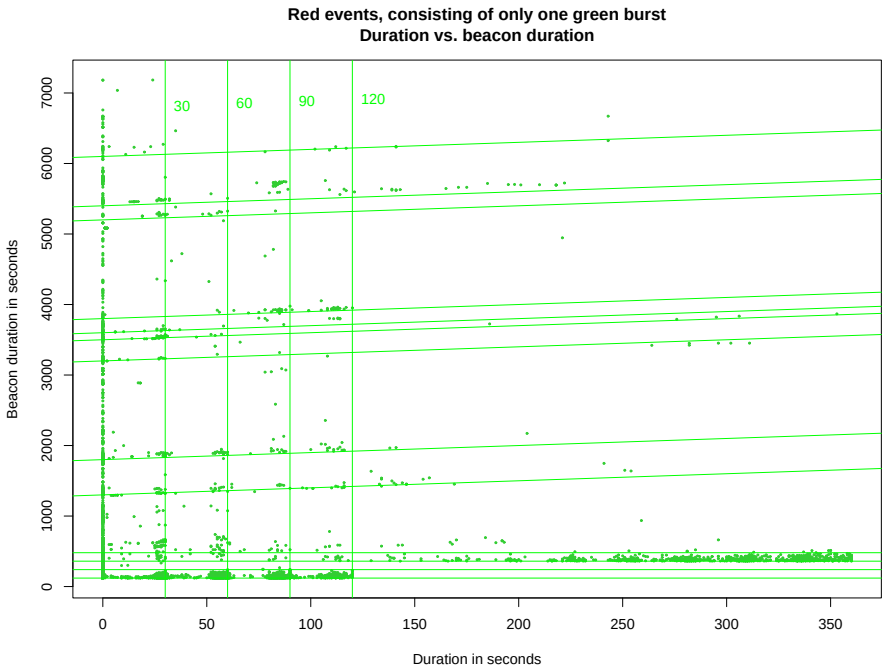

Figure 4.15: Solitary green bursts in red events.

green. Instead, they were classified as red, because the first update reached the peer too late. The reasons for high beacon latency in A-events will be analyzed below.

Notice that all bursts with duration over 120 seconds have a beacon duration over 360 seconds. This is because green bursts in announce events have to have a duration of up to 120 seconds, but red withdraw events have a beacon duration greater than 360 seconds. Therefore, all green bursts with beacon duration between 361 and 480 seconds and with duration between 121 and 360 seconds are green bursts in withdraw events. They failed to become green withdraw events by up to 120 seconds because of a high beacon latency, i. e. the bursts started too late in the two-hour interval. The points in this rectangle, duration of 121 to 360 seconds and beacon duration up to 480 seconds, represent 11.7 % of the single-green-burst red events.

Taken together, red events that failed to be classified green by just 120 seconds account for 17.9 % of the red events: 58.0 % and 11.7 % of the single-green-burst red events. This is almost 1 % (0.8 %) of all events. Those red

events would have been part of the green class if I had chosen slightly different parameters for the green class. Therefore, this part of the red class can surely be considered harmless.

Another notable oddity in Figure 4.15 are the green bursts with burst duration of 0 seconds: 32.9% of the considered bursts have duration 0 and therefore, supposedly, consist of exactly one update. Their beacon durations covers the whole time period, but one third of them arrive between 121 and 240 seconds after the beacon event. This third represents 11.5 % of the lone green bursts in red events.

Other clusters of bursts can be seen at multiples of the MRAI timer of 30 seconds. This is why I added the vertical lines at 30, 60, 90 and 120 seconds to Figure 4.15. The y values of the clusters show their beacon latency: A burst with duration 100 seconds and beacon duration 1 300 seconds will be represented with a point at y value 1 400. The vertical lines above 1 000 seconds beacon duration have a gradient such that they represent beacon latencies of 1 300, 1 800, 3 200, 3 500, 3 600, 3 800, 5 200, 5 400 and 6 100 seconds. Most clusters observed in Figure 4.15 seem to have one of these beacon latencies. This can only be explained with route-flap damping. The values for the beacon latency cannot be explained by the RIPE recommendation only. In fact, current implementations for route-flap damping do not seem to insert fixed damping intervals. Instead, the damping interval depends on a penalty value that is computed from the number of updates received and decays with time. Additionally, multiples of the MRAI timer may be added to the damping period. This results in a great variety of beacon latency values.

In order to observe route-flap damping, one may assume that one needs two bursts rather than a single green one. Damping effects might not be expected to be visible in single-green-burst red events. Apparently, the first set of updates are not able to get through to the observed peer. I can offer one presumed explanation: Let us assume that we observe the two-hour interval of an announce event. At the beginning of this interval, the beacon prefix is not reachable as a result of a previous withdraw event. Right before the first update on behalf of the beacon prefix, the observed peer sends some other BGP update to its neighbour. Since the MRAI timer is implemented on a per-peer basis, this MRAI timer is started at this time. During the current MRAI

interval, the router receives several updates for the beacon prefix. But due to the MRAI timer, it will not send an update to the neighbour. If the number of beacon updates is sufficient to activate route-flap damping, the route to the prefix will be damped, i. e. withdrawn internally. Note that this requires a significant number of updates within one MRAI interval[14], 30 seconds. Since this newly resolved unreachability is actually the same stable state as before – the beacon prefix was unreachable 30 seconds ago –, no update for the beacon prefix needs to be sent after the expiration of the MRAI timer. In this way, the MRAI timer together with route-flap damping can cause a router to swallow the first update burst. Only after the release of the damping, *one* update burst is observed. Its beacon latency is influenced by route-flap damping. This explains the reasons for green bursts with very high beacon duration.

However, route-flap damping is not responsible for beacon durations between 121 and 240 seconds. Recall that those green bursts represent 58.0 % of the single-green-burst red events. What causes a green burst to be delayed up to two minutes? Presumably, the length of the AS path to a beacon could influence this value. In the worst case, every AS that propagates a BGP update delays it by the order of the MRAI timer. To analyze the reasons behind long beacon durations, I compare the AS path length of solitary green bursts in red A-events with beacon duration up to 240 seconds with the AS path length of green bursts of green A-events. I consider only A-events because I compare the AS path length of the stable state for green bursts with different beacon duration. In green bursts in W-events, the stable state corresponds to a withdrawal, and withdrawals do not have an AS path.

I consider only a subpart of the update trace used in this chapter: Only the beacon updates from the 12 peers of the collector RRC00. I selected 3 sets of green bursts: green bursts with beacon duration up to 60 seconds, those with beacon duration between 61 and 120 seconds and green bursts that end between 121 and 240 seconds after the beacon event. Note that the last set corresponds to green bursts in red events. The former two sets are subsets of the green bursts. Figure 4.16 on the next page plots the cdf for the AS path length for each of the three burst sets. The AS path length counts the number of unique AS numbers on the AS path, i. e. AS path length 5 implies an AS path

[14] About 6 updates in a short time period can trigger route-flap damping[4].

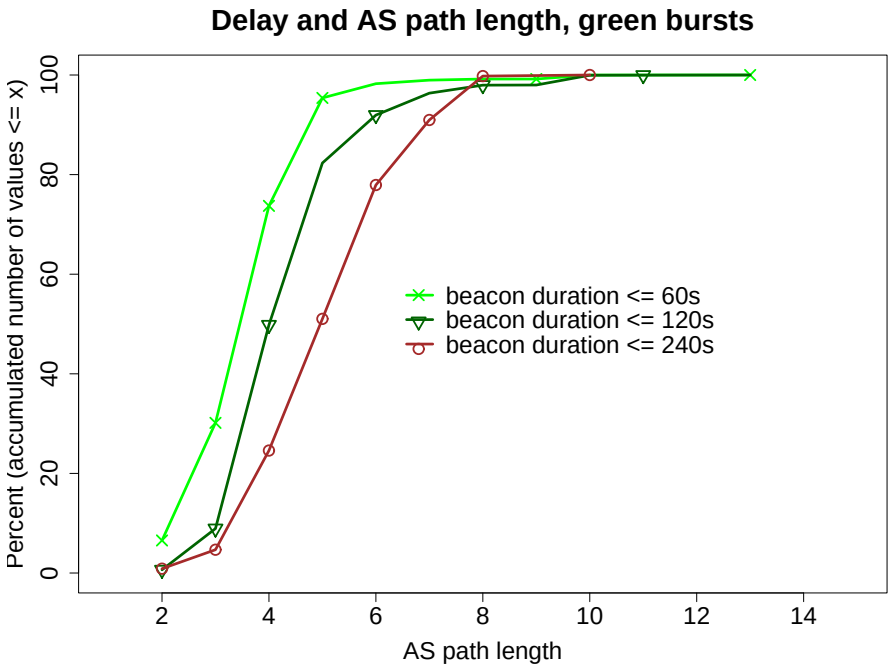

Figure 4.16: AS path length in earlier and later green bursts.

of 5 unique ASNs. The green line represents the cdf of the AS path lengths of the green bursts that lead to a beacon duration of up to 60 seconds. In other words, these are green A-events that converge within 1 minute. 70 % of the quickly-converging green A-events end on an AS path of length 4 or shorter. The dark green line shows the cdf for green bursts that end between 61 and 120 seconds after the beacon event. It thus shows the AS path lengths for green A-events that took somewhat longer to converge. Only 50 % of those bursts converge on an AS path of length up to 4. The brown line corresponds to data from green bursts that have so much delay that the beacon event is classified as red: The beacon duration here is between 121 and 240 seconds. Only 25 % of those bursts have an AS path length below 5.

It is impressive how much the observed delay in terms of beacon duration is influenced by the AS path length. One component of the beacon duration is the beacon latency. It is possible to create a similar plot for the same data set, separated according to beacon latency. The same trend is observable, but the differences are not as pronounced as in Figure 4.16. The figure

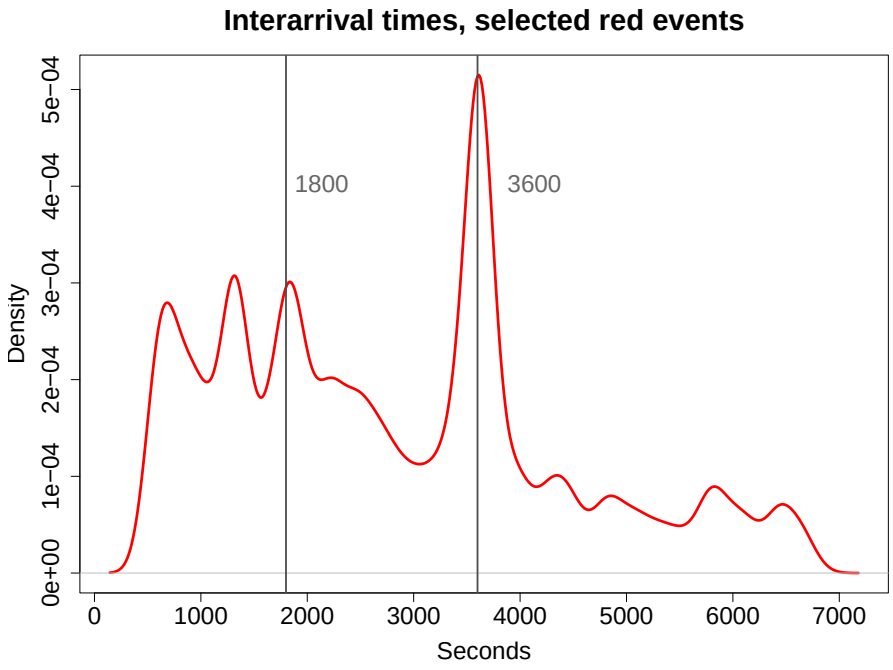

Figure 4.17: Interarrival times between green bursts in red events.

clearly demonstrates the correlation between convergence delay and the AS path length, i. e. with the number of ASes the BGP data has to cross on its way from originating AS to observed AS. In what ways this depends on other aspects of the topology still needs to be clarified.

