Internet asymmetric routing and BGP traffic engineering challenges

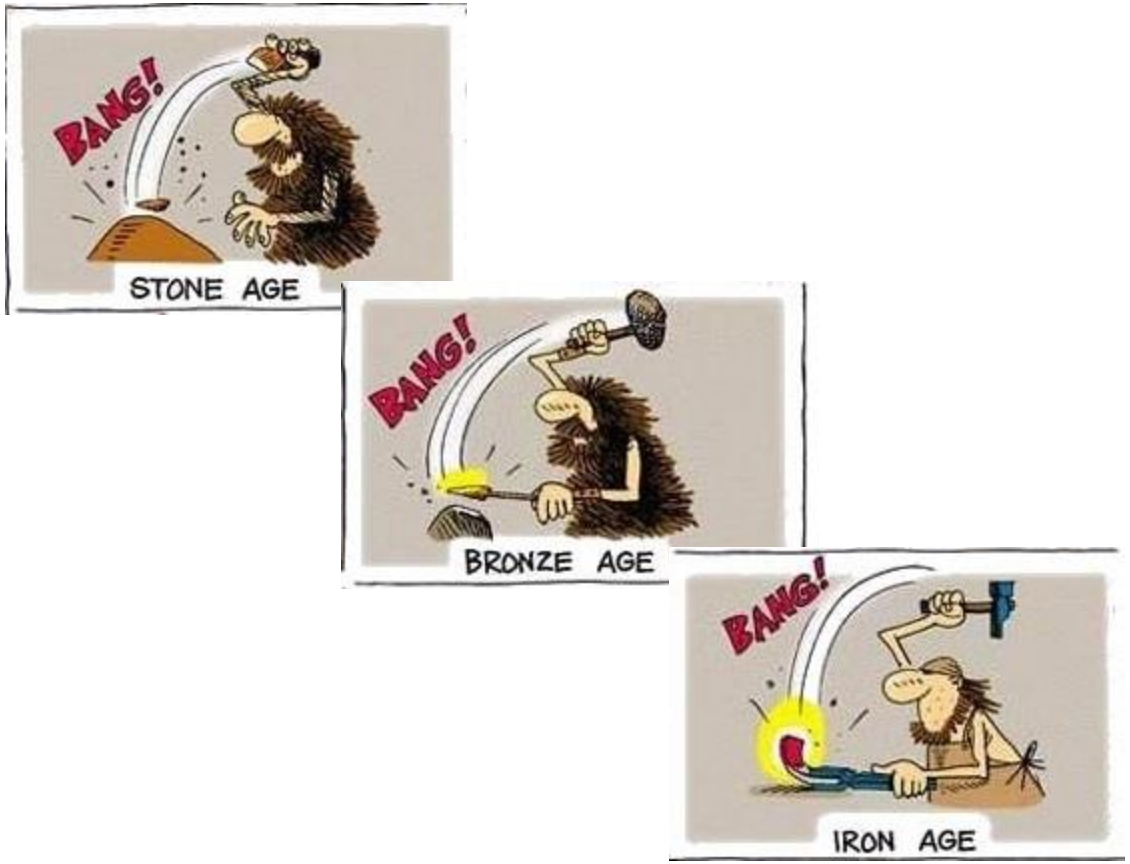
Ulsbold Enkhtaivan
IP network manager (MobiCom Corporation, Mongolia)

Introduce myself

- IP network manager at Mobicom Corporation,
- 15 years of experience in network engineer
- Program committee and founding member of mnNOG.
- APNIC RCT since 2022

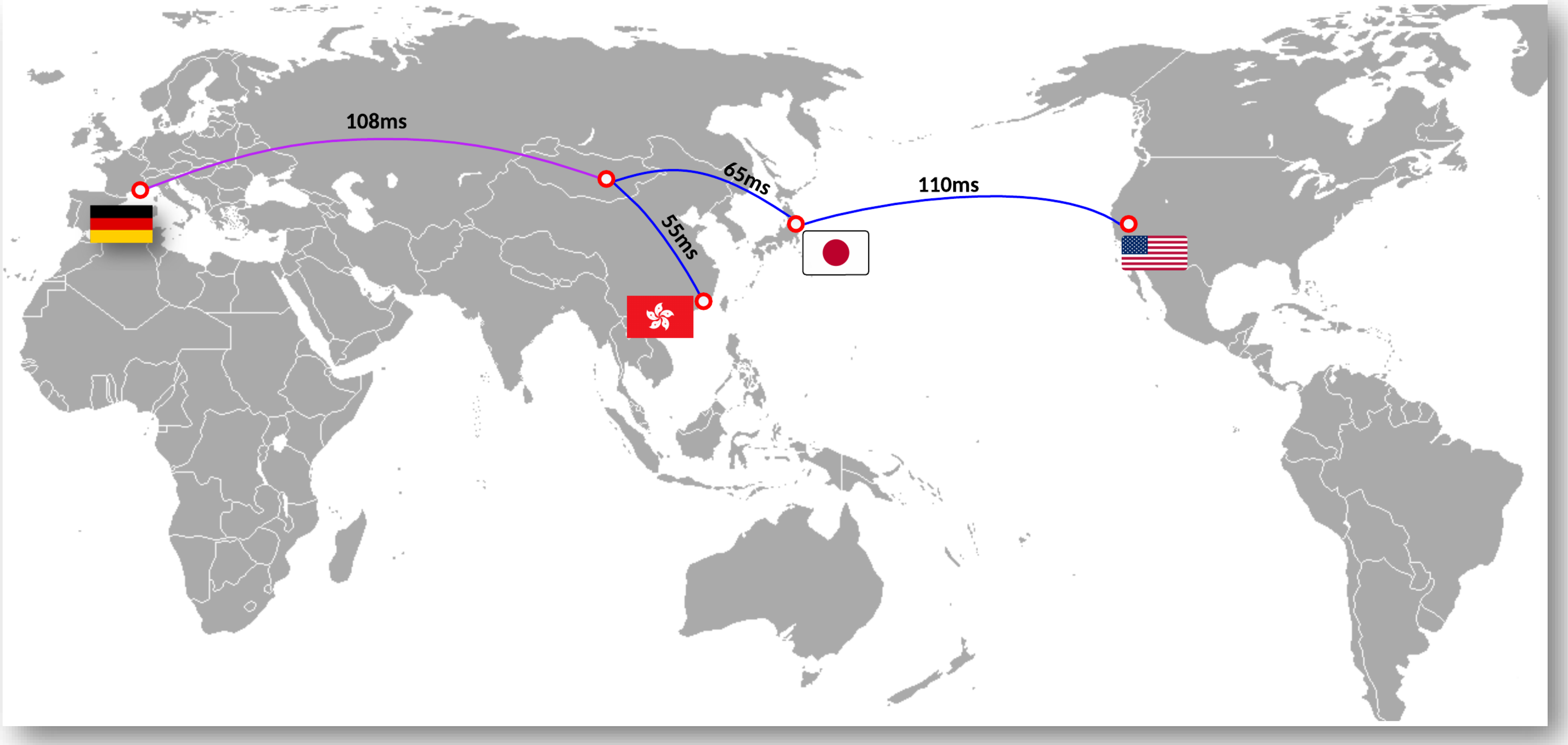
Agenda

- Early
- Middle
- Now

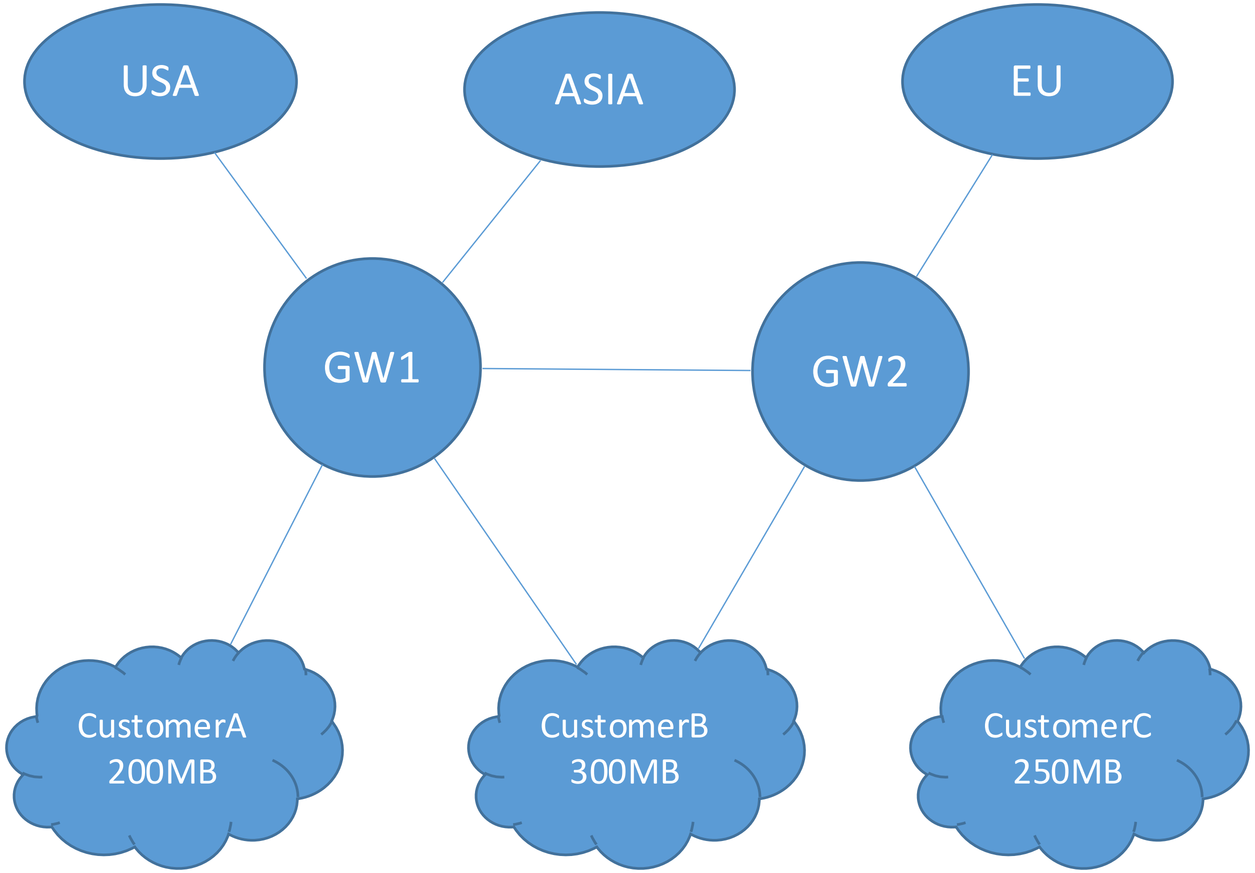


Landlocked country

- Connecting to good internet connectivity need to reach far far away.



