November 2020
Geoff Huston

IETF 109

For a group that works on network technologies it was always a bit odd that the IETF met in person three times a year. Didn't we have enough trust in the efficacy in the technologies that we work on? I don't think that is the case. I think the bandwidth of in-person meetings is exceptionally high, and we just cannot cram all that into a virtual world. In this rather exceptional year the IETF has joined its conference brethren in virtual meetings. The latest, IETF 109, was held in mid-November. I'm going to pick just one presentation from each of a small collection of the week's working group meetings and explore that topic in a little more detail.

IEPG and Fibre Deployment

Jared Mauch talked about building his own local fibre network (https://datatracker.ietf.org/meeting/109/materials/slides-109-iepg-sessa-so-i-had-to-start-a-telco-getting-fiber-to-jareds-house-00). He lives in the semi-urban fringes of Ann Arbor Michigan and had, quite reasonably, expected that fibre optic cable would eventually reach his property. It's a developed community in a wealthy country. If any country can afford to install fibre infrastructure to its office and residences across the entire country, then surely the United States can. However, it's not exactly happening everywhere for everyone. After more than a decade of waiting it was evident that nothing was happening for Jared. So he looked at building his own networking, connecting his neighbours and running a small-scale fibre ISP in the area. He polled the neighbours and ultimately got a 70% signup rate from passed homes (this is a large uptake rate when talking about fibre deployments). It's a small-scale operation, with just under 40 connections, but at the same time the total cost for Jared was just $128,000. And this is not a high-density urban deployment, or even a suburban environment. It's in the semi-rural fringes and in an area where the larger scale infrastructure providers saw no reason to undertake a fibre deployment.

I find it interesting that Jared was able to identify a window of opportunity even though he had no economies of scale and no access to a pool of infrastructure investment monies. How is it that small scale fibre deployments with limited capital budgets can construct and operate fibre-based infrastructure networks and operate at a profit, yet the larger picture of infrastructure upgrades is happening at a pace that can only be described as glacial. Infrastructure operators are just not making these investments.

The existing copper access networks in many countries, including the United States, is ageing and some would claim the copper network is now well past its use-by date, yet its replacement is not happening. Industry actors claim that such projects are uneconomic and are trying to make the case that they need access to some regulatory or financial concessions to change this picture. At the same time, they would rather dissuade governments from taking peremptory action and undertaking their own deployment, such as the National Broadband Network in Australia, which was a clear signal of terminal dissatisfaction with the incumbent national telco in 2008.

So why aren't there more Jareds out there? If it can work for a community of 40 subscribers, then it can work for a municipality of 400,000 or more! I suspect that in the United States, and in many other countries as well, the telco incumbents would rather spend their time and money in lobbying politicians and in engaging in legal actions in the local courts to prevent municipalities taking direct action and

installing fibre access infrastructure themselves. We take it for granted that we can deploy water, power and all the other service utilities using quite conventional infrastructure investment models, yet somehow the telco sector is "special". So, while I can only applaud Jared's direct action in solving his neighbourhood's issues, I am dismayed by the degree of brokenness in this industry that he was driven to the point that there was no alternative to this direct action.

SIDR OPS and ASPA

One of the presentations of interest in SIDR OPS was on the efforts to perform some form of AS Path validation using the Autonomous System Provider Authorization (ASPA) object (https://datatracker.ietf.org/meeting/109/materials/slides-109-sidrops-aspa-ietf109-alexander-azimov-01). This is a variation of the earlier soBGP AS adjacency object, where with ASPAs topological AS adjacency is combined with inter-AS policy. It relies on a coarse taxonomy of inter-AS relationships where, within a particular routed protocol, two AS's can either be "peers" to each other or one can be a provider for the other. In the ASPA framework each "customer" AS lists the complete set of "providers" that are authorised to propagate routes that this AS has passed to it as a provider.

If you assume that no AS is willing to incur costs on behalf of non-customer AS neighbours, then we can arrive at the "valley-free" view of AS Paths. If you think about as AS Path as an ordered sequence of paired relationships, and we use the concept of "elevation" to describe the customer provider relationship, then we can define a policy topology of an AS Path. An AS Path that is a customer-to-provider pair increases the elevation, a provider-to-customer pair decreases the elevation, and a peer-to-peer relationship does not change the elevation. It we apply the principle of incurred costs here, then every AS path is a sequence of customer-to-provider relationships, followed by at most one peer-to-peer relationship and then a sequence of provider-to-customer relationships. In elevation terms every AS path is a "mountain", and no AS path contains a "valley". If we had complete knowledge of every inter-AS relationship and its policy setting, then we could apply this rule to every AS path and filter out all non-conformant paths.

