経路制御の課題と対策

Matsuzaki ‘maz’ Yoshinobu

<maz@iij.ad.jp>

今日のトピック

- BGPにはscopeが無い
- 期待される制御を経路フィルタ等で実装
  - 契約
  - 慣習
  - 経費

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

BGPで到達性確保

- 入方向の到達性
  - 隣接ASへ集約経路を経路広報
- 出方向の到達性
  - 隣接ASから彼らの経路を受信

BGPで到達性担保

- ピアと接続断しても到達性が欲しい
  - 何らかASを跨いだ迂回路が必要
  - →トランジットの調達

- 入方向の到達性
  - 上流ASへ経路広報
- 出方向の到達性
  - 上流ASからフルルート/デフォルト経路を受信

BGPでトラヒック制御

- 入方向の制御
  - 集約経路や分割した経路を隣接ASに広報
    - AS_PATH prepend
    - BGP community
    - MED
    - Large BGP community
- 出方向の制御
  - 隣接ASから受信した経路に重み付け
    - MED
    - Local preference
    - IGP cost

広報した経路には、届けたい範囲がある

- 到達性を担保する経路
  - グローバルに届いて欲しい
- 到達性を確保する経路
  - 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

範囲がズレると問題発生

- 到達性を担保する経路
  - グローバルな到達性
- 到達性を確保する経路
  - 隣接配下への到達性
  - 意図せぬ通信経路
- トラヒックを制御する経路
  - 隣接配下への到達性
  - 意図せぬ通信経路

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)

