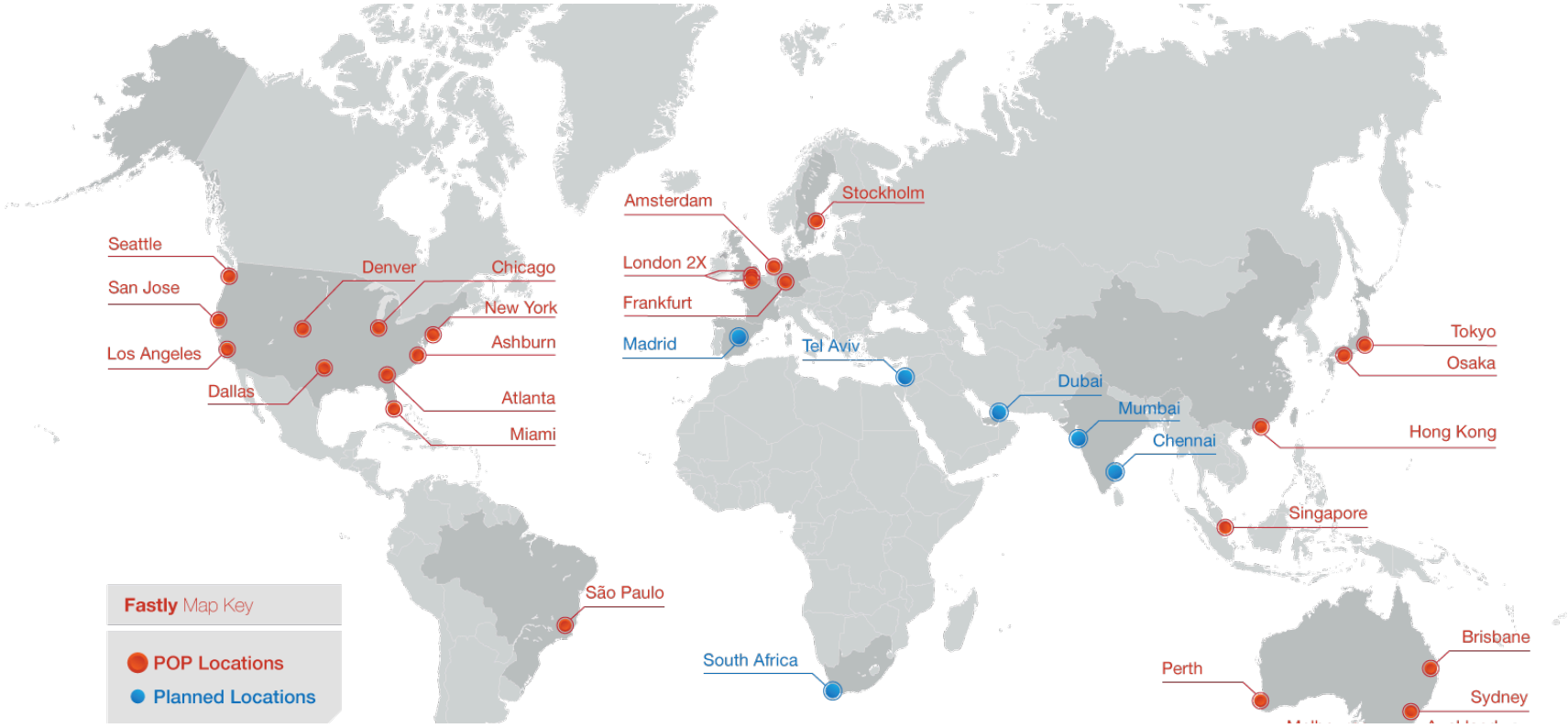
Minimum Viable FIB

Tom Daly
VP, Infrastructure
tjd@fastly.com

Outline

- Evolution of Fastly Egress Routing
- Conditions of Peering at IXPs
- Derivation of a Minimum Viable FIB
- Results

I work for a CDN



Constraints

- "Big" routers: NOT OK
- "Big" load balancers: NOT OK
- Linux Everywhere: OK
- Cache Servers: OK
- Top of Rack Switches: OK
- Stupid DNS Tricks: OK
- Stupid BGP Tricks: Also OK.

Switch Hardware 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

Forwarding Information Base (FIB): Prefixes / routes programmed directly into a routers forwarding plane (ASICs)

