The ISP Column

An occasional column on things Internet

June 2006

Geoff Huston

The BGP Report for 2005

So how's the Internet's inter-domain routing system getting along these days? Some time back in this column I looked at the state of inter-domain routing, and speculated as to how it could evolve (see "The State of Inter-Domain Routing", March 2004, and the earlier RFC (a "Commentary on Inter-Domain Routing in the Internet", RFC 3221). At the time it looked as if we'd be seeing some very real scaling problems with inter-domain routing, where the routing system was growing at a rate that appeared to outstrip router hardware capabilities, and the two forward trend lines of routing requirements and router capabilities would meet sometime around 2003 to 2005 (IETF Plenary Presentation, March 2001)

That was some years ago, and now its 2006.

So what's changed since then, and where are we with inter-domain routing?

The first piece of news, and maybe its not so surprising, is that its still a BGP version 4 inter-domain routing world as far as the Internet is concerned, and nothing substantive has changed in the protocol we use today over what was in use over 12 years ago. The larger these systems become the more inertial mass they accumulate, and fundamental change becomes harder to deploy. So I'll hazard the guess that nothing much in inter-domain routing technology is going to change in the near future. While the Internet used to take some comfort in its ability to perform feats of rapid deployment of innovative technologies up and down the protocol stack to address various forms of growing pains, these days the lower layers of the protocol stack are accreting significant levels of inertia, and it's the upper levels of the stack are left to carry the innovation burden. Routing is, perhaps unfortunately, an inhabitant of one of these lower levels of the protocol stack, while much of the innovative agenda is taking place at the application level.

The Border Gateway Protocol really has not changed at all in its almost two decades of deployment. BGP remains a classic distance vector protocol, using an explicitly enumerated path vector as a combined path metric and loop detector. Indeed the introduction of 32-bit AS numbers to BGP could be argued as one of the larger forthcoming changes to the BGP protocol since the introduction of explicit address prefix masks (Classless Inter-Domain Routing, or "CIDR") back in 1994, and even this change is a relatively minor change to the protocol. Given that its just plain old BGP, and given that we're likely to be stuck with it for some years to come, whether its an IPv4, IPv6 or mixed protocol world, than now is as good a time as any to ask how BGP is going, and to see if we can make some guesses as to what kind of routing load BGP will be required to cope with in the coming years.

There are a large number of measurements of the BGP routing table that can describe the dimensions and dynamic characteristics of the inter-domain internet. Here I'd like to concentrate on the use and behaviour of the protocol itself, so in this article I will take a look at BGP across the year of 2005, and see how well BGP fared.

BGP The Protocol

To recap from last month's article, BGP is a distance vector routing protocol, as distinct from a link-state routing protocol or a map-based routing protocol. BGP is a distributed computation that uses address prefixes as its basic unit of routing. Each BGP speaker maintains a set of tables (Routing Information Bases, or RIBs) – one for each BGP neighbour and one for its own internal use for forwarding. BGP keeps a copy of all prefixes and associated routes that have been advertised by its peers (Adjacency-RIB-IN). It selects the "best" of these routes to use for its local forwarding decisions (Local-RIB), and sends a copy of this "best" route to all its peers (Adjacency-RIB-OUT). Like any distance-vector routing protocol, BGP operates as a loosely synchronized distributed computation based on partial information forwarding.

A BGP peer session uses TCP a reliable transport protocol, so that periodic re-flooding of the route tables, so beloved by the interior routing protocol RIP, is not required in BGP. BGP is a far more parsimonious protocol where once a BGP session has been set up and the initial route set is exchanged, then the subsequent protocol traffic is limited to notification of a prefix that is no longer reachable, or when the characteristics of the local "best" route have changed and the local BGP instance wants to inform its neighbouring peers. This information is passed in a BGP update message. This protocol message contains a collection of route attributes, and a list of prefixes that share this attribute set (announcements) and a set of prefixes that are no longer reachable (withdrawals).

If the entire network is perfectly stable, with no changes of any form, then BGP would be a very quiet protocol, with only the intermittent (30 second by default) exchange of keepalive messages to indicate any activity at all. On the other hand, a large dynamic network where prefixes are appearing and disappearing, and where paths are created and lost, such as in the Internet, is capable of generating a relatively impressive set of updates in very small time intervals.