Red events consisting of several green bursts

The next interesting question arises regarding the interarrival times between bursts. This subsection considers red events that contain several bursts, or more specifically the special case that a red event contains only green bursts. The hope is that in this special case, the influence of route-flap damping can be separated from other aspects. The data set contains the red events consisting of more than one green burst. They represent 26.3 % of the red events. By considering histories with more than one green burst, one can compute the interarrival times between bursts. I expect that those interarrival times show effects of route-flap damping since these effects were observed even in the single-burst red events.

Figure 4.17 shows a smoothed density plot of the interarrival times between green bursts. The two vertical lines, at 1 800 and at 3 600 seconds (half an hour and an hour) mark two peaks of probable damping values (parameters recommended by RIPE [29]). The other two peaks of the density, 700 and 1 300, do not correspond to values that are necessarily used for damping. But if the AS path is reasonably long, the delay may consist of one damping period of either 600 seconds (a RIPE recommended parameter) or twice this amount, 1 200 seconds plus several times the value of the MRAI timer. This can sum up to a total delay of roughly 700 or 1 300 seconds. An alternative, but rather unlikely explanation for a 1 300 second interarrival time is that two peers successively invoke damping. In this case, the release of the damping at the first peer may trigger enough updates to cause damping at the second peer.

Figure 4.17 thus affirms that red events in part are caused by ill-applied route-flap damping. I call it *ill-applied* because the damping does not suppress a flapping route. Instead, the convergence process of a prefix is delayed. However, route-flap damping does not seem to explain all observed interarrival times. So unfortunately, it is as yet impossible to unequivocally separate damping effects from other influences.

One of the reasons why damping leaves no clear trace is the implementation of route-flap damping in common routers. It is implemented in a progressive fashion, with penalties that sum up linearly and decay exponentially. The damping interval depends on the number of updates a peer receives and on their interarrival times. Additionally, an unknown number of waiting periods due to the MRAI timer as well as network and processing delay will result in an unidentifiable mixture. Presumably, route-flap damping will be easier to identify in the interarrival times if the damping peer is close to the collector.

So why not look at the data on a per-peer basis? Similarly to filtering out session resets from the update trace, looking at interarrival times on a per-peer basis can only identify damping parameters of immediate peers. But as soon as the damping AS is some AS hops away, the damping intervals cannot be identified as easily any more. A methodology to identify the damping policy of an AS will be introduced in Section 5.3.

Summary of red events

A rather large fraction of the red events represent bursts that have been delayed. If they had not been delayed, they would have been green events. Another clear cause of red events is route-flap damping which fuels the apprehension that damping should be considered harmful as suggested by R. Bush et al. [4]. At least in the way it is used today, it all too often –wrongly– leads to delayed convergence. This is shown by my study and is consistent with the findings of Z. Mao et al. [26].

It is not clear, however, if the routers that cause these effects abide by the RIPE recommendation [29]. If they do, it is surprising that the number of updates in one update burst becomes as large that a peer believes to receive four flaps although the original event is a single update. Some routers do not use the MRAI timer of 30 seconds. This causes a larger number of updates to be sent in one convergence process. If, additionally, the number of alternative paths to a prefix is large, this can result in many successive updates with different best path [24]. It seems worthwhile to improve the parameters for route-flap damping to account for this possibility, since route-flap damping contributes significantly to the long convergence times in BGP.

4.3.3 Orange events

Orange events can be considered as green events with the wrong outcome: Their timing is green, but the last update is of the "wrong" type. If we know that a beacon was announced at its specified time, and that the network converged quickly (below two minutes), but the last update seen at a peer is a withdrawal, then it would be good to understand what happened. Therefore, this special case is captured in a class of its own, the orange class. Orange events account for 1.8 % of all beacon events.

There are 11 026 orange events in the beacon trace, of which 6 240 are orange A-events and 4 786 are orange W-events. More A-events than W-events with short convergence time lead to a wrong stable state. Not all beacons contribute equally to the orange events: Beacon R_3 contributes the biggest slice. It is responsible for more than one third, more than 4 500 of the orange events. The next beacon in terms of contribution is Beacon 3 with 2 000

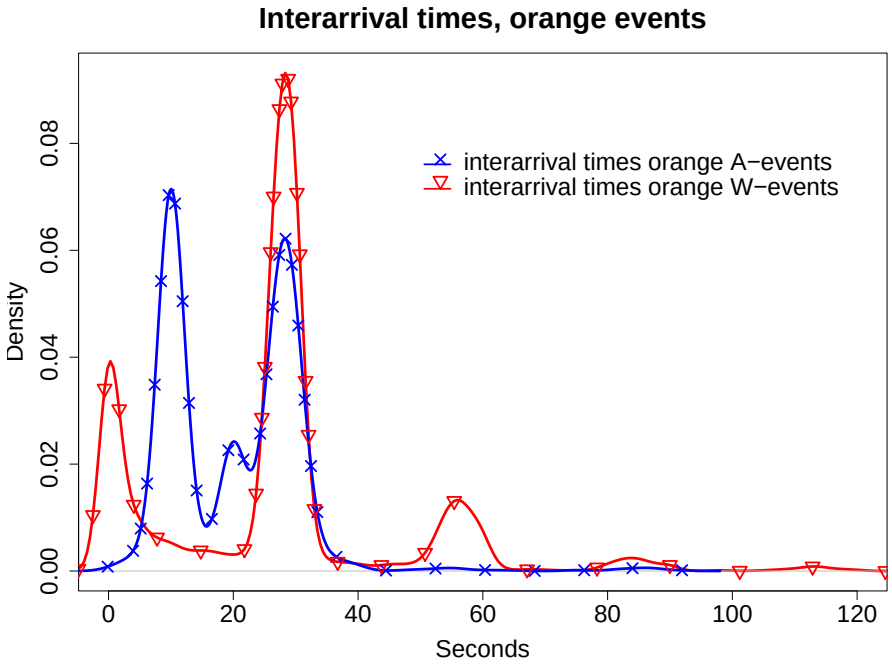

Figure 4.18: Interarrival times of orange events.

orange events. Beacon R_5 adds more than 1 100 events. Each of the other beacons add less than 1 000 orange events.

What could be the reason for orange A-events? Resulting from the previous W-event, the beacon prefix is unreachable at the beginning of the beacon event. Shortly after the A-event, the observed peer receives BGP updates for the beacon prefix, but this results in an update burst that ends in a withdrawal. This update burst may be explained by route-flap damping somewhere close to the beacon: If the only route to the beacon is damped, this results in unreachability in the rest of the Internet. However, damping is usually released within at most 60 minutes, but no further update can be observed in the following two hours. It is improbable that orange A-events result from route-flap damping, but the possibility is worth to be mentioned.

The typical orange A-event consists of one withdrawal: The median number of updates is one, by definition a withdrawal. The maximum number of updates within an orange A-event is 6 updates, and 5 353 orange A-events (85.8 %) do not contain any announcement. The median number of updates within an orange W-event is 2 updates, with a maximum of 56 updates. Most

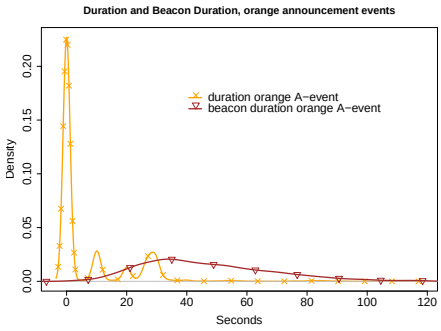

Figure 4.19: Duration and Beacon Duration: orange A-event.

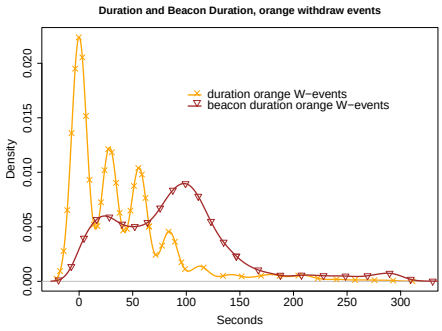

Figure 4.20: Duration and Beacon Duration: orange W-event.

orange W-events (96.6 % of 4 786 orange W-events) do not contain a withdrawal.

The density of the interarrival times of orange events can be seen in Figure 4.18 on the previous page. The density of interarrival times in orange W-events (red line) has its highest peak at about 30 seconds, and additional peaks of decreasing height at roughly 0, 60 and 90 seconds. This again shows the influence of the MRAI timer.

This influence is confirmed by the durations in orange W-events which are shown in Figure 4.20. Their density also peaks at multiples of the MRAI timer. The density of beacon durations of orange W-events in the same figure, plotted in maroon, shows a rather even distribution over the time scale, except that two peaks can be distinguished.

Without knowledge of the actual instability event, it seems impossible to distinguish green A-events from orange W-events: They converge quickly, and both end in an announcement. Some of the orange W-events shown here (with limit 6 minutes) have too long a burst duration to be classified as green bursts (limit 2 minutes). But apart from this subtlety, the bursts in orange W-events have the same characteristics as green bursts in A-events.

After observing that all interarrival time density plots shown so far are dominated by peaks at multiples of the MRAI timer, it is apparent from Figure 4.18 that orange W-events (blue curve) depart from this well-examined form: The largest peak in the blue density is at about 10 seconds. In addition, a peak at 20 seconds is also visible. Almost all values thus represent an interarrival time of 10, 20 or 30 seconds, multiples of 10 seconds instead of

the common 30. The density of durations of orange announce events in Figure 4.19 verifies those multiples: While the highest peak of the density for orange A-events is at 0 seconds, the remaining peaks are at multiples of 10 seconds. The beacon durations also shown in this figure appear to be evenly distributed over the time frame.

The multiples of 10 seconds observed in the interarrival times and durations of orange announce events can be explained with interactions between routers within the same AS. In BGP, routers from different ASes exchange reachability information. But of course, different border routers within one AS also have to exchange the newly acquired information. This is done using *I-BGP*, BGP for use within an AS. I-BGP requires that all border routers stay in contact with each other continuously. In theory, a full mesh of TCP connections is needed. Obviously, this approach does not scale in autonomous systems where dozens or even hundreds of border routers are at work: With n border routers, a full mesh requires in the order of n^2 TCP connections. A typical technique to avoid the full mesh of TCP connections is *route reflection* [40]. In this scenario, every border router is assigned to a cluster, each cluster has at least one route reflector and all route reflectors inside the AS are fully meshed. Border routers themselves usually are no route reflectors since this would impose additional overhead.

Similarly to the MRAI timer for E-BGP[15], there is also a minimum per-prefix advertisement timer for I-BGP, the I-BGP MRAI timer. The default value for this timer in the routers dominating the market today is 5 seconds [6].

When using route reflection, it takes an update two times the minimum, usually two times 5 seconds, to reach a border router from another border router. This value, 10 seconds, is reflected in the density of durations in orange A-events. In orange A-events, a previously unreachable prefix is withdrawn after its announcement at the origin, and traces of I-BGP can be seen in the data. The orange A-events may thus represent a bug in the interaction of I-BGP with E-BGP.