Egress Routing Generation #1

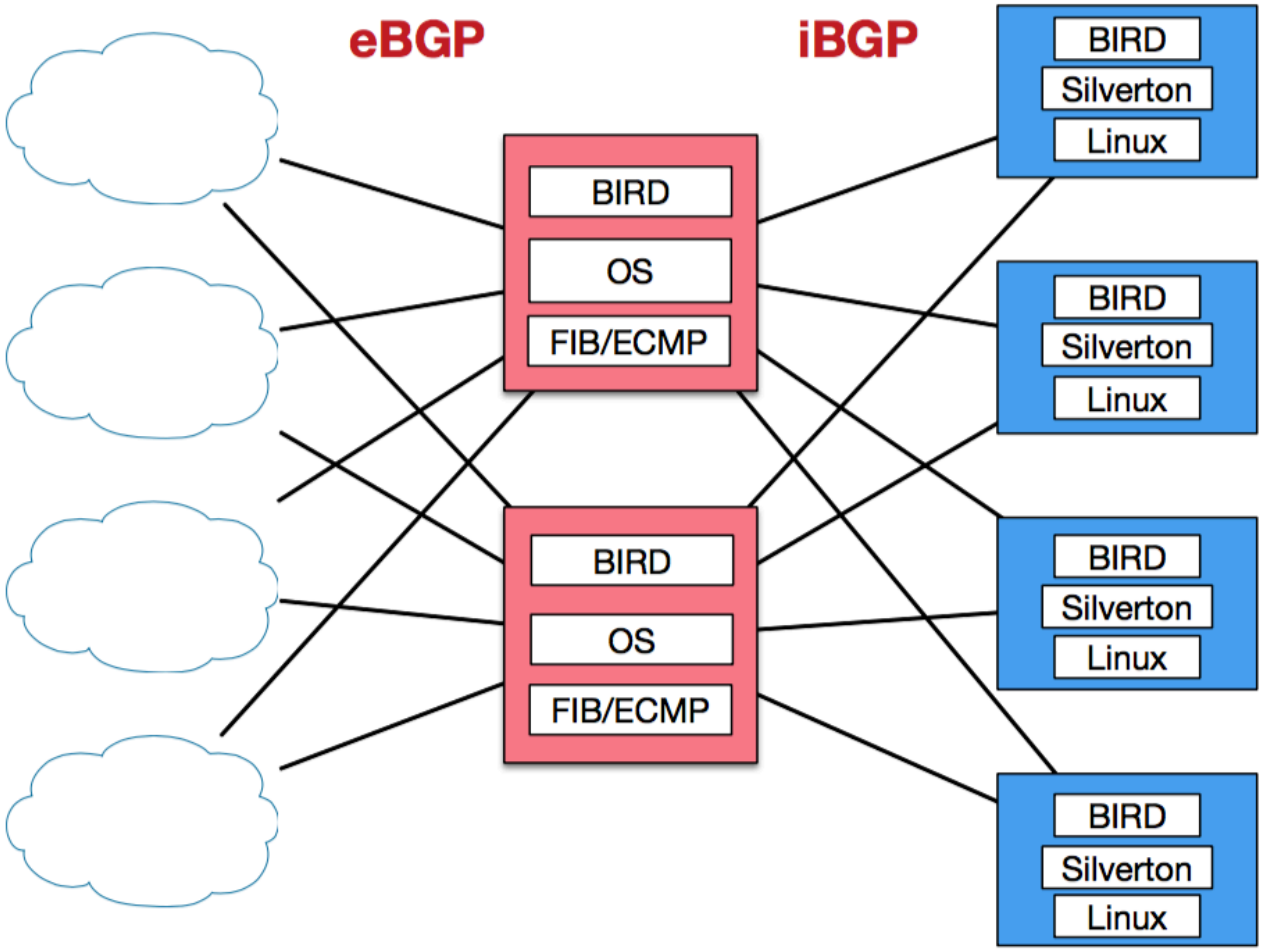
- Caches directly connected to ISPs:
  - doesn't scale
  - not cost effective
  - expensive to make redundant.

