E6998-02: Internet Routing

Lecture 15
Border Gateway Protocol, Part IV

John Ioannidis

AT&T Labs – Research

ji+ir@cs.columbia.edu

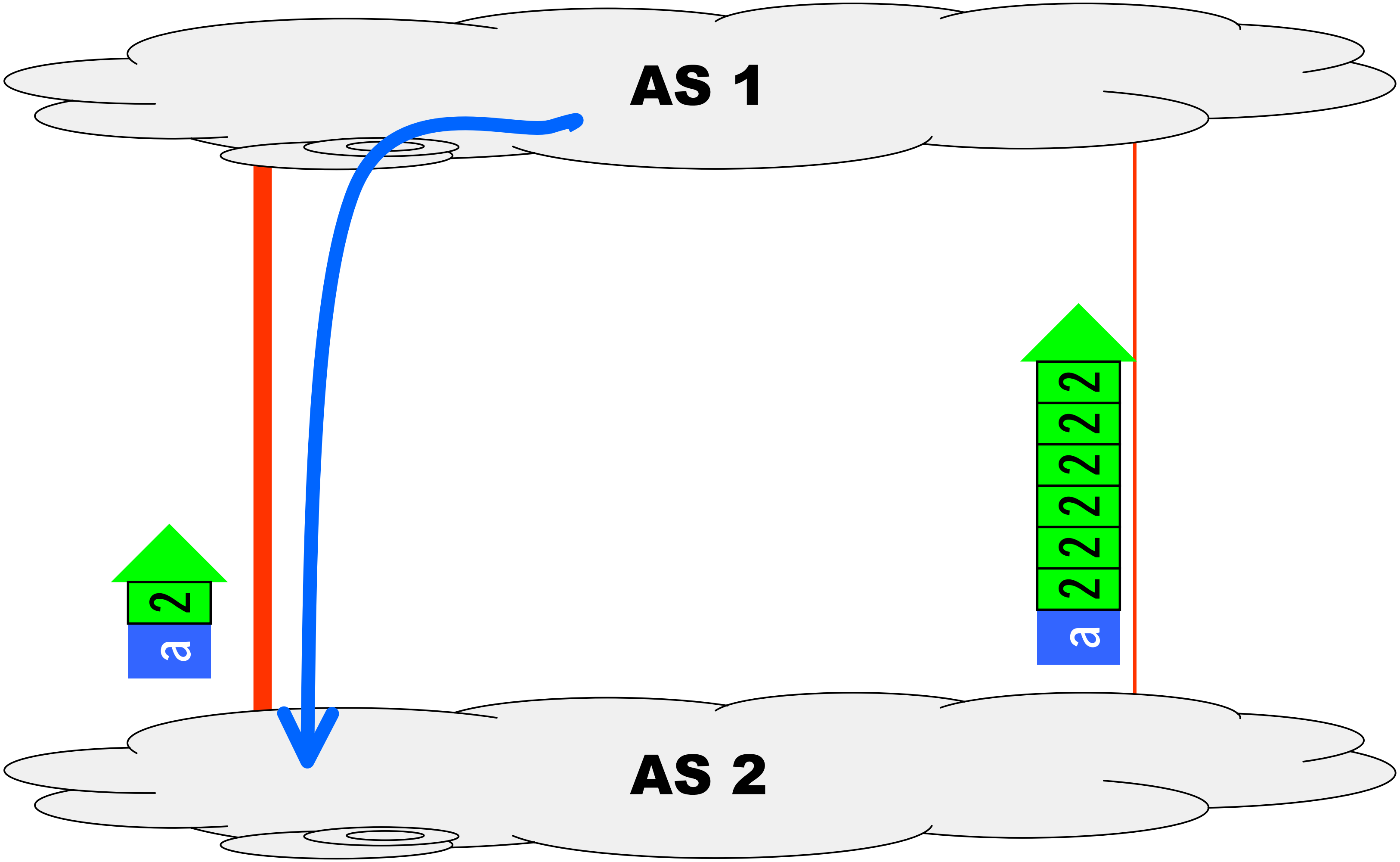
Announcements

Lectures 1-15 are available.

