Controlled IPv6 deaggregation by large organizations

draft-van-beijnum-grow-controlled-deagg-00

IETF 91
10 november 2014, Honolulu

Iljitsch van Beijnum

The IPv6 routing table today

- Size of the routing table:
  - Currently ~ 19000 prefixes
  - Growing at about 4000 prefixes/year
- However, more specifics are growing at 57% per year:
  - Jan 2013: 3049 of 11500: 27%
  - Jan 2014: 4799 of 16100: 29%

Source: http://www.potaroo.net/presentations/2014-02-09-bgp2013.pdf

An example... (1)

```
*   2001:2B8::/32      0 6939 9957 17832 i
*   2001:2B8:2::/48    0 6939 9957 17832 i
*   2001:2B8:11::/48   0 6939 9957 17832 i
*   2001:2B8:16::/48   0 6939 9957 17832 i
*   2001:2B8:17::/48   0 6939 9957 17832 i
*   2001:2B8:19::/48   0 6939 9957 17832 i
*   2001:2B8:20::/48   0 6939 9957 17832 i
*   2001:2B8:21::/48   0 6939 9957 17832 i
*   2001:2B8:22::/48   0 6939 9957 17832 i
*   2001:2B8:26::/48   0 6939 9957 17832 i
*   2001:2B8:28::/48   0 6939 9957 17832 i
*   2001:2B8:30::/48   0 6939 9957 17832 i
*   2001:2B8:31::/48   0 6939 9957 17832 i
*   2001:2B8:32::/48   0 6939 9957 17832 i
*   2001:2B8:35::/48   0 6939 9957 17832 i
*   2001:2B8:36::/48   0 6939 9957 17832 i
```

An example... (2)

```
*    2001:2B8:37::/48   0 6939 9957 17832 i
*    2001:2B8:39::/48   0 6939 9957 17832 i
*    2001:2B8:40::/48   0 6939 9957 17832 i
*    2001:2B8:43::/48   0 6939 9957 17832 i
*    2001:2B8:45::/48   0 6939 9957 17832 i
*    2001:2B8:48::/48   0 6939 9957 17832 i
*    2001:2B8:49::/48   0 6939 9957 17832 i
*    2001:2B8:50::/48   0 6939 9957 17832 i
*>   2001:2B8:51::/48   0 6939 9957 17832 i
*>   2001:2B8:52::/48   0 6939 9957 17832 i
*>   2001:2B8:53::/48   0 6939 9957 17832 i
*    2001:2B8:90::/48   0 6939 9957 17832 1237 i
*    2001:2B8:94::/48   0 6939 9957 17832 1237 i
*    2001:2B8:9A::/48   0 6939 9957 17832 1237 i
*    2001:2B8:9C::/48   0 6939 9957 17832 1237 i
*    2001:2B8:9D::/48   0 6939 9957 17832 1237 i
```

An example... (3)

```
*   2001:2B8:A0::/48   0 6939 9957 17832 1237 i
*   2001:2B8:A4::/48   0 6939 9957 17832 1237 i
*   2001:2B8:B0::/48   0 6939 9957 17832 1237 i
*   2001:2B8:B2::/48   0 6939 9957 17832 1237 i
*   2001:2B8:B4::/48   0 6939 9957 17832 1237 i
*   2001:2B8:B6::/48   0 6939 9957 17832 1237 i
*   2001:2B8:B8::/48   0 6939 9957 17832 1237 i
*   2001:2B8:BA::/48   0 6939 9957 17832 1237 i
*   2001:2B8:BC::/48   0 6939 9957 17832 1237 i
*   2001:2B8:BE::/48   0 6939 9957 17832 1237 i
*   2001:2B8:C0::/48   0 6939 9957 17832 1237 i
*   2001:2B8:C2::/48   0 6939 9957 17832 1237 i
*   2001:2B8:C4::/48   0 6939 9957 17832 1237 i
*   2001:2B8:C6::/48   0 6939 9957 17832 1237 i
*   2001:2B8:C8::/48   0 6939 9957 17832 1237 i
*   2001:2B8:CA::/48   0 6939 9957 17832 1237 i
```

