Removing IPv4 infrastructure addressing from Meta's edge network

RIPE 88
Kraków

Matt Kirkland
Network Engineer

FACEBOOK Infrastructure

Agenda

Introduction and Motivation

Approach

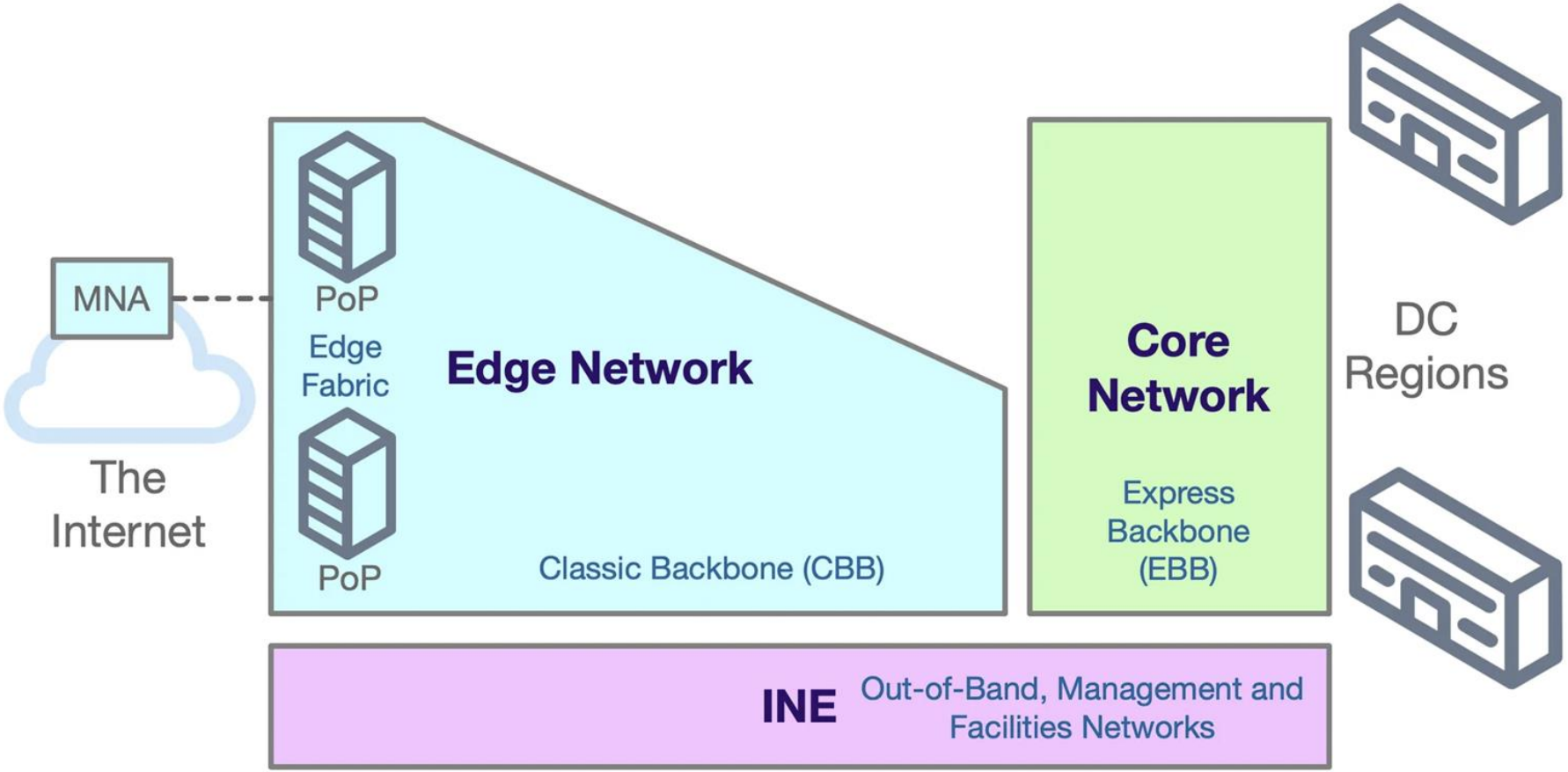
Lessons Learned

What next?

Q&A

Introduction and Motivation

Meta Networks



37%

Traffic between users and Meta is over IPv6. Edge network dual-stacked.

