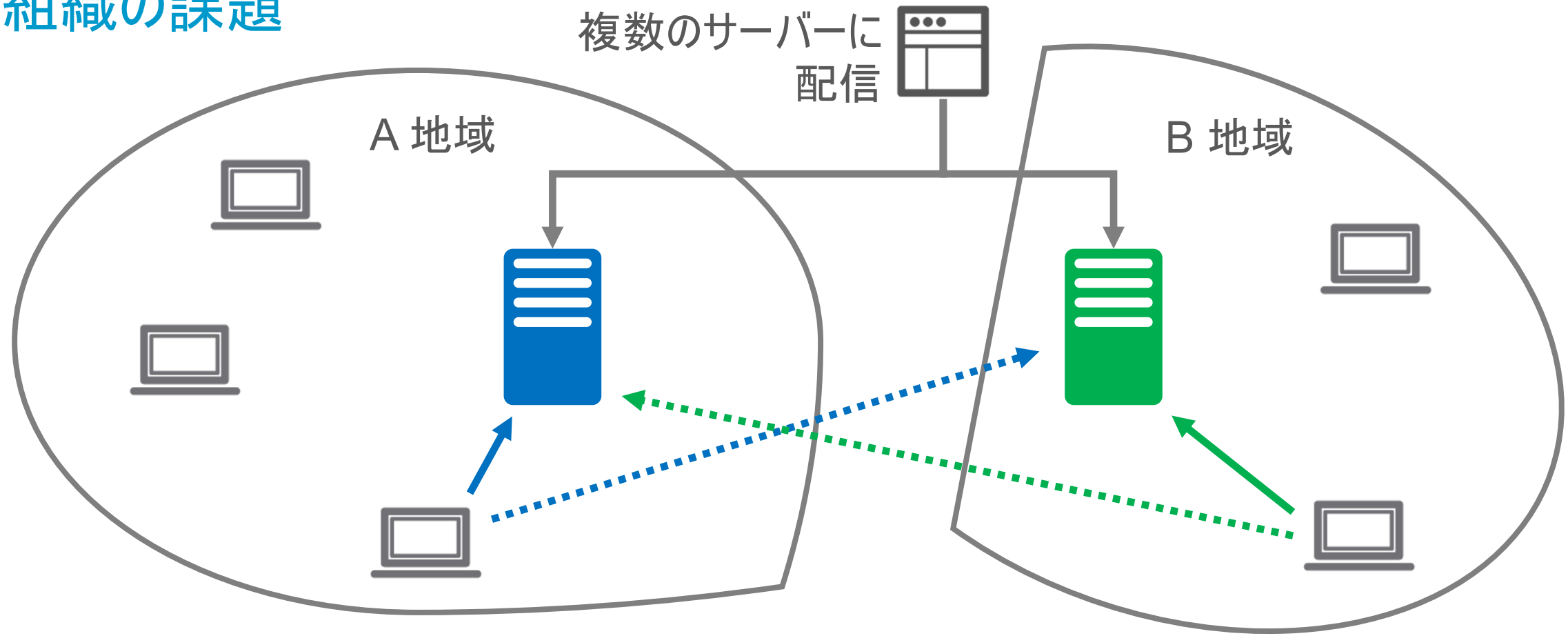
EDNS Client Subnet (ECS)

松本 陽一
アカマイ・テクノロジーズ合同会社

Akamai
Experience the Edge

このプレゼンテーションにおいてなされるステートメントは作成者個人の見解を示すものであり、アカマイ・テクノロジーズの見解を示すものではありません。提供される情報は作成時点において正確なものであると考えておりますが、当該情報についてなんら表明又は保証を行いません。

