The ISP Column
A monthly column on things Internet

June 2019
Geoff Huston

Happy Birthday BGP

The first RFC describing BGP, RFC 1105, was published in June 1989, thirty years ago. By any metric that makes BGP a venerable protocol in the internet context and considering that it holds the Internet together it's still a central piece of the Internet's infrastructure. How has this critically important routing protocol fared over these thirty years and what are its future prospects? It BGP approaching its dotage or will it be a feature of the Internet for decades to come?

Routing protocols have always been fascinating to me. How a collection of mindless automata, pairwise interconnected in random ways and each running precisely the same instructions can self-discover not only the topology of interconnection set, but also identify the optimal path across this topology between any two paths, is quite an achievement. The Internet is glued together not by deliberate human design but as an outcome of the chattering of these mindless routing automata. Work on these *routing protocols* pre-dates the Internet by some decades, with papers published between 1955 and 1958 that described a process of repeated iteration of neighbor-to-neighbor propagation of *best* paths until no further refinement of the path selection is possible. If all the nodes in an inter-connected network are running this algorithm, then the result is a consistent set of local decisions that collectively create a collection of loop-free optimal paths between any two points in the network.

The Internet Protocol suite did not define any particular routing protocol. This space was deliberately left blank to allow routing protocols to evolve, and thereby avoiding painting IP into an evolutionary dead end by selecting a routing protocol that may not have been able to evolve to match the future needs of the network. The *Bellman-Ford* distance vector routing algorithm was an early choice for the emerging Internet. While the Routing Information Protocol (RIP) was documented in an RFC in 1988 (RFC 1058), this document notes that: "This algorithm has been used for routing computations in computer networks since the early days of the ARPANET. … It has become a de facto standard for exchange of routing information among gateways and hosts." [RFC 1058] However RIP had its limitations in terms of the overall size of the managed network. The 1980's ARPANET routing system used a two level hierarchy, where a collection of networks (Autonomous Systems) each maintained an internal network topology with the use of *interior routing protocols*, such as RIP, and address prefix reachability for each of these networks was exchanged between networks using an *exterior routing protocol*. An early exterior routing protocol was developed for the ARPANET by BBN in the early 1980's, subsequently documented in RFC 827 as the Exterior Gateway Protocol (EGP). EGP did have a very critical limitation: "It must also be clearly understood that the Exterior Gateway Protocol is NOT intended to provide information which could be used as input to a completely general area or hierarchical routing algorithm. It is intended for a set of autonomous systems which are connected in a tree, with no cycles. It does not enable the passing of sufficient information to prevent routing loops if cycles in the topology do exist." [RFC 827]

BGP-1

In June 1989 Kirk Lougheed and Yakov Rekhter authored RFC 1105, which was the first version of a more general exterior routing protocol that overcame these limitations of EGP. This new protocol, the *Border Gateway Protocol*, or BGP, created loop-free best paths across arbitrarily interconnected networks, and could do so even if the underlying network interconnection topology could allow routing loops to form. This was a routing algorithm that did not define its own transport mechanism and used a conventional TCP session to support information exchange between two BGP speakers.

The protocol allowed for explicit path enumeration as an attribute of an announced address prefix, which circumvented the issue of count-to-infinity loop detection that occurs in classic distance vector protocols. Each network uses a unique Autonomous System Number (ASN) as an identifier, and when a BGP speaker passes a route advertisement to a neighboring network it attaches its ASN to an AS Path attribute of the route. When a

BGP speaker receives a route from an adjacent network neighbor it looks for its own ASN in the attached AS Path attribute. If it finds it in the AS Path, then it discards the advertisement as a potential routing loop.

BGP-3

In October 1991 Lougheed and Rekhter authored RFC 1267, which specified BGP-3, the third version of the protocol. This was a case of further refinements and clarifications of the protocol, without any substantive changes to the protocol's capabilities or mode of operation.

