February 2014
Geoff Huston

BGP in 2013 – The Churn Report

Last month, in January 2014, I reported on the size of the Internet's inter-domain routing table, and looked at some projection models for the size of the default-free zone in the coming years. At present these projections are looking at relatively modest levels of growth of some 7 – 8% per year with IPv4. Although IPv6 is growing at a faster rate, doubling in size every two years, its relatively modest size of 1/30th of the size of the IPv4 routing table does not give cause for concern at the moment. But size of not the only metric of the scale of the routing space – it's also what BGP does with this information that matters. As the routing table increases in size do we see a corresponding increase in the number of updates generated by BGP as it attempts to converge? What can we see when we look a the profile of dynamic updates within BGP, and can we make some projections here about the likely future for BGP?

BGP is a distance vector routing protocol. This family of routing protocols operates through an iterative process where every BGP speaker informs all neighbouring BGP speakers of its selected best path to a destination. When a BGP speaker obtains an update from its neighbor, it compares the updated information with its current selected best path, and if this information causes the local BGP instance to select a new best path, then it informs its neighbours of this new choice. BGP, like all distance vector routing protocols, can be a very chatty protocol, and the larger the population of BGP speakers, and the denser the level of inter-connectivity between BGP speakers, the greater the potential amount of protocol updates that will occur across the network before it converges to a coherent common state. And of course as the Internet grows, this precisely what is happening to BGP. More Autonomous Systems are being added to the routing system, and the level of interconnectedness continued to rise, as evidenced by the relatively stable position of the average AS path length over time.

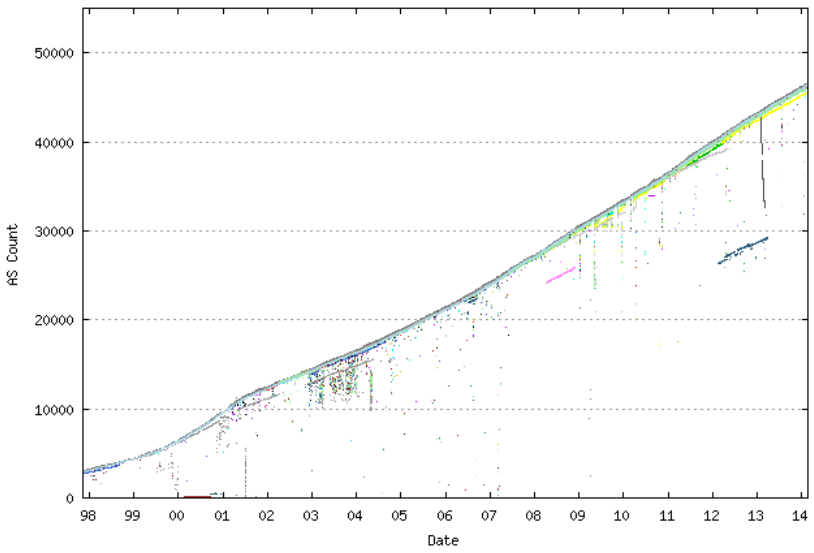

Figure 1 – Number of AS's seen in the BGP routing table by peers of Route Views

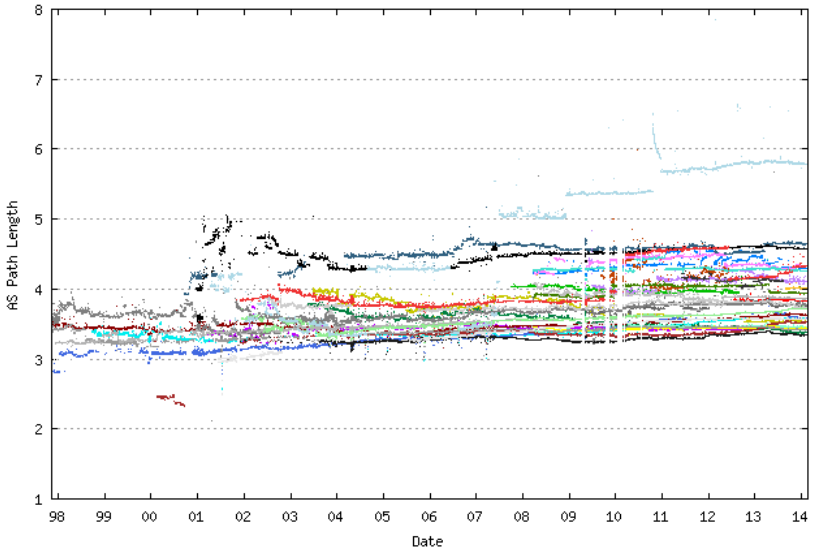

Figure 2 – Average AS Path Length as seen by peers of Route Views

The BGP protocol has two mechanisms than can mitigate the protocol update issue. The first is the use of the AS Path vector, which allows a BGP speaker to detect and discard updates that reflect a routing loop. If a BGP speaker sees its own AS number in an announcement from a BGP neighbour, it will discard that update. The second is the use of the MRAI timer, which requires BGP speakers to wait for an MRAI interval before sending a routing update to the same neighbour about an update to the same route object. When supported, the commonly used value for this timer is a randomly selected value between 27 and 30 seconds. This implies that when a BGP speaker sees a burst of updates from its neighbours within a 30 second interval, it will absorb this burst and send a single update at the expiration of the MRAI timer, dampening the tendency of the protocol to act as an update amplifier. The cost of the second measure is that the protocol can be far slower to converge to a stable state, and some vendor implementations of BGP turn off the MRAI timer by default. However, the observational evidence is that MRAI timers are very widely used in the inter-domain environment, and this remains a major factor in damping protocol updates.