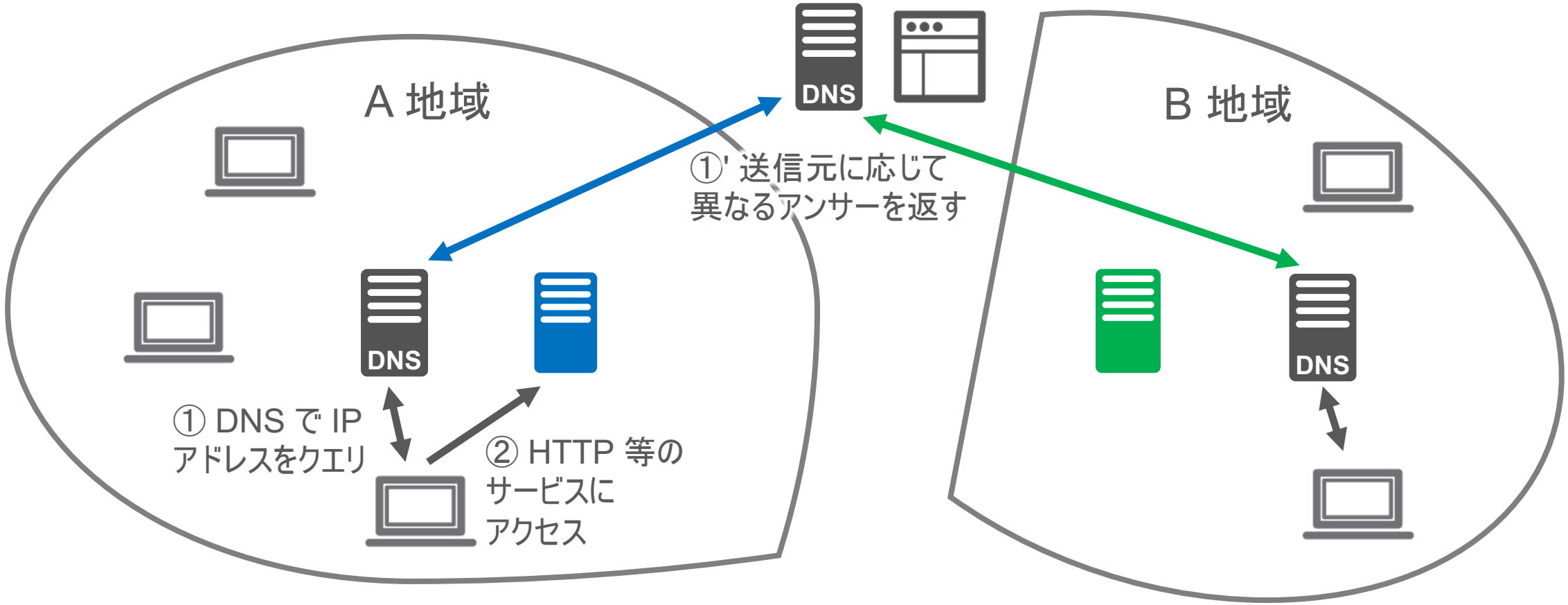
CDN 等、コンテンツやサービスを提供するサーバーを広く分散して持つ組織の課題



同じURLやホスト名によるリクエストに対し、最適なサーバーにアクセスさせたい
(さらには障害や負荷に応じて互いにバックアップ、分散させたい)

Akamai

クライアントに対して最適なサーバーにアクセスさせる方法の一つ
DNS クエリ送信元によるマッピング



権威ネームサーバーで送信元 (フルリゾルバー) の IP アドレスをチェックし、
動的に最適なサーバーの IP アドレスを応答

Akamai

DNS クエリ送信元によるマッピングがうまくいく条件

フルリゾルバーの IP アドレスを知ることがクライアントの位置を知ることと同等、つまり同じフルリゾルバーを使用するクライアント同士がトポロジー的に近くに集まっている

多くの場合、再帰リゾルバは ISP や組織の網内の (近い) ものを指定されるため、この方法はとても有効に働いてきた。(例: GGSN/P-GW は国内数カ所でそれぞれにフルリゾルバーが置かれている)

障害や負荷に応じた割り振りなど、柔軟なコントロールも可能

GSLB (Global Server Load Balancing) や、Traffic Management 等と呼ばれ、権威ネームサーバーとしての製品やサービスとして多く提供されている

