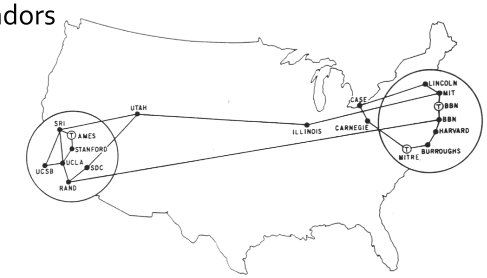
eecs489 Computer Networks

Lecture 17:
Inter-domain Routing and BGP

Routing on the Internet

In the beginning there was the ARPANET:
- route using GGP (Gateway-to-Gateway Protocol), a distance vector routing protocol

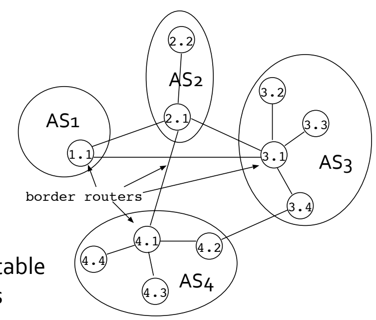
Problems:
- needed "flag-hour" to update routing protocol
- incompatibility across vendors



MAP 4 September 1971

Routing on the Internet

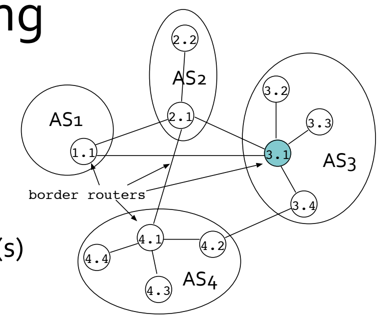
Solution: hierarchical routing
- administrative autonomy:
  - each network admin can control routing within its own network
- internet: network of networks
- allows the Internet to scale:
  - with 200 million hosts, each router can't store all destinations in its routing table
  - route updates alone will swamp the links



Aggregate routers into regions of "autonomous systems" (ASs)

Hierarchical Routing

Gateway/border router
- neighboring ASs interact to coordinate routing
- direct link to router in other AS(s)
- keeps in its routing table:
  - next hop to other ASs
  - all hosts within its AS
- hosts within an AS only keep a default route to the border router

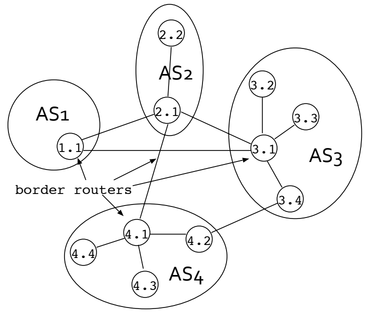


3.1	
dest	next
1.*	1.1
2.*	2.1
4.*	2.1
3.2	3.2
3.3	3.3
3.4	3.4

Hierarchical Routing

Routers in the same AS run same routing protocol

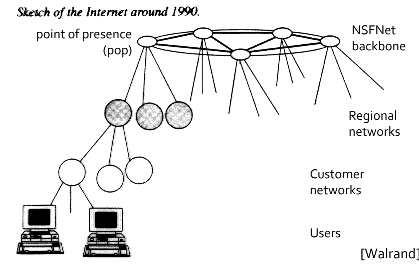
- “intra-AS” routing protocol
- each AS uses its own link metric
- routers in different ASs can run different intra-AS routing protocol
- internal topology is not shared between ASs