BGP-4

At this time (1991 – 1993) the IETF had embarked on the ROAD (Routing and Addressing) program. ROAD was the name of an effort that was looking for solutions to the evident scaling issues in both the addressing and routing space (RFC 1380). The nascent Internet was already set to exhaust the pool of the Class B IP address prefixes within a few years and the exponential growth in the routing space raised concerns that the size of the routing domain would quickly put it beyond the capability of conventional routing hardware. As one commentator observed at the time, the exponential growth trajectory of inter-domain routing was implying that network operators might need to use supercomputers as core routers within a few years!

Address architecture and routing designs are inter-twined topics, and across the period from 1991 to 1993 many approaches were examined. Some were quite radical in approach, some relying on minor changes to the existing protocols. The outcome of this exercise was essentially a conservative one, where in routing terms the imminent exhaustion of Class B addresses was averted by the expedient approach of dropping the Class concept from the IP address architecture, requiring routing protocols to carry a prefix size with every prefix. Equally, IPv6 was a relatively conservative response to address exhaustion by extending the size of the address fields in the IP header.

BGP-4 (RFC1654, then RFC1771) was a minor change to BGP-3, in that it added a length attribute to the prefix field in the protocol, taking an important step away from the class-based implicit address prefix length paradigm that the Internet had used up to that point. This was the introduction of Classless Inter-domain Routing (CIDR) into the Inter-Domain Routing system.

The impact of this simple protocol change was dramatic. We were fortunate that Erik-Jan Bos, then working with Surfnet in the Netherlands, had started measuring the size of the BGP FIB table in SURFNET's BGP routers every hour, starting in January 1994, so we have an excellent record of the impact of the introduction of CIDR on the inter domain routing system. The size of the routing table fell by 10% from 20,000 entries to 18,000 entries within 6 weeks. Another fall was seen following the July 1994 IETF meeting, and another following the September 1994 RIPE meeting (Figure 1).

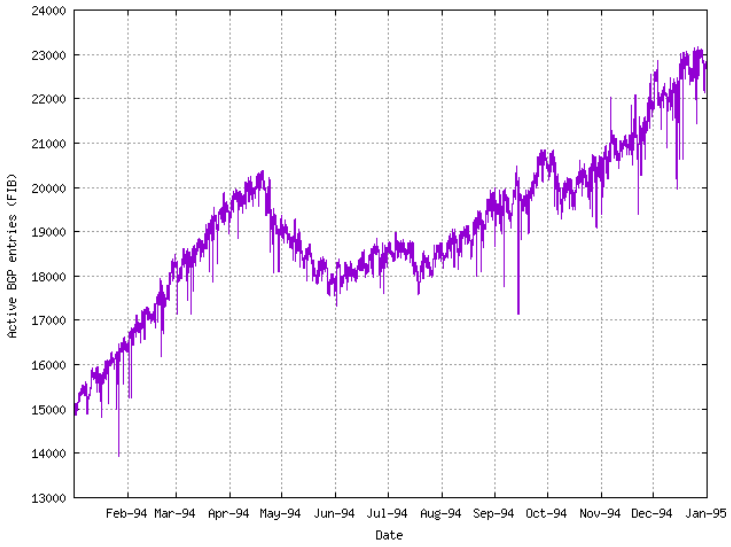

Figure 1 – BGP FIB size in 1994, from bgp.potaroo.net (original data from Erik-Jan Bos)

To illustrate the impact of this minor protocol change, Figure 2 shows the change in the linear model trend of routing table growth in the first quarter of 1994 to the trend of the latter half of 1994.

