DNSOP Working Group
Internet-Draft
Intended status: Standards Track
Expires: May 30, 2015

L. Song
Beijing Internet Institute
D. Ma
ZDNS

November 26, 2014

Using TCP by Default in Root Priming Exchange
draft-song-dnsop-tcp-primingexchange-00

Abstract

DNS payload size limitation constrains operational and policy choice for many years. It is become increasingly notable in IPv6 transition and DNSSEC development, specifically during Priming Exchange process. Given that the load of priming exchange on the root system is relatively low, this memo proposes using TCP by default in Priming Exchange between resolver and root server. This approach is aiming to treat Priming Exchange as a special process which brings a little change to the current root system and provide solution to the long-term problems.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on May 30, 2015.

Copyright Notice

Copyright (c) 2014 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (http://trustee.ietf.org/license-info) in effect on the date of publication of this document. Please review these documents

carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction

The Domain Name System is a typical UDP-based protocol which is connectionless and following single-packet exchange paradigm. One exchange of DNS, which is important but without fully documented, is the Priming Exchange between resolver and root server[Priming-Test]. In simple terms, priming query is a NS query for root zone (qtype=NS, qname='.', qclass=IN). The traffic caused by priming exchange is also called 'ns-for-dot' traffic.

Due to the maximum DNS message size specified in [RFC1035], all the DNS packets including Priming Exchange should fit within 512 octets. It becomes an historical and practical hard DNS protocol limit, even after EDNS0 [RFC6891] was introduced to mitigate this problem. It bring constraints operational and policy choice for many years, and becomes increasingly notable in IPv6 transition and DNSSEC development, specifically for the case of Priming Exchange.

Given that the load of Priming Exchange on the root system is relatively low [TLD-Statues-K] [QTYPE-L], this memo proposes using TCP by default in Priming Exchange between resolver and root server. This approach is aiming to treat Priming Exchange as a special process which brings a little change to the current root system and provide solution to the long-term problems.

Note that this memo is neither protocol novelty nor specification of DNS over TCP which are discussed by [I-D.dickinson-dnsop-5966-bis] [T-DNS]. This memo is to analyze and argue the possibility of using TCP as a default transmission protocol with a small price to pay for it on the occasion.

2. Operational and policy constraints with Priming Exchange

2.1. Full AAAA record with Priming Exchange

To the authors' knowledge, the concept "priming" was first defined in ICANN SSAC/RSSAC serial work [SAC016], [SAC017] and [SAC018]when people consider adding IPv6 AAAA Records to Priming Exchange. For performance and resiliency purpose, there are two major requirements: 1) include all A records for all thirteen root servers, 2) avoid the burden of TCP. So the final operational decision is to return all A/AAAA record in additional section of priming reply, but with special sequencing of records in which type A records precede type AAAA records.

Based on that decision, the AAAA records will be incomplete in small packets (<512B) given that truncation of the additional data section might not be signaled for additional section data (Appendix B of [RFC4472]). Still to date, no more than two AAAA resource records can be included in the Priming response in the absence of EDNS0 support. This operational constraint compromise the resiliency of Root system for IPv6 users, especially under the consideration of IPv6-only deployment [I-D.song-sunset4-ipv6only-dns].

2.2. DNSSEC with Priming Exchange

The data in the additional section of a DNS response (also called referral response) is viewed as "courtesy" and optional. So it is not required to be signed in Priming response and validated when a resolver receive it. In the case of Priming Exchange, resolver only validates the NS RRSet but without validating the corresponding A and AAAA RRSets.

For the performance consideration, most of time the resolver will not issue A or AAAA query for root servers. So there is a risk that A/AAAA RRSet of root server is polluted with another set of data or

invalid data. It is not reasonable. This issue is also addressed in [I-D.ietf-dnsop-resolver-priming] which gives up the possibility of DNSSEC with Priming Exchange since the requirement of DNS payload size even surpasses the largest feasible EDNS0 buffer size.