再帰リゾルバの IP アドレスから適切なサーバーを選択するかはそれぞれ独自の技術。地理情報、ネットワーク品質測定、死活監視、サーバー負荷データ収集…

Akamai

他の方法の例

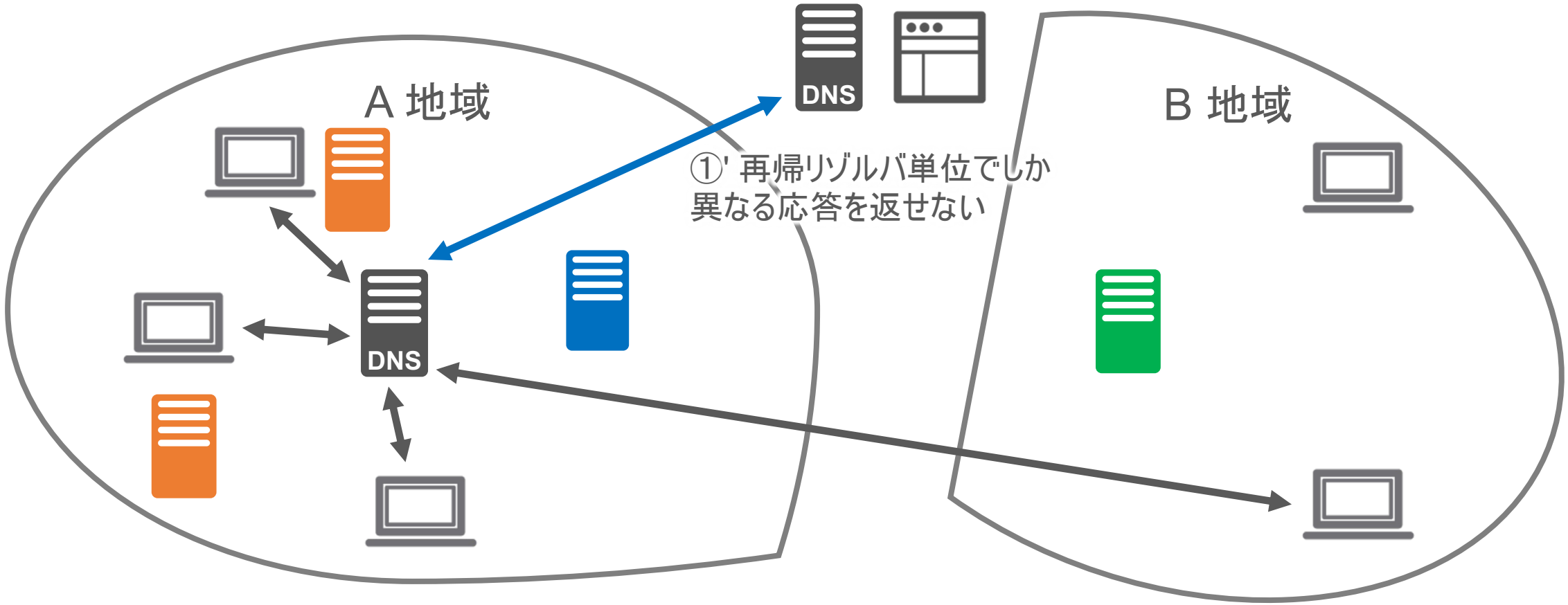
- IP Anycast
  各拠点のサーバーに同じ IP アドレスを持たせ、それぞれのネットワークから経路を広報。動的ルーティングの結果として近いサーバーにアクセスさせる
  - 特長: 権威ネームサーバーで特別な機能が不要
  - 短所: 負荷に応じて割り振るなど細かなコントロールができない、
    切替時に通信中のセッションが断絶、等

- HTTP でのリダイレクション等、アプリケーションプロトコルでの対応
  アプリケーションサーバーで送信元 IP アドレスをチェックし、他に適切なサーバーがある場合は別のサーバーにアクセスしなおさせる等
  - 特長: クライアントの IP アドレスに基づいて決定できる
  - 短所: 時間がかかる、同じホスト名としての通信ではなくなる、等

それぞれ長所、短所あり、複数の方法を組み合わせることも可能

Akamai

DNS クエリ送信元によるマッピングがうまくいかない場面



- 同一のフルリゾルバーを利用しているユーザーが離れた拠点に分散
- (コンテンツ等の) サーバーがフルリゾルバーよりも細かく配置

Akamai

DNS の送信元では不足な場面

公開 DNS（フルリゾルバー) サービスの開始と広がりにより顕在化

例)

- OpenDNS（2005)
- Google Public DNS（2009)
  → 多くの地域に展開し、IP Anycast を用いる等、なるべくクライアントに近いところから権威ネームサーバーにアクセスするようにするとはいえ限界がある

一般の ISP でも..

