Author: Yong Jiang
Affiliation: Telia Research AB, Sweden
Address: Vitsandsgatan 9 B, SE-123 86, Farsta, Stockholm
Telephone: +46 8 713 8125
Fax: +46 8 713 8199
Email: yong.b.jiang@telia.se

Author: Avri Doria
Affiliation: Division of Computer Communication
Institute for System Technology
Lulea University of Technology, Sweden
Mobile: +46 73 029 8019
Office: +46 920 49 3030
Email: avri@sm.luth.se

Author: Daniel Olsson
Affiliation: Operax AB, Sweden
Tel: +46 70 3537111
Email: Daniel.Olsson@operax.com

Author : Fredrik Petterson
Affiliation: Operax AB, Sweden
Tel: +46 70 2595619
Email: Fredrik.Pettersson@operax.com

Inter-domain Routing Stability Measurement

Yong Jiang, Telia Research, Sweden
Avri Doria, Lulea University of Technology, Sweden
Daniel Olsson, Fredrik Pettersson, Operax AB, Sweden

Abstract

Prior inter-domain routing measurements include passive analysis of routing table growth and error injection for convergence measurement. The measurement here is solely based on passive logging of BGP control traffic from measurement points including a major European ISP's backbone networks and some academic networks. Its goal is to understand the failures involved in today's much discussed routing stability issues. As a result, the general status of the Internet reachability and flapping has been observed. Some explanations for the erroneous phenomenon are offered. Though only macro effects from the control plane are measured, part of the future study includes determining work to determine whether these results correlate with the data forwarding plane results.

1 Introduction

There are major shortcomings in the inter-domain routing of the Internet today and this may limit the continuing growth of the Internet [1]. The Internet continues to grow both in terms of its size and in terms of the services running on it. In addition, its continuing scalability places challenges on the routing system's capability to produce a stable view of the overall reachability of the Internet. Various developments in the nature and quality of the services that users want from the Internet are difficult to provide within the current framework as they impose requirements, which were never foreseen by the original architects of the Internet routing system. Remedying these shortcomings will require extensive research to tie down the exact failure modes that lead to these shortcomings and identify the best techniques to remedy the situation.

There is a large body of anecdotal evidence, and experimental evidence and analytic work on the stability of the current main Internet inter-domain routing protocol, BGP, such as its slow convergence [2] and route oscillation [3]. This work studies BGP stability through the passive logging of BGP messages from a major ISP's backbone network and a big academic network, a different perspective from some active measurements done through active error injection [2]. One of the goals of these measurements is to find reliable indicators of network instability. Currently operators only know that a network has become unstable after the fact. It is our basic hypothesis that there will be indicators in the BGP control message flow that indicate when the network is in the process of destabilizing.

2 A Brief Overview of Inter-domain Routing

Domain can mean any collection of systems or domains, which come under a common authority. This common authority determines the attributes that define, and the policies that control that collection. In this study the meaning of domain is used in a strict legacy sense, and is therefore to be understood as meaning Autonomous System (AS).

2.1 Domains and Scalability

In practice, the total number of domains, the number of routes contained in them, and the rate of update of routes have a substantial effect on the continued scalability of the Internet. There is serious concern that if the global rate of growth of these factors, including the propagation of information to domains where it adds no operational value, will cause the current inter-domain routing system to collapse within a few years [1]. New paradigms are needed [4].

2.2 Current Practice in Domain Hierarchy

From a current architectural standpoint, routing systems are divided into interior

gateway protocols and exterior gateway protocols. IGP routing is usually controlled by a single administrative entity. EGPs, which today consist only of Border Gateway Protocol, Version 4 (BGP-4), involve multiple administrative organizations.

Where the basic unit of an IGP domain is an administrative entity such as an enterprise, the core network administrator of a service provider, etc., the basic unit of BGP routing is the Autonomous System (AS).

An AS is a set of routers and prefixes under the control of one or more administrative entities, which present a common routing policy to the Internet.

Most modern IGPs (OSPF, ISIS, EIGRP) have varying levels of internal hierarchy. Typically, there are two levels based on aggregating addresses, although clever configuration can, in effect, introduce more levels of hierarchy based on topology, static routing, constraint characteristics, etc.

