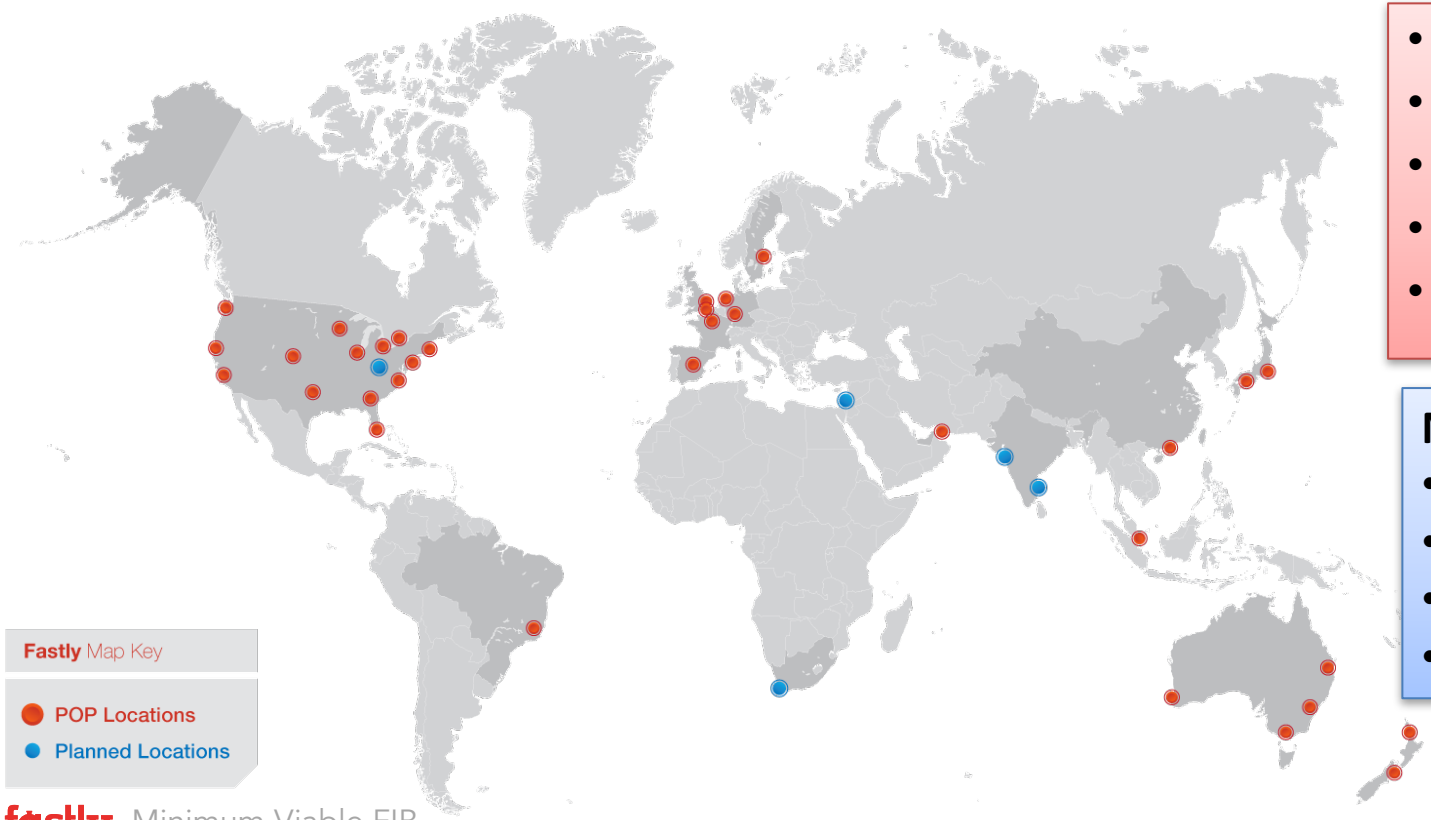
Minimum Viable FIB
a.k.a Route Scale on ToR

