Background

Root operators implement Domain Name Service (DNS) from a range of soft- and hard- ware platforms. The common element which exposes from these services is DNS protocol packets, carried in either Transmission Control Protocol (TCP) or (User Datagram Protocol) UDP, in turn carried over Internet Protocol (IP) packets in either IP version 4 (IPv4) or IP Version 6 (IPv6), including Internet Control Message Protocol (ICMP and ICMPv6) packets.

It is rare for root servers to have to answer DNS queries with responses greater than 1200 octets at this time. There are three points during the forthcoming Key Signing Key (KSK) roll when the root server response to a DNS Security (DNSSEC) priming query exceeds 1400 octets. Referring to "Scoring the DNS Root Server System" [1] it is observed that current root servers appear to handle this form of large response in varying ways, both between roots and within one "letter" instance.

These variances are not well understood. It is not clear when this is deliberate, or accidental, or incurred from server, host or network level upgrades or configuration.

It is anticipated that DNSSEC KSK key rolls will continue and may increase in frequency, and that DNSSEC protocol changes are some way off, as are cipher changes (which would reduce packet sizes in the normal operational cycles of key change). Therefore it is now part of the consideration of reliability and robustness of the root server set, to understand the way in which they pass back large responses from the root zone. Ideally, this information would be acquired before the September key roll events, when new volumes of large packet responses will be expected.

Where in the protocol stack are packet sizes controlled?

DNS packets themselves contain variable parts, including optional Extended DNS (EDNS) flags to indicate maximum DNS message size. This in turn governs the choices of how DNS packets are constructed to be placed in TCP or UDP packets inside IP packets, modulated through design choices in the server code.

However in all cases the IP packets reflect two fundamental configurations at the server (host) side:

- The interface maximum transfer unit (MTU) which governs the basic IP packet fragmentation behavior
- The Maximum Segment Size (MSS) which is negotiated end-to-end in a TCP connection.

These are modulated through the behaviors of ICMP and ICMPv6, regarding "packet too big" (PTB) messages, and path-MTU discovery (pMTU) which is a function both of the end hosts communicating, and of the path:

- End hosts either side may be emitting ICMP or ICMPv6.
- Packet forwarding decisions along the path may determine if they are received.
- Elements along the path may themselves be responsible for emitting ICMP or ICMPv6.

There are different consequences in IPv4 and IPv6 to PTB events, pMTU and fragmentation outcomes, since in IPv4 fragmentation and re-assembly on-the-fly is permitted for packets in transit, whilst in IPv6 fragmentation and re-assembly is only performed by end-hosts, which makes successful transfer of ICMPv6 and application of pMTU discovery all the more important: since no re-assembly is permitted, efficient negotiation of a viable end-to-end transfer unit is essential to avoid either loss of data, or excessive retransmission.

In-DNS behaviors can be understood by packet capture and measurement of packets sent and received: the information is directly visible from the content of the DNS packets. However on-the-wire state is influenced by configuration decisions which affect these behaviors including the interpretation of EDNS size flags, and also server specified limits of the TCP, UDP and IP behaviors.

IP level behavior can only be partially inferred by what is seen: the exact reason larger or smaller packets are received can only partly be understood since any of the endpoint, or elements along the path may have caused ICMP or ICMPv6 to happen (or caused it to be lost) or have caused pMTU discovery to succeed or fail or operate from defined state at the interface MTU level (MSS negotiation can be understood from analysis of the TCP packet flow however).

Without knowledge of the configured state of each node in each Root server any-cast cloud, it is not possible to make complete inferences about the behavior of the system at large: if one node is measured in any-cast routing, no inference can be drawn about either another node at another any-cast route, or the applicability of what is seen at that one node-path combination. To determine the behavior of the system overall, more information is needed.

Ideally, all nodes (hosts, interfaces on hosts which are multi-homed physically or via virtual LAN) would have an understood MTU and MSS, and thus visible effects relating to large packet/fragmentation could be imputed to the host, or the path as relevant. Additional information about host or server specific tuning would help, since kernel and operating system differences can account of behavior seen: if (for instance) host level packet filtering is dropping excessive ICMP above some threshold, there is a chance that for any given attempt to compute end-to-end path, ICMP PTB has been lost at the host.

Relevance and Motivation

This is immediately of concern because the impending DNSSEC KSK roll-over introduces larger DNS packets, which invite use of large IP packets, which invites the incursion of unwanted pMTU, ICMP and ICMPv6 consequences. In IPv4 this probably has low consequence because of on-the-path fragmentation and re-assembly. in IPv6, this is a higher risk situation which it would be beneficial to understand since it directly influences loss of DNS signal, and indeterminate outcomes depending on the node, path outcome.

Scope

The work party will be asked to:

1. Define the set of IP (and ICMP), UDP/TCP and DNS parameters configuration options and on-the-wire outcomes which should be defined, and specified for any host providing root DNS service either stand-alone or in an any-cast configuration. Questions would include:
   - What is the largest un-fragmented payload using UDPv4 and UDPv6?
   - If a server fragments, what is the largest fragment which is sent?
   - When a server truncates DNS answers in UDP, what is the truncation point?
   - Does the host honor ICMPv6 messages in UDP and TCP?
   - Does the host set the DF bit in IPv4?
   - What is the maximal MSS sent by in your DNS?
   - Are all of a given root server letter's instances identical in terms of these behaviors, or do they vary by instance? Can this be enumerated?
2. Define a registry of this parameter/configuration state which is kept up to date, so that inferences about behaviors and changes in behaviors in the DNS can be understood against host- and path- related events.
3. Consider if a defined requirement should exist to set a common MSS/MTU and to define DNS, UDP, TCP, IP, ICMP and ICMPv6 packet size fragmentation and signalling expectations from root server instances, and their connectivity into the global public Internet (it is understood that path effects beyond the point of connect and origination cannot be constrained).
4. Alternatively if it is better to have deliberate variance of these values, to ensure variance of exposure to risk of loss, Can we define what those variances should be, and how they are understood to be applied in each case.

It expected that the work will be carried out reaching out to a wider DNS community than only the root server community. Configuration recommendations in particular will be relevant to gTLD and ccTLD services both of which exist as large anycast clouds.

Deliverable

The final draft of the "RSSACXXX: DNS, UDP, TCP IP and ICMP parameter configurations", numbered Draft-RSSAC-XXX.

Date of Delivery

Final draft submitted to the RSSAC no later than 2017-09-30. Submission prior to the deadline is useful given that DNSSEC key roll-over will introduce large packets during September.

Guidelines

The RSSAC requests that root operators be polled to supply the information based on a template or pro-forma method, to be collated into a single report, and replicated on the web in a public registry which is maintained.

The RSSAC requests the work party leader to report progress on this work to RSSAC as appropriate. In the event of the deadline will not be realized, the work party leader should inform RSSAC without undue delay and provide details of the work that cannot be completed by the deadline.

RSSAC support staff will assist the working party deliberation of the work, including setting up a mailing list for the work party, arranging and supporting regular teleconference calls, taking notes of meetings, drafting background materials of the work, and serving as editors for documents if needed.

References

[1] http://www.potaroo.net/presentations/2017-05-14-scoring-roots.pdf