2.3 How Inter-domain Routing Works

The Internet consists of many ASes (Autonomous System) interconnected to each other. Each runs one or more IGP domains to maintain a stable view of the domain topology. When there are multiple domains, they may be interconnected with BGP used within the AS. Each AS uses BGP to maintain adjacency with other ASes in order to maintain connectivity through the whole Internet.

Internet-wide connectivity is described in BGP routes received from each peer border router. These routes are conceptually stored in the Adj-RIB-In (Adjacent Routing Information Base Input). These routes go through per-peer policy processing, and the remaining routes go through the BGP route selection process. These make up the Loc-RIB (Local Routing Information Base) of BGP. Each entry in the Loc-RIB defines route properties for a specific prefix, such as its next hop address, the AS path, local preference, etc.

Since there can be several routes for a specific prefix, there is a route selection process to determine the active path, which will be installed in the main RIB. Other sources of routing information, such as directly connected hardware and information from IGPs, may be compared with the routes in the Loc-RIB to determine the final active route. The presence of routes from all sources in the RIB is essential to the BGP route selection process. The software design of most commercial routers optimizes the RIB data structuring for efficient updating.

In contrast to the RIB design goals, commercial high-performance routers have one or more separate Forwarding Information Bases (FIB), which contain elements of the main RIB, but in data structures organized for optimal high-speed route lookup. The FIB controls the forwarding behavior of the router. This route selection process is constrained by local domain policies. Through this process a router's forwarding behavior can be derived from monitoring its routing traffic, together with its local policy constrains.

The BGP uses AS as the basic element in route computation. Through the AS path attribute for each route, the router is able to partially build up the AS topology for each specific prefix. However, the topology built by the router may be based on outdated information and may not be correct for the reason that it has been individually processed by each of the ASes on the AS path, without a mechanism to verify correctness. Finer level of topology of the Internet on a link-by-link basis is totally unknown to BGP. Even the next hop attribute of the route is not the direct next interface and may need to be further resolved by an IGP to find the direct next interface so that the route can be installed in the forwarding table.

The BGP traffic exchange also indicates the policy of each domain. If an AS advertises a route to a neighboring AS, it means this AS is willing to accept traffic that is destined to the prefix advertised by that route. If the AS does not originate that prefix, it means this AS is willing to transit traffic for that prefix. When receiving a route, the receiving AS can decide if it will select the sending AS to transit traffic for the announced prefix. Different preference values can be configured to discriminate routes entering through different neighboring ASes, or through different border routers of the same neighboring AS.

Filtering is a mechanism for an AS to implement its policy with its peers and its customers. It uses filters to ensure the right routes have been announced and accepted.

RFC2827 [5] defines the Best Current Practice for ingress filtering. Essentially, it causes an

AS to refuse packets with a source address not assigned to the peer AS.

Classical packet filtering, where the source address is matched against a series of accept/reject rules, has serious problems of scalability when dealing with situations where very large numbers of prefixes are legitimate. Such filtering may be practical on interfaces between customers and edge ISPs, but does not scale to the Internet core.

One workaround to the limits of classic filtering is that major providers trust one another to do universal ingress filtering. A more scalable approach, however, is called reverse path verification. In this method, both the source and destination addresses of packets are looked up in the FIB. If the router has no route to the source address, the packet is assumed to have a forged source address and is dropped.

In addition to the specific validation of prefixes, providers increasingly apply sanity checking to the number of routes received on a particular BGP connection. It is implausible, for example, that an enterprise customer would legitimately advertise thousands of routes.

3 Methodology

Our experiments involved establishing sources of raw data, developing means of inspecting and reducing the data, and then analyzing the data to produce meaningful information on BGP behavior.

3.1 Data Source

Measurement points are established to collect BGP protocol traffic. This means raw session set-up and individual updates. Measurement points are selected at several points at the border of AS 1299 and AS 3301 of Telia's networks in Sweden, and the border of AS1653 at SUNET (the Swedish University Network). The measurement devices passively log all BGP messages announced from the peers.

3.2 Data Inspection

The first efforts involve the creation of scripts for extraction and preparation of the data in the BGP update logs.