An example... (4)

```
*   2001:2B8:CC::/48   0 6939 9957 17832 1237 i
*   2001:2B8:CE::/48   0 6939 9957 17832 1237 i
*   2001:2B8:D0::/48   0 6939 9957 17832 1237 i
*   2001:2B8:D2::/48   0 6939 9957 17832 1237 i
*   2001:2B8:D4::/48   0 6939 9957 17832 1237 i
*   2001:2B8:D6::/48   0 6939 9957 17832 1237 i
*   2001:2B8:DC::/48   0 6939 9957 17832 1237 i
*   2001:2B8:E6::/48   0 6939 9957 17832 1237 i
*   2001:2B8:ED::/48   0 6939 9957 17832 1237 i
*   2001:2B8:EF::/48   0 6939 9957 17832 1237 i
*   2001:2B8:F2::/48   0 6939 9957 17832 i
*   2001:2B8:200::/48  0 6939 9957 17832 i
*   2001:2B8:380::/48  0 6939 9957 17832 1237 i
```

An example... (5)

```
inet6num:       2001:02B8::/32
netname:        NGINET-KRNIC-KR-20010115
descr:          NGInet(Next Generation Internet Network)
descr:          is the national-wide Internet service
descr:          provider for public oganizations
country:        KR
```

What is this?

- Traditionally, types of addresses:
  - Provider Aggregatable (PA): used by ISPs
  - Provider Independent (PI): used by end users
- However, large organizations find it useful to have one big PA-like prefix
- But: their offices connect to different ISPs!
  - because they operate in many countries
  - or they have largely independent subunits

So: deaggregation

- So organizations such as:
  - big multinationals
  - governments
- Become "enterprise LIRs" and obtain a PA prefix
- Then subunits advertise deaggregates / more specifics of that PA block
  - towards different ISPs
  - in different locations

Is this a problem for the internet community?

- Not today!
  - IPv6 table is still small
- But people get large blocks so possible to source many deaggregates
  - no obvious way to filter on prefix length
- IPv6 is going to be around for a long time
- IPv4 has shown that mistakes early on are hard to clean up later

Does this work well for those organizations?

- Mostly
- However, deaggregates may be filtered
  - filtering is inconsistent because there is no agreed "safe" prefix length for IPv6
    - (like /24 in IPv4)

What do we do?

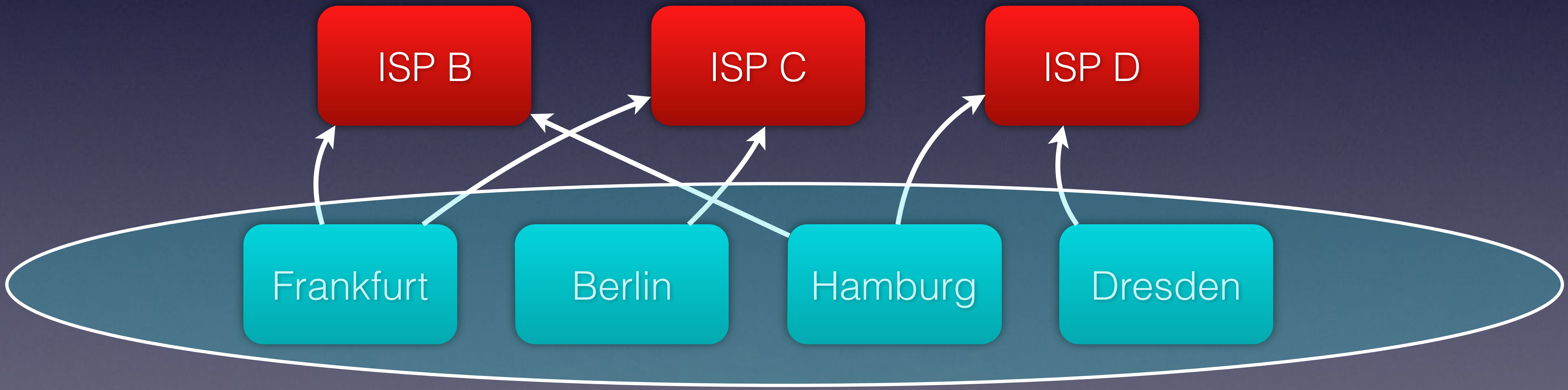
- Nothing?
  - suboptimal for routing table size
  - suboptimal for the organizations involved
  - may even hinder IPv6 deployment?
- Start a conversation between enterprise LIRs and network operators?
  - give enterprise LIRs guidance on what will work
  - give network operators tools to control table size

