BGPと障害と運用と

Matsuzaki ‘maz’ Yoshinobu

<maz@iij.ad.jp>

2017/08/25


ENOG46 Meeting 開催のお知ら
enog.jp/archives/1679
Echigo Network Operators' Group
About ENOG
Mailing List
Meeting
Echigo-IX
お知らせ
活動記録
お知らせ
ENOG46 Meeting 開催のお知らせ
by masakazu • 2017年7月7日 • 0 Comments
開催概要
日時
2017/8/25(金) 15:00〜17:00
会場
柏崎市文化会館アルフォーレ 大会議室
（新潟県柏崎市日石町4番32号）
http://www.artforet.jp/
勉強会
“Cactiが1.xにバージョンアップしているよ” – 外山 文規(株式会社創風システム)
“MTAのヘッダーサイズによる制限について” – 外山 文規(株式会社創風システム)
“RIPE Atlas再び” – 松崎 吉伸 (株式会社インターネットイニシアティブ)
“Peering Asia 1.0 のご案内” – 安部 博文 (BBIX株式会社)
検索 …
最近の投稿
ENOG49 Meeting 開催のお知らせ
ENOG48 Meeting を開催しました
ENOG48 Meeting 開催のお知らせ
ENOG47 Meeting を開催しました
ENOG47 Meeting 開催のお知らせ

観測されている概要

- 2017/08/25 12:22JST頃
  - AS15169が他ASのIPv4経路をトランジット開始
  - 日頃流通しない細かい経路が大量に広報
    - これによりトラヒックの吸い込みが発生
  - 国内の各ASで通信障害を検知
- 2017/08/25 12:33JST頃
  - AS15169がトランジットしていた経路を削除

観測された問題のBGP経路概要

- 経路数
  - 全体で約11万経路 (日本分が約25000経路)
    - /10から/24まで幅広い経路(半数程度が/24)
    - 通常流れていない細かい経路が多かった
- AS PATHは概ね “701 15169 <本来のAS PATH>”
  - 広報元AS番号は正しそう
  - 各ASが直接AS15169と張っているBGP接続では今回の経路広報は観測されていない
- 対象AS
  - 全体で約7000 AS程度 (日本分が約89 AS)