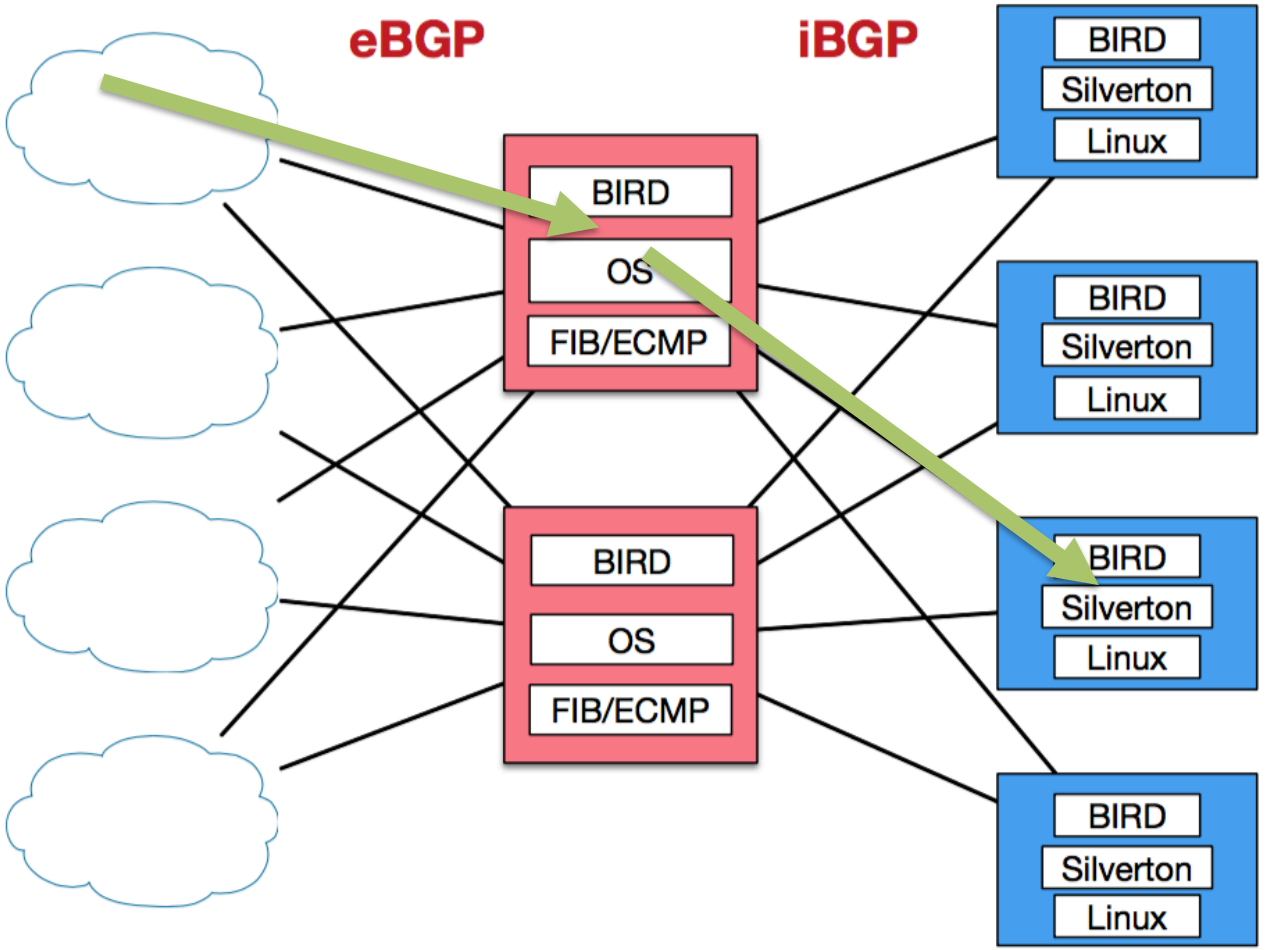
Egress Routing Generations #2-4

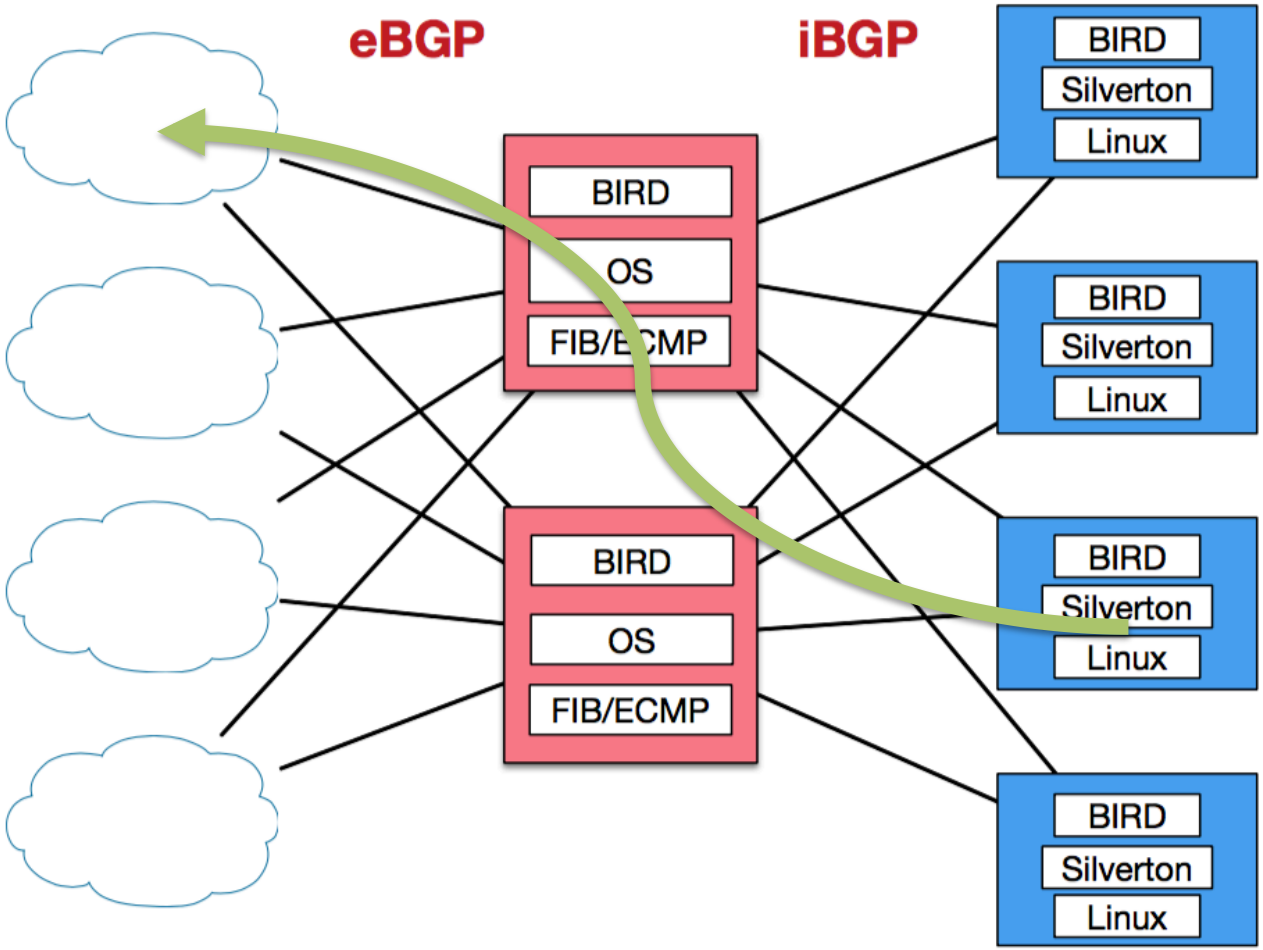
- ToR switches doing dumb things:
  - Virtual Router Redundancy Protocol (VRRP)
  - KeepaliveD
  - BGP
  - Constant east/west congestion problems
  - Ugh. Thrash.