- 物理的に広い国や、バックボーンの帯域が潤沢に確保できない ISP
- コンテンツの多様化による大容量・広帯域化、低遅延の要求により、網内の多拠点にサーバーを配置したり、多拠点で他のネットワークと接続

[解決方法]
ECS（EDNS Client Subnet) - フルリゾルバーから権威ネームサーバーへのクエリに、再帰要求元クライアントの IP アドレスのサブネット情報をつける

Akamai

EDNS (0) - RFC 6891

DNS プロトコルの拡張 – 追加 (Additional) セクション内に「OPT」Pseudo（擬似）RR（41）
OPT は他の一般的な RR と異なるフォーマットで扱われる (クラス部分とTTL部分を再定義)

- UDP による DNS メッセージサイズ（RFC 1035 では 512bit）の拡大
- 新たなフラグ（DO ビット – DNSSEC OK)
- レスポンスコード（RCODE）空間の拡張 – BADVERS（16）

EDNS (OPT) で必然な基本部分

- オプションの追加（オプションコードと可変長のオプションデータのペア)
  各種データを乗せてクライアント - サーバー間でやりとり
  例)
  - EDNS Client Subnet（RFC 7871 / オプションコード 8)
  - DNS Cookie（RFC 7873 / オプションコード 10)
  - Padding（RFC 7830 / オプションコード 12)
  - DNSSEC Key Tag（RFC8145 / オプションコード 14)
  - Client ID / Device ID の類（非標準だがベンダー独自で実装。26946 や 65073 など)

オプショナル

OPT 擬似 RR 自体は 1 つだが、オプションはなくても複数あってもよい

(RFC 2671 / 2673 で定義されたラベルタイプ拡張は RFC 6891 で廃止)

Akamai

EDNS (0)　(続き)

フルリゾルバー – 権威ネームサーバー間では広く利用

- DNSSEC 署名検証では必須 (EDNS でないと DO ビットが立てられない)
- ほとんどの実装でデフォルトで EDNS を利用
- 「DNS Flag Day」ってありましたね

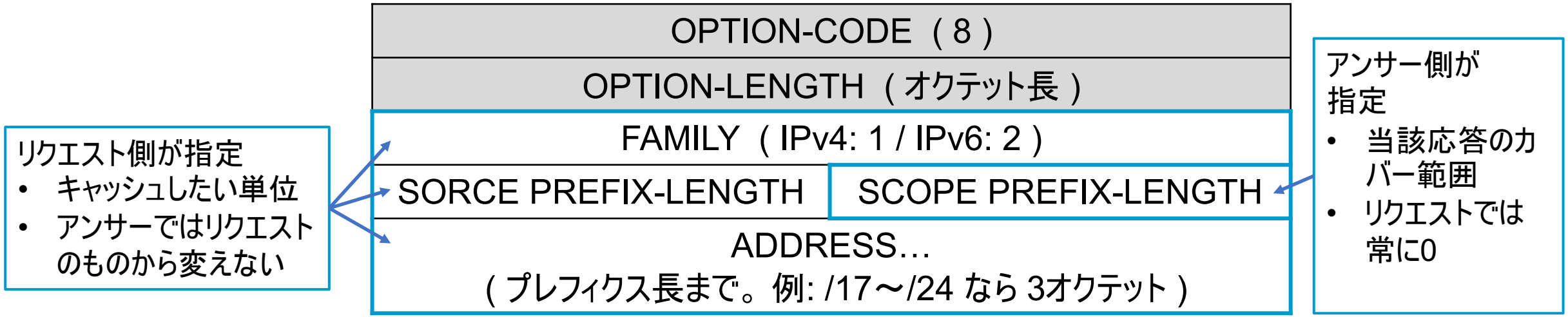
スタブリゾルバー – フルリゾルバー間での利用は限定的

- フルリゾルバーで DNSSEC 署名検証を行っても、利用するスタブリゾルバー側の動作は変わらない (AD フラグや CD フラグには EDNS は必須ではない)
- dig で生成するクエリはデフォルトで EDNS
- DNS-over-TLS、DNS-over-HTTPS 等における Padding

