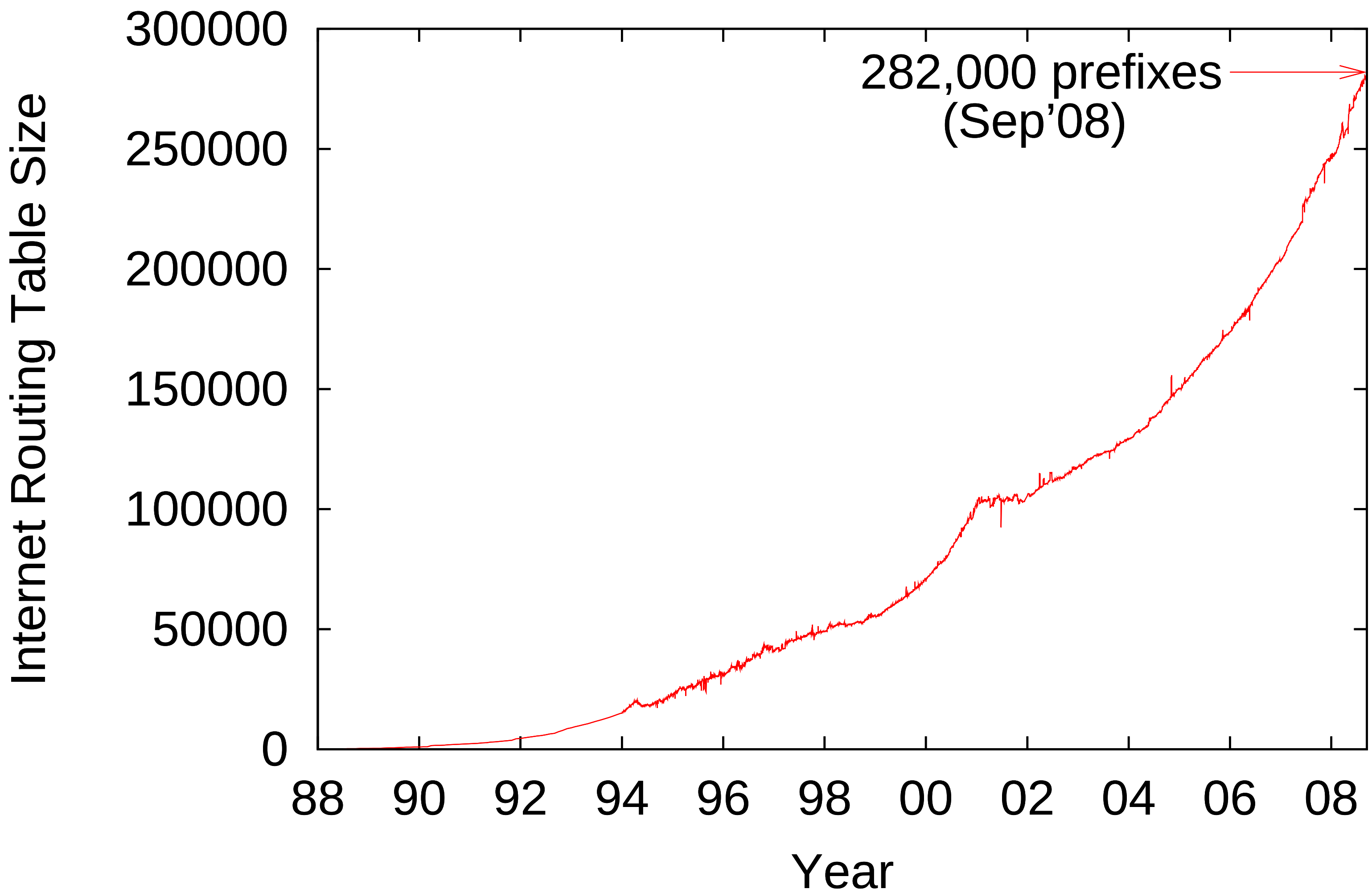
Making Routers Last Longer with ViAggre

Hitesh Ballani, Paul Francis, Tuan Cao and Jia Wang

Cornell University and AT&T Labs–Research

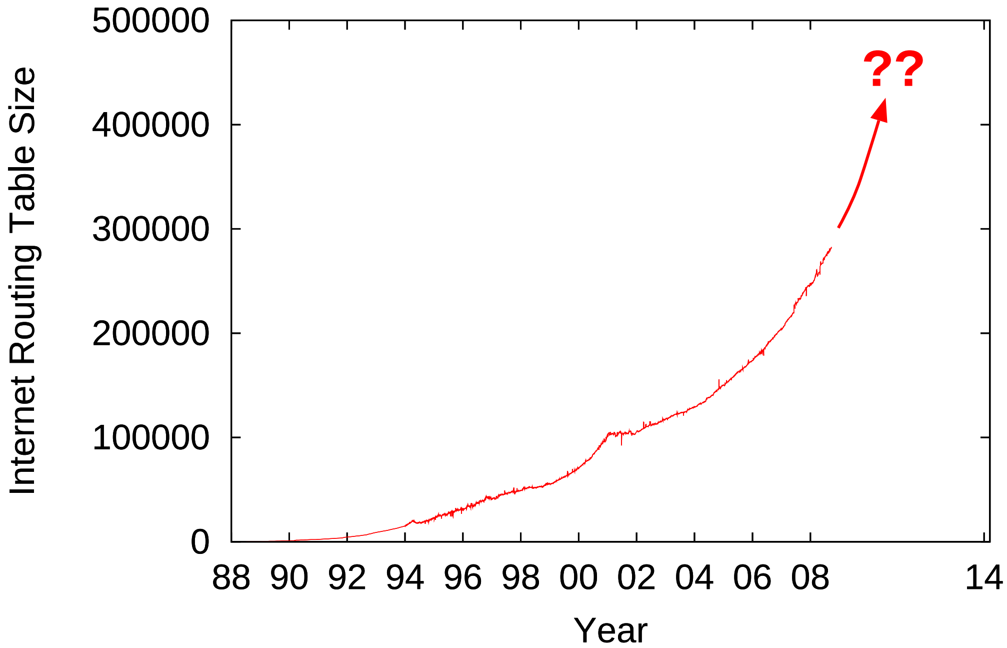
NSDI 2009

Motivation: Rapid Routing Table Growth



[Data Credit: Geoff Huston]

Motivation: Rapid Routing Table Growth



Rapid future growth

- IPv4 exhaustion
- IPv6 deployment

Routing Table stored in Forwarding Information Base (FIB) on Routers

Large Routing Table $\Rightarrow$ More FIB space on Routers

Does FIB Size Matter?

The problem is Scaling Properties of FIB memory (low volume, off-chip SRAM)

Technical concerns

- Power and Heat dissipation problems

Business concerns

- Low-volume, off-chip SRAM does not track Moore's law
- Larger routing table $\Rightarrow$ Less cost-effective networks
  - Price per byte forwarded increases
- Cost of router memory upgrades

Does FIB Size Matter?

Anecdotal evidence shows ISPs are willing to undergo some pain to extend the lifetime of their routers

Virtual Aggregation (ViAggre)

A “configuration-only” approach to shrinking router FIBs

- Applies to legacy routers
- Can be adopted independently by any ISP

Real World Impact

- IETF Standards effort
- Huawei implementing ViAggre into routers

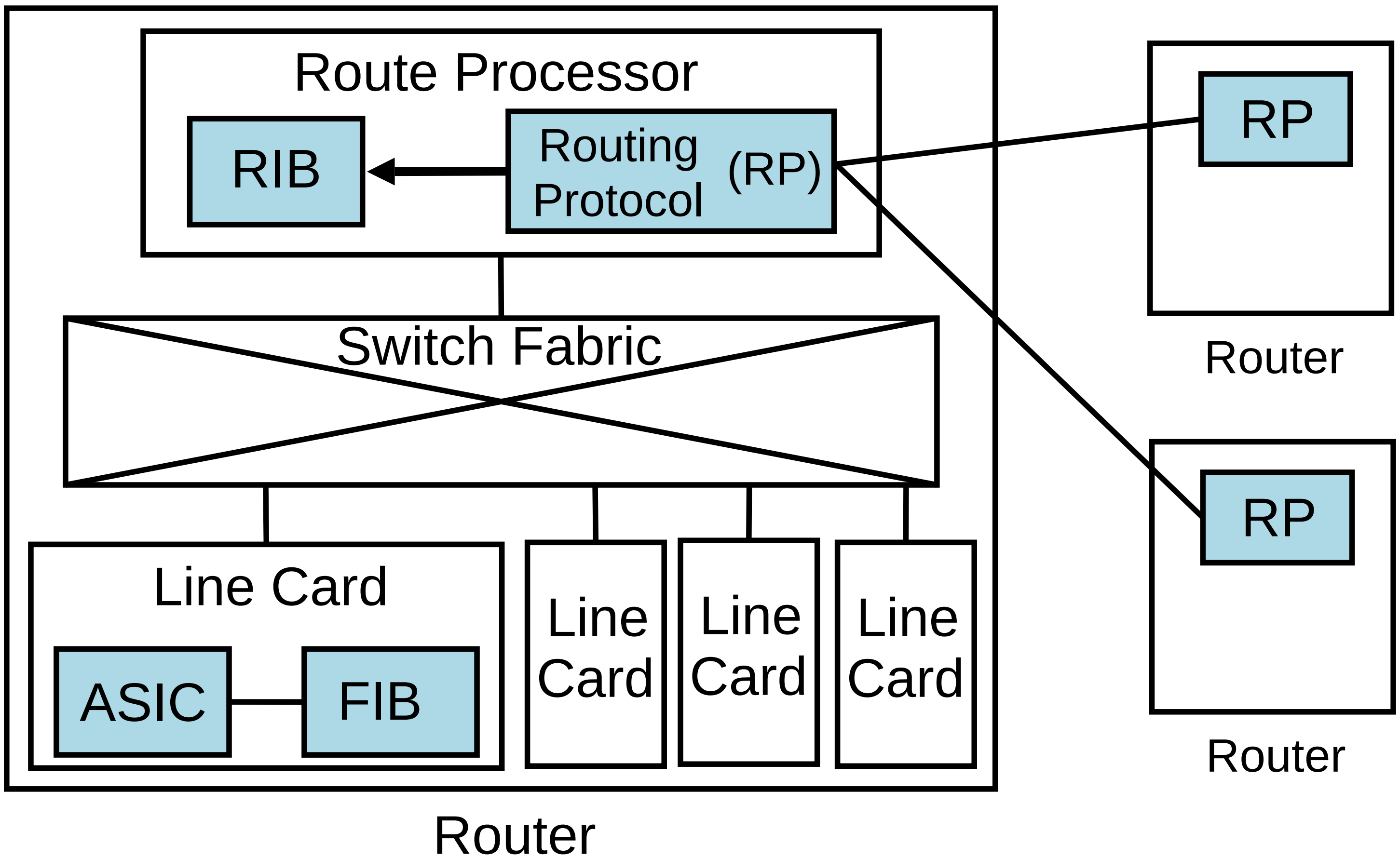
Key Insight: Divide the routing burden

A router only needs to keep routes for a fraction of the address space

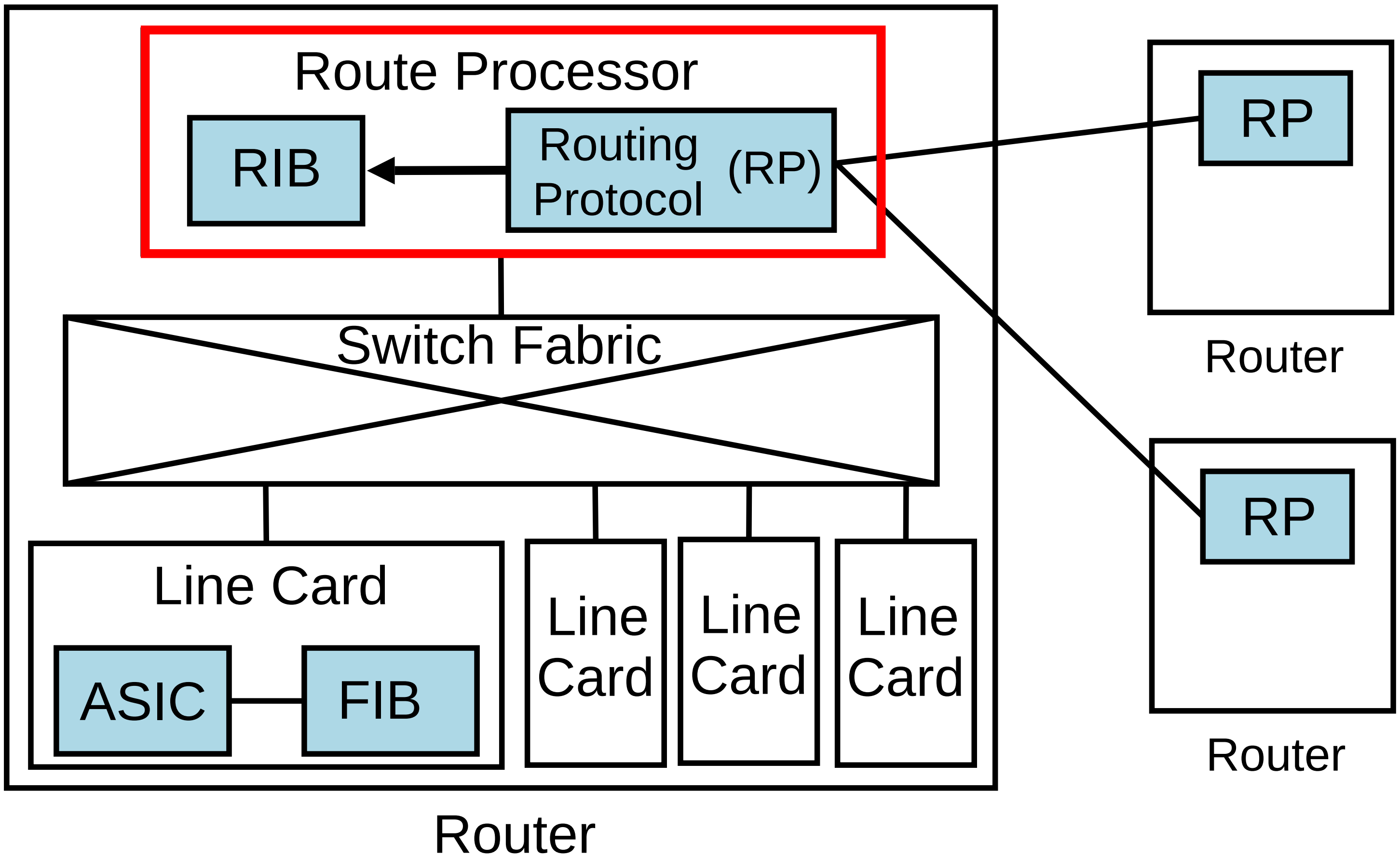
Talk Outline

- Motivation
- Router Innards
- Big Picture
- ViAggre Design
- Design Concerns
- Evaluation
- Deployment

Router Innards



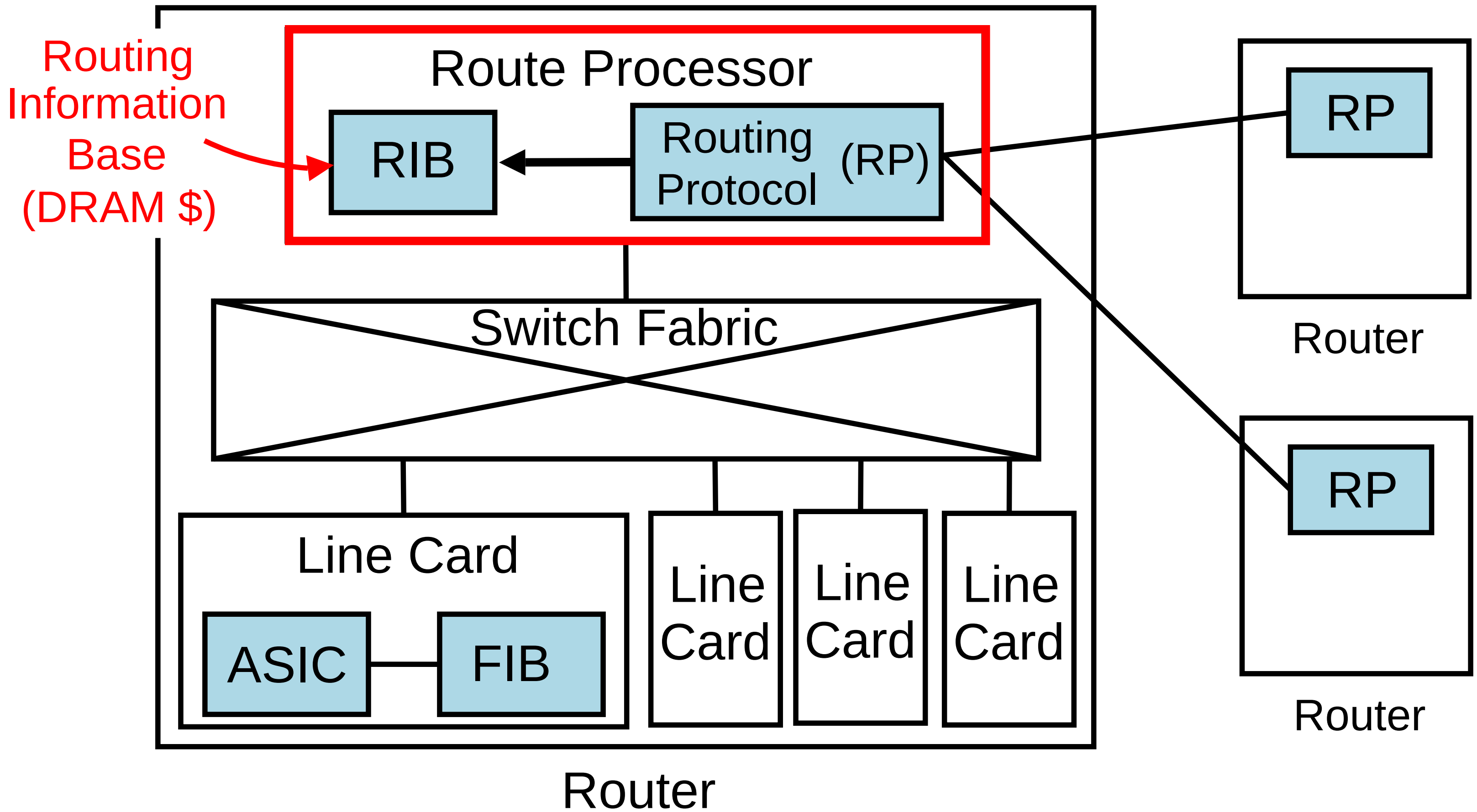
Router Innards



Control Plane

Participates in routing protocol

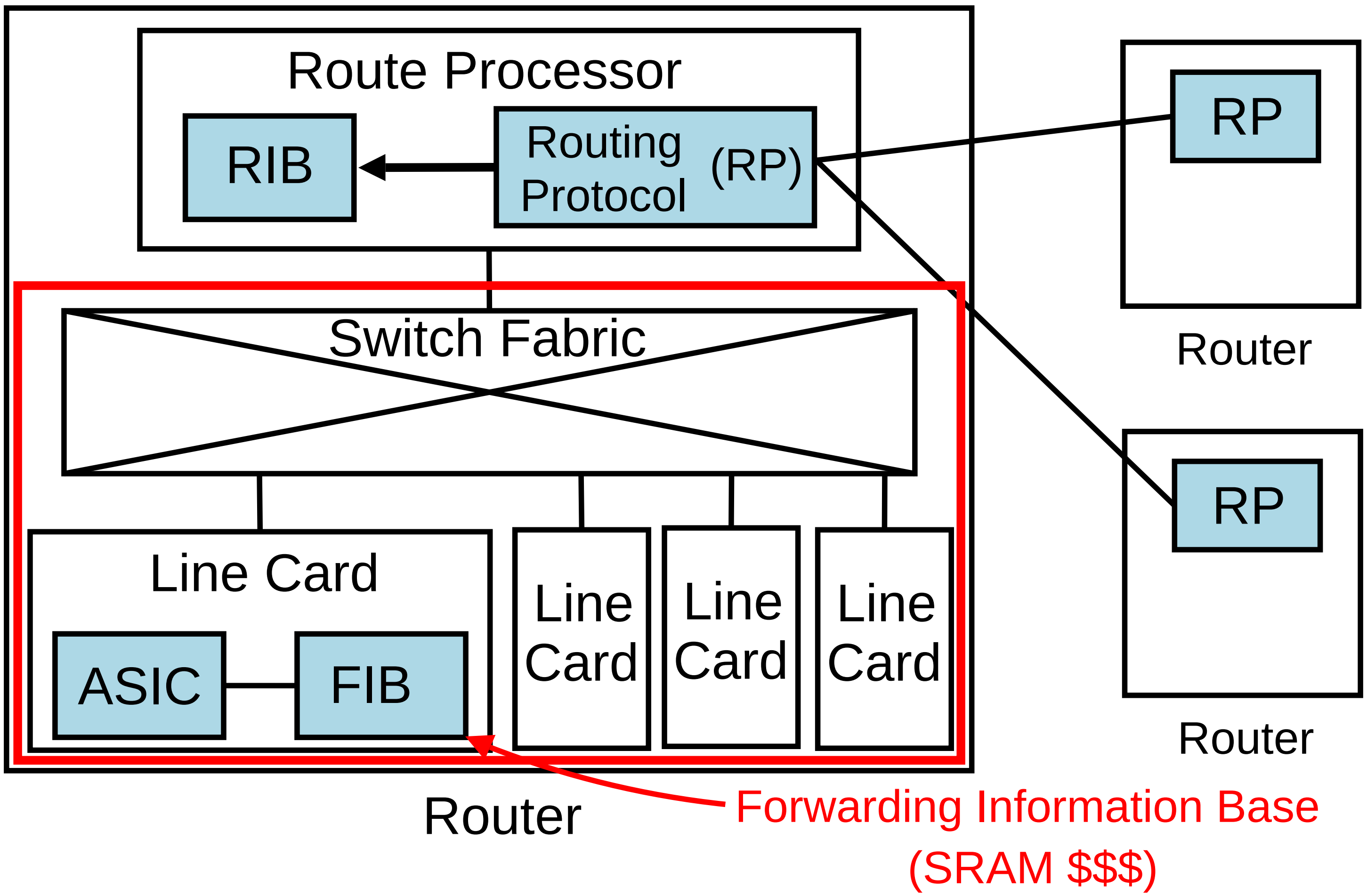
Router Innards



Control Plane

RIB is a table of routes and is stored on slow memory

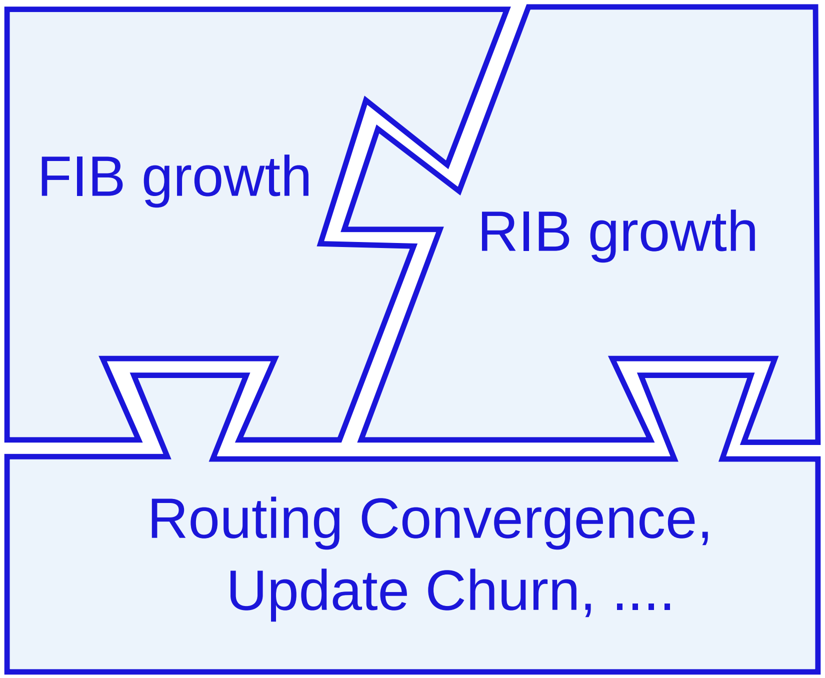
Router Innards



Data Plane

Responsible for sending packets based on FIB (stored in fast memory)

