BGPによるドメイン間経路制御の現状と将来：
障害事例と対策

2017.11.09

インターネット技術第163委員会研究会 (ITRC meet42)
NTTコミュニケーションズ 西塚要

自己紹介：

- 西塚要(にしづか かなめ)

- 2006年 NTTコミュニケーションズ入社
- OCNアクセス系ネットワークの設計に従事した後、大規模ISPの運用サービスを担当。現在は研究開発組織にて、トラフィック分析などISPの課題に関する研究開発に従事。
- メインフィールド
  - トラフィック分析
  - DDoS対策
  - IPv4枯渇対策関連技術

- 社外活動
  - IETF標準化 DOTS WG
  - JPNIC「IPv6教育専門家チーム」

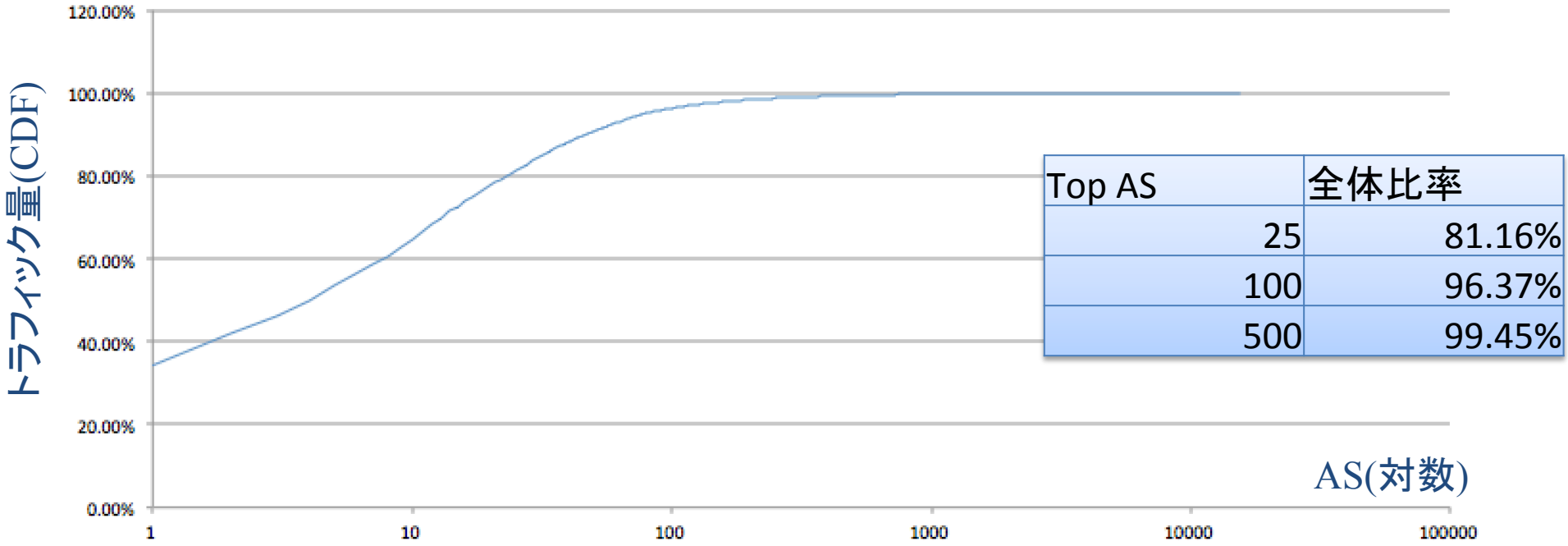
この発表では：

- インターネットを流れるトラフィックの全体的な傾向がわかる(2016-2017)
- ISPの視点で、BGPによってトラフィックがどのような目的で制御されているかがわかる
- CDNによる、BGPを無視した配信のISP視点でのインパクトがわかる
- 8/25に発生したRoute Leak事象の背景がわかる

トラフィック傾向

・複数ISP のフロー情報により算出したトラフィック長期傾向を分析
・送信元 IPアドレスをAS番号に変換→ダウンロードトラフィックがどのオリジンASから流れてくるかを集計
対象期間：2016年9月～2017年8月

ダウンロードトラフィックの寡占化



Top AS	全体比率
25	81.16%
100	96.37%
500	99.45%

- 上位 100 ASで全トラフィックの 約96%を占める
  - 100AS程度みれば、トラフィックのおおよその見積もりには十分

増加傾向のコンテンツ

- シェア = 対象ASからのトラフィック / 全体トラフィック
- 成長幅 = シェア上昇幅(1年での成長の傾き)

rank	asname	yearlyshare	yearlygrowth
1	15169:GOOGLE - Google Inc.	34.24%	4.78%
12	2914:NTT-COMMUNICATIONS-2914 - NTT America	1.75%	2.72%
4	16509:AMAZON-02 - Amazon.com	3.81%	2.36%
9	8068:MICROSOFT-CORP-MSN-AS-BLOCK - Microsoft Corporation	2.10%	2.16%
10	20940:AKAMAI-ASN1	2.05%	1.68%
11	32934:FACEBOOK - Facebook	1.79%	1.39%
14	13414:TWITTER - Twitter Inc.	1.49%	1.23%
23	2906:AS-SSI - Netflix Streaming Services Inc.	0.71%	0.82%
39	3356:LEVEL3 - Level 3 Communications	0.32%	0.77%
37	32590:VALVE-CORPORATION - Valve Corporation	0.39%	0.50%
33	13335:CLOUDFLARENET - CloudFlare	0.46%	0.32%
41	46489:JUSTINTV - Twitch Interactive Inc.	0.30%	0.27%
31	9009:M247	0.54%	0.23%
77	49981:WORLDSTREAM	0.10%	0.21%
5	22822:LLNW - Limelight Networks	3.49%	0.17%
8	23816:YAHOO Yahoo Japan Corporation	2.12%	0.17%
7	38634:DWANGO DWANGO Co.	2.33%	0.16%
27	3549:LVLT-3549 - Level 3 Communications	0.59%	0.15%
57	58209:KPTNETWORK-AS	0.18%	0.15%
59	60781:LEASEWEB-NL Netherlands	0.17%	0.14%
58	54113:FASTLY - Fastly	0.18%	0.14%
29	16276:OVH	0.56%	0.13%
25	2497:IIJ Internet Initiative Japan Inc.	0.66%	0.11%
69	29789:REFLECTED - Reflected Networks	0.14%	0.11%

