CSCI-1680
Network Layer:
Inter-domain Routing

Rodrigo Fonseca

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

Administrivia

- **Midterm moved up from 3/17 to 3/15**
- **IP due on Friday**

Today

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

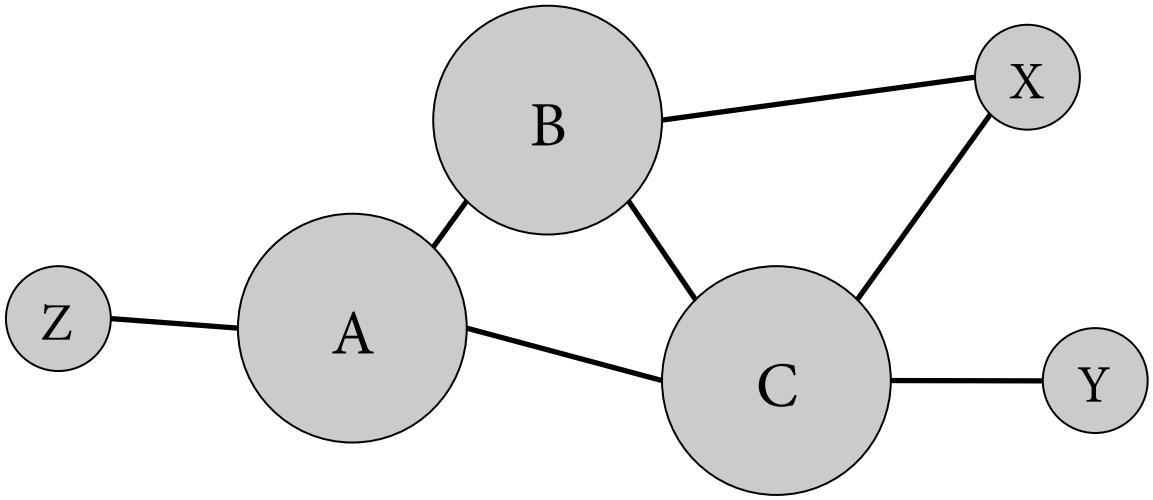
Why Inter vs. Intra

- **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: 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

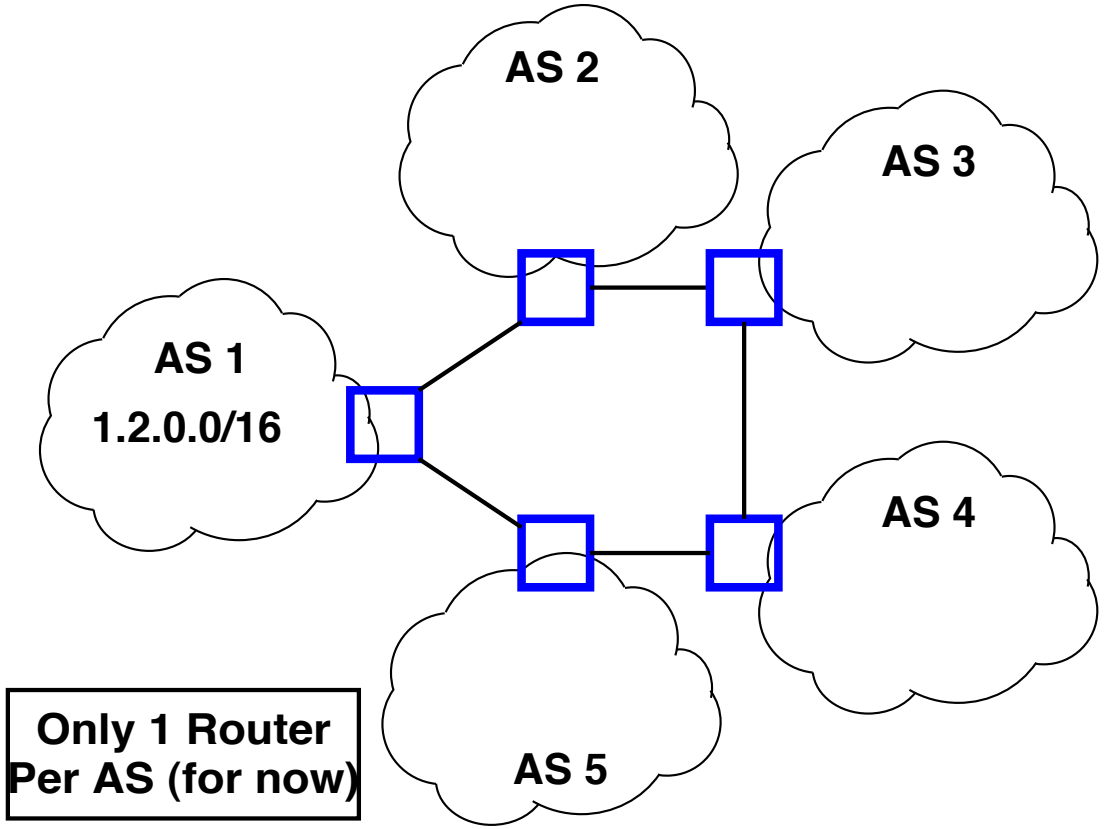
BGP - High Level

- **Single EGP protocol in use today**
- **Abstract each AS to a single node**
- **Destinations are CIDR prefixes**
- **Exchange prefix *reachability* with all neighbors**
  - E.g., "I can reach prefix 128.148.0.0/16 through ASes 44444 3356 14325 11078"
- **Select a single path by routing *policy***
- **Critical: learn many paths, propagate one**
  - Add your ASN to advertised path

