CSCI-1680
Network Layer: Inter-domain Routing

Rodrigo Fonseca

Based partly on lecture notes by Rob Sherwood, David Mazières, Phil Levis, John Jannotti

Today

- **Last time: Intra-Domain Routing (IGP)**
  - RIP distance vector
  - OSPF link state
- **Inter-Domain Routing (EGP)**
  - Border Gateway Protocol
  - Path-vector routing protocol

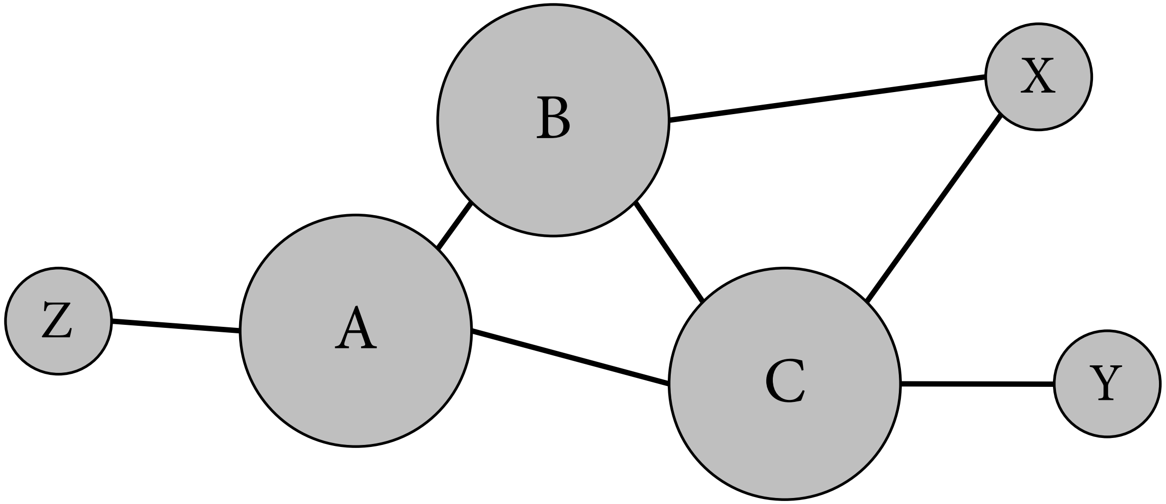
Why Inter vs. Intra

- **Why not just use OSPF everywhere?**
  - E.g., hierarchies of OSPF areas?
  - Hint: scaling is not the only limitation
- **BGP is a policy control and information hiding protocol**
  - intra == trusted, inter == untrusted
  - Different policies by different ASs
  - Different costs by different ASs

Types of ASs

  - Local Traffic – source or destination in local AS
  - Transit Traffic – passes through an AS
- **Stub AS**
  - Connects to only a single other AS
- **Multihomed AS**
  - Connects to multiple ASs
  - Carries no transit traffic
- **Transit AS**
  - Connects to multiple ASs and carries transit traffic

AS Relationships



- **How to prevent X from forwarding transit between B and C?**
- **How to avoid transit between CBA ?**
  - B: BAZ -> X ("B advertises BAZ to X")
  - B: BAZ -> C ? (=> Y: CBAZ and Y:CAZ)

Example from Kurose and Ross, 5th Ed

Choice of Routing Algorithm

- **Constraints**
  - Scaling
  - Autonomy (policy and privacy)
- **Link-state?**
  - Requires sharing of complete information
  - Information exchange does not scale
  - Can’t express policy
- **Distance Vector?**
  - Scales and retains privacy
  - Can’t implement policy
  - Can’t avoid loops if shortest path not taken
  - Count-to-infinity

Path Vector Protocol

- **Distance vector algorithm with extra information**
  - For each route, store the complete path (ASs)
  - No extra computation, just extra storage (and traffic)
- **Advantages**
  - Can make policy choices based on set of ASs in path
  - Can easily avoid loops

BGP - High Level

- **Single EGP protocol in use today**
- **Abstract each AS to a single node**
- **Destinations are CIDR prefixes**
- **Exchange prefix *reachability* with neighbors**
  - E.g., “I can reach prefix 128.148.0.0/16 through ASes 44444 3356 14325 11078”
  - May choose to not advertise some paths to some neighbors
- **Select a single path by routing *policy***
- **Critical: learn many paths, propagate one**
  - Add your ASN to advertised path

