draft-ietf-dnsop-avoid-fragmentation-05

K. Fujiwara, P. Vixie
dnsop WG Interim meeting
2021/9/15
Summary of draft-ietf-dnsop-avoid-fragmentation-05

• Introduction: Fragmented DNS UDP responses have systemic weaknesses

• Proposal
  • Recommendations for UDP responders
    • SHOULD send DNS responses with IP*_DONTFRAG options
    • MAY probe to discover the real MTU value per destination.
    • SHOULD compose UDP responses fit in path MTU (or good value)
  • Recommendations for UDP requestors
    • SHOULD send DNS responses with IP*_DONTFRAG options
    • SHOULD use the requestor's payload size as calculated or good value
  • good values: 1220, 1232, 1400, 1472/1452, or measured

• Additional texts (Minimal responses, IP_MTU getsockopt, tracepath)
<table>
<thead>
<tr>
<th>Source</th>
<th>IPv4</th>
<th>IPv6</th>
</tr>
</thead>
<tbody>
<tr>
<td>RFC 4035 (MUST)</td>
<td>1220</td>
<td>1220</td>
</tr>
<tr>
<td>Software developers / DNSFlagDay2020 propose</td>
<td>1232</td>
<td>1232 (1280-40-8)</td>
</tr>
<tr>
<td>Authors' recommendation</td>
<td>1400</td>
<td>1400 (1500-40-8-some headers)</td>
</tr>
<tr>
<td>Maximum: Ethernet MTU 1500 [Huston2021]</td>
<td>1472 (1500-20-8)</td>
<td>1452 (1500-40-8)</td>
</tr>
<tr>
<td>Measured</td>
<td>MTU-20-8</td>
<td>MTU-40-8</td>
</tr>
</tbody>
</table>
DNS over TCP Considered Vulnerable

• Haya Shulman et al. published a new paper: “DNS over TCP Considered Vulnerable” at ANRW 2021 (July 28, 2021)
  • ICMP attack targets are intermediate routers between resolvers and authoritative servers.
  • They show that some routers accept ICMPv4 "Fragmentation Needed and DF set" to resolvers.
  • 496 of Alexa top-100K domains are vulnerable to fragmentation over TCP.

• How to measure?
  • At IETF 111, there is a comment: 0.5% is small and they are bugs.
  • At ANRW, I cannot get clear answer about IPv6 and who is vulnerable.
  • Fragmentation does not happen on IPv6 at intermediate routers.
  • Recent TCP implementations support RFC 4821 “Packetization Layer Path MTU Discovery” and set IP_DF (Don’t Fragment) bit on IPv4 TCP packets.
    • The DF bit SHOULD be set on the (TCP) fragments (Quoted from Section 8 of RFC 4821).
    → Add texts and reference to the paper at “Weaknesses of IP fragmentation”
Discussions at IETF 111

• Paul Hoffman mentioned he expected a single value in a BCP document, while Viktor Dukhovni is fine with a set of values.
Questions

1. Can we agree with a set of "good" UDP sizes, rather than a single value?
2. What are the good values?
   • 1220, 1232, 1400, 1472/1452?
3. Is it possible to probe good values per destination at UDP requestor?
   • PLPMTUD (RFC 8899) or BIND 9’s way

• Our concern is that when leaf sites are under tunnels and their MTU are small, standardizing a large value (with IP_DF) will prevent communications.
  • Some VPN appliances offer default MTU 1280
  • Leaf site case, software MAY probe MTU size to the Internet and generate good value
Probing good values

• If complexity (PMTU discovery) and insecurity (TCP vulnerability) are to be avoided above all else, then a small EDNS buffer size should be offered. (For example, 1220 or 1232)

• If network efficiency both now in the future is to be maximized, then adaptive retry after silent failure should be done, beginning with a large value and trying smaller values, similar to PLPMTUD (RFC 8899).

• In all cases, fragmentation either by an endpoint or gateway must be avoided; in a definite future something like PLPMTUD and its attendant complexity and state costs will be necessary to take advantage of vastly larger path MTUs of the future.