CSCI-1680
Network Layer:
Inter-domain Routing

Chen Avin

Based partly on lecture notes by David Mazières, Phil Levis, John Jannotti, Peterson & Davie, Rodrigo Fonseca

and “Computer Networking: A Top Down Approach” - 6th edition

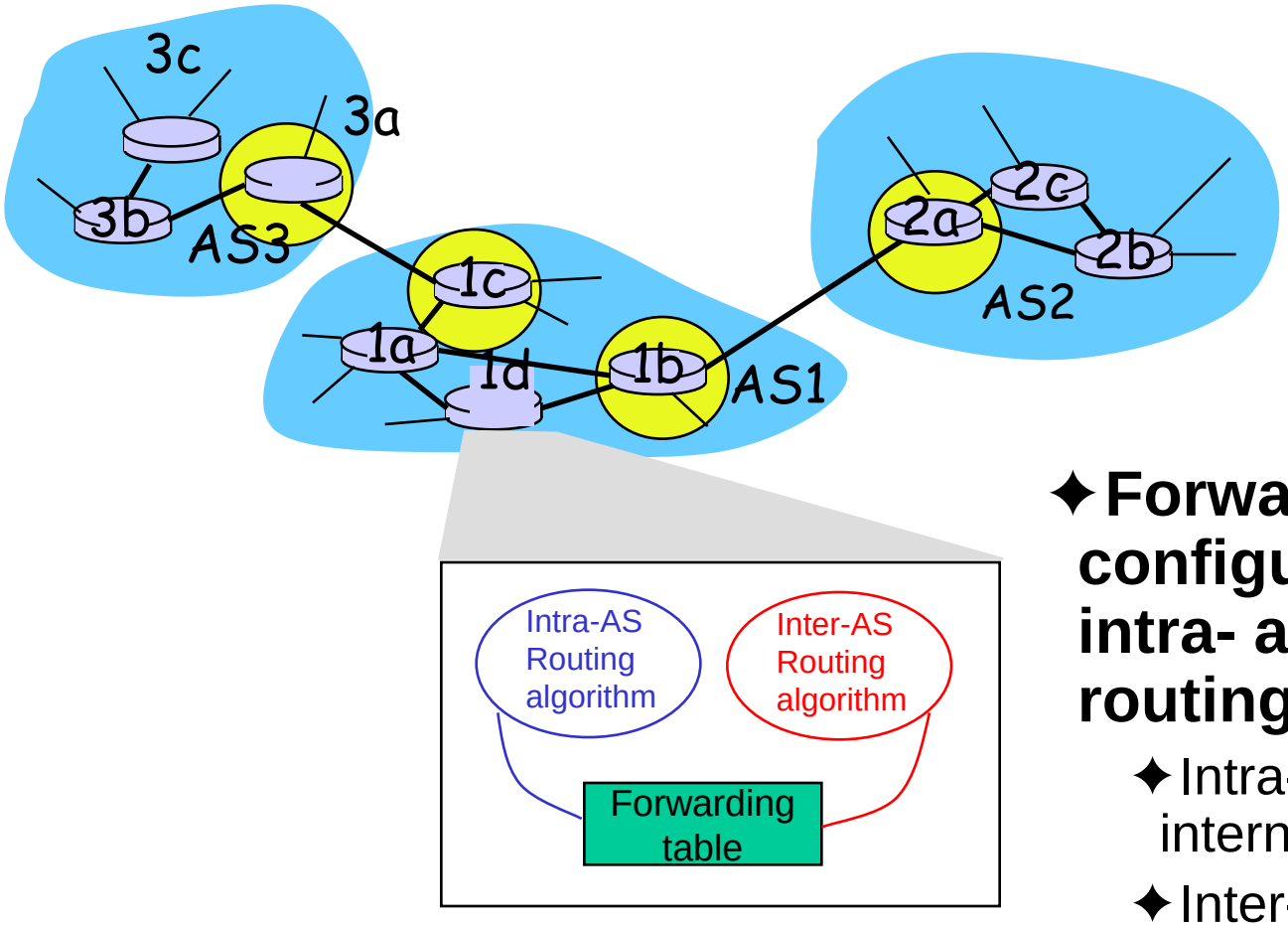
Today

- **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

Interconnected ASes



- **Forwarding table is configured by both intra- and inter-AS routing algorithm**
  - Intra-AS sets entries for internal dests
  - Inter-AS & Intra-As sets entries for external dests

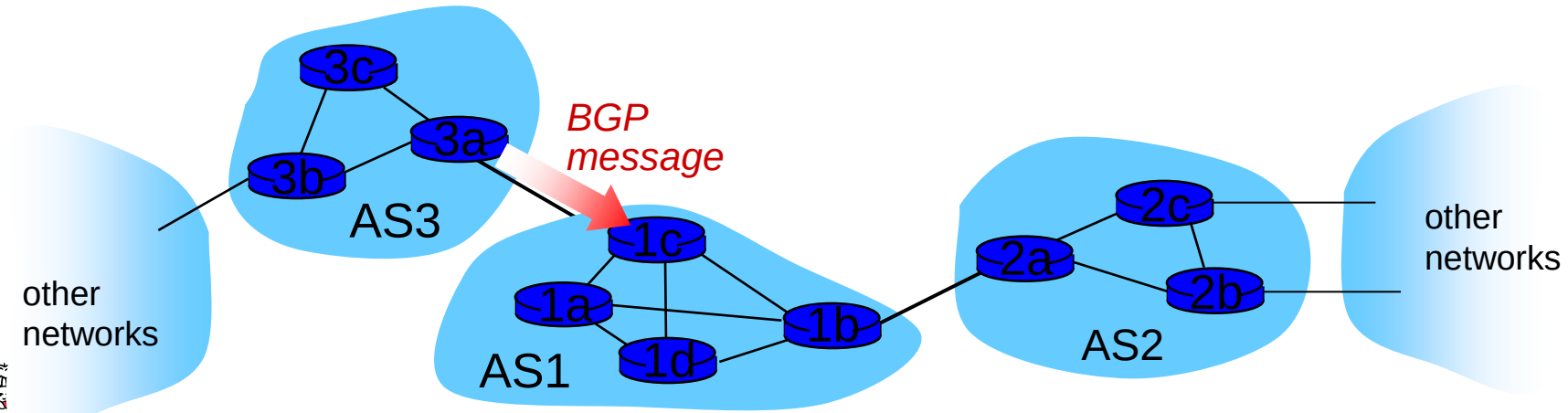
Internet inter-AS routing: BGP

- **BGP (Border Gateway Protocol): *the* de facto inter-domain routing protocol**
  - "glue that holds the Internet together"
- **BGP provides each AS a means to:**
  - eBGP: obtain subnet reachability information from neighboring ASs. BGP "speakers".
  - iBGP: propagate reachability information to all AS-internal routers.
  - determine "good" routes to other networks based on reachability information and policy.
- **allows subnet to advertise its existence to rest of Internet: *"I am here"***

