The ISP Column
A monthly column on things Internet

October 2015
Geoff Huston

DNS OARC Fall Workshop Report

The DNS Operations, Analysis and Research Centre holds a 2 day workshop twice a year. These are my impressions of the Fall 2015 workshop, held at the start of October in Montreal. At the outset I note that there was less of an emphasis on the coopting of the DNS into an attack vector, and while there were some presentations on DDOS mitigation techniques for resolvers, other topics also were aired at the workshop. One of the new themes is the consideration of the amount of information leakage in the DNS, and there were some presentations on efforts to improve the privacy of the DNS, principally concerning efforts relating to the shifting of DNS queries and responses from the current open UDP transactions to ones that are conducted through secure transport sessions, using either TCP or a transport protocol that extensively borrows from TCP.

The full set of presentations can be found at: *https://indico.dns-oarc.net/event/24/timetable/#20151004.detailed*

DNS Privacy - Alison Mankin

> Interestingly, RFC 4033 (The DNSSEC Requirements specification) has no confidentiality requirement. The NSEC3 activity considered zone confidentiality but it went no further. The DNSCurve and DNSCrypt tools offer confidentiality, but were not taken up by IETF DNS mainstream at the time. Since then the pervasive security theme has taken root in the IETF and the focus is back on confidentiality and protection of personal privacy at the protocol level (RFC7258). The IETF's DPRIV WG has pushed out RFC7626 as its first document (the problem statement for the Working Group). This covers various points in the DNS where DNS data could be compromised, inspected by unauthorized third parties or otherwise misused.
>
> The two major initiatives being considered are DNS-over-TLS, and qname minimisation. There is interest in the draft draft-ietf-dprive-dns-over-tls, which assumes DNS queries and responses over TCP, using models of an opportunistic channel security profile and pre-configured security profile to be applied against a specific DNS server. It was noted that implementations of the approach are appearing The Qname Minimisation draft has completed WG Last Call, and will be sent for IESG review anytime soon.
>
> Further possible areas of exposure: NSEC3 is vulnerable to hash attacks by a well-resourced adversary. TLS uses cleartext domain names it its handshake. DHCP is similarly accessible. SNI exposes the domain name of the client, and as a response a reader should look at "domain fronting" in *www.icir.org/vern/meek-PETS-2015.pdf*. More parameters in the RTLS handshake are being protected in this manner.

TLS in DNS – Sara Dickinson

> A number of possible approaches are being considered.
>
> - STARTTLS uses port 53. Its a known technique with incremental deployment. But there is the potential for middlebox confusion when encountering encrypted port 53 traffic, and its also susceptible to a downgrade MITM attack.

- TLS using a new port will not interfere with port 53, but there is no firewall support and this may be an impediment.

- DTLS, which is basically DNS over TLS over UDP. There is the issue of UDP fragmentation and truncation, with a fallback to clear text or TLS. No running code as yet.

At the base here is DNS-over-TCP, used traditionally as a fallback in response to a TC=1 query response (RFC5966). A USC/ISI 2014 paper, "Connection-oriented TCP," shows that TCP/TLS performance is feasible for DNS. The work documented in draft-ietf-dnsop-5966bis contemplates re-use of the TCP TLS session, as you want to amortize the cost of session setup to spam multiple queries (edns-tcp-keepalive for long-lived DNS-over-TCP sessions). This allows a model of conventional query response sequencing or query pipelining with queries in parallel. TCP fast open (RFC7413) uses cookie exchange and data in the first SYN. TLS session resumption (RFC5077) allows abbreviated handshake using a session ticket. There are the usual caveats about system tuning under intense TCP load (including TCP connection state explosion).

The unbound resolver has added TLS in 1.4.14 as a last chance connection attempt, and LDS and NSD have TLS patcheds. The getdns resolver is heading in the same direction, using TLS port 1021, and options for strict TLS, opportunistic TLS and conventional TCP/UDP fallback, with pipelining, out of order processing and a configurable idle session timer.

RFC7525 - BCP for TLS and DTLS, specifies TLS 1.2 as supported and preferred, with recommended cipher suites. There is work on TLS 1.3, which is anticipated to obsolete TLS 1.2

DANE Adoption - Shumon Huque

This presentation was a report on a recent survey on DANE adoption (and DNSSEC of course).

