08/25の通信障害概説

Matsuzaki ‘maz’ Yoshinobu

<maz@iij.ad.jp>

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)

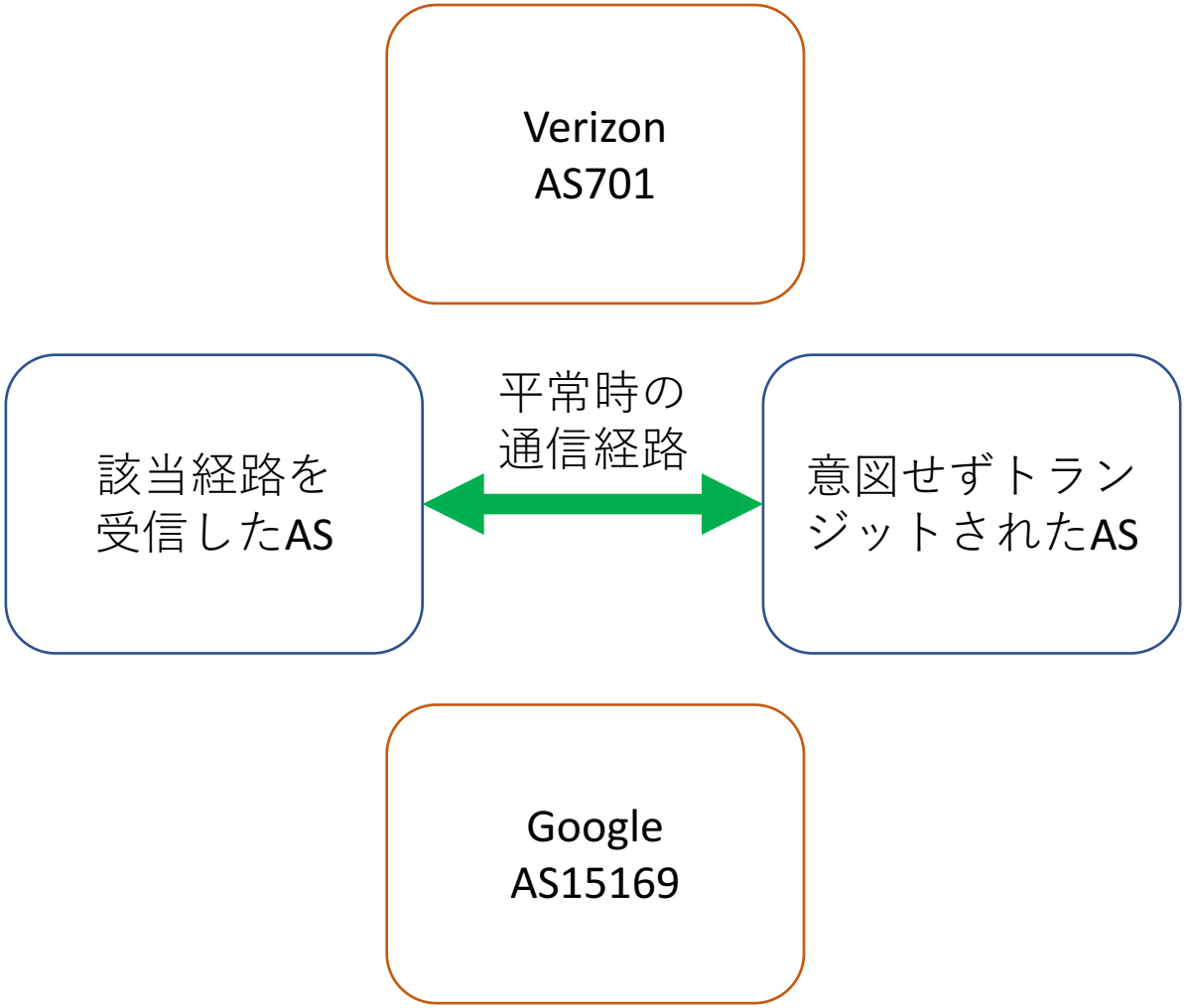
BGPは観測点によって見える情報が異なるのでご注意

その他、AS15169の広報経路