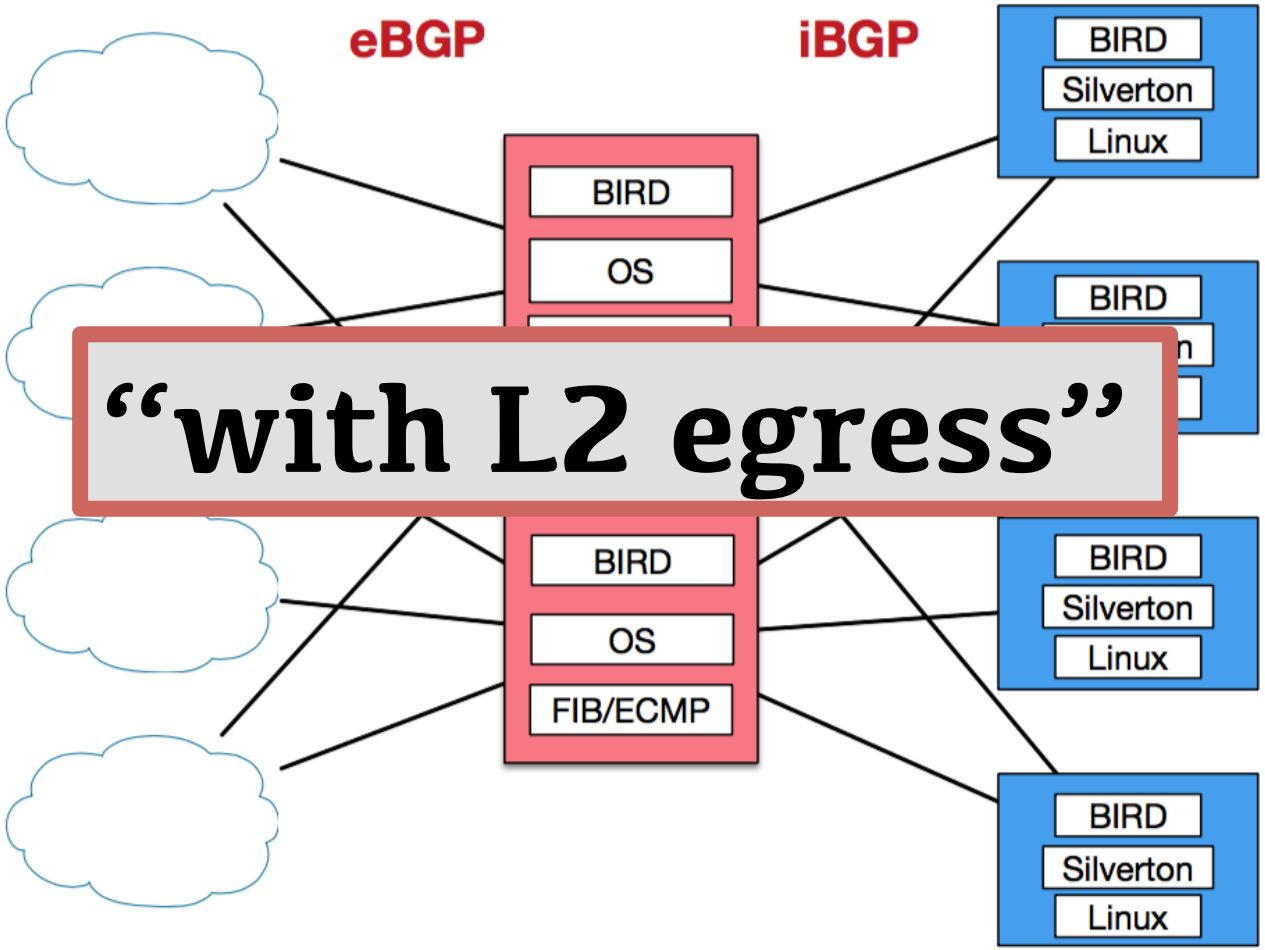
Egress Routing Gen #5:

- Dumb switches + smart caches = smart things.
  - Multipath everywhere **with L2 egress.**
  - Silverton manages egress relationships
    - Next-hop and MAC address 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.