Homework 4 will be available tomorrow, due 11/12.

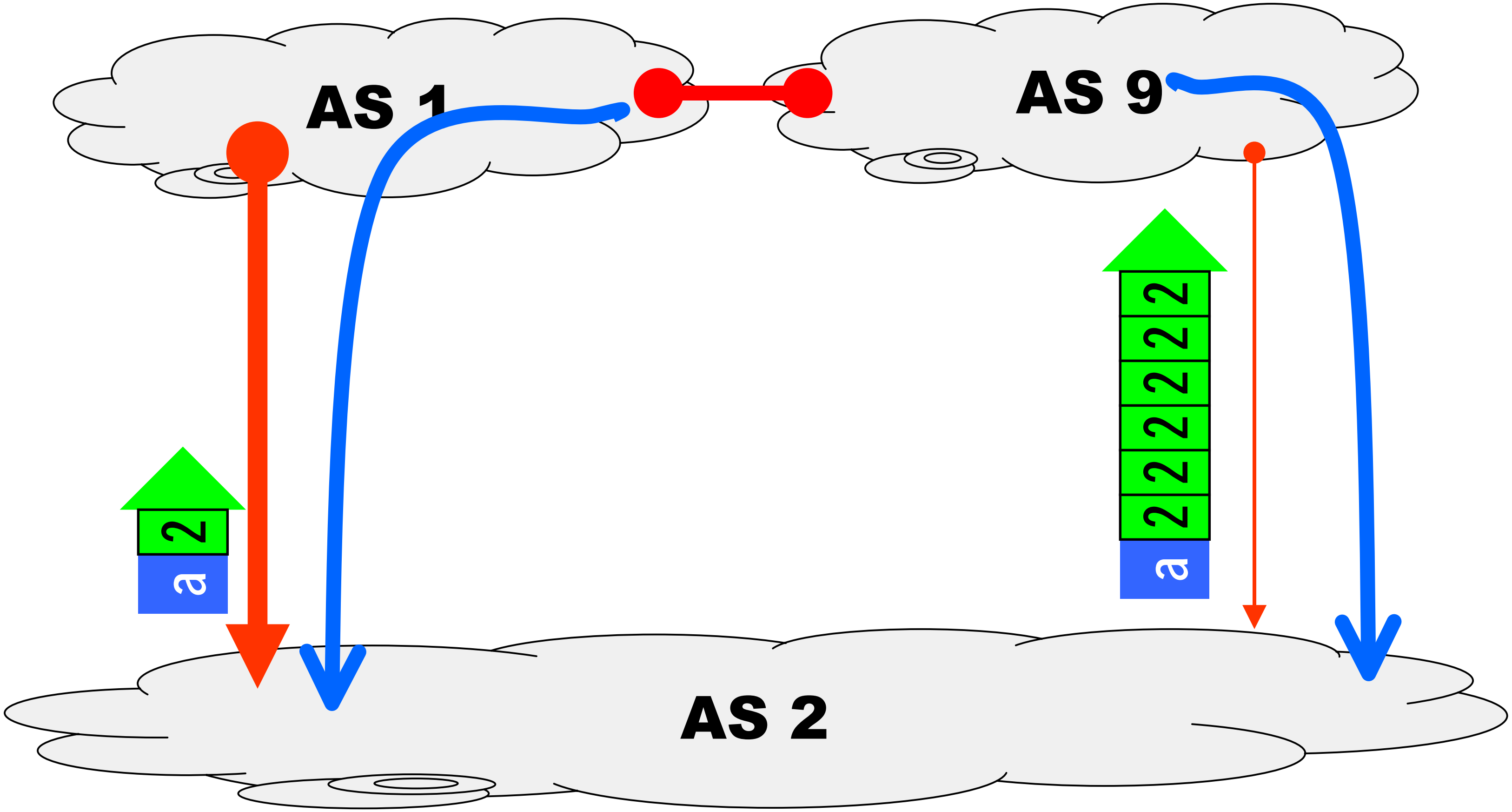
Backup Links (inbound traffic)

- Hack: AS_PATH padding.