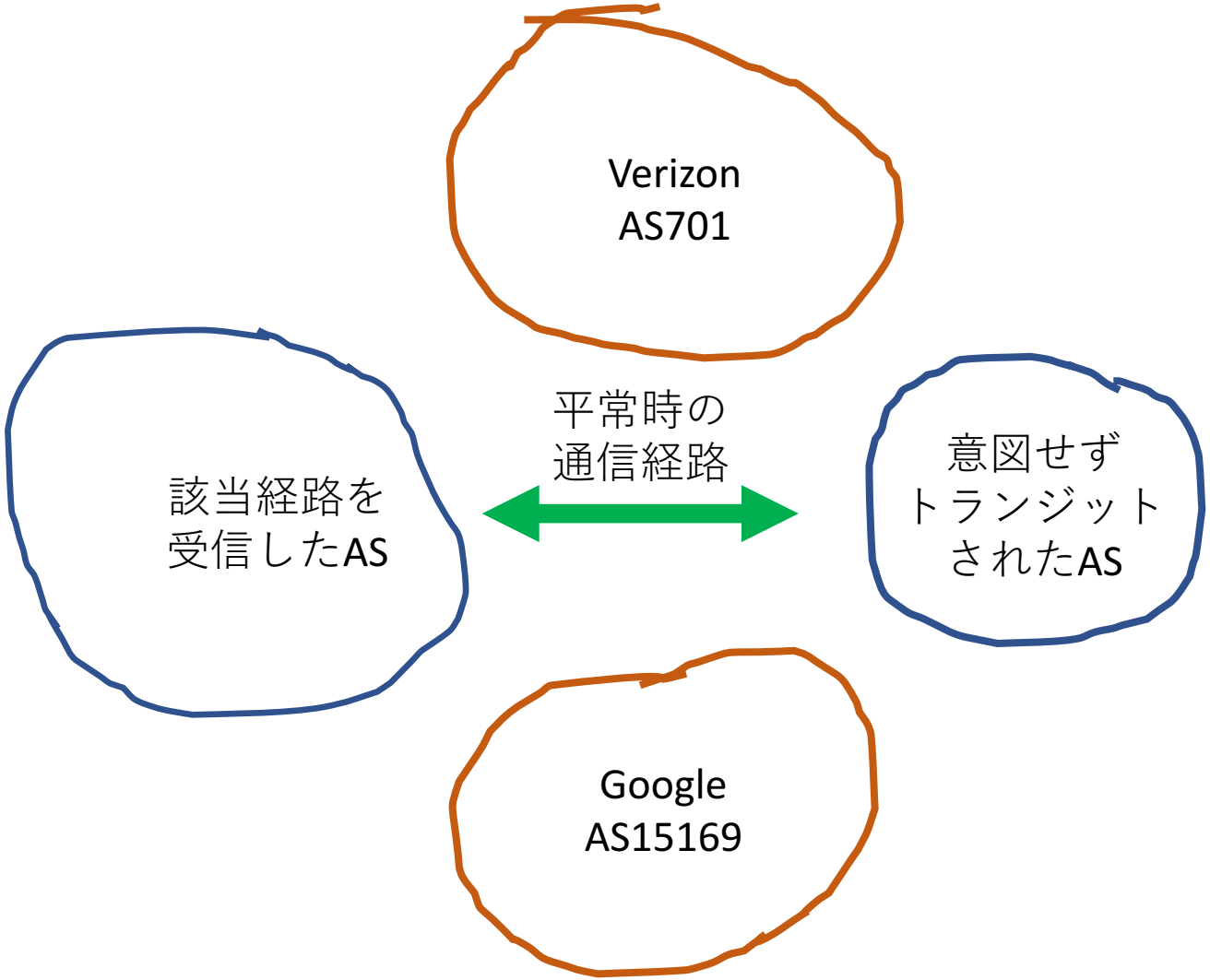
BGPは観測点によって見える情報が異なるのでご注意

トランジットされちゃったAS

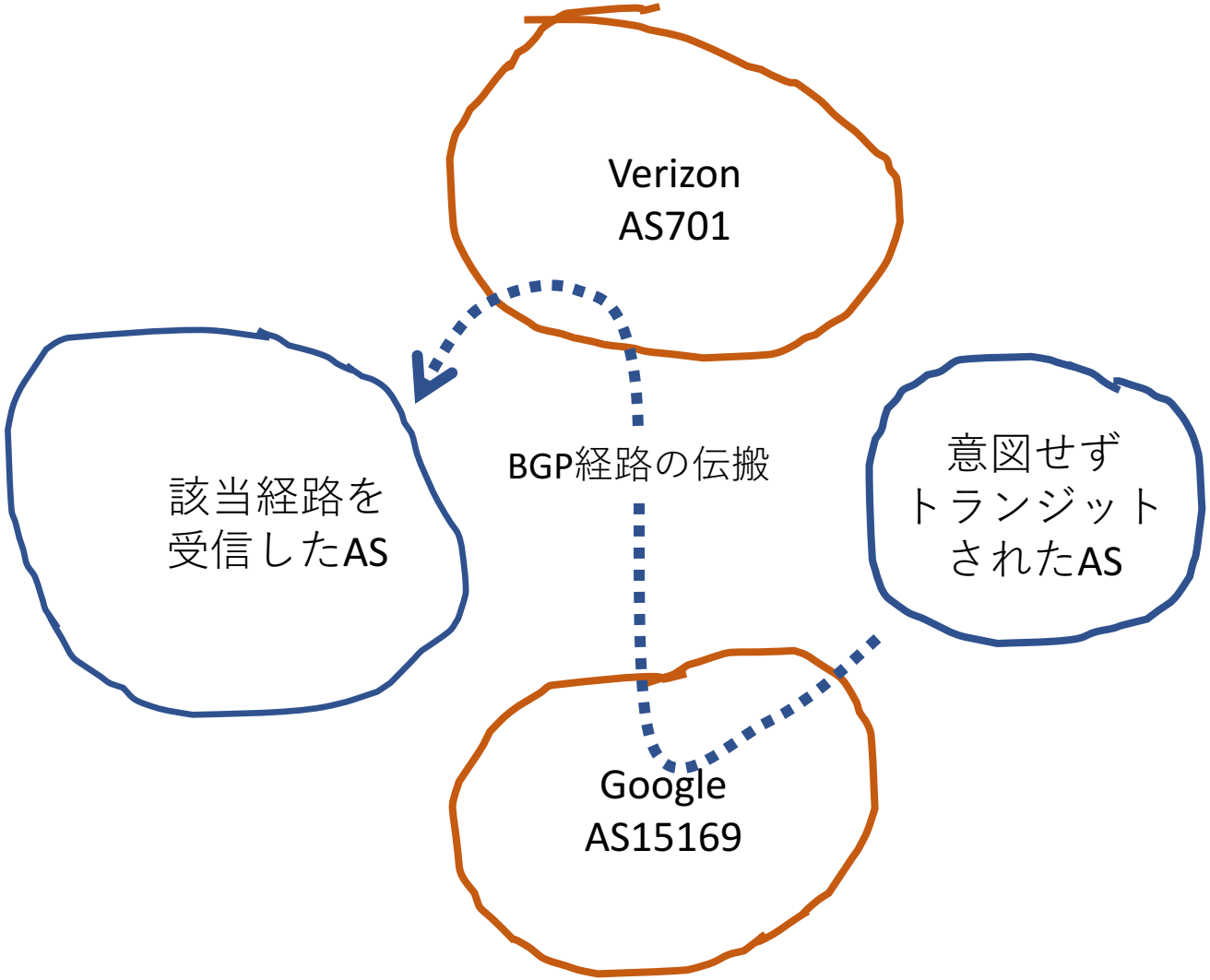
- 世界でおよそ7000 AS程度
  - 内、日本(JPNIC管轄)のものが 89 AS
- 広報されたprefix数のAS別順位
  - OCN/AS4713が大きな影響を受けている

AS番号	prefix数
4713/OCN	24381
7029/WINDSTREAM	7837
8151/UNINET	4639
9121/Turk Telecom	4606
1659/TANet	3106
9394/CTTNET	2137