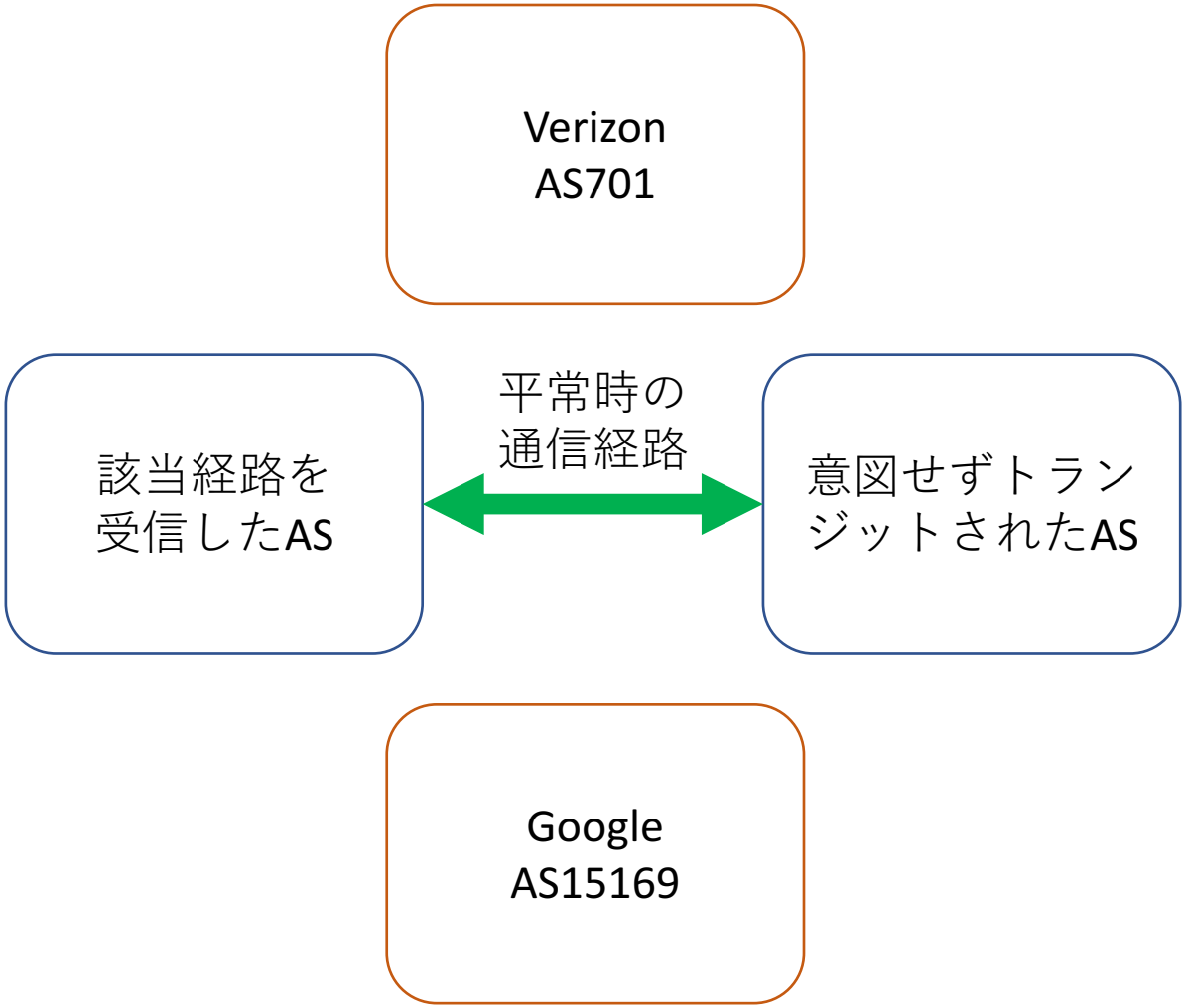
BGPは観測点によって見える情報が異なるのでご注意

その他、AS15169の広報経路

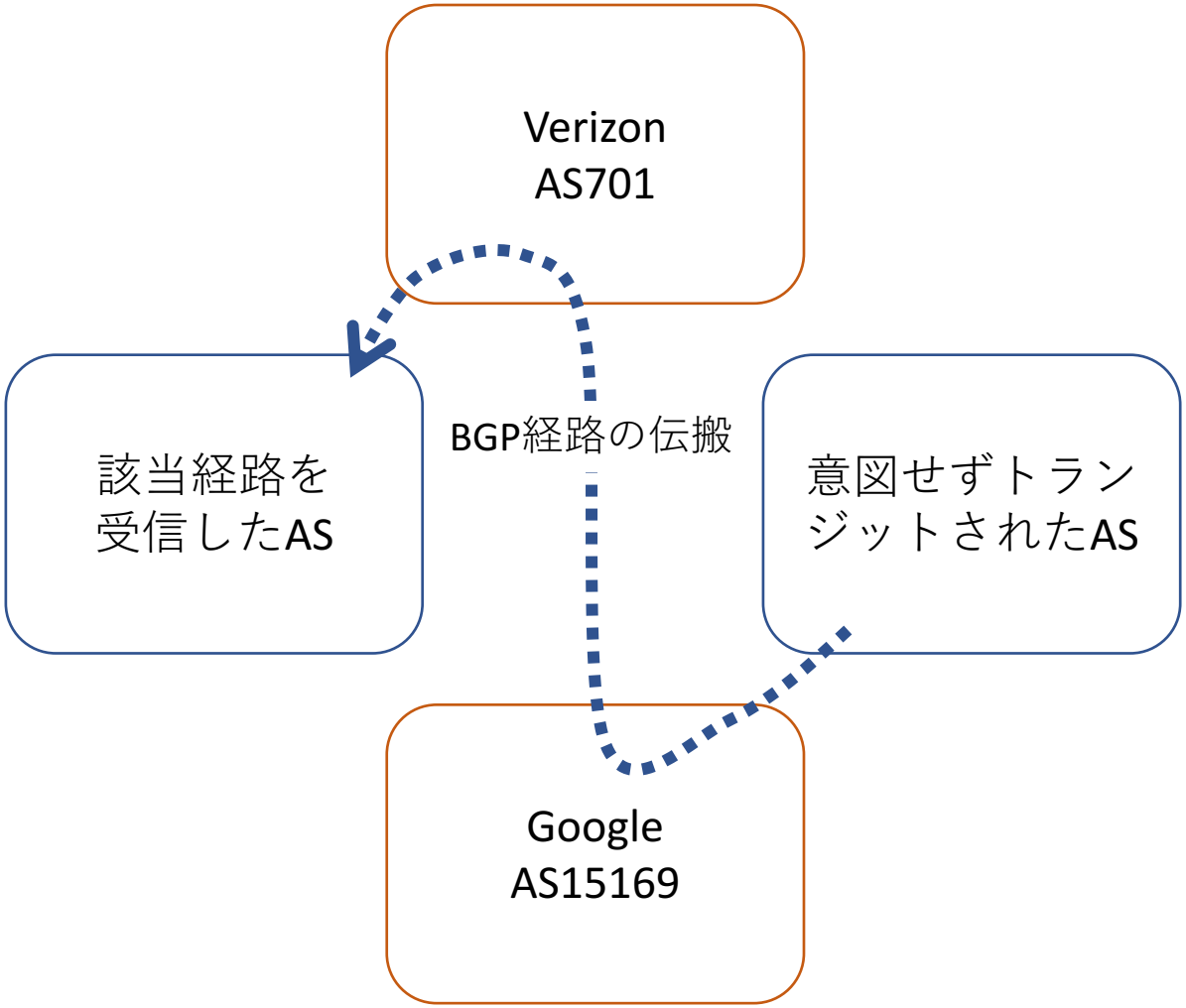
- 期間中、AS15169が経路生成元となる、日頃見えない経路が追加で広告されていた
- Google関連
  - AS15169とその配下ネットワークの細かな経路
  - 654経路
- 世界中のIXPで使われていると思われるsegment
  - 78経路
- その他、まだ判別できていないsegment
  - 2経路