BGP Implications

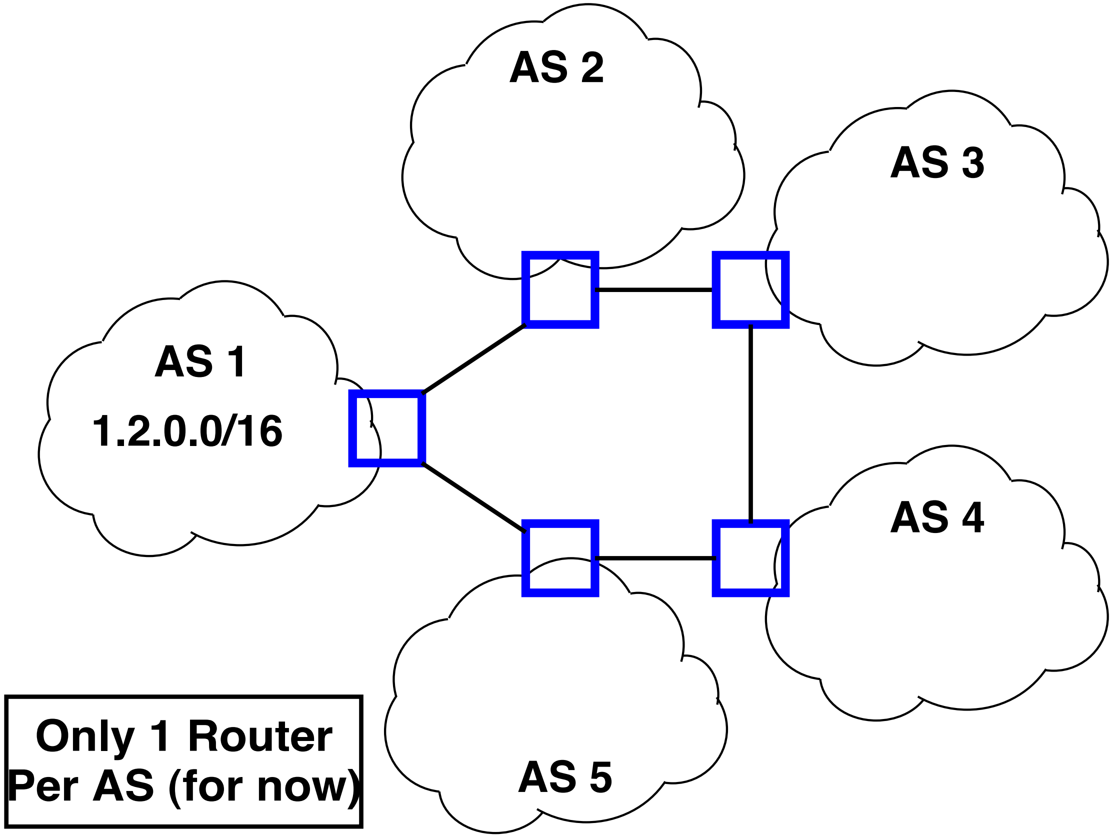
- **Explicit AS Path == Loop free**
  - Except under churn, IGP/EGP mismatch
- **Not all ASs know all paths**
- **Reachability not guaranteed**
  - Decentralized combination of policies
- **AS abstraction -> loss of efficiency**
- **Scaling**
  - 55K ASs
  - 685K+ prefixes
  - ASs with one prefix: 21292
  - Most prefixes by one AS: 5551 (AS4538 ERX-CERNET-BKB - China Education and Research Network Center)

Source: cidr-report 17Oct2017

Why study BGP?

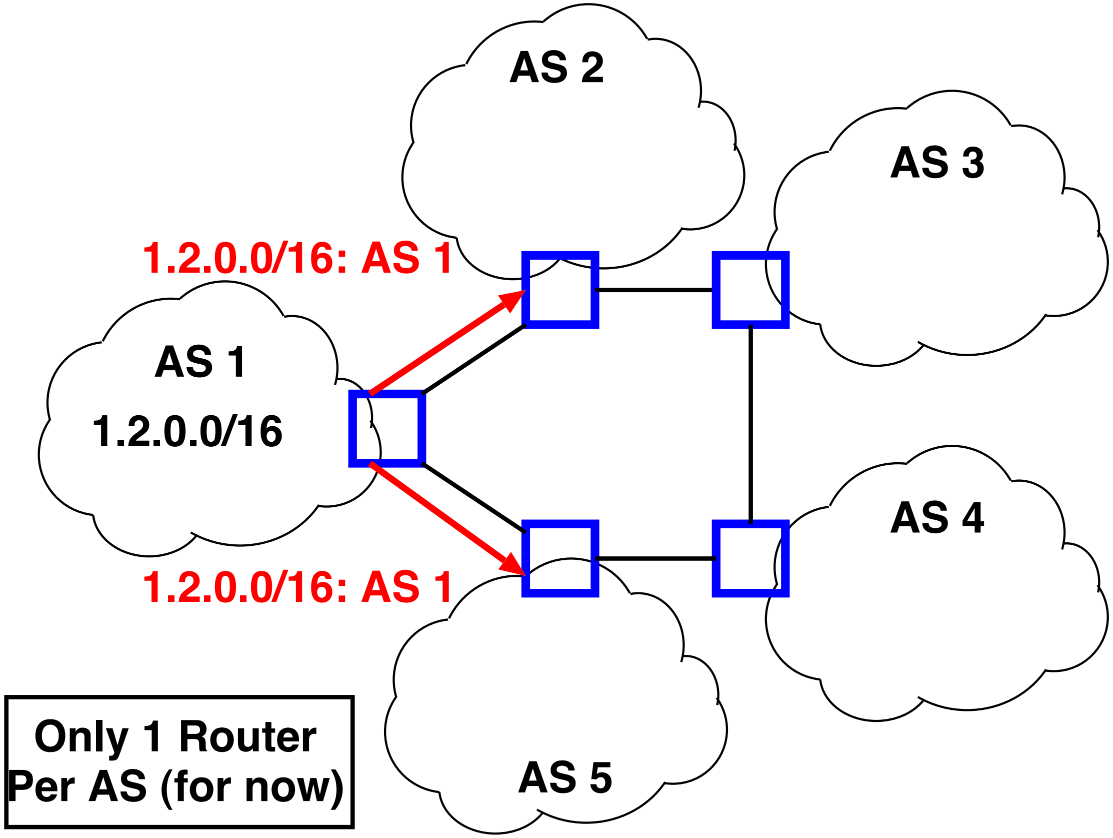
- **Critical protocol: makes the Internet run**
  - Only widely deployed EGP
- **Active area of problems!**
  - Efficiency
  - Cogent vs. Level3: Internet Partition
  - Spammers use prefix hijacking
  - Pakistan accidentally took down YouTube
  - Egypt disconnected for 5 days