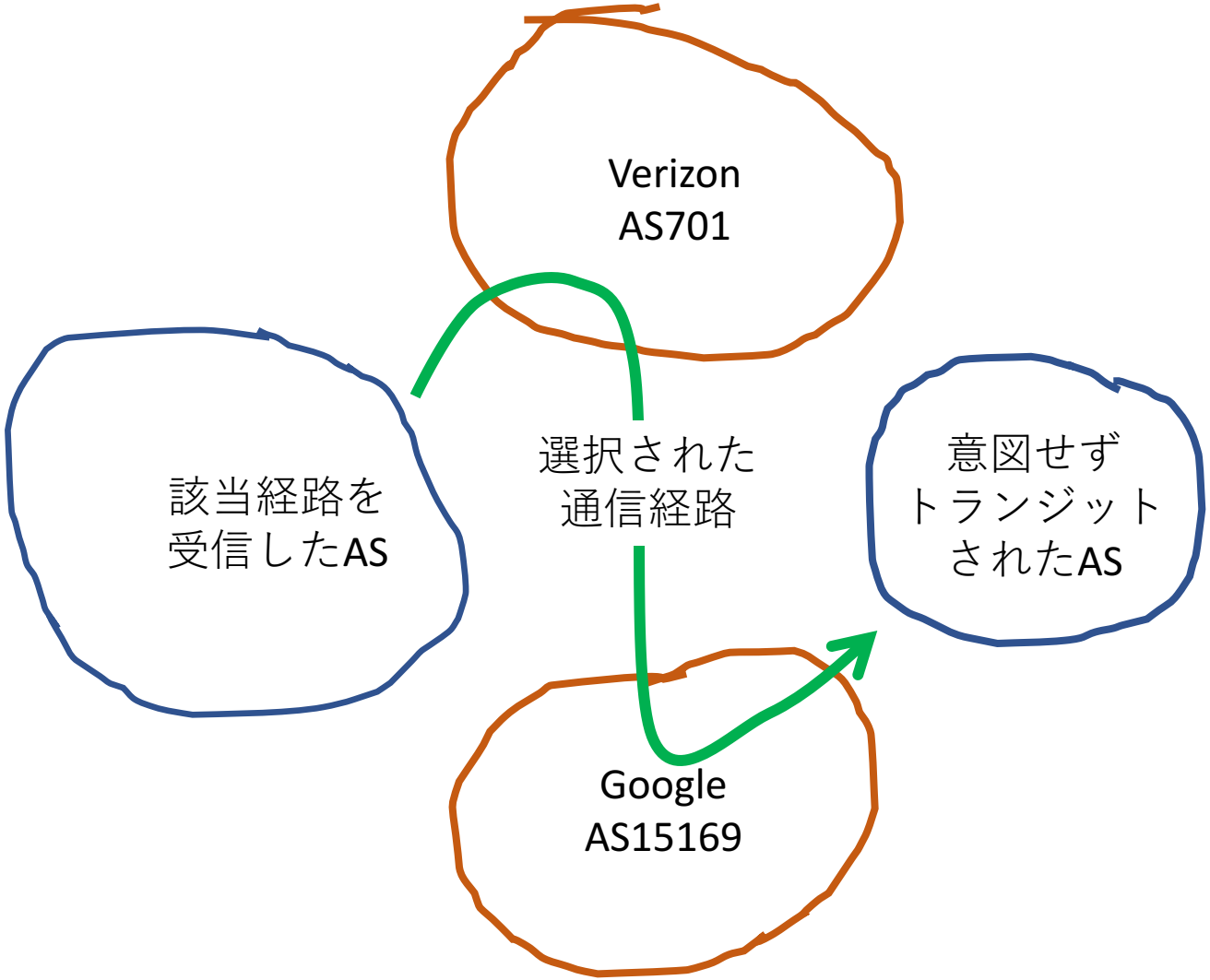
概要図１：平常時


Verizon
AS701
該当経路を
受信したAS
平常時の
通信経路
意図せず
トランジット
されたAS
Google
AS15169

概要図１：平常時


Verizon
AS701
該当経路を
受信したAS
BGP経路の伝搬
意図せず
トランジット
されたAS
Google
AS15169

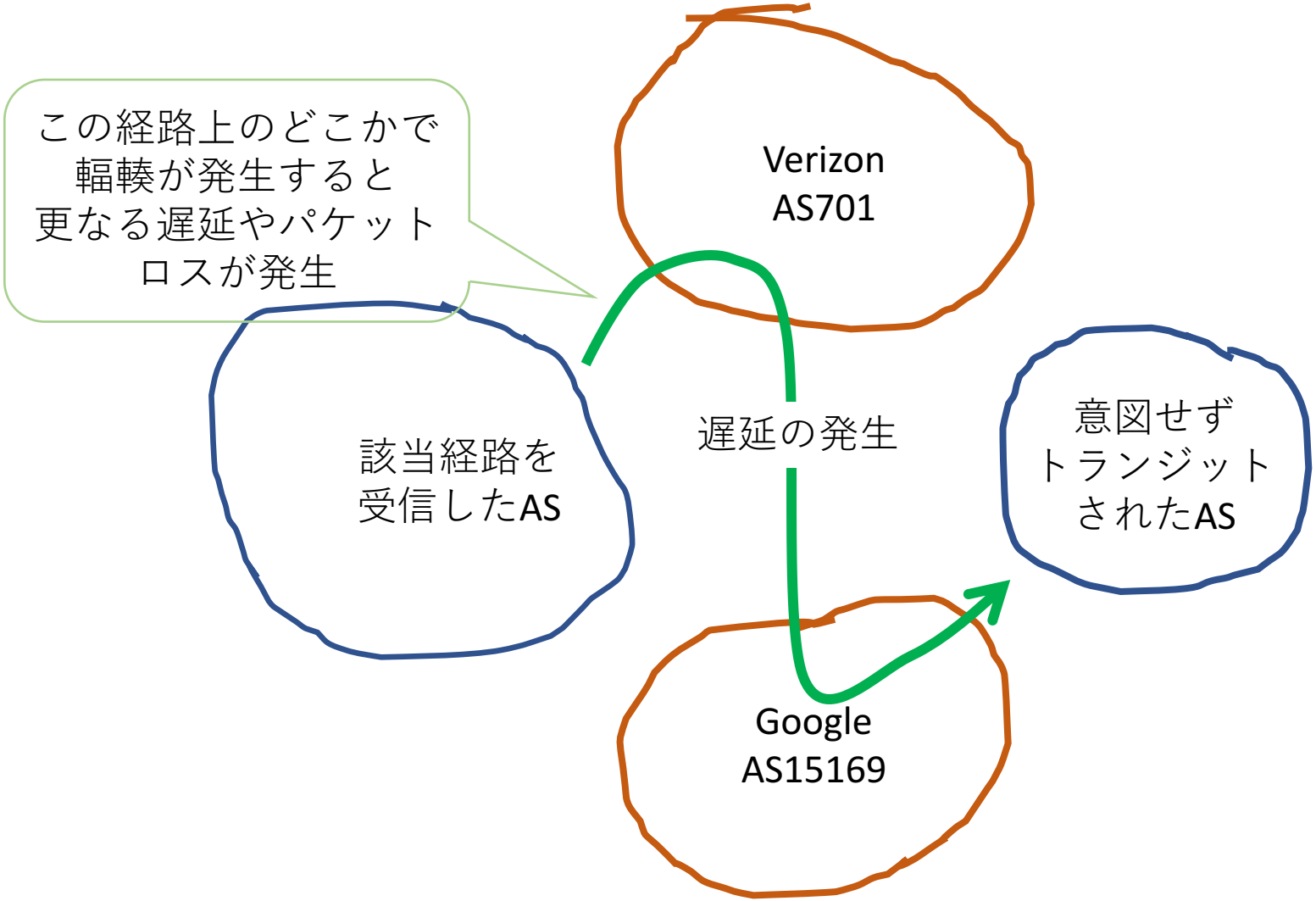
概要図１：平常時


Verizon
AS701
該当経路を
受信したAS
選択された
通信経路
意図せず
トランジット
されたAS
Google
AS15169

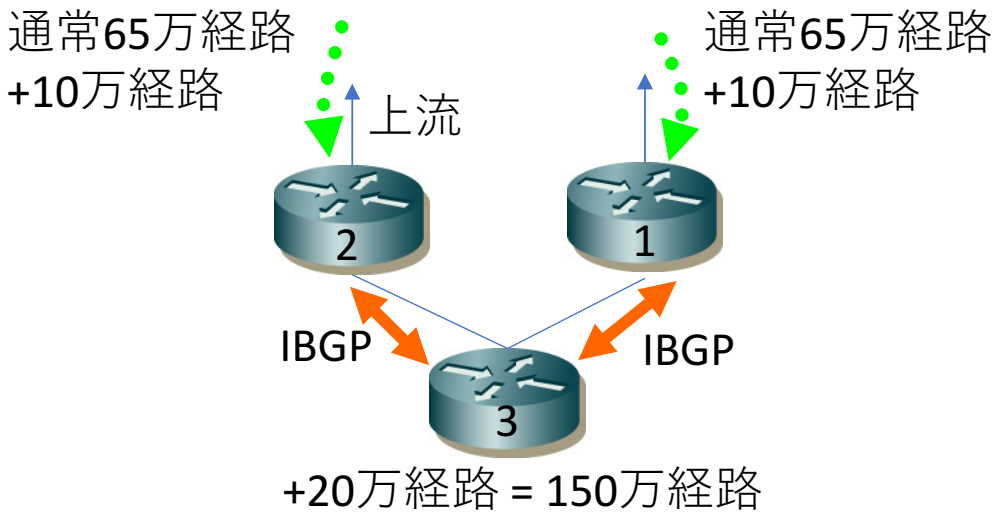
推定される障害や影響

- 広報された宛先向けの通信で遅延が発生した
  - 通信が米国経由になったため
  - 経路上に十分な帯域がない場合は輻輳の発生
- 大量の経路広報で機器が不安定になった
  - 負荷上昇でルータが不安定になった
  - RIB/FIB溢れでルータが不安定になった
  - 何らか機器のbugを踏んだ
- IXP越しの通信が迂回したかも
  - 発生条件が特殊なので、恐らく今回は発生していないのではと考えている

概要図１：平常時


この経路上のどこかで
輻輳が発生すると
更なる遅延やパケット
ロスが発生
Verizon
AS701
該当経路を
受信したAS
遅延の発生
意図せず
トランジット
されたAS
Google
AS15169