Early in our network



- For inbound traffic
  - Customer A advertised only USA
  - Customer B advertised only ASIA
  - Customer C advertised only EU
- For outgoing traffic using **PBR(policy based routing)**
  - Customer A goes to only USA
  - Customer C goes to only ASIA
  - Customer B goes to only EU

- There were no asymmetric routing, easy traffic engineering for administrator,

Why we did that

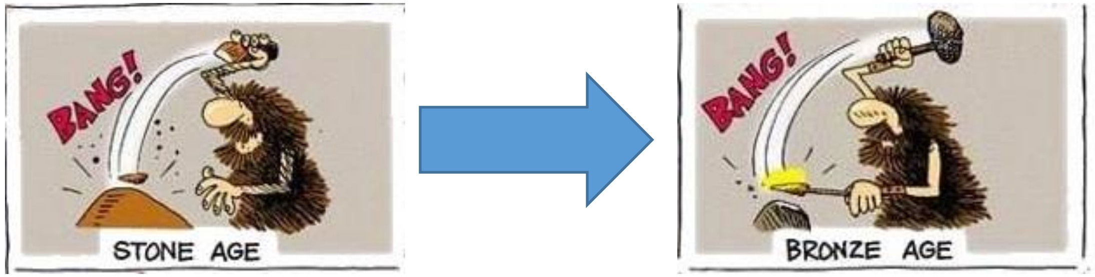
- Only possible reason was upstream link's bandwidth to low. STM1 or STM4 links.
- No CDN's inside in our network.

- issues,
  - Customers complained our internet quality. High delay etc..
  - No redundancy (during in failure we did re-configure all advertisement and PBR one by one)
  - Policy based routing issues. It creates complexity

Early to Middle

- Then we advertise all customer to all upstream, Remove all **PBR’s**.
- But there is no any policy on BGP, all configuration were like default.

- That day we just fix our redundancy.
- Delay problem still there.
  - Need to apply some policy to our network



For Outgoing traffic

- Used APNIC and RIPE database to differentiate them.
- Used regex format and create as-path access-list to match ASN’s
- For example:

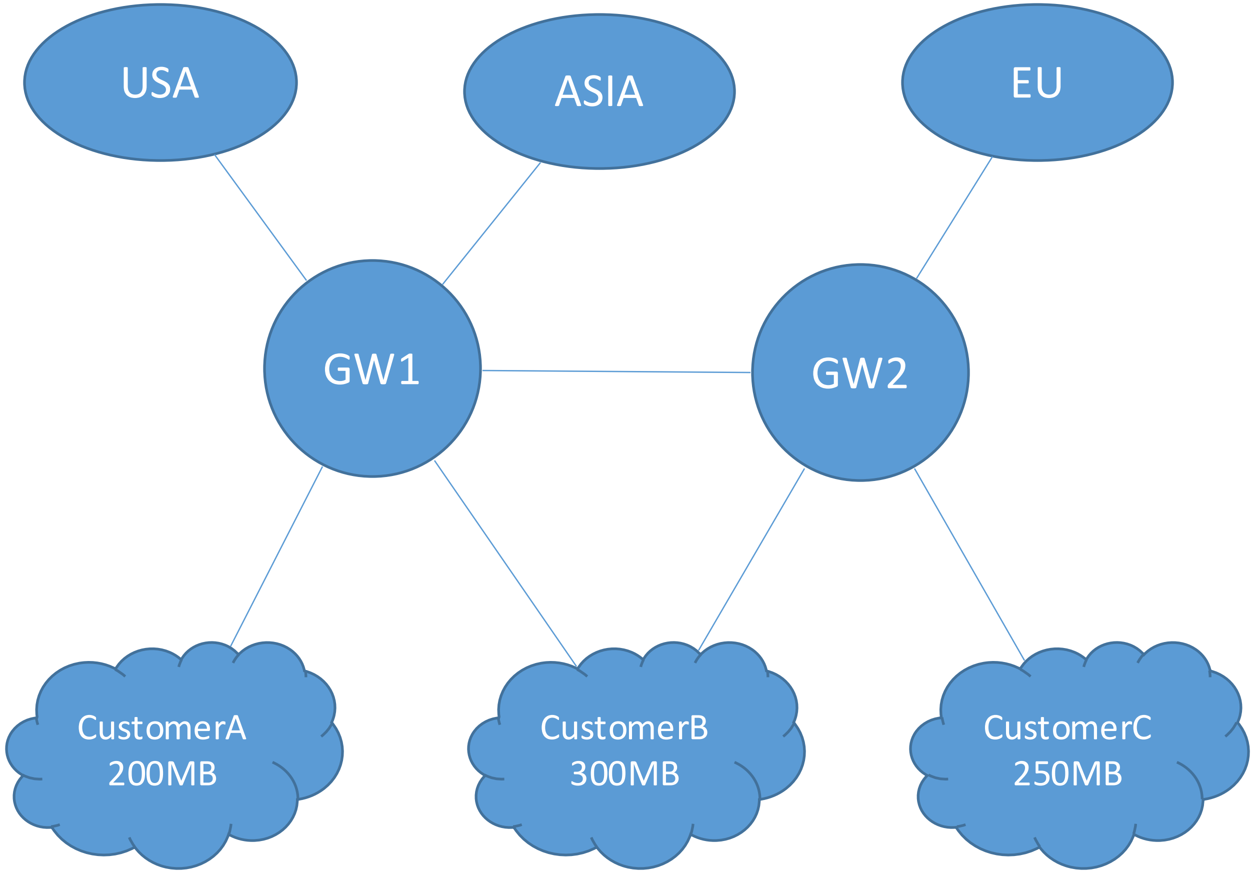
- Example for IOS-XR

```
as-path-set ASIA
  ios-regex '_(3784|3786|3787|3813)$',
  ios-regex '_(3825|3836|3839|3840)$',
  ios-regex '_(3929|3969|3976|4007)$',
  ios-regex '_(4040|4049|4058|4060)$',
  ios-regex '_(4142|4134|4158|4174|4175)$',
  ios-regex '_(4197|4202|4251|4274)$',
  ios-regex '_(4352|4381|4382|4431)$',
  ios-regex '_(4433|4434|4515|4528)$',
  ios-regex '_(4538|4594|4605|4961)$',
  ios-regex '_(5017|5018|5051|5085)$',
  ios-regex '_(5087|5709|6068|6163)$',
  ios-regex '_(6262|6619|6648|7131)$',
end-set
```

- Example for IOS

```
ip as-path access-list 20 permit _(3784|3786|3787|3813)$
ip as-path access-list 20 permit _(3825|3836|3839|3840)$
ip as-path access-list 20 permit _(3929|3969|3976|4007)$
ip as-path access-list 20 permit _(4040|4049|4058|4060)$
ip as-path access-list 20 permit _(4142|4134|4158|4174|4175)$
ip as-path access-list 20 permit _(4197|4202|4251|4274)$
ip as-path access-list 20 permit _(4352|4381|4382|4431)$
ip as-path access-list 20 permit _(4433|4434|4515|4528)$
ip as-path access-list 20 permit _(4538|4594|4605|4961)$
ip as-path access-list 20 permit _(5017|5018|5051|5085)$
ip as-path access-list 20 permit _(5087|5709|6068|6163)$
ip as-path access-list 20 permit _(6262|6619|6648|7131)$
```

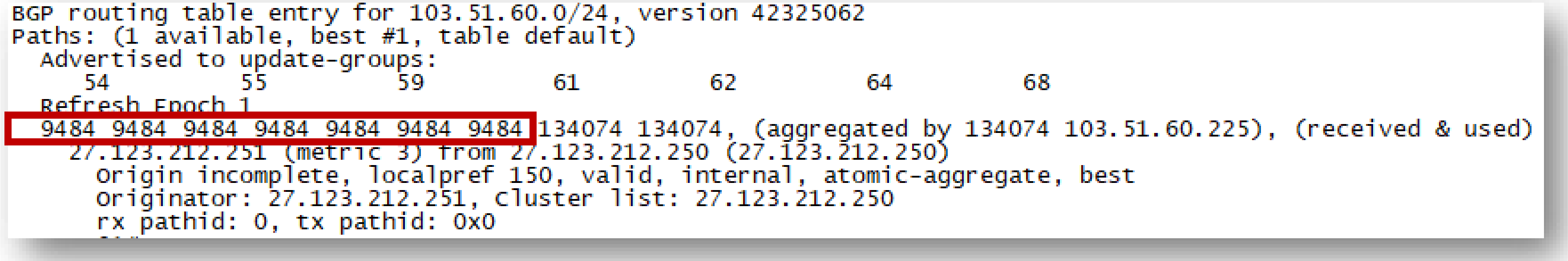
For Outgoing traffic



- Asian ASN's matched in Asian GW and set LP to higher
- EU ASN's matched in EU GW and set LP to higher
- Didn't matched ASN's goes thru USA
  - As-path list was about more than **550 lines** in our configuration
  - Some special case we uses prefix list, that was also more than **180 lines** in our configuration,