BGP Example



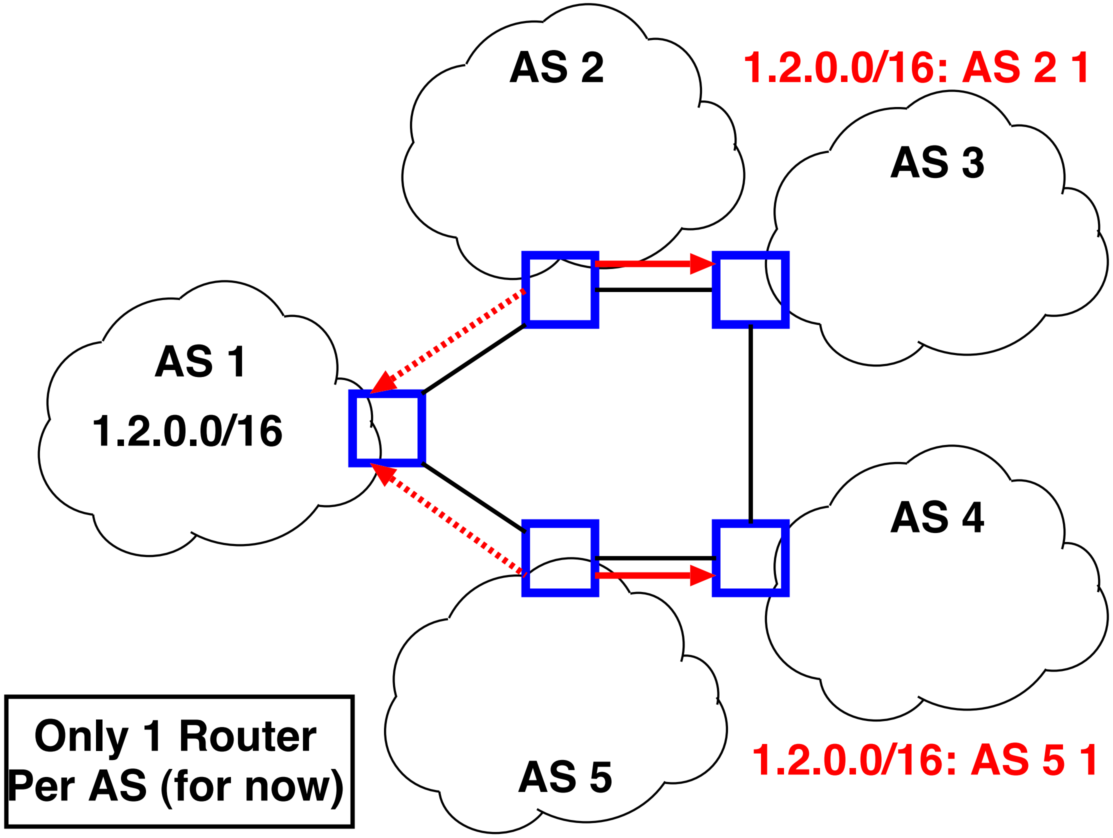
Only 1 Router
Per AS (for now)

BGP Example



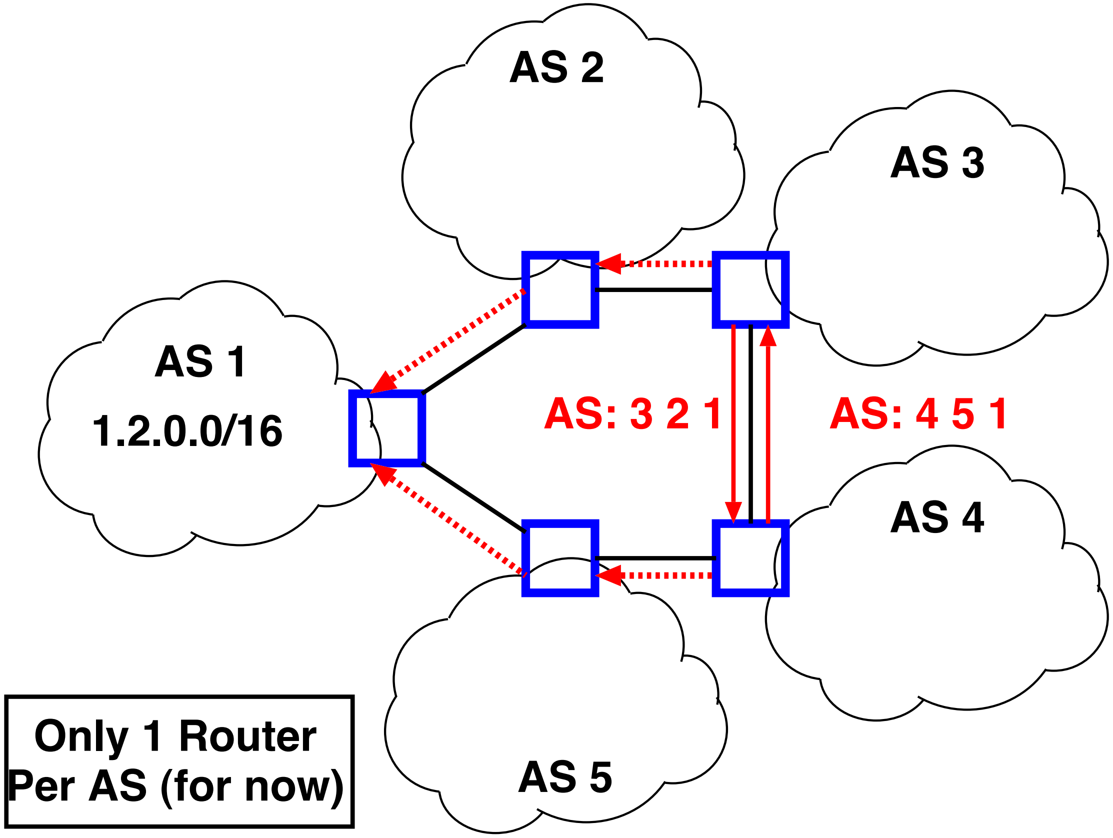
Only 1 Router Per AS (for now)

