


Stairway to Anycast

All the details

Our starting point

The parking lot

-  6connect is a global company
- Our DNS platform should be global too!
- The best way to scale DNS globally is by using Anycast
- Mentioning this to our commercial colleagues resulted in “Oh, we may have some customers for that...”

How to do that?

The roadmap

- Build a prototype
- Set up measurements
- Fine-tune the prototype
- More measurements
- Done!
- ...Right?

Building the prototype

Start with building the car

- Software choices:
 - BIND, Knot DNS, NSD, PowerDNS, which one?
 - All of them! Let's use *dnsmdist*
 - *Bird2* for BGP routing
 - *Ansible* for automation / rollout
 - bash/sed/awk for scripting!

Designing a node

The Haynes™ Manual

- *Dnsdist* provides scripting and monitoring
- Zone sync: Python script to update zone files
- Damocles: Bash script to query *dnsdist* and kill *BIRD* on failure
- Managed using *Ansible*

The first nodes

Finding the on-ramp

- 6connect clusters: Fremont (US), Ljubljana (SI) and Apeldoorn (NL)
 - Not a really good spread, but it's a start
- Which IP resources to use?
 - How many IPv4/IPv6 prefixes?
 - Which AS numbers?
- One ASN announcing 3x /24 IPv4 and 3x /48 IPv6
- 3 nodes, each announces primary IPv4+IPv6 + secondary IPv4+IPv6

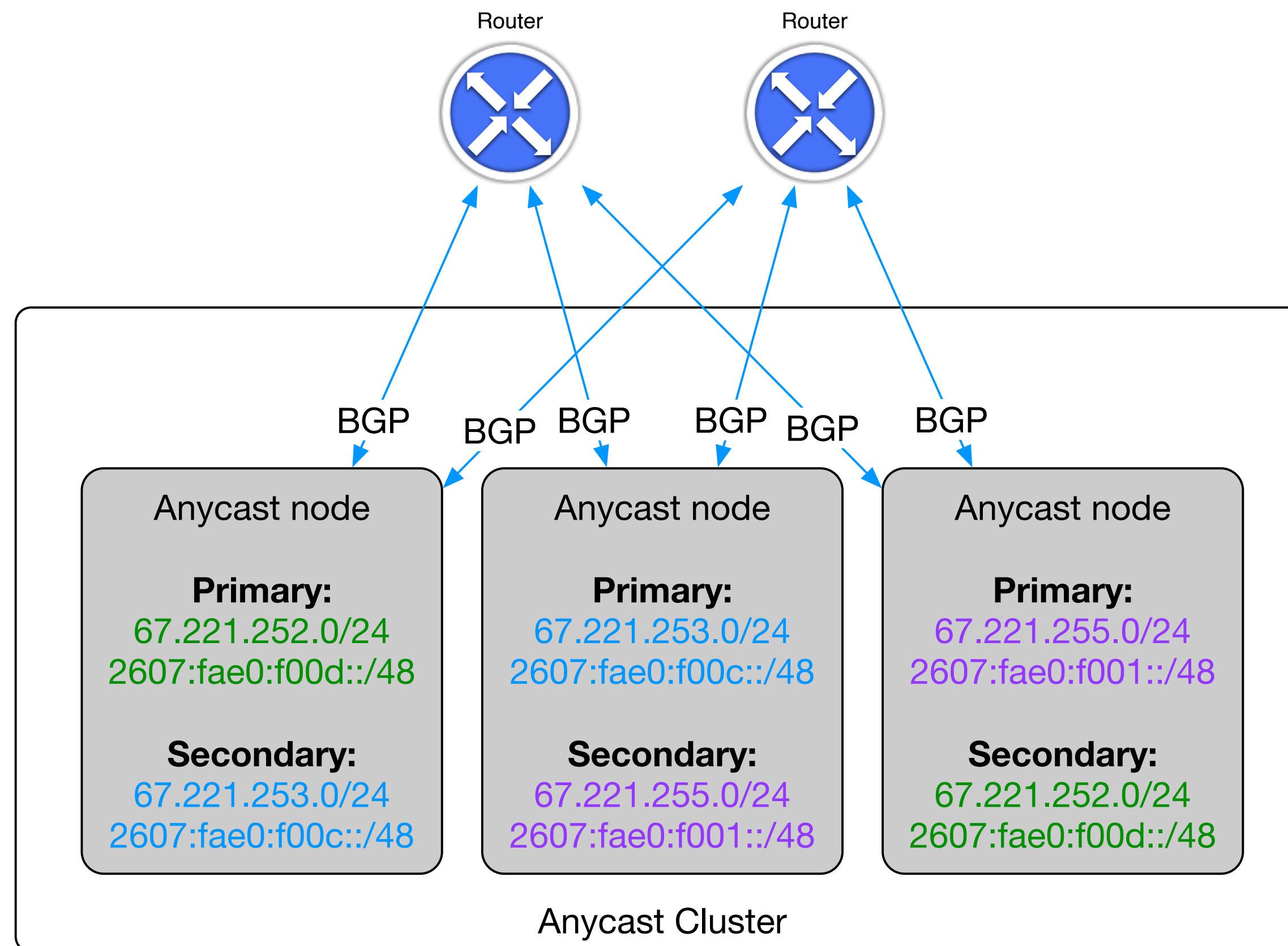
Cluster composition

Turning on the radio

- Anycast nodes announce
 - Primary prefix with high priority
 - Secondary prefix with low priority
- Method depends on relationship to the routers:
 - iBGP uses local-pref
 - eBGP uses prepending

Our Anycast architecture

The layout of one PoP



- One ASN: 8038
- Three Anycast prefixes:
 - 3x IPv4 /24
 - 3x IPv6 /48
- Three DNS nodes per PoP
- 4th node is for measurements

The need for measurements

Unexpected potholes

- It doesn't seem to work as well as expected
- But why/how/when/where?!?!?!?!?
- We need monitoring and measurements
- Route views help a bit
- RIPE Atlas provides some information, but not very detailed

In the meantime

A little detour

- While we are thinking about measurements, let's add more things!
 - We deployed a set of VMs in Tokyo (JP) using Vultr
 - Added them to the Anycast setup
- Moved our 6clabs.com domain to Anycast
 - Eat your own dog food...
 - What could possibly go wrong?

Our own control and monitoring

I think our car needs a speedometer

- We use 6connect ProVision as the control center for Anycast DNS
- Zones are administered and distributed from here to all Anycast DNS servers
- We use LibreNMS to keep track of *dnsdist* queries, performance and uptime
- We should also measure each backend.

More anycast ideas

The scenic route

- Our initial prototype is authoritative DNS
- We can also do recursive DNS, should be easy
- We also offer a cloud-hosted IPAM, can we Anycast that?
 - We'd need a replicated DB (Galera?)
- Having a high-available mail service would be nice
 - Proxmox Mail Gateway as a spam filter
 - Dovecot (dsync?) for replicated mailbox storage

Design decisions

Where's the sat-nav when you need it...?

- Not all services need to be in all Anycast locations
- How many services can we host in one /24 & /48?
 - If one service fails the whole prefix needs to be pulled out
- HAProxy or nginx as the front-end Anycasted load balancer
 - If the load balancer is the only Anycasted service this is a lot easier
 - If a local service fails the load-balancer can send the traffic to another site

Getting TLS certificates

Did you bring your passport?

- Anycasted services need a certificate for the distributed hostname
 - Using Let's Encrypt is more complicated than usual
 - We don't know where the ACME verification is going to be received
- The load-balancer can send all ACME traffic to a central node
 - This node can get the certificate
 - And distribute the keys and certificates to all relevant Anycast nodes
 - This needs to be built...

The dilemma stays

Hello?!? Can anybody hear us? Please tell us where we are...

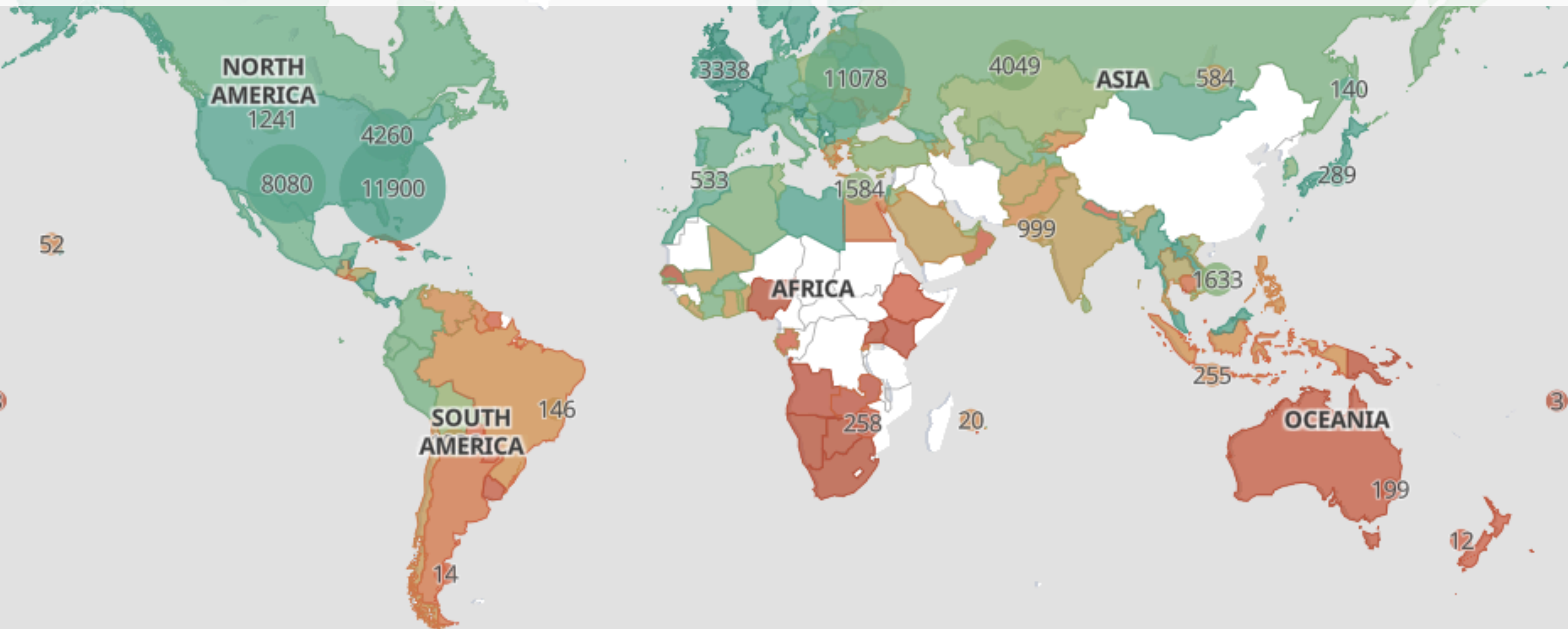
- Are we globally visible?
- Where are the black holes?
- Which networks are sending traffic to which site?
 - Is Asia sending traffic to US?
 - Is Europe sending traffic to Asia?
 - Is traffic going to the nearest site?
- Are we making our users take large detours?

