Lesson Learned from 25th Aug

Matsuzaki ‘maz’ Yoshinobu

<maz@iij.ad.jp>

Observed incident

- 25/Aug/2017 03:22UTC (12:22JST)
  - AS15169 started transiting other ISPs' IPv4 prefixes
  - Mostly de-aggregated prefixes usually not seen in DFZ
    - Traffic to those prefix were routed through US according to the announcements
  - ISPs started to receive many complaints from customers
- 25/Aug/2017 03:33UTC (12:33JST)
  - AS15169 withdrawn those announcements

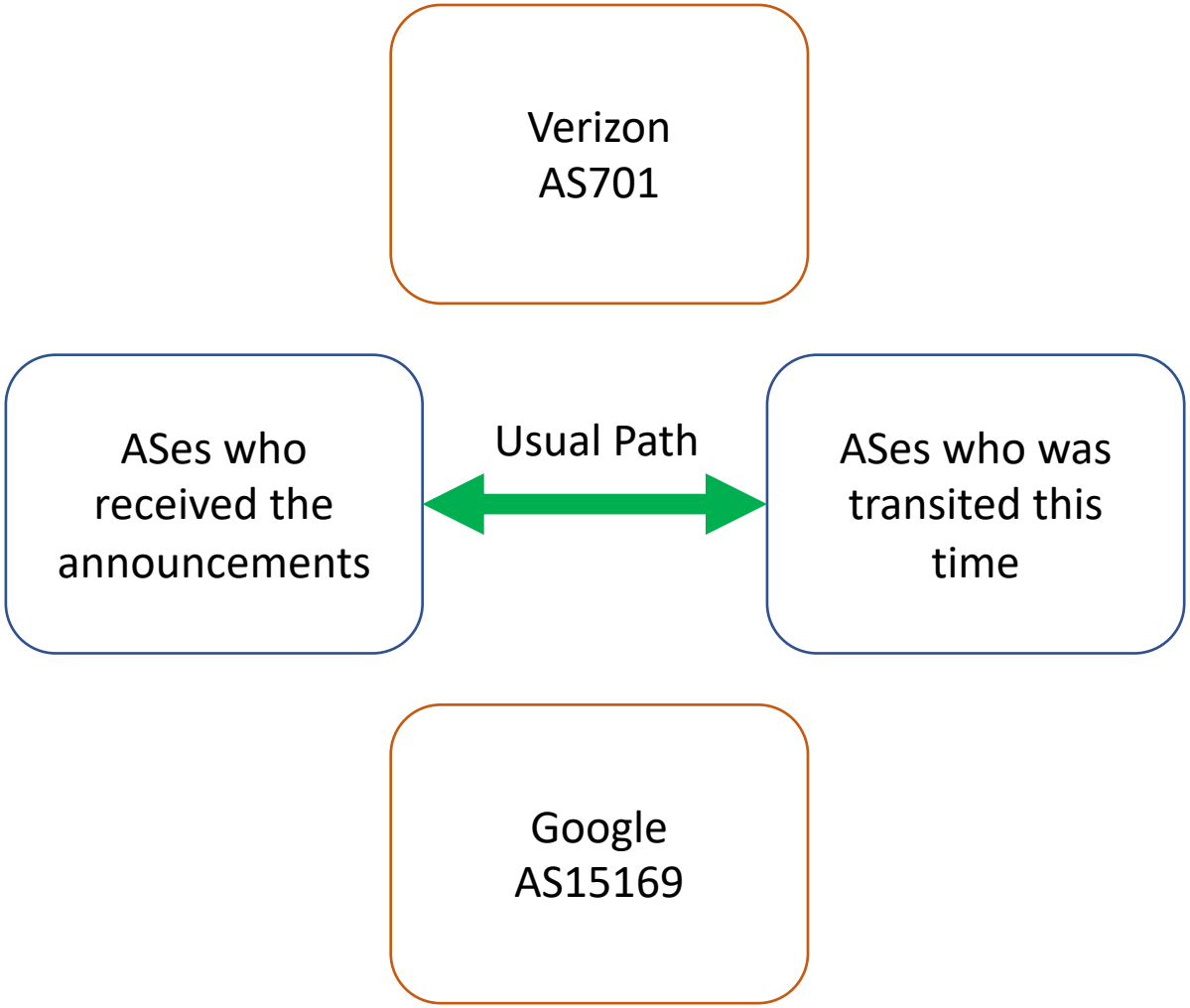
The BGP announcements

- The prefixes
  - About 110K prefixes total (including 25K Japanese ones)
    - From /10 to /24 (about half of them are /24s)
    - Mostly de-aggregated prefixes usually not seen in DFZ
- AS PATH looks like “701 15169 <Usual AS PATH>”
  - The origin AS looks correct
  - We didn’t see the announcements on the direct peering sessions with AS15169
- Transited ASes
  - About 7K ASes total (including 89 Japanese ASes)