BGP basics

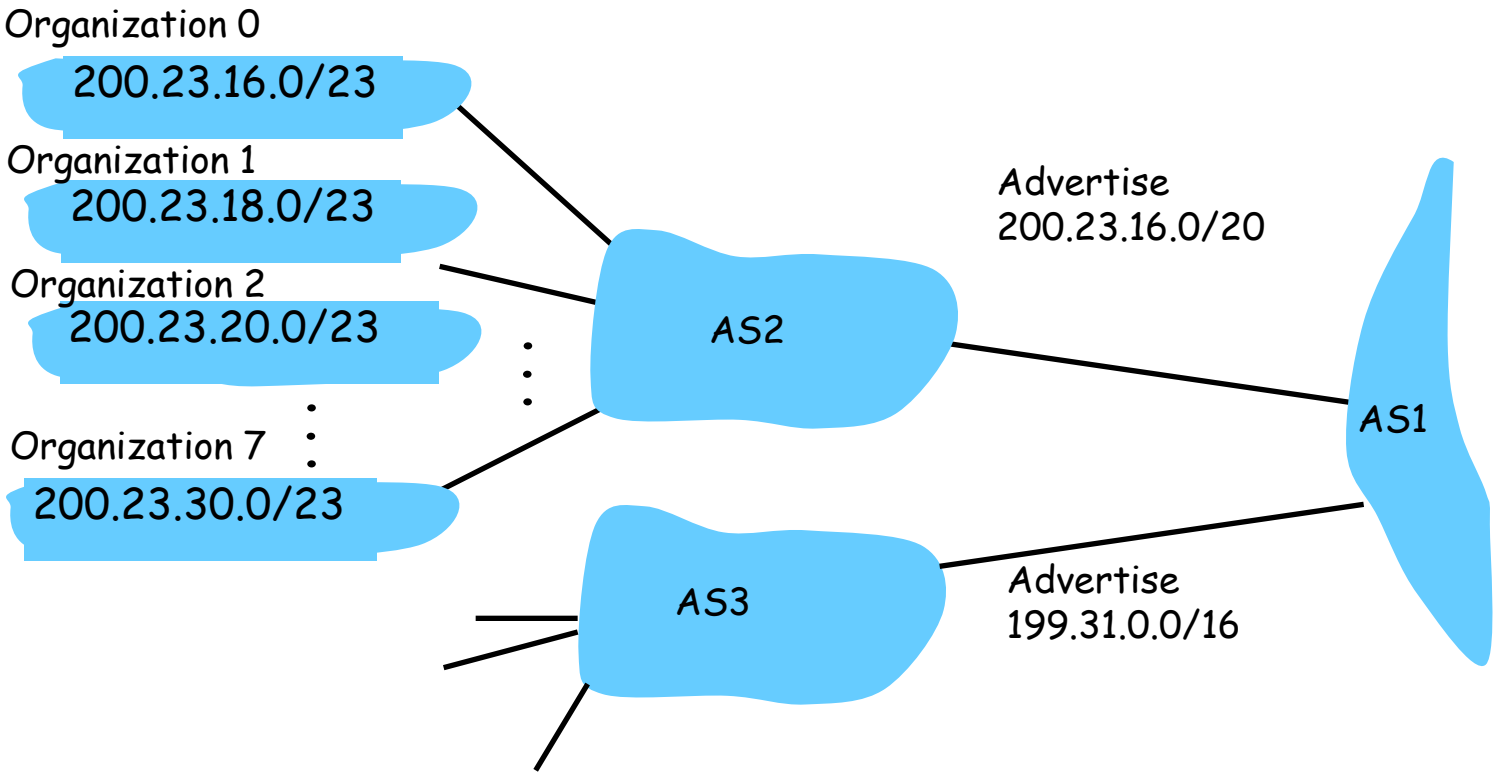
- BGP session: two BGP routers ("peers") exchange BGP messages:
  - advertising *paths* to different destination network prefixes ("path vector" protocol)
  - exchanged over semi-permanent TCP connections

- **when AS3 advertises a prefix to AS1:**
  - AS3 *promises* it will forward datagrams towards that prefix
  - AS3 can aggregate prefixes in its advertisement



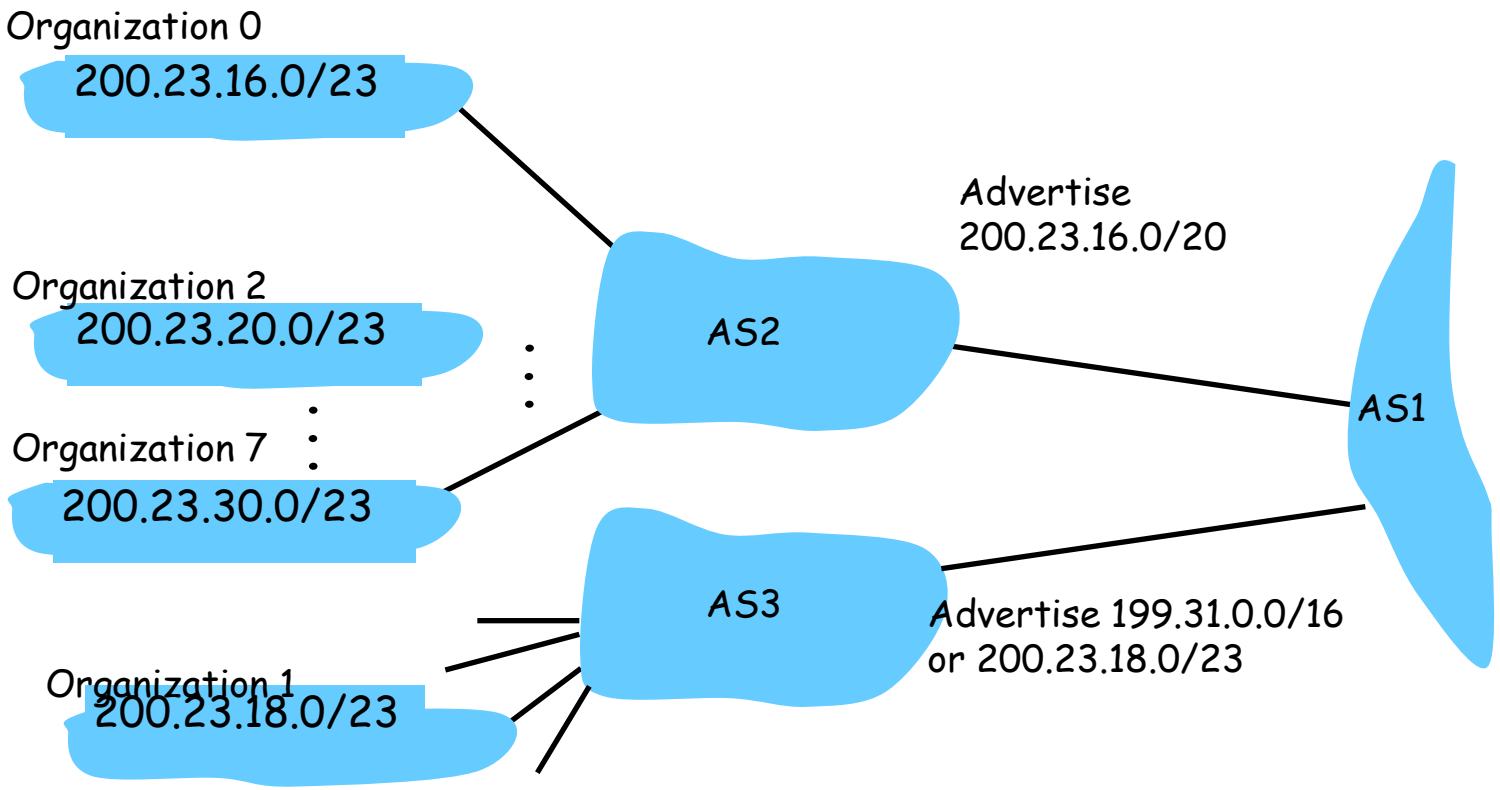
route aggregation

Hierarchical addressing allows efficient advertisement of routing information:



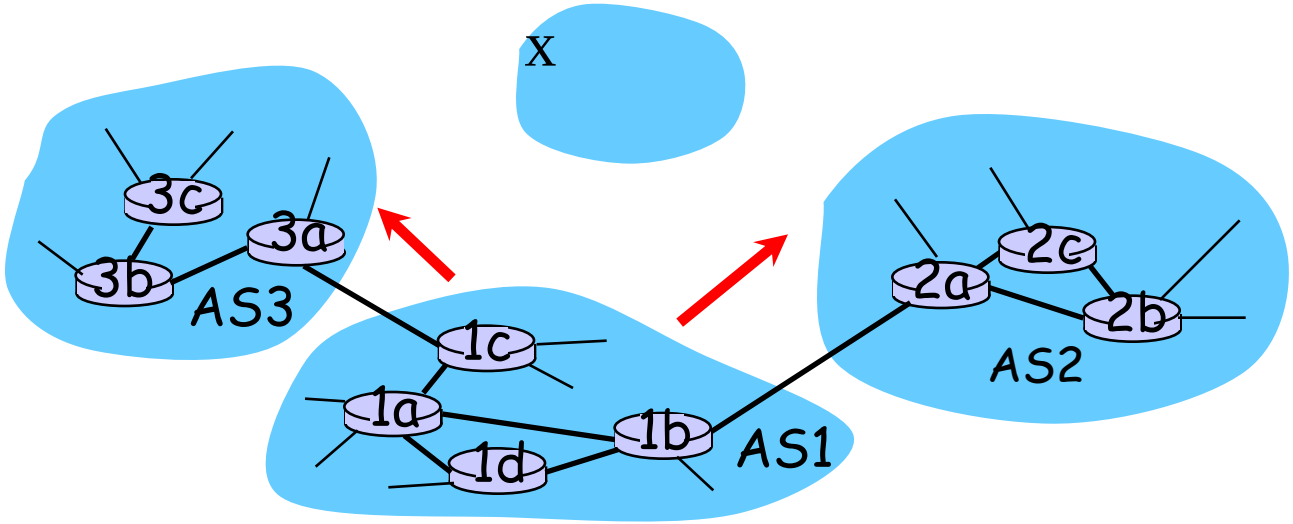
Hierarchical addressing: more specific routes

Will because routers use longest-prefix matching for forwarding.



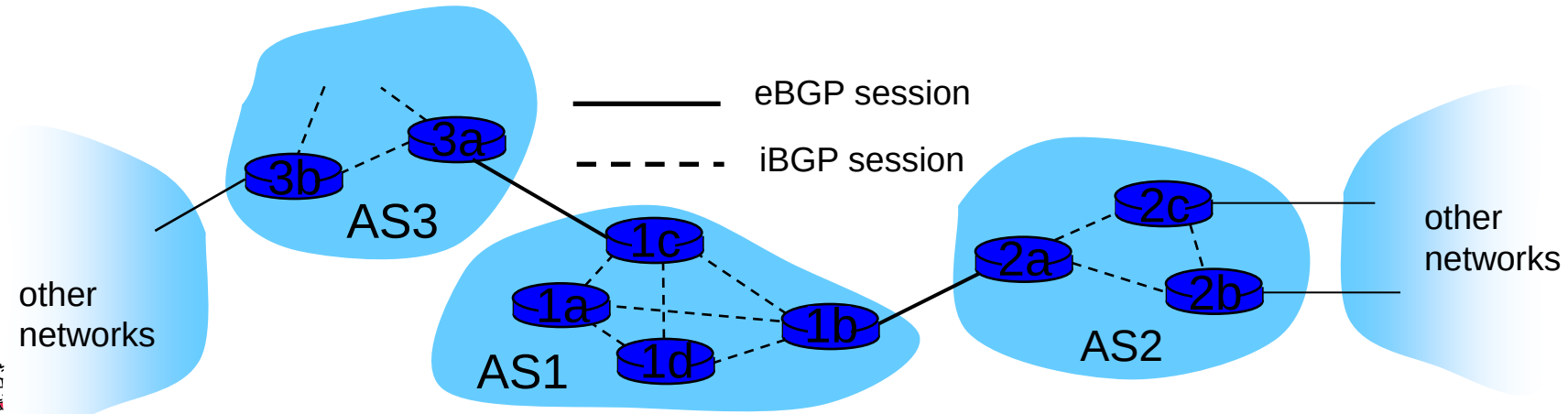
Example: Choosing among multiple ASes

- Now suppose AS1 learns from the inter-AS protocol that subnet x is reachable from AS3 and from AS2.
- To configure forwarding table, router 1d must determine towards which gateway it should forward packets for dest x.
- This is also the job on inter-AS routing protocol!



BGP basics: distributing path information

- **using eBGP session between 3a and 1c, AS3 sends prefix reachability info to AS1.**
  - 1c can then use iBGP do distribute new prefix info to all routers in AS1
  - 1b can then re-advertise new reachability info to AS2 over 1b-to-2a eBGP session
- **when router learns of new prefix, it creates entry for prefix in its forwarding table.**