What we are left with is a protocol that has a tendency to become very chatty as the network grows, but is equipped with some mechanisms that can damp this form of behavior, and its requirement to route across an Internet that appears to be growing inexorably. To what extent are these mitigating mechanisms able to contain the dynamic behavior of BGP? Is BGP still able to route the Internet, or are we approaching areas of increasing vulnerability to some form of protocol overload?

BGP Updates

Each BGP protocol message contains an update section, to announce a "new" path to a destination, and a withdrawal section, to list those destinations no longer reachable. In practice BGP speakers use one of the other update forms, and within each protocol transaction either announce or withdraw a set of address prefixes. However its not really the number of protocol transactions that are the essential metric of protocol load – it's the number of prefixes that are being updated and withdrawn within these transactions that reflect the level of work that a BGP speaker must perform to keeps its local view of the routing space consistent with its BGP neighbours.

If one were to assume that routing noise was equally likely in any part of the internet, then we could model the root cause of routing instability as a probability function that was equally likely to occur within any Autonomous System (AS). The inference is that if this instability probability remains constant, then, as the AS population increases, then the number of BGP updates should increase.

IPv4 Routing Updates

The following figure shows the daily count of the number of prefix update and withdrawals per day since mid 2007. The measurement is made as AS131072, a measurement AS located at the edge of the network that contains a single eBGP speaker.. The figure also shows the daily size of the BGP routing table., and its clear that the level of dynamic activity in BGP is not growing at the same rate as the total number of objects contained in the routing system. Indeed, it seems that the number of updates has been relatively steady over this entire 6 ½ year period.

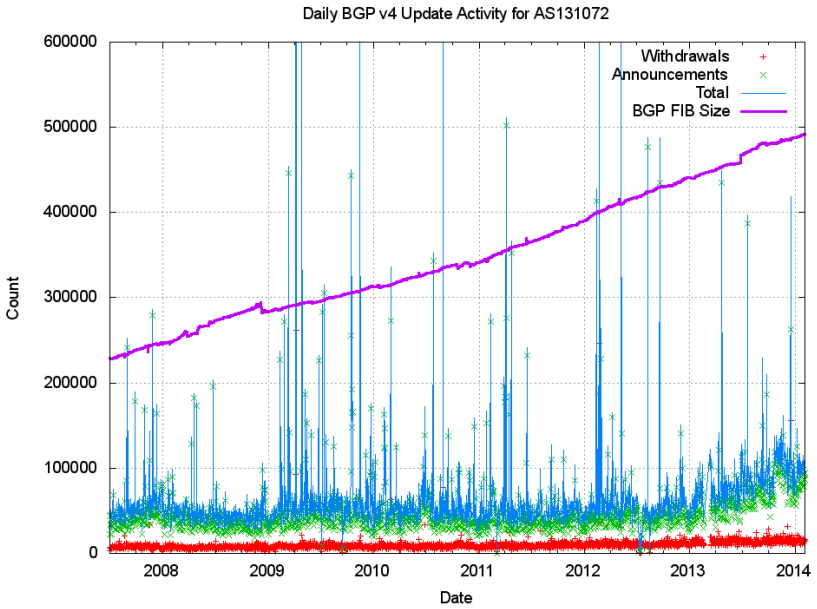

Figure 3 – Daily Total of Prefix Updates as seen as AS131072

This is not an outcome that one would normally expect from a distance vector protocol operating across a continually growing routing space. It calls into question the assumption that routing "noise" is equally likely in any AS. We can drill into this a little further by looking not at the number of updates seen, but at the number of prefixes that were the subject of BGP updates. If the model of routing instability is one that relies on an even distribution of probability of routing instability then we should see this metric rise in proportion to the number of prefixes contained in the routing table. As shown in Figure 4, this is not the case. On those days where the entire table has not been updated, the number of unstable prefixes in BGP has remained relatively constant, and while there is some slow upward trend in the data, the model of growth of this metric is, at best, a linear model of growth

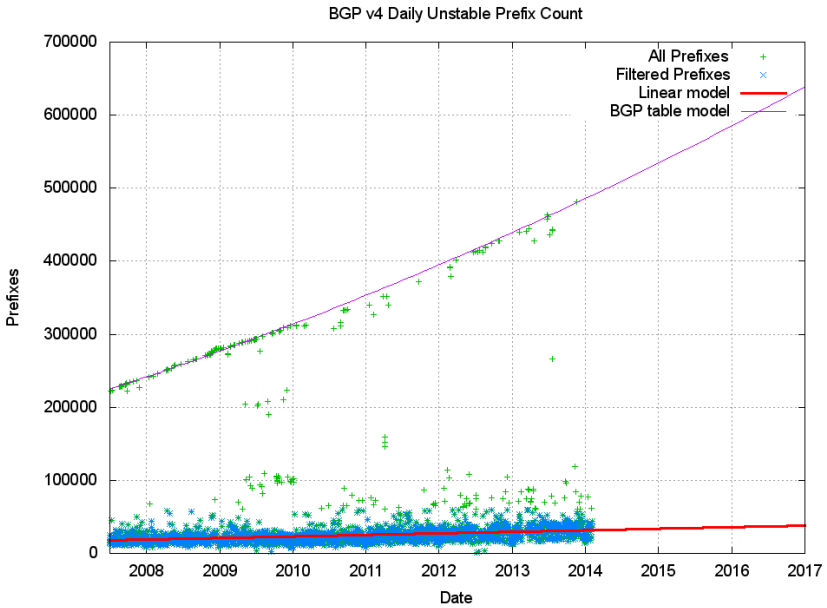

Figure 4 – Daily Total of Updated Prefixes, as seen as AS131072

It is also evident that the increase in updates in 2013, shown in Figure 3, is not the result in some underlying change in the model of routing instability in the Internet, but is more likely a result of the addition of a new upstream provider to the observation AS, and this additional transit AS is adding some further component of update activity for each basic root cause event.

