June 2015
Geoff Huston

Diving into the DNS

The North American Network Operator's Group held its 64th Meeting in San Francisco in early June. Here's my impressions of some of the more interesting sessions that grabbed my attention at this meeting.

Policy: US Regulatory Matters

At the start of the year the US FCC voted to reclassify Broadband Internet access services under Title II of the US Telecommunications ACT, effectively viewing Internet access providers as common carriers, with many of the rights and responsibilities that goes with this classification. This "new" framework of the FCC for these services is described through 3 "bright line" rules: No blocking, No throttling, No Paid Prioritization. There is also a General Conduct rule about no unreasonable interference and no unreasonable disadvantage, and enhancements to the Transparency Rule relating to the nature of disclosures to customers over price and terms of service. The FCC has discretion not to enforce particular terms of the Telecommunications Act where the FCC believes that it is not necessary, and in this case the FCC has exercised forbearance over rate regulation, tariffing and unbundling rules to these Internet service providers. The latter, forbearance over unbundling, may well prove to be inadequate!

In many ways this is not much of a change to the operation of the Internet access services sector in the US. Much of what happens today will continue. There is no direct change in the interconnection rules, nor any other major changes. The comment was made that the FCC is adopting a position of "regulatory humility" and says it will learn more as things progress. That's certainly a unique admission from a telecommunications regulator, but perhaps these are indeed unique times of fundamental change in this sector.

One thing is clear: this FCC action marks the end of the telephone network and its regulatory structure as we knew it. These days voice is just another Internet application.

The issue now is one of where next? To what extent will further legal challenge influence the FCC's rule making? To what extent will the emerging recognition of the Internet as the essential public telecommunication shape future rule making? What is the appropriate regulatory intent?

Network Operations: Automation

It has been pointed out that the Simple Network Management Protocol, SNMP, is around 25 years old, and between that and a plain text command line interface that is still the mainstay of most network operations. SNMP is used to gather data and the Command Line Interface (CLI) is used to push configuration commands back to the network elements. These days we appear to have renamed the data gathering exercise "telemetry", we have started playing with models of explicit "push" rather than the "pull" periodic polling model of SNMP, and taken the leap away from the rather obscure ASN.1 data encoding and headed on to use highly fashionable JSON data encoding. On the CLI side it appears that not an whole lot has changed, in so far as it still a set of scripts and regular expression

parsing and a set of rules, and whether its Ansible, Puppet of any other of these frameworks for codified rules the underlying models is much the same.

But one thing has changed, and that's the push from open software into this space. For many years it appeared that each outfit had its own locally developed network automation scripts and expended much effort in maintaining them. With the push to use one of these open software frameworks much of this local software maintenance effort is no longer required and instead attention can focus on higher level rule sets that attempt to correlate particular observations of network state with remedial actions. In other words we are seeing these systems evolve from simple outage detectors and diagnostic tools into preventative intervention.

However, it is always useful to bear in mind that when you combine even simple network structures, such as load balancing across multiple physical circuits, with edge-managed data flow control as we have with TCP data flows, the results can often be quite surprising. A network problem described by Facebook's Peter Hoose was problem they called "microbursts", where one of these physical circuits was operating at capacity while the other circuits in the bundle were not. The response was to reset the load balancing mechanism at the router in their network management system, which appear to solve the particular problem but unable to prevent its recurrence. It was only when repeated iterations of this form of network response proved to inadequate did they look to the traffic profile, and what they found was that the various TCP flow control algorithms had markedly different responses to small burst periods of relatively small congestion-based packet loss with the Illinois and Cubic flow control algorithms recovering extremely quickly and the Vegas and Reno systems taking far longer. It just shows that you can't look at the network in isolation from the protocols that make use of it.

IPv6: A Shift in Gears

I chaired a panel on IPv6 deployment at NANOG. It appears that the last 12 months has been a critical period in the story of IPv6 deployment, and while the overall Internet-wide numbers have shifted by just a few percentage points, the story in some countries is markedly different, and in Germany and the United States the consumer market access providers in both fixed and mobile services have taken up IPv6 with some serious deployment effort, and this is being matched by a similar uptake in support from content providers. The panel had John Brzozowski from Comcast, the larges access service provider in the US, Gaurav Madan from T-Mobile, the US mobile carrier with a model of IPv6-only mobile service network with an IPv4 overlay, and Paul Saab from Facebook, who have large scale IPv6 support in their network.

Almost one half of Comcast's customer base now is able to use IPv6, and when you add in efforts by T-Mobile, Verizon and AT&T then the surprising observation is that today almost one quarter of the US user base is now using IPv6. The Internet is a classic example of a network effect, where there is safety in numbers and providers tend to follow each other in order to ensure that they are not isolated. So what is the follow on effect when one of the more influential parts of today's Internet performs a relatively rapid deployment of IPv6? Some other markets have also jumped, including Germany where IPv6 penetration is also at approximately one quarter of their national customer base, and Belgium where IPv6 is used by almost one half of the users in that country.

