dnsop
Internet-Draft
Intended status: Informational
Expires: January 7, 2016

C. Contavalli
W. van der Gaast
Google
D. Lawrence
Akamai Technologies
W. Kumari
Google

July 6, 2015

Client Subnet in DNS Queries
draft-ietf-dnsop-edns-client-subnet-02

Abstract

This draft defines an EDNS0 extension to carry information about the network that originated a DNS query, and the network for which the subsequent response can be cached.

IESG Note

[RFC Editor: Please remove this note prior to publication]

This informational document describes an existing, implemented and deployed system. A subset of the operators using this is at http://www.afasterinternet.com/participants.htm . The authors believe that it is better to document this system (even if not everyone agrees with the concept) than leave it undocumented and proprietary.

Table of Contents

1. Introduction

Many Authoritative Nameservers today return different responses based on the perceived topological location of the user. These servers use the IP address of the incoming query to identify that location. Since most queries come from intermediate Recursive Resolvers, the source address is that of the Recursive Resolver rather than of the query originator.

Traditionally, and probably still in the majority of instances, Recursive Resolvers are reasonably close in the topological sense to the Stub Resolvers or Forwarders that are the source of queries. For these resolvers, using their own IP address is sufficient for authority servers that tailor responses based upon location of the querier.

Increasingly, though, a class of Recursive Resolvers has arisen that handle query sources that are often not topologically close. The motivation for a user to configure such a Centralized Resolver varies but is usually because of some enhanced experience, such as greater cache security or applying policies regarding where users may connect. (Although political censorship usually comes to mind here, the same actions may be used by a parent when setting controls on where a minor may connect.) Similarly, many ISPs and other organizations use a Centralized Resolver infrastructure that can be distant from the clients the resolvers serve. These cases all lead to less than desirable responses from topology-sensitive Authoritative Nameservers.

This draft defines an EDNS0 [RFC6891] option to convey network information that is relevant to the DNS message. It will carry sufficient network information about the originator for the Authoritative Nameserver to tailor responses. It will also provide for the Authoritative Nameserver to indicate the scope of network addresses for which the tailored answer is intended. This EDNS0 option is intended for those recursive and authority servers that would benefit from the extension and not for general purpose

deployment. It is completely optional and can safely be ignored by servers that choose not to implement it or enable it.

This draft also includes guidelines on how to best cache those results and provides recommendations on when this protocol extension should be used.

At least a dozen different client and server implementations had been written based on the original specification, first known as draft-vandergaast-edns-client-subnet. While they interoperate for the primary goal, they have varying behaviour around poorly specified edge cases. Known incompatibilities will be described.

2. Requirements Notation

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

3. Terminology

ECS EDNS Client Subnet.

Client A Stub Resolver, Forwarder or Recursive Resolver. A client to a Recursive Resolver or a Forwarder.

Server A Forwarder, Recursive Resolver or Authoritative Nameserver.

Stub Resolver: A simple DNS protocol implementation on the client side as described in [RFC1034] section 5.3.1. A client to a Recursive Resolver or a Forwarder.

Authoritative Nameserver: A nameserver that has authority over one or more DNS zones. These are normally not contacted by Stub Resolver or end user clients directly but by Recursive Resolvers. Described in [RFC1035] Section 6.

Recursive Resolver: A nameserver that is responsible for resolving domain names for clients by following the domain's delegation chain. Recursive Resolvers frequently use caches to be able to respond to client queries quickly. Described in [RFC1035] Section 7.

Intermediate Nameserver: Any nameserver (possibly a Recursive Resolver) in between the Stub Resolver and the Authoritative Nameserver.

Centralized Resolvers: Recursive Resolvers that serve a topologically diverse network address space.

Tailored Response: A response from a nameserver that is customized for the node that sent the query, often based on performance (i.e. lowest latency, least number of hops, topological distance, ...).

Topologically Close: Refers to two hosts being close in terms of number of hops or time it takes for a packet to travel from one host to the other. The concept of topological distance is only loosely related to the concept of geographical distance: two geographically close hosts can still be very distant from a topological perspective, and two geographically distant hosts can be quite close on the network.

4. Overview

The general idea of this document is to provide an EDNS0 option to allow Recursive Resolvers, if they are willing, to forward details about the origin network from which a query is coming when talking to other Nameservers.

The format of this option is described in Section 5, and is meant to be added in queries sent by Intermediate Nameservers in a way transparent to Stub Resolvers and end users, as described in Section 6.1. ECS is only defined for the Internet (IN) DNS class.

As described in Section 6.2, an Authoritative Nameserver could use this EDNS0 option as a hint to better locate the network of the end user and provide a better answer.

Its response would contain an edns-client-subnet (ECS) option, clearly indicating that the server made use of this information, and that the answer is tied to the network of the client.

As described in Section 6.3, Intermediate Nameservers would use this information to cache the response.

Some Intermediate Nameservers may also have to be able to forward ECS queries they receive. This is described in Section 6.5.

The mechanisms provided by ECS raise various security related concerns related to cache growth, the ability to spoof EDNS0 options, and privacy. Section 10 explores various mitigation techniques.

The expectation, however, is that this option will primarily be used between Recursive Resolvers and Authoritative Nameservers that are sensitive to network location issues. Most Recursive Resolvers,

Authoritative Nameservers and Stub Resolvers will never need to know about this option, and will continue working as they had been.