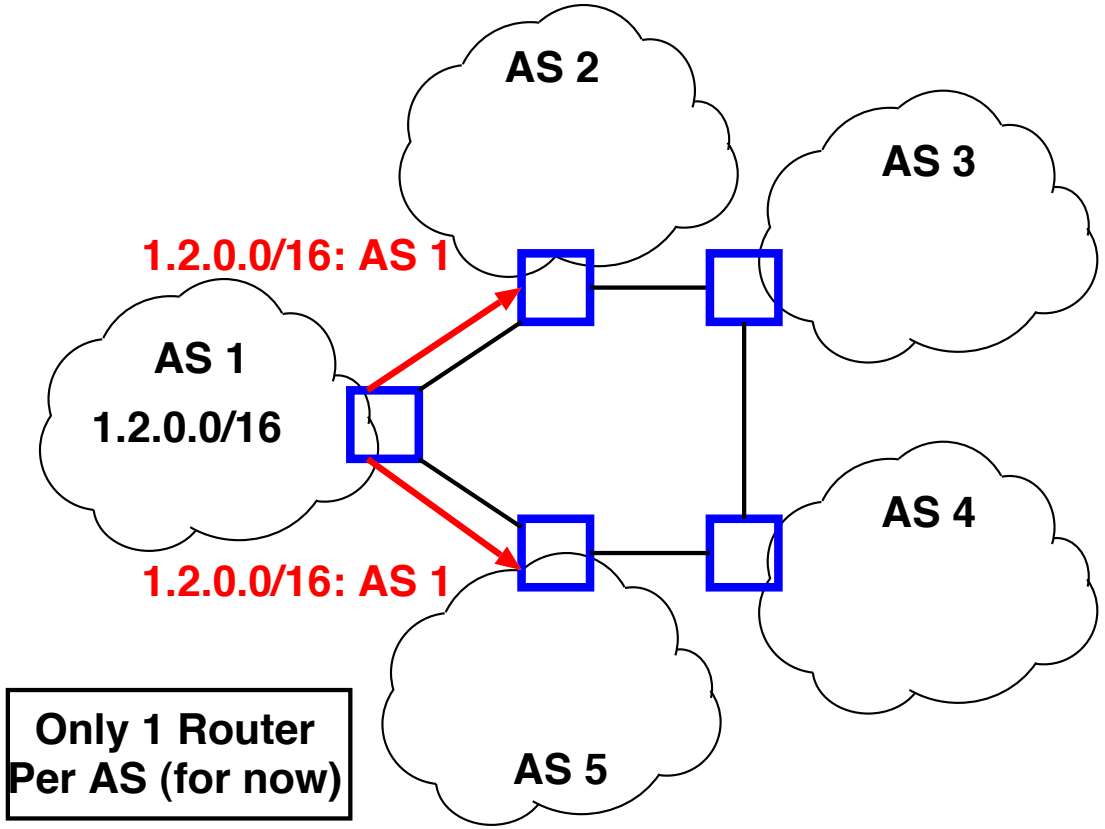
Why study BGP?