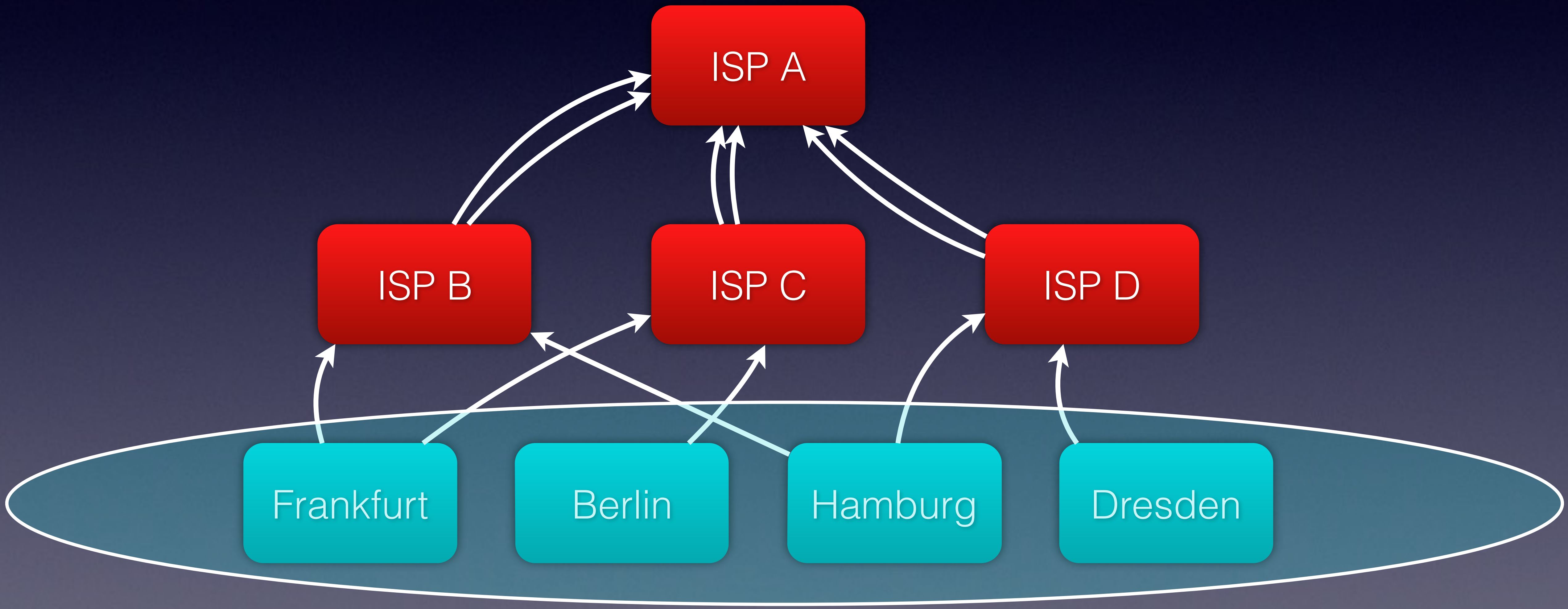
The idea

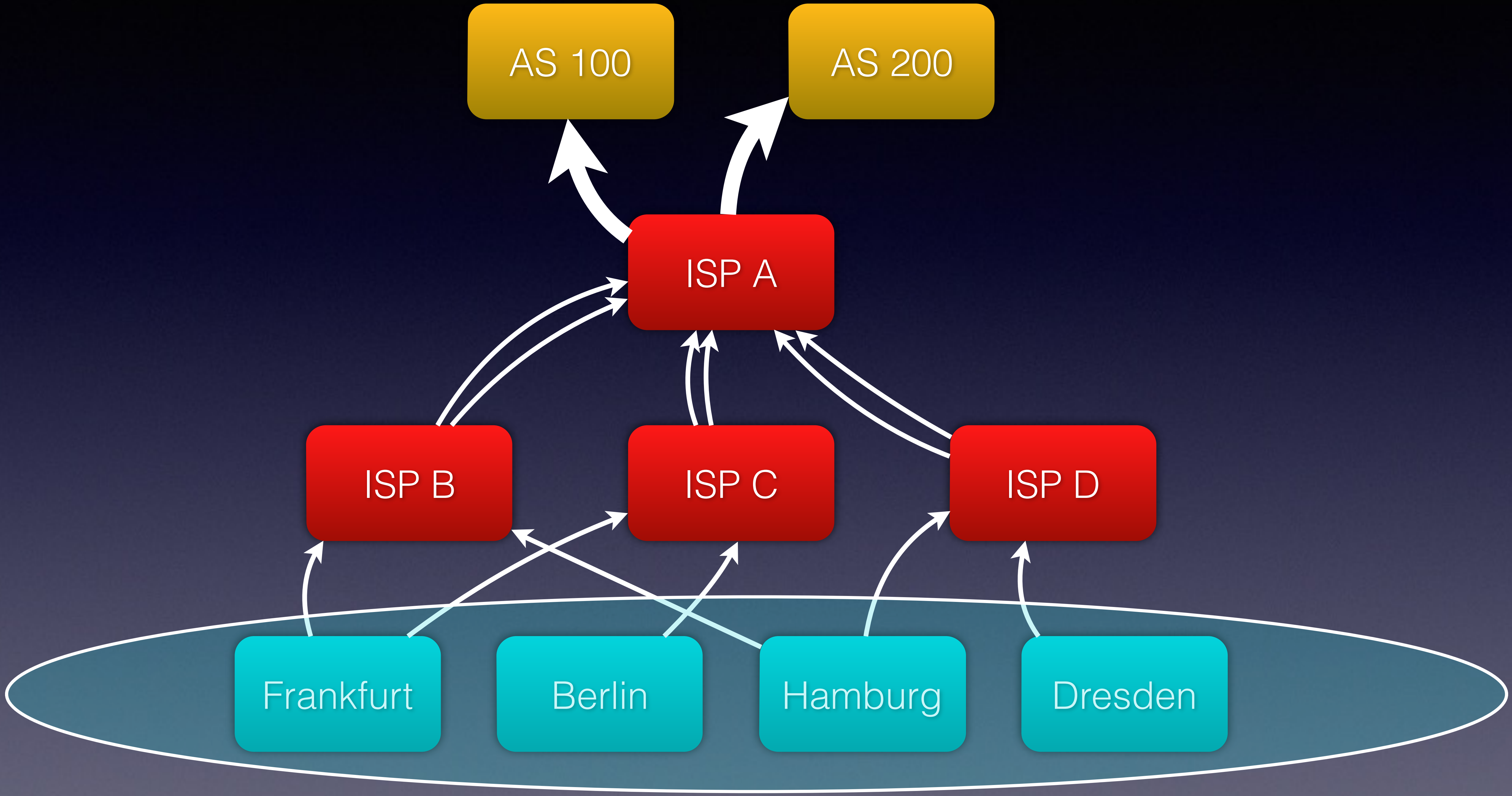
- Allow enterprise LIRs to set up an "aggregate of last resort" (AoLR)
    - so traffic has a place to go if deaggregates are filtered
- Tag deaggregates with BGP communities
    - indicate that it's safe to filter if needed
    - indicate where the deaggregate comes from
        - may want to allow "close" deaggregates but filter ones from far away

