The ISP Column
A monthly column on things Internet

March 2009

Geoff Huston

BGP in 2008

Here in my part of the world the season has well and truly turned from summer to autumn, which means that another year has come and gone. I thought that it might be time to give MTU examination a rest for a month or more and instead review the last 12 months in BGP-land and see what's been happening there.

BGP has been toiling away, literally holding the Internet together, for close to two decades now, and nothing seems to be falling off the edge of the Internet. So why should we be interested in the growth trends for BGP? Here's some possible reasons why this data can be useful for folk in Internet business.

For the ISP network operator, this information may be help in figuring out how big a router should you buy today if you want it to still cope with the full BGP routing load in 3 - 5 years time. Perhaps you might want to work out what FIB size is necessary in that time, and what TCAM size is necessary, in which case you may want to have a conservative estimate of the anticipated number of entries in the routing table over that period. If this applies to customers of routing equipment, the same applies to a vendor of such equipment: How big a router should a vendor build to cope with the BGP load over the next 3 - 5 years? What are the Internet's scaling factors at play here?

Underlying this questions are a more basic set of questions about BGP itself. Is BGP scaling or is it failing? Do we need to develop a new Inter-Domain Routing protocol to take over from BGP? If so, how much time do we have before a new approach is needed? And if we are going to head down this path is the problem simply one of routing over an ever larger and more diverse population, or is this an expression of a more fundamental scaling limitation of the Internet's current concepts of names and addresses? In other words, if we are facing a major problem with routing scalability do we need now to examine alternate models of identity and location separation in order to build truly massive and highly diverse networks? Or, is routing scaling an intractable problem within the confines of the current architecture and we need to shift around the basic building blocks of the Internet architecture in order to allow a different routing architecture that has radically different scaling properties?

These questions were studied by the Internet Architecture Board at its workshop in October 2006, which was written up as RFC4984, and projections of routing table inflation in the coming years were a source of considerable concern:

> The workshop participants believe that routing scalability is the most important problem facing the Internet today and must be solved, although the time frame in which these problems need solutions was not directly specified. The routing scalability problem includes the size of the DFZ RIB and FIB, the implications of the growth of the RIB and FIB on routing convergence times, and the cost, power (and hence, heat dissipation) and ASIC real estate requirements of core router hardware.
>
> It is commonly believed that the IPv4 RIB growth has been constrained by the limited IPv4 address space. However, even under this

> constraint, the DFZ IPv4 RIB has been growing at what appears to be an accelerating rate [DFZ]. Given that the IPv6 routing architecture is the same as the IPv4 architecture (with substantially larger address space), if/when IPv6 becomes widely deployed, it is natural to predict that routing table growth for IPv6 will only exacerbate the situation.
>
> RFC4984: Report from the IAB Workshop on Routing and Addressing, September 2007

At the time the picture was not looking overly optimistic for the longer term prospects of BGP, and the workshop prompted further studies of routing techniques and architectures that were capable of sustaining a greater level of information aggregation.

> First of all, the workshop participants would like to reiterate the importance of solving the routing scalability problem. They noted that the concern over the scalability and flexibility of the routing and addressing system has been with us for a very long time, and the current growth rate of the DFZ RIB is exceeding our ability to engineer the routing infrastructure in an economically feasible way. We need to start developing a long-term solution that can last for the foreseeable future.
>
> RFC4984: Report from the IAB Workshop on Routing and Addressing, September 2007

But I wasn't going to head in that direction of describing the areas of possible direction for future routing systems in this article. The question I'd like to ask here is somewhat more pragmatic in nature: has anything changed in this perspective on BGP? Are the prospects of the medium term collapse of BGP through scaling overload still a realistic option for the routing environment? Should we still be concerned about routing scaling? Is the BGP sky about to fall on our heads?

I can't let that pass without at least a passing reference to Asterix and the chief of the Gaulish village, Abraracourcix (or Vitalstatistix in the English version, if you prefer), who had nothing to fear except for the sky falling on his head! <http://en.wikipedia.org/wiki/Recurring_characters_in_Asterix#Vitalstatistix>

The BGP Measurement Environment

In trying to analyse long baseline data series, the ideal approach is to keep as much of the local data gathering environment as stable as possible so that the changes that occur in the collected data reflect the larger environment and not the local configuration of the data collection equipment. In this case the measurement point being used is a BGP router configured as AS2.0 (or AS131072 if you prefer!). This AS generates no traffic and originates no routes in BGP. It's a passive measurement point that has been logging all received BGP updates since 1 July 2007, and is the successor to an earlier setup located in AS1221. The router is fed with a default-free eBGP feed from AS 4608, which is the APNIC network located in Australia, and AS 4777, which is the APNIC network located in Japan, as AS1280, a RIPE RIS Route Collector. For IPv6 routes the

measurement system is being fed with complete route sets from AS1221 (Telstra), AS1280 (RIPE NCC), and AS5539 (ISC). My thanks to these folk for their willingness to fee me routing data for this work.

What is being used here is a single view of the "edge" of the network, looking at an eBGP perspective, as distinct from a mixed eBGP / iBGP environment. This AS is not an upstream for anyone else, so it has no transit role, and does not have a large set of BGP peers.

There is also no iBGP in this setup. While it has been asserted at various times that iBGP is a major contributor to BGP scalability concerns in BGP, the consideration here in trying to place some data against this assertion is that there is no "standard" iBGP configuration, and each network has its own rather unique configuration of Route Reflectors and iBGP peers. This makes it hard to generate a "typical" iBGP load profile, let alone analyse the general trends in iBGP update loads. In this study the scope of attention is limited to simple eBGP configuration that is likely to be found at a "stub" AS at the edge of the Internet, and the effects of iBGP are not included in these measurements.

The measurement system took a snapshot of the BGP RIB every hour, as well as logging all received BGP updates.

IPv4 BGP Table Data

The following tables show some of the vital statistics for IPv4 in BGP over the past 12 months. In and of themselves the graphs are not that informative. The graphs show relatively stable increase in most of the routing metrics. The discontinuities in March, April December was caused by the measurement environment adding and dropping a BGP peering session with AS1280, rather than any shift in the characteristics of the network itself.

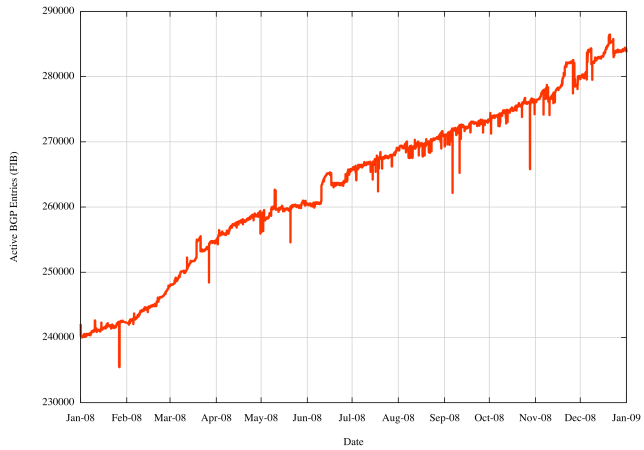

Figure 1 - IPv4 BGP Routing Table Size (RIB)