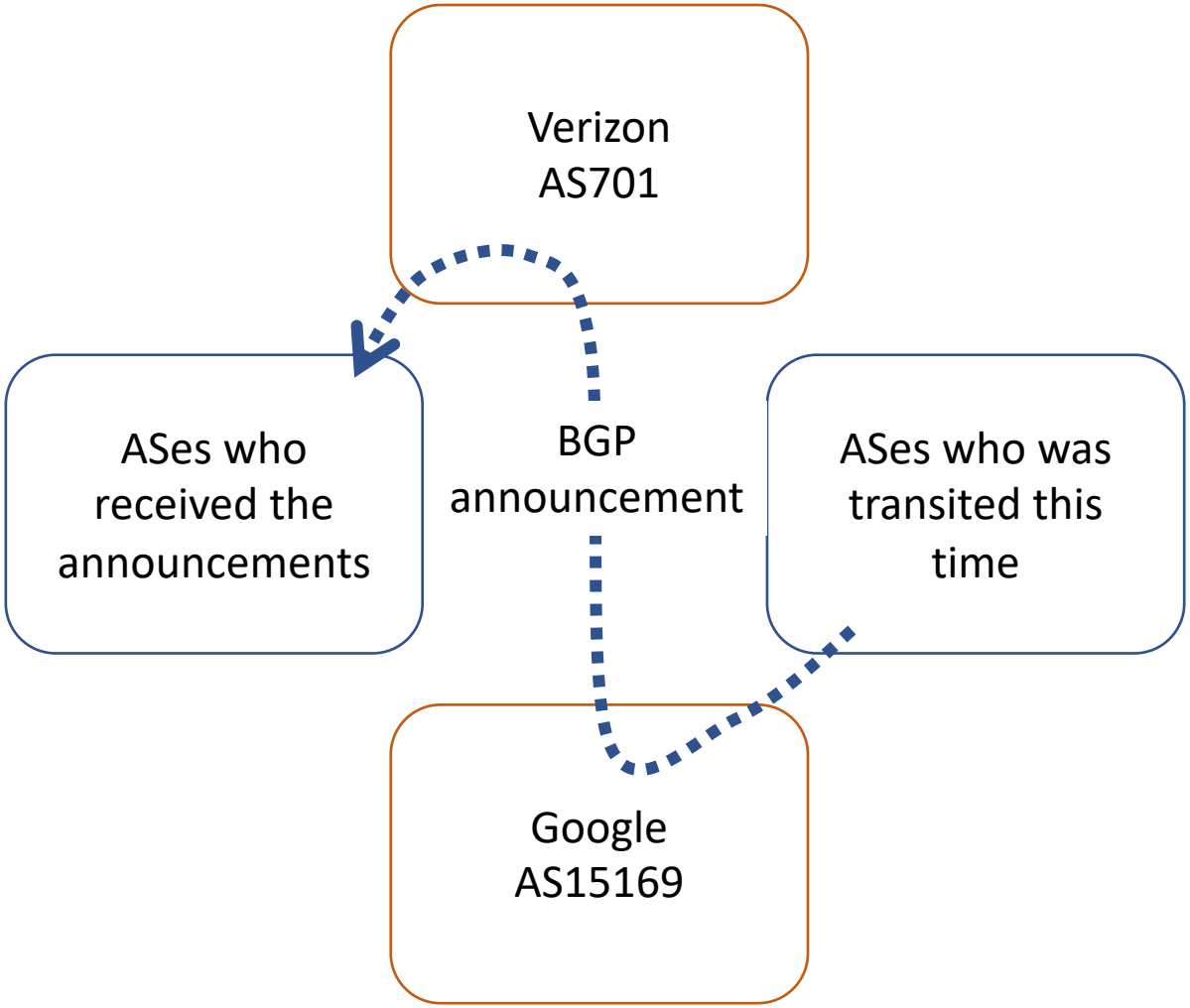
Other AS15169 originating prefixes

- We also observed unusual prefixes originating from AS15169 during the incident
- De-aggregated prefixes
  - AS15169 and its downstreams'
  - 654 prefixes
- IXP segments
  - 78 prefixes
- I can't tell, but probably these are IXP segments
  - 2 prefixes