So if the number of unstable prefixes per day is relatively constant, what about the number of updates required for an instability event to reach a converged state? Is BGP getting any chattier as the network grows in size? Figure 5 shows the daily average of the number of updated received for each instability event. The average number of updates has remained highly stable at 2 since 2008. If we filter out all sequences of 1 of 2 updates, the average of the remainder is a stable value of 2.4. BGP is not getting any chattier as the network grows.

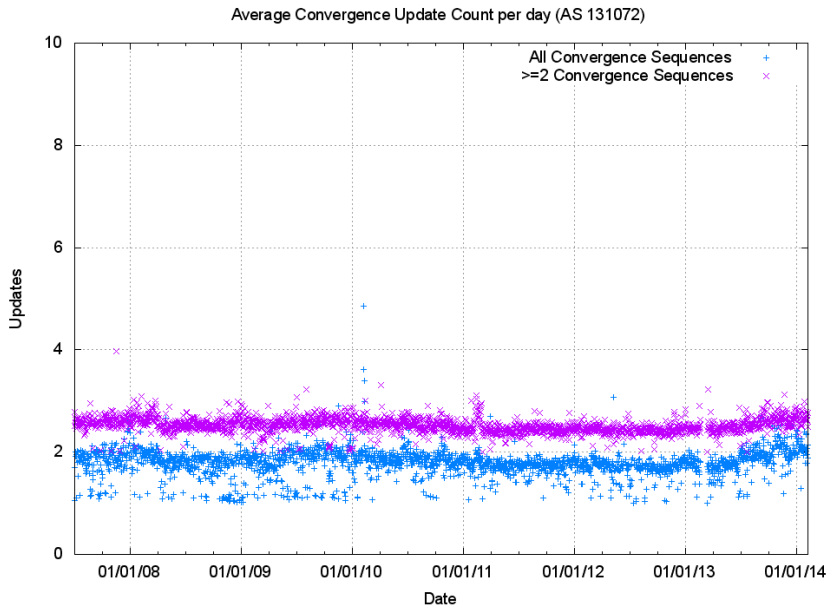

Figure 5 – Average Update count of a Convergence Sequence, as seen as AS131072

Why are we seeing this long term stability in the protocol convergence profile for BGP? What aspect of the Internet's inter-AS topology, coupled with the behavior of BGP, could cause this? The answer lies in part on the widespread use of the MRAI interval, which effectively damps out the propagation of path hunting behavior following a withdrawal at source. The second part of the answer lies in the relatively constant "diameter" of the inter-AS space. Because path hunting behavior tends to produce an exploration of possible paths of AS path length n, then n+1, n-2, and so on, then if the AS Path length is bounded in size, then the path hunting behavior is also bounded. And we have seen an extraordinarily stable average AS Path length in the internet for the past 15 year. Figure 6 shows the average AS path length as provided by each of the peers of Route Views since 1998. It tends to suggest that as the Internet grows, new AS connections tend to connect into the "core" of the Internet, rather than attach at the "edge" of the network, so that the growth can be expressed as an increase in the connectivity "density" in the inner transit parts of the network.

The conclusion from these observations is that the "amplification" factor of BGP updates has not played a significant role in inflating the workload of BGP over time. Each instability event generates some 2.4 updates on average, and as the spacing of updates is that of the MRAI interval, of 27 – 30 seconds, the average time for each routing convergence event is some 70 seconds.

This leaves one outstanding question, however. What lies behind the data presented in Figure 4? Why is the number of unstable prefixes growing at a rate far smaller that the number of prefixes in the routing table? There is no immediately evident answer to this question, so we need to look around in a little

more details to see if there are any clues as to what is happening in the network. One approach to try and understand this is to look at the routing update behavior in IPv6, so see to what extent that far smaller IPv6 network mimics the behavior of the IPv4 network.

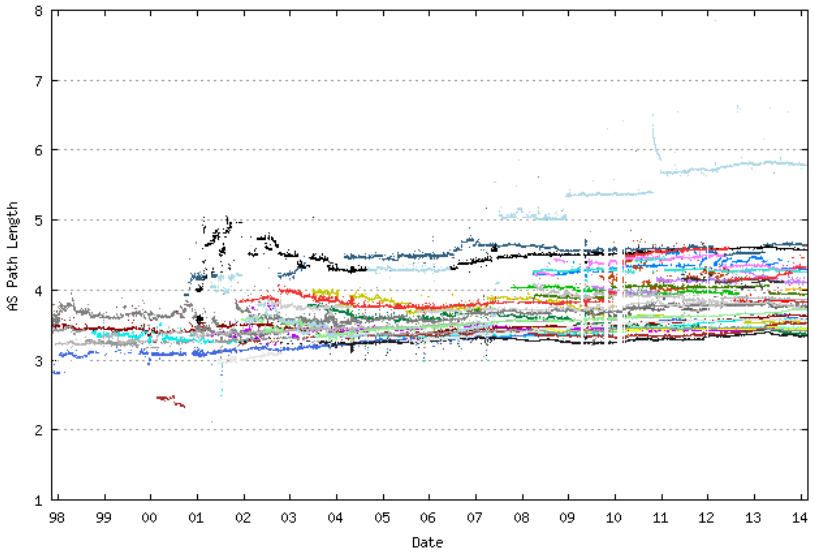

Figure 6 – Average AS Path Length as seen by Route Views Peers

IPv6 Routing Updates

The update profile for IPv6 since 2008, as shown in Figure 7, shows three distinctly different phases.