What we want out of monitoring

The dials we need on the dashboard

- Where are our prefixes visible?
 - How are we visible from around the world?
- Which announcement do users see?
 - In other words: which Anycast cluster do users use?
 - Are clients using the closest node?
 - What is the latency from each user to "their" Anycast cluster?
- Where are the black holes?
 - Which ISPs do we need to talk with?

We need a good view to make it better...
At the end of our road trip we want everything green!



World map

Where are these nodes?

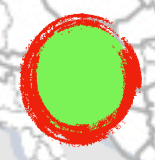
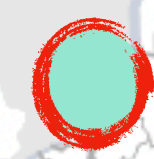
**NORTH
AMERICA**



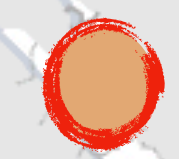
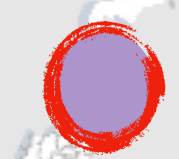
**SOUTH
AMERICA**



AFRICA



ASIA

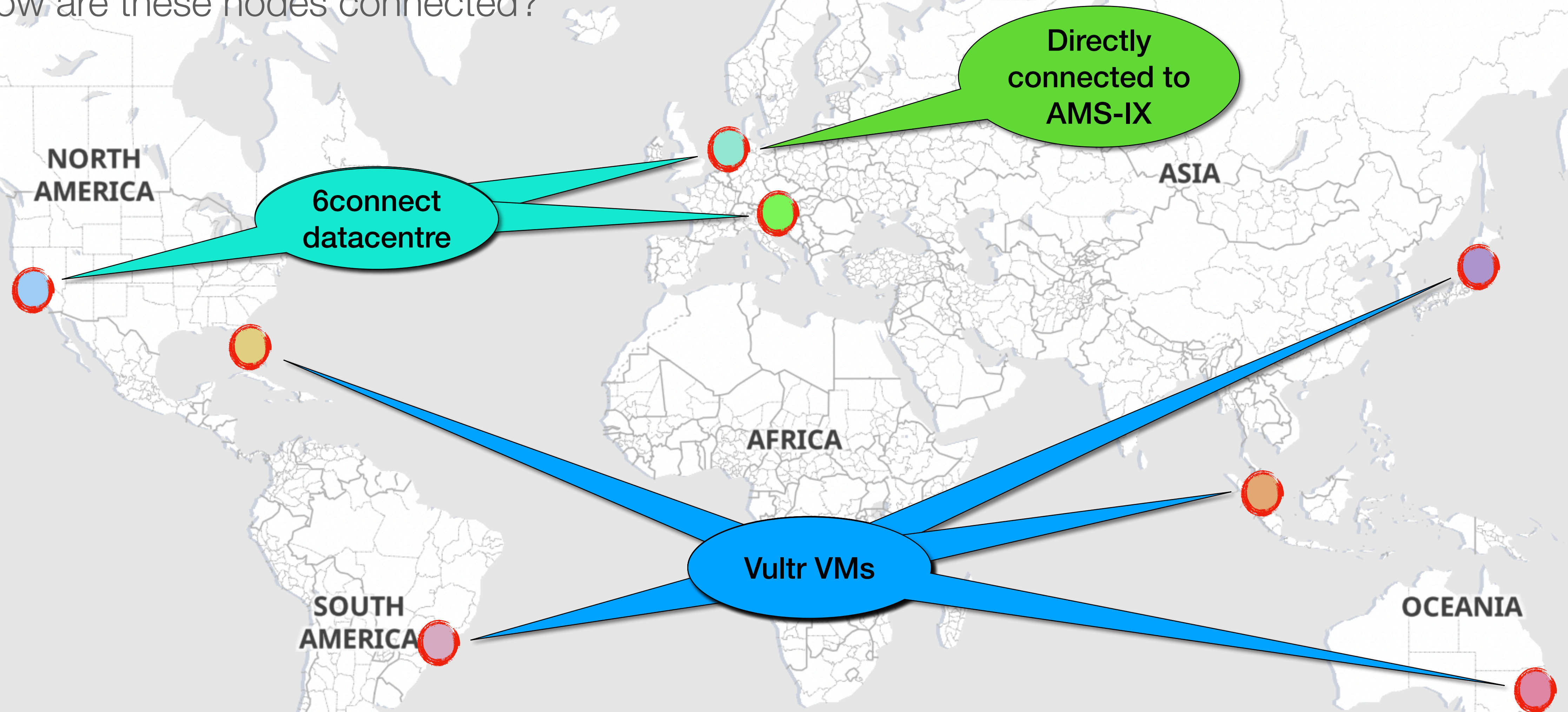


OCEANIA



World map

How are these nodes connected?



Users and latency per PoP

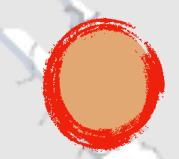
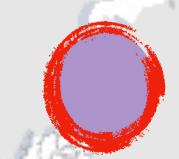
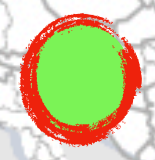
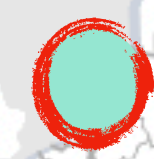
NORTH AMERICA

