Geoff Huston
November 2016

RIPE 73 Meeting Report

RIPE held its 73rd meeting in Madrid in the last week of October. Here are a few of my takeaways from that meeting.

What's behind all those NATs? We suspect that there are at least 10 billion devices connected to today's Internet, and we know that less than two billion individual IPv4 addresses are in active use. Simple maths implies that most of the connected Internet lies behind some form of IPv4 NAT. But discovering exactly where the NATs might be and how many devices lurk behind each NAT is a more challenging question. Philipp Richter reported in some research into the state of infrastructure NATS (or "Carrier Grade NATS" (CGNs) as they are commonly known). Despite the well-known reservations about CGNs, there is a certain level of forced compromise here in that for many network operators there is simply no other viable option other than to deploy these devices in their networks. Philipp reported on a study to detect CGN presence and the properties of these CGNs. One approach used in this study was to analyze the Distributed Hash Tables carried in BitTorrent by crawling the DHT space, asking each visible peer for their neighbour peer table. Where two or more peers are located in shared space behind a common CGN they will report the peer via its private (internal) address. This crawling technique netted some 700,000 peers in 5,000 ASNs. They found a distinct pattern which is best seen in his presentation, reproduced here from Philipp's presentation. NATs at the edge of the net have a 1:1 association between a public and private IPv4 address, while CGNs typically have a "halo" of private addresses surrounding the CGN's public addresses.

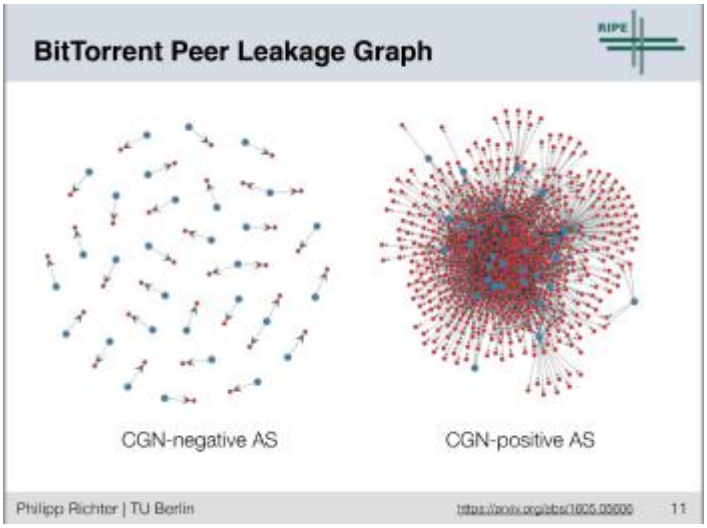

Figure 1 – NAT Classification – from "A Multi-Perspective Analysis of Carrier-Grade NAT Deployment", Philipp Richter

They complemented this DHT analysis with more targetted Netalyzr data, which has information gathered form some 550,000 sessions in some 1,500 ASNs. The estimate that some 94% of mobile data service network deploy CGNs of one form or another, while in other non-mobile networks the

number is a far lower 17%. They also detected a number of mobile providers co-opting address ranges outside of the reserved number pools for CGN "internal" space. Networks 21/8 and 25/8 are popular choices here, but also networks were seen to use 25/8, 26/8, 29/8, 30/8, 33/8, 51/8 and 100/8. For as long as stand-alone IPv6 is not a viable option no doubt this NAT story is going to get more and more convoluted!

IPv4 address transfers was the subject of an interesting plenary presentation. One statistic I noted was the observation that 85% of logged transfer space was not advertised in the routing system before the transfer, but appeared in the routing system after the transfer. Typically, these transferred addresses are announced within six months of the transfer, and once they are transferred they appear to be used more intensively (Figure 2).

