The ISP Column
An occasional column on things Internet

The State of Inter-Domain Routing

May 2004

In May 2004 an Inter-Domain Routing Workshop was held in Amsterdam, just before the RIPE-48 meeting.

> The website is at http://www.tm.uka.de/idrws/2004/ and the organisers have put the proceedings online. There are some pictures of the non-talk-part of the workshop under http://www.tm.uka.de/idrws/2004/story.html ..

This was an interesting workshop, as it brought together a collection of researchers and network operators to look at both current research activities in inter-domain routing, as well as looking at the current state of the operational practices in this domain.

The focus of the inter-domain routing activity has certainly changed over recent years. Going back to around 1999 it was evident that the issue of greatest attention was a rapid expansion of the number of objects being passed into the inter-domain routing space. This expansion showed strong exponential properties, with a doubling rate as fast as 12 months at its peak. The question at the time was whether this expansion rate was faster than the rate of increase in capability in routers. If we really were expanding this space faster than Moore's law then we were about to face some very challenging questions regarding the ability of deployed routers to actually route the Internet. There were related questions as to whether it was the size of the routing table (memory requirements), the behavior of the protocol in terms of the number of updates (protocol) or the ability of the routers to process these updates and still have a coherent forwarding map (cpu). While the Internet crash in 2001 had some considerable range of negative impacts, one positive aspect was that the exponential rate of inter-domain routing table inflation abruptly stopped.

The inter-domain routing agenda then shifted attention to the behavior of the inter-domain routing protocol, BGP, and some excellent studies were published about the verbosity of the protocol implementations, where the background chatter of routing updates was being amplified by various implementation characteristics into a very noisy din. The studies looked at number of updates that carried new information, versus the updates that effectively resulted in no state change in the routing state. From these studies emerged further work on the performance of BGP in the context of the Internet, looking at the time taken to propagate a new route, the time taken to update a route with a new AS path, and the time taken for a route to be completely withdrawn. The observation was that the time taken to converge to a new state was often much longer than was commonly believed. During a route change there is also an intermediate period when BGP explores a set of false paths.

As well as examining protocol behavior, there has also been extensive work in adding further capabilities to BGP, particularly relating to the use of BGP as a signalling protocol for MPLS in the context of supporting Virtual Private Networks. This has entailed augmentations to BGP in terms of support for multiple protocol families, additional objects and additions to the functional capability of the protocol. Of course, as with many other aspects of today's Internet there has been much activity in improving the resilience and security of BGP and the integrity of the route objects that are carried by the protocol. Recently the TCP reset bug has been associated with BGP sessions, primarily due to the fact that BGP sessions are long-held TCP sessions, and there has been a flurry of updates to BGP peering sessions to add MD5 checksums to the BGP exchange in order to prevent vulnerability to this particular third party attack.

This is a very brief potted history of inter-domain routing activity over the past 5 years. The full scope of activity encompasses a diverse community of folk working at all levels of routing, looking at enhancements and improvements to the current state, as well as looking forward in the near and far term to see what may be useful or necessary to support as inter-domain routing evolves.

A two day workshop could not pretend to cover all of the scope of current activity in the inter-domain routing space, but what it did cover was some areas that I found fascinating for various reasons, and in this column I'd like to share just some of the insights I gained from the workshop.

The first of these is a research program that looks at inter-domain routing from a perspective that is entirely different to the mainstream research approach. Most efforts in this area of attempting to characterize protocol behavior grapple with the entire inter-domain routing space, and attempt to sift through a massive set of protocol updates looking for some particular characteristic of the routing system. This program takes the reverse approach, in that it deliberately generates a series of dynamic changes in a particular routing object at predetermined times. The original approach was that of the "route beacon" where a set of route originators periodically announce route objects into the Internet's routing space at pre-determined times and at pre-determined locations. The route update collectors are then used to gather the protocol updates related to these route object transitions. Here the "root cause" of the routing protocol updates are already known, and the initial study concerning these beacons looked at the behavior of the routing protocol at it propagated these simple state changes across the Internet. The work has already exposed some interesting observations about the dynamic behavior of the system. A simple change of a route at the source can generate an extended sequence of withdrawals and updates at a remote point. Indeed so many updates can be generated that a single root cause action may cause a remote BGP instance to penalize the route object due to an excessively high frequency of fluctuation, and damp the route object. This approach has been taken one step further. A set of remote agents have been configured to send a steady stream of packets towards a multi-homed routing beacon, and the disruptions to the packet stream were recorded as the beacon changed state. The question being investigated was to understand the extent to which this inter-domain convergence issue actually affects data streams. This approach is a very compelling way of looking at routing behavior. Not only does it avoid having to analyzes massive amounts of data and guessing root cause from a huge data set of updates, it allows quite precise experimentation to be performed on the routing system. Ultimately, the objective of all this routing protocol activity is to deliver packets to their intended destination. Measuring how BGP state changes actually impact data flows looks like being a very fruitful area of study.