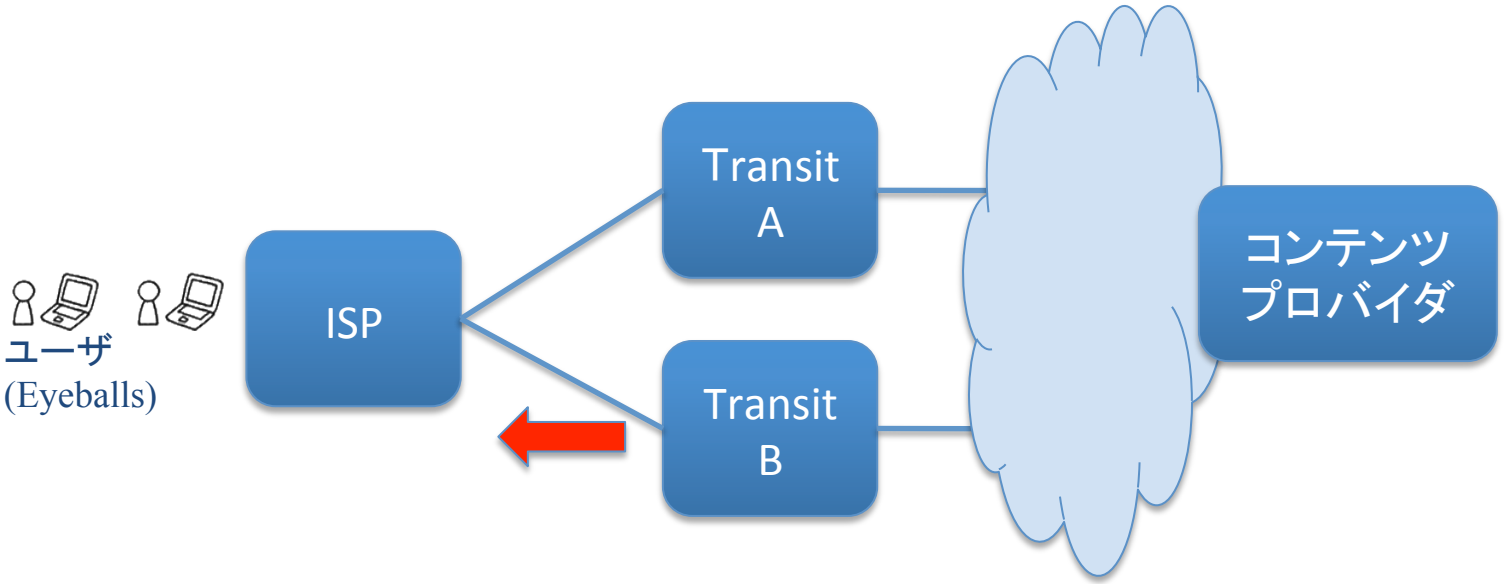
知見(2016年~2017年)

- 上位コンテンツによる寡占化が進む
- Google トラフィックのさらなる伸長
  - シェア・成長幅ともに最大
- GCP, AWS, Azureなどのクラウドサービスのシェアの伸長
- SNS系も順調にシェア増
  - コンテンツのリッチ化
- 動画配信サイトも堅調にシェアを伸長
  - ゲーム配信系が新興
- 新興CDN事業者が見え始めてきた
- 国内ISPからのトラフィックのシェアが減少

トランジット：引きの強さと力学

引きの強さ

- 引きの強さ: あるISPが複数事業者でマルチホームしているときに、下りトラフィックがどちらのトランジットから流れてくるかという強さのこと

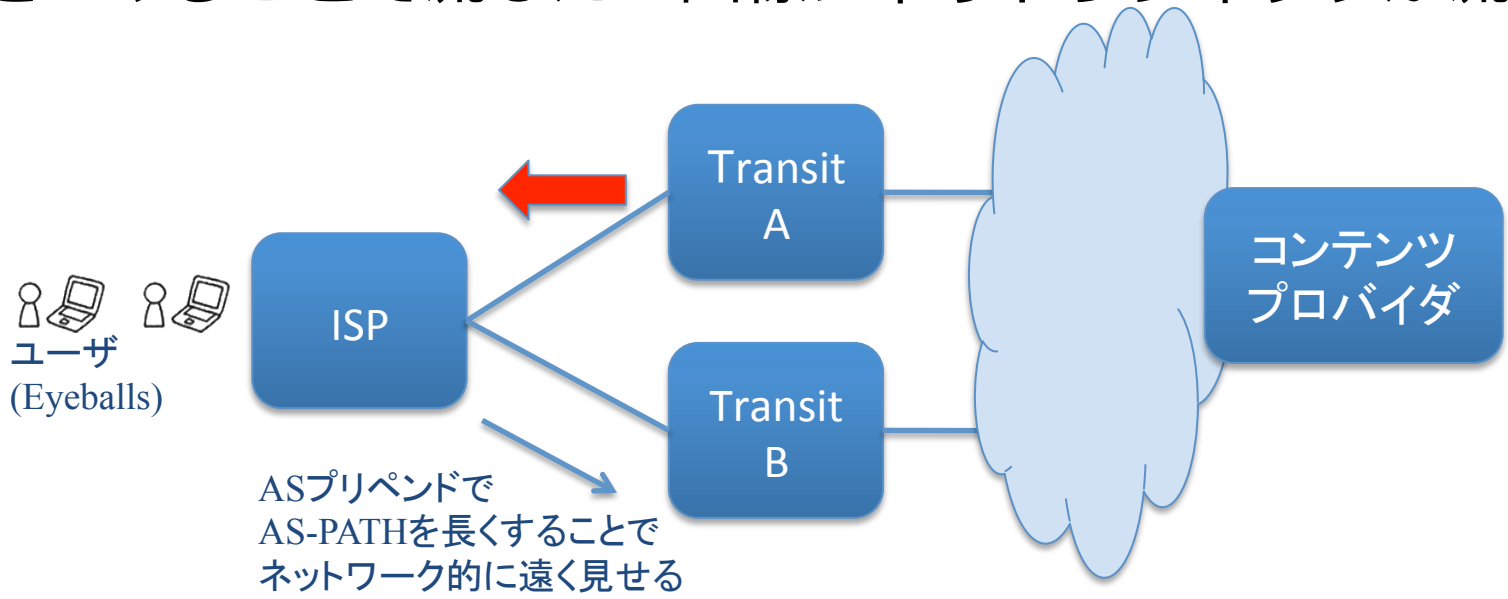


コミット

- トラフィックを何Gbpsまで流してよいのか、というトランジット事業者との契約上の合意
- 10Gの回線をひいたとて、従量制の料金体系(メガ単価: $/Mbps/monthの価格)が基本
- しかし、企業の会計上は従量制の料金体系は処理しにくいので、何Gbpsまでなら固定でいくら、という契約を行う
- 例えば、3Gbpsまで100万円/月というのを5Gbpsまで100万円/月にすると、収入は変わらないが実質の値下げになる
- 顧客ISPは、コミットが増えた分、その回線に流したいが、はたしてどのように流すのか？というのがキーポイント

AS-PATH による制御(BGP)

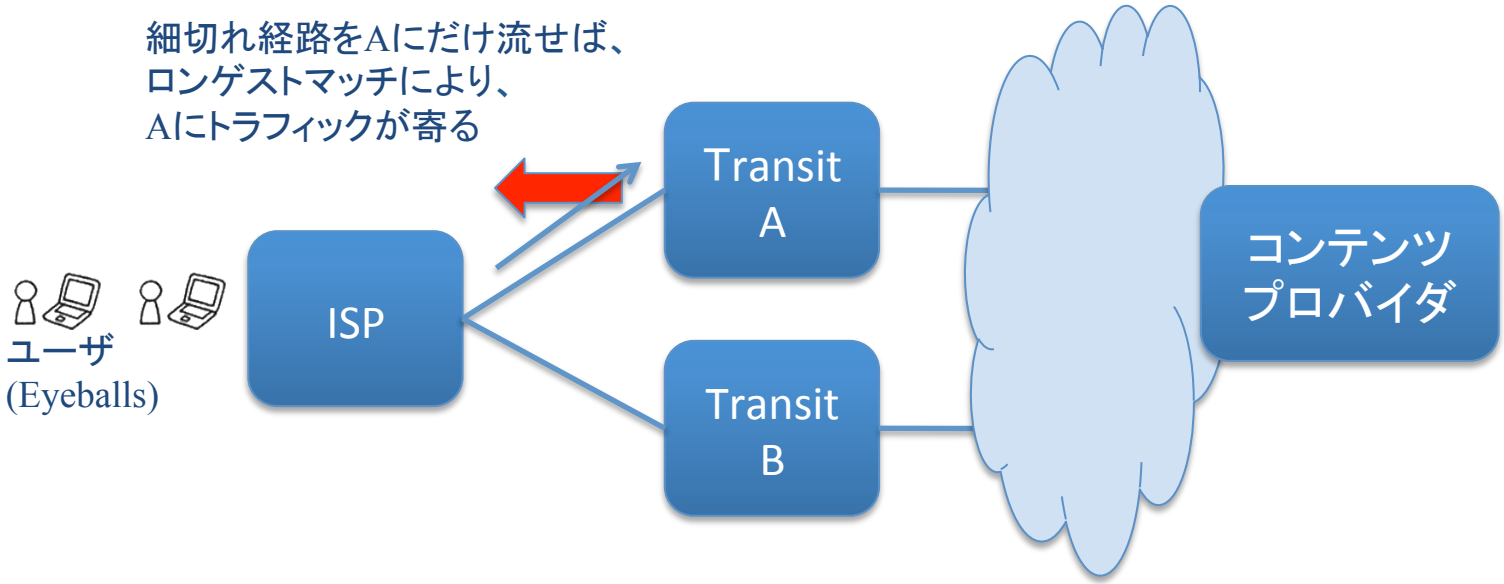
- 楽観的なトラフィック制御では、AS-PATHの長さで制御されるので、流したくないトランジットピアにAS Prependをつけることで流したい回線に下りトラフィックが流れる



