Geoff Huston
February 2016

NANOG 66

NANOG continues to be one of the major gatherings on network operators and admins, together with the folk who work to meet the various needs of this community. Their program committee produces a program that never fails to provide thought provoking interest. Here are my reactions to some of the presentations I heard at NANOG 66, held in San Diego in February.

IPv6

It was always assumed that the exhaustion of the pools of IPv4 addresses would act as a major incentive to the deployment of IPv6, but in the 5 years since the original point of exhaustion, the central pool of IPv4 addresses held by the IANA the deployment of IPv6 has been somewhat lacklustre. The measure of the level of IPv6 used to access Google sites has reached 10% in recent days (http://bit.ly/1WLCVg6), but the global estimate of IPv6-capable users is a little lower than this, currently at some 5% of the total Internet user population (http://bit.ly/1LHZRa3). Another view is the relative proportion of ASNs that announce both IPv4 and IPv6 prefixes. (http://bit.ly/1WLCPow) Of note in this view is the rapid rise in the number of IPv6-anouncing networks in the ARIN region following the exhaustion of the ARIN IPv4 address pool in mid 2016.

Emile Aben presented on as aspect of RIPE's support for IPv6 adoption. For some years RIPE has had a program of IPv6 "stars" that provided a public acknowledgement of networks that go through the steps of obtaining IPv6 addresses, adding them into their service infrastructure and announce an IPv6 service. Of the 13,000 Local Internet Registries (LIRs) in the region served by the RIPE NCC approximately one quarter have no IPv6 stars and one third have one star (received an IPv6 allocation). The remainder have 2 or more stars, corresponding to routing advertisement, reverse DNS registration and route registry entries, and the count of IPv6-capable users in their network. The observation was made that progress has not always been up and to the right. The presentation notes that some 462 entities have stopped announcing IPv6 addresses. The RIPE NCC contacted these folk and found that in many cases this corresponded to a test of IPv6 and the advertisement was withdrawn at the conclusion of the test. Other reasons include lack of infrastructure support, lack of a commercial imperative as seen by explicit customer demand, and some concerns relating to the network's security framework. In the majority of cases there is no definite plan when the IPv6 prefix will be re-announced.

It certainly appears that the "whole of Internet" current picture of IPv6 deployment is not uniform. While there are many signals that show a concerted effort to deploy IPv6 in the United States, comparable activity is not so evident in many other countries., and this presentation shows that in parts of Europe the forward momentum is counter balanced by some evidence of regression to an IPv4-only service by some network operators.

I also presented at NANOG in the topic of measuring IPv6 performance, comparing the outcome of work performed in 2012 with a similar study undertaken across 2015. The good news is that IPv6 performance reliability is improving, and these days the we see a connection failure rate of around 2%. The major change in the intervening four years is the drop in the use of Teredo and 6to4. Both of these

auto tunnelling techniques have largely lapsed in the intervening period, and Teredo has all but disappeared while 6to4 is declining. Which is just as well as these two tunnelling techniques are notoriously unreliable! But what we are left with is this residual connection failure rate of around 1 in 50 connection attempts. IPv4, by comparison has a failure rate of 0.2%, or 1 in 500 connections. So from a reliability perspective IPv6 still has some issues to address in many networks. The other aspect of performance is related to relative speed. Is IPv6 faster or slower than IPv4? Here the answer is a lot better. IPv6 is just as fast as IPv4 for around 70% of all the sample points. So as long as the customer is not seeing the IPv6 Internet through some form of tunnel, then once the connection is established the connection will run at much the same speed as Ipv4. The residual issue is that the odds of setting up a connection are still far better for IPv4.

Network Management

Every network fails, and large networks fail more often. Many times the issue is clearly visible, but every now and then there is something that goes by undetected by device-based network monitoring systems. This talk described Facebook's experience of building a "black-box" fault detection and isolation system for data-center and backbone networks. The heart of the system is "high-rate active probing" component that allows for detection of failures regardless of the underlying cause, with appears to be a fancy description of UDP-based ping and traceroute mechanisms. While there is a lot of attention these days on automated network management approaches, this talk served as a reminder to me that the now venerable approaches of ping and traceroute are still extremely useful tools in the network manager's current toolbox!

But then I can't help but wonder about that. The use of tunnels, particularly with MPLS, is commonplace in many networks, and when coupled with issues such as path asymmetry, multi-path routing and even 5-tuple ware SDN constructs, its easy to construct an "intelligent" network where ping and payload traverse different paths and potentially encounter different network conditions! We appear to be investing a lot in the claim that traceroute exposes what the network is going and ICMP echo requests traverse the same paths as payloads. This may have been the case on the Internet of a couple of decades ago, but these days its not so clear that it still holds, and certainly the case that in some networks it certainly is not the case! But with ping and traceroute what's left to understand what is happening in the network?

