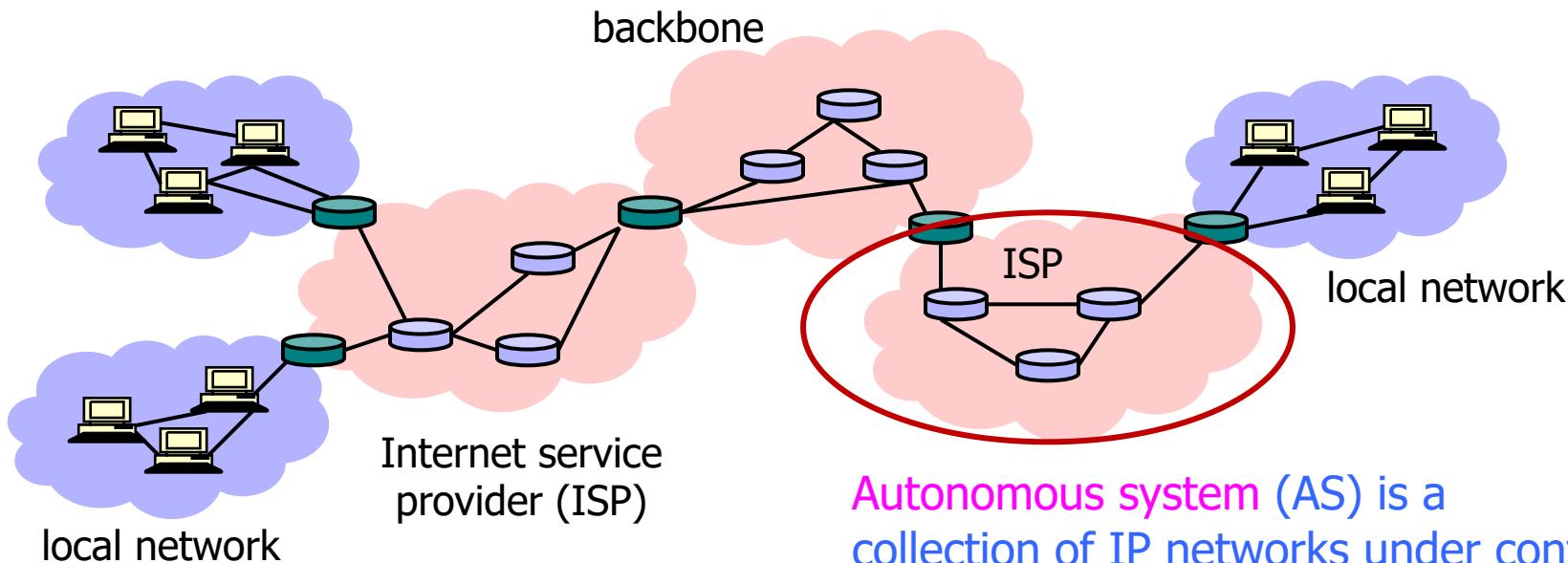


CS 5450

# Inter-Domain Routing

Vitaly Shmatikov

# Internet Is a Network of Networks

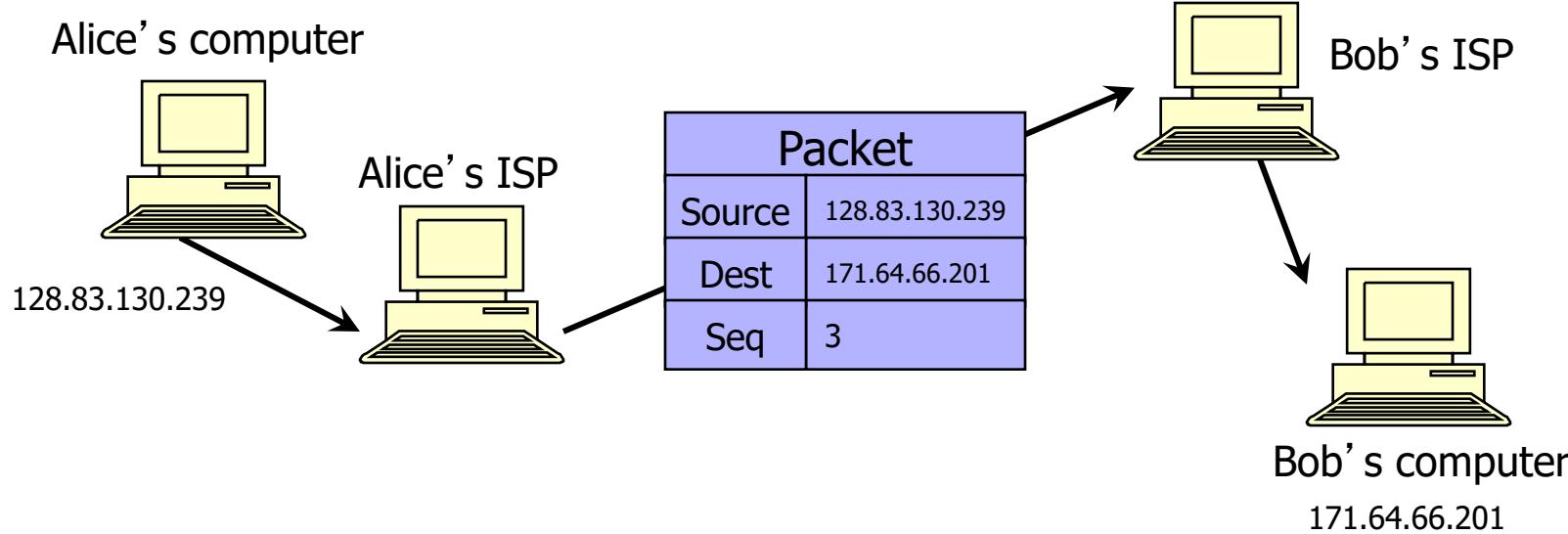


**Autonomous system (AS)** is a collection of IP networks under control of a single administrator (e.g., ISP)

- ◆ TCP/IP for packet routing and connections
- ◆ Border Gateway Protocol (BGP) for route discovery
- ◆ Domain Name System (DNS) for IP address discovery

# IP (Internet Protocol)

- ◆ Connectionless
  - Unreliable, “best-effort” protocol
- ◆ Uses numeric addresses for routing
- ◆ Typically several hops in the route



# IP Routing

---

- ◆ Routing of IP packets is based on IP addresses
  - 32-bit host identifiers (128-bit in IPv6)
- ◆ Routers use a forwarding table
  - Entry = destination, next hop, network interface, metric
  - Table look-up for each packet to decide how to route it
- ◆ Routers learn routes to hosts and networks via routing protocols
  - Host is identified by IP address, network by IP prefix
- ◆ **BGP** (Border Gateway Protocol) is the core Internet protocol for establishing inter-AS routes

# Routing Table vs. Forwarding Table

---

## ◆ Forwarding table

- Used when a packet is being forwarded
- A row in the forwarding table contains the mapping from a network number to an outgoing interface and some MAC information, such as the Ethernet address of the next hop

## ◆ Routing table

- Built by the routing algorithm as a precursor to build the forwarding table
- Generally contains mapping from network numbers to next hops

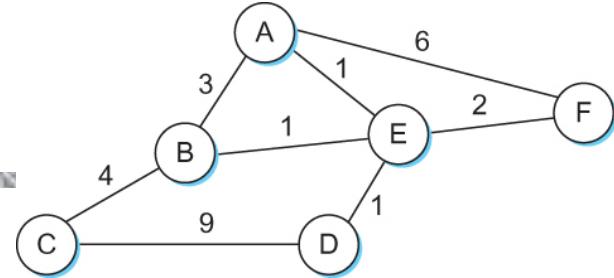
# Routing Table vs. Forwarding Table

(a)		Routing table
Prefix/Length	Next Hop	
18/8	171.69.245.10	

(b)	Prefix/Length	Interface	MAC Address
	18/8	if0	8:0:2b:e4:b:1:2

# Routing Problem



- ◆ Network as a graph
- ◆ Basic problem: find the lowest-cost path between any two nodes
  - Cost of a path = sum of the costs of all the edges that make up the path
- ◆ Topology is dynamic: need a distributed and dynamic protocol
- ◆ Two main classes of protocols
  - Distance Vector
  - Link State

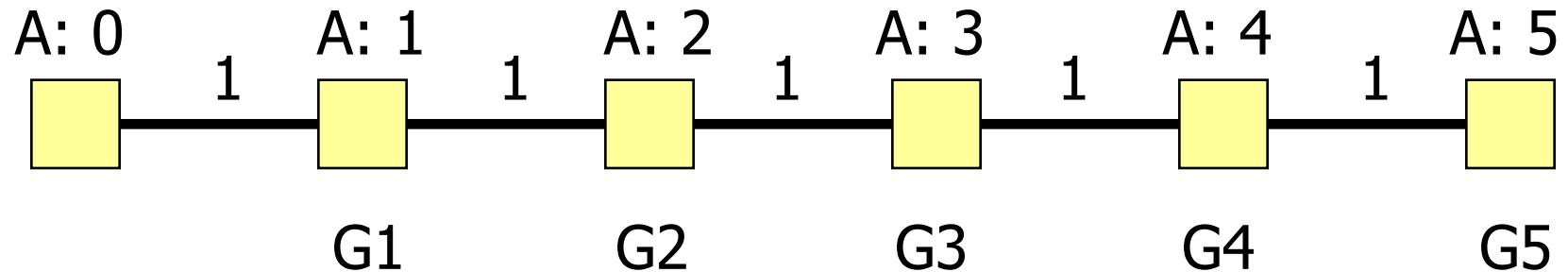
# Distance-Vector Routing

---

- ◆ Each node keeps vector with distances to all nodes
- ◆ Periodically sends distance vector to all neighbors
- ◆ Neighbors send their distance vectors, too; node updates its vector based on received information
  - Bellman-Ford algorithm: for each destination, router picks the neighbor advertising the cheapest route, adds his entry into its own routing table and re-advertises
  - Used in RIP (routing information protocol)
- ◆ Split-horizon update
  - Do not advertise a route on an interface from which you learned the route in the first place!