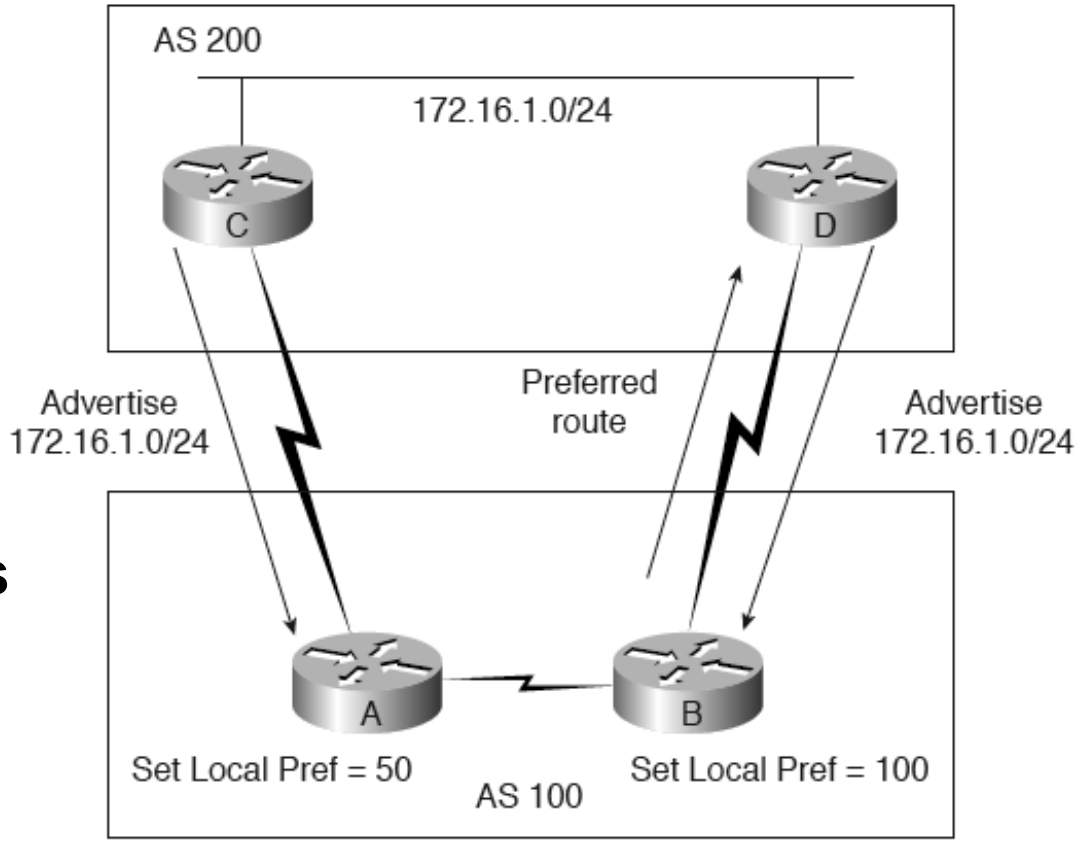
Path attributes & BGP routes

- **When advertising a prefix, advert includes BGP attributes.**
  - prefix + attributes = “route”
- **BGP attributes:**
  - Weight
  - Local preference
  - Multi-exit discriminator
  - Origin
  - AS_path
  - Next hop
  - Community

BGP attributes (1)

- **Local Preference: used to prefer an exit point from the local autonomous system (AS). The local preference attribute is propagated throughout the local AS.**

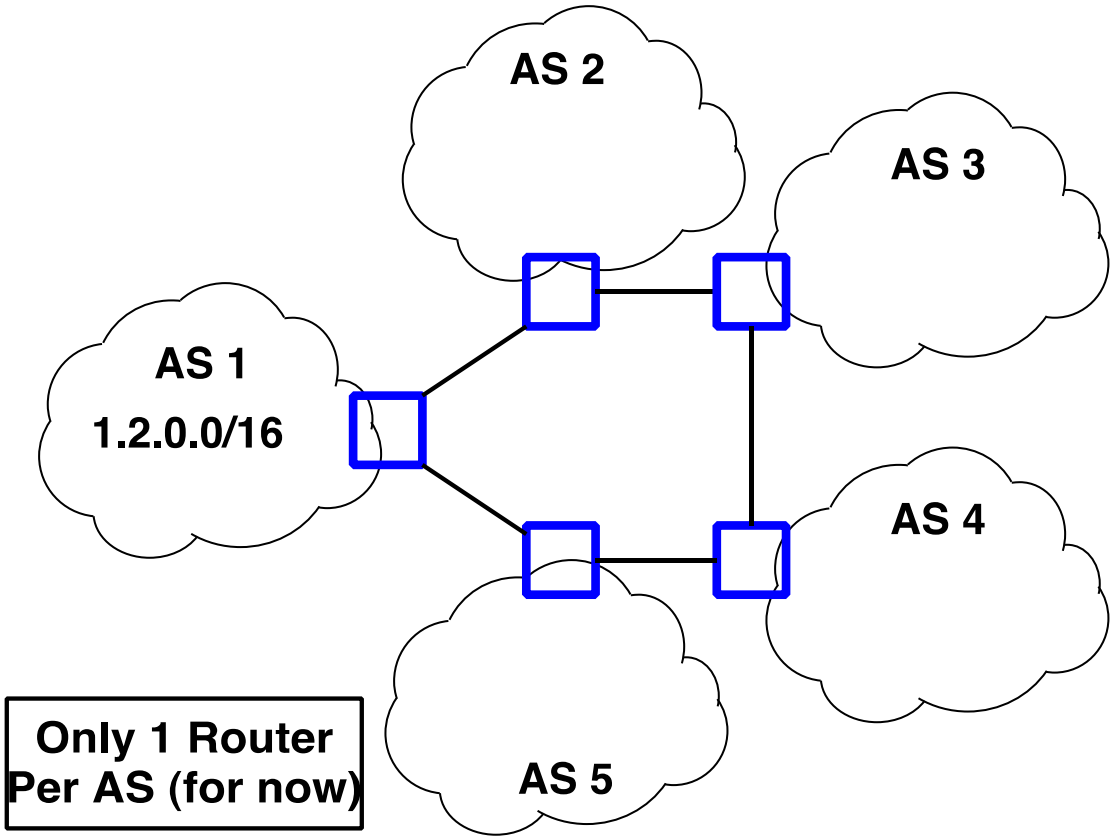
Figure 39-3 BGP Local Preference Attribute



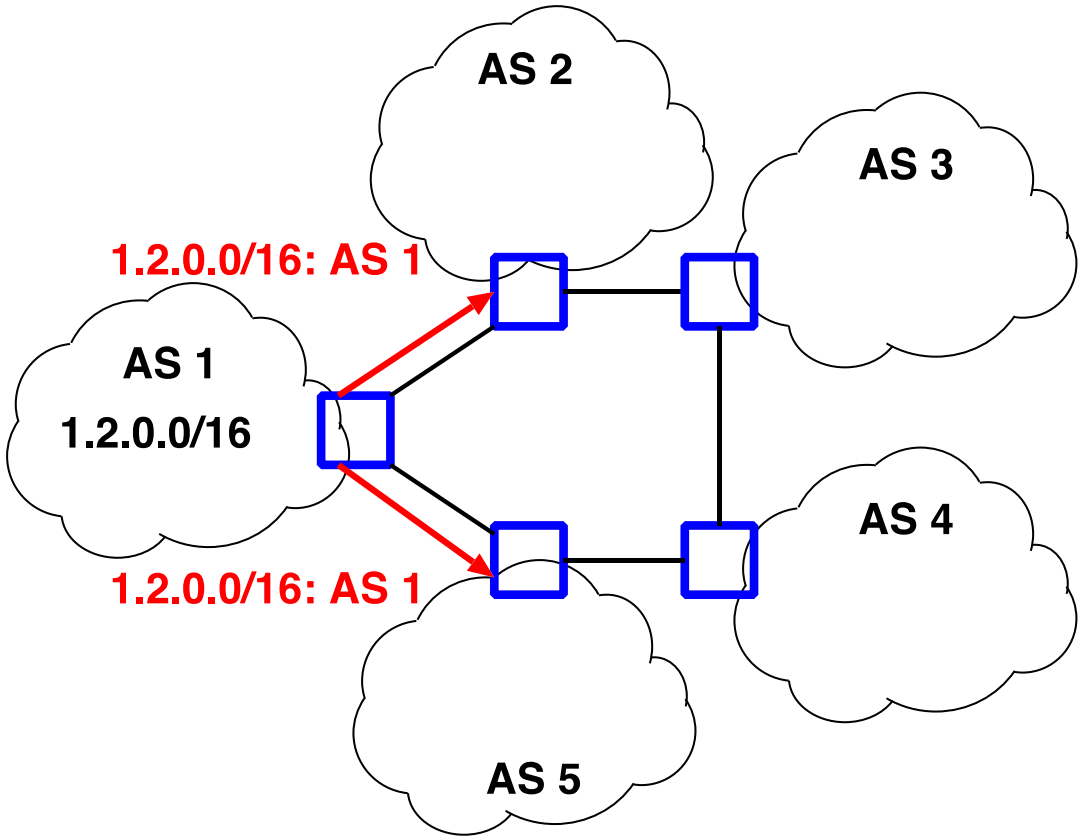
BGP attributes (2)