ASIA

AFRICA

SOUTH AMERICA

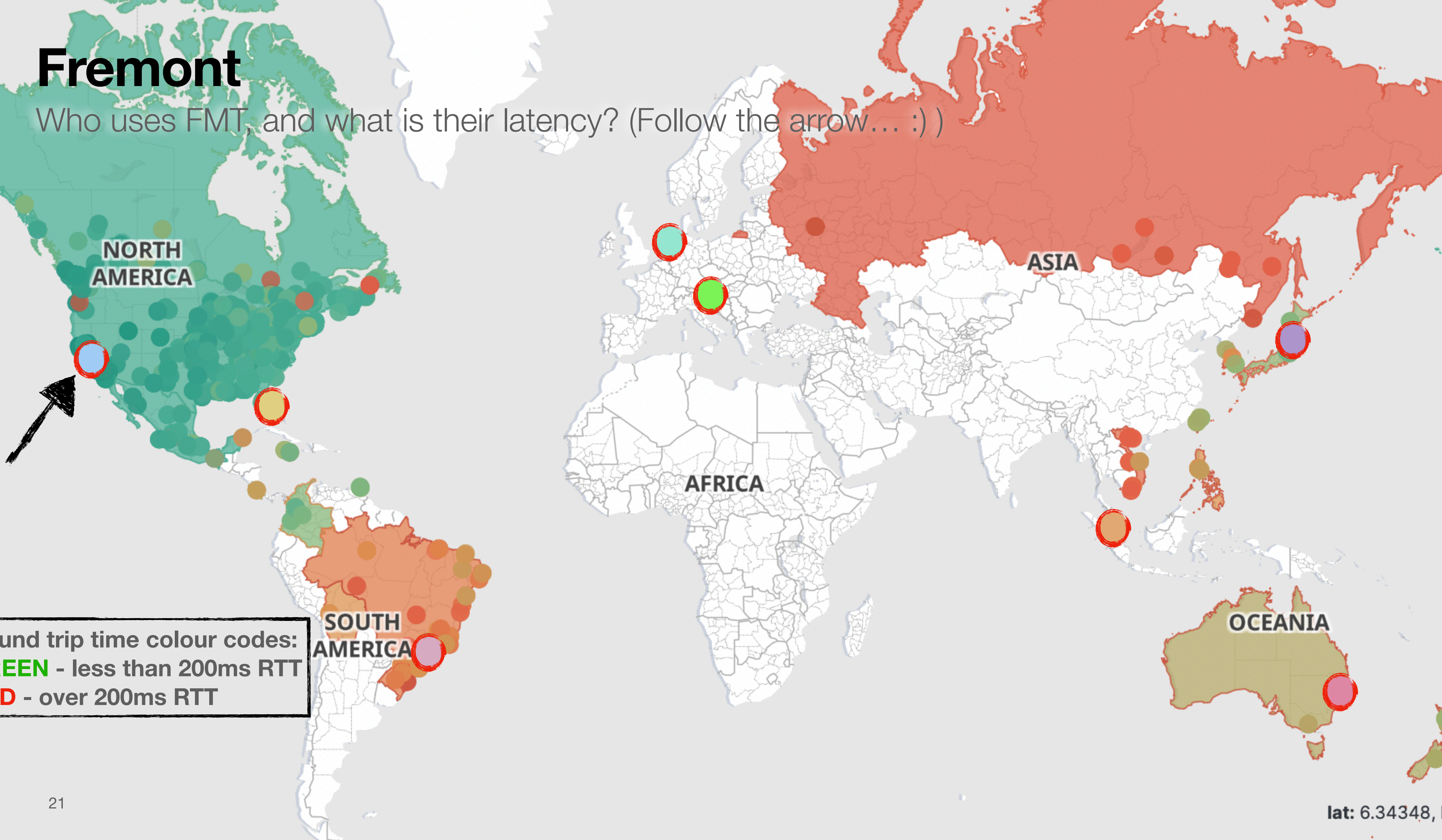
OCEANIA



lat: -7.33434, lon: 108.20814

Fremont

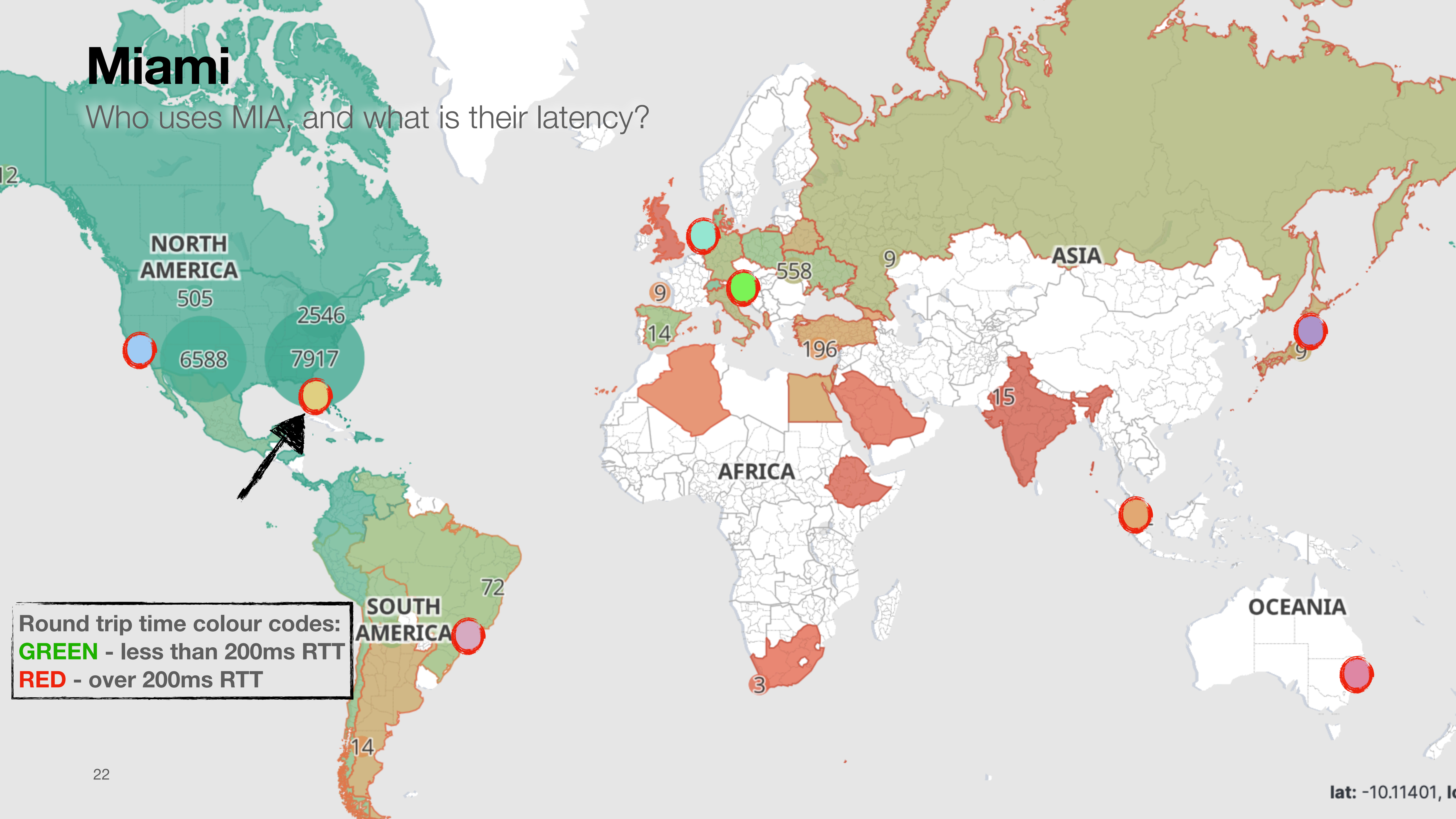
Who uses FMT, and what is their latency? (Follow the arrow... :))



Round trip time colour codes:
GREEN - less than 200ms RTT
RED - over 200ms RTT

Miami

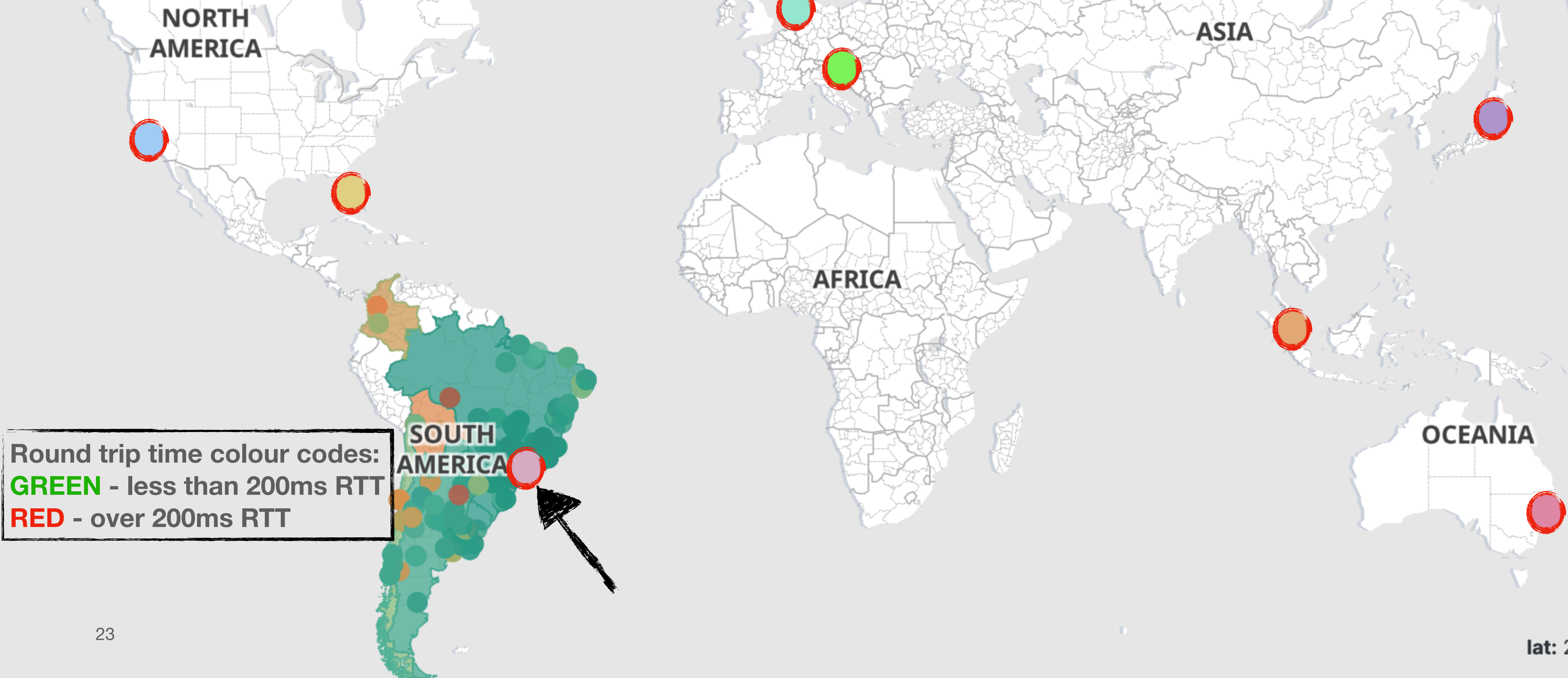
Who uses MIA, and what is their latency?



Round trip time colour codes:
GREEN - less than 200ms RTT
RED - over 200ms RTT

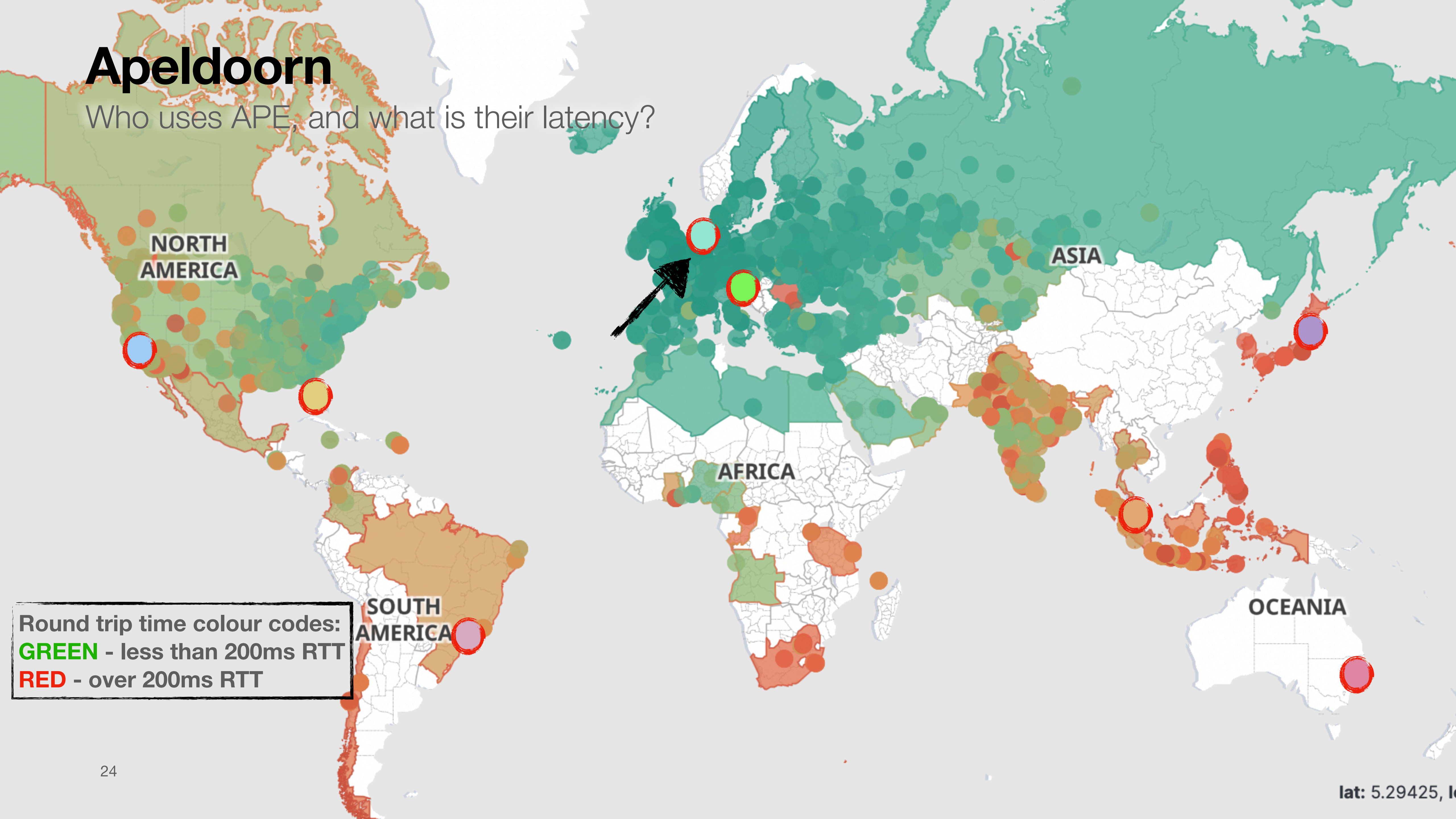
São Paulo

Who uses BRA, and what is their latency?