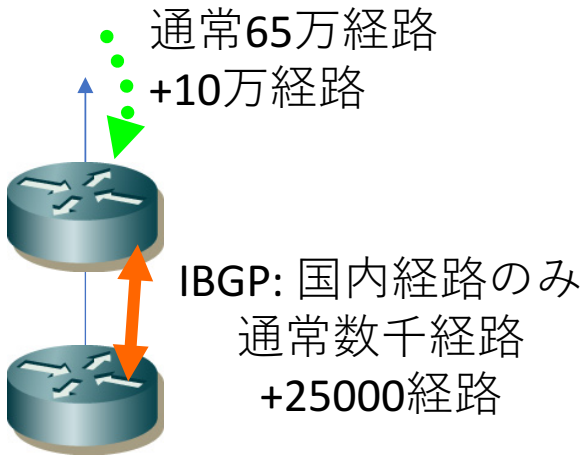
大量の経路追加



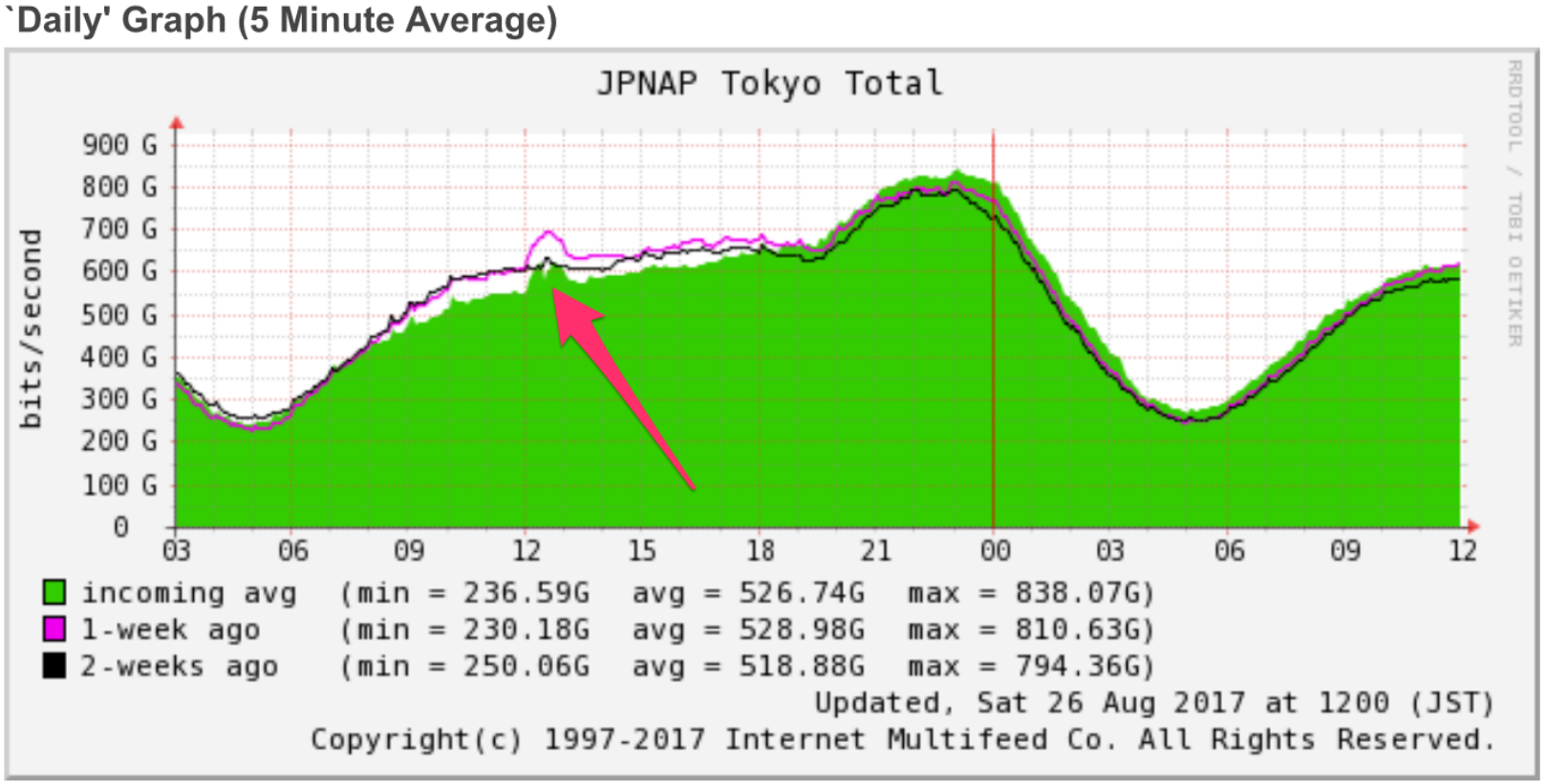
- 現状ざっくり約65万経路
  - 何もしていないと内部のRT3は65万x2で130万経路
- 今回、10万x2追加で150万経路受信していたかも
  - 構成によっては更に多い場合も

経路削減を適用してても

- 非力なルータで運用するため国内経路のみを内部ルータに渡している場合
- AS PATH(_4713_等)で国内経路を識別していた場合、追加で約25000経路
  - 構成によってはもっと多い
  - 通常時の5倍から10倍の経路数が追加された可能性がある
- これら非力なルータが過負荷になるなどの障害が発生した可能性がある



IXPでトラヒック減を観測



http://www.jpnap.net/service/traffic.html

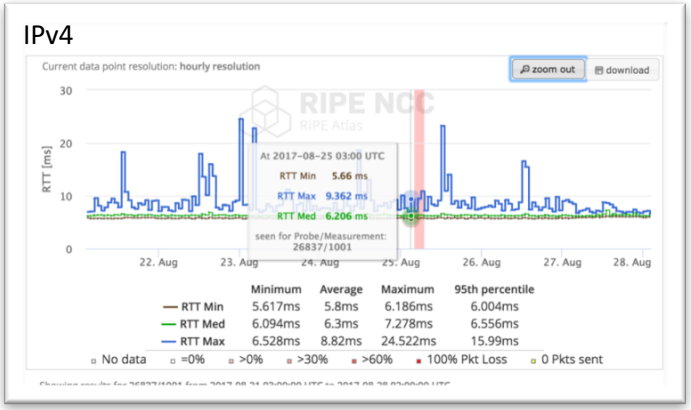
RIPE Atlas Probeで通信影響測定

- RIPE NCCのプロジェクト
  - 世界にProbeを配っていて、エンドユーザ視点での計測が可能
- AS4713のprobeを抽出し、宛先別に影響を推定
  - OCN内で通信が完結する宛先: k.root-servers.net
  - 国内で今回の影響を受けた宛先: m.root-servers.net
  - 海外で今回の影響を受けた宛先: ctr-ams02.atlas.ripe.net