- 実際はAS-PATH長では制御されていないので、ほとんどよらない(後述)

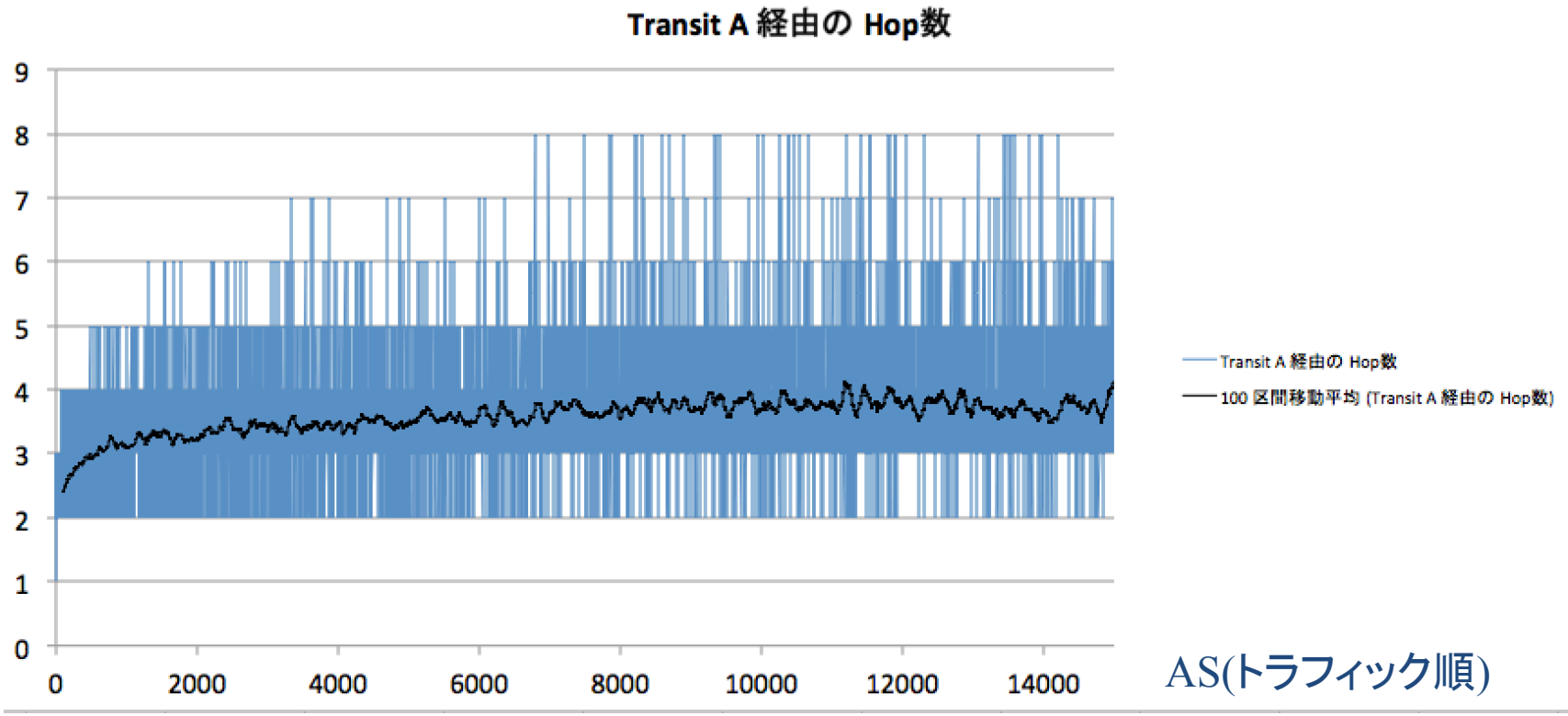
ロンゲストマッチ による制御

- longer 経路を流せば、ほぼ必ず寄ることは期待できる
  - ただし、prefixの分割は禁断の手法
  - 運用の複雑化、レイテンシの増加、トラブルの元、フルルートの増加



コンテンツへのホップ数

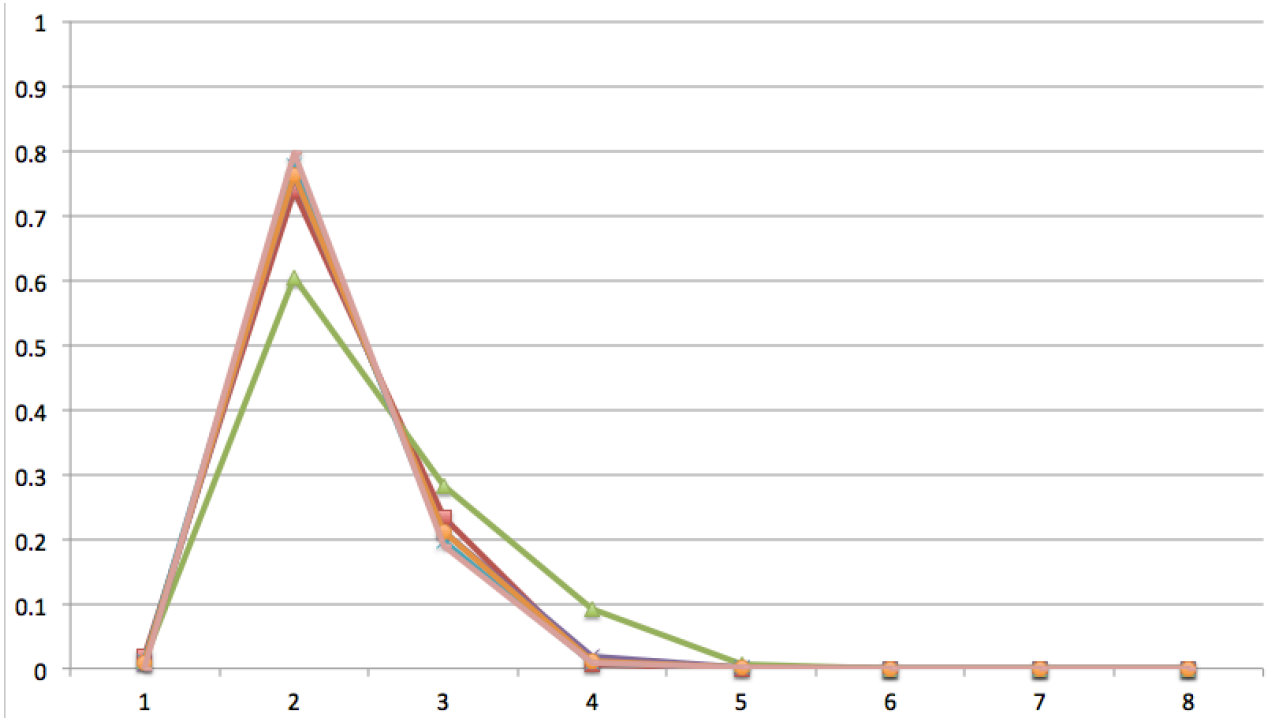
- 国内主要トランジット事業者(6社)について、主要ASに何ホップで到達できるかをBGP経路から算出