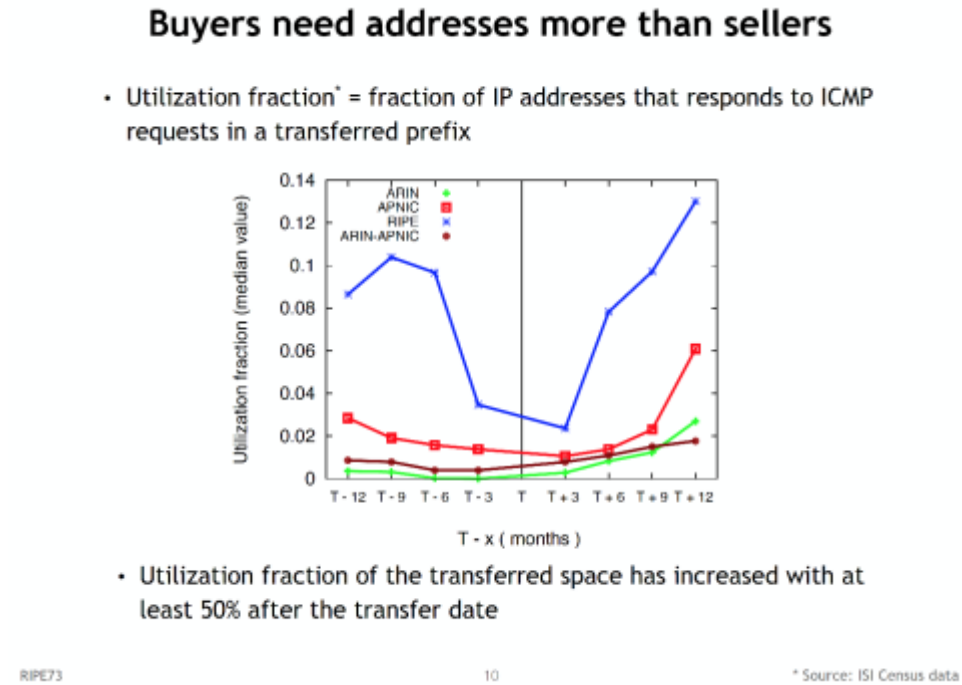

Figure 2 - from "Measuring the IPv4 Transfer Markets", Ioana Livadariu

While Regional Internet Registries have a public log to record the transactions that involve the movement of IP addresses and ASNs, there is a continuing suspicion that not all transfers are visible in these logs. Ioana Livadariu presented on efforts to detect these unreported transfers using BGP data. They used a set of detection rules and checked this against the listed transfers, and reported that some 90% of the listed transfers were detectable in the BGP logs. However, the report of unrecorded transfers was somewhat inconclusive at this stage. Doubtless there will be more reports of the outcomes of this technique in the coming months.

Anycast continues to be a featured topic. Ricardo Schmidt reported on an analysis of four anycast constellations of the DNS root server system to determine if the BGP-based anycast instance selection resulted in a segmentation of the Internet such that individual vantage points were being directed to the "closest" anycast instance, as measured by packet latency. Obviously the outcome of this work depends on the distribution of the vantage points used for testing, and some of their results do not appear to match an informal intuition of what should be happening here. For example, their results indicate that the C-Root anycast constellation, with just 8 instances produces an equivalent median latency measurement to L-Root with 144 instances. From a geographic perspective L-Root has instances in many more locations that C-Root and clearly it provides "closer" service to more populations. But many of these serviced populations are small, particularly when compared to the user populations of the larger markets. Just locating anycast instances in China, India, USA, Japan and Brazil gets you "close" to almost one half of the total estimated Internet user population, but it still leaves

large geographies untouched. So I'm unsure that the methodology used in this study is one that provides that useful a perspective. I'm also uncomfortable with the concept of equating anycast services in the root servers of the DNS with a more general question about anycast and latency. In the case of the root server system the prime objective of the anycast deployment is not to reduce the latency of responding to queries (considering than some 90% to 95% of queries to the root system are junk queries that elicit an NXDOMAIN response, a faster NXDOMAIN answer is of no benefit to anyone!). The objective is to reduce the vulnerability of the system to attack. An individual attacker will "see" just a single instance within an anycast setup, so to mount an effective attack against an entire root server anycast constellation its necessary to enlist a widely distributed and highly capable attack platform. An optimal anycast deployment strategy might want to focus on anticipated loci of attack, as distinct from minimizing latency. Finally, I am worried about the idea that BGP routing produces a latency-related partitioning of the Internet. BGP does not do that. BGP does not route on a latency metric, or even a bandwidth metric, or any other derived cost metric. BGP just attempts to minimize the number of transit AS's to each destination. So comparing a BGP-directed outcome to optimizing latency is somewhat of a comparison of oranges and apples in my opinion: they are just not directly related!