Tom Daly
VP, Infrastructure
tjd@fastly.com

The Days of CDN Life



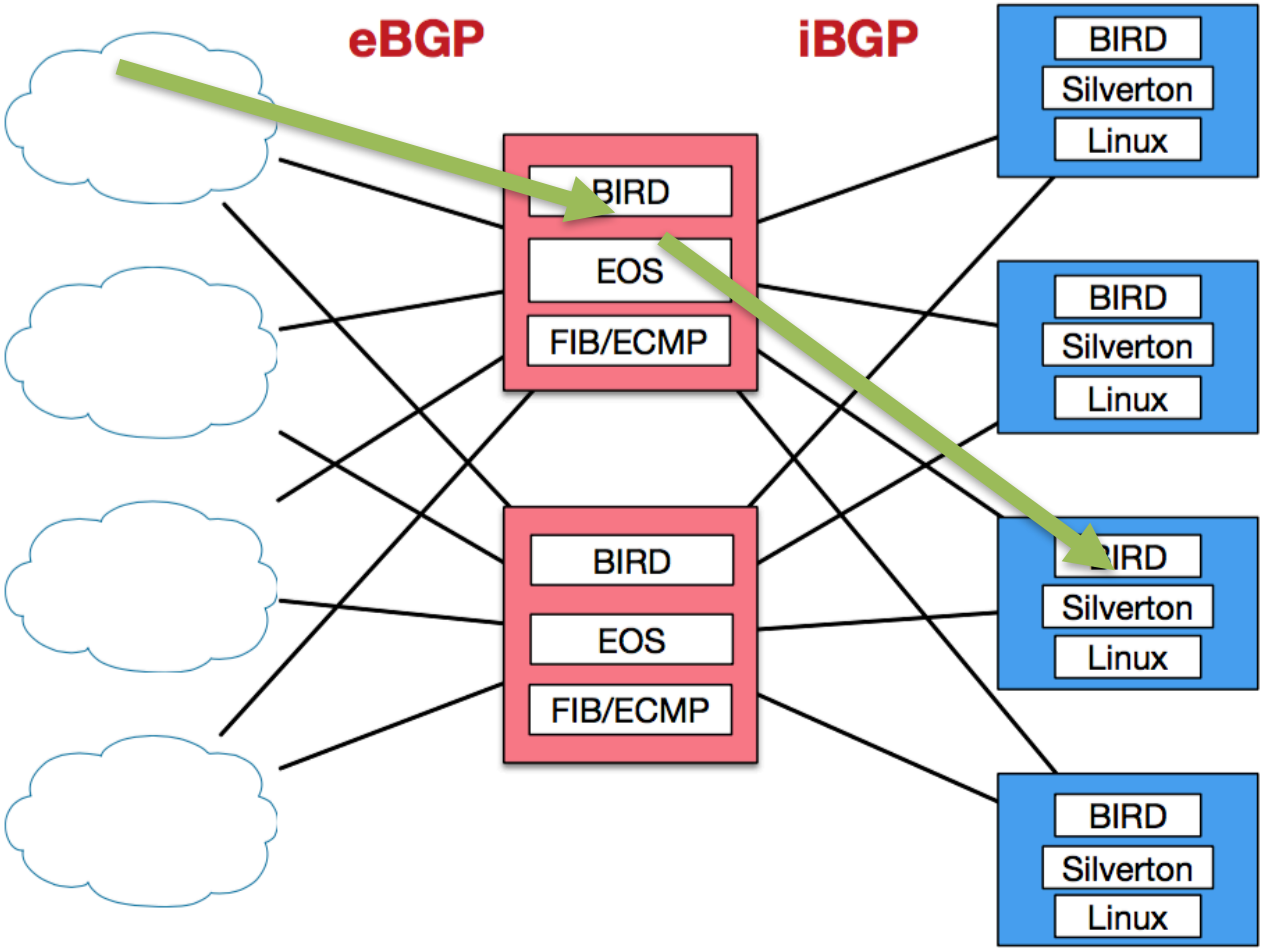
- Routers: NOPE
- LBs: NOPE
- Linux: OK
- ToR: OK
- BGP Tricks: OK.