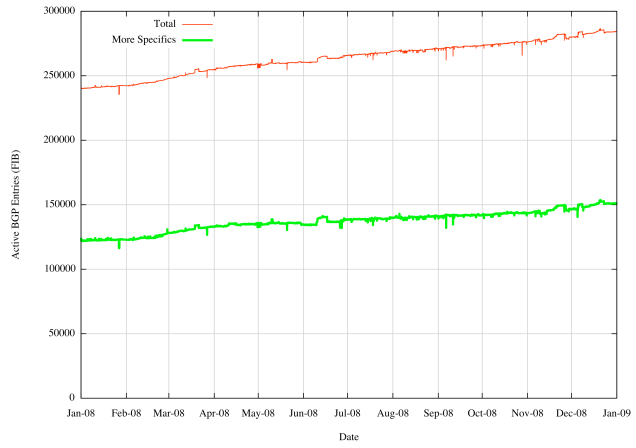

Figure 2 - IPv4 BGP Routing Table - More Specific Entries

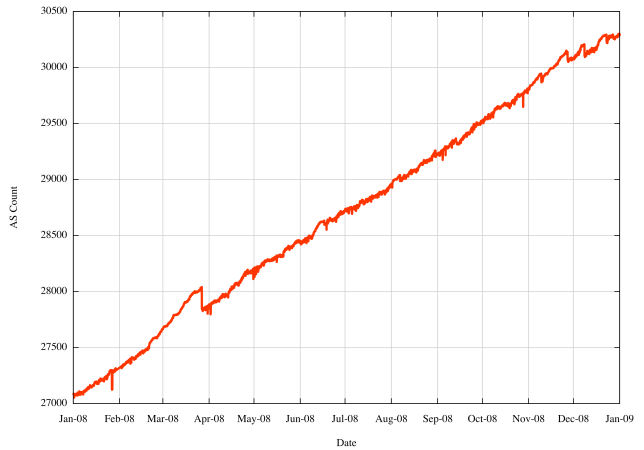

Figure 3 - IPv4 AS Count

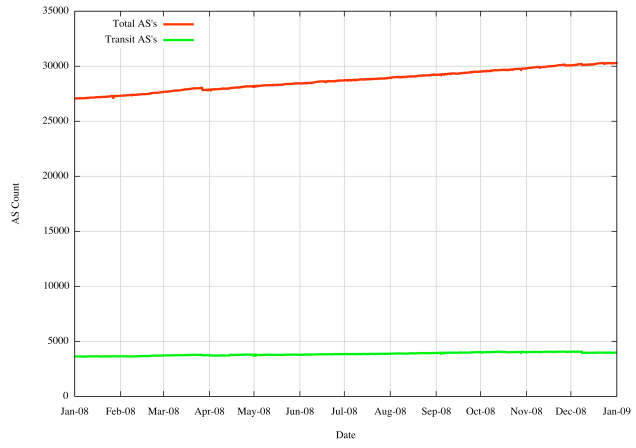

Figure 4 - IPv4 Transit AS Count

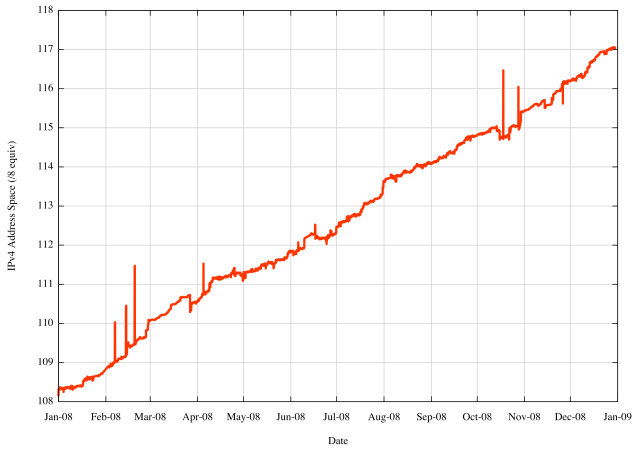

Figure 5 - IPv4 Advertised Address Space

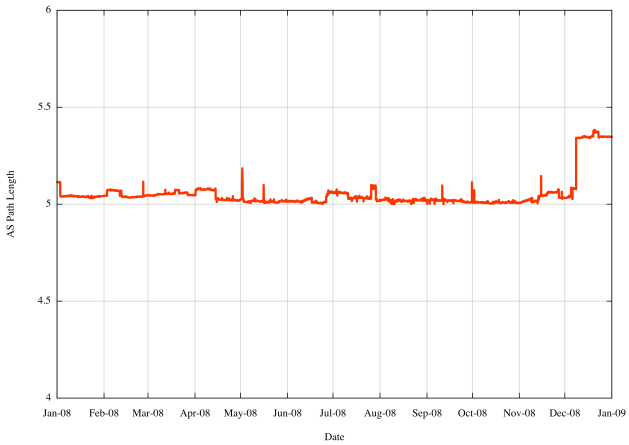

Figure 6 - IPv4 Average AS Path Length

The summary of the IPv4 BGP network for 2008 is

	Jan-08	Dec-08	2008	*2005*
Prefix Count	245,000	286,000	17%	*18%*
Root Prefixes	118,000	133,000	13%	*18%*
More Specifics	127,000	152,000	20%	*18%*
Address Span (/8s)	106.39	118.44	11%	*10%*
AS Count	27,000	30,200	11%	*14%*
Transit AS Count	3,600	4,100	14%	*14%*
Stub AS Count	23,400	26,200	11%	*14%*

What this table indicates is that for the IPv4 Internet the use of aggregates in the routing system has not improved over 2008, nor has it become significantly worse. The average size of advertisements is getting smaller in terms of address span per routing table entry, the span of originating addresses per AS is getting smaller, the average AS path length is constant at around 5 AS hops (which would translate to 4 AS hops if the measurement setup overhead was removed) and the number of AS's is increasing, and the interconnection degree of AS's is getting higher. The implication is that the granularity of the inter-domain routing system continues to get finer and the density of interconnection is getting greater. In other words the growth of the Internet is not "growth at the edge" and the network is not getting any larger in terms of average AS path change. Instead, the growth is happening by increasing the density of the network by attaching new networks into the existing transit structure and peering at established exchange points. This makes for a network whose diameter, measured in AS hops, is essentially static, yet whose density, measured in terms of prefix count, AS interconnectivity and AS Path diversity, continues to increase.

The growth metrics of the routing system in 2008 are not overly different from that of 2005 in terms of the growth of the routing table and the span of announced addresses. The growth rate of the transit ASs is slightly lower than in 2005, but not significantly so.

IPv6 BGP Table Data

A similar exercise has been undertaken for IPv6 routing data, and the comparable figures are shown below.

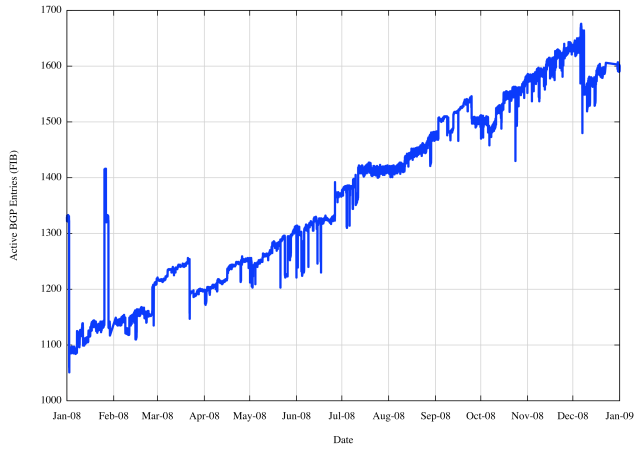

Figure 7 - IPv6 BGP Routing Table Size (RIB)

