March 2022
Geoff Huston

IETF 113 – IEPG Meeting

The IETF met in a hybrid format in March 2022. Here are my impressions from the IEPG meeting at the start of the week.

The IEPG (www.iepg.org) meets on the Sunday at the start of the IETF week. If there is a theme for the diverse collection of presentations here it is perhaps a focus on operational topics, but the particular selection of subjects in these sessions can be quite diverse.

IPv6 Extension Headers

The first couple of presentations at this meeting were on the topic of IPv6 and the handling of Extension Headers. The extent to which various IPv6 extension headers present issues to both network elements and end hosts was noted in RFC7872, published in 2016. That study involved sending various forms of IPv6 packets to well-known servers and noted a fragmented packet drop rate of around 30% for web traffic. There has been sporadic interest in using IPv6 Extension Headers for various forms of signalling, as was the intent of having these headers integrated into the specification of IPv6 in the first place. However, there is an underlying nagging suspicion that once you leave a carefully controlled private IPv6 deployment and enter into the space of the public IPv6 Internet, then all forms of Extension Headers just do not work.

APNIC Labs is performing a similar experiment to that described in RFC7872, but in the opposite direction, by sending specifically crafted IPv6 packets in the direction from the server to the end client (the opposite of the direction measured in RFC7872). In 2017 this experiment measured an average fragmented packet drop rate of 21%. A repeat of this experiment in 2021 and 2022 is showing a marked improvement, with an average 8% packet drop rate when the packet is fragmented. This is a large-scale measurement that spans a collection of some 4.4M end points each day, using an online ad campaign to enrol endpoints into a measurement test. The measurement also includes the measurement of the "atomic fragment" drop rate (an "atomic fragment" is a packet that includes the fragmentation header, but the entire unfragmented IP payload is included in the packet). The drop rate for atomic fragments is around 6%. The average drop rate for packets whose initial fragment size is greater than 1,360 octets is considerably larger than smaller packets. The reason for this elevated drop rate is not clear. The experiment also measures the end-to-end delivery of unfragmented packets with a padding Destination Option Extension Header. The end-to-end drop rate for Destination Option packets was 94.5%. A similar padding Hop-by-Hop Extension Header is also measured and the drop rate is seen to be an unexpectedly high 99.5% in this test.

A different methodology is used in the JAMES measurement. Here some 13 VM hosts are used, and each host sends trace route-like packets towards the other 12 hosts. The packets can be crafted with various forms of extension headers, and the progressive traceroute-like probing should allow a first level diagnosis of where the drops are occurring. The results differ markedly from the APNIC data. Atomic fragmentation shows a 30% drop rate, which is considerably higher than the APNIC data. The Destination Options show an escalating drop rate above 64 octets, and Hop-by-Hop were reported as

"unreliable" without any drop metric being reported. The major issue here is that the small probe set provides only a highly limited set of transit paths to be tested, and the measurement is only about network-level drops, not end client drops. Given that these are VM end points in the tests I suspect that the paths being tested here could well be data center-to-data centre, and completely avoid server-to-client hops. Any samplings-based measurement has the issue of having to generalise from the sample set to the general case, and the smaller the sample set, and this is a very small sample set, the greater the leap of extrapolation from the sample set to any broader inference.

The high order takeaway from both of these measurements is that all IPv6 extension headers are unusable in the public IPv6 Internet and proposals to carry information in such headers are completely unrealistic in this context. The conclusion is simple: If you are wanting to deploy a robust service on the IPv6 public Internet then you should avoid any use of Extensions Headers.

BGPsec Crypto

There was an interesting analysis on BGPsec algorithm efficiency. The analysis noted that BGPsec proposes the use of SHA2 for hashing. This is not expensive computationally, but it operates on 32 byte blocks with a granularity of 4 bytes. This algorithm vectorises well. P-256 for signing and verification is computationally expensive, and verification is significantly more expensive than signing.