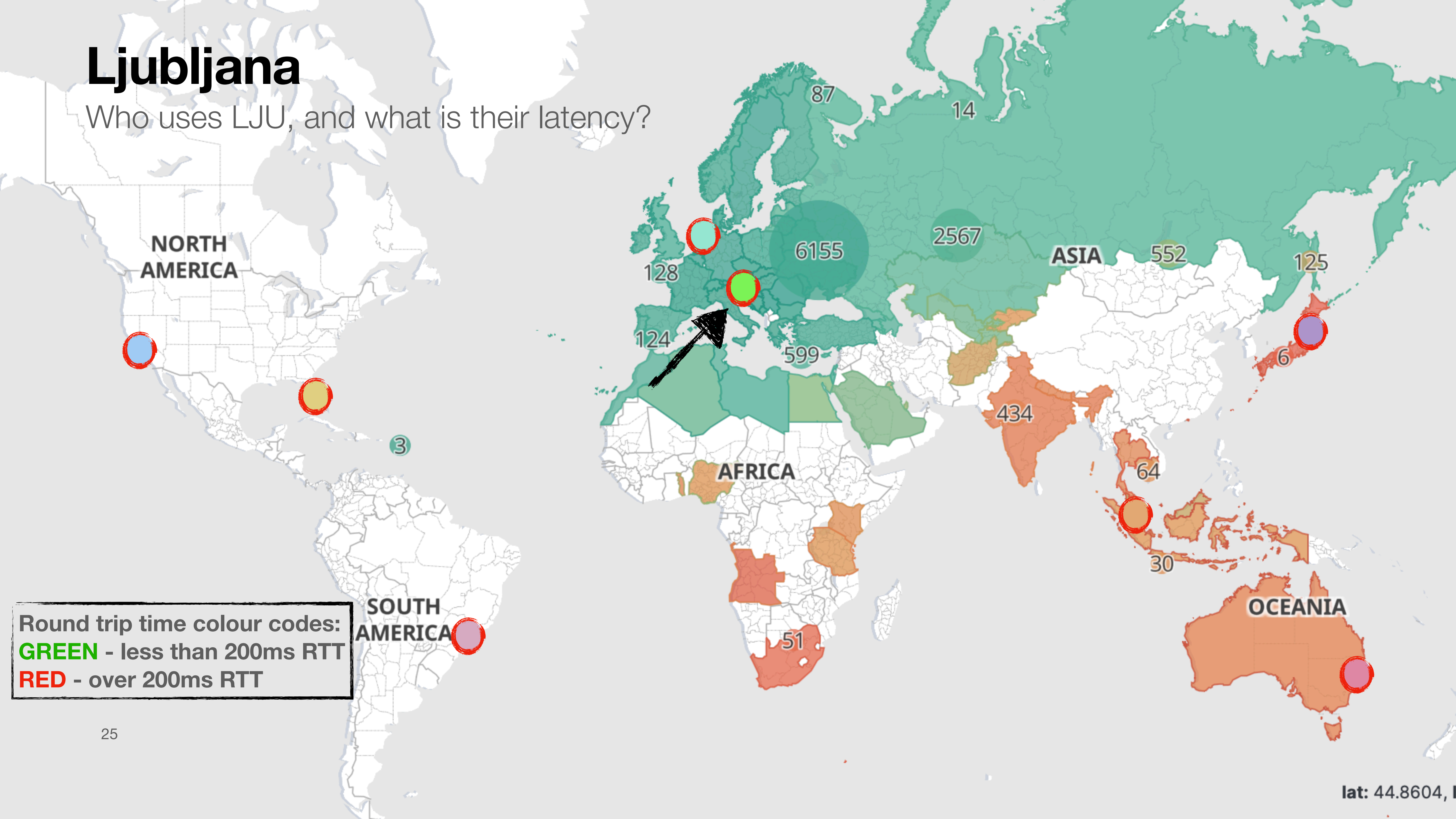
Apeldoorn

Who uses APE, and what is their latency?



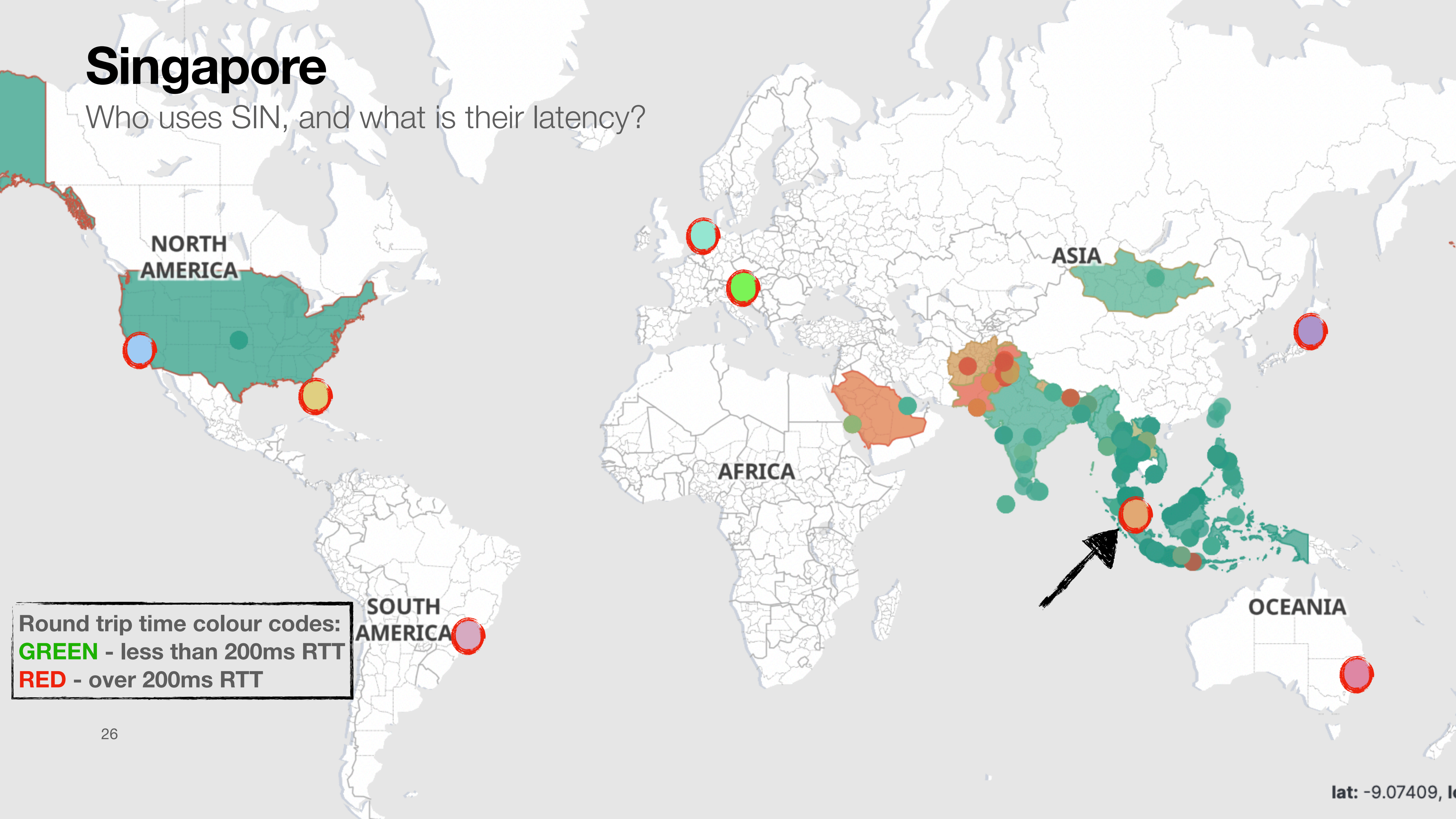
Ljubljana

Who uses LJU, and what is their latency?



Singapore

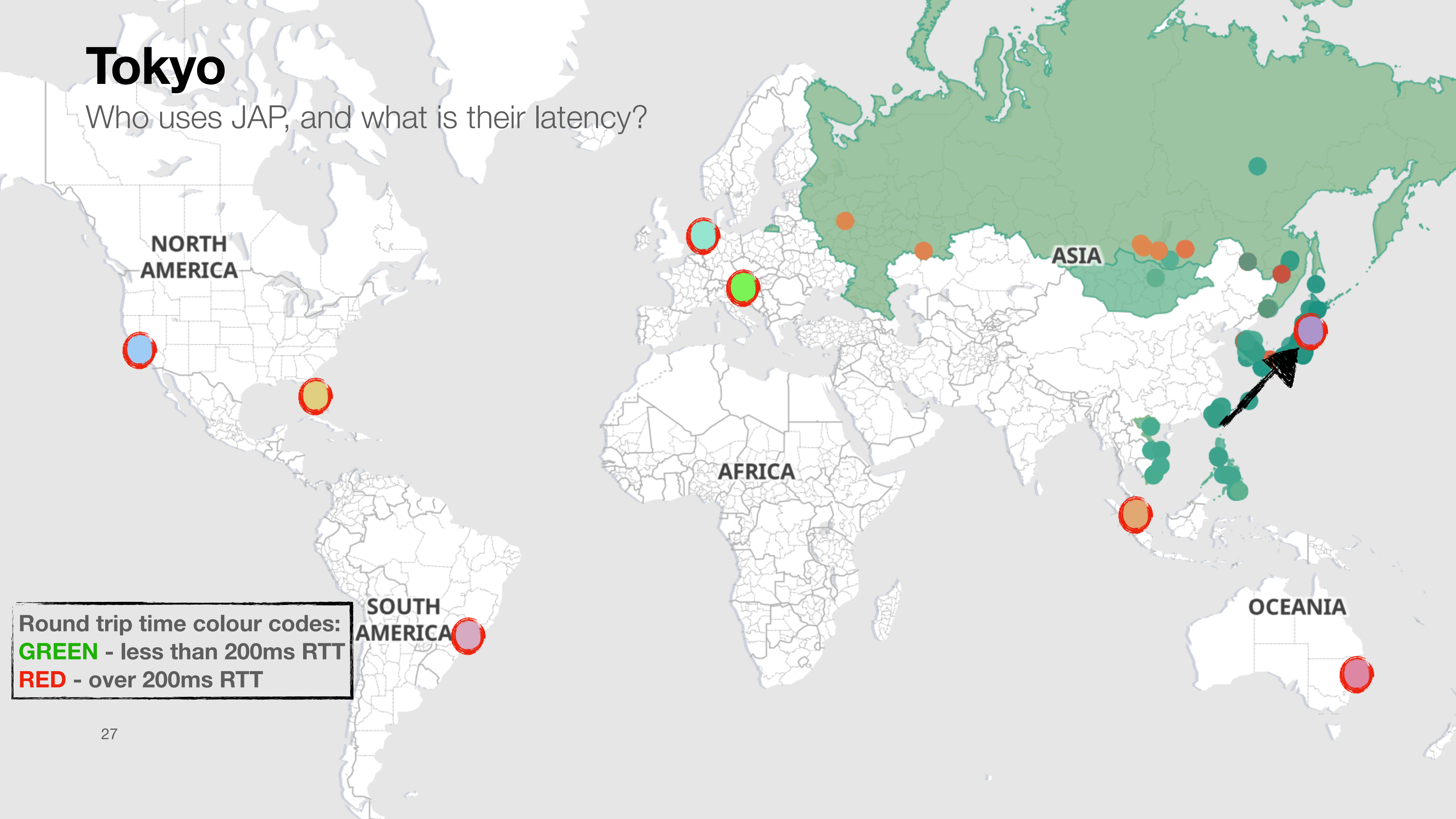
Who uses SIN, and what is their latency?