Source: facebook.com/ipv6

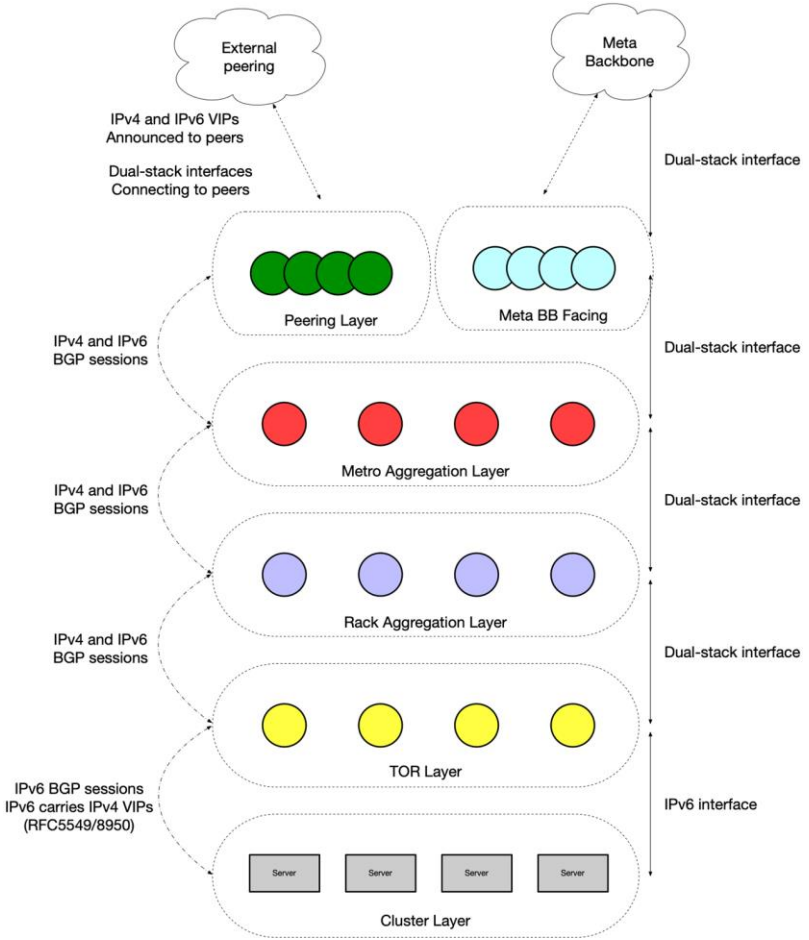
>99%

Internal traffic is over IPv6

Source: Internal report

Edge Network - Dual Stack

- Traffic is a mix of v6 and v4.
- Server to ToR addressing is v6 only, v4 VIPs announced via v6 BGP sessions with v6 next-hop (RFC5549/8950)
- Other infrastructure links are dual-stacked, dedicated v4 and v6 addresses and BGP sessions.
- Links to peers are dual-stacked if peer supports it.




Why do anything
more?

Motivation

Simplification

Maintaining two sets of address families increases engineering and operational overhead.

Scale

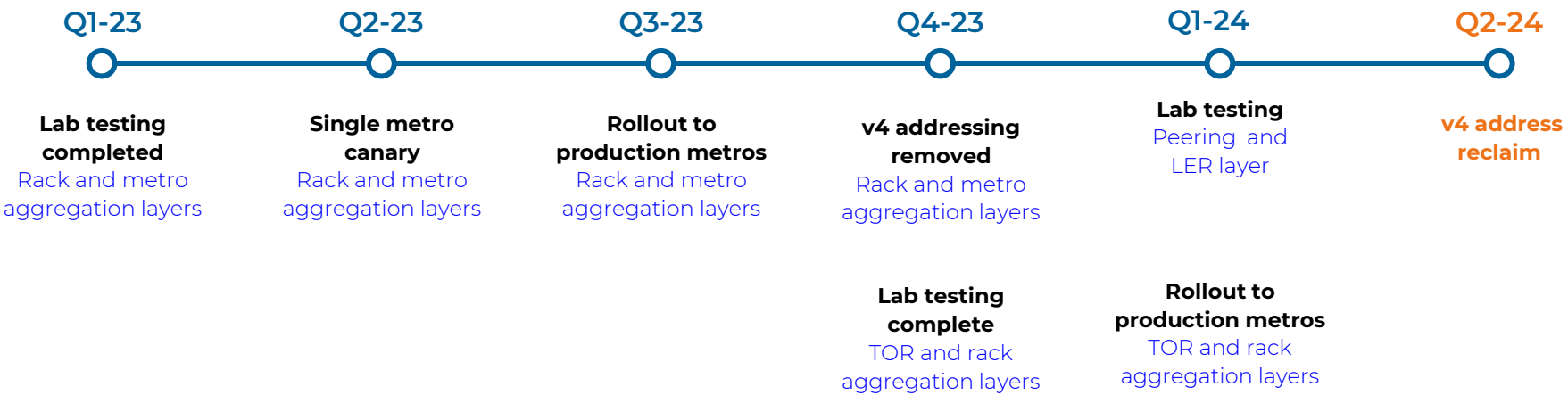
Our edge network infrastructure is sufficiently large that we have run into scaling problems with IPv4 addressing.