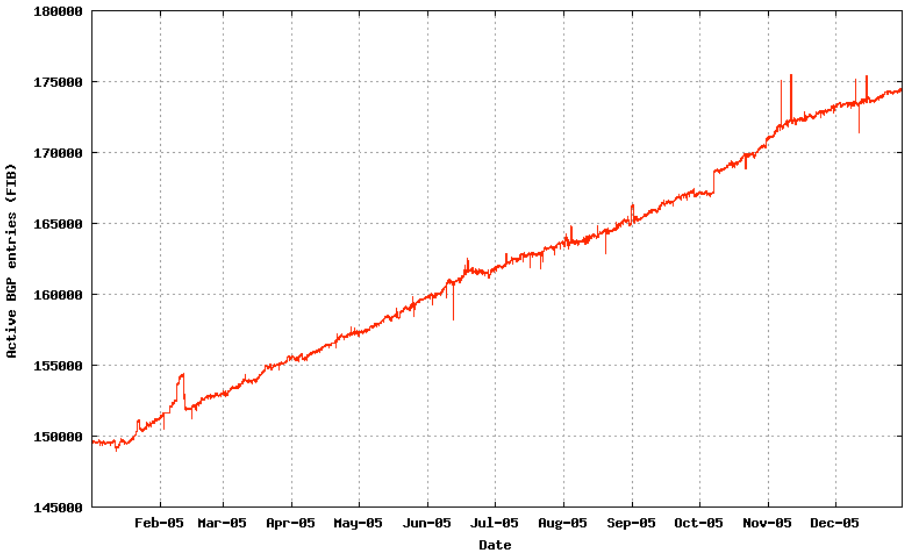
Each received update represents work to be undertaken. The incoming update message causes a change in the Adjacency-RIB-IN. If the information is a prefix withdrawal, then a comparison needs to be made with the local RIB. If there is a match, then all other Adjacency-RIB-Ins need to be scanned and a new "best" route installed into the local RIB, as well as loading new announcement messages in the Adjacency-RIB-OUTs to reflect this local change of best path. If there are no other candidate routes in the other RIB-IN's then the route is withdrawn from the local RIB and a withdrawal message is passed to the BGP Speaker's peers. If the incoming update message is an announcement, then the BGP engine has to update the Adjacency-RIB-IN and then compare this route to the current best path in the Local-RIB. If this new route represents a "better" path, then the Local-RIB is updated and announcement messages are queued in all the Adjacency-RIB-OUTs.

In terms of protocol workload and routing stability its not the size of the BGP routing table that is the critical issue – it's the dynamic characteristics of BGP update messages. The longer the delay in processing update messages the longer the time for the entire system to converge upon a stable routing state that reflects optimised paths across the inter-domain space, and the larger the number of intermediate messages that are generated during this process of convergence, which in turn compounds the problem. At the extreme case the local BGP engine will exhaust its incoming BGP message buffer and fail to process updates. At this stage there is the potential for inconsistent information to be embedded in the routing system, leading to loops and black holes in the routing system. This is the point at while the routing could be said to have "collapsed".

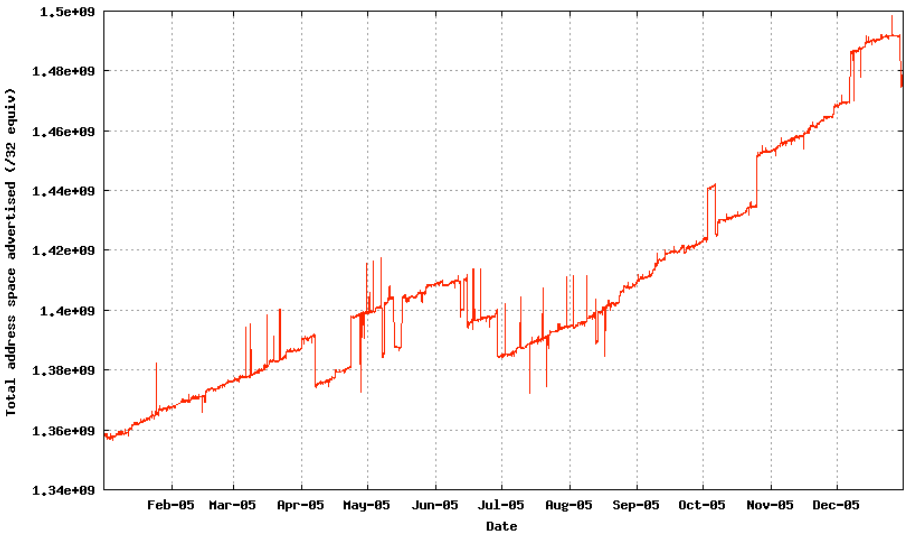
Looking at the BGP update rate, and in particular the relative rates of growth of the BGP routing table as compared to the rates of growth of update messages, and updated prefixes can give us a helpful indicator of the pressures for growth in the routing system, and also an indicator of what size router we'll need to use to cover the Internet's routing system in the coming years.

So what can we say about the Internet and inter-domain routing in 2005? Lets have a look at a number of vital statistics for the year. The following graphs were generated from a stream of one-hourly 'snapshots' of the routing table across 2005, taken from the boundary of AS1221.