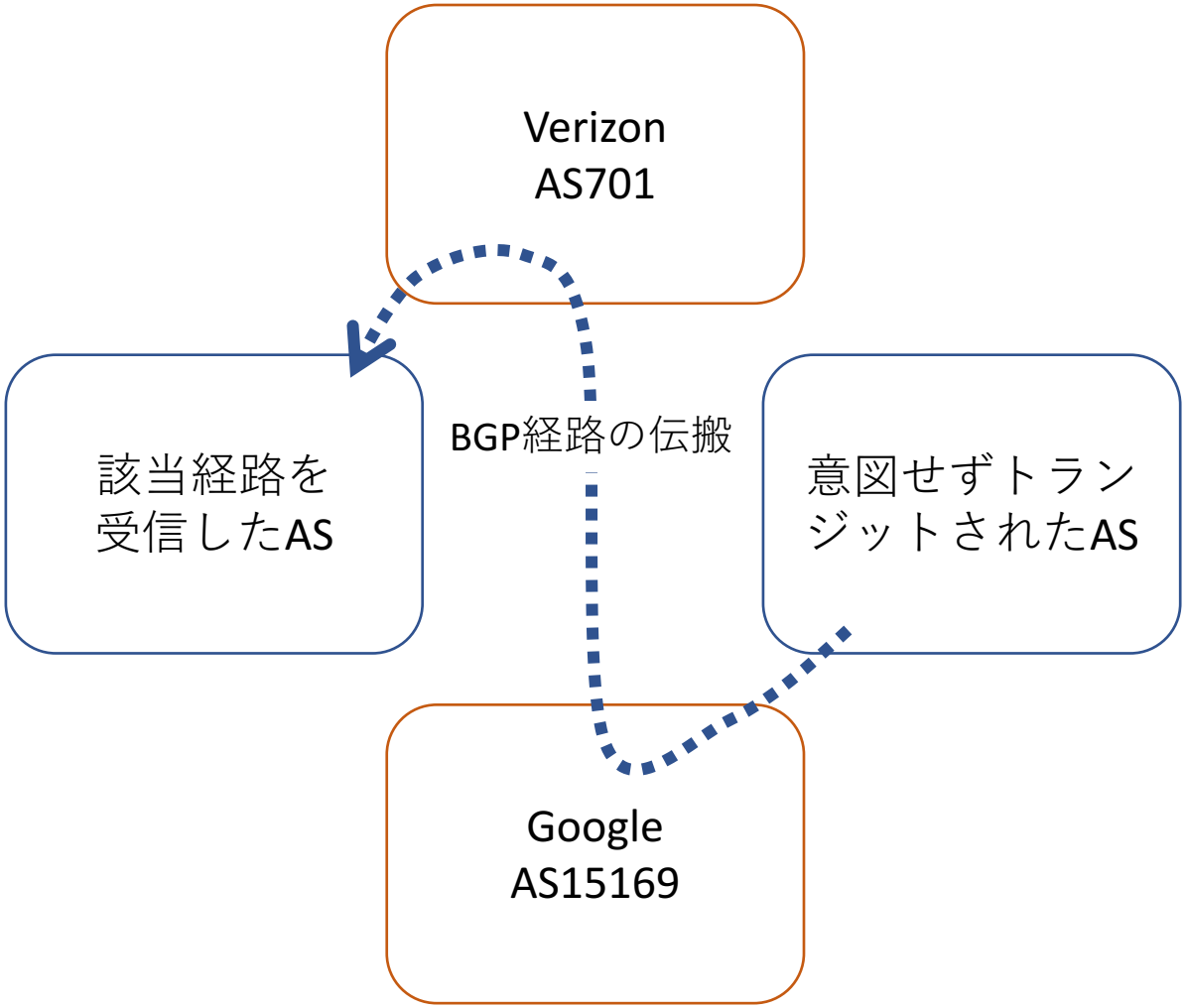
- 期間中、AS15169が経路生成元となる、日頃見えない経路が追加で広告されていた
- Google関連
  - AS15169とその配下ネットワークの細かな経路
  - 654経路
- 世界中のIXPで使われていると思われるsegment
  - 78経路
- その他、まだ判別できていないsegment
  - 2経路