Round trip time colour codes:
GREEN - less than 200ms RTT
RED - over 200ms RTT

Tokyo

Who uses JAP, and what is their latency?



Round trip time colour codes:
GREEN - less than 200ms RTT
RED - over 200ms RTT

Sydney

Who uses SYD, and what is their latency?

NORTH AMERICA



ASIA



AFRICA

SOUTH AMERICA



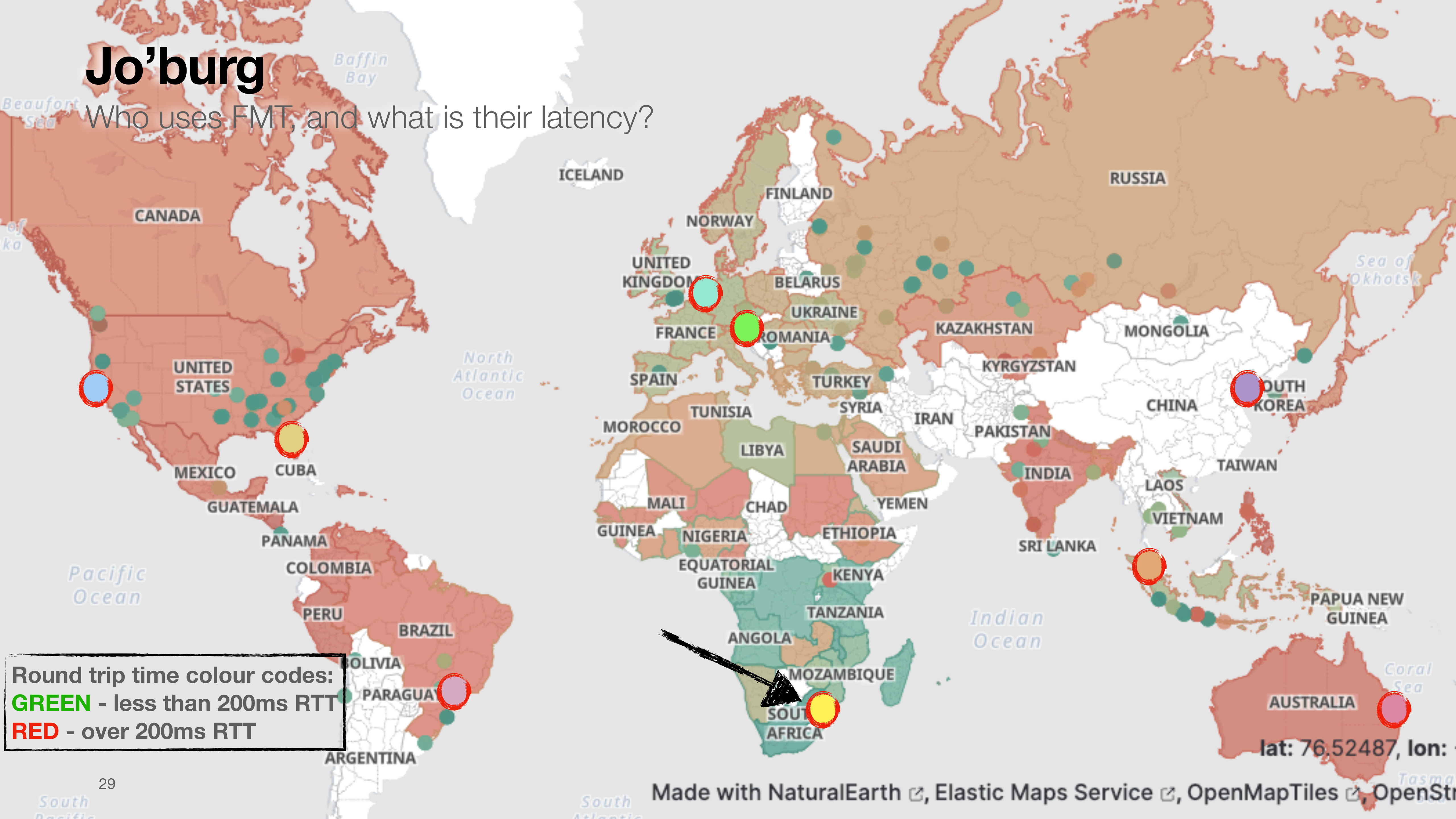
OCEANIA



Round trip time colour codes:
GREEN - less than 200ms RTT
RED - over 200ms RTT

Jo'burg

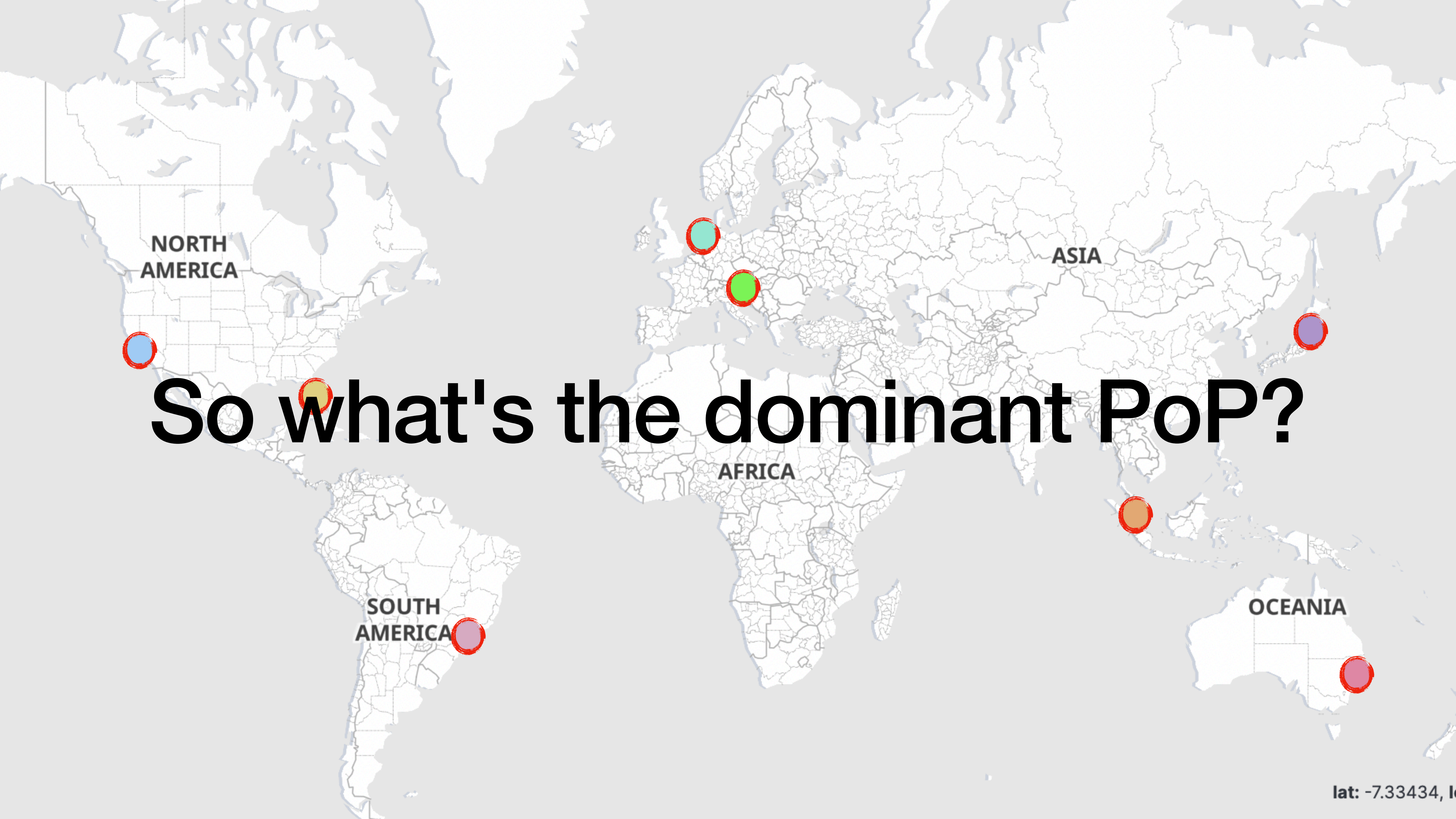
Who uses FMT, and what is their latency?



Those were the snapshots...