BGP Example



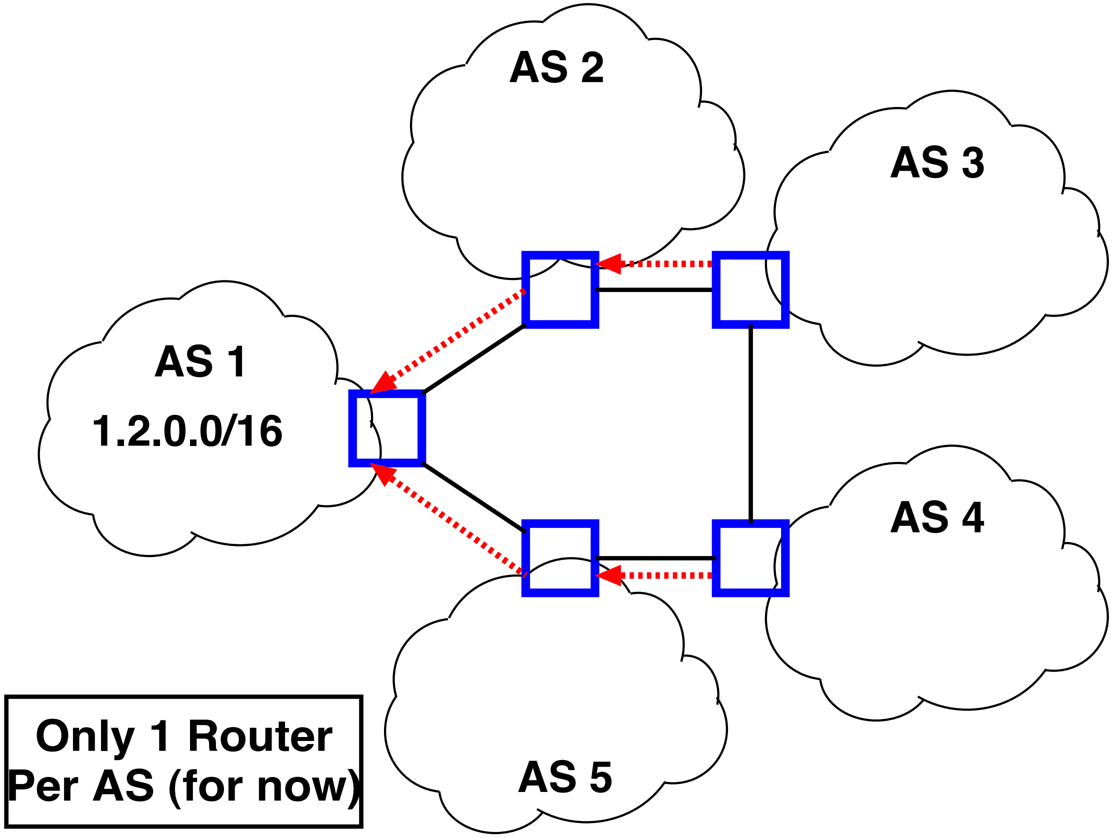
Only 1 Router
Per AS (for now)

BGP Example



Only 1 Router
Per AS (for now)

BGP Example



Only 1 Router
Per AS (for now)

BGP Protocol Details

- **Separate roles of *speakers* and *gateways***
  - Speakers talk BGP with other ASes
  - Gateways are routes that border other ASes
  - Can have more gateways than speakers
  - Speakers know how to reach gateways
- **Speakers connect over TCP on port 179**
  - Bidirectional exchange over long-lived connection

BGP Messages

- **Base protocol has four message types**
  - OPEN – Initialize connection. Identifies peers and must be first message in each direction
  - UPDATE – Announce routing changes (most important message)
  - NOTIFICATION – Announce error when closing connection
  - KEEPALIVE – Make sure peer is alive
- **Extensions can define more message types**
  - E.g., ROUTE-REFRESH [RFC 2918]

Anatomy of an UPDATE

- **Withdrawn routes: list of withdrawn IP prefixes**
- **Network Layer Reachability Information (NLRI)**
  - List of prefixes to which path attributes apply
- **Path attributes**
  - ORIGIN, AS_PATH, NEXT_HOP, MULTI-EXIT-DISC, LOCAL_PREF, ATOMIC_AGGREGATE, AGGREGATOR, …
  - Each attribute has 1-byte type, 1-byte flags, length, content
  - Can introduce new types of path attribute – e.g., AS4_PATH for 32-bit AS numbers