BGPは観測点によって見える情報が異なるのでご注意

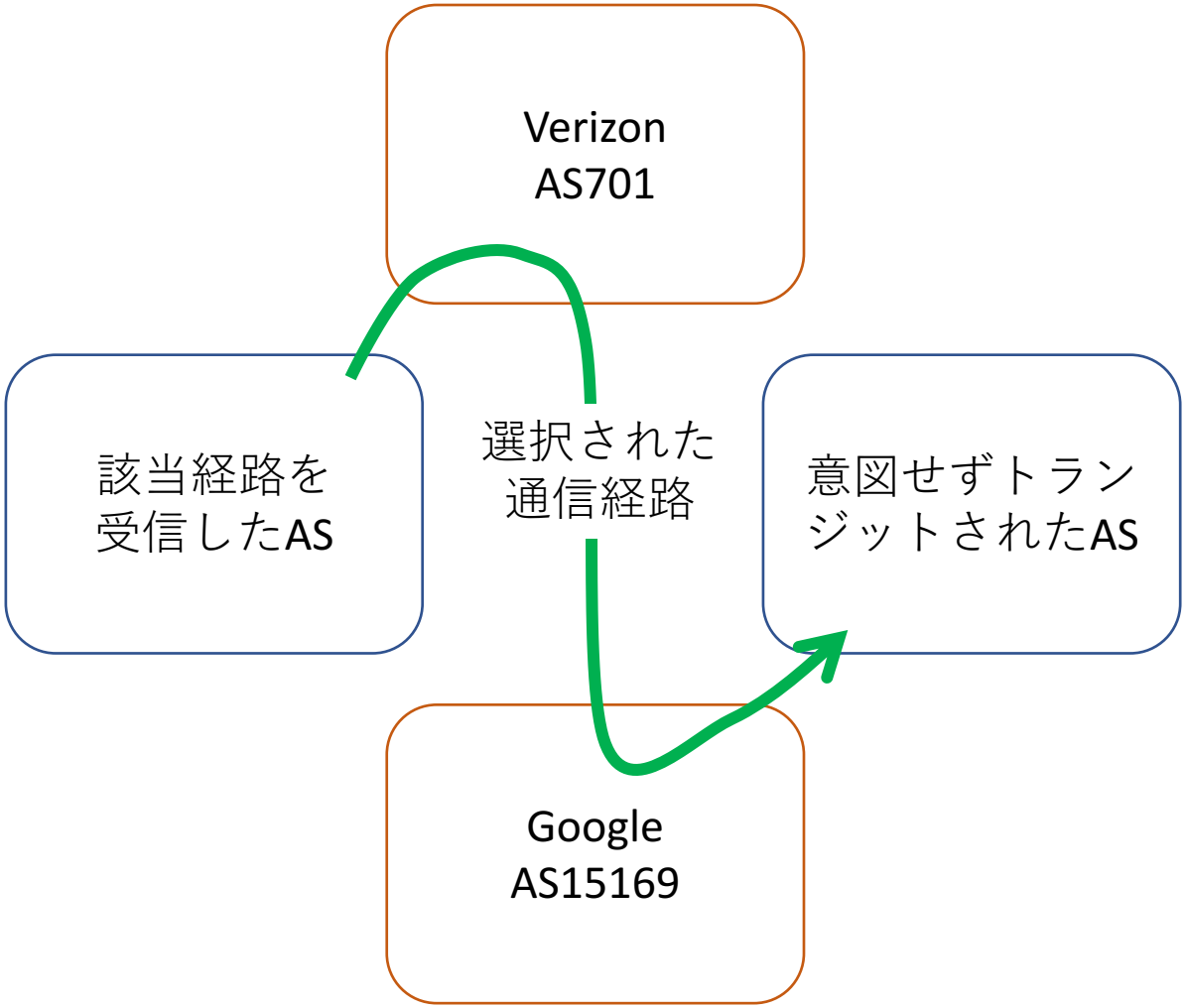
概要図 1 : 平常時


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２：誤トランジット発生



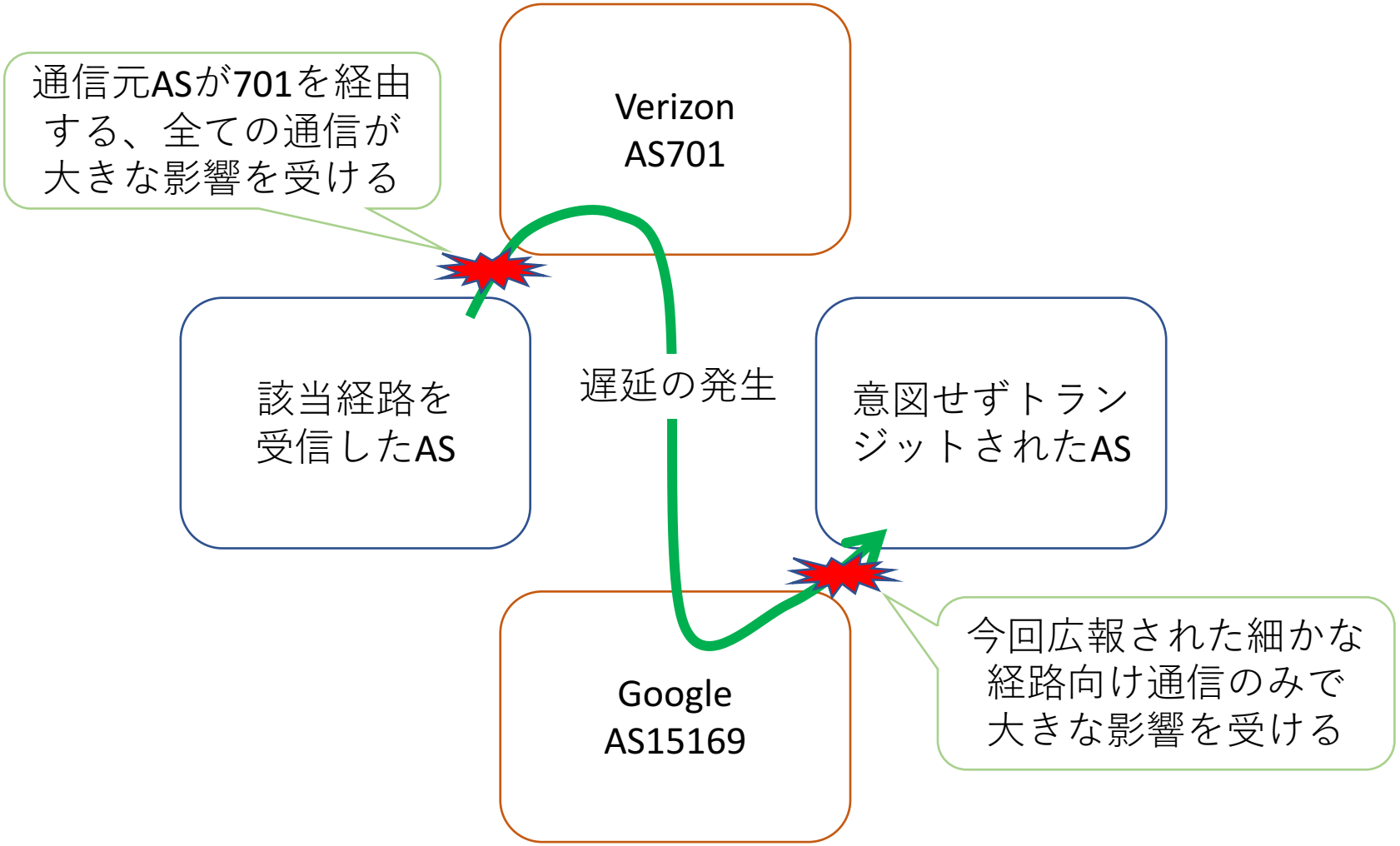
概要図３：障害期間中の通信



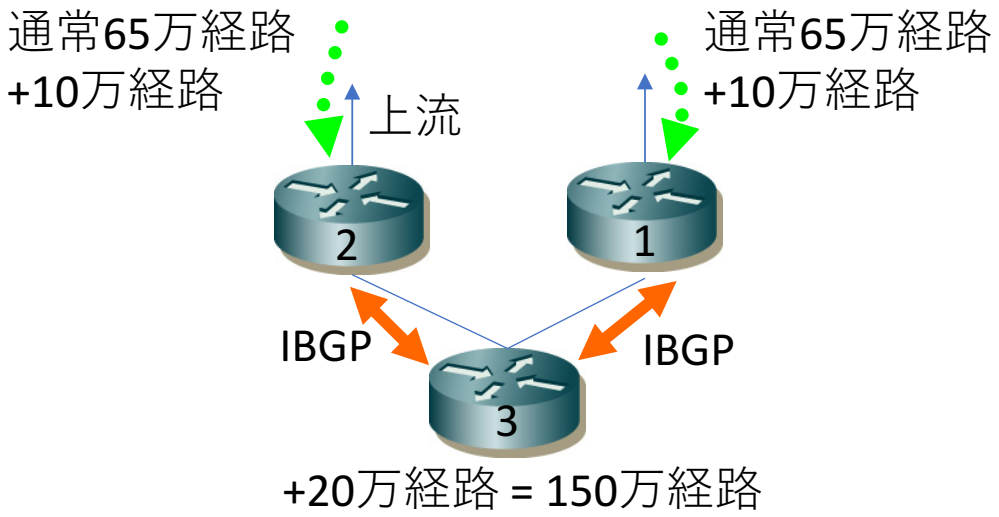
障害や影響の推定

- 広報された宛先向けの通信が米国経由になった
  - 遅延の発生
  - 経路上に十分な帯域がない場合は輻輳の発生
- 大量の経路広報を受信した
  - 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越しの通信が意図しない経路に迂回したかも
  - 発生条件
    - 内部的にIXPセグメントのIPアドレスをNEXTHOPに利用
    - 外部から今回広報されたIXPセグメントの経路を受信
    - しかも内部で優先されてしまう
      - prefix長が一緒だと一般に Connected > EBGP > IBGP な優先度

輻輳箇所と影響