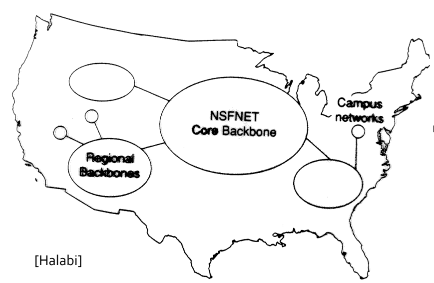


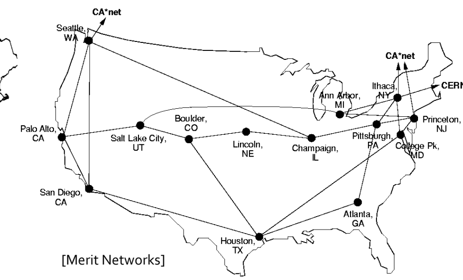
The NSFNet 1989

Area hierarchy:

- backbone/core: NSFNet
- regional networks: MichNet, BARRNET, Los Nettos, Cerfnet, JVCNet, NEARNet, etc.
- campus networks





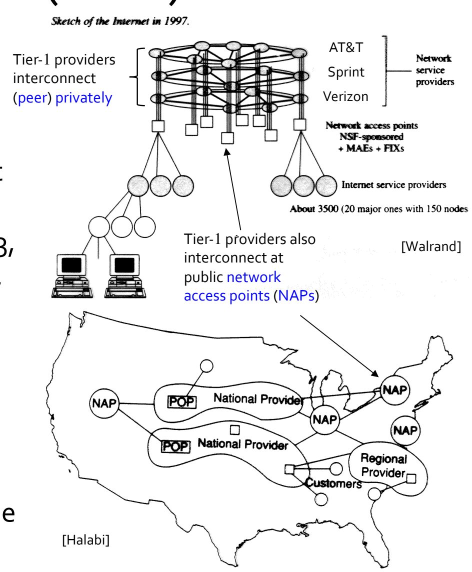


Commercialization (1994)

Roughly hierarchical

At center: “Tier-1” ISPs

- Tier-1 ASs: top of the Internet hierarchy of ~10 Ass: AOL, AT&T, Global Crossing, Level3, Verizon/UUNET, NTT, Qwest, SAVVIS (formerly Cable & Wireless), Sprint, etc.
- full (N^2) peering relationships between Tier-1 providers
- has no upstream provider
- national/international coverage



AS Structure: Other ASs

Lower tier providers

- provide transit service to downstream customers
- but, need at least one provider of their own
- typically have national or regional scope
- includes several thousand ASs

Stub ASs

- do not provide transit service to others
- connect to one or more upstream providers
- includes the vast majority (e.g., 85-90%) of the ASs

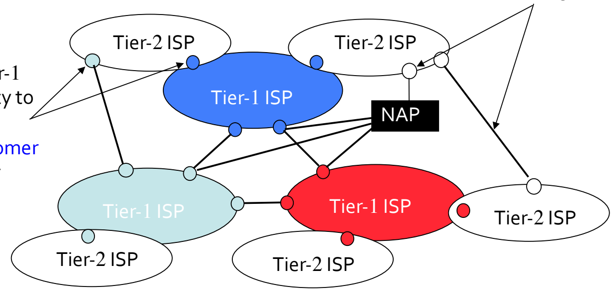
[Rexford]

"Tier-2" ISPs: Smaller (Often Regional) ISPs

Connect to one or more tier-1 ISPs, possibly other tier-2 ISPs

Tier-2 ISPs also peer privately with each other, and interconnect at NAPs

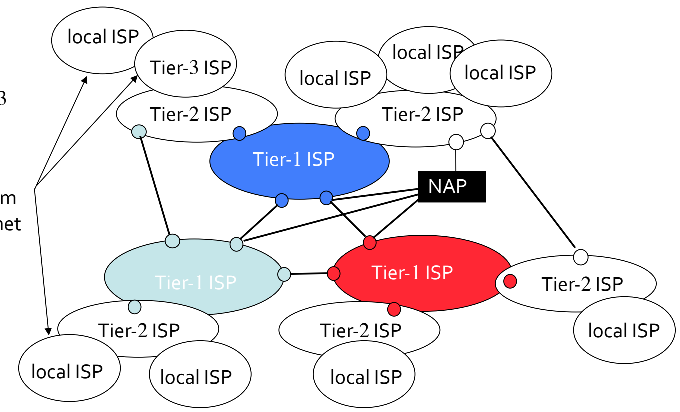
Tier-2 ISP pays tier-1 ISP for connectivity to rest of Internet
- tier-2 ISP is customer of tier-1 provider