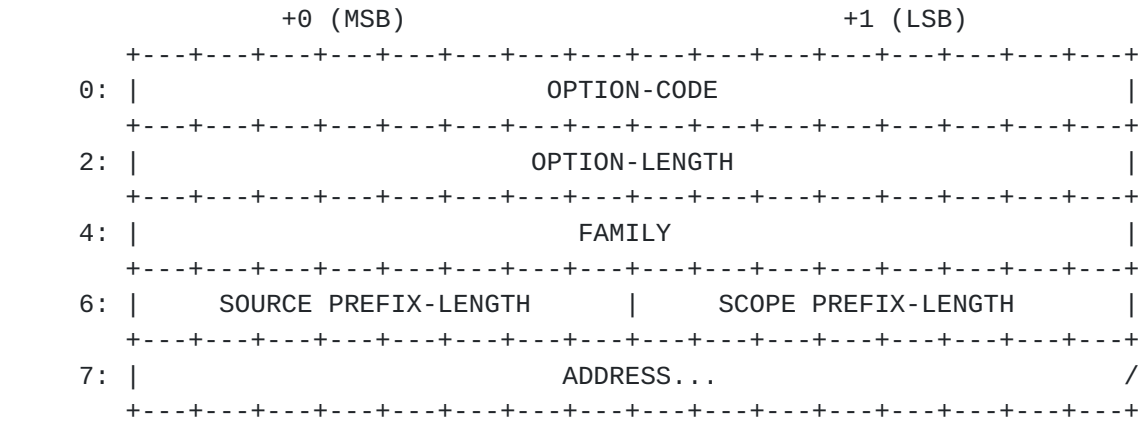
Failure to support this option or its improper handling will, at worst, cause suboptimal identification of client location, which is a common occurrence in current content delivery network (CDN) setups.

Section 6.1 also provides a mechanism for Stub Resolvers to signal Recursive Resolvers that they do not want ECS treatment for specific queries.

Additionally, operators of Intermediate Nameservers with ECS enabled are allowed to choose how many bits of the address of received queries to forward, or to reduce the number of bits forwarded for queries already including an ECS option.

5. Option Format

This protocol uses an EDNS0 [RFC6891]) option to include client address information in DNS messages. The option is structured as follows:



- (Defined in [RFC6891]) OPTION-CODE, 2 octets, for ECS is 8 (0x00 0x08).
- (Defined in [RFC6891]) OPTION-LENGTH, 2 octets, contains the length of the payload (everything after OPTION-LENGTH) in octets.
- FAMILY, 2 octets, indicates the family of the address contained in the option, using address family codes as assigned by IANA in IANA-AFI [2].

The format of the address part depends on the value of FAMILY. This document only defines the format for FAMILY 1 (IP version 4) and 2 (IP version 6), which are as follows:

- SOURCE PREFIX-LENGTH, an unsigned octet representing the leftmost significant bits of ADDRESS to be used for the lookup. In responses, it mirrors the same value as in the queries.

- SCOPE PREFIX-LENGTH, an unsigned octet representing the leftmost significant bits of ADDRESS that the response covers. In queries, it MUST be set to 0.

- ADDRESS, variable number of octets, contains either an IPv4 or IPv6 address, depending on FAMILY, truncated to the number of bits indicated by the SOURCE PREFIX-LENGTH field, with bits set to 0 to pad to the end of the last octet needed. Trailing all-zero octets SHOULD be omitted.

All fields are in network byte order ("big-endian", per [RFC1700], Data Notation).

6. Protocol Description

6.1. Originating the Option

The ECS option should generally be added by Recursive Resolvers when querying Authoritative Nameservers, as described in Section 11. The option can also be initialized by a Stub Resolver or Forwarder.

6.1.1. Recursive Resolvers

The setup of the ECS option in a Recursive Resolver depends on the client query that triggered the resolution process.

In the usual case, where no ECS option was present in the client query, the Recursive Resolver initializes the option by setting the FAMILY of the client's address. It then uses the value of its maximum cacheable prefix length to set SOURCE PREFIX-LENGTH. For privacy reasons, and because the whole IP address is rarely required to determine a tailored response, this length SHOULD be shorter than the full address, as described in Section 10.

If the triggering query included an ECS option itself, it MUST be examined for its SOURCE PREFIX-LENGTH. The Recursive Resolver's outgoing query MUST then set SOURCE PREFIX-LENGTH to the shorter of the incoming query's SOURCE PREFIX-LENGTH or the server's maximum cacheable prefix length.

Finally, in both cases, SCOPE PREFIX-LENGTH is set to 0 and the ADDRESS is then added up to the SOURCE PREFIX-LENGTH number of bits, with trailing 0 bits added, if needed, to fill the final octet. The total number of octets used should only be enough to cover SOURCE

PREFIX-LENGTH bits, rather than the full width that would normally be used by addresses in FAMILY.

FAMILY and ADDRESS information MAY be used from the ECS option in the incoming query. Passing the existing address data is supportive of the Recursive Resolver being used as the target of a Forwarder, but could possibly run into policy problems with regard to usage agreements between the Recursive Resolver and Authoritative Namserver. See Section 11.2 for more discussion on this point. If the Recursive Resolver will not forward the FAMILY and ADDRESS data from the incoming ECS option, it SHOULD return a REFUSED response. An ECS-aware resolver MUST retry the query without ECS to distinguish the response from a lame delegation, which is the common convention for a REFUSED status.