Network Automation

Network element configuration has always been a rather sad backwater of the network management story. It seems that the command line interface (CLI) syntax used in many network devices has a pedigree that dates as far back as the RSX-11 operating system of the early 1980's, and has changed at a pace that is somewhere between glacial and geological! The underlying model is that of a human operator entering a textual configuration that is not all that far removed from conventional English. Admittedly, its a step up from JCL, if anyone remembers that, but frankly its not the best model. But as the number of devices proliferate and the diversity of operational behaviours proliferate this model breaks down - the configurations are long and detailed and the potential for mistakes and inconsistencies are manifold. Its no surprise that there have been a number of initiatives that are intended to remove the human typist from the configuration management picture and replace it with a set of scripted tools. SNMP tried with with the SNMP Write command but that turned out be be an epic fail. So if we are stuck with these textual CLI's and we still remove the human from the loop? Leslie Carr presented a delightfully simple and information presentation on what tools like Ansible, Chef and Puppet are attempting to achieve. It starts with Git as a central repository of the configurations. (https://try.github.io), and then proceeds with the advice to load the complete set of device configurations into git as a means of pulling all the configuration information into a single location. The next observation is that any such collection of configuration files normally has a large amount of information that is common and a small amount that its specific to the individual device.

One logical answer is to parameterize the configurations, and reduce the common elements to a template and the specific information to a set of variable value that are applied to these templates.

Exchange Interactions

In a network everything is connected to everything else - which is either a somewhat spooky or incredibly mundane claim! Daniel Kopp presented on a particular case of inter-dependence where a rather large operational incident that caused a large traffic drop at the AMSIX exchange in Amsterdam (one of the major Western European Internet Exchanges) caused a simultaneous traffic drop at the DE-CIX exchange in Frankfurt (another one of major Western European Internet Exchanges). This is not an intuitively obvious outcome. The theory goes that in a richly interconnected environment the drop of one set of inter-AS connections would see routing, and subsequently traffic, switch to another set of inter-AS connections, and DE-CIX would presumably see an increase in traffic. The critical missing part of this presentation is an analysis of BGP behaviour as seen by peers of AMSIX, and, dare I say it, traceroutes that occurred across the AMSIXC exchange at the time of the outage. Without that additional data I am left wondering if what was seen was a number of units actually falling over and performing a full reset at the time the original AMSIX incident, which may be a potential cause of a cascading failure that would impact other exchanges.

ARIN Policy Consultation Session

The elephant in the ARIN Address (and Registry) Policy room (and indeed in the equivalent rooms of all the Regional Internet Registries) is considering the issues of how to keep the registry function complete and accurate with respect to address disposition. Omitting many fine points of detail, the original model was that the RIR was the sole source of addresses and the registry simply recorded the RIR's action. But in the IPv4 registry this no longer applies and we are seeing an aftermarket emerge. The subsequent transactions have a number of subtleties, and its evident that not all transactions are simple unconditional sales. There are leases, caveats, options and rights of use being traded as well.

These are significant challenges to the more traditional perspective of the registry function. It raises the distinction between the current "user" of an address, the "beneficial controller" of an address and the "owner" of an address, and even pulls in the concept of a caveat on the title over an address. This may all be bread and butter to a land title registry, but its new territory to the RIR policy processes and its a challenge to ensure that the registry remains complete, accurate and above all useful in such circumstances. At this point I gathered the distinct impression that the policy folk in this session appeared to be trying to leave leasing and similar matters of more complex structures of mutual interest in address as an unaddressed matter (if you will pardon the poor pun) and thereby allow it to be subsumed into areas of 'creative ambiguity' of interactions between entities and the registry operator. This is probably not an optimal long term approach.

Network operators are now in the difficult position of accepting so-called "Letter of Authorization" from a presumed current user of an address as a means of legitimating the way in which a route in entered into the routing system. To put it more crudely, the current way to have an address routed is just a matter of ascii artwork, and its no surprise that this practice is being abused by address hijackers.

Measurement

Dave Clark and kc claffey presented an interesting session on the intersection of network measurements and public regulation. AT&T is merging with the Direct TV, and according to their own publicity machine this would create the largest pay TV provider not only in the US but in the entire world, and noting that this merger would "set [the combined entity] apart from the competition". Such claims are open invitations to any competent and attentive regulator, and evidently the FCC is indeed paying attention, as were the competition of course! The FCC noted that broadband Internet access providers have the ability to use terms of interconnection to disadvantage edge providers and that

consumers' ability to respond to unjust or unreasonable broadband access practices are limited by switching costs. As part of the merger conditions, AT&T has agreed to develop, in conjunction with an independent expert, a methodology for measuring performance of its Internet traffic exchange, and regularly report these metrics to the FCC. the measurements concern latency, packet drop and link utilization levels. CAIDA has been identified as the independent expert in this context. The details are in the presentation pack, and I won't repeat them in this summary.