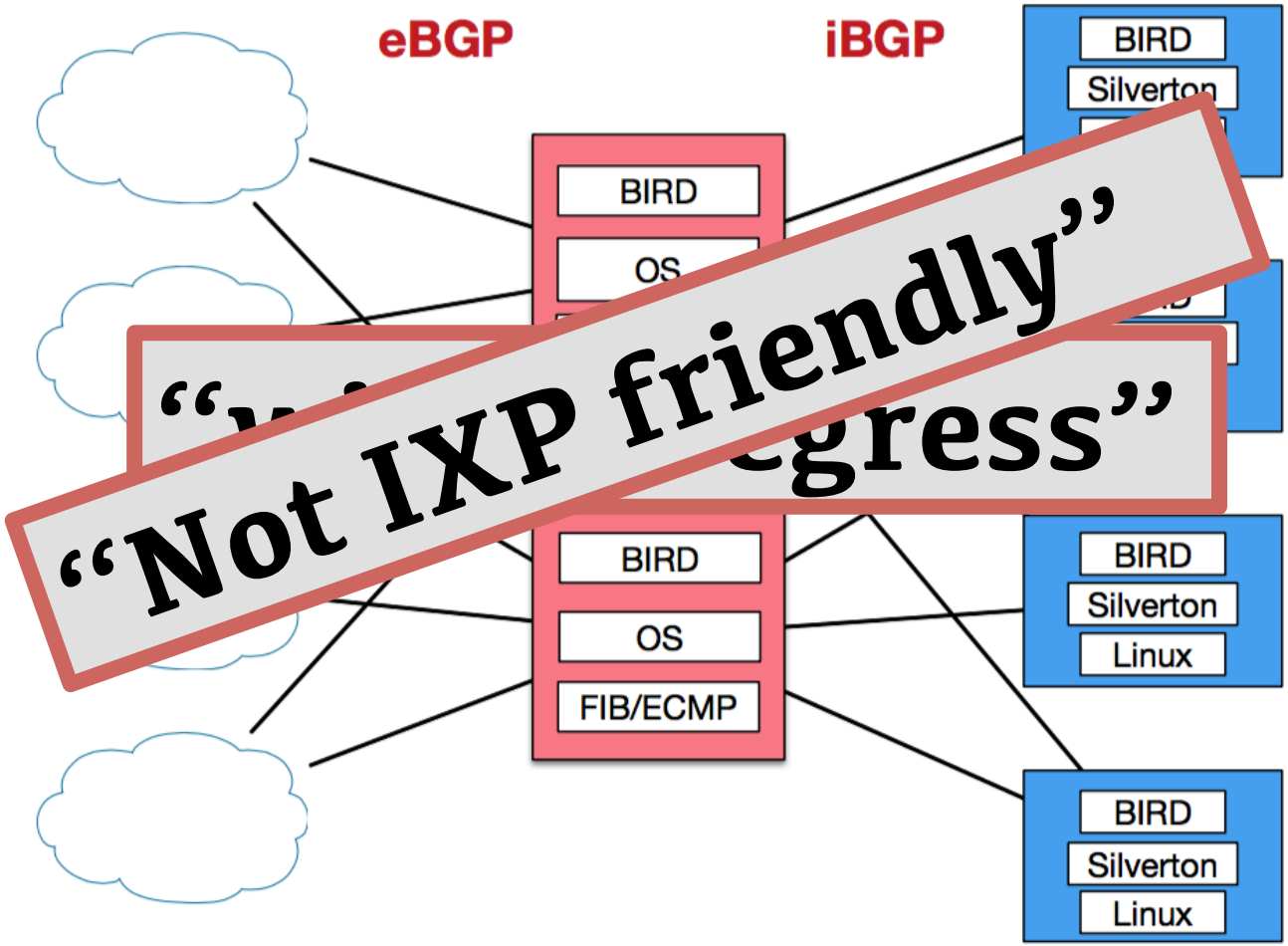

eBGP
iBGP
BIRD
OS
FIB/ECMP
BIRD
OS
FIB/ECMP
BIRD
Silverton
Linux
BIRD
Silverton
Linux
BIRD
Silverton
Linux
BIRD
Silverton
Linux


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


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



“with L2 egress”


eBGP
iBGP
BIRD
Silverton
BIRD
OS
FIB/ECMP
BIRD
Silverton
Linux
BIRD
Silverton
Linux
"Not IXP friendly"

IXP Rules (stolen from LONAP)

- Members must not proxy-arp on any peering VLANs.
- 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
- The public peering VLANs and permitted ethertypes:
  - 0x0800 (IPv4) - must have a unicast destination mac
  - 0x0806 (ARP) - must only reply to ARP requests for their assigned IPs on the peering lan
  - 0x86dd (IPv6)
- Members must ensure that their use of IXP is not detrimental to the use made by other Members.

Problems

- L2 egress:
  - Multiple Source MACs
  - Can't spoof switch SMAC
  - Can leak ARP, broadcast, multicast, etc (yes we outbound ACL this)
  - Generally bad exchange hygiene
- L3 egress:
  - FIB too small
  - EU MPLE???
  - Large BLP peers with many prefixes
- IPv6
  - Must support IPv4 and IPv6

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

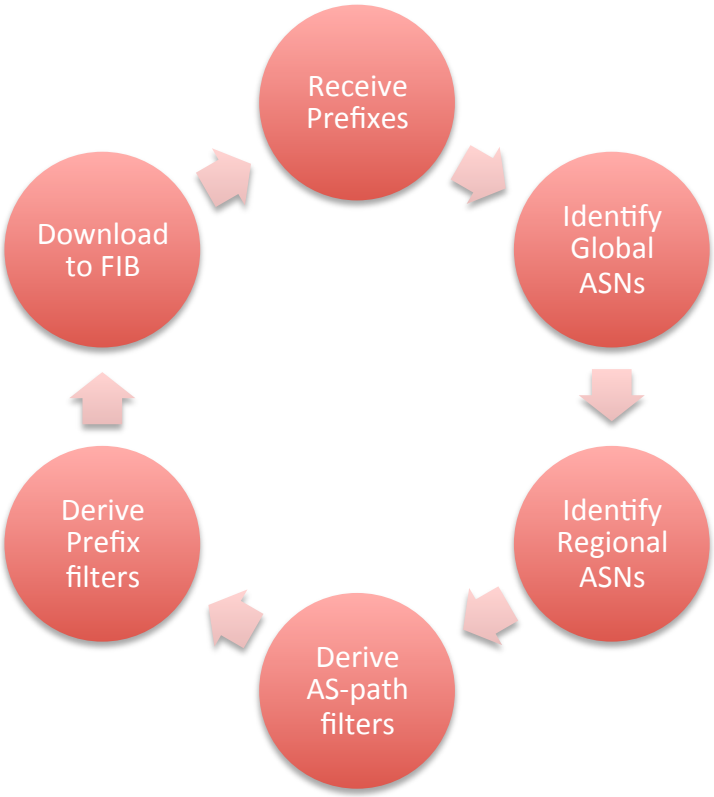
Workarounds

- OpenFlow / DirectFlow:
  - Rewrite SMAC/DMAC as “comply”
- Multiple VRFs:
  - Each VRF with a “default route” towards a peer
  - Needs ingress policy based routing
- Use Overlay Network:
  - IPv4 DSCP / MPLS / VXLAN / NVGRE
  - Make routing decision at cache; apply “label”; pop label at switch and forward.
- We decided to stay simple: manage the FIB