通信元ASが701を経由する、全ての通信が大きな影響を受ける
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
今回広報された細かな経路向け通信のみで大きな影響を受ける

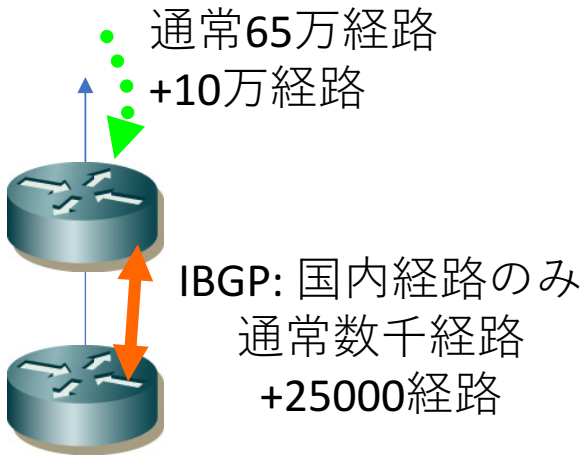
大量の経路追加



- 現状DFZに約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



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

BGPは観測点によって見える情報が異なるのでご注意

4713が生成している経路

平常時(内78prefixが影響)

prefix長	prefix数
/10	1
/11	3
/12	7
/13	9
/14	6
/15	12
/16	38
/17	11
/18	5
/19	5
/20	15
/21	11
/22	21
/23	9
/24	67

今回、追加で流通した経路

prefix長	prefix数
/10	
/11	
/12	
/13	1
/14	1
/15	3
/16	29
/17	10
/18	15
/19	79
/20	868
/21	1764
/22	3035
/23	2432
/24	16594