Subsequent queries to refresh the data MUST, if unrestricted by an incoming SOURCE PREFIX-LENGTH, specify the longest SOURCE PREFIX-LENGTH that the Recursive Resolver is willing to cache, even if a previous response indicated that a shorter prefix length was sufficient.

6.1.2. Stub Resolvers

A Stub Resolver MAY generate DNS queries with an ECS option set to indicate its own level of privacy via SOURCE PREFIX-LENGTH. An Intermediate Nameserver that receives such a query MUST NOT make queries that include more bits of client address than in the originating query.

A SOURCE PREFIX-LENGTH of 0 means the Recursive Resolver MUST NOT add address information of the client to its queries. The subsequent Recursive Resolver query to the Authoritative Nameserver will then either not include an ECS option or MAY optionally include its own address information, which is what the Authoritative Nameserver will almost certainly use to generate any Tailored Response in lieu of an option. This allows the answer to be handled by the same caching mechanism as other queries, with an explicit indicator of the applicable scope. Subsequent Stub Resolver queries for /0 can then be answered from this cached response.

A Stub Resolver MUST set SCOPE PREFIX-LENGTH to 0. It MAY include FAMILY and ADDRESS data, but should be prepared to handle a REFUSED response if the Intermediate Nameserver that it queries has a policy that denies forwarding of the ADDRESS. If there is no ADDRESS set, FAMILY MUST be set to 0.

6.1.3. Forwarders

Forwarders essentially appear to be Stub Resolvers to whatever Recursive Resolver is ultimately handling the query, but look like a Recursive Resolver to their client. A Forwarder using this option MUST prepare it as described in the Section 6.1.1 section above. In particular, a Forwarder that implements this protocol MUST honor SOURCE PREFIX-LENGTH restrictions indicated in the incoming query from its client. See also Section 6.5.

Since the Recursive Resolver it contacts will essentially treat it as a Stub Resolver, the Forwarder must be prepared for a REFUSED response if the Recursive Resolver does not permit incoming ADDRESS information. The Forwarded MUST retry with FAMILY and ADDRESS set to 0.

6.2. Generating a Response

6.2.1. Authoritative Nameserver

When a query containing an ECS option is received, an Authoritative Nameserver supporting ECS MAY use the address information specified in the option in order to generate a tailored response.

Authoritative Nameservers that have not implemented or enabled support for the ECS option ought to safely ignore it within incoming queries, per [RFC6891] section 6.1.2. Such a server MUST NOT include an ECS option within replies, to indicate lack of support for it. Implementers of Intermediate Nameservers should be aware, however, that some nameservers incorrectly echo back unknown EDNS0 options. In this protocol that should be mostly harmless, as SCOPE PREFIX-LENGTH should come back as 0, thus marking the response as covering all networks.

A query with a wrongly formatted option (e.g., an unknown FAMILY) MUST be rejected and a FORMERR response MUST be returned to the sender, as described by [RFC6891], Transport Considerations.

An Authoritative Nameserver that implements this protocol and receives an ECS option MUST include an ECS option in its response to indicate that it SHOULD be cached accordingly, regardless of whether the client information was needed to formulate an answer. (Note that the [RFC6891] requirement to reserve space for the OPT record could mean that the answer section of the response will be truncated and fallback to TCP indicated accordingly.) If an ECS option was not included in a query, one MUST NOT be included in the response even if the server is providing a Tailored Response -- presumably based on the address from which it received the query.

The FAMILY, SOURCE PREFIX-LENGTH and ADDRESS in the response MUST match those in the query, unless the query specified only the SOURCE PREFIX-LENGTH for privacy (with FAMILY and ADDRESS set to 0). Echoing back these values helps to mitigate certain attack vectors, as described in Section 10.

The SCOPE PREFIX-LENGTH in the response indicates the network for which the answer is intended.

A SCOPE PREFIX-LENGTH value longer than the SOURCE PREFIX-LENGTH indicates that the provided prefix length was not specific enough to select the most appropriate Tailored Response. Future queries for the name within the specified network SHOULD use the longer SCOPE PREFIX-LENGTH.

Conversely, a shorter SCOPE PREFIX-LENGTH indicates that more bits than necessary were provided, and the answer is suitable for a broader range of addresses. This could be as short as 0, to indicate that the answer is suitable for all addresses in FAMILY.

As the logical topology of any part of the network with regard to the tailored response can vary, an Authoritative Nameserver may return different values of SCOPE PREFIX-LENGTH for different networks.

Since some queries can result in multiple RRsets being added to the response, there is an unfortunate ambiguity from the original draft as to how SCOPE PREFIX-LENGTH would apply to each individual RRset. For example, multiple types in response to an ANY metaquery could all have different applicable SCOPE PREFIX-LENGTH values, but this protocol only has the ability to signal one. The response SHOULD therefore include the longest relevant PREFIX-LENGTH of any RRset in the answer, which could have the unfortunate side-effect of redundantly caching some data that could be cached more broadly. For the specific case of a CNAME chain, the Authoritative Nameserver SHOULD only place the CNAME to have it cached unambiguously appropriately. Most modern Recursive Resolvers restart the query with the canonical name, so the remainder of the chain is typically ignored anyway. For message-focused resolvers, rather than RRset-focused ones, this will mean caching the entire CNAME chain at the longest PREFIX-LENGTH of any RRset in the chain.