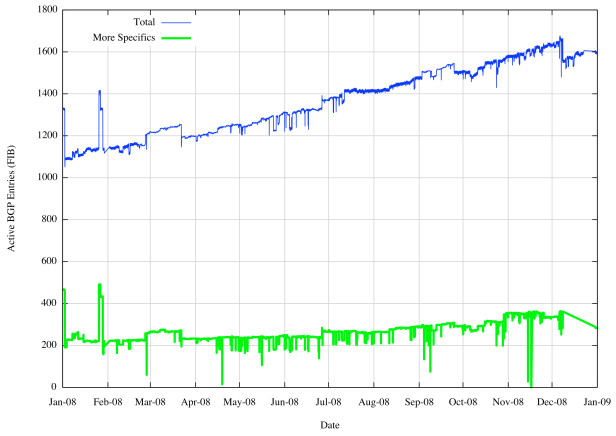

Figure 8 - IPv6 BGP Routing Table - More Specific Entries

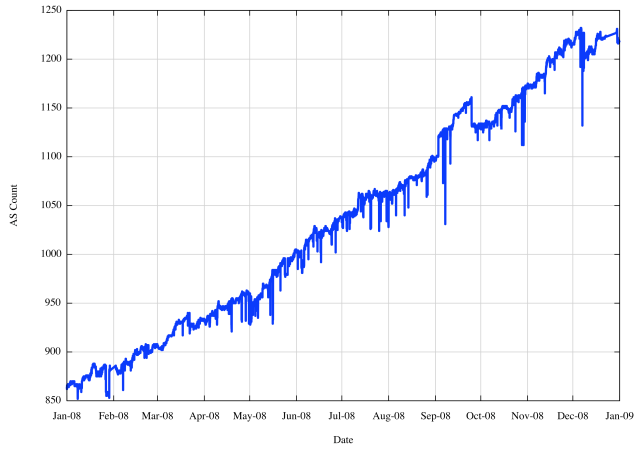

Figure 9 - IPv6 AS Count

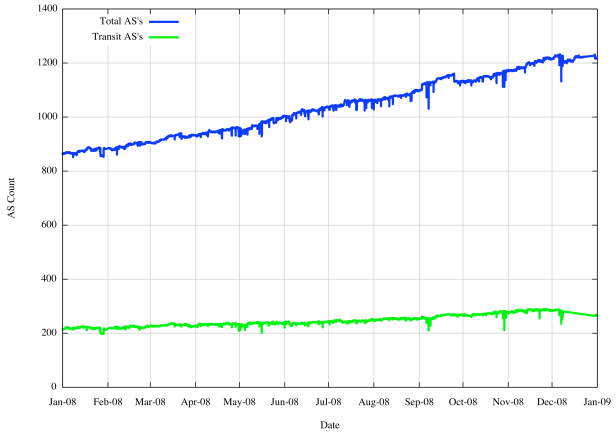

Figure 10 - IPv6 Transit AS Count

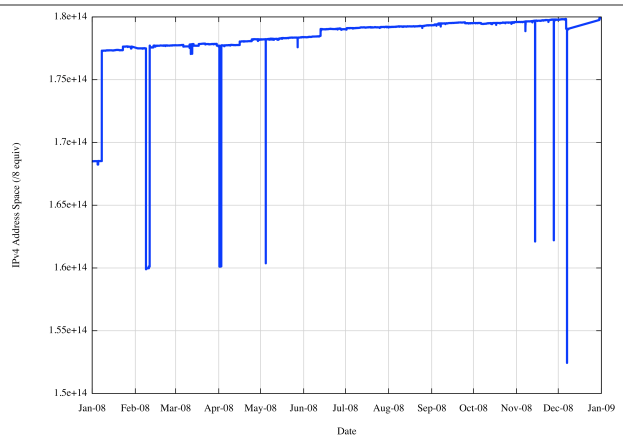

Figure 11 - IPv6 Advertised Address Space

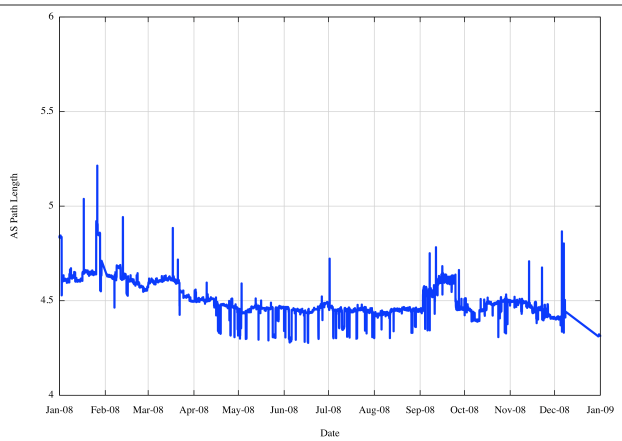

Figure 12 - IPv6 Average AS Path Length

The summary of the IPv6 Internet for 2008 is as follows:

	Jan-08	Dec-08	2008	*2005*
Prefix Count	1,050	1,600	52%	*21%*
Root Prefixes	840	1,300	55%	*15%*
More Specifics	210	300	43%	*51%*
Address Span	/16.67	/16.65	1%	*50%*
AS Count	860	1,230	43%	*20%*
Transit AS Count	240	310	29%	*21%*
Stub AS Count	620	920	48%	*18%*

It is harder to make generalizations about the trends in the IPv6 network over 2008, as the IPv6 network is simply not large enough to show any overall trend behaviour as yet. Certainly the rate of pick up is higher than the comparable statistics in the IPv4 network, and the annual rate of increase is higher than was seen in 2005. This is encouraging news if you are looking for positive signs of IPv6 update in the Internet, but in absolute terms the metrics still fall far short of the comparable metrics of the IPv4 Internet.

Projecting the BGP Size

What can this data tell us in terms of projections of the future of BGP in terms of BGP table size?

The technique used here is to take the hourly snapshots of the BGP table size and firstly filter the data to remove some anomalous entries related to additional routes visible from AS12054 but not globally visible, then apply a filter that generates a daily average table size, then applies a smoothing function across the data, using a 60 day value as the parameter to the multi-day smoothing function. This has been done using an extended data set that cover the past 60 months. The result of this function applied to the IPv4 BGP table is shown in the following figure.

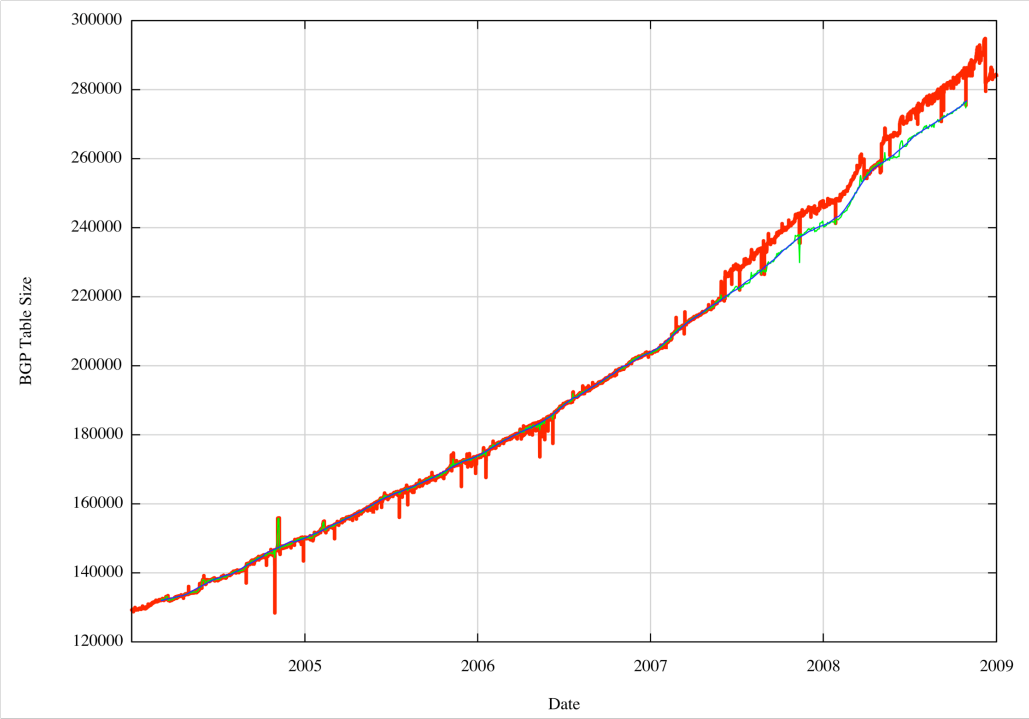

Figure 13 - Smoothed IPv4 BGP Table Size

The first order differential of the smoothed data is then taken, as shown in the red line the following figure.

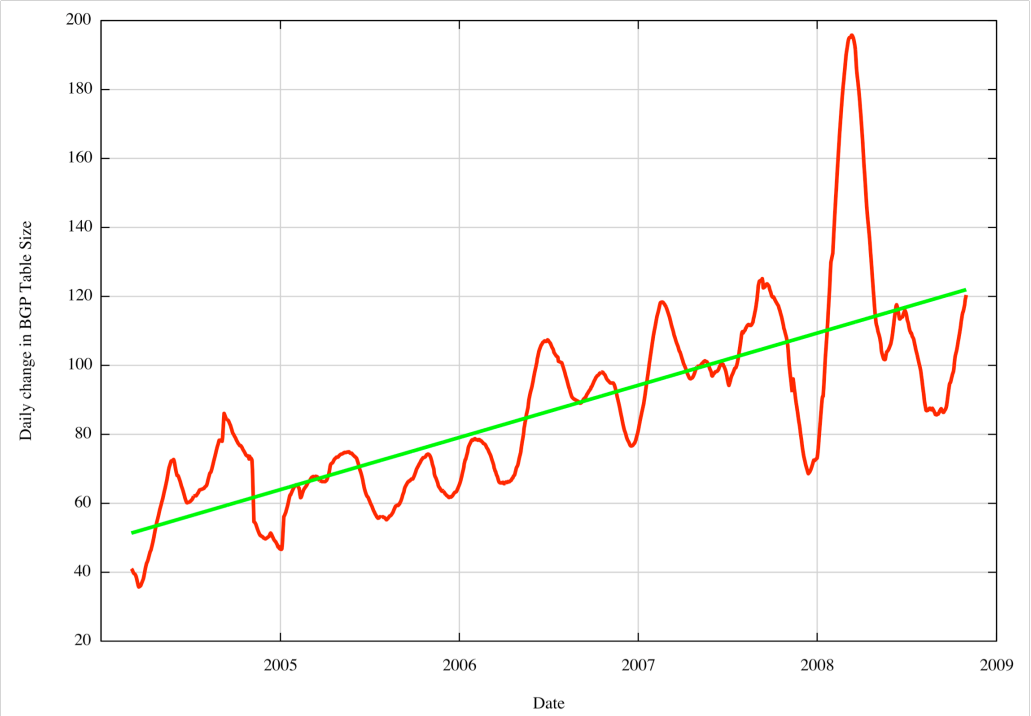

Figure 14 -First Order Differential of Smoothed IPv4 BGP Table Size

The longer term trend of this first order differential is a linear function, shown in green in the above figure. A linear first order differential (dy/dx = ax+b) implies a fit of a quadratic function to the data (y = a/2 x**2 + bx + c).

This quadratic function can then be used to create a forward projection of the table size, shown as the blue line in the following figure.