- **AS-PATH: contains ASs through which prefix advertisement has passed: prevent loops**
  - Well-known, mandatory.
  - If forwarding to internal peer:
    - do not modify AS_PATH attribute
  - If forwarding to external peer:
    - prepend self into the path.

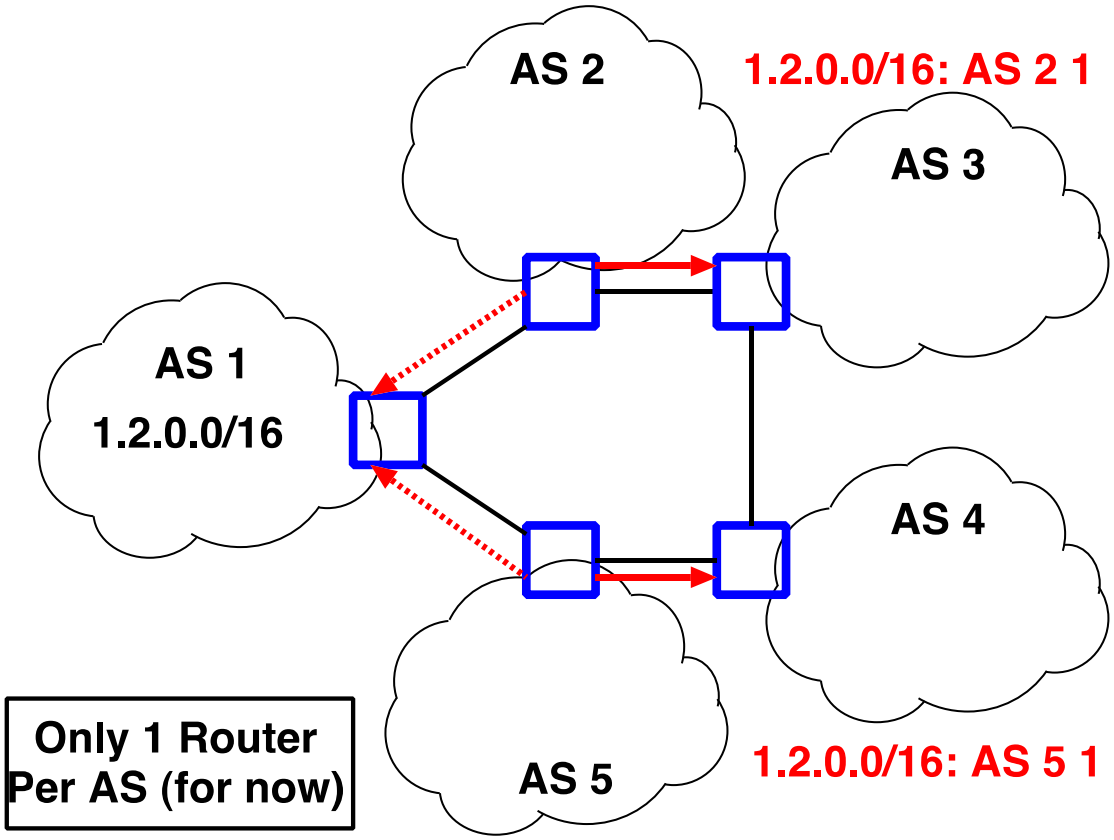
BGP Example



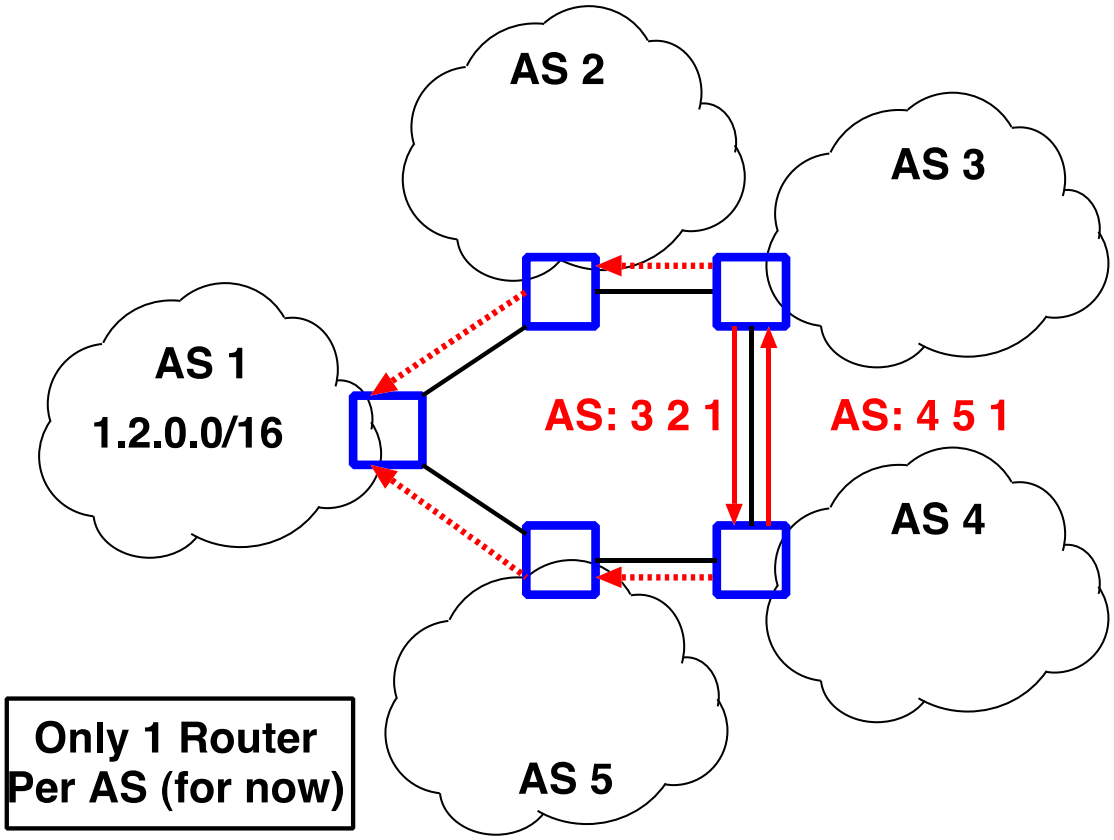
BGP Example



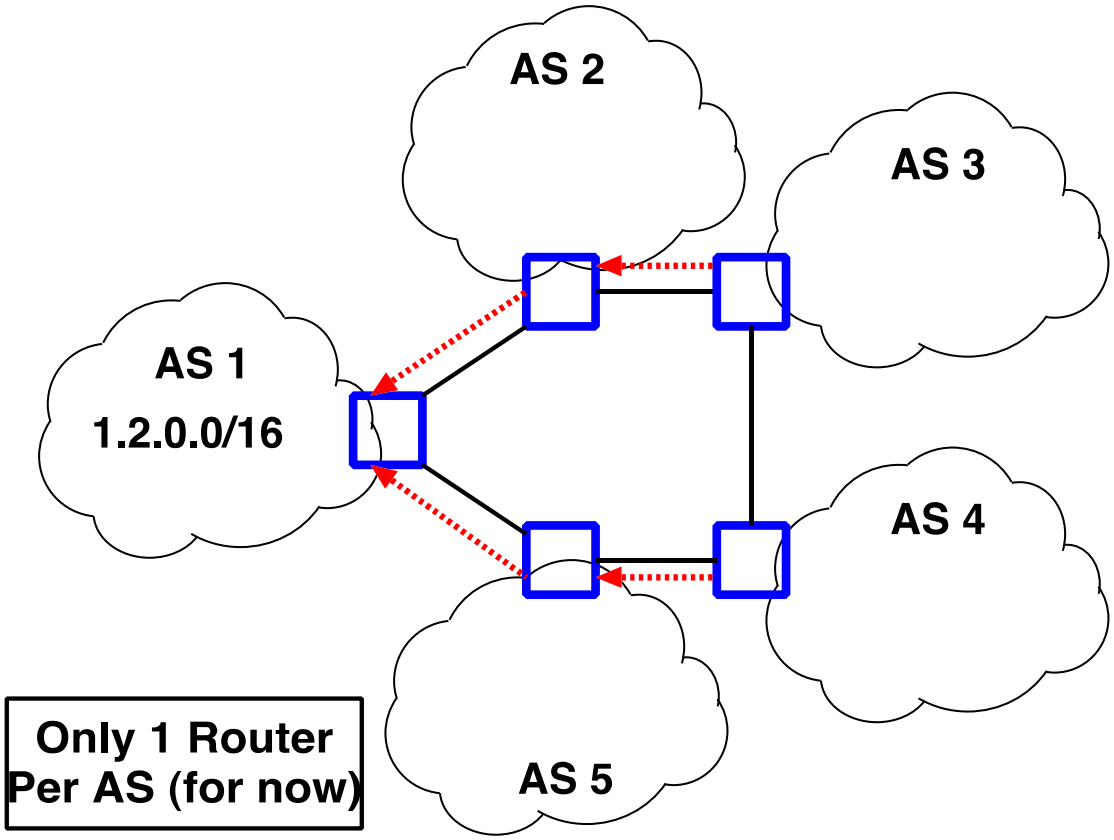
BGP Example


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)

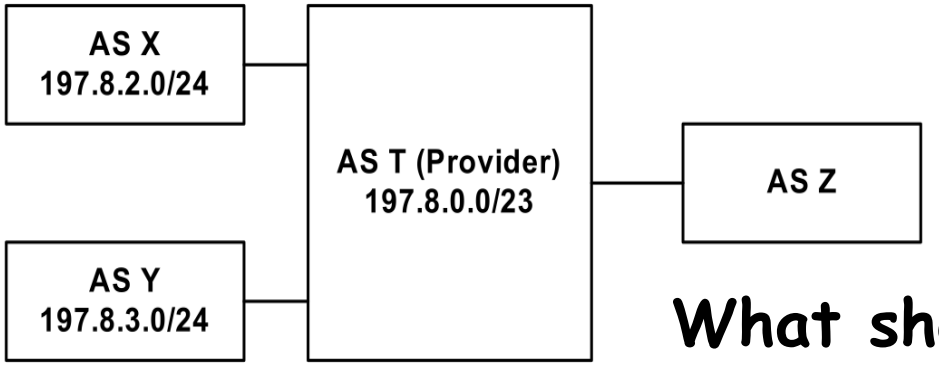
BGP Example


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



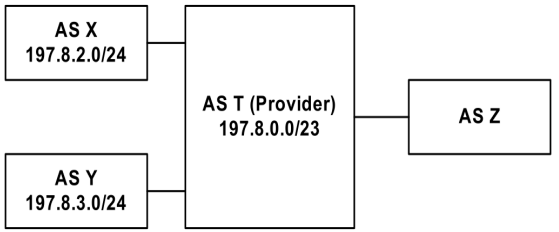
CIDR and BGP



What should T announce to Z?

- **Advertise all paths:**
  - Path 1: through T can reach 197.8.0.0/23
  - Path 2: through T can reach 197.8.2.0/24
  - Path 3: through T can reach 197.8.3.0/24
- **But this does not reduce routing tables! We would like to advertise:**
  - Path 1: through T can reach 197.8.0.0/22

Sets and Sequences



- **Problem: what do we list in the route?**
  - list T: omitting information - not acceptable, may lead to loops
  - list T, X, Y: misleading, appears as 3-hop path
- **Solution: restructure AS Path attribute as:**
  - Path: (Sequence (T), Set (X, Y))
  - if Z wants to advertise path:
    - Path: (Sequence (Z, T), Set (X, Y))

BGP attributes (3)

- **NEXT-HOP:** Indicates specific internal-AS router to next-hop AS. NEXT_HOP is always the IP address of the first router in the next autonomous system. (There may be multiple links from current AS to next-hop-AS.) NEXT_HOPs are only changed across eBGP sessions, but left intact across IBGP sessions.
- **Community**
  - no-export—Do not advertise this route to EBGP peers.
  - no-advertise—Do not advertise this route to any peer.
  - internet—Advertise this route to the Internet community; all routers in the network belong to it.

Routing information bases (RIB)