Planning Overhead

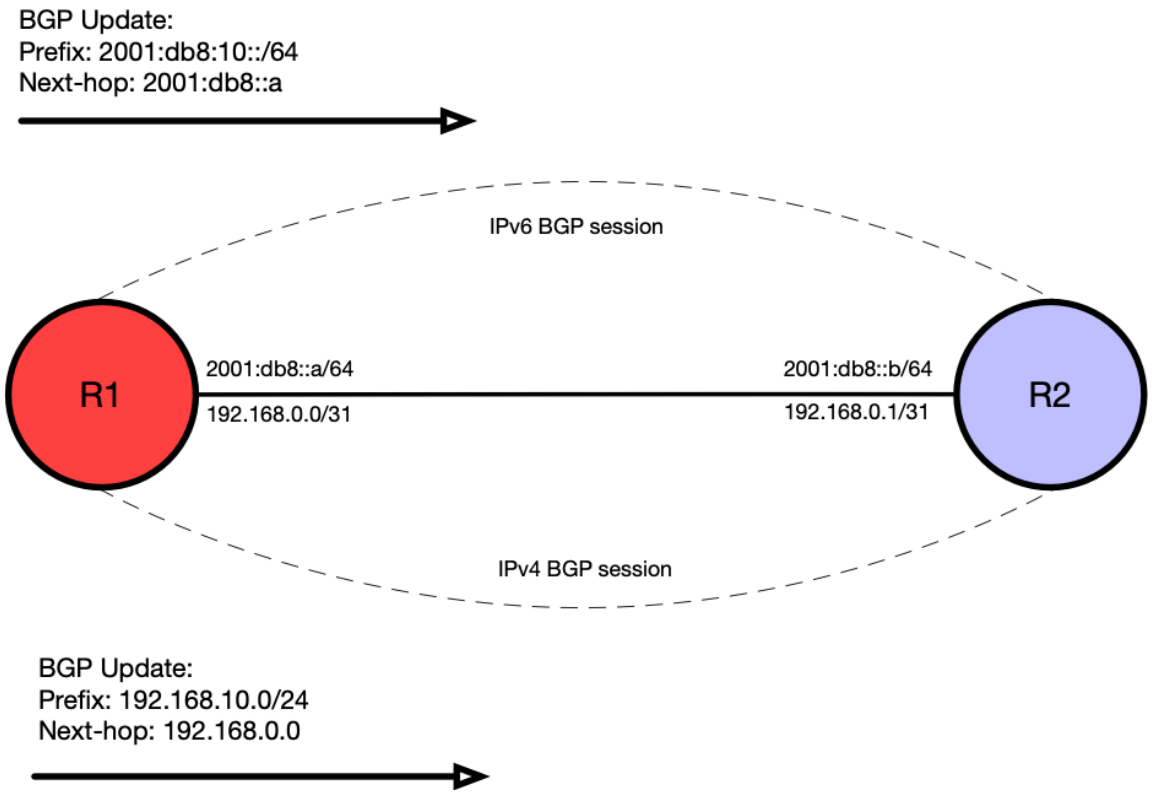
IPv4 is a valuable and finite resource, wherever used it needs to be carefully planned. Avoiding using it removes this need entirely.