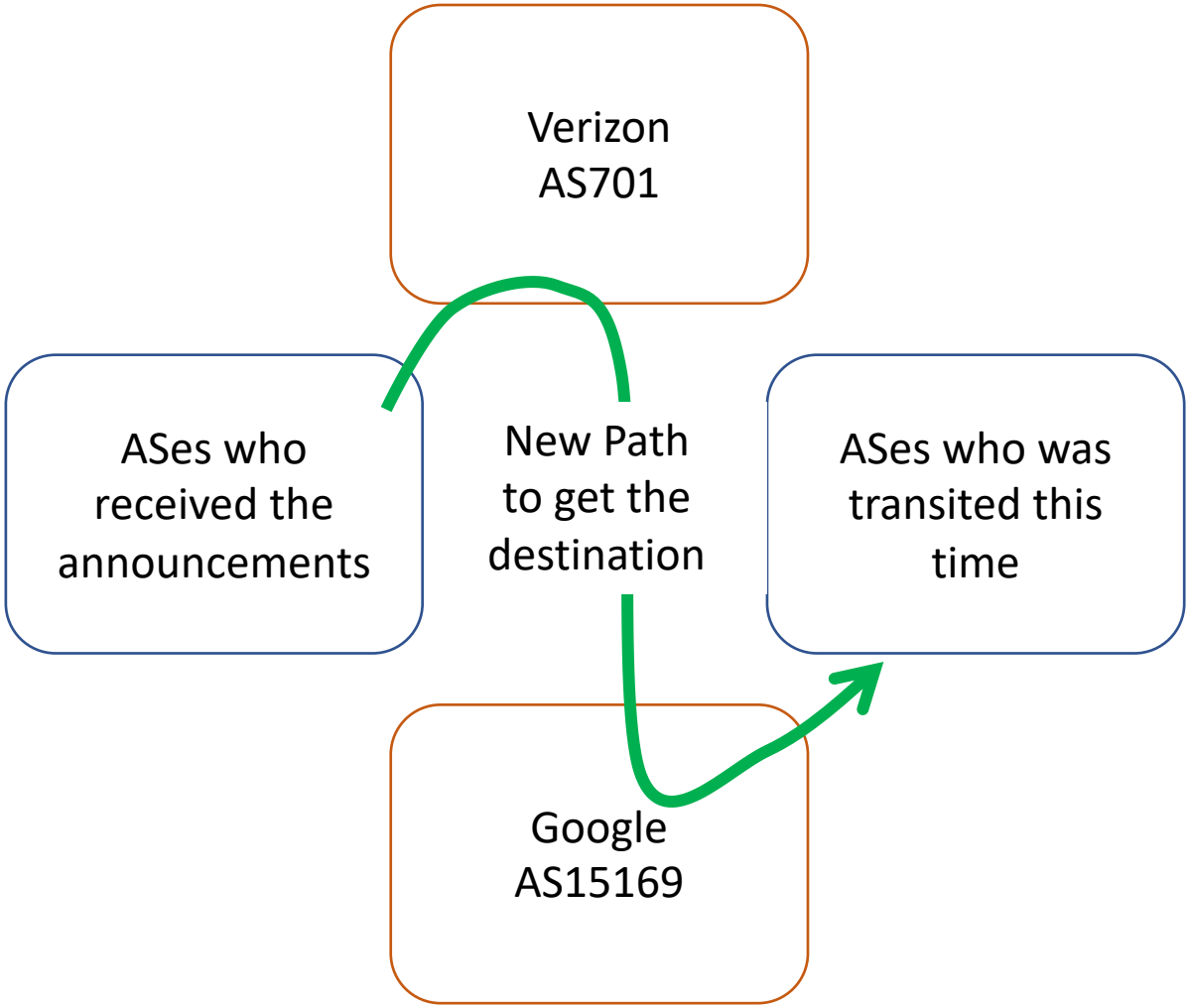
Usual forwarding path



Unexpected transit happened



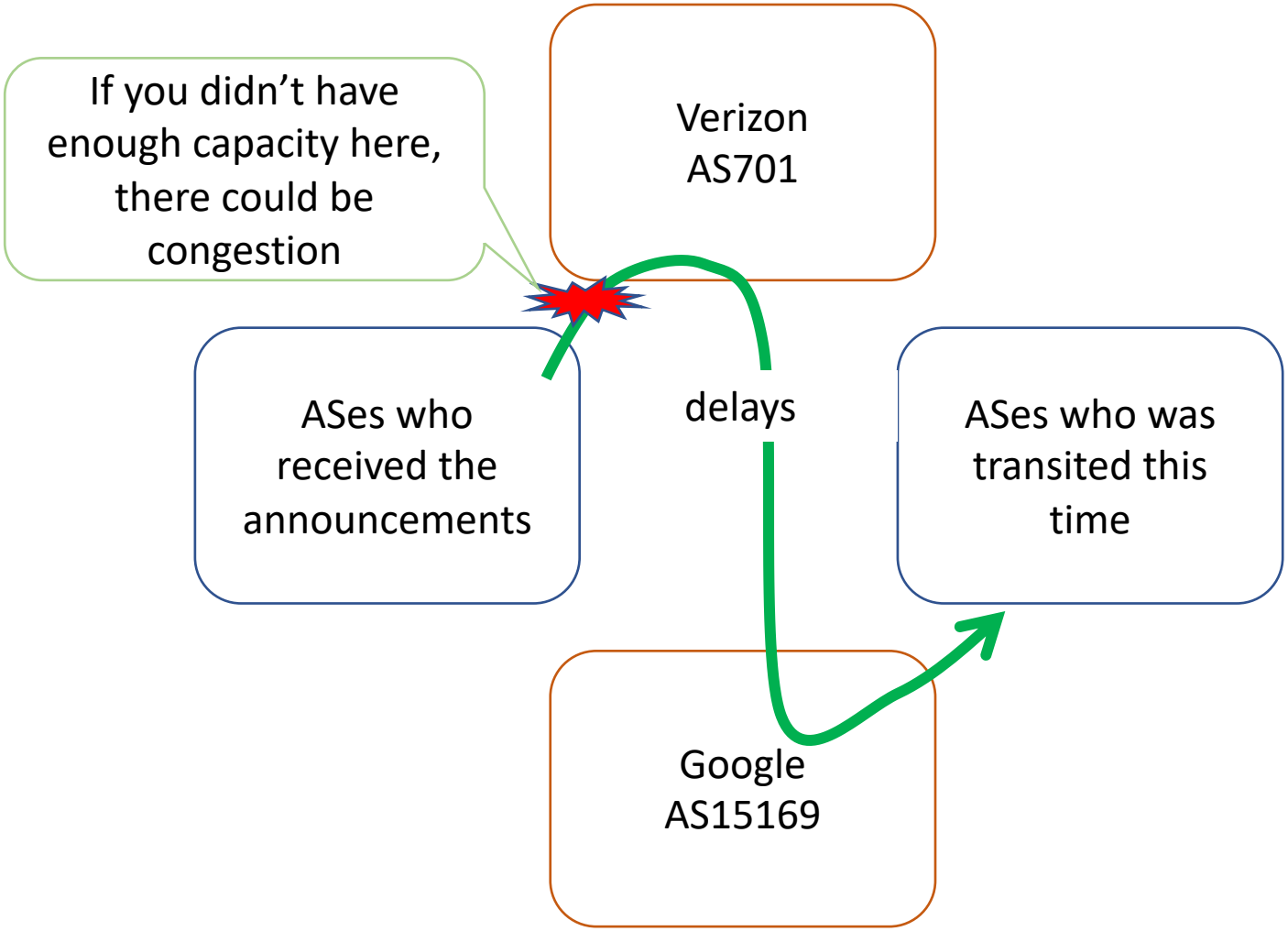
Traffic flowed accordingly



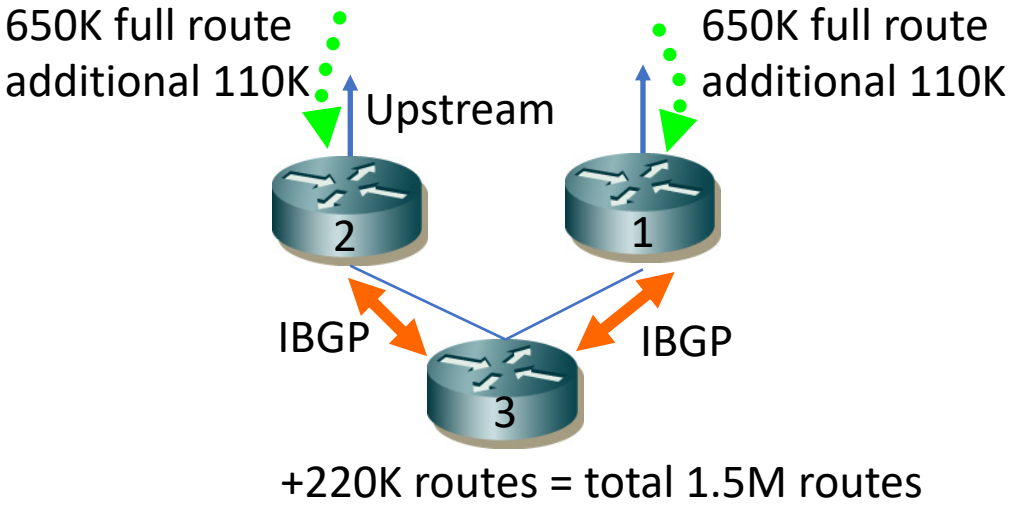
Estimation of the effects

- Traffic were routed through US to get those announced destination
  - Increased delay
  - Might cause congestion
- Traffic over IXP might be affected if you are careless
  - In case the announcements affected your BGP nexthop lookup for IXP peering partner
- Routers got unstable because of the additional 110K announcements
  - Poor routers

Possible delay and congestion



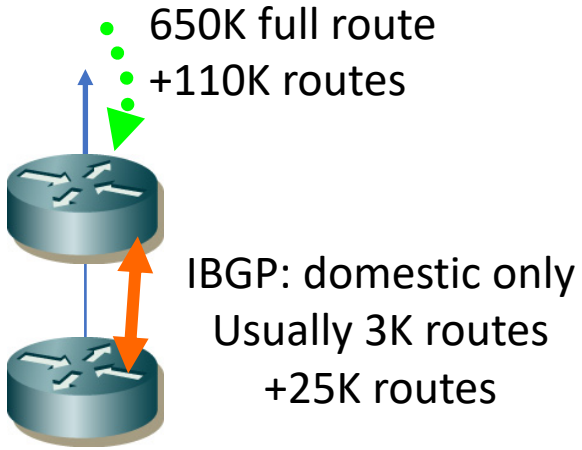
Additional 110K prefixes