BGPは観測点によって見える情報が異なるのでご注意

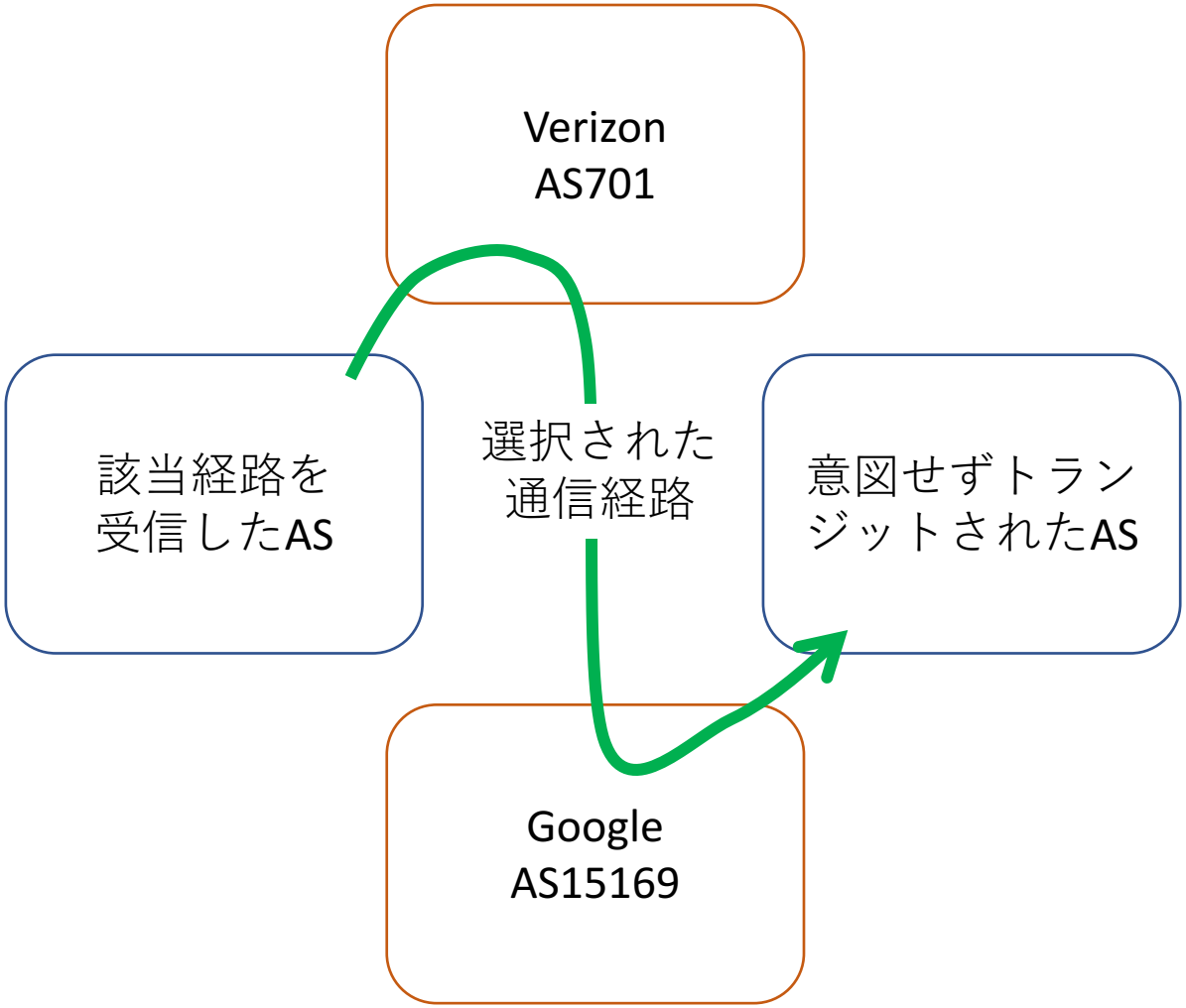
概要図１：平常時


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



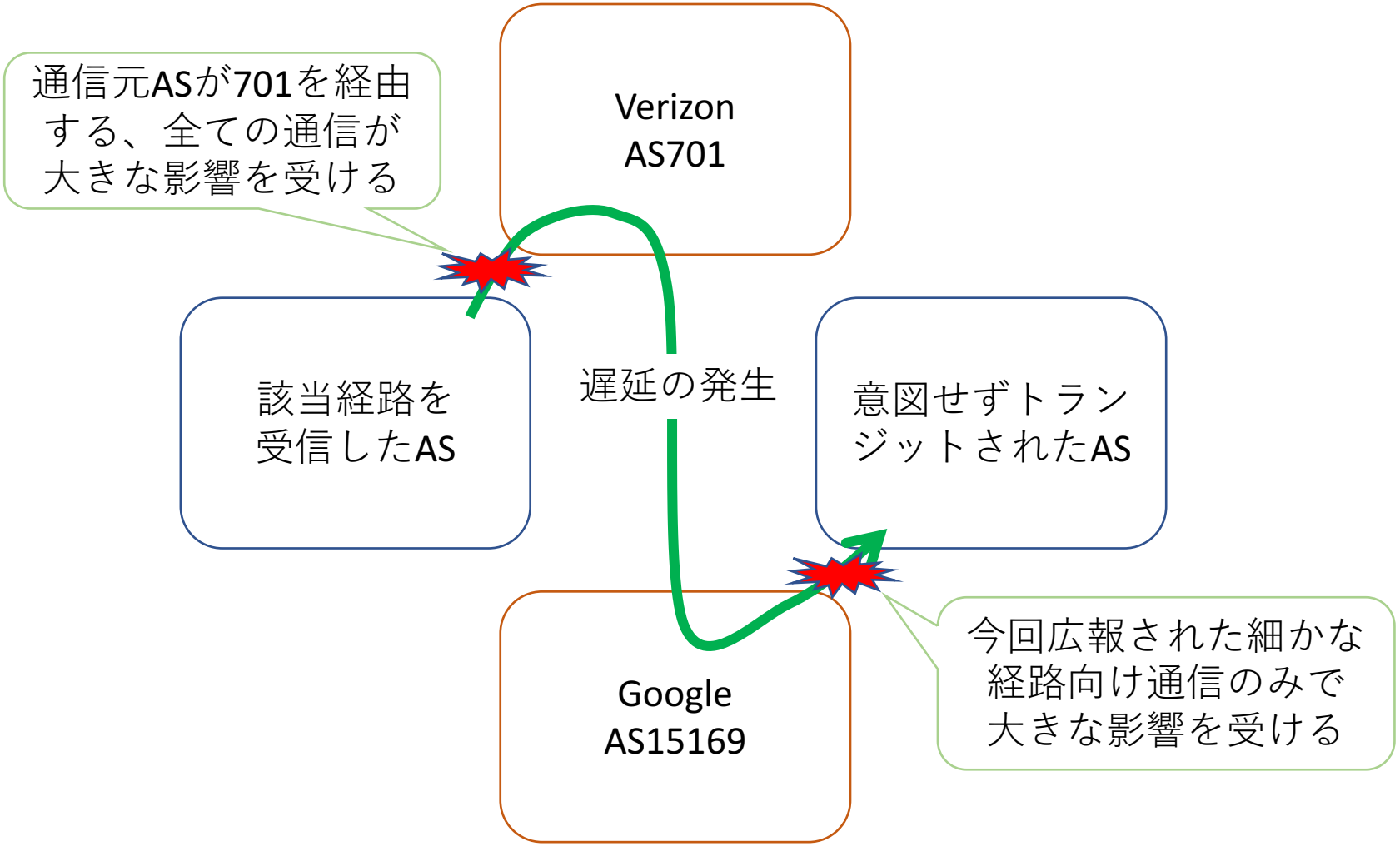