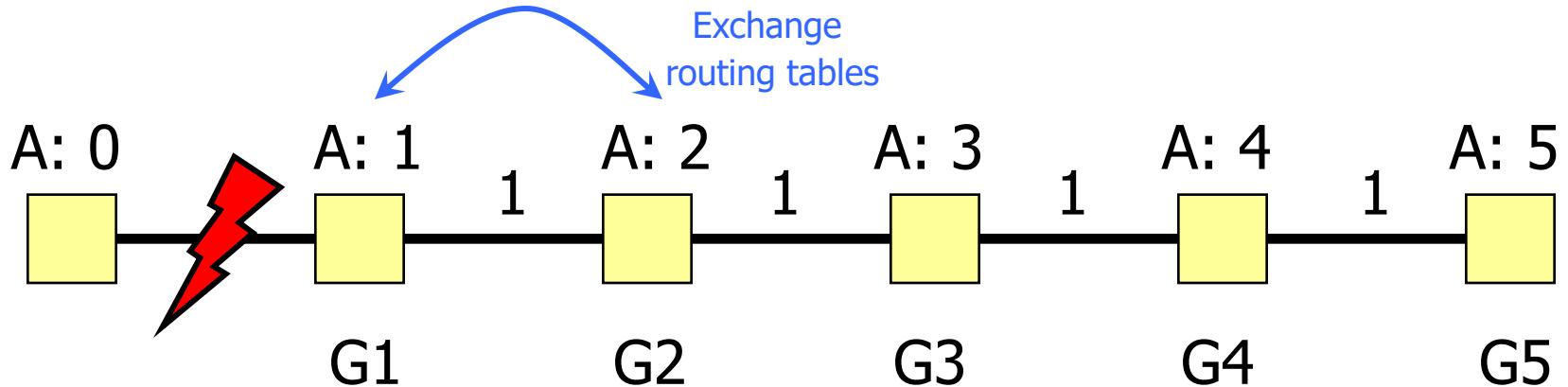
# Good News Travels Fast

---



- ◆ G1 advertises route to network A with distance 1
- ◆ G2-G5 quickly learn the good news and install the routes to A via G1 in their local routing tables

# Bad News Travels Slowly



- ◆ G1's link to A goes down
- ◆ G2 is advertising a pretty good route to G1 (cost=2)
- ◆ G1's packets to A are forever looping between G2 and G1
- ◆ G1 is now advertising a route to A with cost=3, so G2 updates its own route to A via G1 to have cost=4, and so on
  - G1 and G2 are slowly counting to infinity
  - Split-horizon updates only prevent two-node loops

# Link-State Routing

---

- ◆ Strategy: send to all nodes (not just neighbors) information about directly connected links (not entire routing table)
- ◆ Link-state packet (LSP)
  - ID of the node that created the LSP
  - Cost of link to each directly connected neighbor
  - Sequence number (SEQNO)
  - Time-to-live (TTL) for this packet

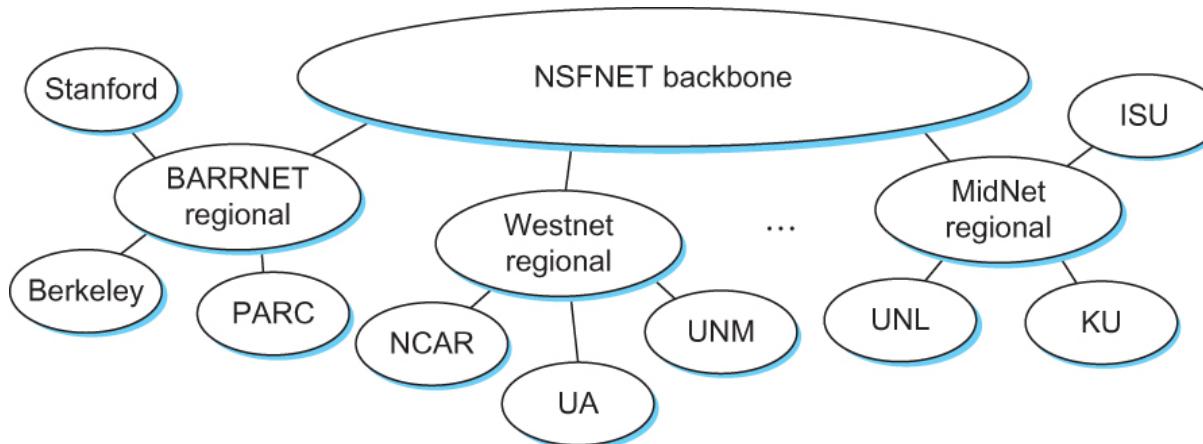
# Reliable Flooding

---

- ◆ Store most recent LSP from each node
- ◆ Forward LSP to all nodes but one that sent it
- ◆ Generate new LSP periodically; increment SEQNO
- ◆ Start SEQNO at 0 when reboot
- ◆ Decrement TTL of each stored LSP, discard when TTL=0

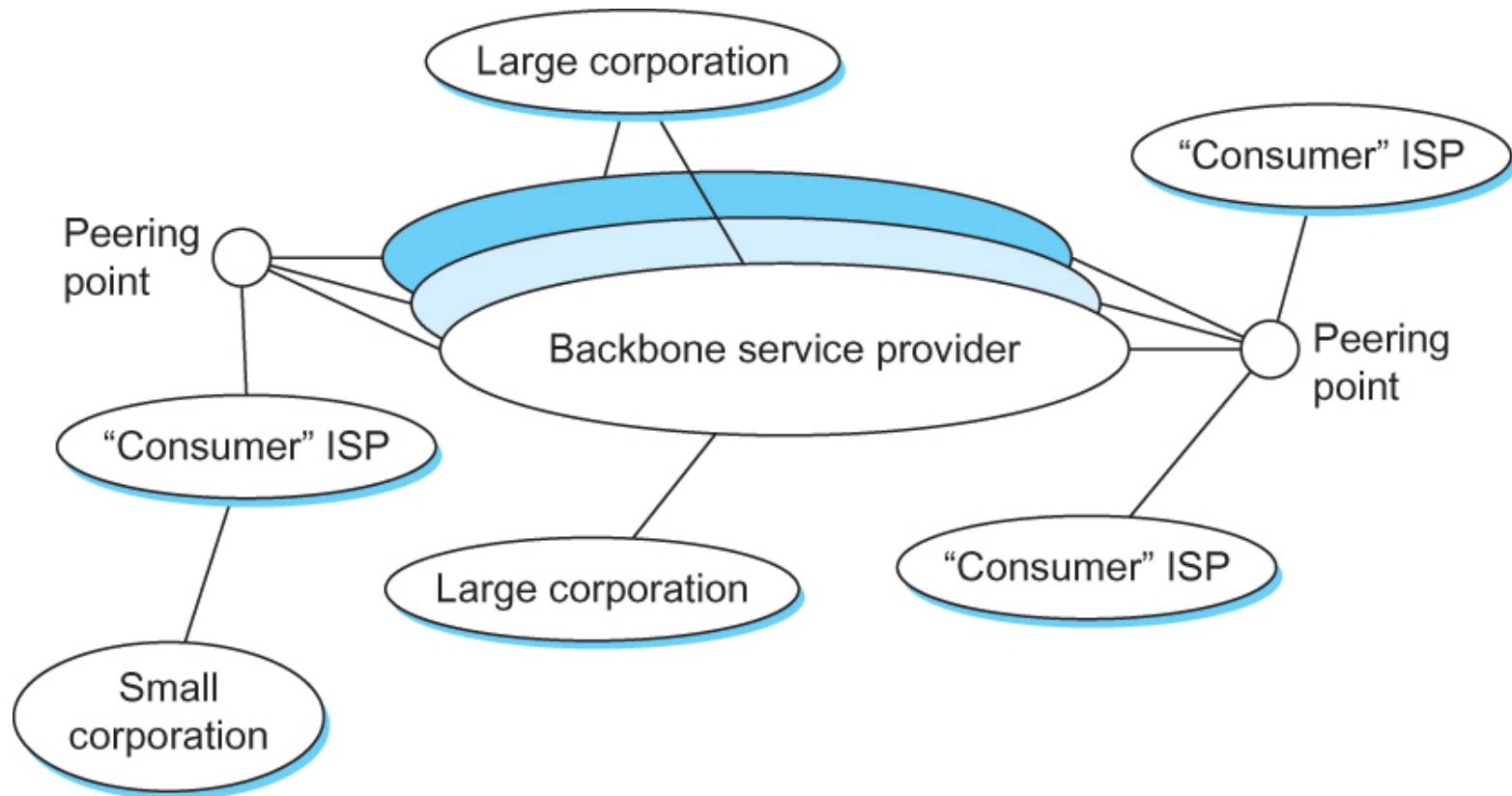
# Inter-AS Routing

- ◆ Neither DV, nor LS scaled even to the Internet as it looked like in the 1990s



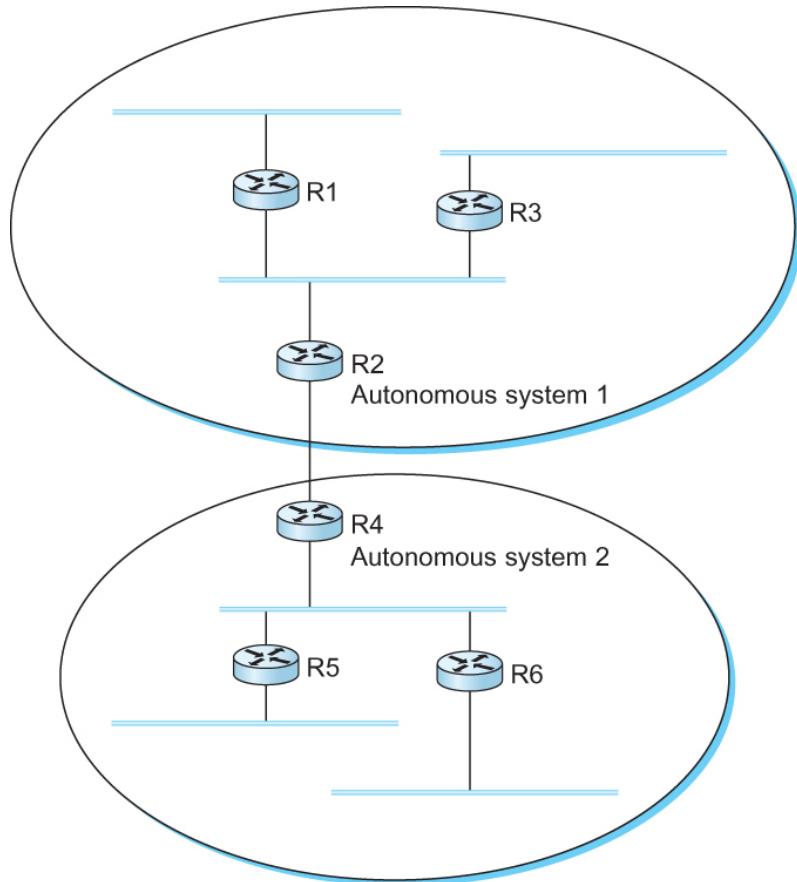
- ◆ Today need to handle hundreds of thousands of networks, billions of end nodes

