May 2014
Geoff Huston

NANOG 61

This was a pretty typical NANOG meeting, with a plenary stream, some interest group sessions, and an ARIN Public Policy session. The meeting attracted some 898 registered attendees (!), which was the biggest NANOG to date. No doubt the 70 registrations from Microsoft helped in this number, but even so the interest in NANOG continues to grow, and there was a strong European contingent, as well as some Japanese and a couple of Australians. The meeting continues to have a rich set of corridor conversations in addition to the meeting schedule. These corridor conversations are traditionally focused on peering, but these days there are a number of address brokers, content networks, vendors and niche industry service providers added to the mix. The meeting layout always includes a large number (20 or so) of round tables in the common area, and they are well used. NANOG manages to facilitate conversations extremely well, and I'm sure that the majority of the attendees attend for conversations with others, while the presentation content takes a second place for many.

That said, the program committee does a great job of balancing keynotes, vendor presentations, operator experience talks and researchers.

On the Hidden Nature of Complex Systems – Dave Meyer

Large scale engineered systems often borrow from nature, and in nature we see systems that are apparently complex, yet are surprisingly robust in some aspects and at the same time fragile in others. Why is this? And perhaps also its useful to ask: what is complexity? Where does it come from? How do you identify it? There is no quantitative foundation to the concepts behind the simple mantra of "simplicity is good"

What is complexity and why is it hidden? Dave asserts that complexity is about hidden structure that provides a system with robustness in the face of component uncertainties. Its hidden because of the intention of layering interfaces, that expose functions, but not mechanisms of implementation.

Dave asserts that complexity is a necessary element of robustness, but its a function where robustness rises, then trails off as the complexity of the system increases. Robustness is a generalized description of system outcomes, encompassing, scalability, reliability, efficiency and modularity. Formally, its the preservation of some property or outcome in the presence of component uncertainty. Fragility is the opposite of robustness, but it is possible for a system to be robust in the face of some perturbations, yet fragile in the face of others. So there is a well understood concept of Robust Yet Fragile (RYF) in systems.

Dave points to a recent result that suggests that there is some form of overall conservation, where increases in robustness in one aspect of a system creates increased fragility in another. Also, fragility and scaling are related. i.e. a big event may be "worse" in terms of the impact on a system than the sum of a set of small events. Understanding RYF is the challenge. it pops up everywhere. The challenge is to figure out the incidence of catastrophic events within the realm of the operation of a system. Its easy to create robust features, but far harder to understand the creation of fragility here.

Dave then heads to the Internet and the hourglass architecture. And today this exists nowhere, and everywhere. Instead of the "clean" layered model of system composition, we see tradeoffs - metal vs virtual, hardware vs software, convergence time vs held state. Are there fundamental generalizations that can be made about the tradeoffs in play here? Dave asserts that the hourglass model is the way we respond to this, whether its IP, OS, SDN or open source. There is on many of these systems a single "waist" which functions as the core adaptation / messaging / translation function that takes otherwise incompatible elements and allows them to be integrated into a single system.

Can you build better networks if you take this into account?

The architecture we have right now is a byproduct of coping with these tradeoffs involved in building large scale systems that are robust. Dave is heading with John Doyle into a mathematical model of this aspect of engineering as the next step - I think he is heading to a direction that posits that if you can bend the curve of complexity vs robustness and push it out, you can create systems with greater levels of complexity that still manage to increase robustness without a corresponding penalty in fragility.

Optics and End to End with Len Bozak

Len Bozak has been in this industry for many decades (a co-founder of Cisco in the late '80s), and spoke at NANOG on his perspectives on scaling, optics and end-to-end.

