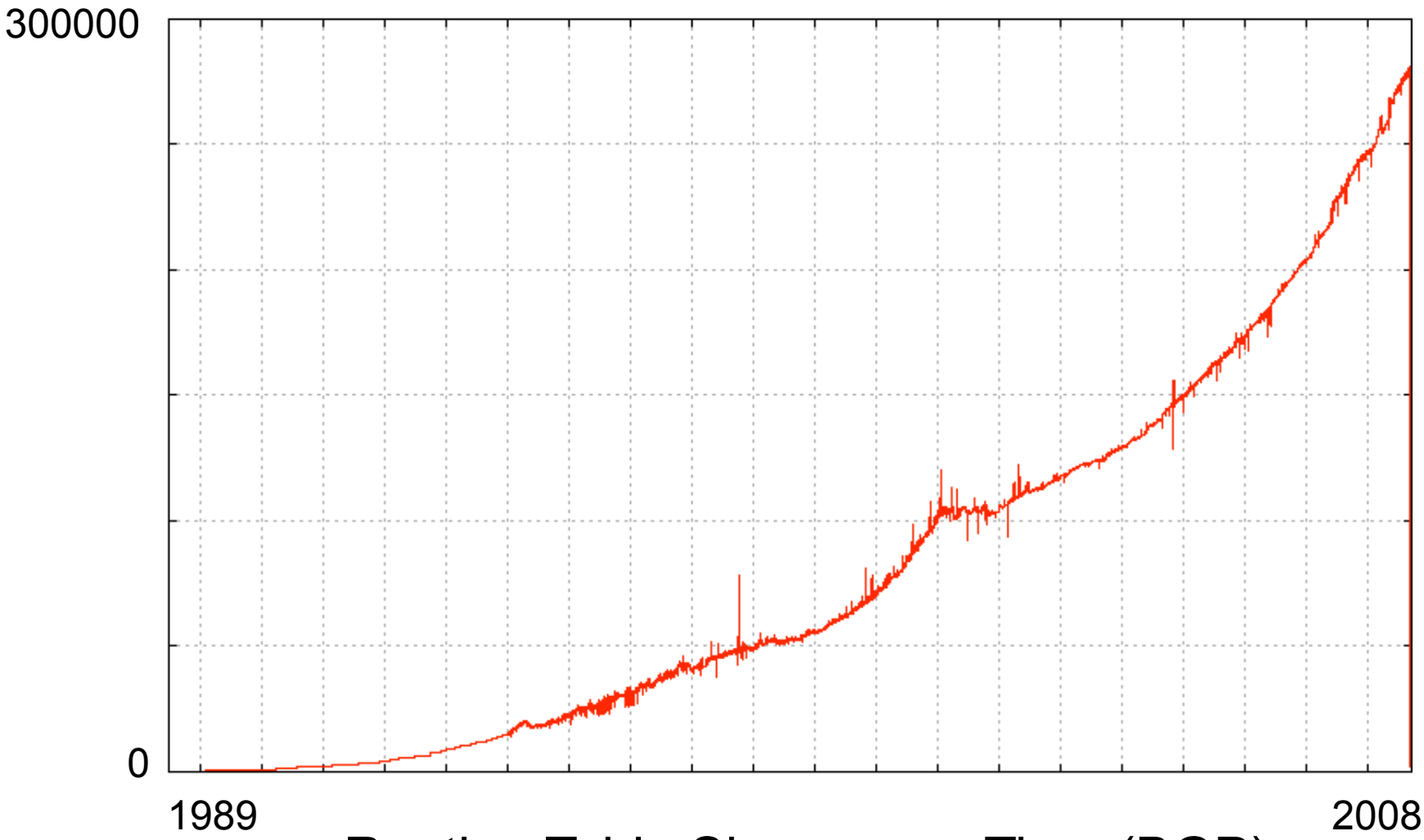
A DIRTY-SLATE APPROACH to Scaling the Internet

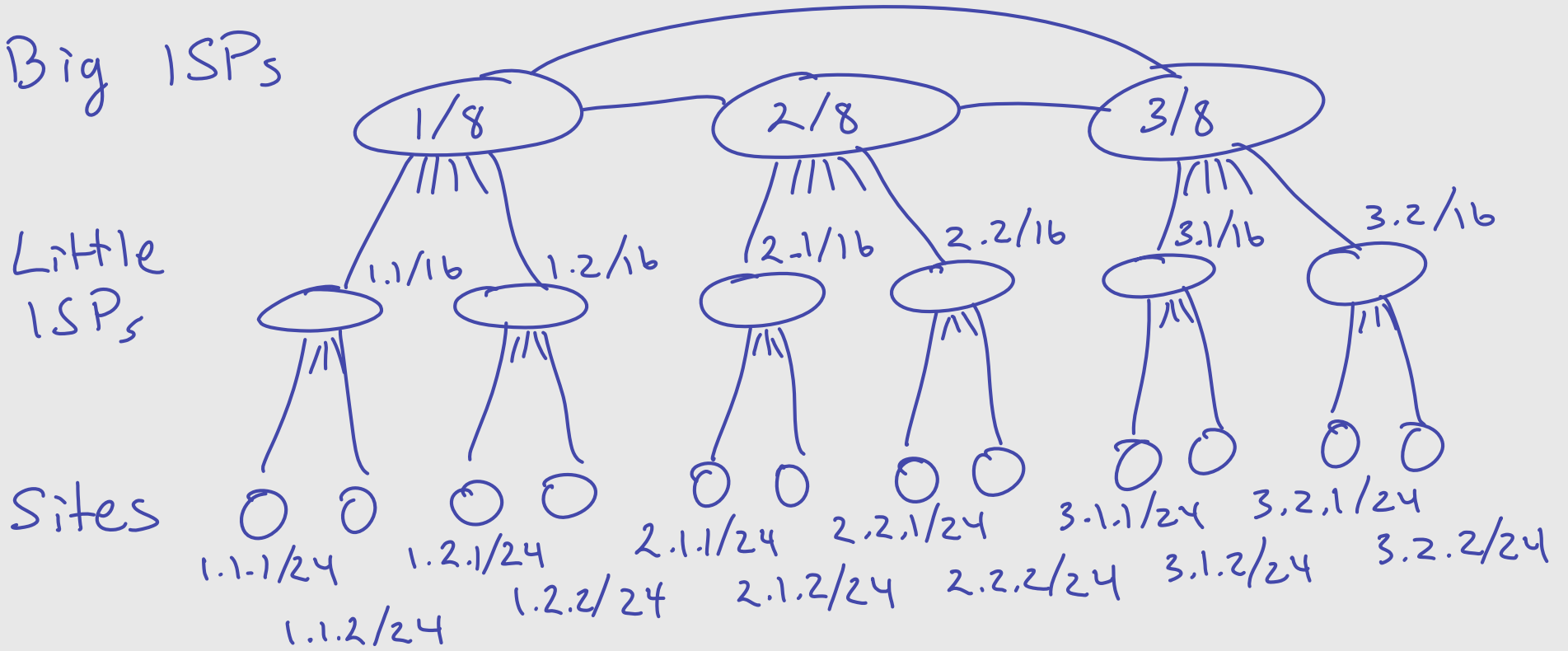
Paul Francis
Hitesh Ballani, Tuan Cao
Cornell



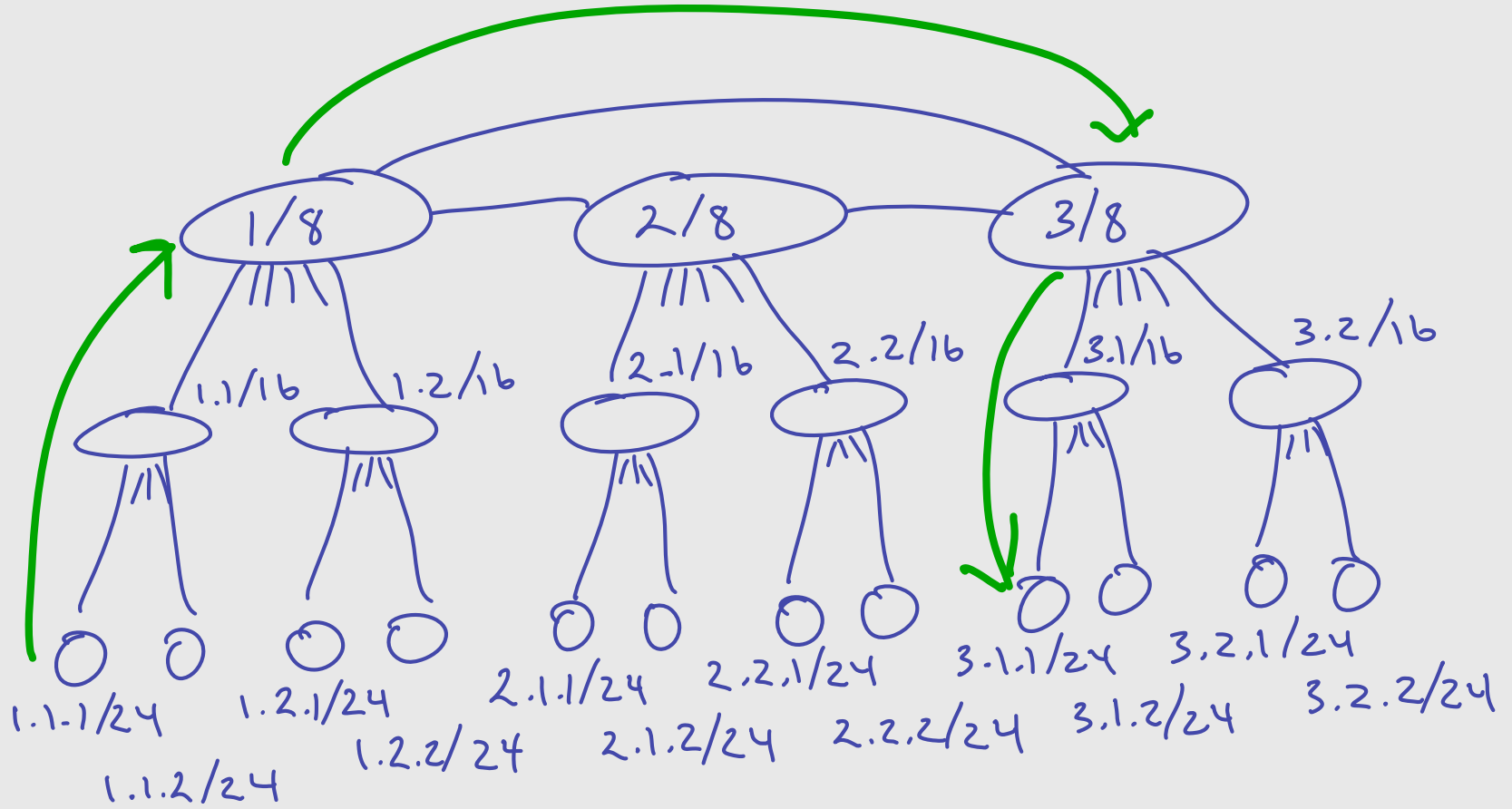
Routing Table Size versus Time (BGP)

BGP data from AS65000, http://bgp.potaroo.net/as2.0/bgp-active.html

Internet is ***addressed like*** a forest of Autonomous Systems (AS)



Routing: Up, Across, Down

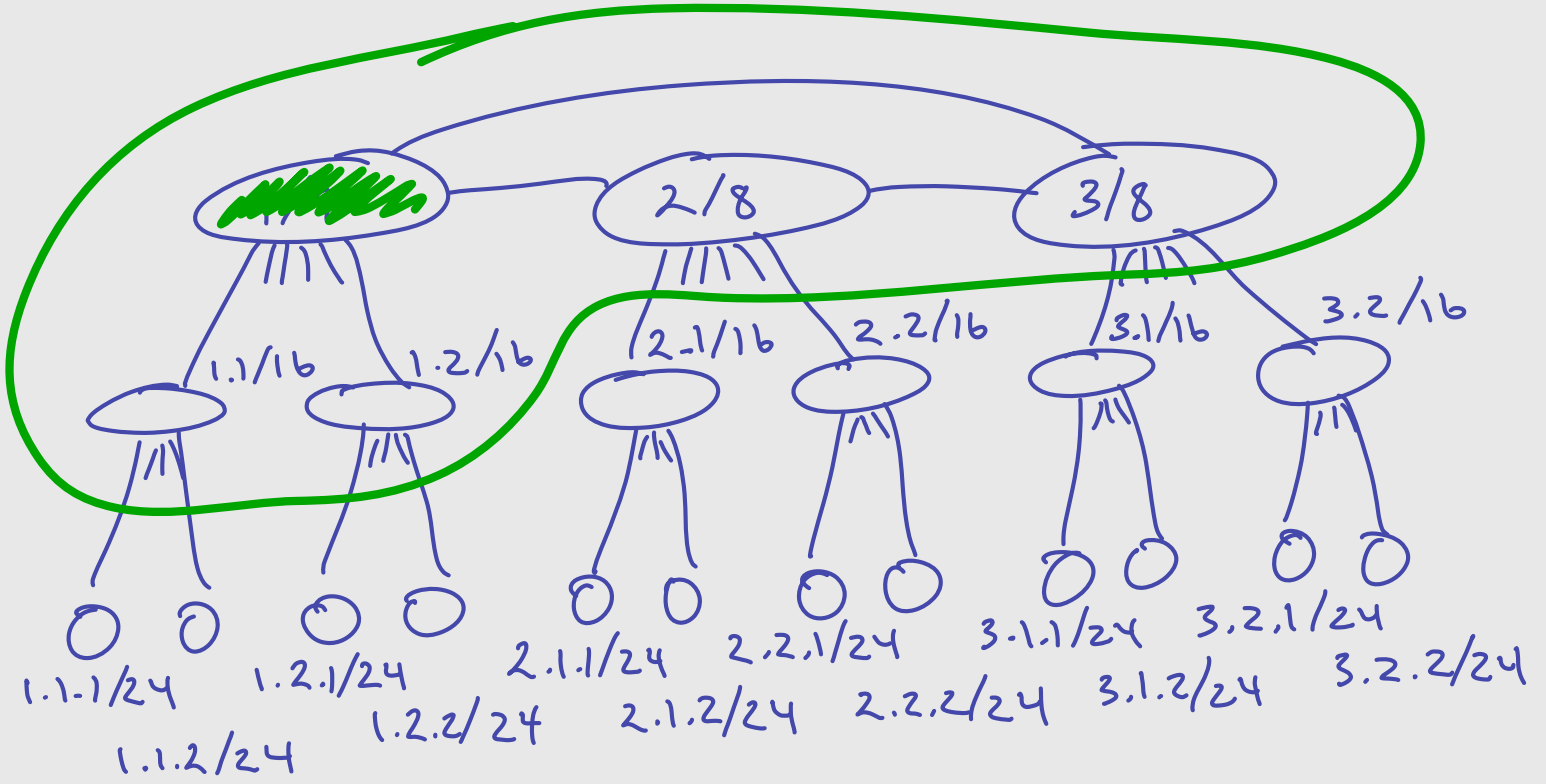


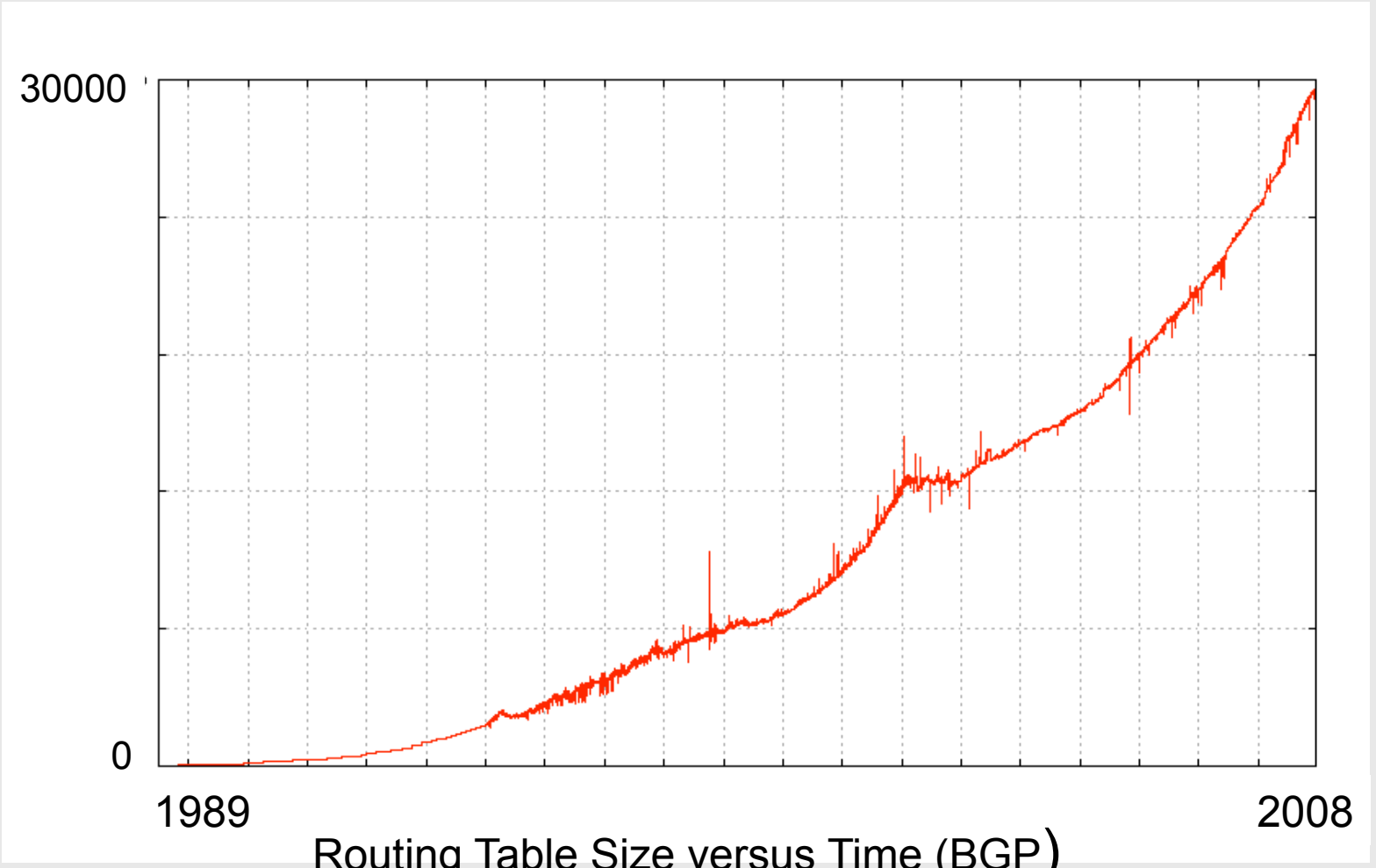
Should scale by:

Number of top-level AS's,

and

Size of Fan-out

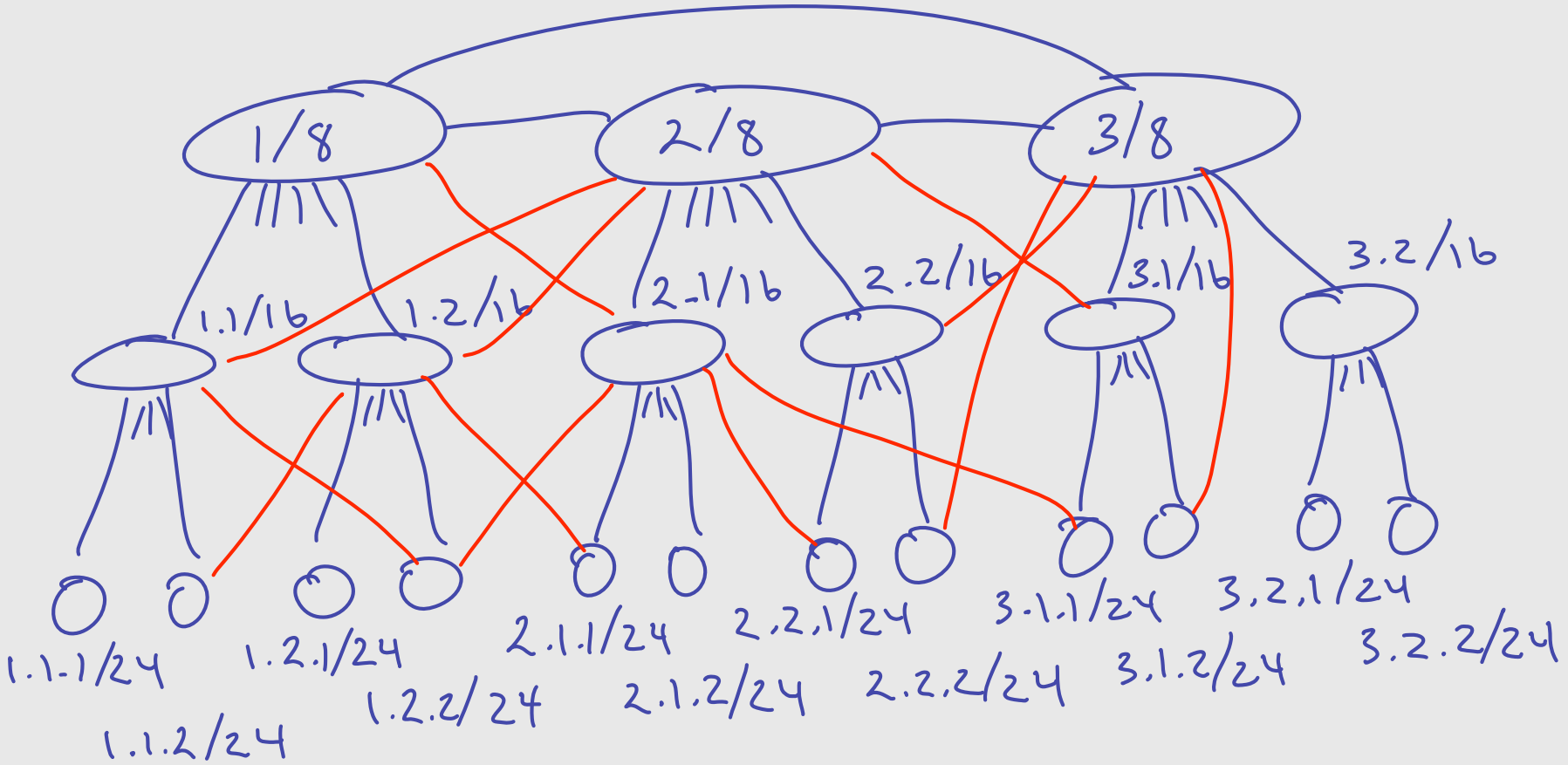




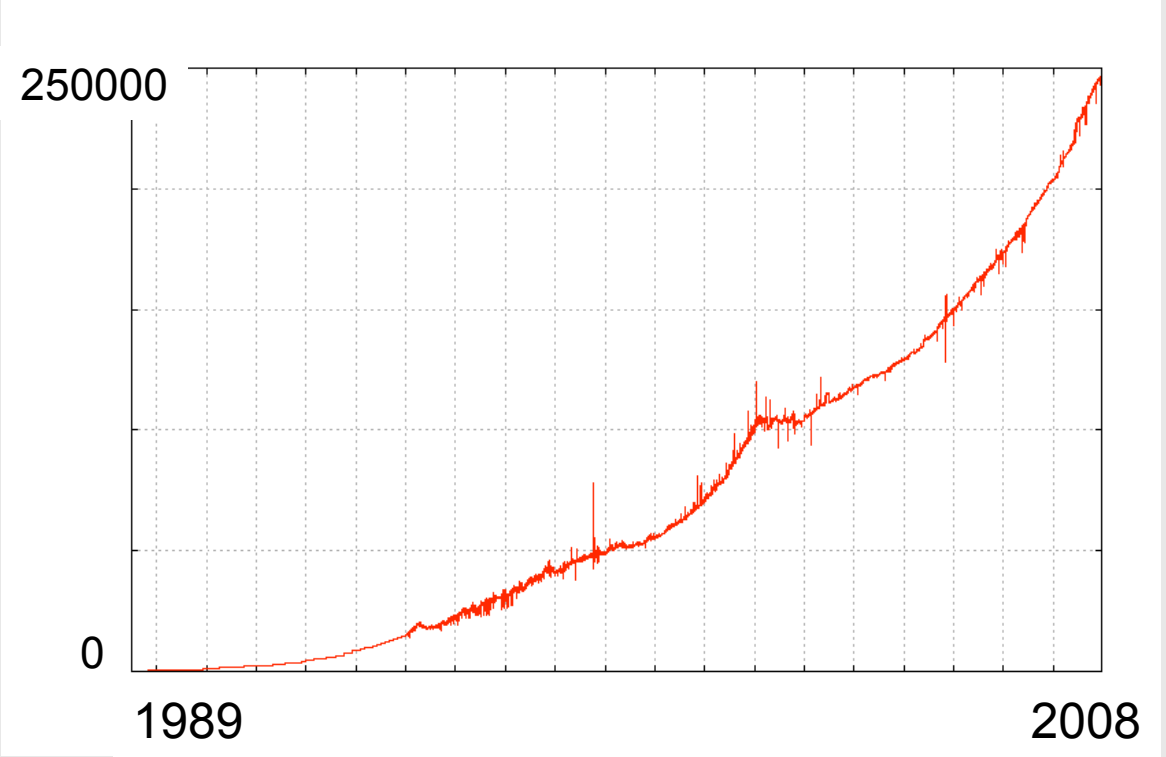
Routing Table Size versus Time (BGP)

What went wrong???

Multi-homing (sites and ISPs)



Address ↔ Topology Mismatch

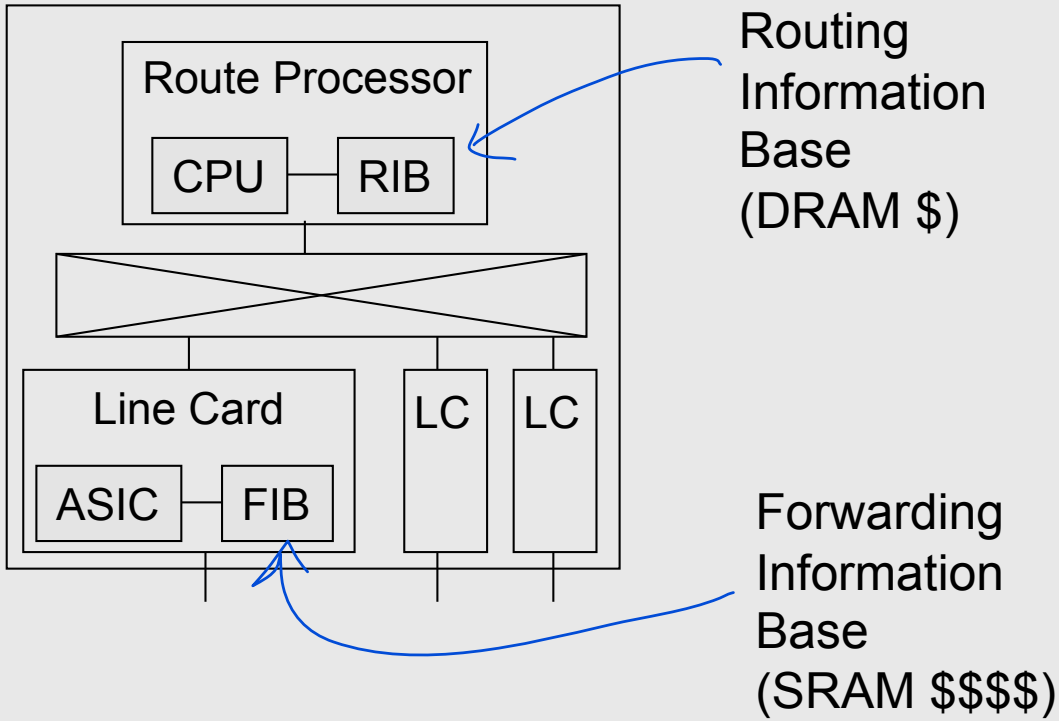


Large routing table managed using:

- Brute Force
- Engineering Constraints

Brute Force:

Large physical memory (FIB)

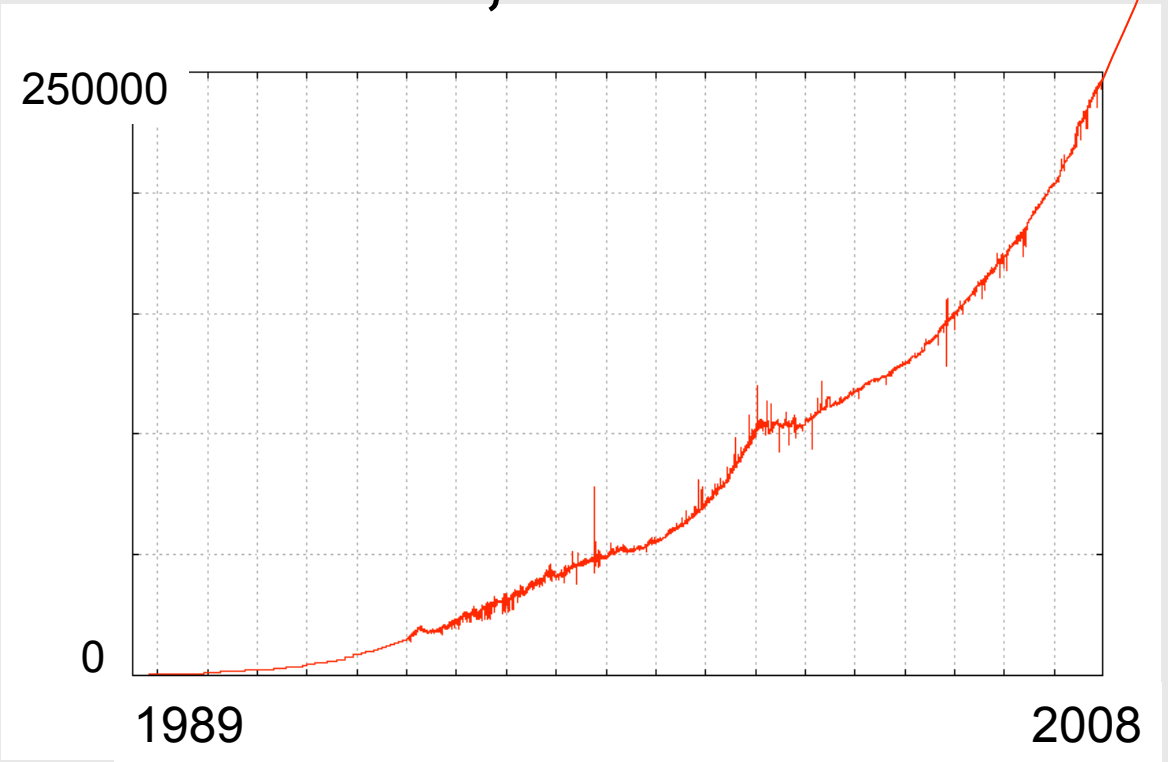


Engineering Constraints

- Prefix size (e.g. /24)
- Frequency/Delay of BGP updates
- Route flap hold-down

Growth rate may increase!

IPv4 address exhaustion, leading to increased address fragmentation



Uptake of IPv6

OUTLINE

An Overview of Previous Approaches

Geographic Addressing

Indirection and Tunneling Schemes

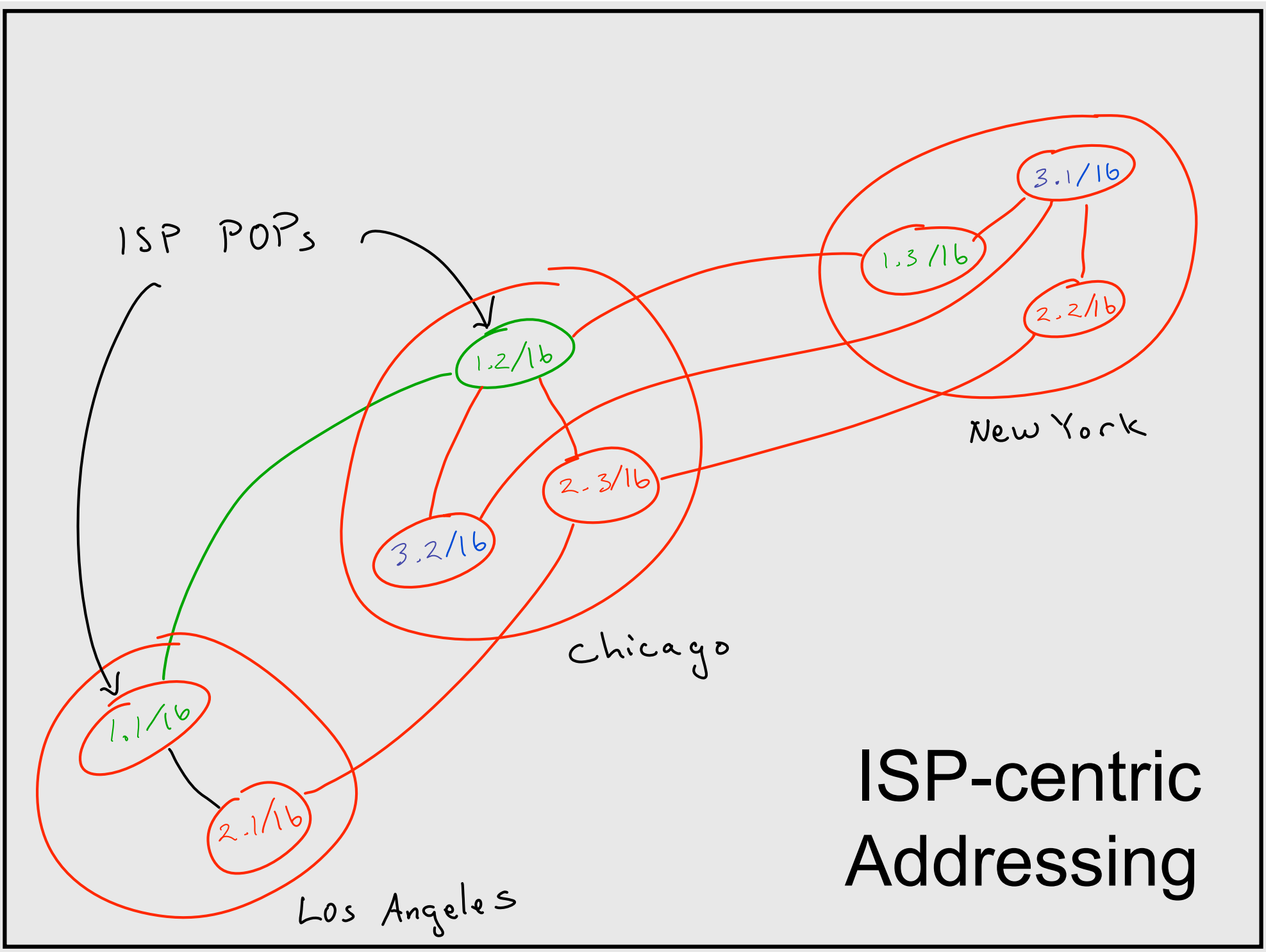
Virtual Aggregation

Current IP addressing is ISP-centric

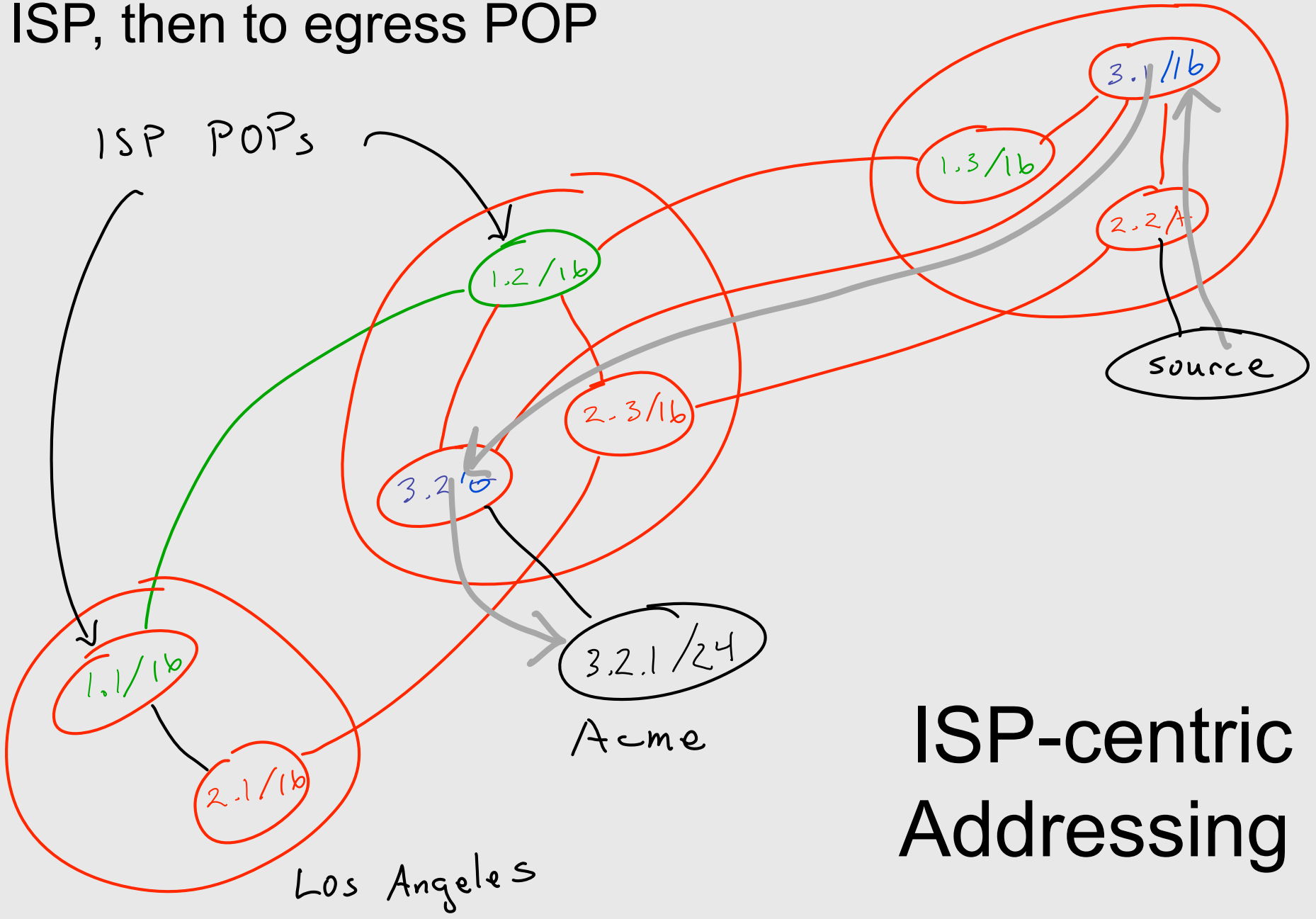
What about geographical addressing?

ISP POPs tend to cluster around metro areas

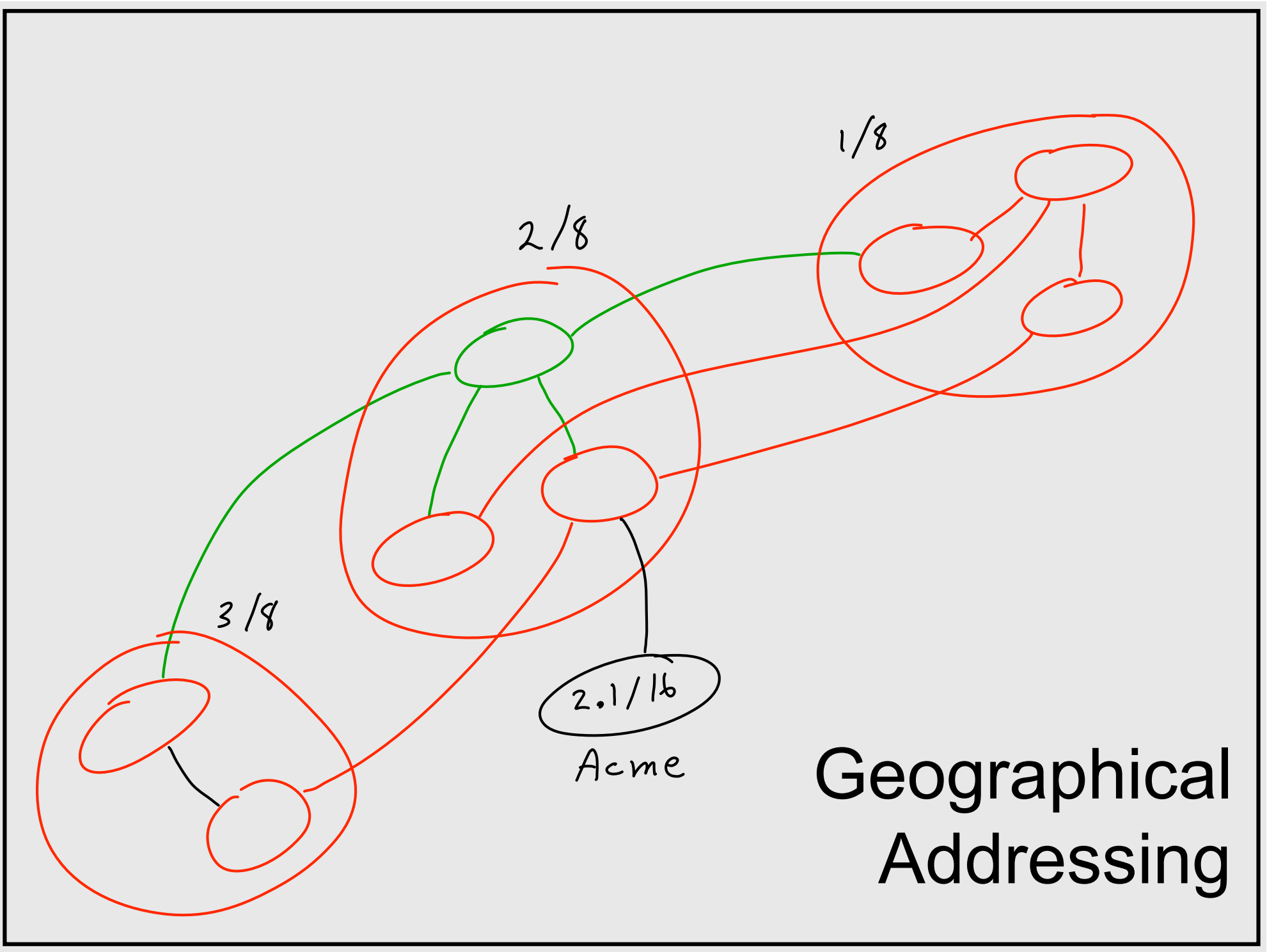
- Multihomed site will almost always connect to ISPs in a given metro area