# Simple Model of Global Internet



# Interdomain Routing

- ◆ Internet is organized as autonomous systems (AS), each of which is under the control of a single administrative entity
- ◆ Autonomous System (AS) corresponds to an administrative domain
  - Examples: university network, corporate internal network, backbone network, network of a single Internet service provider

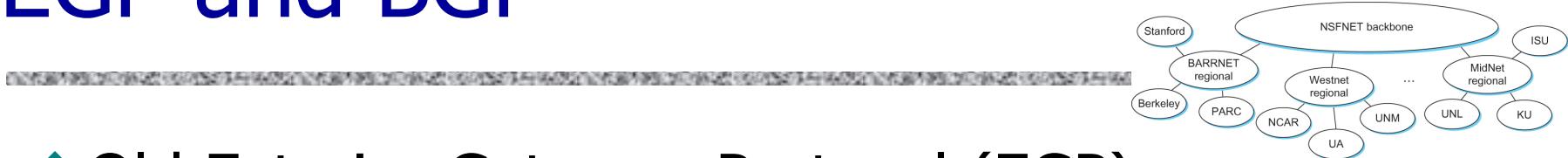


# Route Propagation

---

- ◆ Idea: to improve scalability, hierarchically aggregate routing information in a large internet
- ◆ Divide the routing problem in two parts
  - Routing within a single autonomous system
  - Routing between autonomous systems
- ◆ Two-level route propagation hierarchy
  - Inter-domain routing protocol (Internet-wide standard)
  - Intra-domain routing protocol (each AS selects its own)

# EGP and BGP

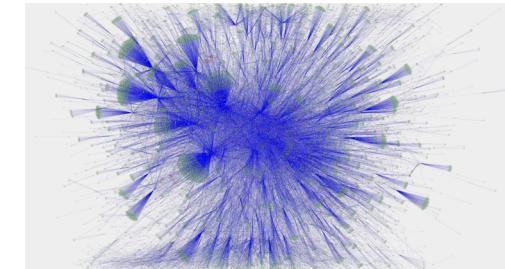


## ◆ Old Exterior Gateway Protocol (EGP)

- Forced a tree-like topology onto the Internet
  - There is a single backbone and autonomous systems are connected only as parents and children and not as peers
- Did not allow for the topology to become general

## ◆ Border Gateway Protocol (BGP)

- Assumes that the Internet is an arbitrarily interconnected set of ASs
  - Today's Internet consists of an interconnection of multiple backbone networks (service provider networks operated by private companies rather than the government)
  - Sites are connected to each other in arbitrary ways



# Inter-AS Connections

---

- ◆ Some large corporations connect directly to one or more of the backbones, while others connect to smaller, non-backbone service providers
- ◆ Many service providers exist mainly to provide service to “consumers” (individuals with computers and devices), and these providers must connect to the backbone providers
- ◆ Often many providers arrange to interconnect with each other at a single “peering point”

# Types of Autonomous Systems

---

- ◆ **Local traffic** = traffic that originates at or terminates on nodes within an AS
- ◆ **Transit traffic** = traffic that passes through an AS
- ◆ Three types of ASs
  - **Stub AS** has only a single connection to one other AS; such an AS will only carry local traffic (small corp)
  - **Multihomed AS** has connections to more than one other AS, but refuses to carry transit traffic (large corp)
  - **Transit AS** has connections to more than one other AS, and is designed to carry both transit and local traffic (backbone provider)

# Goals of BGP

---

- ◆ The goal of Inter-domain routing is to find any path to the intended destination that is loop-free
- ◆ Reachability rather than optimality!
  - ◆ Finding path anywhere close to optimal is considered to be a great achievement (why?)

# BGP Challenges

---

- ◆ Scalability: a backbone router must be able to forward any packet to anywhere in the Internet
  - Routing table must have a match for any valid IP address
- ◆ Full autonomy of the domains
  - Impossible to calculate meaningful cost for a path that crosses multiple ASes
  - A cost of 1000 might imply a great path from one provider and unacceptably bad from another provider
- ◆ Providers may not trust each other
  - One provider may not believe another's route advertisements

# BGP Speakers

---

Each AS has...

- ◆ One BGP **speaker** that advertises:
  - local networks
  - other reachable networks (transit AS only)
  - path information
- ◆ One or more **border gateways** (need not be the same as speakers)
  - Routers through which packets enter and leave the AS

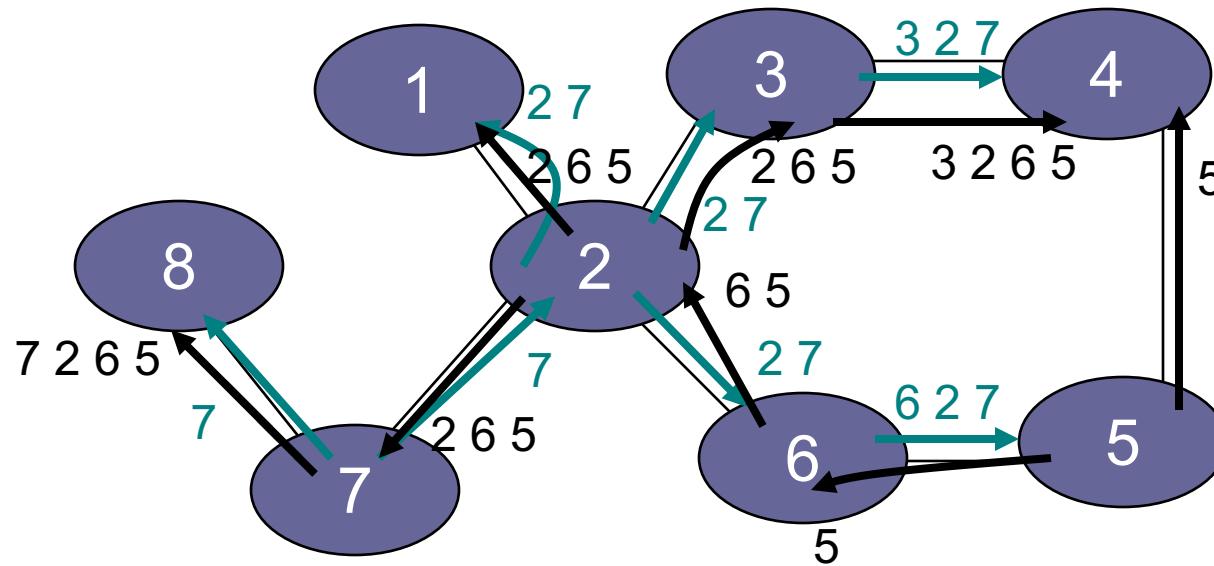
# Overview of BGP

---

- ◆ BGP is a **path-vector** protocol between ASes
- ◆ Just like distance-vector, but routing updates contain an actual path to destination node
  - The list of traversed ASes and the set of network prefixes belonging to the first AS on the list
- ◆ Each BGP router receives update messages from neighbors, selects one “best” path for each prefix, and advertises this path to its neighbors
  - Can be the shortest path, but doesn’t have to be
    - “Hot-potato” vs. “cold-potato” routing
  - Always route to the **most specific prefix** for a destination

# BGP Example

[Wetherall]



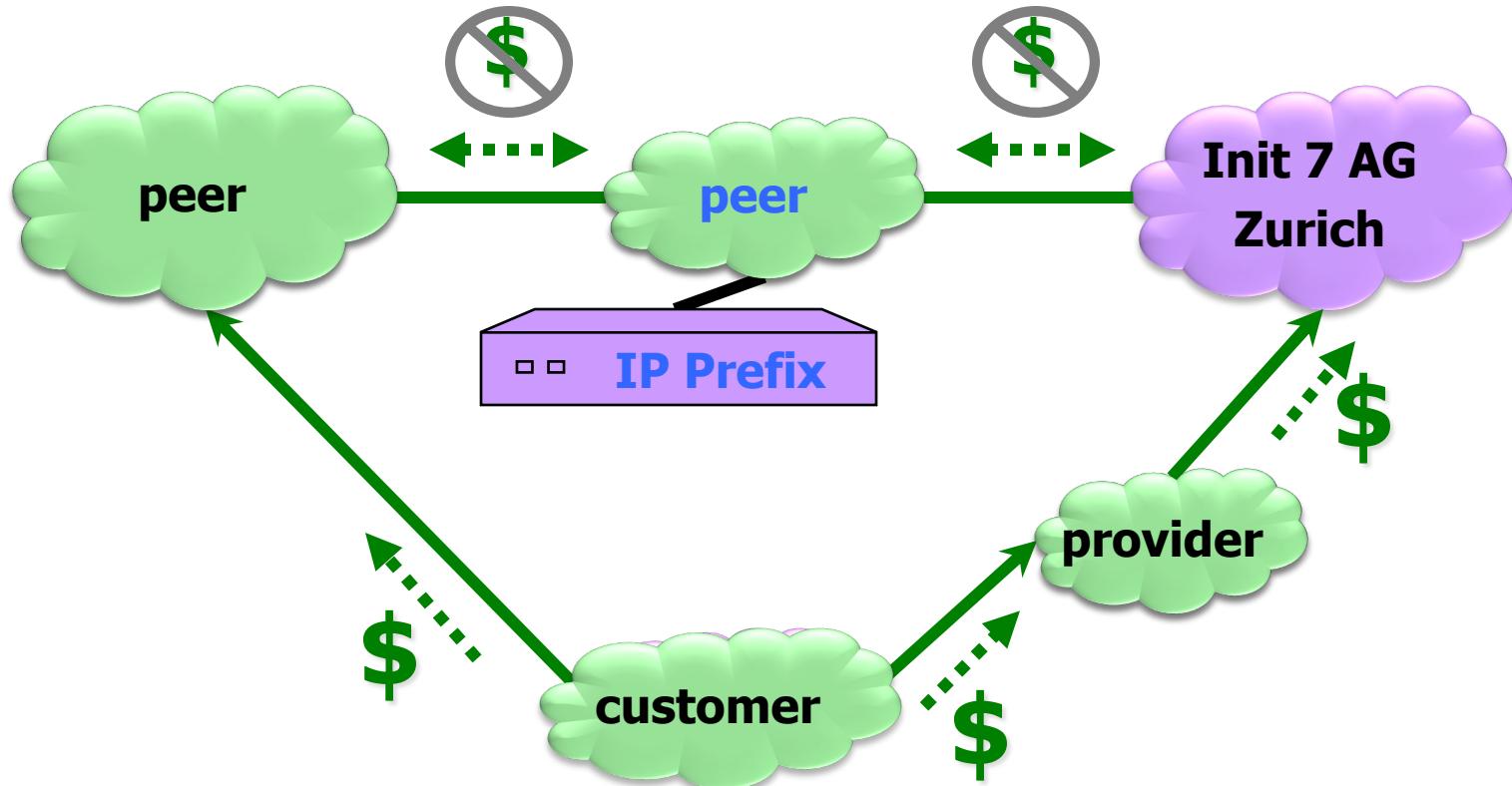
- ◆ AS 2 provides **transit** for AS 7
  - Traffic to and from AS 7 travels through AS 2