Example

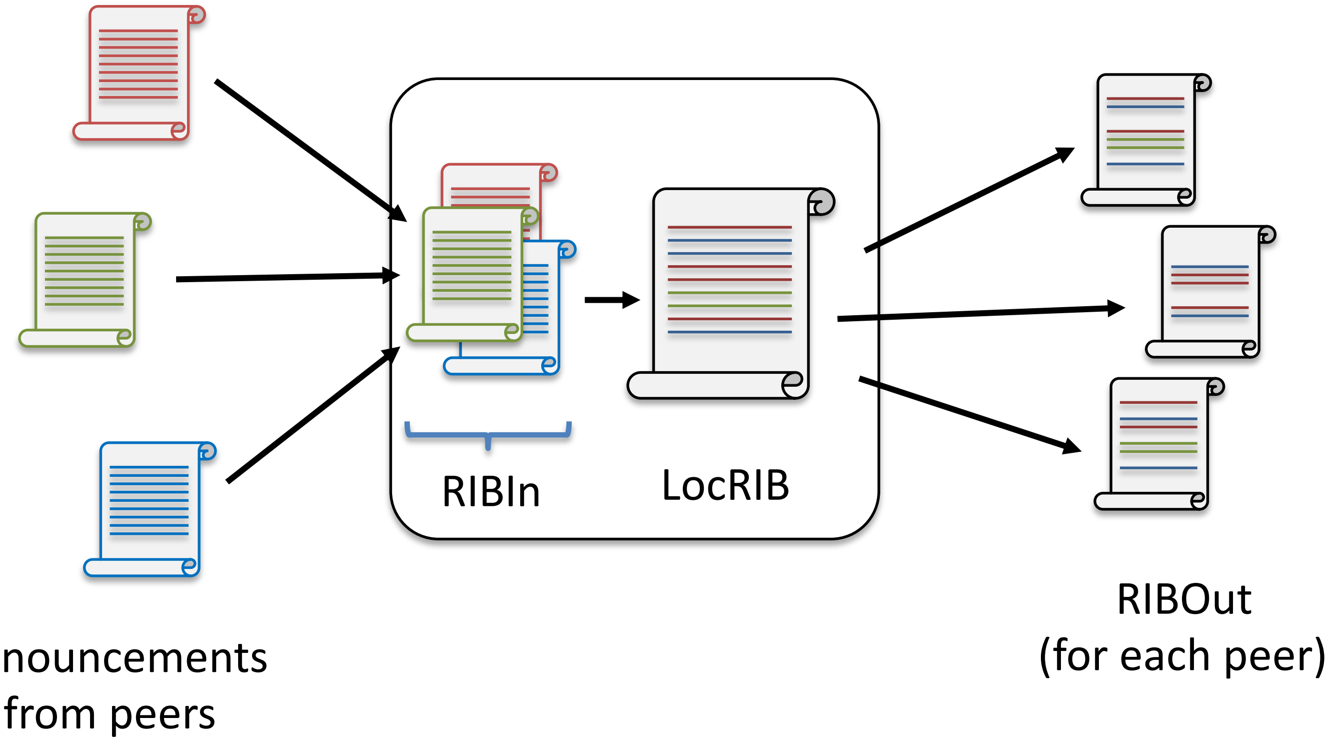
- **NLRI: 128.148.0.0/16**
- **AS Path: ASN 44444 3356 14325 11078**
- **Next Hop IP: same as in RIPv2**
- **Knobs for traffic engineering:**
  - Metric, weight, LocalPath, MED, Communities
  - Lots of voodoo

BGP State

- **BGP speaker conceptually maintains 3 sets of state**
- **Adj-RIB-In**
  - "Adjacent Routing Information Base, Incoming"
  - Unprocessed routes learned from other BGP speakers
- **Loc-RIB**
  - Contains routes from Adj-RIB-In selected by policy
  - First hop of route must be reachable by IGP or static route
- **Adj-RIB-Out**
  - Subset of Loc-RIB to be advertised to peer speakers
  - Can be different for each peer