Routing Scalability Problem Space

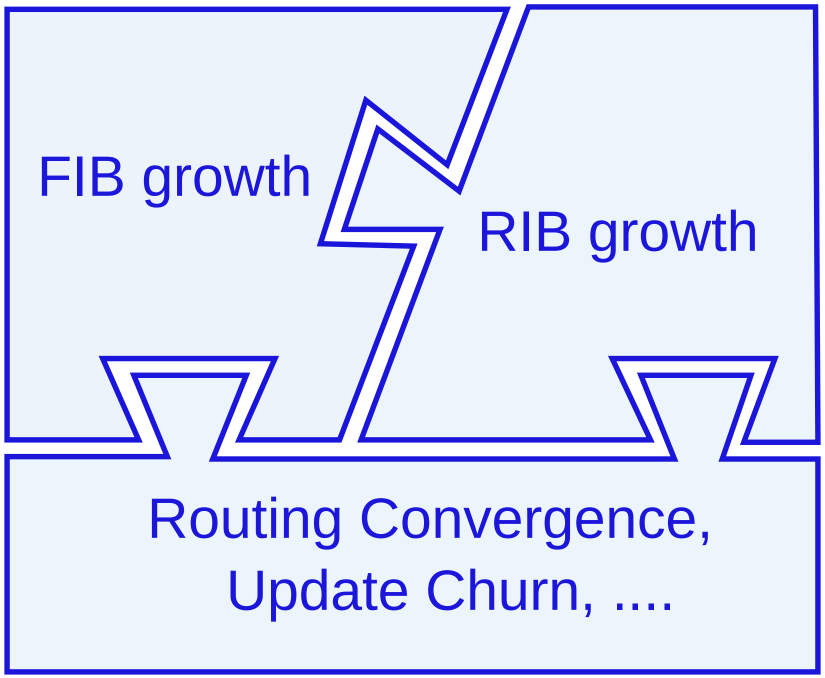


[MapEncap'96]
[GSE, ID'97]
[Atoms, '04]
[CRIO, ICNP'06]
[LISP, ID'07]
[SIRA, ID'07]
[TRRP, '07]
[APT, ID'07]
[Six/One, MobiArch'08]

[Francis, CNIS'94]
[Deering, ID'00]
[Hain, ID'02]
[Krioukov, Arxiv'05]
[Shim6, ID'07]
[Multipath, '08]

A few problems afflict Internet routing scalability
Lots of work to address these problems

Routing Scalability Problem Space

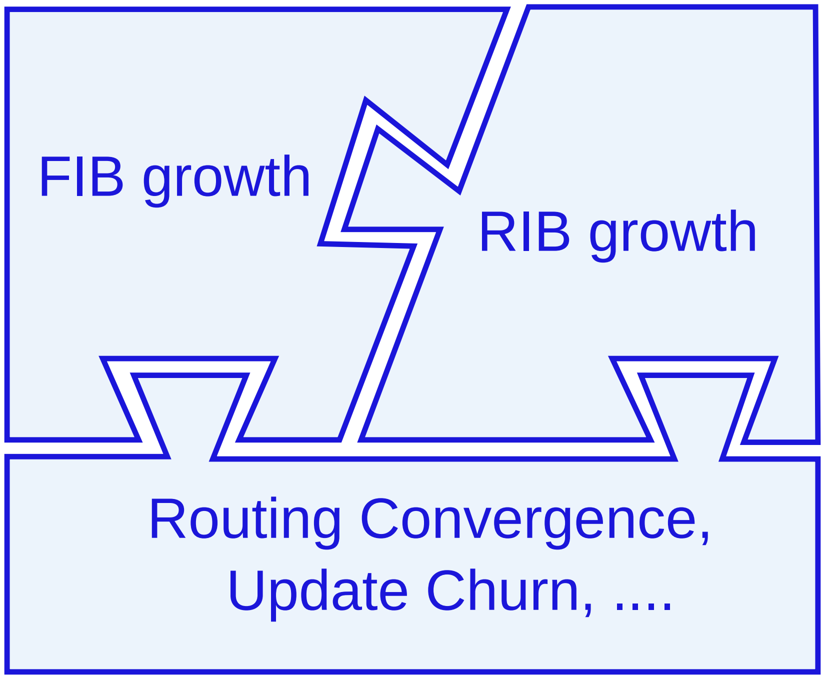


[MapEncap'96]
[GSE, ID'97]
[Atoms, '04]
[CRIO, ICNP'06]
[LISP, ID'07]
[SIRA, ID'07]
[TRRP, '07]
[APT, ID'07]
[Six/One, MobiArch'08]

[Francis, CNIS'94]
[Deering, ID'00]
[Hain, ID'02]
[Krioukov, Arxiv'05]
[Shim6, ID'07]
[Multipath, '08]

Separate edge from the core

Routing Scalability Problem Space



[MapEncap'96]
[GSE, ID'97]
[Atoms, '04]
[CRIO, ICNP'06]
[LISP, ID'07]
[SIRA, ID'07]
[TRRP, '07]
[APT, ID'07]
[Six/One, MobiArch'08]

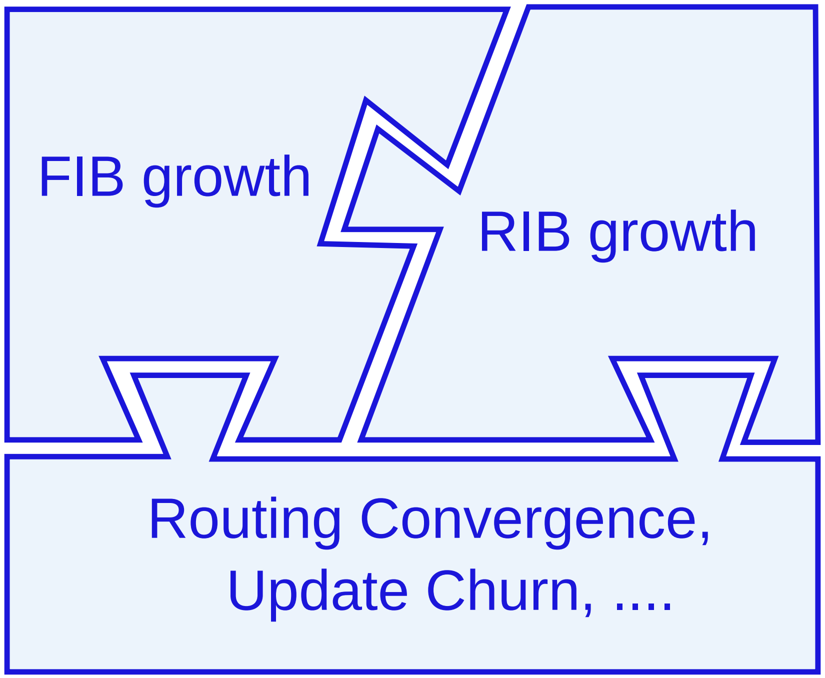
[Francis, CNIS'94]
[Deering, ID'00]
[Hain, ID'02]

[Krioukov, Arxiv'05]

[Shim6, ID'07]
[Multipath, '08]

Geographical routing

Routing Scalability Problem Space

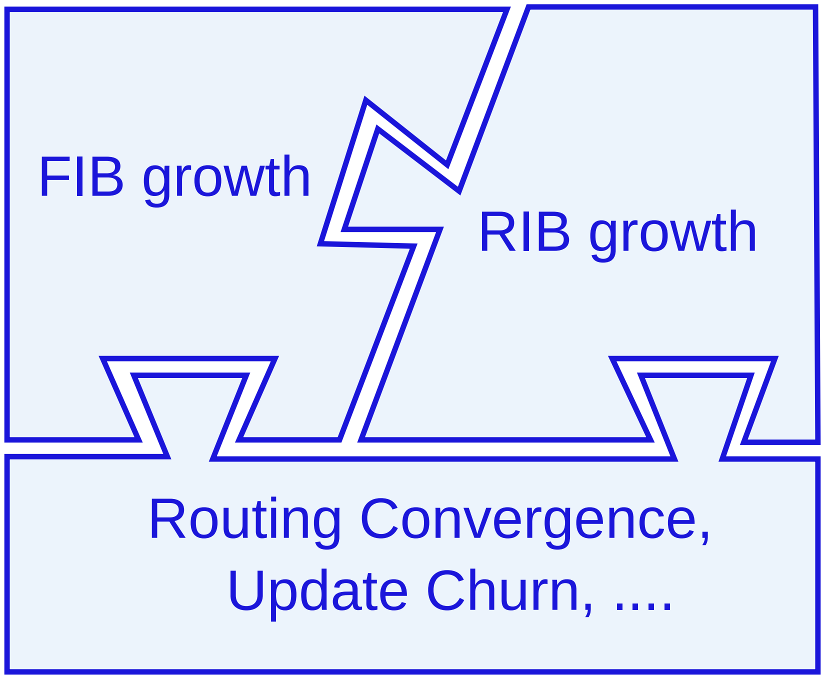


[MapEncap'96]
[GSE, ID'97]
[Atoms, '04]
[CRIO, ICNP'06]
[LISP, ID'07]
[SIRA, ID'07]
[TRRP, '07]
[APT, ID'07]
[Six/One, MobiArch'08]

[Francis, CNIS'94]
[Deering, ID'00]
[Hain, ID'02]
[Krioukov, Arxiv'05]
[Shim6, ID'07]
[Multipath, '08]

Compact routing

Routing Scalability Problem Space

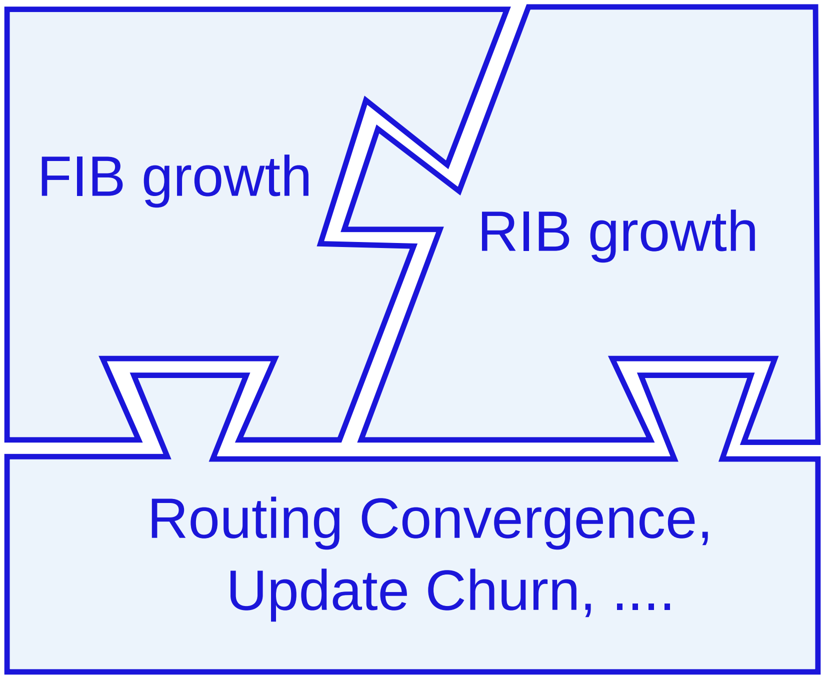


[MapEncap'96]
[GSE, ID'97]
[Atoms, '04]
[CRIO, ICNP'06]
[LISP, ID'07]
[SIRA, ID'07]
[TRRP, '07]
[APT, ID'07]
[Six/One, MobiArch'08]

[Francis, CNIS'94]
[Deering, ID'00]
[Hain, ID'02]
[Krioukov, Arxiv'05]
[Shim6, ID'07]
[Multipath, '08]

Elimination Approaches

Routing Scalability Problem Space



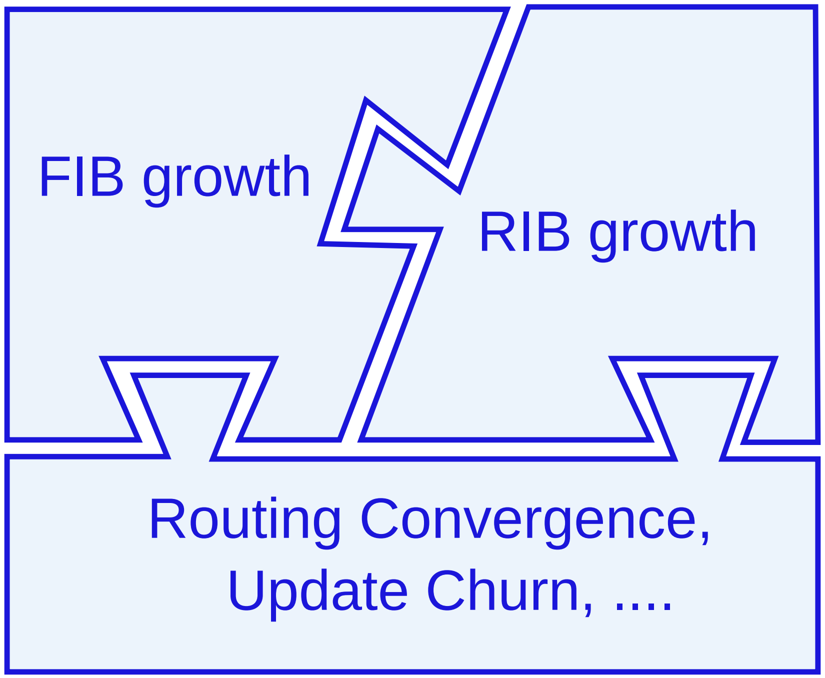
[MapEncap'96]
[GSE, ID'97]
[Atoms, '04]
[CRIO, ICNP'06]
[LISP, ID'07]
[SIRA, ID'07]
[TRRP, '07]
[APT, ID'07]
[Six/One, MobiArch'08]

[Francis, CNIS'94]
[Deering, ID'00]
[Hain, ID'02]
[Krioukov, Arxiv'05]
[Shim6, ID'07]
[Multipath, '08]

All require architectural change

So many good ideas, so little impact!

Routing Scalability Problem Space

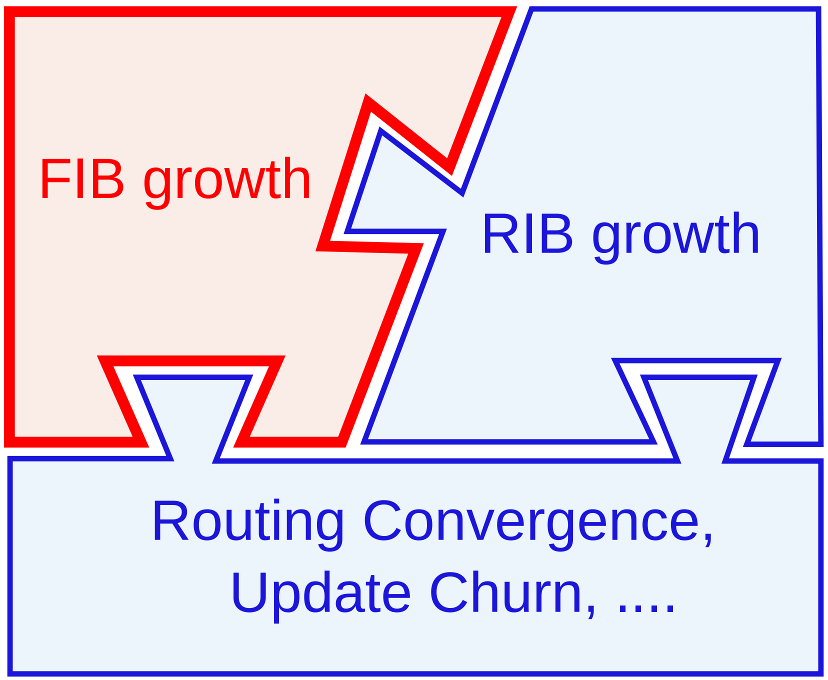


[MapEncap'96]
[GSE, ID'97]
[Atoms, '04]
[CRIO, ICNP'06]
[LISP, ID'07]
[SIRA, ID'07]
[TRRP, '07]
[APT, ID'07]
[Six/One, MobiArch'08]

[Francis, CNIS'94]
[Deering, ID'00]
[Hain, ID'02]
[Krioukov, Arxiv'05]
[Shim6, ID'07]
[Multipath, '08]

Can we devise an incremental solution by focusing on a subset of the problem space?

Routing Scalability Problem Space



[MapEncap'96]
[GSE, ID'97]
[Atoms, '04]
[CRIO, ICNP'06]
[LISP, ID'07]
[SIRA, ID'07]
[TRRP, '07]
[APT, ID'07]
[Six/One, MobiArch'08]

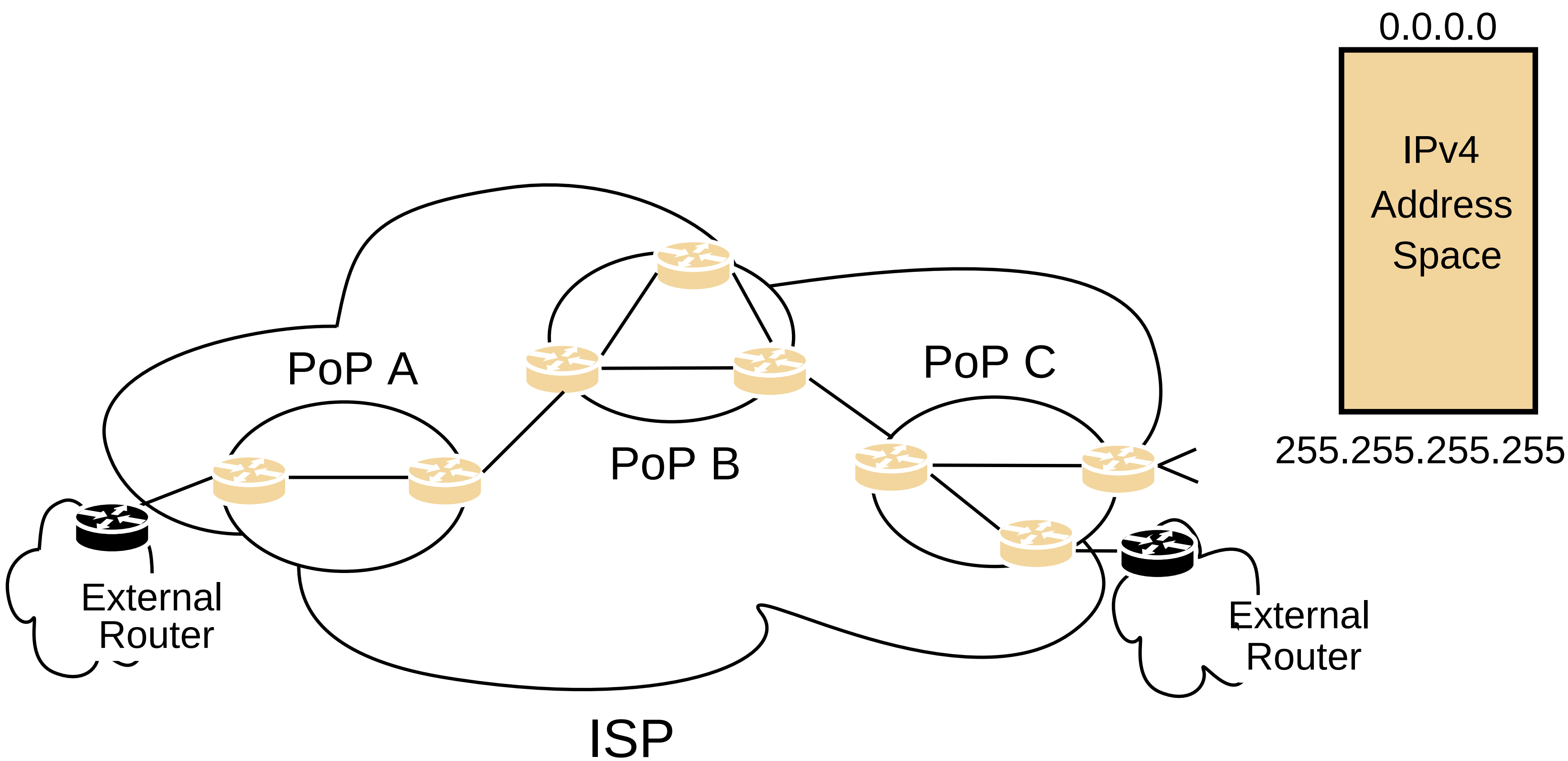
[Francis, CNIS'94]
[Deering, ID'00]
[Hain, ID'02]
[Krioukov, Arxiv'05]
[Shim6, ID'07]
[Multipath, '08]

