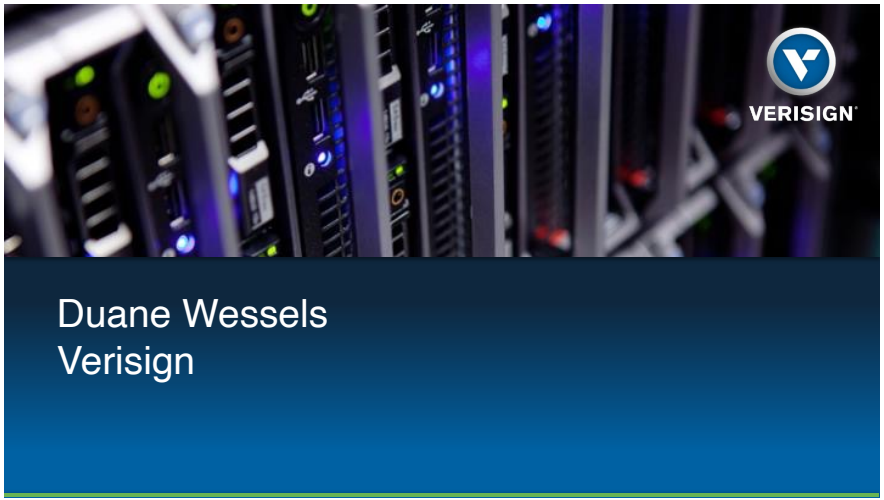
Real world challenges with large responses, truncation, and TCP

Ralf Weber
Akamai

8.2.2023



Duane Wessels
Verisign

DNS-OARC 42
Charlotte, NC
February 8, 2024

How it all began

- A year ago at DNS-OARC 40 in Atlanta
  - I gave a talk: *How Ready is the global DNS for IPv6?*
  - Main problem was IPv6 glue missing at the parent
  - A lot of domains experiencing it were from CDN

Here’s a slide from the talk

Top 10 IPv4 only domains

akadns.net.

Ooops (it can happen)

trafficmanager.net.

g.aaplimg.com.

fastly.net.

bytefcdn-oversea.com.

ovscdns.net.

wsdvs.com.

v.aaplimg.com.

ms-acdc.office.com.

ha.office365.com.

```
; <<>> DiG 9.16.1-Ubuntu <<>> akadns.net @d.gtld-servers.net
;; global options: +cmd
;; Got answer:
;; ->>HEADER<<- opcode: QUERY, status: NOERROR, id: 4033
;; flags: qr rd; QUERY: 1, ANSWER: 0, AUTHORITY: 10, ADDITIONAL: 6
;; WARNING: recursion requested but not available

;; OPT PSEUDOSECTION:
; EDNS: version: 0, flags:; udp: 4096
;; QUESTION SECTION:
;akadns.net.            IN  A

;; AUTHORITY SECTION:
akadns.net.     172800  IN  NS  a3-129.akadns.net.
akadns.net.     172800  IN  NS  a7-131.akadns.net.
akadns.net.     172800  IN  NS  a11-129.akadns.net.
akadns.net.     172800  IN  NS  a1-128.akadns.net.
akadns.net.     172800  IN  NS  a9-128.akadns.net.
akadns.net.     172800  IN  NS  a5-130.akagtm.org.
akadns.net.     172800  IN  NS  a28-129.akagtm.org.
akadns.net.     172800  IN  NS  a13-130.akagtm.org.
akadns.net.     172800  IN  NS  a18-128.akagtm.org.
akadns.net.     172800  IN  NS  a12-131.akagtm.org.

;; ADDITIONAL SECTION:
a3-129.akadns.net.  172800  IN  A   96.7.49.129
a7-131.akadns.net.  172800  IN  A   23.61.199.131
a11-129.akadns.net. 172800  IN  A   84.53.139.129
a1-128.akadns.net.  172800  IN  A   193.108.88.128
a9-128.akadns.net.  172800  IN  A   184.85.248.128

;; Query time: 0 msec
;; SERVER: 2001:500:856e::30#53(2001:500:856e::30)
;; WHEN: Mon Feb 13 04:11:51 UTC 2023
;; MSG SIZE  rcvd: 344
```

Let’s make this better!