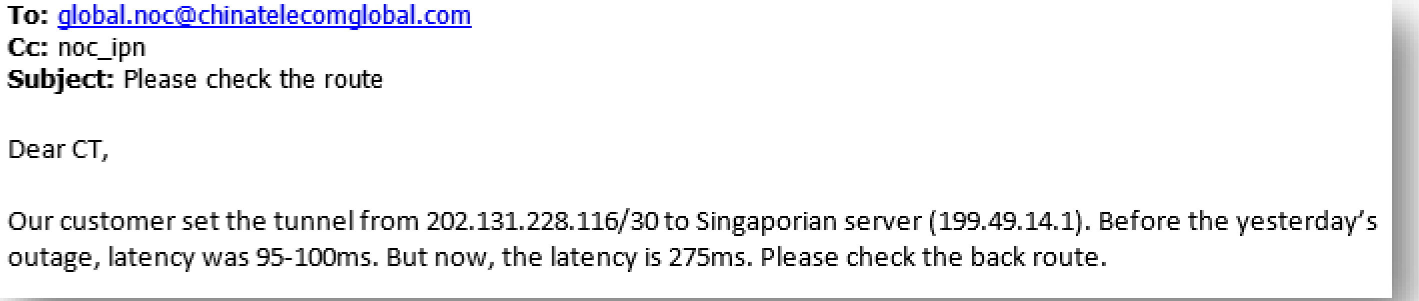
For inbound traffic

- We trying to use as-path prepend our some customers prefixes.
- But this technic is not efficient. Own as-path prepend impact’s all incoming traffic,



```
BGP routing table entry for 103.51.60.0/24, version 42325062
Paths: (1 available, best #1, table default)
  Advertised to update-groups:
     54         55         59         61         62         64         68
  Refresh Epoch 1
  9484 9484 9484 9484 9484 9484 9484 134074 134074, (aggregated by 134074 103.51.60.225), (received & used)
    27.123.212.251 (metric 3) from 27.123.212.250 (27.123.212.250)
      Origin incomplete, localpref 150, valid, internal, atomic-aggregate, best
      Originator: 27.123.212.251, Cluster list: 27.123.212.250
      rx pathid: 0, tx pathid: 0x0
```

- Also we requested to our upstream provider’s for change return path.
  - Due to we can’t influence the our upstream or tier1 providers routers RIB.
- This method is not for optimal for us.
  - Wasting time to conversation with upstream NOC.
  - Mostly result was unsuccessful.

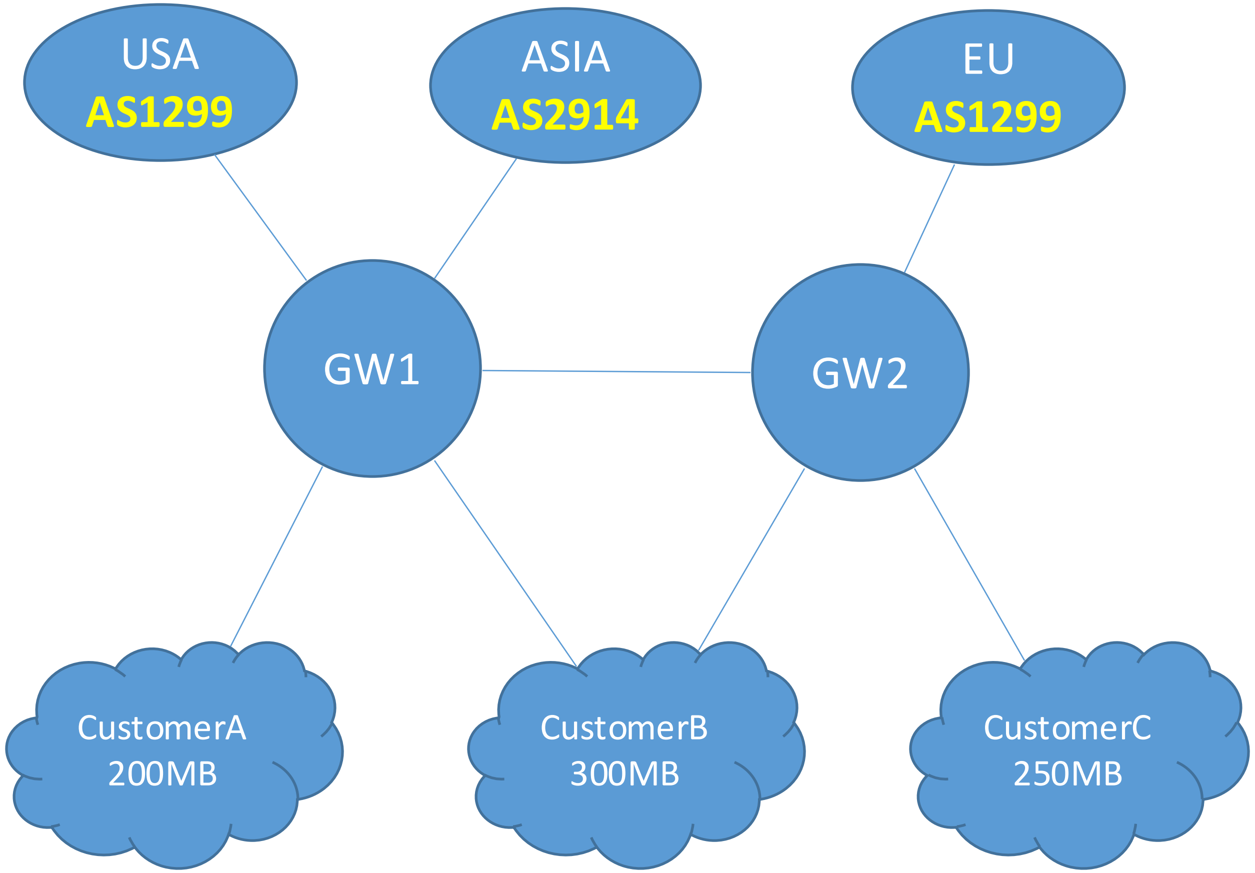


To: global.noc@chinatelecomglobal.com
Cc: noc_ipn
Subject: Please check the route

Dear CT,

Our customer set the tunnel from 202.131.228.116/30 to Singaporian server (199.49.14.1). Before the yesterday’s outage, latency was 95-100ms. But now, the latency is 275ms. Please check the back route.

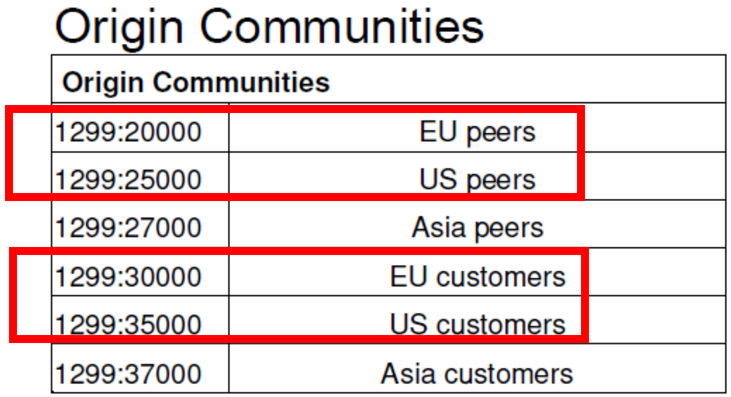
Then



- We use the BGP community of our upstream providers,
- That's powerful and popular tool for BGP traffic engineering.

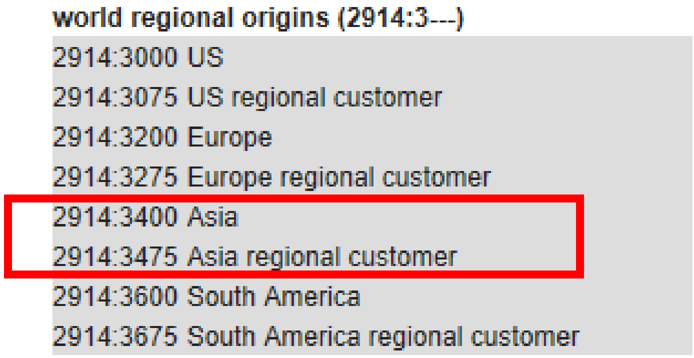
Outgoing traffic

- Our EU and USA upstream provider is (Telia), current name is Arelion. ASIA’s NTT
- We are using our upstream’s BGP origin communities for outgoing traffic
- *550 line as-path and 180 prefix list line replaced to just 3 line community*



Origin Communities

Origin Communities	
1299:20000	EU peers
1299:25000	US peers
1299:27000	Asia peers
1299:30000	EU customers
1299:35000	US customers
1299:37000	Asia customers



world regional origins (2914:3---)

2914:3000	US
2914:3075	US regional customer
2914:3200	Europe
2914:3275	Europe regional customer
2914:3400	Asia
2914:3475	Asia regional customer
2914:3600	South America
2914:3675	South America regional customer