The specific logic that an Authoritative Nameserver uses to choose a tailored response is not in the scope of this document. Implementers are encouraged, however, to consider carefully their selection of SCOPE PREFIX-LENGTH for the response in the event that the best tailored response cannot be determined, and what the implications would be over the life of the TTL.

If the Authoritative Nameserver operator configures a more specific (longer prefix length) Tailored Response within a configured less specific (shorter prefix length) Tailored Response, then implementations can either:

1. Deaggregate the shorter prefix response into multiple longer prefix responses, or,

2. Alert the operator that the order of queries will determine which answers get cached, and either warn and continue or treat this as an error and refuse to load the configuration.

Implementations SHOULD document their chosen behavior.

6.2.2. Intermediate Nameserver

When an Intermediate Nameserver uses ECS, whether it passes an ECS option in its own response to its client is predicated on whether the client originally included the option. Because a client that did not use an ECS option might not be able to understand it, the server MUST NOT provide one in its response. If the client query did include the option, the server MUST include one in its response, especially as it could be talking to a Forwarder which would need the information for its own caching.

If an Intermediate Nameserver receives a response which has a longer SCOPE PREFIX-LENGTH than the SOURCE PREFIX-LENGTH that it provided in its query, it SHOULD still provide the result as the answer to the triggering client request even if the client is in a different address range. The Intermediate Nameserver MAY instead opt to retry with a longer SOURCE PREFIX-LENGTH to get a better reply before responding to its client, as long as it does not exceed a SOURCE PREFIX-LENGTH specified in the query that triggered resolution, but this obviously has implications for the latency of the overall lookup.

The logic for using the cache to determine whether the Intermediate Nameserver already knows the response to provide to its client is covered in the next section.

6.3. Handling ECS Responses and Caching

When an Intermediate Nameserver receives a response containing an ECS option and without the TC bit set, it SHOULD cache the result based on the data in the option. If the TC bit was set, the Intermediate Resolver SHOULD retry the query over TCP to get the complete answer section for caching.

If the FAMILY, SOURCE PREFIX-LENGTH, and SOURCE PREFIX-LENGTH bits of ADDRESS in the response don't match the non-zero fields in the corresponding query, the full response MUST be dropped, as described in Section 10. For a response to query which specified only the SOURCE PREFIX-LENGTH for privacy masking, the FAMILY and ADDRESS fields should contain the appropriate non-zero information for caching.

If no ECS option is contained in the response, the Intermediate Nameserver SHOULD treat this as being equivalent to having received a SCOPE PREFIX-LENGTH of 0, which is an answer suitable for all client addresses. See further discussion on the security implications of this in Section 10.

6.3.1. Caching the Response

In the cache, all resource records in the answer section MUST be tied to the network specified by the FAMILY, ADDRESS and SCOPE PREFIX-LENGTH fields, as limited by the Intermediate Nameserver's own configuration for maximum cacheable prefix length. Note that the additional and authority sections from a DNS response message are specifically excluded here. Any records from these sections MUST NOT be tied to a network. See more at Section 6.4.

Records that are cached as /0 because of a query's SOURCE PREFIX-LENGTH of 0 MUST be distinguished from those that are cached as /0 because of a response's SCOPE PREFIX-LENGTH of 0. The former should only be used for other /0 queries that the Intermediate Resolver receives, but the latter is suitable as a response for all networks.

Although omitting network-specific caching will significantly simplify an implementation, the resulting drop in cache hits is very likely to defeat most latency benefits provided by ECS. Therefore, when implementing this option for latency purposes, implementing full caching support as described in this section is strongly recommended.

Enabling support for ECS in an Intermediate Nameserver will significantly increase the size of the cache, reduce the number of results that can be served from cache, and increase the load on the server. Implementing the mitigation techniques described in Section 10 is strongly recommended.

6.3.2. Answering from Cache

Cache lookups are first done as usual for a DNS query, using the query tuple of <name, type, class>. Then the appropriate RRset MUST be chosen based on longest prefix matching. The client address to

use for comparison will depend on whether the Intermediate Nameserver received an ECS option in its client query.

- If no ECS option was provided, the client's address is used.
- If there was an ECS option, the ADDRESS from it MAY be used if local policy allows. Policy can vary depending on the agreements the operator of the Intermediate Nameserver has with Authoritative Nameserver operators; see Section 11.2. If policy does not allow, a REFUSED response must be sent.

If a matching network is found and the relevant data is unexpired, the response is generated as per Section 6.2.

If no matching network is found, the Intermediate Nameserver MUST perform resolution as usual. This is necessary to avoid Tailored Responses in the cache from being returned to the wrong clients, and to avoid a single query coming from a client on a different network from polluting the cache with a Tailored Response for all the users of that resolver.

6.4. Delegations and Negative Answers

The prohibition against tying ECS data to records from the Authority and Additional section left an unfortunate ambiguity in the original specification, primarily with regard to negative answers. The expectation of the original authors was that ECS would only really be used for address records, the use case that was driving the definition of the protocol.

The delegations case is a bit easier to tease out. In operational practice, if an authoritative server is using address information to provide customized delegations, it is the resolver that will be using the answer for its next iterative query. Addresses in the Additional section SHOULD therefore ignore ECS data, and the authority SHOULD return a zero SCOPE PREFIX-LENGTH on delegations. A recursive resolver SHOULD treat a non-zero SCOPE PREFIX LENGTH in a delegation as though it were zero.