BGP State

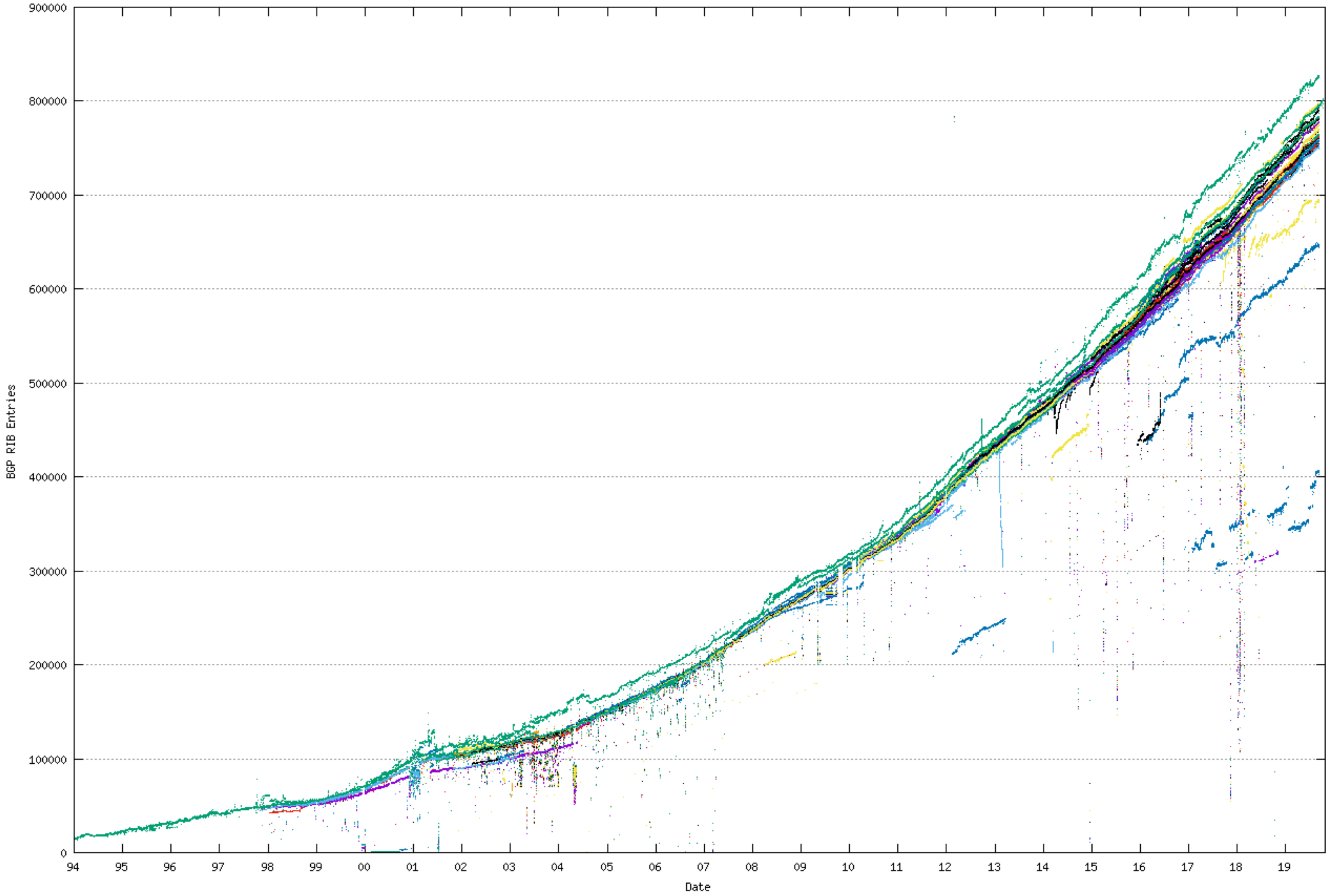
- RIBIn ⊃ LocRIB ⊃ RIBOut



Prefixes

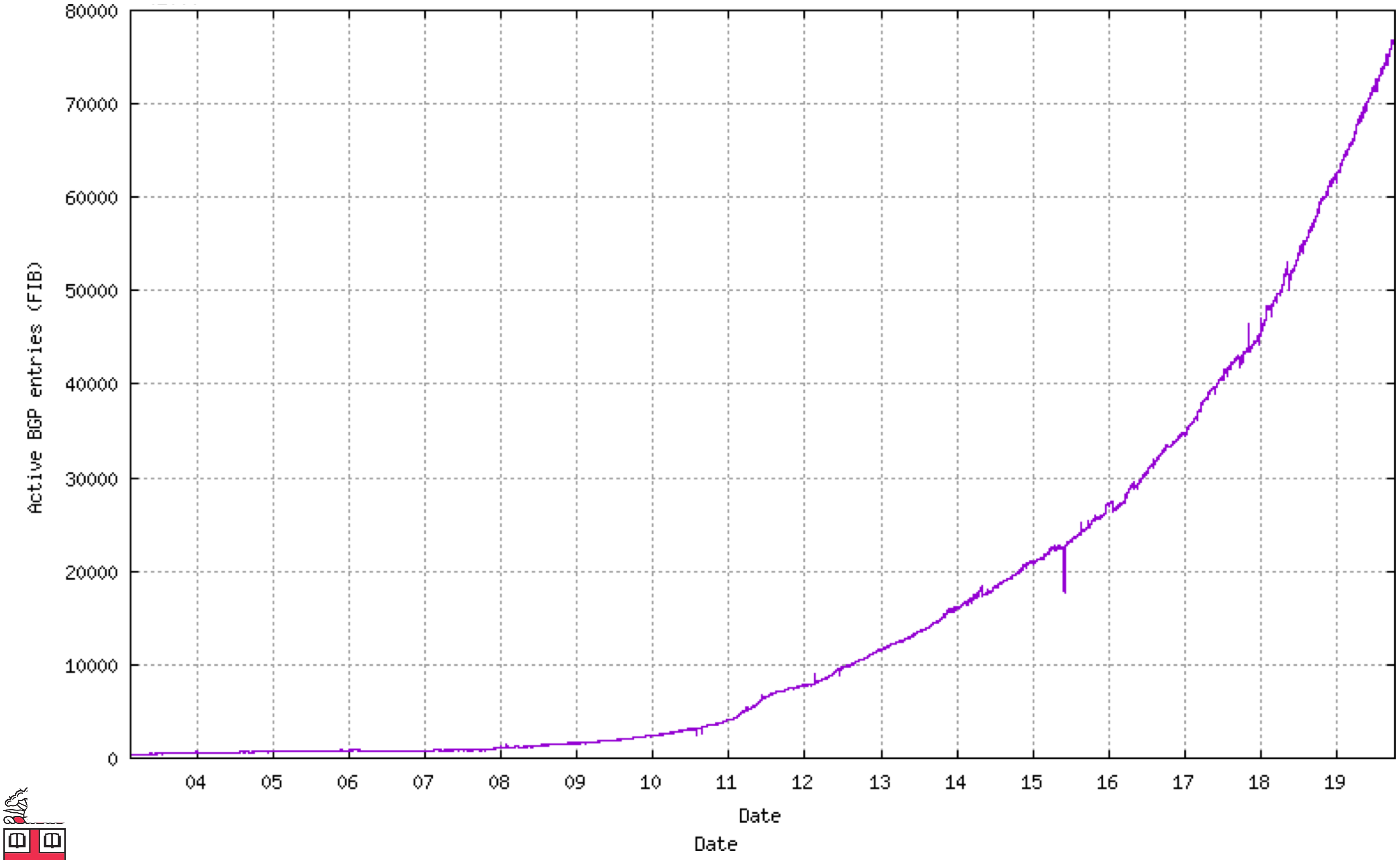
- **Nodes in local network share prefix**
  - Key to decide whether to send message locally
- **Prefixes can also aggregate multiple networks**
  - E.g., 100.20.33.128/25, 100.20.33.0/25 -> 100.20.33.0/24
- **If networks connected hierarchically, can have significant aggregation**
  - What happens if not?