- Example for IOS

```
ip community-list 11 permit 1299:20000
ip community-list 11 permit 1299:30000

route-map EU-IN permit 10
match community 11
set local-preference 300
route-map EU-IN permit 20
set local-preference 50
```

```
ip community-list 12 permit 1299:25000
ip community-list 12 permit 1299:35000

route-map USA-IN permit 10
match community 12
set local-preference 300
route-map USA-IN permit 20
set local-preference 50
```

```
ip community-list 13 permit 2914:3400
ip community-list 13 permit 2914:3475

route-map ASIA-IN permit 10
match community 13
set local-preference 300
route-map ASIA-IN permit 20
set local-preference 50
```

Outgoing traffic

- Example for IOS-XR

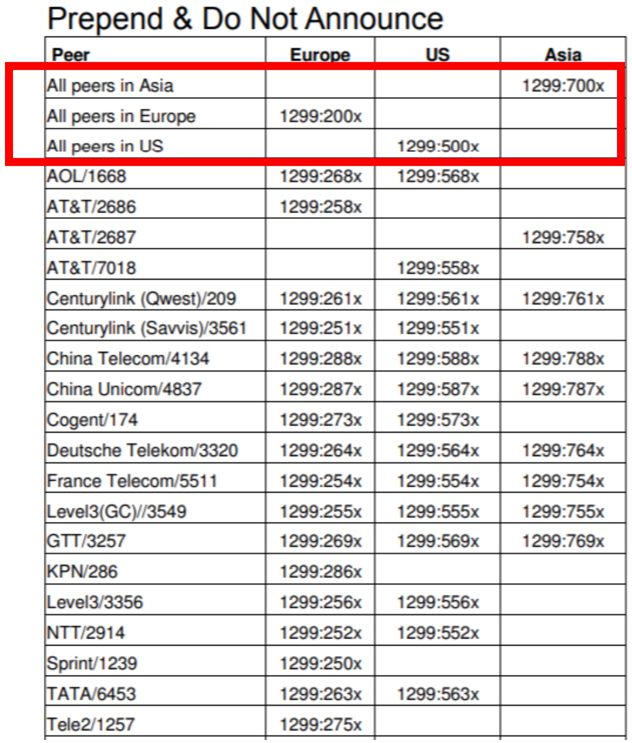
```
community-set TELIA-EU
  #EU-peers
  1299:20000,
  #EU-customers
  1299:30000
end-set

route-policy EU-IN
  if community matches-any TELIA-EU then
    set local-preference 300
  else
    set local-preference 50
  endif
end-policy
```

```
community-set TELIA-USA
  #USA-peers
  1299:25000,
  #USA-customers
  1299:35000
end-set

route-policy USA-IN
  if community matches-any TELIA-USA then
    set local-preference 300
  else
    set local-preference 50
  endif
end-policy
```

```
community-set NTT-ASIA
  #Asia
  2914:3400,
  #Asia-customers
  2914:3475
end-set

route-policy ASIA-IN
  if community matches-any NTT-ASIA then
    set local-preference 300
  else
    set local-preference 50
  endif
end-policy
```

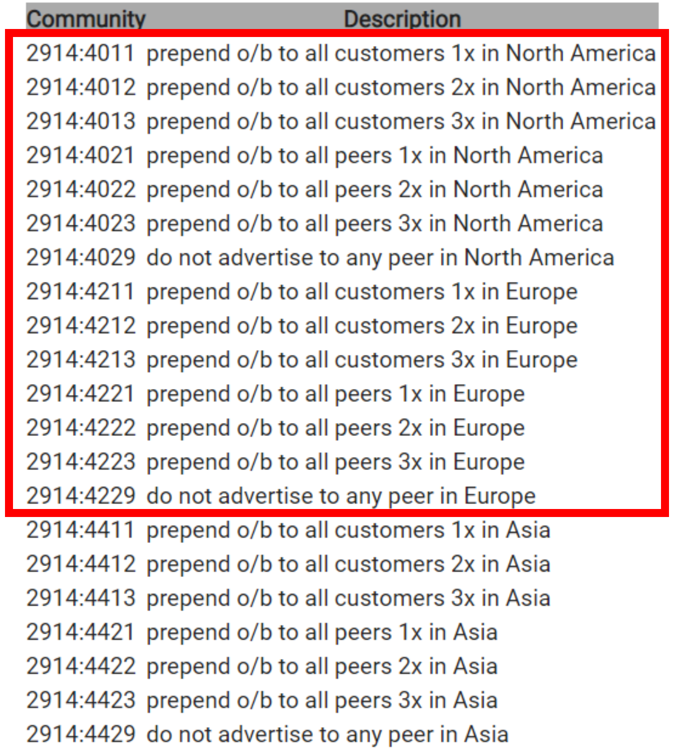
incoming traffic

- We don’t need to contact our upstream for change the routing,
- We have using upstream predefined bgp communities to prepend or deny our announcements to specific geolocations and *peer or their customers.*
  - NTT and Teliacarriers action community



Prepend & Do Not Announce

Peer	Europe	US	Asia
All peers in Asia			1299:700x
All peers in Europe	1299:200x		
All peers in US		1299:500x	
AOL/1668	1299:268x	1299:568x	
AT&T/2686	1299:258x		
AT&T/2687			1299:758x
AT&T/7018		1299:558x	
Centurylink (Qwest)/209	1299:261x	1299:561x	1299:761x
Centurylink (Savvis)/3561	1299:251x	1299:551x	
China Telecom/4134	1299:288x	1299:588x	1299:788x
China Unicom/4837	1299:287x	1299:587x	1299:787x
Cogent/174	1299:273x	1299:573x	
Deutsche Telekom/3320	1299:264x	1299:564x	1299:764x
France Telecom/5511	1299:254x	1299:554x	1299:754x
Level3(GC)//3549	1299:255x	1299:555x	1299:755x
GTT/3257	1299:269x	1299:569x	1299:769x
KPN/286	1299:286x		
Level3/3356	1299:256x	1299:556x	
NTT/2914	1299:252x	1299:552x	
Sprint/1239	1299:250x		
TATA/6453	1299:263x	1299:563x	
Tele2/1257	1299:275x		



Community	Description
2914:4011	prepend o/b to all customers 1x in North America
2914:4012	prepend o/b to all customers 2x in North America
2914:4013	prepend o/b to all customers 3x in North America
2914:4021	prepend o/b to all peers 1x in North America
2914:4022	prepend o/b to all peers 2x in North America
2914:4023	prepend o/b to all peers 3x in North America
2914:4029	do not advertise to any peer in North America
2914:4211	prepend o/b to all customers 1x in Europe
2914:4212	prepend o/b to all customers 2x in Europe
2914:4213	prepend o/b to all customers 3x in Europe
2914:4221	prepend o/b to all peers 1x in Europe
2914:4222	prepend o/b to all peers 2x in Europe
2914:4223	prepend o/b to all peers 3x in Europe
2914:4229	do not advertise to any peer in Europe
2914:4411	prepend o/b to all customers 1x in Asia
2914:4412	prepend o/b to all customers 2x in Asia
2914:4413	prepend o/b to all customers 3x in Asia
2914:4421	prepend o/b to all peers 1x in Asia
2914:4422	prepend o/b to all peers 2x in Asia
2914:4423	prepend o/b to all peers 3x in Asia
2914:4429	do not advertise to any peer in Asia

incoming traffic

- Example for IOS

```
route-map EU-OUT permit 20
 match ip address prefix-list CUSTOMER-PREFIXES
set community 1299:7003 1299:5002
```

```
route-map US-OUT permit 20
 match ip address prefix-list CUSTOMER-PREFIXES
set community 1299:7003 1299:2002
```

```
route-map EU-OUT permit 20
 match ip address prefix-list CUSTOMER-PREFIXES
set community 2914:4013 2914:4023 2914:4123 2914:4223
```

- Example for IOS-XR

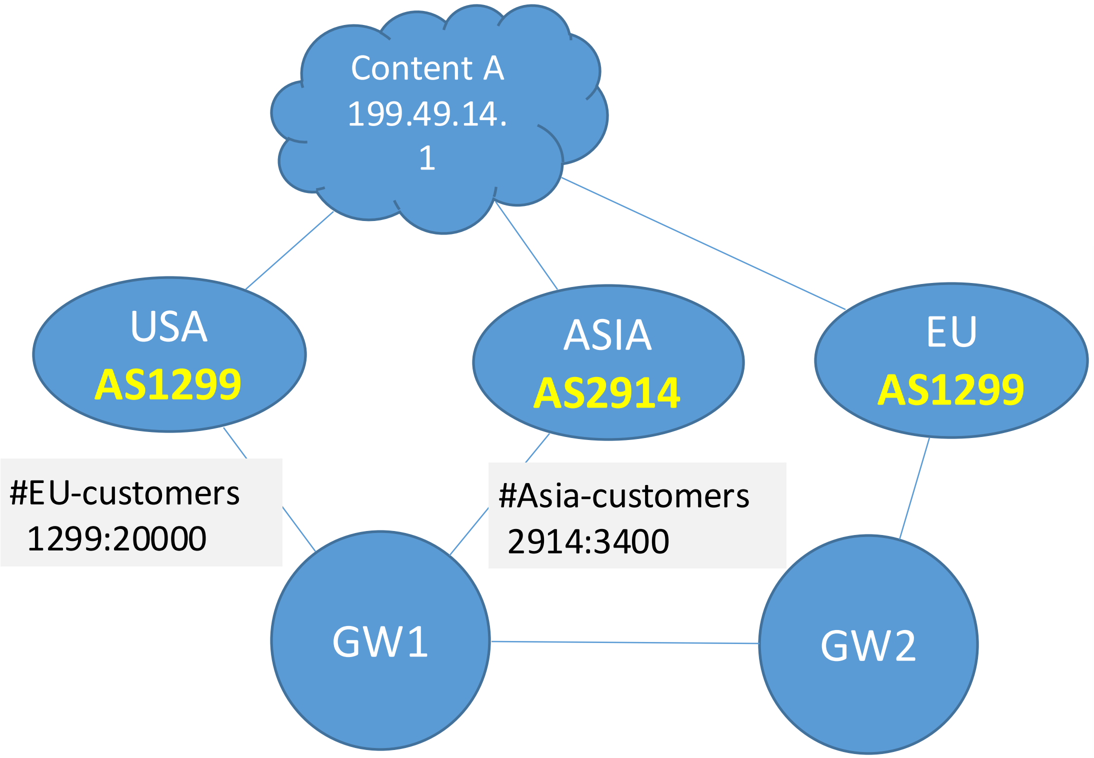
```
community-set PREPEND-TELIA-EU
  #Prepend-Asia3x
  1299:7003,
  # Prepend-US2x
  1299:5002
end-set

route-policy EU-OUT
  if destination in CUSTOMER-PREFIXES then
    set community-set PREPEND-TELIA-EU
  endif
end-policy
```