Add IPv6 glues…

All looks good now…

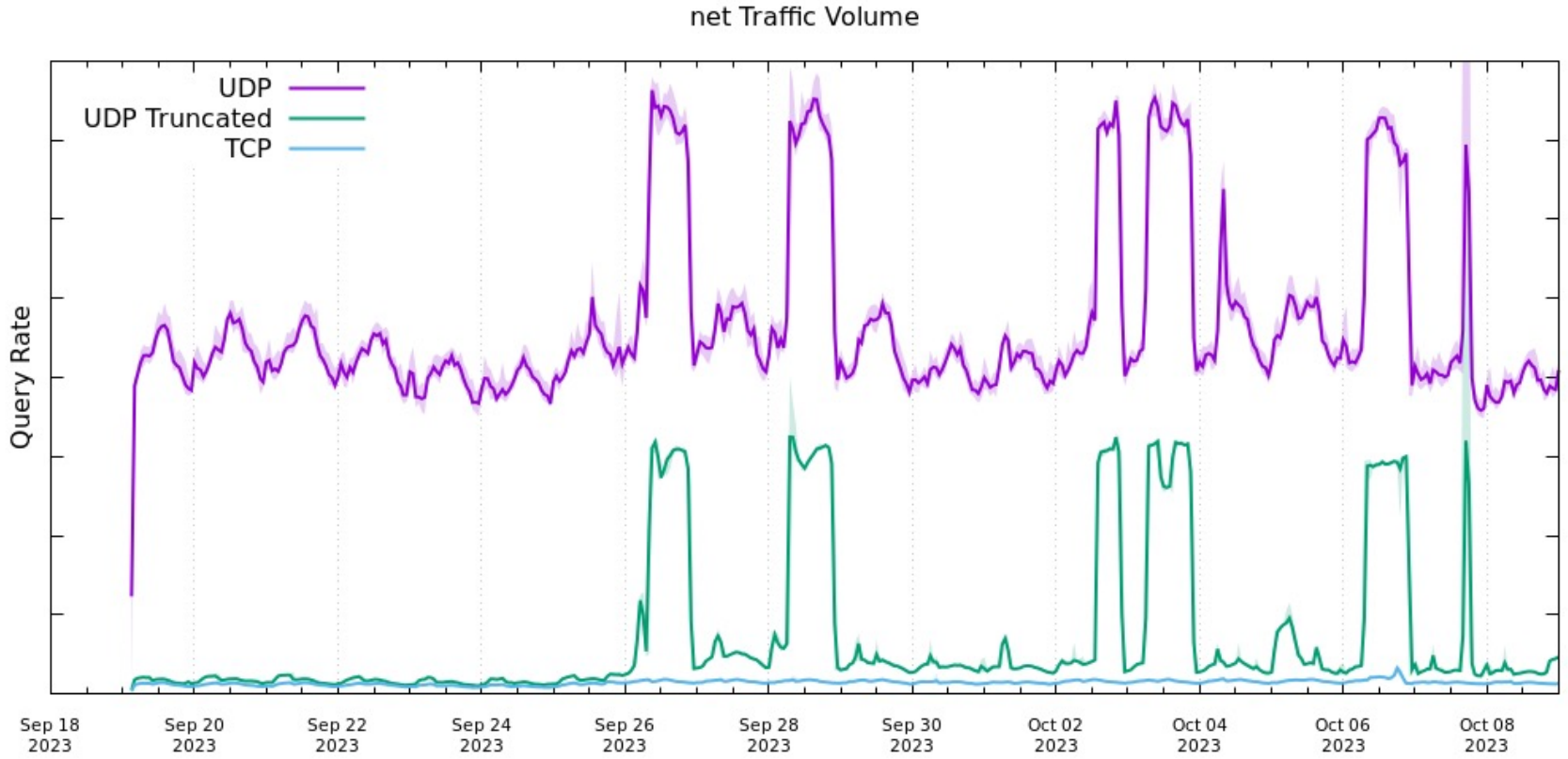
or?

```
;; AUTHORITY SECTION:
aka-ns.net.  172800 IN NS g-n2-a2.aka-ns.net.
aka-ns.net.  172800 IN NS g-n2-a5.aka-ns.net.
aka-ns.net.  172800 IN NS g-n2-a6.aka-ns.net.
aka-ns.net.  172800 IN NS g-n2-a13.aka-ns.net.
aka-ns.net.  172800 IN NS g-n2-a18.aka-ns.net.
aka-ns.net.  172800 IN NS g-n2-a22.aka-ns.net.
aka-ns.net.  172800 IN NS g-n2-a28.aka-ns.net.
aka-ns.net.  172800 IN NS g-n2-a10.aka-ns.net.
aka-ns.net.  172800 IN NS g-n2-a9.aka-ns.net.
aka-ns.net.  172800 IN NS g-n2-a14.aka-ns.net.
aka-ns.net.  172800 IN NS g-n2-a24.aka-ns.net.
aka-ns.net.  172800 IN NS g-n2-a26.aka-ns.net.
aka-ns.net.  172800 IN NS g-n2-a1.aka-ns.net.
A1RT98BS5QGC9NFI51S9HCI47ULJG6JH.net. 86400 IN NSEC3 1 1 0 -
A1RTLNPGULOGN7B9A62SHJE1U3TTP8DR NS SOA RRSIG DNSKEY NSEC3PARAM
A1RT98BS5QGC9NFI51S9HCI47ULJG6JH.net. 86400 IN RRSIG NSEC3 8 2 86400