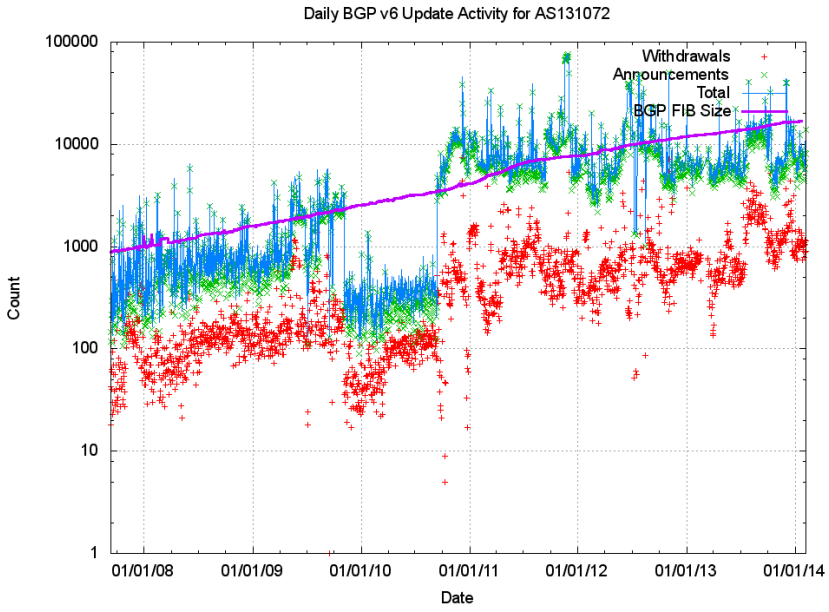

Figure 7 – Update Profile for IPv6

In the period 2007 to 2010, the number of updates (announcements and withdrawals) grew at a rate that approximately tracked the size of the routing table. The local connectivity at AS131072 was altered in 2010, and the upstream BGP speaker was applying a very strict damping profile to IPv6 updates. With the connection of further IPv6 transit providers to our observation AS in the third quarter of 2010, the update profile returned to what appears to be a conventional profile. However, there is no further growth in the update volume, and since late 2010 this profile is similar to IPv4. The number of prefix updates appears to be steady at some 10,000 updates per day, and some 1,000 withdrawals per

day, while over the same period the size of the IPv6 routing table has risen from 3,000 to 16,000 entries.

If we look at the number of unstable prefixes on each day we see a similar outcome (Figure 8). The count of updated prefixes is growing at a far smaller rate than the growth in the size of the routing table.

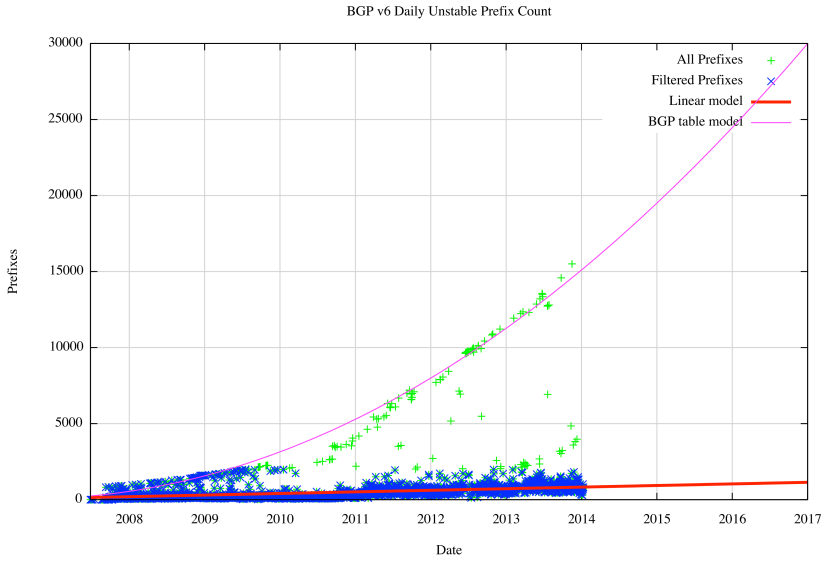

Figure 8 – Updated Prefixes per day for IPv6

What is perhaps not so apparent from this figure is that since the start of 2011 the number of unstable prefixes per day has risen from 1,200 per day to 1,500 per day while the overall IPv6 BGP table size has risen from 4,000 to 16,000 entries. As with the IPv4 routing system, the model of instability in IPv6 is not one of a uniform distribution of probability of instability. The time to reach convergence is equally bounded in IPv6, corresponding to an average update count of some 3 updates per convergence sequence, as shown in Figure 9.