Once the data has been extracted, it is subjected to simple statistical inspection. That is, the frequency, means and standard deviations of various values are calculated and inspected for patterns and anomalies. Visual inspection of the data is also done using assorted graphical techniques. In order to refine the methods of inspection, scripts were developed and will evolve as the program continues.

3.3 Continuing Data Analysis

As the study continues, the data will be subjected to correlational analysis and various multivariate statistical techniques to form hypotheses and determine their suitability.

Tools were developed to process the huge amount of data collected from various measurement points. Results are extracted from the raw data and stored into the database.

4 Measurement Results

Our measurements produced statistics on the load produced by various kinds of routing messages, on the non-local effects of errors and route flapping, on AS and prefix reachability, and on convergence time.

4.1 Routing Messages

2 kinds of BGP messages have been passively logged, the BGP announcement and the BGP withdrawal message. An announcement message carries the next hop address and AS path information for a specific prefix. Receiving it means the route for this prefix is up. A withdrawal message shows that a route for a specific prefix is down. Whenever a BGP message is received, the BGP peer needs to process it and update its RIB appropriately. It is our hypothesis that the load of the routing messages that arrive at a point in the network is a good indication of the general status of the network stability. If a lot of messages are received, it indicates the network is rather instable. It also reflects how much the BGP process has been loaded.

The graph below is a plot of the total number of BGP announcement and withdrawal messages per hour for over one month's period from an EBGP peer at the border of AS 1299.

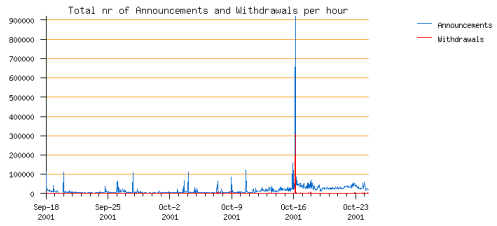

Figure 1 BGP Message Load

Many spikes are observed in the graph. The most pronounced one is on October 16. Most of these spikes are because an ISP's networks are flapping during that period. On investigation it was found that the reason for the spike on Oct.16 is due to the fact that the EBGP peer had been up and down many times on that day, and thus had resulted in announcements and withdrawals of the entire routing tables to its peers.

On average the BGP routing process receives a few hundred messages every 30 seconds for each peering session. 30 seconds is the default value for the MinRouteAdvertisementInterval timer in the Cisco router, and routers send messages every time the timer expires.

4.1.1 Duplicate Announcements

If an BGP announcement message has exactly the same route attributes for a specific prefix as a previous announcement, such as the next hop and the AS path attributes, and this BGP announcement comes from the same session, then the message is considered a duplicate. From the messages received at different probes, there was a discovery that about 40% of all BGP announcements are duplicates.

One explanation offered for this high ratio is that a BGP peer will send duplicates to synchronize the MinRouteAdvertisementInterval timer whenever there are no updates for a time interval. However, given the fact that within every 30 seconds there are hundreds of messages received, there does not seem to be a need to send duplicates for this reason.

Another, we believe more credible explanation is due to the use of confederation in some ISPs or large enterprise networks. When the BGP peer chooses a different internal confederation area path for a specific route, this means that the change in the AS path occurs at the private ASs part, and when the ASs border router announces this update to its EBGP peers, the border router will drop the part of the private AS path, which are internal confederated ASes. This message will then appear externally as a duplicate BGP message. It is suspected that in the current BGP implementations the BGP process just replaces the existing BGP message with the received one, without comparing them to see if there is really any change. And if this route is selected as the active one, the announcement will be further propagated to other peers, further propagating duplicate BGP updates.

Vendors have explained that it is very expensive to store outbound BGP messages. However since at least inbound BGP messages are stored, and if the BGP implementation can have a simple verification process to see if there is really some change whenever it receives an update, this can prevent the further propagation of duplicate BGP updates to the whole Internet, thus greatly reducing the number of duplicate BGP messages circulating in the Internet.

4.1.2 Flapping Announcements

A route flap is defined as the rapid withdrawal and announcement of a route. A route flap is not a problem until a route is flapped several times in close succession. This causes negative repercussions throughout the Internet. The explicit withdrawals show prefixes being up and down. The implicit withdrawals show prefixes choosing between the best and the backup paths.