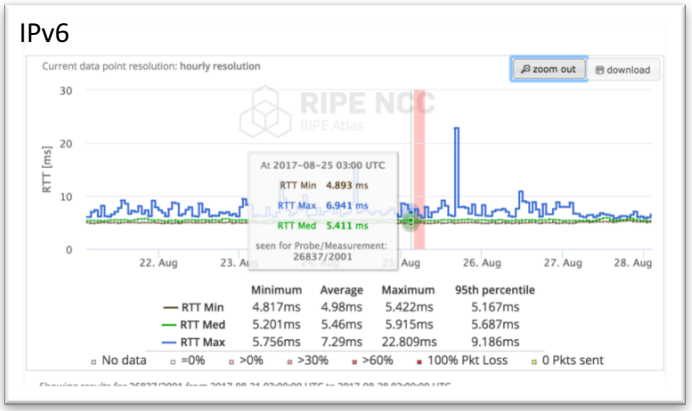
RIPE Atlasで見る：OCN内

Probe26837

IPv4

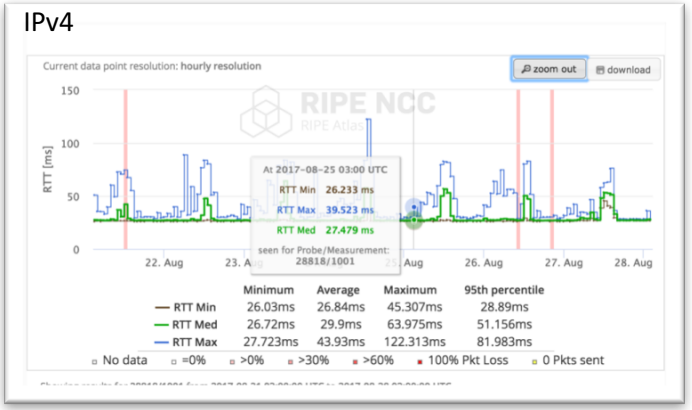


IPv6

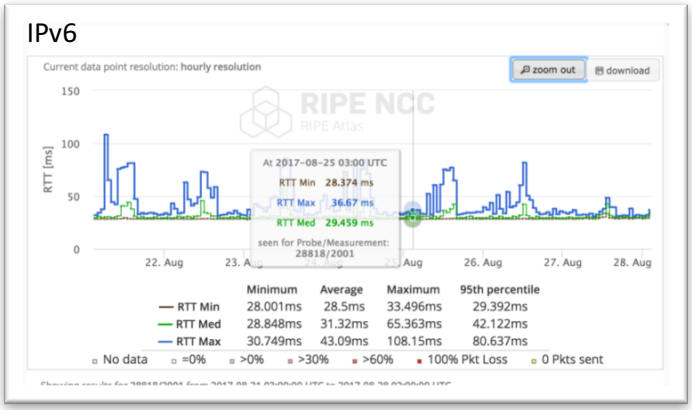


Probe28818

IPv4



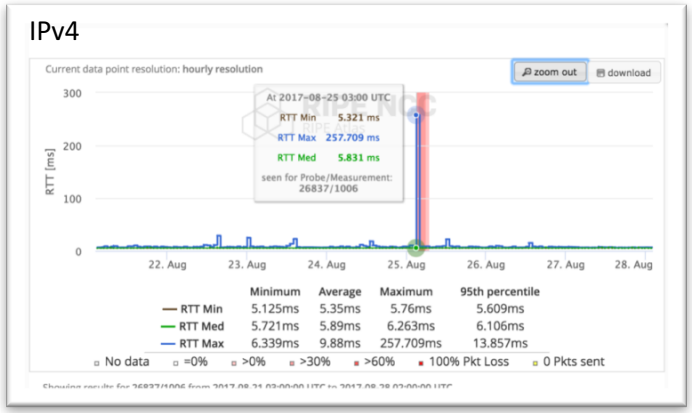
IPv6



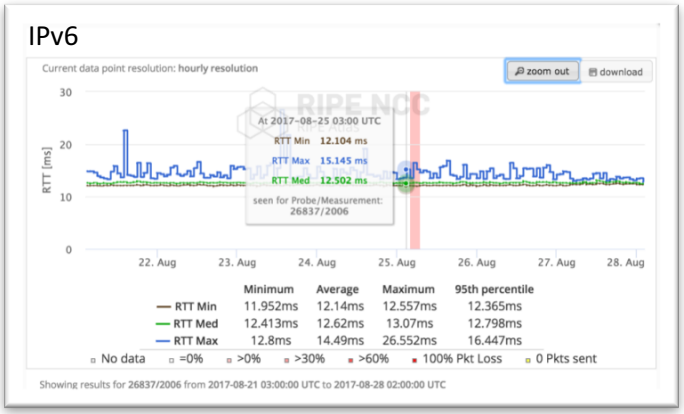
RIPE Atlasで見る: OCNと国内

Probe26837

IPv4

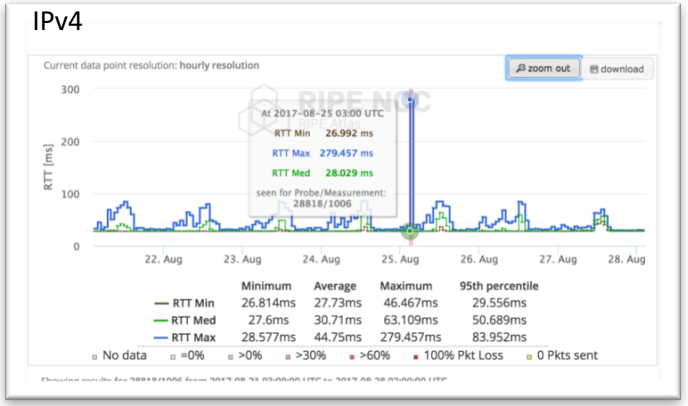


IPv6

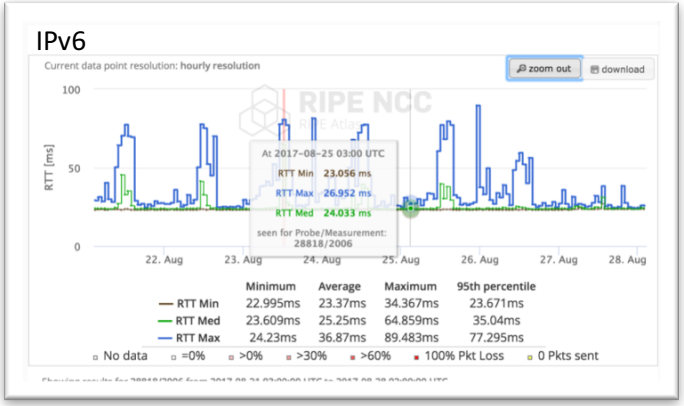


Probe28818

IPv4



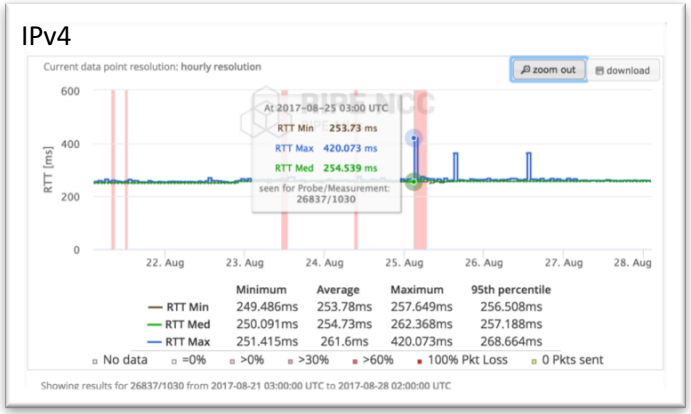
IPv6



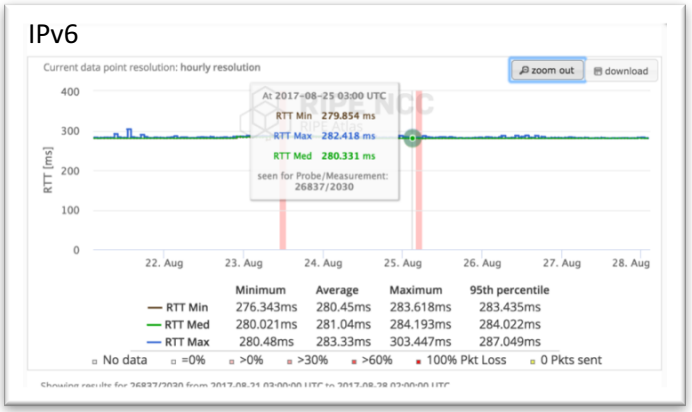
RIPE Atlasで見る: OCNと海外

Probe26837

IPv4