Deriving the Minimum Viable FIB

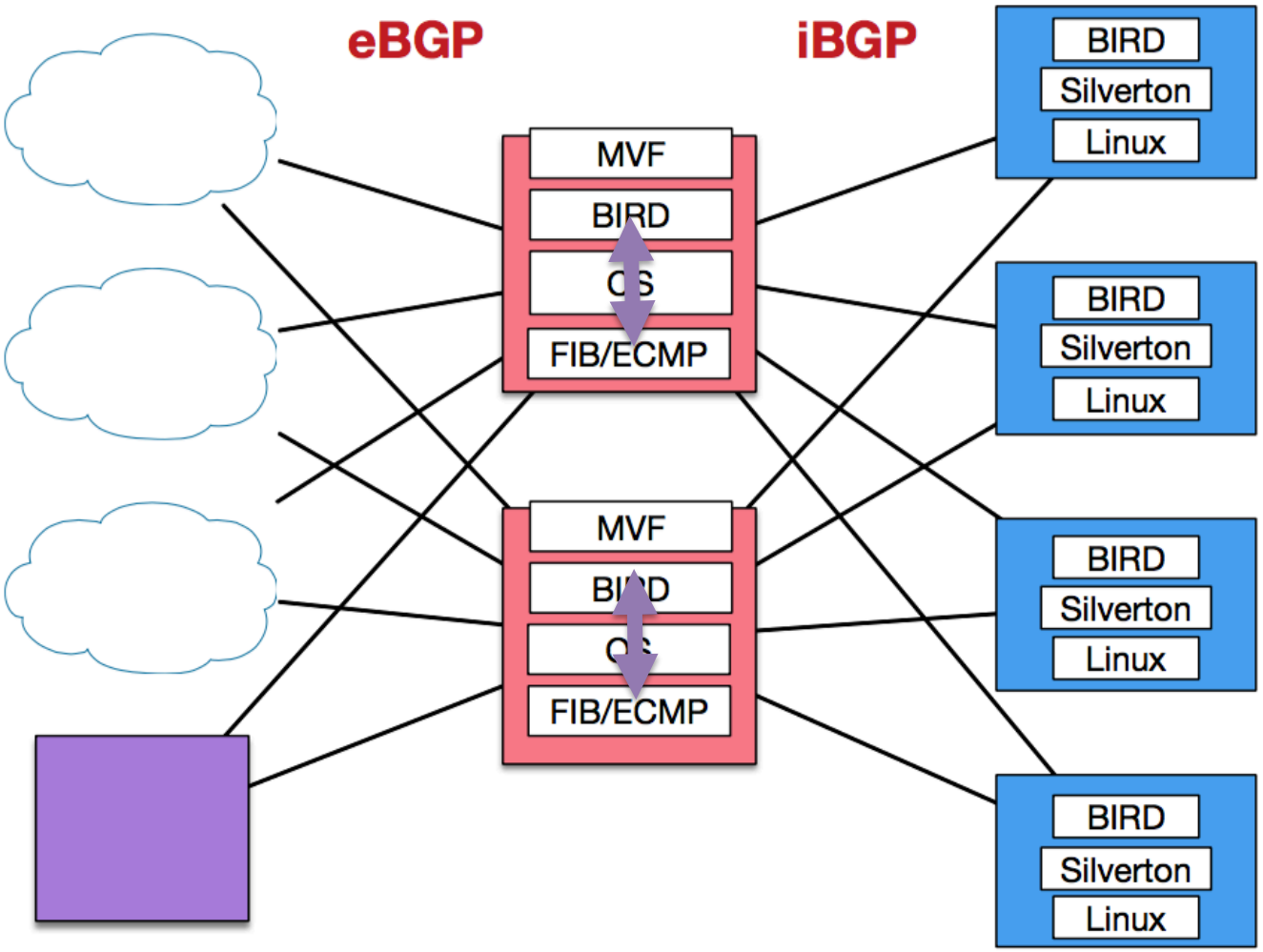
- High value ASNs and prefixes
  - Scoped to point of interconnect
    - Shaped through DNS and BGP anycast
  - Valued through traffic utilization
  - Valued through performance
- Backed up via transit, always.
- Therefore: Selective Peering