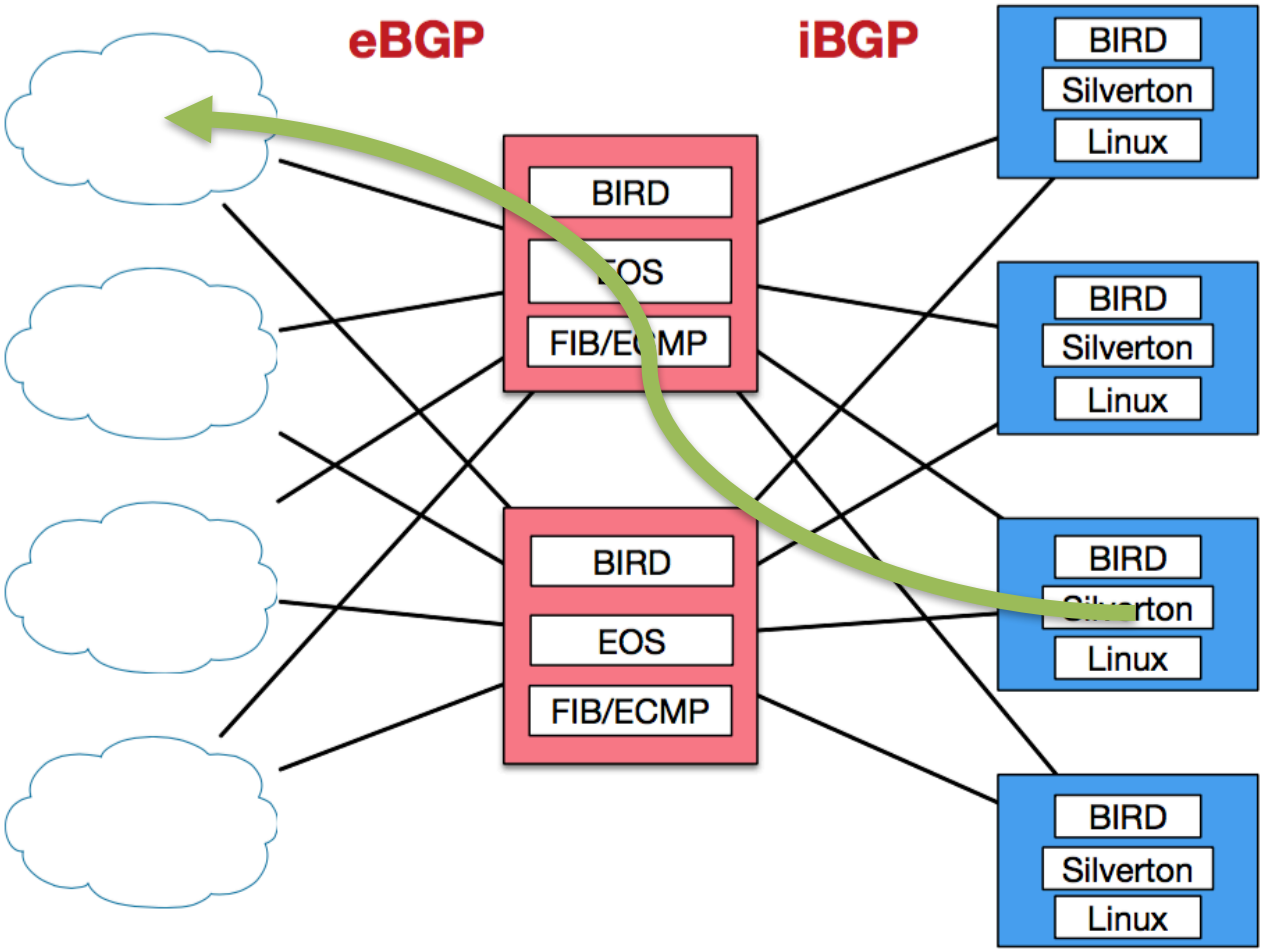
Network Statistics

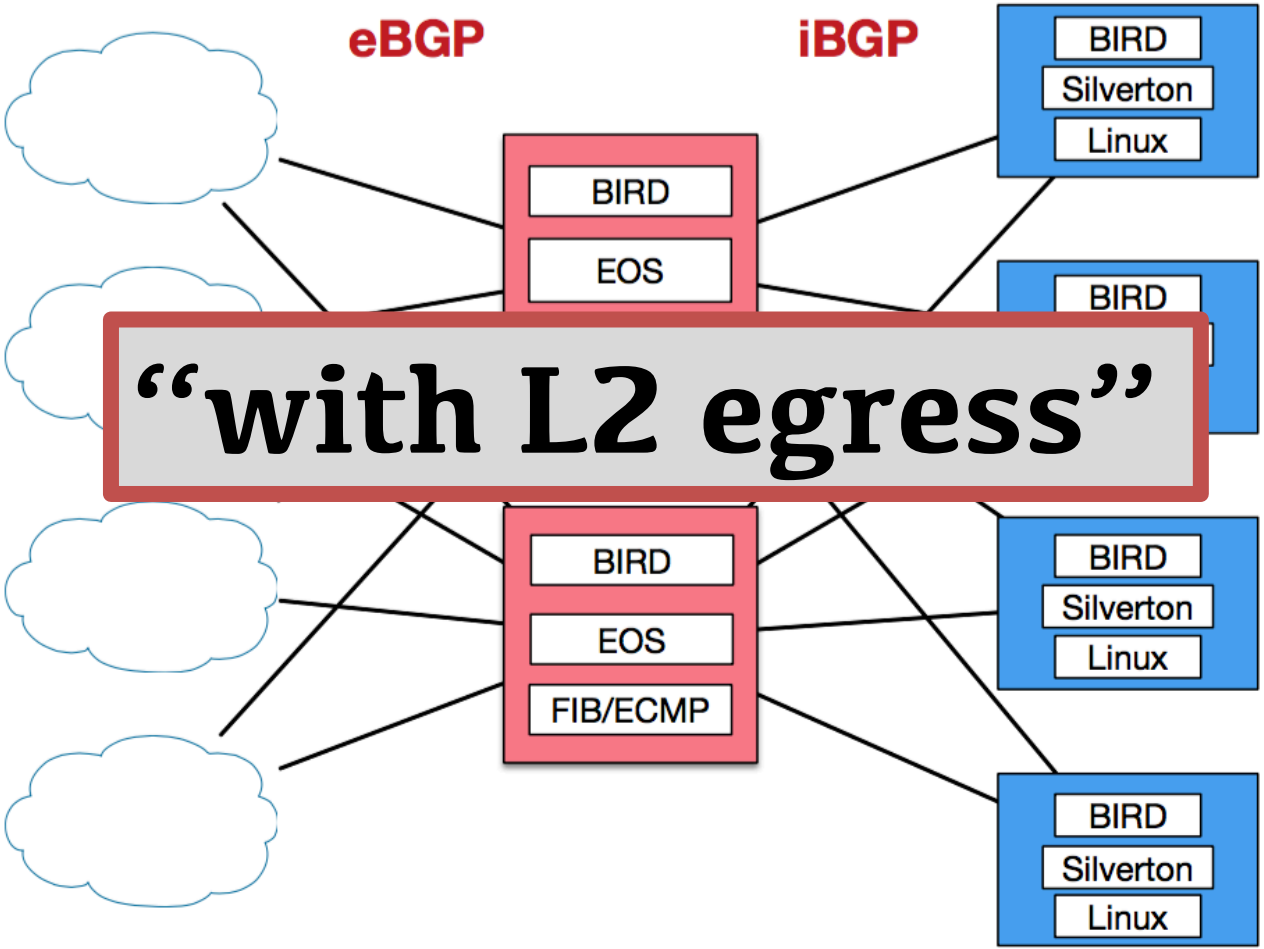
- 35 POPs
- 10 Tbps Edge
- ~200 Peers
- ~37 IXPs

Cache Egress Routing at Fastly

- Dumb switches + smart caches = smart things.
  - ECMP load balance with L3 ingress.
    - Software to manage ECMP consistent hashes and health.
  - Multipath outbound **with L2 egress.**
    - Software to manage egress NH and MAC changes
  - BIRD in switch linux userland:
    - Terminate peer eBGP sessions in a sane location.
    - Route reflect via iBGP to every cache.
  - Switch FIB: Multiple default routes for backup.