The orange events actually point out problems in BGP: It seems unnecessary BGP traffic that is hard to explain. Something seems to go wrong in those

[15] E-BGP (as opposed to I-BGP introduced above), External BGP, is the BGP that I talk of in all the rest of the thesis.

beacon events. However, the reasons may be found in BGP implementations with the help of router testing, and this goes beyond the scope of this thesis.

4.4 Invisible events

There are beacon events where no update shows up at the peer during the fixed two-hour interval. They account for a large fraction of the theoretically visible events. By considering 4 months of data from up to 90 peers for 13 beacons, data from roughly 1.6 million peer-beacon events[16] should be available. Only 620 000, or roughly 40 % of them, have an echo in the update trace and were classified as green, red or orange events in the previous section.

A beacon event that is invisible at a peer is termed **grey event**. Grey events account for more than 60 % of the theoretically observable beacon events, a sizable percentage especially since at first sight there is no reason for it. Possible causes for grey events are:

1. Aggregation
2. Failure of the collector
3. Failure of the peer
4. Failure at the beacon
5. Failure at the upstream provider[17]
6. Filtering

If a beacon is subsumed into a bigger prefix via aggregation, the sole purpose of the beacon is gone, namely to provide data for research. But for a regular prefix, there is no harm done if the aggregation is done correctly.

Points two and three can usually be discerned quite easily: If suddenly all peers of a collector do not send any updates any more, it is clearly a failure of the collector, as there is a constant noise of BGP updates in BGP traffic. If only one of several peers of a collector shows such behaviour, then the peering session must be down.

Points four and five are hard to distinguish. If at some point in time, all peers stop sending updates for a beacon, but continue sending updates for some other beacons, then the beacon or its upstream provider (or some link

[16] i. e. an event from one beacon seen at one peer
[17] Note that 12 out of 13 beacons are single-homed, i. e. they have only one upstream provider

Beacon event	Peer 1	Peer 2	...	Peer 12
Jan 1, 0:00 (A-event)	...			
Jan 1, 2:00 (W-event)		...		
...			...	
Jan 10, 22:00 (W-event)				...

Table 4.5: Schematic representation of proposed visualization technique.

in between) must have some kind of problem. By examining the data for other prefixes issued by the same provider, the question whether it is the beacon or the provider can be answered.

Filtering is a technique used at routers to make sure the stored reachability information is valid and, at the same time, the routing tables remain manageable. Large routing tables may slow down the data packet processing of a router. In 1997, at least one large ISP implemented a filtering policy to ignore /20 and longer prefixes, and other ASes followed suit [19]. Although controversial at the time, filtering at the /19 prefix length helped in controlling routing table growth. As the beacon prefixes are /24 networks, they may be filtered out at some ASes.

4.4.1 Visual patterns

It is not easy to decide of how much interest grey events may be. Since real updates seemed more interesting, I gave them only a cursory look. Usually, a peer either sees most of the beacon events of one beacon or none at all. Most of the grey events are caused by peer-beacon pairs where the peer never sends an update about the beacon. Other grey events happen around session resets of peers or collectors or when a beacon is down for a short time.

When building up tables for visualisation of the type (colour) of events (e. g. for each beacon one table, with rows for events, columns for peers), each field can be coloured according to the class to which the beacon event seen at that peer belongs. Table 4.5 gives an idea of how such a table will look like. An actual screenshot is presented in Figure 4.21 on the next page and will be discussed later.

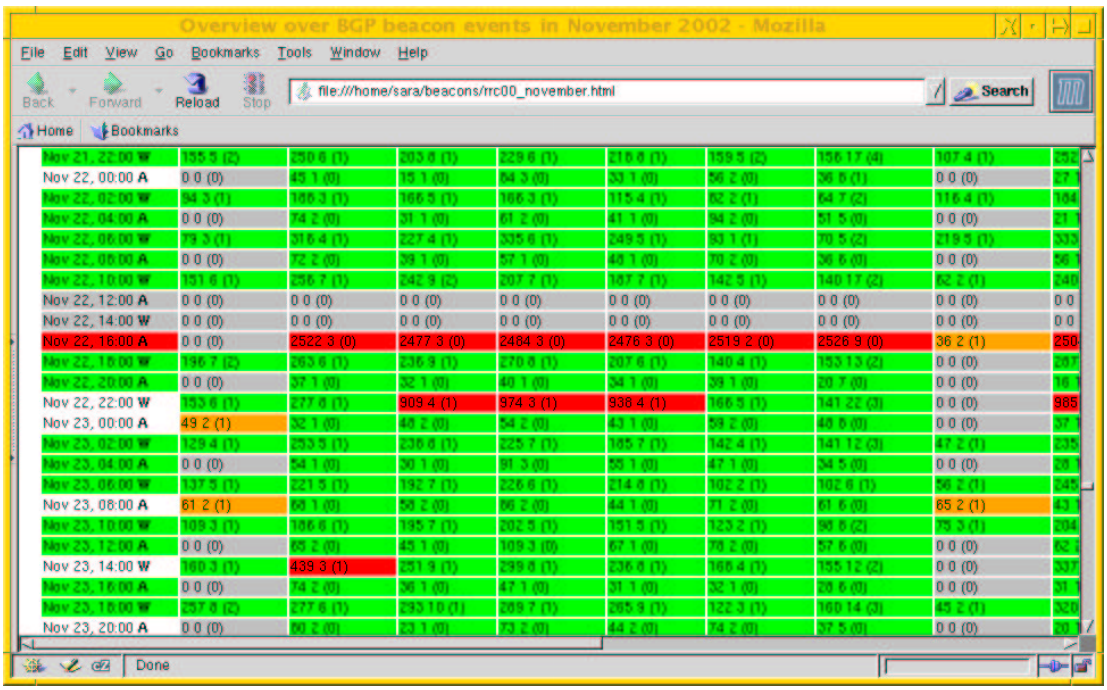

Figure 4.21: Screenshot: Sample visualization table.

Pattern red-grey-red

A typical pattern seen in these tables is a red row followed by a small number of grey rows, followed by another red row. Since more than 90% of the visible events are green, the remaining rows in the table typically are mainly green with a few red fields. The red-grey-red pattern can easily be identified as a failure of either the collector or the beacon: At one point in time, one whole row (i. e. the same event at all peers) is full of red events for this beacon. In the following events, there is no data available about that beacon, and that is why the rows are grey. Once the failure is repaired, the resulting updates are usually sent in the middle of some beacon event. That is why another row is coloured red, but the next beacon event already reflects normal behaviour. The failure is of the beacon if, for another beacon, updates can be observed at the collector for the same time period.

An example for this pattern is shown in the screenshot in Figure 4.21. It is a beacon failure since other beacons do not exhibit the same pattern in their table for this time period. The screenshot shows a part of the visualization table for beacon R_1 in November 2002 for collector RRC00. Each row repre-

sents a different beacon event. Its time is shown in the first column together with the type of the event (W for W-event, A for A-event). There is one column for each peer. The screenshot shows a subset of the 12 peers. The fields are coloured according to the class of the beacon event. This failure pattern only has a red row at its end but no red row at the beginning. The reason for it is that the failure happened during a W-event: At the time of the failure, the beacon prefix had already been withdrawn successfully. Therefore, no further updates were necessary to account for the beacon failure.

The table in this screenshot (Figure 4.21) includes additional information for each peer-beacon event: The three numbers in each field, in format $x\ y\ (z)$, are beacon duration, total number of updates and, in braces, the number of withdrawals. I noticed that if a row has multiple red entries, the beacon durations often have almost the same value.

The example screenshot shows a table built on a per-beacon basis. Another possibility is to build tables on a per-peer basis. Each row represents again the beacon event, but the columns represent beacons. This is how resets of peering sessions can be identified. In some cases, it is necessary to have access to both kinds of tables, for all beacons and for all peers.

Further insight might be gained by characterizing global patterns in the data like the one described above. There are more mysterious patterns, for example where the red-grey-red pattern appears on a subgroup of the peers for one or several beacons, but all other peers show normal behaviour. If subgroups of the peers are affected, this can point to route-flap damping. It could prove interesting to analyze the topological relationships between those peers that are affected and those that are not.

Pattern (green-grey)$^+$

There are peer-beacon pairs where the column of events has the encoding green-grey-green-grey-(...), i. e. green W-events interchanged with grey A-events. An example can be seen in the screenshot in Figure 4.21 in parts of the second and ninth column, i. e. at two peers for beacon R_1 in November.

This appears to be unnecessary traffic: If the beacon is not announced in the A-event, it is not necessary to withdraw it in the W-event either. Maybe this can be explained by some aggregation mechanism, or else it could be an

interaction of IGP with BGP. I leave the question unanswered.

Sudden appearance

Sometimes, updates for beacons that usually are not seen on most peers may suddenly be observable for one event. For example, Beacon R_7 is observed for only one event during the month of October on RRC00. Close examination of the relationships with global BGP events or excessive BGP traffic did not resolve this mystery. More likely, the sudden appearance was caused by a temporal change of the filtering or aggregation policy at some intermediary AS.

The sudden appearance of Beacon R_7 takes place on October 7, 2002. The schedule of the beacon requires an A-event at 8 o'clock. Some updates are observed at every peer. The last update is always a withdrawal, and the beacon duration is similar on all peers. All remaining events of Beacon R_7 in this month are invisible.

4.4.2 Summary of grey events

Grey events account for a large part (60 %) of the theoretically possible peer-beacon events. Most grey events are due to aggregation, in usual operation a desired feature. A rather more important question is why solitary grey events appear between green events: When an A-event is propagated and the following W-event is not, this leads to the wrong reachability information: The prefix appears reachable even though we know it is not the case. This may be considered dangerous because the router forwards packets that cannot reach their destination.

But the same is true when an A-event that follows a withdrawal is not propagated. The router cannot forward packets that could indeed reach their destination.

It is difficult to judge which of the two cases is more serious. From the prefix' point of view, which wants to be reachable at all times, an announce event should converge immediately. In a withdraw event, the prefix may have an alternative arrangement to connect to the Internet with another medium. If the unreachability is unknown, the alternative connection cannot be used. In this grey withdraw event, the unreachability is unnecessary. From the point

of view of a customer trying to reach the prefix, he may prefer to know for sure if the prefix is really reachable and even accept unreachability, but he does not want to send data and wait for an answer in vain.

If one sums up the periods of validity of a route from the update trace, some routes to beacons remain announced more than half of the time in total. However, the beacon by definition is reachable only half of the time. Therefore, packets sent to this destination would often be sent in vain, some of the time without the router realizing it. This can be due to grey W-events after a green A-event. An alternative explanation are update losses in the collection software.

Visual patterns as described above provide a higher level of BGP analysis. The usual methodology up to now mainly consists of analysis of single peer-prefix pairs. In contrast, a visualization table gives an overview over a slightly larger group of peers or beacons. This can lead to the identification of inter-relationships between groups of peer-beacon pairs. Additionally, the method provides a way to get a very good impression of the overall behaviour in BGP beacon traffic.

It is surprising that 60 % of the theoretically visible beacon events do not result in updates observed in the update trace. Some of the grey events can be explained easily, others point to interesting problems. Since no updates are observed in the grey events, it is rather difficult do draw any conclusions of those events. To check any further inferences, it is necessary to investigate in various data sources, e. g. policies of different ASes. To this end, interactions with network operators would be necessary. This is beyond the focus of this thesis.