- **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

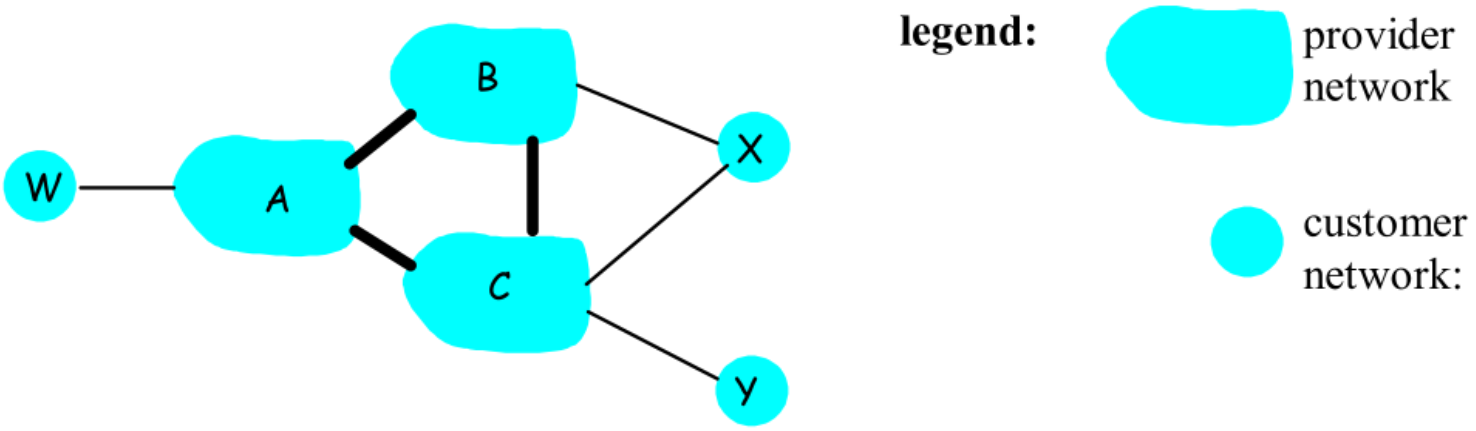
BGP route selection

- **Router may learn about more than 1 route to some prefix. Router must select route.**
- **Elimination rules:**
  1. Local preference value attribute: policy decision
  2. Shortest AS-PATH.
  3. Closest NEXT-HOP router: hot potato routing.
  4. Additional criteria

BGP messages

- **BGP messages exchanged using TCP.**
- **BGP messages:**
  - OPEN: opens TCP connection to peer and authenticates sender
  - UPDATE: advertises new path (or withdraws old)
  - KEEPALIVE keeps connection alive in absence of UPDATES; also ACKs OPEN request
  - NOTIFICATION: reports errors in previous msg; also used to close connection
- **Extensions can define more message types**
  - E.g., ROUTE-REFRESH [RFC 2918]

BGP routing policy

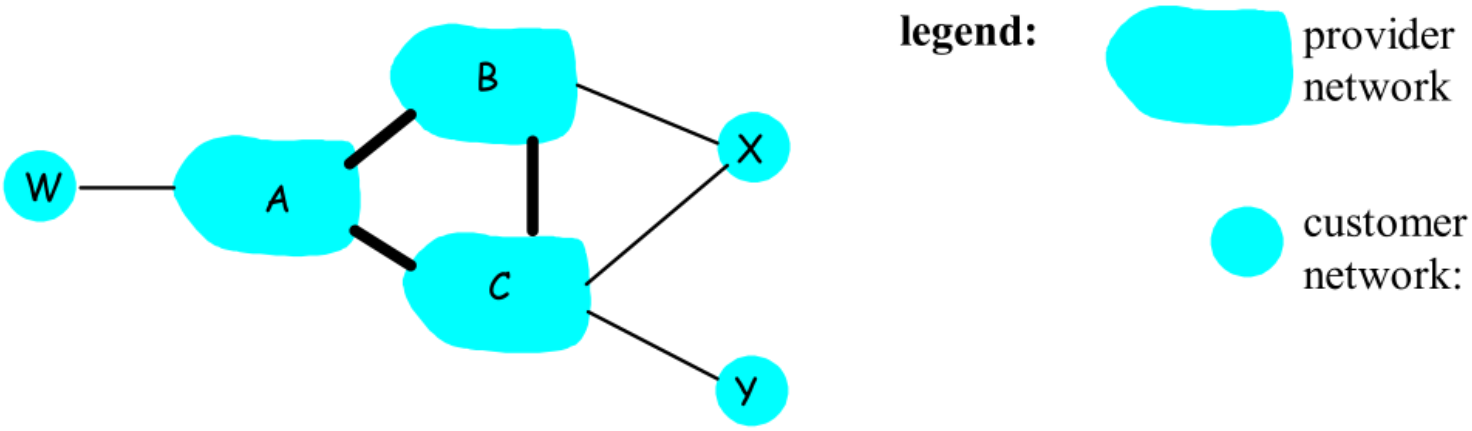


- A,B,C are provider networks
- X,W,Y are customer (of provider networks) - stub networks
- X is dual-homed: attached to two networks
- X does not want to route from B via X to C
- .. so X will not advertise to B a route to C

AS categories

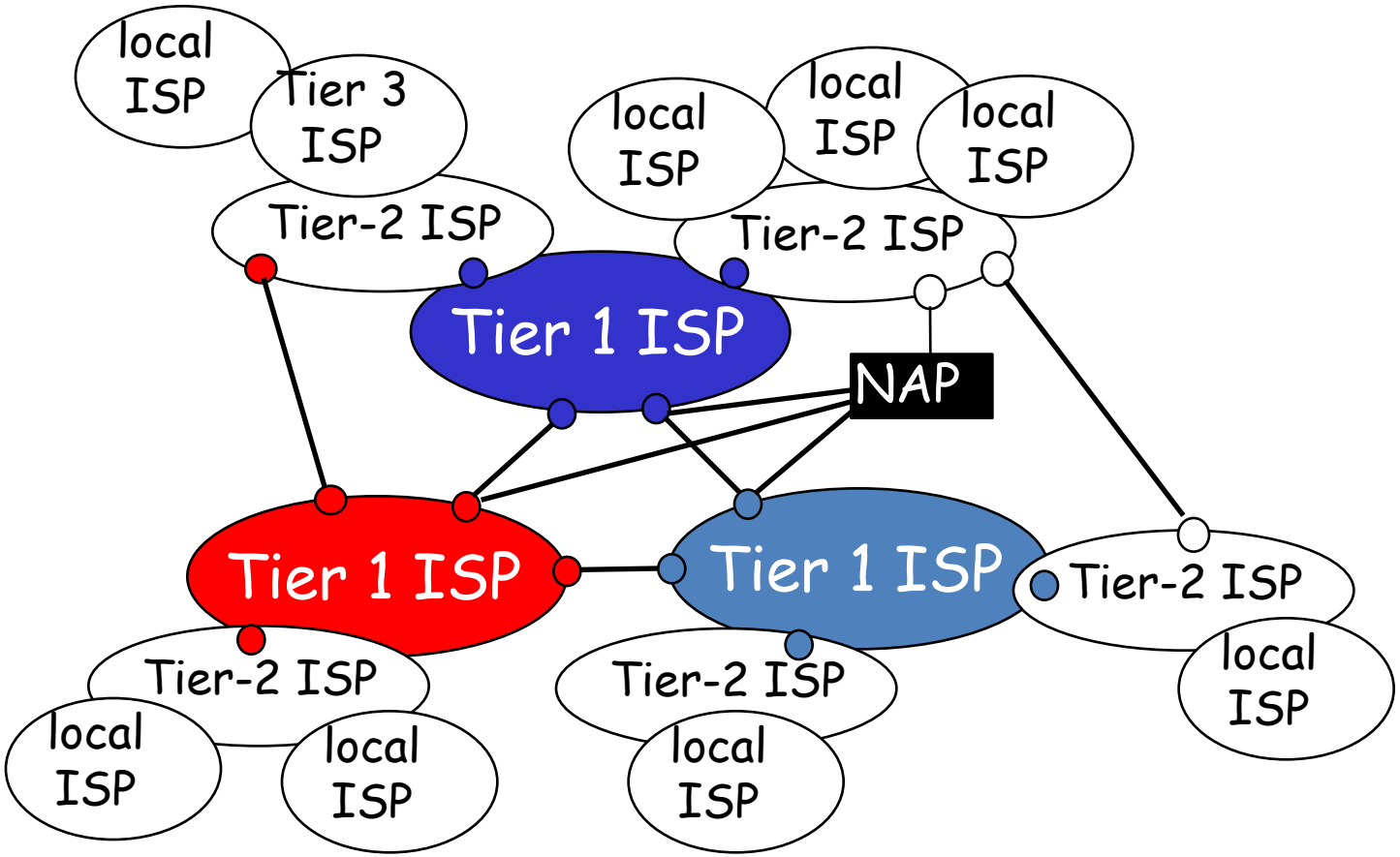
- Stub: an AS that has only a single connection to one other AS - carries only local traffic.
- Multihomed: an AS that has connections to more than one AS, but refuses to carry transit traffic
- Transit: an AS that has connections to more than one AS, and carries both transit and local traffic (under certain policy restrictions)