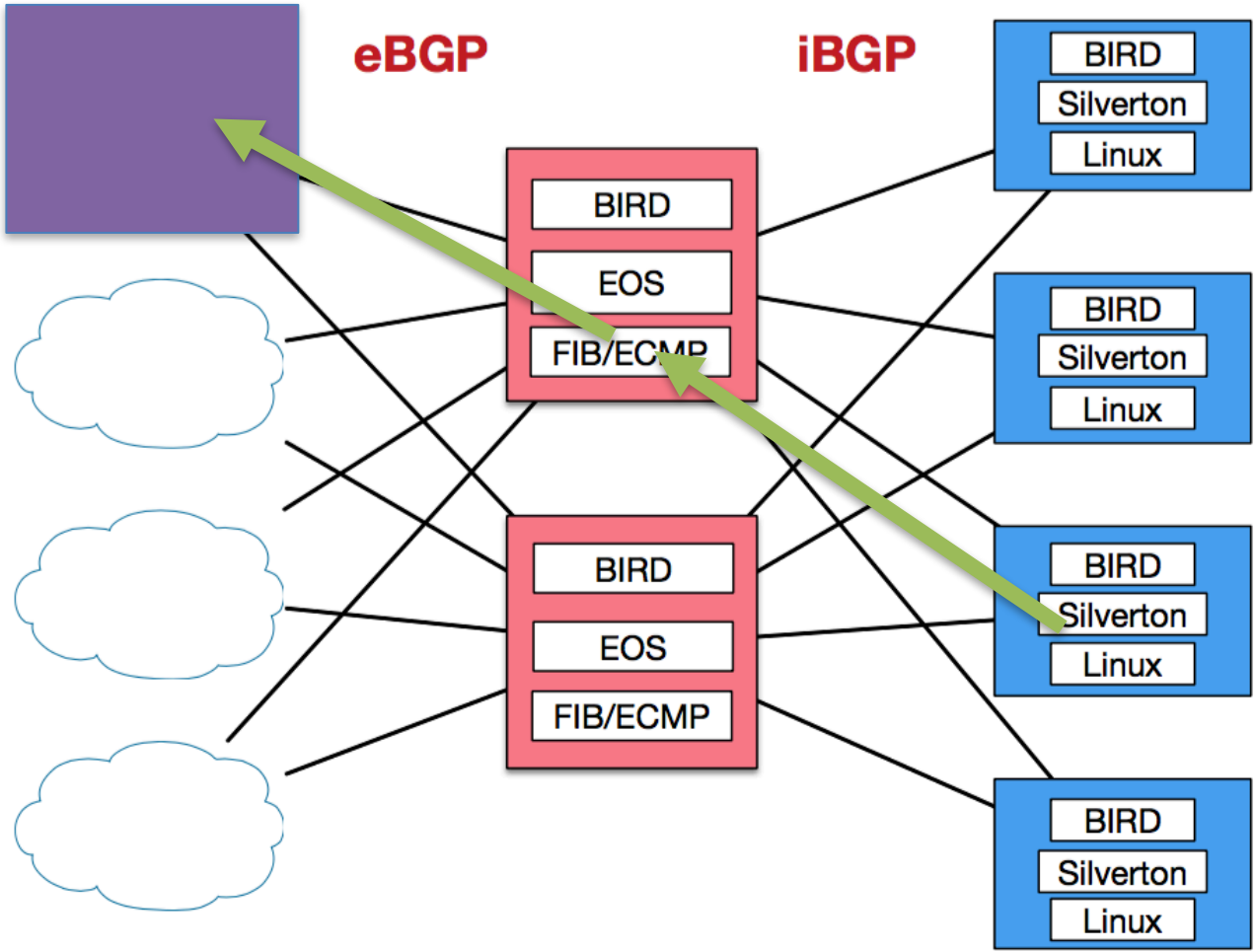

eBGP
iBGP
BIRD
EOS
FIB/ECMP
BIRD
EOS
FIB/ECMP
BIRD
Silverton
Linux
BIRD
Silverton
Linux
BIRD
Silverton
Linux
BIRD
Silverton
Linux


eBGP
iBGP
BIRD
EOS
FIB/ECMP
BIRD
EOS
FIB/ECMP
BIRD
Silverton
Linux
BIRD
Silverton
Linux
BIRD
Silverton
Linux
BIRD
Silverton
Linux


eBGP
iBGP
BIRD
EOS
BIRD
Silverton
Linux
BIRD
BIRD
EOS
FIB/ECMP
BIRD
Silverton
Linux
BIRD
Silverton
Linux
"with L2 egress"

IXP Rules (stolen from LONAP)

- Members must only **respond to ARP requests for addresses which they have been assigned.**
- Members **must only present one source mac-address** on any one physical port they are connected
- Members must ensure that their use of IXP **is not detrimental** to the use made by other Members.


eBGP
iBGP
BIRD
EOS
FIB/ECMP
BIRD
EOS
FIB/ECMP
BIRD
Silverton
Linux
BIRD
Silverton
Linux
BIRD
Silverton
Linux
BIRD
Silverton
Linux

Route Servers means Blowing Up!