4.5 Summary

Analyzing BGP beacon traffic gives a good picture of BGP's convergence behaviour for simple, non-flapping instability events. It can be observed that BGP generally converges within 2 minutes for announce events, and within 6 minutes for withdraw events. Considering the size and complexity of today's Internet, this seems to be reasonable for BGP convergence.

But there are BGP beacon events with long convergence times. To identify reasons behind rather long and therefore unexpected convergence be-

haviour, beacon events are classified into three classes.

The green class contains quickly-converging beacon events. With more than 93 % of the beacon events, the green class represents prevalent convergence behaviour. The main delay ingredients are the MRAI timer, and router and/or network delay. Burst durations within green events are usually multiples of the typical MRAI timer value, 30 seconds. But additional delays can add several seconds to the total convergence time.

Beacon events with long convergence times are assigned to the red class. Red events and the number of BGP updates during the same time period are uncorrelated and represent more than 4 % of the beacon events. Most of the red events have slightly larger convergence delays than the green ones. This can be explained by long AS paths and/or other additional delays. However, although the beacon event consists of one single update, some beacon events are subject to route-flap damping. This implies that route-flap damping suppresses valid, non-flapping routes to prefixes. The reason for this is that several incrementive updates, inherent in a BGP system of this size, can –with the parameters currently in use– trigger damping too easily. This leaves us with only a negligible percentage of the red events whose origin cannot be explained.

The orange class contains beacon events that converge quickly, but in the wrong state. For example, the beacon event is an announcement, but the update burst that is observed results in a withdrawal of the route to the prefix. The orange class represents 1.8 % of the beacon events. There are indications that orange A-events may be caused by I-BGP interaction.

60 % of the beacon events do not leave any trace in the update collections. Aggregation, filtering and failures do not fully explain these missing events. To give an overview over beacon behaviour, I use a methodology that involves coloured tables. Several interesting patterns are observable in these tables. The methodology could be a good approach to interpret the interrelationships of different observation points via the differences in observed convergence behaviour.

The BGP beacons are an important innovation to gain a better understanding of BGP in general. However, 13 BGP beacons are toys that cannot really reflect all of BGP's complexity. In the following chapter, the classifica-

tion derived from beacon traffic is therefore transferred on global BGP traffic to gain insight into the overall convergence processes.

Chapter 5

Convergence processes in global BGP data

The results from Chapter 4 indicate that within quickly-converging prefixes, the main contributors to the convergence time are the MRAI timer together with some additional delay at the routers or links. Other reasons, such as route-flap damping, contributed to long convergence time. In the beacon study, I separated beacon events with quick and long convergence times. In this chapter, I want to test this classification on general BGP updates and to examine if BGP beacons give a good picture of BGP convergence behaviour in general. To this end, different aspects of BGP convergence are considered to see if and where results from the beacon study can be applied to general BGP traffic.

5.1 Data sets

The data set considered in this chapter is from the same time period as the data set used for the beacon analysis in Chapter 4. The data set consists of four months of update traces starting on October 1, 2002 and ending on January 31, 2003. Therefore, results from both studies can be compared. Due to the magnitude of the data set, I used only the update trace from RIPE's collector RRC00 [34] instead of the data from all collectors. Spot checks show that the characteristics of this data set are representative for all collectors. The trace is referred to as *global trace* and contains more than 346 million updates from 12 peers.

Results from the beacon study in Chapter 4 will be revisited for comparisons with results from the overall BGP data. The data set used for the beacon study is described in Section 4.2 and is referred to as *beacon trace*. Often, only a subpart of the beacon trace is used in the comparisons, the beacon updates from collector RRC00. These updates form the intersection of the beacon trace with the global trace, and the trace is denoted *RRC00 beacon trace*.

In contrast to the beacon trace that contains only the 13 beacon prefixes, the global trace contains reachability information for 181 081 prefixes. This rather large number of prefixes needs an explanation. Today's global BGP routing tables usually consist of 120 000 prefixes. (One routing table snapshot of the BGP forwarding table on one peer.) As different peers have different views on the Internet, the prefix count from several routing tables from different peers yields a larger number. Furthermore, as the trace covers 4 months of data, some cases of deaggregation, valid only for a limited time, will increase the total number of prefixes.

As in the last section of the beacon analysis, updates are combined into update bursts using a timeout value of 500 seconds. This results in 100 435 284 bursts in the global trace and 706 392 bursts in the beacon trace. Recall that the global trace contains the updates for the beacon prefixes that appear on the RRC00 collector, i. e. the updates of the RRC00 beacon trace. The RRC00 beacon trace contains 145 631 update bursts. Expressed as a fraction of the bursts in the global trace, the update bursts of the RRC00 beacon trace account for 0.145 % of the bursts. This is the same ratio for bursts as well as for updates: The updates of the beacon trace account for 0.14 % of all updates.

5.2 Transfer of the classification

For an arbitrary non-beacon update or burst, we usually do not know the instability that caused it. Only for BGP beacons it is known at what time what kind of update was sent, and where the instability event occurred. In the classification in Section 4.3.2, bursts were assigned a colour depending on the duration of the beacon event of the respective two-hour bin. The assumption of the beacon chapter (Chapter 4) is that instabilities in beacon traffic only

occur at the originating AS of the prefix, and only at the times specified by the beacon schedule.

In the global trace, instability events can occur at the network that presides over the prefix, at the originating AS, or somewhere between the originating and the observed AS. Furthermore, policy changes that modify the best route selection somewhere in the Internet may also cause an update burst at the observed peer. In short, time and type of the instability event that cause a burst are usually unknown.

Since an update burst of the global trace uses a timeout value of 500 seconds, the stable state for the prefix remains valid for at least 500 seconds. The way in which the update bursts are classified thus needs to be changed. The stable state is represented by the last update in the burst. In the green class of the beacon study, which represents more than 90 % of the observed beacon events, the stable state at the end of the update burst is of the same type as the underlying instability. Therefore, I use the following assumption for the global trace: The type of the last update of an update burst indicates the instability type that caused it. Bursts that end with a withdrawal are referred to as *W-bursts* since they are interpreted as withdraw instabilities in the same way that W-events are in the beacon study. Bursts that end with an announcement are called *A-bursts* accordingly.

It is clear from the results in the previous chapter that this is not always the case: The beacon trace contains 1.8 % orange events, which converge to another state than the type of the instability. In addition, route-flap damping can lead to a withdrawal of a prefix if no alternative route is available. However, the knowledge about instability time and type is simply not available in the global trace. Therefore, the use of this heuristic seems appropriate.

The observed bursts are classified into two classes, a green class and a red class. The parameters are the same as in Section 4.3.2:
A burst is classified as **green** if:

- the burst ends with an announcement and lasts up to two minutes or
- the burst ends with a withdrawal and lasts up to six minutes.

All other bursts are classified as **red**.

The analysis of the red and green update bursts of the update trace are discussed next.

5.2.1 Update burst statistics

	A-bursts	W-bursts	all bursts	% of W-bursts
green bursts	73 419 350	5 155 748	78 575 098	6.6 %
red bursts	20 765 693	1 094 493	21 860 186	5.0 %
total bursts	94 185 043	6 250 241	100 435 284	6.2 %
% of red bursts	22.0 %	17.5 %	21.8 %	

Table 5.1: Burst statistics for the global trace.

	A-bursts	W-bursts	all bursts	% of W-bursts
green bursts	72 566	71 471	144 037	49.6 %
red bursts	1 064	530	1 594	33.2 %
total bursts	73 630	72 001	145 631	49.4 %
% of red bursts	1.4 %	0.7 %	1.1 %	

Table 5.2: Burst statistics for the RRC00 beacon trace.

Table 5.1 shows the distribution of the red and green bursts in the RRC00 update trace, while Table 5.2 shows the same data for the RRC00 beacon trace. The second table is a subset of Table 5.1, namely those bursts that refer to beacon prefixes. The second row in Table 5.1 contains information about green bursts. Only 6.56 % of all 78 575 098 green bursts end with a withdrawal. The third row presents the same information for red bursts (those that last longer). The fourth row sums the red and green bursts, while the last row shows the percentages of red bursts among all bursts. The column "W-bursts" shows the total number of bursts ending with a withdrawal, while the last column, "%W of bursts" shows their percentage in the total row.

About one fifth (21.8 %) of the total bursts in the global trace are coloured in red. This percentage is significantly larger than for the RRC00 beacon trace: Only 1.1 % of the beacon bursts are red. This percentage is surprising as 4 % of the beacon events were classified red. Two aspects are responsible: First, red events can contain up to 10 bursts. In fact, only 11 % of the bursts in red events are red bursts. Second, a single red burst can span more than one event and it can cause several events to be classified as red: The red burst starts during one beacon event and ends in another later event. Since there are

updates late during the first event, this causes the first event to be classified as red. If the burst spans more than the first 2 to 6 minutes of the next event, the next event is also classified red, etc. As mentioned in Section 4.3.2, 0.4 % of all bursts in the beacon trace span over more than one event (each event corresponds to a two-hour bin).

Surprisingly, only 0.74 % of the W-bursts in the RRC00 beacon trace are red. This again indicates that withdraw events take a bit longer to converge, 6 minutes instead of 2, but on the other hand, the possibility that a withdraw event does not converge in the specified limit is very low. This possibility doubles for A-bursts, but 1.45 % of red A-bursts is still a rather low percentage.

Overall, the percentages for the global trace are quite different from those for the RRC00 beacon trace. Many more of the beacon bursts end with withdrawals, and significantly fewer bursts are coloured red. This may raise some doubts about the portability of results from the beacon study to global BGP data. Yet the first concern is easy to explain: 50 % of all beacon events are withdraw events. Since the beacon prefix is unreachable for 2 hours, this will result in a W-burst. The prevalent beacon behaviour includes one burst per event which is observed at the beginning of the two-hour time period for the event. For 50 % of all beacon events, this is an A-burst, and for 50 % a W-burst. In contrast, most instabilities in the Internet represent changes in reachability. Only a small fraction of instabilities are due to unreachable prefixes and should therefore end in a withdrawal. Most customers today are multi-homed, i. e. they have more than one provider. When a link to a provider fails, the customer is not unreachable. Rather, the route to its prefix has to be changed. This is reflected in the low percentage of only 6 % W-bursts.