DNSSEC: TLDs are largely (84%) DNSSEC-signed (https://www.huque.com/app/dnsstat/), but SLD sites (such as the Alexa sets) are largely unsigned. .com reports 518K out of 118M (0.44%), .net report 94K out of 15M (0.63%) and .org reports 53K out of 11M (0.51%)

DANE: In the .com and .net SLDs, a zone crawl saw TLSA records for HTTPS (443), SMTP (465) and XMPP (5269). TLSA records are around 1 to 5% of the DNSSEC-signed zones.

Benchmarking Authoritative servers and DNSSEC impact - Thomas Hlavecek

This is a report of a test of authoritative servers with the tcpreplay tool, applied to captured DNS queries. In particular the report is looking at the ratio of queries sent and queries answered under increasing levels of query load. He is looking at a simple model of threshold saturation, where responses match queries 1:1 to a break point which corresponds to server saturation, at which point the response rate falls as the query rate increases.

Unknown EDNS Option Handling – Vicky Risk

ednscomp.isc.org

Some 90% of resolvers are EDNS-aware, but a lower number (~60%) are compliant with all parts of the EDNS specification. This URL includes a number of compliance tests of boundary cases in

EDNS. ISC is wanting resolvers to test against this script to see if the issue is improving over time.

Benchmarking and Profiling Resolvers – Robert Edmonds

This is an interesting exercise in tool construction from open source components. He uses trafgen, ifpps (from netsniff-ng), tc (traffic control), perf-events and mpstat as tool kit. trafgen can set up >1Mqps (provides trafgen profile), and he feeds this to tc to control the actual traffic level he passes to resolvers under test.

Set up Bind, Knot, Unbound and dns-echo (control, and plots system utilisation vs queries answered per second

He then uses perf-events to find the code hot spots in the resolver implementations – which is really rather cool!

DNS over TCP – Joao Damas

This study took a UDP query set from two authoritative servers over a 1 day period (average ~300 qps) and fed this into a TCP simulator, varying the number of simultaneous TCP sessions supported. He looked at connection reuse levels when varying the number of connections, and saw the highest reuse with 300 active connections. In his sample set he saw no responses over 1500 octets (!) (including DNSSEC-signed responses.)

The study noted burstiness of the query traffic, and noted that once the TCP capacity is sufficiently large the reuse of connections is quite efficient.

There are some options here for treatment of TCP in resolvers and authoritative servers, and one option follows a guideline that I originally heard from Bert Hubert of Power DNS – effective recursive resolvers run hot! One is to dispense almost completely with TCP session idle timers and leave the maximal number of TCP sessions open that the host can maintain efficiently. The session set can be managed using conventional cache management strategies, and when a new client wants to open a session the least desireable session is killed to allow the new session to take its place.

DNS over QUIC - Kazunori Fujiwara

This presentation was a thought bubble of "what if" noting that at present he can't readily set this up, ass QUIC is relatively complex and there are no ready-to-use code libraries at this point. It looks promising in that QUIC is roughly the same as TLS 1.3. The presentation detailed his work to date in trying the set up a test bed on a FreeBSD platform.

Managing DDOS Attacks - Brian Somers

Uses a taxonomy of:

- Accidental - single source with repeated queries
- Amplification - falsified source with payload bloat
- NXDOMAIN - targeting an authoritative server with unique queries

Rate limiting based on common source address is a common response from the authoritative server. This talk is about combining query data from a set of resolvers to generate a distributed "freeze list" of domains to be rate limited. I'm not sure if this is much more than another brick in the wall of the defense armament escalation. It shows that there are potential benefits in pooling intense query patterns across a set of recursive resolvers to apply a common rate limit rule set, but it strikes me as another step in a path rather than the end point here.

Right now they (OpenDNS) receive around 95 billion queries per day, and drop 15 billion per day as suspected malicious query traffic.