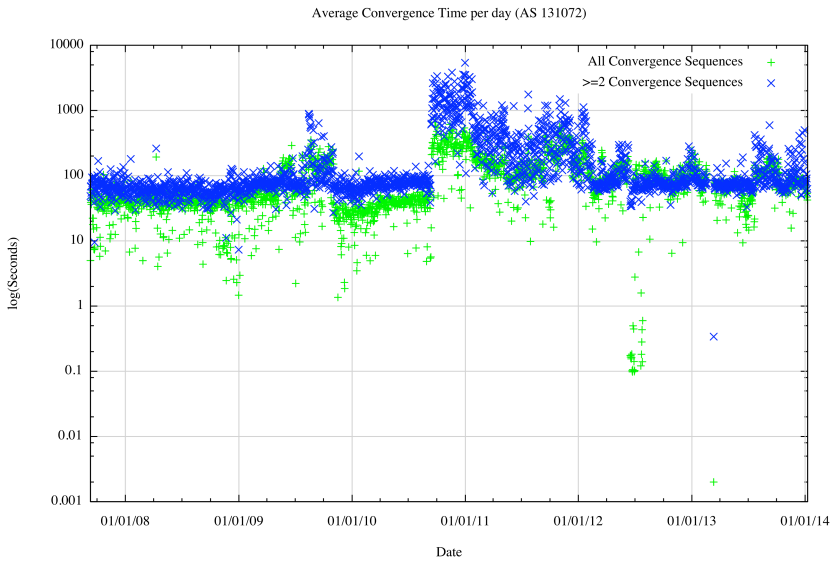

Figure 9 – Average time to reach convergence per day for IPv6

This is slightly higher than the comparable figure for IPv4. A possible reason for this may be found in the comparison of the average path length of the observation AS, AS131072, with that seen from the various IPv6 peers of Route Views. Figure 10 shows the solid blue line, of the observation AS sees an

average AS Path Length around 1 AS longer than most other Route Views peers. Its possible that the further one is located away from the "core" of the network, the greater the amount of routing traffic that is associated with routing convergence.

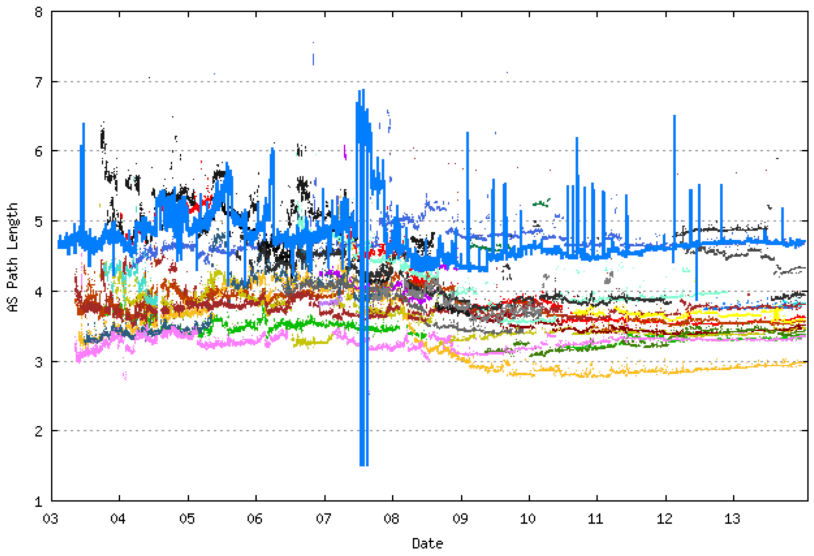

Figure 10 – Average AS Path Length of Routing Views IPv6 Peers

More Specifics and Route Stability

In trying to gain a better understanding of what form of prefixes contribute to routing instability, it may be useful to compare the routing behaviour of aggregate routes and more specifics. Intuitively one might expect that more specifics used to engineer traffic loads along certain transit paths may be adjusted more frequently than aggregate routes. Can we see this form of behaviour in the profile of routing updates for 2013? In this section we will concentrate on the profile of BGP announcements in the IPv4 Internet.

Figure 11 shows the last decade of the routing table history, showing the number of more specifics as well as the total table size. It is clear that the two time series are closely tracking each other. When we re-plot this to show the more specifics as a fraction of the total table size (Figure 12) we see the somewhat unexpected result that the number of more specifics in the IPv4 Internet has remained at a very stable 50 % of the total table size for the past decade.