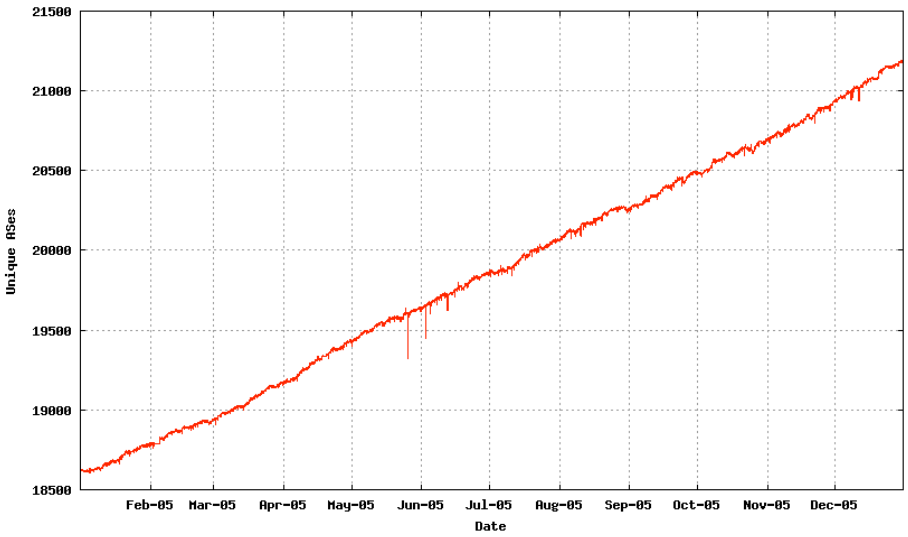
1. The number of IPv4 BGP Prefixes



2. The total span of IPv4 address space in the routing table


Total address space advertised (/32 equiv)
1.5e+09
1.48e+09
1.46e+09
1.44e+09
1.42e+09
1.4e+09
1.38e+09
1.36e+09
1.34e+09
Feb-05
Mar-05
Apr-05
May-05
Jun-05
Jul-05
Aug-05
Sep-05
Oct-05
Nov-05
Dec-05
Date

3. The number of AS numbers in the routing table



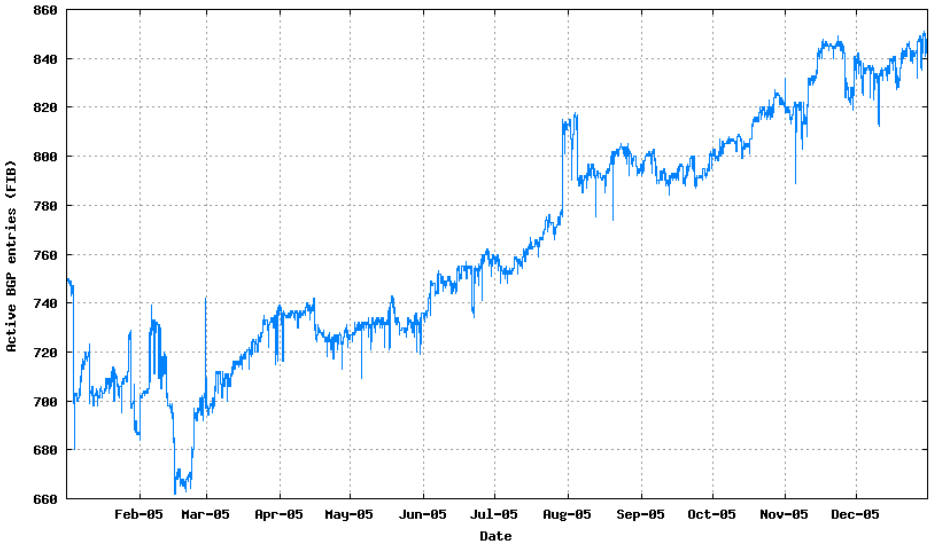
The IPv4 data can be summarised as follows:

	Prefixes	148,000 – 175,400	+18%	+26,900 entries
	Prefix Roots	72,600 – 85,500	+18%	+12,900 entries
	More Specifics	77,200 – 88,900	+18%	+14,000 entries
	Addresses	80.6 – 88.9 (/8s)	+10%	+8.3 /8s
	ASNs	18,600 – 21,300	+14%	2,600 ASNs

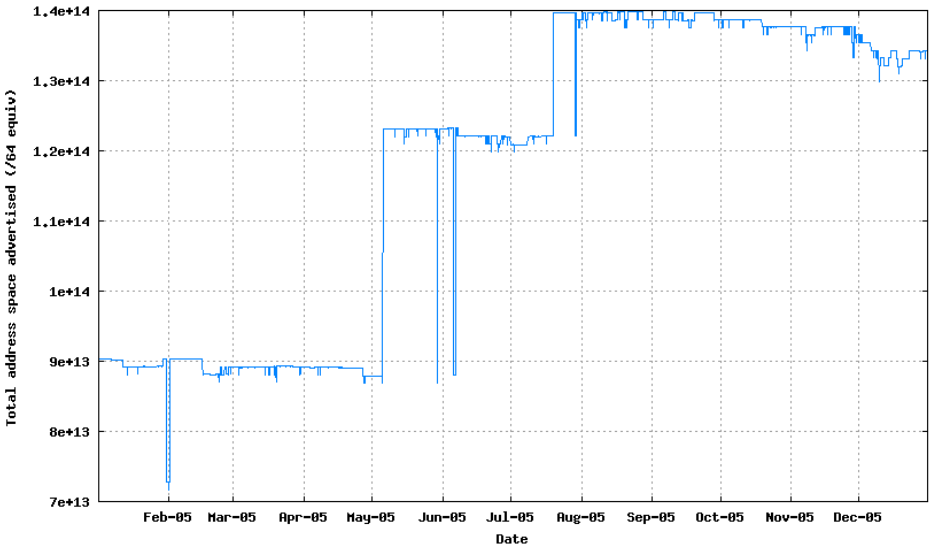
What this table indicates is that for the IPv4 Internet the use of aggregates in the routing system has not improved. The average size of advertisements is getting smaller in terms of address span per routing table entry, the span of originating addresses per AS is getting smaller, the average AS path length is constant at around 3.5 AS hops 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. For a distance vector protocol such as BGP is not heartening news.