For negative answers, some independent implementations of both resolvers and authorities did not see the section restriction as necessarily meaning that a given name and type must only have either positive ECS-tagged answers or a negative answer. They support being able to tell one part of the network that the data does not exist, while telling another part of the network that it does.

Several other implementations, however, do not support being able to mix positive and negative answers, and thus interoperability is a problem.

This issue is expected to be revisited in a future revision of the protocol, possibly blessing the mixing of positive and negative answers. There are implications for cache data structures that developers should consider when writing new ECS code.

6.5. Transitivity

Generally, ECS options will only be present in DNS messages between a Recursive Resolver and an Authoritative Nameserver, i.e., one hop. In certain configurations however, for example multi-tier nameserver setups, it may be necessary to implement transitive behaviour on Intermediate Nameservers.

It is important that any Intermediate Nameserver that forwards ECS options received from their clients MUST fully implement the caching behaviour described in Section 6.3.

Intermediate Nameservers supporting ECS MUST forward options with SOURCE PREFIX-LENGTH set to 0 (that is, completely anonymized). Such options MUST NOT be replaced with more accurate address information.

An Intermediate Nameserver MAY also forward ECS options with actual address information. This information MAY match the source IP address of the incoming query, and MAY have more or fewer address bits than the Nameserver would normally include in a locally originated ECS option.

If for any reason the Intermediate Nameserver does not want to use the information in an ECS option it receives (too little address information, network address from a range not authorized to use the server, private/unroutable address space, etc), it SHOULD drop the query and return a REFUSED response. Note again that a query MUST NOT be refused solely because it provides 0 address bits.

Be aware that at least one major existing implementation does not return REFUSED and instead just process the query as though the problematic information were not present. This can lead to anomalous situations, such as a response from the Intermediate Nameserver that indicates it is tailored for one network (the one passed in the original query, since ADDRESS must match) when actually it is for another network (the one which contains the address that the Intermediate Nameserver saw as making the query).

7. IANA Considerations

IANA has already assigned option code 8 in the "DNS EDNS0 Option Codes (OPT)" registry to ECS.

The IANA is requested to update the reference ("draft-vandergaast-edns-client-subnet") to refer to this RFC when published.

8. DNSSEC Considerations

The presence or absence of an [RFC6891] EDNS0 OPT resource record containing an ECS option in a DNS query does not change the usage of the resource records and mechanisms used to provide data origin authentication and data integrity to the DNS, as described in [RFC4033], [RFC4034] and [RFC4035]. OPT records are not signed.

Use of this option, however, does imply increased DNS traffic between any given Recursive Resolver and Authoritative Nameserver, which could be another barrier to further DNSSEC adoption in this area.

9. NAT Considerations

Special awareness of ECS in devices that perform Network Address Translation (NAT) as described in [RFC2663] is not required; queries can be passed through as-is. The client's network address SHOULD NOT be added, and existing ECS options, if present, SHOULD NOT be modified by NAT devices.

In large-scale global networks behind a NAT device (but for example with Centralized Resolver infrastructure), an internal Intermediate Nameserver might have detailed network layout information, and may know which external subnets are used for egress traffic by each internal network. In such cases, the Intermediate Nameserver MAY use that information when originating ECS options.

In other cases, Recursive Resolvers sited behind a NAT device SHOULD NOT originate ECS options with their external IP address, and instead rely on downstream Intermediate Nameservers to do so. They MAY, however, choose to include the option with their internal address for the purposes of signaling a shorter, more anonymous SOURCE PREFIX-LENGTH.

If an Authoritative Nameserver on the publicly routed Internet receives a query that specifies an ADDRESS in [RFC1918] or [RFC4193] private address space, it SHOULD ignore ADDRESS and look up its answer based on the address of the Recursive Resolver. In the response it SHOULD set SCOPE PREFIX-LENGTH to cover all of the relevant private space. For example, a query for ADDRESS 10.1.2.0

with a SOURCE PREFIX-LENGTH of 24 would get a returned SCOPE PREFIX-LENGTH of 8. The Intermediate Nameserver MAY elect to cache the answer under one entry for special-purpose addresses [RFC6890]; see Section 10.3.

10. Security Considerations

10.1. Privacy

With the ECS option, the network address of the client that initiated the resolution becomes visible to all servers involved in the resolution process. Additionally, it will be visible from any network traversed by the DNS packets.

To protect users' privacy, Recursive Resolvers are strongly encouraged to conceal part of the IP address of the user by truncating IPv4 addresses to 24 bits. 56 bits are recommended for IPv6, based on [RFC6177].

ISPs should have more detailed knowledge of their own networks. That is, they might know that all 24-bit prefixes in a /20 are in the same area. In those cases, for optimal cache utilization and improved privacy, the ISP's Recursive Resolver SHOULD truncate IP addresses in this /20 to just 20 bits, instead of 24 as recommended above.

Users who wish their full IP address to be hidden can include an ECS option specifying the wildcard address (i.e. SOURCE PREFIX-LENGTH of 0). As described in previous sections, this option will be forwarded across all the Recursive Resolvers supporting ECS, which MUST NOT modify it to include the network address of the client.

Note that even without an ECS option, any server queried directly by the user will be able to see the full client IP address. Recursive Resolvers or Authoritative Nameservers MAY use the source IP address of queries to return a cached entry or to generate a Tailored Response that best matches the query.