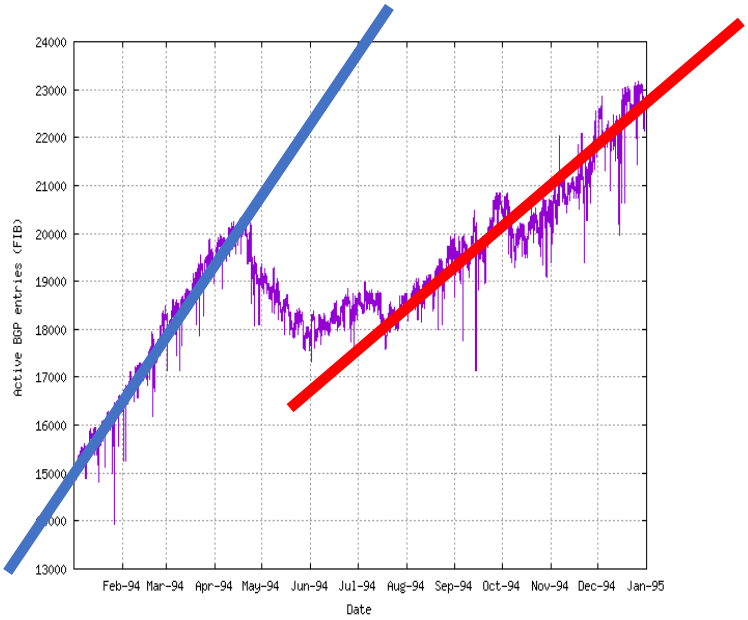

Figure 2 – Trend in BGP FIB size in 1994, from bgp.potaroo.net (original data from Erik-Jan Bos)

The longer-term prospects of averting the worst impacts of the so-called "routing table explosion" were equally dramatic, replacing the exponential growth trajectory of the FIB size of the early Internet between 1990 to 1994 with a linear growth model that prevailed for a further five years, up to the first internet boom and bust in 1999. (Figure 3)

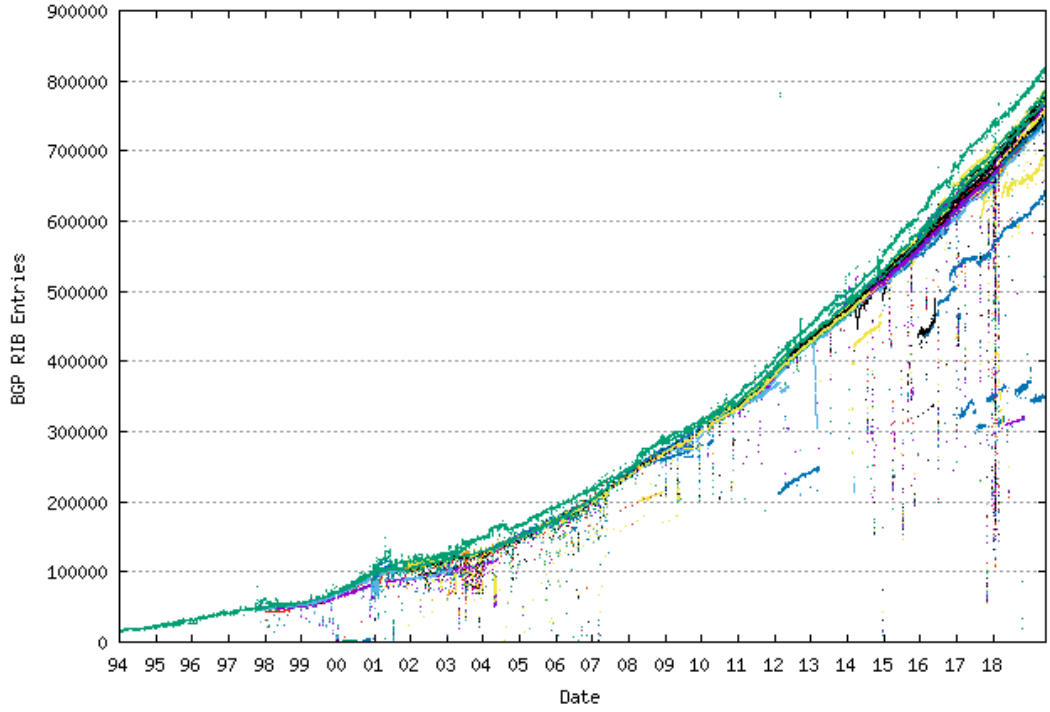

Figure 3 – BGP FIB size in 1994 - 2019, from bgp.potaroo.net (data from Erik-Jan Bos, Geoff Huston and Route Views archive)

IPv6: MP-BGP

The next major change in BGP was the extension to route IPv6 prefixes relatively painlessly. BGP does not apply much in the way of semantics to the address field in the protocol payload. BGP needs to understand what a *match* is when comparing two prefixes, and understand the concepts of *disjoint*, *more specific* and *covering aggregate* as a part of the route selection process, because BGP by default prefers more specific address prefixes over aggregate address prefixes. With these concepts defined BGP can, in theory at any rate, route almost anything!