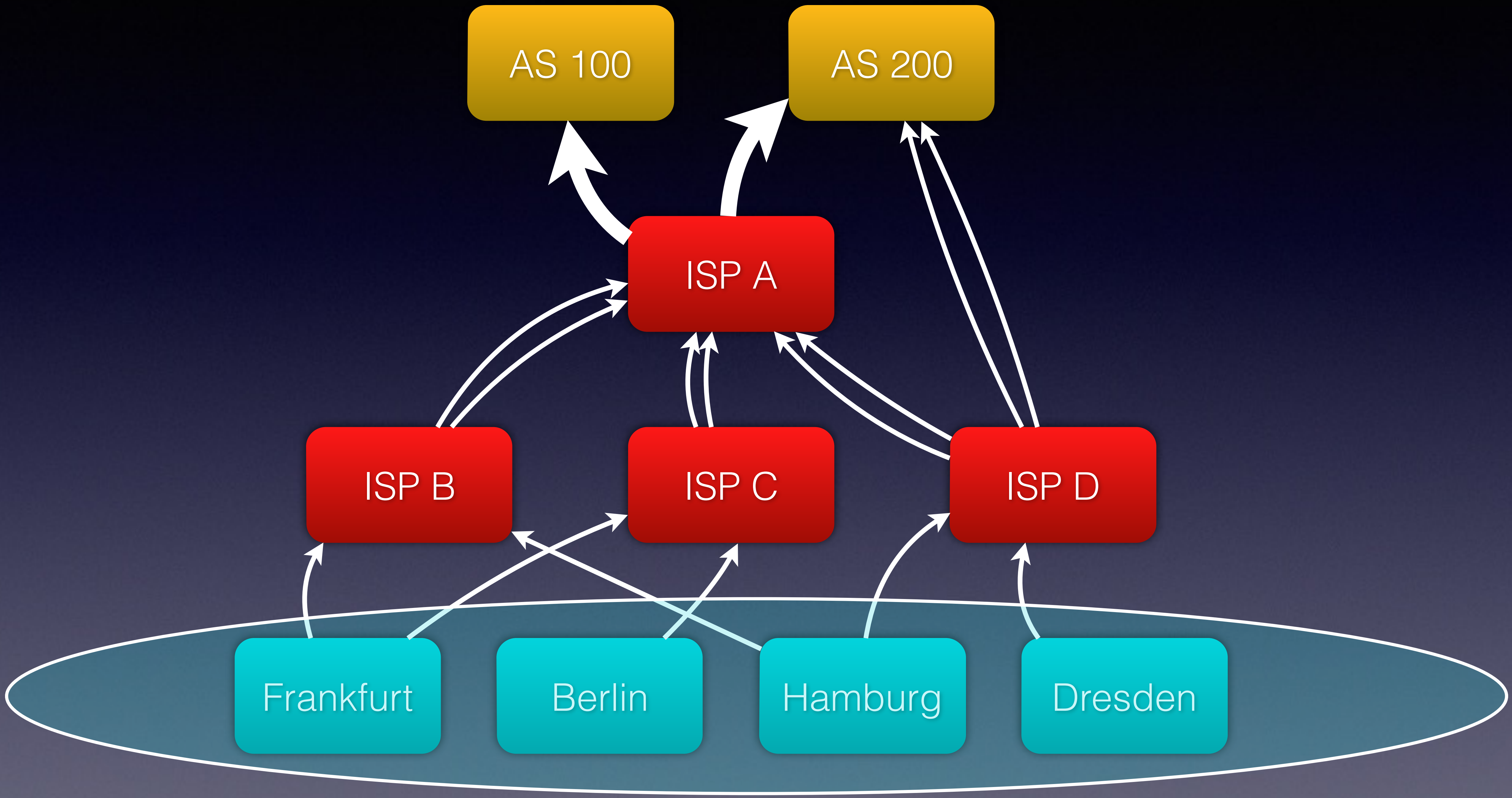
Aggregate of last resort

- ISP A injects the entire prefix in BGP
- ISPs B, C, D, ... (and maybe A) provide connectivity towards subunits of the organization
- B - D interconnect with A
- A accepts the deaggregates from B - D
- So the rest of the internet delivers packets to A
- A hands over the packets to B - D
    - so A only carries the packets a relatively short distance


ISP B
ISP C
ISP D
Frankfurt
Berlin
Hamburg
Dresden


ISP A
ISP B
ISP C
ISP D
Frankfurt
Berlin
Hamburg
Dresden


AS 100
AS 200
ISP A
ISP B
ISP C
ISP D
Frankfurt
Berlin
Hamburg
Dresden


AS 100
AS 200
ISP A
ISP B
ISP C
ISP D
Frankfurt
Berlin
Hamburg
Dresden

Aggregate of last resort (2)

- This works well if A is a large world-wide network
- However, B - G can be smaller regional or national networks
- A would have to be paid to provide this service
  - But can now be held accountable!
- (Multiple ISPs can provide the AoLR service if desired)
- Rest of the internet can safely filter the deaggregates

Location in BGP community

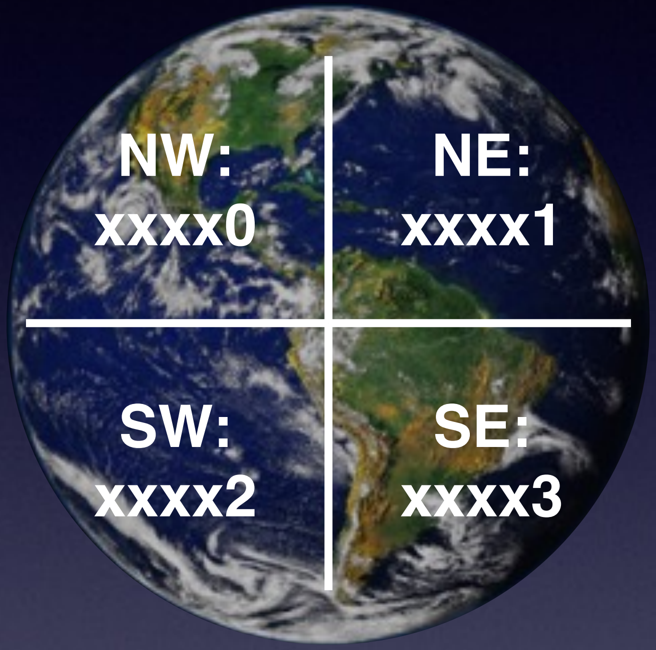
- A BGP "community" is simply a label attached to a prefix
  - 702:120 or NO_EXPORT
- In Europe we probably don't care about Korean deaggregates
- We Europeans just send the traffic in the general direction of Korea and once the packets get closer, the deaggregates will be there

Location in BGP community (2)

- GPS coordinates in BGP communities
  - Precision is 1 degree, ~ 100 km
- Not subject to change or political controversy!
- *Somewhat* human readable/understandable
- Maybe express filters as geographic areas in the future if router vendors add this to their routers
- But can work today!