10.2. Birthday Attacks

ECS adds information to the DNS query tupe (q-tuple). This allows an attacker to send a caching Intermediate Nameserver multiple queries with spoofed IP addresses either in the ECS option or as the source IP. These queries will trigger multiple outgoing queries with the same name, type and class, just different address information in the ECS option.

With multiple queries for the same name in flight, the attacker has a higher chance of success to send a matching response with the SCOPE PREFIX-LENGTH set to 0 to get it cached for all hosts.

To counter this, the ECS option in a response packet MUST contain the full FAMILY, ADDRESS and SOURCE PREFIX-LENGTH fields from the corresponding query. Intermediate Nameservers processing a response MUST verify that these match, and SHOULD discard the entire response if they do not.

That requirement to discard is "SHOULD" instead of "MUST" because it stands in opposition to the instruction in Section 6.3 which states that a response lacking an ECS option should be treated as though it had one of SCOPE PREFIX-LENGTH of 0. If that is always true, then an attacker does not need to worry about matching the original ECS option data and just needs to flood back responses that have no ECS option at all.

This type of attack could be detected in ongoing operations by marking whether the responding nameserver had previously been sending ECS option, and/or by taking note of an incoming flood of bogus responses and flagging the relevant query for re-resolution. This is more complex than existing nameserver responses to spoof floods, and would also need to be sensitive to a nameserver legitimately stopping ECS replies even though it had previously given them.

10.3. Cache Pollution

It is simple for an arbitrary resolver or client to provide false information in the ECS option, or to send UDP packets with forged source IP addresses.

This could be used to:

- pollute the cache of intermediate resolvers, by filling it with results that will rarely (if ever) be used.
- reverse engineer the algorithms (or data) used by the Authoritative Nameserver to calculate Tailored Responses.
- mount a denial-of-service attack against an Intermediate Nameserver, by forcing it to perform many more recursive queries than it would normally do, due to how caching is handled for queries containing the ECS option.

Even without malicious intent, Centralized Resolvers providing answers to clients in multiple networks will need to cache different

responses for different networks, putting more memory pressure on the cache.

To mitigate those problems:

- Recursive Resolvers implementing ECS should only enable it in deployments where it is expected to bring clear advantages to the end users. For example, when expecting clients from a variety of networks or from a wide geographical area. Due to the high cache pressure introduced by ECS, the feature SHOULD be disabled in all default configurations.

- Recursive Resolvers SHOULD limit the number of networks and answers they keep in the cache for any given query.

- Recursive Resolvers SHOULD limit the number of total different networks that they keep in cache.

- Recursive Resolvers MUST never send an ECS option with a SOURCE PREFIX-LENGTH providing more bits in the ADDRESS than they are willing to cache responses for.

- Recursive Resolvers should implement algorithms to improve the cache hit rate, given the size constraints indicated above. Recursive Resolvers MAY, for example, decide to discard more specific cache entries first.

- Authoritative Nameservers and Recursive Resolvers should discard ECS options that are either obviously forged or otherwise known to be wrong. They SHOULD at least treat unroutable addresses, such as some of the address blocks defined in [RFC6890], as equivalent to the Recursive Resolver's own identity. They SHOULD ignore and never forward ECS options specifying other routable addresses that are known not to be served by the query source.

- Authoritative Nameservers consider the ECS option just as a hint to provide better results. They can decide to ignore the content of the ECS option based on black or white lists, rate limiting mechanisms, or any other logic implemented in the software.

11. Sending the Option

When implementing a Recursive Resolver, there are two strategies on deciding when to include an ECS option in a query. At this stage, it's not clear which strategy is best.

11.1. Probing

A Recursive Resolver can send the ECS option with every outgoing query. However, it is RECOMMENDED that Resolvers remember which Authoritative Nameservers did not return the option with their response, and omit client address information from subsequent queries to those Nameservers.

Additionally, Recursive Resolvers SHOULD be configured to never send the option when querying root, top-level, and effective top-level domain servers. These domains are delegation-centric and are very unlikely to generate different responses based on the address of the client.

When probing, it is important that several things are probed: support for ECS, support for EDNS0, support for EDNS0 options, or possibly an unreachable Nameserver. Various implementations are known to drop DNS packets with OPT RRs (with or without options), thus several probes are required to discover what is supported.

Probing, if implemented, MUST be repeated periodically, e.g., daily. If an Authoritative Nameserver indicates ECS support for one zone, it is to be expected that the Nameserver supports ECS for all of its zones. Likewise, an Authoritative Nameserver that uses ECS information for one of its zones, MUST indicate support for the option in all of its responses to ECS queries. If the option is supported but not actually used for generating a response, its SCOPE PREFIX-LENGTH MUST be set to 0.

11.2. Whitelist

As described previously, it is expected that only a few Recursive Resolvers will need to use ECS, and that it will generally be enabled only if it offers a clear benefit to the users.

To avoid the complexity of implementing a probing and detection mechanism (and the possible query loss/delay that may come with it), an implementation could use a whitelist of Authoritative Namesevers to send the option to, likely specified by their domain name. Implementations MAY also allow additionally configuring this based on other criteria, such as zone or query type.

An additional advantage of using a whitelist is that partial client address information is only disclosed to Nameservers that are known to use the information, improving privacy.