To support IPv6, BGP was augmented in RFC 2283 in 1996 to support multi-protocol extensions. This extension introduced the concept of the Address Family Identifier and the Subsequent Address Family Identifier to BGP (AFI/SAFI). AFIs are used to identify IPv4 and Ipv6, while the SAFI is used to distinguish btween unicast and multicast. For instance, AFI 1 with SAFI 2 means BGP is carrying IPv4 multicast routing information, and AFI 2 with SAFI 1 specifies a payload of IPv6 unicast routing information. (The AFI/SAFI concept was further generalized to support various forms of logical network segmentation, as is commonly used to host VPN-based services.) An MP-BGP speaker will tell its MP-BGP neighbor which AFI/SAFI combinations it intends to use in the OPEN message at the beginning of a BGP session, and the session will continue if the neighbor indicates that it also can handle this AFI/SAFI combination.

As well as incorporating the capability to multiple protocols in the payload of BGP exchanges there are two other places where this multi-protocol topic intrudes in BGP. The first is the use of next-hop addresses. In BGP this next-hop address is the identification of the exit point of the route from this network, and the assumption is that as the traffic traverses the interior of the network the next-hop is the Interior routing target of this traffic. In theory it is not strictly necessary to use the same AFI for both the route objects and the next-hop targets, although operationally it has been considered prudent to avoid mixing of protocols in this way. The second is the protocol used to support the TCP session. In theory it is perfectly possible to use an IPv6 TCP session to carry an IPv4 BGP session and equally possible to do the reverse and configure the EBGP session to use an explicitly configured BGP next-hop address. Again, operational prudence suggests that this is an unnecessary complication and IPv4 routing information should be exchanged over an IPv4 connection, and the same for IPv6.

BGP Evolution

This picture of an unchanging BGP protocol is not entirely accurate, and the protocol that is used today has had some significant changes to the protocol that was used in 1994, despite the constant use of version number 4 in the protocol. BGP-4 has shown a sufficient level of flexibility in a number of its aspects that allows such incremental changes:

- The initial session negotiation accommodates the use of incorporating new capabilities.
- The ability to define new update attributes and pass them through BGP speakers that do not understand their meaning as opaque attributes has been important.
- The use of TCP as BGP's transport protocol has meant that BGP can be flexible with BGP message sizes.
- The use of TCP allows BGP to assume a reliable hop-by-hop information propagation model and not implement a protocol-specific information reliability mechanism.
- TCP also provides a flow control mechanism, preventing a BGP speaker from overwhelming its neighbor with updates. The receiver can close its TCP receive window when its input buffers have been filled, pushing back on the sender to throttle the message flow rate.
- There is no specific dependency on specific timer values for interoperation.
- The hop-by-hop protocol model coupled with opaque attribute handling allows for various forms of piecemeal deployment of new extensions to BGP within necessitating a flag day or any form of large scale coordinated action by all BGP speakers.

A significant example here of BGP's flexibility is the response to the exhaustion of the 16-bit AS number pool some ten years ago. This was as important to BGP as the forecast exhaustion of IPv4 addresses in the IP space. Use of the hop-by-hop information propagation model, capability signaling in the session negotiation and the use of opaque transitive community attributes allowed a backward-compatible transition of deployed BGP speakers from 2-byte to 4-byte AS numbers on a piecemeal basis, avoiding the need for flag days or other forms of coordinated orchestration within the operational community.