A similar exercise has been done for IPv6 for 2005:

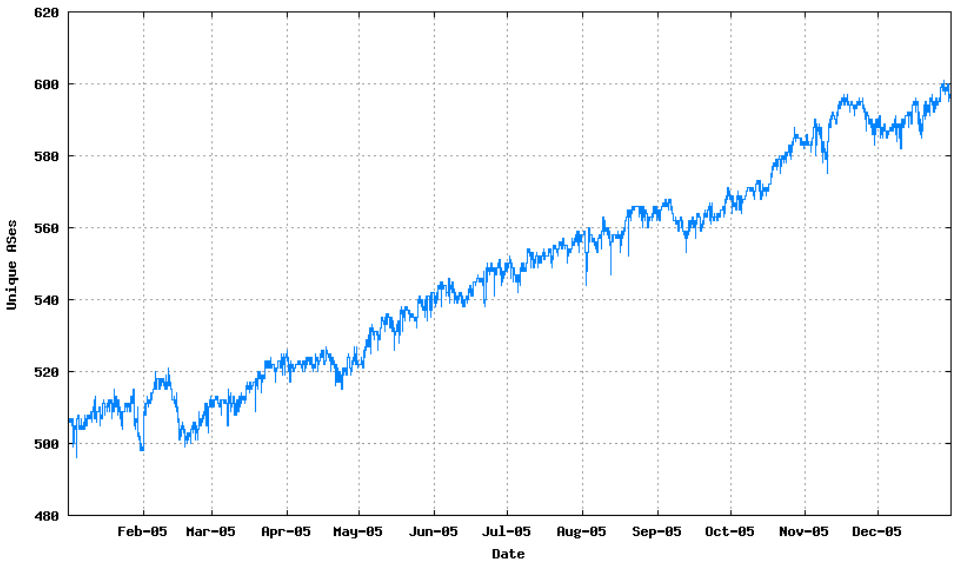
4. The number of IPv6 BGP Prefixes



5. The total span of IPv6 address space in the routing table



6. The number of AS numbers in the routing table



The IPv6 data can be summarised as follows:

	Prefixes	700 – 850	+21%	+150 entries
	Prefix Roots	555 - 640	+15%	+185 entries
	More Specifics	145 - 210	+51%	+65 entries
	Addresses	9.0 – 13.5 (10^{13} / 64s)	+50%	+4.5 (10^{13} / 64s)
	ASNs	500 - 600	+20%	100 ASNs

Its far harder to make generalizations about the trends in the IPv6 network over 2005, as the IPv6 network is simply not large enough to show any overall trend behaviour as yet.

However the IPv4 trends for 2005 are a source of some concern. How big can the Internet grow in the coming years? Will we continue to be able to deploy routers in the default-free routing zone of the Internet that can comfortably route the Internet. Can we add additional functionality into the routing system and still stay within comfortable limits of the capability of the routing system and the routers? If you are an ISP and are considering purchasing new 'core' routers what capabilities should you specify for an operational lifetime of 2 years? How about for the next 5 years? And if you are a router vendor designing routing products for the market 3 or 5 years in the future what capacity should you build into the router? How much processing capacity should you plan for to support default-free BGP? How much memory is necessary?

These are all relevant questions, of course, so the next question is what data can we gather to attempt to provide some likely answers? These snapshots give us some rough information about likely trends, but to provide a more reasoned response its useful to take a more detailed examination of BGP over the year.

Perhaps the best question to pose here is: how have these overall trends manifested themselves in the operation of the BGP protocol?

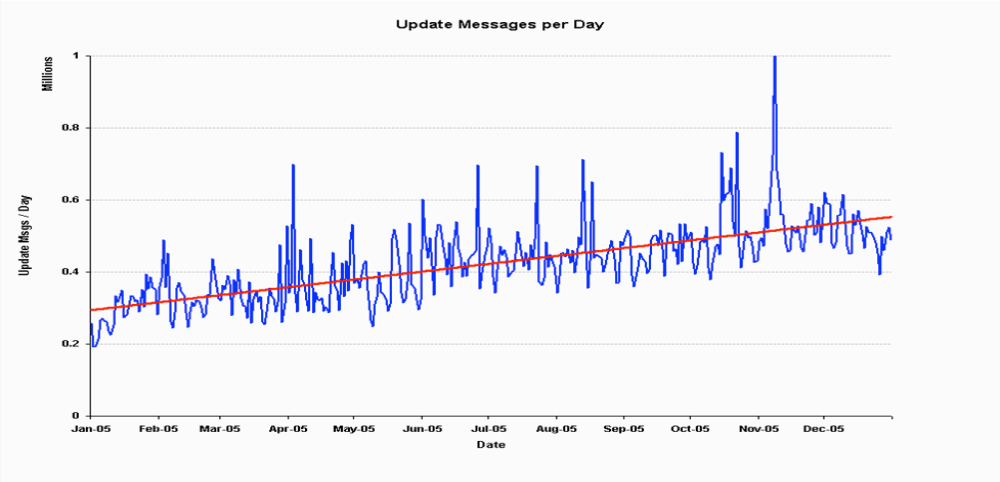
For this exercise a BGP measurement point was set up inside AS1221, and all BGP protocol messages (or "updates") that were passed within that network were recorded with a timestamp on a logging host. The update data was processed to eliminate the internal routing changes and the set of exterior

BGP updates was analysed. Only the IPv4 BGP traffic is reported here. The aim here is to see if there are some trend data that we can extract from the assembled update logs for the year and make some predictions about overall BGP capacity requirements in the coming years.

