Design Expectations vs. Deployment Reality in Protocol Development

The Border Gateway Protocol, 25 years on

Geoff Huston
APNIC

Some 25 years ago, in mid-1994, the IETF published RFC1694, the initial specification of the inter-domain routing protocol BGP-4. A few months prior to the RFC publication date, in April 1994, BGP-4 was deployed on the Internet.

BGP-4 was one of the products of the IETF's ROAD (Routing and Addressing) program, which was looking for solutions to the evident scaling issues in both the addressing and routing space (RFC 1380). BGP-4 was a minor change to BGP-3 in that it added a length attribute to the prefix field in the routing protocol, taking a step away from the class-based address prefix 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 change was dramatic.

We were fortunate that Erik-Jan Bos, then of 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 o fate 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).

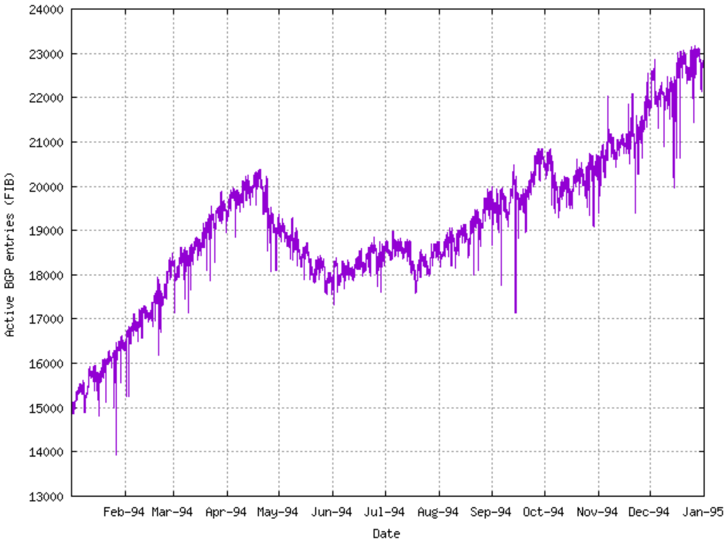

Figure 1 – BGP FIB size in 1994, from bgp.potaroo.net (data from Erik-Jan Bos)

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 until the first internet book in 1999. (Figure 2)

Effective as it was, the adoption of CIDR was not considered to be the final solution to the concerns over the growth pressures being placed on the routing system, just as NATs were never thought of as a long-term response to IPv4 address exhaustion issues. CIDR was a means to buy more time, both in terms of the pace of address consumption and bloat of the routing table space. CIDR would give the IETF additional time to work on successor inter-domain routing technologies that would provide stable and scalable routing platforms looking forward.

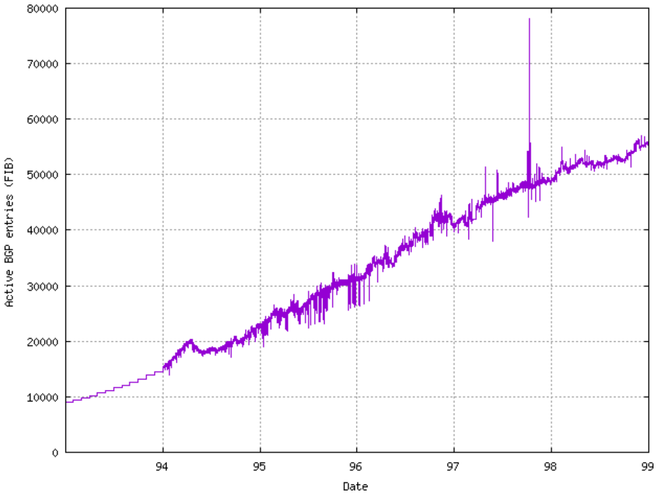

Figure 2 – BGP FIB size in 1993 - 1998, from bgp.potaroo.net (data from Erik-Jan Bos and Geoff Huston)

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 visible future. That is not to asset 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 signalling capability within a BGP session
- lack of ability to separate the concepts of topology maintenance, policy negotiation and adequate support for mobility

At various times the IETF has supported work to consider a new inter domain routing protocol, including as part of the the ROAD work in 1992 - 1993, and the Locator/ID separator work in 2006 (at the IAB workshop on Routing and Addressing).

However, despite these and other efforts over this time, no novel inter-domain routing protocol or even a novel 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 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 Internet. 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.

One reasonably 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 would lead to the 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.

That picture of a constant BGP-4 is not entirely accurate, and the protocol that is used today has had some significant changes to the protocol that was used in 1994. 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
- no dependency on specific timer values for interoperation
- a hop-by-hop protocol model

A significant example here of BGP's flexibility is the response to the pending exhaustion of the 16-bit AS number pool. Use of the hop-by-hop information model, capability signaling in the session negotiation and the use of opaque transitive community attributes allowed a 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.

BGP Design Expectations vs Deployment Reality

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 budget solutions (such as TTL hacking) and more complex solutions (TCP MD5) are both used in the network to protect the session.

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, 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 FSM 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.

BGP's Longevity

The key question here is perhaps less 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 25 year old protocol designed to manage a topology of some 1,000 networks and 20,000 address prefixes to manage a topology of 70,000 networks and rapidly approaching 1 million prefixes.

Two 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.

Secondly BGP's choice to use TCP as its transport protocol provided both reliable information transfer and elasticity in the definition of protocol objects.

BGP is by no means the 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.