Other changes, such as Add Path and Fast Reroute have also been facilitated by the same underlying flexibility in BGP's protocol design, and there are even efforts to carry link state attributes in BGP as an efficient form of link state flooding.

The key aspect here is the capability negotiation that occurs when a BGP session is fired up. This allows a BGP speaker to offer a set of supported capabilities and allow the neighbor to indicate their level of support for the same capability.

BGP's Weaknesses

A successor routing protocol has not replaced BGP-4 in the past 25 years, and there is no prospect of any such replacement in the foreseeable future. That is not to assert that BGP-4 is free from many issues. The opposite is the case, and the perceived operational problems of the protocol have included:

- Insecurity of both the payload and the sessions.
- Dynamic instability and the consequent inability of the protocol to exhibit rapid convergence.
- Lack of signaling capability within a BGP session.
- Limited ability to recover from loss of synchronized routing state.
- Lack of ability to separate the concepts of topology maintenance, policy negotiation and adequate support for mobility.
- Limited scaling capabilities.

At various times the IETF has supported work to consider a new inter domain routing protocol, such as the Locator/ID separator work in 2006 (following an IAB workshop on Routing and Addressing). From time to time we see proposals to use geo-based addressing schemes and gain aggregation efficiencies through routing these geo-summaries rather than fine-grained prefixes. However, despite this and many other efforts over time, no novel inter-domain routing protocol or even a novel addressing and routing architecture has emerged in the past 25 years that has been a viable replacement for BGP-4. During this same time the size of the set of BGP-routed objects in the inter-domain space has risen from 20,000 objects to a total of some 880,000 objects in the default-free zone of the IPv4 Internet, and the comparable size of the IPv6 routing table is currently at 70,000 entries. The number of distinct Autonomous System numbers in this routing system has risen from 1,000 to 65,000 ASNs. Despite these metrics of significant growth in the set of objects managed by the inter-domain routing system, the BGP-4 protocol itself is essentially unaltered. Even efforts to 'clean up' a large amount of apparently useless routes (where more specific route advertisements carry precisely the same attributes as a covering aggregate) lapse into inattention and disuse. The marginal cost of adding routes to BGP appears to be sufficiently small that there is no pushback. While the routing table contains some 880,000 entries, the information content of that collection of routes could be rephrased efficiently with less than 300,000 entries. But the gains from such an effort appear to be outweighed by inertial barriers and further inflating the size of the BGP tables appears to be the path that has the less common friction.

One explanation of this apparent stasis is that incumbency generates its own inertial resistance, and the larger the system the greater the level of this inertial resistance. This view leads to a conclusion that the Internet is now too big to contemplate a change to its inter-domain routing protocol, and that BGP will remain the Internet's inter-domain routing protocol for the foreseeable future.

But the inexorable growth in the size of the routed space leads to some unrealistic projections. The issue here is that adopting CIDR as the solution to the routing explosion issue was not in fact a solution at all. In BGP-4 the semantics of addresses and the related routing system was unaltered, and for routing the same risks of the routing table expanding beyond the point of viability is still as big a risk today as it was back in 1993. The IPv6 address plan used by the Regional Address Registries appear to set the minimum allocation prefix size at a /32 for each registry, and there are 4.3 billion such prefixes in IPv6. It is just not possible to conceive that BGP as we know it could cope if it had to manage 4.3 billion prefixes. Address exhaustion in IPv4 is creating similar route fragmentation pressures, and the size of the IPv4 routing table continues to grow despite the hiatus in the supply of 'new' addresses. The growth is due to the ever-decreasing size of routing advertisements in IPv4, and here the prospect of achieving a routing table of a billion or more small prefix entries should not be completely discounted. These large numbers seem, in some intuitive sense, to be well beyond the capabilities of the protocol. Could we ever contemplate a session restart if that implied reloading billions of route entries? How would a high-speed router be designed if the per-packet decision space is encompassed by a decision tree containing more than a billion entries? This is not just an IPv4 issue. It is also observed that one half of the IPv6 routing table uses /48 more specific announcements. There are some 281 trillion (10^{12}) such /48 prefixes in IPv6, and again this is a number that is way outside our imaginings of BGP's capabilities, and way outside our current concepts of router design.