概要図３：障害期間中の通信



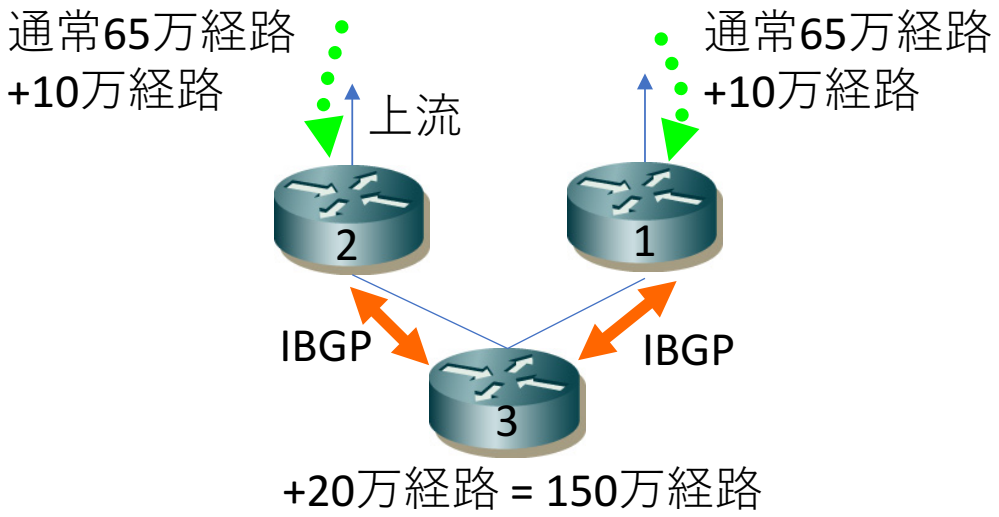
障害や影響の推定

- 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
- 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