ISP POPs
1.2/16
3.1/16
1.3/16
2.2/16
New York
2.3/16
3.2/16
Chicago
1.1/16
2.1/16
Los Angeles
ISP-centric
Addressing

Routing: First get packet to ISP, then to egress POP

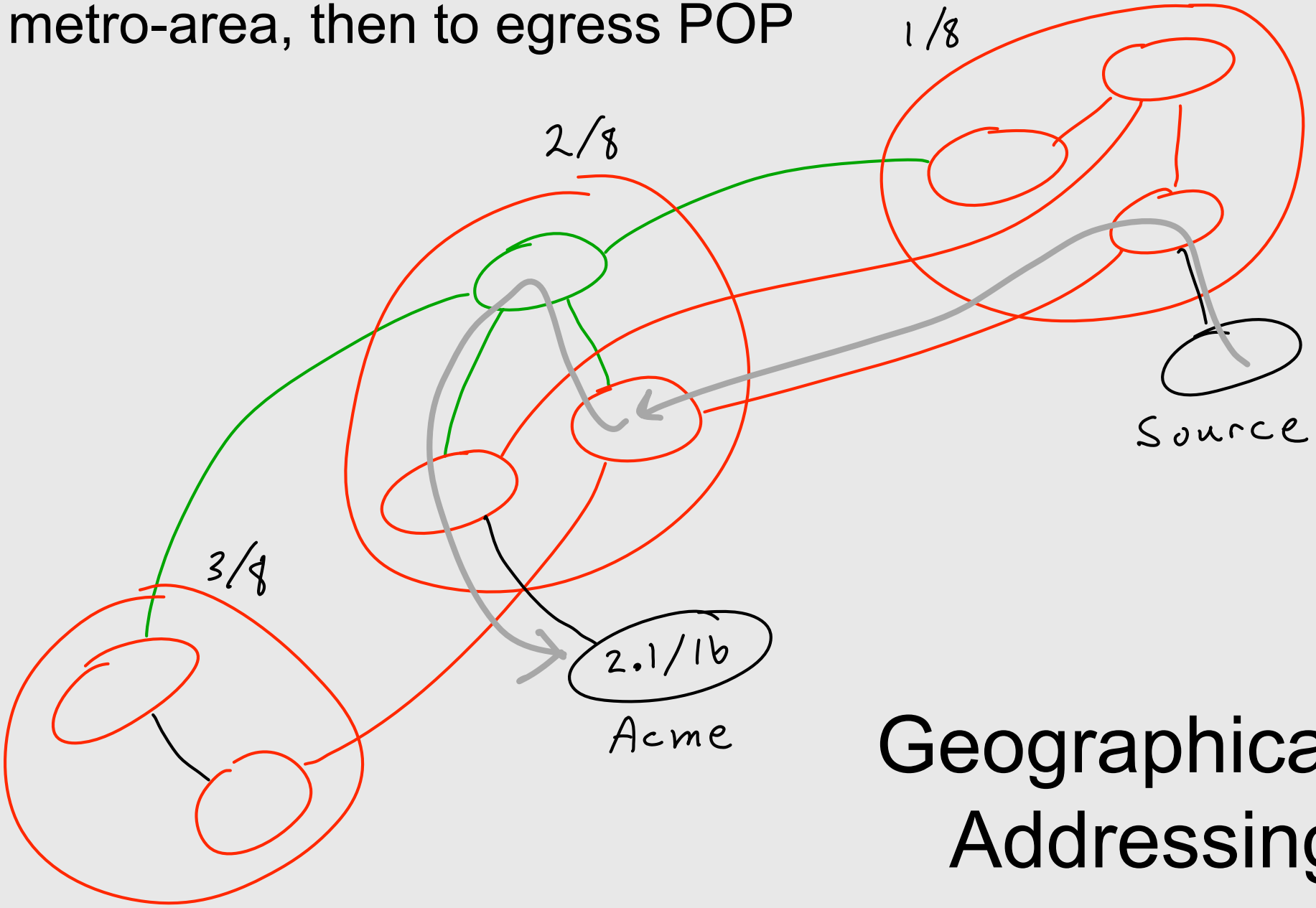


ISP-centric Addressing



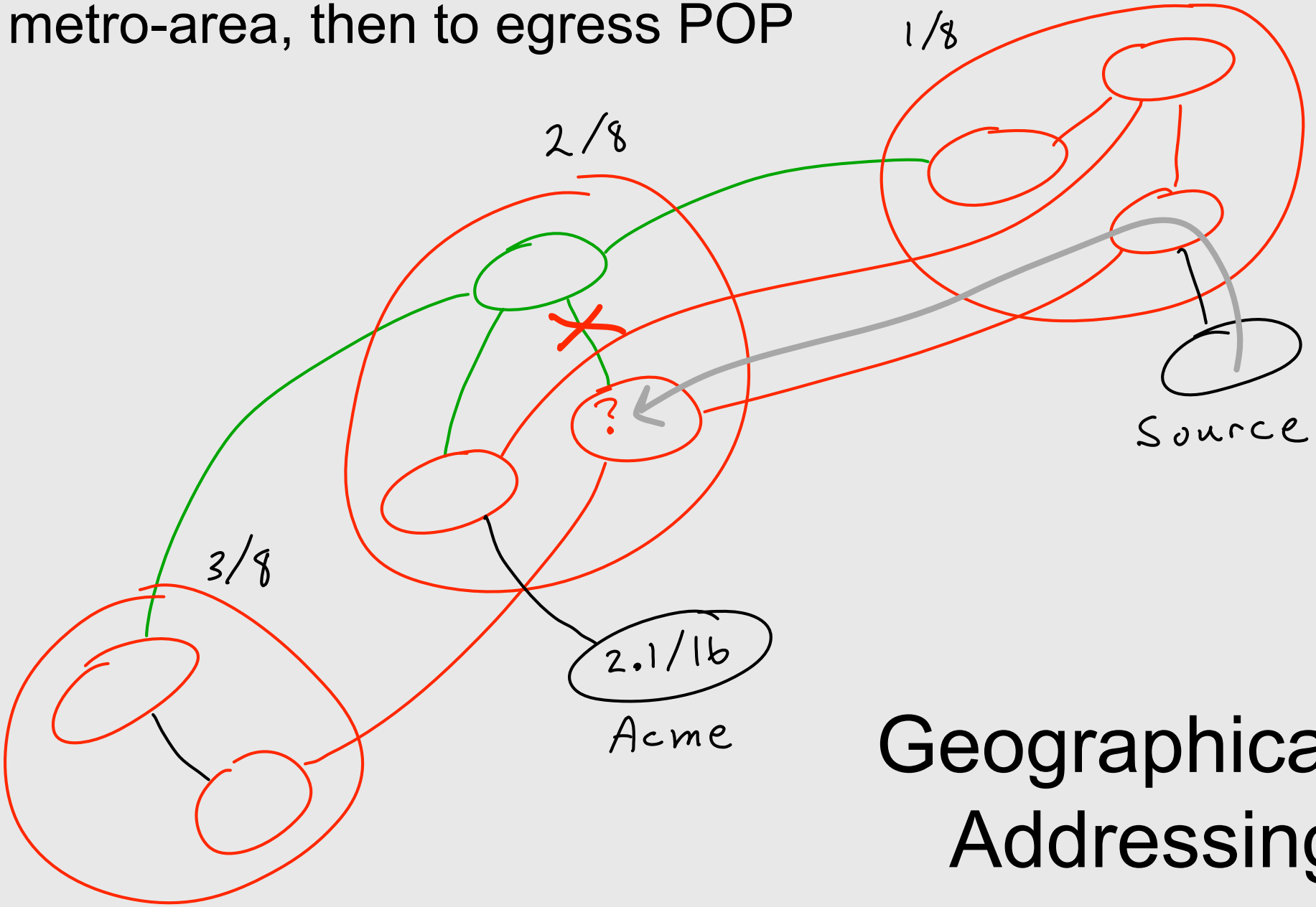
Geographical Addressing

Routing: First get packet to metro-area, then to egress POP

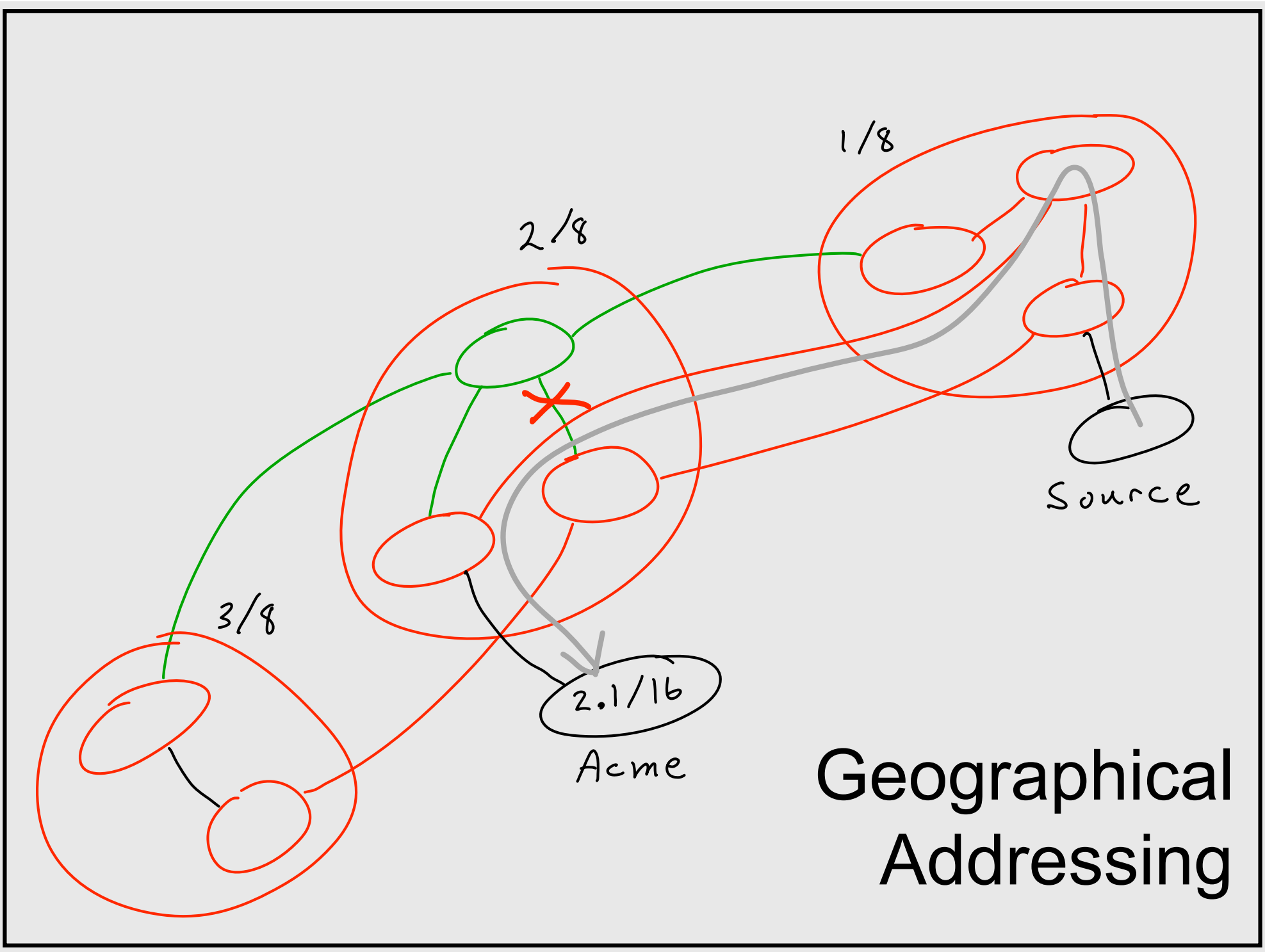


Geographical Addressing

Routing: First get packet to metro-area, then to egress POP



Geographical Addressing


1/8
2/8
3/8
Source
2.1/16
Acme
Geographical
Addressing

Geographical versus ISP-centric

ISP-centric: Robust intra-ISP topology
Geographical: Robust intra-metro topology

Need to re-think address assignment: Metro-oriented (but ISP-centric within a metro??)

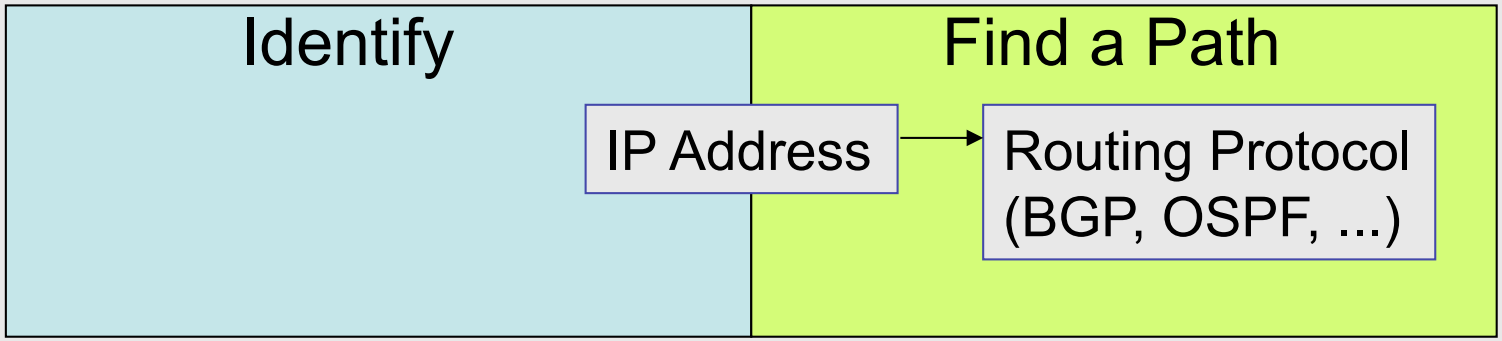
Need some political entity to manage topology and addresses in each metro

True in 1994 [F94], still true today

Indirection

IP addresses both IDENTIFY and LOCATE

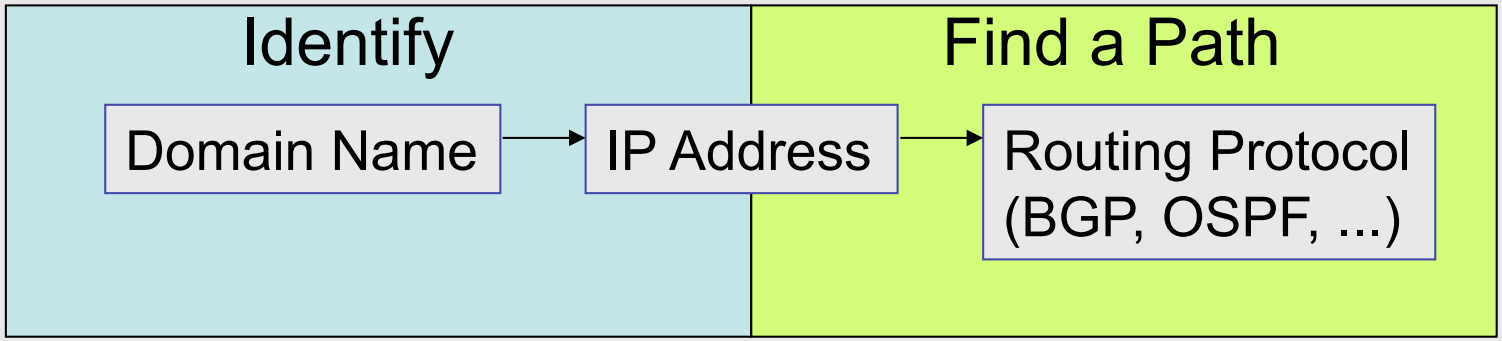
Should be singular, unique, and stable



Indirection

IP addresses both IDENTIFY and LOCATE

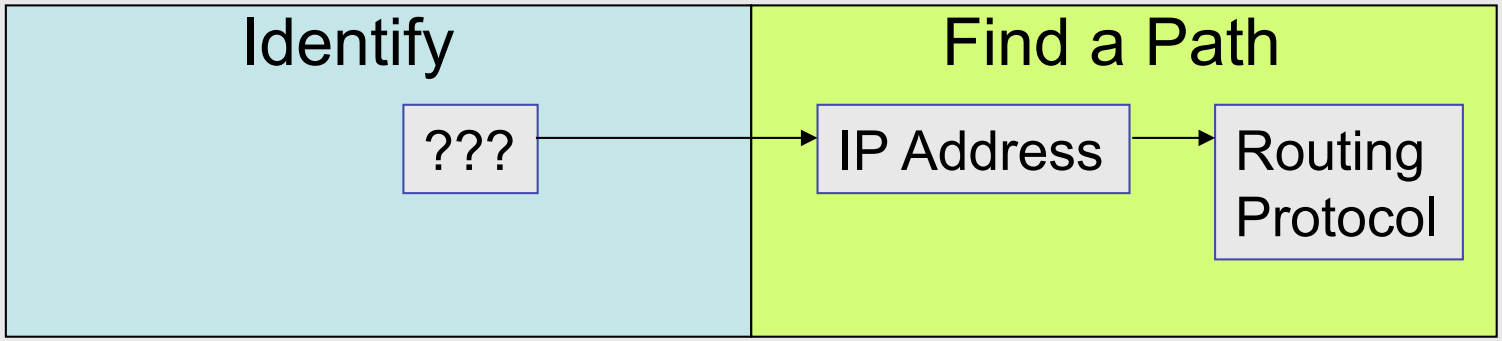
Should be singular, unique, and stable



Indirection

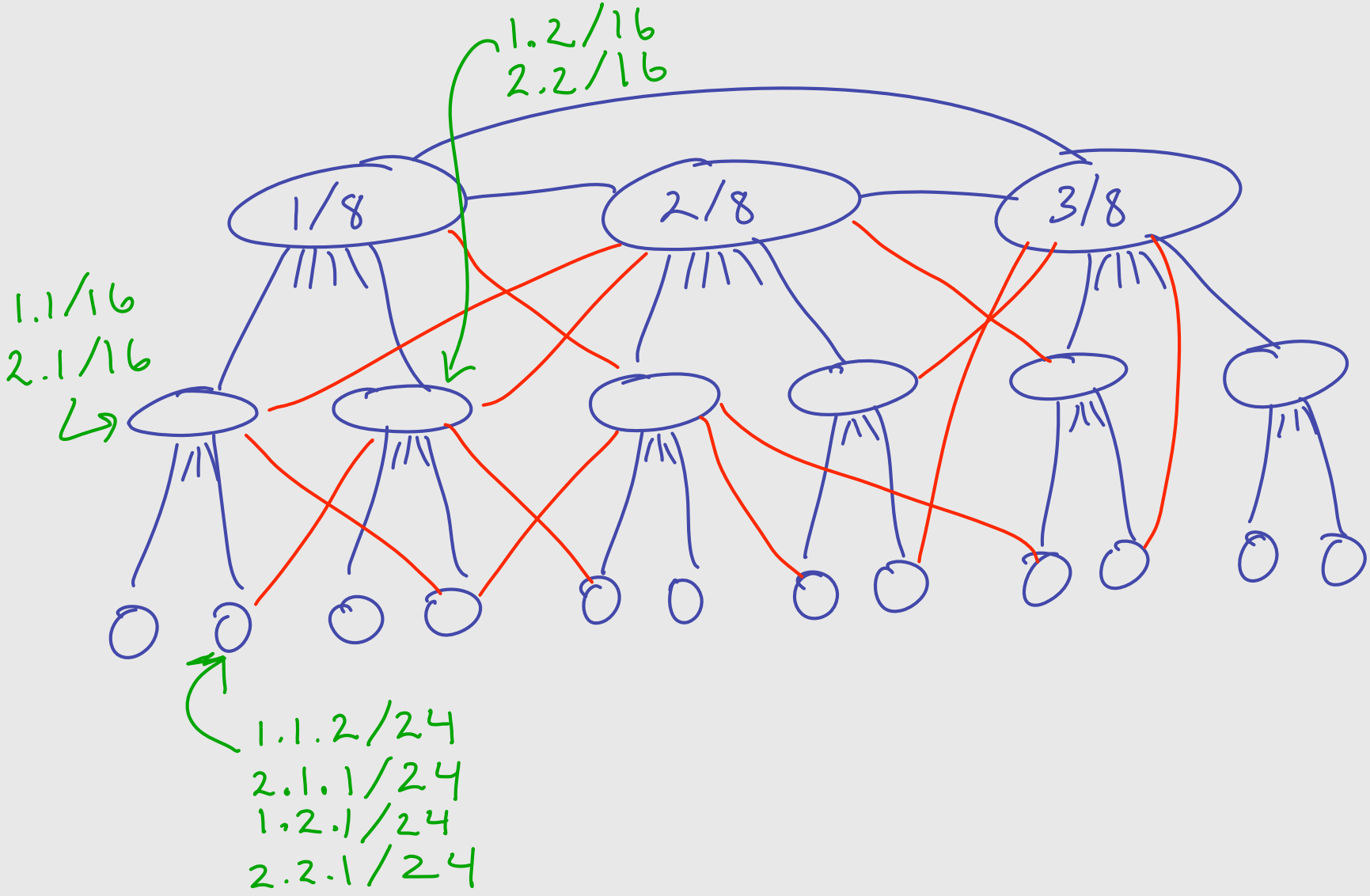
What if we limit the role of IP?