Approach

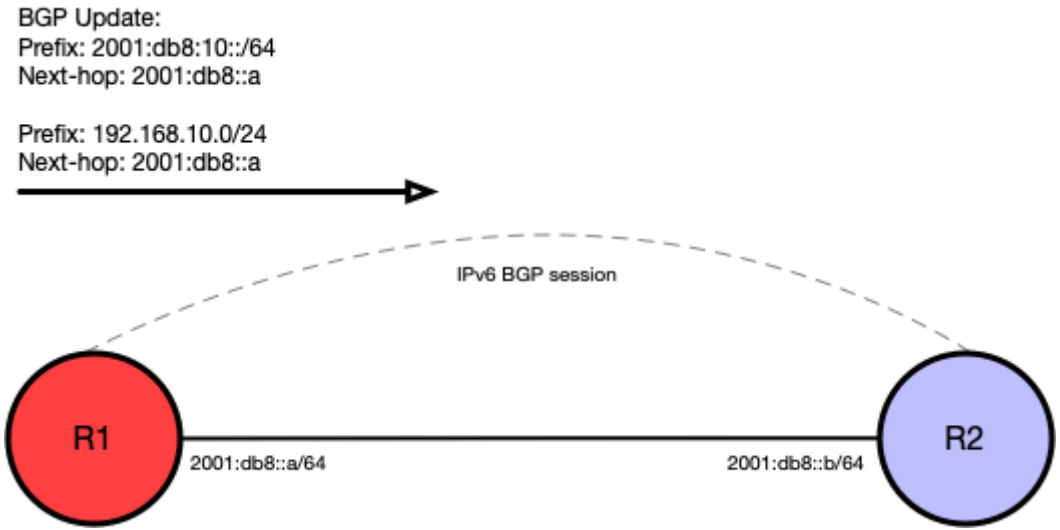
Timeline


Q1-23
Lab testing completed
Rack and metro aggregation layers
Q2-23
Single metro canary
Rack and metro aggregation layers
Q3-23
Rollout to production metros
Rack and metro aggregation layers
Q4-23
v4 addressing removed
Rack and metro aggregation layers
Lab testing complete
TOR and rack aggregation layers
Q1-24
Lab testing
Peering and LER layer
Rollout to production metros
TOR and rack aggregation layers
Q2-24
v4 address reclaim

Dual Stack

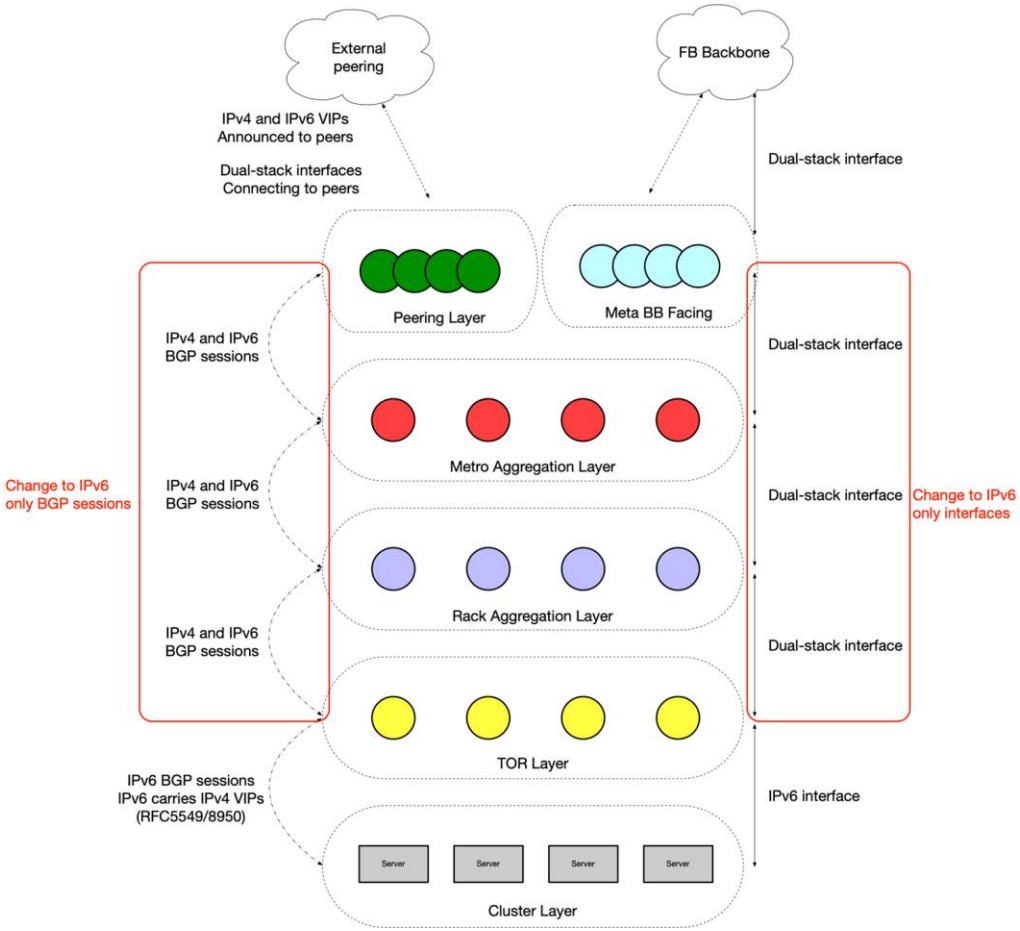

BGP Update:
Prefix: 2001:db8:10::/64
Next-hop: 2001:db8::a
IPv6 BGP session
R1
2001:db8::a/64
192.168.0.0/31
2001:db8::b/64
192.168.0.1/31
R2
IPv4 BGP session
BGP Update:
Prefix: 192.168.10.0/24
Next-hop: 192.168.0.0

