The ISP Column
*A monthly column on things Internet*

October 2015
Geoff Huston

# NANOG 65 Report

NANOG 65 was once again your typical NANOG meeting: a set of operators, vendors, researchers and others, meeting for 3 days, this time in Montreal in October. Here's my impressions of the meeting.

## Keynote

The opening keynote was from Jack Waters from Level 3, which looked back over the past 25 years of the Internet, was interesting to me in its reference to the "Kingsbury Letter" (https://www.washingtonpost.com/news/the-switch/wp/2013/12/19/this-100-year-old-deal-birthed-the-modern-phone-system-and-its-all-about-to-end/). As the referenced Washington Post article points out, in exchange for this government-sponsored monopoly, AT&T agreed in 1913 to operate as a public utility, eventually providing high-quality phone service to the vast majority of Americans regardless of income or geography. Kingsbury's commitment to President Wilson would later be formalized and expanded by Congress into the legal obligations that still bind the modern successors of the old AT&T: Verizon, Sprint, and the entity that today carries the AT&T name. At the time the US Government had undertaken anti-trust actions against monopolies in the rail, oil, and tobacco industries, and were contemplating similar action against AT&T. AT&T managed to take the initiative and create in effect a government sanctioned privately owned monopoly.

Things are again changing. The Open Internet Order from the FCC is putting the former unregulated Internet business under more overt regulatory purview, and while the current rules are vague and not highly prescriptive, the FCC is now equipped with the clear ability to proceed with rule making in response to poor behaviours by Internet access service operators. We have already seen TWC's proposed merger with Comcast being blocked, and instead TWC has bulked up with Charter. It certainly appears that the number of providers in this sector is shrinking as market expansion stops and operators start to look hard at their costs, and the three way tussle between carriage, content and regulation appears to be one that will keep a set of lawyers employed for many years.

Another interesting perspective I picked up from the presentation was that Jack felt that silicon switch capability was outpacing fibre. He pointed out the issues of fibre with 100G per usable channel and the resorting to use of ECMP and LAG in routers as a means of synthesis of higher capacities that are not being provided on the fibre plant. What he was claiming was that silicon switches are more "dense" than WDM in fibres, so the string is now skinnier than the terminations. He is sceptical about the ability of this year's SDN fashion, or NFV for that matter, to bridge this gap. It's hard to reconcile these calls for ever high aggregate channel capacity out of fibre and the issues of memory speeds and the limitations imposed by memory on packet rates that come from higher channel capacities. One other observation he made that struck me was that the dynamics of peering and interconnection have altered in recent times and this is due in part to John Oliver's "expose" of this business. Has the role of determination of communications policy in the United States now a role that has passed from the Federal Courts of Appeal to the host of a satirical comedy television show?

Finally, the Security elephant in the room: with an Internet of Things who's responsibility is it to fix it when the things inevitably turn toxic?

## DDOS

Roland Dobbin's presentation on DDos for Bitcoins was an interesting perspective of the activity of a particular criminal element performing online extortion. One interesting aspect here was that it was not clear from the presentation if this was a very active individual or an organised group.

The later DDOS tutorial was a walk through the toxic behaviour taxonomy. I'm not sure that there was anything "new" in terms of attack behaviour or the nature of the weaknesses that these attacks are exploiting. The hard truth is that the Internet of Stupid Things is the protein of Internet toxicity, and there is no real incentive for remediation here. Its not a happy outlook given the poor quality of software that we are embedding into these things.

## SFLOW

Avi Freedman presented on the many uses of sflow. Many flow tools present aggregated information that is already "combed" by the flow tool. The disturbing part of this presentation is the concept of pervasive network level snooping on the data flows. It seems that part of the reason why routers have computing grunt is to tap the packets as they fly past and perform a basic flow analysis to spit out an aggregated view of user traffic. Frankly, if you are at all concerned about the level of unauthorised and uncontrolled surveillance of users' online activity, then much of this talk is about as good as advertisement for using encrypted connections for all user connections as any I have seen, or for adoption of encrypted VPN services. On a similar vein it to me that Google's decision to use UDP as QUIC is an even better idea simply to confound the current TCP fume sniffers that sit out there in the network's infrastructure!