BGPは観測点によって見える情報が異なるのでご注意

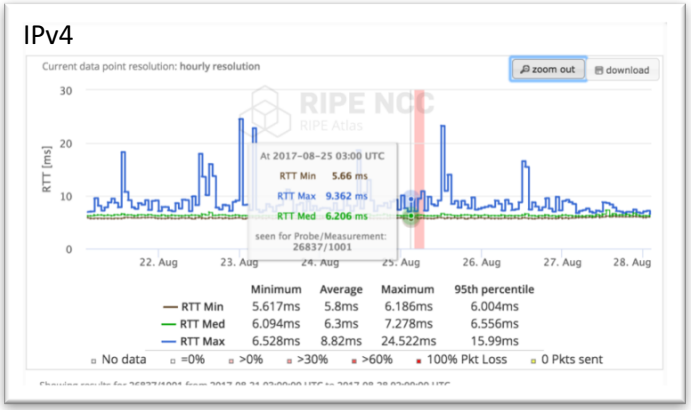
RIPE Atlas Probe

- 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

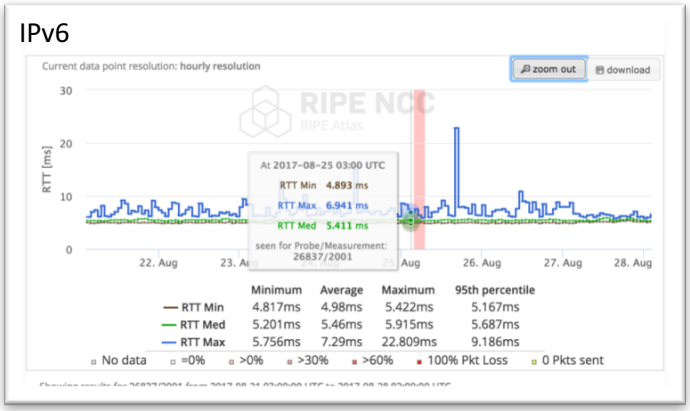
RIPE Atlasで見る：OCN内

Probe26837

IPv4

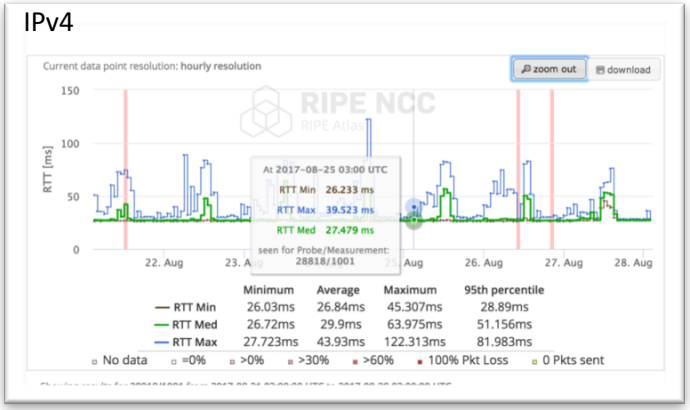


IPv6

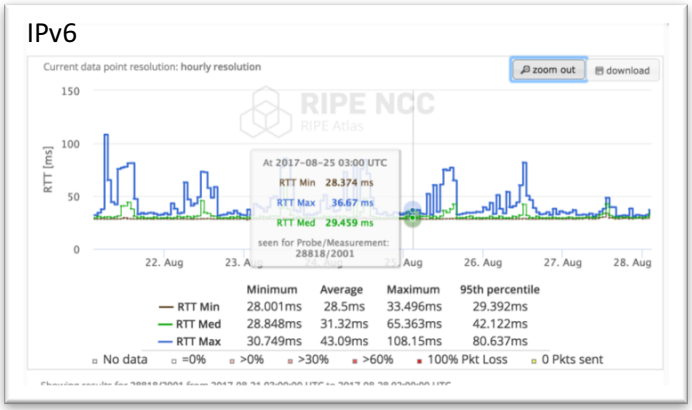


Probe28818

IPv4



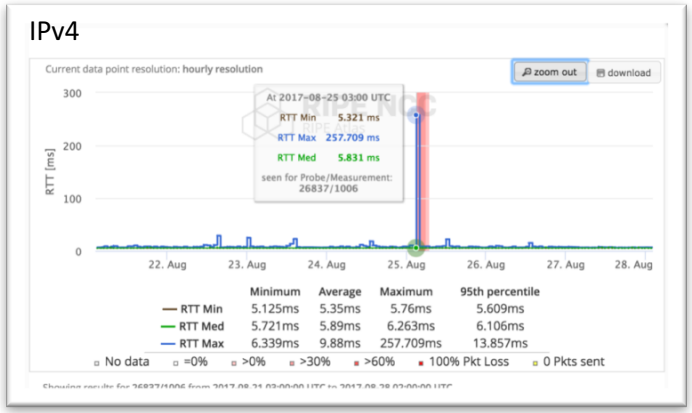
IPv6



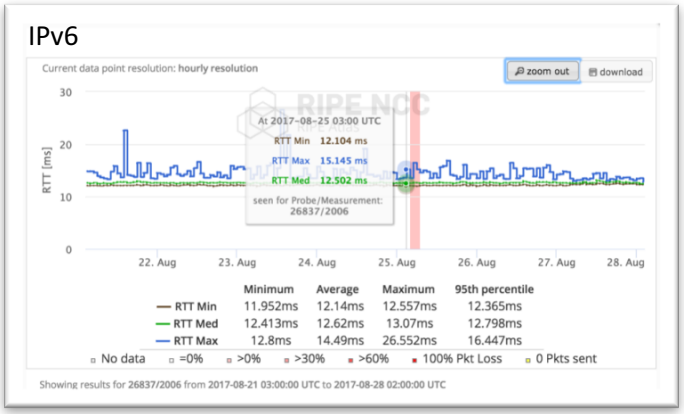