# Some (Old) BGP Statistics

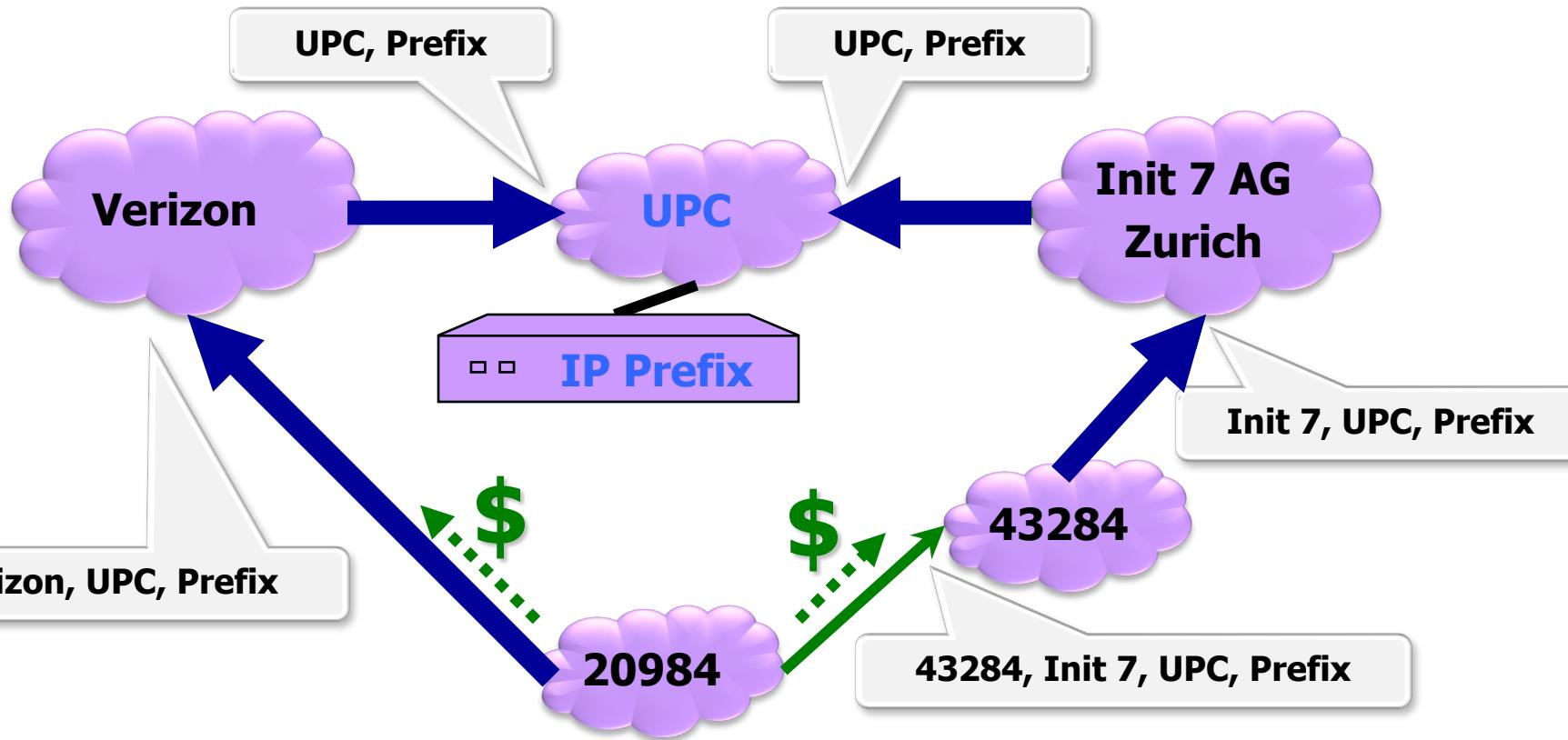
---

- ◆ BGP routing tables contain about 125,000 address prefixes mapping to about 17-18,000 paths
- ◆ Approx. 10,000 BGP routers
- ◆ Approx. 2,000 organizations own AS
- ◆ Approx. 6,000 organizations own prefixes
- ◆ Average route length is about 3.7
- ◆ 50% of routes have length less than 4 ASes
- ◆ 95% of routes have length less than 5 ASes

# Illustration

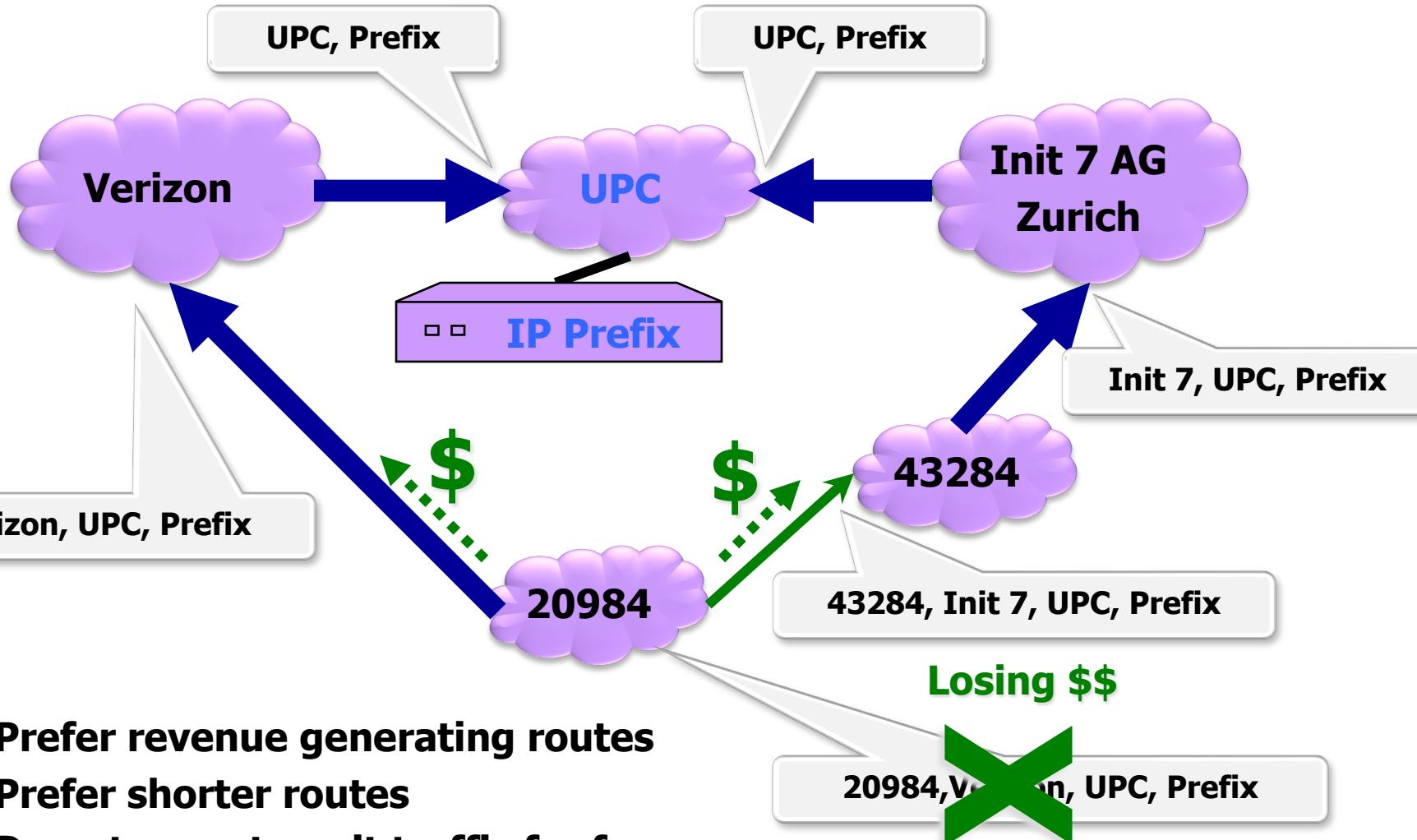


# Routing with BGP

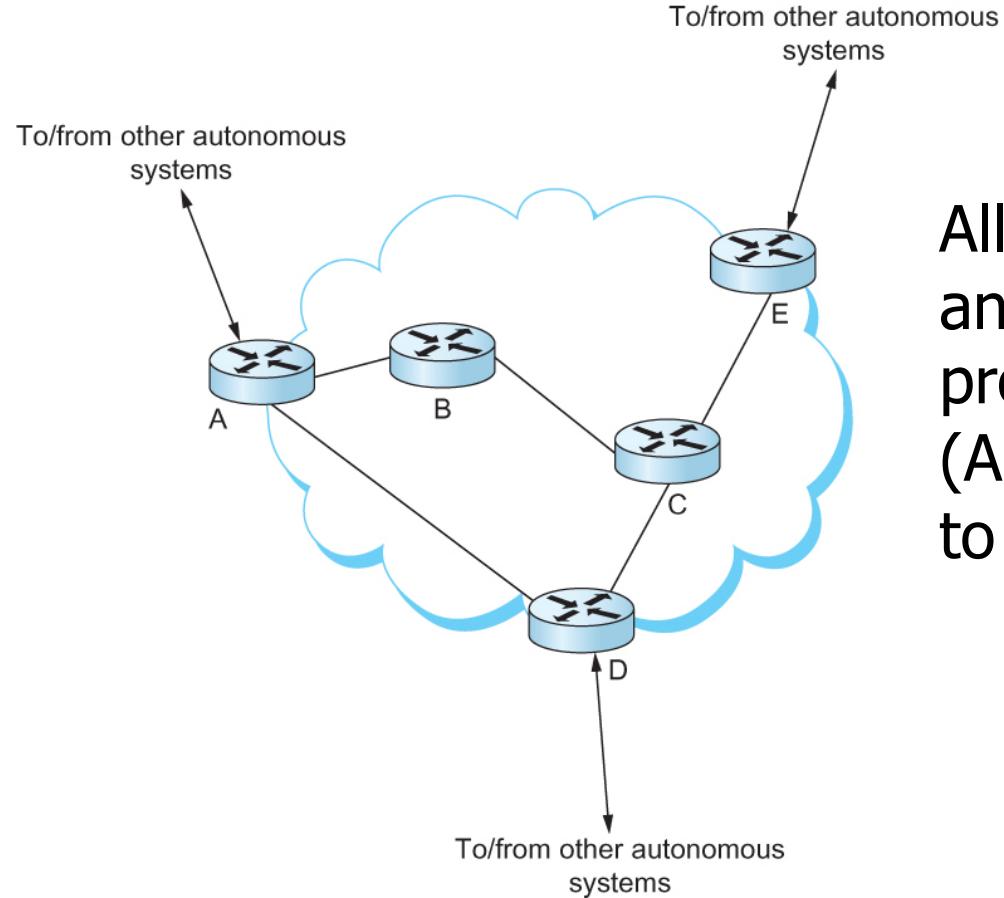


- 1) Prefer revenue generating routes
- 2) Prefer shorter routes

# Routing with BGP



# Integrating Inter- and Intra-Domain



All routers run iBGP and an intradomain routing protocol. Border routers (A, D, E) also run eBGP to other ASs.