Location in BGP community (2)

- Use 4 blocks of 216 communities:
  - xxxx: to be filled in by IANA



- Then encode latitude (2 digits) and longitude (3 digits) rounded to whole degrees
  - (some magic for > 64° north/south)

Examples

Honolulu, US	21° 17′ N, 157° 50′ W	xxxx0:21158
Berlin, DE	52° 31′ N, 13° 23′ E	xxxx1:53013
Chicago, US	41° 50′ N, 87° 41′ W	xxxx0:42088
Mumbai, IN	18° 58′ N, 72° 49′ E	xxxx1:19073
Rio de Janeiro, BR	22° 54′ S, 43° 11′ W	xxxx1:19073
Saint Petersburg, RU	59° 57′ N, 30° 18′ E	xxxx1:60030
Spitsbergen, NO	78° 45′ N, 16° 00′ E	xxxx1:01796
McMurdo Station, Antarctica	77° 51′ S, 166° 40′ E	xxxx3:16787

Why not extended?

- Another way to go: extended communities
- Upside: ???
- Downside: AFAIK, no default representation
    - so would have to wait for vendors to catch up!

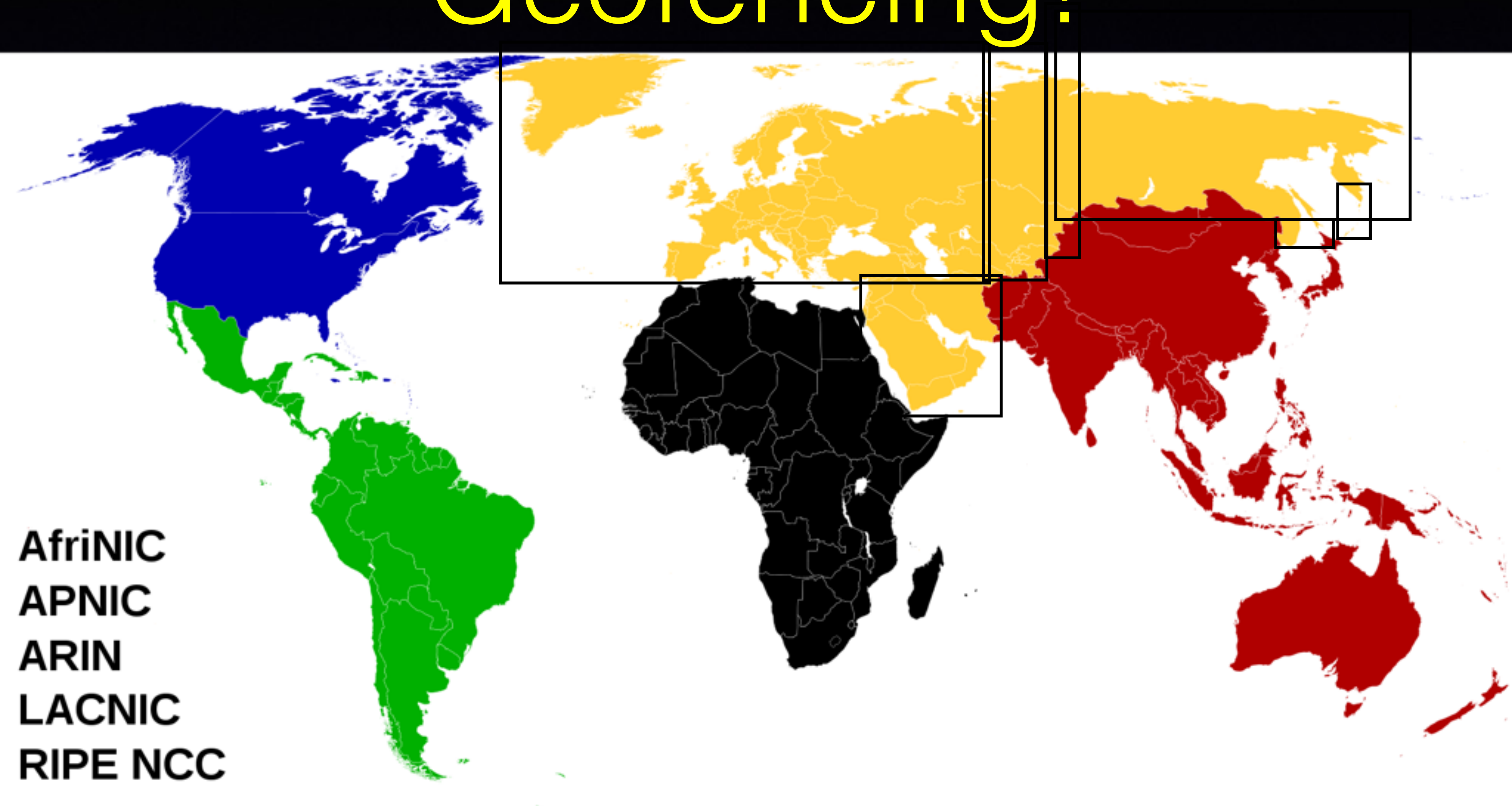
Selective filtering

- Everyone decides which deaggregates to carry
  - Big routers? Maybe carry them all
  - Small routers? Maybe carry none of them
  - Regional network? Maybe only carry deaggregates announced in the region
  - World-wide network? Maybe each router only carries deaggregates announced in the same region
    - so the network as a whole carries all deaggregates
    - but individual routers don't

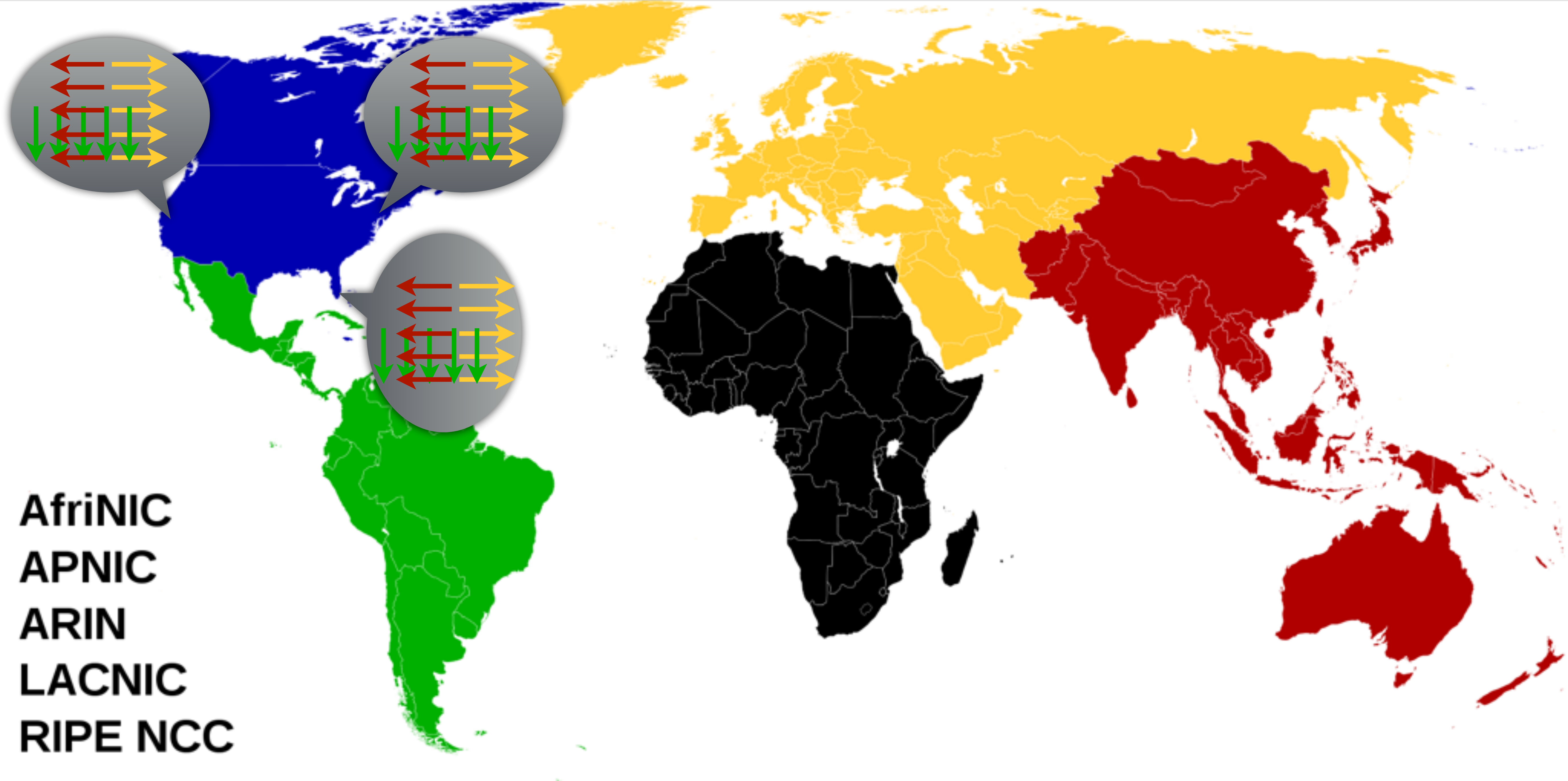
Selective filtering (2)