20231013065811 20231006054811 39455 net.
[…]

;; ADDITIONAL SECTION:
g-n2-a2.aka-ns.net. 172800 IN AAAA 2600:1480:7000::80
g-n2-a2.aka-ns.net. 172800 IN A 95.100.174.128
g-n2-a5.aka-ns.net. 172800 IN AAAA 2600:1480:b000::80
g-n2-a5.aka-ns.net. 172800 IN A 95.100.168.128
g-n2-a6.aka-ns.net. 172800 IN A 23.211.133.128
g-n2-a6.aka-ns.net. 172800 IN AAAA 2600:1401:1::80
g-n2-a13.aka-ns.net. 172800 IN A 2.22.230.128
g-n2-a13.aka-ns.net. 172800 IN AAAA 2600:1480:800::80
g-n2-a18.aka-ns.net. 172800 IN AAAA 2600:1480:4800::80
g-n2-a18.aka-ns.net. 172800 IN A 95.101.36.128
g-n2-a22.aka-ns.net. 172800 IN A 23.211.61.128
g-n2-a22.aka-ns.net. 172800 IN AAAA 2600:1480:7800::80
g-n2-a28.aka-ns.net. 172800 IN AAAA 2600:1480:d800::80
[…]

;; Query time: 0 msec
;; SERVER: 2001:503:a83e::2:30#53(2001:503:a83e::2:30)
;; WHEN: Fri Oct 06 11:03:47 UTC 2023
;; MSG SIZE  rcvd: 1454
```