Zooming out a bit

- Routing changes over time
- Let's look at a slightly longer timeframe
 - I have arbitrarily chosen a week



NORTH AMERICA

ASIA

AFRICA

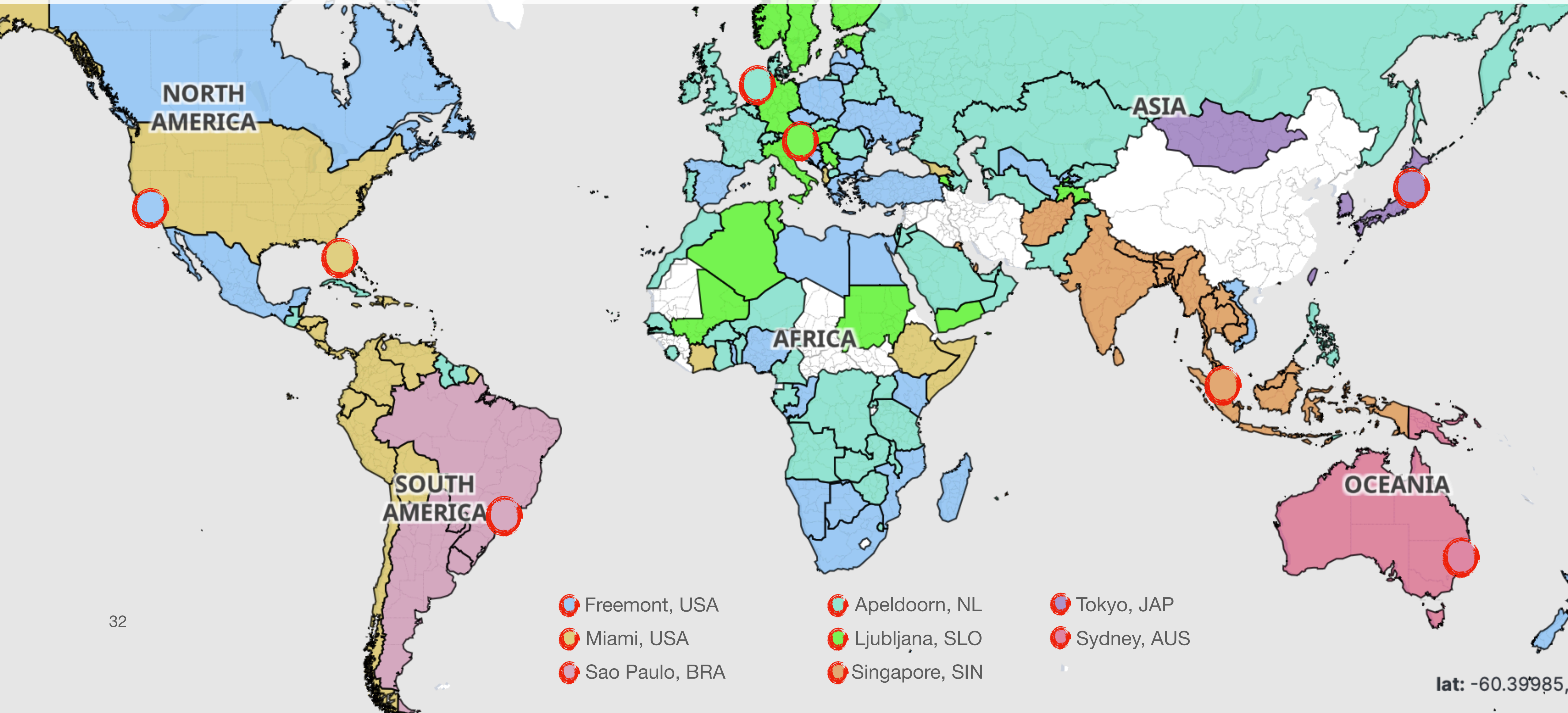
SOUTH AMERICA

OCEANIA

So what's the dominant PoP?

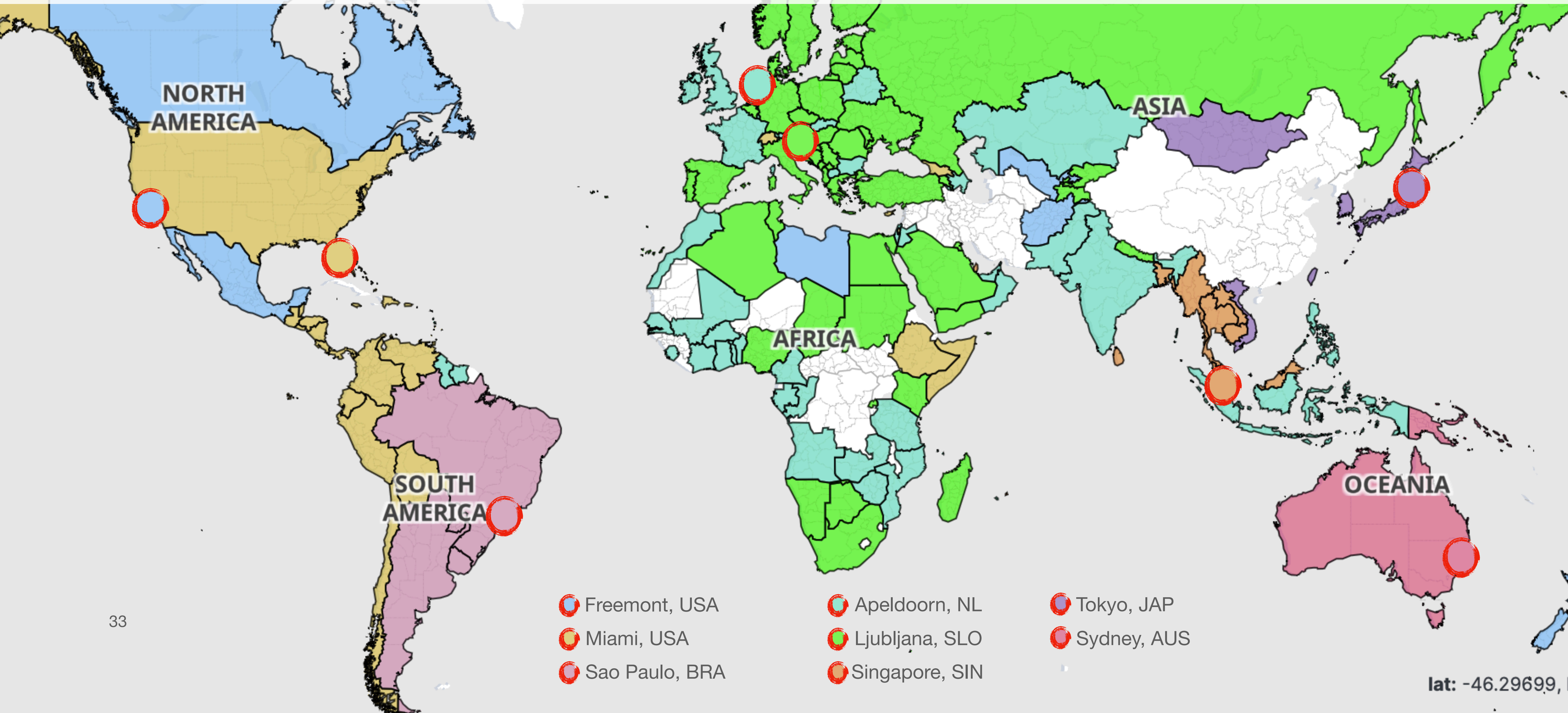
Global routing changes (preferred node, not RTT :)

Many weeks ago - which country prefers which node/PoP ?



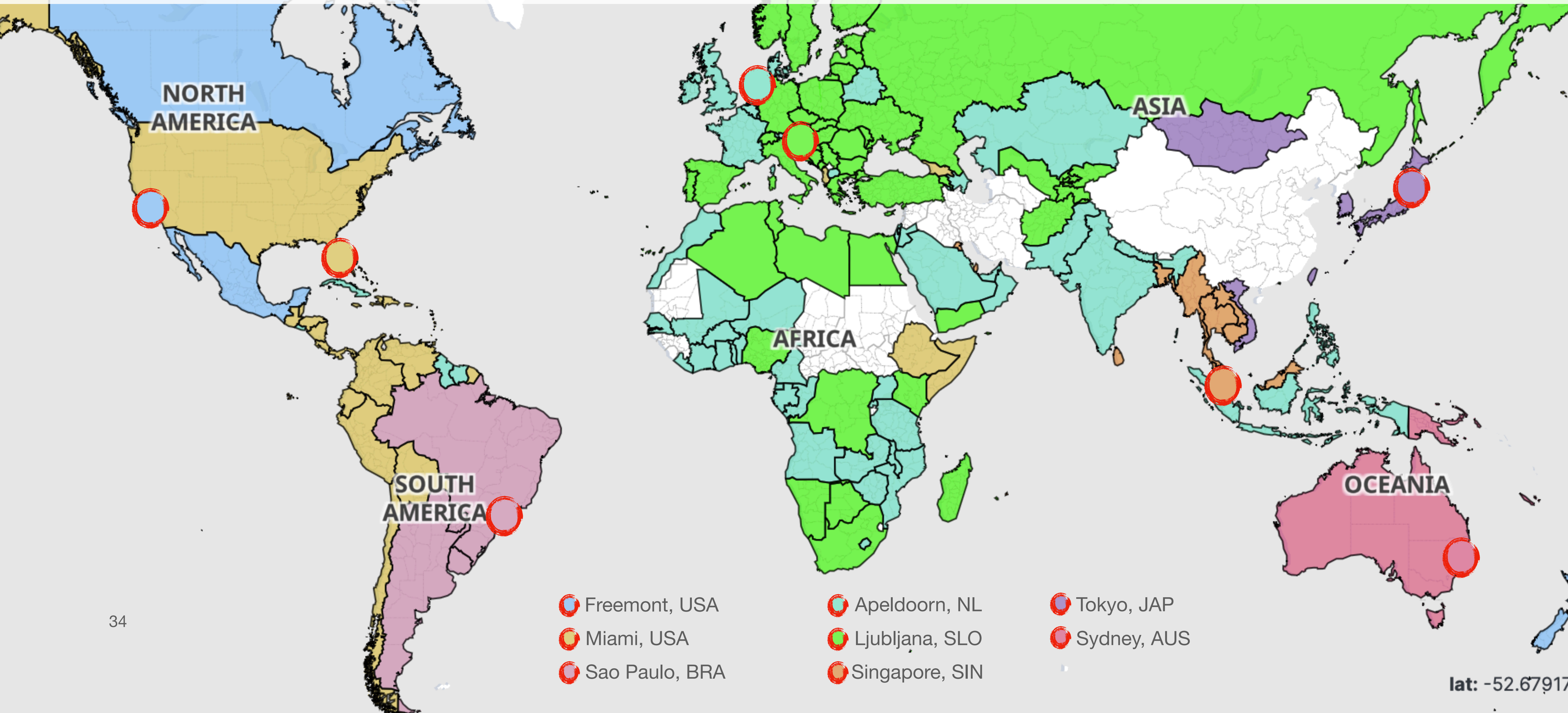
Global routing changes (preferred node, not RTT :)

A week later - which country prefers which node/PoP ?