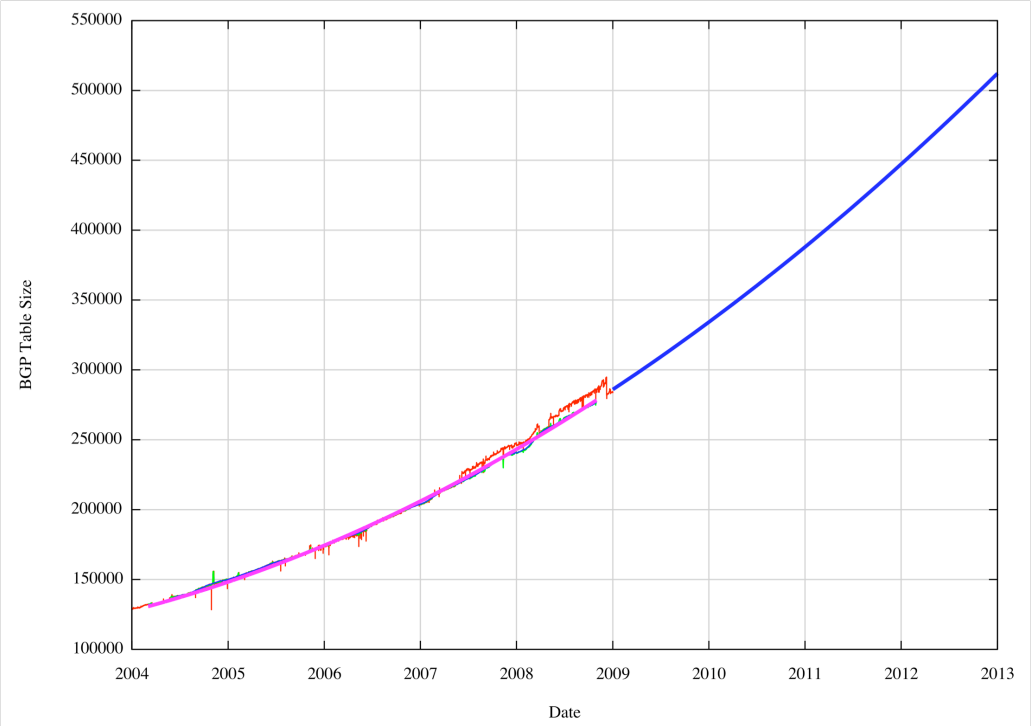

Figure 15 -Prediction of IPv4 BGP Table Size

This same predictive exercise was undertaken in January 2006, and the following table shows the predictions generated from the current data and those generated using the same approach three years earlier

	Jan 2009 prediction	Jan 2006 prediction
Jan 2009	285,000	275,000
Jan 2010	335,000	322,000
Jan 2011	388,000	370,000
Jan 2012	447,000 *	
Jan 2013	512,000 *	

There is a relatively good correlation between the numbers predicted by a quadratic growth model of BGP table size in 2006 with the data in the period 2006 - 2009, and a reasonably good correlation of predictions for the next two years.

With the caveat that this prediction is based on the assumption that tomorrow is a lot like today and that the influences that shape tomorrow have already shaped today, then its reasonable to predict that the routing table in two years time, at the start of 2011, will contain an additional 100,000 entries, making a total for IPv4 of some 388,000 entries.

However I'm not anywhere near as confident in making predictions beyond 2011, and certainly not all that confident in the predictions generated by this model for January 2012 and January 2014. The problem is that another predictive model, that of the consumption of as-yet unallocated IPv4 addresses, predicts the effective exhaustion of the unallocated IPv4 number pool in 2011 / 2012. It is not possible to use the current models of BGP growth to peer into this post-exhaustion IPv4 routing environment, so the numbers given in the table above for January 2012 and 2013 are extremely uncertain.

Perhaps there is another way of looking at this. If one assumes that the major objective here is to ensure that the "unit cost" of routing continues to decline over time, or at least remain constant, what benchmark could be used to compare the BGP prediction against in terms of a constant unit cost curve?

One possible model that could be used as a benchmark of a prediction of constant unit cost in terms of this form of routing and packet forwarding hardware in packet networks is Moore's Law <http://en.wikipedia.org/wiki/Moore%27s_law>. Here the general assumption is that as long as the growth parameters of the routing table sit within the parameters of Moore's law then the expectation is that the unit cost of routing and switching hardware should not escalate to any appreciable extent. The following figure compares the quadratic projection model of the size of the BGP default free zone with an exponential model of doubling every two years, as used in Moore's Law. As can be seen in the figure below there is no real cause for alarm at this stage, and the BGP table size appears to fit comfortably within these parameters within the current projection model.

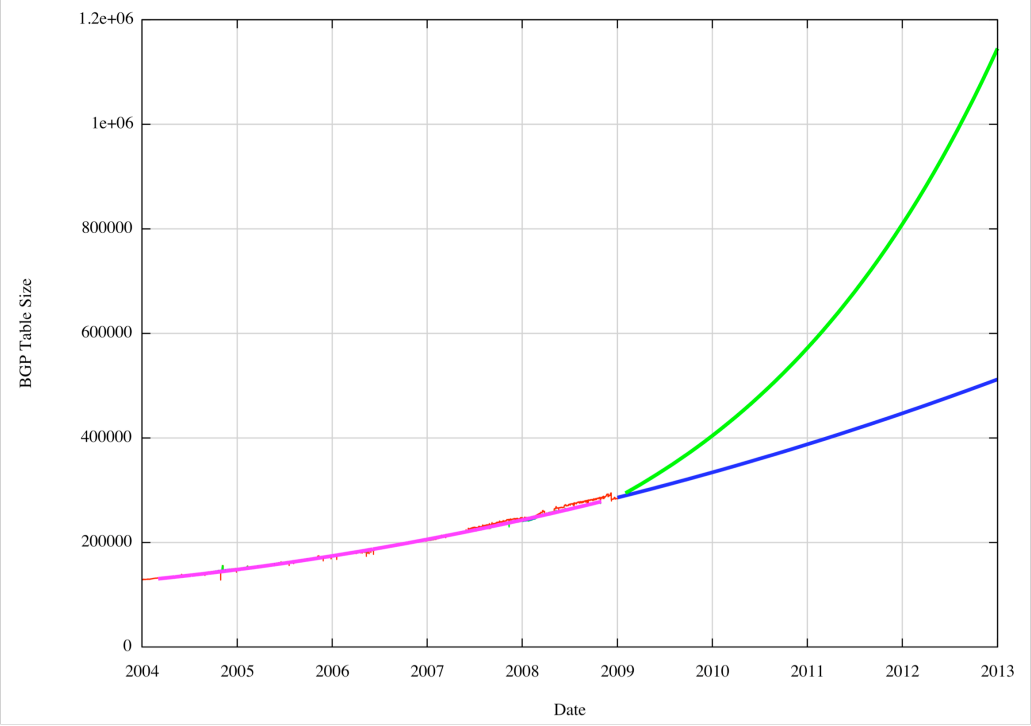

Figure 16 -Comparison of BGP RIB prediction to Moore's Law Growth

Of course, if address exhaustion causes a rapid doubling on the routing table across 2011, inflating the routing table by early 2012 to a size of 1 million entries or more, then this would represent a somewhat different scenario. What could potentially drive this rapid inflation scenario is some form of IPv4 address redistribution function that as focussed solely on the public addressing requirements in IPv4 NAT scenarios, probably increasing the prevalence of global routes at the /24 level, or potentially at even smaller sizes, coupled with a scenario of very rapid level of uptake across the global IPv4 BGP routing table. Such scenarios are related to the levels of speculation concerning the industry reaction to the exhaustion of the existing mechanism of IPv4 address distribution, and at this point in time the level of speculation about the nature of the redistribution function and the pressures placed on the routing space in consequence is extremely high indeed.

Measuring BGP Updates

Whenever this discussion about routing scalability takes place, there is a related discussion about what aspect of scaling is being discussed. Is it really the size of the routing space that is the topic of deep concern, or is it the dynamic properties of the routing system? Should we be looking at the average time to reach convergence? Or the volume of BGP update message per unit time?

Part of this measurement exercise has been to collect every BGP update. The figure below shows the number of BGP updates per day, or to be more precise, the number of prefix updates per day over 2008. This is shown in the following figures.

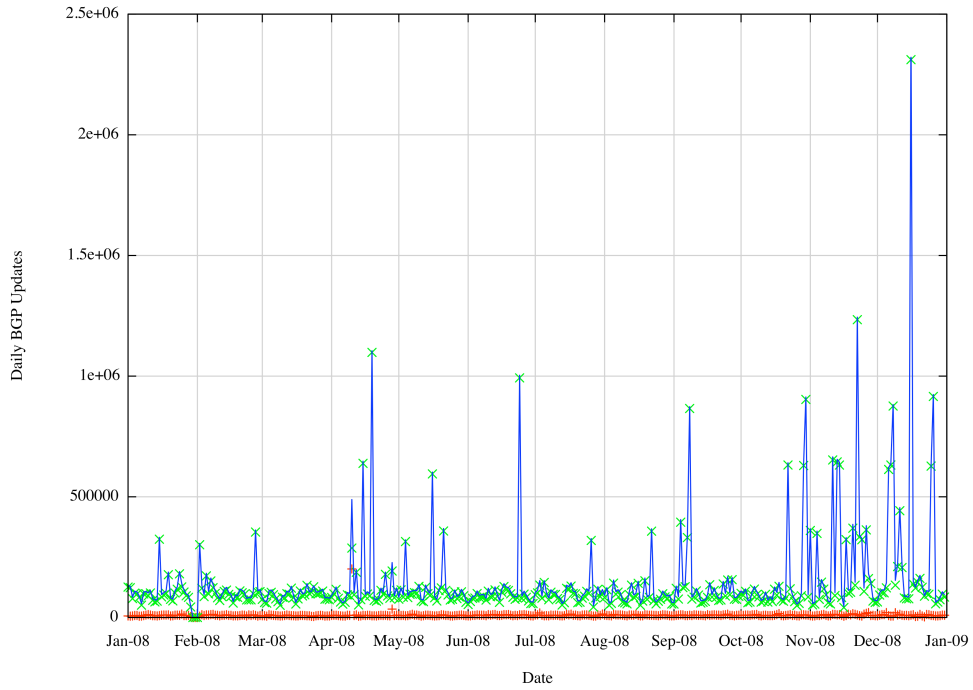

Figure 17 -Daily BGP Updated Prefix Counts for 2008

The first view is the number of updated prefixes per day in BGP (Figure 17). At this scale, the daily withdrawal rate is relatively constant, while the number of updates per day shows a number of extreme outliers.