Deployment of Nameserver Infrastructure - Christian Petrasch

A presentation on the automation of the staging infrastructure for DENIC authoritative nameservers with orchestration by Ansible etc. This was more of a presentation around operational process management and its automation than something that was intrinsic to the operational management of DNS servers per se.

Neutering ANY Queries - Olafur Gudmundsson

This is a presentation on the ANY query and its treatment by Cloudflare. I have always been impressed by the ability of the folk at Cloudflare to innovate in the DNS and provide efficient scalable solutions to serving signed DNS records in novel ways, and this was no exception. ANY is an unreliable meta-type. Is ANY meant to be "all?" Or "Some?" Or "Many?" Or even just "A & AAAA" please? In Cloudflare's case they find ANY is an expensive question to answer if ANY is the same as "all", as not all the various components of a name are available all of the time at the point when the DNS response is being assembled.

Cloudflare initially responded by saying NOTIMP ("not implemented" response code) as a response to ANY. However NOTIMP was perhaps an over-reaction, as ANY is used as a query type. For example Firefox used ANY as a shorthand for "A and AAAA." Qmail used ANY as a means of potential optimization, and it would fall back to explicit A and AAAA queries. In trying to optimize the approach to responding to ANY queries, an option here is to answer with what's convenient. They have observed that ANY answers do not pre-populate caches with specific RR types that may have been included in the ANY response, but only populate the cache with the ANY lookup query type. The problem is the open resolver zombies without a cache (the open resolver problem). They have no cache and there are users behind them.

So the approach taken was a readable cacheable response that can be readily synthesized as it contains mo domain-specific data of any form. They opted for a fixed HINFO response as a response to ANY. This can be generated on the fly, and signed on the fly with their ECC signer. There was a comment in the room agreeing with this approach, noting that HINFO is already so compromised that it has no serious purpose any more, so why not throw HINFO under the bus, and just use that as the ANY response.

There was a question about using DNSSEC OPTOUT to even stop signing HINFO!

Canadian Content Hosting - Jim Cowie

Once upon a time we boasted that the Internet blurred geography. We boasted that a web site located in Oonadatta could be seen by an entire world of Internet users, and an Internet user could browse every part of a world without borders. There were no "long distance charges" any more – it was one big flat Internet. Somewhere between the stunningly inappropriate mid-twentieth century models of content distribution that tries to insist that everyone has a binding obligation not to pass their precious content across these somewhat invisible national borders, and a world that is increasingly driven by a variety of nationalistic paranoia bought about through the rampant enthusiasm for broad scale surveillance, we apparently now want to erect political boundaries in the network, and place border controls on the flows of data. Do Swedish users communicate with other Swedish users along conduits and through servers only located on Swedish soil? Even 10 years ago it's a question that would've elicited more or less polite laughter in response, but these days it's asked seriously, and the answers pondered with equal gravitas. So when this presentation asks what proportion of the most popular content sites used by users in Canada are hosted inside Canada, it's a question, and an answer, that some would like to take seriously. Maybe its part of this post-Snowden world that there are such levels of concern about where the national political boundaries really exist in the Internet. So when this presentation observes that some 78% of domain names in the .CA top level domain are hosted by enterprises located in the US are we to congratulate the canny Canadians for finding a good deal, and equally congratulate these entrepreneurial US enterprises for their zeal in servicing international export markets, or are we to deplore this situation as major threat to the integrity of citizens' online data?

I am not sure what to conclude from this presentation!

Caching and Root Server TTLs - Matthew Thomas

This presentation was based on the RSSAC 003 report (*https://www.icann.org/en/system/files/files/rssac-003-root-zone-ttls-21aug15-en.pdf*). You would expect large TTLs to have less queries at the root, but this is not the case. The work was based on an analysis of the DITL data sets, and looks at the elapsed time between queries to the same root server for the same query. There are local peaks in the CDF of inter-query delay in the DITL at 12 and 24 hours.

The second part of this talk was an active measurement talk where a domain was set up with a 10 day TTL, and then open recursives were queries in a structured manner. The popular open recursives appear to perform time-nased TTL reduction, but they refresh their cache in 1 day (Ultra, Dyn). OpenDNS is a little different with one record having a long lived cache time with a number of more frequently refreshed cached records. Google's PDNS exhibited some jitter, but principally stored the cached records with a 6 hour TTL.