# BGP Issues

---

## ◆ BGP convergence problems

- Protocol allows policy flexibility
- Some legal policies prevent convergence
- Even shortest-path policy converges slowly

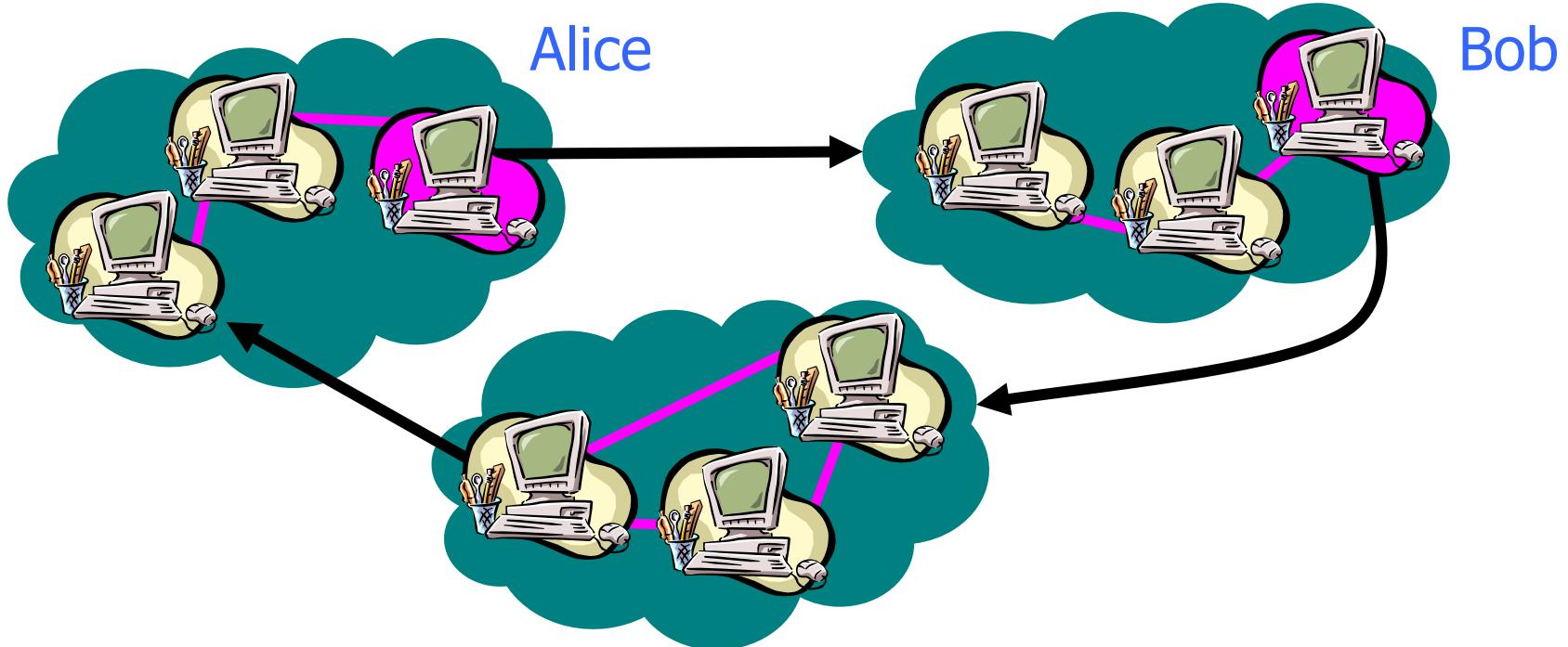
## ◆ Incentive for dishonesty

- ISP pays for some routes, others free

## ◆ Security problems

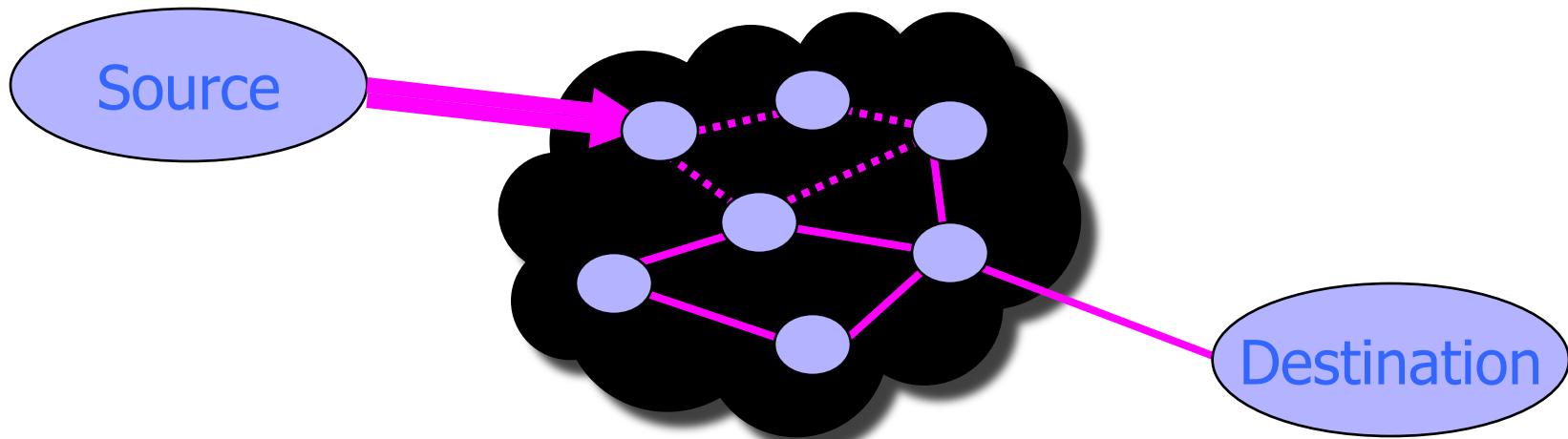
- Potential for disruptive attacks

# Evidence: Asymmetric Routes



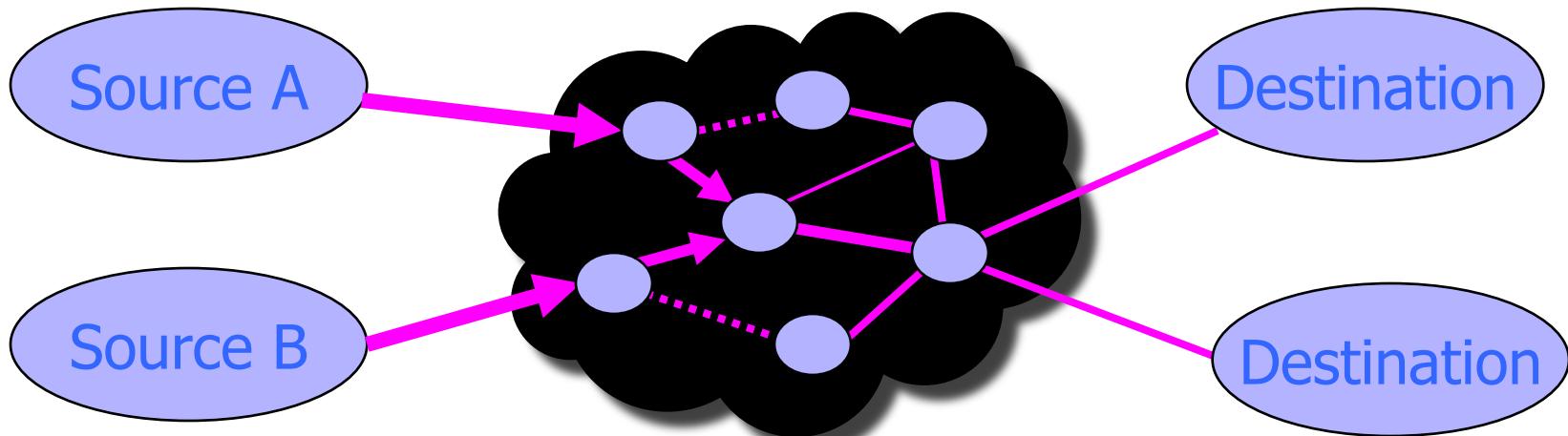
- ◆ Alice, Bob use cheapest routes to each other
  - Cheapest is not always shortest!
- ◆ Asymmetric routes are prevalent
  - AS asymmetry in 30% of measured routes
  - Finer-grained asymmetry far more prevalent

# Side Note: TCP Congestion Control



- ◆ If packets are lost, assume congestion
  - Reduce transmission rate by half, repeat
  - If loss stops, increase rate very slowly
  - Design assumes routers blindly obey this policy

# Protocol Rewards Dishonesty



- ◆ Amiable Alice yields to boisterous Bob
  - Alice and Bob both experience packet loss
  - Alice backs off
  - Bob disobeys protocol, gets better results

# BGP Misconfiguration

---

- ◆ Domain advertises good routes to addresses it does not know how to reach
  - Result: packets go into a network “black hole”
- ◆ April 25, 1997: “The day the Internet died”
  - AS7007 (Florida Internet Exchange) de-aggregated the BGP route table and re-advertised all prefixes as if it originated paths to them
    - In effect, AS7007 was advertising that it has the best route to every host on the Internet
  - Huge network instability as incorrect routing data propagated and routers crashed under traffic