- We have 650K routes in DFZ
  - Internal routers need to handle 1.3M usually
- The internal router needed to handle 1.5M RIB in this configuration
  - Or +110K prefixes might affect poor routers simply

Even you have route reduction

- Some routers can't handle full BGP table anymore
  - Some ops feed domestic prefixes only to those poor routers to reduce FIB/RIB size
- If you picked 'domestic routes' by AS PATH like _4713_, the router received additional 25K routes
  - almost 10times bigger than usual
- Those poor routers might cause such a long recovery time



Transited ASes

- About 7000 ASes total
  - Including 89 Japanese ASes
- # of transited prefixes per AS
  - OCN/AS4713 was transited the most

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

AS4713 originating prefixes

Usual (78prefixes were affected)

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

Additional prefixes that were transited

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

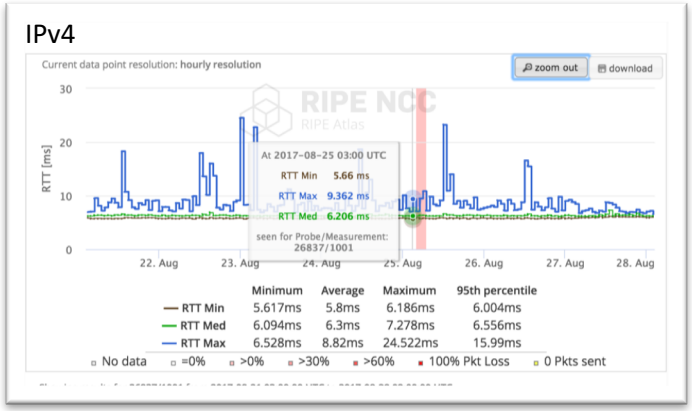
RIPE Atlas Probe

- A measurement infrastructure run by RIPE NCC
  - Probes are distributed around the world
  - It has common measurements against popular sites like root nameservers
- Picked probes in AS4713 to evaluate the impact
  - AS4713 internal: k.root-servers.net
  - Domestic and affected: m.root-servers.net
  - International: ctr-ams02.atlas.ripe.net