RIPE Atlasで見る: OCNと国内

Probe26837

IPv4

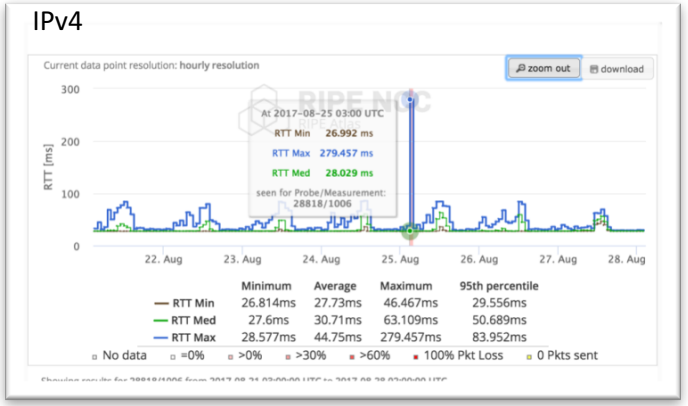


IPv6

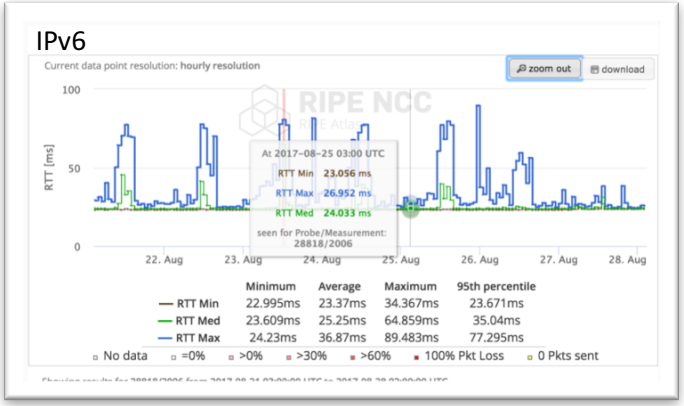


Probe28818

IPv4



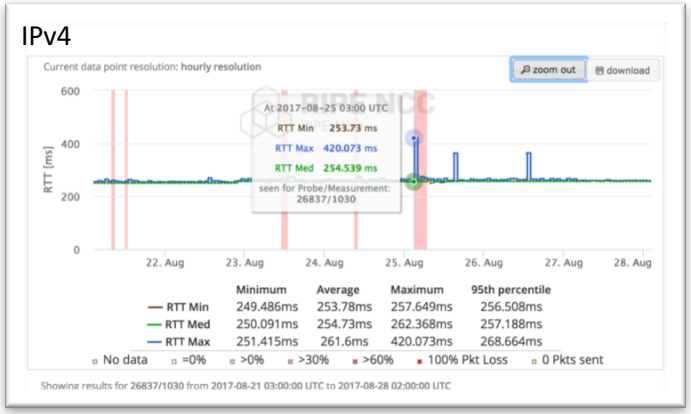
IPv6



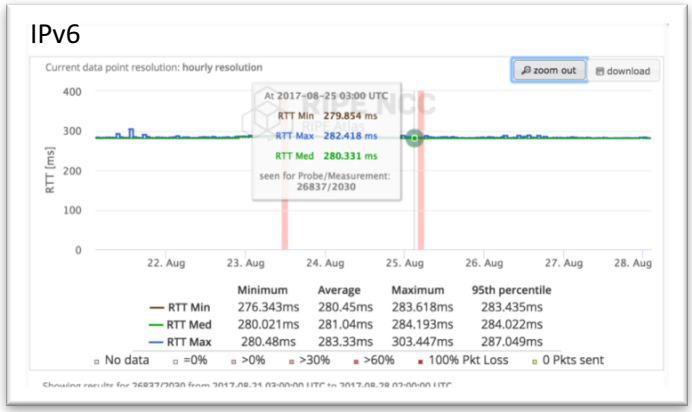
RIPE Atlasで見る: OCNと海外

Probe26837

IPv4

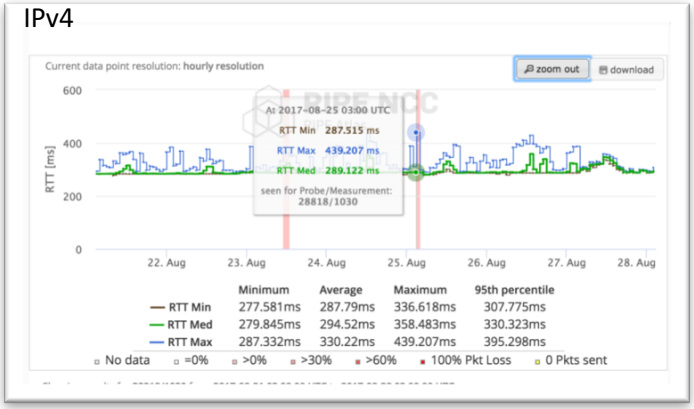


IPv6

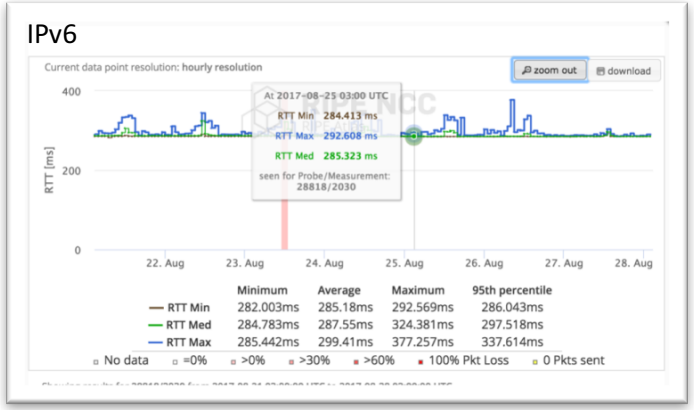


Probe28818

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