This Talk: Focuses on reducing FIB size

Talk Outline

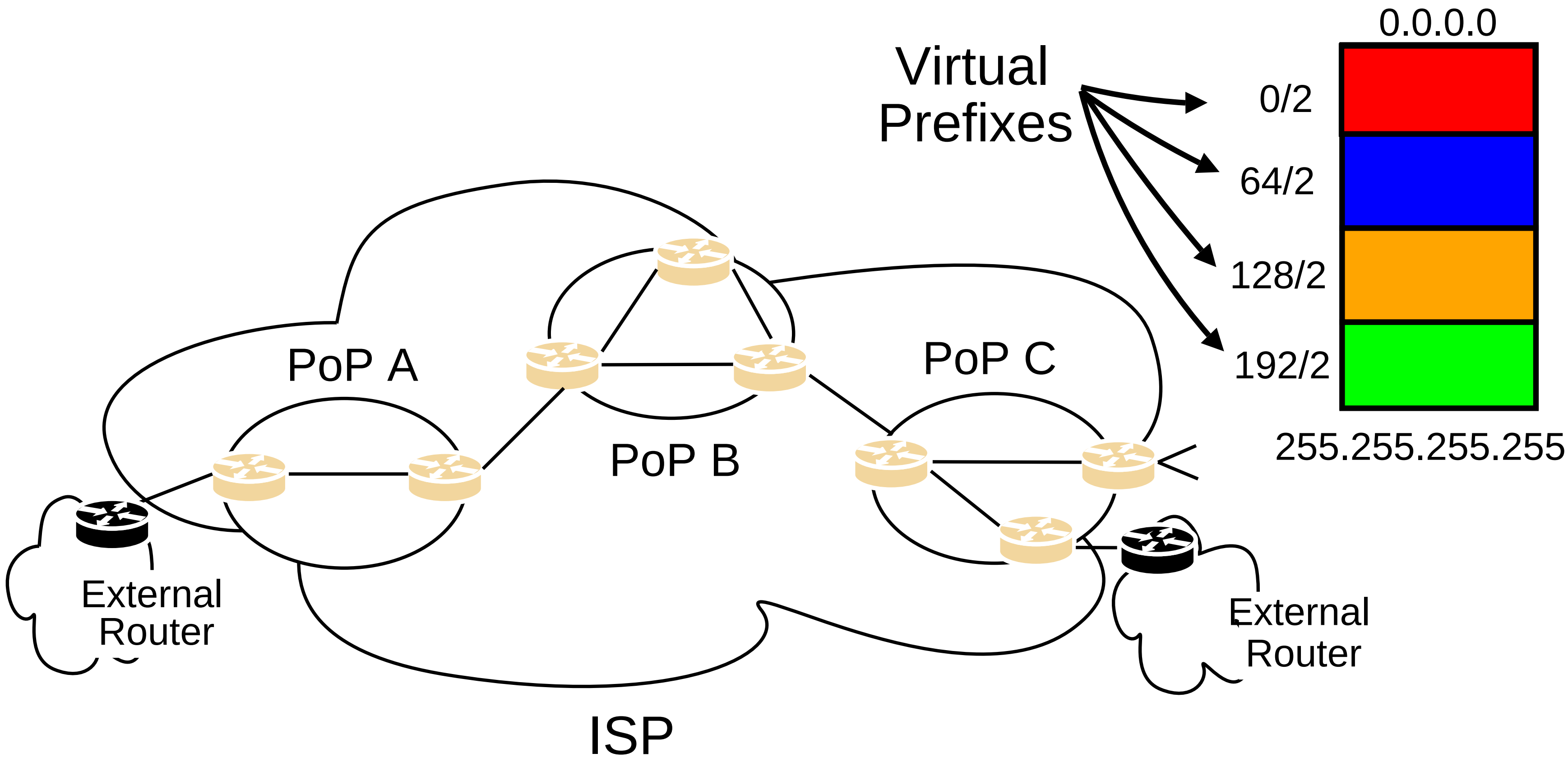
- Motivation
- Router Innards
- Big Picture
- ViAggre Design
- Design Concerns
- Evaluation
- Deployment

ViAggre: Basic Idea



Today: All routers have routes to all destinations

ViAggre: Basic Idea

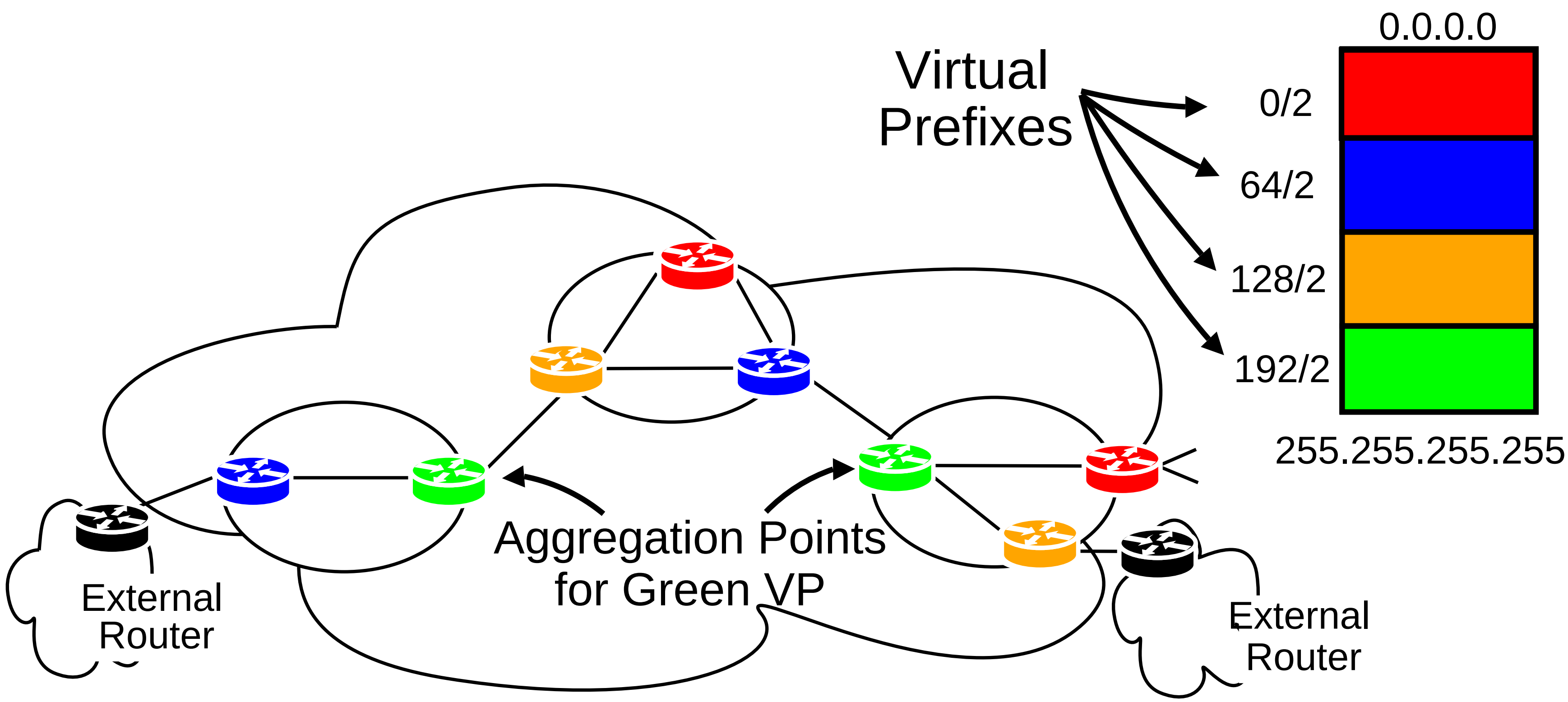


Divide address space into Virtual Prefixes (VPs)

Notation: "/2" implies that the first two bits are used to group IP addresses. "0/2" represents addresses starting with 00.

i.e. $0/2 \Rightarrow 0.0.0.0/2 \Rightarrow$ [0.0.0.0 to 63.255.255.255]

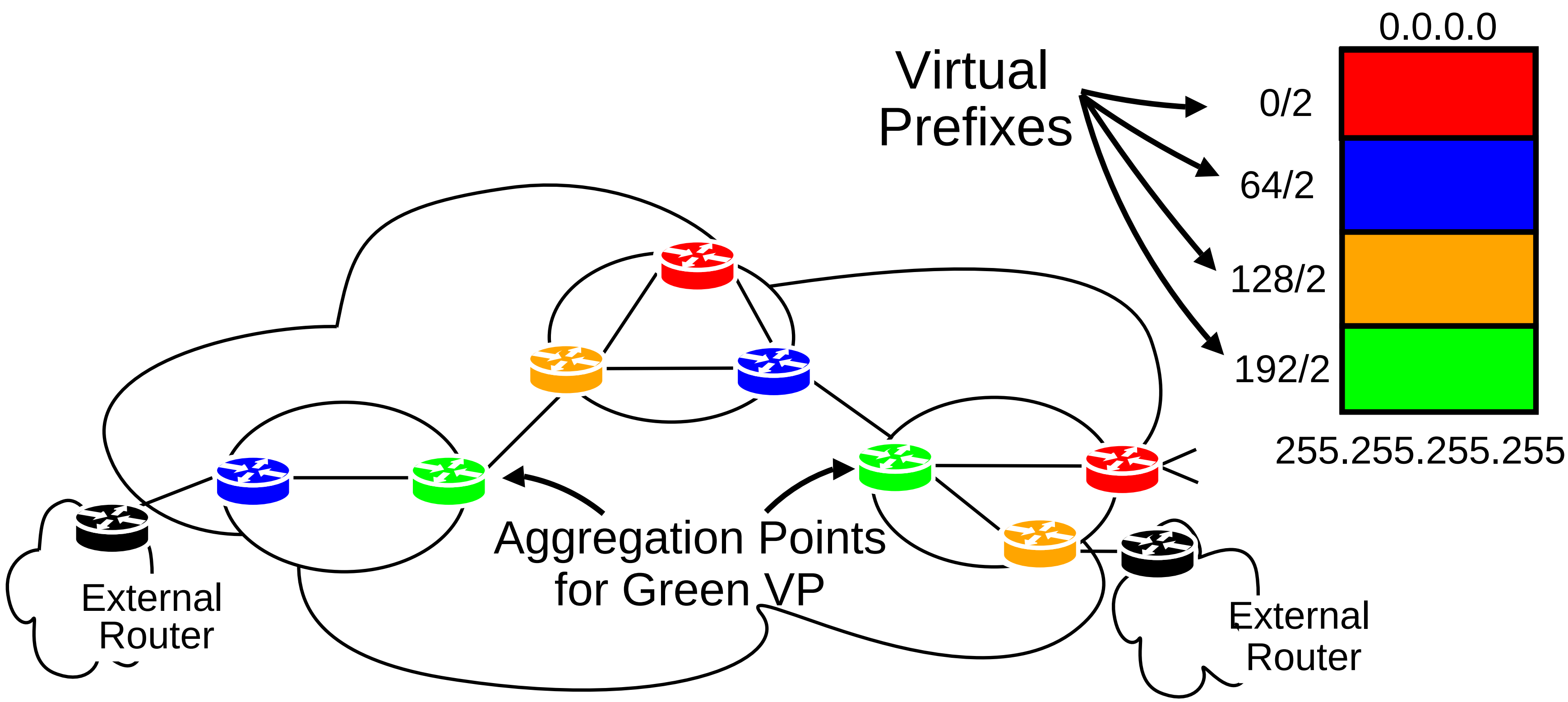
ViAggre: Basic Idea



Assign Virtual Prefixes to the routers

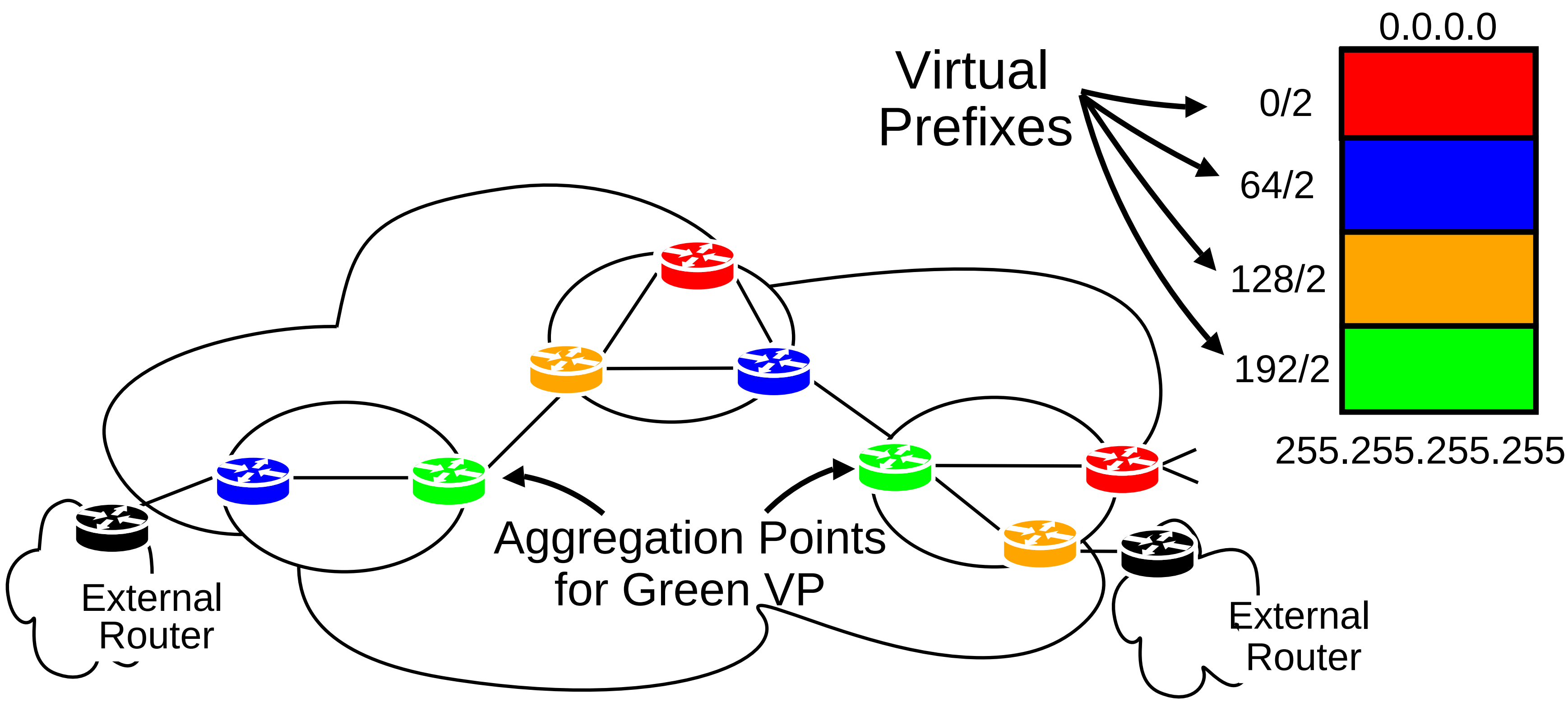
Green Aggregation Points maintain routes to green prefixes

ViAggre: Basic Idea



Routers only have routes to a fraction of the address space

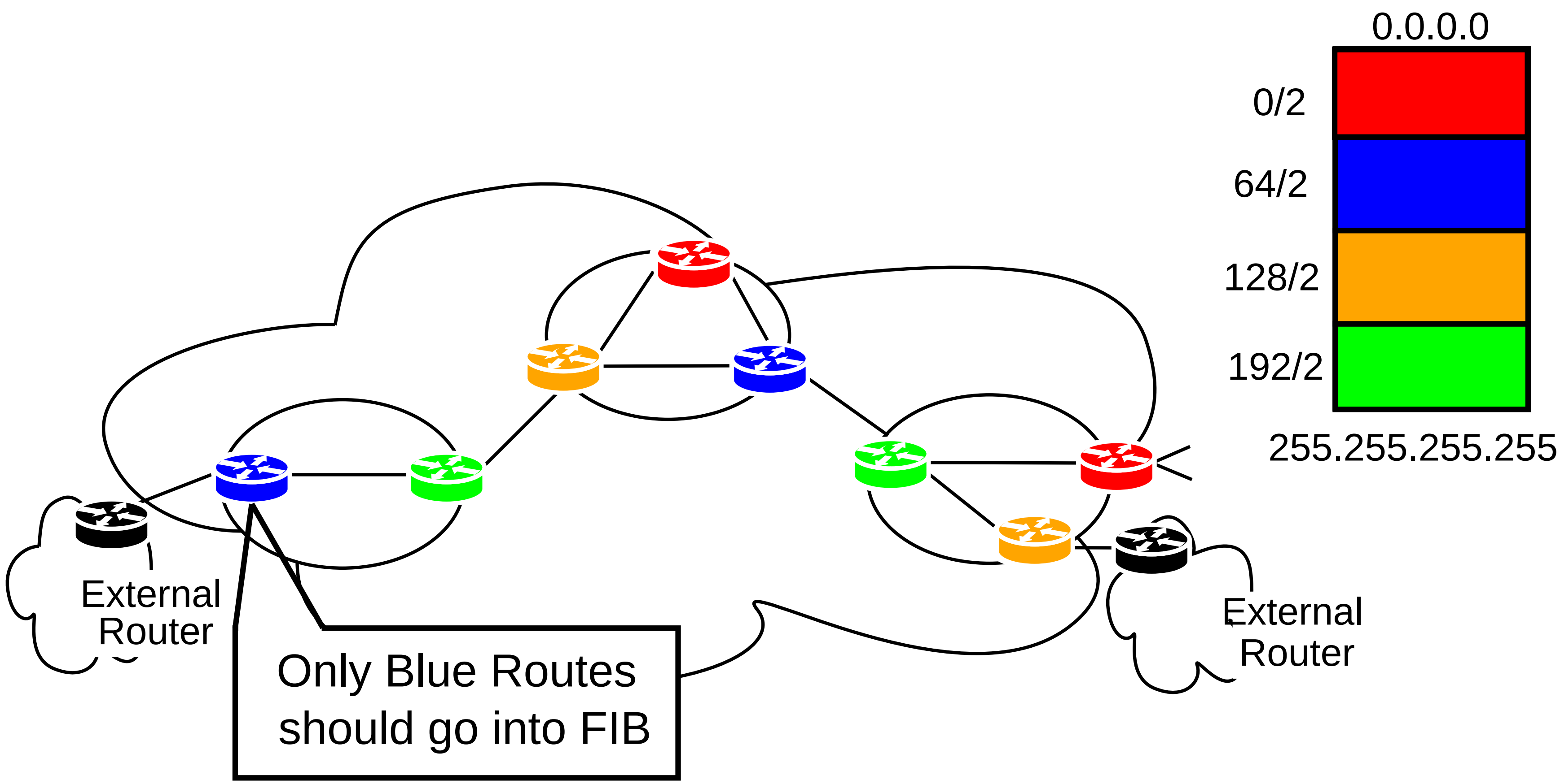
ViAggre: Basic Idea



1. How to achieve such division of the routing table without changes to routers and external cooperation?

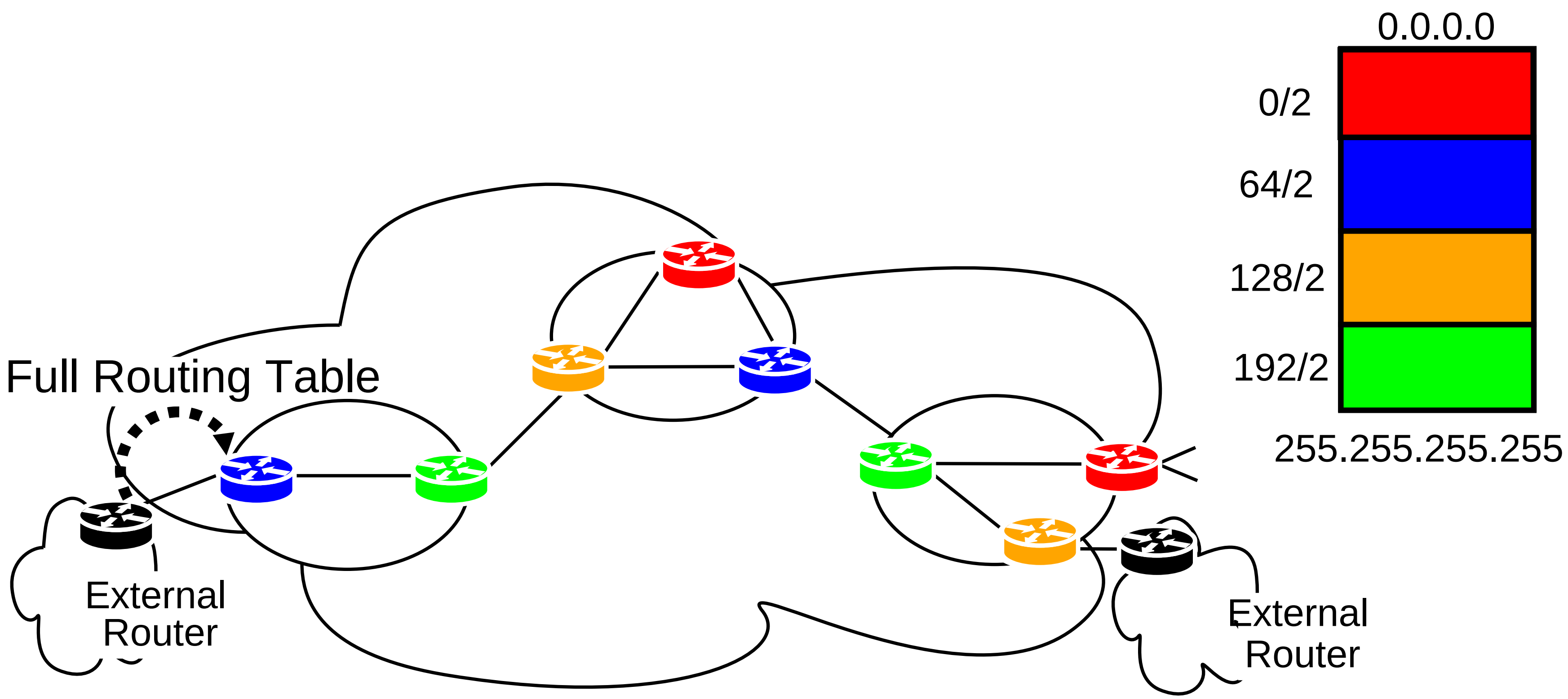
2. How do packets traverse even though routers have partial routing tables?