- Equinix Singapore:

```
bird> show route protocol ixp_eqix_mlp_1 count
29965 of 60003 routes for 30028 networks
```

- AMS-IX Amsterdam:

```
bird> show route protocol ixp_amsix_mlp_1 count
17872 of 239680 routes for 222257 networks
```

Jan 15 20:36:14 spine-XXX2001 StrataL3: %ROUTING-3-HW_RESOURCE_FULL: Hardware resources are insufficient to program all routes

Switch RIB / FIB Limits

	DRAM	IPv4 FIB	IPv6 FIB
FM6000 (10G)	4GB	70K	18K
SC Trident 2 (10G)	4GB	16K / 144K	8K / 77K
MC Trident 2 (10G)	32GB	16K / 144K	8K / 77K
ARAD (10G/100G)	4GB	64K	12K
Tomahawk (10/100G)	8GB	104K	56K
Jericho (10/25/100G)	8GB+	~1MM+	~1MM+

Our Numerous IXP Problems

- L2 egress:
  - Multiple Source MACs
  - Can't spoof switch SMAC
  - Can leak ARP, broadcast, multicast, etc.
  - Generally bad exchange hygiene
- L3 egress:
  - FIB too small
  - Large Route Servers
  - Large BLP peers with many prefixes
- IPv6
  - Must support IPv4 and IPv6

Enter the Minimum Viable FIB

- Not all prefixes on an MLPE are valuable
- Manage the FIB size as related to "value" of prefixes:
  - High value ASNs and prefixes
  - Geographically scoped to point of interconnect
  - Valued through traffic utilization
  - Valued through performance
- Backed up via transit default routes.

Fastly BigPeer:

- Internal Peering Database
  - Preference BLP over MPLE routes
- sFlow:
  - Find high volume prefixes
- IRR (bgpq3):
  - To carefully watch for max-prefix / prefix origin
- BGP:
  - Collect routes and feedback to dampen churn in selection

BigPeer Outputs:

```
[tjd@switch-sin6901 configs]$ head -n 15 bird-filters.conf
define toptalkers = [
1.0.128.0/19, # MLP+
1.0.128.0/24, # MLP+
1.0.129.0/24, # MLP+
1.0.131.0/24, # MLP+
1.0.132.0/22, # MLP+
1.0.136.0/24, # MLP+
```

```
[tjd@edge-ams4011 configs]$ head -n 15 bird-filters.conf

define toptalkers = [
1.0.176.0/20, # MLP+
1.0.200.0/22, # MLP+
1.10.128.0/18, # PEER-
1.10.164.0/24, # PEER+
1.10.192.0/19, # PEER+
1.10.213.0/24, # MLP+
```