Backup Links (inbound traffic)

- AS_PATH padding does not shut off all traffic.
- AS 9 has higher LOCAL_PREF for customer routes.
- Some traffic from AS9 still flows through the backup link.

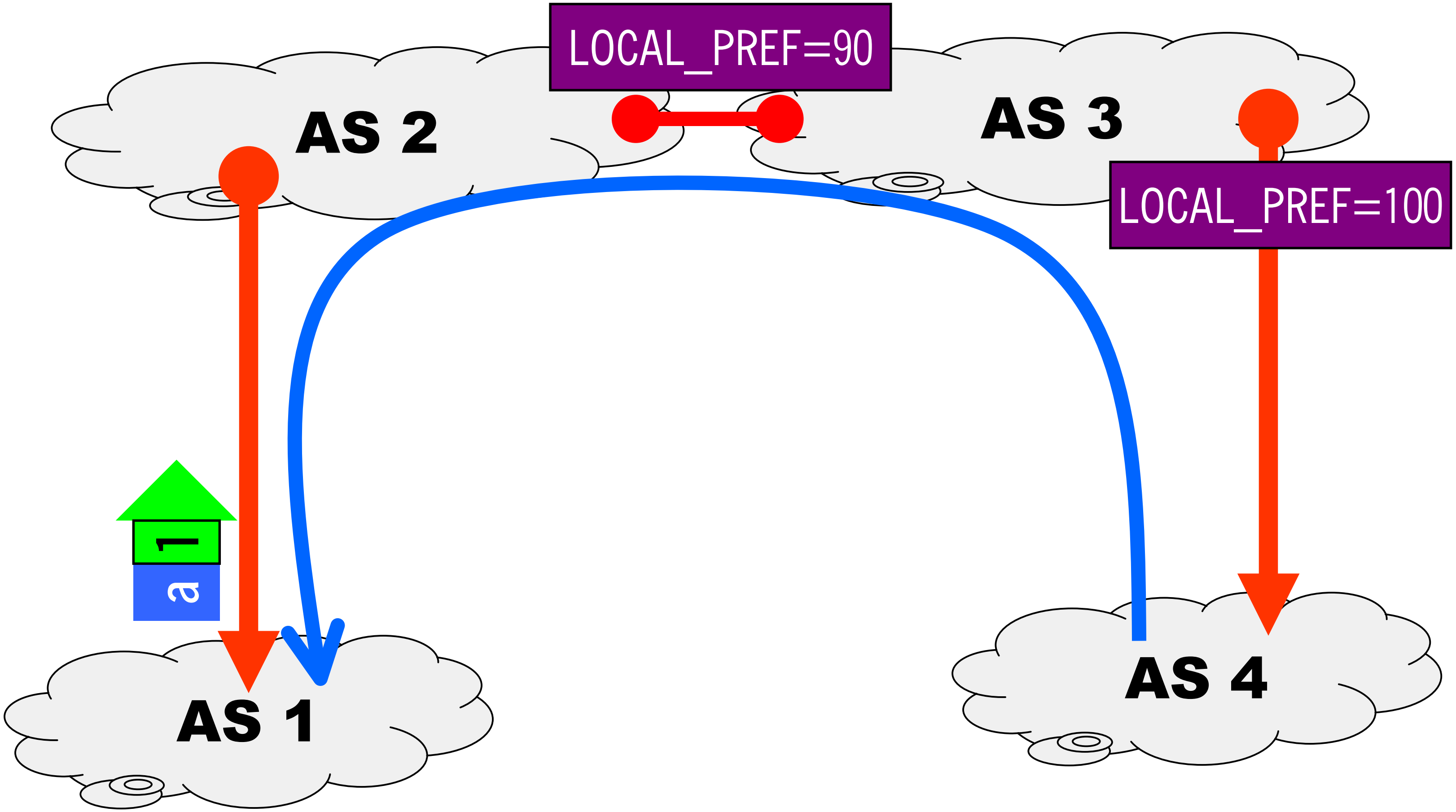


Backup links (inbound traffic)

- COMMUNITY to the rescue!
- AS9 has LOCAL_PREF = 100 for customer and 90 for peer.
- AS9 has the following import policy:
  - If 9:90 in community, set local_pref to 90.
  - If 9:80 in community, set local_pref to 80.
  - If 9:70 in community, set local_pref to 70.
- AS2 advertises its routes (over the backup link to AS9) with community 9:70.
- Now peer has higher local pref and traffic flows as intended!