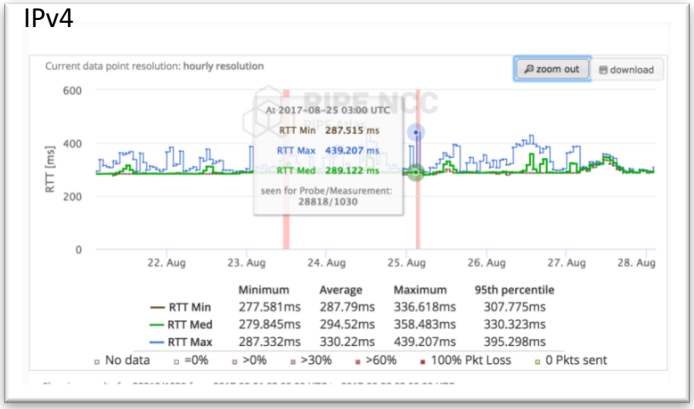


IPv6

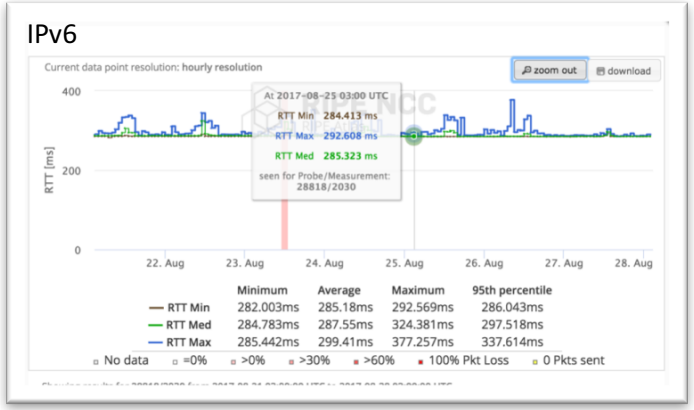


Probe28818

IPv4



IPv6



RIPE Atlasから見えること

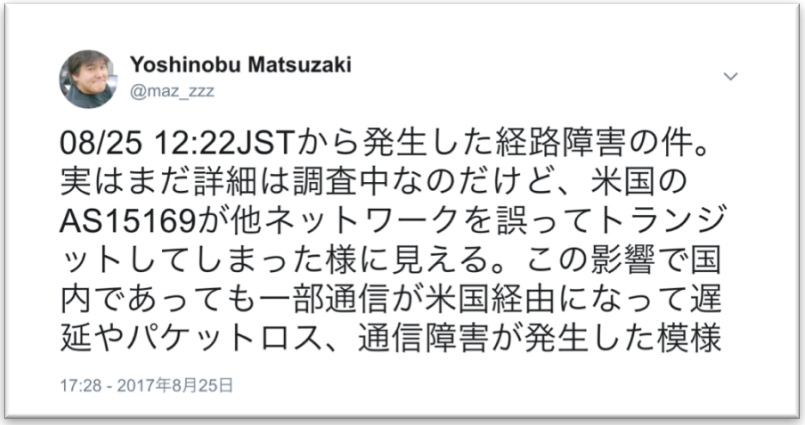
- 該当Probeでは国内、海外のIPv4通信に遅延の増加やパケットロスを観測
- IPv6へ直接の影響はほとんどなかった模様
  - IPv4のBGP経路が対象であったため
  - probe26837ではIPv4/IPv6で宛先に寄らずパケットロスが観測されているためProbe近傍のどこかで輻輳が発生していたかもしれない

影響時間の推定

- 迂回による遅延や輻輳の影響
  - 今回広報された経路に関わる通信
  - 経路やRIPE Atlasなどの記録からすると20分以内には収束したはず
- 大量の経路広報で不安定になった機器の影響
  - 影響した機器があった場合、そこを通過する通信
  - 収束時間はその運用組織の対応に依存

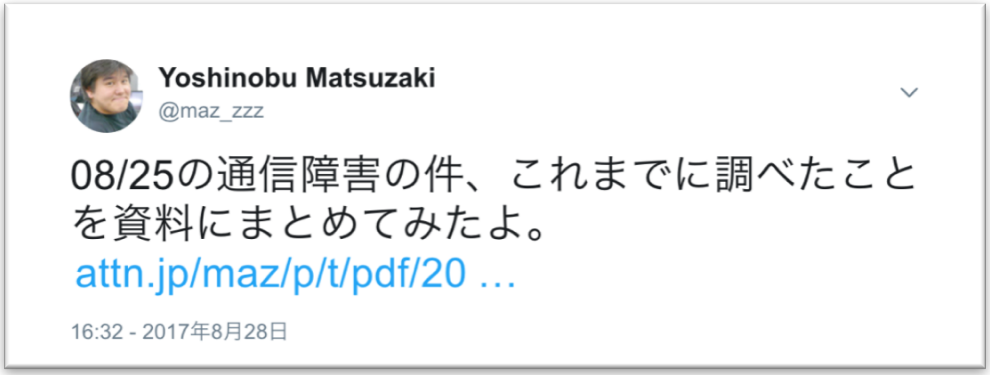
僕がやってたこと 08/25(金)