- トラフィックの多いASほどホップ数は近い(世の中のピアリング戦略が有効に働いている)

凝縮するインターネット

- ホップ数とトラフィックシェア



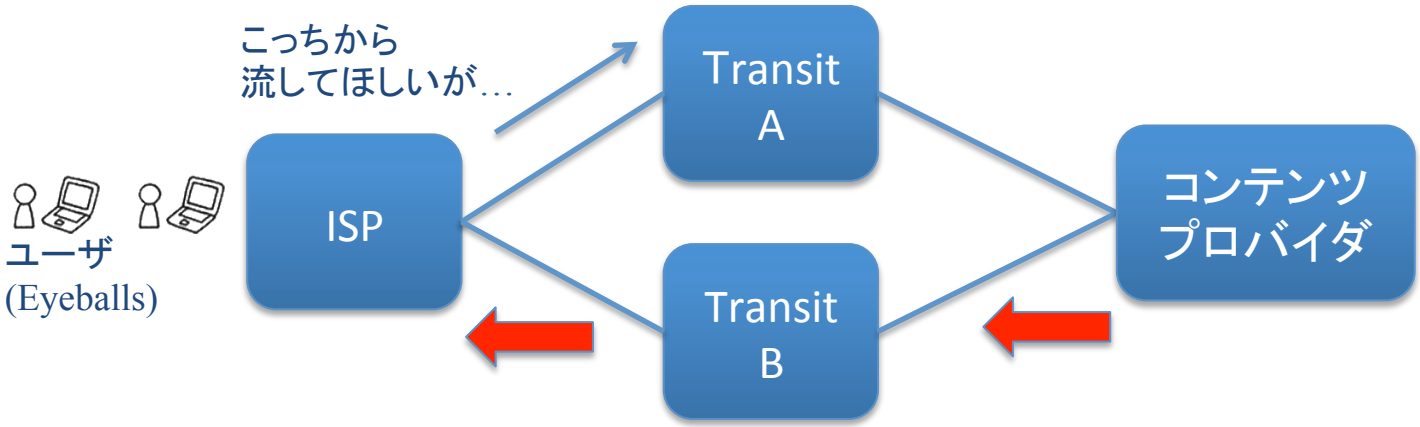
トラフィックシェアで
重みづけした
平均ホップ数

Transit A	2.22
Transit B	2.46
Transit C	2.23
Transit D	2.20
Transit E	2.23
Transit F	2.20

- 国内でトランジットを購入すると、6～8割のトラフィックは、2ホップで到達

ISPのトラフィックエンジニアリングは、効かない

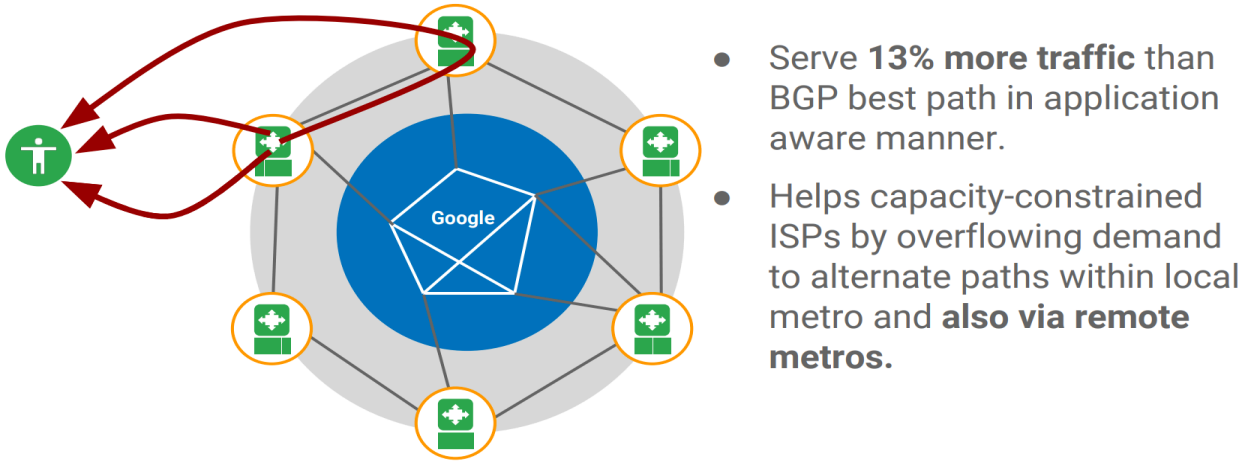
- 主要コンテンツはほとんど2ホップ先



- コンテンツ側に、特定のトランジットに流すインセンティブがあると、ISP側のトラフィックエンジニアリングは効かない
  - コンテンツ事業者にとっての金銭的インセンティブ
  - アプリケーションアウェアルーティング(cf. Google Espresso)

- 上記のシチュエーションでは、ASプリペンドによる操作は効かない
  - Transit B にとって、ISPは顧客なので、AS-PATH長よりもLocal Preference値 (通常 customer > peer) が優先される

Google のトラフィックエンジニアリング(TE)およびEspresso



- トラフィックエンジニアリングの入力
  - 1) BGP経路
  - 2) ユーザの利用帯域と品質
  - 3) ピアのリンク使用率
  - と、コストなどのハイレベルなポリシ(論文では強調していないが)
- Googleにとっての効用:
  - グローバルトラフィック最適化:ピアのリンク使用率の向上
  - Application-aware TE: End-to-End のQoEの向上
  - コストダウン

Espressoにおける非ベストパス(Overflow)