Here is an example of a flapping /24 network.

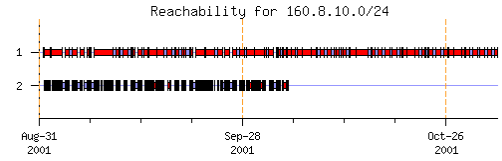

Figure 5 A Flapping Prefix

As shown in the figure, there are 2 AS paths for the prefix. AS Path 1 is the preferred one compared with AS Path 2, and AS Path 2 is the backup path. At any point in time only one of these two can be the active path. The figure shows that the active path has been jumping back and forth between these 2 AS Paths. The reason is that AS Path 1 has been flapping quite a bit. It is constantly withdrawn and then announced. Whenever AS Path 1 is withdrawn, AS Path 2 becomes the active path. And whenever AS Path 1 is announced as being up, it becomes the active path again. If the flapping rate of AS Path 1 had been taken into

consideration in the route selection process, AS Path 2 would have been selected as the active path even when AS Path 1 was announced as being up, and this could result in a much more stable route.

Below is a figure plotting the number of times that a prefix has been announced and withdrawn. We can see that those spikes are due to routes being announced and withdrawn tens of thousands of times within a single month.

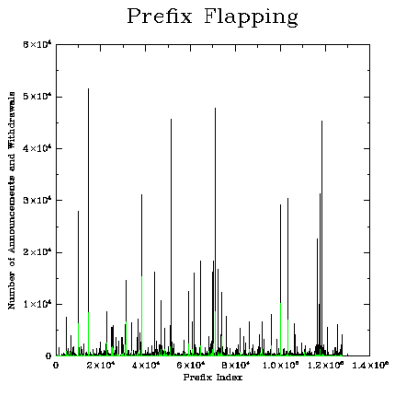

Figure 6 Route Flapping

Flapping can be seen here as either prefixes being up and down, or prefixes being implicitly withdrawn. An implicit withdrawal means a prefix chooses a different route. If a prefix has been implicitly withdrawn many times, it means that it flaps between the best path and the alternative path. The use of flap damping is encouraged. This can greatly reduce the number of BGP messages announced and stabilize some backup routes. Flap damping means that the local policy takes into account the stability of each route and the backup route becomes preferable because it is more stable.

4.1.3. Erroneous Announcement

Using private AS numbers has many legitimate applications, such as confederations and multihoming. When these applications are properly implemented, the private AS number should stay local to the registered AS number.

Unfortunately, improper implementation not infrequently causes private AS numbers to be announced into the public Internet. This is also observed at the measurement points. This is for the reason that some networks use private AS numbers to run BGP to connect to their transit providers, and these transit providers don't strip off the private AS numbers when announcing them into the Internet. This can create great confusion for networks using confederation, since confederation also uses private AS numbers and can thus mislead routers to interpret the AS path as a forwarding loop.

Approximately 4% of prefixes announced into the Internet are up less than 20% of the time. As mentioned in the AS Reachability section, these prefixes are usually short-lived, only appearing for a very short period. Some of them come from unused address space. Some of them are intentionally announced by parties not owning that address space. This can lead to traffic destined to those address spaces being mis-routed to other places. This problem is hard to detect today, except for when the owner of the address space finds their networks cannot be reached by some parts of the world. If the DNS root servers and some other golden addresses are ill-announced, this can lead to disastrous effect on the functioning of the Internet.

Why are there so many short-lived illegal prefix announcements? This seems to indicate that the current AS-based filtering mechanism between many peering providers does not look into the prefix level due to the huge amount of prefixes exchanged. These peering parties trust each other when announcing prefixes to each other. This makes it vulnerable for announcement of unused prefixes and private AS origin, and stealing others' address space. Since AS-based filtering is used, BGP doesn't look into the prefixes level. This means that peers cannot prevent each other from sending erroneous prefix announcements.

4.2 Non-locality of Effects of Instability and Misconfiguration

There have been a number of instances of a mistake in BGP configuration in a single peripheral AS propagating across the whole Internet and resulting in misrouting of most of the traffic in the Internet.