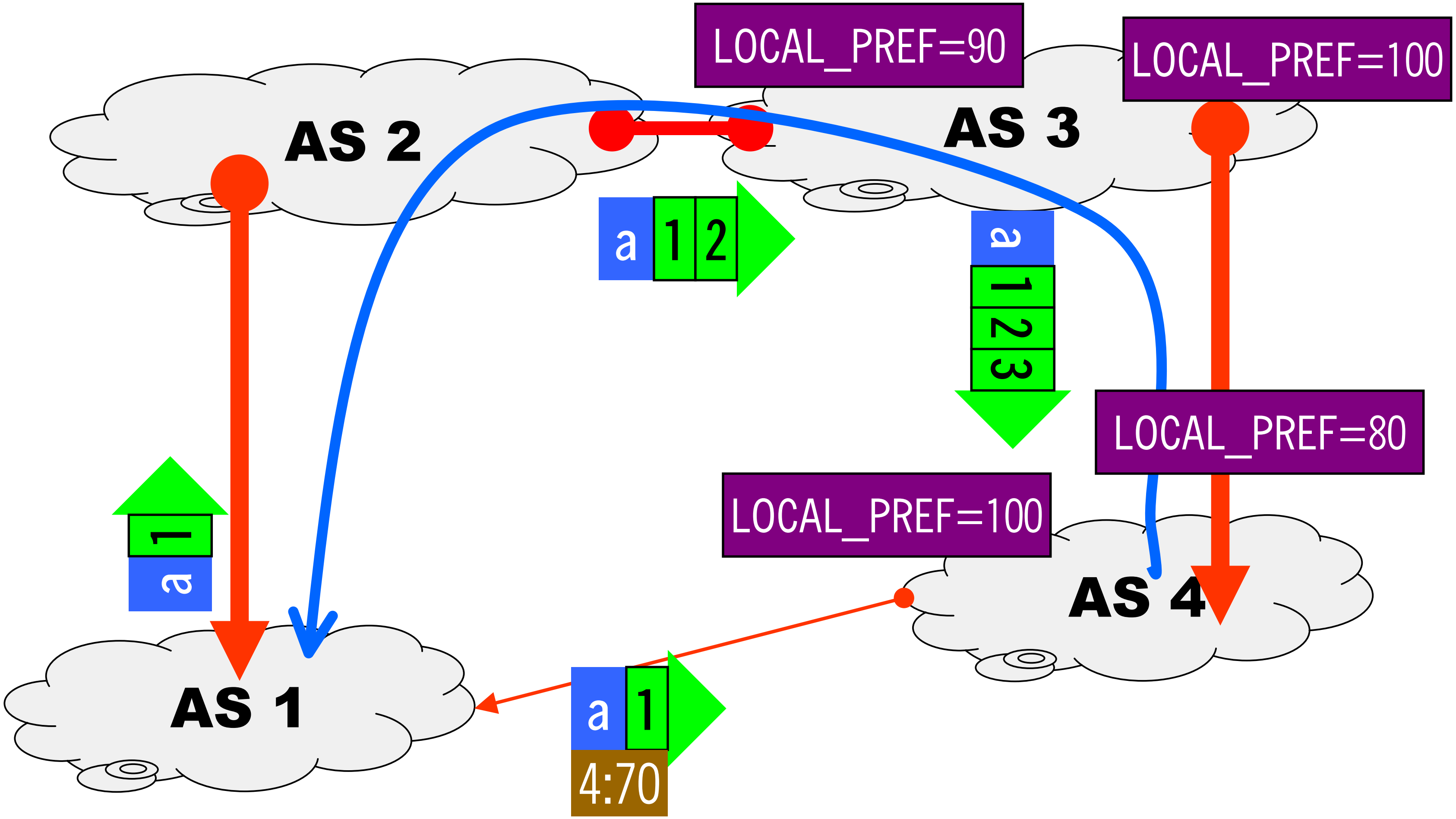
Policy Interaction

- Example: backup route with community hack.
- AS4 advertises prefix a over its (only) link.