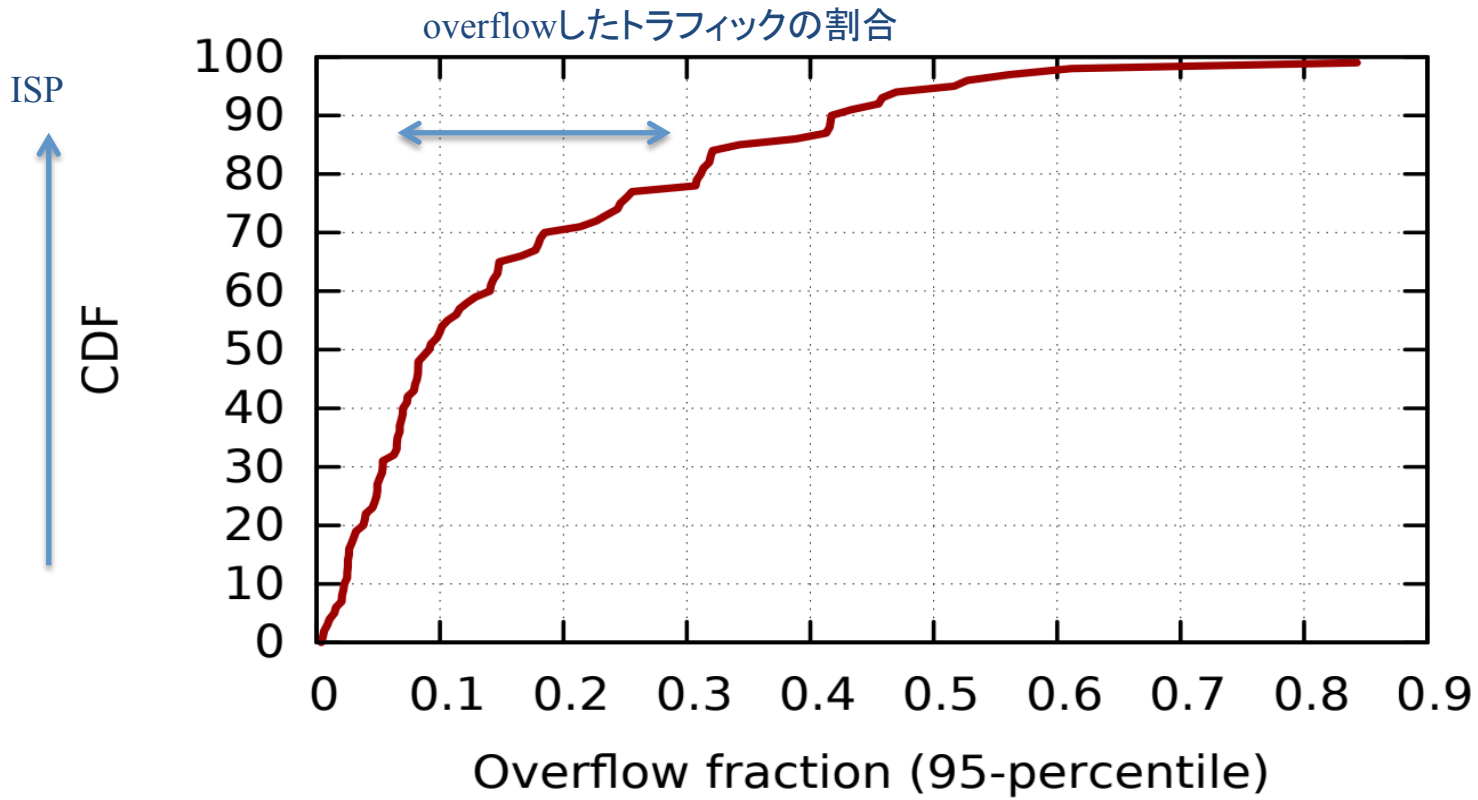


Figure 7: CDF of overflow (over a day) managed by GC for top 100 client ISPs. X axis shows the fraction of total traffic to the ISP that is served from a non best location.

- 50%以上のISPで、10%以上のトラフィックが、他に回ったほうがいいとなった
- ごく少数のISPでは、ほとんどのトラフィックがOverflowした

ISP視点では、迷惑な話

8/25 経路漏れ(Route Leak)事象

事象の概要：何が起きたのか？

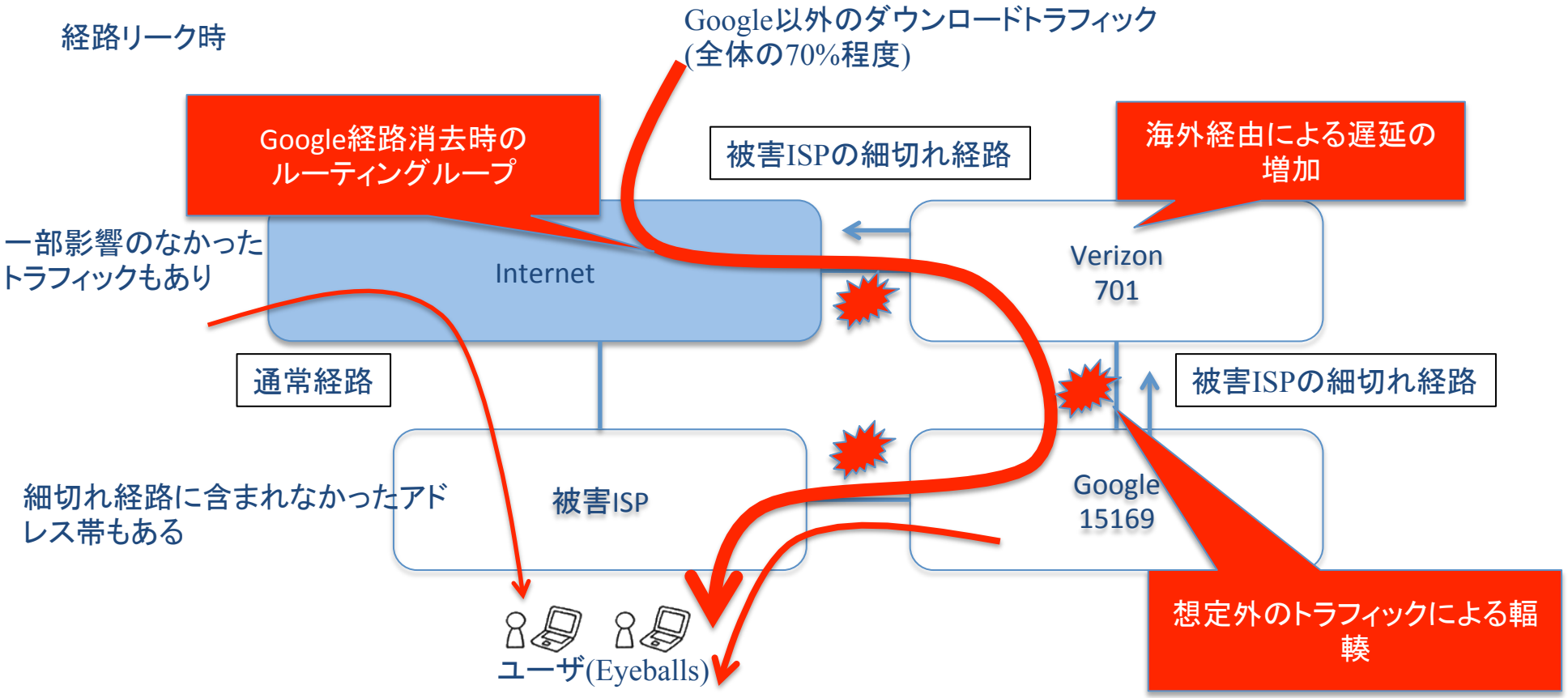
- Route Leak from Google
  - Googleからベライゾン経由の経路漏れ
  - Googleがピアから受け取った経路を誤って広報してしまった
  - ベライゾンはそれを受け取ってしまった
  - 経路は、ISPがインターネットに通常広報している経路よりも細切れだった

- Recommended Articles
  - 8月25日に発生した大規模通信障害をまとめてみた
  - OCNはとばっちり、米グーグルによる大規模ネット障害の真相
  - パブリックデータから経路リークを探る
  - 08/25の通信障害概説(pdf)
  - Large BGP Leak by Google Disrupts Internet in Japan
  - BGP leak causing Internet outages in Japan and beyond

先日も…

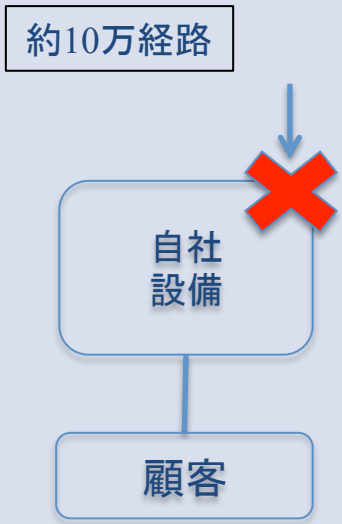
- Route Leak from Level3 (11/6)
  - 北米のインターネットが約90分間不安定に
  - 内部のComcast細切れ経路(18AS/3000 prefix)の漏洩