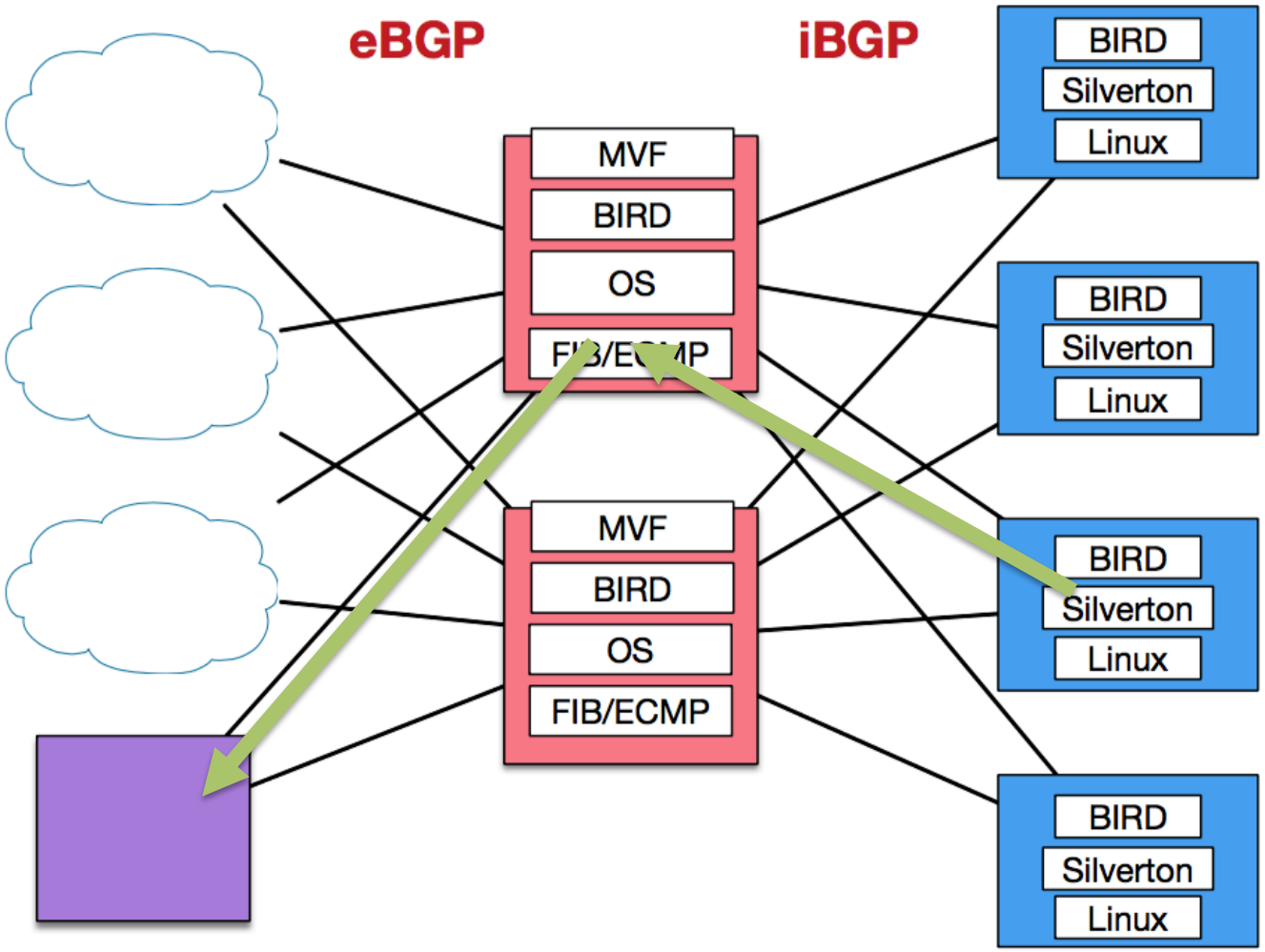
Minimum Viable "Peering" FIB: Process

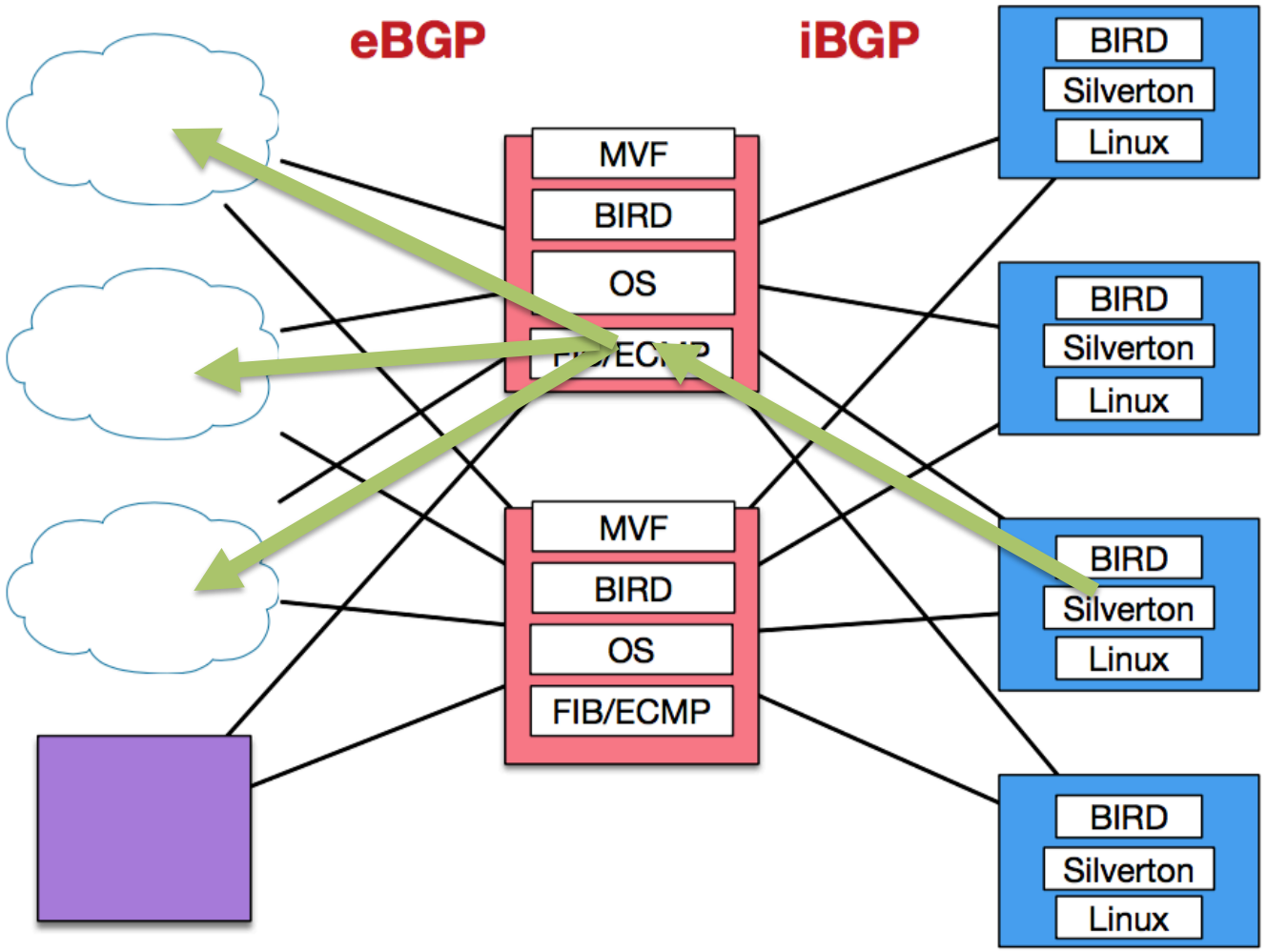


Automated Selection Criteria

- Internal Peering Database
  - Preference BLP over MPLE routes
- sFlow:
  - Find high volume prefixes (see “sir” from David Barroso)
- IRR:
  - To carefully watch for max-prefix / prefix origin
- BGP:
  - Collect routes; dampen churn in selection