## DNS

The DNS track was a re-run of some presentations from the DNS OARC meeting held immediately prior to the NANOG meeting. I have already written on these presentations (http://www.potaroo.net/ispcol/2015-10/oarcf2015.html) so I won't do so again here.

## BGP

Being NANOG there was no shortage of BGP-related presentations

A session looked at the state of efforts to secure BGP. OpenDNS has taken over BGPmon. It's a routing anomaly detector that attempts to discern the difference between routing updates and the normal cut and thrust of link level unreliability, and the intended target of the anomaly detector is routing hijacks and other hostile forms of manipulating the routing system. Of course some routing attacks are such that everyone sees the information, but other more subtle attacks are directed to a particular network, and the propagation of the false information is deliberately limited. Its not easily determined just how fine a granularity is used in the collection of BGPmon data, which means that its unclear to what extent focussed routing attacks are detectable by this setup. The session had another presentation on IRR tools, and a repeat of a proposal currently circulating round the traps to try and find a consistent subset of IRR route objects. There is a salutary lesson here, and its probably more about human nature rather than technology: it may be easy to solve your particular problem by building just another instance of a routing registry that meets your particular needs, but in the end such actions contribute to the entropy of the Internet, not reducing it! There is a long term underlying issue with BGP security that we have been working on for a couple of decades now. That's a solid indicator that its a difficult problem, and what was more evident from this session was not the specifics of the various approaches considered here, but the more general observation that we are just not finding a feasible approach to detecting "good" from "bad" in inter-domain routing that meshes with practical constraints of a diverse set of players in the routing system. These operational responses strike me as more band aid. We are still searching for some form of approach that supports piecemeal deployment, accommodates a variety of operational models, does not get bogged into a morass of heavy weight crypto.

BGP BMP (BGP Monitor Process) has been slowly gestating in the GROW WG of the IETF for many years now. The specification has now celebrated its 10th anniversary of sitting the IETF drafts repository! I suspect, sadly, that this is not a record for today's IETF. BGP monitoring used to either put the BGP speaker into debug mode and push out as debug messages all the input and output BGP messages, or just sit as a BGP peer and collect the post-processing state updates of the BGP speaker being monitored. BMP is a standardisation of the debug approach, allowing a BGP monitor to track the state of the BGP speaker, tracing both the incoming BGP updates and the state changes that occur within the BGP speaker. This presentation reported on combining BMP with Apache Kafka, which is an open source message bus. The presentation looked at a number of scenarios where the combination of BMP and Kafka could be used as an effective local BGP anomaly detector, including route leak detection and inconsistent origination.

Much has been said about the growth in the BGP routing tables, and much more will be said no doubt. The size of the IPv4 BGP routing table is now at 575,000 routes, and there are no signs that this growth is slowing down anytime soon. Of course while all these route entries are jammed into high speed memory of every 'core' routing in the Internet, the issue is that most of this information is unused and is unnecessary. At any point in the network a BGP speaker carries some 250,000 superfluous routes that are more specific routes of covering aggregate routes that reflect the same intended path to the destination. DRAGON is an attempt to try and perform proxy aggregation in a router to try and remove much of this superfluous information. Other than an academic exercise I'm really not sure what the value is. If you really wanted to reduce the route decision space a FIB compression algorithm has a far greater potential outcome in terms of reduction of the size of the decision space, or if you really wanted to compress the routing state then perhaps looking at actual data would help. As Brian Field of Comcast reported at NANOG 64, in a 6 day period in their network they observed that some 415K entries had no traffic at all. Some 90% of the data traffic handled by the routers was directed to 3,156 distinct routing prefixed, and 99% of the traffic was sent to 25,893 prefixes. One possible response is to load the in-line FIBs with a far smaller "core" of active IPv4 prefixes, and send a default route via a tunnel to a nearby Internet egress point.