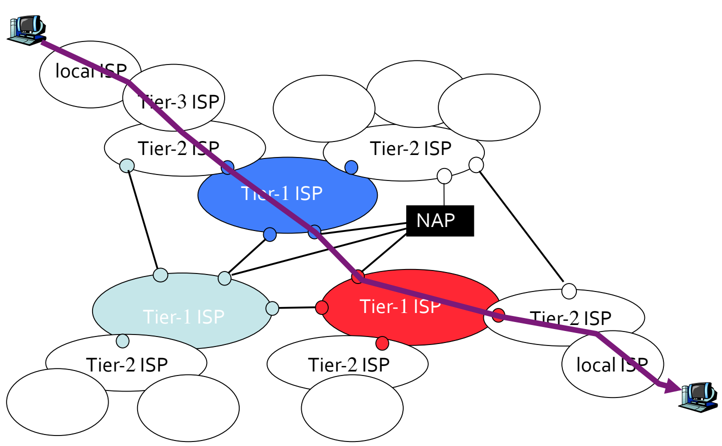
"Tier-3" ISPs and Local ISPs

Last hop ("access") network (closest to end systems)

Local and tier-3 ISPs are customers of higher tier ISPs connecting them to rest of Internet



A Packet Passes Through Many Networks



AS Number Trivia

AS number is a 16-bit quantity
- 65,536 unique AS numbers

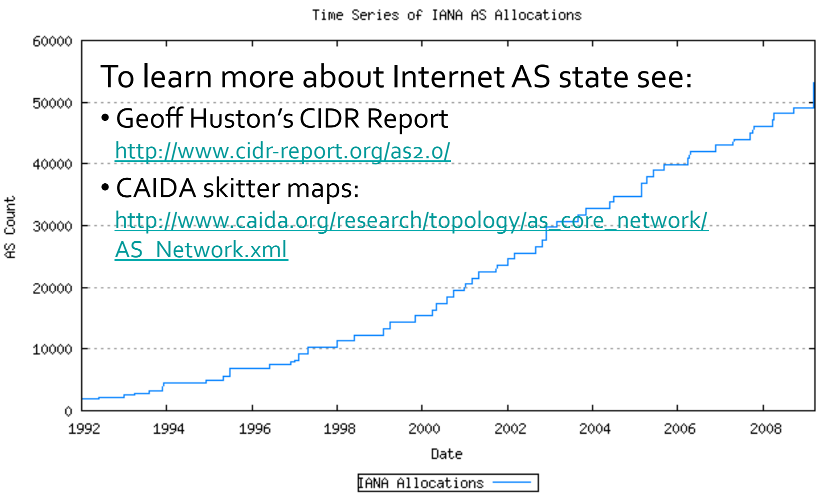
Some are reserved numbers (e.g., for private ASs)
- only 64,510 are available for public use

Managed by Internet Assigned Numbers Authority (IANA)
- gives blocks of 1,024 to Regional Internet Registries
- RIRs assign AS numbers to institutions
- 49,649 AS numbers in visible use (Feb '15)

In 2007 started assigning 32-bit AS #s

[Rexford]