BGP Table Growth



Source: bgp.potaroo.net

BGP Table Growth for v6



Source: bgp.potaroo.net

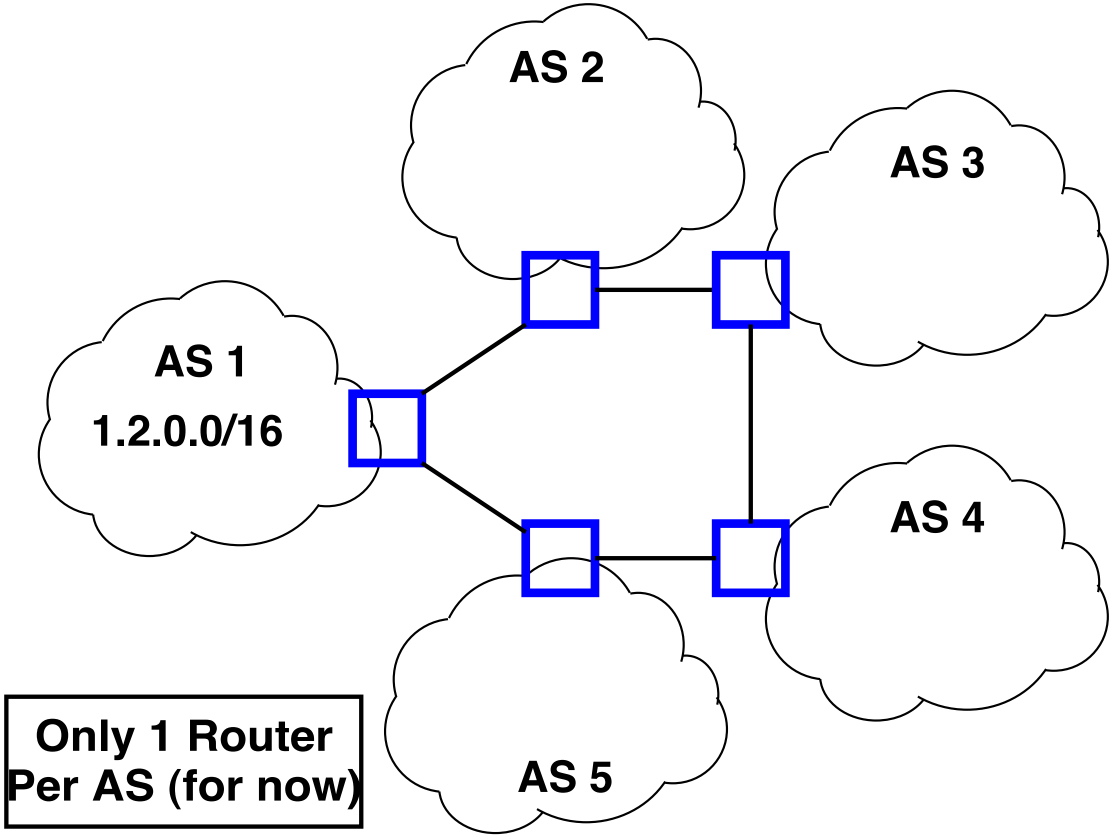
What can lead to table growth?

- **More addresses being allocated**
- **Fragmentation**
  - Multihoming
  - Change of ISPs
  - Address re-selling

Integrating EGP and IGP

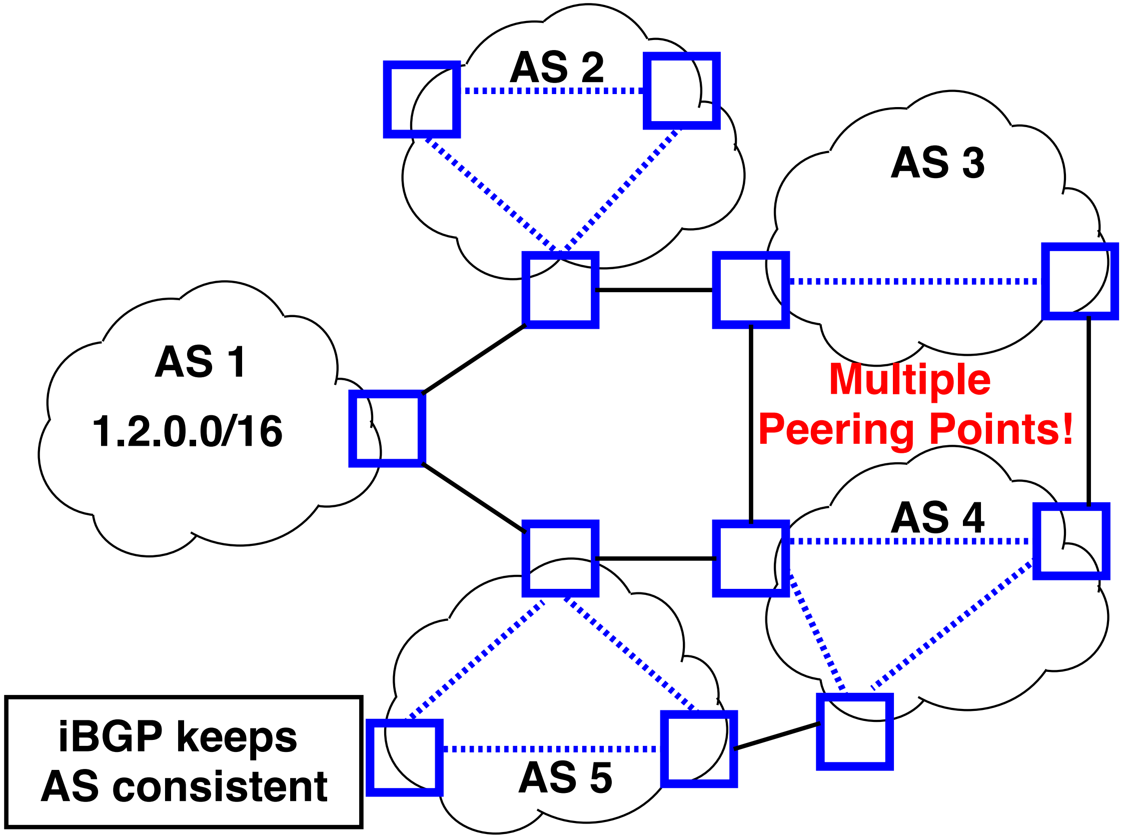
- **How to get internal nodes to route towards external prefixes?**
- **Stub ASs**
  - Border router clear choice for default route
  - Inject into IGP: "any unknown route to border router"
- **Inject specific prefixes in IGP**
  - E.g., Provider injects routes to customer prefix
- **Backbone networks**
  - Too many prefixes for IGP
  - Run internal version of BGP, iBGP
  - All routers learn mappings: Prefix -> Border Router
  - Use IGP to learn: Border Router -> Next Hop