Akamai

EDNS Client Subnet（ECS）– RFC 7871

- EDNS (0) オプション（Option Code: 8）を利用してクライアントの IP アドレスのサブネット情報 (アドレスとプレフィクス長) をやりとりする
- 原則としてフルリゾルバーと権威ネームサーバーの間で使われる
  RFCではクライアントから ECS を受け取った場合についても議論
- フォーマット



EDNS Client Subnet (ECS)

dig コマンドでも (最近のものなら) 付加可能

```
$ dig （中略） +subnet=192.0.2.0/23
 (中略)
;; OPT PSEUDOSECTION:
; EDNS: version: 0, flags:; udp: 512
; CLIENT-SUBNET: 192.0.2.0/23/24
;; QUESTION SECTION:
```

Wireshark / tshark で (最近のものなら)
解析表示可能

```
▼ Additional records
  ▼ <Root>: type OPT
      Name: <Root>
      Type: OPT (41)
      UDP payload size: 512
      Higher bits in extended RCODE: 0x00
      EDNS0 version: 0
    ▶ Z: 0x0000
      Data length: 11
    ▼ Option: CSUBNET – Client subnet
        Option Code: CSUBNET – Client subnet (8)
        Option Length: 7
        Option Data: 00011718c00002
        Family: IPv4 (1)
        Source Netmask: 23
        Scope Netmask: 24
        Client Subnet: 192.0.2.0
```


Akamai

ECS の利用

みんなが対応することを目指すものではなく、権威ネームサーバー側もフルリゾルバー側もそれぞれ選択的に用いる

- フルリゾルバー側
  付けたい時のみ権威ネームサーバー宛のクエリに付加する
  - 権威 DNS 側が応答を返せない場合もある
  - プライバシー
    → クエリ送信先権威ネームサーバのホワイトリスト
- 権威ネームサーバー側
  - 対応したくなければ ECS は無視 (オプションを削って応答) してよい
  - 送信元 IP アドレスや ECS に基づいて異なる IP アドレスを返すこと自体が各々独自
  - 虚偽のアドレスをつけるクライアントは信頼したくない (スキャンの恐れ) → クエリ送信元のホワイトリスト

権威ネームサーバー側とフルリゾルバー側との合意、調整が必要
(公開 DNS では原則として ECS をつけるものもある)

Akamai

ECS を複雑にしているもの..あらゆる境界的なパターン

- 権威ネームサーバーは SOURCE と異なる SCOPE を返してもよい
  - SCOPE を長くして返す -> 本当はより細かいサブネットの情報が欲しい
    (フルリゾルバーは応じてリクエストを再送してもよい)
  - SCOPE を短くして返す -> より荒い情報で十分
    (192.0.2.0/16 というのはネットワークアドレス表記としてはおかしいのだが)

  SOURCE と SCOPE のプレフィクス長の短い方（アドレス空間の広い方）でキャッシュ

  フルリゾルバー側ではその後どう対応するか？モニターの必要性？
- フルリゾルバーへのクエリにクライアントが ECS をつけてもよい
  - クライアントが Forwarding Resolver の場合
    フルリゾルバーは受け入れるか? 受け入れないなら REFUSED
    例) クライアントのつけてきた ECS が自社の管理するネットワーク外
  - オプトアウト (SOURCE-PREFIX-LENGTH=0)

Akamai

フルリゾルバーにおける扱い

- ECS つきの応答の情報はサブネット情報と共にキャッシュに保存され、対象サブネット内からリクエストを受けた場合のみ使われる
  ECS がついていないものや SCOPE=0 の応答は通常通り全体向けにキャッシュ
- サブネットが細かい (プレフィクス長が長い) ほど…。
  - サブネットの数が多くなる = キャッシュ容量 (メモリー) が増える可能性
  - サブネットあたりのクライアントが少なくなる = キャッシュミスが増える可能性
    (権威側へのクエリの増加、平均応答時間の増加)
- ホワイトリストの単位 (権威ネームサーバー、ドメイン、その他)
- クライアントから受信した ECS のポリシー

将来 ECS を利用することを考えると、ネットワーク内の IP アドレスの割り当てにおいて、ネットワークトポロジーや地理的位置をなるべく短いプレフィックスに反映させるのがよいかもしれない