    - Probeから2ホップ以内ではパケットロスを観測せず

観測されている事象

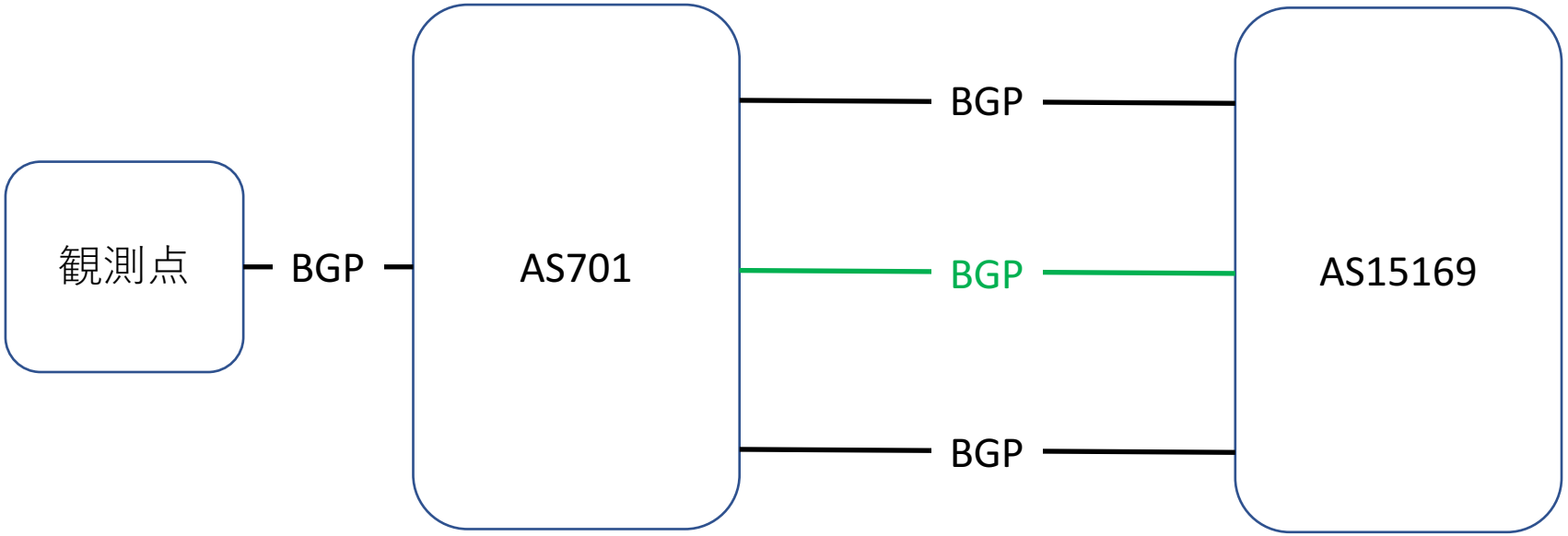
- 2017/08/25 12:22JST頃
  - AS15169が他ASのIPv4経路をトランジット開始
  - 日頃流通しない細かい経路が大量に広報
    - AS15169内部の細い経路も広報
- 2017/08/25 12:33JST頃
  - AS15169がトランジットしていた経路を削除
  - AS15169 originのいつもの経路情報の全てが追加でUPDATE

UPDATEからの推定

- まず、BGP UPDATEの発生要因のおさらい
  - 何らかAS15169側で制御可能な属性値が変更
    - MED, BGP community, その他transitiveな値
    - BGP nexthop
- AS701とAS15169間は複数の接続があると予想