However, the growth of the BGP domain is relatively slow and so far we have been able to deploy equipment that can easily handle the load that is associated with some million or so routing entries, and it is not impossible today to conceive of routing system that could manage some ten million such entries without resorting to an entirely new approach to routing and packet switching. At this point in time, the prospect of a BGP melt-domain appears

to be a theoretic one, and operational reality points to BGP being used to route the Internet for many years to come.

BGP Design Expectations vs Deployment Reality

As part of this review of thirty years of BGP, I'd like to look at a comparison of our assumptions and expectations back in the early 90's with the learning experience gathered over the ensuing 30 years. There are some aspects of BGP where the initial design assumptions of BGP appear to be at some difference with deployment requirements. Here are some examples of this variance.

1. **Session Longevity**
   Design: The BGP TCP sessions were never intended to be long-lived. The expectation in the design was that sessions would be restarted in an integral of days or weeks.

   Deployment: BGP sessions are kept up as long as possible. Session lifetimes are measured in months or years. The very high cost of session restart means that network operators strive to maintain session integrity. The result is that there are an unknown number of 'ghost' routes in the routing system where the withdrawal of routes has not propagated across the entirety of the routing space. Ghost Routes were identified in the early days of the IPv6 routing table, when the table was sufficiently small to allow detailed examination of the history of all routing entries. Regular route flushing would address this behavior, but the original design parameters included an implicit assumption of regular session restart

2. **Session Security**
   Design: The protocol is intended to pass public routing information, so there is little to be gained by attempting to secure the BGP session.

   Deployment: BGP sessions can be readily disrupted by RST injection into the TCP stream or even session hijacking. Low impact solutions (such as TTL hacking) and more complex solutions (TCP MD5) are both used in the network to protect the session, but the basic operational approach is to avoid multi-hop eBGP sessions wherever possible, and limit BGP sessions to direct interconnection wherever possible.

3. **Payload Security**
   Design: BGP was conceived as a hop-by-hop protocol and no form of content security was incorporated into the design.

   Deployment: BGP shows a constant stream of routing mishaps. Some of these are the result of deliberate efforts to inject false information into the routing domain, and BGP remains vulnerable to such efforts to distort the routing space. Other forms of synthetic information injected into the routing system (such as AS Path poisoning) are used by operators to implement their traffic engineering or policy requirements, and the distinction between hostile injection of routing information and the intentional manipulation of routed objects is at times challenging to define.

4. **Convergence Behaviour**
   Design: The protocol was designed to minimize the number of updates generated as the system hunted for a stable converged state.

   Deployment: Convergence speed is considered to be more important than update message volumes in certain contexts, and vendor implementations vary. The result is somewhat chaotic in terms of protocol convergence performance.

5. **Error Handling**
   Design: The protocol had no error handling capability. Conditions that generated error states, such as unknown messages or inconsistent state transitions in the BGP Finite State Machine cause the BGP speaker to drop the session.

   Deployment: Operational considerations require that session shutdown be avoided wherever possible, and that the impacts of session restart be mitigated wherever possible.

6. **Traffic Engineering**

Design: The protocol has very rudimentary capabilities to control the distributed route selection algorithm.

Deployment: Some 50% of the objects in the BGP routing table do not add to the basic reachability of advertised address space, but instead attempt to qualify that reachability by expressing a preference for certain forwarding paths.

Reasons for BGP's Longevity

The key question here is perhaps less about those areas where the protocol design is not well aligned to operational requirements, but more what aspects or aspects of the design have allowed a 30 year old protocol designed to manage a topology of some 500 networks and 10,000 address prefixes scale up to manage a topology of 70,000 networks and rapidly approaching 1 million address prefixes.

Three technical aspects of BGP appear to be important for BGP in providing flexibility to adapt the protocol to meet new requirements.