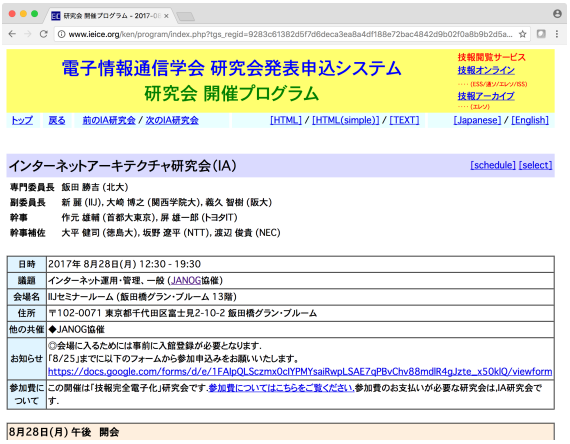
- 14時台
  - ENOG46会場に向かい中、某社SOCより連絡
  - 調査開始して、関係各所に情報共有
- 15時台
  - JANOG MLに投稿 [janog:14007]
  - ENOG46で発表
- 17時台
  - AS15169 NOCに情報共有
  - Twitterに投稿
- 18時台
  - ENOG46懇親会



僕がやってたこと 08/28(月)

- その日のIA研@IIJで事象を概説することになり、慌てて資料作成
- 何とかできたので公開

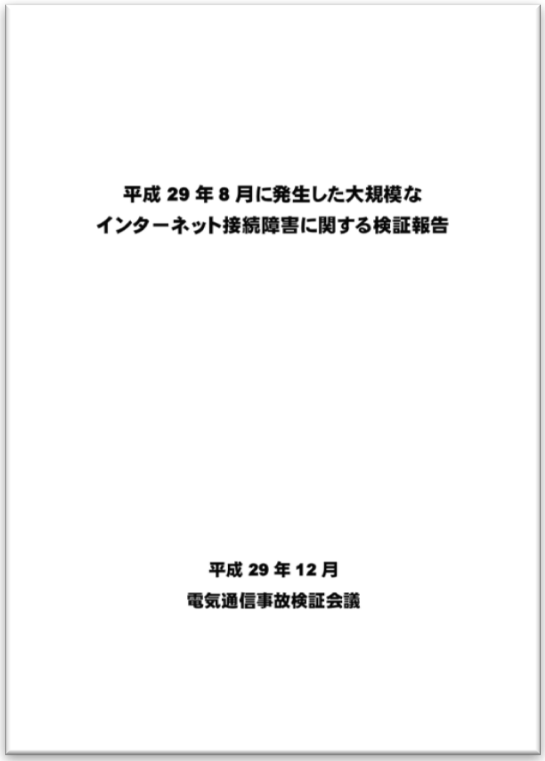




https://www.attn.jp/maz/p/t/pdf/20170825-routeleakage.pdf

https://search.yahoo.co.jp/realtime/search?p=%E9%80%9A%E4%BF%A1%E9%9A%9C%E5%AE%B3&ei=UTF-8

電気通信事故検証会議の報告



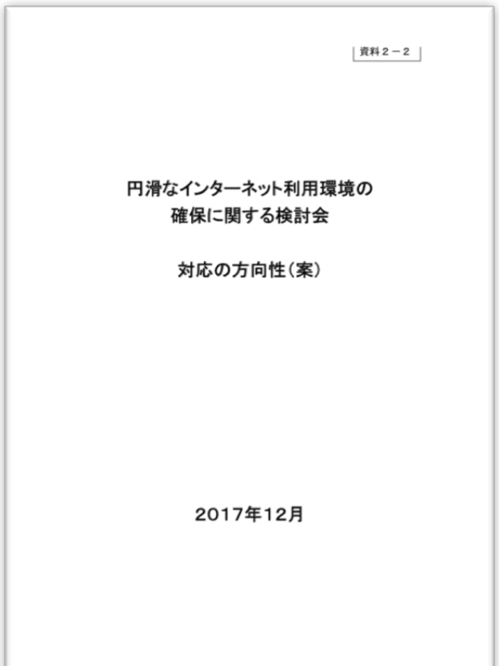
- 人為的ミスの未然防止
- 誤送信された経路情報の受信防止及び不要な経路情報の送信防止
- 障害に関する情報の電気通信事業者間での共有
- 利用者周知

16 ベストエフォートサービスにおいては、停止が
ない限りは品質の低下と判断していないことから、
総務省は、本事案については電気通信事業法上の
事故には該当しないと判断している。

http://www.soumu.go.jp/main_content/000523153.pdf

円滑なインターネット利用環境の確保に関する検討会

http://www.soumu.go.jp/main_sosiki/kenkyu/smooth_internet/


資料２－２

円滑なインターネット利用環境の確保に関する検討会

対応の方向性（案）

2017年12月

検討会(第２回)配布資料: 対応の方向性(案)

- 誤送信された経路情報の受信防止及び不要な経路情報の送信防止を図るためのフィルターやリミッターの設定といった仕組み自体は既に存在しているが、広く電気通信事業者に、それぞれのネットワーク構成や運用の考え方に合わせて柔軟かつ適切に対応することを求めるためには、例えば、「情報通信ネットワーク安全・信頼性基準」にルータのフィルター設定について規定する等、適切な制度的対応を検討する必要がある。
- また、総務省が本障害のような事象を迅速かつ的確に把握し、情報共有の結節点となるためには、これまで総務省への報告対象ではなかったインターネットの障害についても、サイバー攻撃に起因するものも含めて報告対象とする等、総務省への報告の在り方を含め、障害に関する情報共有体制の整備を検討する必要がある。

http://www.soumu.go.jp/main_content/000525811.pdf

情報集約と展開？

- 日本経済新聞、1月23日紙面
  - サービスで障害が発生した場合、第一報を接続事業者から総務省に報告
  - 同省が…関係する業者に必要な情報を送る
  - 対象となる障害の度合いやどこまで報告するかは業者の意見を踏まえて今後詰める

情報共有、周知への課題

- 会社事情
  - 会社ポリシー、NDA
  - 「公式見解」による硬直化
- 利用者対応
  - 即時性と不確かな”障害”情報
- 報道対応
  - 報道は犯人探しになりがち