eBGP
iBGP
MVF
BIRD
OS
FIB/ECMP
MVF
BIRD
OS
FIB/ECMP
BIRD
Silverton
Linux
BIRD
Silverton
Linux
BIRD
Silverton
Linux
BIRD
Silverton
Linux


eBGP
iBGP
MVF
BIRD
OS
FIB/ECMP
MVF
BIRD
OS
FIB/ECMP
BIRD
Silverton
Linux
BIRD
Silverton
Linux
BIRD
Silverton
Linux
BIRD
Silverton
Linux


eBGP
iBGP
MVF
BIRD
OS
FIB/ECMP
MVF
BIRD
OS
FIB/ECMP
BIRD
Silverton
Linux
BIRD
Silverton
Linux
BIRD
Silverton
Linux
BIRD
Silverton
Linux

Global Peer

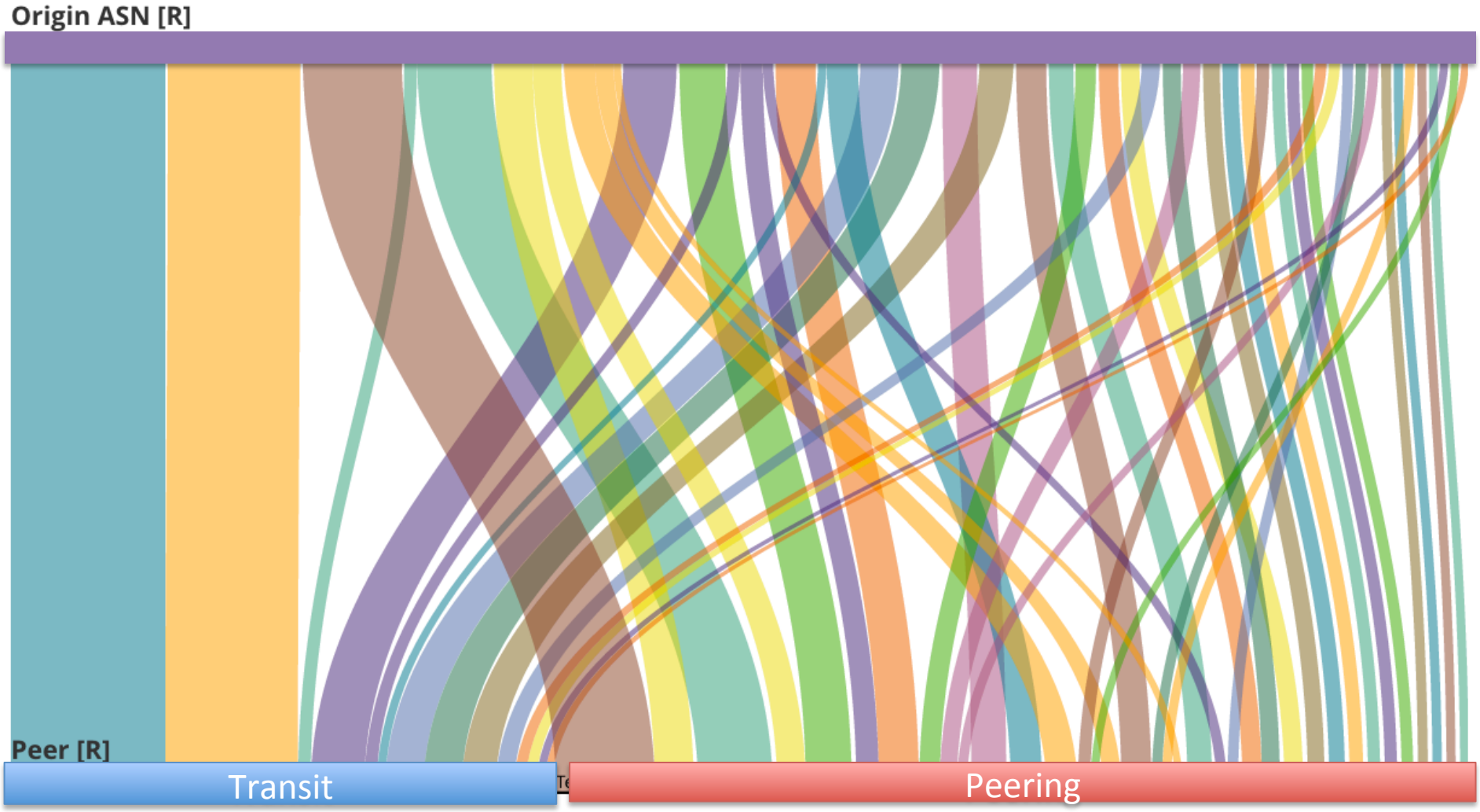
Results

From MPLE:

```
bird> show protocols all ixp_mlp_1
Routes: 46001 imported, 1685 filtered, 22 exported
```

Injected to FIB:

```
bird> show protocols all ixp_injector
Routes: 0 imported, 43347 exported
```


Origin ASN [R]
Peer [R]
Transit
Peering

Questions?

http://www.fastly.com/peering

http://as54113.peeringdb.com

Tom Daly
VP, Infrastructure
tjd@fastly.com

Key to Peering Policy:

- https://www.fastly.com/peering
- “Scope our routing announcements according to the point of interconnect; providing access to local market POP specific prefixes as well as selected regional and/or global anycast prefixes.”
- “Do not have a backbone, and therefore do not announce a consistent set of prefixes across IXPs. Peers will not receive a consistent routing table from Fastly across multiple points of interconnection.”