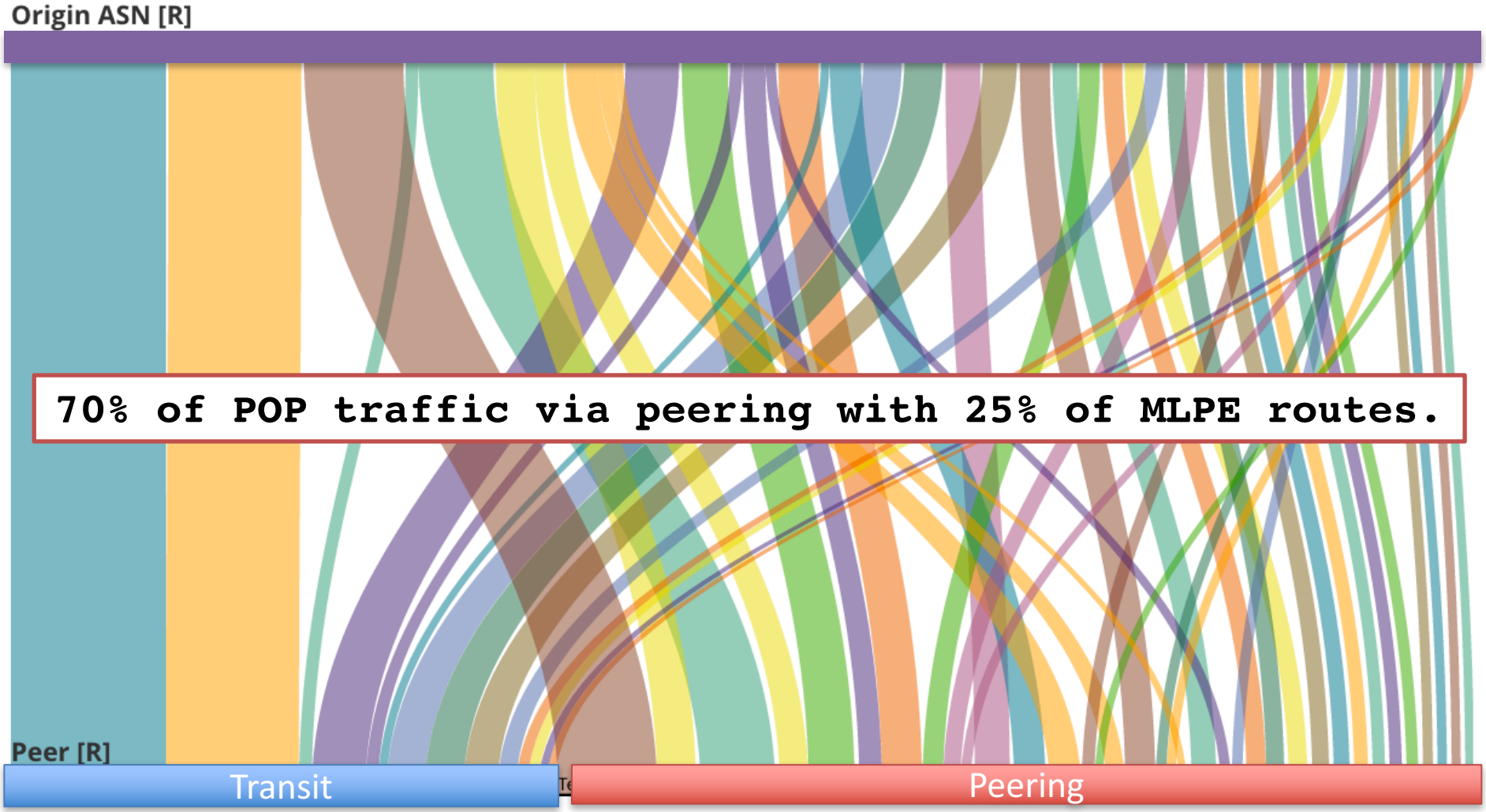
Final FIB Switch State

```
switch-sin6901#show platform
fm6000 l3 summary

IPv4 unprogrammed/parent drop
routes: 0/0
IPv6 unprogrammed/parent drop
routes: 0/0
Lpm routes in Bst: 25492
Total unicast routes in Bst:
25897/61380
Total Hw routes in Vrf default:
25966
```

```
edge-ams4011#show platform arad
ip route summary

Total number of VRFs: 1
Total number of routes: 37706
Total number of route-paths:
22619
Total number of lem-routes:
15095
Total number of /24 routes in
lem: 13884
Total number of /32 routes in
lem: 1211
```


Origin ASN [R]
70% of POP traffic via peering with 25% of MLPE routes.
Peer [R]
Transit
Peering

Variant Work - SlashPeer

- SlashPeer manages PNI bandwidth constraints:
  - Set desired utilization rate for interface
  - Use sFlow data at prior peaks to extract valuable prefixes
  - Generate prefix filers against BLP session
  - Win! No congestion on PNIs

Future Considerations

- Jericho / Jericho+ Hardware
  - Claims up to 1/2MM+ FIB
  - Does this obsolete these tools?
- No:
  - Existing install base
  - BGP add-path for fast reconvergence / multipath
  - IPv4 / IPv6 growth room
  - SlashPeer keeps PNIs running warm, not hot.

Questions?

http://www.fastly.com/peering

http://as54113.peeringdb.com

Tom Daly
VP, Infrastructure
tjd@fastly.com