Addressing could be more flexible



Multiple, dynamic addresses

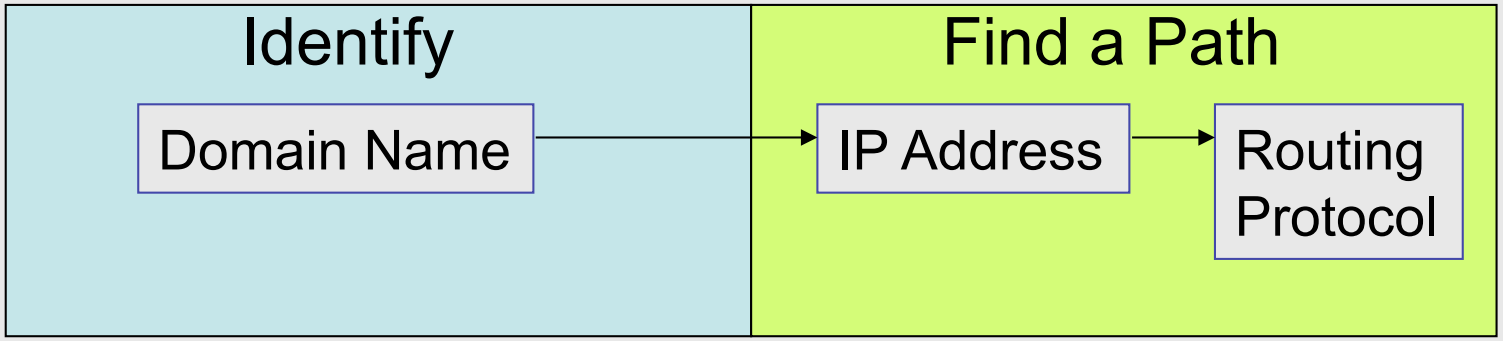
Helps with hierarchy and mobility



Map DNS names to IP addresses

Can be used to select an ISP [F91]

Later adopted by IPv6, subsequently rejected because site renumbering is hard

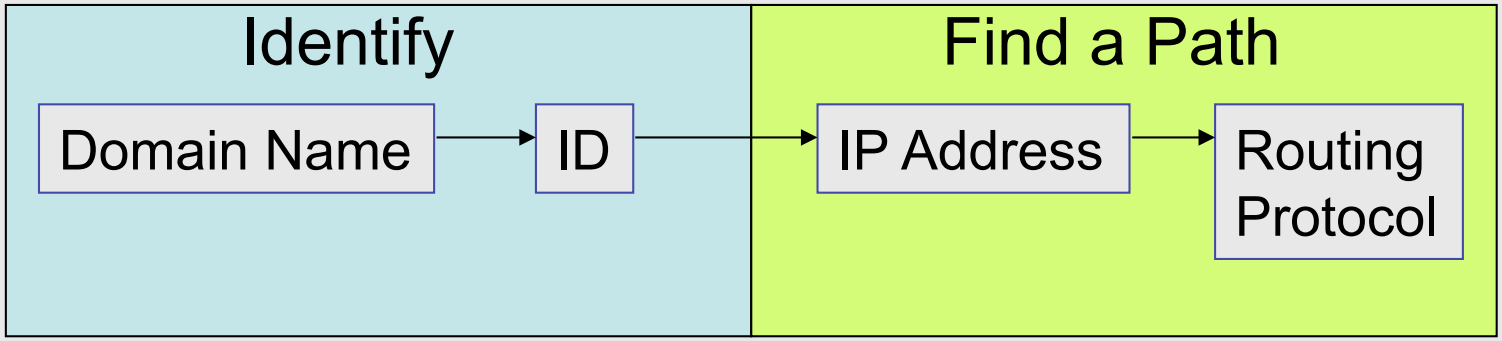


I'm fond of name-based approaches (IPNL [GF01] and NUTSS [GF07])

Flat identifier in packet headers

Early proposal for IPng (Pip) [FG94]

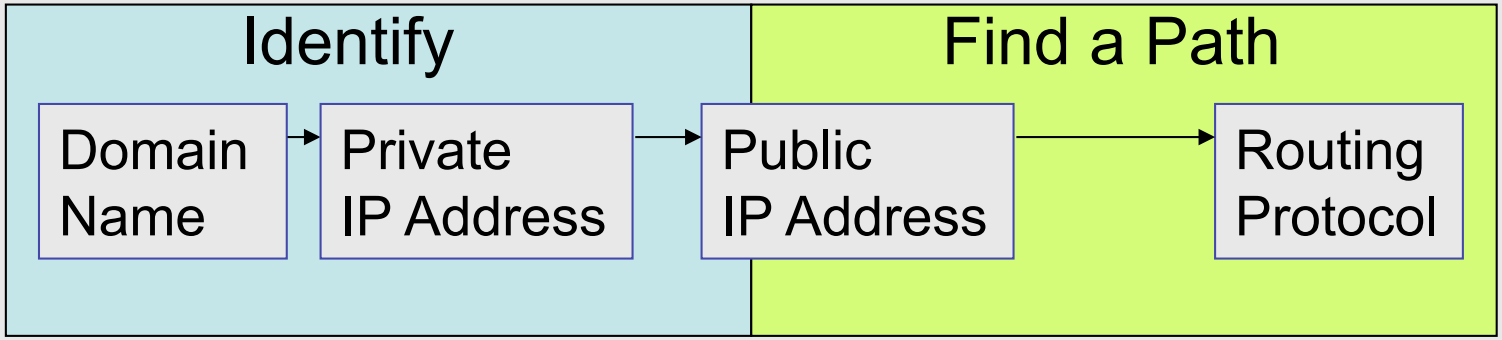
(Though naive because lacked encryption)



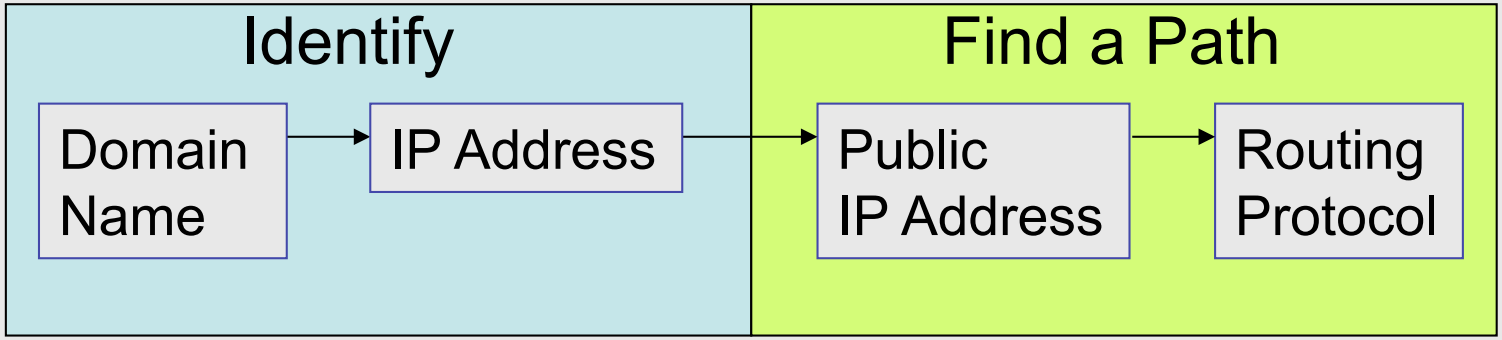
Recent: HIP, SHIM6 (IETF) and various research papers (i3, DONA . . .)

Map IP addresses to IP addresses

Network Address Translation (NAT) [FE93]

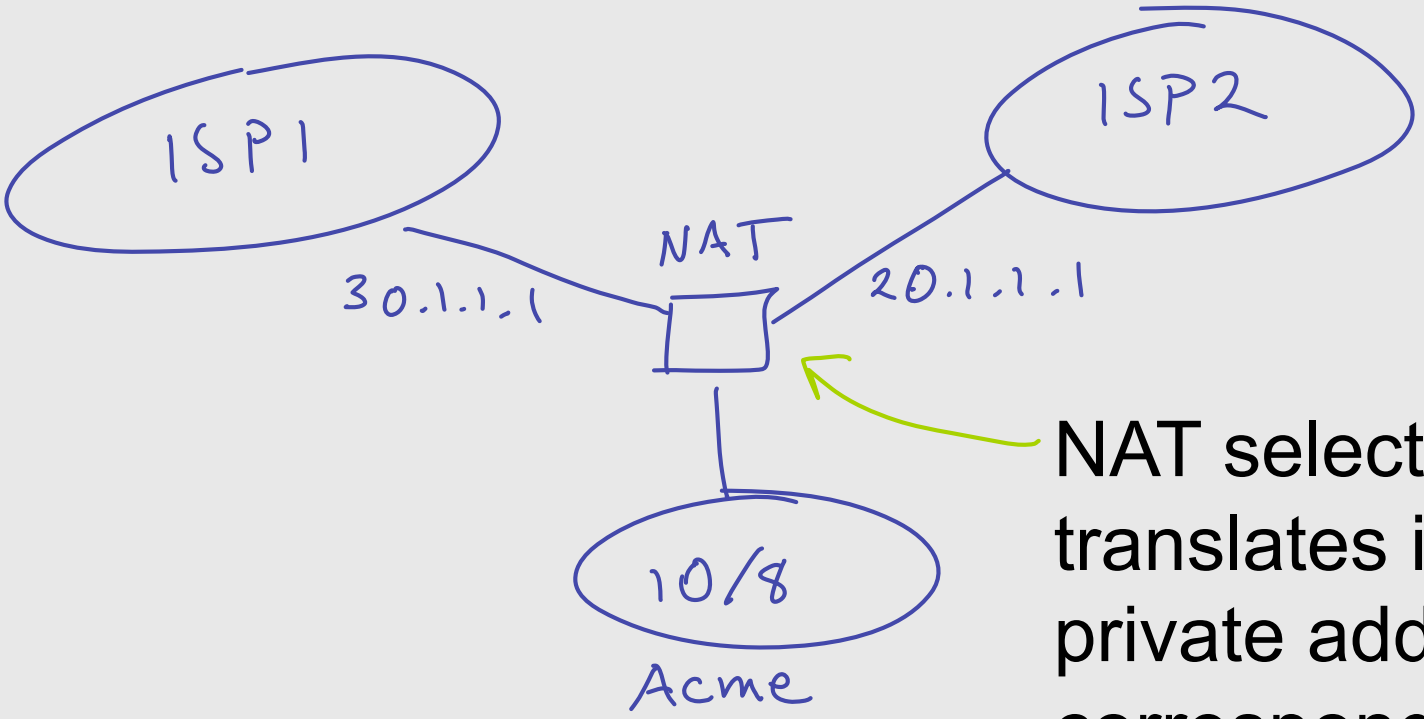


Tunnels



NAT commonly used for multi-homing

Including load balance

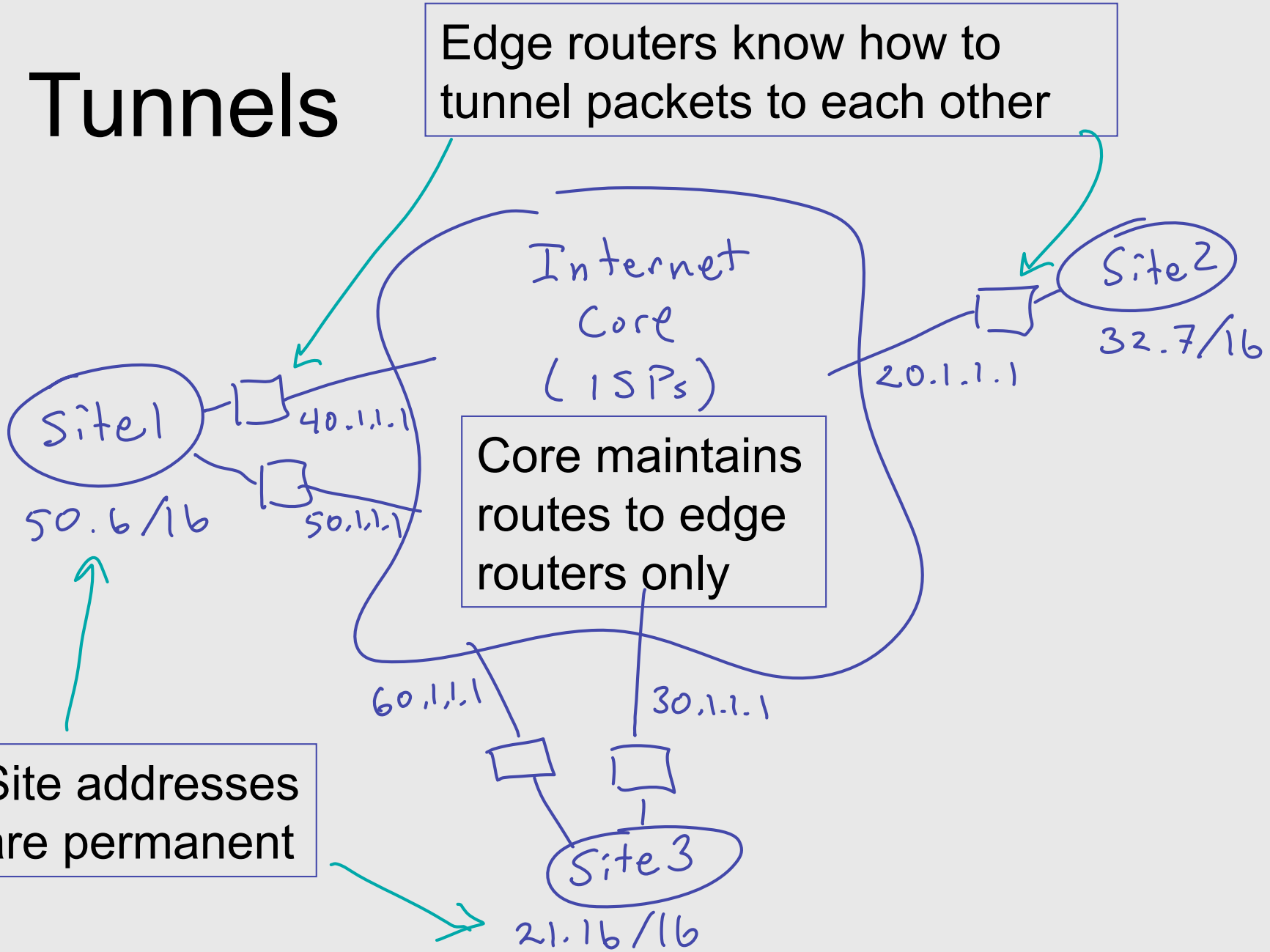


NAT selects a link, translates internal private address into corresponding global public address

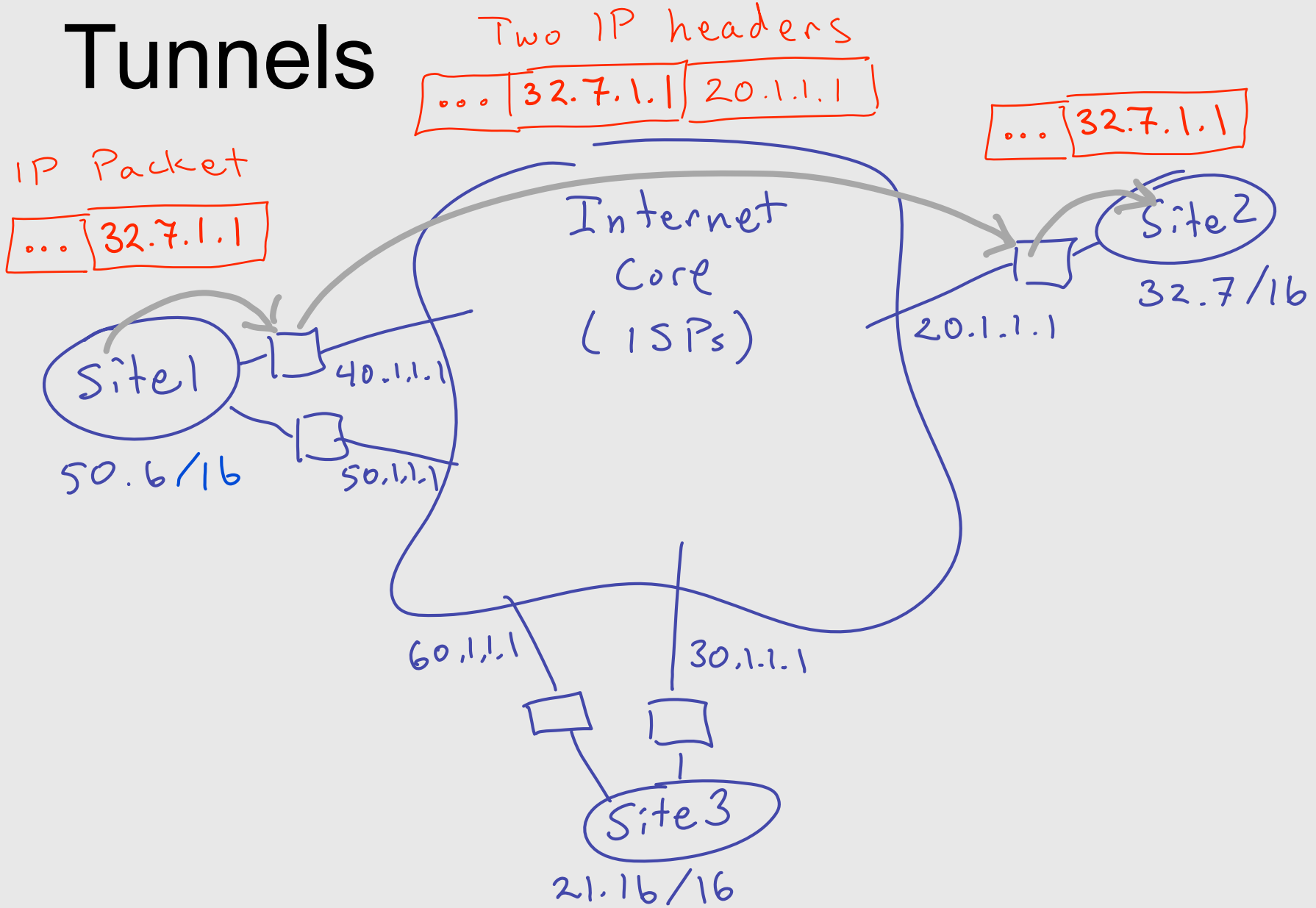
Many Problems...