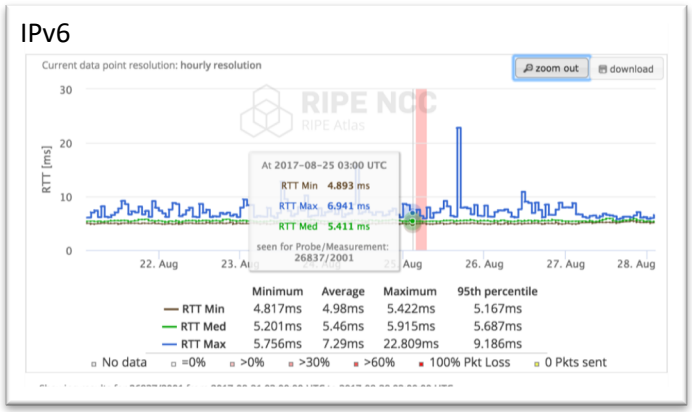
OCN/AS4713 Internal

Probe26837

IPv4

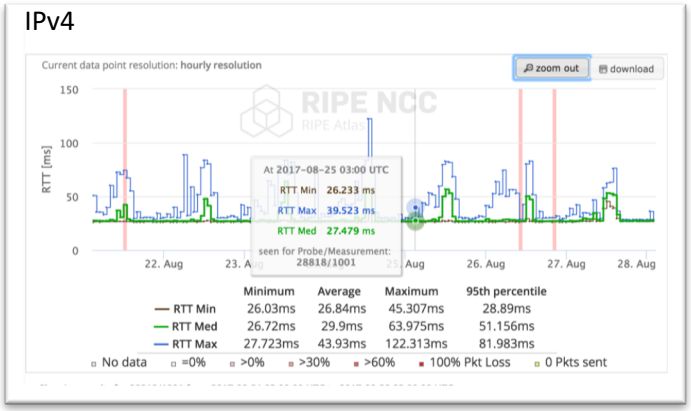


IPv6

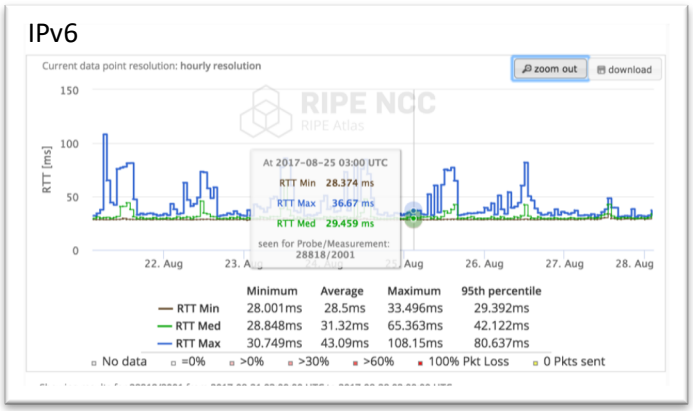


Probe28818

IPv4



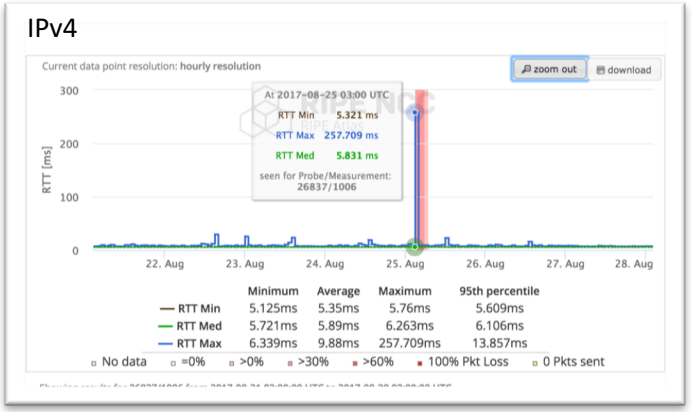
IPv6



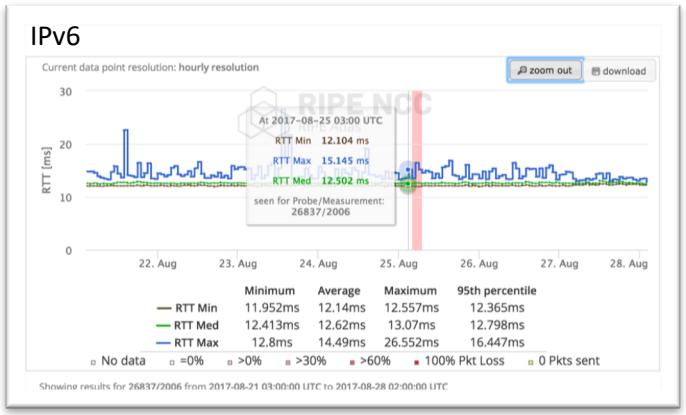
OCN/AS4713 and domestic

Probe26837

IPv4

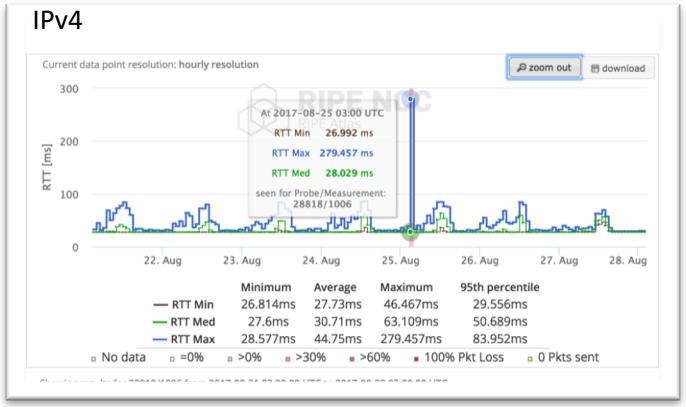


IPv6

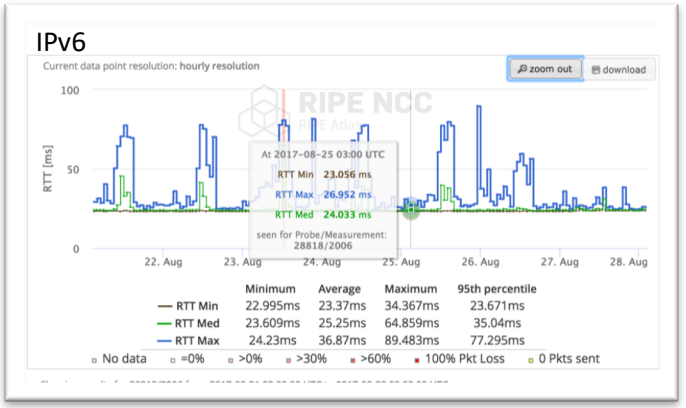


Probe28818

IPv4



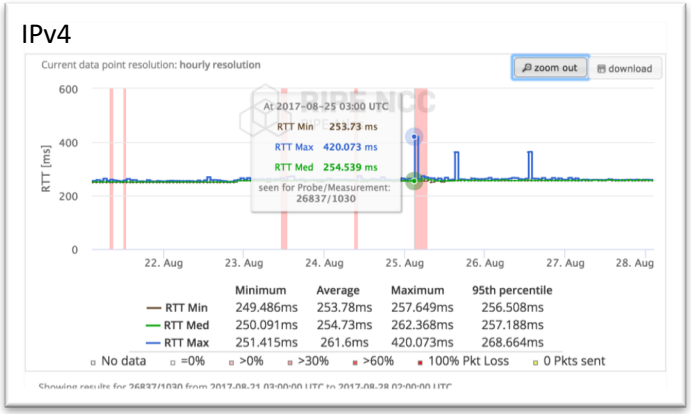
IPv6



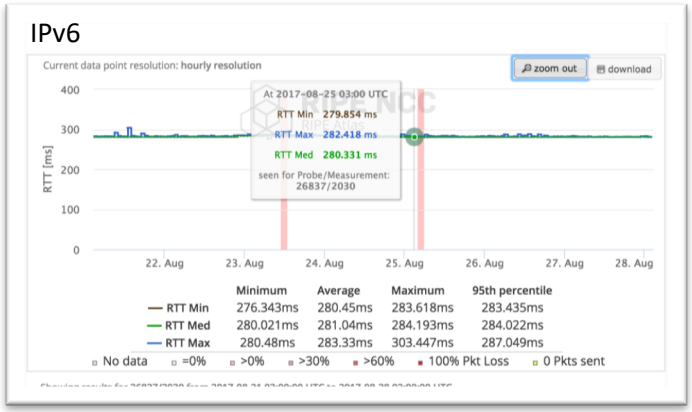
OCN/AS4713 and international

Probe26837

IPv4

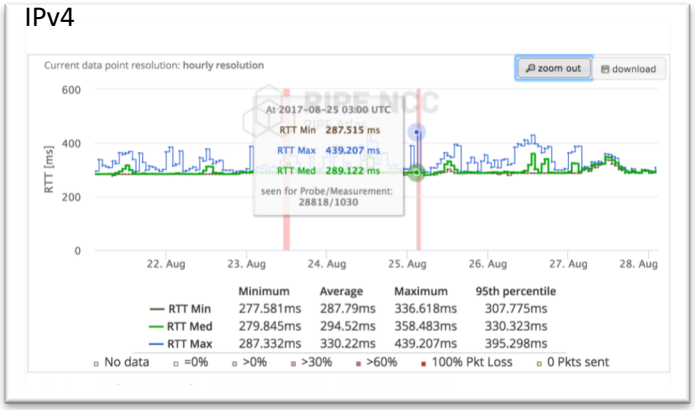


IPv6

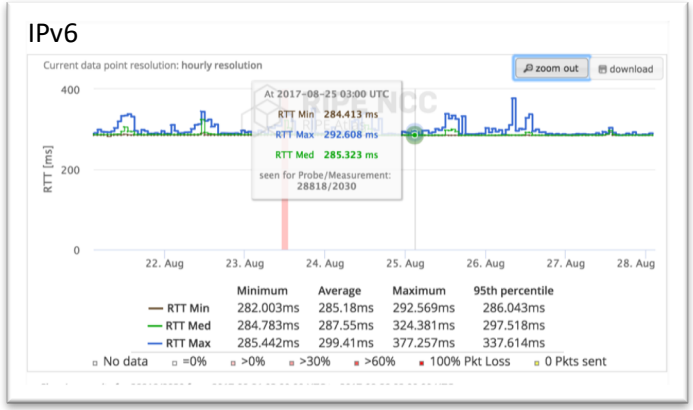


Probe28818

IPv4



IPv6



From the probes' view

- Domestic and international IPv4 communications were affected a bit
  - delays and some losses
  - at lease they had reachability
- No direct impact for IPv6 communications
  - The announcements were IPv4 only this time
  - probe26837 probably had a congestion point close to the probe as we can observe losses regardless of destination

It created hysteria in Japan

- Users expect ‘the perfect’ services here
- Lunch break is one of the major peak times of internet use now days