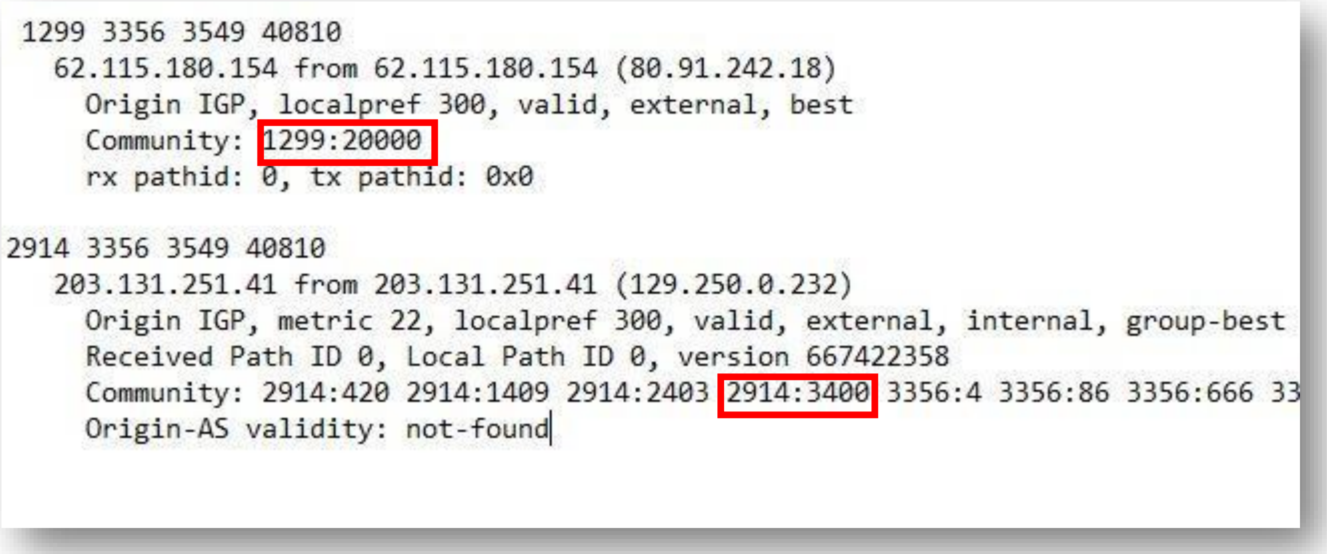
```
community-set PREPEND-TELIA-US
  #Prepend-Asia3x
  1299:7003,
  # Prepend-EU2x
  1299:2002
end-set

route-policy US-OUT
  if destination in CUSTOMER-PREFIXES then
    set community-set PREPEND-TELIA-US
  endif
end-policy
```

```
community-set PREPEND-NTT-ASIA
  #Prepend-US3x
  2914:4013,
  2914:4023,
  # Prepend-EU3x
  2914:4123
  2914:4223
end-set

route-policy ASIA-OUT
  if destination in CUSTOMER-PREFIXES then
    set community-set PREPEND-NTT-ASIA
  endif
end-policy
```

Some challenges on origin community

- Some of origin community’s duplicated on two region.
  - That will be caused our outgoing traffic

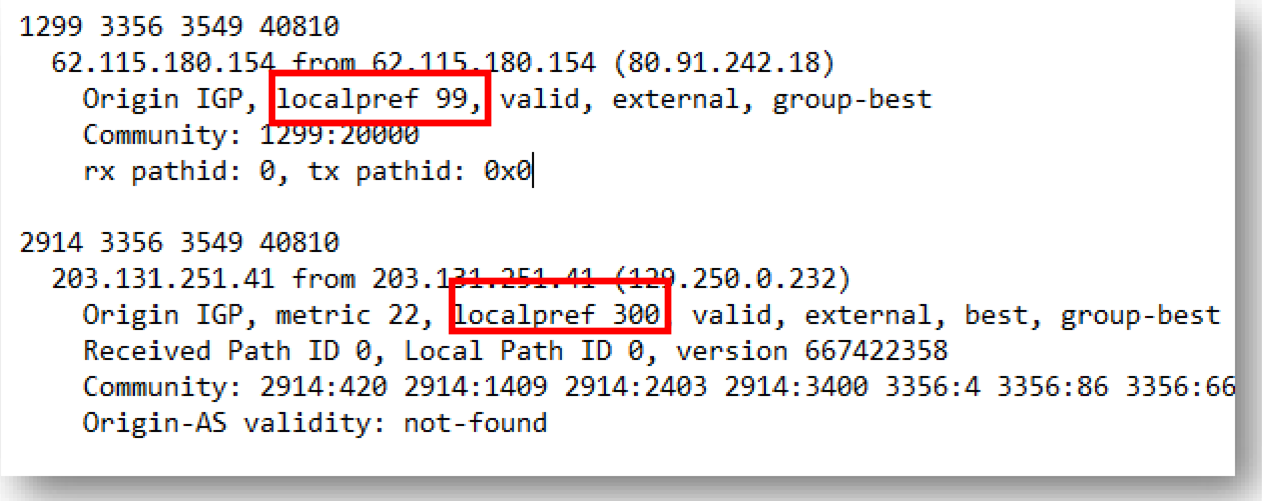




```
1299 3356 3549 40810
  62.115.180.154 from 62.115.180.154 (80.91.242.18)
    Origin IGP, localpref 300, valid, external, best
    Community: 1299:20000
    rx pathid: 0, tx pathid: 0x0

2914 3356 3549 40810
  203.131.251.41 from 203.131.251.41 (129.250.0.232)
    Origin IGP, metric 22, localpref 300, valid, external, internal, group-best
    Received Path ID 0, Local Path ID 0, version 667422358
    Community: 2914:420 2914:1409 2914:2403 2914:3400 3356:4 3356:86 3356:666 33
    Origin-AS validity: not-found
```

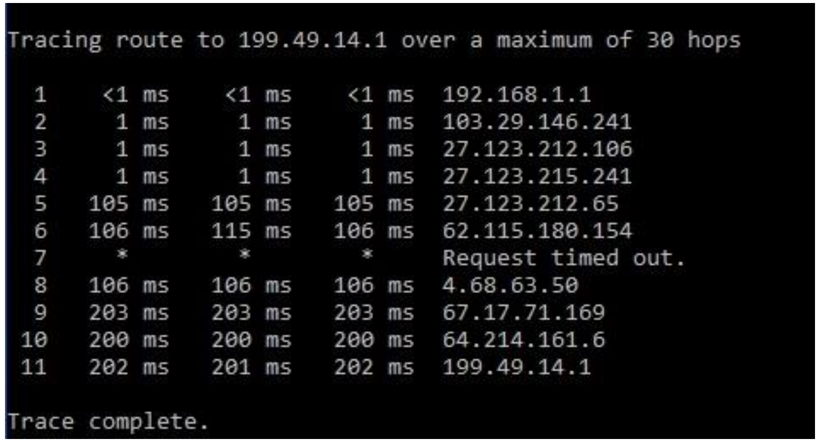
Some challenges in origin community

- This case we still using as-path and prefix list for setting incorrect origin community's to worst path



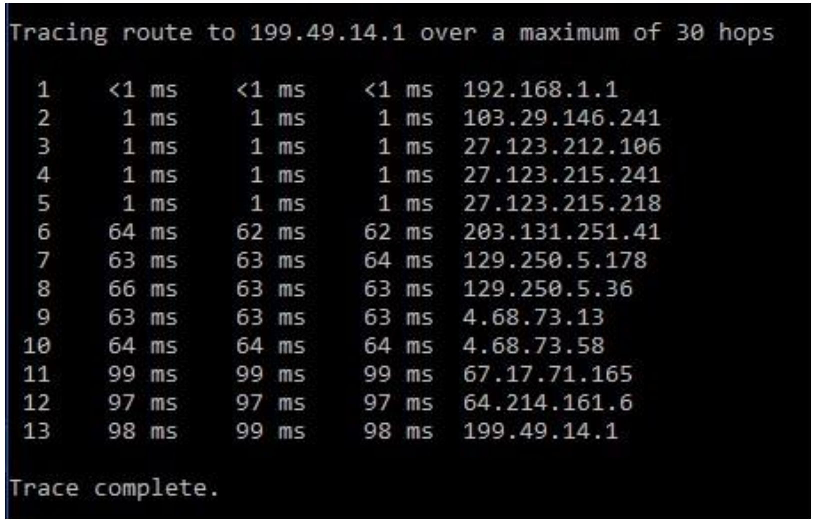
```
1299 3356 3549 40810
  62.115.180.154 from 62.115.180.154 (80.91.242.18)
    Origin IGP, localpref 99, valid, external, group-best
    Community: 1299:20000
    rx pathid: 0, tx pathid: 0x0

2914 3356 3549 40810
  203.131.251.41 from 203.131.251.41 (129.250.0.232)
    Origin IGP, metric 22, localpref 300, valid, external, best, group-best
    Received Path ID 0, Local Path ID 0, version 667422358
    Community: 2914:420 2914:1409 2914:2403 2914:3400 3356:4 3356:86 3356:66
    Origin-AS validity: not-found
```