- どうすれば、うまくやれるだろうか？

経路制御でやりたいこと

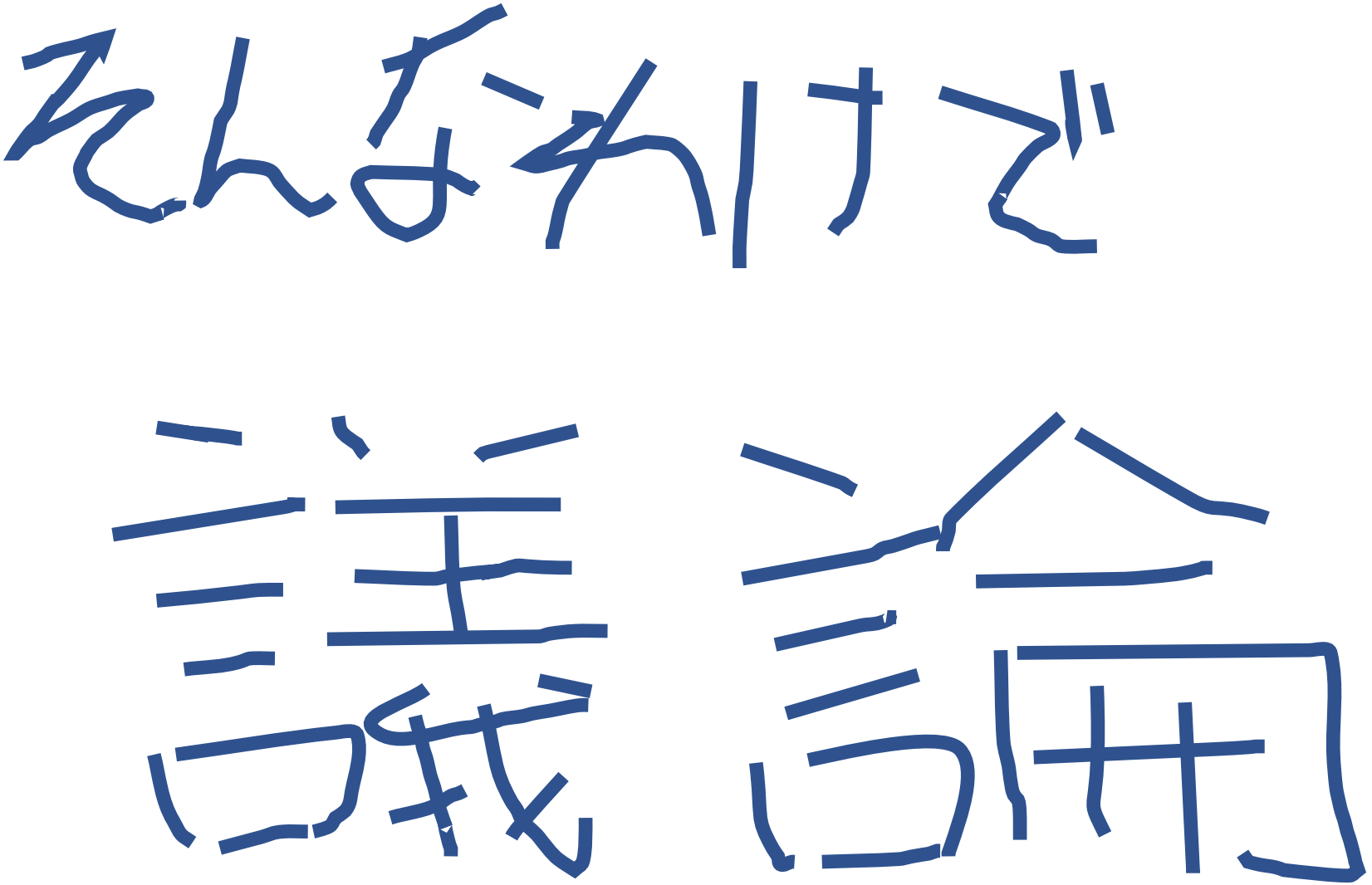
- **到達性確保**
- **トラヒック制御**

- 付随して確保したいことは他にもあるけど
  - 冗長性
  - 拡張性
  - 運用容易性

自身が広報するBGP経路には届けたい範囲(スコープ)がある

- 到達性を担保する経路(トランジット向け広報)
  - グローバルに届いて欲しい
- 到達性を確保する経路(ピア向け広報)
  - 隣接ASとその配下のみに伝搬して欲しい
- トラヒックを制御する経路
  - 隣接ASとその配下のみに伝搬して欲しい
  - 場合によっては隣接ASだけでも良い

- でも、これらは自分では制御できない
  - 一旦広報した経路は、受信側のポリシで制御される


そんなわけで
議論

情報の集取と展開

- みんな何から情報集めてますか？
- どこで情報提供するのが嬉しいですか？

- それっぽい情報を出す元はどこが良い？
  - 個人
  - JANOG WG
  - ICT-ISAC
  - JPNIC
  - お役所

BGP運用どうします？

- フィルターやリミッターの設定？
  - リミッターはかなり覚悟が必要な設定
  - 今回の件は防げない
- 相互接続を推進、もしくは縮退？
  - 接続先を増やしてAS_PATH長で勝てる状態を作る
  - 接続先を減らして漏洩の可能性を減らす
- leak community？
  - https://datatracker.ietf.org/doc/draft-heitz-idr-route-leak-community/

局面に応じた情報共有

- 発生直後
  - 事象把握のために事業者間で情報交換
    - JANOG等で、信頼関係を構築しておきましょう
  - 硬直化前にできるだけ早く
- 利用者周知
  - 障害認知の公表、状況の適宜更新
  - 権威ある周知情報があると参照しやすい
- 報道対応
  - 僕たちが伝えたいことと、報道価値は別モノかも

送出経路ポリシは基本２つ

- ピア/上流に広報する経路
  - 自身と配下の経路
- 顧客ASに広報する経路
  - フルルート/デフォルトルート

- BGPでトラヒック制御しようとすると複雑になっていく
  - 例えば、特定ピアのみ細かな経路の広報などなど

BGPでトラヒック制御しない

- 到達性の維持に努める
- 必要なところに必要な帯域を用意する
- 必要なASと適宜ピアを拡充する

- ピア、上流に常に同一の経路広報を行う
  - 誰かが経路流出させても、周辺ASと多拠点でピアできていれば、流出経路はAS PATH長で採用されない可能性が高まり、影響範囲が極小となる

複雑化するBGP

- 経路に局所性が出てきている
  - トラヒック制御のため
  - 隠すため
- BGPは見ているポイントで経路が異なる
  - 漏れ出した経路は、特定のネットワークだけに届くかも

経路削減は延命策 not 解決策

- 経路フィルタで削減された経路数は現状のもの
  - 将来やトラブル発生時の経路数は未知
- max-prefixなどで受信経路数を制限できるけど
  - BGPピアが切断しても大丈夫？
  - alert受信して、機器が不安定になる前に対処可能？

- 十分余裕を持った運用を心がけるしかない
  - あるいは複数の安全策の併用

BGP運用でできること

- 統一した広報ポリシ
  - AS間のトラヒック制御に細かい経路を利用しない
- 受信ポリシの検討
  - 受信経路で酷い影響が出ないようにする
- 余裕を持った機材での運用
- 事象を第三者が検証できるように、経路アーカイブ等へ経路情報を提供しておく

peerlockingの適用

- Peerlocking
  - https://www.nanog.org/sites/default/files/Snijders_Everyday_Practical_Bgp.pdf
- あるASと堅牢に相互接続できてるなら、そのAS関連の経路を他から受信しない
  - 他ASが経路を漏洩しても影響を受けない
  - 迂回路が直接接続のどれかに限られるため、事故が起こった際の影響はより酷いことになるかも
  - 運用は結構大変

peerlocking模式図

- AS_PATHにAS-YYYが含まれている経路を他のASから一切受信しない
  - 上流、ピア、顧客