Peer Leak: ユーザへの影響について



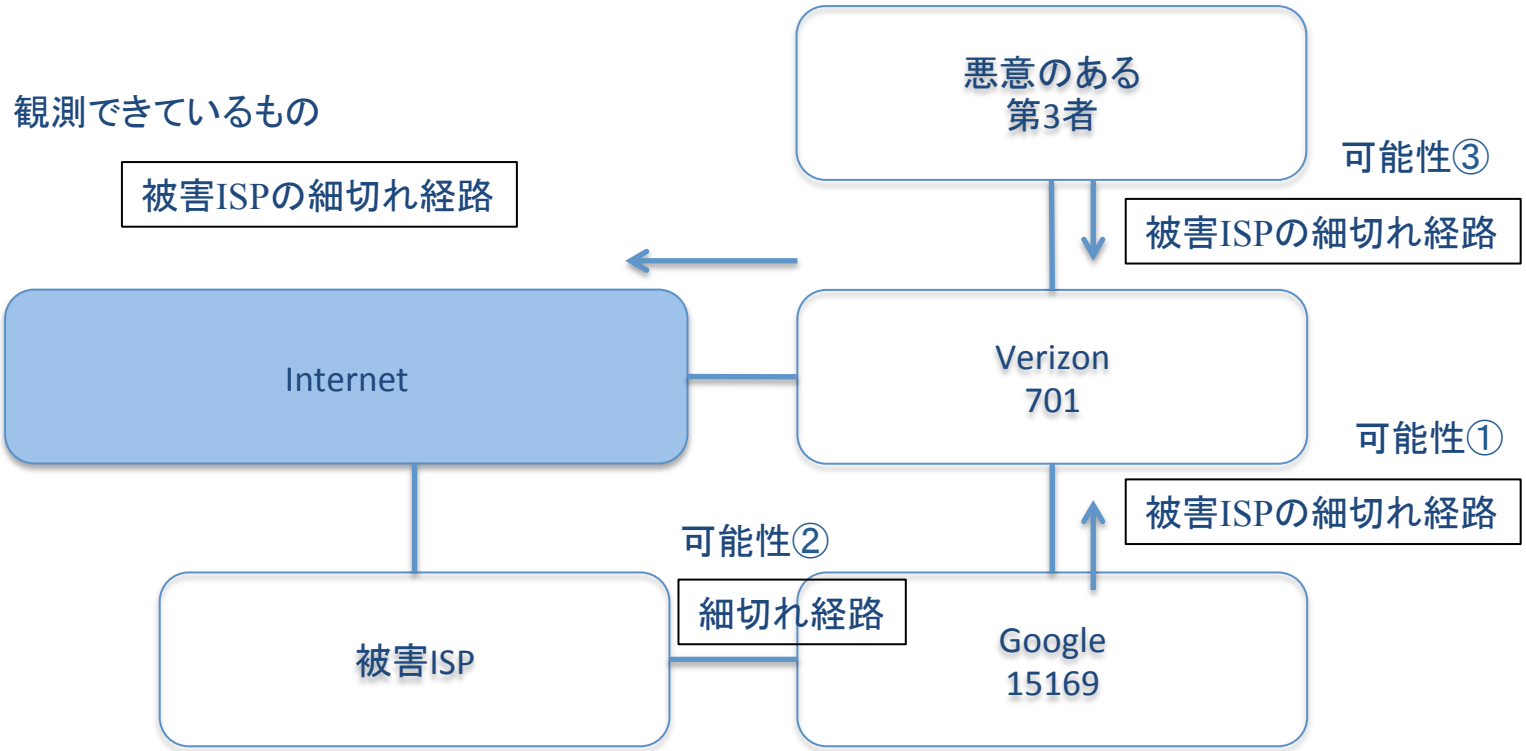
- ほとんどがVerizon-> Google 経由に
  - 影響1: 海外経由による遅延の増加
  - 影響2: トラフィックの輻輳
  - 影響3: Googleが経路を約10分後に消去した時のルーティングループ

リークの影響に加えて経路数の影響があった

リーク影響型	経路影響・自社型	経路影響・透過型
 リークされた経路のユーザ		
経路のリークによって、ユーザがつながりにくい事象 1. 海外経由による遅延の増加 2. トラフィックの輻輳 3. 経路消去時のルーティングループ(※)	トランジットから流入した経路増大により、自社設備に影響がでて、自社や顧客のサービスに影響	トランジットから流入した経路増大により、自社設備には影響がなくても、顧客のNW機器(あるいはDCなど)が経路増大で影響

※経路消去(withdraw)時のループに関しては、IRS26の資料参照

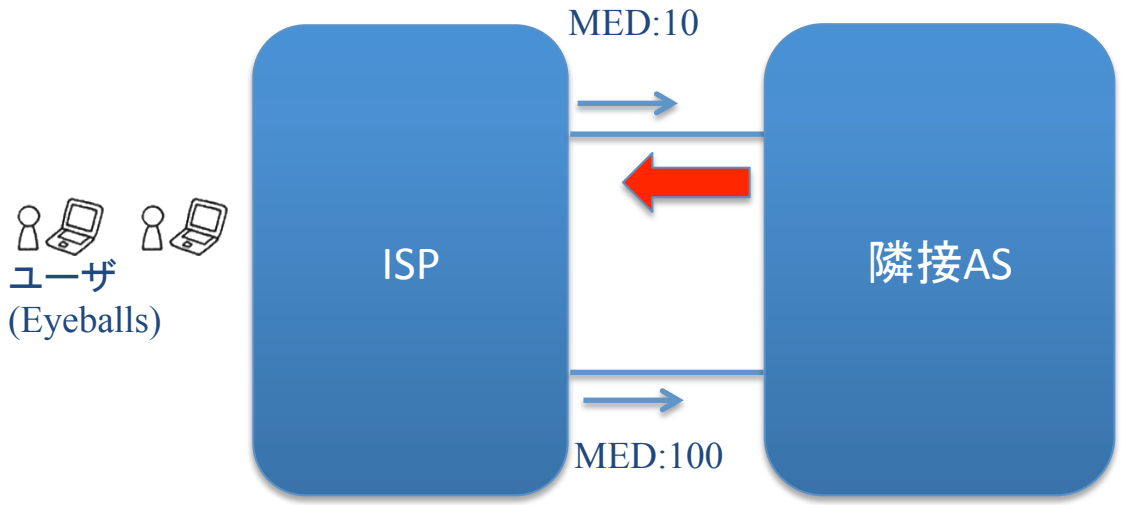
経路はどこから来たのか？



1. GoogleがISPから受け取った経路を内部的に分割していた(？)
2. ISPがGoogleに細切れ経路を何らかの理由で渡していた(？)
3. 第3者が詐称して注入した(❌)
   - Googleが自身のミスを後日認めたためこの可能性はない(BGP経路は、誰が注入したのかはわからないので、事象発生時点では、Googleのミスかどうかは同定できなかった)

隣接ASとのトラフィックエンジニアリング(BGP)

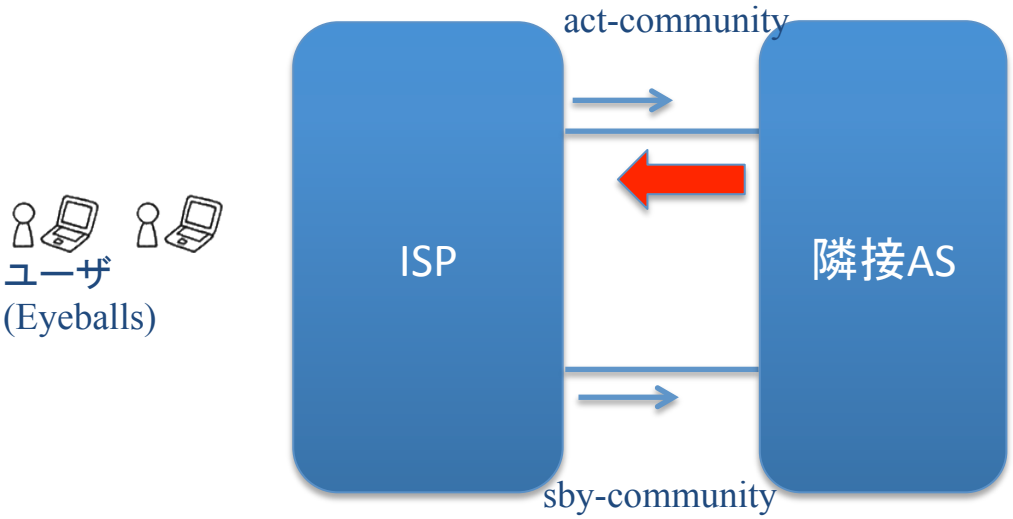
- MEDによる重み付け



- MEDによって、どのピアからトラフィックを出してほしいか、要望することができる
- MEDは受信AS側で上書きすることができる
- インターネットでは、入ってくるトラフィックは、根本的には制御できない
  - 双方に合意がなければ