Similarly, route flap in a single peripheral AS can require route table recalculation across the entire Internet. This has been observed from all the measurement points. From the extreme spike of the Figure 1, the BGP Message Load, there was suddenly a huge amount of

announcement and withdrawal messages received. This is due to the reason that a customer's EBGP peer of the measured provider's AS was up and down many times and this effect was spread into the provider's whole domain, affecting all BGP routers, and very likely further spread into other Ases, unless of course route flap damping had been properly configured.

This non-locality of effects is highly undesirable, and it would be a considerable improvement if such effects were naturally limited to a small area of the network around the problem.

4.3 Reachability

BGP is intended to maintain Internet-wide network connectivity. Since a router relies on the BGP control traffic to establish its routing table, it is, theoretically, possible to know the reachability of any network in the Internet from the measurement point through logging the exchanged BGP messages.

4.3.1 AS reachability

An AS is defined as being up from the first time there is a BGP announcement with that AS as the origin AS. An AS is defined as being completely down when all the prefixes originating from the AS have been withdrawn. Measurement found about 8% of the ASes had been completely down for some time during a period of one month.

From the total number of BGP announcements and withdrawals for prefixes originated from each AS, it's possible to know the stability status for each AS. Below is a graph to plot the withdrawal messages for each AS. As shown here, some ASes have extremely high amount of withdrawals and thus a low degree of stability.

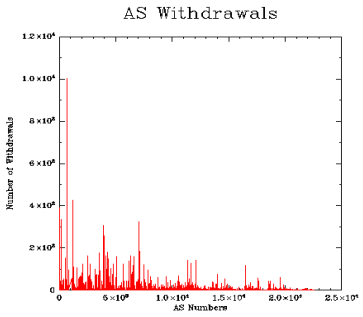

Figure 2 AS Withdrawal Messages Count

A lot of these ASes have an uptime below 20%. The reasons are that these AS numbers rarely appear in the whole period. There is only a short period, such as a few hours on a specific day, when some short-lived prefixes are announced and then withdrawn. These short-lived prefixes are usually from the unallocated address space, or with a very short prefix length.

Here is a typical unstable AS. It's a network with 8 /24 prefixes. 10425 announcement and 1871 withdrawal messages have been received with prefixes originating from this AS, and it has been completely down for 251 times within a period of 1 month.

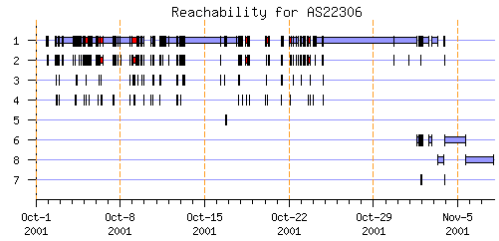

Figure 3 An AS's Reachability

4.3.2 Prefix Reachability

A prefix is up when there is a BGP announcement message received. A prefix is down when there is a withdrawal message arrived. The time between an announcement message and a withdrawal message is the time that prefix is reachable. And the time between a withdrawal message and an announcement message is the time that prefix is unreachable. Here is a plot of the reachability for all prefixes observed in the measurement points.

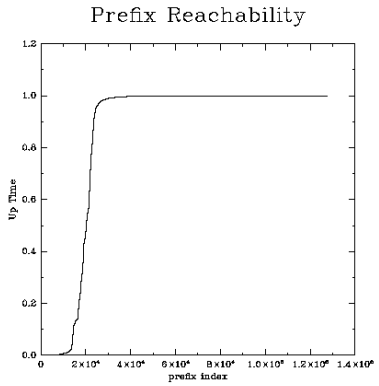

Figure 4 Prefix Reachability

As shown in the figure, 13% of the total prefixes have a low reachability of 20%. These are usually due to erroneous announcements. 5% of the total prefixes have a reachability between 20% to 90%. These are prefixes that usually flap a lot. Even a 90% reachability is not a very good degree of reachability.

Prefix reachability doesn't necessarily mean network reachability, especially for the reason that a specific prefix can be contained within an aggregated route. However, an aggregated route is usually statically configured in the origin AS, and will never be withdrawn from the origin AS even though the more specific prefixes are no longer reachable in the origin AS. In this sense, the reachability of a specific prefix more realistically represents the reachability of the network than an aggregated one.