The relationship between Content Distribution Networks and other parts of the Internet has never been completely smooth. The work on Client Subnet in the DNS is one example of efforts to make the DNS a little chattier about who is asking the question to help the responder give the "right" answer. But of course the relationship between two addresses on the Internet depends on the routing state, and this is the topic of the "Whack-a-Mole" routing presentation. BGP does not use "performance" or even "latency" as a routing metric. It tries to minimise the AS path between any two addresses and sometimes that's not optimal for some consumers of the outcome (such as the CDNs in this case). To date there is no clear answer - routing does not allow the form of differentiated forwarding that is being called for, and no Virginia, SDN won't help here. One very common response for the larger CDN providers is to place replicated content sources everywhere, so that long line routes are minimised simply by making every client "close" to an instance of the data source. Or you can alter the content application to be more tolerant of longer latency and higher jitter with exposing all of the vagaries of the long line connection to the end user. Or you can complain at NANOG that BGP routing doesn't do exactly what you want it to do!

## Tutorials

WiFi: See slide 5 of the presentation (http://www.nanog.org/sites/default/files/NANOG65 - WiFi training - FINAL-2.pdf) - enough said! Actually both the slides and the presentation itself are worth looking through - there is some excellent material here to consumers of WiFi services just as much to designers of WiFi service networks.

I sat in on Rafal Szarecki's presentation in the internals of packet buffering inside routers (https://www.nanog.org/sites/default/files/wednesday_tutorial_szarecki_packet-buffering.pdf). He explains that sometimes the external behaviour of a switching system may be completely anomalous, such as when a flow through the unit sees packet loss even when the ingress and egress port utilisation rates are less than 1% of the port capacity. Internally routers are multi-processing systems, and at the simplest level can be considered as a collection of ingress packet forwarding engines connected to a collection of egress packet forwarding engines, connected by a switch fabric. Evidently these days routers use "virtual output queuing", where instead of queuing packets against each output interface after they have been switched, each ingress element has as many virtual input queues as there are egress ports, and each egress grants switch access to requesting input ports. When coupled with a crossbar switching element, that can switch multiple inputs to multiple outputs simultaneously, the result is highly efficient. However, this can lead to some surprising pathologies under load, where high flow rates across some ports can lead to high drop rates on supposed unrelated ports. The presentation pack is well worth a review if you are into this aspect of networking.

## Cables and Carriage

The Telegeography presentation by Tim Stronge was exceptionally interesting. He talked on the current state of undersea cables. Notably, there is the rise of purely machine-driven communication. When you look at the use of trans-Pac and trans-Atlantic over time, the large content providers now haul more volume of data than ISPs. People are not the reason why the bulk of data moves across these systems - its data storage systems priming their local storage, so, as Tim puts it, the major traffic component now machine-to-machine. ISPs barely exceed 50% of capacity. Buying power on cables has shifted away from carriage resellers (carriers) to buyers (content). Telegeography estimate that there is some 12Tbs of traffic due to ISP operations, and 14Tbps of traffic due to content loading on the aggregate of the Trans-Atlantic systems. The content folk are now in a position to purchase fibre pairs of their own, or be a submarine cable partner. Secondly, he observed the disruptive issues of price erosion on cables. The Monet cable is significant as an illustrating erosive price pressures on cable systems. The price today Miami-Sao-Paolo is four times the price of London-NY, LA-Tokyo and HK-Tokyo per 10G (these are selling at approximately $10K per month per 10G), so the massive price premium on the Miami - Brazil run exposes a build opportunity, which is being taken up by Monet. This is a content provider core anchor cable. Some time back Google was a major in Unity with a handfull of carrier partners and the carrier partners dumped their capacity onto the market. The LA-Tokyo transPac prices post-Unity for 10Gbs went down from $70K p.m. to $30K p.m. due to the strong influence of these dumping actions have over the market. Google is a core owner in Monet, and this time the other 3 are Algar Telecom (Brazilian SP consortium) Angola Cables (wanted to link this to a BR - Angola cable which is frozen so its a stranded asset) and Antel Uruguay (dramatically over-provisioned), so it looks like Monet will also dump into the market and Miami-Sao Paolo prices will plummet.