隣接ASとのトラフィックエンジニアリング(BGP)

- CommunityおよびLocal Preferenceによる重み付け



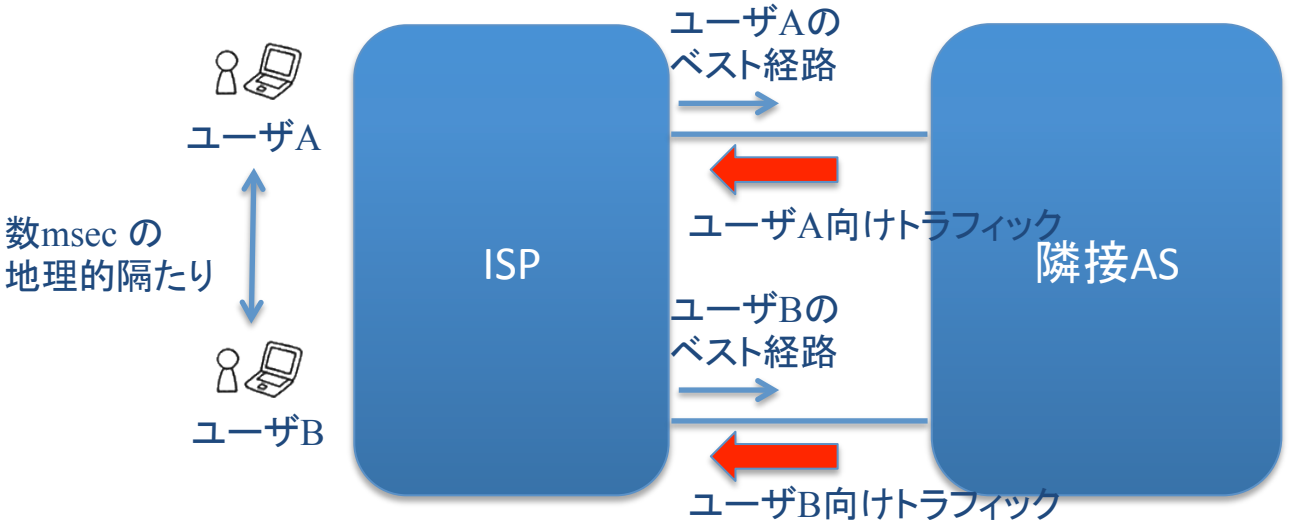
ntt.netの例

Community	Local-pref	Description
(default)	120	customer
65520:nnnn	50	only within country origin <nnnn> (see country origin list below)
65530:nnnn	50	only within region <nnnn> (see region list below)
2914:435	50	only beyond the connected country
2914:436	50	only beyond the connected region
2914:450	96	customer fallback
2914:460	98	peer backup
2914:470	100	peer
2914:480	110	customer backup
2914:490	120	customer default
2914:666		blackhole

- 隣接AS側が、Communityに応じたLocal Preference値の実装を公開している場合がある
- Communityによってどのピアからトラフィックを出してほしいか、要望することができる
- 双方に合意があるとみなせる

地理的広がりとプレフィックス分割

- 地理的広がりがあると、レイテンシが効いてくる



- ISPには地理的隔たりに合わせて、プレフィックスを分割して、隣接ASに経路を広告するモチベーションがある
  - 特に、隣接ASがコンテンツプロバイダで、
  - 低レイテンシが鍵となるアプリケーションである場合
- 隣接ピア向けとそれ以外で経路の細かさが異なる場合がある

Peer leaks

- 最近では、GoogleおよびLevel 3の事例
- 隣接ピアが、誤って経路を全インターネットあるいは部分に対して広告してしまうリスク
- 特定のピア”だけ”に細かい経路を流していると、リスクが高まる

Secure BGP

経路漏れに関して、BGP元来の脆弱性を補う提案がいくつかある

- BGP Prefix Origin Validation
  - Route Origin Authorizations(ROAs)を配布し検証する
  - しかし、Origin ASが正しい今回のようなハイジャックには効果がなかったと思われる
    - maxlen(maximum prefix length)は有効かもしれない
- BGP Path Validation
  - AS-PATHに対して、署名検証する
  - しかし、MITMは防げるが、今回のようなRouteLeakは防げなかったと思われる

しかし、そもそも、世界中のASが対応しないと効果がない

実装上・プロトコル上の対策提案

- [RFC8212]Default External BGP (EBGP) Route Propagation Behavior without Policies
  - 何も考えずにeBGPのピアを張ったときに、全経路を送出しようとする動作をやめる
- [I-D.]Route Leak Prevention using Roles in Update and Open messages
  - ピアを張るときに、そのピアの用途を規定する
  - “ピア経由の経路”という transitive attributeをつける