Atlas and F-Root Anycast Selection - Ray Bellis

These days anycast drives the expansion of the root zone, and most (but not all) of the 13 root servers are in fact a constellation of synchronised servers that are intended to respond identically. But from the inside there is an interesting couple of questions: how effective is an existing root server constellation? If the anycast constellation were to be expanded, then where would additional anycast servers make sense? Both of these are difficult questions, and one suspects that the current anycast placement processes are more about "pins in a map" than a deep understanding of DNS query flows.

Ray has looked at RIPE measurement 10304: every root server, every 240s "hostname.bind CH TXT" query. He is mapping with the Atlas geo data to visualise the "client constellation" of each F Root anycast instance, and, in particular, identify those locations that are seeing DNS query response delays in excess of 200ms. Obvious anomalies are visible in this approach, such as French users using the Atlanta F root instance.

However the "what if" question is extremely difficult to answer from this approach. You can tell what is sub-optimal, but that's not quite the same as an optimal placement of a further element in the anycast constellation.

But this leads to the next question: Is anycast an eloquent expression of failure?

This gets back to the question of what is the problem that we are wanting to solve, and is this the problem we appear to be solving. One can't help but wonder if faster NXDOMAIN responses really matters. It appears that the actual benefit is not faster NXDOMAIN, but attack minimization. There were times when single instance root servers were DDOS'ed off the net, and large anycast constellations certainly help in minimizing the effectiveness of attack, even if the attack is a widely distributed attack.

But at this stage the thought process heads into ways to leverage DNSSEC to alter the entire root zone distribution architecture and see if we can eliminate even these 13 anycast constellations as critical points. But that is heading way beyond Ray's presentation.

13 years of Old J Root - Duane Wessels

Almost 20 years later (set up in 1997) is still seeing J Root query traffic being directed to the old IPv4 address (198.41.0.10). They are wondering if they can turn off this resolver at this IP address. Without any current root zone hints file data pointing to this address, there is still a query pattern to this address. There is this theory that the DNS is in fact an infinite memory machine and any information that enters the DNS, be it a query or the address of a root server is in fact never forgotten! This presentation appears to be another piece of evidence to substantiate this view.

(One cute clue pointer for a tool: fpdns is a fingerprint tool to get the resolver version)

Mapping Clients to Resolvers - Matt Larson

Matt used javascript embedded in a small number of web sites, and the result is data that is heavily skewed to the US.

Clients per recursive resolver is an interesting view where he plotted number of clients vs % of seen resolvers as a CDF.

EDNS Client Subnet - Brian Hartvigsen

This is OpenDNS' view of END0 Client subnet signalling - they continue to use a manually maintained whitelist. The talk was about experience gained as an open recursive resolver when they switched on EDNS0 client subnet.

DNSTAP-whoami - Robert Edmonds

This is in the category of "clever DNS tricks", along the lines of whoami.akamai.net, porttest.dns-oarc.net, and rs.dns-oarc.net. This one has a response that pulls in the resolver address, the TCP/UDP query port, the EDNS buffer and the EDNS0 options and packages all this into a response. However the response is encoded, so there is an associated response decoder.

e.g. `$ dig +short @8.8.8.8 whoami.dnstap.info NULL | dnstap-ldns -xy`

Software:

- Reference decoding tool – https://github.com/dnstap/dnstap-ldns Ā
- Custom nameserver – https://github.com/dnstap/dnstap-evldns
- Protobuf schema – https://github.com/dnstap/dnstap.pb
- Ray Bellis, for his “evldns” DNS server framework – https://github.com/raybellis/evldns

Happy DNS Eyeballs - Geoff Huston

A talk on what we observe when a NS has both V4 and V6 connectivity: i.e. what do resolvers do when given the choice of querying an authoritative name server over V4 or V6? The presentation reported on a recent experiment undertaken to observe resolver behaviour and noted that in general resolvers appear to either not use V6 at all, or only over V6 when there is no V4 (i.e. Google) or appears to favour V4 over V6. The report also indicated that Bind had around 55% market share of recursive resolvers.

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

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