ViAggre Control-Plane



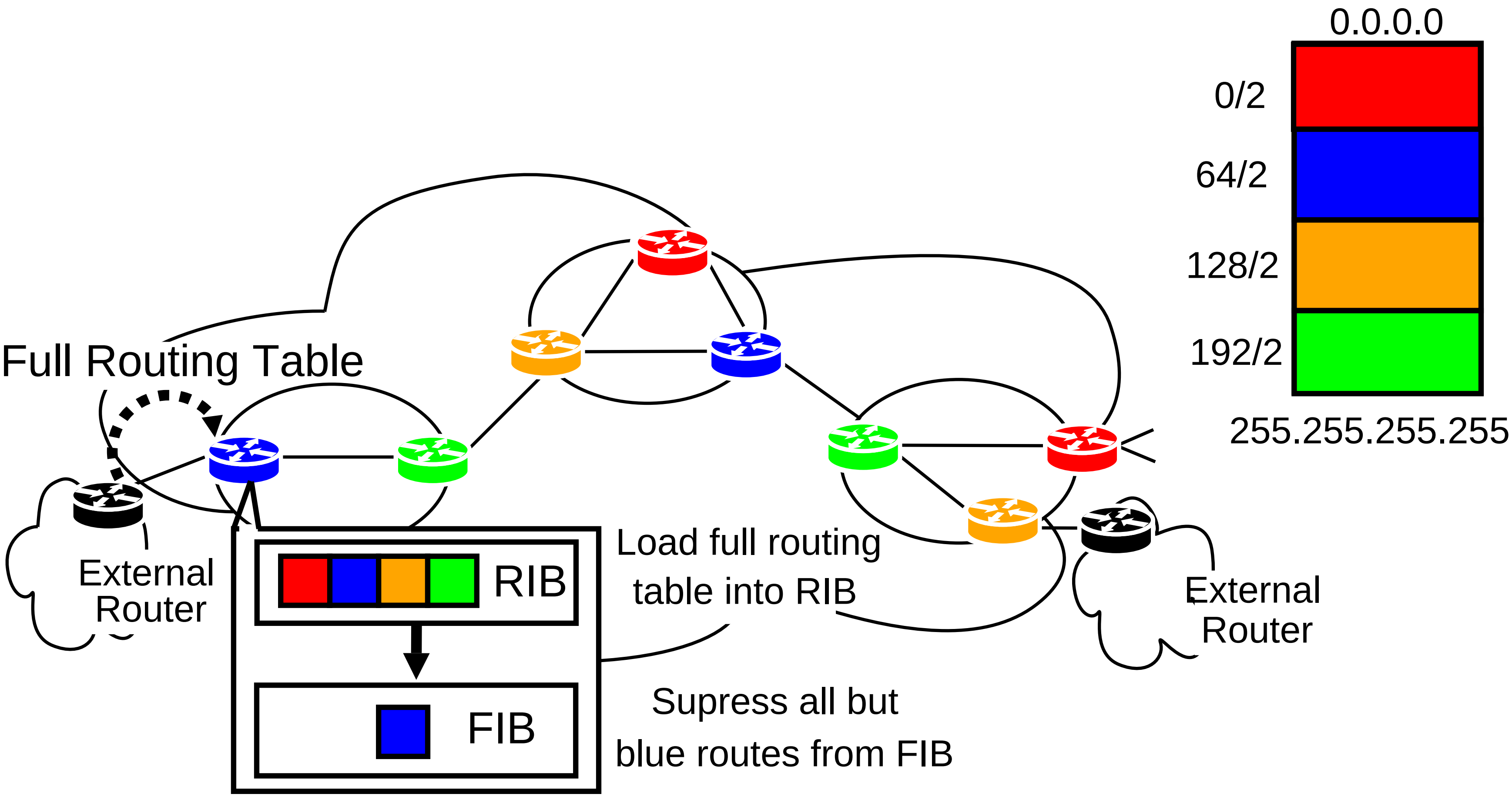
Control-plane needs to ensure that a router's FIB only contains routes that the router is aggregating

ViAggre Control-Plane



External BGP Peers may advertise full routing table

ViAggre Control-Plane

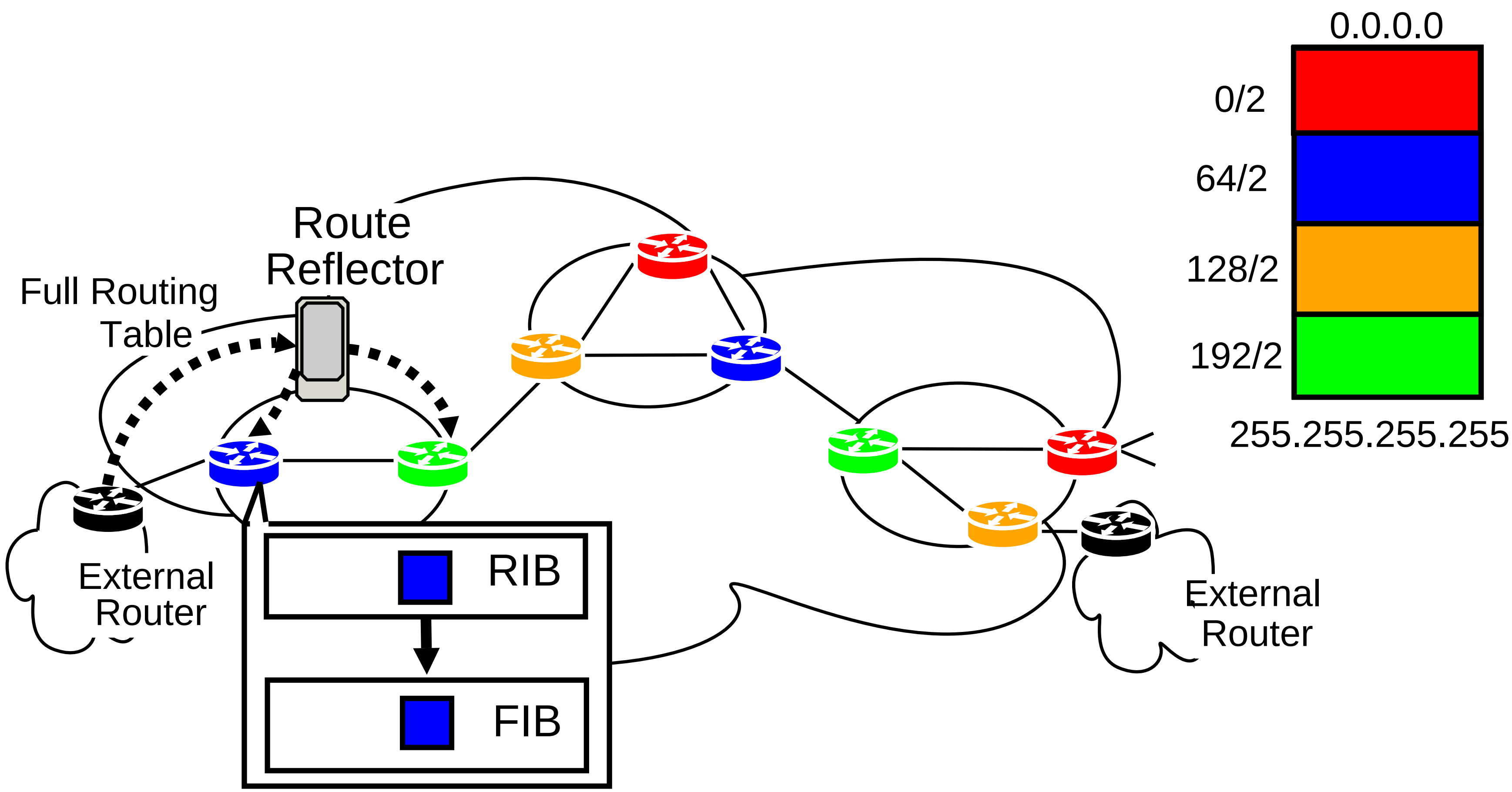


Simple Approach: FIB Suppression

Routers can load a subset of the RIB into their FIB

High Performance Overhead

ViAggre Control-Plane

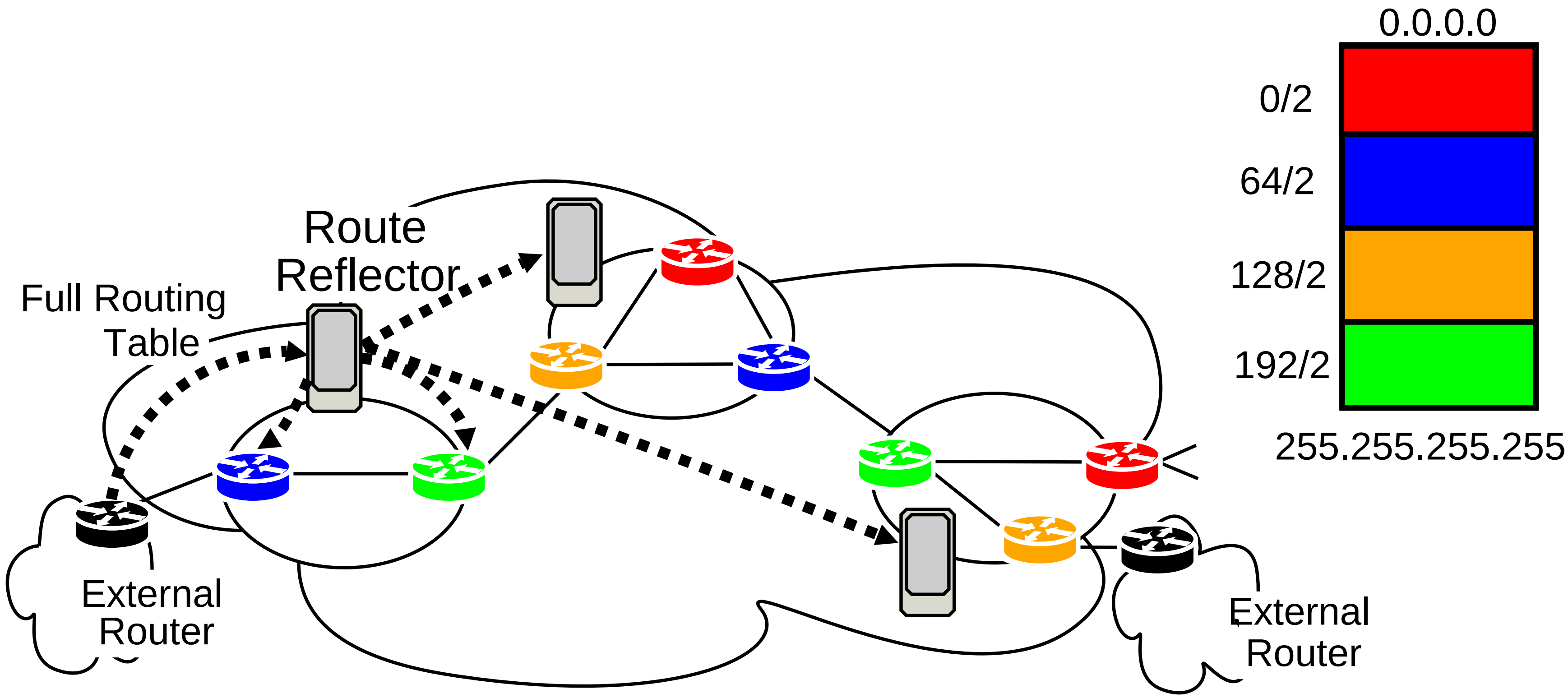


Practical Approach: Route-reflector Suppression

External router peers with a route-reflector

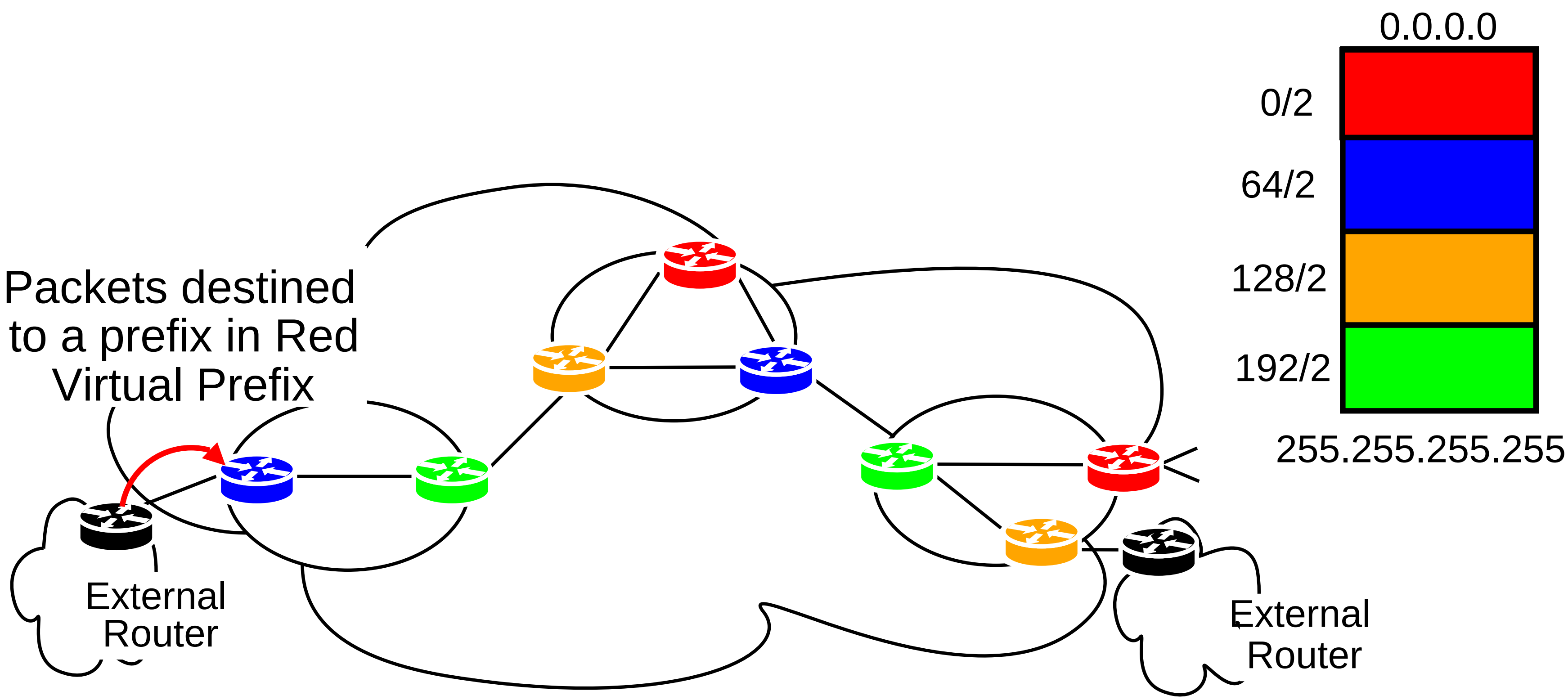
Blue router receives only blue routes

ViAggre Control-Plane



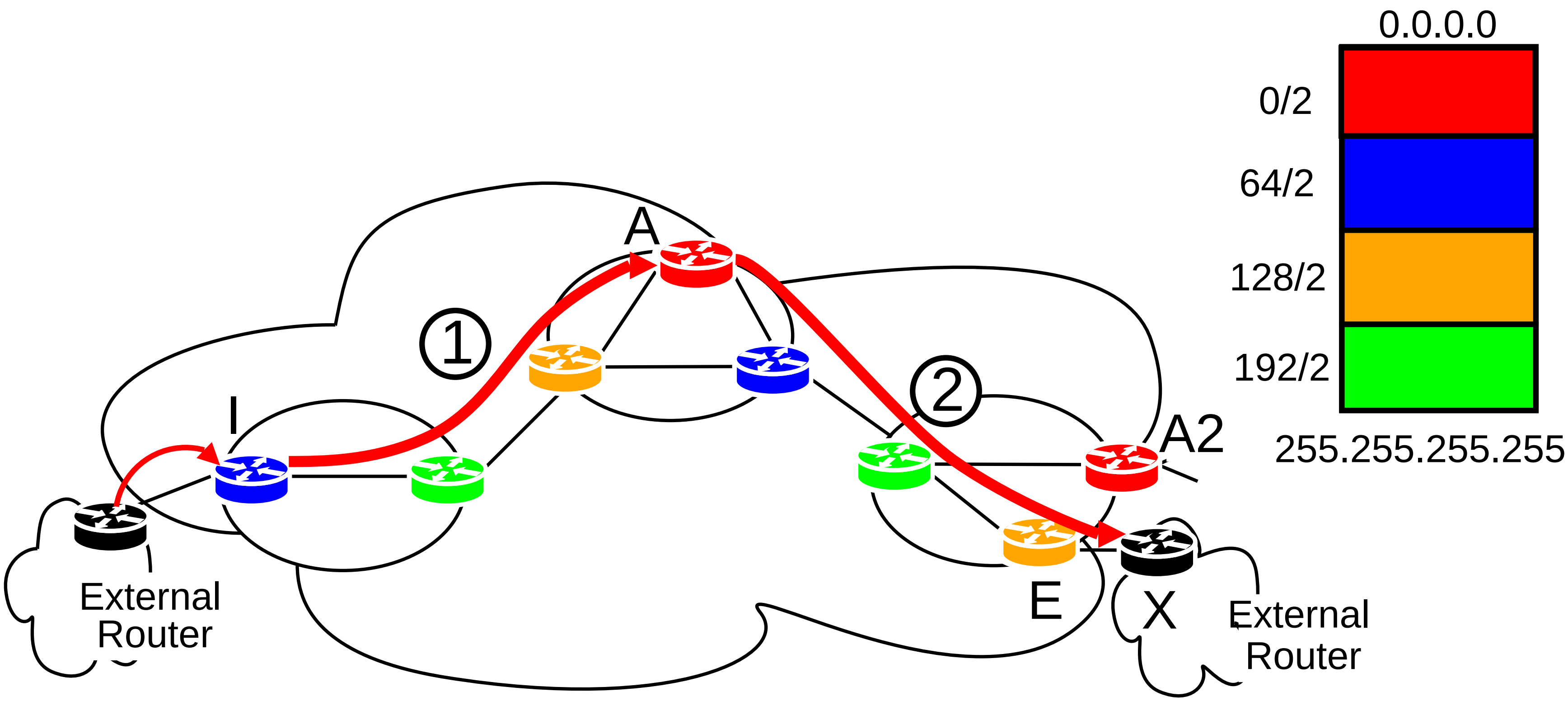
Practical Approach: Route-reflector Suppression
Route-reflectors exchange routes with each other

Data-Plane paths



Consider packets destined to a prefix in the red VP

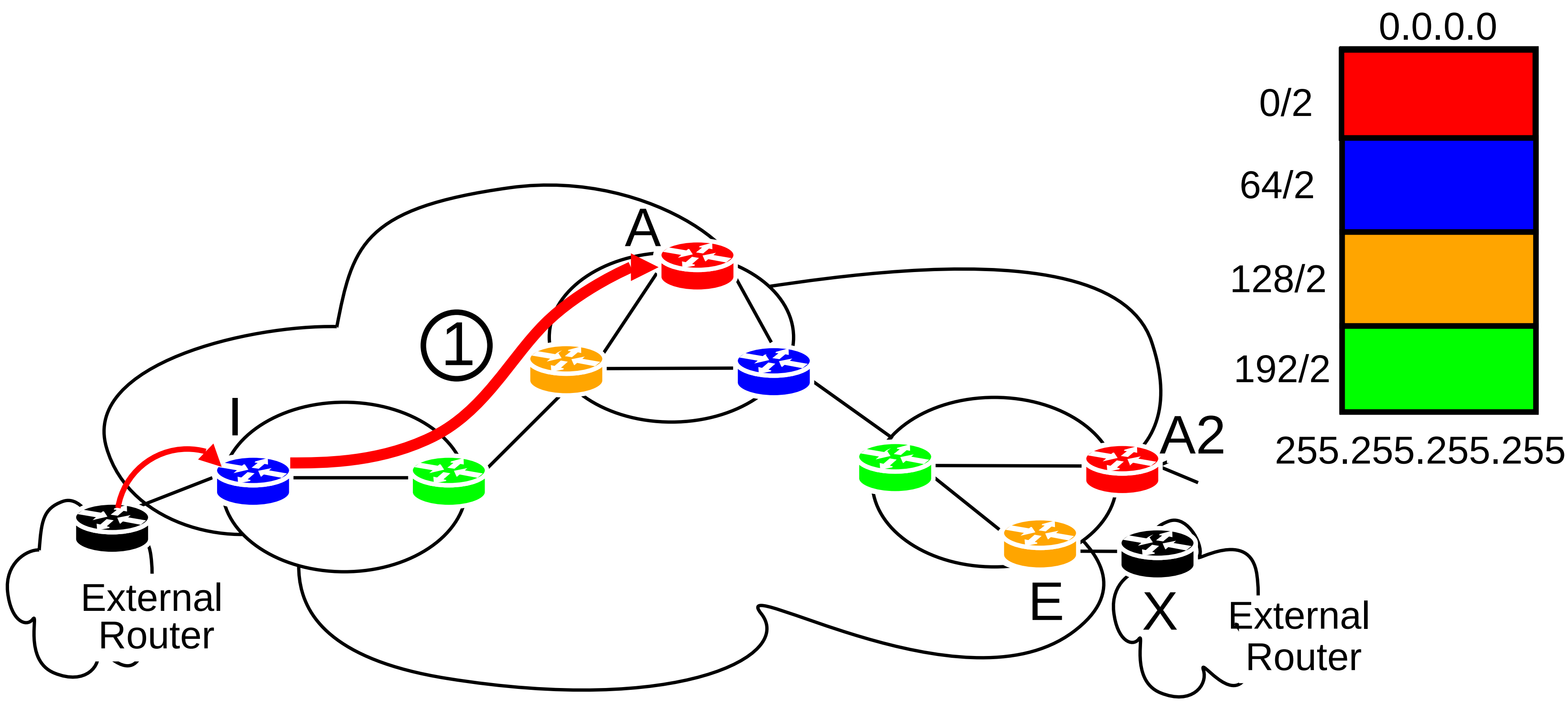
Data-Plane paths



ViAggre path

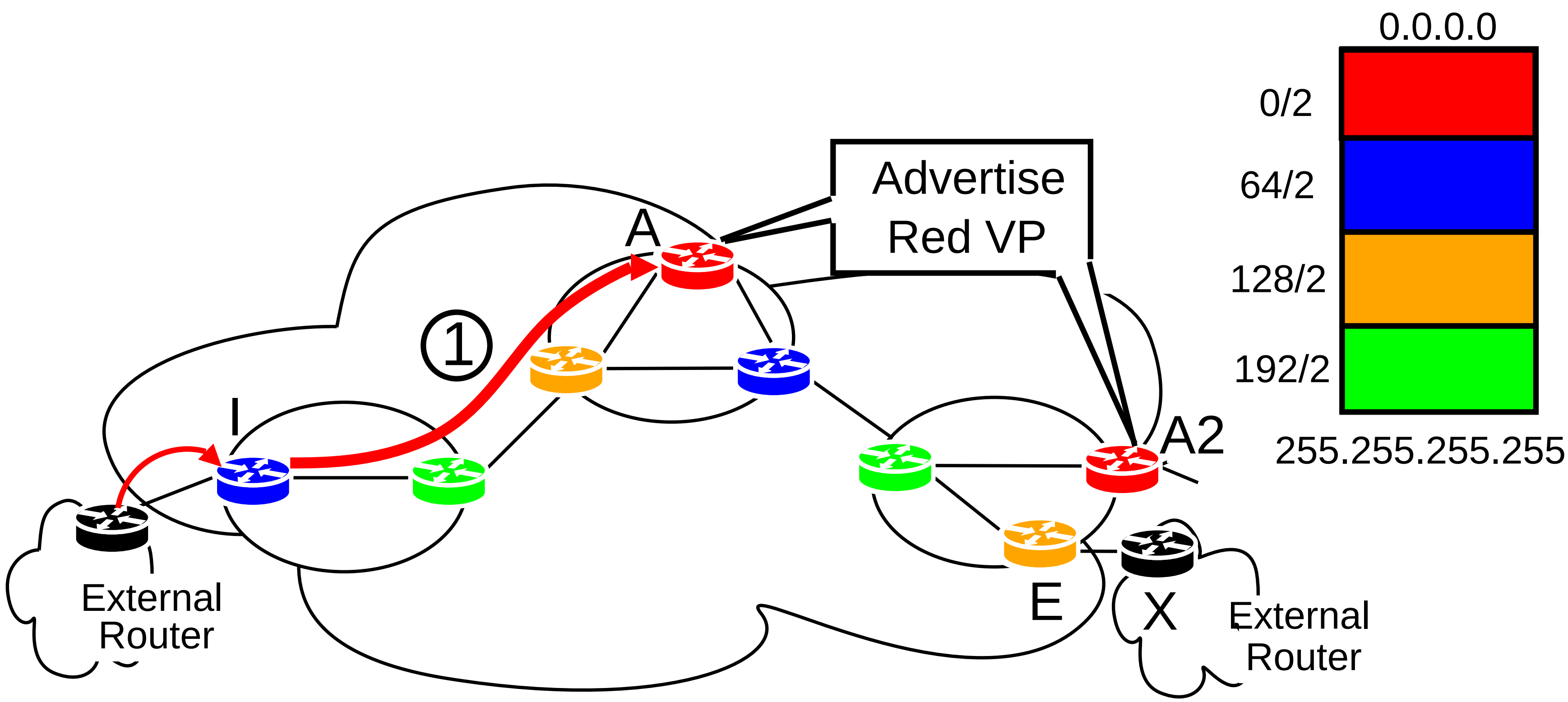
Ingress (I) $\longrightarrow$ Aggregation Pt (A) $\longrightarrow$ Egress (E)