All three panel members report that IPv6 provides a superior customer experience. Perhaps its due to the relative lack of network middleware compared to the extent of network manipulation that occurs in IPv4. It could be the lack of NAT state creation and maintenance. Or any one of a number of related factors. But in a world where milliseconds of delay can be measured in hundreds of millions of dollars in keeping a user's attention, then this observation that IPv6 shaves away some elements of delay, then no wonder Facebook is keenly interested in supporting this protocol. I can't help but wonder if this entire IPv6 saga is finally moving from a prolonged state of a seemingly endless onslaught of various forms of vapourware to one of real substantive progress - finally!

Protocols: Google's QUIC

I also learned something else in the IPv6 Panel session. In the mobile space there appears to be a significant level of competition for control of the user and the user experience. We have the mobile carriers, WiFi access operators, the handset manufacturers, the handset operating system and of course the applications themselves. Most applications make use of the underlying services from the host operating system, including the DNS name resolution service and the local TCP protocol stack. Facebook has taken a position of what I can only describe as a "paranoid" application and reduces its level of dependence on the host operating system to the bare minimum, and instead has loaded its own DNS resolution libraries and its own TCP flow control protocol into its own application, so that when the Facebook application communicates with Facebook services its Facebook itself that is driving both ends of the communication. We are going to see more of this.

But this is not the only case where the application itself is defining the way in which it chooses to communicate. Google has also headed into this space with QUIC.

QUIC is a reliable multiplexed encrypted UDP based transport. It subsumes the functionality that is provided with the Transport Layer Security function (TLS 1.2), the data reliability and flow control functionality normally provided by TCP and the multi-streaming parts of HTTP/2.

These days using UDP as a substitute for TCP falls foul of various forms of NAT middleware. TCP does not have the same problem as TCP uses explicit end of session signalling with the FIN and RST flags. A NAT can remove its session binding for a TCP session when it sees a session complete with these flags. But UDP has no concept of a session, and NATs are left with guessing that there is a session and when a session has ended. Furthermore NATs have very different UDP behaviours which only exacerbate the issues in trying to create a reliable protocol on UDP in a world of NATs. QUIC assumes that a NAT will keep a UDP binding active for a minimum of 30 seconds of idle time, and will continue to maintain the UDP binding as long as there is active traffic. More aggressive NAT behaviour will break QUIC. It was also noted that QUIC requires a 1350 octet sized MTU.

QUIC uses CUBIC flow control, with a number of additional TCP behaviours, including Forward ACK recovery (FACK), Tail Loss Probing (TLP) and Forward Retransmission Timeout recovery (F-RTO). However QUIC uses some additional behaviours that step outside of conventional TCP. For example, retransmission uses a new sequence number to disambiguate instances retransmission from the original data stream. Also coming is interleaved simple FEC, which would allow a received to perform reconstruction of a lost packet within a burst, support for multipath connections and lowering handshake overhead.

QUIC eliminates much of the TCP + TLS handshakes that occur on conventional session start. If a QUIC client has previous conversed with a QUIC server it can start sending data in the initial packet. Google claim that 75% of Web connections can make use of this cold session restart capability. Google claim that approximately one half of the sessions between Chrome browsers and Google servers are made using QUIC. They also note that 70% of the traffic from Google is encrypted. It is noted that the entire QUIC header is essentially a UDP payload, and this means that it is included in the encrypted envelope.

This is an interesting development in the so-called "protocol wars". By walking away from TCP Google are essentially removing the entire stream from direct visibility on the network. Middleware cannot perform TCP header inspection and attempt to manipulate the flow rates by manipulating TCP window sizes. Middleware cannot insert RST packets into the data flow. Indeed almost nothing is left visible to the network in a QUIC session other than the IP addresses and UDP port 443. This is one answer to the claim that the pervasive deployment of intrusive middleware was throttling the Internet: simply remove the network's visibility of the entire end-to-end data stream control headers and leave behind just a flow of UDP with encrypted payloads. Its an interesting development in the tensions between the "ends" of the end-to-end Internet and the "middles" of the network packet carriers.

Scaling Forwarding

Comcast has undertaken a lot of work in deploying IPv6 in recent months, and they are now looking to the point where IPv4's position will be waning in their network, and what this will mean.
One aspect that they have been looking at is forwarding hardware. Comcast run without a default route in their cores, and this means that they are running their interfaces with some 580K entries in their line card FIBs. And of course this number continues to grow. But if the traffic intensity of IPv4 is going to wane at some near term time then can they look forward to using line cards in the future with drastically smaller FIBs?

Comcast's Brian Field shared one very interesting observation: 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. This would reduce the size, power and cost of line cards in large amounts of Comcast's infrastructure and at a performance cost of tunnelling the traffic destined to little used destinations. It's an interesting thought at this point in time.

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