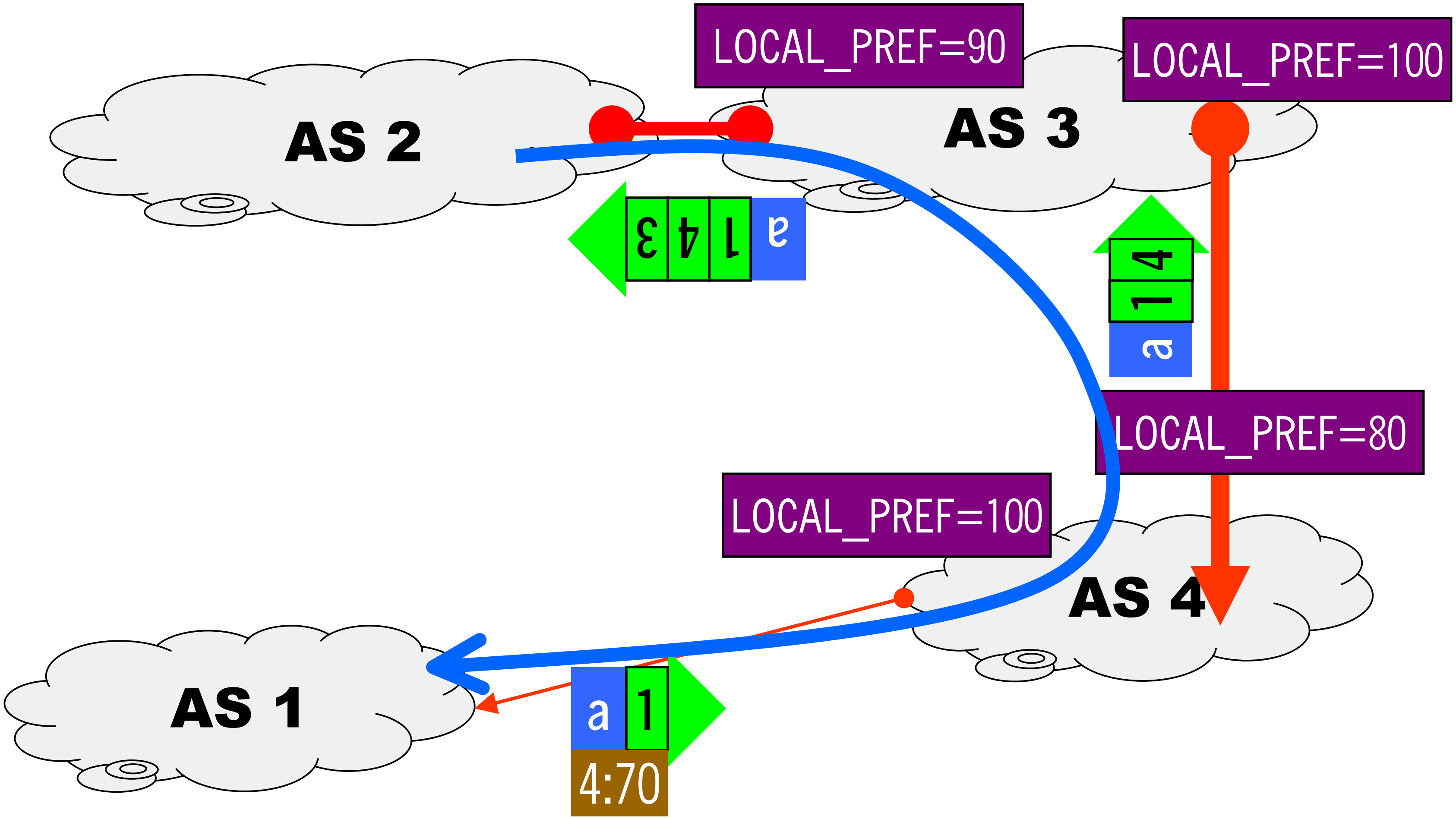
Policy Interaction cont'd

- Backup link gets installed, AS1 advertises community 4:70.
- AS4 still prefers route via AS3 (highest local_pref).