Ingress → Aggregation Point



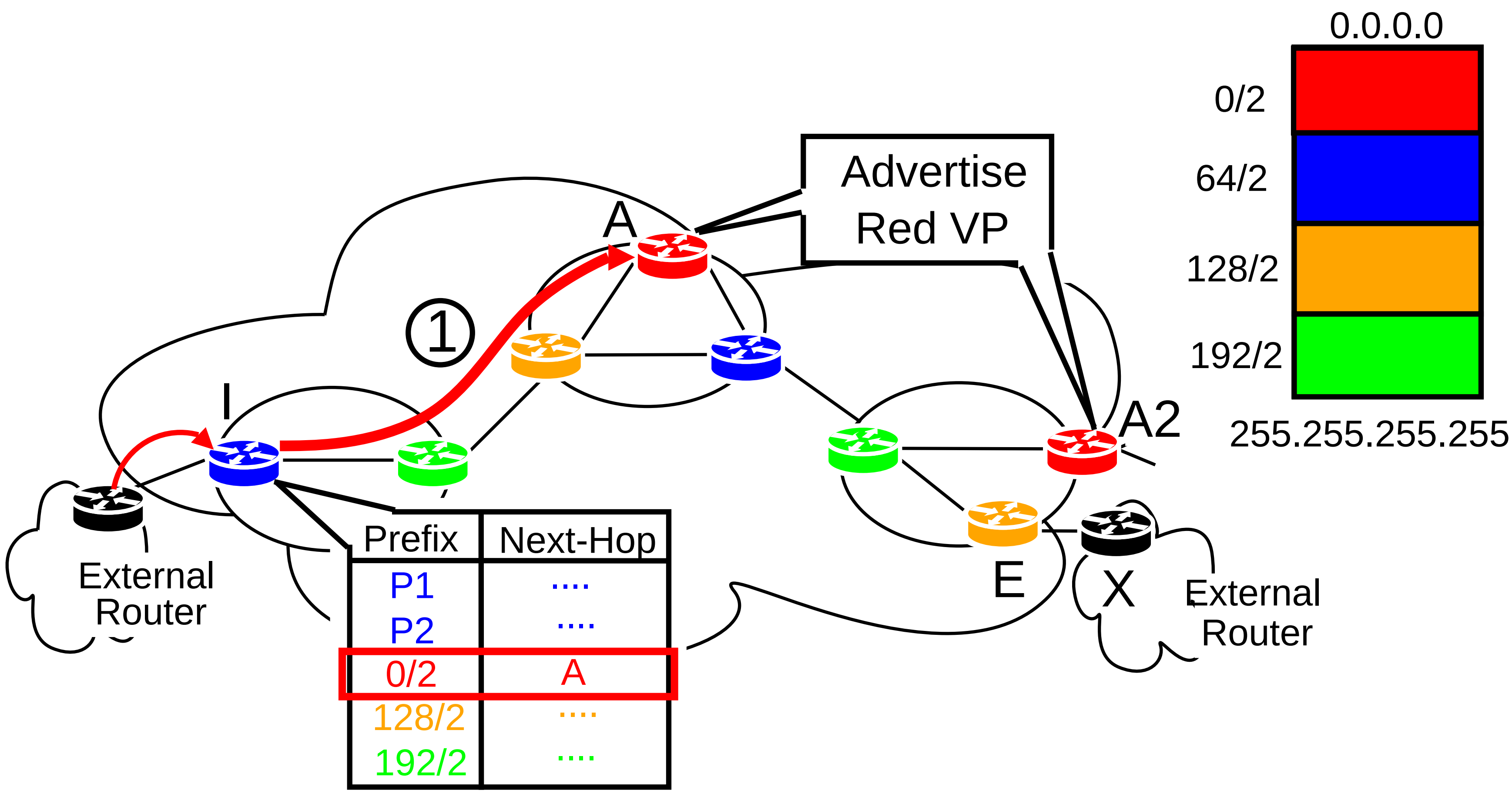
Router I doesn't have a route for destination prefix

Ingress $\longrightarrow$ Aggregation Point



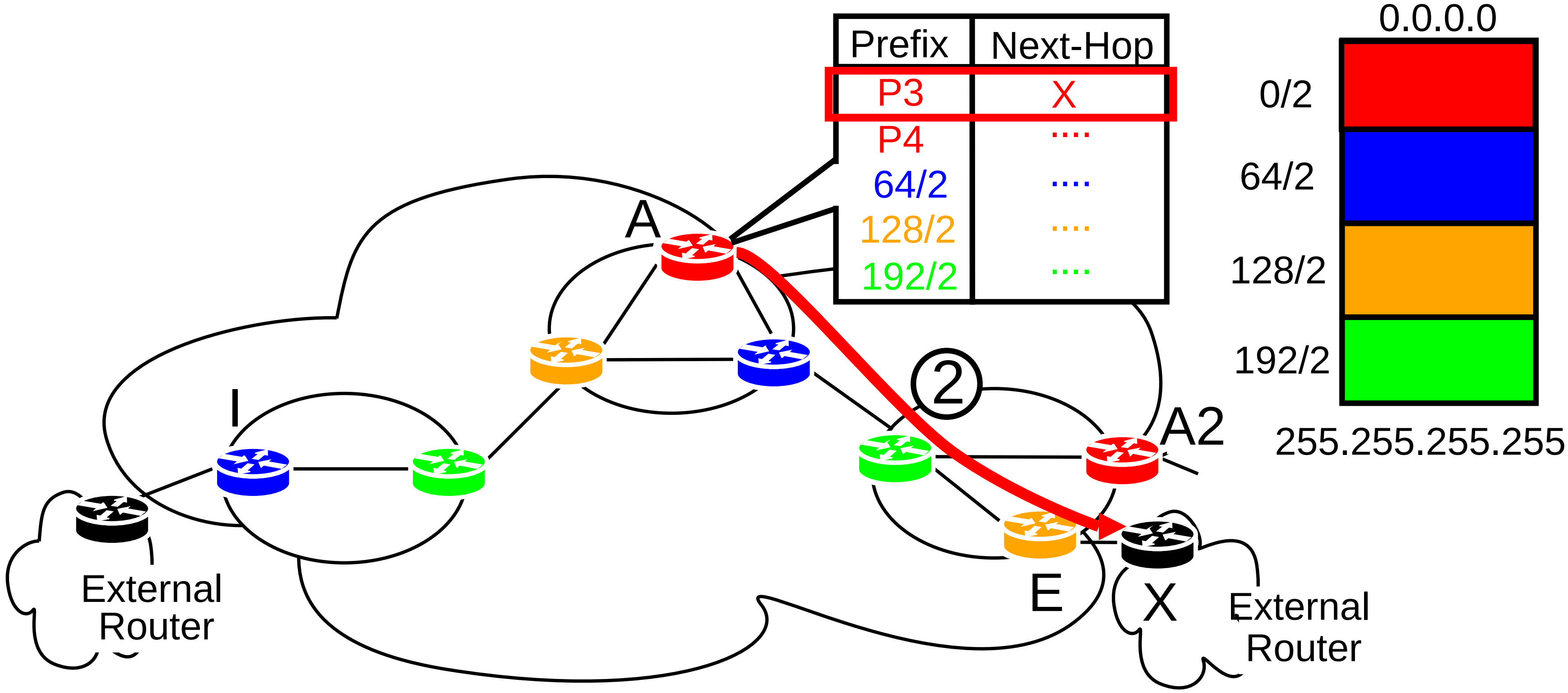
Aggregation Points advertise corresponding Virtual Prefixes

Ingress → Aggregation Point



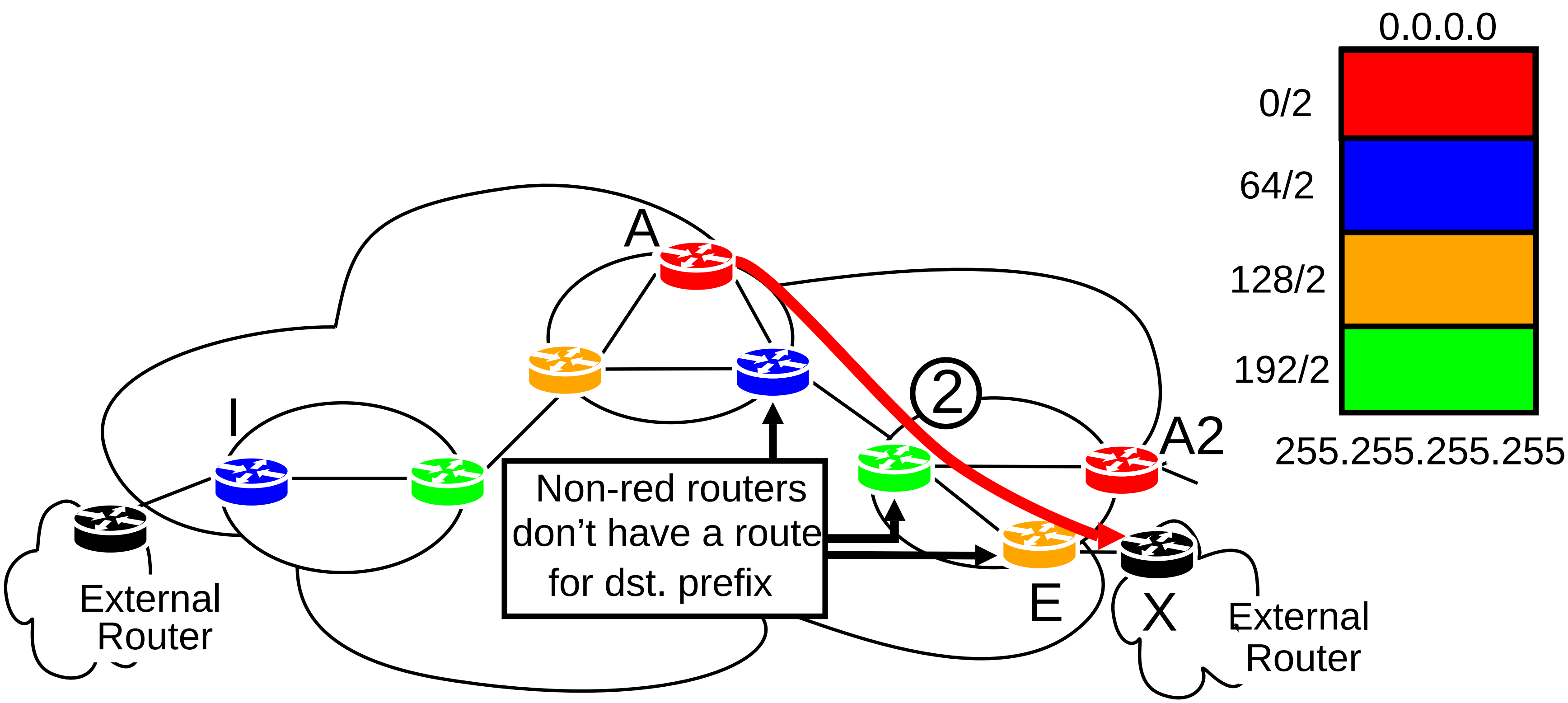
Blue router has a route for the red Virtual Prefix

Aggregation Point → Egress



Aggregation Pt. A has a route for destination prefix

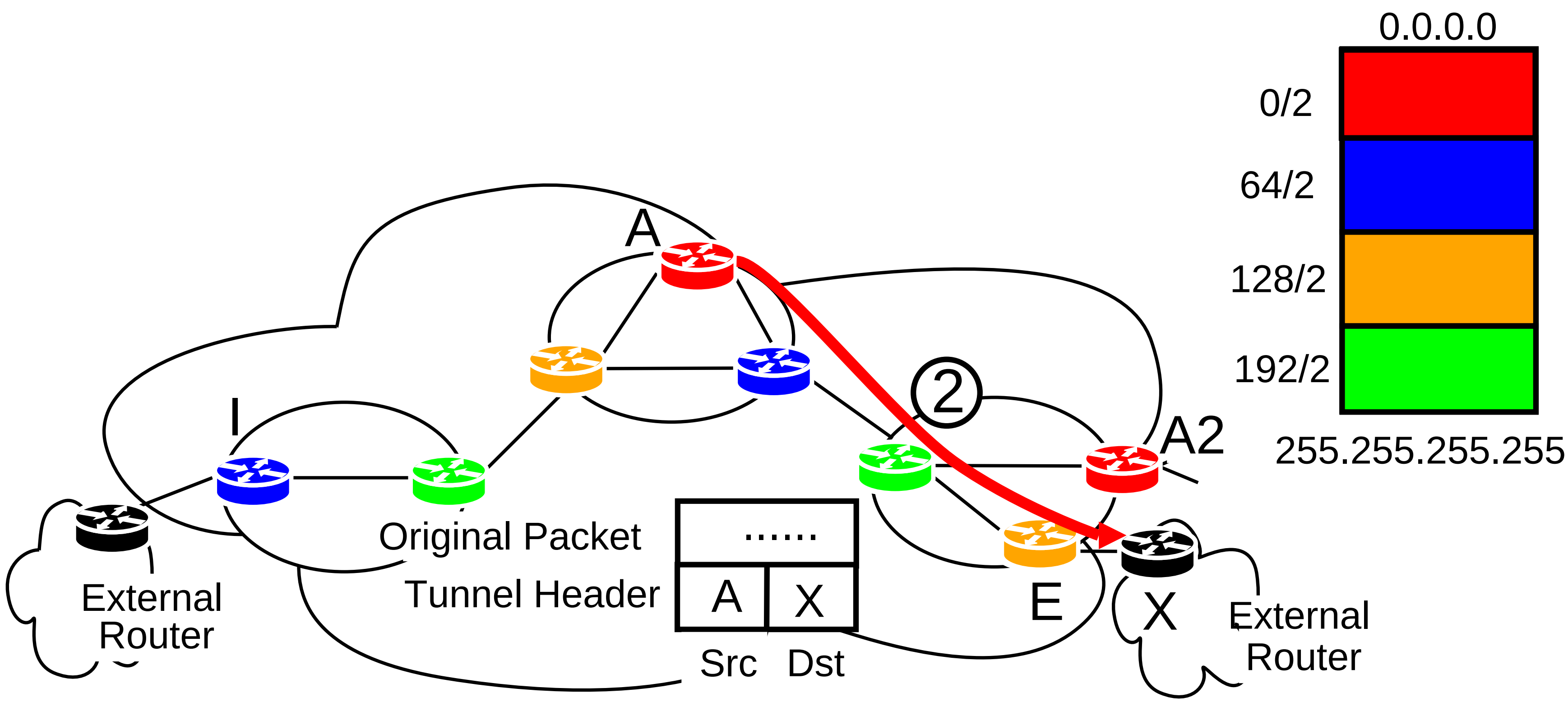
Aggregation Point → Egress



Router A **tunnels** packet to external router as intermediate routers don't have route to dst. prefix

Original packet is encapsulated in tunnel header with X as dst.

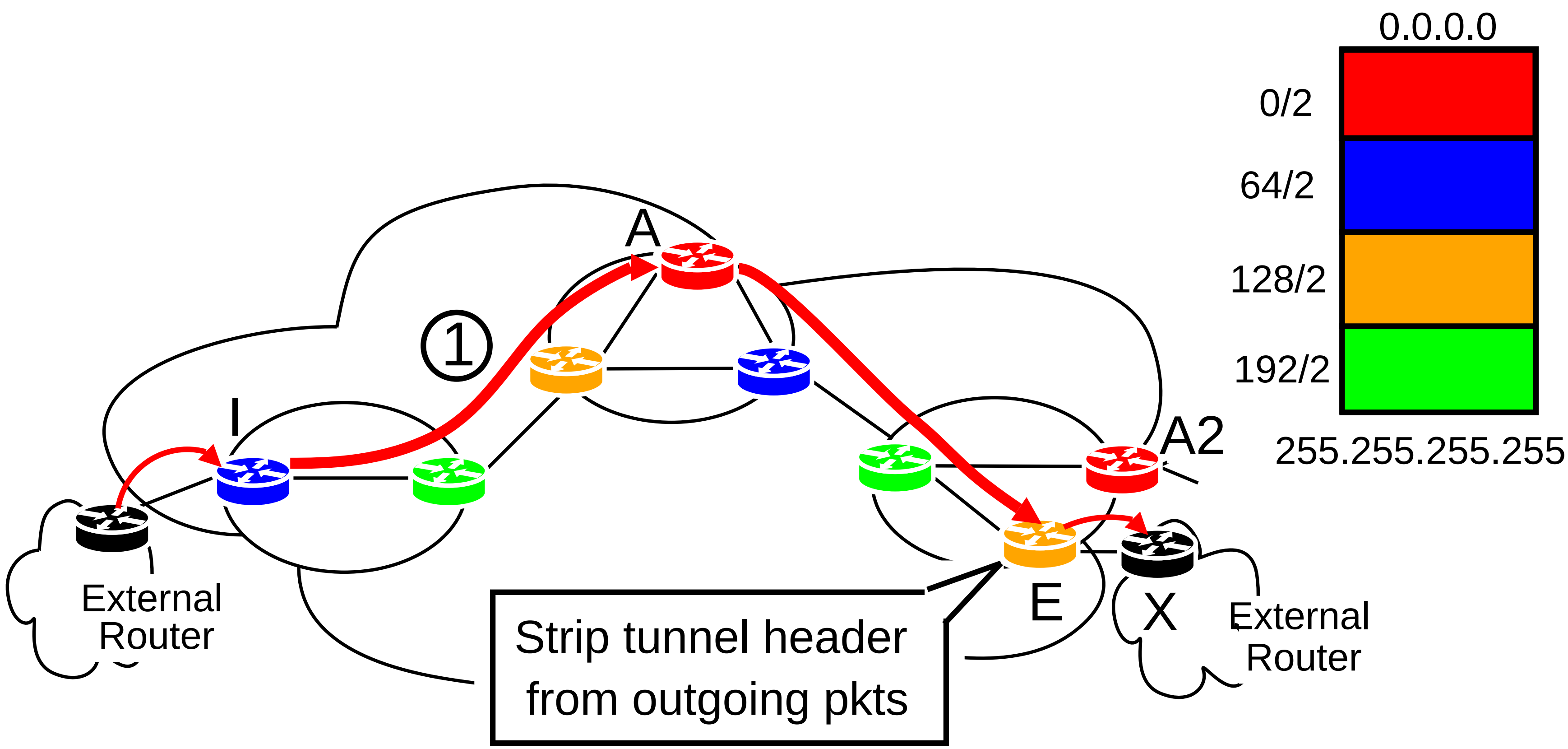
Aggregation Point → Egress



Router A **tunnels** packet to external router as intermediate routers don't have route to dst. prefix

Original packet is encapsulated in tunnel header with X as dst.