Optical transmission has finite bounds - there is a limited set of wavelengths where glass is highly transparent, and there are some fundamental limits based on signal to noise imposed by Shannon and Nyquist. What was for me somewhat surpassing to realize was that fiber systems used a simple on/off encoding for years, and this scaled up to 10G without need to explore ways to improve the ways to use symbol encoding to increase the level of bits per symbol. Contrast this with the analogue voice modem world, where the original 300baud model was quickly replaced by increasing utilization of phase and amplitude shifting to get to 56K from the same basic voice system. Optics is heading there now, with the use of polarization and Fourier space encoding to improve the capacity of a wavelength. 100G systems are not particularly challenging, in so far as it uses only two polarization planes and limited QAM to keep the signal within the dispersion characteristics of cable systems and within the sensitivities of today's DSPs. But increasing from this requires the square of the QAM for each doubling of bit capacity. While 100G uses 4QAM states, 200G requires 16QAM, and 400G requires 256QAM. This latter target is a stretch in terms of current understanding of capability, and of course management of the increasing power spectrum being pumped into the cable.

Len made the point that the original 9600baud modems were a feat of putting FFD algorithms on a chip that used the basic 300 symbols per second rate. We are being faced with the same challenges of placing increasing power into physically limited spaces in order to increase the bits per symbol in optical systems.

The next question is - how can you terminate a 400G optical system? A termination could be constructed, but none has done so, so far. the option exists for 4 x 100, but how does this scale further into Tb? We need to improve the data density - we will see 1T / 2T systems in the predictable near term future, but the termination and ancillary support are open questions - bit rates per pin are a real issue here.

So there is a certain level of competitive pressure at play in switch design. Increased functionality in the units equates to increased power and cost. At the same time increased carriage capacity per wavelength requires increased capacity per pin, requires ever increasing levels of complexity, and the power and cost scaling functions are an issue here. One approach, say en, is to keep it simple, keep costs down, and allow improvements in modulation to drive improvements in transmission. And also, a warning to be extremely careful in feature requests at this level, as even simple requests can lead to complex (and potentially fragile) solutions!

Netflix - Traffic Engineering with Open Source

CDN issues as seen by Netflix. Netflix has grown quickly, and while Netflix used to use a CDN (Limelight) in its early days, it has now pulled this in-house into its own feeder network. Netflix uses the model of loading hw caches that are located within ISP networks. These are a collection of "island" caches, and where there is a content flow demand then the response is a dynamic response to determine which cache will be used to service the demand i.e. a user initiates a content stream request and the CDN needs to assign a cache to the request that is capable to supporting this stream..

They have a conventional BGP system with netflow, and are adding BGP multi path as a means of feeding a more complete picture of the edge router's choices back to the data collection function. Netflix is using pmacct - an open source project that can take in net flow data (IPFIX, sFlow, libpcap and BGP dumps) and store the data in various database formats. (www.pmacct.net)

Can be used with net flow to construct traffic matrices, which can be used for capacity management. This is combined with the BGP add-paths attribute to combine netflow, pmacct, and add-paths.

Internet Atlas

This is a project that is intended to maintain Atlas of the Physical Internet, namely a catalog/database of the geolocation of nodes and links, intended to use for visualizations and analysis of incidents to network service

Paul refers to Sean Gorman's PhD work, reported in Washington Post articles, where Gorman's 2003 work took publicly available information about US infrastructure — power grids, fiber optic networks, etc. — maps it all, and uses algorithms to find the weak points. the compendium was, evidently, classified. Interesting story about how using technology to crunch together a bunch of unclassified information results in aggregated knowledge that the government has an interest in classifying. 2003 was a time of heightened sensitivity to such exercises, but Paul's work seems extremely similar to me, and I wonder is the same sensitivities will be exercised here, but I am also reminded of an earlier exercise in the amalgamation of a variety of public sources to produce a design of a nuclear explosive device where the result was, obviously, classified.

The procedure here is to search data sources to establish locations of physical plant, then transcribe to a uniform data format and then verify against current network maps. The upside is that this data is not highly dynamic - it grows but does not change once added into the catalogue. This project is aiming for high resolution - street address granularity where possible. It heads all the way from inter-continental to national or even regional. The presentation is about the potential of taking public information sources to create a highly accurate picture of network infrastructure plant. (internetatlas.org)

There are some obvious parallels here to the RIPE MAT Geloc project, but some pretty obvious differences as well, including the use in RIPE for crowd sourcing, and the imposed limit of a city level of granularity in the GeoLoc coordinates.