Growth of AS numbers



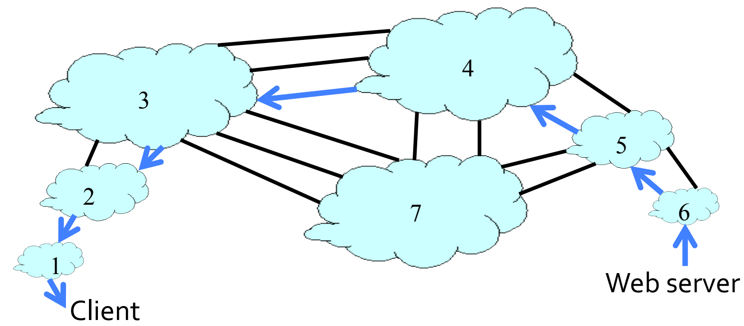
To learn more about Internet AS state see:

- Geoff Huston's CIDR Report
  http://www.cidr-report.org/as2.0/
- CAIDA skitter maps:
  http://www.caida.org/research/topology/as_core_network/AS_Network.xml

Interdomain Routing

AS-level topology

- destinations are CIDR address prefixes (APs, e.g., `12.0.0.0/8`)
- nodes are Autonomous Systems (ASs)
- edges are business relationships



[Rexford]

Challenges for Interdomain Routing

Scale

- address prefixes (APs): 200,000 and growing
- ASs: ~50,000 visible ones, and 60K allocated
- routers: at least in the millions

Proprietary information:

- ASs don't want to divulge internal topologies
- nor their business relationships with neighbors

Policy

- no Internet-wide notion of a link cost metric
- need control over where you send traffic
- and who can send traffic through you

[Rexford]

Why SPF is not Suitable

Topology information is flooded

- high bandwidth and storage overhead
- nodes must divulge sensitive commercial information

Entire path computed locally per node

- high processing overhead in a large network

Route computation minimizes some notion of total distance

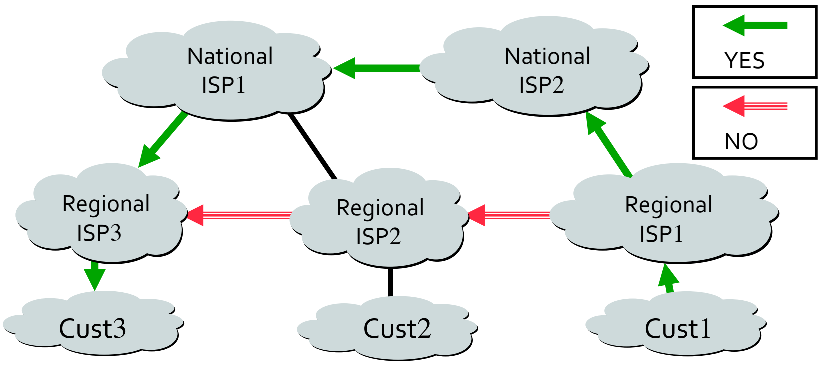
- all traffic must travel on shortest paths

[Rexford]

Why SPF is not Suitable

All nodes need common notion of link costs
- works only if policy is shared and uniform

Incompatible with commercial relationships



[Rexford]

Why Not Distance Vector?

Advantages
- hides details of the network topology
- nodes determine only "next hop" toward the destination

Disadvantages
- route computation still entails minimization of some notion of total distance, which is difficult in an inter-domain setting
- slow convergence due to reliance on counting-to-infinity to detect routing loop

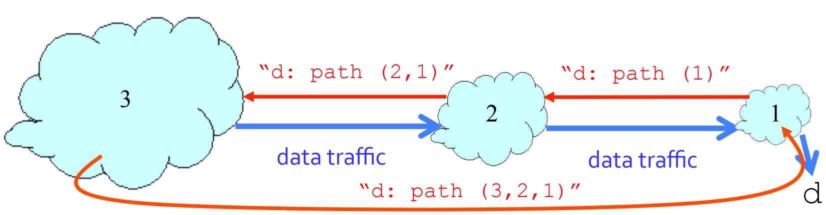
Instead use path vector
- easier loop detection

[after Rexford]

Path-Vector Routing

Avoid counting-to-infinity by advertising entire path
- distance vector: send distance metric per destination
- path vector: send the entire path for each destination