- Before



```
Tracing route to 199.49.14.1 over a maximum of 30 hops

  1    <1 ms    <1 ms    <1 ms  192.168.1.1
  2     1 ms     1 ms     1 ms  103.29.146.241
  3     1 ms     1 ms     1 ms  27.123.212.106
  4     1 ms     1 ms     1 ms  27.123.215.241
  5   105 ms   105 ms   105 ms  27.123.212.65
  6   106 ms   115 ms   106 ms  62.115.180.154
  7     *        *        *     Request timed out.
  8   106 ms   106 ms   106 ms  4.68.63.50
  9   203 ms   203 ms   203 ms  67.17.71.169
 10   200 ms   200 ms   200 ms  64.214.161.6
 11   202 ms   201 ms   202 ms  199.49.14.1

Trace complete.
```

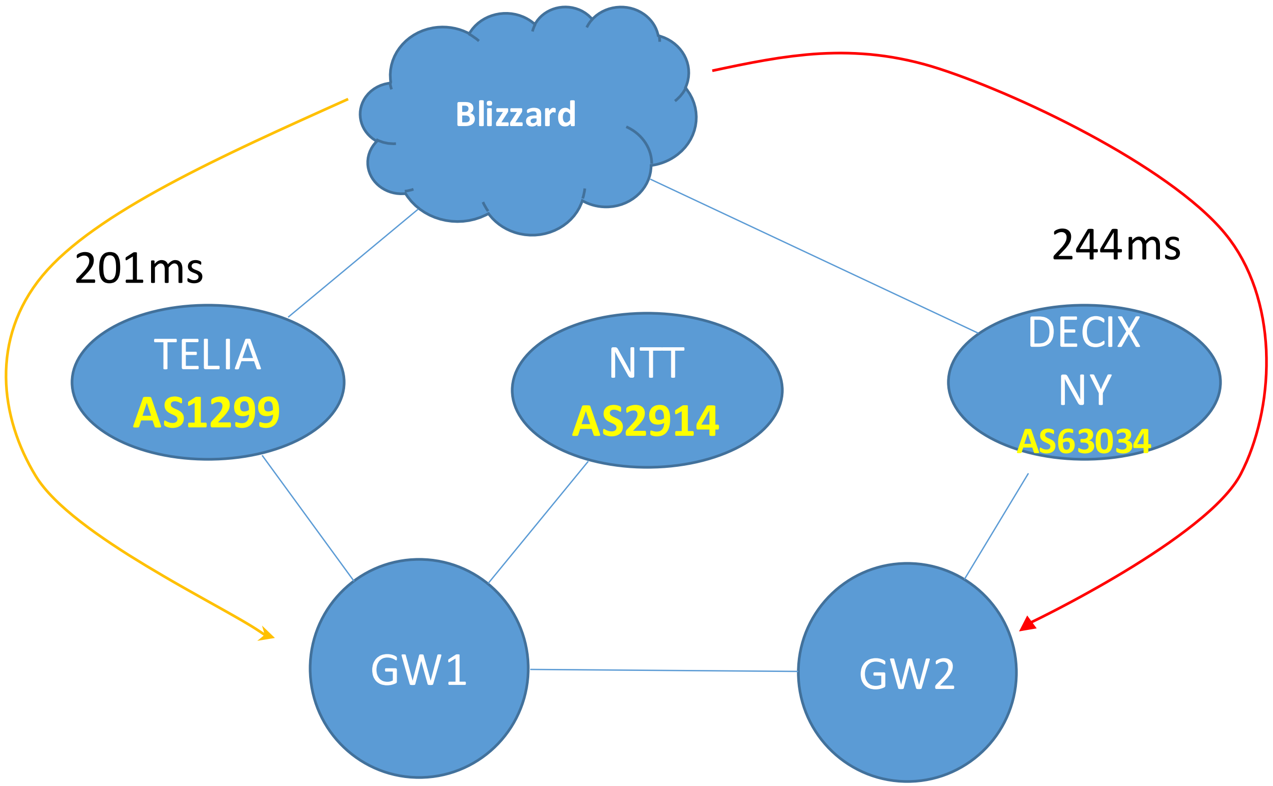
- After



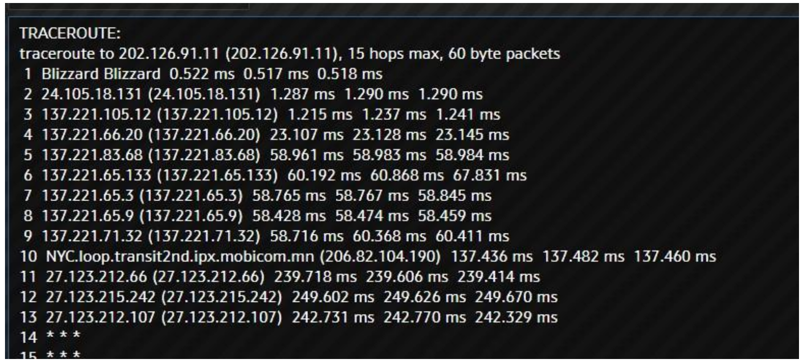
```
Tracing route to 199.49.14.1 over a maximum of 30 hops

  1    <1 ms    <1 ms    <1 ms  192.168.1.1
  2     1 ms     1 ms     1 ms  103.29.146.241
  3     1 ms     1 ms     1 ms  27.123.212.106
  4     1 ms     1 ms     1 ms  27.123.215.241
  5     1 ms     1 ms     1 ms  27.123.215.218
  6    64 ms    62 ms    62 ms  203.131.251.41
  7    63 ms    63 ms    64 ms  129.250.5.178
  8    66 ms    63 ms    63 ms  129.250.5.36
  9    63 ms    63 ms    63 ms  4.68.73.13
 10    64 ms    64 ms    64 ms  4.68.73.58
 11    99 ms    99 ms    99 ms  67.17.71.165
 12    97 ms    97 ms    97 ms  64.214.161.6
 13    98 ms    99 ms    98 ms  199.49.14.1

Trace complete.
```

Example WOW US server

- In this cases destination located in US but incoming traffic goes thru EU
- We added community in our advertisement about **do not announce blizzard AS57976** to DECIX-NY
- After that incoming traffic goes thru US and latency decreased.

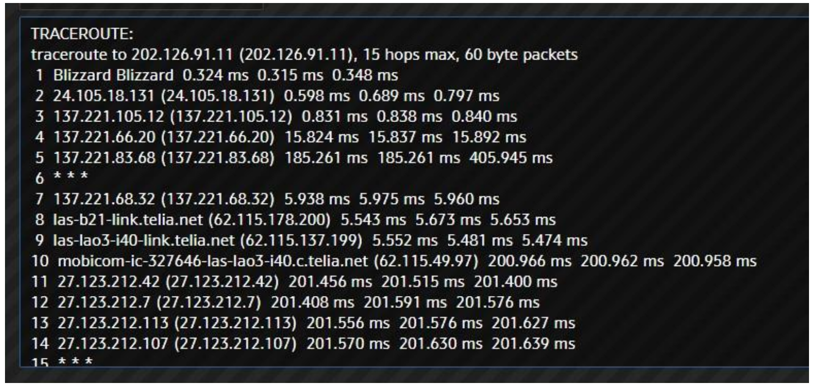


- Before



```
TRACEROUTE:
traceroute to 202.126.91.11 (202.126.91.11), 15 hops max, 60 byte packets
 1  Blizzard Blizzard  0.522 ms  0.517 ms  0.518 ms
 2  24.105.18.131 (24.105.18.131)  1.287 ms  1.290 ms  1.290 ms
 3  137.221.105.12 (137.221.105.12)  1.215 ms  1.237 ms  1.241 ms
 4  137.221.66.20 (137.221.66.20)  23.107 ms  23.128 ms  23.145 ms
 5  137.221.83.68 (137.221.83.68)  58.961 ms  58.983 ms  58.984 ms
 6  137.221.65.133 (137.221.65.133)  60.192 ms  60.868 ms  67.831 ms
 7  137.221.65.3 (137.221.65.3)  58.765 ms  58.767 ms  58.845 ms
 8  137.221.65.9 (137.221.65.9)  58.428 ms  58.474 ms  58.459 ms
 9  137.221.71.32 (137.221.71.32)  58.716 ms  60.368 ms  60.411 ms
10  NYC.loop.transit2nd.ipx.mobicom.mn (206.82.104.190)  137.436 ms  137.482 ms  137.460 ms
11  27.123.212.66 (27.123.212.66)  239.718 ms  239.606 ms  239.414 ms
12  27.123.215.242 (27.123.215.242)  249.602 ms  249.626 ms  249.670 ms
13  27.123.212.107 (27.123.212.107)  242.731 ms  242.770 ms  242.329 ms
14  * * *
15  * * *
```