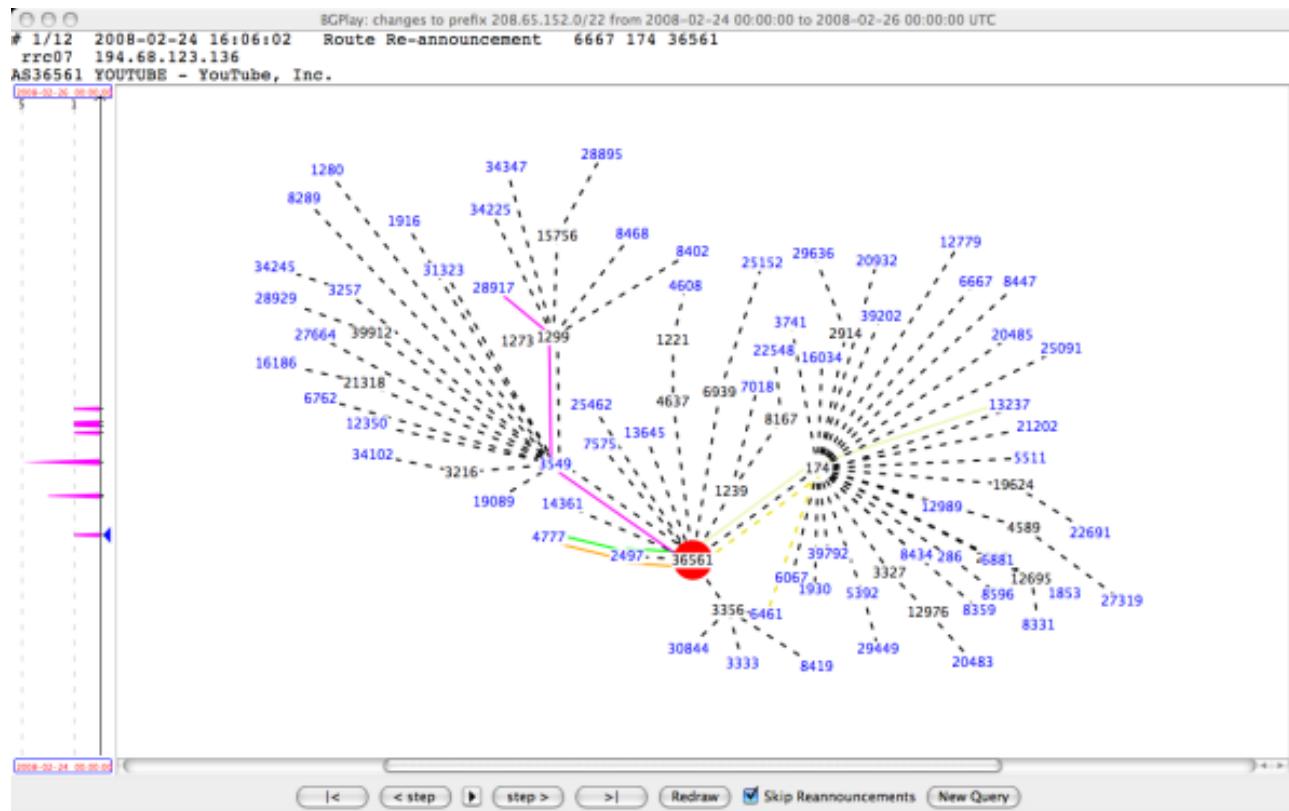
# BGP (In)Security

---

- ◆ BGP update messages contain no authentication or integrity protection
- ◆ Attacker may falsify the advertised routes
  - Modify the IP prefixes associated with a route
    - Can blackhole traffic to certain IP prefixes
  - Change the AS path
    - Either attract traffic to attacker's AS, or divert traffic away
    - Interesting economic incentive: an ISP wants to dump its traffic on other ISPs without routing their traffic in exchange
  - Re-advertise/propagate AS path without permission
    - For example, a multi-homed customer may end up advertising transit capability between two large ISPs

# YouTube (Normally)

- ◆ AS36561 (YouTube) advertises 208.65.152.0/22

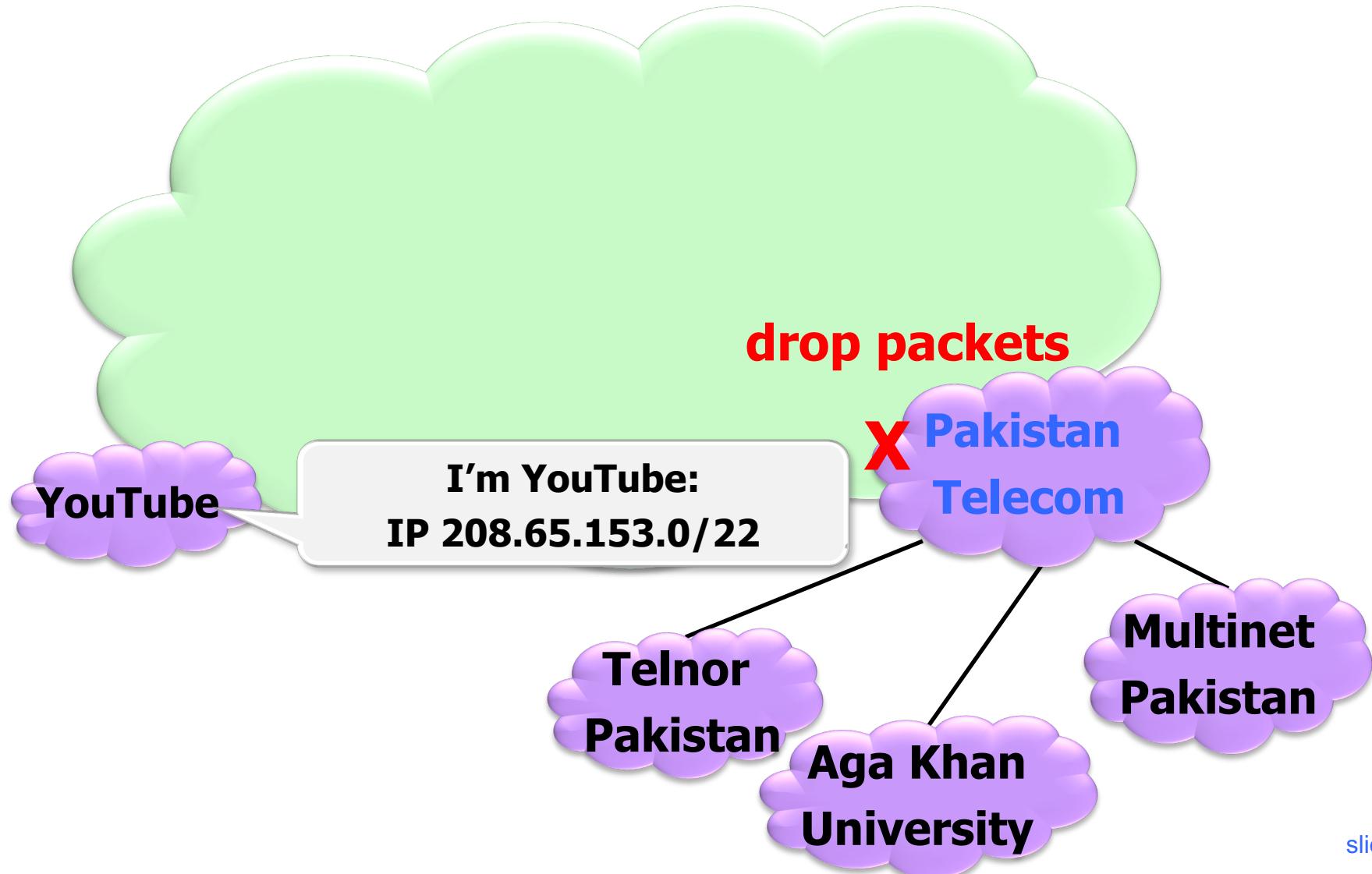


February 24, 2008

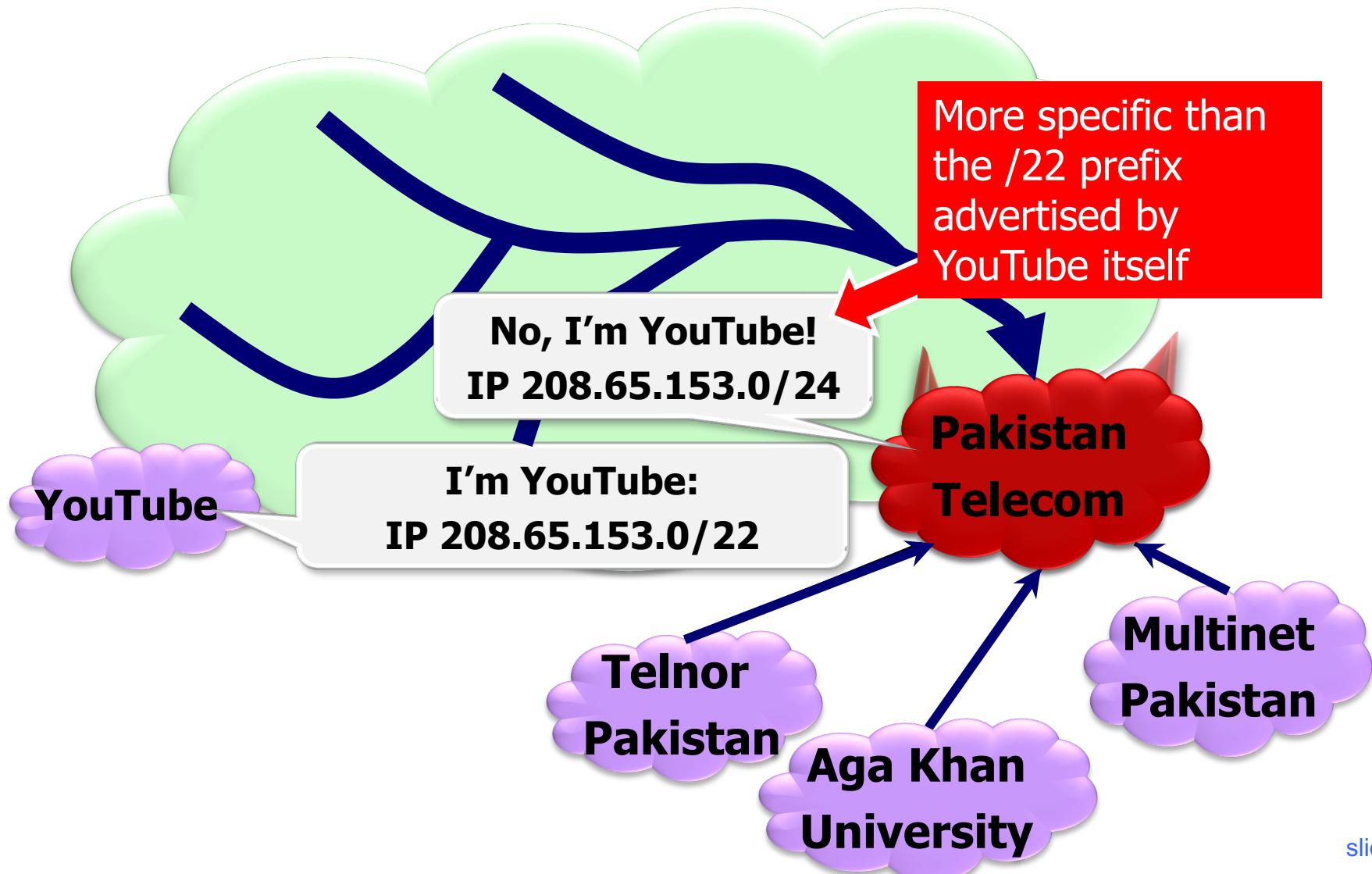
# ◆ Pakistan government wants to block YouTube