Loop detection:
- each node looks for its own node identifier in advertised path
- and discards paths with loops
- e.g., node 1 sees itself in the path (3, 2, 1) and discards the path

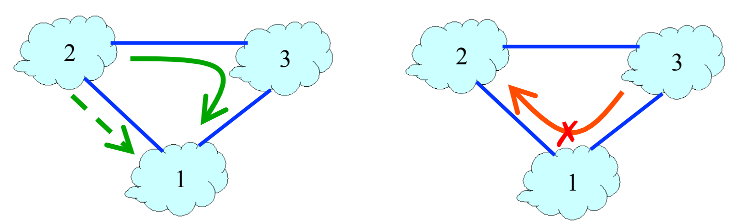


[Rexford]

Other Advantage: Flexible Policies

Each node can apply local policies
- path selection: which path to use?
- path export: which paths to advertise?

Examples
- node 2 may prefer the path "2, 3, 1" over "2, 1"
- node 1 may not want node 3 to hear of the path "1, 2"



[Rexford]

Internet inter-AS Routing: BGP

BGP (Border Gateway Protocol) is the de facto standard for inter-AS routing

- 06/89 v.1
- 06/90 v.2 EGP (Exterior Gateway Protocol) to BGP transition
- 10/91 v.3 BGP installed
- 07/94 v.4 de facto standard

Internet inter-AS Routing: BGP

BGP provides each AS a means to:

- use prefix-based path-vector protocol
- propagates AP reachability to all routers inside the AS
- obtains AP reachability from neighboring ASs
- determines "good" routes to APs based on reachability information and policy
- Inter-AS routing is policy driven, not load-sensitive, generally not QoS-based

When an AS advertises an AP to another AS, it is promising to forward any packets the other AS sends to the AP

- an AS can aggregate CIDR APs in its advertisement

BGP runs over TCP

Pairs of BGP routers (BGP peers) establish semi-permanent TCP connections: BGP sessions

- advantage of using TCP: reliable transmission allows for incremental updates: updates only when changes occur
- disadvantage: TCP congestion control mechanism slows down route updates that could decongest link!

Failure detection:

- TCP doesn't detect lost connectivity on its own
- instead, BGP must detect failure
  - sends `KEEPALIVE` packets every 60 seconds
  - hold timer: 180 seconds

BGP sessions do not correspond to physical links, but rather business relationship

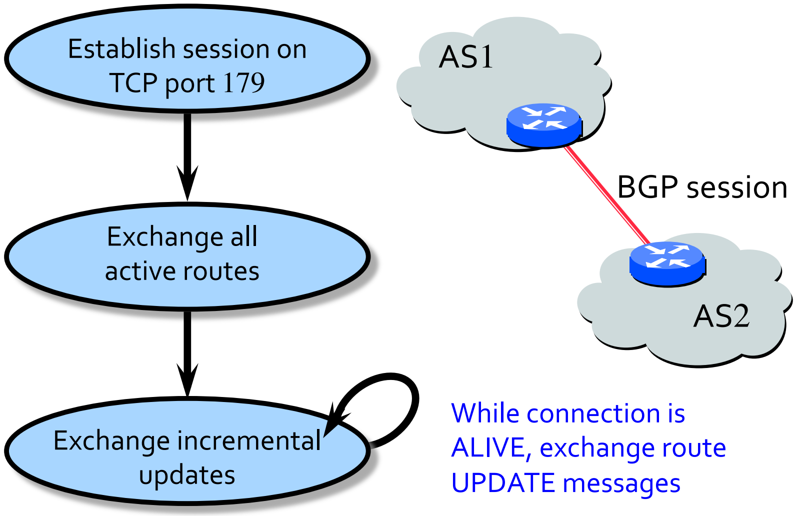
BGP Messages

BGP messages:

- `OPEN`: opens TCP connection to peer and authenticates sender
- `UPDATE`: advertises a new active path (or withdraws one no longer available)
- `KEEPALIVE`: keeps connection alive in the absence of `UPDATE`s; also acknowledges `OPEN` request
- `NOTIFICATION`: reports errors in previous message; also used to close connection

[after Rexford]

BGP Operations



[Rexford]

Path Attributes & BGP Routes

When advertising an AP, advertisement includes BGP attributes

Two important attributes:

- AS-PATH: the path vector of ASs through which the advertisement for an AP passed through
- NEXT-HOP: the specific internal-AS router to next-hop AS (there may be multiple exits from current AS to next-hop-AS)

Path Attributes & BGP Routes

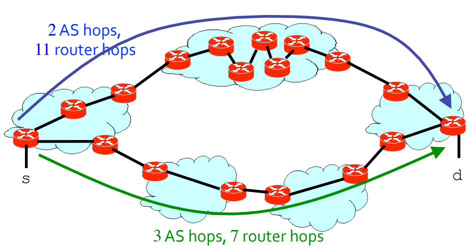
Sample BGP entry:

```
destination         NEXT-HOP       AS-PATH
198.32.163.0/24     202.232.1.8    2497 2914 3582 4600
```

- address range 198.32.163.0/24 is in AS 4600
- to get there, send to next hop router at address 202.232.1.8
- the path there goes through ASs 2497, 2914, 3582, in order

AS path chosen may not be the shortest AS path

Router path may be longer than AS path



[after Rexford]

Causes of BGP Routing Changes

Topology changes

- equipments going up or down
- deployment of new routers or sessions

BGP session failures

- due to equipment failures, maintenance, etc.
- or, due to congestion on the physical path

Changes in routing policy

- changes in preferences in the routes
- changes in whether the route is exported

Persistent protocol oscillation

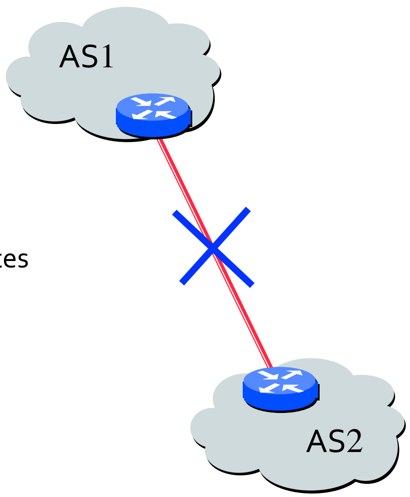
- conflicts between policies of different ASs

[Rexford]

BGP Session Failure

Reacting to a failure

- discard all routes learned from the neighbor
- send new updates for any routes that change
- overhead increases with # of routes
  - reason why many Tier-1 ASs filter out prefixes longer than /24



[Rexford]

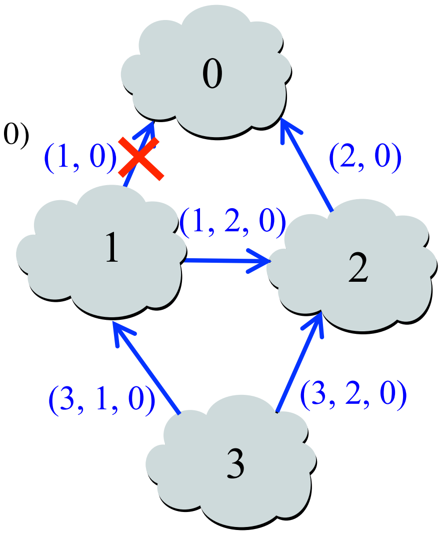
Routing Change: Before and After

AS1

- delete the route (1, 0)
- switch to next route (1, 2, 0)
- send route (1, 2, 0) to AS3

AS3

- sees (1, 2, 0) replace (1,0)
- compares to route (2, 0)
- switches to using AS2



[Rexford]