(though most problems go away with IPv6)

Tunnels


Edge routers know how to tunnel packets to each other
Internet Core (ISPs)
Site1
50.6/16
40.1.1.1
50.1.1.1
Site2
32.7/16
20.1.1.1
Core maintains routes to edge routers only
60.1.1.1
30.1.1.1
Site3
21.16/16
Site addresses are permanent

Tunnels

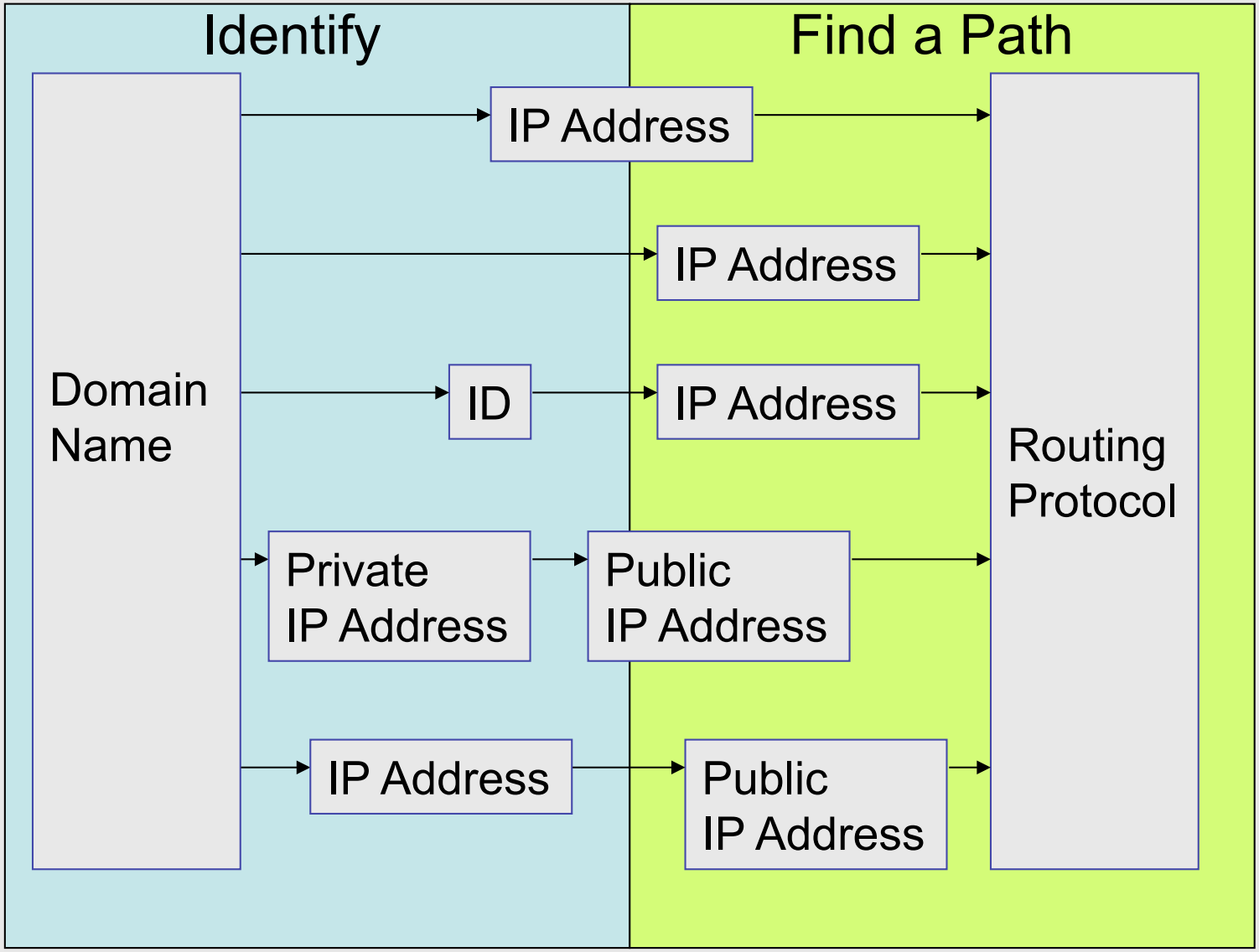

Two IP headers
... 32.7.1.1 20.1.1.1
IP Packet
... 32.7.1.1
... 32.7.1.1
Internet
Core
(ISPs)
Site1
50.6/16
40.1.1.1
50.1.1.1
20.1.1.1
Site2
32.7/16
60.1.1.1
30.1.1.1
Site3
21.16/16

Tunnels

Recently suggested for global routing

- Routing Research Group (RRG) in IRTF (many proposals)
- Main issue is how to distribute global mapping table
- Cache versus full, push versus pull, failure recovery . . .

So many ideas, so little impact!



So many ideas, so little impact!

Industry impact in networking is hard

All players must see short term $$$$

- Standards: IETF, IEEE, ITU . . .
- Vendors: OS, host, network gear . . .
- Providers: ISP, enterprise, data center . . .

Virtual Aggregation

Reduces routing table size

- Easily order-of-magnitude

Negligible performance penalty

- Latency and load

CRIO [ZF06]

No software or protocol changes

- Config changes only

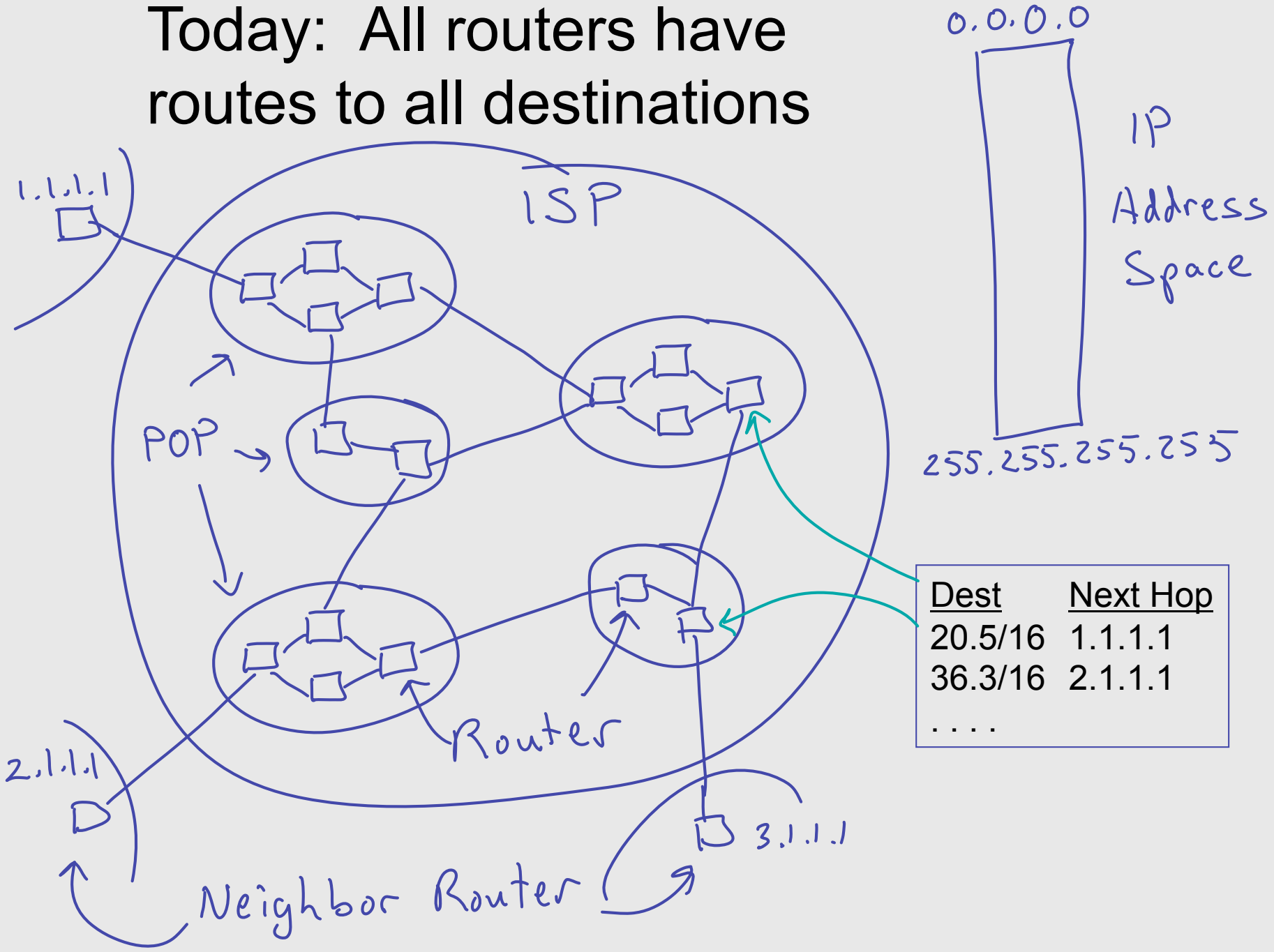
ISPs can independently and autonomously deploy

[BF08]

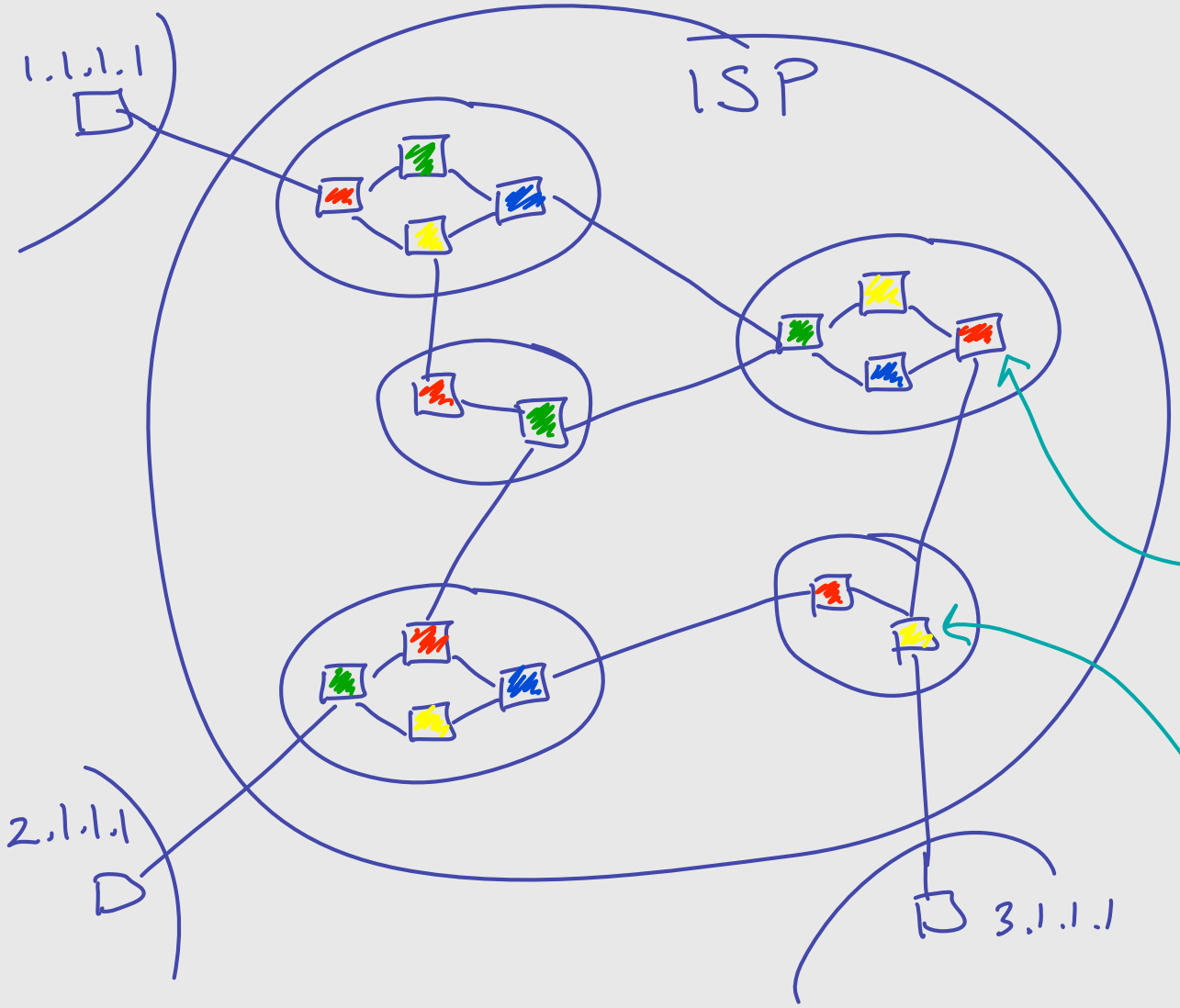
Chance of Impact?

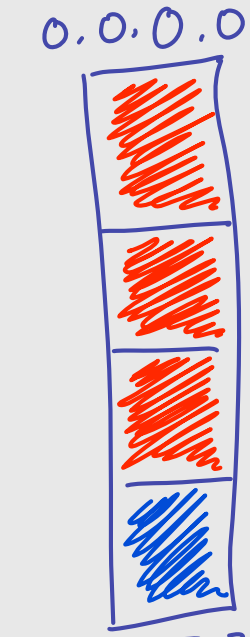
- Only one player involved (ISP)
  - No standards or vendors
- Addresses specific pain point
  - ISPs need to upgrade router due to FIB size
  - (Note that this may hurt router vendors)
- Never-the-less, disrupts network operation
  - New configuration must be error-free
  - ISPs are risk averse

Today: All routers have routes to all destinations



Virtual Aggregation: Routers have routes to only part of the address space



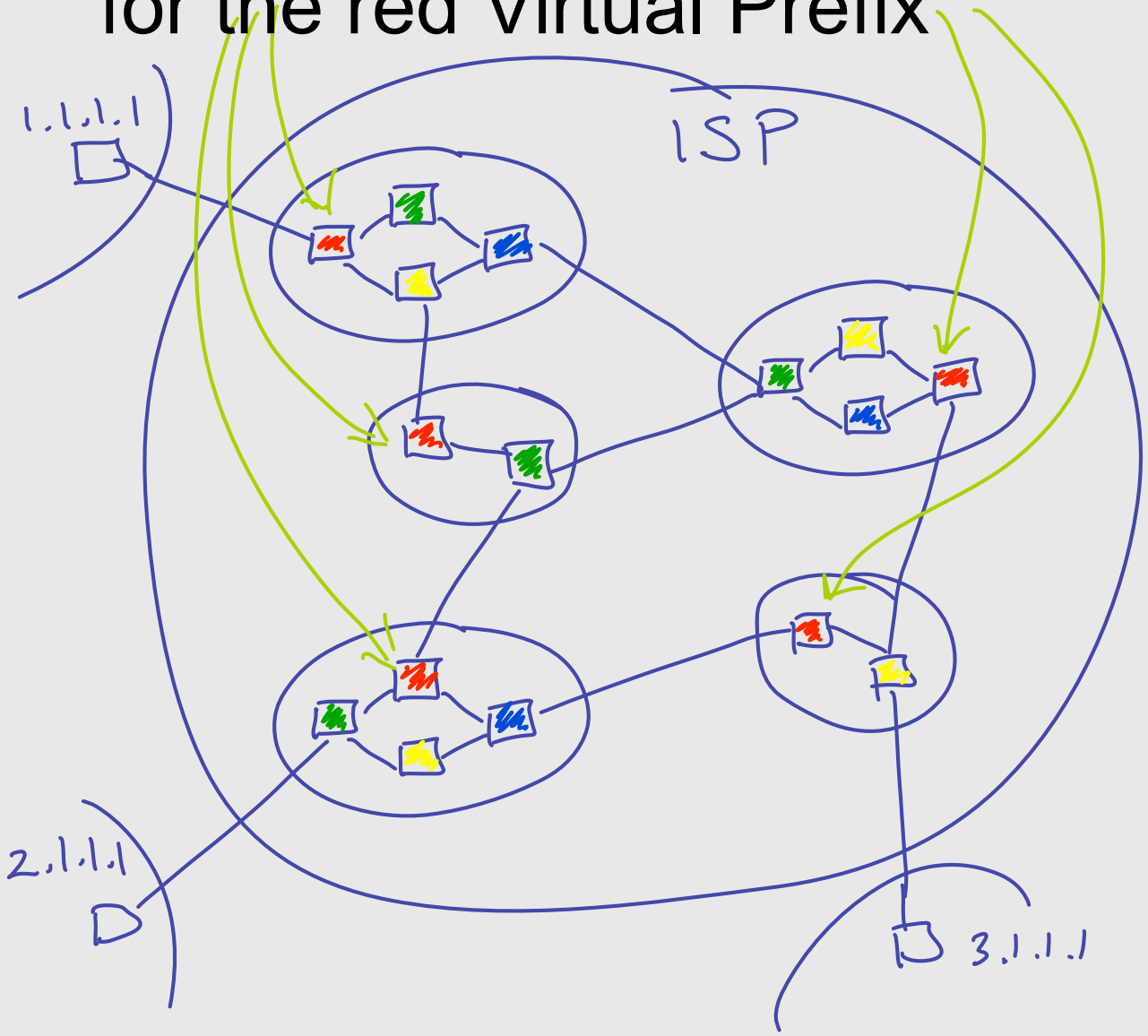


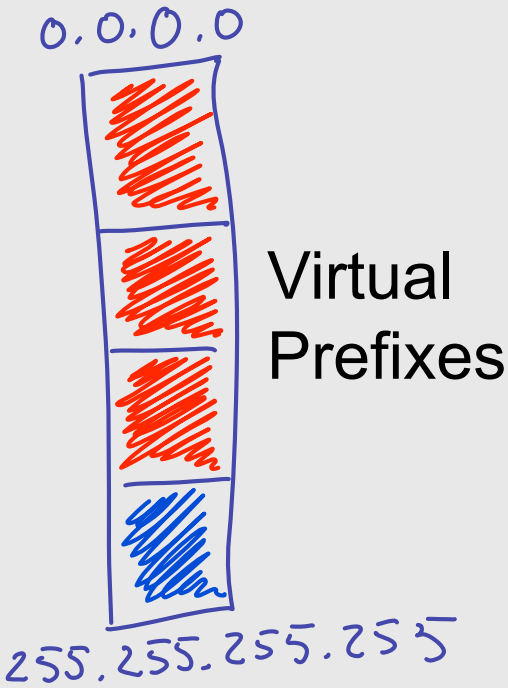
Virtual Prefixes

Dest	Next Hop
20.5/16	1.1.1.1
. . . .	

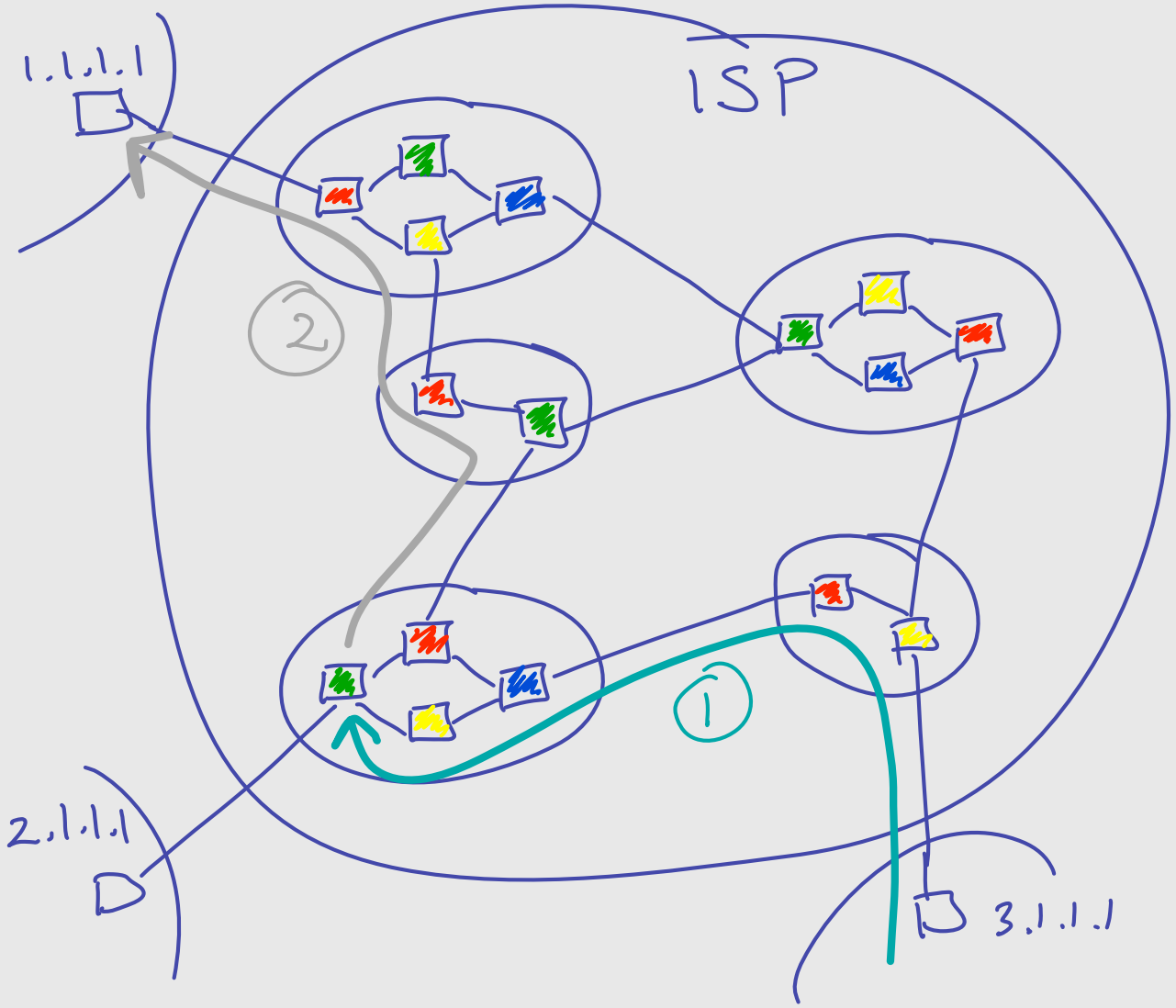
Dest	Next Hop
188.3/16	2.1.1.1
. . . .	