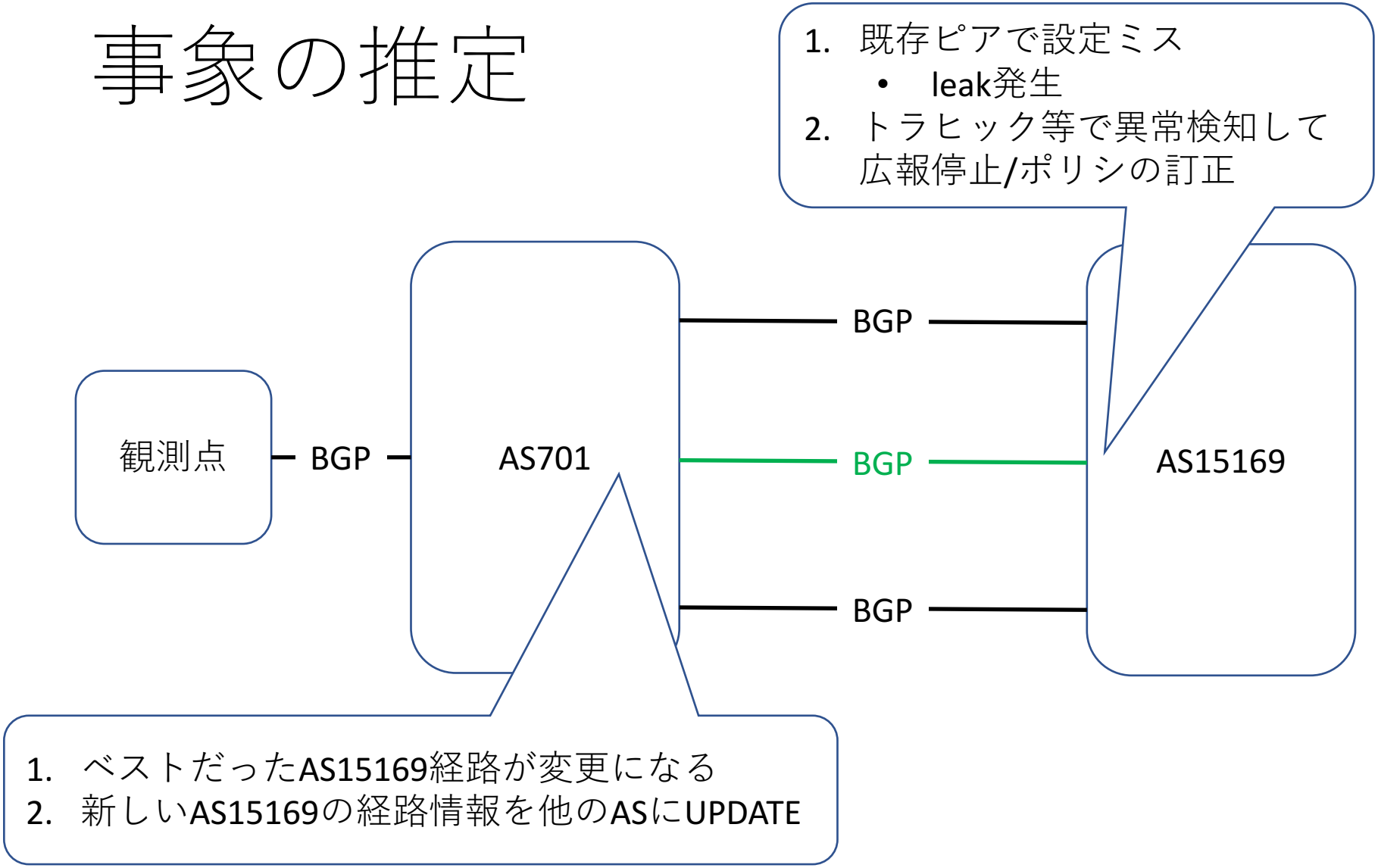
- 想定されること
  - 既存の相互接続でポリシを更新して問題発生

BGPと挙動



- 新規ピアで問題発生とか、shut/no shutしたとかは無さそう

事象の推定


1. 既存ピアで設定ミス
• leak発生
2. トラヒック等で異常検知して広報停止/ポリシの訂正
観測点
BGP
AS701
BGP
BGP
BGP
AS15169
1. ベストだったAS15169経路が変更になる
2. 新しいAS15169の経路情報を他のASにUPDATE

事象のサマリ

- 本来隣接ASとのトラヒック制御を意図していた経路が外部に流出したと考えられる
- USの特定ASにのみ流出してトラヒックの吸い込みが発生したたため、日本では遅延等が発生
- 長時間の障害が発生していたネットワークでは、一時的な経路数増加による影響を受けていたと考えられる

BGPで出来ないこと

- 経路を届けたい範囲が制御できない
  - 到達性確保
  - トラヒック制御
- 他人が広報する経路を制御できない
  - 生成する経路
  - トランジットする経路

- 人は失敗する
  - 設定ミス
  - 自動化はすごく助けになるが、完全では無い

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

受信経路フィルタ必須

- BGP nexthopの経路など、内部制御に影響する経路は絶対に受け取らない
  - 自身のPA
  - IXPのセグメント
- トランジットを提供する場合
  - 顧客ASからの広報に厳密なフィルタを適用
  - 意図しない経路でも、受け取ってしまうと世界に情報が流通する

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、十分ピアできていればAS PATH長で勝てるので、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

- 可能な限り、第三者検証できる状態を作っておくのが大事
  - 公開経路アーカイブなどへの協力が必要

まとめ

- 到達性確保とトラヒック制御が期待される
  - BGPのトラヒック制御は他ASの運用に依存
  - 設定ミスが発生すると、意図しない状態となる
- 性能に余裕を持った機材で運用
  - 扱える経路数など
- 経路フィルタでの防衛
  - 意図しない経路を送出、受信しない
- 経路ポリシでの防衛
  - ポリシを単純化し、充実した相互接続を進めることで他ASの設定ミスの影響を限定的にする