- **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

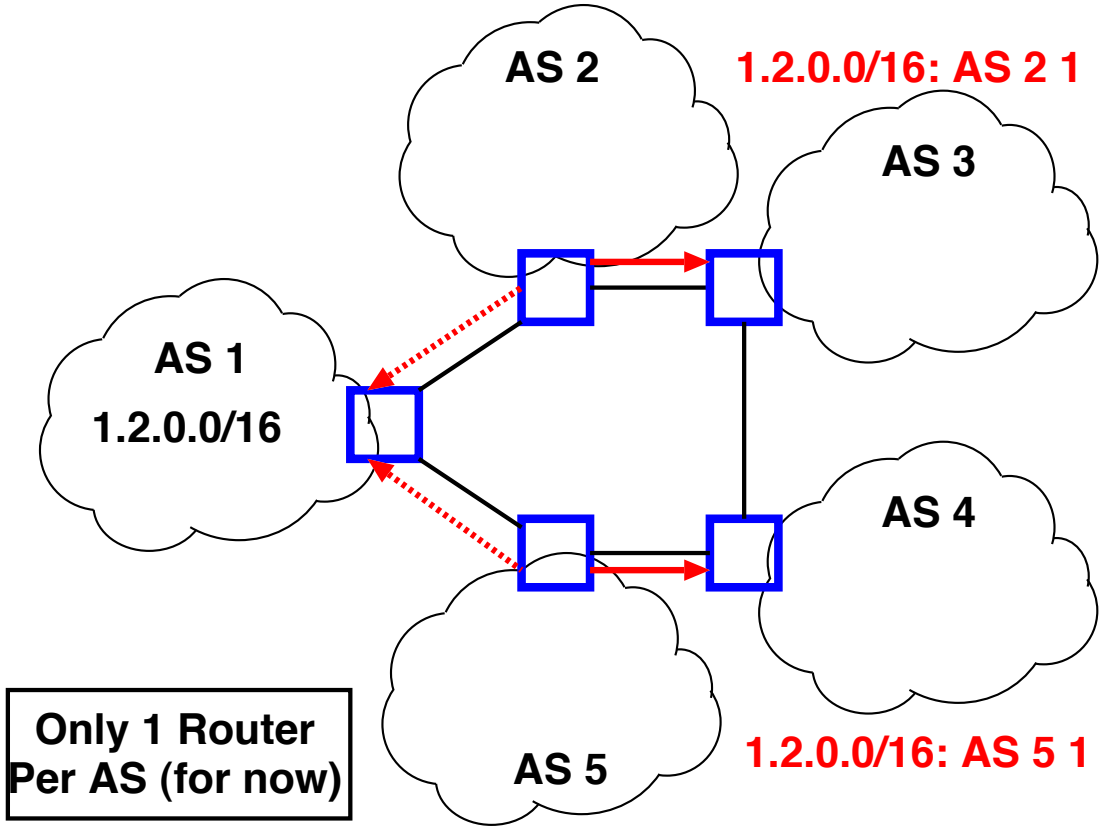
BGP Example



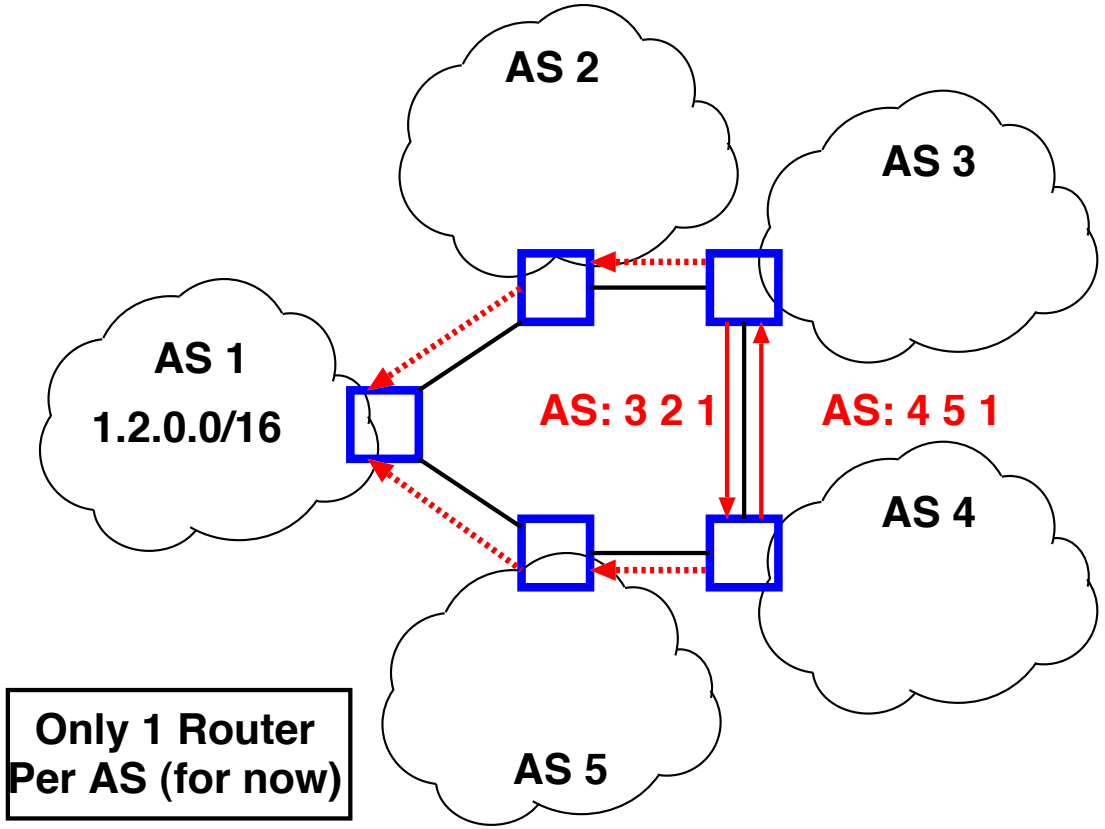
BGP Example


AS 2
AS 3
1.2.0.0/16: AS 1
AS 1
1.2.0.0/16
AS 4
1.2.0.0/16: AS 1
AS 5
Only 1 Router
Per AS (for now)