The analysis noted that the most efficient memory alignment of blocks for SHA2 processing is not the wire format used by BGP. The presentation suggested that re-aligning the information on the wire in BGP could offer some performance improvements in hashing, signing and verification by routers run BGPsec. This is not the first time we have headed down this path. Back in 2003 pr so when the IETF was looking as the costs of running sBGP there were a number of proposals that attempted to reduce the crypto processing load on routers, as the routers at the time simply could not keep up with the processing load. At the time there was the notion that silicon would ultimately deliver, and Moore's Law would deliver higher capacity processing capabilities in routers that could manage the crypto load. What happened was that BGP has kept on growing as well, and the crypto load on today's routers still exceeds the capacity of deployed routers.

We ask a lot from our security tools. We want them to add zero cost to our operational environment. We want them to not introduce any new vulnerabilities or risks. We want them to operate so quickly that there is no incremental time penalty when we turn them on. And we want good, fast and cheap. And two out of three is just not enough.

Is BGPsec broken?

From a narrow perspective of protocol correctness where a BGP speaker is wanting to assure itself that the AS Path in a BGP update has been propagated by routers in each of the ASes in turn, and the path has not been manipulated or unaltered in any other way, then no. BGPsec is not broken.

From the perspective of being a robust, efficient, cheap and useful tool that does not add any incremental cost and risk to the network and is capable of delivering useful benefits even in the face of piecemeal deployments, then, yes. BGPsec is about as broken as sBGP was twenty years ago.

I suspect that the efforts to salvage BGPsec are way beyond rearrangement of information within the wire formats of BGP protocol elements. I suspect that meeting the objectives of good, fast and cheap for securing BGP require us to re-think what we want from such security tools.

Regional Internet Blocking Considerations

I suppose someone had to go there. With all the current talk of applying sanctions in today's badly messed up world the Internet is an obvious place to apply such sanctions. This presentation was not about blocking measures and their likely efficacy, or not. This was a plug for an Internet draft, draft-giuliano-blocking-consderations that goes into further detail.

It's a difficult conversation for many reasons. A lot depends on context I guess.

We've seen in various natural disasters that disconnecting the physical wires can be effective in isolating a community. The earthquake near Tonga in January of this year took out that island's connectivity for a few days as the submarine cables that the island used were snapped by the earthquake. However, there are many ways a regional community can diversify its connection options, and few large communities, even island communities such as Australia or New Zealand, depend on just one wire, or even a handful of wires. Routing is another potential option, and existing upstream and peering sessions can be shutdown. Of course, the same observation about diversity of connection options applies equally to the routing level as it does to the physical plant, and the consequent game of cat and mouse chasing through the diversity of options in the routing system probably has no end point! Then there is the DNS. And the X.509 certificate system. And routing registries, prefix filters and RPKI.

However, it would be remembered that the objective for much of the Internet infrastructure was not to construct single potential control points, but to build a high diverse system that had no such critical points. Single points of failure in such systems are an anathema to the robust design of common infrastructure of our public services and attempting to install such points of control seems somewhat counter-intuitive. I find that I can agree with one important point in the document: "The Internet is, by design, a decentralized system of interconnections. Thus, it is nearly impossible to completely block Internet access for a region. Simply put, there will always be ways to circumvent any blocking attempts by sufficiently motivated parties."

An RPKI Beacon

How can you evaluate the extent of update of a technology on the Internet? One approach uses a technique of "beacons". The first step is to create a test condition and a control, and the second step is to direct transactions to the test condition to see how many transactions succeed or fail.

In the context of ERPKI through entail advertising four address prefixes, two each in IPv4 and IPv6. One prefix would be advertised in such a way that Route Origination Validation would classify the advertisement as "invalid", with the other would be "valid". A variation of this would also use more specific prefixes. At this point Willem Toorop's group at the University of Amsterdam is looking at various alternatives involving RIPE Atlas probes and a set of beacons, working out how to create a useful tool.

We have been down a similar path at APNIC a couple of years ago, with a similar objective of trying to measure the extent of use of RPKI and Route Origination Validation over time. We devised a slightly different approach using a single IPv4 prefix, a rudimentary anycast setup with 4 locations spread across the Internet, and we used time to flip between "valid" and "invalid" states. To test this approach, we use an online advertisement campaign to direct users to retrieve an object that sits in this beacon. That way we can enrol a very large collection of test points. The data we generate using this system can be found at https://stats.labs.apnic.net/RPKI.