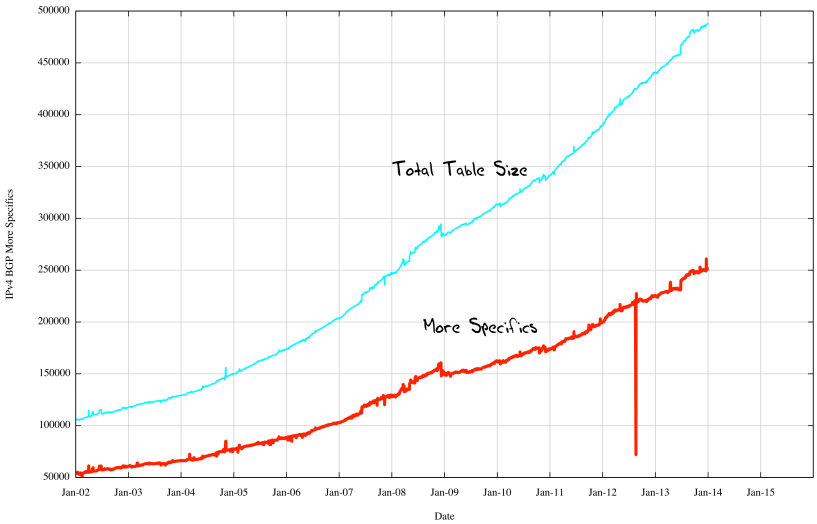

Figure 11 – IPv4 More Specifics

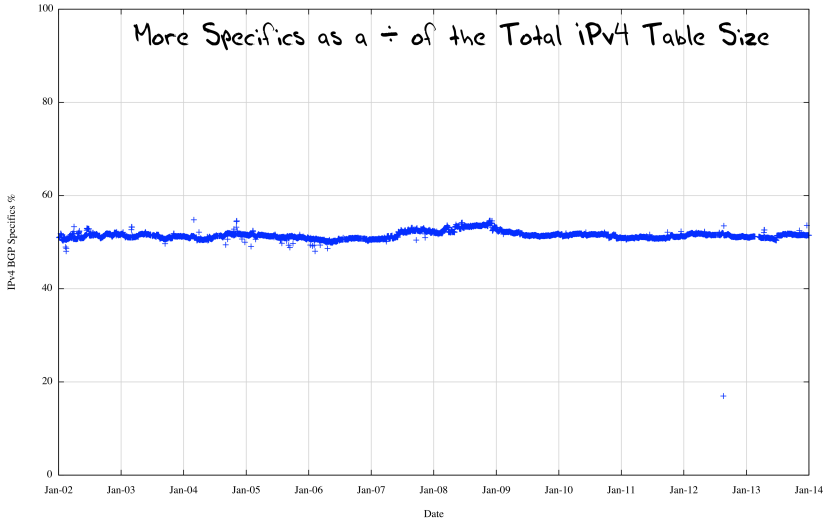

Figure 12 – IPv4 More Specifics

It is interesting to look at the profile of more specifics. Not every origin AS announces more specifics. Indeed some 55% of the 47,000 Origin AS's announce no more specifics at all, and some 458 Origin AS (some 1% of the total number of AS's) announce 133,688 more specifics, or some 54% of the total number of more specifics. It appears that the distribution of who announces more specifics in the IPv4 network is highly skewed, and a cumulative distribution plot bears this out (Figure 13)

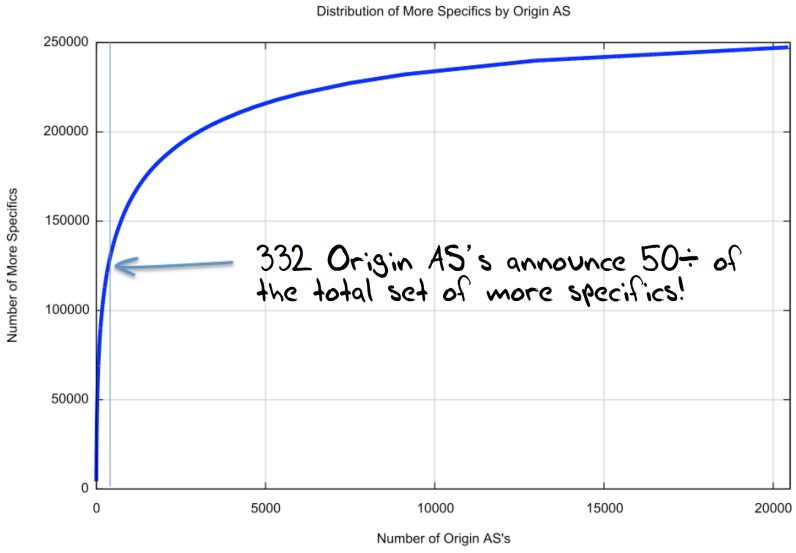

Figure 13 – Cumulative Distribution of IPv4 More Specifics

What can we say about more specific advertisements and BGP updates. Do these more specifics, which make up 50% of the IPv4 routing table, represent 50% of the prefix updates? Or is the proportion higher or lower? In other words, are more specifics more or less stable than aggregate announcements?

Perhaps its useful to look at the stability of this set of more specifics to see of there is any change in the profile of with Origin AS's announce more specifics over the past three years. The following figure (Figure 14) shows the day-by-day record of those Origin ASes who have been in the daily top 10 of advertisers of more specifics, and tracks their record of advertising more specifics over this period. There are a number of different behaviours, ranging from a steady state of the same number of more specifics announced over the period, a profile of rising or falling gradually over time, and some abrupt step changes. The set of more specifics appears to be one that is constantly changing over time. So while the total number of more specifics appears to accurately track a metric of 50% of the total number of entries in the routing table, the individual components of more specifics announced per

origin AS, particularly for those AS's with the highest number of announced more specifics, shows a much higher degree of variability.

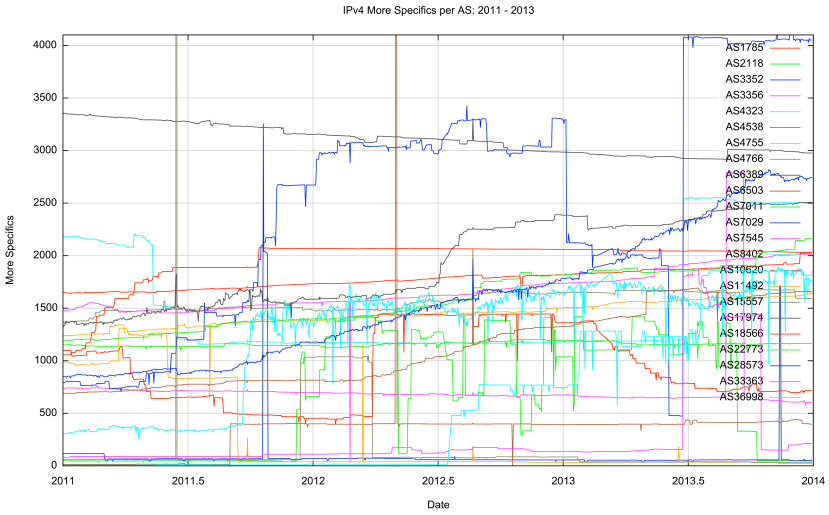

Figure 14 – More Specifics announced per Origin AS 2011 - 2014

We can use a simple taxonomy to categorise these more specifics, based on the relationship of the AS Path of the more specific to its aggregate. One such taxonomy uses three categories:

- The more specific has exactly the same AS Path as its covering aggregate.
- The more specific has the same Origin AS, but a different AS Path, which is strongly suggestive of a traffic engineering advertisement.
- The more specific has a different Origin AS and a different AS Path, which is suggestive of a form of "hole punching" where the more specific is used by a different entity who has a distinct and different routing policy than the aggregate.