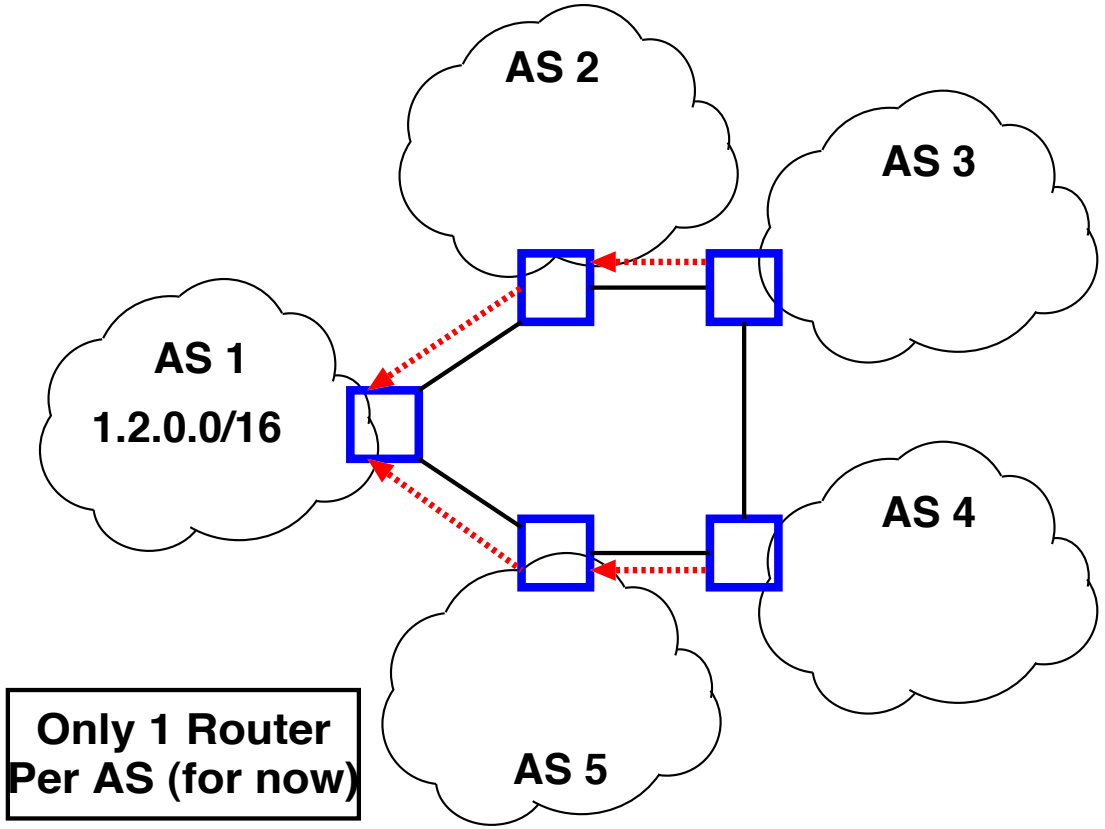
BGP Example


AS 2
1.2.0.0/16: AS 2 1
AS 3
AS 1
1.2.0.0/16
AS 4
AS 5
1.2.0.0/16: AS 5 1
Only 1 Router
Per AS (for now)

BGP Example


AS 2
AS 3
AS 1
1.2.0.0/16
AS: 3 2 1
AS: 4 5 1
AS 4
AS 5
Only 1 Router
Per AS (for now)

BGP Example



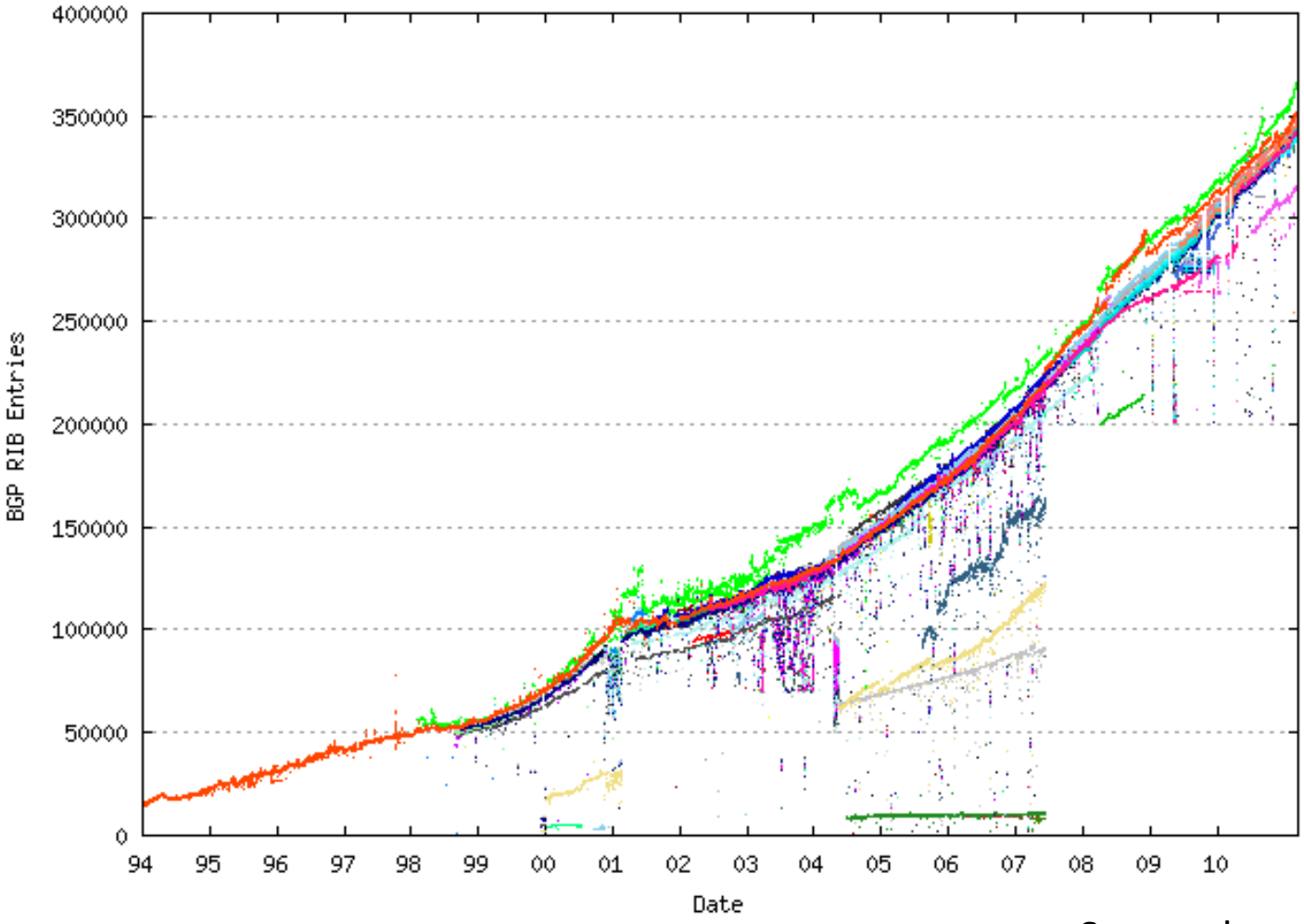
BGP Protocol Details

- **Separate roles of *speakers* and *gateways***
  - Speakers talk BGP with other ASs
  - Gateways are routes that border other Ass
  - 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 Implications

- **Explicit AS Path == Loop free**
  - Except under churn, IGP/EGP mismatch
- **Reachability not guaranteed**
  - Decentralized combination of policies
- **Not all ASs know all paths**
- **AS abstraction -> loss of efficiency**
- **Scaling**
  - 37K ASs
  - 350K+ prefixes
  - ASs with one prefix: 15664
  - Most prefixes by one AS: 3686 (AS6389, BellSouth)