Measuring behaviours in the routing system has a number of idiosyncrasies, as compared to measuring end-to-end behaviours in the network. For example, if a transit network performs Route Origination Validation and drops invalid routes, then all the client networks that use this network as a transit will also behave as if they were performing drop-invalid as well. If the networks that are attached directly to the beacon points perform drop-invalid than it will appear that every network is performing drop-invalid! With the APNIC measurement we concluded that the best form of beaconing was by using a relatively dense anycast network, and in our case we use the Cloudflare system to host beacons.

Perhaps there is a deeper question here as well. What's the desired end point here? Do we really want every network to perform drop-invalid? Or do we simply want to prevent users from reaching prefixes

that are advertised in a ROV-invalid manner? In the latter case then perhaps it's far more critical to look at the uptake of RPKI Origin Validation in transit networks.

Higher Levels of Address Aggregation

Back in 2006 there was an emerging belief that the internet's routing space was growing faster than we could construct routers to cope. The IAB convened a workshop on the topic (RFC 4984). There was much discussion about the basic cause of the issues, including the semantic overloading of IP addresses with both location and identity roles, traffic engineering issues, policy overlays, and an address overlay was independent of topology, to name a few factors at play.

Part of the perceived problem at the time was the rampant use of more specifics in routing. Of then then 200,000 routes in the default-free routing tables around one half of them were more specifics of existing routes. There were many proposed responses at the time, but one that was interesting was to compare the telephone numbering system with the Internet's addressing and routing system. The idealised model of the telephone system was to divide the network into two levels: a collection of national networks each with one or more international gateways and an international network that connected these gateways (this perhaps portrays it as being far more sophisticated than it really was, in that most international arrangement were bilateral between national networks as multi-party transit arrangements were beyond the crude financial and technical tools used in the telephone system.) The critical observation was that all of the end points in each country could be aggregated into a single address prefix. For example, "+61" was a routing prefix for Australia, "+49" for Germany and so on.

With the adoption of CIDR in 1994 we could perform IP address aggregation to arbitrary levels, so the thought was that if we could geo-localise IP address allocations that that each region used addresses drawn from a common prefix, then the inter-region routing system need only contain this small set of regional prefixes.

I've never been convinced that this is practical in the context of the Internet, despite the obvious existence proof in the telephone network. The Internet eschewed two fundamental principles that were a feature of the international telephone system. Firstly, the issue of a uniform quality of service. Any transit provider had to handle voice calls to a minimum defined level of quality, which in turn implied that all transit providers essentially offered an undistinguished product, and any transit provider offerred a service that was the same as any other transit provider. Secondly, the call accounting financial settlement structure meant that call revenues were evenly balanced between the call originator and the call terminator, and providers were meant to be unable to leverage revenue from each other. That way each transit provider in the telephone world essentially provides the same service at the same cost.

This has never been the case for the Internet. There is no common definition of a service in terms of quality metrics, and there is certainly no concept of call accounting financial settlement. If I advertise your routes, even as part of advertising a regional aggregate route, then somebody needs to pay me to carry the consequent traffic. If we use regional aggregate routing, then all local networks that are part of this aggregate need to engage all regional transit providers as a provider. We've never been able to devise a way to achieve this outcome without it looking like a criminal extortion racket!

Since 2006 the routing table was grown from 200K entries to around 1M IPv4 routes and 150K IPv6 routes, and more specifics sill account for around one half of all these routes. Somehow, we've achieved this without regional aggregation, and without redefining the financial or the technical foundations of the Internet

Resources

The presentations from this meeting can be found at https://www.iepg.org/2022-03-20-ietf113/index.html, and the recording to the meeting can be found at https://www.youtube.com/watch?v=u2wVZnJ3fo0

Disclaimer

The above views do not necessarily represent the views or positions of the Asia Pacific Network Information Centre.

Author

Geoff Huston AM, B.Sc., M.Sc., is the Chief Scientist at APNIC, the Regional Internet Registry serving the Asia Pacific region.

www.potaroo.net