Actually, one may argue about calling 6 % W-bursts a low percentage: On the one hand, 6 % is a small percentage. On the other hand, when claiming that most customers are multi-homed, 6 % may appear rather large. Some customers are still single-homed, i. e. there are failures where no alternative route exists. If the time to repair is significantly more than 500 seconds, such a failure will cause a W-event, and the repair will cause an A-event. Another reason for W-bursts are route-flap damping: A damping peer will withdraw a route it believes to be flapping. Usually, an alternative route is available, but if there is no alternative route, route-flap damping will result in a W-burst,

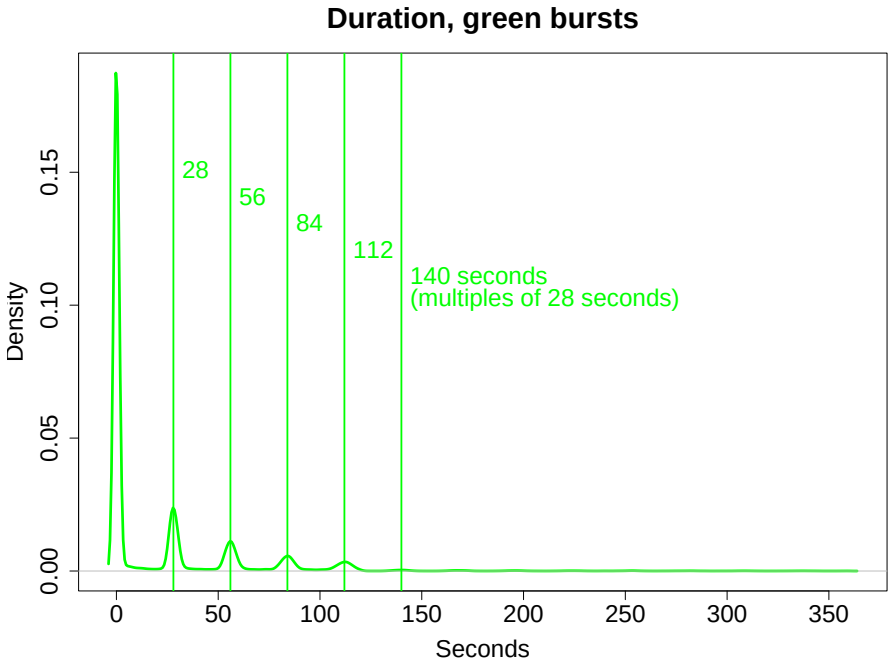

Figure 5.1: Duration of green bursts.

followed by an A-burst after the release of the damping.

In sharp contrast to the percentages in the beacon traffic, the global data shows more than 20 % red bursts, a rather alarming number. Therefore, the characteristics of the red and green bursts are investigated more thoroughly in the following sections.

5.2.2 Green bursts

Green bursts represent 78.2 % of the global trace, but only 6.56 % of those green bursts are W-bursts. In other words, 73.1 % of the update bursts converge in up to 2 minutes (green A-bursts), 5.13 % converge in up to 6 minutes (green W-bursts).

Figure 5.1 plots the smoothed density of the duration of green bursts in the global trace. For technical reasons, the density was computed by sampling the data: one value out of 100 was selected randomly. The x scale is not cut off. It contains all possible durations of green bursts, since the maximum burst duration is 360 seconds (6 minutes) by definition.

The density shows that most green bursts have a duration of 0 seconds, which is also the median duration, and longer bursts usually have a duration that is a multiple of the jittered 30 second MRAI timer. Vertical green lines were included for multiples of 28 seconds. These coincide perfectly with the peaks of the density. The mean duration of green bursts is 21 seconds and the 90 % quantile is 79 seconds. The maximum duration is 360 seconds, the same value as the maximum duration of green bursts by definition.

The density of the duration of global green bursts shows the same characteristics as the duration of green beacon events in Figures 4.3 and 4.4 on page 34. This is a clear indication that quick convergence processes in BGP are well reflected by the behaviour of green beacon events. Furthermore, we know from the beacon study that router and network delay will have a share in the total convergence time. This will result in larger convergence time, and each peer will have a different time to first update. This time to first update in the previous chapter is captured by the beacon latency. For non-beacon bursts, this value can usually not be identified.

Beacon traffic is thus a good basis for understanding simple BGP convergence processes. Through the interconnectivity in the Internet, BGP peers tend to get many different routes, and together with the MRAI timer, some BGP peers need to update their best route several times. The interarrival time between updates for the best route corresponds to the value of a jittered MRAI timer. This leads to MRAI timer cycles, but usually, for announcements, the convergence process is finished within 2 minutes, while the withdrawal instability events tend to take longer (up to 6 minutes).

5.2.3 Red bursts

Red bursts represent 21.8 % of the global data, of which 5 % are withdraw bursts, i. e. most cases of long convergence times are caused by announcements. Announcements represent new routes, alternative routes or modified attributes. And in one fifth of all announcement bursts take a rather long time. They last for more then 2 minutes.

There are a lot of red bursts with very long duration. The maximum duration that was recorded for a red burst in the data was 4 939 790 seconds, i. e. one prefix caused updates for 57 days, 4 hours, 9 minutes and 50 seconds,

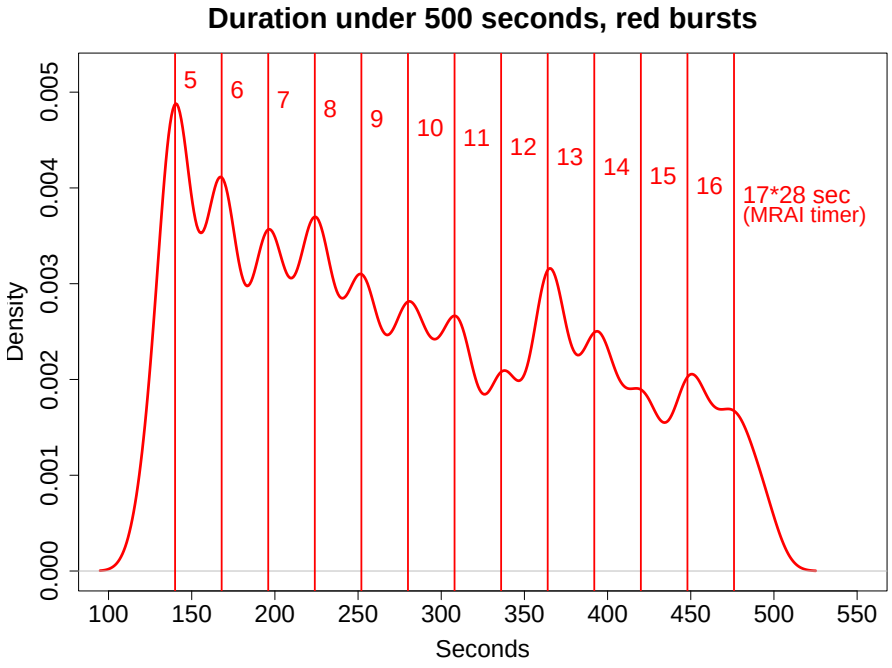

Figure 5.2: Duration of red bursts with maximum length 500 seconds.

without a break larger than 500 seconds. Looking at the density of the durations of red bursts, one can only perceive that most durations are well below 2 000 seconds, while the remaining red bursts show durations in the whole range up to the maximum value, i. e. 4 939 790 seconds.

To show more detail, Figure 5.2 plots the density of red burst durations that are less than 500 seconds. In this plot, the MRAI timer peaks are still perceivable, and the red vertical lines mark $5, \ldots, 17 \times 28$ seconds, the equivalent of multiples of the jittered MRAI timer. Although there are peaks at multiples of the MRAI timer, there are further points to be noted: The density does not stick to these multiples very closely, and there are a lot of durations that do not fit this pattern.

Besides, there is no strict decrease of the peaks in the density density per MRAI multiple as can be observed in Figure 5.1. Instead, the peak at 13×28 seconds is much higher than those at $10, \ldots, 12 \times 28$ seconds. Additionally, after 350 seconds, the vertical lines showing 28-second multiples stop hitting the peaks as they did for smaller values. One possible explanation is that there

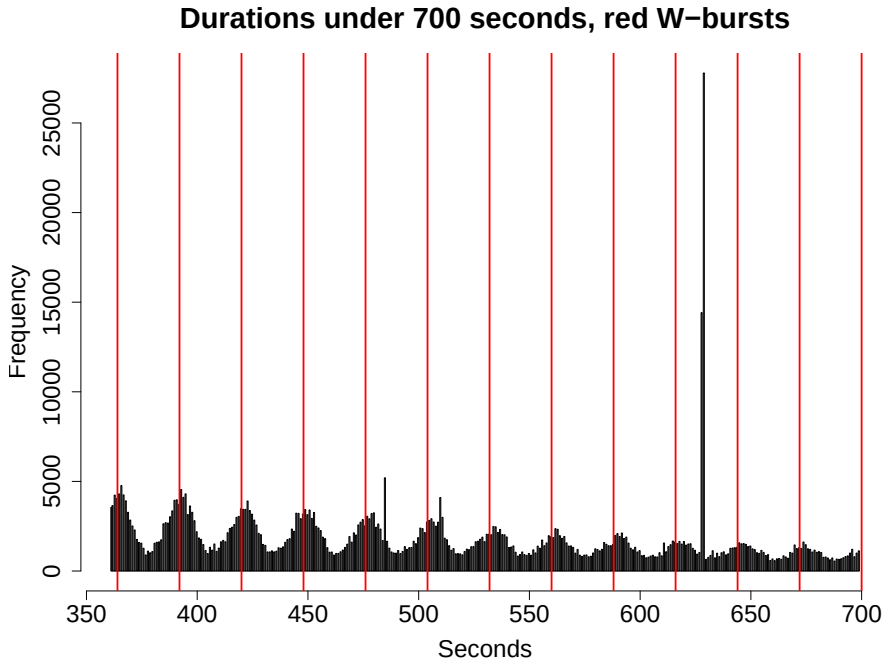

Figure 5.3: Duration of red W-bursts with maximum length 700 seconds.

are damping cycles with small damping intervals of about 360 seconds. This is a good explanation for the reappearing pattern that the 13×28 seconds peak is high and the following have decreasing height, just as the first and following peaks do and as the first and following MRAI peaks do in Figure 5.1.

Since one may suspect that the peak at about 360 seconds could be caused by red W-events that failed the limit for green bursts only by a small measure, I further analyzed the durations of red bursts. The density of durations of red A-bursts has the same characteristics as the density of Figure 5.2. The density seems almost independent from the durations of red W-events. Note that W-events account for only 5.0 % of the red bursts.

Figure 5.3 plots the histogram of the durations in red W-bursts up to 700 seconds. Vertical lines at multiples of 28 seconds are again introduced in the picture. The influence of the MRAI timer is again clearly visible. The histogram shows surprisingly large bars at around 630 seconds, and two smaller outliers at 485 and 510 seconds. The bars at around $10\frac{1}{2}$ minutes are 42 203 durations of 628 and 629 seconds. One immediately suspects those values

to be damping effects since the values are exactly 10 minutes plus one MRAI timer. As a matter of fact, the two bars are the result of three back-to-back BGP messages from one peer. Those three messages contain 40 342 prefix withdrawals concerning prefixes that were mentioned for the first time either 628 or 629 seconds earlier. Only one rather improbable explanation has come to my mind. Theoretically, a larger prefix p is deaggregated, and a lot of more-specific prefixes are announced. Prefix p is re-announced almost instantly, but this flap leads another router to damp prefix p. After 10 minutes, the damping is released, and all more-specific prefixes are withdrawn. In this scenario, the durations of 40 342 W-bursts are influenced by route-flap damping parameters. This scenario is rather implausible since aggregation needs to be configured in the router and is usually not done automatically. However, it appears worth to be mentioned.

The total median duration in red and green bursts remains 0 seconds. This is the same value as the median duration in the green bursts. The reason for it is that there are so many green bursts of duration 0 seconds, and that red bursts are only 21.8 % of all bursts.

The median number of updates in red bursts is 4, the 90 % quantile is 9 updates, and a maximum of 205 404 updates were seen in one red burst. This maximum number occurs in a red withdraw burst that lasted 3 735 125 seconds (more than 43 days). The maximum number of updates in red A-bursts is 186 575. Note that the burst with the maximum number of updates and the one with maximum duration are not identical. But the maximum number of updates also occurs in a long-lasting red A-burst. Its duration is 3 386 697 seconds (more than 39 days).