Wouter de Vries reported on work that looked at anycast-based network segmentation. This study constructed a custom anycast platform with 10 instances that had a broad geographic spread. The measurement appears to use source address spoofing using the anycast address as the ping source, such that they generated some 10 million pings (one for each /24 in the advertised Internet. Apart from indicating that a large proportion of the addresses in the IPv4 Internet that are not hidden behind NATs are located in North America and Europe, and Atlas nodes are highly concentrated in Europe, there were some "anomalies" in the anycast distribution (such as European IP addresses having their responses passed to an anycast instance in Miami. But perhaps what is shows is more about the interconnectivity mesh in BGP than it shows about anycast. The Internet is not "long and stringy". It's "fat and dense". The average AS Path Length from one "edge" of the Internet to all advertised destinations is just 4. This means that there is little distinction in BGP between "far away" and "nearby", and its entirely expected that BGP will take an anycast constellation and produce a segmentation outcome that has some geographically surprising outcomes.

Merging two or more networks into a single network can be tricky, particularly if the networks have a large number of external BGP peer sessions. Alexander Azimov explored a couple of approaches in BGP that attempt to simplify this process. One approach is described in RFC7705, a look as AS migration mechanisms, that uses a Local AS setting in an eBGP speaker. In this case the migrating network configures the eBGP speaker with a Local AS value of the AS being retired, and "points" this local AS to the existing eBGP peers. The AS Paths are also manipulated such that the local AS is used on outbound announcements, and it is stripped on inbound announcements. Another approach is to use AS Confederations (RFC5506). Confederations allow a set of internally connected networks, each using a different ASN to look to the external BGP environment as a single network with a single ASN. In many ways Confederations are a more general mechanism than the Local AS mechanism described in RFC7705. It is likely to be the case that many of the simpler forms of AS migration will be more readily managed using the Local AS approach, but of course every case is different, and Confederations provide solutions to scenarios not readily encompassed by the Local AS migration approach.

Ondřej Surý presented in the Knot recursive resolver, and the use of filtering rules in the resolver. The Knot resolver is an open source resolver written in C and LuaJIT, featuring an extensible design that allows scripting and modules to be added. The most recent version of this resolver has an HTTP/2 interface and a DNS firewall. this gives the administrator considerable flexibility as to which queries will be accepted by the resolver. AS a recent entrant to the set of recursive resolvers Knot is certainly an interesting offering, and well worth a closer look.

Jaap Akkerhaus gave a progress report in an analysis of the impact of the introduction on the stability and security of the root system. Since late 2013 the number of entries in the root zone of the DNS has grown from 350 in late 2013 to 1,510 today. The overall question is whether this expansion has impacted on the DNS root service in any way. There are some interesting observations over time about the nature of the queries seen at the root. in 2012 more than one half of the queries seen at the root were for names that used delegated top level domain names. These days that has dropped to some 40% of queries. For these queries, it appears that a rule of thumb is that the number of queries that relate to a particular TLD is around 10 times the number of delegated domains in that tld. Aside from this, there is no obvious signs that the new gTLDs have had any impact at all on the operation of the root zone.

Johan Ihren provided his perspective on the changing DNS environment. To Quote from his presentation: "Once upon a time the DNS was simple […] However, things went south over time: good guys started doing bad things (split-DNS, strange forwarding setups, policy-based responses, lots of rope everywhere), and bad guys showed up doing bad things (also with DNS). The major reason that DNS is becoming a "problem" is that there is not sufficient revenue to match the increasing cost of operation." DNS system complexity is exploding with anycast, feature creep, DDOS mitigation and behavioural variation. Johan sees this drive to greater complexity continuing, but like electronic mail services, DNS services will be provided by a smaller number of large scale providers. As to the DNS name resolution service – it's becoming an API not a protocol!