Ethernet VPNs

This is more along the lines of a typical vendor pack of problem/solution style, espousing the virtues of this particular technology. It is one more approach in the area of VLANs and MPLS and is little distinguished from the earlier generation of X.25, FR and ATM approaches, where the network enforces a logical segmentation into communicating subsets. It can be point-to-point or broadcast, or multicast - the basic architectural approach is the same in all cases. The difference over time is the point of application, where the original network-centric approach of such virtual circuit technologies has evolved into the space of high speed, high density data centers and their interconnection.

So this particular approach to L2 VPNS uses MP-BGP to propagate MAC addresses, rather than flooding, and the assertion is that this measure increases scaling and flexibility of the resultant set of VPNs and their common host. This can also leverage on the topology maintenance function of BGP to eschew spanning tree and potentially TRILL I suppose. The underlying data plane is not fixed in this model - it could be MPLS, or provider backbone bridges or network virtualization overlays.

Yes, its different to other L2VPN approaches, but does this produce substantially superior outcomes in terms of cost and scale? Or is this bit twiddling? It's not an giant leap forward in networking architecture in any case, but for data center operators and exchange point operators this approach may represent a better set of tradeoffs than the other VPN approaches.

Cyber Risk Insurance Panel

The actuarial sector closely tracks the rest of society, and in particular tracks its calamities, real or imagined. So in addition to the conventional portfolio of life, house, car, fire, food, earthquake and public liability insurance, one can also obtain insurance against cyber attacks. While this may not be a matter of concern for individuals, it is proving popular in the corporate world, where the leaks of customer databases, and the customers' credit card details, passwords, and similar, can represent a significant risk for any company.

So the increasing visibility, regulatory oversight and understanding of risk have created a window for the insurance industry to offer cyber risk insurance offerings. These focus on the potential victims, and the coverage includes provision for notification, rectification and litigation arising from data breaches.

There are some interesting stats in the material presented in the panel session. Claims in the area of data loss / compromise broken down by root cause: hacking: 24%, rogue employee: 20%, human error - 14%, privacy policy - 9, lost / stolen hardware - 22%. There is also variance across sectors - e.g. health care, retail, tech or professional services. The closing slides had one memorable quote, attributed to Mike Tyson: "Everyone has a plan - until they get punched in the face."

Heartbleed - Mike Sullivan (Cloudflare)

A presentation that Mike states was actually about "how certificate revocation is broken and endangered our network." Cloudflare is a web hosting intermediary, with pops all over the place, and an any cast routing system. They say that they see some 800M IP visitors per month, and this represents some 1.6 billion endue users, or 75% of the Internet's user population per month. They use a conventional array of technologies, including the DNS, HTTP and of course HTTPS, based on OpenSSL, which leads straight to Heartbleed.

Heartbleed introduced into OpenSSL on the 14 March 2012, version 1.0.1 included "TLS/DTLS heartbeat support" The memory leak from the server is completely invisible in retrospect at the server. Whatever was in memory at the time might have been included. A server might have disclosed decrypted data stored in memory of concurrent or recently running images. Potentially this includes private key values.

So test this theory Cloudflare decided to crowdsource the test, and they set up a challenge on a target system to crack a private key on a heartbleed-vulnerable host system, and 10 hours later the private key was cracked. This exploited a second vulnerability in OpenSSL, where the memory used by OpenSSL for private key manipulation was not cleaned up, so the key value was left available in memory.

Given this revelation, Cloudflare then revoked all of its certificates in a comprehensive emergency key roll. There was massive traffic spike in consequence, not only to retrieve a copy of the new certificate, but also because Cloudflare's CRL of current, but revoked certificates grew from 22K to 5M, so the relying parties (including browsers of course) needed to update their local copy of this CRL. Mike related that in the case of Chrome this required a code path for Chrome.