Anomalous Traffic Events



What we learned...

- Spikes began within 24 hours of Akamai delegation changes
- Spike source IPs are resolvers for a European ISP
- Verisign receives TCP SYN packets from these sources, but not the final SYN+ACK
- Outside spike events, Verisign observes occasional successful TCP transactions from these sources
- Other resolver IPs also began exhibiting UDP truncation retry behavior, but to a lesser extent

European ISP

- Verisign attempted outreach through email and telephone
- Akamai reached out via customer channels
- Eventually learned that the ISP uses Linux iptables with connection tracking, which sometimes become full
- TCP SYN were permitted outbound, but apparently rejected returning SYN+ACK
- When state tracking is full, resolver retries aggressively over UDP
- Unknown what triggers or resolves the condition

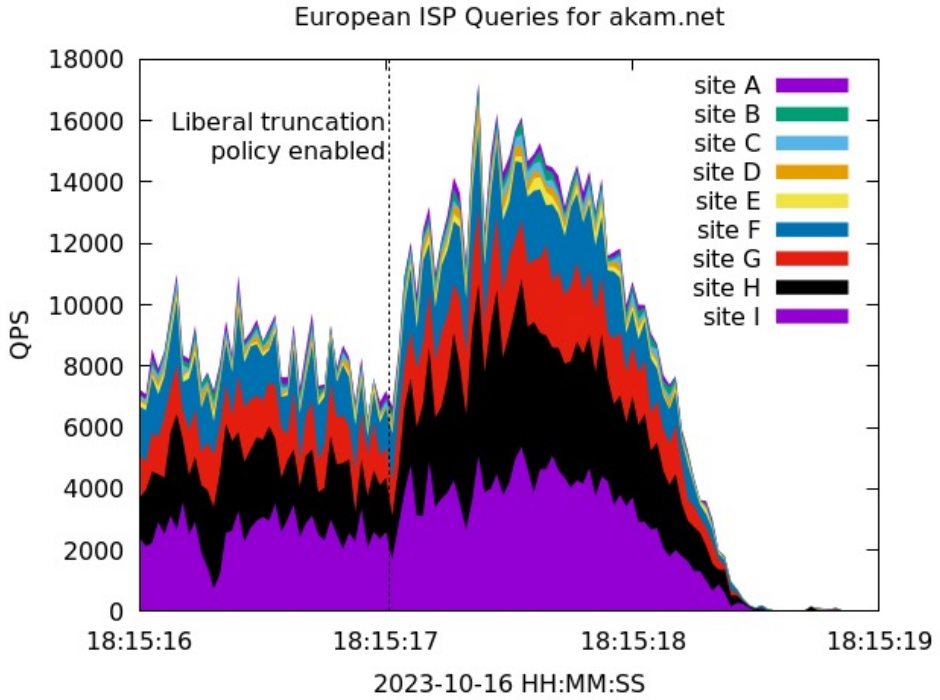
How We Got Here

- European ISP: unable to reliably use DNS-over-TCP
- Verisign: Strict truncation policy
  - RFC 9471: DNS Glue Requirements in Referral Responses
- Akamai: Large delegation responses