RFC 5549/8950



Edge Network - v6 only linknets

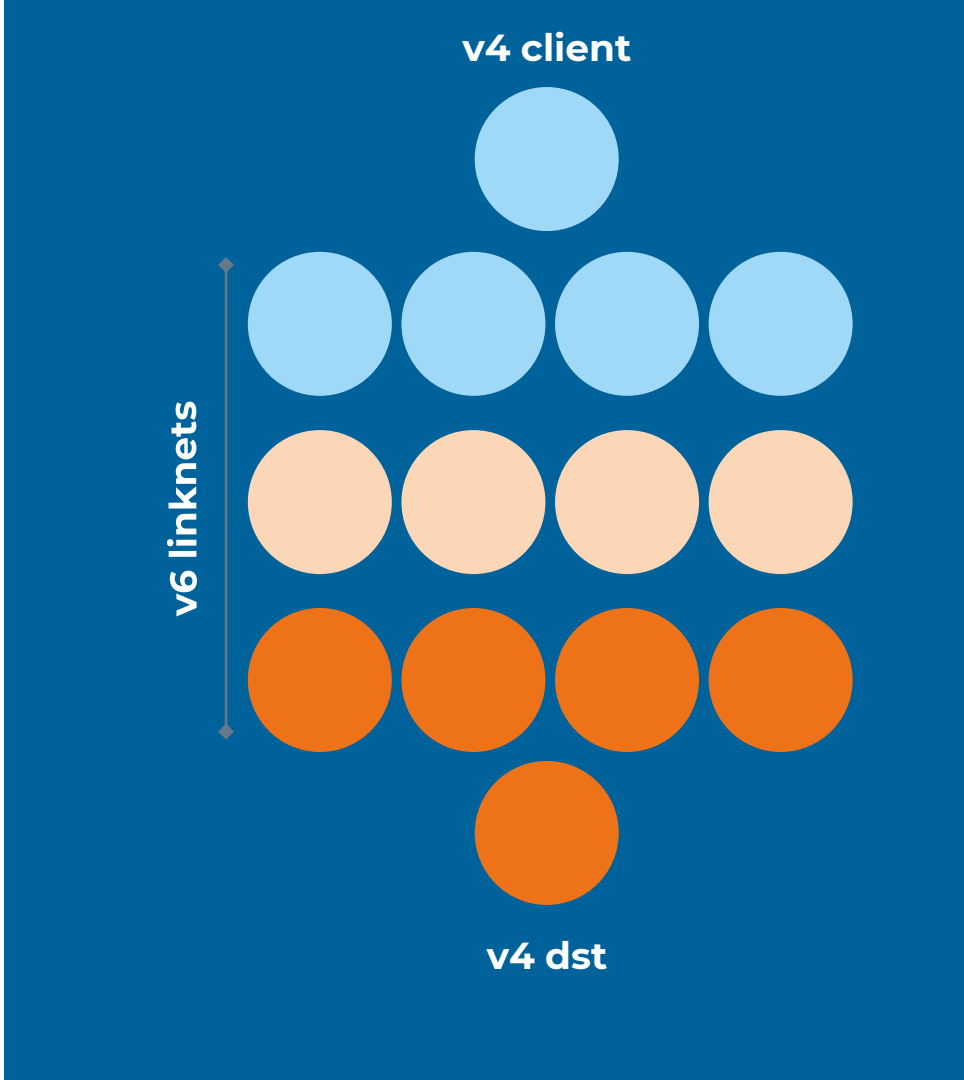
- Server to ToR addressing is v6 only already, nothing further needed.
- Enable v4 address family over existing v6 sessions.
- Remove IPv4 BGP sessions and IPv4 addressing from all affected links.