- Big names make the news $valuable for presses
- The government is getting sensitive about infrastructure security as they are preparing for 2020

Tweets

- # of tweet that has “network disturbance” keywords

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

BGP is a hiding protocol

- Some prefixes can be seen at specific ASes only
  - Only the best path can be propagated
  - We have many route filtering to enforce our routing policy for peering, downstream and upstream partners
- Feed your full BGP table to public route archives
  - We need more visibility from different instances
  - 3rd party can check routing based on such data

Possible counter measures: filtering

- Route filtering
  - AS701 should have a decent filtering there
  - prefix based or AS PATH based
- Maximum prefix setting
  - Inbound setting is getting common
  - Outbound could be useful as well
  - ‘shutdown’ is radical operation for some peering sessions

Possible counter measures: Secure BGP

- Path Validation by Secure BGP
  - It seems the neighboring relationship looks correct in this particular case, so we couldn’t prevent that
- Maximum prefix length of ROA
  - If the ISPs want to announce de-aggregate prefixes to neighbors, they can not set it strictly

Possible counter measures: TE

- Avoid de-aggregated prefixes to do traffic engineering
  - We could minimalize the impact, if ASes announced the same prefixes on all of their EBGP sessions
- Use /25 or longer for traffic engineering :P
  - Many ops tend to filter prefixes longer than /24, so these shouldn’t be able to propagate :P

Possible counter measures: detection and communication

- Anomaly detection
  - AS PATH or traffic
  - Expecting too many false positives
- Better communication among operators
  - Multiple channel
  - Trusted personal and/or business relationships

Summary

- The BGP announcements affected some users' IPv4 communication a bit
  - Delay and losses depending on source and destination
  - But that should be fixed around in 20min at the most
- That might trigger other stuff
  - Poor routers might go unstable
    - These should be upgraded in advance
  - Users might react against the incident
    - Checking service availabilities, try to connect services
  - I suppose those made the effect looks !bigger!