Global routing changes (preferred node, not RTT :)

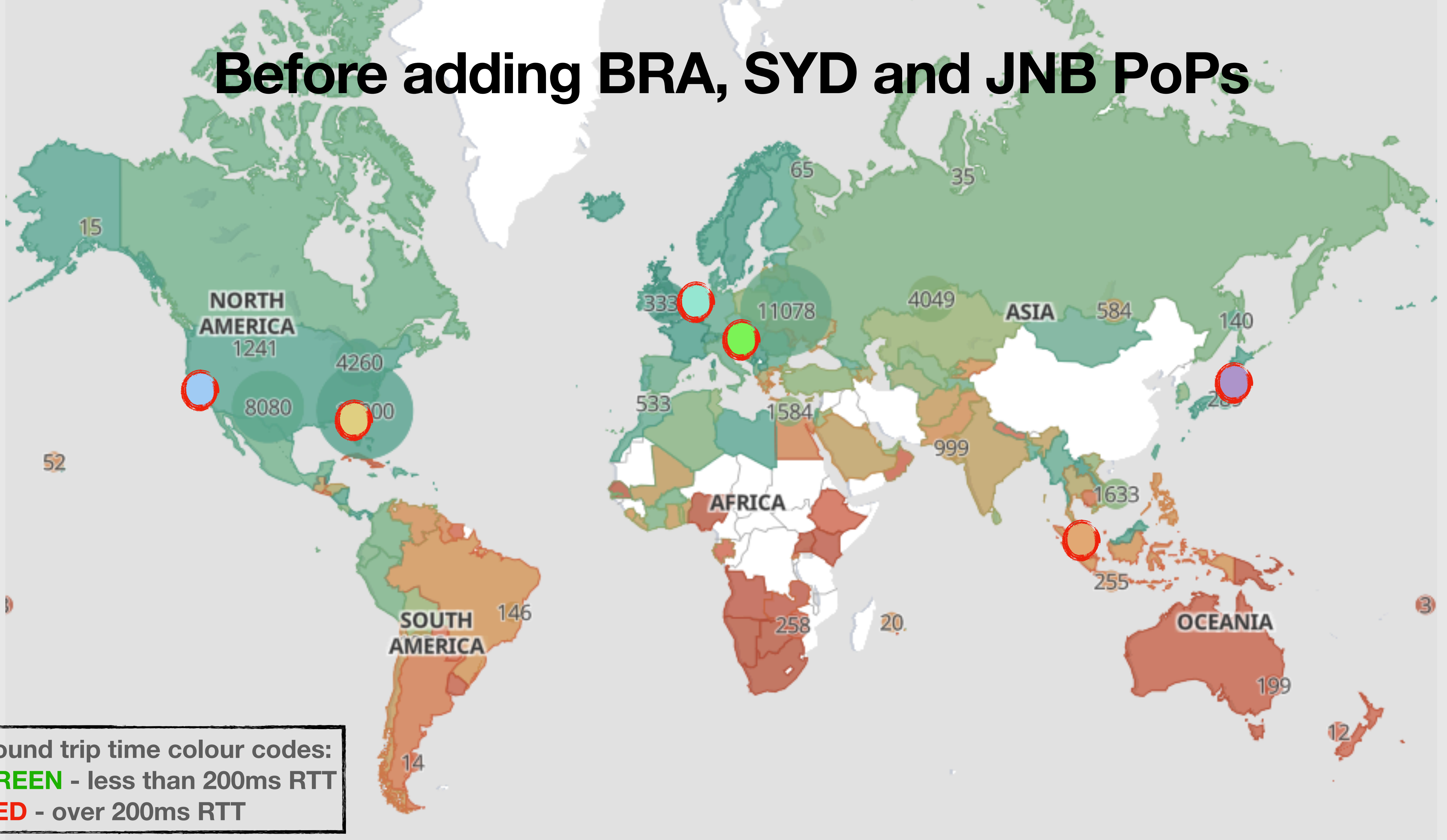
Two weeks later - which country prefers which node/PoP ?

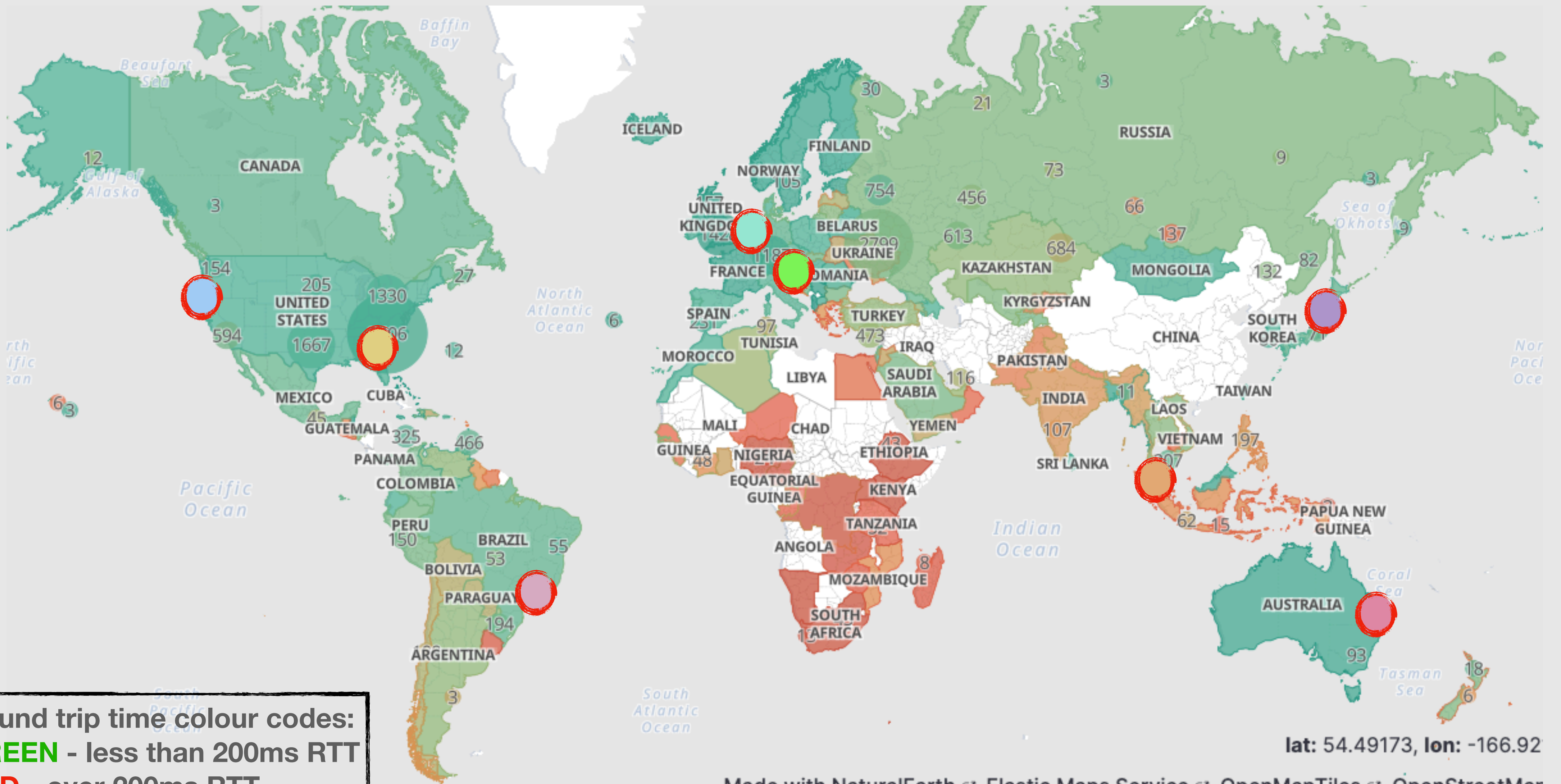


How the RTT map changed over time

(adding new PoPs)