Akamai

ECS 以外のアプローチ

フルリゾルバーをクライアントが分散している単位毎に展開する
(他網やデータセンタとの接続点の単位等ネットワークトポロジー的に)

- クライアントの位置に応じてフルリゾルバーを割り当てる
- 公開 DNS 等、単一の IP アドレスで展開する場合はサービス用（クライアント側）インターフェイスを IP Anycast で分散する方法
- サーバーの台数が増える可能性もあるが、従来と同じ方法で課題を解決
- 細かい単位でアドレスがフラグメントしてばらまかれている場合は ECS よりも正確なマッピングが可能な場合も
- キャッシュがフラグメントされることによるキャッシュ効率の低下も防ぐ

Akamai

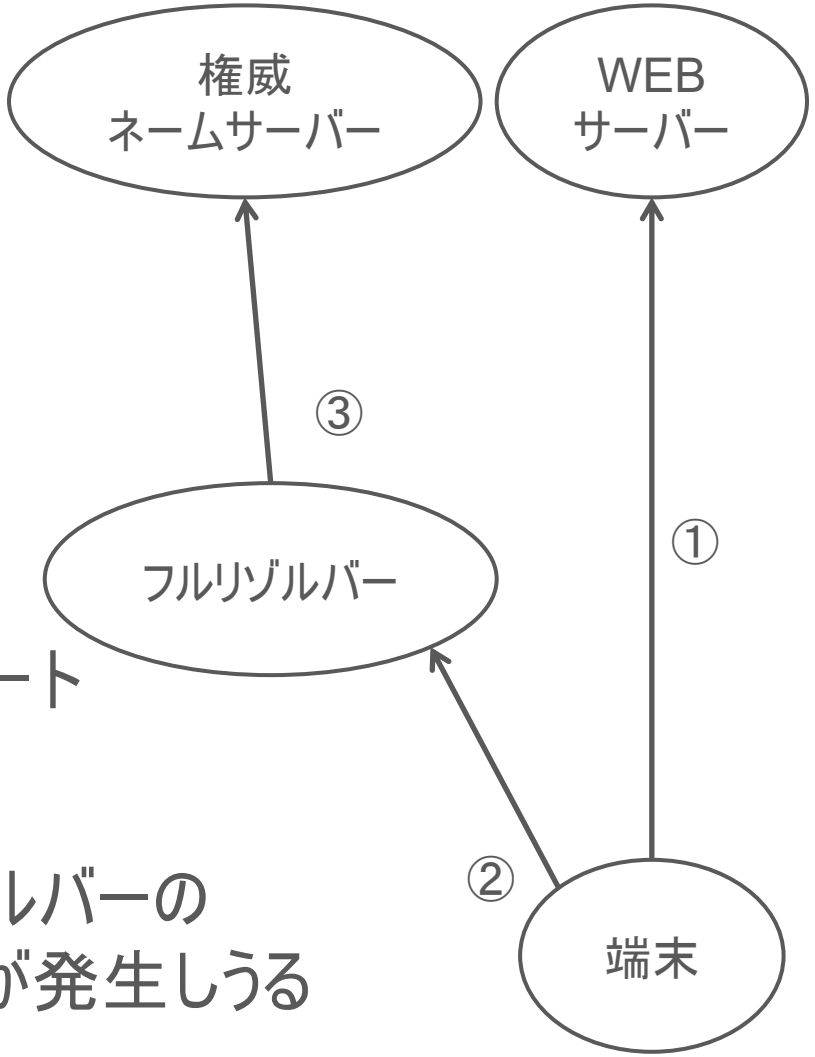
IPv6 とマッピング、ECS

下記のアドレスファミリはそれぞれ独立しており異なる
可能性があるので意識が必要

① HTTP等、目的の通信

② クライアント (スタブリゾルバー) とフルリゾルバーとの間の
トランスポート

③ フルリゾルバーと権威ネームサーバーとの間のトランスポート



- IPv6 通信のためのアドレス (①) のマッピングをフルリゾルバーの
  IPv4 アドレス (③) によって行なうこと (またはその逆) が発生しうる
- ECS (②) はさらに異なる可能性がある
- ECS を受け取ると権威ネームサーバーは ② と ③ 両方使用できる