A major drawback is scalability. The operator needs to track which Authoritative Nameservers support ECS, making it harder for new Authoritative Nameservers to start using the option.

Similarly, Authoritative Nameservers can also use whitelists to limit the feature to only certain clients. For example, a CDN that does not want all of their mapping trivially walked might require a legal agreement with the Recursive Resolver operator, to clearly describe the acceptable use of the feature.

The maintenance of access control mechanisms is out of scope for this protocol definition.

12. Example

1. A stub resolver, SR, with IP address 192.0.2.37 tries to resolve www.example.com, by forwarding the query to the Recursive Resolver, RNS, from IP address IP, asking for recursion.

2. RNS, supporting ECS, looks up www.example.com in its cache. An entry is found neither for www.example.com, nor for example.com.

3. RNS builds a query to send to the root and .com servers. The implementation of RNS provides facilities so an administrator can configure it not to forward ECS in certain cases. In particular, RNS is configured to not include an ECS option when talking to TLD or root nameservers, as described in Section 6.1. Thus, no ECS option is added, and resolution is performed as usual.

4. RNS now knows the next server to query: the Authoritative Nameserver, ANS, responsible for example.com.

5. RNS prepares a new query for www.example.com, including an ECS option with:

   * OPTION-CODE, set to 8.

   * OPTION-LENGTH, set to 0x00 0x07 for the following fixed 4 octets plus the 3 octets that will be used for ADDRESS.

   * FAMILY, set to 0x00 0x01 as IP is an IPv4 address.

   * SOURCE PREFIX-LENGTH, set to 0x18, as RNS is configured to conceal the last 8 bits of every IPv4 address.

   * SCOPE PREFIX-LENGTH, set to 0x00, as specified by this document for all queries.

* ADDRESS, set to 0xC0 0x00 0x02, providing only the first 24 bits of the IPv4 address.

6. The query is sent. ANS understands and uses ECS. It parses the ECS option, and generates a Tailored Response.

7. Due its internal implementation, ANS finds a response that is tailored for the whole /16 of the client that performed the query.

8. ANS adds an ECS option in the response, containing:

   * OPTION-CODE, set to 8.

   * OPTION-LENGTH, set to 0x00 0x07.

   * FAMILY, set to 0x00 0x01.

   * SOURCE PREFIX-LENGTH, set to 0x18, copied from the query.

   * SCOPE PREFIX-LENGTH, set to 0x10, indicating a /16 network.

   * ADDRESS, set to 0xC0 0x00 0x02, copied from the query.

9. RNS receives the response containing an ECS option. It verifies that FAMILY, SOURCE PREFIX-LENGTH, and ADDRESS match the query. If not, the message is discarded.

10. The response is interpreted as usual. Since the response contains an ECS option, the ADDRESS, SCOPE PREFIX-LENGTH, and FAMILY in the response are used to cache the entry.

11. RNS sends a response to stub resolver SR, without including an ECS option.

12. RNS receives another query to resolve www.example.com. This time, a response is cached. The response, however, is tied to a particular network. If the address of the client matches any network in the cache, then the response is returned from the cache. Otherwise, another query is performed. If multiple results match, the one with the longest SCOPE PREFIX-LENGTH is chosen, as per common best-network match algorithms.

13. Contributing Authors

The below individuals contributed significantly to the draft. The RFC Editor prefers a maximum of 5 names on the front page, and so we have listed additional authors in this section

Edward Lewis
ICANN
12025 Waterfront Drive, Suite 300
Los Angeles CA 90094-2536
USA
Email: edward.lewis@icann.org

Sean Leach
Fastly
POBox 78266
San Francisco CA 94107

Jason Moreau
Akamai Technologies
8 Cambridge Ctr
Cambridge MA 02142-1413
USA

14. Acknowledgements

The authors wish to thank Darryl Rodden for his work as a co-author on previous versions, and the following people for reviewing early drafts of this document and for providing useful feedback: Paul S. R. Chisholm, B. Narendran, Leonidas Kontothanassis, David Presotto, Philip Rowlands, Chris Morrow, Kara Moscoe, Alex Nizhner, Warren Kumari, and Richard Rabbat from Google; Terry Farmer, Mark Teodoro, Edward Lewis, and Eric Burger from Neustar; David Ulevitch and Matthew Dempsky from OpenDNS; Patrick W. Gilmore and Steve Hill from Akamai; Colm MacCarthaigh and Richard Sheehan from Amazon; Tatuya Jinmei from Internet Software Consortium; Andrew Sullivan from Dyn; John Dickinson from Sinodun; Mark Delany from Apple; Yuri Schaeffer from NLnet Labs; Duane Wessels from from Verisign; Antonio Querubin; and all of the other people that replied to our emails on various mailing lists.

15. References

15.1. Normative References

[RFC1034] Mockapetris, P., "Domain names - concepts and facilities", STD 13, RFC 1034, November 1987.

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

[RFC1700] Reynolds, J. and J. Postel, "Assigned Numbers", RFC 1700, October 1994.

[RFC1918] Rekhter, Y., Moskowitz, R., Karrenberg, D., Groot, G., and E. Lear, "Address Allocation for Private Internets", BCP 5, RFC 1918, February 1996.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC4033] Arends, R., Austein, R., Larson, M., Massey, D., and S. Rose, "DNS Security Introduction and Requirements", RFC 4033, March 2005.

[RFC4034] Arends, R., Austein, R., Larson, M., Massey, D., and S. Rose, "Resource Records for the DNS Security Extensions", RFC 4034, March 2005.