Aggregation Point → Egress

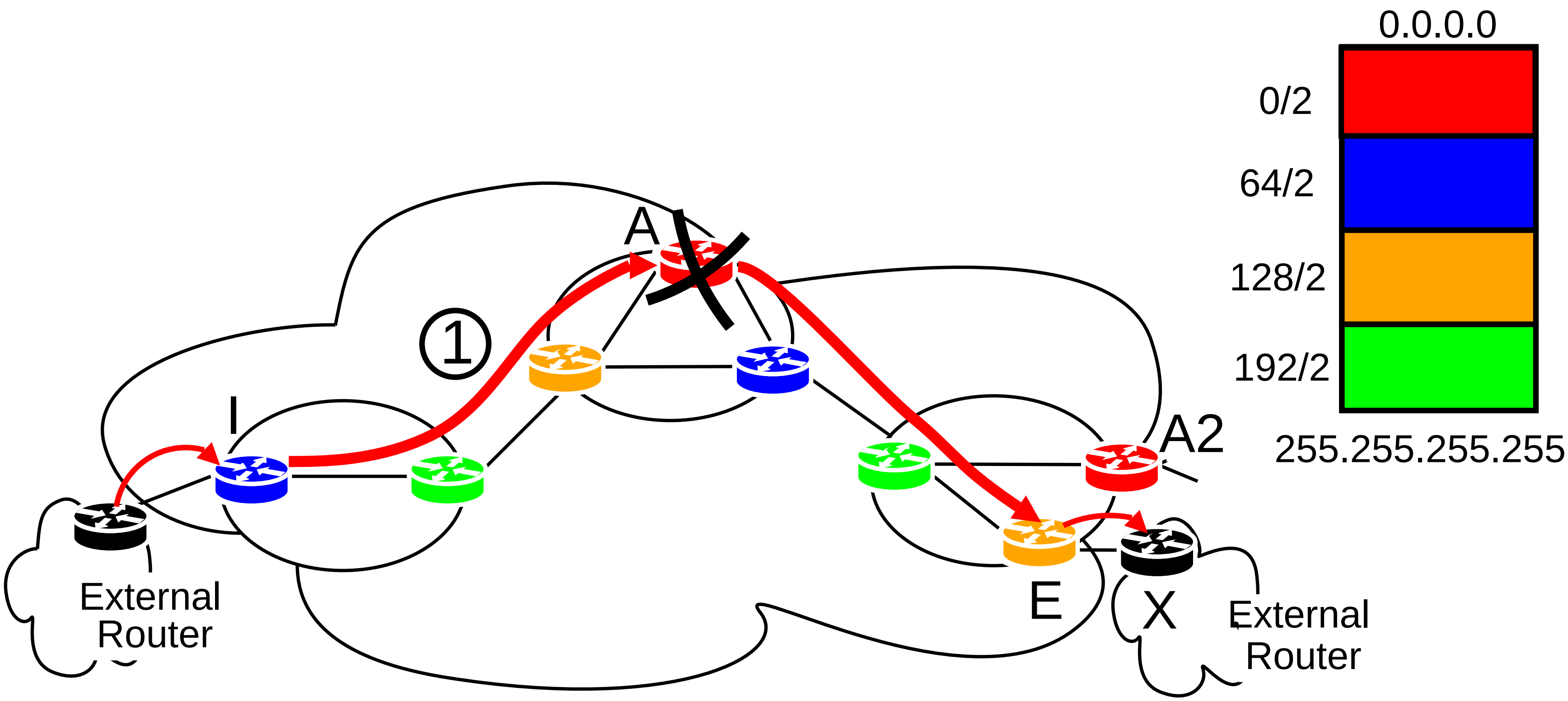


Egress Router strips the tunnel header off outgoing packets

Talk Outline

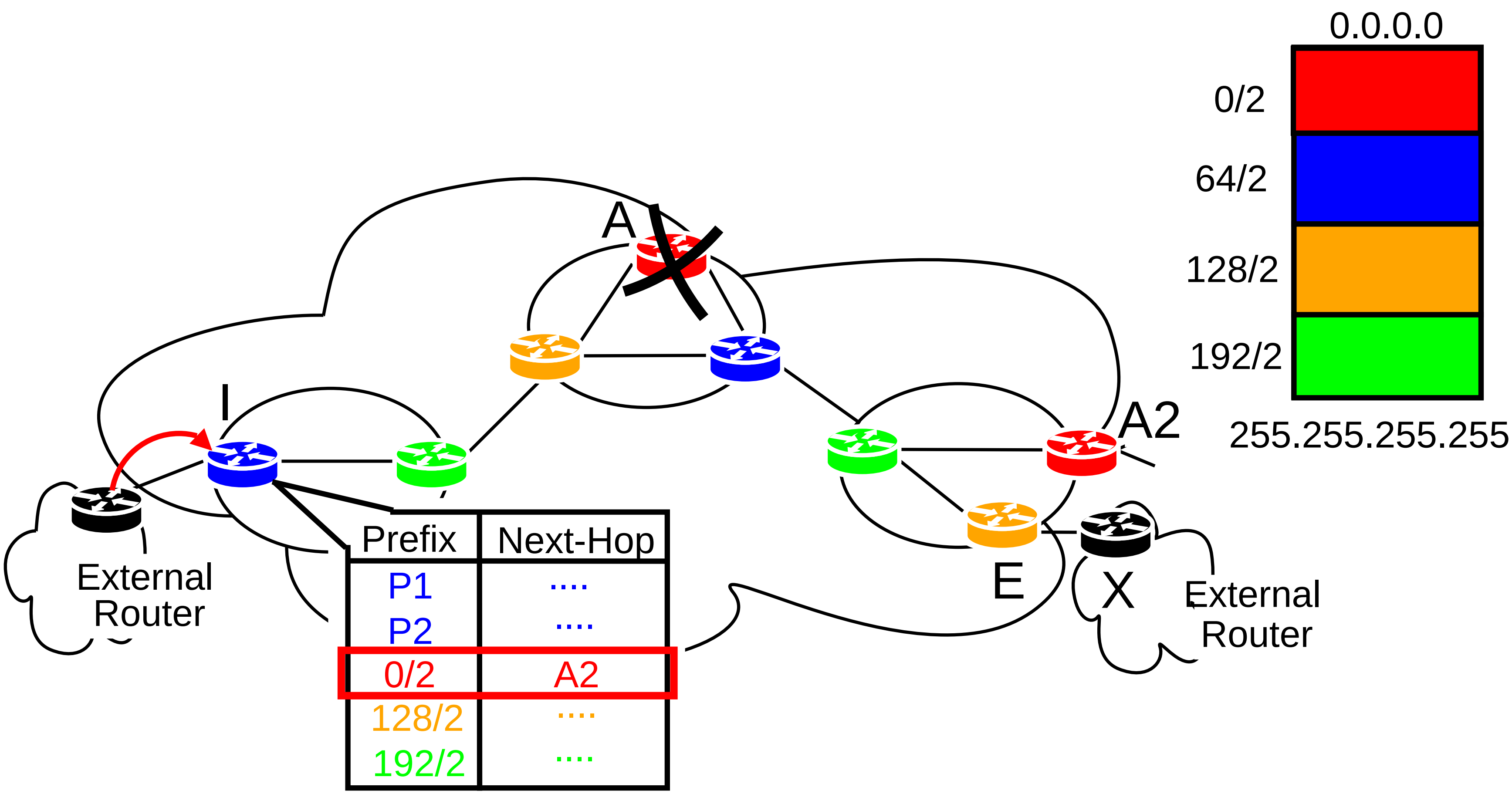
- Motivation
- Router Innards
- Big Picture
- ViAggre Design
- Design Concerns
- Evaluation
- Deployment

Failure of Aggregation Point



What if Aggregation Pt. A fails?

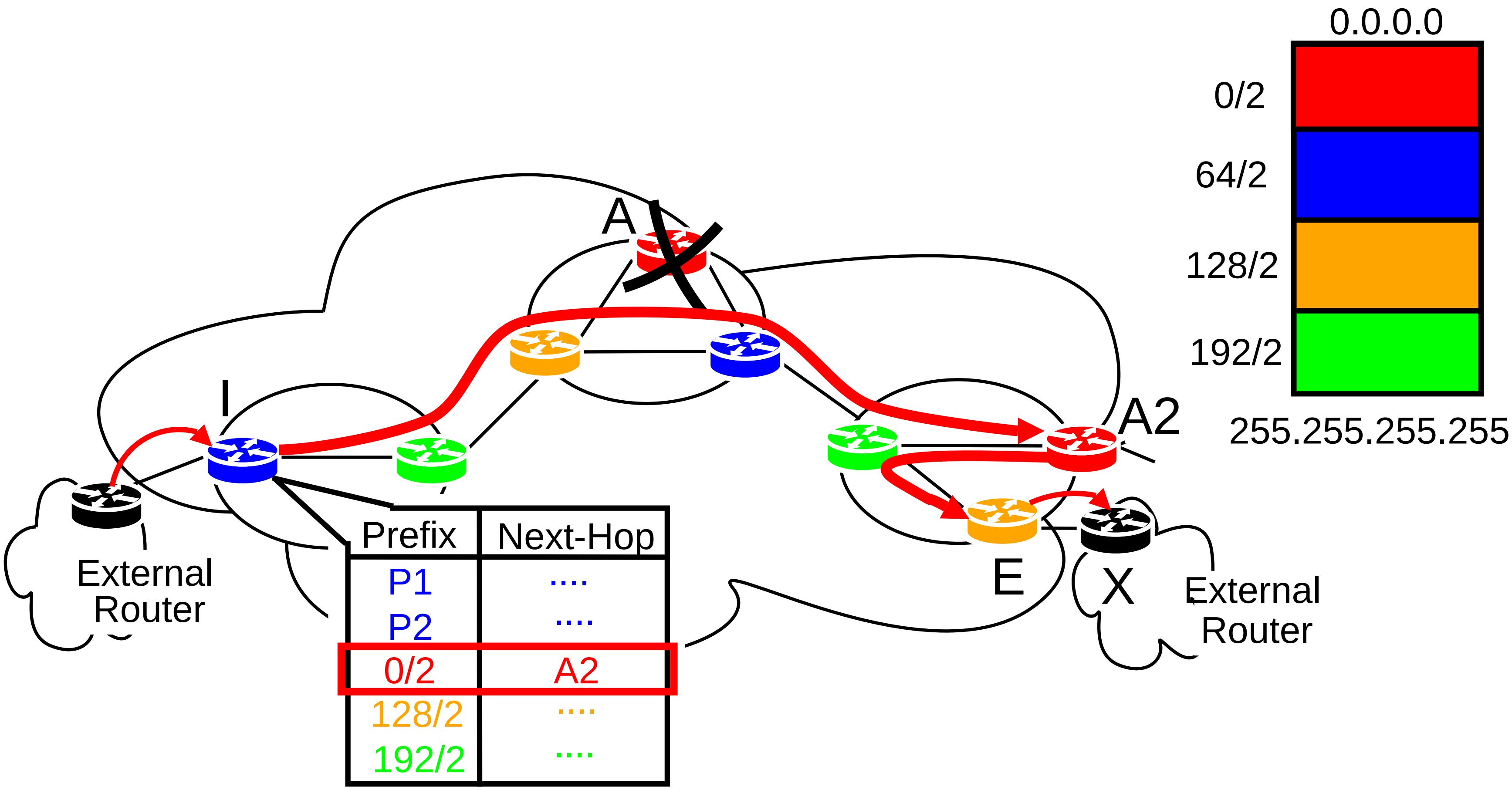
Failure of Aggregation Point



Prefix	Next-Hop
P1	....
P2	....
0/2	A2
128/2	....
192/2	....

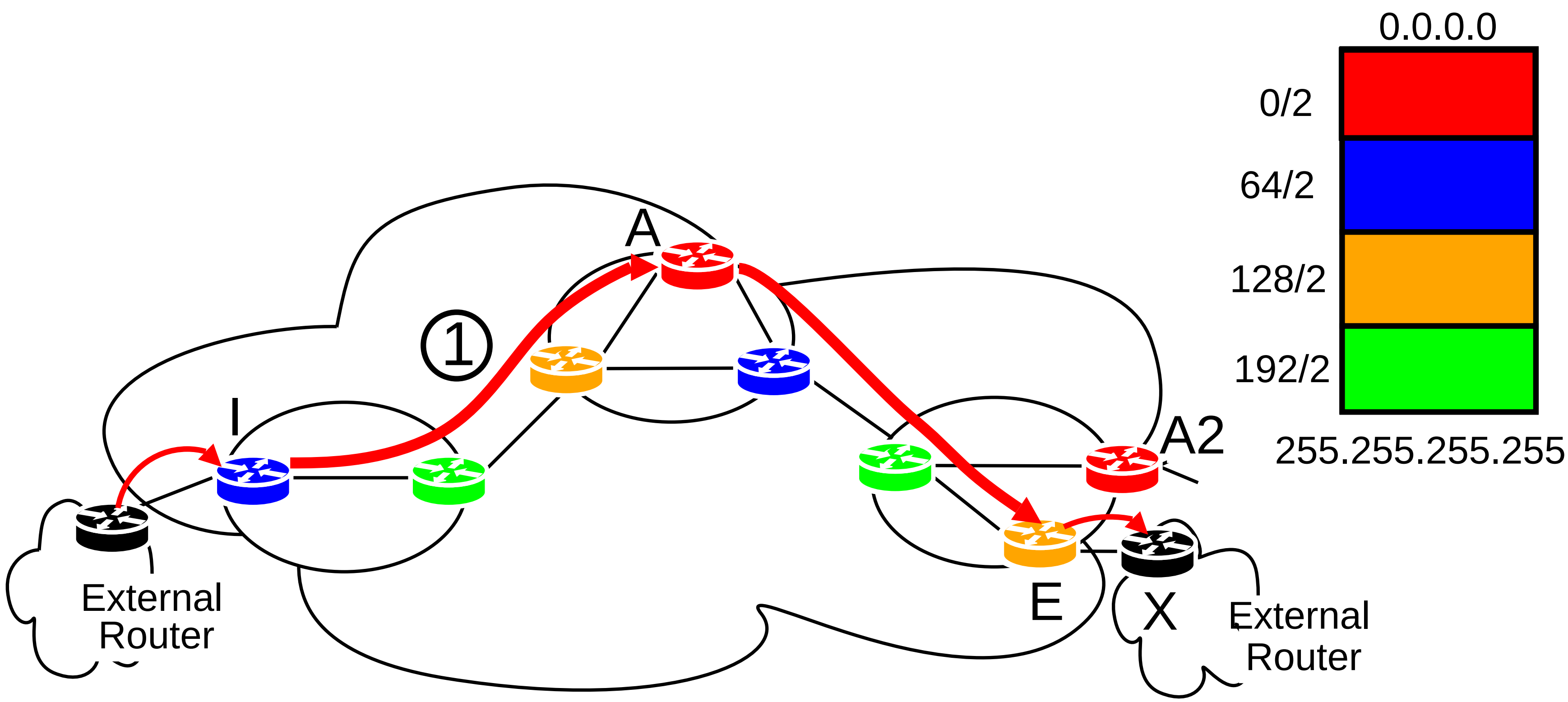
Router I installs the route advertised by A2

Failure of Aggregation Point



Packets are re-routed appropriately

ViAggre's impact on ISP's traffic



ViAggre paths can be longer than native paths

Traffic stretch, increased router and link load, etc.

Popular Prefixes

Traffic volume follows power-law distribution

- 95% of the traffic goes to 5% of prefixes
- Has held up for years

Install “Popular Prefixes” in routers

- Stable over weeks
- Mitigates ViAggre’s impact on the ISP’s traffic

Talk Outline

- Motivation
- Router Innards
- Big Picture
- ViAggre Design
- Design Concerns
- Evaluation
- Deployment

ViAggre's impact on adopting ISP

Positive	**Negative**
Reduction in *FIB Size* (% of global routing table)	Increase in path length (*Stretch* in msec) *Load Increase* (Increase in traffic carried by routers)

ViAggre's impact on adopting ISP

Positive	**Negative**
Reduction in *FIB Size* (% of global routing table)	Increase in path length (*Stretch* in msec) *Load Increase* (Increase in traffic carried by routers)

ViAggre deployment options

- Choosing Virtual Prefixes
- Choosing Aggregation Points
- Choosing Popular Prefixes

ISP can make these choices to tune +ves Vs -ves

ViAggre's impact on adopting ISP

Positive	Negative
Reduction in *FIB Size* (% of global routing table)	Increase in path length (*Stretch* in msec) *Load Increase* (Increase in traffic carried by routers)

ViAggre deployment options

- Choosing Virtual Prefixes
- Choosing Aggregation Points
- Choosing Popular Prefixes

ISP can make these choices to tune +ves Vs -ves

Choosing Aggregation Points

Assigning more routers to aggregate a virtual prefix

- Reduces Stretch imposed on Traffic (as there is a close-by aggregation point to send traffic to)
- Increases FIB size (as more cumulative FIB space is used)

ISP can choose aggregation points to trade-off

FIB Size Vs Stretch

Aggregation Point Assignment Problem

$$\begin{array}{ll} \min & \text{Worst FIB Size} \\ \text{s.t.} & \text{Worst Stretch} \leq \textit{Constraint} \end{array}$$

Constraint on Worst Stretch ensures

- ISP's Service Level Agreements not breached
- Latency-sensitive traffic not hurt too much

Worst FIB Size

- Important for provisioning routers

Aforementioned Constraint Problem

- Can be mapped to MultiCommodity Facility Location
- NP-hard problem
- Logarithmic approximation algorithm [Ravi, Sinha, SODA'04]

Tier-1 ISP Study

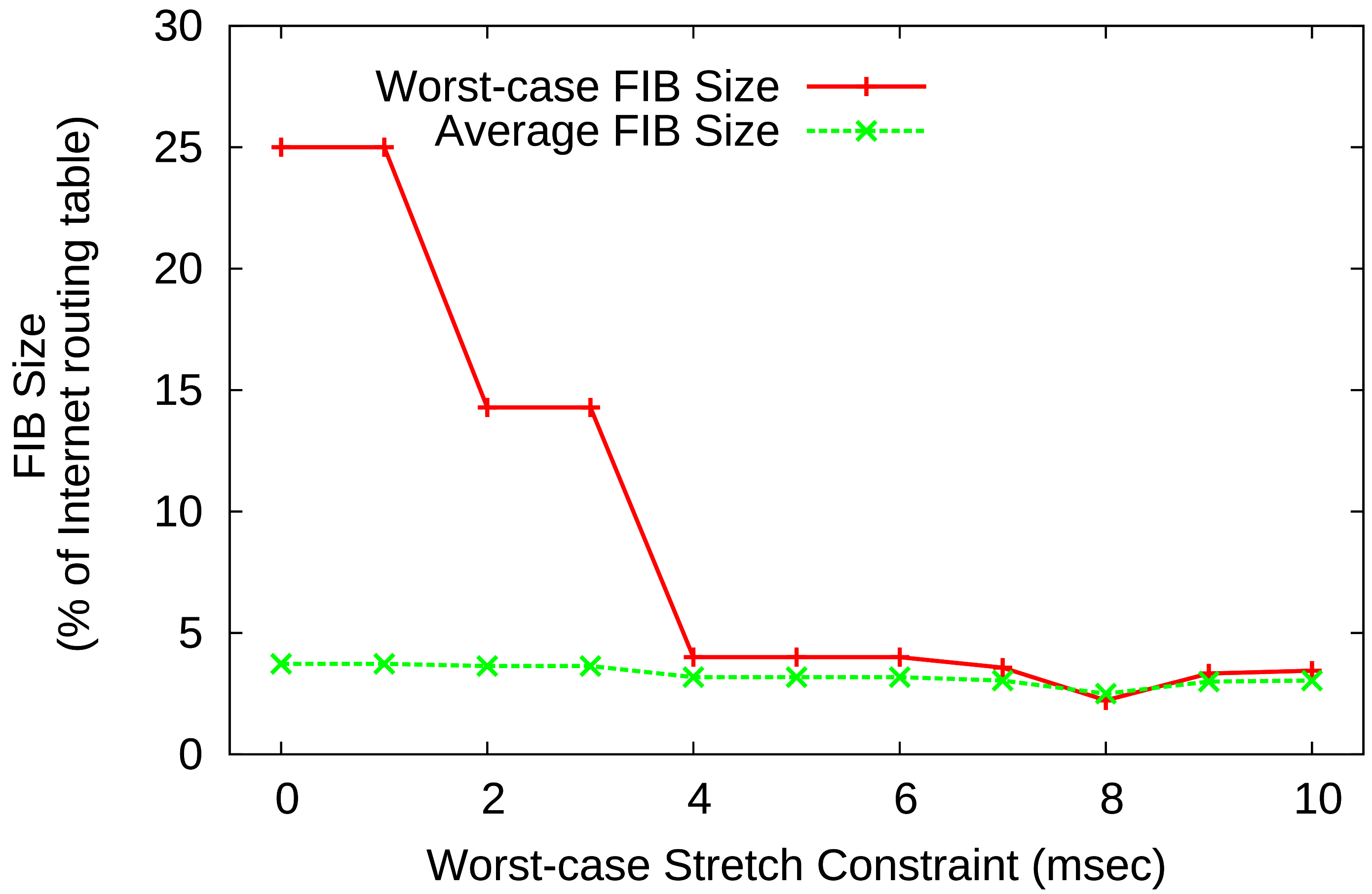
We implemented a greedy approximation algorithm

Algorithm Input: Data from tier-1 ISP

- Topology, Routing tables, Traffic matrix

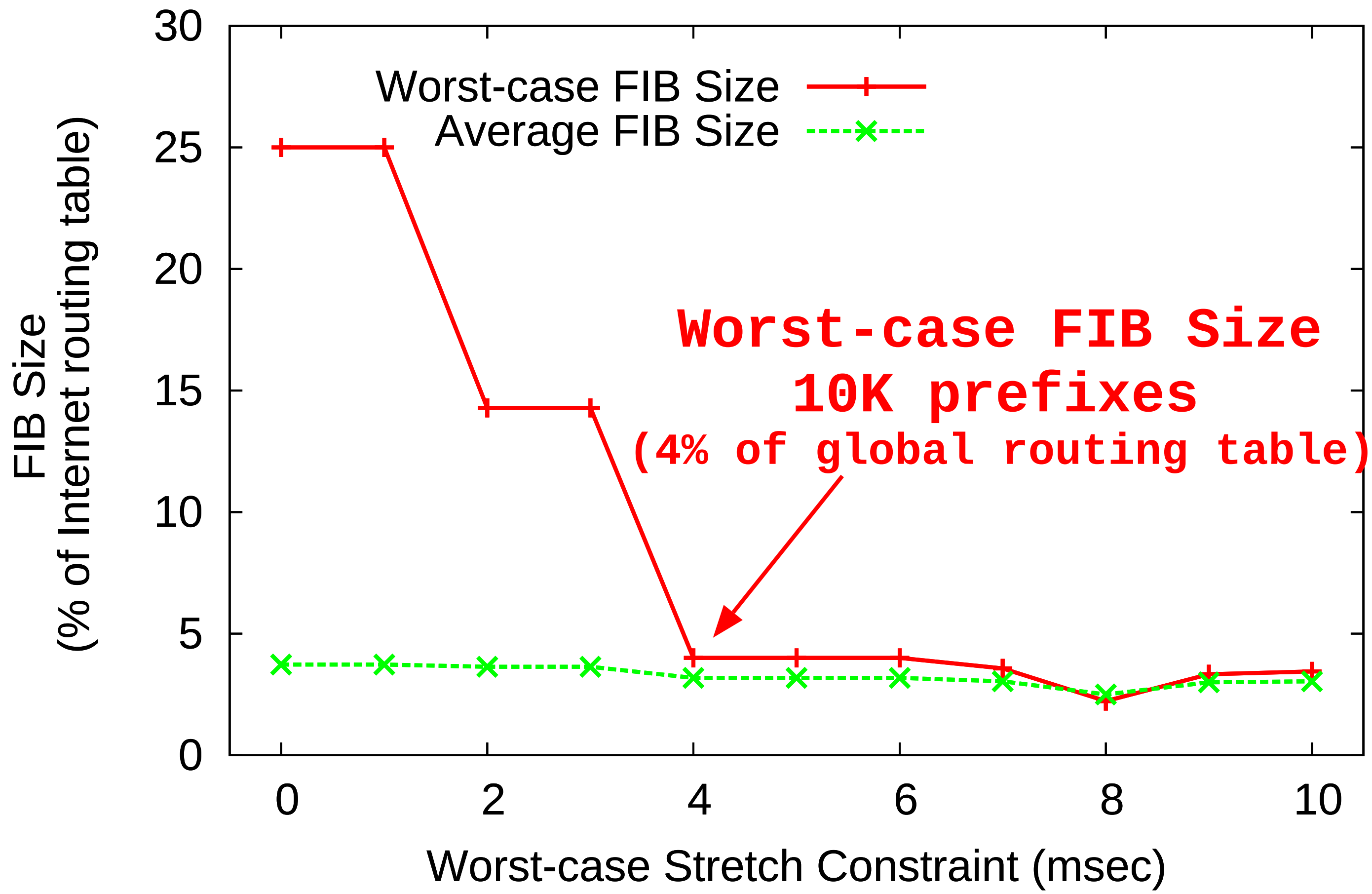
Used our algorithm with varying stretch constraints