Akamai

ECS とプライバシー

通常は権威ネームサーバーには見えないクライアントの情報が見えるようになってしまうのでプライバシーの後退ではないかという議論

- ホワイトリスト
- オプトアウト - RFC 7871 はクライアントが SOURCE-PREFIX-LENGTH=0 の ECS を送ることでオプトアウトする方式を記述
- Mozilla の Trusted Recursive Resolver ポリシー
  https://wiki.mozilla.org/Security/DOH-resolver-policy

Akamai

調査に役立つ QNAME

`whoami.(ipv4|ipv6|ds).akahelp.net IN TXT`

とある (ECS に対応した) 公開 DNS

```
$ dig @X.X.X.X whoami.ipv4.akahelp.net TXT +short
"ns" "198.51.100.1" ← 権威ネームサーバーへのクエリ送信元（つまりフルリゾルバー）のアドレス
"ecs" "192.0.2.0/24/0" ← 権威ネームサーバーが観測した ECS。 /24 であることが分かる
"ip" "192.0.2.14" ← ECS の範囲内のアドレスがランダムで返る（権威ネームサーバーには本当のIPアドレスは分からない）
$
```

とある (ECS に対応した) フルリゾルバー

```
$ dig @Y.Y.Y.Y whoami.ipv4.akahelp.net TXT +short
"ns" "203.0.113.1"
"ecs" "192.0.2.0/24/0"
"ip" "203.0.113.1" ← "ns" と同じ
$
```

アカマイの権威ネームサーバーは後者の ECS を受け取り解釈しているが、
使用していないことが分かる

調査に役立つ QNAME (続き) 権威ネームサーバーに直接クエリしてみる

```
$ dig @Z.Z.Z.Z whoami.ipv4.akahelp.net TXT +subnet=10.0.0.0/8 +norec +noall +comments +answer

; <<>> DiG 9.14.0 <<>> @Z.Z.Z.Z whoami.ipv4.akahelp.net TXT +subnet=10.0.0.0/8 +norec +noall +comments
+answer
;  (1 server found)
;; global options: +cmd
;; Got answer:
;; ->>HEADER<<- opcode: QUERY, status: NOERROR, id: 61295
;; flags: qr aa; QUERY: 1, ANSWER: 3, AUTHORITY: 0, ADDITIONAL: 1

;; OPT PSEUDOSECTION:
; EDNS: version: 0, flags:; udp: 4096 ← OPT 擬似 RR に ECS オプションデータが載っていない
;; ANSWER SECTION:                        = フルリゾルバーはサブネットを限定しないでキャッシュするべき（SCOPEが/0と同等）
whoami.ipv4.akahelp.net. 20      IN      TXT     "ns" "192.0.2.101"
whoami.ipv4.akahelp.net. 20      IN      TXT     "ecs" "10.0.0.0/8/0"
whoami.ipv4.akahelp.net. 20      IN      TXT     "ip" "192.0.2.101" ← "ns" と同じ。
$                                                                    ECS を無視していることを示す（前頁）
```


Akamai

調査に役立つ QNAME (続き 2) IPv6…

NS に A と AAAA があった際、再帰リゾルバが権威ネームサーバーにアクセスするソースは、再帰リゾルバに依存。

```
$ dig @X.X.X.X whoami.ds.akahelp.net TXT +short
"ns" "2001:DB8:XXXX::104" ← IPv4だったりIPv6 だったり
"ecs" "192.0.2.0/24/0" ← ECSはクライアントが再帰リゾルバに問い合わせた時のソースIPアドレスを元にしている
"ip" "192.0.2.14" ← ECSの範囲内のアドレスがランダムで返る（権威ネームサーバーには本当のIPアドレスは分からない）
$
```

```
$ $ dig @W.W.W.W whoami.ipv6.akahelp.net TXT +noall +answer +comment

; <<>> DiG 9.14.0 <<>> @Z.Z.Z.Z whoami.ipv6.akahelp.net TXT +noall +answer +comment
;  (1 server found)
;; global options: +cmd
;; Got answer:
;; ->>HEADER<<- opcode: QUERY, status: SERVFAIL, id: 43045 ← この再帰リゾルバはIPv4でしかクエリを出さない？
;; flags: qr rd ra; QUERY: 1, ANSWER: 0, AUTHORITY: 0, ADDITIONAL: 1
 (略)
```


Akamai
Experience the Edge