For comparison: The median number of updates in one green burst is 1, the 90 % quantile is 3 updates in a green burst, and the maximum number of updates in one green burst is 62.

The number of updates in all red bursts are close to those for red A-bursts. But since red A-bursts account for 95 % of all red bursts, this is not really surprising. The median number of updates in a red W-burst is 7, and the 90 % quantile is 17, i. e. W-bursts contain about twice the number of updates as A-bursts. But at the same time, the median duration of red A-bursts (W-bursts) is 327 (629) seconds, and the 90 % quantile is 793 (1 698) seconds. Roughly

speaking, the number of updates per minutes is the same. Further analysis confirmed this rough conclusion: All red bursts have an update-per-minute rate of 0.6 (median) to 0.8 (mean) updates per minute. The rate is slightly larger for red A-bursts. They also have a higher maximum, at 44.2 updates.

The values for updates per minute in green bursts appear to be larger, e. g. 4 updates per minute is the median for green A-bursts. But the numbers do not include green bursts with duration 0: One cannot compute the updates per minute for a duration of 0 minutes. However, those short bursts with duration 0 account for roughly two thirds (64.3 %) of the bursts. The values for updates per minute for green bursts are therefore not very representative.

By the way, 0 second bursts usually contain exactly 1 update as one may suspect. Only a very low percentage, 0.59 % of the green A-bursts with duration 0 seconds contain more than one update. The maximum of updates is 12. The percentage is even lower for green W-bursts: Only 0.40 % of all green W-bursts with duration 0 contain more than one update (maximum 8).

In summary, even in red bursts, the MRAI timer cycles can be identified very clearly, but the reasons for red bursts could not be revealed by this statistical analysis. I have been arguing in the last chapter that route-flap damping is dangerous. However, in the global trace, with more than 20 % red bursts that can last a very long time, it is obvious that route-flap damping is not widely deployed. Route-flap damping would reduce the number of red bursts significantly. It is not clear how much harm this would do and how many more update bursts would be seen. An estimation of the impact of route-flap damping is provided in the next section as well as a discussion of its assets and drawbacks.

Note that 81.4 % of the red bursts have a duration of under 600 seconds. In other words, almost 96 % of all bursts (red and green) have a duration less than 10 minutes. While there is an alarming number of red bursts (20 %), only 4 % of all bursts have a duration larger than 10 minutes. This implies that the situation is much less dramatic. After all, 10 minutes may be an acceptable convergence time.

Interarrival times between updates in the global trace will not be shown. The interarrival times of global BGP data were analyzed already in [25], and the distributions in my data set are similar. In accordance to the interarrival

times of green and red beacon events, no difference between densities of interarrival times in red and green bursts can be perceived.

5.3 Route-flap damping

The beacon study shows that route-flap damping occurs even if the original instability event consists of only one update. Therefore, we expect to see route-flap damping for other prefixes as well. Can we identify reasons for route-flap damping or find some indication which AS is damping the observed route?

Studying interarrival times in BGP beacon traffic proved useful to identify route-flap damping. To identify the impact of route-flap damping, interarrival times between bursts in the global trace are analyzed.

The RIPE recommendation for route-flap damping [29] suggests different damping parameters for prefixes of different length. The meaning behind this idea is: Longer prefixes represent smaller networks, i. e., in case of false positives, fewer hosts are affected. Damping of shorter prefixes affects many hosts. In addition, the history of routing in the Internet indicates that smaller networks are responsible for a larger percentage of instabilities than the number of IP addresses under their control would suggest. This may be due to less experienced network administrators and their "beginner's mistakes".

Figure 5.4 on the next page shows the density of interarrival times between bursts of the same peer and prefix. Three sets of prefix length ranges were chosen to be compared with respect to their interarrival times: Short prefixes with mask length up to 15, prefix lengths from /16 to /23, and /24- and longer prefixes. Only interarrival times up to 10 000 seconds are shown. While much larger values exist, they are not relevant for damping and would blur the other information.

The interarrival times for prefixes up to netmask length 15 are shown fully (about 750 000 values). For technical reasons, the remaining data sets were sampled: For /16 to /23 prefixes, each value is chosen with probability $\frac{1}{50}$ (resulting in about 800 000 values), for the longer prefixes with a probability of $\frac{1}{100}$ (resulting in about 500 000 values). Note that short prefixes tend to have slightly shorter (up to 30 minutes or 1 800 seconds) interarrival times relative

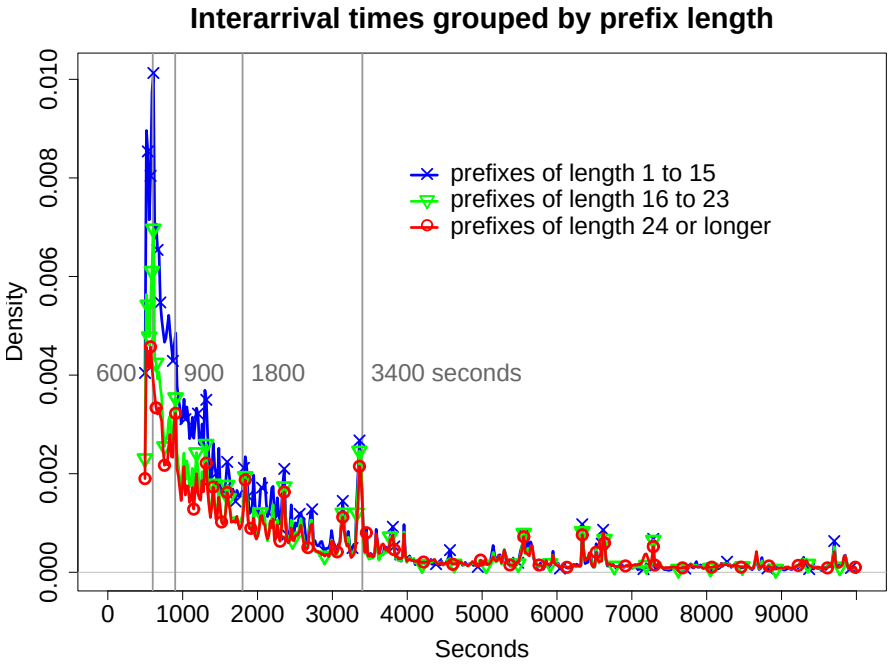

Figure 5.4: Interarrival time between bursts, separated by prefix length .

to the total number than longer prefixes do. This may suggest a higher long-term stability for shorter prefixes. Since there are no major differences in the curves, Figure 5.4 shows us that current BGP implementations do not take advantage of the flexibility to use different damping parameters for different prefix lengths.

The densities of interarrival times show small peaks at values that are "suspicious" for route-flap damping. Vertical lines are inserted at 600, 900, 1 800 and 3 400 seconds. There are peaks in all three densities which can be interpreted as typical interarrival times with route-flap damping.

However, these densities do not reveal if the damping was well-applied. Since there are many red bursts of long duration, I can presume that not all ASes deploy route-flap damping. Furthermore, the beacon study showed that if damping is enabled, this can lead to ill-applied route-flap damping. Therefore, route-flap damping will be analyzed further in the following.

5.3.1 Preferred route

In general, time and type of routing instabilities are usually unknown. Without this knowledge, it is not possible to distinguish flapping routes from non-flapping routes. Yet, the route-flap damping algorithm tries to make this distinction. Since route-flap damping leaves some trace in observed updates, it should be possible to estimate the impact of route-flap damping in the global trace. To explain the methodology, let me recall how best routes are chosen and in which cases damping is involved and observed.

In BGP, peers exchange their current best routes to prefixes. Due to the good connectivity in the Internet, a router usually has many alternative paths available for each prefix. BGP's path-selection process is responsible for selecting one route from the available routes as its best route. This process is well-defined and deterministic. It will always select the same route as long as the routes and the policies remain do not change.

However, in the course of a convergence process after an instability, the available routes to a prefix may change. This can result in changes of the route to a prefix, and this information has to be propagated to the peers. An instability can create an update burst in a peering session that triggers route-flap damping on the receiving side of the peering session. I will discuss what happens in this case: For each neighbouring peer, the router computes penalties for all prefixes reachable via this peer. If too many updates are received from one peer for one prefix, this route to the prefix will be damped, i. e. no route via this peer will be used as best route for some time period. If more updates for this prefix from this peer are received during the damping period, the damping period is extended.

When a router damps a route that is *not* its best route, this cannot be observed in the BGP update traffic since no BGP updates are generated. Computing the penalties for non-best routes simply constitutes additional processing overhead at the damping router.

When the router damps its best route, this leads to the announcement of an alternative route[1]. Consequently, all neighbours[2] will receive an update. At least some updates are generated. But the neighbours will only up-

[1] Assume that an alternative route is available.
[2] All neighbours concerned according to the policy.

date their best route if their best route utilizes the damped route or if they prefer the newly-announced route. Therefore, only some instances of route-flap damping are observable at the BGP collectors. Indications of route-flap damping, e. g. in interarrival times, can only be observed if the usual path to a prefix includes an AS that employs route-flap damping. I am going to refer to this route as the *preferred route*: A preferred route at a router is the route to the prefix that is selected as best route whenever the prefix is reachable and this route is available. This preferred route is not always the same as the best route. For example, damping can lead to an alternative best route. However, this alternative route will be valid for the duration of a damping interval, and this is exactly why damping can be observed in interarrival times.

Whenever a popular AS that routes traffic to many destinations damps its preferred route, the release of this damped route will lead to a relatively large number of updates in the global BGP system, since all downstream ASes for these routes have to change their BGP routing tables. For popular ASes, it is thus critical to avoid ill-applied damping.

As a matter of fact, the number of updates would be much lower if BGP reachability information did not contain the full path: ASes that change their best route without changing the next-hop AS on the AS path do not have to change their forwarding table. However, BGP is involved, and potentially many ASes need to update their BGP routing tables.

The preferred route is useful to estimate the impact of route-flap damping. I used the following heuristic to determine the preferred route for a peer-prefix pair: If there is a route that is identified as best route in the BGP routing table, and it remains best route at least 10 times as long as some other route, this route is interpreted to be the preferred route from the peer to the prefix. This heuristic is applicable to bursts as well as to updates. Bursts, defined with an appropriate timeout value, can be used to join together all updates of the convergence process. The result of the convergence process is the last update and should be the preferred route.

This allows me to identify for each AS which peer-prefix pairs use this AS in their preferred route. Not all ASes employ route-flap damping, and many ASes use different parameters. Theoretically, the route-flap-damping parameters of an AS can be identified by inspecting interarrival times of bursts for

all peer-prefix pairs whose preferred route uses this AS.

5.3.2 Example damping scenario

To find an AS that is supposed to use route-flap damping, I further analyzed interarrival times between bursts in red events. Recall that in Figure 4.17 on page 55, the density of interarrival times of green bursts peaks at 1 300 seconds, a rather strange value. I turned to analyze red events that contain exactly two green bursts in the hope to encounter examples of route-flap damping. I found that half of the interarrival times of those bursts are caused by a single peer-prefix pair.

In other words, there is a peer-prefix pair for which, in several announce events, two green bursts are observable. Those A-events are classified as red because their beacon duration is larger than 2 minutes. The burst history observed at peer p shows two green bursts, and their interarrival time is roughly 1 300 seconds. I consider the stable paths that result from the two bursts. The result of the first burst is called *route* R_1. The result of the second burst is called *route* R_2, accordingly. Further analysis shows that route R_1 is usually different from route R_2. More specifically, route R_2 usually includes an ASN that is not used in route R_1. I will call it *AS* d in the following.