"Aggregation Point" routers for the red Virtual Prefix





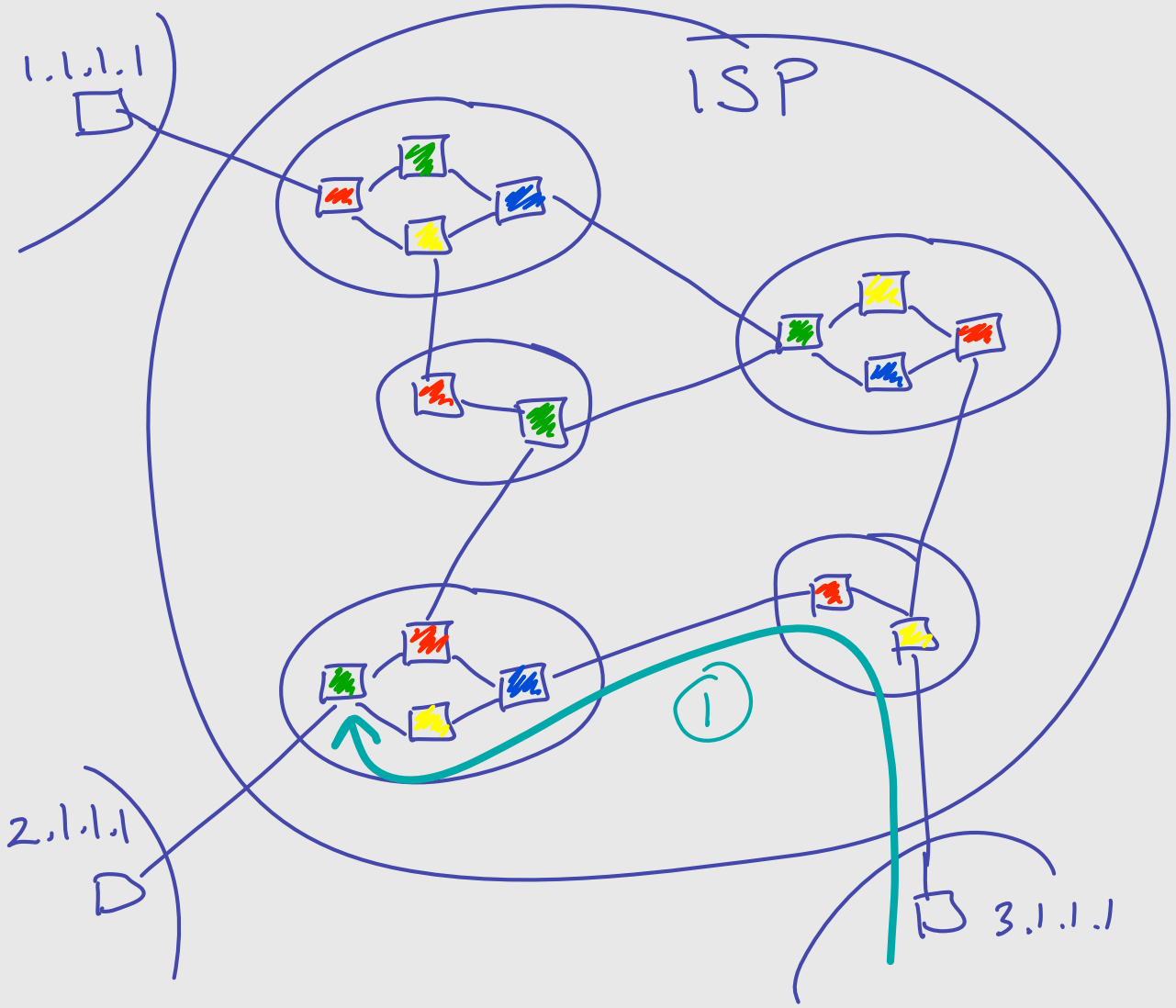
Paths through the ISP have two components:



1: Route to a nearby Aggregation Point

2: Tunnel to the neighbor router

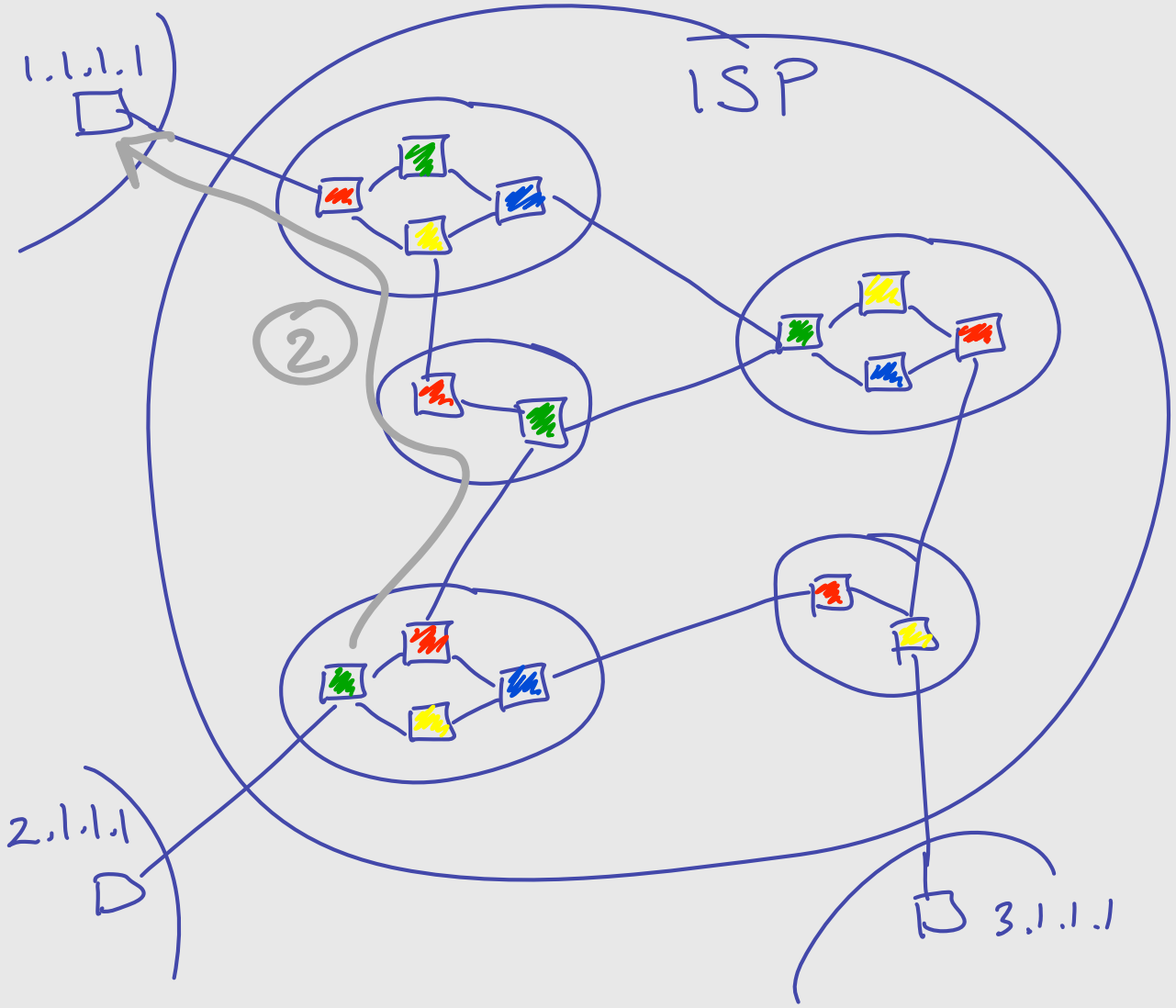
1: Routing to a nearby Aggregation Point



Configure Aggregation Point with static route for the Virtual Prefix

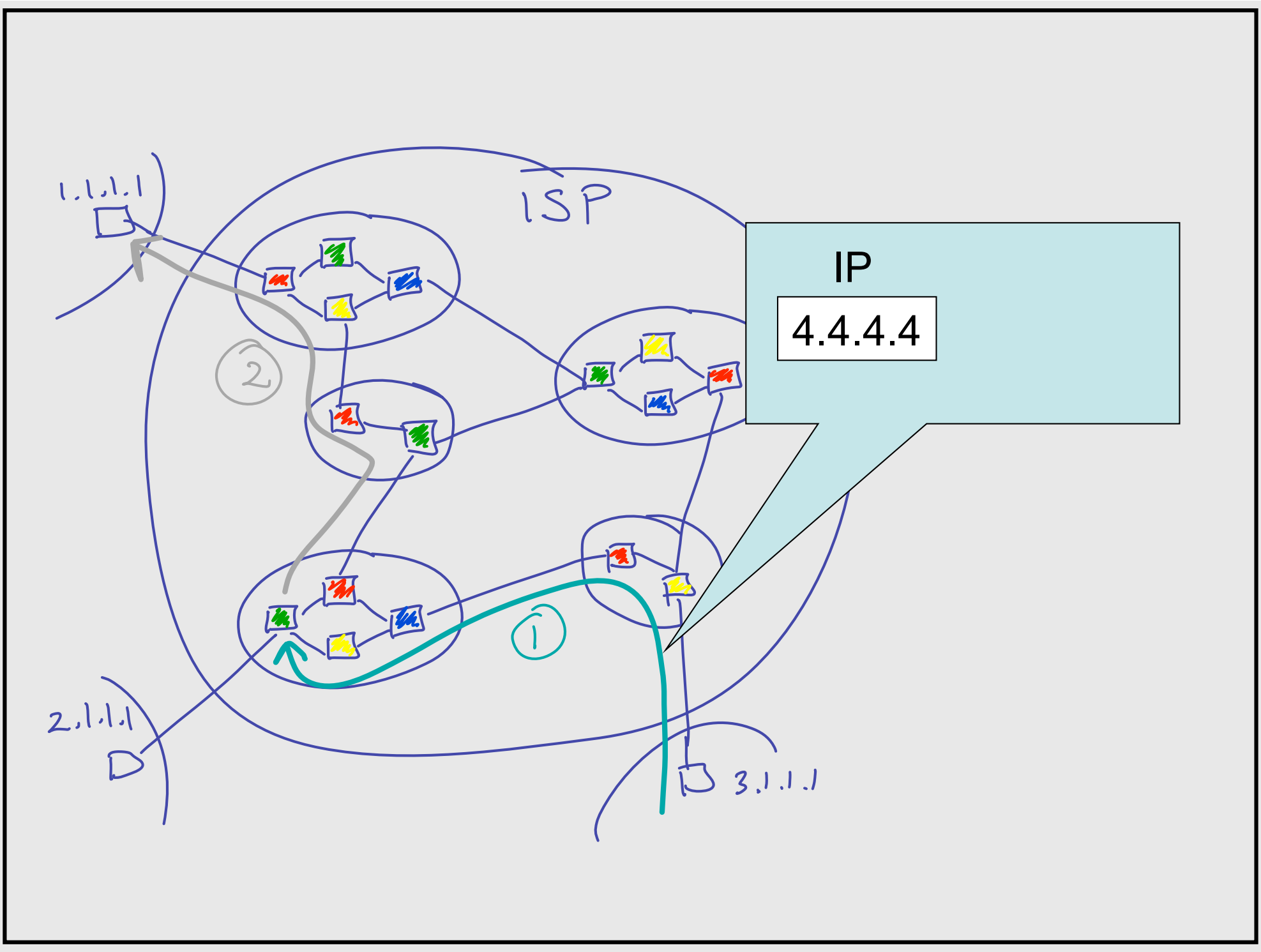
Virtual Prefix is advertised into BGP

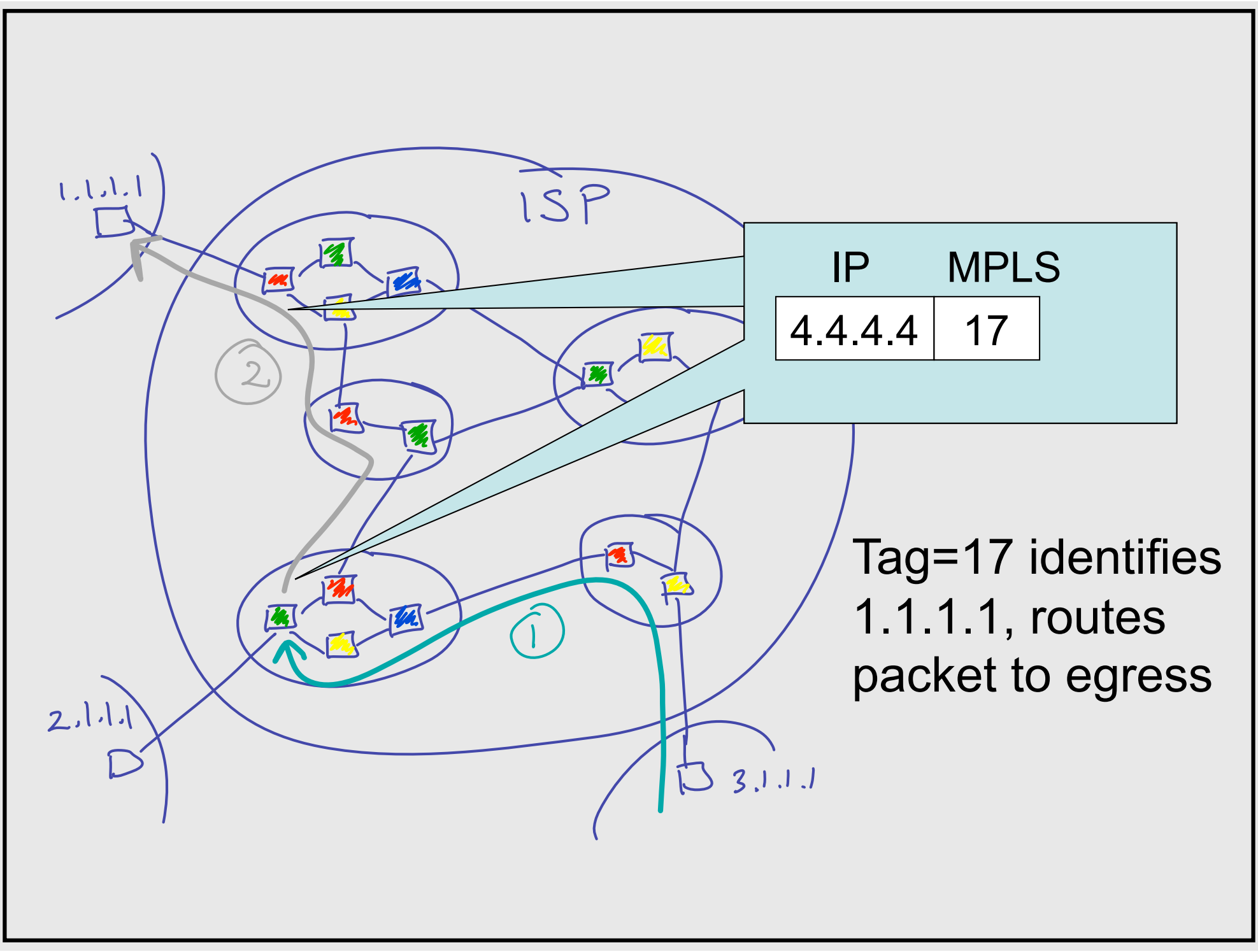
2: Tunnel packet to neighbor router (MPLS)

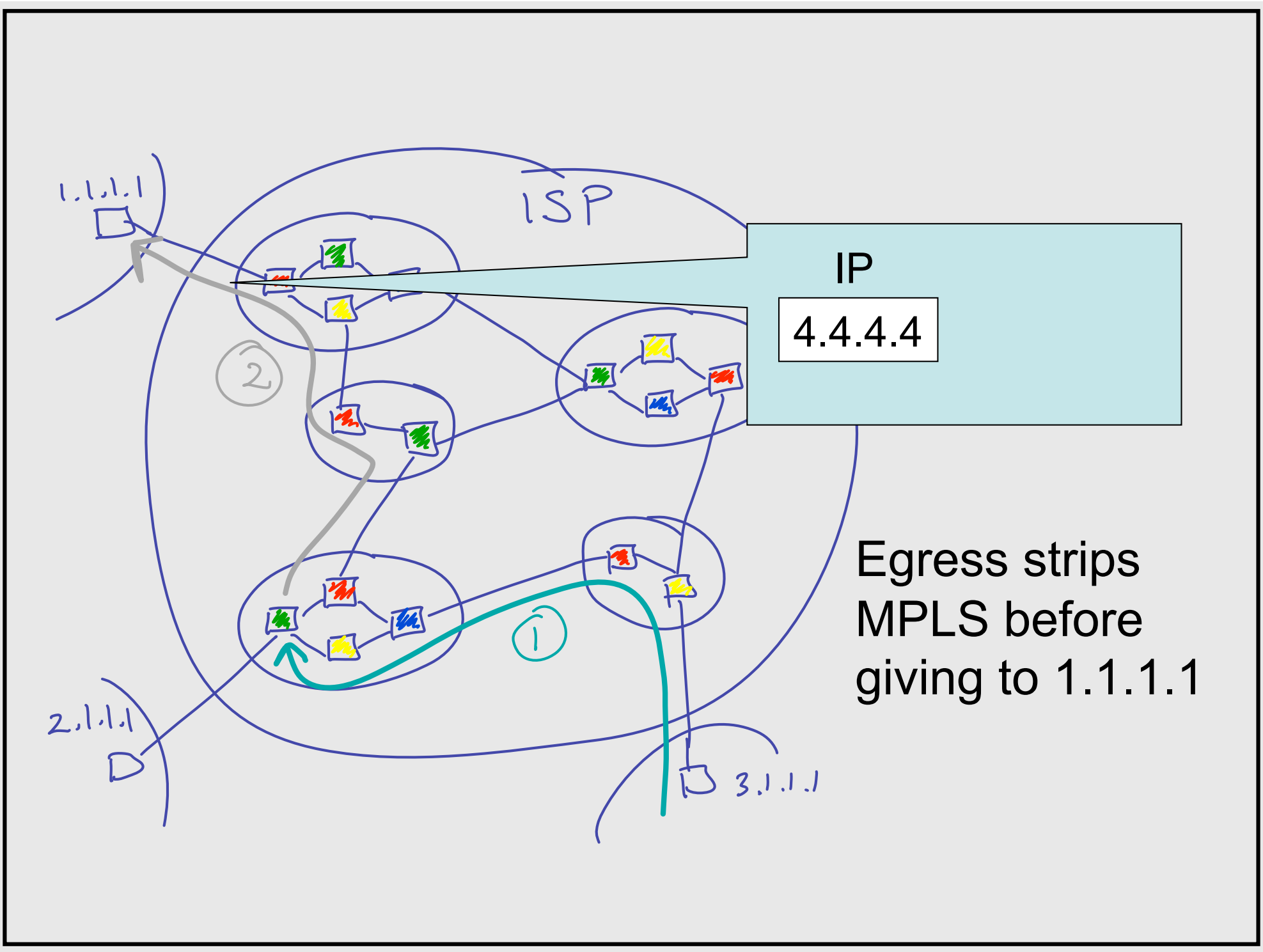


Static routes for all neighbors are imported into OSPF

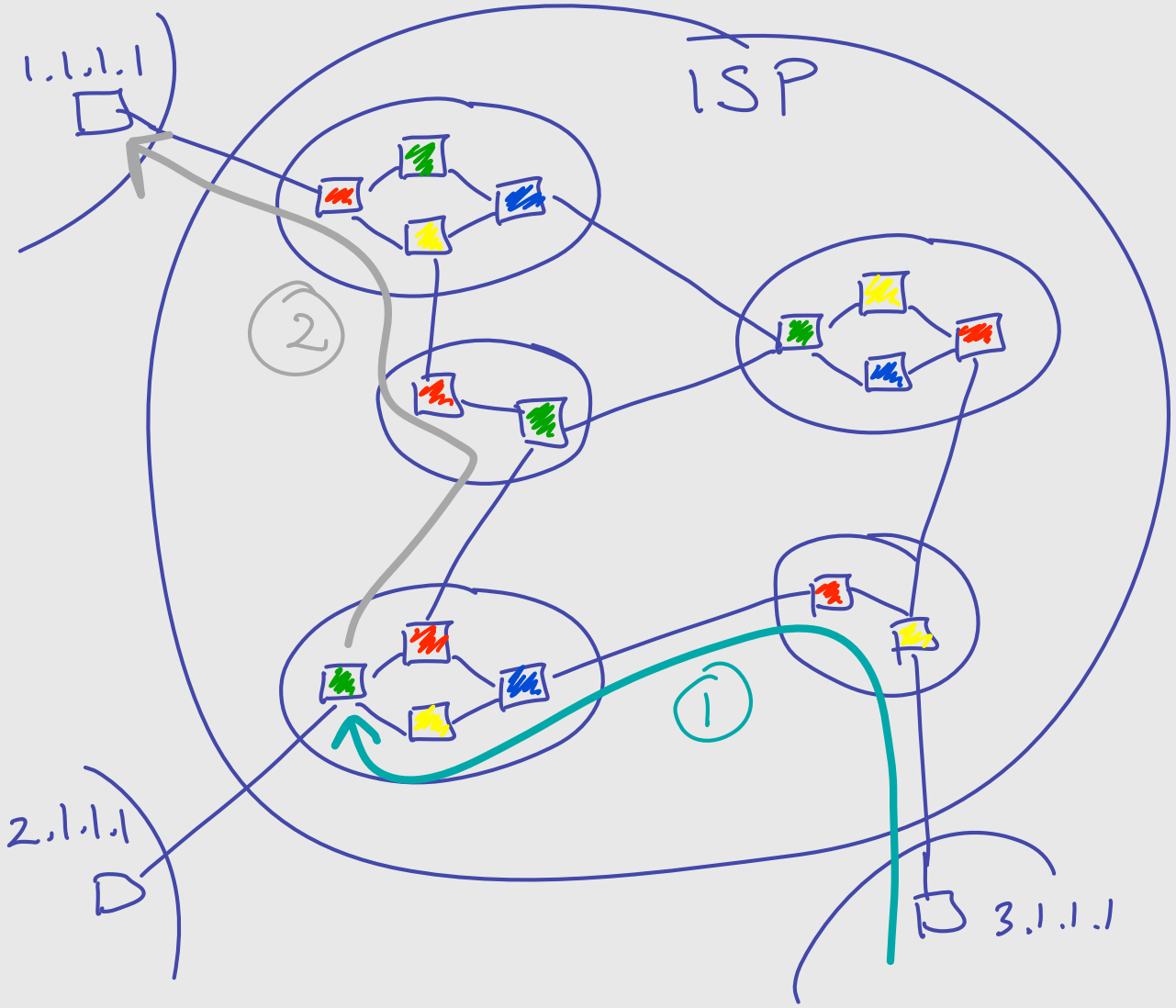
MPLS LDP creates tunnels to every neighbor router