- Having different prefixes in different routers in the same AS:
  - requires prefix filters for iBGP
    - not great, but no reason this can't work

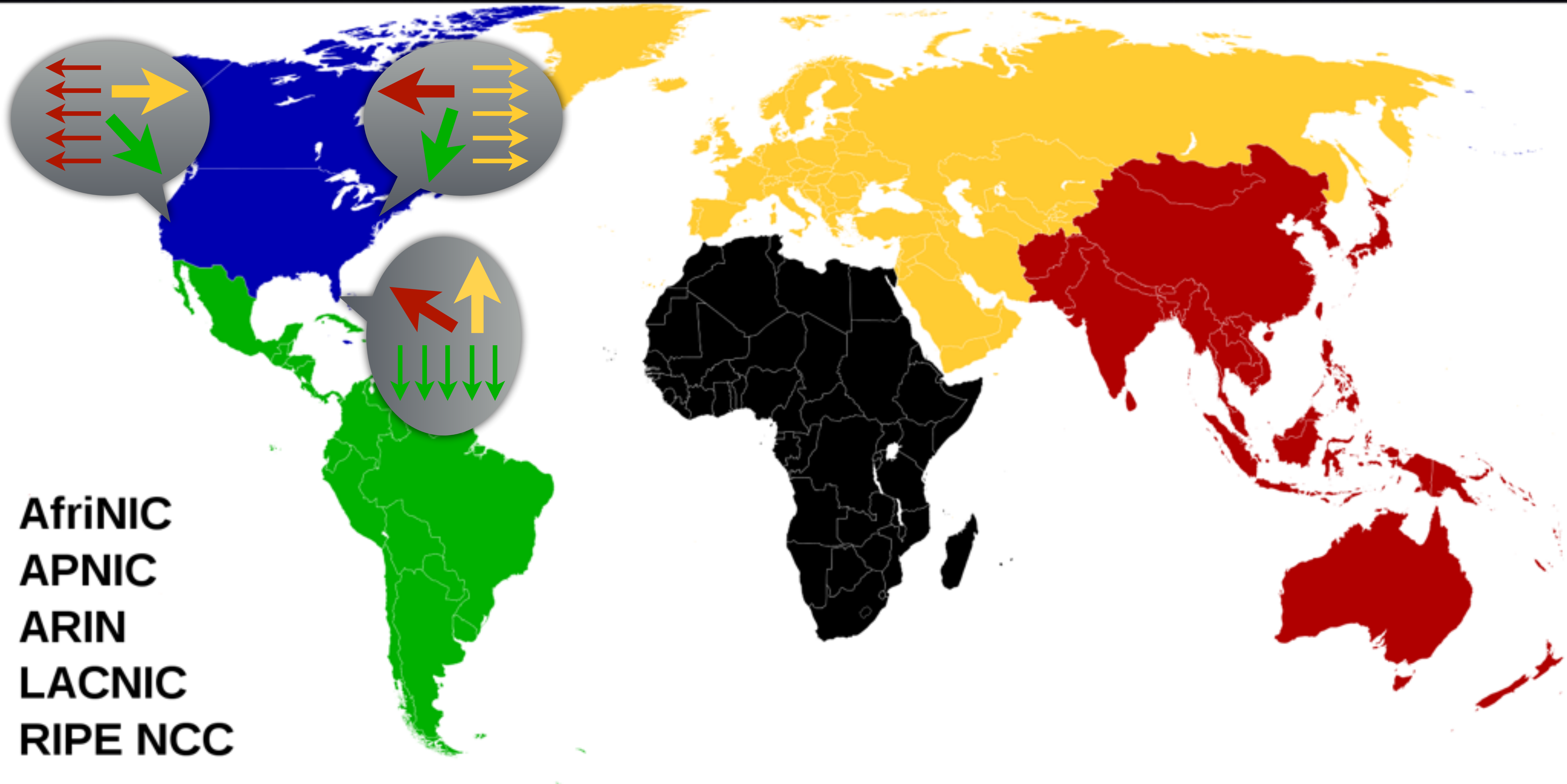
Geofencing!



Without aggregation



With aggregation



AfriNIC
APNIC
ARIN
LACNIC
RIPE NCC

What now?

- RIPE BCOP interested in the best practice part
- Defining communities needs to happen in an RFC
    - perhaps in a document that also defines other new well-known communities

- Questions?