4.4 Convergence

Within the context of BGP, there are convergence times with three different scopes: internet-wide, single AS, and single router. There has been extensive research done already concerning internet-wide convergence [2]. From the measurement points it's easy to discover that the same issues are apparent at different measurement points within a time span of a few minutes.

The IETF Benchmarking Working Group (BMWG) is working on the single-router convergence time issue [6][7]. In addition to defining terminology and methodology, experiments by some of the coauthors have shown that such things as the sorting of updates by prefix length, the number of prefixes per update, etc., can have significant effects on single-router convergence. Effects seen in single routers can reasonably be expected to propagate to the larger scopes of the AS and the Internet.

5 Conclusion

The Internet is a meshed structure with many ASes interconnected to each other. BGP as a path-vector protocol makes the instability at any point of the Internet an accumulated result of the whole Internet, multiplied by the AS-topology and different policies. This means that the instability at any point of the Internet will not only be spread into your network, but will do as as a multiplied result of all the networks from the origin of the instability to your network.

The measurements done here offer some glimpses into the Inter-domain routing stability. Though people have talked a lot about network stability, there is no consensus about how to measure it and what to measure. Logging the BGP control traffic is one good way of measurement since a router uses these messages to establish its routing tables and further determines how data traffic will be forwarded. Actually Inter-domain routing traffic is a practical approach through which a perception of the stability of the whole Internet can be achieved. A lot of efforts have been focused on measuring the forwarding traffic itself. Due to the forwarding traffic's huge volume it's difficult to cover the whole Internet though progress is being made in that direction. Through studying the control plane effects it's possible to study if there is any correlation with the data plane effects. If there is, then it will be possible to predict instability and possibly take corrective action in a more dynamic manner.

A BGP logging approach has been used and described here. Internet stability has been studied from the perspective of the message arrival process, the network up time, and the network flapping rate. Some discoveries have been made, such as the large number of duplicate announcements, the propagating effect of network flapping, and the erroneous announcements. All these are good indicators of network instability. It is also worth noting that the current BGP protocol is vulnerable to all these kinds of misbehaviors.

Though flap damping can greatly reduce the flap rate, operators are conservative in using it since they want to reach as many networks as possible. However, in the presence of backup path, the current BGP decision algorithm should take the stability of a specific route into account in its route selection process. This can lead to more stable routes, and a more stable Internet.

The measurement results shown here are mainly the results of inspection and some initial analysis results. More analysis results will come in the area of convergence time, correlation between data plane and control plane, and finding more indicators of network instabilities.

6 Acknowledgement

The measurement done here is the result from the Babylon project, a collaboration between Telia Research AB, Nortel Networks, Utfors, Royal Institute of Technology, and Luleå University of Technology. Daniel Olsson and Fredrik Pettersson were affiliated with Telia Research AB when they participated in this project. The authors would like to thank Magnus Larsson, Anders Bergsten, Thomas Eriksson, Olle Pers, Howard Berkowitz and many others in the project for their valuable comments and contribution.

References

[1] Geoff Huston, Commentary on Inter-Domain Routing in the Internet, RFC 3221, December 2001

[2] Craig Labovitz, et al, Delayed Internet Routing Convergence, Sigcomm2000

[3] Danny McPherson, et al, BGP Persistent Route Oscillation Condition, draft-ietf-idr-route-oscillation-01.txt, work in progress, February 2002

[4] Avri Doria, et al, Future Domain Routing Requirements, draft-irtf-routing-reqs-groupb-00.txt, work in progress, February 2002

[5] P. Ferguson, D. Senie, Network Ingress Filtering: Defeating Denial of Service Attacks which employ IP Source Address Spoofing, RFC 2827, May 2000.

[6] Howard Berkowitz, et al, Benchmarking Methodology for Basic BGP Device Convergence, draft-ietf-bmwg-bgpbas-01.txt, work in progress, February 2002

[7] Howard Berkowitz, et al, Terminology for Benchmarking External Routing Convergence Measurements, draft-ietf-bmwg-conterm-01.txt, work in progress, February 2002