Figure 15 shows the relative proportion of these three types of more specific over the past three years.

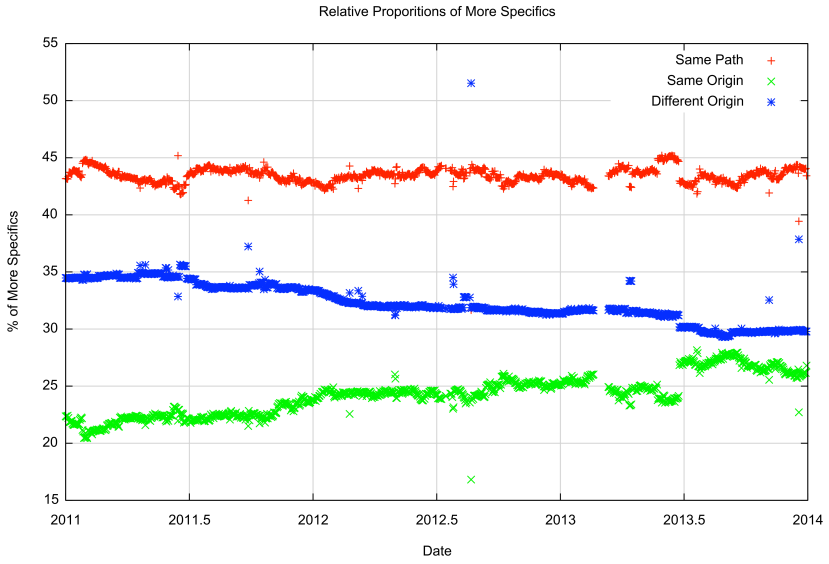

Figure 15 – Relative Proportion of More Specifics categories 2011 - 2014

The relative number of same path more specifics has remained steady at some 45% of the total number of more specifics throughout this three year period, while the relative proportion of traffic engineering prefixes (same Origin AS) has risen from 21% to 26%, with a corresponding fall in the number of "hole punching" (different Origin AS) more specifics.

Are more specifics noisier than aggregates? Figure 16 shows the relative proportion of updates that relate to more specifics and aggregate prefixes. While some 50% of the total population of routes prefixes are more specifics, they constitute some 80% of the total number of BGP updates on any day. It would appear that more specifics are some 4 times noisier than aggregates from the perspective of routing updates.

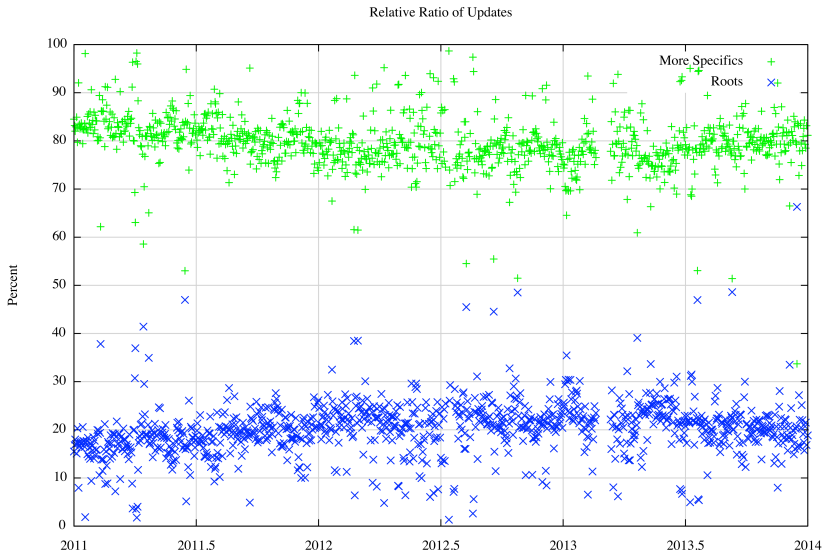

Figure 16 – Relative Proportion of Updates between More Specifics and Aggregates: 2011 - 2014

This figure appears to support an intuitive assumption that more specifics are less stable than aggregates. An assumption here is that more specifics are used to refine an existing prefix advertisement and this refinement may be temporary, based on requirements associated with routing policies and traffic engineering as distinct from basic reachability.

If this is indeed the case, then we might also expect that the traffic engineering prefixes would represent a relatively higher proportion of BGP updates than other forms of more specifics. However this is not the case. Figure 17 shows that the category of more specifics that are disproportionally over-represented in the update profile are those that "hole punch" the aggregate, and use a different origin AS than that used by the aggregate.