FIB Size Vs Stretch



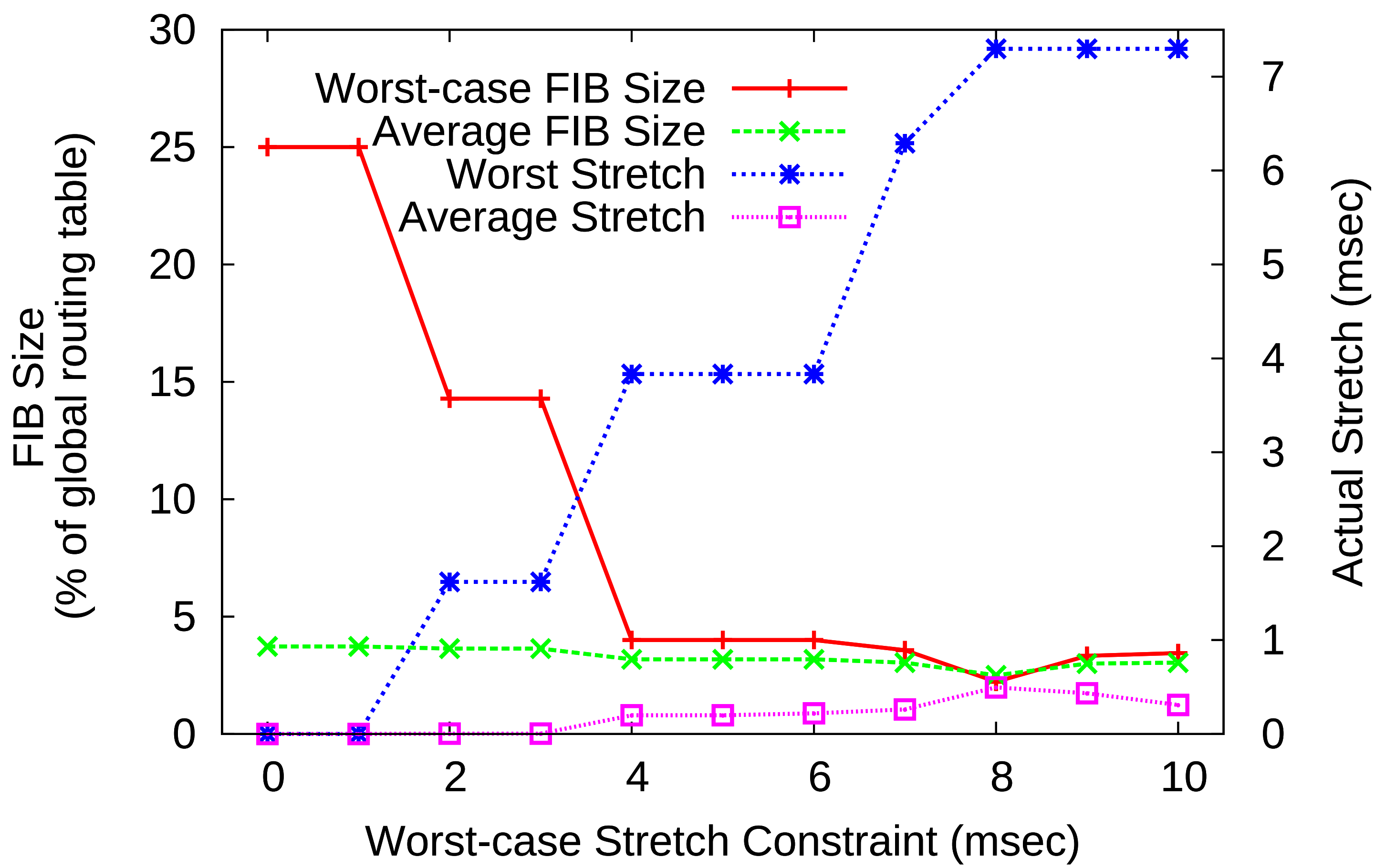
FIB Size reduces as Stretch constraint is relaxed

FIB Size Vs Stretch



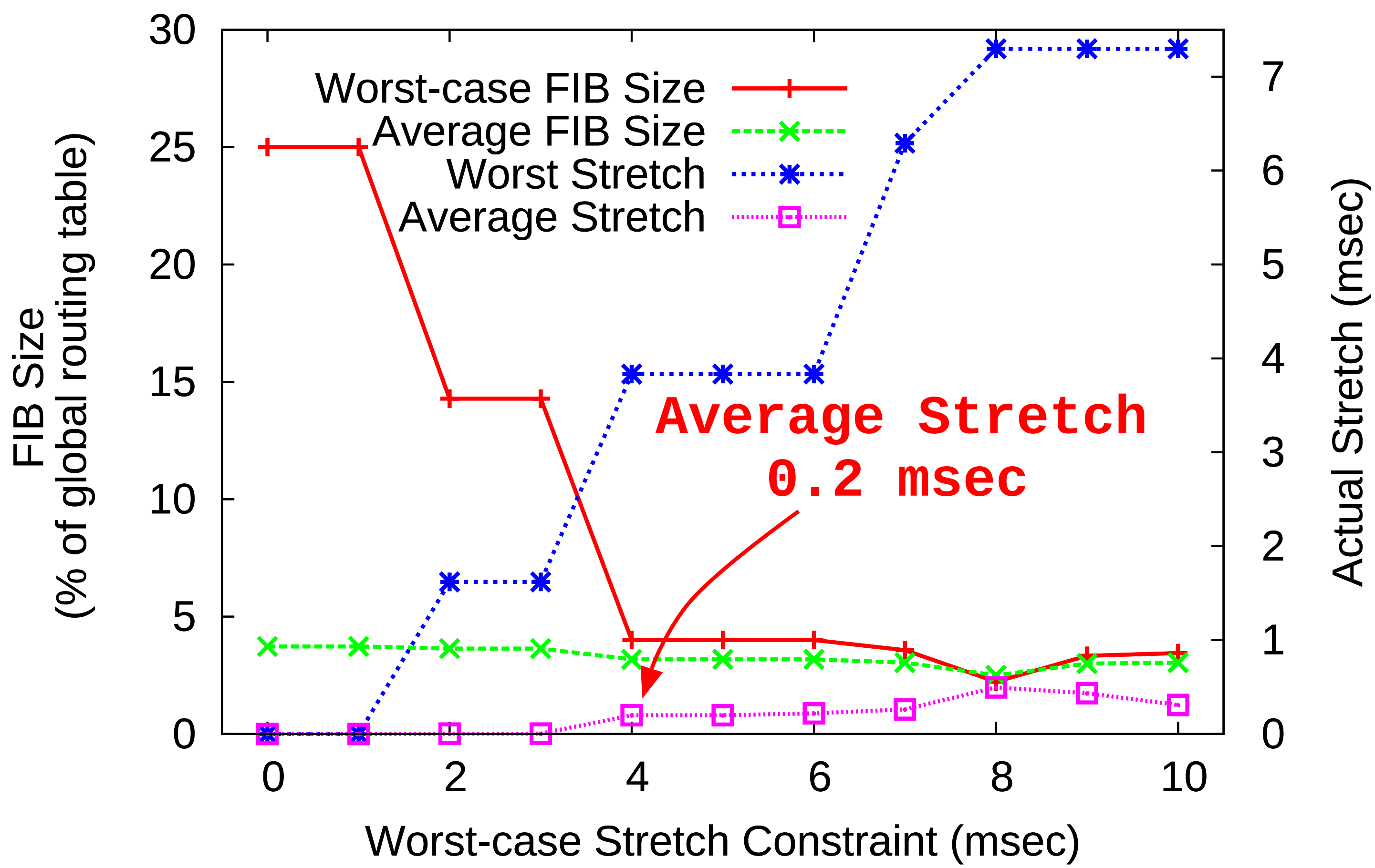
FIB Size reduces as Stretch constraint is relaxed

FIB Size Vs Stretch



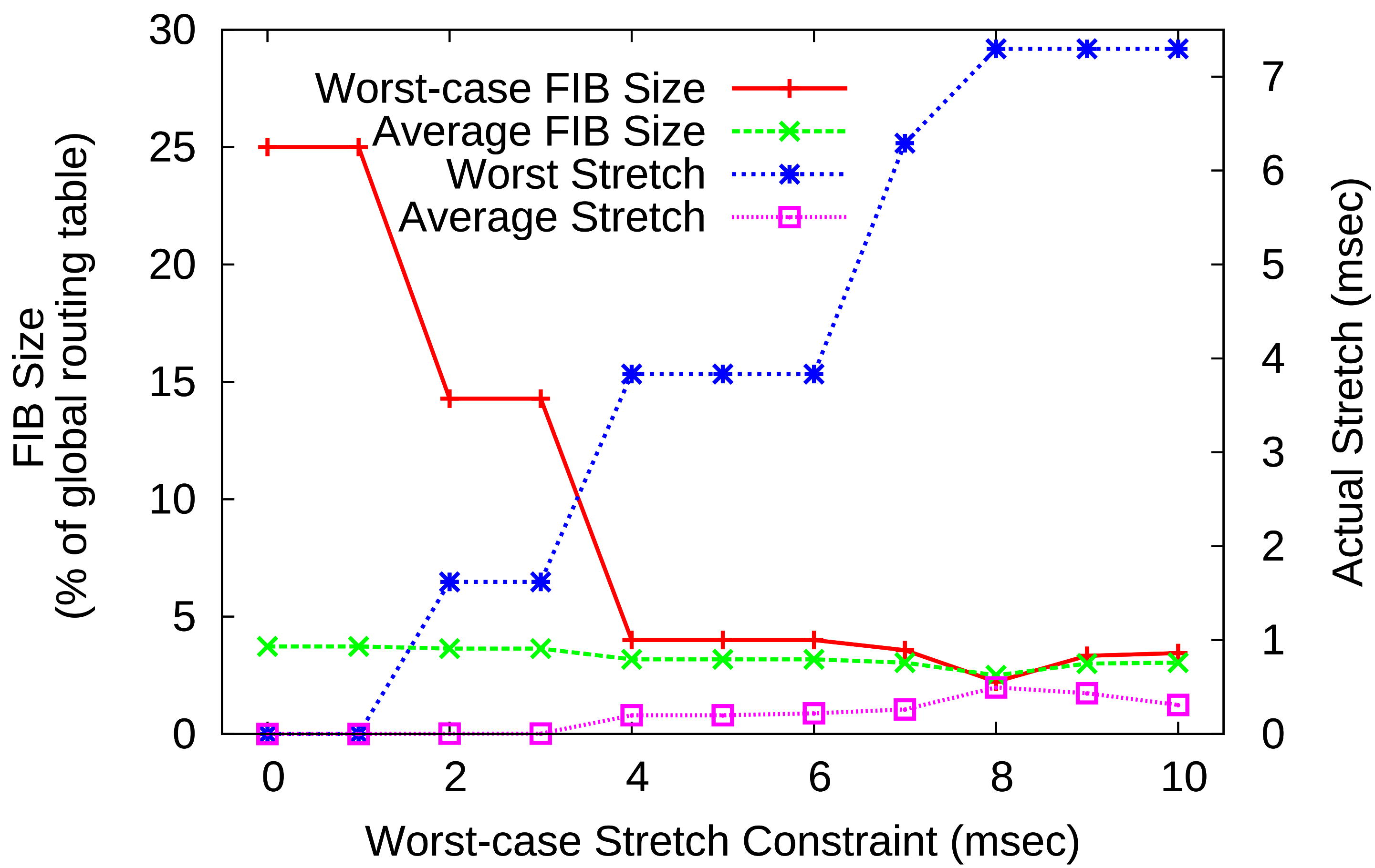
Average Stretch is negligible

FIB Size Vs Stretch



Average Stretch is negligible

FIB Size Vs Stretch



ViAggre can extend lifetime of outdated routers by 7-10 years while imposing no stretch

(Worst-case Stretch Constraint = 0ms)

Router Load

Naïve ViAggre deployment

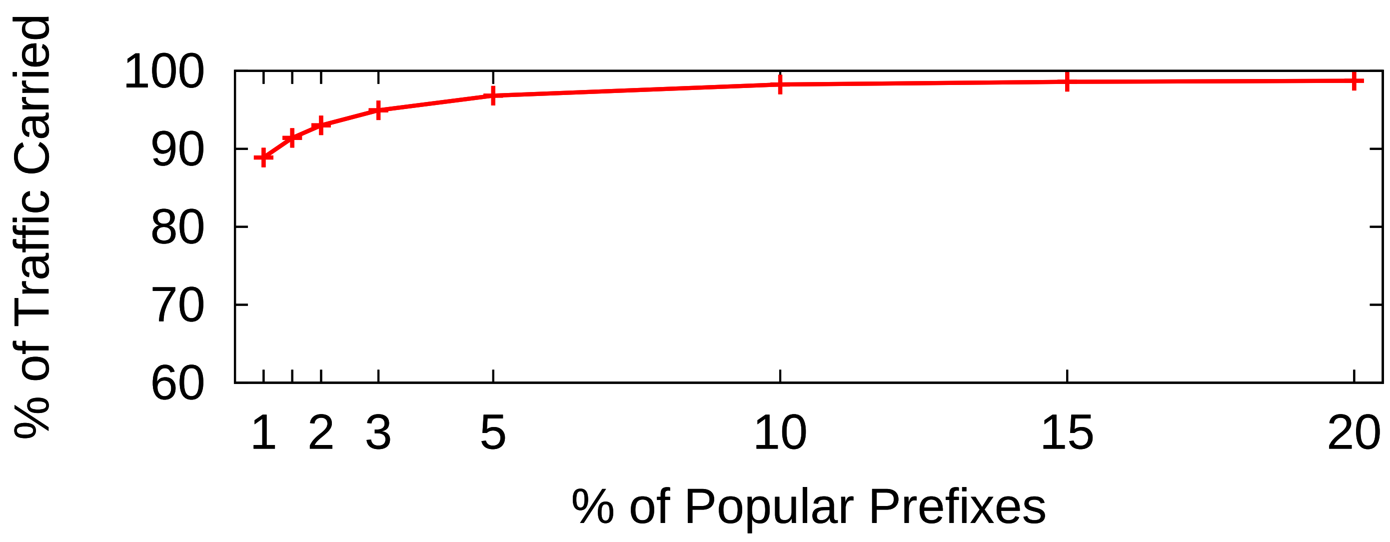
- Traffic routed through aggregation points
- Can lead to substantial load increase across routers
- Alleviative: Use of Popular Prefixes

Router Load

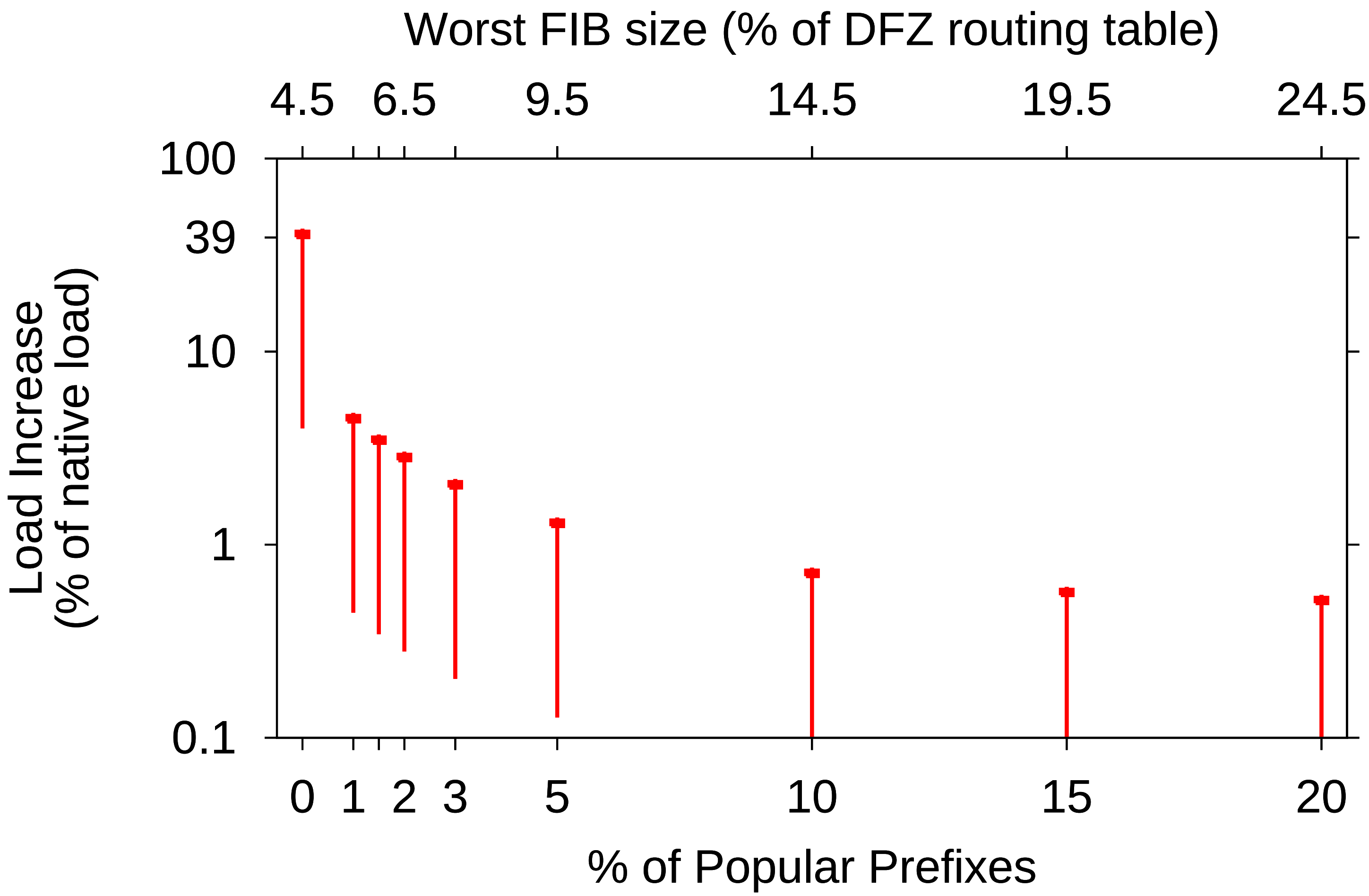
Naïve ViAggre deployment

- Traffic routed through aggregation points
- Can lead to substantial load increase across routers
- Alleviative: Use of Popular Prefixes

A lot of traffic destined to popular prefixes



Router Load



Popular prefixes populated in all routers

5% Popular prefixes $\Rightarrow$ Max. Load Increase= 1.38%

ViAggre Pros

10x reduction in FIB Size

- Negligible Traffic Stretch (<0.2 msec)
- Negligible Increase in Load (<1.5%)

Advantages

- Can be incrementally deployed
- Can be deployed on a limited-scale
- Incentive for deployment
- No change to ISP's routing setup
  - Does not affect routes advertised to neighbors
  - Does not restrict routing policies

ViAggre Cons

Control-plane hacks can impact

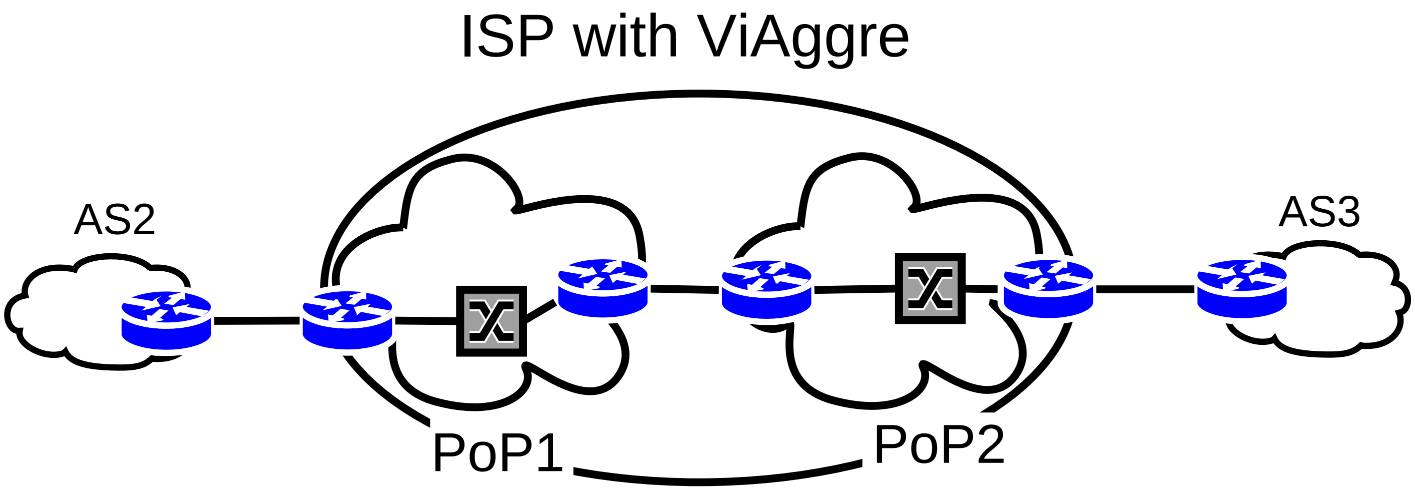
- Installation Time
- Convergence Time
- Failover Time

Planning Overhead

- Choosing virtual prefixes
- Assigning aggregation points
- Assuring network robustness

Configuration overhead of a configuration-only solution

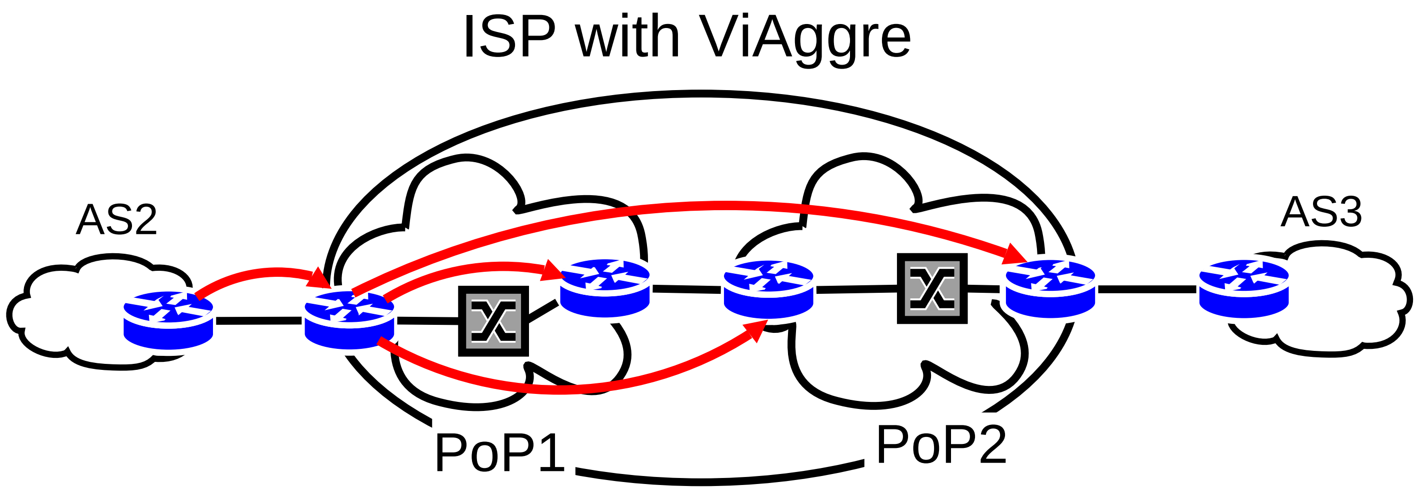
ViAggre Deployment on WAIL



Routes propagated using

- Status Quo
- ViAggre (prefix lists for selective advertisement)

ViAggre Deployment on WAIL

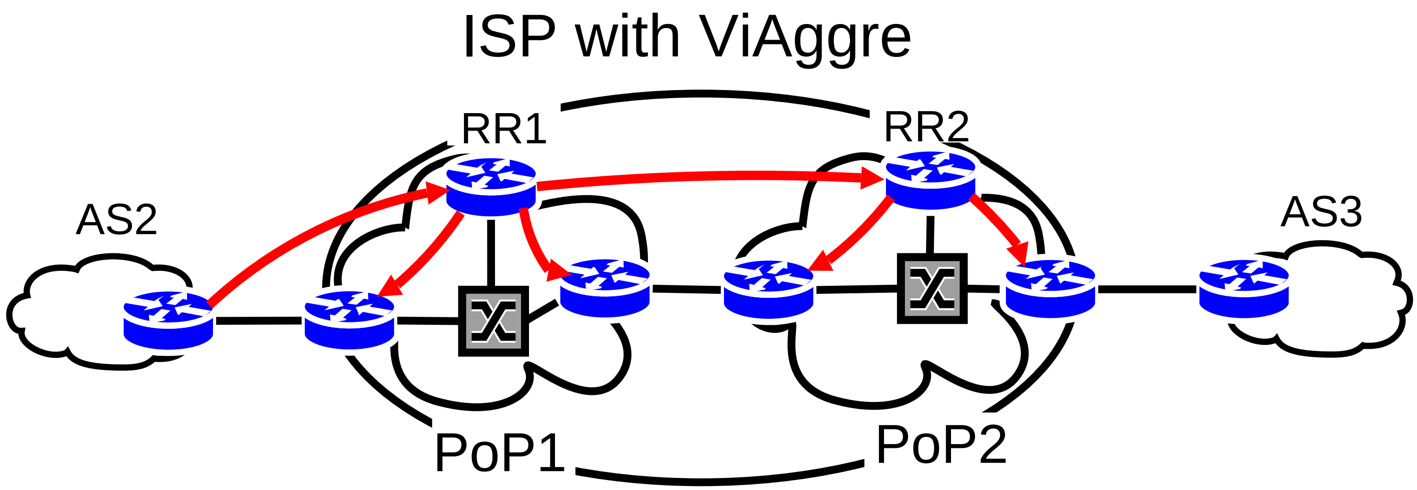


Routes propagated using

- Status Quo
- ViAggre (prefix lists for selective advertisement)