On investigation, these outliers are attributable to session resets in the local measurement setup, where the local BGP system performs a reset and is re-fed the complete route set. ON some occasions there were multiple resets in the day, including one day where the BGP table was reloaded 9 times. These local session reset updates can be filtered out from the data set, to produce the following view of the number of updated prefixes per day in BGP for 2008, and a best fit can be applied to the data, using a least squares best fit.

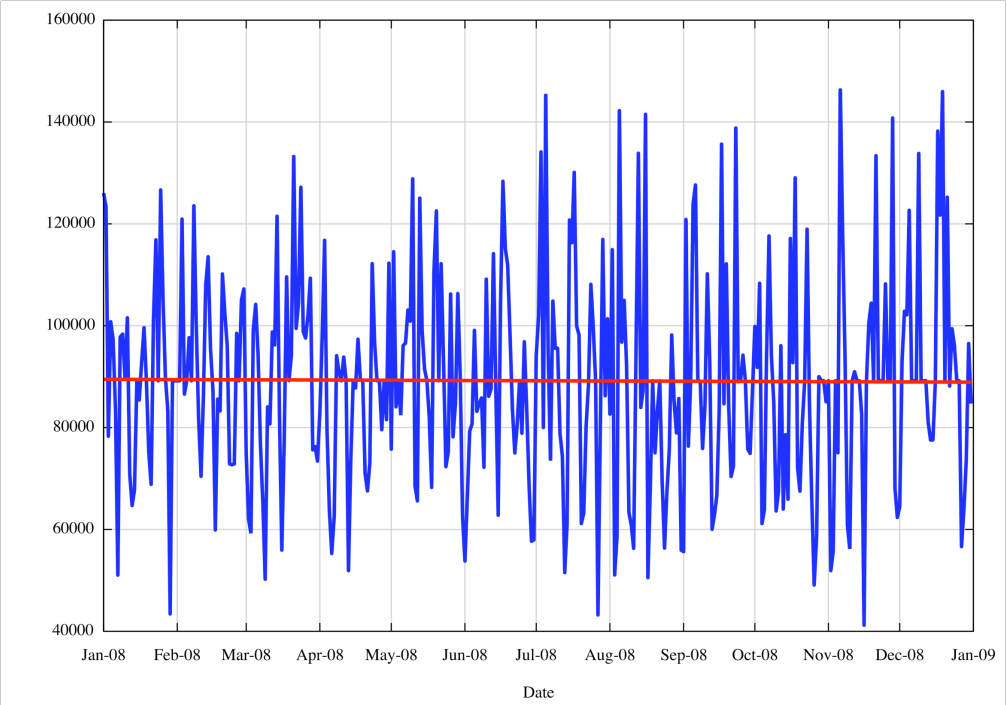

Figure 18 -Daily BGP Updated Prefix Counts for 2008

This data shows a daily rate of 89,000 updated prefixes per day, or an average prefix update rate at a level of slightly over 1 a second. Obviously this has very little relationship to the actually update rate that a BGP speaker is likely to see, but it is a useful metric in looking at the order of scale of the processing load imposed by the flow of BGP updates. This update data can be extended back in time using data collected from previous years, again using the same techniques of filtering out local BGP reset traffic and applying a least squares best fit to the data.

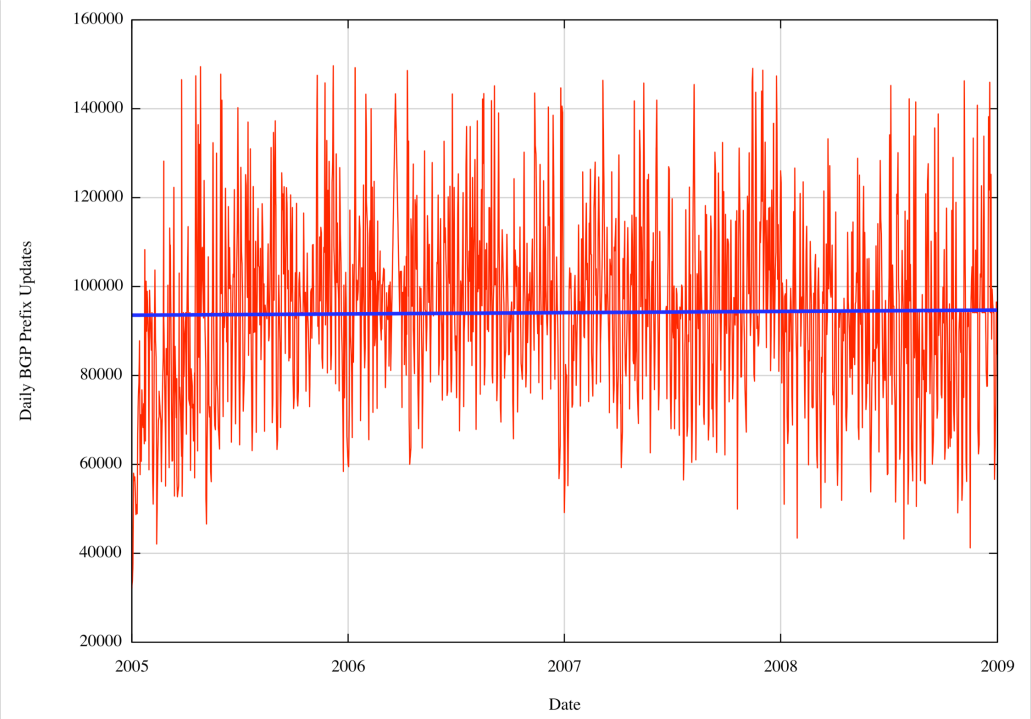

Figure 19 -Daily BGP Updated Prefix Counts for 2005 - 2008

The forward projection of the number of BGP prefix updates is shown in the following figure, this time using a linear function derived from a least squares best fit to the daily data.

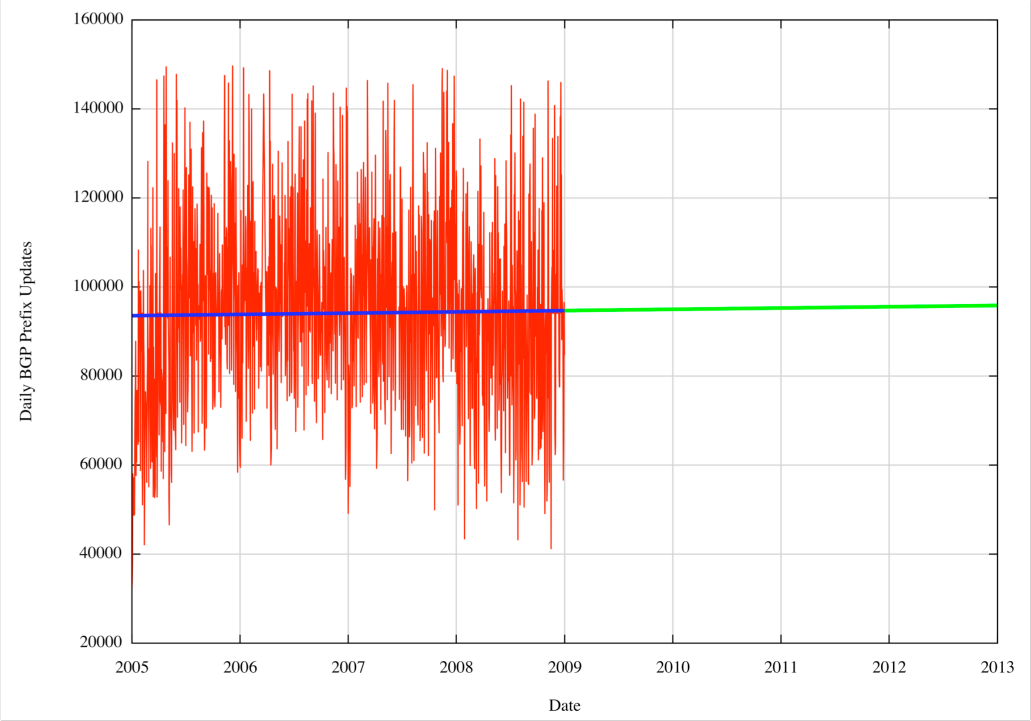

Figure 20 -Daily BGP Updated Prefix projection

The update data shows a surprisingly consistent view of BGP updates with very slight growth projected in the coming years, based on the data from previous years. If there is a looming issue with BGP update processing loads in the coming years, the rate of eBGP updates that are unrelated to local BGP session resets does not appear to be a strong contributor to any such issue.