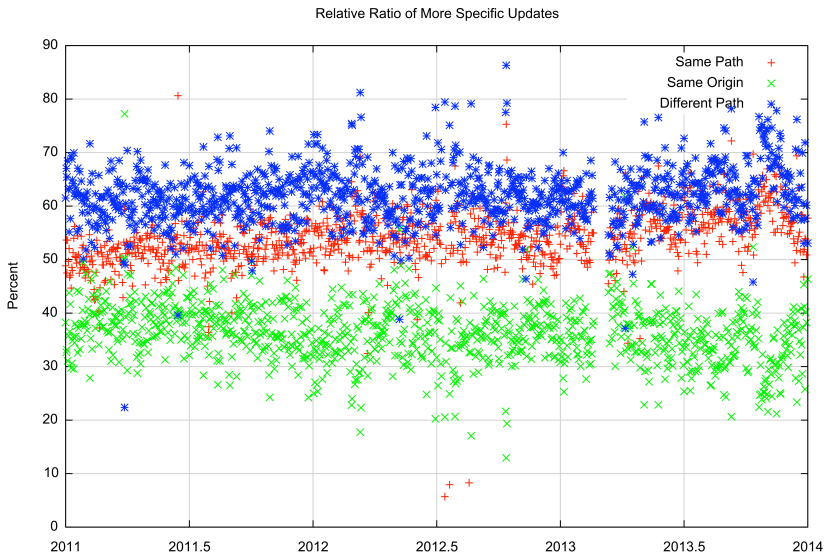

Figure 17 – Relative Proportion of Updates between More Specifics: 2011 - 2014

It's possible that the reason for the higher level of activity of hole-punching more specifics is due to a longer AS path of the more specific, or it could be that these forms of more specifics are more unstable (i.e. have a higher probability of instability) than the other forms of more specifics. One way to try and distinguish between these two cases is to look at the relative ratio of instability for each of the three prefix types. In this case we are looking at the difference between the relative proportion of the occurrence of each type of more specific and the relative level proportion of each type of prefix that was updated on any particular day. A positive value indicated that this type of prefix was relatively less stable than the other parts, while a negative value indicates a higher level of stability. Figure 18 shows this data for the period 2100 to the present.

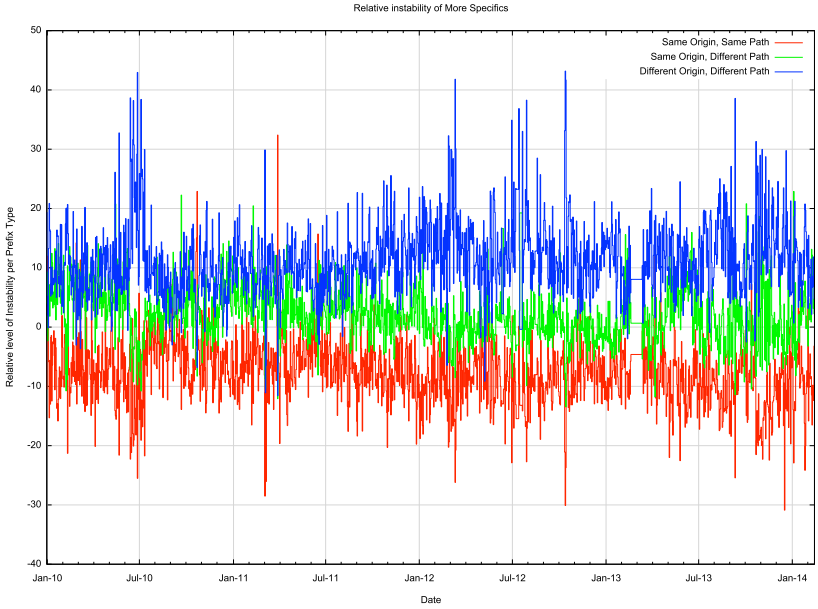

Figure 18 – Relative Instability of different types of More Specifics: 2010 – 2014

The set of so-called "hole punching" more specifics, that show a different origin AS and different AS Path are some 10% more likely to show instability, as compare to the other forms of more specifics. More specifics that share a common AS Path with the covering aggregate are, relatively speaking, more stable,, by around the same level of some 10%, while the "traffic engineering" prefixes show a level of instability that is comparable to their relative occurrence in the routing table.

Observations about Routing Churn

It is commonly believed that more specifics are the source of a disproportionally high level of BGP routing activity., and the data gathered in this exercise tends to support this supposition. While more specifics make up some 50% of the routing table, they are the subject of 80% of the routing updates.

We can take this a little further, and observe that those specifics that use exactly the same Origin AS and AS Path as their covering aggregate tend to be more stable, and more specifics that use a different Origin AS and different AS Path are, relatively speaking, less stable than other forms of more specifics.

In other respects, the routing table defies conventional expectations. The number of unstable prefixes each day, and the number of BGP updates required to reach a stable converged state has remained uncannily constant over many years. This has resulted in the surprising observation that the number of routing updates has remained stable for some years, despite a continued growth in the number of prefixes being advertised in the routing table.

Part of this apparent anomaly can be explained by the topology of the expanding network: as the network continues to grow, the pattern of new connected AS's tends to repeat the overall topology of the Internet. In other words the growth of the Internet is one of increasing density, rather than

increasing size. The result is a relatively constant AS Path length, which appears to limit the extent to which BGP will perform path hunting in order to reach a stable state.

But the other part of this stability is harder to explain. Why is the daily number of unstable prefixes so stable? What is the underlying common constraint that limits this level of routing churn to some 20,000 prefixes per day? That's a question that still has no clear answer. So whatever we are doing with the growth of the Internet has been extremely effective so far, and while the number of routing entries continues to grow, the metrics of routing instability have been held constant. This means that BGP continues to be effective as a routing protocol, and the cost of routing continues to drop over time, both of which are highly fortuitous outcomes. So whatever we are doing so well in BGP, we should continue to do. However, perhaps it would be more comforting to understand exactly what it is we are doing so well!

Disclaimer

The above views do not necessarily represent the views or positions of the Asia Pacific Network Information Centre.

Author

Geoff Huston B.Sc., M.Sc., is the Chief Scientist at APNIC, the Regional Internet Registry serving the Asia Pacific region. He has been closely involved with the development of the Internet for many years, particularly within Australia, where he was responsible for the initial build of the Internet within the Australian academic and research sector. He is author of a number of Internet-related books, and was a member of the Internet Architecture Board from 1999 until 2005, and served on the Board of Trustees of the Internet Society from 1992 until 2001.

www.potaroo.net