BGP Table Growth

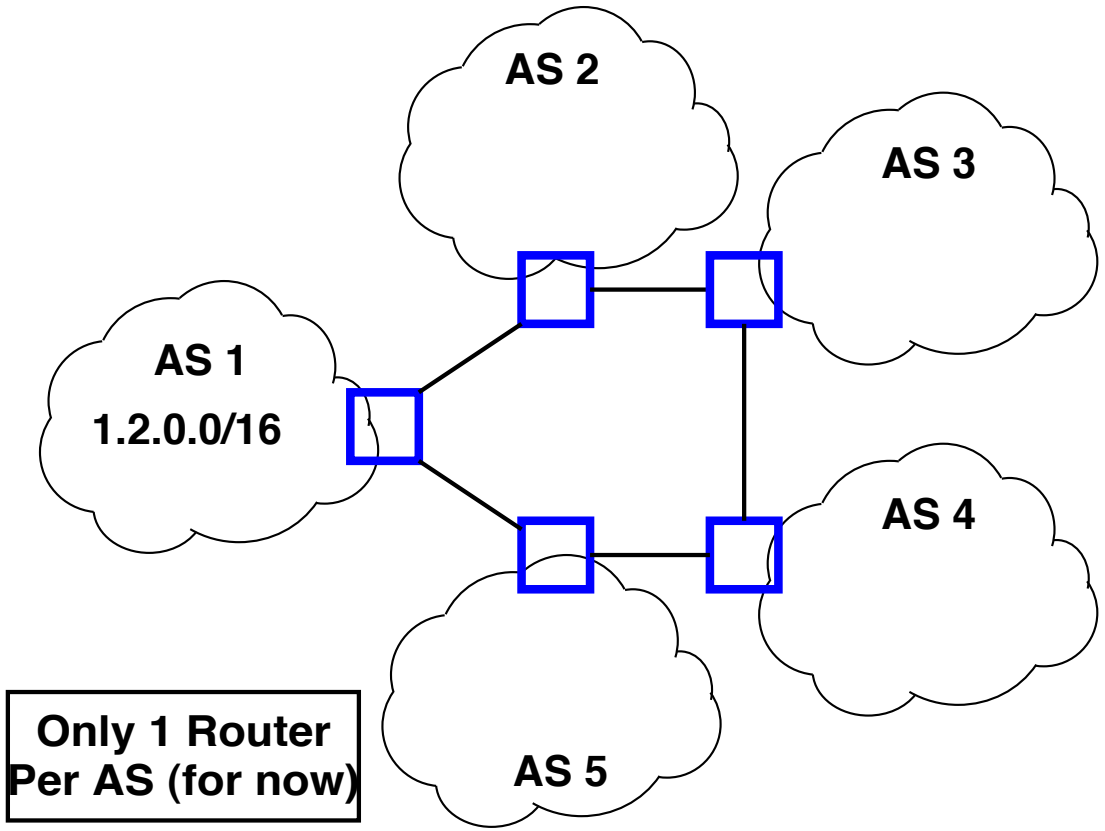


Source: bgp.potaroo.net

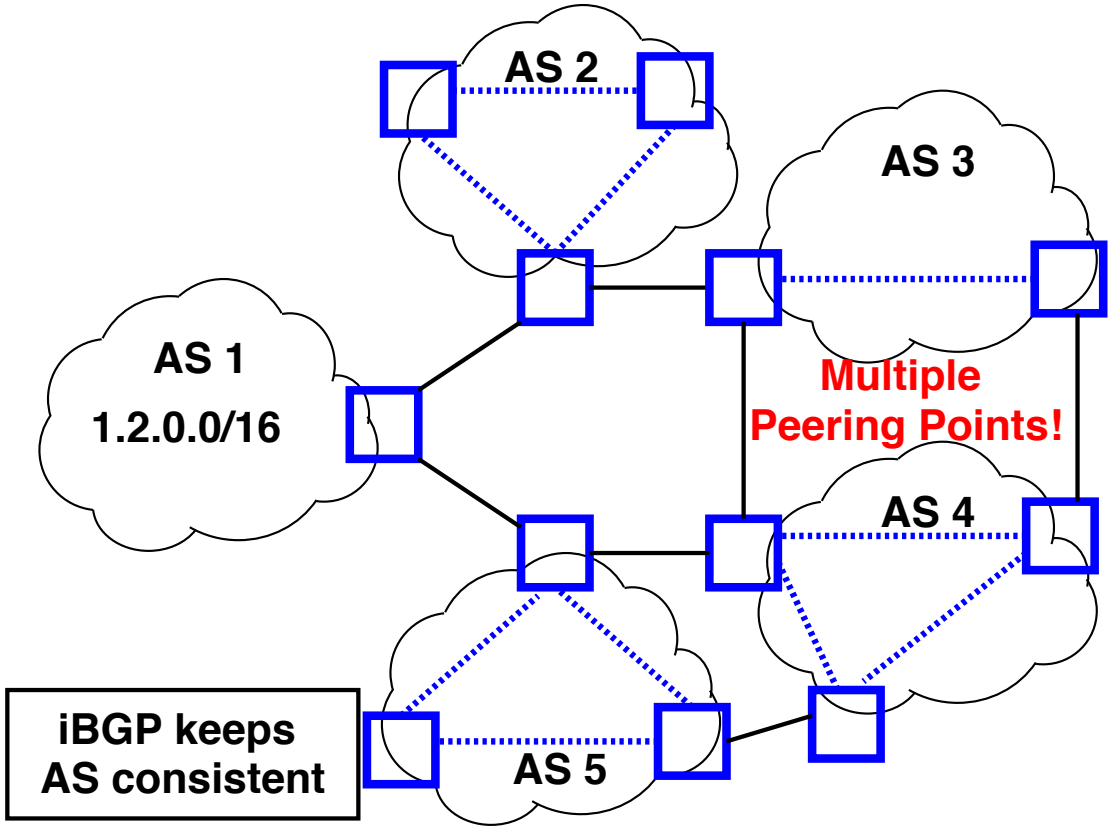
Integrating EGP and IGP

- **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



iBGP


AS 2
AS 3
AS 1
1.2.0.0/16
Multiple
Peering Points!
AS 4
AS 5
iBGP keeps
AS consistent