BGP Update messages per Day

The data set is admitted large – some 146 million BGP update messages were recorded for the entire year. One way of breaking down this data is looking at the number of BGP Update messages per day. On a daily basis the number of update messages appears to have almost doubled for 2005, starting from some 260,000 update messages per day at the start of 2005 to some 550,000 update messages per day by the end of the year. Considering that even by the end of the year there were only 170,000 prefixes in the global routing table, to have this routing population generate 550,000 updates messages per is an impressive achievement. This is a growth rate that is much higher than the growth in the table size. Either the network is far less stable than we'd like to believe, or some other factor is driving up the BGP update rate. The increasing density of interconnection in the inter-domain space may be relevant to this very high growth rate.

7. BGP Update messages per Day



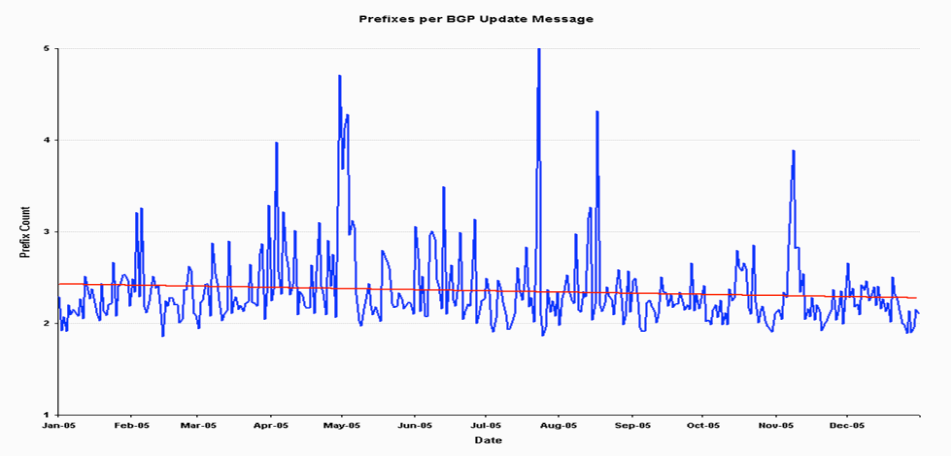
The other interesting observation is that BGP has 'good' days and "bad" days – one day in November recorded 1 million update messages in a single day. This is a very high level of variation, and it indicates a level of instability in the Internet that is not clearly evident at the user level, where most users tend to see a relatively stable and reliable Internet service.

Prefixes per Update Message

Why has the number of Update messages increased so significantly? The daily update rate has doubled over the year, while the size of the routing table itself increased by a much smaller growth factor of 18%. Each BGP update messages contains a number of prefixes. One question to ask is whether the number of prefixes in each update message is increasing or decreasing on average. The daily average number of prefixes per update message is The next area of interest is the average number of prefixes contained in each update message. On average there were between 8.1 and 8.3 prefixes per originating AS across 2005, and if it is really the case that prefixes are managed in a manner such that each AS has a single coherent routing policy then we would expect to see a

relatively consistent number of prefixes in each BGP update message. This is not the case, and the number of prefixes per update message declined over the year.

8. Daily Average number of Prefixes per Update Message



The inevitable conclusion here is that the "unit" of inter-domain routing appears to be converging closer to the level of an individual prefix than to an individual AS. The implication here is that if we wish to contemplate a new routing system based on inter-AS connectivity then we need to understand the extent of the number of unique routing policies that must be encompassed in such an environment, and their dynamic behaviour.

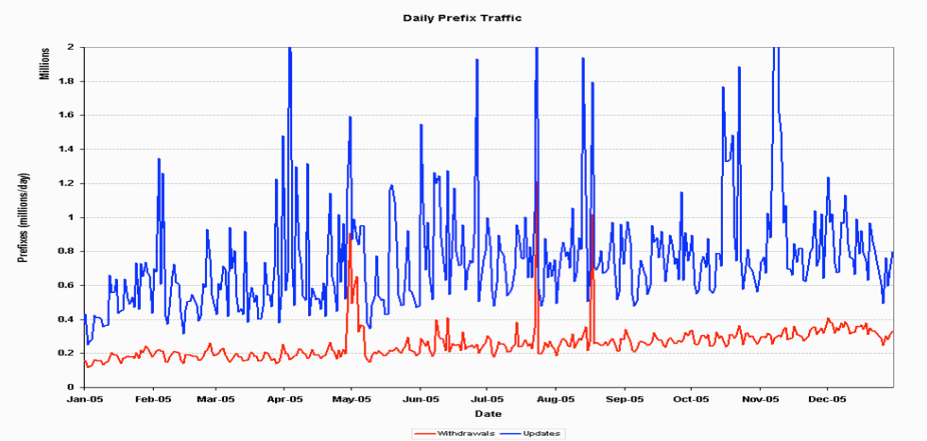
Again the level of daily variation in this average is very high, and while a least squares best fit indicates an overall downward trend for 2005 from 2.4 prefixes per update message at the start of the year to 2.3 prefixes per update message at the end of the year. The high 'spikes' of this measure on some individual days indicates some form of BGP session resets, where a number of peering sessions may have been reset on a day and the resultant reconstruction of the BGP peering session would normally use dense packing of a large number of prefixes in each update message. But there are on average some 8 prefixes per AS, and the average of a little over 2 prefixes per update message appears to indicate a use of fine-grained routing policies at a level finer than an AS. It would appear that the 'unit' of a BGP routing policy is more fine-grained than an AS, and is now heading towards the level of each advertised prefix having individual routing policies and individual attributes. This implies that the efforts of BGP to compress the update load by grouping prefixes into bundles is no longer as effective as it may have been in the past as a measure of assisting in making BGP an efficient routing protocol.