The observation behind ASPA is that we don't need this much information to apply this rule. As long as every AS lists all its providers then every valid AS path is a sequence of customer-to-provider relationships, at most one unattested AS path followed by a sequence of provider-to-customer relationships.

The problem with this approach is the problem with many systems that overlay a security framework to an underlying data set. Security can tell you what's "good", but it can't tell you what's bad. As long as every good data item has its own security credentials then it's easy to infer that the lack of valid credentials is "bad". But in an environment of partial deployment then the missing elements confuse our efforts to categorise AS Paths as either "good" or "bad". Despite this limitation associated with partial deployment of ASPA credentials is this still of marginal benefit? Let's put this a different way: If I publish attestations of all my providers in ASPA objects can a hostile party still use my AS in a synthetic AS Path? The answer is "yes" but it is now in a more limited context. An attacker can still generate as AS Path that falsely represents an AS as one of my customers or peers, but it can only do so in more constrained circumstances.

The entire effort here is not to prevent all potential forms of route synthesis, but to limit the set of possible synthetic paths. Furthermore, the greater the level of adoption of credentials the greater the constraints that are imposed on credible falsehoods. If we can't stop all forms of attack in the routing system, then a reasonable objective is to make the attack harder to mount and minimize its impact. In this respect ASPA appears to be a useful refinement of the original soBGP approach, if we ever get around to adopting it. Origination validation is just one part of the overall story in improving the resilience of BGP. I've observed previously that if RPKI route origin validation is the only available routing protection mechanism, then it's little different to wearing a supposedly impregnable defensive shield made of wet lettuce leaves! Origination needs to be coupled with some form of mechanism that

make it harder to lie in the AS Path. The BGPSEC model of AS Path protection was never going to get deployed in my opinion, and the earlier soBGP approach of certifying AS peer pairs looked like a pragmatic approach to the problem. ASPA is a useful refinement to this approach, introducing the element of customer/provider policy to picture. I'm not sure if it will gain traction, but it seems to me that there are few alternatives on the table right now to address the path issue.

GROW and AS Prepending

Not only does BGP maintain an inter-AS topology, but also BGP is used to overlay that topology with preferences to perform traffic engineering. While a network can have its internal preferences to steer outbound traffic in certain ways, controlling incoming traffic is a little more involved. The network has to advertise routes in such a way that it constrains or biases the outbound choices made in neighbouring networks. BGP has a number of rules that determines the preference of outbound paths, and one of these selection criteria is to select the outbound next hop whose AS Path is the shortest of the candidate next hops. Now it's difficult to make AS Paths shorter, but by use of prepending it's possible to make AS paths longer. A BGP speaker who wishes to depreference a path would make this path longer in AS Path length terms, and the best way to do this is by prepending the AS Path with multiple instance of the local AS to the outbound BGP route object. AS Prepending is very common in the Internet's routing system. From one BGP vantage point some 25% of all visible routes have prepended AS Paths, with an average prepend level of 2.8 (https://www.potaroo.net/ispcol/2019-10/prepending.html).

A presentation to the GROW WG proposed that the practice of AS Path Prepending be described in a BCP (https://datatracker.ietf.org/meeting/109/materials/slides-109-grow-grow-as-path-prepending-00). The document is intended to describe the risks that are exposed when overly long AS prepending sequences are used, and also to describe where prepending can be useful and where other forms of traffic preference signalling (through various communities) can be more effective.

It's hard to say whether routing administrators use prepending because it is observed to work as intended, or because they simply hope that it might work as intended! A BCP describing AS Prepending in a little more detail may be helpful here.

DNSOP and Private Use Domain Names

The saga of the evolution of the name space of the Internet spans the entire life of the Internet, and it's way too long to even summarise here. But we seem to have landed up in the uncomfortable position of having a couple of bodies who appear to believe that they have some role in the polities that are used to define what names are used as top-level domains on the Internet. Back in 2000 the IETF came to an arrangement with ICANN where ICANN would assume the responsibility for the delegation of all top-level domain names in the DNS, with the exception of "assignments of domain names for technical uses" (RFC 2860) which the IETF retained control of.

This was largely a dormant subject until the IETF opened up a "Special Use" domain name registry (RFC 6761) and populated it with some new entries, including ".local" (RFC 6762) and ".onion" (RFC 7686). This stirred up some fundamental questions as to how the DNS name space was being managed, given that two bodies now had two different registries and appeared to be operating without much in the way of clear coordination. A number of questions were raised about how to cope with this, which were summarised in an excellent article by Peter Koch in the IETF Journal in April 2016 (https://www.ietfjournal.org/special-use-domain-names-a-registry-under-review/). There was also an extensive Problem Statement prepared at the time, published as RFC 8244. It was noted in that document that "There is strong resistance within the IETF to assigning domain names to resolution systems outside of the DNS".