2.3. The number of NS server with Priming Exchange

Still due to the DNS payload limitation, the number of NS server of root zone is strictly limited. The number 13 is never changed since 1997 when the last Root letter "M" was assigned. This fixed number of root NS servers introduces a topic constantly incurring arguments and misunderstanding (Maybe the rarity means the value, who knows). Some discussion and proposal [I-D.lee-dnsop-scalingroot] try to extended 13 to millions for pervasive distribution of Root zone.

From the technical point of view, the number of NS server should be treated as a parameter of the root system, rather than a constant making the root server as some kind of rare resource of Internet which only incurs constant disputes with Internet governance issue. With TCP enhancement in Priming Exchange, the root zone operator is able to decide and control the value of this parameter based on practical and technical requirements. It is also promising that the non-technical disputes with Internet governance will cease from this very topic.

<<subsection 2.3 may be a little controversial and out of the scope of IETF discussion. But it fit the topic of policy constraint brought out by technical issues. This text may be amended largely according to the feedback of the community. >>

3. Review of DNS over TCP

The Domain Name System is a typical UDP-based protocol which is connectionless. TCP is also designed in the original DNS specification [RFC1035] but it is only used in zone transfer and act as a backup when DNS packet gets truncated with large payload [RFC1123]. In addition, DNS over TCP is commonly viewed as expensive and not optimal compared to connectionless UDP. This "community consensus" and operational practice are held since early days of DNS.

Today the DNS system based on UDP has run into troubles in privacy, security and constraints both on policy and operation choices. The limitation of pay load size for example, is one of most notable issues. Besides the long-tail adoption, any mechanism supporting larger UDP packets such as EDNS0 will cause unexpected risk of IP-fragmentation [Fragment-Problem][Fragment-Poisonous]. That issue becomes increasingly notable concerning DNSSECwith key rollover process which involves much larger response packets.

As to the very issue, some discussions focus on the problem of 512B limitation and the firewall/middle-box misbehavior. Some propose the way to work around [I-D.ietf-dnsop-respsize]. This memo benefits mainly from the discussion of DNS over TCP [I-D.dickinson-dnsop-5966-bis][I-D.hzhwm-dprive-start-tls-for-dns][Question-DNS], especially from the extensive work of [T-DNS]. In short, they all identify that TCP is an extant part of current DNS implementation and has its inherent property overcoming the issue brought about by DNS over UDP.

The author of this memo is the advocate for DNS over TCP, especially on the very occasions when it is needed. In this memo, the priming exchange is identified as one of these occasions.

4. Requirement of priming exchange over TCP

The priming exchange and TCP are introduced respectively in the previous sections. In this section some requirements and considerations are proposed on how to switch on priming exchange over TCP on both sides of resolver and root server. Note that all the following discussion is based on the assumption that the root zone operator (ICANN/IANA) decides to adopt TCP in its priming exchange in order to break through the constraint explained in section 2

The requirements for recursive server are:

1)Be aware that more information with larger payload will be introduced into the priming exchange and make sure the TCP function of resolver is ok when a resolver issues "ns-for-dot" query to root server.

2)Make a configuration or a necessary patch to the recursive server in order to send a priming query on TCP by default. (The little change does not affect other UDP Qtype like A/AAAA/SOA)

3)Similar to the suggestion of [I-D.ietf-dnsop-resolver-priming], do not send and re-send priming query over TCP to root server every often. Considering the TTL of the NS records is currently six days (518400 seconds), 3~4 day is a proper setting for the interval of re-priming.

4)Without actions above, be aware of the large possibility of risk due to truncated UDP packets or fragmentation.

The requirements for root server:

1)Be aware any changes may be introduced to Root zone and priming exchange, if the root zone operator (ICANN/IANA) decides to adopt priming exchange over TCP proposal.

2)Given the rise of TCP connections from priming exchange, proper setting on maximum concurrent TCP connections should be carefully considered. This setting will affect the response time of priming exchange over TCP [RIPE-TCP-Test]. Now common setting of maximum concurrent TCP connections is 10-100.