So if we want to look at the trends in BGP, perhaps we should be looking at the update and withdrawal rates of individual prefixes, rather than looking at the level of BGP protocol update messages. So what data is available for the number of prefix updates across 2005?

Prefix Update and Withdrawal Rates

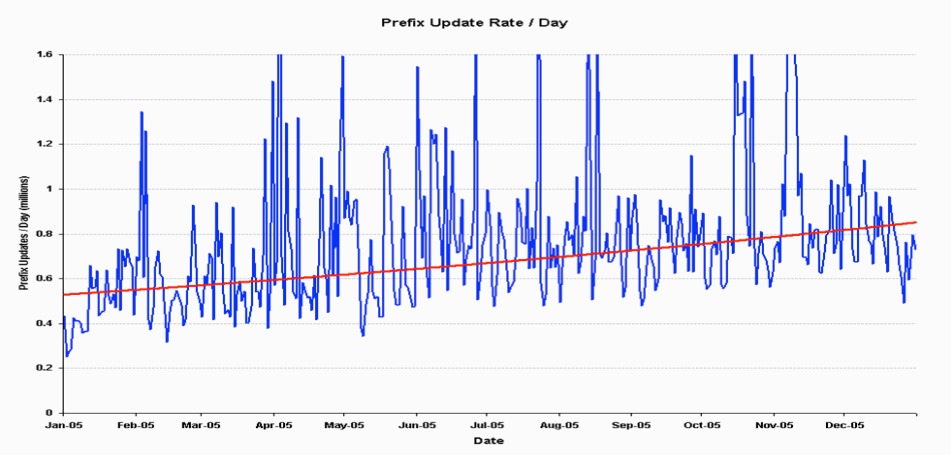
A similar approach has been made to look at the average number of prefixes that are updates each day in BGP. As Prefixes may be withdrawn or updated, the following graph shows the update and withdrawals per day, counting the number of prefixes in each category

9. Daily average prefix count of updates and withdrawals



Again the high level of daily variation is visible, and there is now a clearer indication of when there were full BGP session resets without backup paths (high withdrawal and update counts) and BGP re-routing (high update count without a corresponding high withdrawal count).

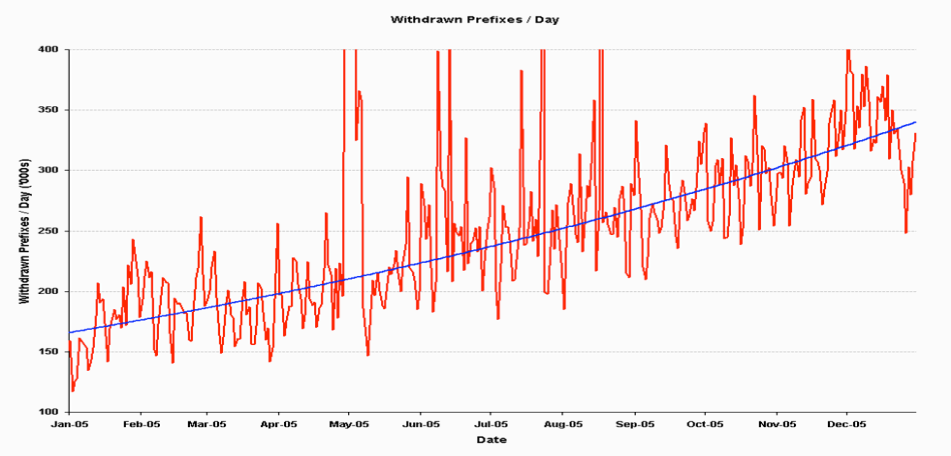
10. Prefix Update Counts



Here the trend across 2005 is visible for updates. The trend line here is an exponential curve best fit, with an overall growth trend from 570,000 prefixes updated per day at the start of the year to some 850,000 prefixes being updated each day by the end of the year. Again that is a very high growth rate, and it should also be remembered that there are, on average some 165,000 unique prefixes in the Internet's routing table. Clearly some prefixes are evidently generating a very high number of updates on a daily basis.