- Any one party could “solve” this particular problem

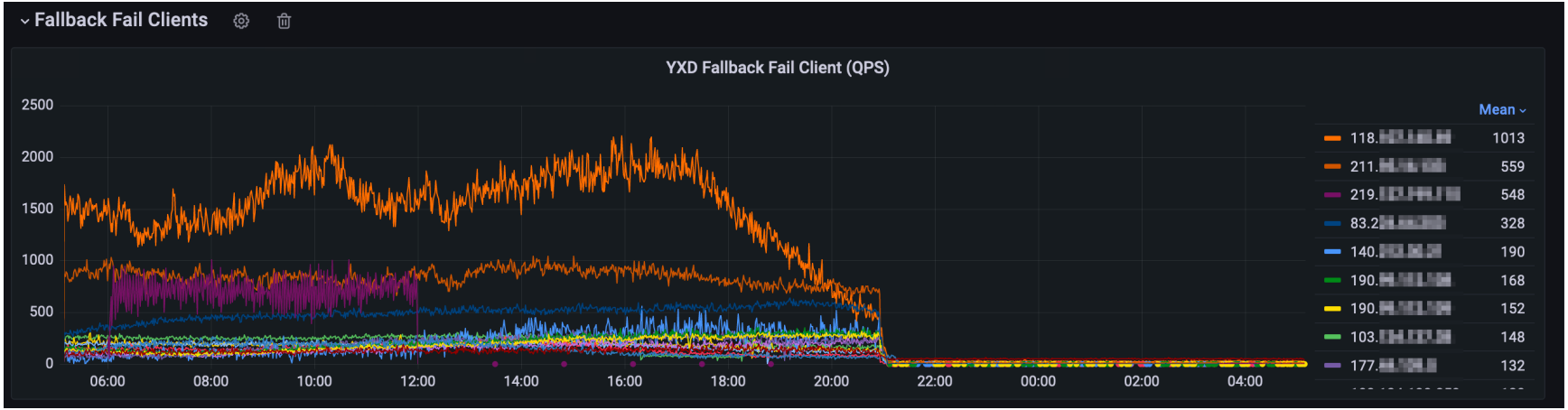
Glue Truncation Policy Experiment

- For the NET algorithm rollover, numerous aggressively retrying ISPs could become a problem
- Verisign tested a less strict glue truncation policy on a single site during a spike event
- Aggressive query traffic dropped to zero across all Verisign sites within two seconds of making the change



Effect of Reducing Delegation Response Size

- On Oct 24, 2023 Akamai removed some delegation name server / glue records
- Immediate effect on resolvers with high UDP truncation



What's Causing the Increases? A small experiment

Understand the impact on Akamai domains

- Querying Akamai or customer domains on our CDN
- Cold *and* semi hot cache scenarios (after our target TTL of 20s expired)

Method:

- Query www.akamai.com (Cold Cache)
  - Wait 60s
- Query www.apple.com (Cold Cache, but CDN mostly hot)
  - Wait 60s
- Query www.akamai.com (Hot Cache, except CDN target)
  - Wait 60s
- Query www.apple.com (Hot Cache, except CDN target)
- Use a "normal" resolver and a resolver that can't switch to TCP

To TCP or not to TCP

TCP incapable resolver

TCP capable resolver

- www.akamai.com
  - 106 packets
  - 4 TCP sessions (44 packets)
- www.apple.com
  - 22 packets
- www.akamai.com
  - 12 packets
- www.apple.com
  - 8 packets

To TCP or not to TCP

TCP capable resolver

- www.akamai.com
  - 106 packets
  - 4 TCP sessions (44 packets)
- www.apple.com
  - 22 packets
- www.akamai.com
  - 12 packets
- www.apple.com
  - 8 packets

TCP incapable resolver

- www.akamai.com
  - 398 udp packets
  - 192 tcp attempts
- www.apple.com
  - 70 udp packets
  - 32 tcp attempts
- www.akamai.com
  - 384 udp packets
  - 192 tcp attempts
- www.apple.com
  - 64 udp packets
  - 32 tcp attempts

Summary of Experiment Results

- Nearly 10x traffic for 4 queries when TCP not available (148 / 1364)
- For the TLD it's even worse as in the hot cache case
  - With working TCP they get no traffic
  - Without working TCP every query SERVFAILS
  - And results in all packets for the resolution go to the TLD
    - UDP and TCP attempts for all name servers
  - Negative (SERVFAIL) caching helps, but only is a couple of seconds
    - A lot different domains are on our CDN and will issue new queries
- Conclusion TCP has to work for a resolver
  - Don’t try stateful anything for DNS

Concluding Thoughts

- Should RFC 9471 have exceptions to strict truncation?
  - So that authoritative servers can protect themselves from aggressive / broken resolvers?
- Perhaps truncation policy change triggered by DNS Response Rate Limiting (RRL)?

Questions?


Akamai


VERISIGN®
© 2024 VeriSign, Inc. All rights reserved. VERISIGN and other trademarks, service marks, and designs are registered or unregistered trademarks of VeriSign, Inc. and its subsidiaries in the United States and in foreign countries. All other trademarks are property of their respective owners.