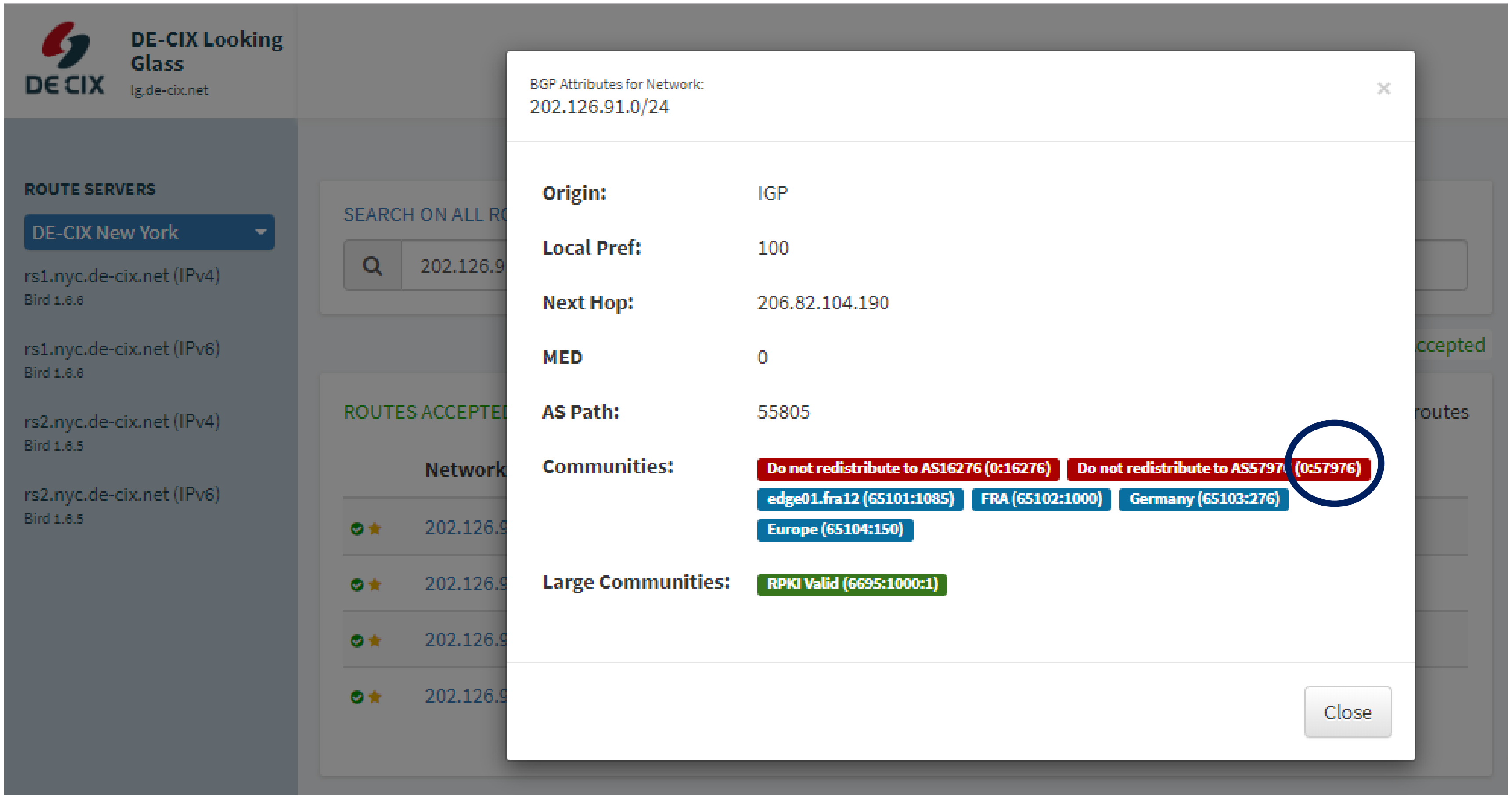
- After



```
TRACEROUTE:
traceroute to 202.126.91.11 (202.126.91.11), 15 hops max, 60 byte packets
 1  Blizzard Blizzard  0.324 ms  0.315 ms  0.348 ms
 2  24.105.18.131 (24.105.18.131)  0.598 ms  0.689 ms  0.797 ms
 3  137.221.105.12 (137.221.105.12)  0.831 ms  0.838 ms  0.840 ms
 4  137.221.66.20 (137.221.66.20)  15.824 ms  15.837 ms  15.892 ms
 5  137.221.83.68 (137.221.83.68)  185.261 ms  185.261 ms  405.945 ms
 6  * * *
 7  137.221.68.32 (137.221.68.32)  5.938 ms  5.975 ms  5.960 ms
 8  las-b21-link.telia.net (62.115.178.200)  5.543 ms  5.673 ms  5.653 ms
 9  las-lao3-i40-link.telia.net (62.115.137.199)  5.552 ms  5.481 ms  5.474 ms
10  mobicom-ic-327646-las-lao3-i40.c.telia.net (62.115.49.97)  200.966 ms  200.962 ms  200.958 ms
11  27.123.212.42 (27.123.212.42)  201.456 ms  201.515 ms  201.400 ms
12  27.123.212.7 (27.123.212.7)  201.408 ms  201.591 ms  201.576 ms
13  27.123.212.113 (27.123.212.113)  201.556 ms  201.576 ms  201.627 ms
14  27.123.212.107 (27.123.212.107)  201.570 ms  201.630 ms  201.639 ms
15  * * *
```

Example WOW US server

- Checked our action community on looking glass



BGP Attributes for Network:
202.126.91.0/24

Origin: IGP

Local Pref: 100

Next Hop: 206.82.104.190

MED 0

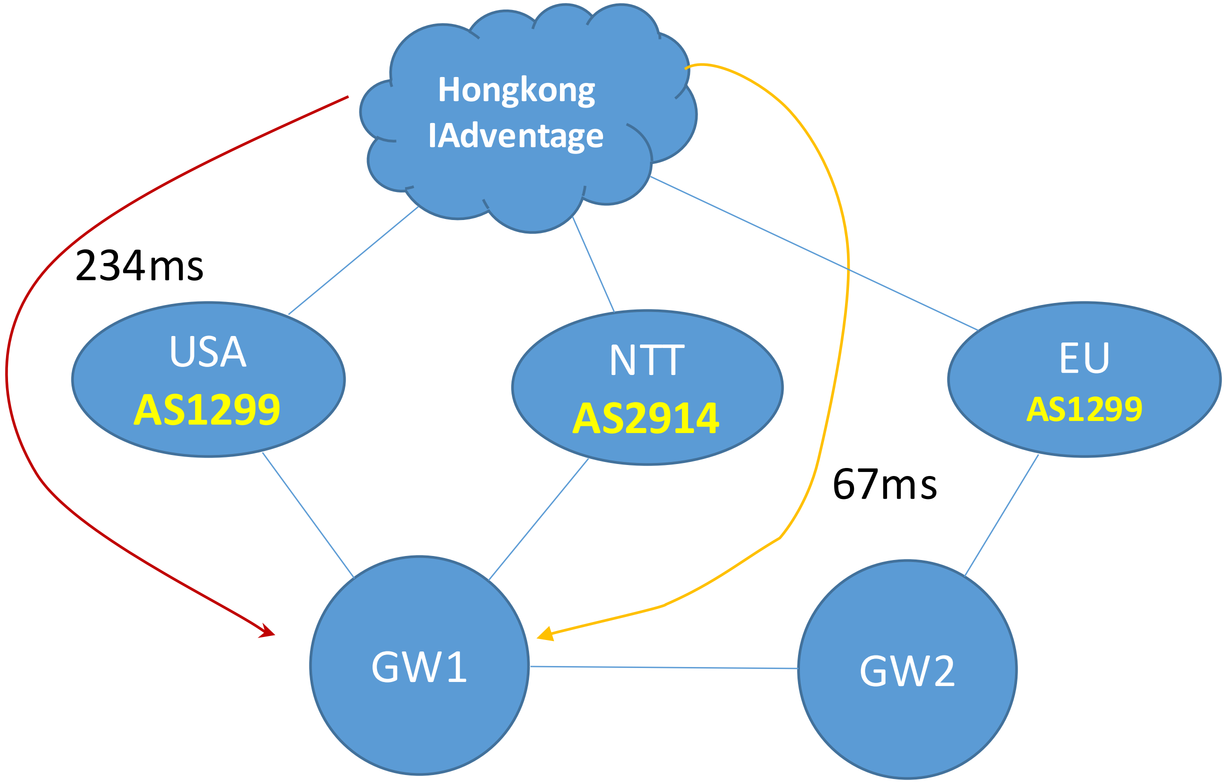
AS Path: 55805

Communities: Do not redistribute to AS16276 (0:16276) Do not redistribute to AS57976 (0:57976) edge01.fra12 (65101:1085) FRA (65102:1000) Germany (65103:276) Europe (65104:150)

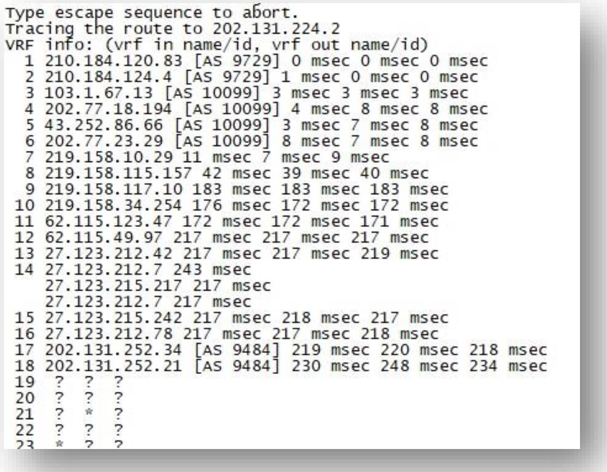
Large Communities: RPKI Valid (6695:1000:1)

Example Hong Kong traffic

- From Hongkong AS9729 incoming traffic was goes thru US and latency was high
- We added community in our advertisement about **prepend 3x All ASIA peer** to TELIA-US but it doesn't affect the traffic.
- Then we send the community about **Do not advertise chine Unicom**.
- After that incoming traffic goes thru NTT and latency decreased.