Challenges

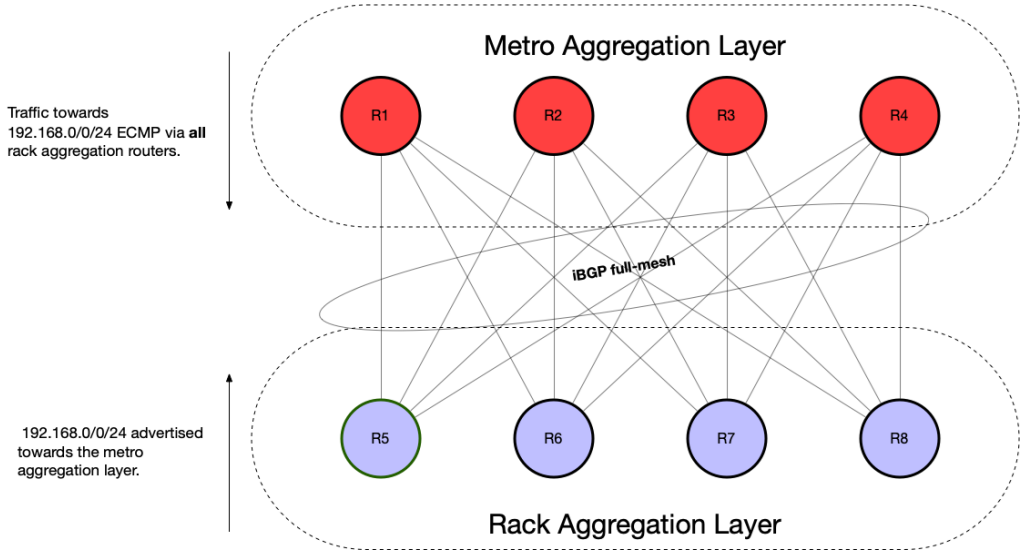
What about traceroute?

- Typically routers will send TTL expired message sourced from the IP address associated with the outbound interface towards the sender.
- If the interface no longer has an IPv4 address, what happens?
  - The router will reply using the loopback address.
- RFC8335 and RFC5837 improving ping/traceroute.



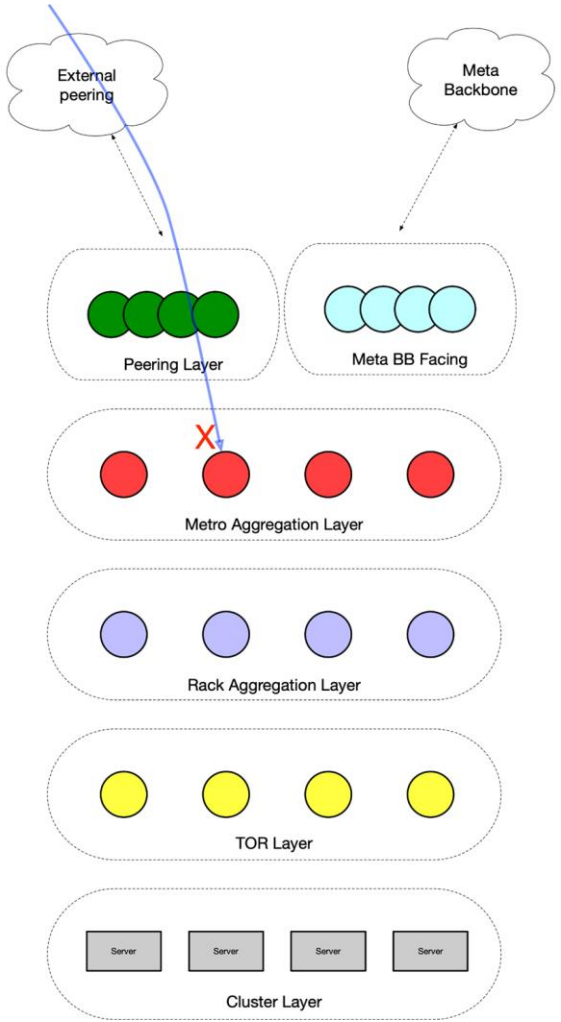
ECMP not that equal

- Inter-layer connectivity is fully-meshed, many ECMP paths.
- *Some routers would not do ECMP between paths learnt with v4 and v6 next-hop, even if all other BGP attributes matched. Vendor specific behaviour.*
- Needed to increase the weight of the routes with v4 next-hop, until all v4 NLRI had been learnt via v6 sessions.
- Three vendors, three different approaches.



Some v4 just dropped

- Some router platforms would drop v4 packets if they didn't have a v4 address configured.
- Needed additional command to forward v4 traffic without a v4 address.
- This was not consistent across platforms, even from the same vendor.



Interface counters not consistent

- Platforms reported counters in different ways.
- Some platforms would report v6/v4 counters in a single direction only.
- Some would report total and v6 so we needed to subtract the v6 number from total to derive v4.
- Important to test this!