Firstly, BGP is a distance vector protocol which forces it to be a hop-by-hop protocol. Hop-by-hop protocols are often more flexible in supporting partial deployment of capability, in so far as a new behavior needs only to define how to tunnel though sequences of "old behaviour" in a transparent manner. This permits innovations to be deployed in a piecemeal and loosely coordinate manner, which matches the characteristics of the inter-domain operational community. In other words, BGP can evolve to suit changing requirements, and do so in a manner that does not require universal adoption, flag days or any other form of internet-wide coordinated actions.

Secondly, BGP's choice to use TCP as its transport protocol provided both reliable information transfer and elasticity in the definition of protocol objects. This design choice implied that BGP could safely assume that all information sent to the remote BGP neighbour was received and processed by that neighbour. This greatly simplified the protocol and had a major bearing on the scalability of the protocol as well.

Thirdly, and perhaps a little more controversially, I believe that the use of the Minimum Route Advertisement Interval (MRAI) timer has been an important factor in BGP's continued ability to route an ever-larger Internet. A denser mesh of interconnectivity would normally drive a distance vector protocol into a large volume of incremental updates as updated reachability information travels at subtly different speeds across the higher-interconnected network. The MRAI interval damps that high frequency instability trading a longer time to converge against a highly damped protocol update load.

The other aspect of BGP's longevity in the field is that BGP is extremely well suited to the business environment of inter-provider interaction. A network takes in reachability, makes a set of internal decisions about which prefixes and paths to use based on local policy and propagates the outcome. A network does not have to expose its policies to any external party. BGP in this role is a negotiation protocol, where route advertisements reflect a network's import preferences and the selection between otherwise equivalent route advertisements reflect a network's export preferences.

BGP has also benefitted from the business environment. It appears that many networks prefer to avoid long chains of connection and prefer to obtain service either directly from a so-called Tier-1 provider, or as close as possible to a Tier-1 provider. The resultant network is not "long and stringy" but instead its "short and dense". Short dense networks behave very efficiently in terms of convergence performance. As the time for a routing uptake to converge to a stable state is dependent on the average AS path length rather than the number of interconnected networks, then this connection pattern allows the network to grow through increasing density, and thereby preserve some consistency with convergence performance.

The effects of this "short and dense" interconnection preference is shown in Figure 4.