The interarrival time of the two bursts is a value suspicious for route-flap damping. Presumably, this is what happens in those announce events: The beacon prefix is announced at the beginning of the beacon event. This results in a number of BGP updates. Recall that, at peer p, a green burst is observed at the beginning of the beacon event. Some router of AS d is lead to believe that its route to the beacon is flapping and suppresses it. This is the reason why route R_1 of peer p does not include AS d. After about 1 300 seconds, the damping at AS d is released. Since the downstream ASes all prefer the route R_2 that includes AS d, this causes another update burst. This second burst can be observed at peer p. The second stable path, R_2, is presumed to be the preferred route for the peer-prefix pair.

Furthermore, since d is an AS whose BGP damping policy is publicly available on the world-wide web, one can confirm that the observed damping effects are consistent with the announced policy.

In a next step, I identified peer-prefix pairs whose preferred route includes

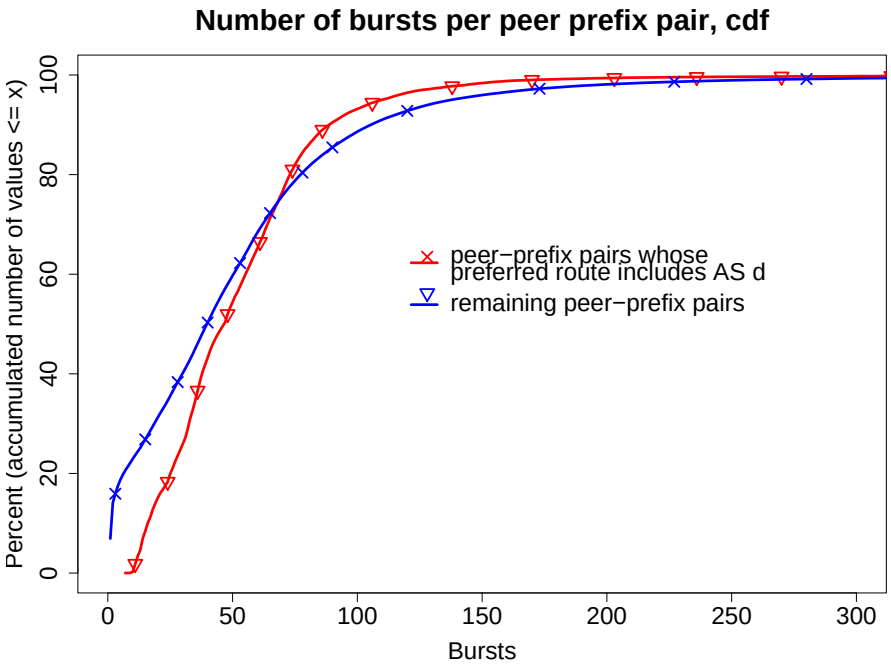

Figure 5.5: CDF: Number of bursts per peer-prefix pair.

AS d with the heuristic introduced in the previous section. I calculated the preferred routes according to the heuristic for 10 days at the beginning of the global trace (Oct 1 to Oct 9 including one routing table at the beginning) and for 12 days at its end (Jan 20 to Jan 31, also with a routing table from the beginning of this interval). I included one routing table at the beginning of each considered time period to make sure that the computation accounts for routes that remain unchanged in the routing table.

The first interval lead to 163 019 peer-prefix pairs, the second to 178 508, but the intersection contains only 109 204 peer-prefix pairs whose preferred route includes AS d. This intersection is computed over the peer-prefix pairs, ignoring the exact preferred route. In fact, it is a bit disappointing that the pairs do not seem to remain very stable.

The intersection of the two sets of peer-prefix pairs is the set S of peer-prefix pairs. S contains all peer-prefix pairs whose preferred route includes AS d both at the beginning and at the end of the global trace. I consider the total number of bursts observed in the global trace for the peer-prefix pairs.

In other words, I compute the number of bursts per peer-prefix pair in four months of BGP updates. Figure 5.5 shows two cdfs of the number of bursts seen for a peer-prefix pair. The red line presents those numbers for the peer-prefix pairs in S whose preferred route includes AS d. The blue line shows the same information for all other peer-prefix pairs in the global trace, i. e. the peer-prefix pairs not in set S.

It is remarkable that for small numbers of bursts, the peer-prefix pairs not in S give a better result: There are more peer-prefix pairs with few bursts observed in the time period. But at 68 bursts in four months of BGP traffic, the two cdfs meet and for burst numbers between 68 and approximately 400 bursts seen in four months, AS d has a better statistic: There are fewer peer-prefix pairs with a large number of bursts.

This indicates that route-flap damping can do harm: Only few peer-prefix pairs including AS d in their preferred route produce a low number of bursts. But on a larger scale, route-flap damping also seems to give a benefit. The result is thus ambiguous, but it can be seen that the damping parameters employed by AS d tend to be too harsh. The parameters chosen by AS d are rather close to the RIPE recommendation [29], so it is necessary to change the recommendation.

5.3.3 Estimated impact of damping in the global Internet

We have seen above that route-flap damping is used in the Internet. Some routes are subject to damping. The beacon study showed that not all instances of route-flap damping are well-applied. Of course, the beacon analysis cannot tell us what percentage of the damped routes in the Internet are indeed flapping since beacon prefixes do not flap.

The percentage of damping in BGP beacon traffic that is ill-applied is close to 100 %, but it is not possible to transfer this to the global BGP traffic as the number of flapping routes in the Internet is unknown. Therefore, one would like to know how much damping happens in the Internet. In the following analysis, I will try to produce an upper bound for the impact of route-flap damping in the Internet. This upper bound is also a rough upper bound for ill-applied damping and may help in the discussions about route-flap-damping parameters and algorithms.

The timeout value of 500 seconds for update bursts was chosen in Section 4.3.2 on page 38 to separate the different bursts seen in route-flap damping. In the following, I will call those bursts *small bursts*. This timeout value was reused to be able to compare the results of the two studies. But to be able to estimate the impact of route-flap damping, I regroup the bursts using a larger timeout to be able to estimate the impact of route-flap damping. The larger timeout is chosen such that different small bursts stemming from the same instability event that were separated by route-flap damping will be grouped together.

In the RIPE recommendation [29], the suggested maximum for damping of a route is 60 minutes (3 600 seconds). The data indicates that most ASes use values less than or equal to 3 600 seconds as their maximum damping interval. To account for additional delays, a value above 60 minutes, 4 000 seconds, was chosen to group bursts together to study damping effects.

On the update level, grouping together updates with timeout 4 000 seconds yields the same result as grouping together small bursts using this timeout value: After all, the interarrival times in small bursts are less than 500 seconds. If there is an interarrival time of 3 000 seconds between two small bursts, then this interarrival time is present in the updates as well. The end and beginning of each burst represent timestamps of actual updates.

By grouping together small bursts instead of updates, the number of small bursts that are grouped together in one big burst is preserved. This also gives me a way of validating my timeout value for small bursts. If most big bursts contain more than one small burst, this indicates a bad choice in the earlier timeout value.

Figure 5.6 on the next page shows the cdf of the number of 500-second bursts that are grouped together into a 4 000-second burst. The number of bursts on the x axis is plotted on a logarithmic scale. The maximum number of small bursts grouped together in one big burst is 6 084. 70 % of all big bursts contain only one small update burst, another 20 % contain 2 small bursts, and 5 % big bursts contain 3 small bursts. The median number of bursts is thus 1, the 90 % quantile is 2, the 95 % quantile is 3, and only 5 % of all big bursts contain more than 3 small bursts. As the logarithmic scale on the x axis is not very intuitive, I introduces vertical lines at 2 to 6 bursts.

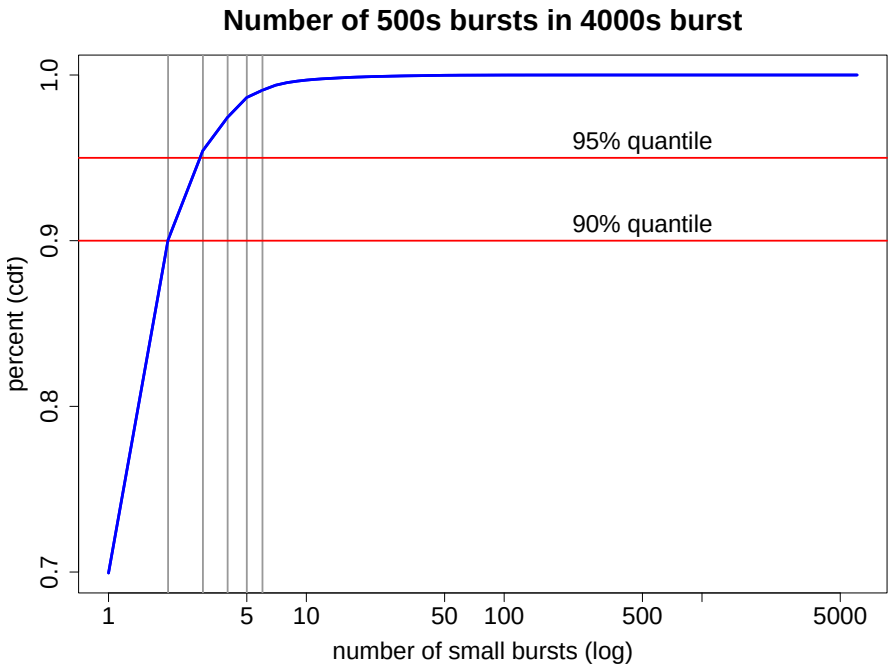

Figure 5.6: CDF: Number of 500-second bursts in one 4 000-second burst .

These numbers can be interpreted in the following way: 20 % of the 4 000-second bursts ("big" bursts) are due to route-flap damping of stage 1, i. e. a route was damped at one point and released later. This results in two small bursts that are close in time. And another 5 % suffered from route-flap damping stage 2. The route was damped at two peers in the Internet uncoordinated. The release at the first peer resulted in damping at the second peer.

Another explanation for big bursts with two small bursts are failure and repair. If the failure was repaired within one hour, this can also result in two small bursts close in time. Accordingly, the numbers only provide an upper bound the impact of route-flap damping: Strictly less than 30 % of all instabilities suffer from damping. It is below 30 % because 70 % of the big bursts contain only one small burst. 20 % minus the percentage where a failure gets repaired within one hour might be a rough estimate for the lower bound. This upper bound of 30 % will need to be lowered to give a more specific number. But it is a first step to identify route-flap damping's total impact on BGP traffic.

5.4 Summary

The classification of red and green bursts can be transferred from the beacon study to the analysis of overall BGP convergence. Relying on four months of observed BGP updates, I show that BGP beacons reflect important aspects of overall BGP convergence: The convergence time usually is in the order of minutes, and burst durations reflect the typical value of the MRAI timer.

However, there are significant differences between the observed behaviour of BGP beacons and of general prefixes: Many more withdraw bursts are observed in beacon traffic. This is due to the nature of the BGP beacons, they are unreachable half of the time. The fact that only 6 % of the bursts in the global trace are withdraw bursts reflects the good connectivity in the Internet: It appears that, in cases of failure, there usually is an alternative path available. Furthermore, BGP beacons show only a very low percentage of red bursts. In contrast, 20 % of all bursts observed in the global trace are red. This indicates that prefixes in the global Internet exhibit more complex instabilities than one update every two hours at the originating AS.