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

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 path are learned internally (iBGP)**
- **Not a production device**

Route Selection

- **More specific prefix**
- **Next-hop reachable?**
- **Prefer highest weight**
  - Computed using some AS-specific local policy
- **Prefer highest local-pref**
- **Prefer locally originated routes**
- **Prefer routes with shortest AS path length**
- **Prefer eBGP over iBGP**
- **Prefer routes with lowest cost to egress point**
  - Hot-potato routing
- **Tie-breaking rules**
  - E.g., oldest route, lowest router-id

Customer/Provider AS relationships

- **Customer pays for connectivity**
  - E.g. Brown contracts with OSHEAN
  - Customer is stub, provider is a transit
- **Many customers are multi-homed**
  - E.g., OSHEAN connects to Level3, Cogent
- **Typical policy: prefer routes from customers**

Peer Relationships

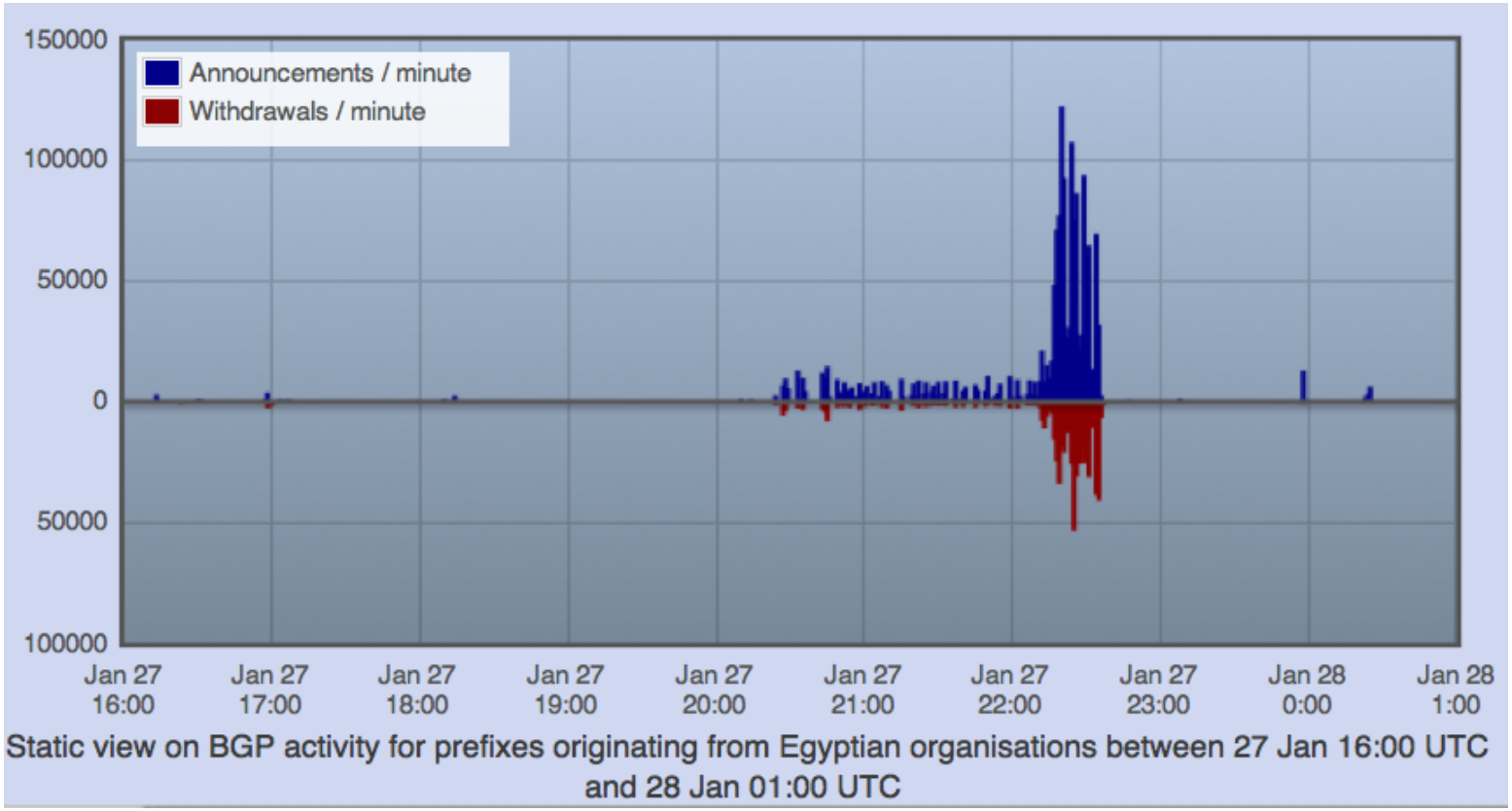
- **ASs agree to exchange traffic for free**
  - Penalties/Renegotiate if imbalance
- **Tier 1 ISPs have no default route: all peer with each other**
- **You are Tier *i* + 1 if you have a default route to a Tier *i***