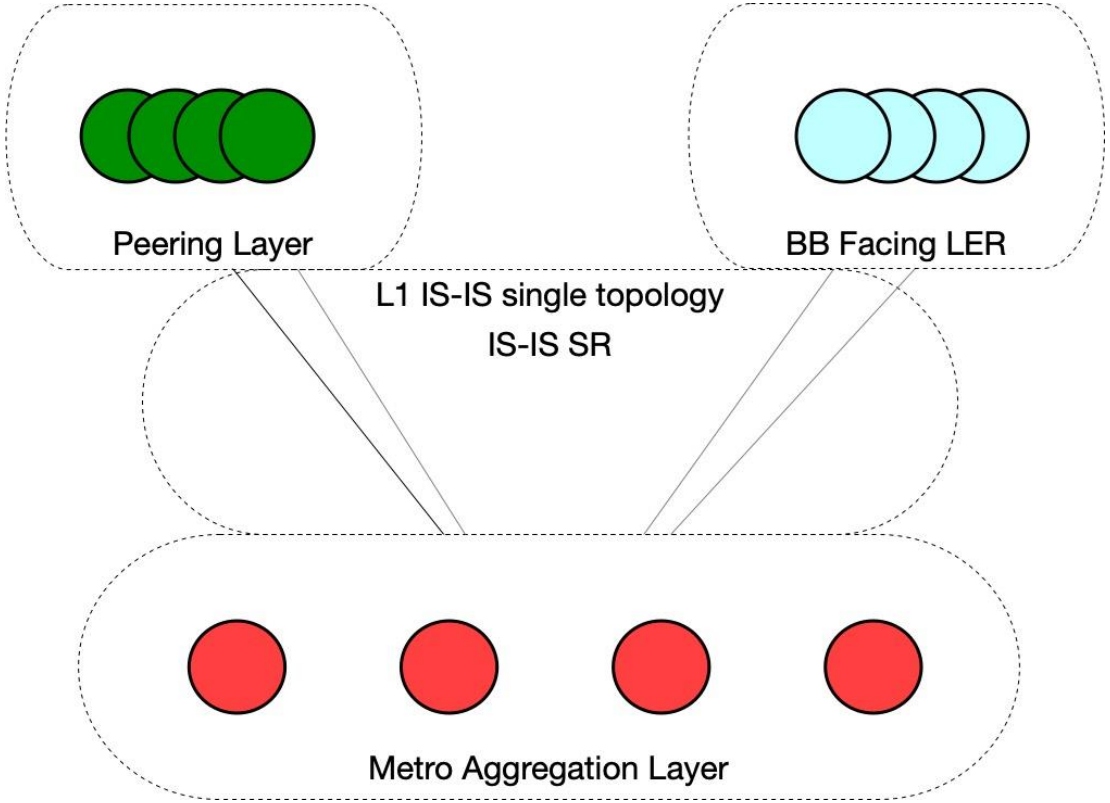
Mixed v6/v4 bgp policy

- Different vendors handle mixed address family policy differently
- Some vendors allow distinct policies per address family
- Other vendors just have policy per peer and the policy must include all rules for all address families

```
!
route-map RACK-AGG-TO-RACK-OUT-V6 deny 10
   description "do not leak interconnects"
   match ipv6 address prefix-list INTERCONNECT-V6
   match community DIRECT
!
route-map RACK-AGG-TO-RACK-OUT-V6 deny 20
   description "do not leak interconnects"
   match ip address prefix-list INTERCONNECT-V4
   match community DIRECT
!
```