But now the focus shifted to ICANN, which was completing the first round of new top-level domain names and looking to open up a second round of namespace expansion. The problem with delegating new domain names is that there is an extensive history of name squatting out there and these locally used

names that were used without any permission may "collide" with the use of a delegated name space that used the same label, which may create vulnerabiliti4es for unsuspecting users (RFC 8203).

In the same way that the IPv4 address plan reserved some addresses for private use (RFC 1918) IPv6 made a similar reservation (RFC 4193) and the AS number registry contains a similar private use reservation (RFC 6696), it was felt that one possible response to name squatting was the reservation of common "private use" name prefix.

However, after numerous iterations of the proposal in the DNSOP working group (https://tools.ietf.org/html/draft-ietf-dnsop-alt-tld-11) it appears that the conclusion was that this was not within the remit of the DNSOP Working Group nor even within the IETF and perhaps ICANN should consider this. ICANN's Security and Stability Advisory Committee published SAC113 (https://www.icann.org/en/system/files/files/sac-113-en.pdf), which recommended that a string be reserved at the top level of the DNS for private use and never be delegated.

Then DNSOP was presented with a separate draft (https://tools.ietf.org/html/draft-arends-private-use-tld-02) which proposes that private use top level labels be drawn from the user assigned code elements of the ISO 3166-1 alpha-2 ascii two letter code registry (the two-letter country code registry). Which (finally) brings us to the DNSOP discussion in IETF 109.

It's totally unclear to me that we know what we are doing in domain name policy, now or previously! Is this an ICANN issue? An IETF issue? Is the concept of a common private use top level domain even useful in any case? Is this a "special use" in the sense of the RFC 6761 Special Use name registry? Irrespective of the decision process to nominate one or more private use top level domain names in the DNS, will it mitigate name squatting? Or is this just a palliative measure as a prelude to auctioning off the highly valued domain names of .home, .corp and .mail?

Confused? I know I am!

MAPRG and DNS Centrality

We used to think of the Internet as a diverse and highly competitive environment of providers. The discipline of vibrant competition in open markets honed the service offerings to efficiently meet customer requirements. But that's a view of the Internet that's a decade or two out of touch. These days a small set of providers totally dominate the Internet. That's what the market capitalization financial figures are telling us, and that's certainly what we think of as a convention al wisdom, but can we measure this "centrality" of the Internet? This is a question posed in a presentation to the MAPRG research group (https://datatracker.ietf.org/meeting/109/materials/slides-109-maprg-clouding-up-the-internet-how-centralized-is-dns-traffic-becoming-sebastian-castro-00). As the presentation points out this can be challenging to measure for a broad cross-section of the Internet. The approach used in this study begins with the observation that every network transaction starts with a DNS query, and if we look at who is asking these queries then even if we don't necessarily know what they are asking for then we might get some measurable dimension of the centrality of DNS resolvers.

I happen to agree with the basic approach but as usual with the DNS the answers you get often depends on where you look! In this study they used the query logs from one root server and two top level domain servers (.nl and .nz). The logs show completely different profiles, where the two top level domain servers see queries from Amazon, Cloudflare, Facebook, Google and Microsoft represent between one quarter to one third of their query volume, yet it is less than 1 tenth of the query volume seen at B root. I'm not sure that I can place this data into a coherent context. The issue with l;ooking at queries to servers in the DNS is that you really don't know where they have been. There is no clear knowledge of why the query is being made. At APNIC we've been undertaking a similar study of resolver concentration (https://www.potaroo.net/ispcol/2019-09/centrality.html) and we see a somewhat different picture. The difference may well line in the nature of the queries being examined. Rather than look at all queries, the

APNIC work deliberately seeds the DNS with names to resolve, and then looks for precisely and only these query names at the authoritative server, looking at the recursive e resolver that passed the query to the authoritative server. Using this methodology, we then categorise the IP addresses of the resolver that passed us the query. On the way we are able to detect the difference between a "current" query and a query that appears to be based on a replay of a query log file. On some days we observe that 90% of all queries are log file replays. Given this relative massive overhang of queries that reflect some selective subset of prior user behaviours it is challenging to simply take a collected query set that includes both current and replay queries and infer some conclusion about the current behaviour of users, or the current state of DNS infrastructure.