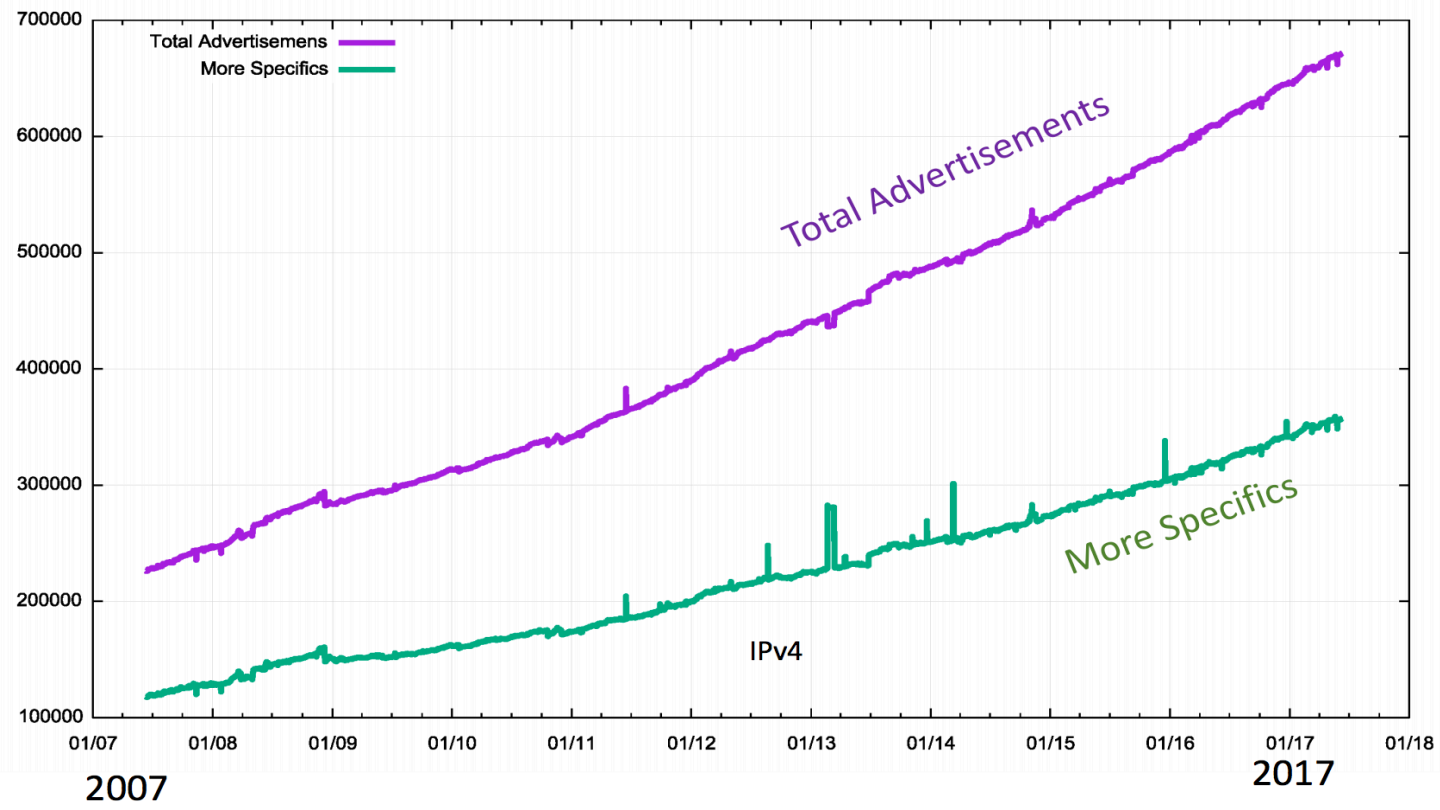
Espressoのインパクト

- Espresso以前
  - Googleに対して、プレフィックス単位での制御(トラフィックエンジニアリング)を期待して、細かい経路を渡す
- Espresso以降
  - ISP側がどのような単位で経路広告をしようが、Google側が、内部処理的に品質単位でプレフィックス分割をして、トラフィックエンジニアリングを実施する
- 今後はISPは細切れ経路を出さなくて済むのか？
  - Googleに対してはそうかもしれないが、全てのコンテンツプロバイダがGoogleと同様の仕組みを持つとは思えない

細切れされていく経路

細切れ経路(more specific)

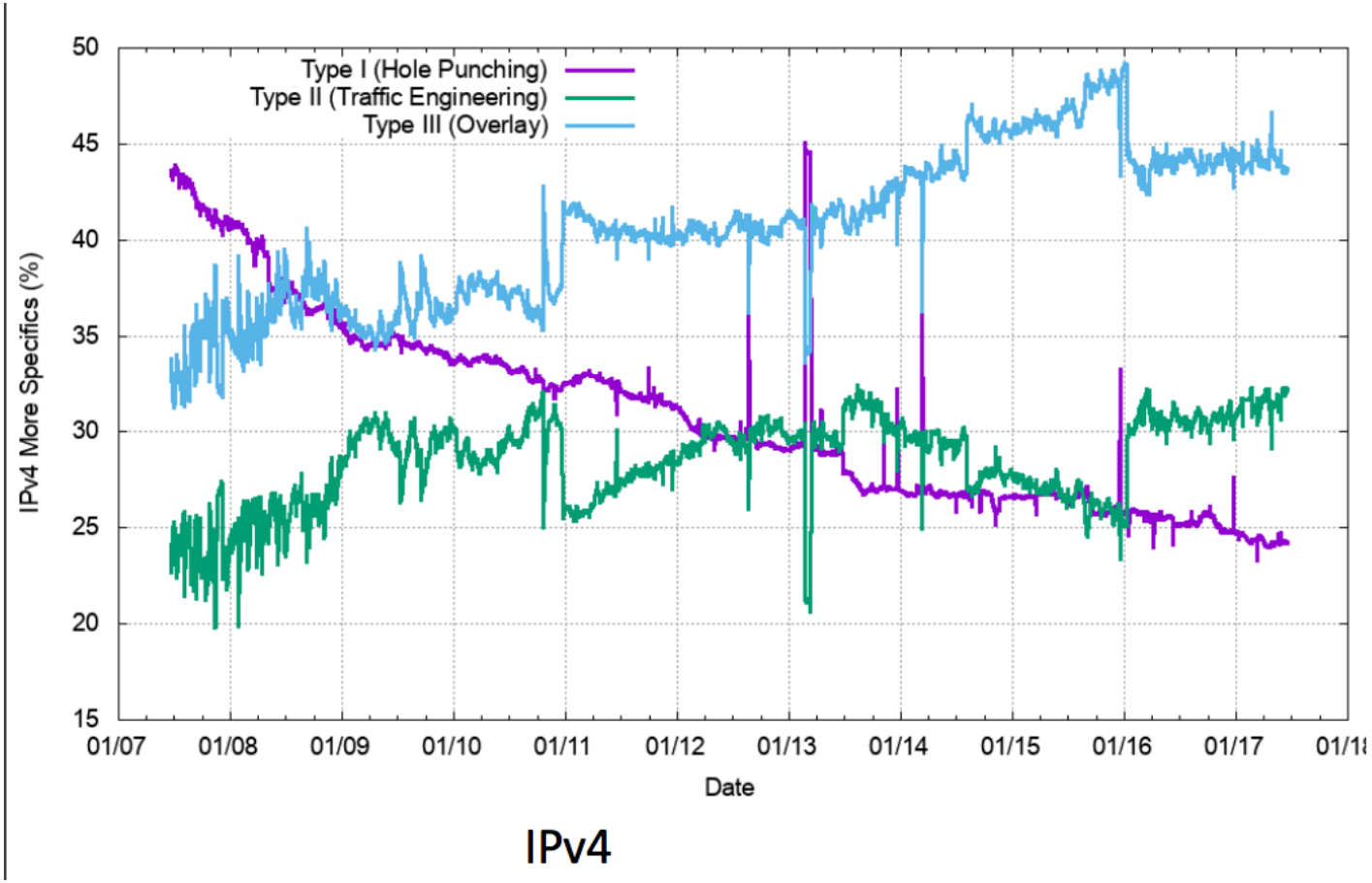
- IPv4総経路のうち約50%は、(本来不要な)細切れ経路



http://www.iepg.org/2017-07-16-ietf99/2017-07-16-bgp-more-specifics-gih.pdf

細切れ経路(more specific)

- 経年で割合が増えているわけではないが、用途が変わってきている



http://www.iepg.org/2017-07-16-ietf99/2017-07-16-bgp-more-specifics-gih.pdf

細切れ経路を出す理由

- トランジットとのトラフィックエンジニアリング
  - パンチングホール
  - マルチホーム時のトラフィック操作
- ピアとのトラフィックエンジニアリング
  - ピアへのみ広報
    - グローバルインターネットでは観測されない
    - これが漏れてしまうとシビアな経路漏れ事象になる
- 経路漏れ/経路ハイジャック対策
- DDoS対策

経路漏れ対策としての細切れ経路

- 経路漏れされてしまったら
  - 経路漏れ経路よりも細かい経路を流す
    - 同一の長さの経路でも、AS-PATH長で勝てるかもしれない
- 経路漏れされないようにするには
  - 基本無理。いかに気づくかが大事
  - 全てのASに送付する経路を一致させる
    - ピアのオペミスの影響を最小化
  - 常時最小単位に細切れする(前述のOverlay Typeの広報)

まとめ

- 昨今のRoute Leak事象の背景には、コンテンツプロバイダでのSDNによる粒度の細かい制御のために、経路を分割したいという要求がある
- BGP上にSDN制御をアドオンするという形式が引き続き続くとすると、インターネット上で観測される経路は、トラフィックエンジニアリング(DDoS対策含む)のためにさらに分割されていくと予想される