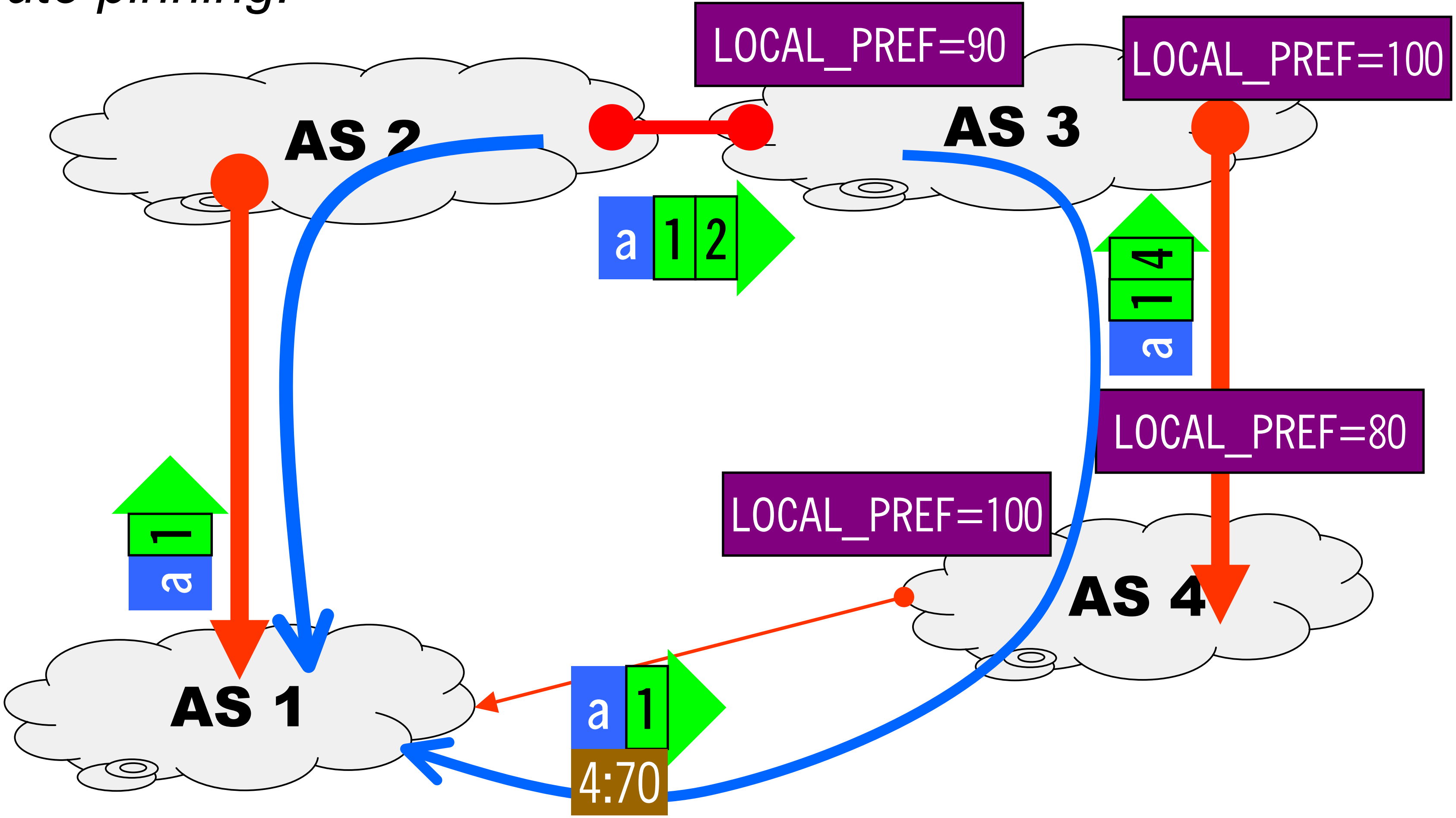
Backhoe Severs Primary Link

- AS2 withdraws route to a.
- Backup link takes over.