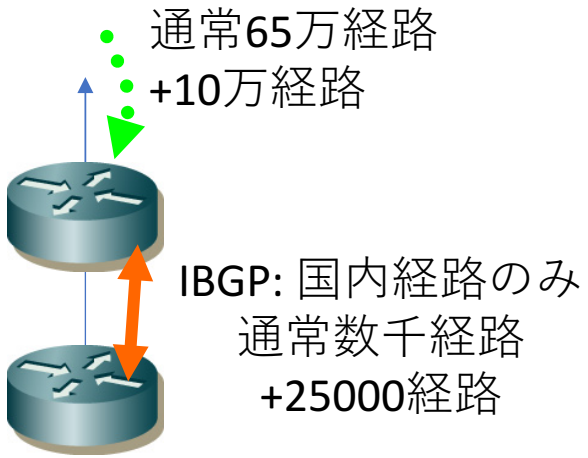
大量の経路追加



- 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

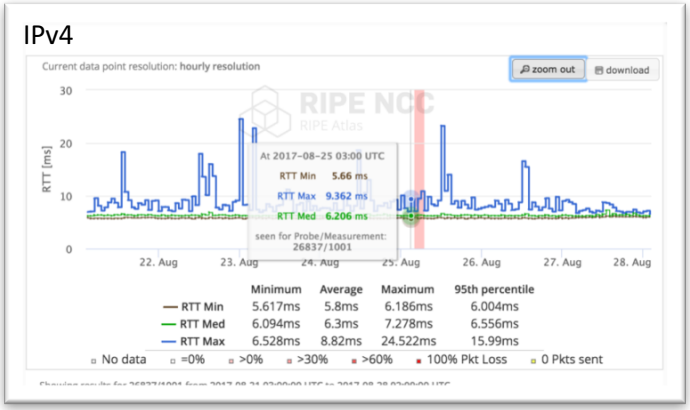
RIPE Atlas Probe

- 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

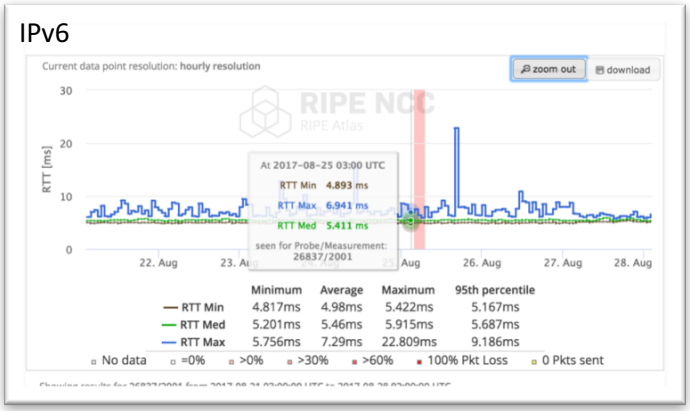
RIPE Atlasで見る：OCN内

Probe26837

IPv4

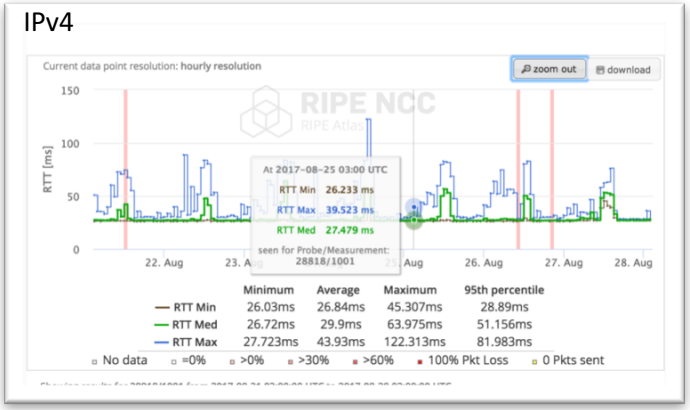


IPv6

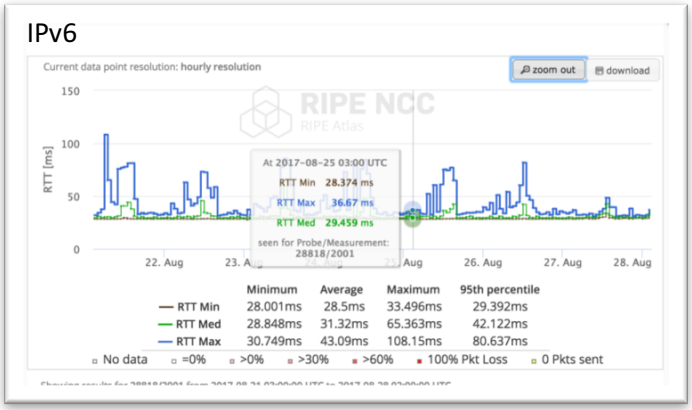


Probe28818

IPv4



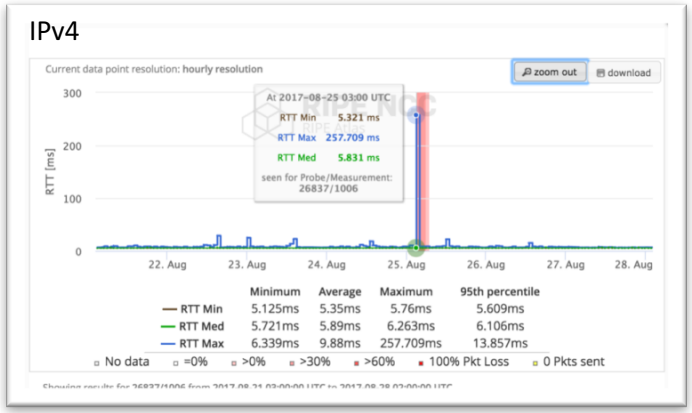
IPv6



