Fragging DNS

Vixie
2020-02

draft-fujiwara-dnsop-avoid-fragmentation

- Comments and concerned from an authority server operator
- “For IPv6, we support DNS queries over TCP and UDP; for IPv4, just UDP
  - Attacks on the DNS infrastructure via IPv4 are frequent. Mitigation services requires limiting the attack surface
  - TCP state on the server is expensive. Millions of QPS using UDP, even moving a fraction of that load to TCP would be disastrous. UDP frag has no sender state.
  - Our IPv4 ranges are anycasted over multiple networks and datacenter locations.
    - no guarantee that establishing a TCP connection is even possible – flows often incoherent.
  - DNSSEC works perfectly with the standard EDNS0 size setting of 4096.
  - Your proposal breaks this working mechanism. We noticed this as a large access provider followed your suggestion, and reduced the EDNS0 bufsize. They have disabled that change for now to unbreak their DNS resolvers.”

Tyranny of IEEE 802.3 MTU

- *bps* ethernet defined min 64 bytes, max 1500 bytes, IRG, trailers
  - Same for , , , , ,
  - Max PPS has grown to , , , , ,
  - *bps* scales with density
  - PPS scales with power, heat, complexity
- Eventually the MTU will have to increase
  - Should have been before now
  - May be ~ in our lifetimes
- We should not be hardwiring payload sizes
  - Look at the route MTU (LAN, MAN, WAN)

512 vs. 1232

- TCP knows payload max size when assembling a segment
    - IP + options, TCP + options, and far-end MSS, are all known
    - Next data in the stream, and PSH signaling, are all known
- UDP senders do not know payload max size
    - DNS client or server must estimate before sending
- Minimum MTU for IPv4 is 68, for IPv6 is 1280
    - Minimum reassembly capacity for IPv4 is 576, for IPv6 is NaN
    - 68 = 60 + 8; 576 = 68 + 512
    - 1280 – 40 – 8 = 1232
        - “and smaller still if additional extension headers are used.” (RFC 2460)

1232 vs. 1400

- UDP with 1232-octet payload only works by accident
  - IPv6 extensions aren’t uncommon
    - “Each extension header is an integer multiple of 8 octets long”
  - If MTU 1280 was happening, UDP would often fail
  - Because MTU is usually ~1500 or larger, UDP on IPv6 usually works
- 1232 assumed that PMTUD would be done, which doesn’t work
  - We should allow DNS UDP senders to check the route’s MTU
    - LAN route might be ~9K, default route will likely be 1500
- We should admit that actual Internet MTU is ~1500
  - And we should leave some slop for tunnel and extension overhead
  - Therefore ~1400 would be a perfectly safe DNS UDP max payload size