A similar story relates to withdrawals. The daily count of withdrawals over 2008 is shown below, and the projection into the coming years is also shown in the following figures.

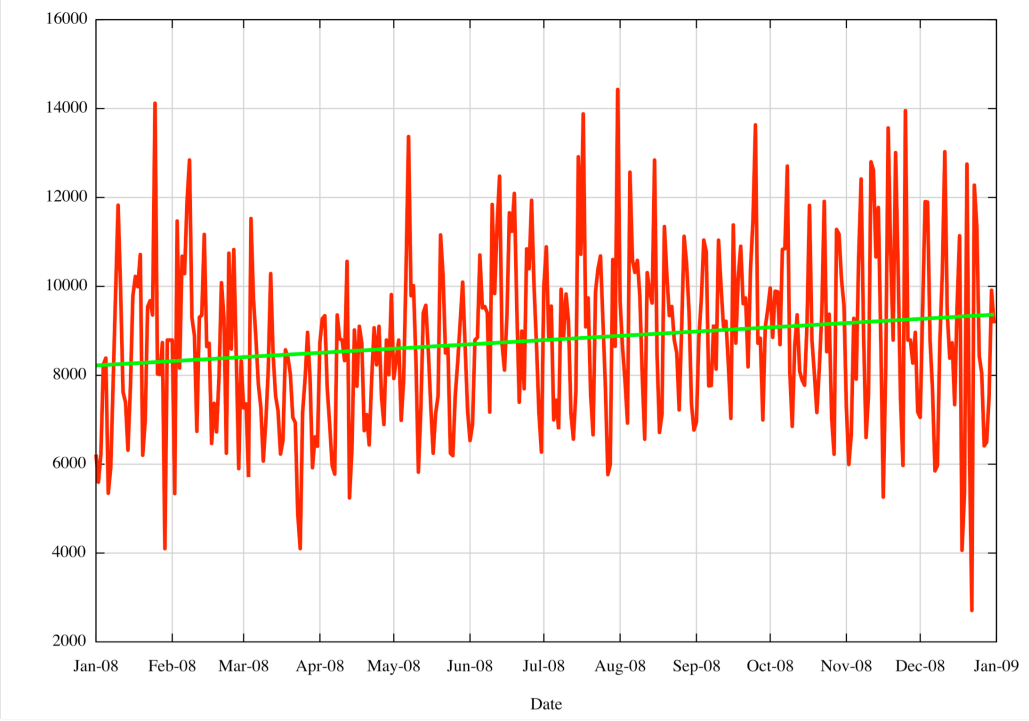

Figure 21 -Daily BGP Prefix Withdrawal Counts for 2008

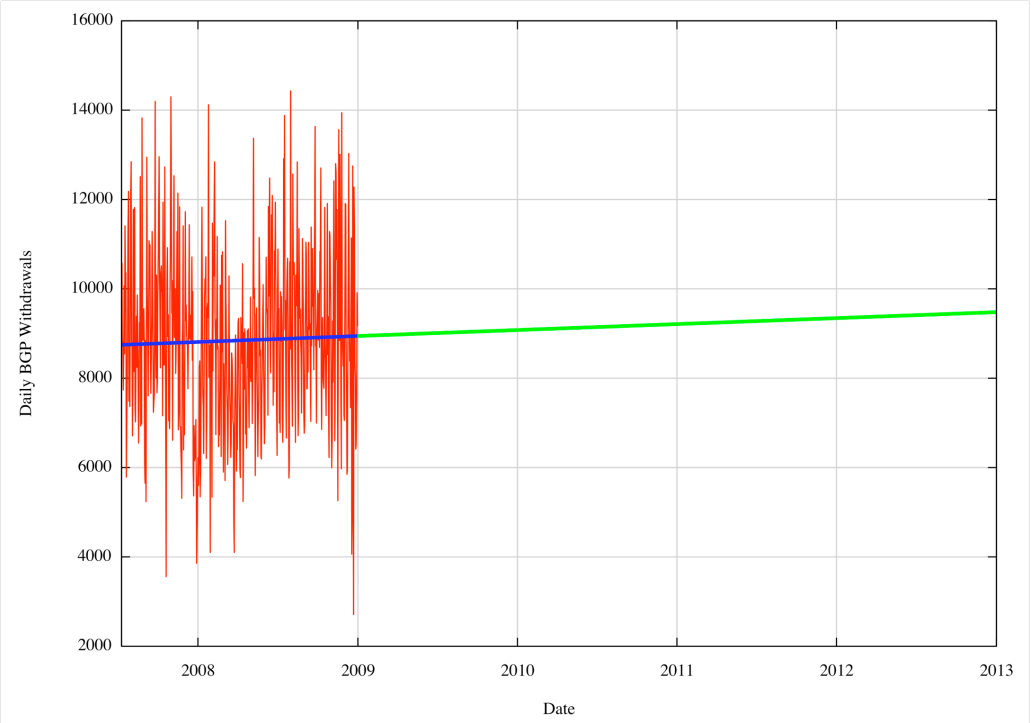

Figure 22 -Projected Daily BGP Prefix Withdrawal Counts

Here a growth trend is visible, to some extent, and while the 2008 data may suggest a 10% annual growth rate, an 18 month window of the data suggests a more conservative view of growth in the number of withdrawals at less than 5%, with the number of withdrawals per day rising from an average of 8,500 a day to 9,800 in the next four years.

Why is this projected growth rate so much smaller than the projects for the growth in the BGP table size?

Surely a more richly connected, larger routing space would generate more routing protocol update traffic. Wouldn’t there be more prefixes? And wouldn’t each prefix generate more updates as a result of BGP's distance vector algorithm attempting to reach convergence? Wouldn't the interaction between a larger routing space and the Minimum Router Advertisement Interval (MRAI) default timer settings on commonly deployed routing equipment delay convergence as the network itself grew?

One way of looking at this is to look at the average number of BGP updates required to reach a converged, or stable, routing state, and the average amount of time taken for routing to reach convergence. Here "stability" is defined as no further updates for 130 seconds or longer. If these suspicions about the behaviour of BGP have any substance, then some form of inflation of these two metrics should be visible in the 2008 data.

The following figures show the base data, which is the daily number of 'instability events' where a prefix took two or more updates before reaching a converged state, and the daily average of the number of updates seen before a prefix is considered stable, and the average amount of time taken. Here the single update events where a prefix moves from a stable state to a new stable state are not included in the data set. This data looking at update sequences of two or more updates, including withdrawals) using this 130 second definition of 'stability.

The number of these 'instability' events appears to be relatively constant on a daily basis. In other words the network as a whole appears to be no more or less unstable at the end of 2008 as it was at the start of 2008, with around 23,000 to 35,000 such events per day.