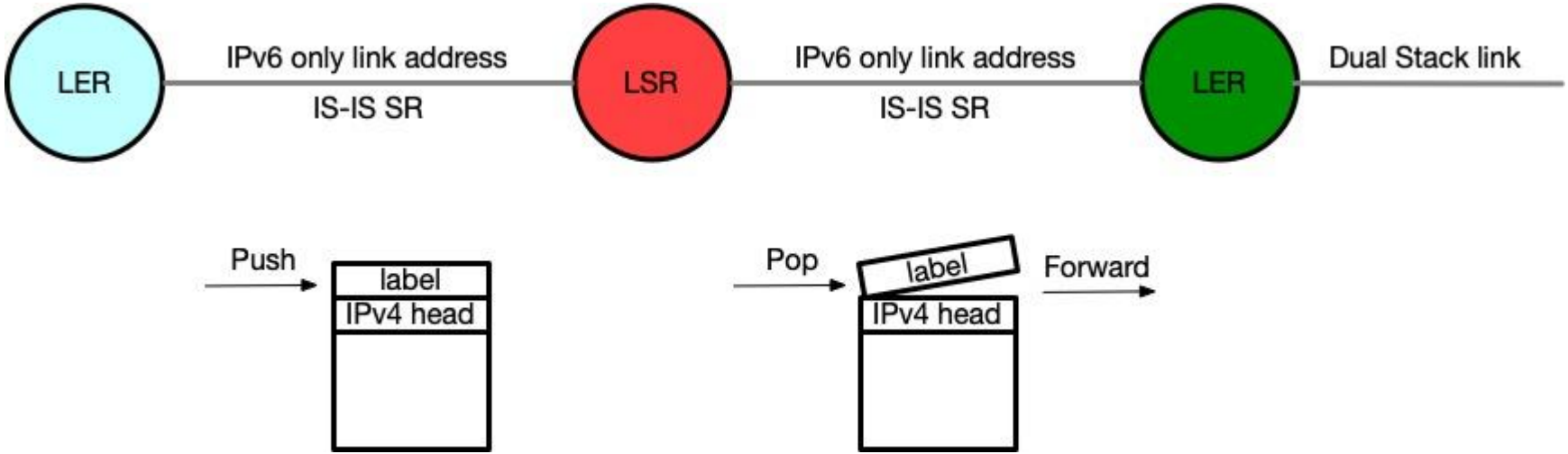
Peering layer migration

- Peering, BB and aggregation layer are part of IS-IS L1 domain.
- IS-IS Segment routing is enabled.
- Removing v4 link addresses, some vendors require we move to MT IS-IS, while others do not.
- This makes migration operations very disruptive.



Labelled forwarding

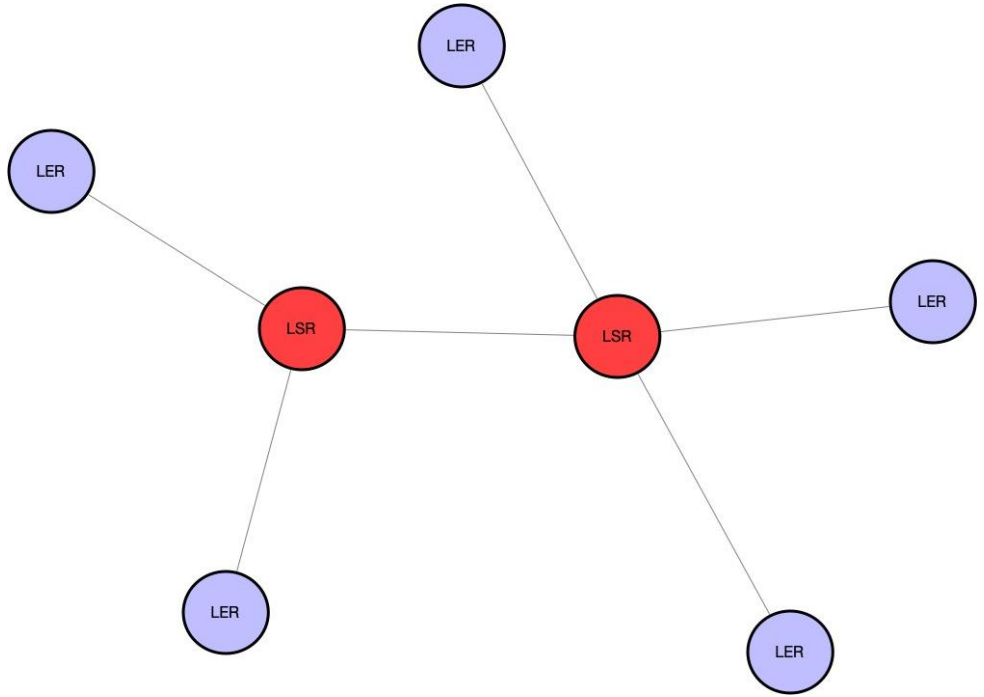
- Using IS-IS segment routing on IPv6 only links
- V4 packet is encapsulated
- imp-null(3) is used to forward packet at the penultimate hop
- Not universally supported, some vendors perform a check on the penultimate hop and drop the v4 packet because the link is v6 only



What next?

V6 only across the core

- All links are addressed in v4 and v6
- RSVP-TE core network
- Vendor implementations of RSVP are IPv4 only, despite RFC3209 supporting IPv6
- RFC5549 BGP prefix exchange works.
- Not a lot of excitement to re-engineer the backbone.



Summary

01

RFC5549/8950 works

v4 NLRI with v6 next-hop works, we're running it in production

02

Need to test

There were some hurdles along the way, no show stoppers but important to understand platform and vendor behaviour

03

It's worth it

Simplified configurations, provisioning workflows and planning

Thank You!

Questions?