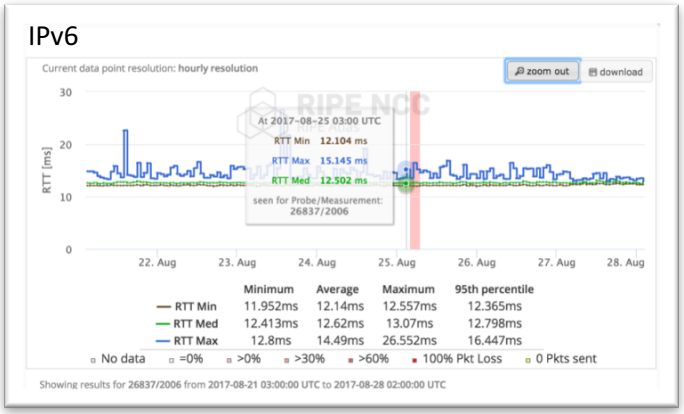
RIPE Atlasで見る: OCNと国内

Probe26837

IPv4

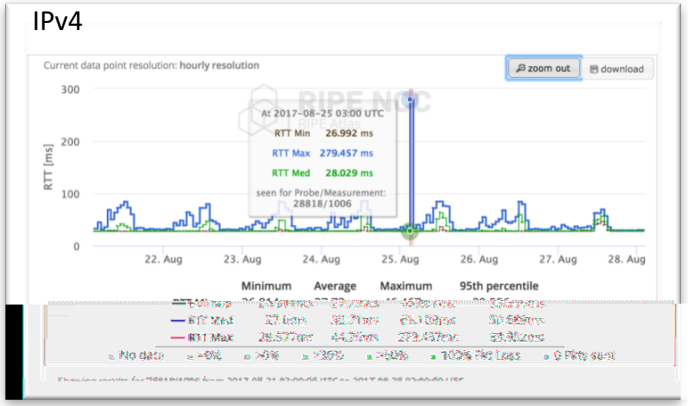


IPv6

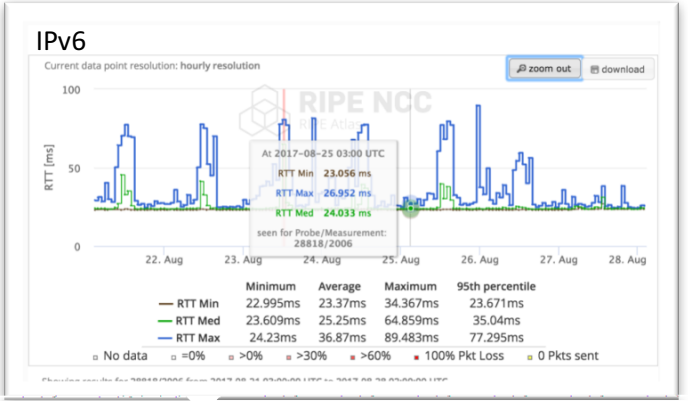


Probe28818

IPv4



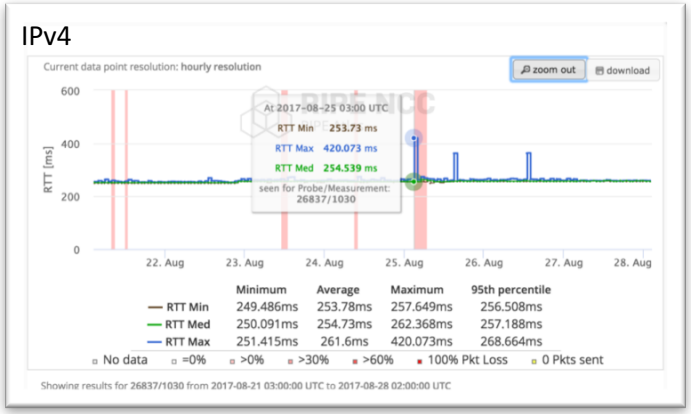
IPv6



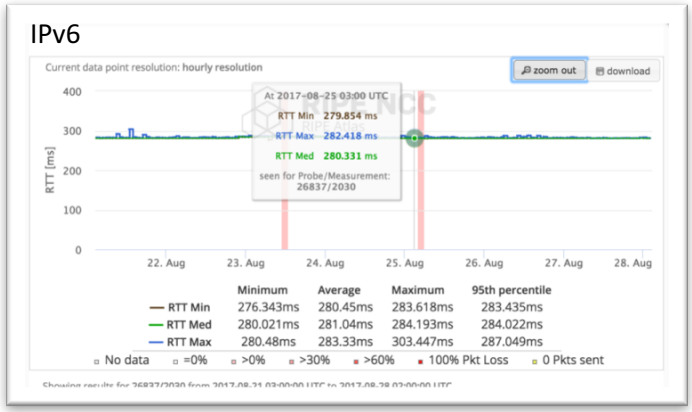
RIPE Atlasで見る: OCNと海外

Probe26837

IPv4

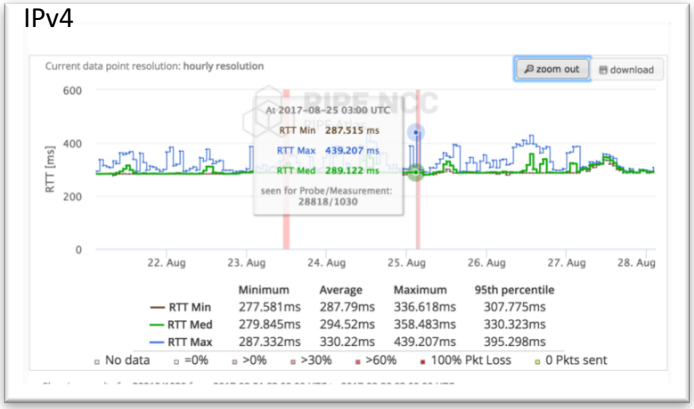


IPv6

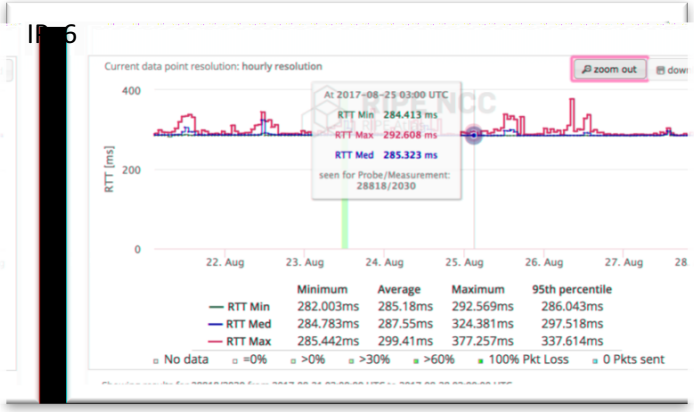


Probe28818

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