Routes propagated using mesh of internal BGP peerings

ViAggre Deployment on WAIL

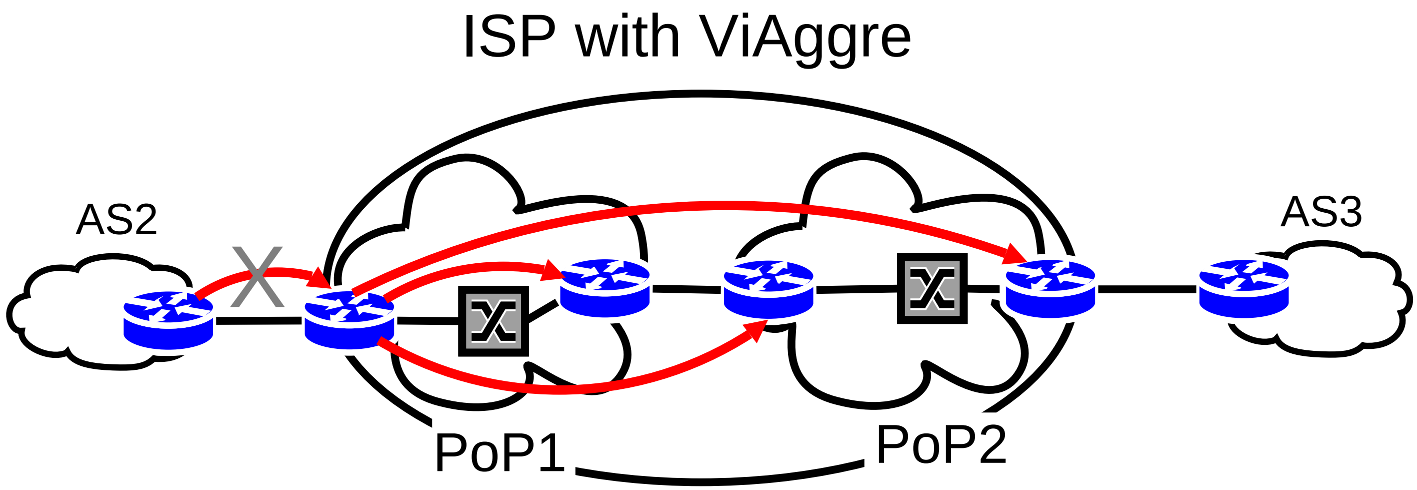


Routes propagated using

- Status Quo
- ViAggre (prefix lists for selective advertisement)

Prefix List size depends on # of popular prefixes

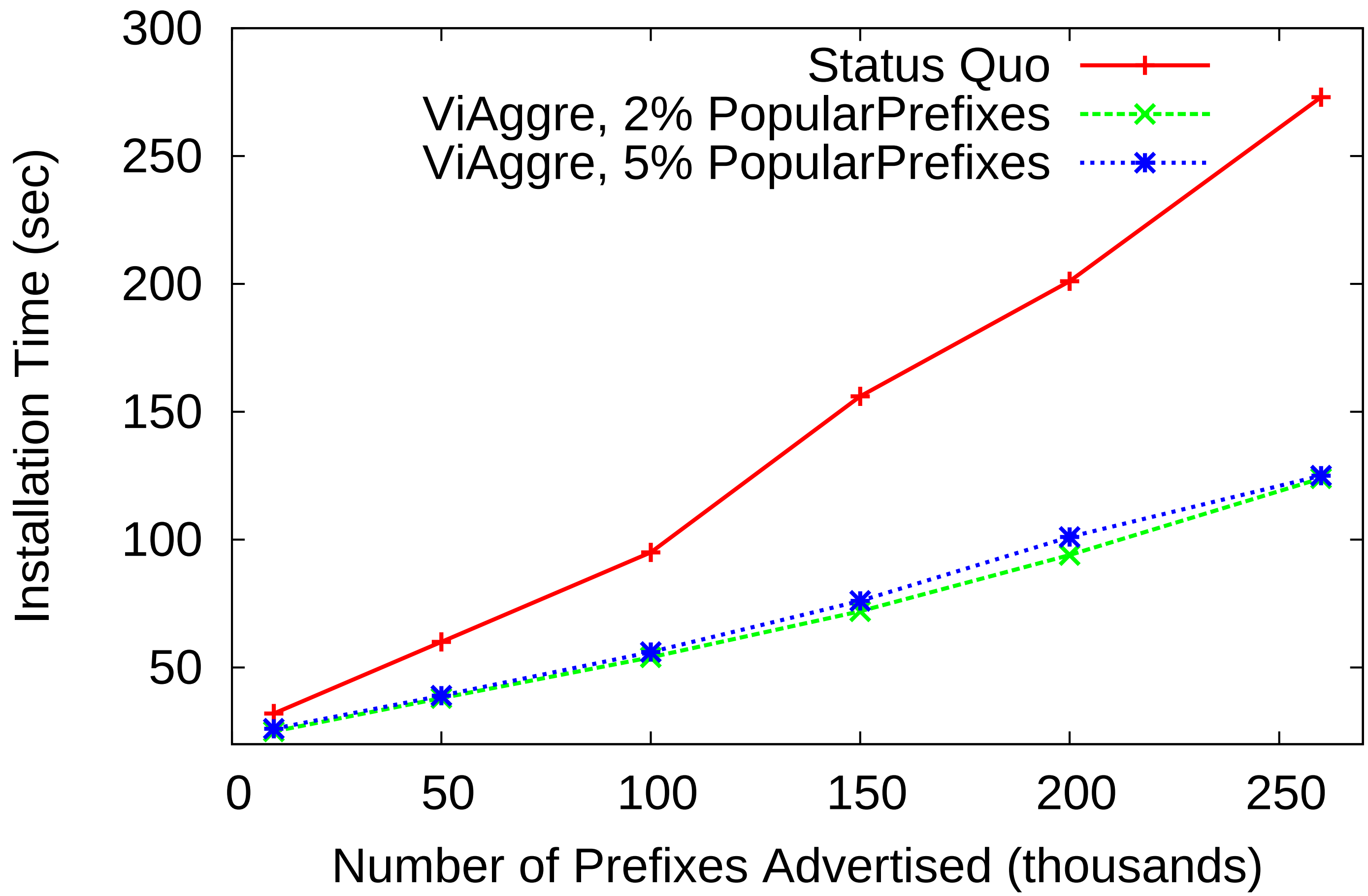
ViAggre Deployment on WAIL



Measuring Control-Plane Overhead

Restart external peering
Measure Installation Time

Installation Time on WAIL



ViAggre reduces Installation Time

Full Routing Table Installation Time

Status Quo=273sec, ViAggre (2% Popular Prefixes)=124sec

ViAggre management overhead

Developed Configuration Tool

- $\sim$330 line python script
- Extracts information from existing configuration files
- Generates ViAggre configuration files
- Planning component in the works

Working with a router vendor (Huawei)

- Implement ViAggre natively
- IETF Draft

ViAggre Conclusion

ViAggre shrinks the FIB on routers

- Can be used by ISPs **today!**
- 10x reduction in FIB Size
- Negligible traffic stretch
- Negligible load increase

ISPs can extend lifetime of their routers

- Outdated routers can be used for 7-10 years

Is this a “complete” solution? No

- A simple and effective first step

Thank You!

Does FIB Size Matter?

Yes

Tony Li [IAB Workshop'06]
Vince Fuller [APRICOT'07]
IAB Workshop [RFC 4984]
. . .

No

DefaultOff [HotNets'05]
AIP [SIGCOMM'08]
. . .

Does FIB Size Matter?

Yes	Maybe	No
Tony Li [IAB Workshop'06]	**Me**	DefaultOff [HotNets'05]
Vince Fuller [APRICOT'07]		AIP [SIGCOMM'08]
IAB Workshop [RFC 4984]		. . .
. . .		

Does FIB Size Matter?

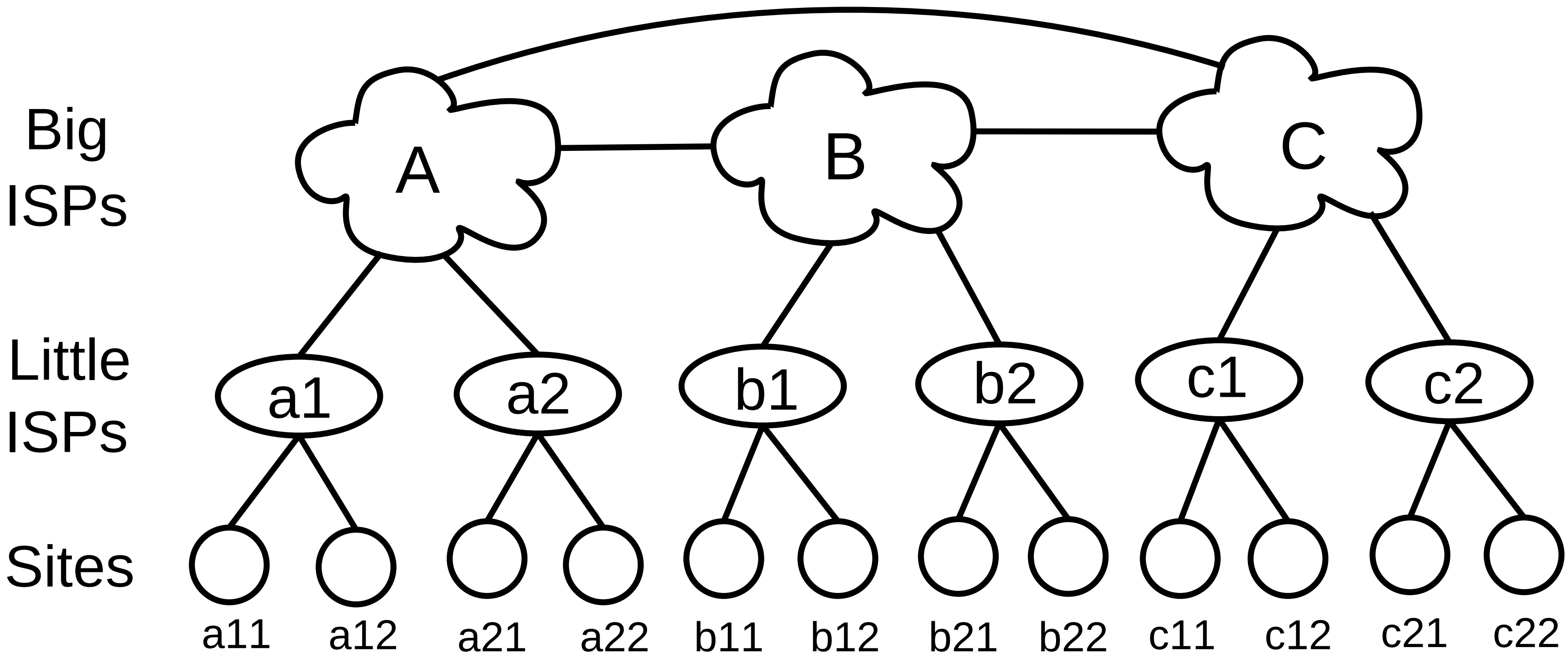
Yes	Maybe	No
Tony Li `[IAB Workshop’06]`	**Me**	DefaultOff `[HotNets’05]`
Vince Fuller `[APRICOT’07]`		AIP `[SIGCOMM’08]`
IAB Workshop `[RFC 4984]`		...
...		

Other reasons to reduce FIB Size

- Rapid future multihoming
- To facilitate commodification of ISP business

Anecdotal evidence shows ISPs are willing to undergo some pain to extend the lifetime of their routers

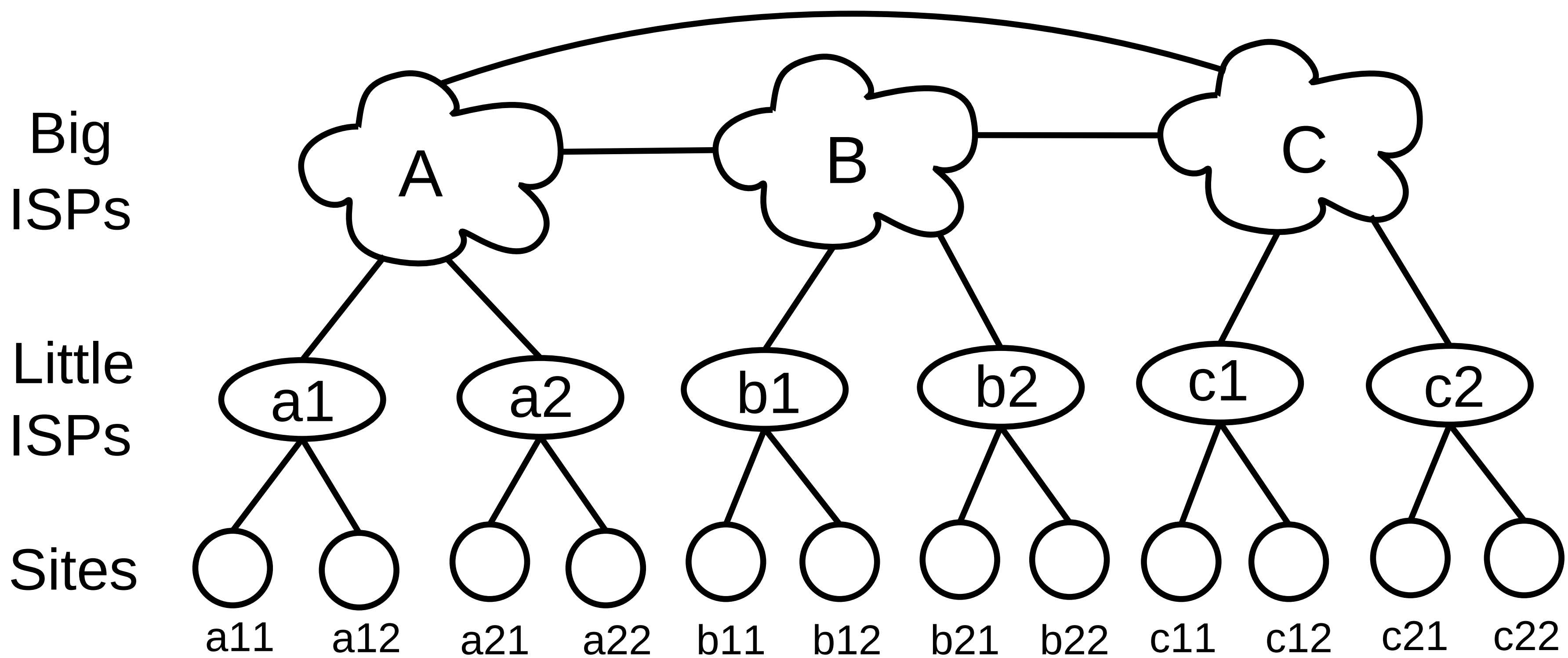
Rapid Routing Table Growth



Internet Routing Scalability is based on hierarchy

Requires addressing to be aligned with topology

Rapid Routing Table Growth

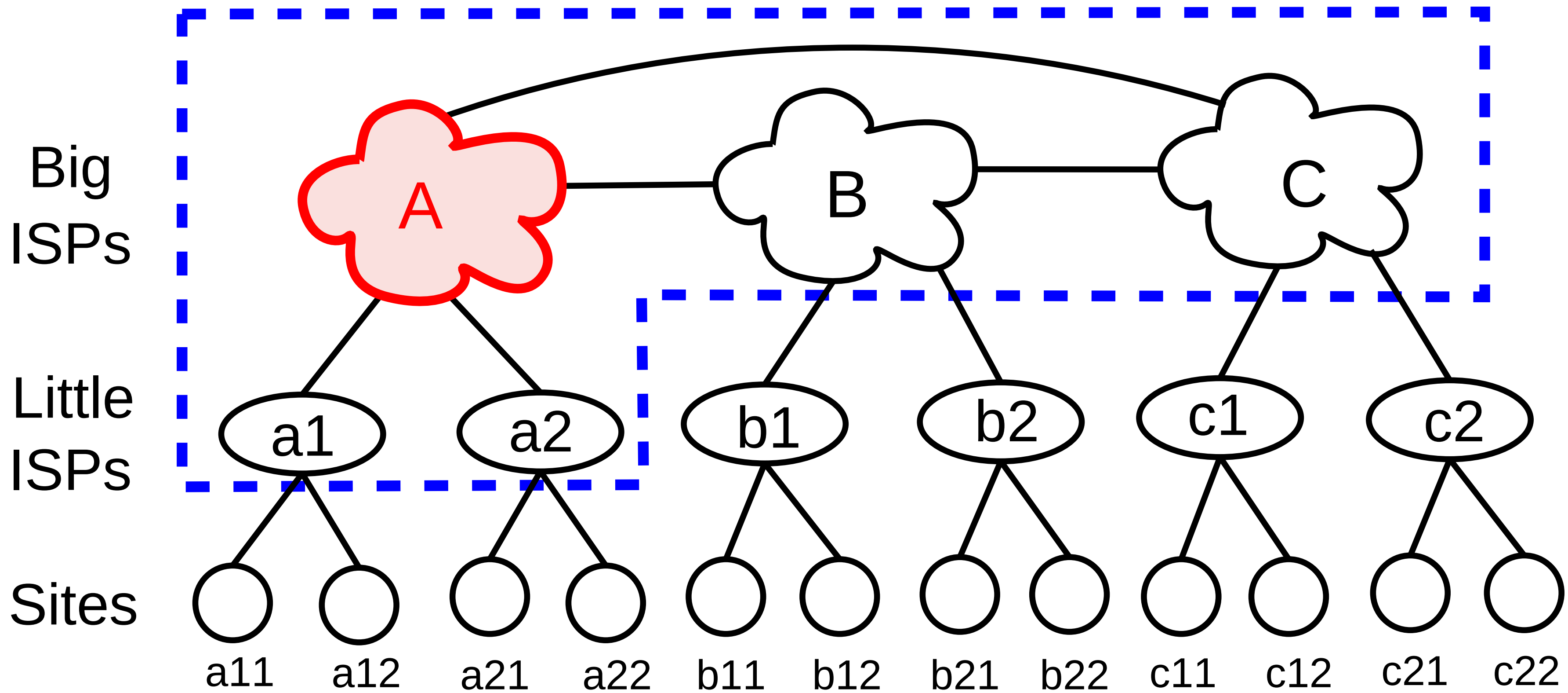


Address $\rightleftarrows$ Topology Match

Sites a11 and a12 are addressed from the address block of a1 which is addressed from the address block of A

$$\{a11, a12\} \subset a1 \subset A$$

Rapid Routing Table Growth

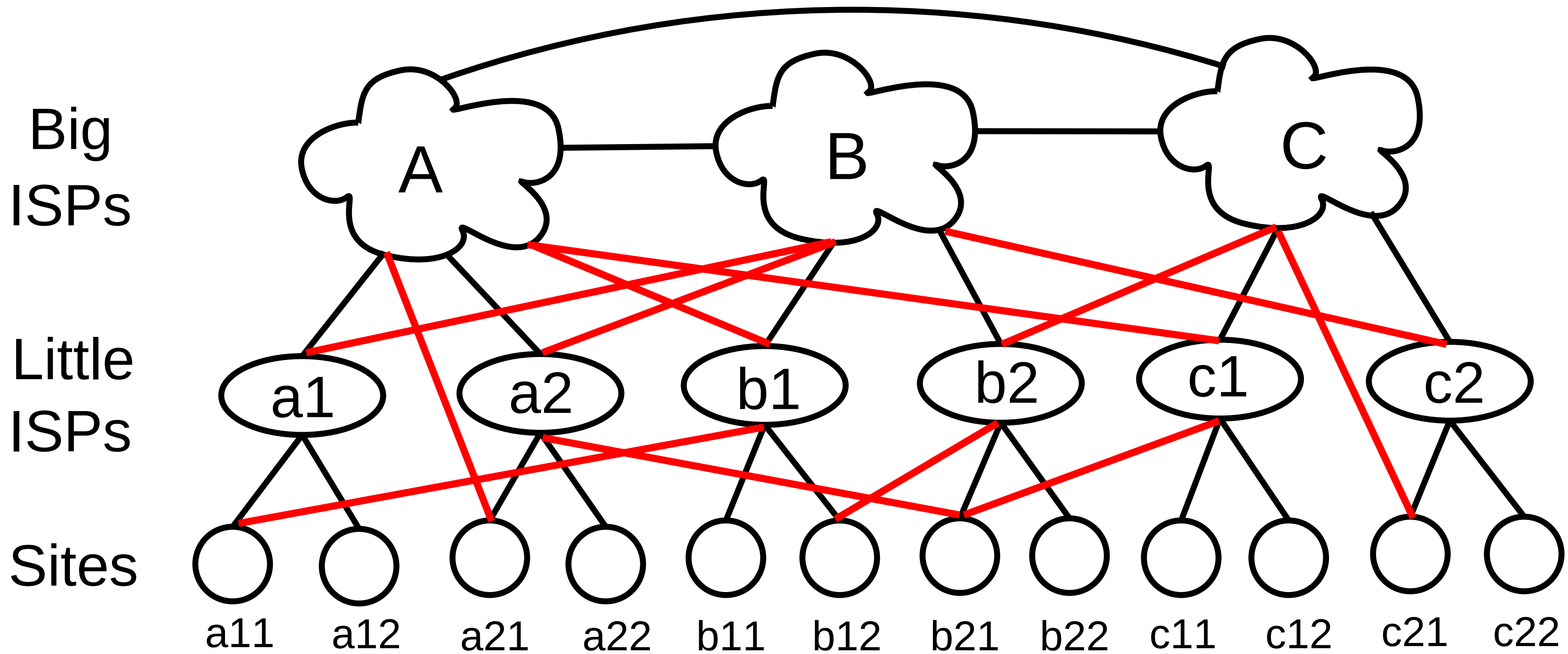


Routing should scale by:

Number of top-level ISPs and Fan-out

Routing state on A: {B, C, a1, a2}

Rapid Routing Table Growth



Address $\rightleftarrows$ Topology Mismatch

Multihoming, Load Balancing, Address Fragmentation, Bad Operational Practices