I found this a highly useful and appropriate response from the FCC. Markets, and the role of regulation of markets, depend on a thorough understanding of behaviours and outcomes. In many cases network measurement is regarded as a private function and the outcomes are folded into the corpus of private data and never disclosed beyond each individual network operator. By withholding that information from public view we push the regulatory function into rule making with incomplete knowledge and while this does not necessarily deter regulatory action, it may impact on the quality and efficacy of such levels of intervention. So measurement helps. What appears to be lost in the argument about treating measurement data as private data is that while today's public Internet has been constructed largely with private capital, it is still a public endeavour, and the public communications function is still a public service, not a private one. Public measurements help us all in this context, including the investors, operators, regulator and the consumer, to assist in understanding the true nature of this public communications environment.

Rethinking Path Validation

RPKI and its role in efforts to define a "secure" version of BGP has been subject to a number of second thoughts in recent years, and Russ White's presentation was another in this vein. Russ was one of the co-authors of the soBGP, and many of the concepts in this earlier work are visible in this presentation. The essential change from the BGPSEC specification currently with the SIDR Working Group of the IETF is that instead of performing a full set of interlocking AS signatures to protect the integrity of an AS Path in BGP, this approach uses a collection of pairwise AS adjacency attestations (or "connectivity certificates") that essentially indicates that the pair of AS's are directly adjacent. IN a world of comprehensive deployment of this form of certification, a received AS Path can be broken into pairs, and each AS Pair can be matched against a valid connectivity certificate. The model can be extended in a number of ways, including the statement of routing policy that applies to that particular inter-AS connection. What this means is that while route objects may be synthesized, the synthetic route needs to match an extant ROA and the AS's listed in the AS Path must be specified by connectivity certificates, and match any applicable policies. In other words, the synthetic route object must correspond to a plausible propagation vector through the network. This dramatically reduces the attack surface on a routing attack, while operating with far lower overheads than the overlocking signatures proposed by BGPSEC.

There is probably some way to go with securing of the routing system, but it appears to me that the approach described in BGPSEC simply has too much emphasis on protocol correctness over pragmatism, and defines many moving parts. There has to be a simpler, and potentially more robust way of doing this, and I suspect that the techniques described in this presentation are part of any revision of the approach.

CDN Routing

There have been two mainstream approaches developed over the years to steer users to the closest instance of replicated content: using the DNS or using anycast routing. This presentation from Nick Holt explored how to use both. The DNS approach is as old as Netscape, if not older. When a query arrived at an authoritative name server it attempts to determine the location of the querier, and provides an answer that refers to the closest instance of the content. This approach relies on a number of assumptions, not the least of which is the assumption that the actual user who will receive the answer is located close to the resolver asking the question. In this age of large public name resolvers

(Google, OpenDNS, Microsoft, Level 3 and Versign all operate such a service) its by no means clear that the user is located anywhere near their resolver. There is also the issue that the DNS response is cached and shared with other users of the same public resolver. It also assumes a relatively good geo-location address data set, which also has its fuzzy edges. So the DNS approach has some limitations here as a distributor for replicated content. The other approach is to place the same content at the same address in multiple locations. In this case all users get the same DNS response, and its left to the routing system to direct the user to the closest instance of the replicated content. This is in most cases a more effective approach, but again it has fuzzy edges. The routing system can sometimes generate lengthy paths, and it cannot re-distribute load. If too many users are pull content from one server, while other servers are idle it is not readily possible to redistribute the load via routing adjustments. This presentation proposed using both approaches, relying on the routing system to perform a base level distribution of load, but then using variant DNS responses with CNAME records when required to perform a redistribution of load.

Want More?

This is just my impressions on a subset of the presentations at NANOG 66. All the presentations and videos can be found at https://www.nanog.org/meetings/nanog66/agenda

Author

Geoff Huston B.Sc., M.Sc., is the Chief Scientist at APNIC, the Regional Internet Registry serving the Asia Pacific region. He has been closely involved with the development of the Internet for many years, particularly within Australia, where he was responsible for building the Internet within the Australian academic and research sector in the early 1990's. He is author of a number of Internet-related books, and was a member of the Internet Architecture Board from 1999 until 2005, and served on the Board of Trustees of the Internet Society from 1992 until 2001 and chaired a number of IETF Working Groups. He has worked as an Internet researcher, as an ISP systems architect and a network operator at various times.

www.potaroo.net

Disclaimer

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