- AS701向け輻輳の影響はまだ不明
  - 残念ながらAS701経由の良い計測が見つけられていない

事象の推定

- 12:22JST頃
  - AS701向けのピアで経路広報ポリシを何か変更
  - 内部経路が全てAS701に漏れ出す
    - 本来はEBGPに広報しない経路
    - 他のEBGPピアから聞いた経路
    - AS15169内部用途な経路
- 12:33JST頃
  - 問題のあったピアで経路ポリシを修正、もしくは該当ピアをshutdown
    - ^701 15169$な追加のBGP UPDATEが見えているので、何らかAS701内部で見えるBGP属性値が変更になっているはず

対策案１：経路フィルタ

1. AS701がAS15169向けに受信の経路フィルタやってれば良かった
   - IRRで!gAS15169引いたら、結構しびれるオブジェクト群だね
   - AS PATHベースのフィルタでも大丈夫だったかも
2. AS15169が多段の安全策を持っていれば良かった？
   - 自動管理になればなるほど、多段の安全策もさくっと通り抜けるかもね

対策案２：Secure BGP?

1. Secure BGPでPath Validationできていれば良かった？
   - 今回、広報の隣接関係自体は正当。トランジット対象かどうかみたいな検証ってPath Validationでできる予定なんだっけ？ -> できなさそう
2. RPKIのROAを使ったOrigin Validationで Maximum Length値を厳密に設定して検証できていれば良かった？
   - 今回、Origin側が細かい経路を特定のピア(AS15169)にだけは広報していたとすると、そもそも Maximum Length値もそれに合せて緩やかにしないといけない気がするので、やっぱり駄目だったかも

対策案３：経路制御ポリシ

1. 特定のピアだけ細かい経路でトラヒック制御するのがイケてない？
   - みんなが全ピアに同じポリシで経路広告していれば恐らく影響はかなり限定的だったはず