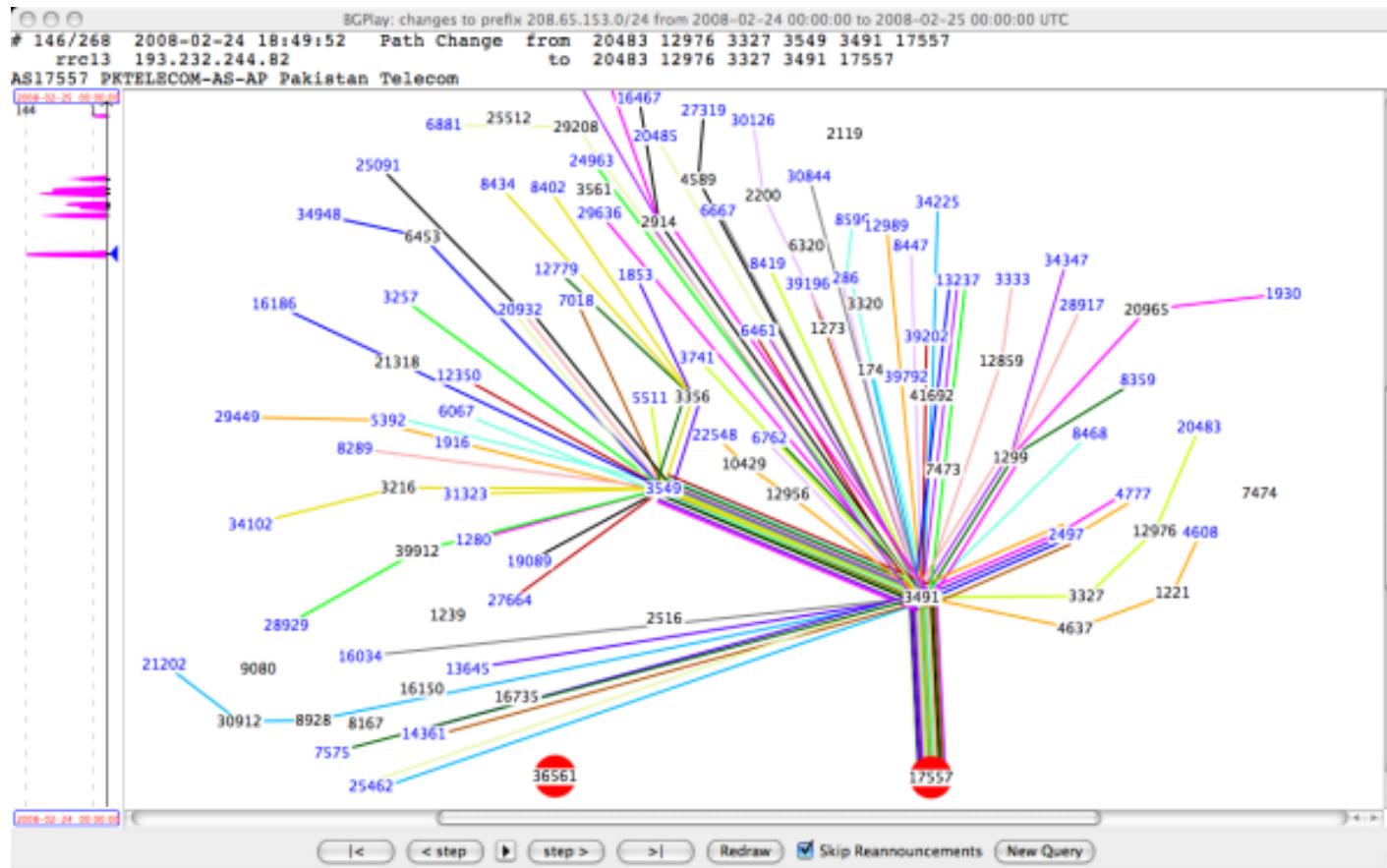
# What Should've Happened



# What Did Happen



# Two-Hour YouTube Outage

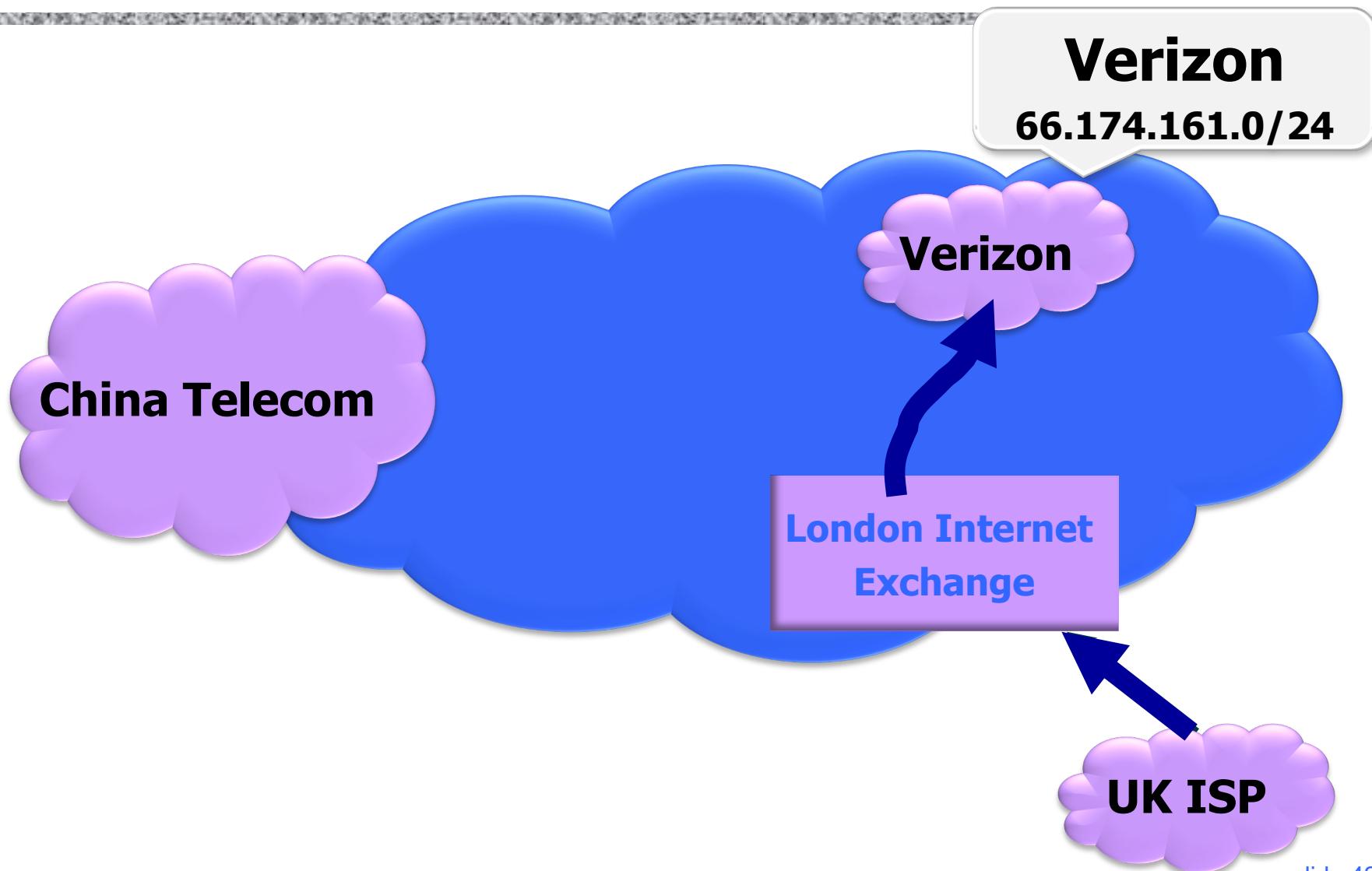


# Other BGP Incidents

---

- ◆ May 2003: Spammers hijack unused block of IP addresses belonging to Northrop Grumman
  - Entire Northrop Grumman ends up on spam blacklist
  - Took two months to reclaim ownership of IP addresses
- ◆ Dec 2004: Turkish ISP advertises routes to the entire Internet, including Amazon, CNN, Yahoo
- ◆ Apr 2010: Small Chinese ISP advertises routes to 37,000 networks, incl. Dell, CNN, Apple
- ◆ Feb-May 2014: Someone uses BGP to hijack the addresses of Bitcoin mining-pool servers, steals \$83,000 worth of Bitcoins