I suspect that the study presented at MAPRG has this issue, that the extent to which replay queries are dominating the results, has this weakness. In the APNIC work, looking only at "current" queries we have observed that a little under 30% of users will have their queries passed to one of the set of popular open DNS resolvers, and of these Google's Public DNS is the overwhelming choice. Some 25% of all users may pass their query to Google's Public DNS resolver server. One resolver service serving one quarter of the entire Internet user population. If you are looking for measurements of infrastructure centrality in the DNS, then I guess that you need to look no further!

V6OPS and that /64 boundary

The address plan for IPv6 has always been a confusing story. IPv6 did not change all that much from IPv4, and it could be argued that the small changes that were made have not been all that successful. One of these was to do away with the Address Resolution Protocol (ARP) and instead to a multicast method of establishing the common 64-bit network prefix and self-assigning a 64-bit interface identifier. The address architecture has a clear view of the distinction between a host and a router.

But what is a mobile phone these days? More to the point what is surrounding that SIM card. It could be a phone, but equally it could be a car, or a more complex end environment. With the increasing bandwidth available in mobile networks there is an increasing desire to locate multiple networks behind the same connection and manage them in such a way that they do not fate share in a single local subnet. But at this point we run into the issues associated with the address plan. If this was managed in an analogous manner to IPv4 we could take a /64 prefix and use subnets with smaller prefixes. But we are reluctant to shift the /64 Interface Identifier to a smaller boundary and unsure about interoperability with the existing deployed base. On the other hand, we are reluctant to assign large prefixes to end points to cover the possibility that the end point is in fact a gateway to a more complex local network topology. Equally, we are reluctant to contemplate NATs in IPv6. The problem statement (https://datatracker.ietf.org/meeting/109/materials/slides-109-v6ops-slaac-with-prefixes-of-arbitrary-length-in-pio-variable-slaac-a-problem-statement-01) is not unexpected, but ready-made scalable answers are not forthcoming.

We need to steer a careful path between enthusiastic over-consumption of the IPv6 address pool where large-scale mobile providers end up with a requirement for a /8 allocation, but there is also a strong desire to avoid NATs, and a strong desire to preserve the 64-bit length of the interface identifier field. There is a reason why this presentation considers a problem statement and not a potential solution path. It's clear what the problem is, but not so clear how a simple scalable non-NAT backward compatible solution might look like in the IPv6 environment.

ICCRG and the Congestion Control Census

You might think that some 40 years later we would understand how TCP works. We don't. Yes, the TCP headers are all the same, modulo a timestamp option or two, but what is an enduring area of research is the management of the TCP flow rate. There were a number of interesting reports about current activities in congestion control in TCP at the meeting of the research group at IETF 109. There was a progress report on the less than best effort LEDBAT flow control protocol, work on BBR, TCP Prague and Data Centre TCP. I was struck by a presentation by Auysh Mishra from the University of Singapore on a flow

control census of today's most popular web sites (https://datatracker.ietf.org/meeting/109/materials/slides-109-iccrg-the-great-internet-tcp-congestion-control-census-00).

Over the past 30 years we've seen some 30 variations of the TCP congestion control algorithm, and every decade or so someone is motivated to try and perform a census to report on which algorithms are popular. This is such an effort. The tool used to perform this census is "Gordon". This is a simple tool in that it performs bursts of packet drop to detect the sender's count of the allowable number of in-flight packets (or "cwnd"). To test this with existing sites they crawled the Alexa lists to find larger web pages and used the smallest possible MTU. This meant a large sequence of packets that they could work with. The bench-tested nine congestion control protocols and developed congestion window profile "signatures" of each of these protocols. They then ran this profiler across the Alexa top 20,000 web sites. CUBIC (the default congestion control platform on Linux platforms) and BBR are the most popular, popular, as well as Microsoft's CTCP. Taking a restricted view of the Alexa top 250 sites BBR and CUBIC appear to account for half of these web sites.

What is evident here is that we are seeing loss-based and rate-based control systems in competition for network resources. Rate-based control paradigms are fundamentally different to the loss-based systems, and the concepts of fairness probably need to be re-examined. There are some very interesting research questions in this work and questions of the future of congestion control algorithms that are widely used the Internet in terms of evolutionary pressures. In that context I found this census very interesting.

Virtual IETF 109

It's still taking the IETF community some time to get used to operating in this virtual void, and while the conferencing tool is getting better, the week still saw a few hiccups. Timezones are always a problem for these global sessions and while the choice of Bangkok as the "home" timezone was very convenient for my location on the East Coast of Australia many others were in a far more challenging sleep-deprived situation. It was a meeting with a lot more "show and tell" when compared to the face-to-face meetings and I'm sure many folk missed that spark of personal interactions. The first meeting of 2021 will also be an online meeting, but after that, like many other things in this world right now, we just can't tell right now!

Disclaimer

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

Author

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

www.potaroo.net