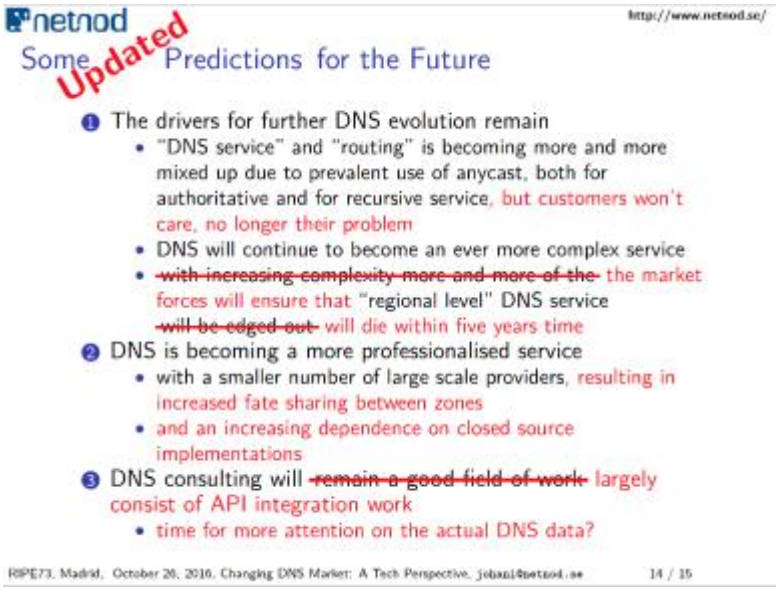

Figure 3 – from "The Changing DNS Market – A Tech Perspective", Johan Ihren

The Internet of Things still raises far more questions than answers. In the discussion at RIPE 73 there was an effort by Marco Hogewoning to try and sort these questions into a number of areas: It's reasonable to assume that "things" will use radio spectrum, but which spectrum? Will we go down the WiFi path and use unlicensed bands, which raises questions about scalability, accountability and reliability, or will we follow the 5G path and use licensed spectrum with attendant imposition of costs of mobile data from the spectrum owner and operators. What is the identity management regime? Is this another case of Ethernet, with IEEE EUI-48/64 MAC layer identifiers? Or will we follow the established mobile operators and use IMSI or IMEI numbers? Or even reapply the telephone numbering system and use E.164 numbers? Do IPv6 addresses play a role here?

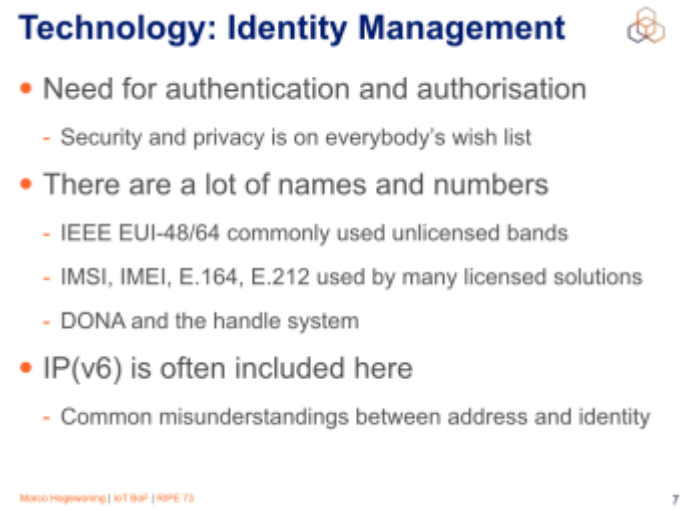

Figure 4 – from "IoT BoF", Marco Hogewoning

Will we use IP? If so which version? IPv4 is the dominant incumbent, but these days its a case of judging to what extent NATs and port sharing can continue to absorb further device growth. If it's IPv6, then the small deployed base of availability is a genuine concern. Of course we could ignore IP altogether, in the manner that RFID already is a layer 2 mechanism that has no IP adaption layer. AS another example, the 3GPP work on narrow band LTE is intended to work with minimal power drain and send small amounts of data in enclosed private realms, and there is no particular reason to use IP in this context other than supply chain availability in chip sets. Is the Internet of Things actually a collection of data-centres of data from things? What are the connectivity requirements of the field devices compared to the access requirements for the data generated by these devices? In this space questions are common. Useful answers, less so!

As usual for RIPE meetings, the program was interesting, provocative at times, informative and always fun! I'm looking forward to RIPE 74 in May next year in Budapest.

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