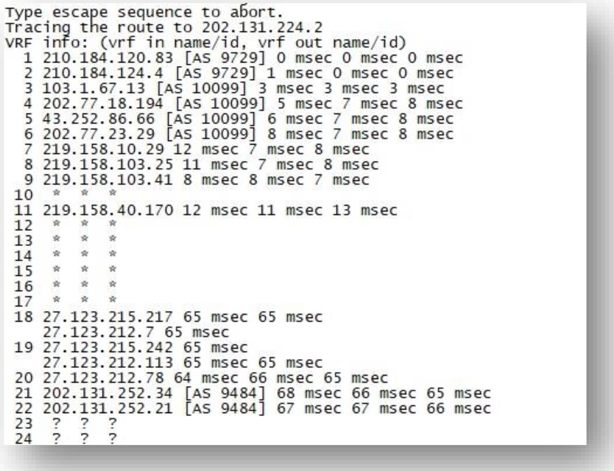


- Before



```
Type escape sequence to abort.
Tracing the route to 202.131.224.2
VRF info: (vrf in name/id, vrf out name/id)
  1 210.184.120.83 [AS 9729] 0 msec 0 msec 0 msec
  2 210.184.124.4 [AS 9729] 1 msec 0 msec 0 msec
  3 103.1.67.13 [AS 10099] 3 msec 3 msec 3 msec
  4 202.77.18.194 [AS 10099] 4 msec 8 msec 8 msec
  5 43.252.86.66 [AS 10099] 3 msec 7 msec 8 msec
  6 202.77.23.29 [AS 10099] 8 msec 7 msec 8 msec
  7 219.158.10.29 11 msec 7 msec 9 msec
  8 219.158.115.157 42 msec 39 msec 40 msec
  9 219.158.117.10 183 msec 183 msec 183 msec
 10 219.158.34.254 176 msec 172 msec 172 msec
 11 62.115.123.47 172 msec 172 msec 171 msec
 12 62.115.49.97 217 msec 217 msec 217 msec
 13 27.123.212.42 217 msec 217 msec 219 msec
 14 27.123.212.7 243 msec
    27.123.215.217 217 msec
    27.123.212.7 217 msec
 15 27.123.215.242 217 msec 218 msec 217 msec
 16 27.123.212.78 217 msec 217 msec 218 msec
 17 202.131.252.34 [AS 9484] 219 msec 220 msec 218 msec
 18 202.131.252.21 [AS 9484] 230 msec 248 msec 234 msec
 19  ?  ?  ?
 20  ?  ?  ?
 21  ?  *  ?
 22  ?  ?  ?
 23  *  ?  ?
```

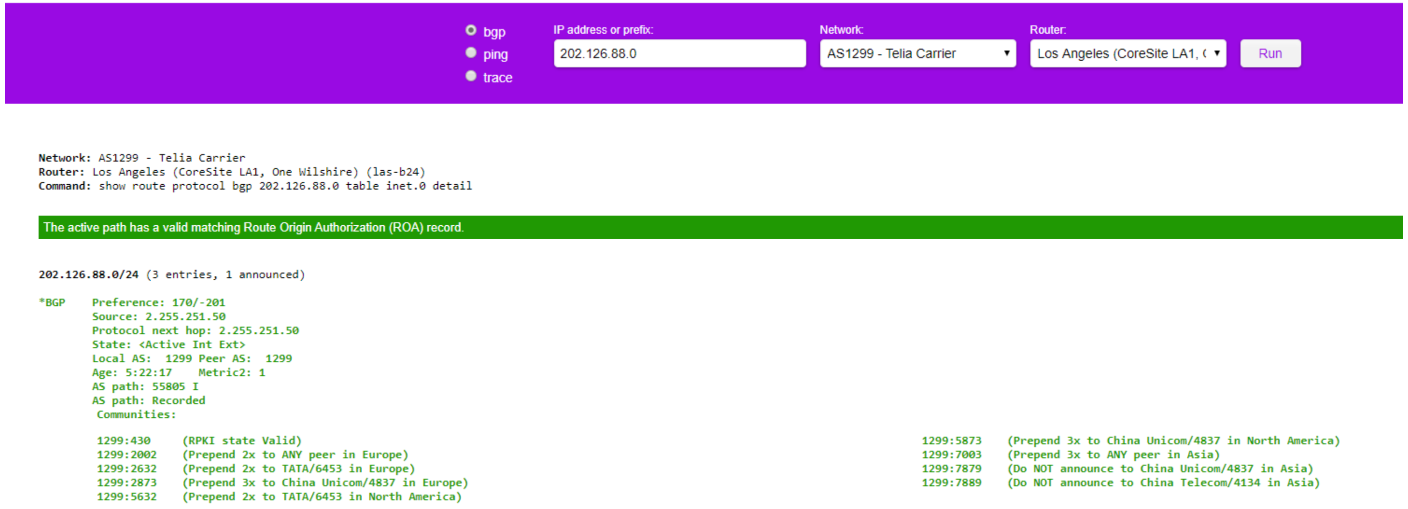
- After



```
Type escape sequence to abort.
Tracing the route to 202.131.224.2
VRF info: (vrf in name/id, vrf out name/id)
  1 210.184.120.83 [AS 9729] 0 msec 0 msec 0 msec
  2 210.184.124.4 [AS 9729] 1 msec 0 msec 0 msec
  3 103.1.67.13 [AS 10099] 3 msec 3 msec 3 msec
  4 202.77.18.194 [AS 10099] 5 msec 7 msec 8 msec
  5 43.252.86.66 [AS 10099] 6 msec 7 msec 8 msec
  6 202.77.23.29 [AS 10099] 8 msec 7 msec 8 msec
  7 219.158.10.29 12 msec 7 msec 8 msec
  8 219.158.103.25 11 msec 7 msec 8 msec
  9 219.158.103.41 8 msec 8 msec 7 msec
 10  *  *  *
 11 219.158.40.170 12 msec 11 msec 13 msec
 12  *  *  *
 13  *  *  *
 14  *  *  *
 15  *  *  *
 16  *  *  *
 17  *  *  *
 18 27.123.215.217 65 msec 65 msec
    27.123.212.7 65 msec
 19 27.123.215.242 65 msec
    27.123.212.113 65 msec 65 msec
 20 27.123.212.78 64 msec 66 msec 65 msec
 21 202.131.252.34 [AS 9484] 68 msec 66 msec 65 msec
 22 202.131.252.21 [AS 9484] 67 msec 67 msec 66 msec
 23  ?  ?  ?
 24  ?  ?  ?
```

Example Hong Kong traffic

- Checked our action community on looking glass



○ bgp
○ ping
○ trace

IP address or prefix: 202.126.88.0

Network: AS1299 - Telia Carrier

Router: Los Angeles (CoreSite LA1, (

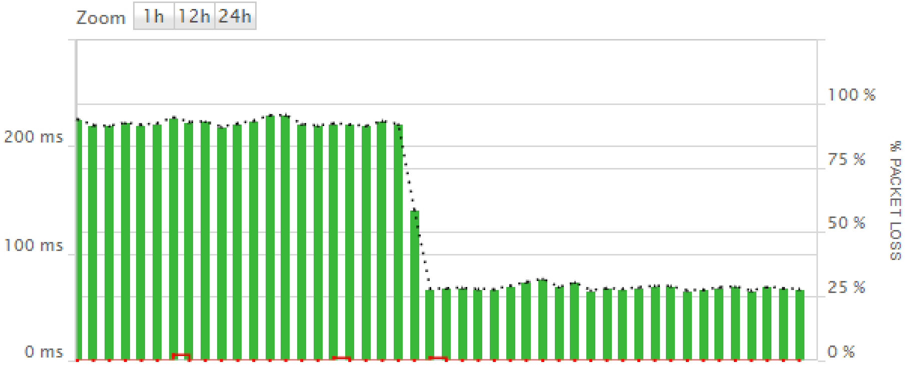
Run

```
Network: AS1299 - Telia Carrier
Router: Los Angeles (CoreSite LA1, One Wilshire) (las-b24)
Command: show route protocol bgp 202.126.88.0 table inet.0 detail
```

The active path has a valid matching Route Origin Authorization (ROA) record.

```
202.126.88.0/24 (3 entries, 1 announced)

*BGP    Preference: 170/-201
        Source: 2.255.251.50
        Protocol next hop: 2.255.251.50
        State: <Active Int Ext>
        Local AS:  1299 Peer AS:  1299
        Age: 5:22:17    Metric2: 1
        AS path: 55805 I
        AS path: Recorded
         Communities:

         1299:430     (RPKI state Valid)                                    1299:5873    (Prepend 3x to China Unicom/4837 in North America)
         1299:2002    (Prepend 2x to ANY peer in Europe)                    1299:7003    (Prepend 3x to ANY peer in Asia)
         1299:2632    (Prepend 2x to TATA/6453 in Europe)                   1299:7879    (Do NOT announce to China Unicom/4837 in Asia)
         1299:2873    (Prepend 3x to China Unicom/4837 in Europe)           1299:7889    (Do NOT announce to China Telecom/4134 in Asia)
         1299:5632    (Prepend 2x to TATA/6453 in North America)

         1299:1000 1299:35000 1299:35400

        Localpref: 200
        Router ID: 2.255.251.50
```

- Performance on monitoring



Conclusion

- Very complex for fix the asymmetric routing without the BGP community.
  - Specially inbound traffic. Due to some action needed to another side.
  - For Outgoing traffic, you can control using any way and any technic.
- BGP community can help quickly fix the asymmetric routing,
- In IXP case bi-lateral peering can help to improve, stabilizing your routing.

Thank you

- Any question?