The impact of route-flap damping in the global Internet is not as easy to identify as in the beacon traffic. The concept of a preferred route is introduced to be able to analyze the route-flap-damping policy of a single autonomous system. Using this concept, I show that the deployment of route-flap damping influences the observed properties of BGP convergence: On the one hand, route-flap damping creates a larger number of update bursts for seemingly stable prefixes. But on the other hand, the number of update bursts for seemingly unstable prefixes is reduced.

By analyzing the interarrival times of bursts, an upper bound for the impact of route-flap damping can be estimated: At most 30 % of the convergence processes in the global Internet are influenced by route-flap damping.

Chapter 6

Conclusion and future work

This work is part of the effort to understand the dynamics of the inter-domain routing protocol BGP. It analyzes four months of BGP beacon traffic in detail and transfers the approach to the analysis of all BGP updates. The results show that BGP beacon traffic accurately reflects BGP convergence processes, but it cannot capture all of BGP's complexity.

The goal of this thesis was to identify factors for convergence time. The MRAI timer, additional router and/or network delay and route-flap damping, represent the main ingredients to the total BGP convergence time. BGP beacons are an important tool to identify their interaction.

Most beacon events (95 %) converge rather quickly, within a few minutes. More specifically, the beacon study demonstrates that convergence time for announce events usually is under 2 minutes. Withdraw events take somewhat longer, 6 minutes. Since withdraw events represent only 6 % of the instabilities in the Internet, this longer convergence time is of minor concern. However, longer convergence times are observed throughout. Some delays are caused by long AS paths. Other delays (1 % of the beacon events) are due to route-flap damping. Since beacon prefixes are no flapping prefixes, this result affirms that route-flap damping should be considered harmful.

Effects of route-flap damping are observable both in delayed BGP beacon convergence as well as in the interarrival times of global BGP updates. In fact, the impact of route-flap damping can be bounded from above by 30 % of the instabilities in the Internet. Since 20 % of the update bursts in global BGP data display long burst durations, the need for a regulating element like route-flap damping can be confirmed. However, current implementations

cause a larger-than-necessary number of update bursts for seemingly stable prefixes. The benefit of route-flap damping can only be observed for seemingly unstable prefixes, i. e. those that have more than 68 instabilities in the four months of analyzed updates. The parameters should be adjusted to reduce the harm of route-flap damping. With different parameters, the advantages of route-flap damping may prevail over the disadvantages of delayed convergence.

In the case of quick convergence, the primary delay contributor is the MRAI timer. Its impact can be observed in beacon traffic as well as in global BGP data. Most burst durations are at a multiple of the typical MRAI timer value, 30 seconds. Reducing the value of the MRAI timer should improve convergence time. However, it seems difficult to settle on a globally optimal MRAI interval [12] because the optimum depends on the topology. One sensible approach may be to choose the value of the MRAI timer locally, i. e. depending on the neighbourhood in the global AS topology. Reduction of the value of the MRAI timer could result in quicker convergence time. But since this would augment the number of BGP updates, this could also increase the impact of ill-applied route-flap damping. Both factors need to be considered together if one wishes to reduce convergence time in BGP.

The beacon study further reveals that an additional delay in the order of seconds to minutes is introduced by the network and/or the routers. The exact causes for this additional delay can only be identified with the help of router testing. It seems also necessary to analyze the interaction with E-BGP, the subject of this study, with I-BGP. The latter may be causing inconsistent routing states in about 1.8 % of the beacon events (i. e. the orange events).

This thesis gives a good overview over the interaction of different delay contributors. But instead of answering all the questions about BGP convergence, it points out open research questions. I will list some open research questions that were not mentioned above:

How can instabilities be located in the Internet? A primary advantage of BGP beacons is that time, type and location of the observed instability are known. In contrast, in global BGP updates, it is usually impossible to know if an update burst was generated by one or several instabilities. To be able to infer more information from observed BGP updates, it would be helpful to

identify the location of the instability in the topology. It would also enable network operators to decide if there is a problem in their field of responsibility. One way to approach this question is the correlation of different observation points. In case one group of observation points shows an instability and another does not, topological analysis could show the location of the instability.

What is the impact of flapping routes? In global BGP updates, 20 % of all update bursts have a long duration. Are flapping routes the reason? Route-flap damping was proposed to relieve routers from too many updates. The underlying assumption is that BGP updates require a lot of processing time at the routers. Today, processing time may not be a big problem any more. If routers do not suffer from high numbers of BGP updates, the MRAI timer value can be reduced to improve convergence time. To this end, route-flap damping needs to be adjusted to the new settings. It is still a goal to damp prefixes that flap heavily, i. e. for a long time.

How stable are best routes? I found that preferred routes do not remain very stable. In terms of traffic engineering, it would be helpful to understand how stable the chosen best routes remain. An approach to this question is further statistical analysis. Maybe the preferred routes are rather stable as soon as less- and more-specific prefixes are considered together.

Bibliography

[1] T. Bates, Y. Rekhter, R. Chandra, and D. Katz. Multiprotocol extensions for BGP-4, June 2000. RFC 2858.

[2] D. Beard, S. Murphy, and Y. Yang. Generic Threats to Routing Protocols. Internet Draft, work in progress, February 2003. URL: `http://www.ietf.org/internet-drafts/draft-ietf-rpsec-routing-threats-00.txt`.

[3] BGP Beacon Info. URL: `http://psg.com/~zmao/BGPBeacon.html`.

[4] Randy Bush, Tim Griffin, and Zhuoqing Morley Mao. Route flap damping: harmful? Talk at RIPE 43, September 2002. URL: `http://www.ripe.net/ripe/meetings/archive/ripe-43/presentations/ripe43-routing-flap.pdf`.

[5] Di-Fa Chang, Ramesh Govindan, and John Heidemann. An Empirical Study of Router Response to Large BGP Routing Table Load. Technical report, USC/Information Sciences Institute, December 2001.

[6] Cisco Systems BGP Configuration Guide. URL: `http://www.cisco.com/en/US/products/sw/iosswrel/ps1828/products_configuration_guide_chapter09186a00800ca571.html`.

[7] Sally Floyd and Van Jacobson. The synchronization of periodic routing messages. *IEEE/ACM Transactions on Networking*, 2(2):122–136, 1994.

[8] Avi Freedman. Industry/government infrastructure vulnerability assessment: Background and recommendations. Talk at NANOG 25 in Toronto, June 2002. URL: `http://www.nanog.org/mtg-0206/avi.html`.

[9] V. Fuller, T. Li., J. Yu, and K. Varadhan. Classless Inter-Domain Routing (CIDR): an Address Assignment and Aggregation Strategy, September 1993. RFC 1519.

[10] Ramesh Govindan and Anoop Reddy. An analysis of Internet inter-domain topology and route stability. In *Proc. IEEE INFOCOM*, April 1997.

[11] Barry Raveendran Greene. BGP security update. June 2002. URL: `http://www.nanog.org/mtg-0206/ppt/barry.pdf`.

[12] Timothy G. Griffin and Brian J. Premore. An experimental analysis of BGP convergence time. In *Proc. International Conference on Network Protocols*, 2001.

[13] Timothy G. Griffin and Gordon Wilfong. An analysis of BGP convergence properties. In *Proc. ACM SIGCOMM*, 1999.

[14] Bassam Halabi. *Internet Routing Architectures*. Cisco Press, 1997.

[15] International Earth Rotation Service. URL: `http://hpiers.obspm.fr/eop-pc/`.

[16] James F. Kurose and Keith W. Ross. *Computer Networking: A Top-Down Approach Featuring the Internet*. Addison-Wesley, 2001.

[17] C. Labovitz, A. Ahuja, A. Bose, and F. Jahanian. Delayed Internet routing convergence. In *Proc. ACM SIGCOMM*, 2000.

[18] Craig Labovitz. Multithreaded Routing Toolkit. In *Merit Technical Report to the National Science Foundation*, 1996.

[19] Craig Labovitz. *Scalability of the Internet Backbone Routing Infrastructure*. PhD thesis, University of Michigan, 1999.

[20] Craig Labovitz, A. Ahuja, and Farnam Jahanian. Experimental study of Internet stability and wide-area network failures. In *Proc. International Symposium on Fault-Tolerant Computing*, June 1999.

[21] Craig Labovitz, Rob Malan, and Farnam Jahanian. Internet routing instability. *IEEE/ACM Trans. Networking*, 6(5):515–558, October 1998.

[22] Craig Labovitz, Rob Malan, and Farnam Jahanian. Origins of Internet routing instability. In *Proc. IEEE INFOCOM*, March 1999.

[23] Olaf Maennel. Generating realistic routing tables in a test lab. Master's thesis, Universität des Saarlandes, Germany, 2002.

[24] Olaf Maennel. Observed properties of BGP convergence. Talk at RIPE 45, May 2003. URL: `http://www.ripe.net/ripe/meetings/ripe-45/presentations/ripe45-routing-bgp-convergence/`.

[25] Olaf Maennel and Anja Feldmann. Realistic BGP Traffic for Test Labs. In *Proc. ACM SIGCOMM*, 2002.

[26] Zhuoqing Morley Mao, Ramesh Govindan, George Varghese, and Randy Katz. Route Flap Damping Exacerbates Internet Routing Convergence. In *Proc. ACM SIGCOMM*, 2002.

[27] D. McPerson, V. Gill, D. Walton, and A. Retana. BGP Persistent Route Oscillation Condition. RFC 3345.

[28] John Moy. *OSPF: Anatomy of an Internet Routing Protocol*. Addison-Wesley, 1998.

[29] Christian Panigl, Joachim Schmitz, Philip Smith, and Cristina Vistoli. RIPE Routing-WG Recommendation for Coordinated Route-flap Damping Parameters , October 2001. URL: `http://www.ripe.net/ripe/docs/ripe-229.html`.

[30] S. Ramachandra, Y. Rekhter, R. Fernando, J.G. Scudder, and E. Chen. Graceful Restart Mechanism for BGP. Internet Draft, work in progress, November 2001. URL: `http://www.ietf.org/proceedings/01aug/I-D/draft-ietf-idr-restart-01.txt`.

[31] Y. Rekhter and P. Gross. Application of the border gateway protocol in the Internet, March 1995. RFC 1772.

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

[33] Réseaux IP Européens. URL: `http://www.ripe.net/`.

[34] RIPE's routing information service raw data page. URL: `http://data.ris.ripe.net/`.

[35] University of Oregon Route Views Project. URL: `http://www.routeviews.org/`.

[36] RIS Routing Beacons. URL: `http://www.ripe.net/ris/beacon.html`.

[37] Saargate. URL: `http://www.saargate.de/`.

[38] Aman Shaikh, Lampros Kalampoukas, Rohit Dube, and Anujan Varma. Routing stability in congested networks: Experimentation and analysis. In *SIGCOMM*, 2000.

[39] W. Richard Stevens. *TCP/IP Illustrated, Volume 1: The Protocols*. Addison-Wesley Professional Computing Series. Addison-Wesley, Sydney;Amsterdam;Tokyo, 1994.

[40] John W. Stewart. *BGP4: Inter-Domain Routing in the Internet*. Addison-Wesley, 1999.

[41] C. Villamizar, R. Chandra, and R. Govindan. BGP route flap damping, 1998. RFC 2439.

[42] Zebra routing software. URL: `http://www.zebra.org/`.

Index