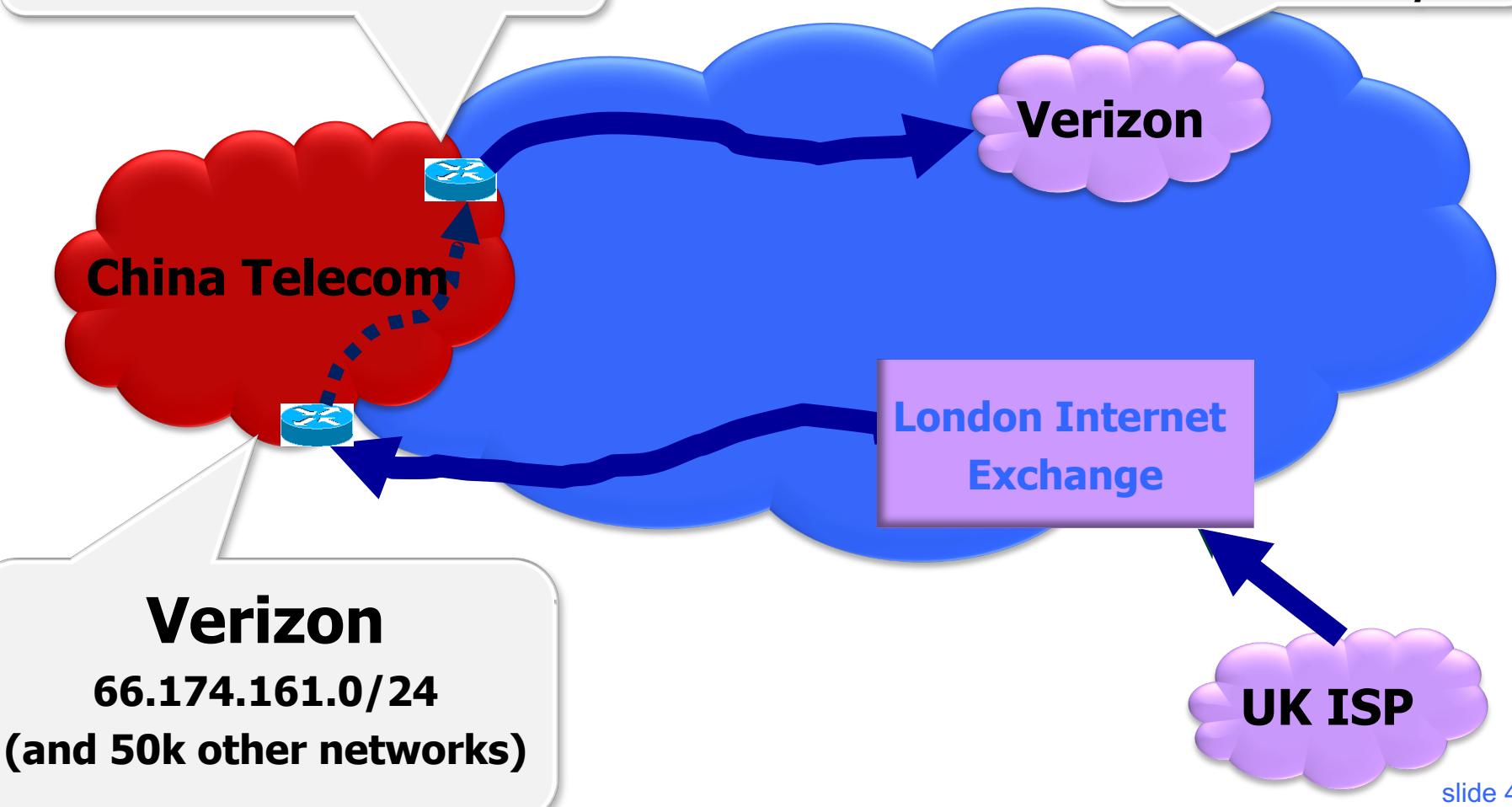
# China Telecom Incident



# China Telecom Incident

This packet is destined for Verizon.

**Verizon**  
66.174.161.0/24



# IP Address Ownership & Hijacking

---

## ◆ IP address block assignment

- Regional Internet Registries (ARIN, RIPE, APNIC)
- Internet Service Providers

## ◆ Proper origination of a prefix into BGP

- By the AS who owns the prefix
- ... or, by its upstream provider(s) in its behalf

## ◆ However, what's to stop someone else?

- Prefix hijacking: another AS originates the prefix
- BGP does not verify that the AS is authorized
- Registries of prefix ownership are inaccurate

# Hijacking is Hard to Debug

---

- ◆ The victim AS doesn't see the problem
  - Picks its own route
  - Might not even learn the bogus route
- ◆ May not cause loss of connectivity
  - E.g., if the bogus AS snoops and redirects
  - ... may only cause performance degradation
- ◆ Or, loss of connectivity is isolated
  - E.g., only for sources in parts of the Internet
- ◆ Diagnosing prefix hijacking
  - Analyzing updates from many vantage points
  - Launching traceroute from many vantage points

# Preventing Prefix Hijacking

---

## ◆ Origin authentication

- Secure database lists which AS owns which IP prefix

## ◆ soBGP

- Digitally signed certificates of prefix ownership

## ◆ Prefix hijacking is not the only threat... in general, BGP allows ASes to advertise **bogus routes**

- Remove another AS from a path to make it look shorter, more attractive, get paid for routing traffic
- Add another AS to a path to trigger loop detection, make your connectivity look better

# Bogus AS Paths

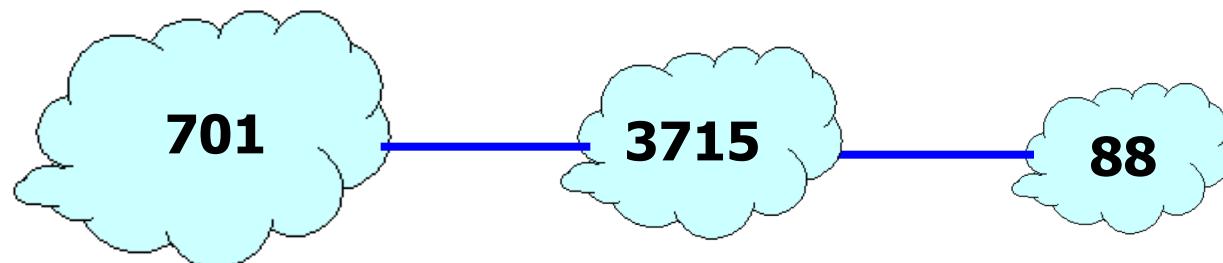
---

- ◆ Remove ASes from the AS path

- E.g., turn “701 3715 88” into “701 88”

- ◆ Possible motivations

- Make the AS path look shorter than it is
  - Attract sources that normally try to avoid AS 3715



# Bogus AS Paths

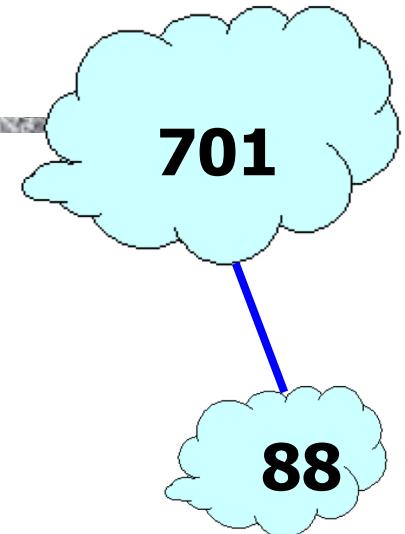
---

- ◆ Add ASes to the path

- E.g., turn “701 88” into “701 3715 88”

- ◆ Possible motivations

- Trigger loop detection in AS 3715
  - Make your AS look like it has richer connectivity



# Protecting BGP

---

- ◆ Simple authentication of packet sources and packet integrity is not enough
- ◆ Before AS advertises a set of IP addresses, the owner of these addresses must authorize it
  - Goal: verify path origin
- ◆ Each AS along the path must be authorized by the preceding AS to advertise the prefixes contained in the UPDATE message
  - Goal: verify propagation of the path vector

# S-BGP Protocol

[Kent, Lynn, Seo]

## ◆ Address attestation

- Owner of one or more prefixes certifies that the origin AS is authorized to advertise the prefixes
- Need a public-key infrastructure (PKI)
  - X.509 certificates prove prefix ownership; owner can then delegate his “prefix advertising rights” to his ISP

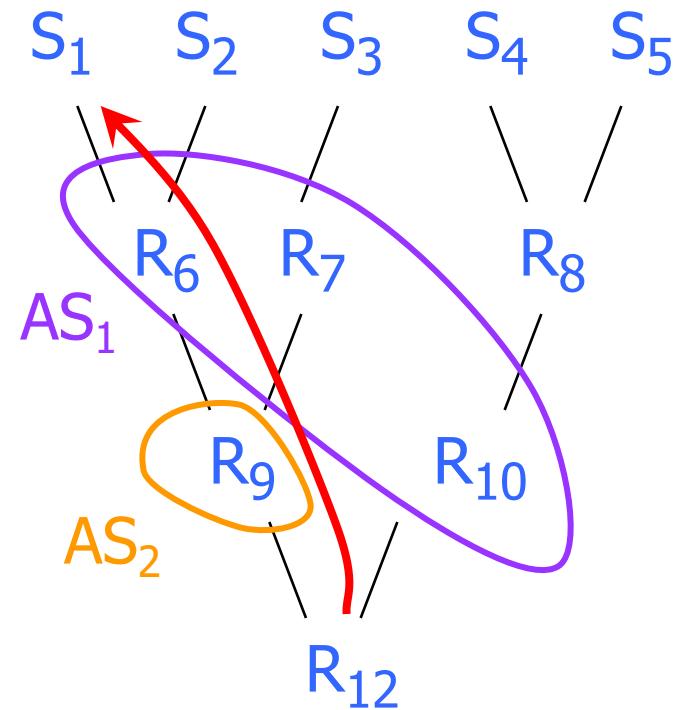
## ◆ Route attestation

- Router belonging to an AS certifies (using digital signatures) that the next AS is authorized to propagate this route advertisement to its neighbors
- Need a separate public-key infrastructure
  - Certificates prove that AS owns a particular router

# S-BGP Update Message

◆ An update message from  $R_9$  advertising this route must contain:

- Ownership certificate certifying that some  $X$  owns IP address  $S_1$
- Signed statement from  $X$  that  $AS_1$  is authorized to advertise  $S_1$
- Ownership certificate certifying that  $AS_1$  owns router  $R_6$ 
  - If AS is represented by a router
- Signed statement from  $R_6$  that  $AS_2$  is authorized to propagate  $AS_1$ 's routes
- Ownership certificate certifying that  $AS_2$  owns router  $R_9$
- Lots of public-key operations!



# Securing BGP

---

- ◆ Dozens of proposals, various combinations of cryptographic mechanisms and anomaly detection
  - Example: Secure BGP (S-BGP)
    - Origin authentication + entire AS path digitally signed
    - Can verify that the route is recent, no ASes have been added or removed, the order of ASes is correct
  - Also: IRV, SPV, psBGP, PGBGP, PHAS, Whisper...
- ◆ How many of these have been deployed?

None

- ◆ No complete, accurate registry of prefix ownership
- ◆ Need a public-key infrastructure
- ◆ Cannot react rapidly to changes in connectivity
- ◆ Cost of cryptographic operations
- ◆ Not deployable incrementally