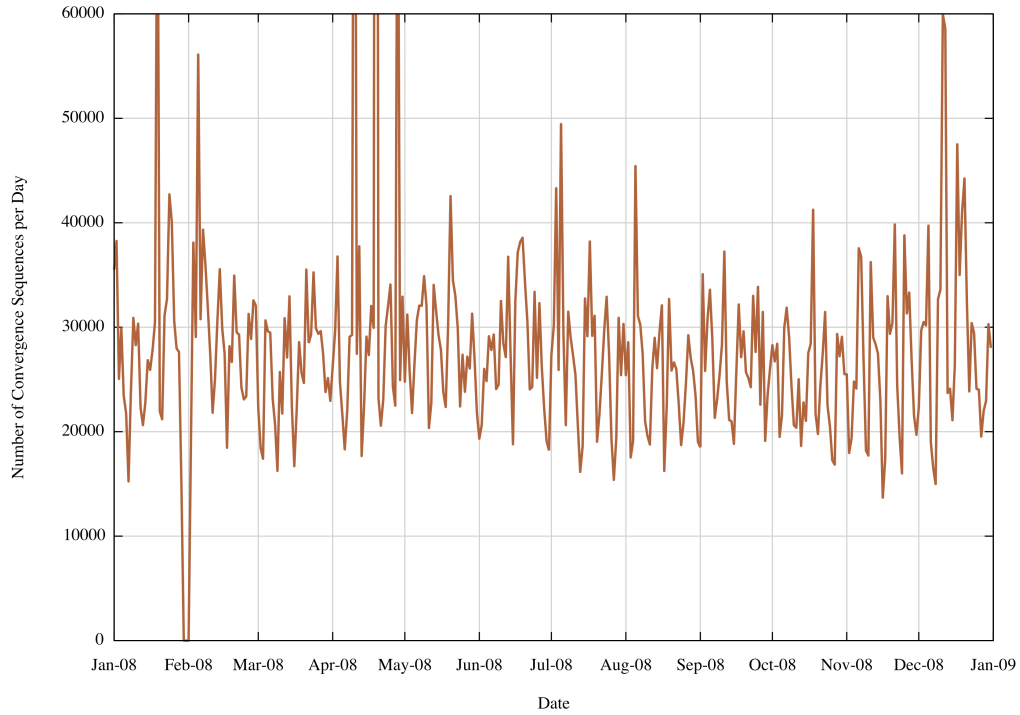

Figure 23 Number of discrete BGP Update sequences per day

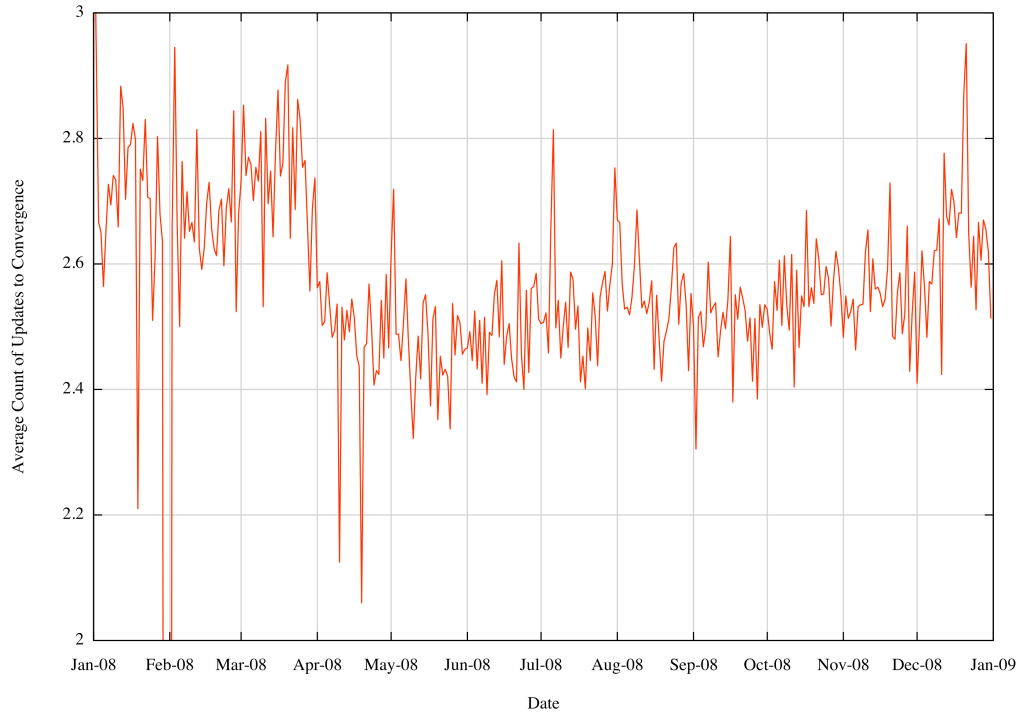

Figure 24 -Daily Average of BGP Updates to reach convergence

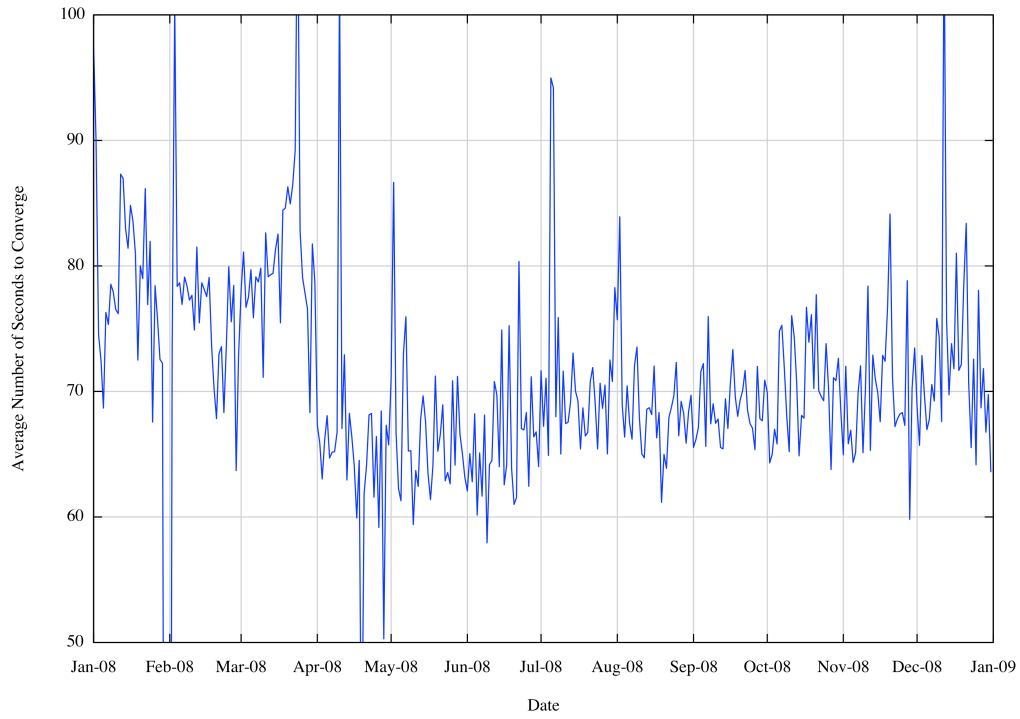

Figure 25 - Daily Average of elapsed seconds to reach convergence

There is a marked anomaly in the data on 1 April, and the network convergence times were significantly improved after that date. On that date the BGP peering with AS1280 was shut down. It was restored a month later, and shut down again in mid December 2009. It appears that this peering session was adding some additional instability into the measurement environment in the first four months of the year.

The next three figures show a least squares best fit across the data. In the case of the number of instability events this is drawn across the full year, and in the case of the number of updates and the average convergence time it is drawn across the period from April to December.