Peering Drama

- **Cogent vs. Level3 were peers**
- **In 2003, Level3 decided to start charging Cogent**
- **Cogent said no**
- **Internet partition: Cogent's customers couldn't get to Level3's customers and vice-versa**
  - Other ISPs were affected as well
- **Took 3 weeks to reach an undisclosed agreement**

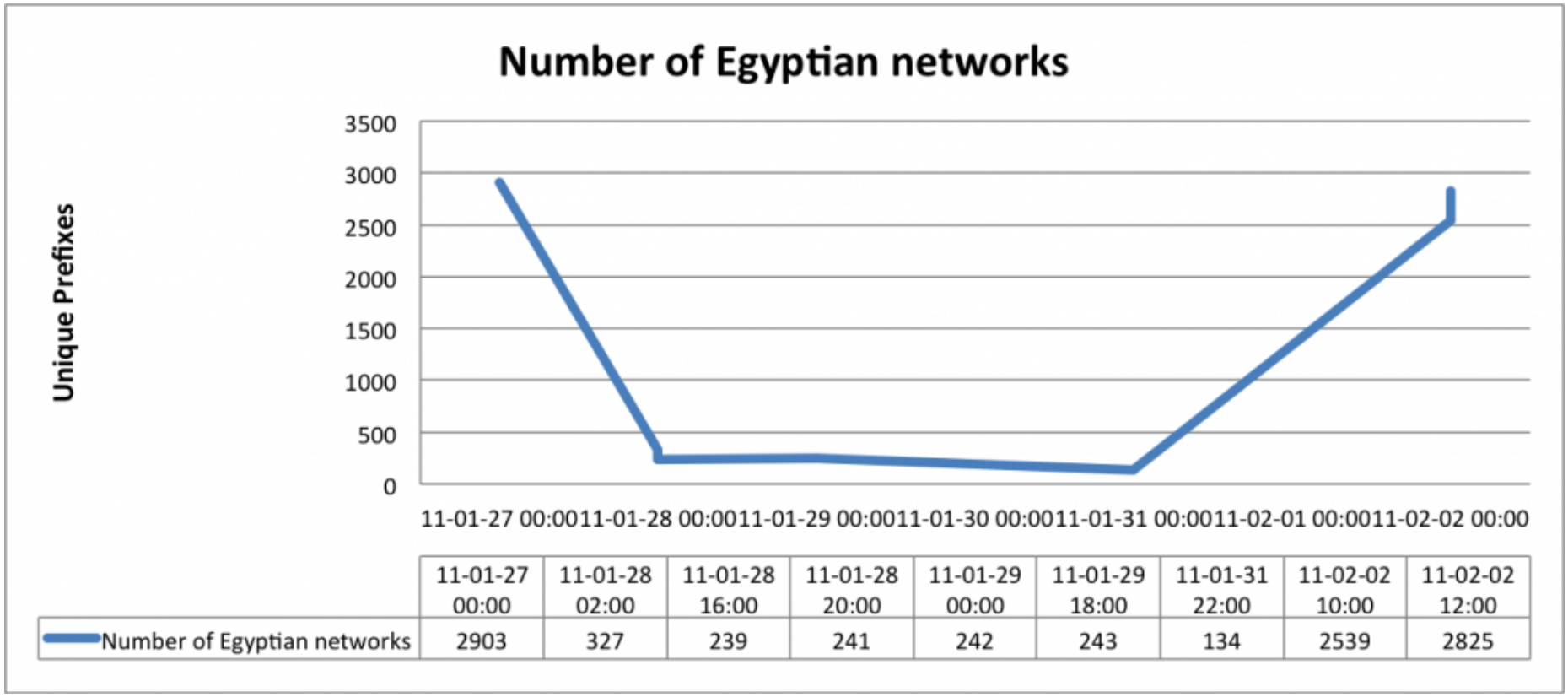
"Shutting off" the Internet

- **Starting from Jan 27th, 2011, Egypt was disconnected from the Internet**
  - 2769/2903 networks withdrawn from BGP (95%!



Source: RIPEStat - http://stat.ripe.net/egypt/

Egypt Incident



	11-01-27 00:00	11-01-28 02:00	11-01-28 16:00	11-01-28 20:00	11-01-29 00:00	11-01-29 18:00	11-01-31 22:00	11-02-02 10:00	11-02-02 12:00
Number of Egyptian networks	2903	327	239	241	242	243	134	2539	2825

Source: BGPMon (http://bgpmon.net/blog/?p=480)