iBGP



Only 1 Router
Per AS (for now)

iBGP

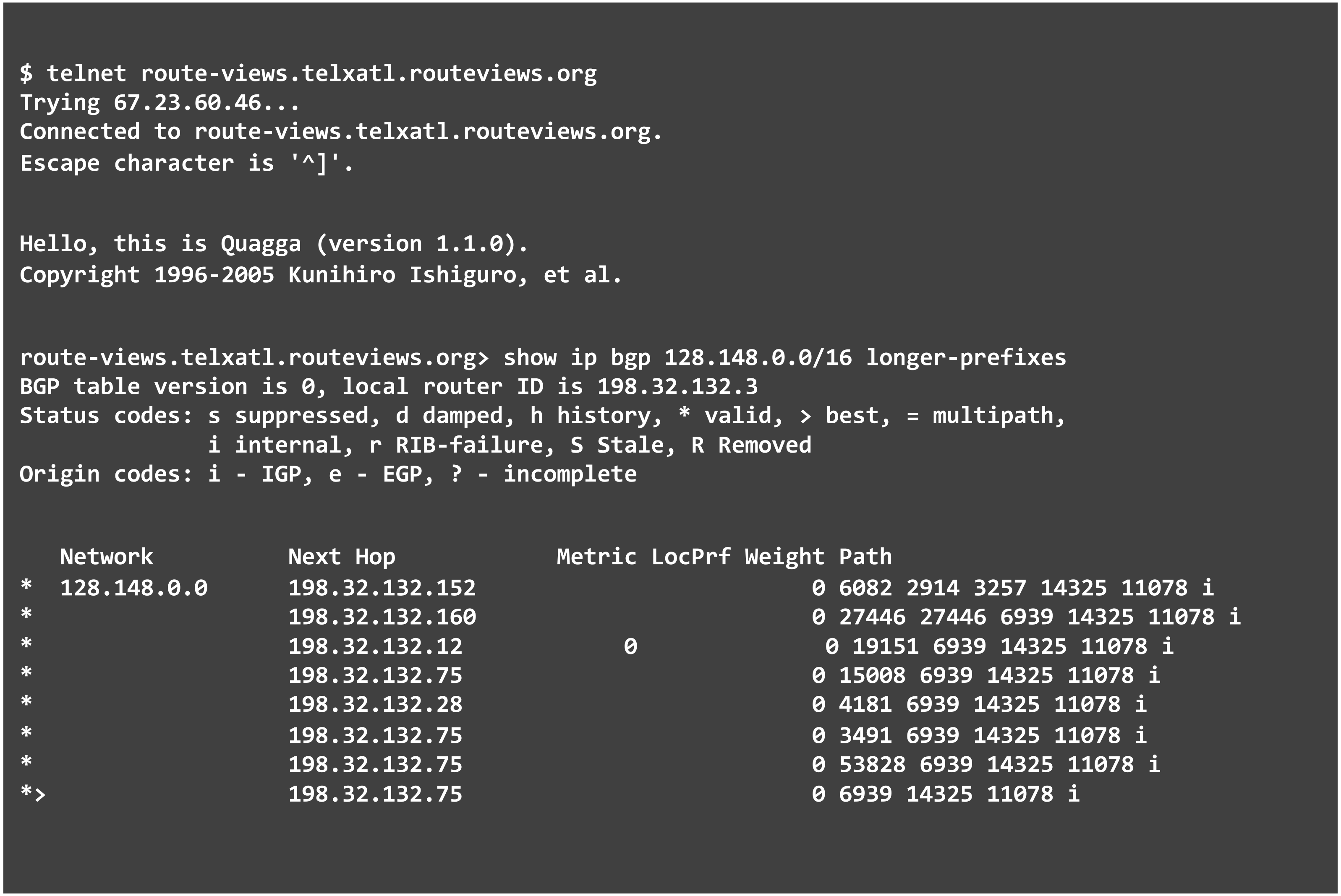


Multiple Peering Points!

iBGP keeps AS consistent

Demo

- **Route views project: http://www.routeviews.org**
  - telnet route-views.linx.routeviews.org
  - show ip bgp 128.148.0.0/16 longer-prefixes
- **All paths are learned internally (iBGP)**
- **Not a production device**



```
$ telnet route-views.telxatl.routeviews.org
Trying 67.23.60.46...
Connected to route-views.telxatl.routeviews.org.
Escape character is '^]'.


Hello, this is Quagga (version 1.1.0).
Copyright 1996-2005 Kunihiro Ishiguro, et al.


route-views.telxatl.routeviews.org> show ip bgp 128.148.0.0/16 longer-prefixes
BGP table version is 0, local router ID is 198.32.132.3
Status codes: s suppressed, d damped, h history, * valid, > best, = multipath,
              i internal, r RIB-failure, S Stale, R Removed
Origin codes: i - IGP, e - EGP, ? - incomplete


   Network          Next Hop            Metric LocPrf Weight Path
*  128.148.0.0      198.32.132.152                         0 6082 2914 3257 14325 11078 i
*                   198.32.132.160                         0 27446 27446 6939 14325 11078 i
*                   198.32.132.12            0              0 19151 6939 14325 11078 i
*                   198.32.132.75                          0 15008 6939 14325 11078 i
*                   198.32.132.28                          0 4181 6939 14325 11078 i
*                   198.32.132.75                          0 3491 6939 14325 11078 i
*                   198.32.132.75                          0 53828 6939 14325 11078 i
*>                  198.32.132.75                          0 6939 14325 11078 i
```

11078 is Brown’s ASN

14325 is Brown’s Provider, OSHEAN

Next class

- BGP Policy Routing and Security