The inter-domain routing space, and BGP in particular can be viewed as a very large distributed computation. There are at present some 140,000 entries in the routing table and the computation is distributed across some 18,000 Autonomous Systems. Each network administrator can configure their intentions within their local domain, and can instrument their network to show the outcomes of this action. But how to do see further afield, and how can you obtain a view of your incoming traffic before it gets to your domain? The RIPE NCC has been running a Routing Information Service for some years now, and this service gathers a comprehensive set of routing updates from a number of route collectors scattered across the Internet. The Roma Tre University Computer Networks Research Group have teamed with the RIPE NCC to create an "animated, interactive graphical display makes it easier to interpret how BGP updates affect the routing of a specific prefix than by analyzing the updates themselves." This is a Java applet that provides a replay of routing updates, and the

combination of the RIS route collector servers and this visualization tool is one of the more fascinating BGP tools I've seen. If the question is "what just happened then with BGP?" then this is an extremely interesting tool.

These days BGP has turned from a relatively simple protocol to a highly expressive policy language. In practice BGP reaches convergence when it has found a solution to a set of simultaneous constraints. These constraints are not only specified by the originator of a route object, but also involve an interplay with routing domains that act in a transit capacity for the route. This is observed to be similar to solving a polynomial equation where, depending on the degree of the equation there may be multiple solutions. With BGP this is also the case that there may be multiple solutions within a set of configuration constraints, but with an additional constraint that BGP will, in most cases, converge to a single state. There are documented instances of oscillation where BGP flips between two or possibly more unstable states, but here the work was looking at stable states. The workshop heard of a class of BGP scenarios where there are multiple stable "solutions" and a set of trigger actions (link up/down transitions) can 'flip' BGP from one of these states to another. Once BGP has 'flipped' it may not be obvious to any single local network operator working within an AS what actions are required to 'flip' BGP back into the desired state.

The class of configurations examined here appear to involve some preference setting , where a local network administrator is attempting to create a primary / backup scenario across two or more upstream providers. While it may be possible to configure the system to have all the incoming traffic across the primary link and non across the backup, an outage and subsequent restoration of the primary link may not result in moving all the traffic off from the backup link. Sometimes how to correct the problem, without undue disruption to the continuity of service may not be possible, and the BGP system becomes "wedged" into the wrong state.

There is continuing research into attempting to use BGP in subtly, or not so subtly, different ways. One of the perennial questions is whether BGP presents enough information to allow the creation of end-to-end paths that have some defined service characteristic. This is commonly phrased as QoS routing, and the question is normally phrased in the context of support for voice and video trunk streams over the Internet. The objective in this effort is to create a "reservation" along a network path, where subsequent data packets that form part of the QoS stream are switched according to strict parameters of packet loss rates, imposed jitter, and QoS path bandwidth. QoS presents many interesting issues, both in detail with particular approaches, and more generally. At the most general level the question is whether QoS is a case of 'dumb' applications that use a 'smart' network, smart applications that automatically to a 'dumb' passive network, or, perhaps the most confusing case, 'smart' adaptive applications operating across a 'smart' adaptive response network. In the context of BGP, the observation is that BGP is a distance vector algorithm, and within the process of propagation of routing information at each point the best path information is extracted and propagated further, while information relating to all other potential paths is suppressed. While it is possible to envisage QoS paths being supported in a routing protocol that supports a richer set of network information being propagated, the question is whether there is a workable solution within the constraints of BGP. This appears to be a topic of further through, research and experimentation.

The question left on the table in the final session of the workshop was to speculate on the nature of inter-domain routing in 15 years time. Will we see the scaling problem reappear with a new explosion of routing table entries that inevitably challenge the deployed router set with its sheer size? Will we see the same outcome of a massively large routing space, but achieved over the 15 years with inexorable steady inflation of the routing space? Will we see a process of continual 'ornamentation' of the protocol's capabilities, creating an ever-richer set of protocol primitives that allow many more expressive ways to describe routing policy? As a personal; observation is does appear that the larger the deployed base of any system the harder it becomes to make fundamental changes and the more we succumb to the temptation of incremental feature creep instead. The result is ever more knobs, levers and widgets, and 15 years represents a pretty large amount of ornamentation! It may still be BGP, but its not clear that the essential simplicity of the protocol will remain visible within this time span.

There were many presentations and conversations across a busy two days, and the commentary above touches only a small fraction of the program. The area of inter-domain routing remains, for me, an exciting area of activity and the interchange of ideas between researchers and practitioners continues to spark some very interesting and challenging interchanges. Now all I have to do is figure out how to get invited to next year's workshop!

Geoff Huston

Disclaimer

The above views do not represent the views of the Internet Society, nor do they represent the views of the author's employer, the Telstra Corporation. They were possibly the opinions of the author at the time of writing this article, but things always change, including the author's opinions!

About the Author

GEOFF HUSTON holds a B.Sc. and a M.Sc. from the Australian National University. He has been closely involved with the development of the Internet for the past decade, particularly within Australia, where he was responsible for the initial build of the Internet within the Australian academic and research sector. Huston is currently the Chief Scientist in the Internet area for Telstra. He is also the Executive Director of the Internet Architecture Board, and is a member of the APNIC Executive Committee. He was an inaugural Trustee of the Internet Society, and served as Secretary of the Board of Trustees from 1993 until 2001, with a term of service as chair of the Board of Trustees in 1999 and 2000. He is author of a number of Internet-related books.