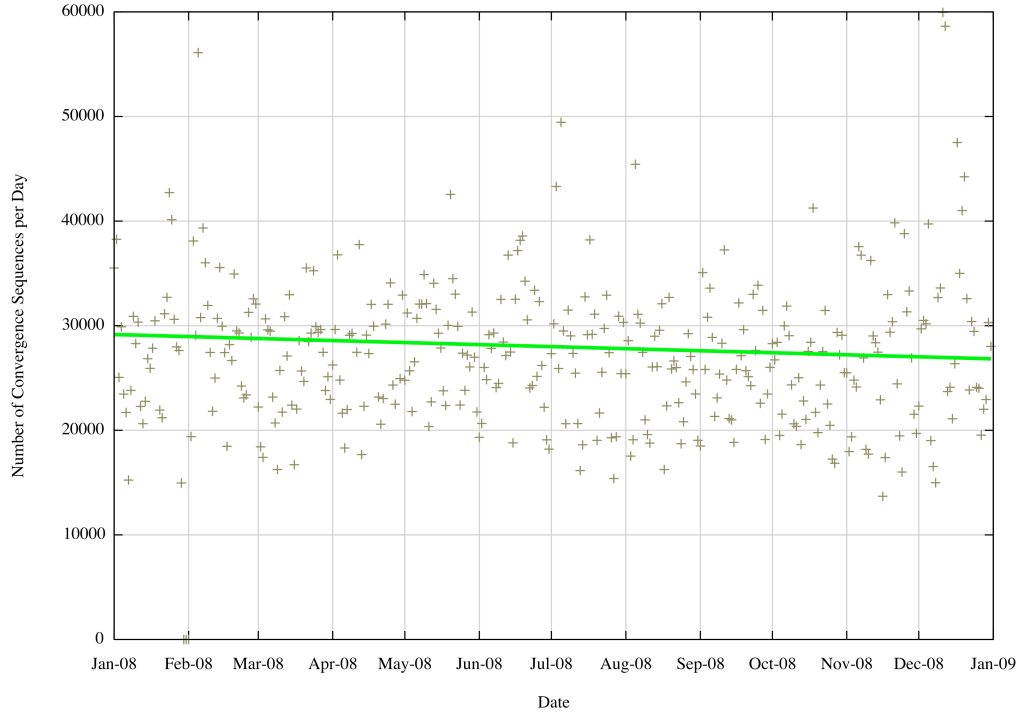

Figure 26 - Trend of number of BGP convergence sequences

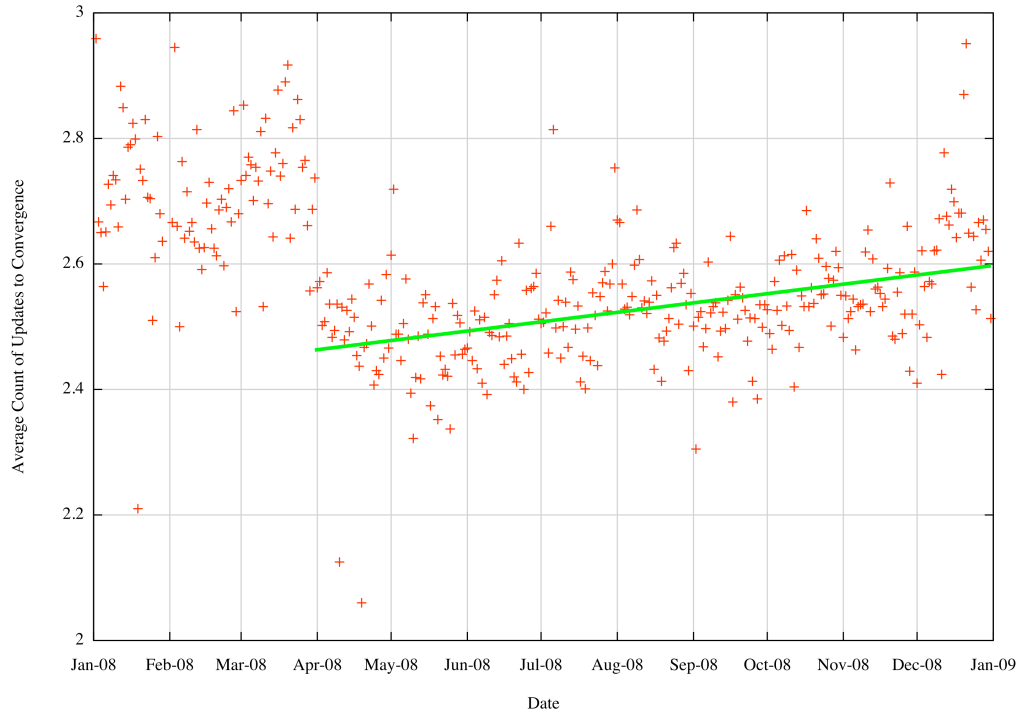

Figure 27 - Daily Average of elapsed seconds to reach convergence - fit to linear model

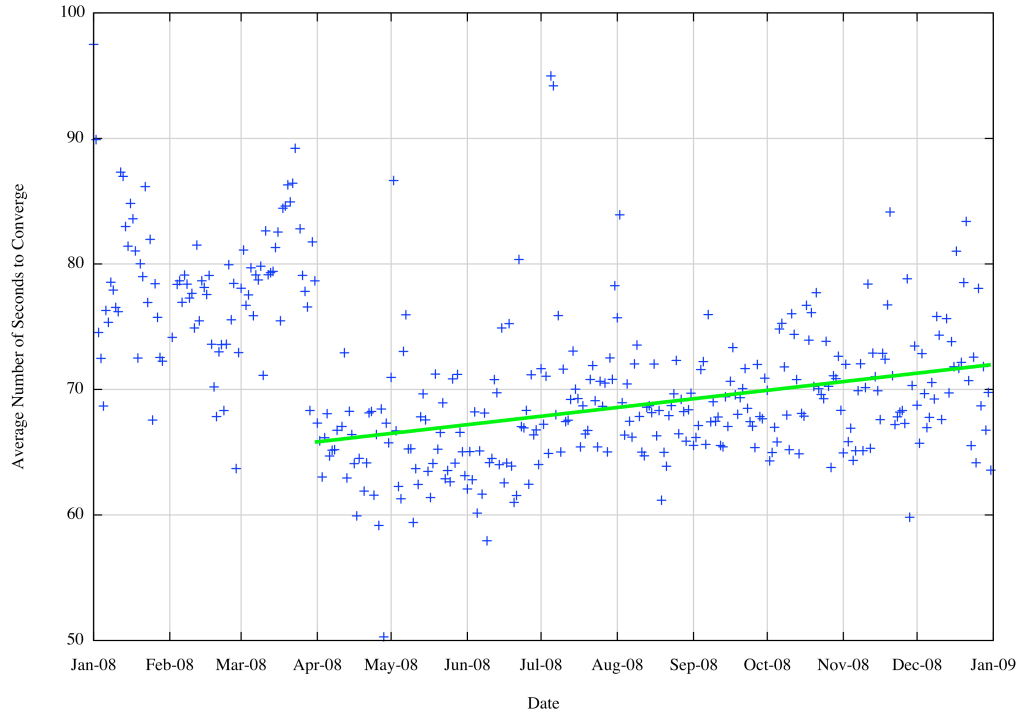

Figure 28 - Daily Average of elapsed seconds to reach convergence - fit to linear model

This shows an increase of 6 seconds, from 66 seconds to 72 seconds over a 9 month period for this convergence metric, and a comparable increase in the average update count from 2.46 to 2.59 updates.

I suspect that the underlying relationship here is between routing convergence and average AS Path length. While the AS Path length remains constant then the dynamic behaviour of BGP to propagate information remains bounded in some sense.

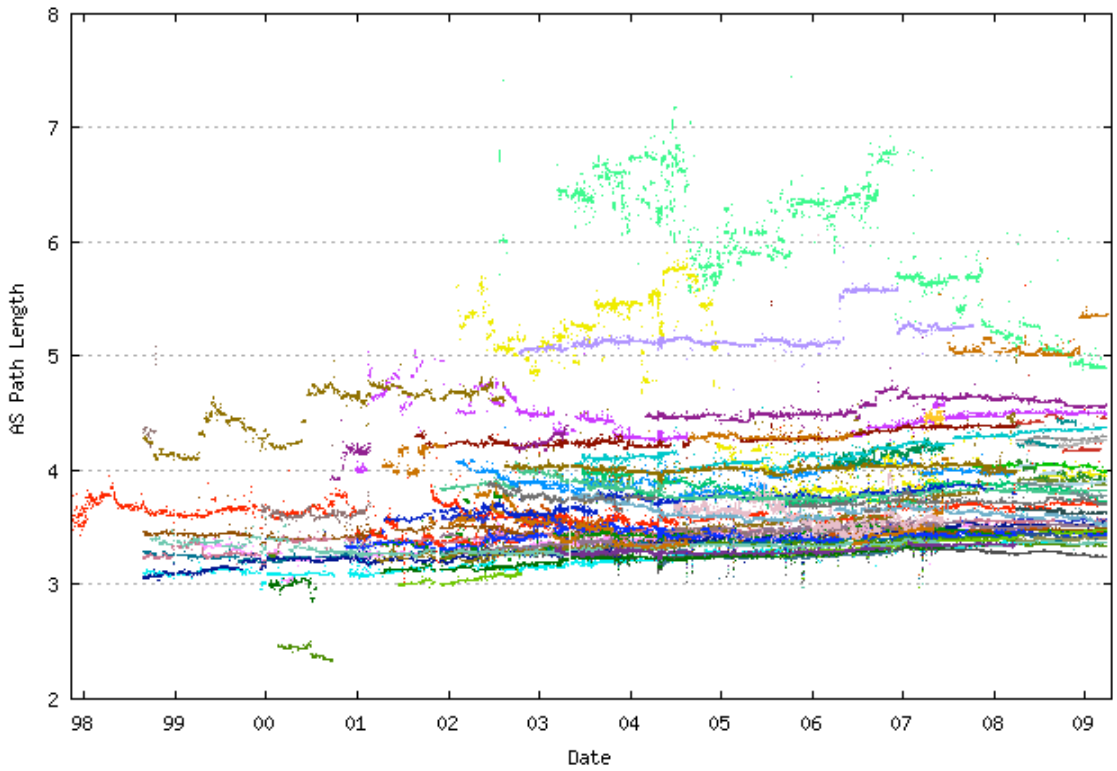

Figure 29 - Average AS Path Length, as seen by each Route Views BGP Peer

As shown in the figure above, the Internet has been remarkably steady in terms of the average AS Path Length metric for more than 10 years. Over the same period the number of routing entries has grown from 50,000 to 300,000 entries, yet the "diameter" of the Internet has remained relatively constant, and all the routing domain growth has increased its "density" of interconnection rather than its radial length. In such an environment BGP has been able to scale very effectively, as the limits to the amount of update traffic required for BGP to reach convergence appears to be more strongly related to the "radial length" of the Internet, in terms of AS hop count, than it is related to the "density" of the Internet, in terms of AS interconnectivity metrics.

As long as this network characteristic is preserved, it appears the BGP can continue to function very effectively.

BGP in 2008

I'm not sure I could say that BGP is on a sure path to perdition, based on the data collected in 2008 relating to the growth in the routing system and the dynamic behaviour of BGP. None of the metrics indicate that we are seeing such an explosive level of growth in the routing system that it will fundamentally alter the viability of carrying a complete eBGP routing table in the near future, nor do the characteristics of convergence behaviour show any sign of the Internet entering into a phase of uncontrollable route instability.

At least for 2008 the BGP sky did not fall on our heads, and the signs for 2009 are looking good, so far!

Disclaimer

The above views do not necessarily represent the views or positions of the Asia Pacific Network Information Centre, nor the Internet Society.

About the Author

GEOFF HUSTON is the Chief Scientist at APNIC, the Regional Internet Registry serving the Asia Pacific region. he graduated from the Australian National University with a B.Sc, and M.Sc. in Computer Science. 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.

http://www.potaroo.net