<<This part is the initial consideration and loosely composed on the requirement of how to switch on priming exchange over TCP on both resolver and root server. It may amended largely after receiving comments from the community.>>

5. Performance analysis

There are mainly two performance concerns when TCP is introduced by default in priming exchange: response time and load of root system.

5.1. Response time

There was a test in 2012 by RIPE NCC comparing TCP and UDP Response times of DNS root servers [RIPE-TCP-Test], which demonstrated that the response time of TCP query is twice or three times of UDP response time. Although the measurements was done on the round-trip (RT) values of DNS queries for SOA, it is valid for estimate of priming response time.

There are also some discussions and proposal of DNS over TCP with Fast Open technology [I-D.ietf-tcpm-fastopen] which will largely reduce end-to-end TCP latency [T-DNS].

Because the frequency of priming exchange is not much high, the resolver has plenty of time to re-priming exchange with Root in any case whenever it boots up or its TTL timeout. So it is convincible for the author of this memo that twice or three times latency is endurable in the very case of priming exchange.

5.2. Load of system

According to the monitoring system and tools developed by root server operators [TLD-Statues-K] [QTYPE-L], the information of priming load of Root server is public in a real-time manner. At the time of this memo drafted, the average load of "ns-for-dot" query is less than 1k qps (out of 30k qps in total) in K-root which is composed of 17 instances which are sharing the load. The "ns-for-dot" query load in the case of L-root is 1.4k qps (out of 50k qps in total) with 152 instances sharing the load. It seems that the load of priming exchange traffic is not much high in this regard.

Note that it's necessary and important for the root zone operator to conduct a carefully designed test of the impact of introduction of TCP into priming exchange in order to detect any latent anomalies.

<<To be extended, comments are welcome!>>

6. Pros and cons

6.1. Pros

Because TCP is a stream protocol in its nature, there is no longer any payload size limitation for DNS transmission. This property is beneficial in following aspects:

1)Adding full A/AAAA address of Root server in priming response, which will enhance the resiliency of Root system for IPv6 users, especially under the consideration of IPv6-only deployment [I-D.song-sunset4-ipv6only-dns].

2)DNSSEC validation for priming exchange. The priming exchange plays a key role guiding the traffic to the unique identity system. There is little evidence or argument that priming exchange does not need integrity production, given that DNS becomes increasingly the attack vector of the Internet.

3)Making DNS Priming Exchange scale to more extensions and applications, especially providing a possibility that a larger range of root zone hosts could be employed in a controlled manager following the current hierarchical architecture, which offers pervasive distribution of root service.

6.2. Cons

No change is cost free, even with a little changes in the current system. It should be prudent and fully consider the impact of using TCP by default in priming exchange.

1)Bring unexpected degradation in the performance of root services, latency, system load or other aspects which are needed tests in a large scale. This may involve a lot of discussion and argument.

2)This memo requires recursive name servers be aware of Priming Exchange is a special process and the relating mechanism should hence be triggered somehow as for software implementations.

3)There is a concern about the capability of TCP for some resolver. Evidence [Question-DNS]show that a significant set of resolver cannot operate correctly over TCP even when TCP is required.

4)If there is no enough incentive or push, it will take time for resolver to adopt the changes due to the common inertia. It may incur both time and labor expense.

<<To be extended, comments are welcome!>>

7. Security Considerations

TBD

8. IANA Considerations

TBD

9. Acknowledgements

Special tanks to Professor John Heidemann and his team at USC/ISI. Their work on T-DNS is the major source of inspiration for this memo.

10. References

10.1. Normative References

[RFC1035] Mockapetris, P., "Domain names - implementation and specification", STD 13, RFC 1035, November 1987.

[RFC1123] Braden, R., "Requirements for Internet Hosts - Application and Support", STD 3, RFC 1123, October 1989.

[RFC4472] Durand, A., Ihren, J., and P. Savola, "Operational Considerations and Issues with IPv6 DNS", RFC 4472, April 2006.