1.1.1.1
ISP
IP
4.4.4.4
2
1
2.1.1.1
D
3.1.1.1


1.1.1.1
ISP
IP
MPLS
4.4.4.4
17
2
1
2.1.1.1
D
3.1.1.1
Tag=17 identifies 1.1.1.1, routes packet to egress


1.1.1.1
ISP
IP
4.4.4.4
2
1
2.1.1.1
3.1.1.1
Egress strips
MPLS before
giving to 1.1.1.1

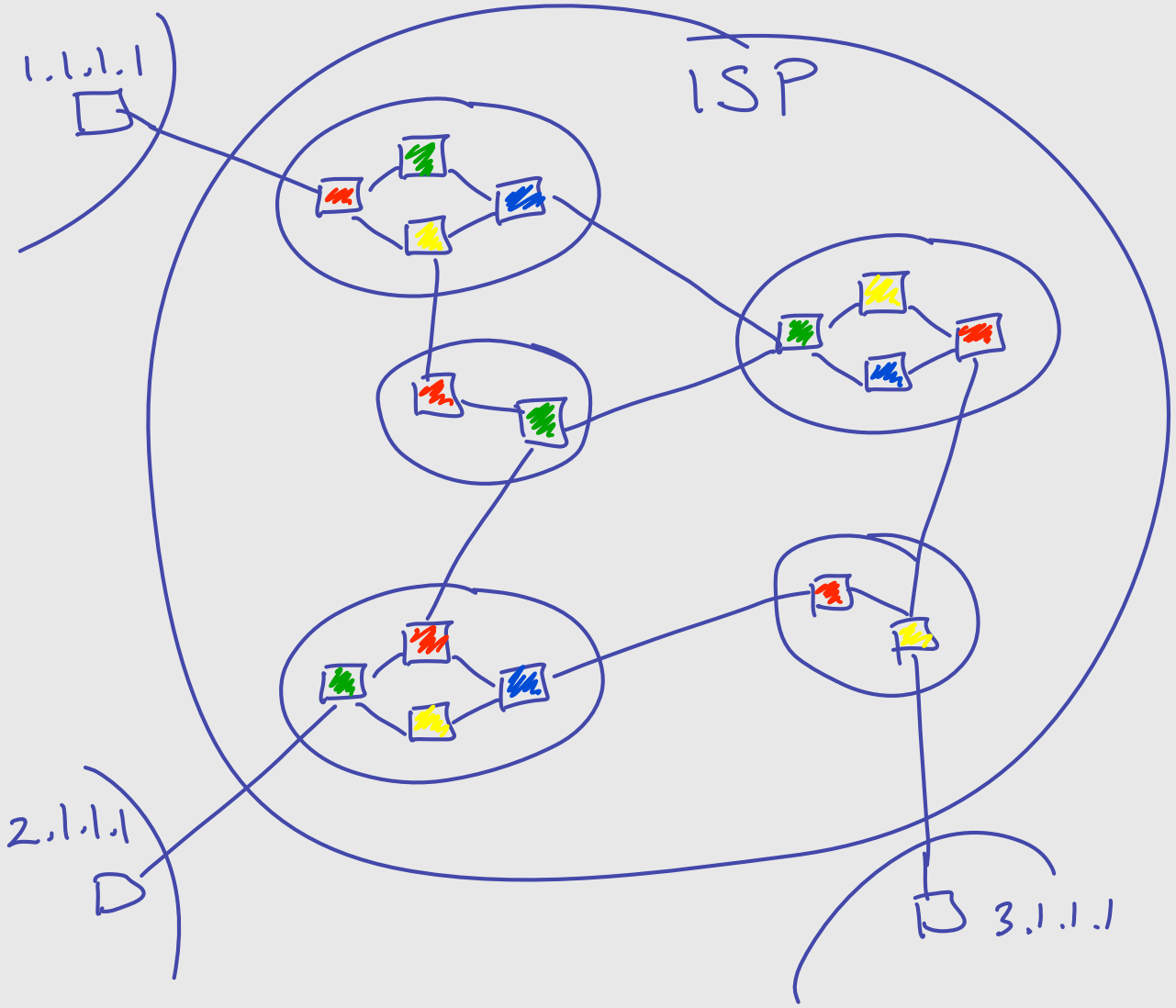
Virtual Aggregation paths can be longer than shortest path



Adds to both latency and load

Basic table-size versus overhead trade-off

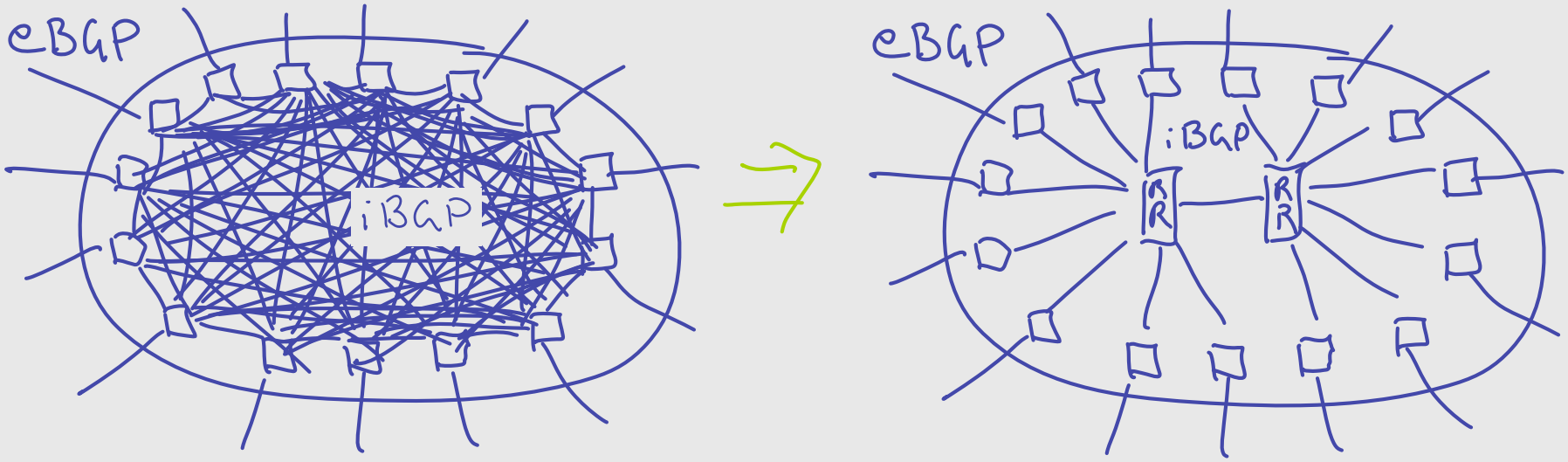
Neighbor routers require full routing tables!



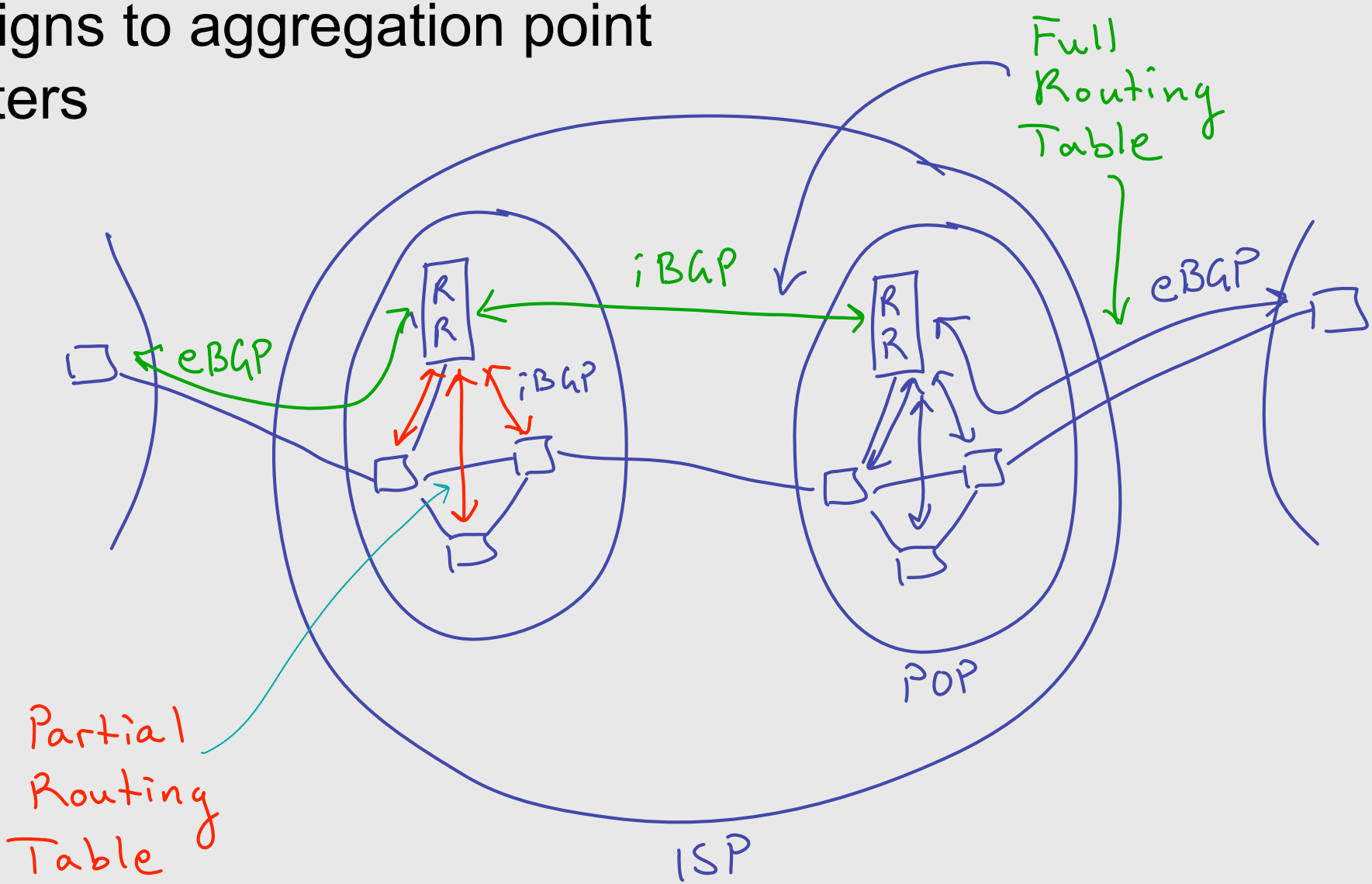
How can an Aggregation Point router peer with a neighbor router?

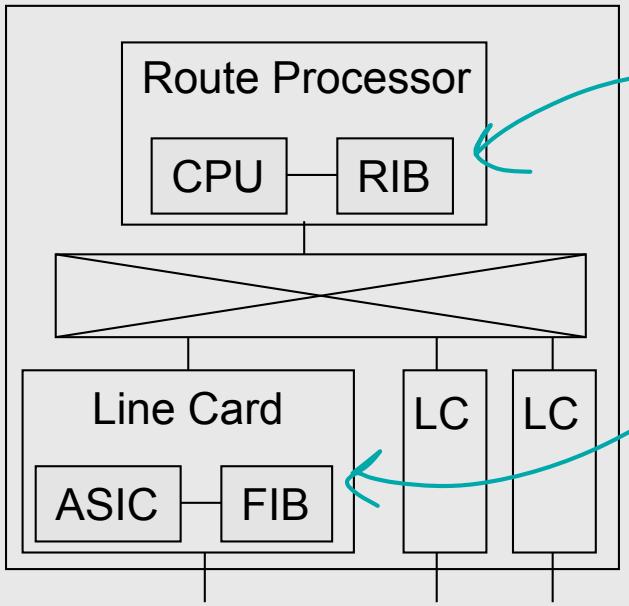
Use a Route Reflector (RR) to peer with neighbors

Hierarchy of RR's are used by ISPs today to help scale iBGP (interior BGP)



Route Reflectors (RR) filter out prefixes from neighbors and assigns to aggregation point routers





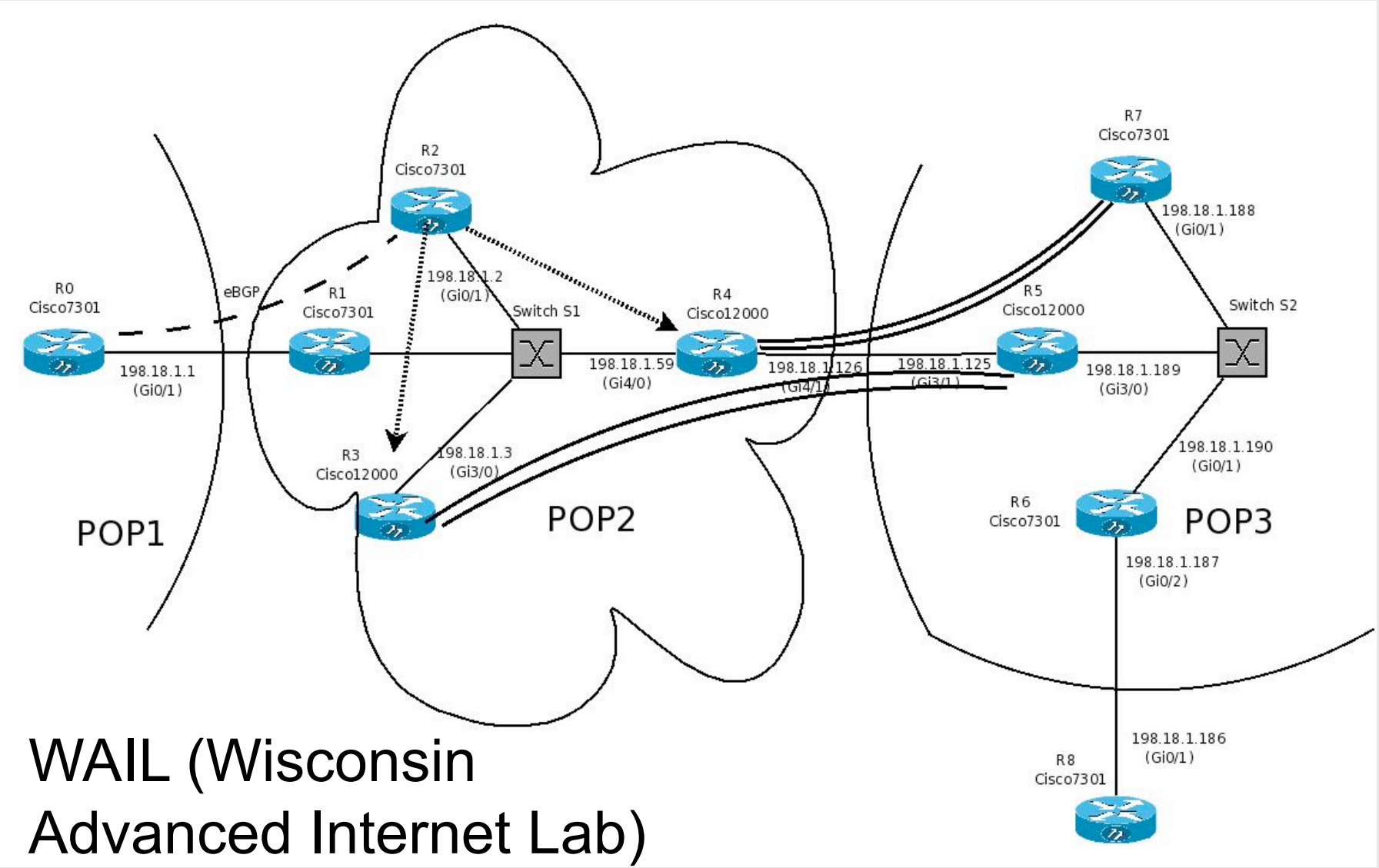
RR doesn't need fast FIB memory---full routing table stored in RR RIB only

Routers need fast FIB, but only need to store partial routing table

RR's don't forward packets, so don't need (expensive) line card FIB memory

RR's scale by number of neighbors (hierarchical organization)

Configuration Testbed



WAIL (Wisconsin Advanced Internet Lab)

Minimizing Overhead

- Traffic volume follows a power-law distribution
  - 95% of traffic goes to 5% of prefixes
  - This has held up for years
- Install “Popular Prefixes” in routers
  - On a per-POP or per-router basis
  - Different POPs have different popular prefixes
  - Popular prefixes are stable over weeks

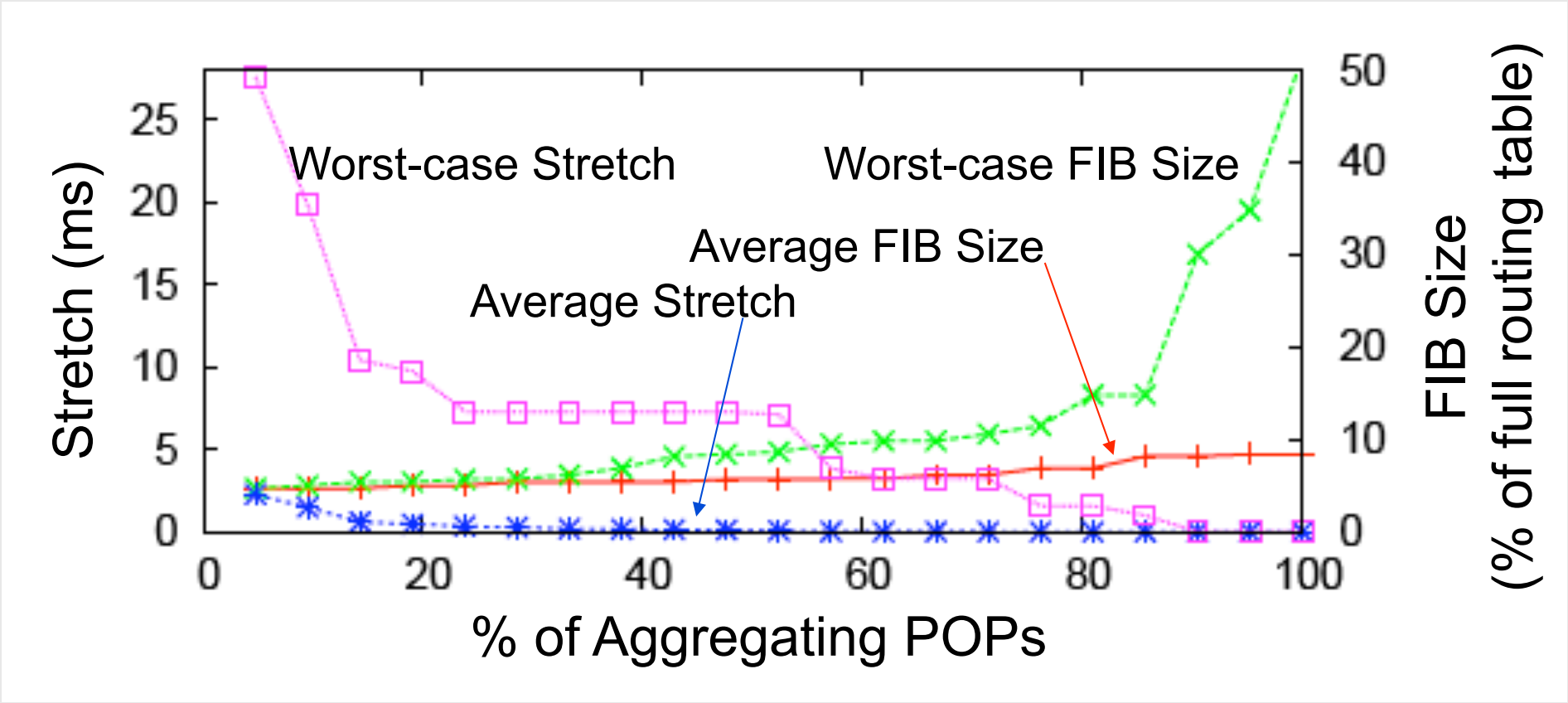
Performance Study

Data from a large tier-1 ISP

- Topology and traffic matrix

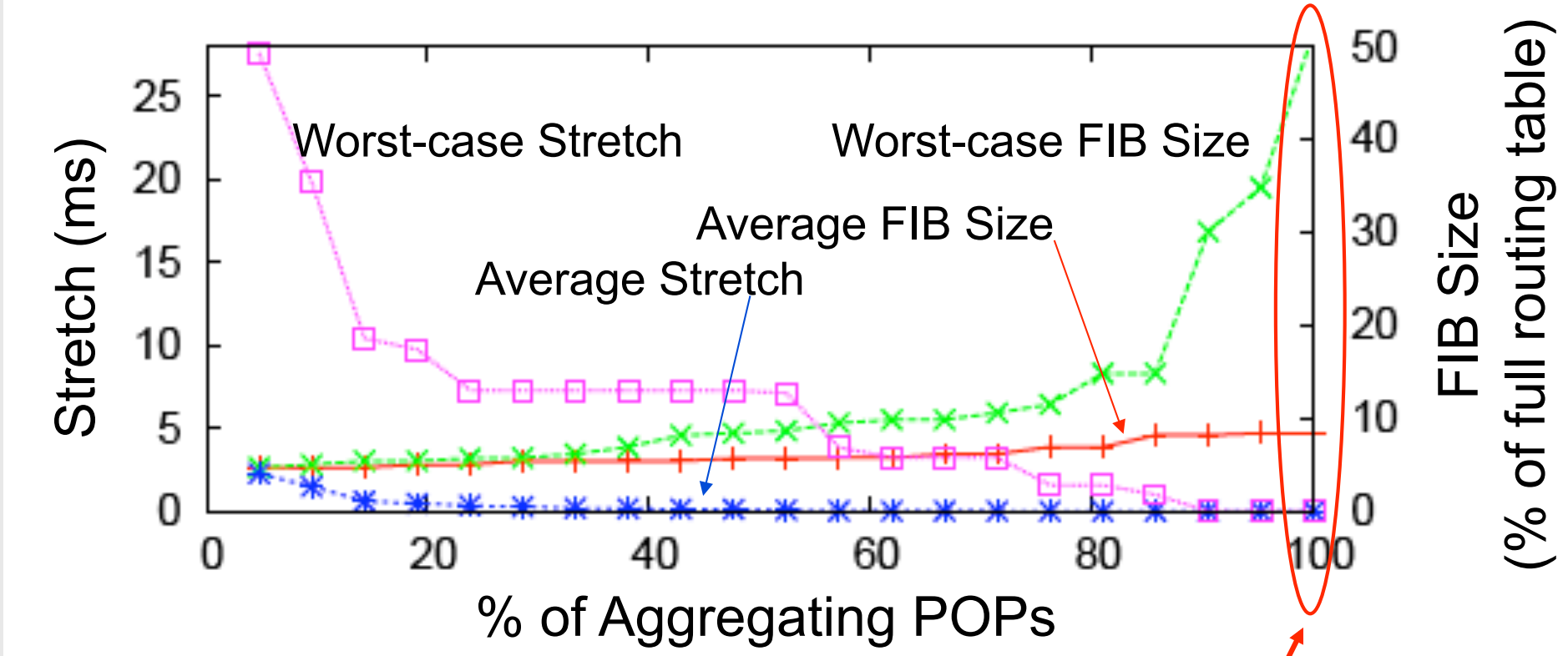
Vary number of Aggregation Points (AP) and number of popular prefixes