Some BGP Challenges

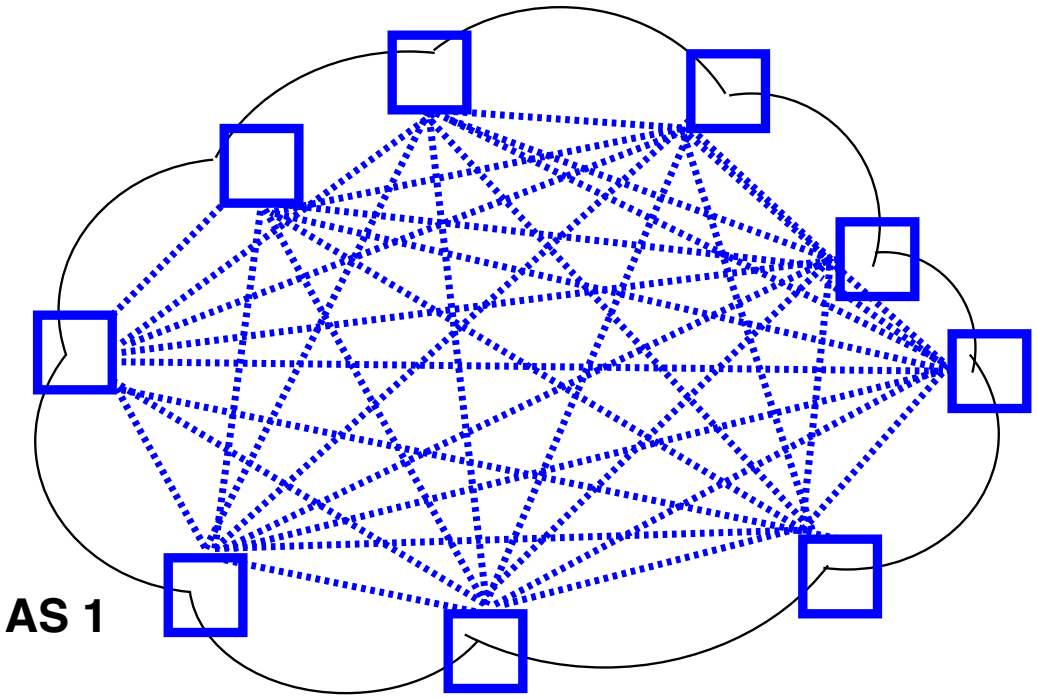
- **Convergence**
- **Scaling (route reflectors)**
- **Traffic engineering**
  - How to assure certain routes are selected
- **Security**

Convergence

- **Given a change, how long until the network re-stabilizes?**
  - Depends on change: sometimes never
  - Open research problem: “tweak and pray”
  - Distributed setting is challenging
- **Easier: is there a stable configuration?**
  - Distributed: open research problem
  - Centralized: NP-Complete problem!
  - Multiple stable solutions given policies (e.g. “Wedgies”, RFC 4264)

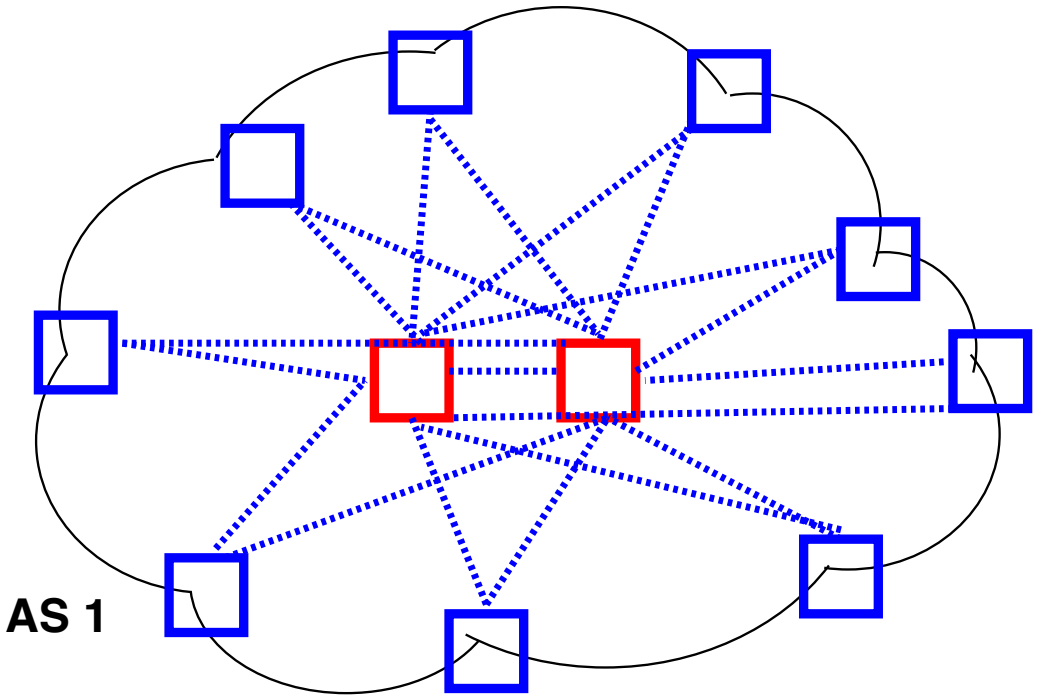
Scaling iBGP: route reflectors

iBGP Mesh == O(n^2) mess



Scaling iBGP: route reflectors

Solution: Route Reflectors
O(n*k)



Route Engineering

- Route filtering
- Setting weights
- More specific routes: longest prefix
- AS prepending: “477 477 477 477”
- More of an art than science

BGP Security

- **Anyone can source a prefix announcement!**
  - To say BGP is insecure is an understatement ☺
- **Pakistan Youtube incident**
  - Youtube's has prefix 208.65.152.0/22
  - Pakistan's government order Youtube blocked
  - Pakistan Telecom (AS 17557) announces 208.65.153.0/24 in the wrong direction (outwards!)
  - Longest prefix match caused worldwide outage
- **http://www.youtube.com/watch?v=IzLPKuAOe50**

Many other incidents

- **Spammers steal unused IP space to hide**
  - Announce very short prefixes
  - For a short amount of time
- **China incident, April 8th 2010**
  - China Telecom's AS23724 generally announces 40 prefixes
  - On April 8th, announced ~37,000 prefixes
  - About 10% leaked outside of China
  - Suddenly, going to www.dell.com might have you routing through AS23724!
- **Secure BGP is in the works**

BGP Recap

- Key protocol that holds Internet routing together
- Path Vector Protocol among Autonomous Systems
- Policy, feasibility first; non-optimal routes
- Important security problems

Next Lecture

- **Network layer wrap-up**
  - IPv6
  - Multicast
  - MPLS
- **Next Chapter: Transport Layer (UDP, TCP,…)**