Cryptech - Randy Bush

A presentation on the issues involved in open sourcing a trustable HSM. I have seen previous versions of this presentation, but some details are clearer here, including the entire concept of assurance in the tools, compilers, hardware, that are used to construct the HSM. The issue is not just in constructing the hardware, but being assured that all the components and the tools used to construct the components operate with integrity and have not left vulnerabilities in the constructed artifact. This was an interesting insight into the design of an HSM built using an FPGA, including the measures relating to being tamper-proof. This is a planned 3 year project, and the group is 1/2 way through year 1 it appears. (http://cryptech.is)

DDOS and Extortion

Extortion demands are now getting involved in DDOS attacks. in the case presented in this session a hosting ISP is a victim of a DDOS that included a demand for payment to stop the attack. The operation hosts webcams, and are used to large traffic volumes, so the attacks were not significant at the start, but they quickly escalated. The first major attack was a large syn flood against the webcam site and the front end load balancers immediately fell over. The second phase of the attack was a SYN attack on the name servers. they fell over. The initial response was to block tcp on the name servers, then worked to construct a large ACL of the attack address profile. The attacks continued, and they responded on a case-by-case. The attacks increased in intensity, rising to 50G, with millions of packets per second of SYN attacks. Certainly this presented the victim with a dilemma, as the cost of the continuing response to the operation was larger than the ransom being demanded, but in any case such as this, its highly likely that any form of payment would ensure subsequent attacks and subsequent extortion demands.

This is a relatively ugly space that is evidently liked to various forms of criminal activity., and relating this presentation back to that of the cyber risk insurance session, its unclear where this heads. The credit card industry appears to have taken the position to writing off losses from credit card fraud, and instead defrays this cost against the fees associated with credit card use. Its an open question as to where we are heading with this cyber extortion activity.

DNS and Root Zones

The combination of DNSSEC with a signed root zone brings up the question of the role of the root zone in the overall picture of DNS operation. Warren Kumari spoke to an internet draft that espoused an alternative approach that a recursive resolver could simple obtain the root zone public key, perform an AXFR of the root zone, and use the key value to validate its integrity, and then be able to provide all the response directly that would be provided were it to forward queries onto a root zone server. Google's PDNS does this, and it appears that others so as well. There are some unresolved issues, noted in the draft about handling the unsigned elements in the root zone, but these issues do not appear to be major obstacles. The presentation attracted some critical comment, but I couldn't help but notice that the critical questions came from existing operators of DNS root instances. The issue of the selection of the anointed root zone operators has been a recurring pain point in debates over DNS governance, and while the wide scale deployment of any cast clouds has largely addressed the engineering substance of scaling the root zone, the cachet of being a root zone operator appears to persist, as does a certain level of continued debate over DNS governance in matters of the root zone.

Knot DNS - Ondřej Surý

One of the reasons why the Heartbleed vulnerability has such a widespread impact is that it appears that almost everything that uses security functionality relies on OpenSSL. This monoculture itself creates critical vulnerability, in so far as any issue with OpenSSL affects a massive number of systems

using secure services. A similar argument has been made about BIND, and in response we've seen NSD, Unbound, PowerDNS, and Knot. I like Knot. Indeed after listening to Sean Kerr speak on the decline and Fall of Bind 10 last month (https://ripe68.ripe.net/presentations/208-The_Decline_and_Fall_of_BIND_10.pdf) I like Knot even more. It is not all things to all people. It is an authoritative name server, But its good And it gets better. Knot 1.6 is copying with online DNSSEC signing (set up a pointer to the keys in the conf file and set it to sign as it answers). It is coming out with a really neat way of synthesizing the reverse IPv6 records, in a signed reverse zone. They are also looking at switching from OpenSSL to GnuTLS, and not because its better, or worse, but because its not what everyone else uses. I like Knot, and this presentation was a good illustration of why.

dnstap

There has always been a lot of interest in DNS analytics. Not only does the DNS provide a window on the entire Internet, the DNS itself is a valid subject of study as a massive distributed system. However the conventional tools for logging are frustrating. Packet captures require extensive analysis to reconstruct queries and responses, and frustration with this lead to other approaches, notably from Florian Weimar. So this is the equivalent of a packet capture at the IP level - its a DNS transaction (query and response) capture. Its a verbatim wire format capture tool, so there is no overhead of text formatting and state reconstruction (basically its the tee tool for DNS).