[RFC4035] Arends, R., Austein, R., Larson, M., Massey, D., and S. Rose, "Protocol Modifications for the DNS Security Extensions", RFC 4035, March 2005.

[RFC4193] Hinden, R. and B. Haberman, "Unique Local IPv6 Unicast Addresses", RFC 4193, October 2005.

[RFC6177] Narten, T., Huston, G., and L. Roberts, "IPv6 Address Assignment to End Sites", BCP 157, RFC 6177, March 2011.

[RFC6890] Cotton, M., Vegoda, L., Bonica, R., and B. Haberman, "Special-Purpose IP Address Registries", BCP 153, RFC 6890, April 2013.

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

15.2. Informative References

[RFC2663] Srisuresh, P. and M. Holdrege, "IP Network Address Translator (NAT) Terminology and Considerations", RFC 2663, August 1999.

15.3. URIs

[1] http://www.iana.org/assignments/address-family-numbers/

Appendix A. Document History

[RFC Editor: Please delete this section before publication.]

-02 to -03 (IETF)

- Clean up the open issues, mostly by saying that they were out of scope for this document.

- How in the world did no reviewers note that "Queries" had been spelled as "Querys" in the title? (Aaron Falk did.)

-01 to -02 (IETF)

- Note ambiguity with multiple RRsets appearing in reply, eg, for an ANY query or CNAME chain. (Duane Wessels)

- Open issue questioning the guidance about resolvers behind a NAT. How do they know they are? What real requirement is this imposing? (Duane Wessels)

- Some other wording changes based on Duane's review of an earlier draft.

-00 to -01 (IETF)

- <David> Made the document describe how things are actually implmented now. This makes the document be more of a "this is how we are doing things, this provides information on that". There may be a future document that describes additional funcationality.

- NETMASK was not a good desription, changed to PREFIX-LENGTH (Jinmei, others). Stole most of the definition for prefix length from RFC4291.

- Fixed the "SOURCE PREFIX-LENGTH set to 0" definition to include IPv6 (Tatuya Jinmei)

- Comment that ECS cannot be used to hand NXDOMAIN to some clients and not others, primarily because of interoperability issues. (Tatuya Jinmei)

- Added text explaining that implmentations need to document thier behavior with overlapping networks.

- Soften "optimized reply" language. (Andrew Sullivan).

- Fixed some of legacy IPv4 cruft (things like 0.0.0.0/0)

- Some more grammar / working cleanups.

- Replaced a whole heap of occurances of "edns-client-subnet" with "ECS" for readability. (John Dickinson)

- More clearly describe the process from the point of view of each type of nameserver. (John Dickinson)
- Birthday attack still possible if attacker floods with ECS-less responses. (Yuri Schaeffer)
- Added some open issues directly to the text.

A.1. -00

- Document moved to experimental track, added experiment description in header with details in a new section.
- Specifically note that ECS applies to the answer section only.
- Warn that caching based on ECS is optional but very important for performance reasons.
- Updated NAT section.
- Added recommendation to not use the default /24 recommendation for the source prefix-length field if more detailed information about the network is available.
- Rewritten problem statement to be more clear about the goal of ECS and the fact that it's entirely optional.
- Wire format changed to include the original address and prefix length in responses in defence against birthday attacks.
- Security considerations now includes a section about birthday attacks.
- Renamed edns-client-ip in ECS, following suggestions on the mailing list.
- Clarified behavior of resolvers when presented with an invalid ECS option.
- Fully take multi-tier DNS setups in mind and be more clear about where the option should be originated.
- A note on Authoritative Nameservers receiving queries that specify private address space.
- A note to always ask for the longest acceptable SOURCE prefix length, even if a prior answer indicated that a shorter prefix length was suitable.

- Marked up a few more references.
- Added a few definitions in the Terminology section, and a few more aesthetic changes in the rest of the document.

A.2. -01

- Document version number reset from -02 to -00 due to the rename to ECS.
- Clarified example (dealing with TLDs, and various minor errors).
- Referencing RFC5035 instead of RFC1918.
- Added a section on probing (and how it should be done) vs. whitelisting.
- Moved description on how to forward ECS option in dedicated section.
- Queries with wrongly formatted ECS options should now be rejected with FORMERR.
- Added an "Overview" section, providing an introduction to the document.
- Intermediate Nameservers can now remove an ECS option, or reduce the SOURCE PREFIX-LENGTH to increase privacy.
- Added a reference to DoS attacks in the Security section.
- Don't use "network range", as it seems to have different meaning in other contexts, and turned out to be confusing.
- Use shorter and longer prefix lengths, rather than higher or lower. Add a better explanation in the format section.
- Minor corrections in various other sections.

A.3. -02

- Added IANA-assigned option code.

Authors' Addresses

Carlo Contavalli
Google
1600 Amphitheater Parkway
Mountain View, CA 94043
US

Email: ccontavalli@google.com

Wilmer van der Gaast
Google
Belgrave House, 76 Buckingham Palace Road
London SW1W 9TQ
UK

Email: wilmer@google.com

David C Lawrence
Akamai Technologies
8 Cambridge Center
Cambridge, MA 02142
US

Email: tale@akamai.com

Warren Kumari
Google
1600 Amphitheatre Parkway
Mountain View, CA 94043
US

Email: warren@kumari.net