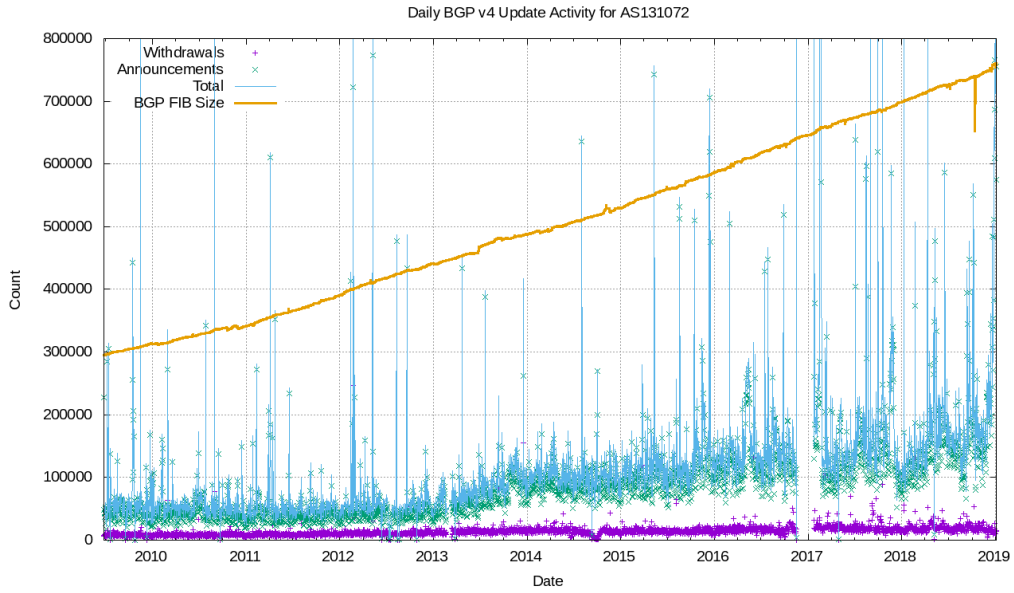

Figure 4 – BGP Table size and update and withdrawl daily activity

While the number of FIB entries has almost trippled in the past decade, the number of updates hads grown at a far slower rate, and the number of withdrawals is relatively constant. The capped update rate is partly due to a network whose average AS path length has remained constant. This figure also shows a relatively constant rate of withdrawals over time. The reasons for this capped behaviour is not clearly understood.

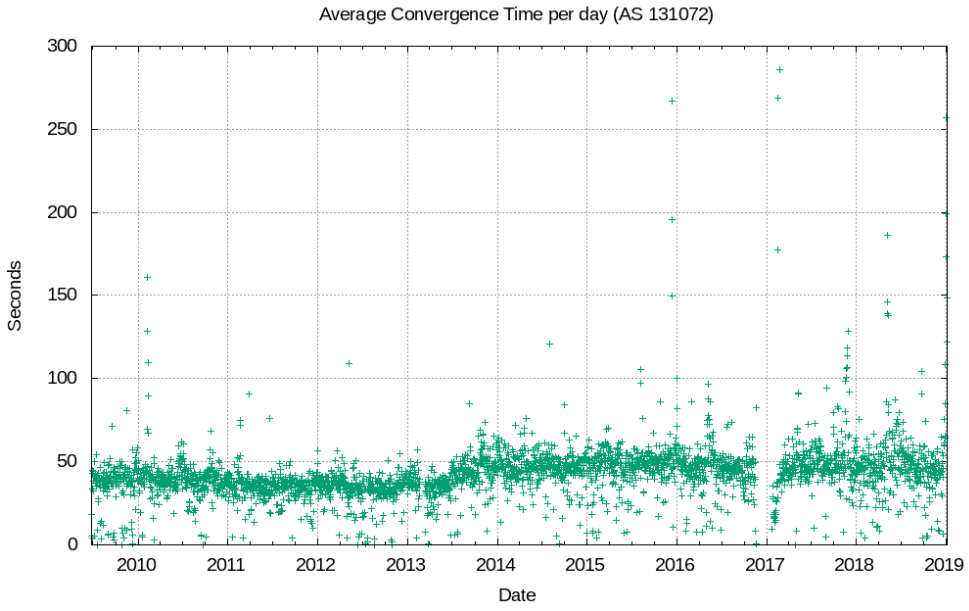

Figure 5 – Average BGP convergence times

The "short and dense" network produces the performance outcome shown in Figure 5. The average time to converge in BGP has remained constant for the past decade. Again, this is related to the stability of the average AS Path length in BGP overtime, which itself is related to the business model of connecting into the network as close as possible to the so-called Tier One 'core' of the network.

BGP Looking Forward

BGP might not be the absolutely perfect interdomain routing protocol for the Internet, but its longevity is a testament to the observation that the effort required to address its shortcomings through incremental changes to the protocol is far less effort than would be required to define and deploy an entirely novel inter-domain routing protocol.

Will BGP be around for another 30 years?

That's a tough question, but the odds are in BGP's favor. The cost of change to something as fundamental as the internet's routing protocol are extremely high, so any new protocol will need to generate massive improvements in cost and performance to overcome the stasis of BGP's entrenched incumbency.

It is also the case that the business models of interconnection and BGP are now closely intertwined. Any replacement interdomain rioting protocol will need to support the same interconnection attributes, including selective obscurity of local policy settings, and the negotiation of tensions between the import and export preferences of each network. In other words, any successor protocol to BGP looks like it had better behave in the same way as BGP behaves, at least while the current business models of Internet service provision still hold sway!

It may well be that BGP will now last for as long as the Internet will last.

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.

www.potaroo.net