DNS Record Injection Attacks in CPE

Home CPE equipment has a rich and broken history in the Internet. They were an integral part of the deployment of the first wave of DSL and cable based services, and now are ubiquitous. These systems typically provide connectivity services to an upstream provider, and also provide a local NAT, security services, and a local DNS resolver. They are high volume low cost units and are typically purchased and installed and operate in an unmanaged mode. This is a recipe for disaster, and its no surprise that the open resolver project's 30 million open DNS resolvers points a large accusing finger at CPE devices as being the major component of this problem. This presentation points to ways that the operation of these DNS resolvers can be readily compromised by simple forms of attack from the Internet side.

Facebook Infrastructure

Many large service providers see their internal service architecture as a commercially sensitive asset, and Facebook is no exception to this observation. This is a presentation that attempted to describe some of the approaches to the way Facebook perform capacity management on their network, but to do so without providing much in the way of specific detail, and certainly nothing in the way of commercially sensitive material. The presentation easily achieved the latter objectives, but I found it challenging to see that it also achieved the first objective as well. It appears that Facebook use a form of virtual circuit-based modeling where the unit of traffic management are edge-to-edge flows across their infrastructure, where individual flows have clearly stated traffic requirements. At this stage I am think "this is the computer science knapsack problem", but the presentation was at an abstract level, and the particular algorithms used to achieve an efficient loading of demand onto available capacity were not described here.

Submarine Cable Panel

This session featured presentations from Hibernia on the cable commissioning process, and Telefonica on cable maintenance.

There are a number of issues of jurisdiction issues associated with this, and the relevant treaties extend back to the original transatlantic cables and the UnderSea Cable Protection Treaties of 1984. There is also additional protocols of the High Seas from 1958, and the 1982 Law of the Sea (ratified by most countries with the notable exceptions of the United States!) What night look like "the se" is actually distinguished between the high sea, the continental shelf, the exclusive economic zones, the 12 mile

limits, the 3 mile limits and the 24 mile contiguous zones. The Hibernia used an example of their recent UK to US cable to highlight the steps in the process of commissioning a cable within this well populated space.

Oddly enough Google Earth is the starting point with a cable proposal. If there is interest the cable proposer would commission a desktop study, which would visit the designated cable landing sites, initiate coordination with marine and land authorities, examine fishing, shipping and marine zones, seafloor composition and sea surface depth. This desktop study would result in a cable design that can then be passed over to submarine surveyors for a detailed survey of the proposed cable route. At the same time the cable operation would initiate the process to obtain the various operating licenses, permits in principle, operational permits to undertake installation . This can be an extensive process, and the relevant authorities are based on the location of the cable. The requires on land to the high water mark are different to the requirements between high and low water, and from low water to the 3 mile point, thence to the 12 mine point, thence to the 200 mile point, to the boundary of the Exclusive Economic Zone. Evidently no permits are required for the high seas once the cable leaves an EEZ.

Cable maintenance operations are generally based on a geographic zone rather than per cable. There are a set of so-called consortium zone cable maintenance arrangements, covering the Atlantic, the Mediterranean, etc., and in addition there are a set of private cable maintenance arrangements. A cable owner purchases from one of these arrangements, and the maintenance operator maintains vessels. The presentation included an interesting map of the incidence of cable faults over the world, with cable damage "hot spots" off Japan and the Luzon Strait (earthquake), Singapore, Aden, Alexandria and Malta (shallow and high shipping levels). in the coastal regions, and in parts of the high seas an undersea cable has to contend with clam and scallop dredgers doing sea bed scouring, and beam drawers that can scour up to 0.5 m deep on the sea floor. There is also submarine drilling and deep sea mining, offshore wind farms, oil and gas wells, fish farms and shipping - the sea is an exceptionally crowded place these days!.