A similar trend is visible in the prefix withdrawal counts for 2005

11 Prefix Withdrawal Counts

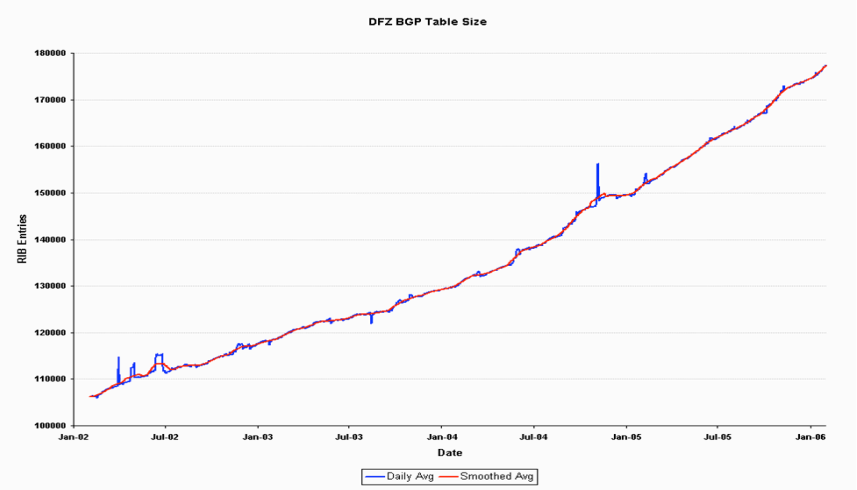


Again an exponential curve best fit trend has been plotted against the withdrawal counts, and the withdrawal count has grown from some 160,000 prefixes being withdrawn on a daily basis at the start of the year to some 340,000 withdrawn prefixes per day by the end of the year.

Trend Behaviour in BGP

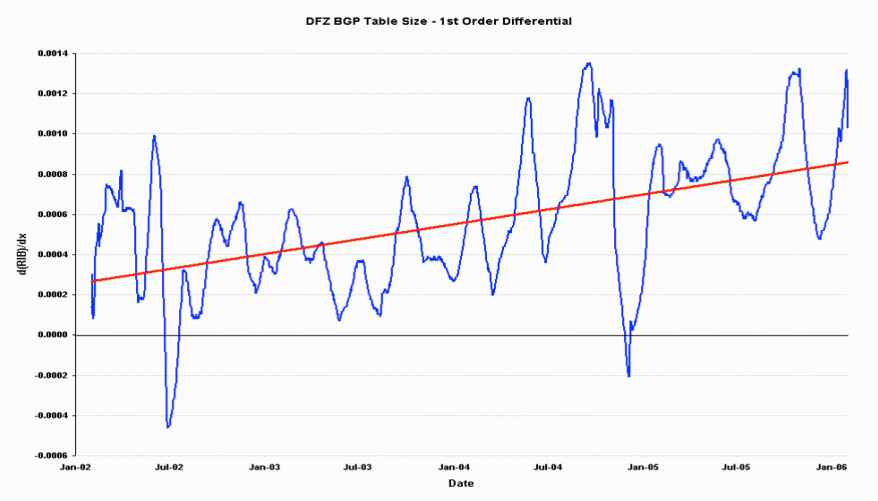
The next question is to relate these prefix update and withdrawal rates against the BGP table size, and look at the likely trends of the load of the BGP protocol in terms of prefix update and withdrawal rates against the trend of the projections of growth of the BGP table itself. The BGP table size over the period from 2002 until the start of 2006 is shown in the following figure.

12. BGP Prefix Table Size



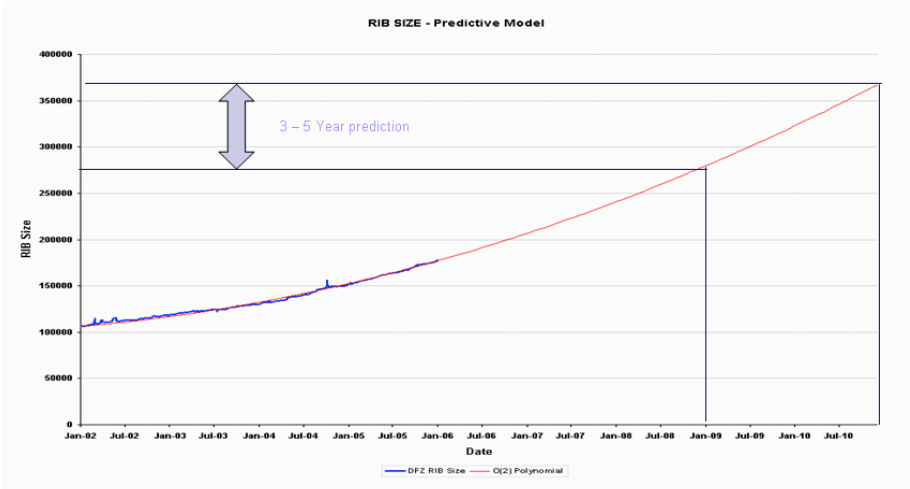
In this figure the raw data of hourly snapshots (the blue line) has been smoothed as part of the first step in generating a trend projection. The next step is to take the first order differential of the smoothed data series

13. First order differential of BGP Table Size



The linear approximation of the first order differential can be fitted to a trend of an O(2) polynomial trend in the BGP table size. This allows a trend projection in the BGP table over the next 3 – 5 years using this O(2) polynomial, as shown in the figure below.