Before adding BRA, SYD and JNB PoPs

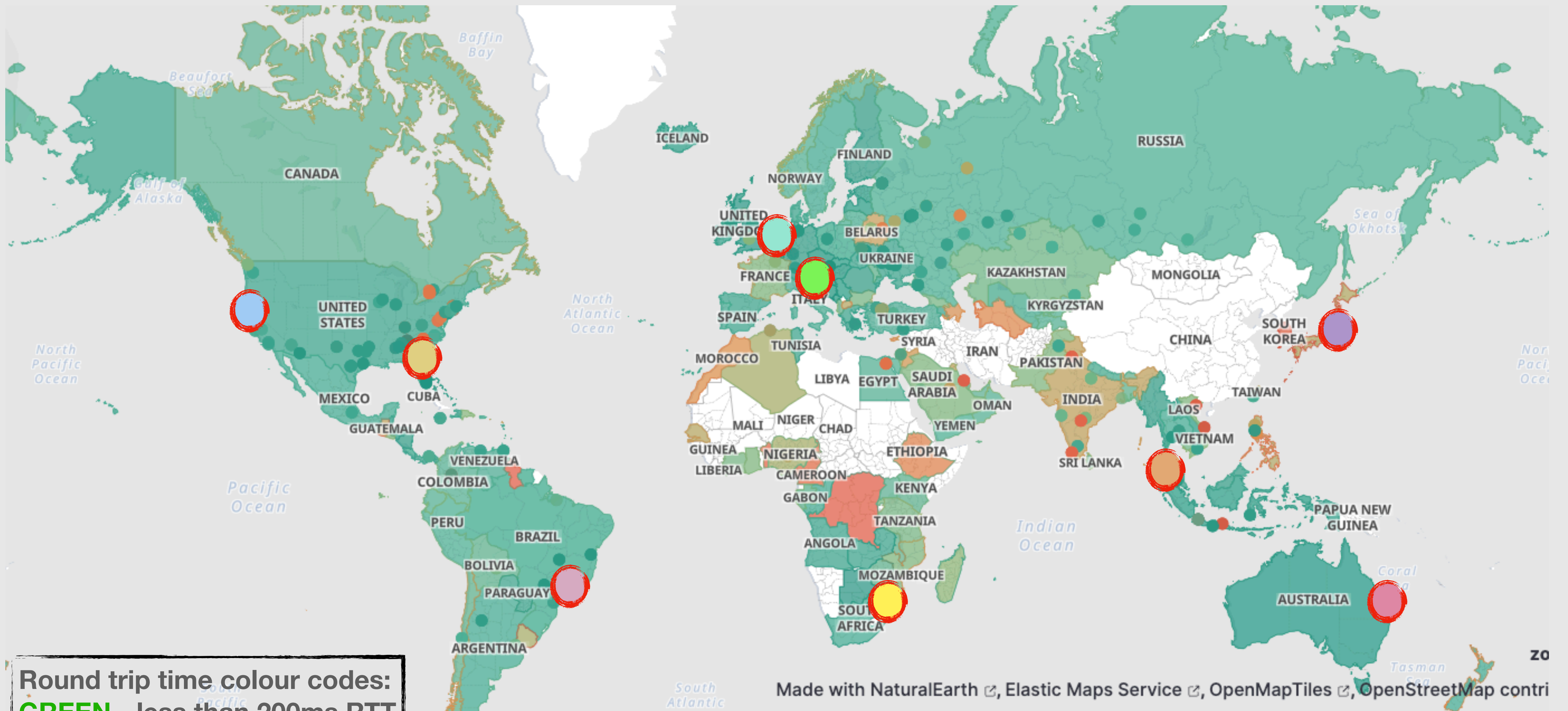


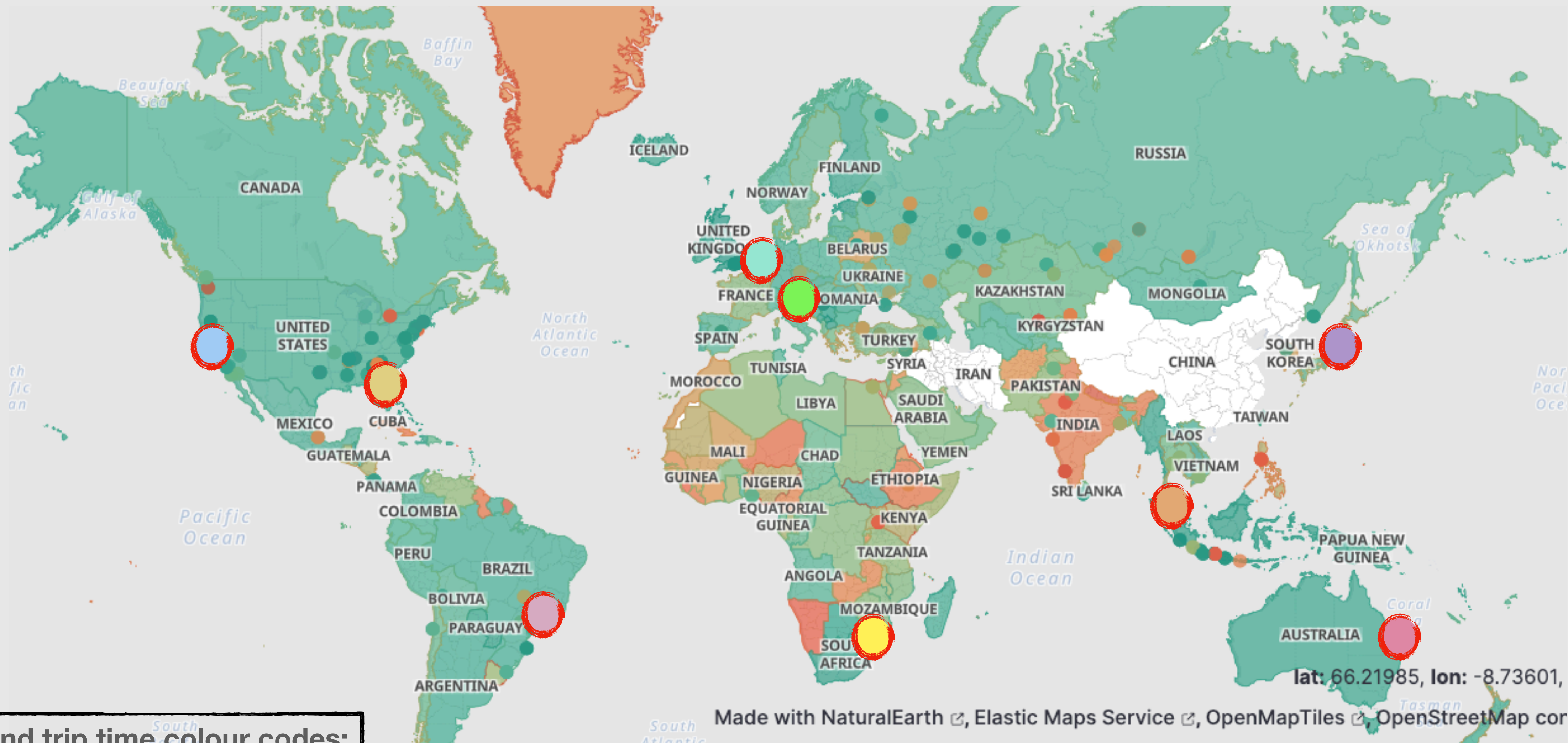


Round trip time colour codes:
GREEN - less than 200ms RTT
RED - over 200ms RTT

lat: 54.49173, lon: -166.92

Made with NaturalEarth, Elastic Maps Service, OpenMapTiles, OpenStreetMap





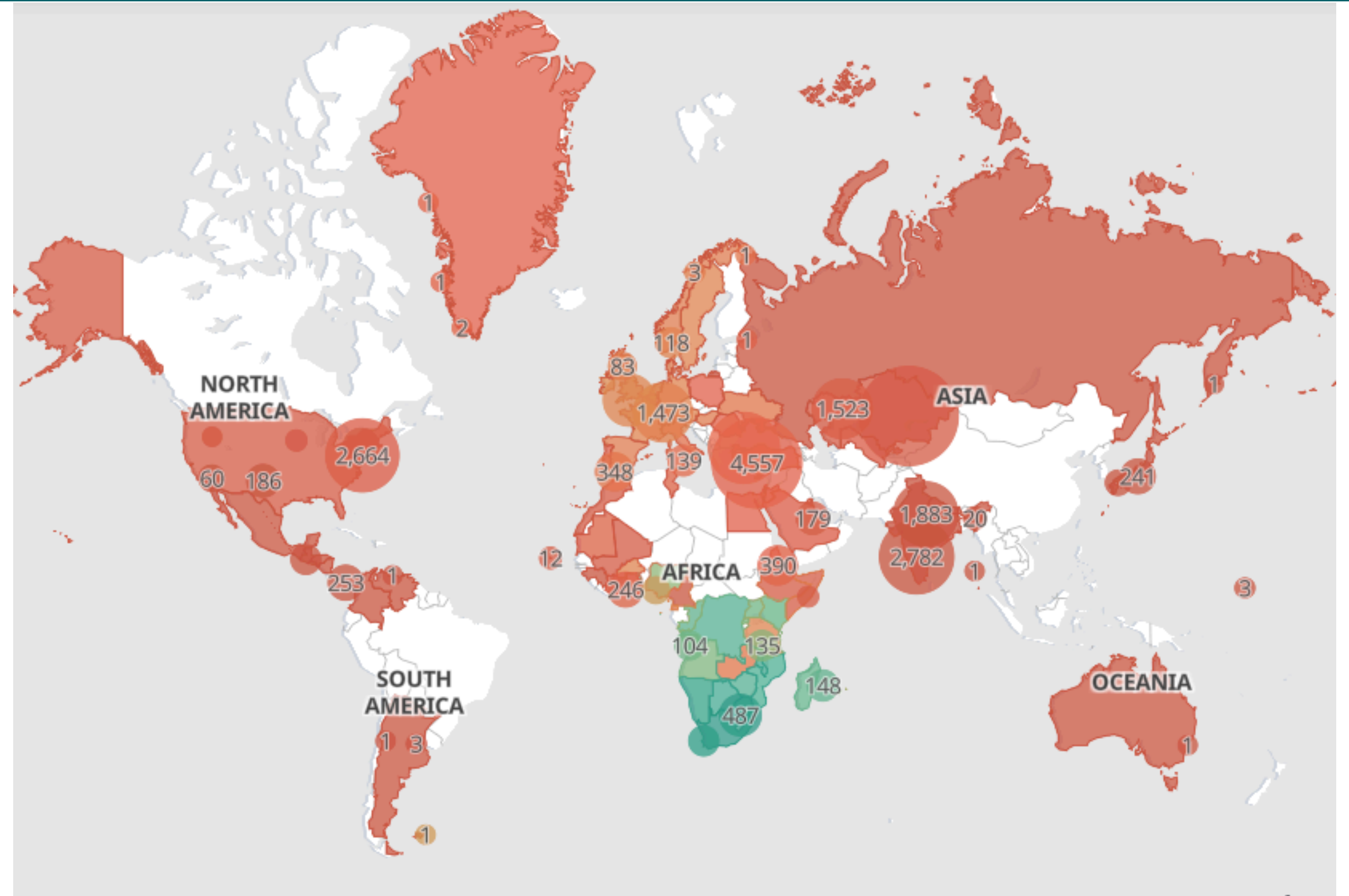
Round trip time colour codes:
GREEN - less than 200ms RTT
RED - over 200ms RTT

Too well connected peer...

Node in Jo'burg is connected to extremely well peered connectivity provider, hence the BGP announcement is preferred far too wide globally.

How to fix that?

- BGP communities (specific to upstream)
- Pre-pending



Present state?

- Running Anycast DNS services for 6connect (6connect.com, 6clabs.com, all the reverse DNS zones, etc...)
- Hosting Ukraine .UA TLD (64 zones, .ua., com.ua., kyiv.ua., etc...)
- Building distributed and anycasted email system for our use
- Designing Anycast IPAM cloud service that we'll try to build in the future.
- Simplifying of the anycast node - this is maybe too complex as it is:
 - Maybe one DNS daemon per node, three different on site
 - No dnsdist...

Questions?

Jan Žorž - jan@6connect.com
<https://6connect.com/>