Cable operators with optical amplifiers have been performing regrooming of the surface optics and electronics to extract higher capacities from the cable. Cables can be regroomed from 10G lambdas to 100G, and with a suitable dispersion profile 200G is achievable. A reduction in the size of the WDM guard bands also allows additional lambdas. SAM-1, for example, was installed as a 1.9T system in 2000 with an intended 25 year operational life. 14 years on it is now a 19.2T system, and there is the prospect of further re-grooming to support 200G lambdas and a 40T overall system.

I suppose I am betraying a deep personal interest in the engineering of submarine cable systems when I say that this session was a fascinating insight into a highly specialized area of our industry.

Internet Governance Panel

I'm not sure that this needs much in the way of further words, given that many many words have been said on this topic in many venues in recent times, particularly relating to the issues around the "IANA transition." I did hear a number of new observations, and one, from a former IANA manager, who noted that the current NTIA process step in the production of the DNS root zone was a sanity check on ICANN's administration of the root zone of the DNS as a part of checks and balances. The governmental relationship also allowed for issuing exceptional licenses for US nationals working for IANA to perform functions in the service of countries that are listed by UN Sanctions. And of course the oversight function was as a contingency in the event that ICANN, as the IANA function operator, was captured by sectional interests and ceased to fulfill its obligations in a fair and balanced manner. Obviously any transitional mechanism would need to provide similar capabilities and provide similar forms of balances and contingency capabilities, in an environment, where there are, to quote the speaker, "large governments, who presumably have nuclear weapons, and large private interests, who presumably have their own nuclear weapons" . No doubt much more will be said on this topic in the coming weeks and months in other venues, but I thought this session was pragmatic and forthright, particularly relating to the role of the IETF and its relationship with the entirety of the IANA function.

464XLAT

And you thought that NANOG would have no IPv6 sessions this time around? No! Cameron Byrne of TMobile explained the motivations that led TMobile to an approach that avoided dual stack, and instead uses an all-IPv6 wireless access network, and performs the IPv6 to/from IPv4 translation at the boundary of their access network on the one side, and as a software module (CLAT) in the handsets at the other side. It also involves DNS interception and translation as part of the stateless address translation capability. And yes, it evidently all Just Works, for most forms of communication. There are some issues with radio to wifi handover and cache flushing, but the experience of TMobile is very positive.

Open Router OS architecture - Brian Field (Comcast)

Commercial routers and switches are generally packaged in an opaque manner. Operators can direct functionality via the unit's configuration, but they cannot add functionality or replace selected functionality. However one vendor is offering a unit that supports an open architecture using a Linux kernel, and this presentation described how they added segment routing to unit. What they did was to add a control user process to the router OS that processed incoming packets where the segment routing header referenced this local loopback and performed the segment routing transforms on the IP packet. As Brian pointed out, this allows others to create develop and deploy new control plane protocols on production routing platforms. It has fascinating possibilities.

Embracing Failure

This was a presentation on resiliency engineering and the approach of injection of faults and the analysis of the outcomes in order to improve the robustness of the system. There are various forms of the potential for failure of a large at scale system such as Netflix - bad code, byzantine failures, etc. The talk was predominately an informative about procedures used by Netflix to inform the process of network management. (netflix.github.com). I can't help but contrast this presentation with the opening presentation by Dave Meyer about the tradeoffs between robustness, complexity and the introduction of fragility. The impression gained from this presentation was one of increasing robustness without any downside, but I can't help but wonder about the perspective that there is a law of conservation going on here and increasing robustness in one are has a cost in fragility elsewhere.

There were numerous other sessions in the three days of NANOG 61, but I did not manage to attend everything. Archives, videos, etc as usual are at www.nanog.org

NANOG is one of the best technical meetings on the entire global calendar if your interest in in the network itself. If you share such an interest, and you would not have made it to the end of this article if you didn't, then I would recommend that if you can, show up at a NANOG meeting and engage with other who share the same passion for building and operating networks.

Disclaimer

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

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 the initial build of the Internet within the Australian academic and research sector. 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.

www.potaroo.net