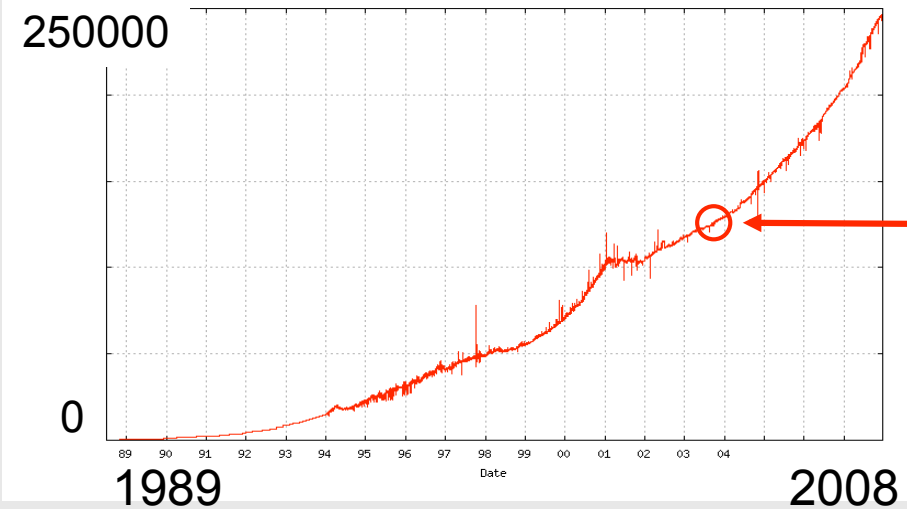
- Naive AP deployment: A POP has either (redundant) AP's for all virtual prefixes, or no virtual prefixes
- Naive popular prefixes deployment: same popular prefixes in all routers



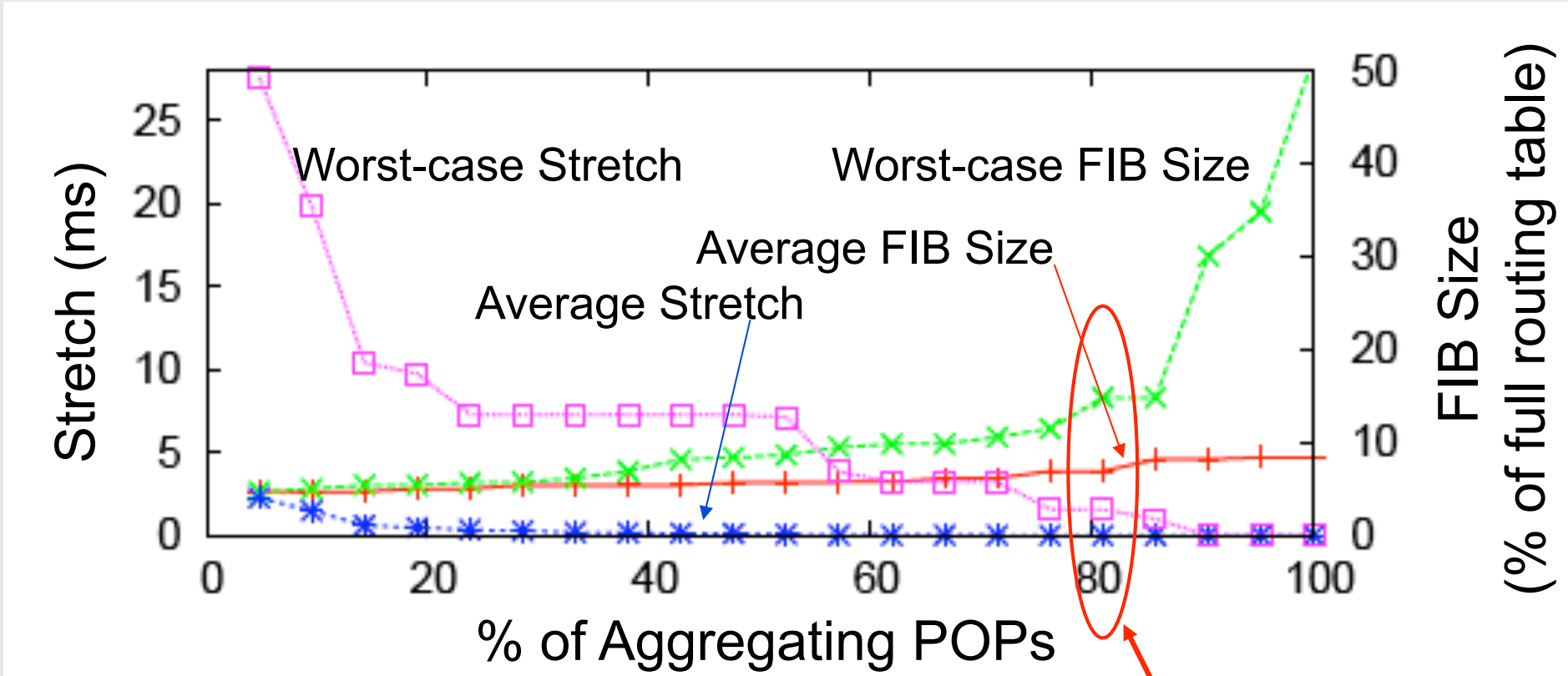
Install 1.5% of popular prefixes in all routers

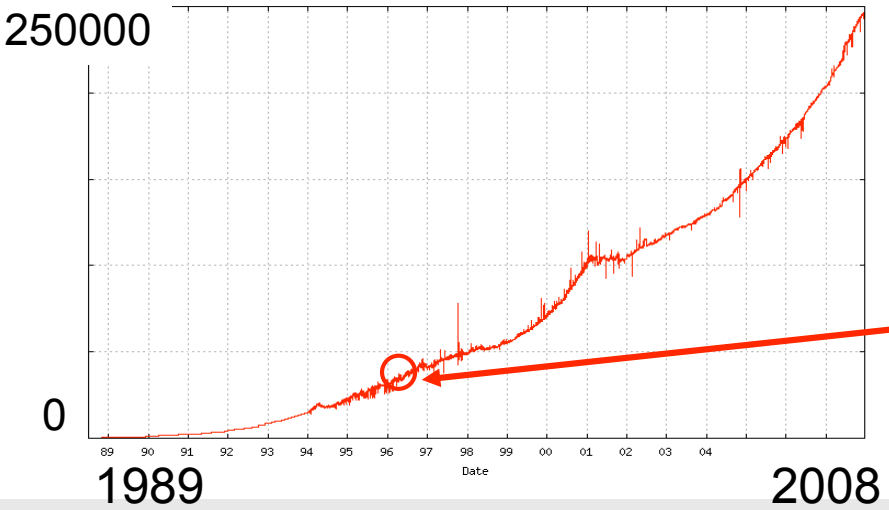
Stretch versus FIB size



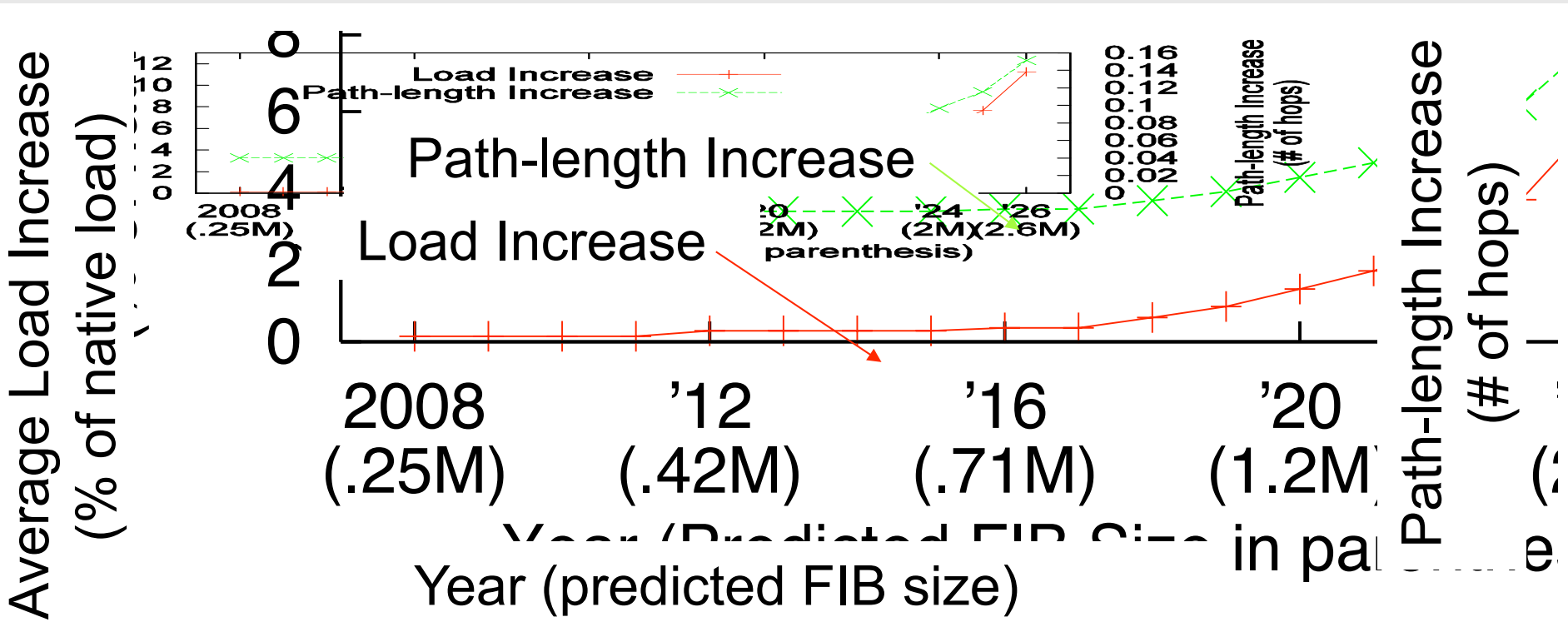


Cuts FIB in half (2004 level), virtually no stretch



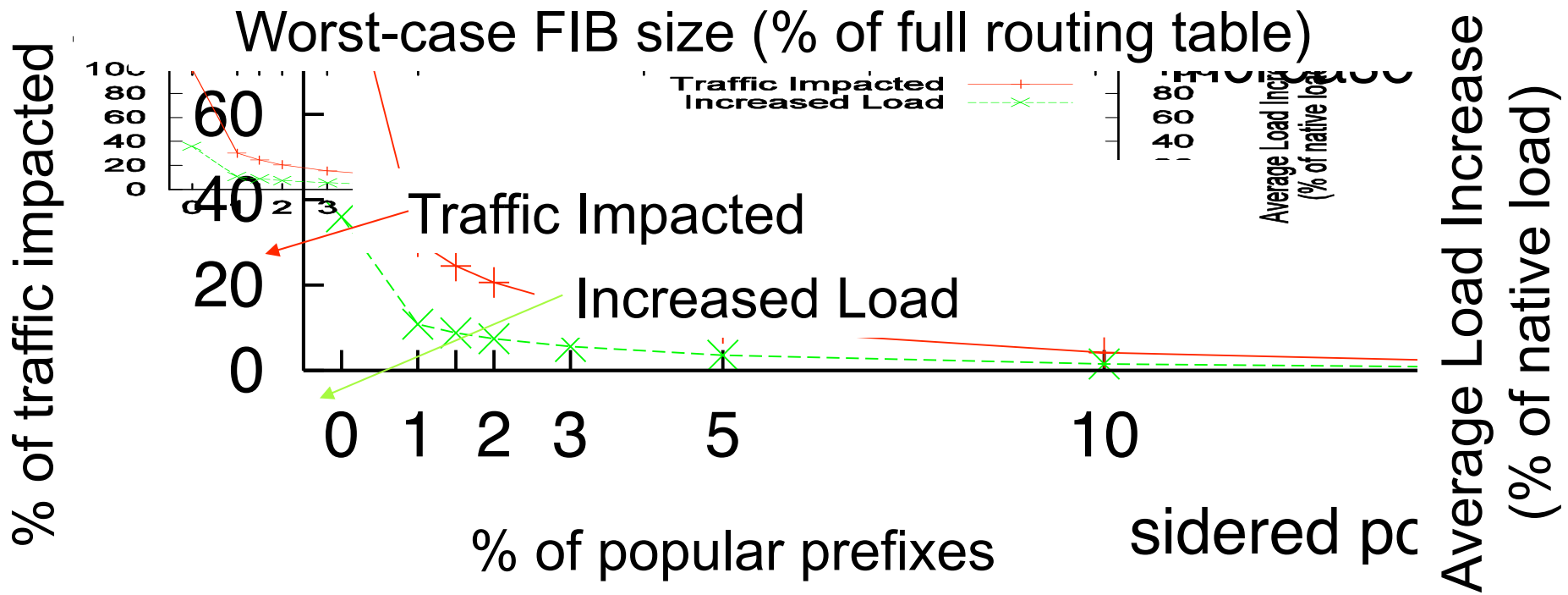


Cuts FIB five times (1996 level), worst case 2ms stretch



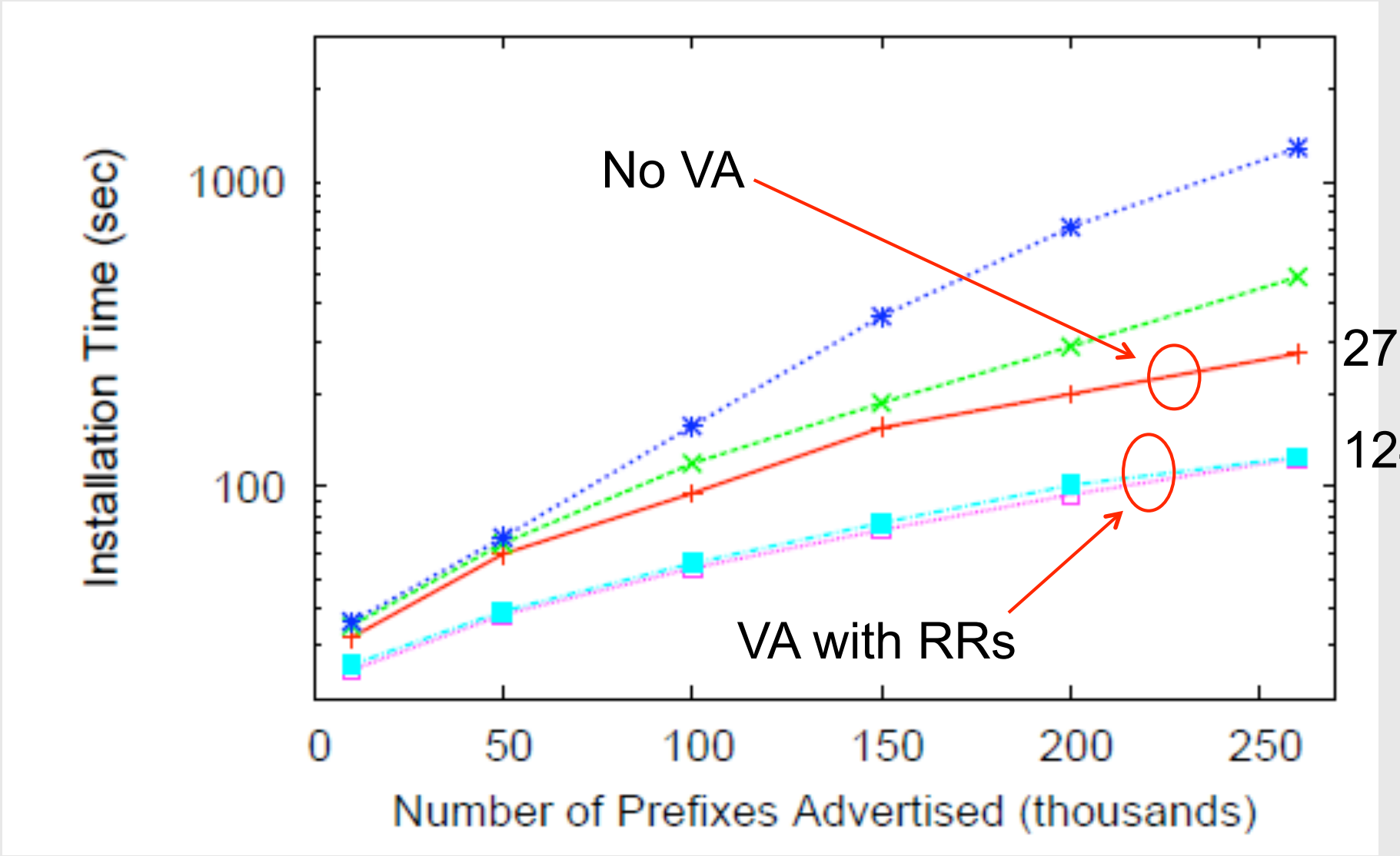
Assume 240K FIB entries (current routers)

Load and path-length over time



Roughly 50 % aggregating POPs

Load versus FIB size



Time to install full routing table

Is this a “real” solution???

Reduction is not log(N)

- Rather, we reduce slope of growth

RR’s still require full routing table, ISPs exchange full routing table

- Global convergence time and update frequency unchanged

Dependency on traffic matrix unfortunate

Current status and thinking

Router vendor (Huawei) is implementing VA natively

- Pushing in IETF
- http://tools.ietf.org/html/draft-francis-intra-va-00

Next: Use similar “divide and conquer” approach to shrink RIB size and processing

[F91]		"Efficient and Robust Policy Routing using Multiple Hierarchical Addresses," SIGCOMM 91
[FE93]	Tony Eng	“Extending the Internet through Address Reuse," SIGCOMM CCR 1993
[F94]		"Comparison of Geographical and Provider-rooted Internet Addressing," Computer Networks and ISDN Systems 27(3)437-448, 1994
[FG94]	Ramesh Govindan	"Flexible Routing and Addressing for a Next Generation IP," SIGCOMM 94
[GF01]	Ramakrishna Gummadi	“IPNL: A NAT-Extended Internet Architecture,” SIGCOMM 2001
[ZF06]	Joy Zhang, Jia Wang	“Scaling Global IP Routing with the Core Router-Integrated Overlay,“ ICMP 2006
[GF07]	Saikat Guha	"An End-Middle-End Approach to Connection Establishment," ACM SIGCOMM 2007
[BF08]	Hitesh Ballani, Tuan Cao, Jia Wang	“ViAggre: Making Routers Last Longer," ACM Hotnets 2008, NSDI 2009