It should be noted that this business is not just a technology business, but a financial venture business. If the wildest expectations of a cable's backers are achieved then they will be in a position to calibrate release of capacity in a cable at a level that continues to leave unmet demand, so that the price of access to a cable will include a hefty scarcity premium, and if all goes according to plan the cable owners will generate a revenue stream that repays the original capital investment within two years, and continues to pay solid returns for up to two decades. But today such stories of financial windfalls are few and far between, and the environment has its far share of failed venture partners who are forced into fire sales of stranded assets. (e.g. Angola Cables). Despite dumping and price slumps, undersea fibre pairs are disappearing - book capacity is now being sold off by the cables. Before today's data centre content priming rush noone needed that much capacity, but these days the fibre pairs are being sold off quickly, so new cable development is likely in 2016 - 2017 (there has been no new cable trans Atlantic since 2003). These days there are new cables coming up but its Microsoft, Facebook, Google, etc driving this. They are driving this not as resellers, but as owner/operators, and are not interested in creating artificial scarcity

conditions, and the result is that resellers are seeing dwindling margins as the content industry pushes a regime of cost base pricing with marginal rates of return on the original build capital.

In this business time can be crucial. If you can be faster relative to others then you can sell this speed differential. High Frequency Traders (HFTs) found a microwave link between Chicago and New York was faster than fibre because of the inherent advantage of propagation time over air (slightly lower than c) vs propagation time in fibre (0.65 c). Hibernia trenched and armored the trans-Atlantic run across the heavily fished Grand Banks. Telegeography calculate the advantage they managed to get from this slightly more direct route to be 2.9ms. HFTs will pay up to 20x (yes, twenty times!) the ISP rate for the fastest route. But HFTs do not have infinite demand, so the cable consortium needs to sell its remaining book inventory, which they will do with an induced latency via an extra spool on the cable so that the HFTs retain their relative time edge that they paid for, and ISPs will see a "longer" cable with propagation times comparable to other cable systems.

Carriers don't know to price 100G capacity. Submarine cables were slow to the market compared to land prices, and the factor Teleglobe are using here to reach this conclusion is the price multiplier between 10G and 100G services. i.e. Is 100G 10x more expensive than 10G? Or 6x? They observe a lot of variance in the multiplier at present. There is a view that this multiplier for 100G services will stabilise at 4x to 5x 10G prices.

It appears that new sources of demand for international capacity (the current large scale demands for content priming) is leading to new infrastructure builds that displace traditional carrier resellers with owner/operators who use the capacity within their own business. The dumping of excess capacity on the reseller market disrupts this market leading to continued price erosion.

### IPv6 at Telus

Telus, the connectivity provider to the meeting gave a presentation on their experience leading to their deployment of IPv6 in their retail products. Their results are clearly visible (http://stats.labs.apnic.net/ipv6/AS852). They use MPLS on V4 in the core network, and implemented V6 as 6PE, which was a very minor change to their network architecture. However, as a consumer ISP, so a large part of this is a story of the high investment by their consumer ISP division in process and in having staff familiar with the process. In this context IPv6 is not the same as IPv4, and it can require careful consideration in some cases. There was a long lead time for this work: Telus started including IPv6 peering in 2012, some 3 years ago. Interestingly they used their conference provision of IPv6 for IETF, NANOG and ARIN taught them some valuable lessons about IPv6 in their production environment. They upgraded their managed Home Gateways, and sequenced this in a way that did not produce a large scale flood, using DHCPv6 as their prefix distribution mechanism.

They are doing V6 in their mobile network and evidently using NAT64/DNS64 for Apple iOS and 464XLAT for Android - though it appears that this is not a completed deployment - he cited a rather unique problem of Canadian users roaming too close to the US border and getting a handover dropout that lasted for minutes. They are working on this. They also had a rather strange problem of the PTMU packets being too big and getting dropped!

### Lightning Talks

Quick talks for quick takeaways: My takeaway, the CZ.NIC folk, who have done great work with BIRD for BGP and the Knot Authoritative server are working on Knot as a DNSSEC-validating resolver (https://gitlab.labs.nic.cz/knot/resolver). QNAME minimisation is included, and not only is it on by default - you can't turn it off!

As usual with NANOG, the presentations, both as slides and the video are all archived. https://www.nanog.org is a good starting place to explore these for yourself.

## 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. He has worked as a an Internet researcher, as a ISP systems architect and a network operator at various times.

*www.potaroo.net*

## Disclaimer