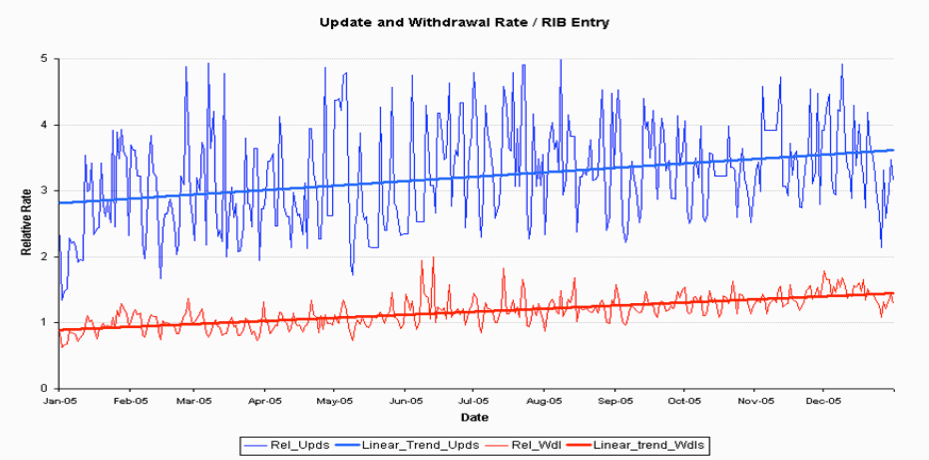
14. BGP Table Size Projection



If current trends in BGP continue for the next 3 – 5 years then this model predicts that the BGP routing table will grown from the level of some 176,000 entries at the end of 2005 to 275,000 entries at the end of 2008 and some 370,000 prefixes by the end of 2010.

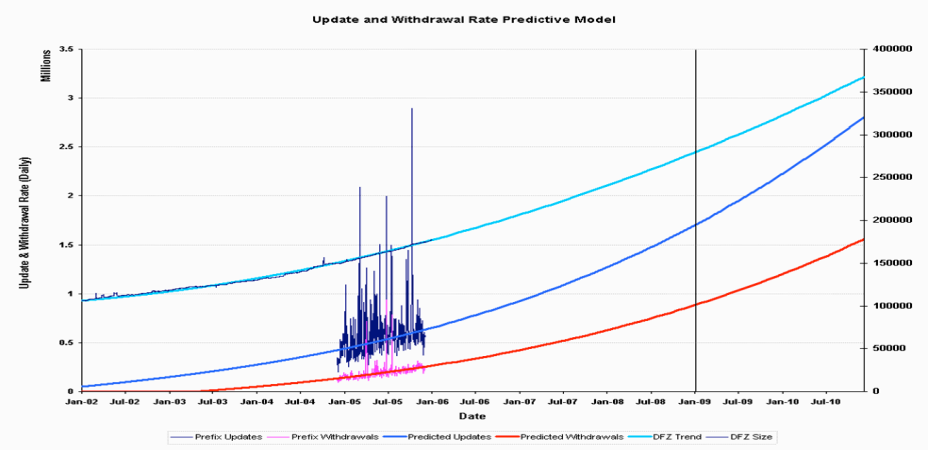
It is possible to use this predictive model to also forecast the amount of BGP update activity. In this model the starting point is the trend of the number of prefix updates and withdrawals per BGP routing table entry across 2005

15. Relative Prefix Update and Withdrawal Rates per BGP Table Entry



These trend lines can then be applied to the BGP projection model, as shown in the next figure.

16. Prefix Update Rate Projection



The projections of BGP activity from this model indicate a growth rate of some 1.7 million prefix updates per day by the end of 2008 and 2.8 million prefix updates per day by the end of 20010. That's four times the update rate as of the end of 2005. A similar growth trend is forecast for prefix withdrawal rates, to 0.9 million withdrawals per day by the end of 2008 and 1.6 million withdrawals by the end of 2010. This implies a CPU processing load that will increase by a similar factor over this 3 to 5 year period.

These projections are summarized in the following table:

Date	BGP Table Size	Daily Prefix Updates	Daily Prefix Withdrawals
End 2005	176,000	700,000	400,000
End 2008	275,000	1,700,000	900,000
End 2010	370,000	2,800,000	1,600,000

Some Observations

Any projection of this nature is ultimately a guess about a potential future here, but irrespective of the precise values in these projections it is evident that there are some accelerating factors within BGP that tend to suggest that the 'load' of BGP, in terms of processing update messages and in terms of processor cycles (update-related processing) is growing faster than the memory requirements and the forwarding decision structure (table size-related aspects). It appears that the combination of finer levels of granularity of routing information in the routing system, denser levels of interconnectivity in the network, greater levels of policy discrimination in the routing system are all combining to create the picture of a system that is increasingly sensitive to perturbation and increasingly difficult to discover and stabilise on a new converged state following each dynamic change. It would appear that these factors of BGP 'load' are growing far faster than the relatively simple metric of number of advertised prefixes in the BGP Routing Table. There is a further multiplicative factor in the load projection that appears to indicate that as the routing system grows, the level of routing overhead grows at a far higher rate.

The other significant factor here is one of peak capacity as compared to average capacity in the routing system. BGP appears to be a very chaotic system in terms of burstiness of traffic, and the peak per-second rate of updates within BGP can be some 1,000 times greater than the daily average. The implication here is that the components of the system should be able to handle very short term peak loads rather than extended average loads in order to preserve any reasonable form of convergence in the routing system.

In addition, how the routing system could cope with adding additional functionality, such as with additional processing overheads relating to improving the overall security in BGP, or with adding further policy-based functions to direct route propagation remains to be seen.

It would appear that if the original question was about the capacity of a routing engine to cope with the anticipated routing load over the coming 3 to 5 years, the basic answer is that very much bigger than what we are using today is very definitely better!

Disclaimer

About the Author

Geoff Huston B.Sc., M.Sc., 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 has been active in the Internet Engineering Task Force for many years.

www.potaroo.net