BGP routing policy (2)



- A advertises to B the path AW
- B advertises to X the path BAW
- Should B advertise to C the path BAW?
  - No way! B gets no "revenue" for routing CBAW since neither W nor C are B's customers
  - B wants to force C to route to w via A
  - B wants to route only to/from its customers!

Internet structure: network of networks



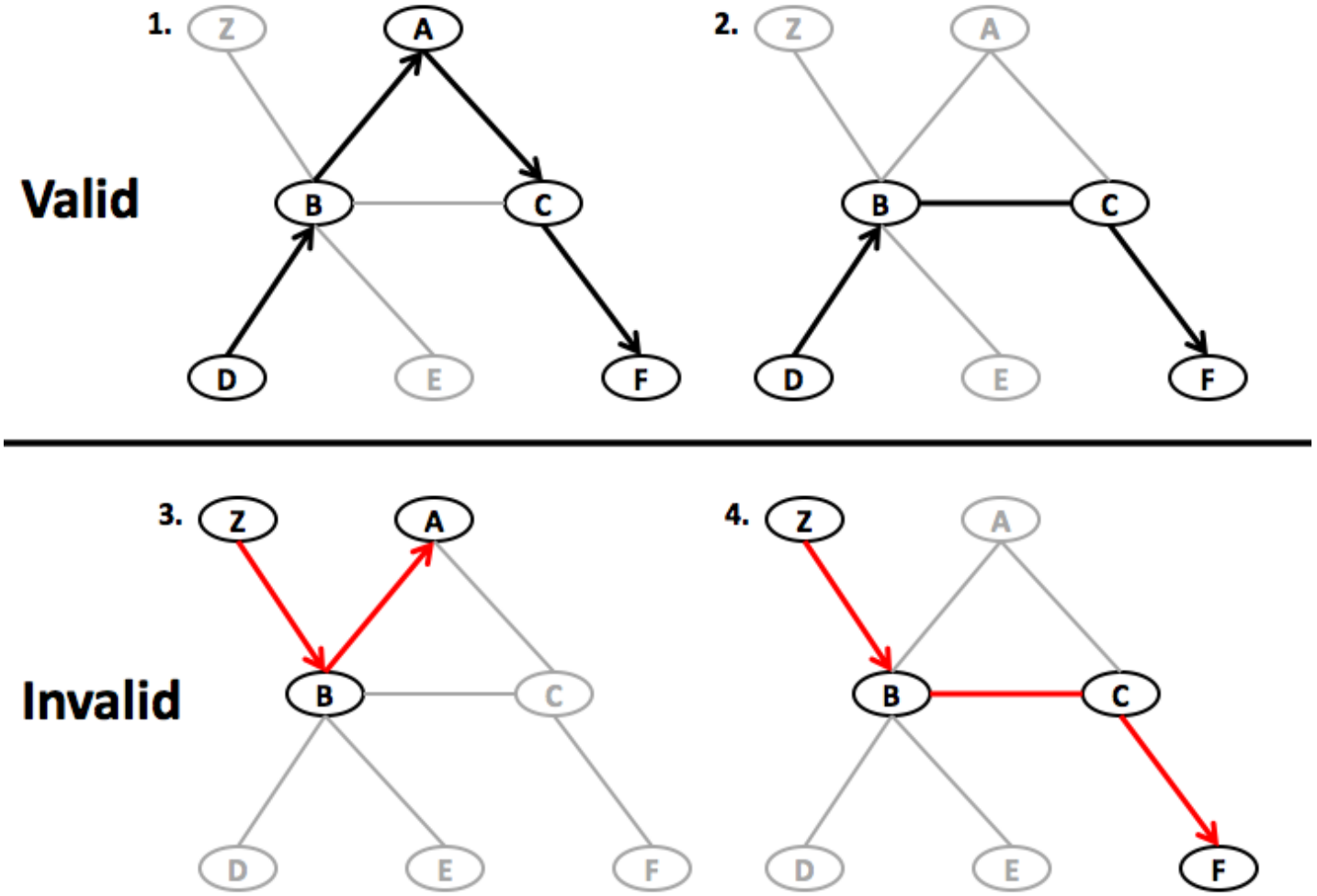
Structure of ASs

- **3 Types of relationships (Customer, Provider, Peer)**
  - Customer-Provider: customer AS pays provider AS for access to rest of Internet: provider provides transit service
    - End customers pay ISPs, and ISPs in lower “tiers” pay ISPs in higher tiers
  - Peers: ASs that allow each other transit service
    - ISPs on same tier, usually involvesno fees
- **Customer-Backup Provider**: Provider if primary provider fails. May be peers otherwise

AS BGP Policies

- AS Policy for its customers - an AS gives its customers transit services toward all of its neighboring ASes.
- AS Policy for its providers - an AS gives its providers transit services only toward its customers.
- AS Policy for its peers - an AS gives its peers transit services only toward its customers.
- “Valley free” paths.

“Valley free”



Why different Intra- and Inter-AS routing ?

Policy:

- Inter-AS: admin wants control over how its traffic routed, who routes through its net.
- Intra-AS: single admin, so no policy decisions needed

Scale:

- hierarchical routing saves table size, reduced update traffic

Performance:

- Intra-AS: can focus on performance
- Inter-AS: policy may dominate over performance

News

GIZMODO

CYBERWAR

China's Internet Hijacking Uncovered

Cybercrime experts have found proof that China hijacked the Internet for 18 minutes last April. China absorbed 15% of the traffic from US military and civilian networks, as well as from other Western countries—a massive chunk. Nobody knows why.

BY JESUS DIAZ NOV 17, 2010 10:00 AM

Share +1 Like 1k 90,770 460

CNET › News › Security

Report: China hijacked U.S. Internet data

by Lance Whitney | October 22, 2010 10:27 AM PDT

Follow

A Chinese state-run telecom provider was the source of the redirection of U.S. military and corporate data that occurred this past April, according to excerpts of a draft report sent to CNET by the U.S.-China Economic and Security Review Commission.

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 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)

Next class

- **More Network layer as time permitted**
  - BGP issues
  - IPv6
  - Mobile IP
  - Mulicast