[RFC6891] Damas, J., Graff, M., and P. Vixie, "Extension Mechanisms for DNS (EDNS(0))", STD 75, RFC 6891, April 2013.

10.2. Informative References

[Fragment-Poisonous]
Herzberg, A. and H. Shulman, "Fragmentation Considered Poisonous", 2012.

[Fragment-Problem]
Kent, C. and J. Mogul, "Fragmentation considered harmful", 1987.

[I-D.dickinson-dnsop-5966-bis]
Dickinson, J., Bellis, R., Mankin, A., and D. Wessels, "DNS Transport over TCP - Implementation Requirements", draft-dickinson-dnsop-5966-bis-00 (work in progress), October 2014.

[I-D.hzhwm-dprive-start-tls-for-dns]
Zi, Z., Zhu, L., Heidemann, J., Mankin, A., and D. Wessels, "TLS for DNS: Initiation and Performance Considerations", draft-hzhwm-dprive-start-tls-for-dns-00 (work in progress), October 2014.

[I-D.ietf-dnsop-resolver-priming]
Koch, P. and M. Larson, "Initializing a DNS Resolver with Priming Queries", draft-ietf-dnsop-resolver-priming-04 (work in progress), February 2014.

[I-D.ietf-dnsop-respsize]
Vixie, P., Kato, A., and J. Abley, "DNS Referral Response Size Issues", draft-ietf-dnsop-respsize-15 (work in progress), February 2014.

[I-D.ietf-tcpm-fastopen]
Cheng, Y., Chu, J., Radhakrishnan, S., and A. Jain, "TCP Fast Open", draft-ietf-tcpm-fastopen-10 (work in progress), September 2014.

[I-D.lee-dnsop-scalingroot]
Lee, X., Vixie, P., and Z. Yan, "How to scale the DNS root system?", draft-lee-dnsop-scalingroot-00 (work in progress), July 2014.

[I-D.song-sunset4-ipv6only-dns]
Song, D., Vixie, P., and D. Ma, "Considerations on IPv6-only DNS Development", draft-song-sunset4-ipv6only-dns-00 (work in progress), October 2014.

[Priming-Test]
RIPE NCC, "A Look at DNS Priming Queries to K-root", 2007, <https://labs.ripe.net/Members/emileaben/content-look-dns-priming-queries-k-root>.

[QTYPE-L] "The "NS for Dot" query for L-root (DNS queries by QTYPE)", 2007, <http://hedgehog.dns.icann.org/hedgehog/hedgehog.html>.

[Question-DNS]
Huston, G., "a question of DNS Protocols", September 2013, <http://www.potaroo.net/ispcol/2013-09/dnstcp.html>.

[RIPE-TCP-Test]
"Comparing TCP and UDP Response Times of DNS Root Servers", <https://labs.ripe.net/Members/bwijnen/tcp-udp-dns-soa-rt-ratio>.

[SAC016] ICANN Security and Stability Advisory Committee, "Testing Firewalls for IPv6 and EDNS0 Support", 2007.

[SAC017] ICANN Security and Stability Advisory Committee, "Testing Recursive Name Servers for IPv6 and EDNS0 Support", 2007.

[SAC018] ICANN Security and Stability Advisory Committee, "Accommodating IP Version 6 Address Resource Records for the Root of the Domain Name System", 2007.

[T-DNS] Zhu, L., Hu, Z., and J. Heidemann, "T-DNS: Connection-Oriented DNS to Improve Privacy and Security (extended)", 2007, <http://www.isi.edu/~johnh/PAPERS/Zhu14b.pdf>.

[TLD-Statues-K]
"TLD's Status for k-root", 2007, <http://k.root-servers.org/statistics/ROOT/tlds.html>.

Authors' Addresses

Linjian Song
Beijing Internet Institute
2508 Room, 25th Floor, Tower A, Time Fortune
Beijing 100028
P. R. China

Email: songlinjian@gmail.com

Di Ma
ZDNS
Beijing
P. R. China

Email: madi@zdns.cn