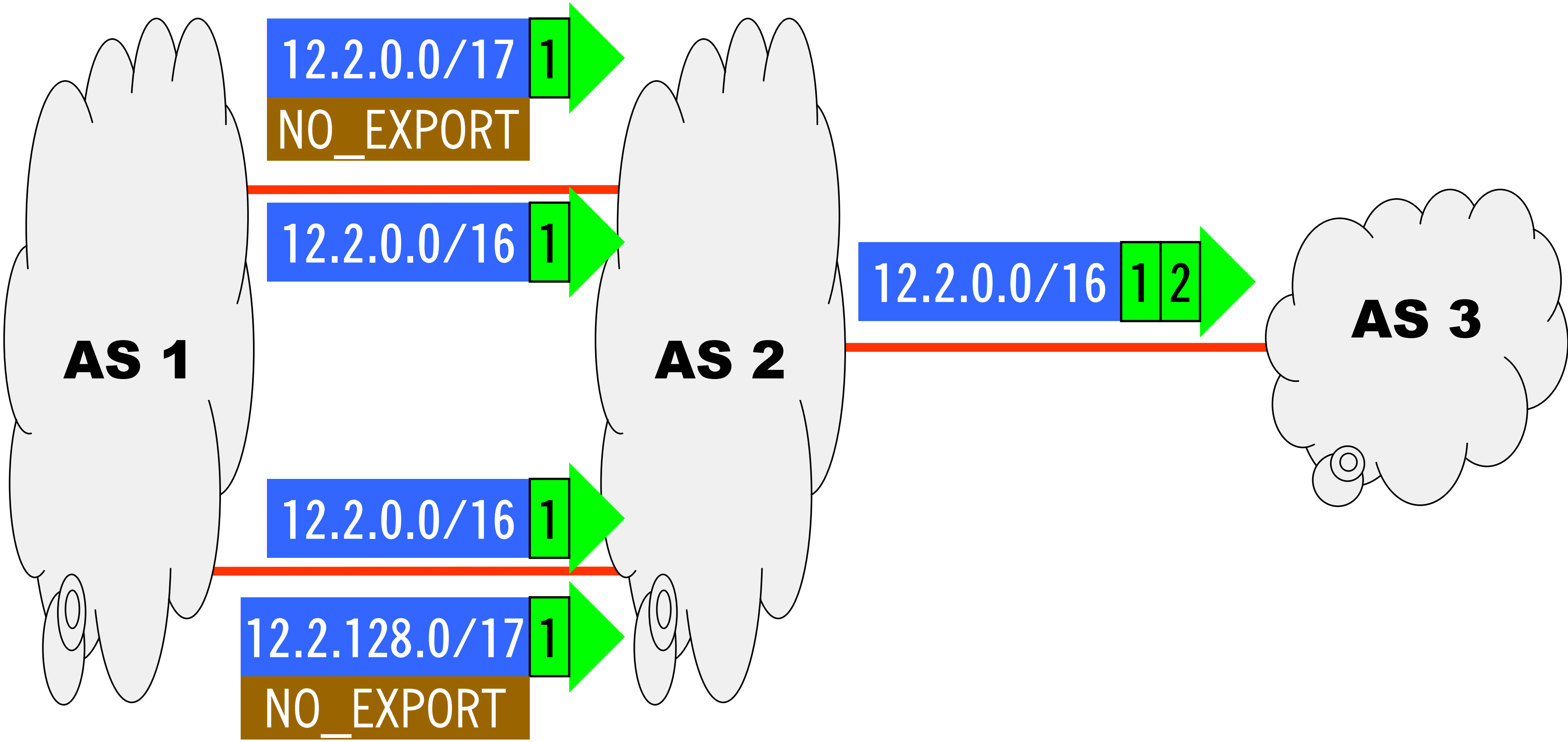
Primary link restored

- AS4 is still advertising route to AS1.
- Route from AS2 has lower local pref, gets ignored!
- *Route pinning.*



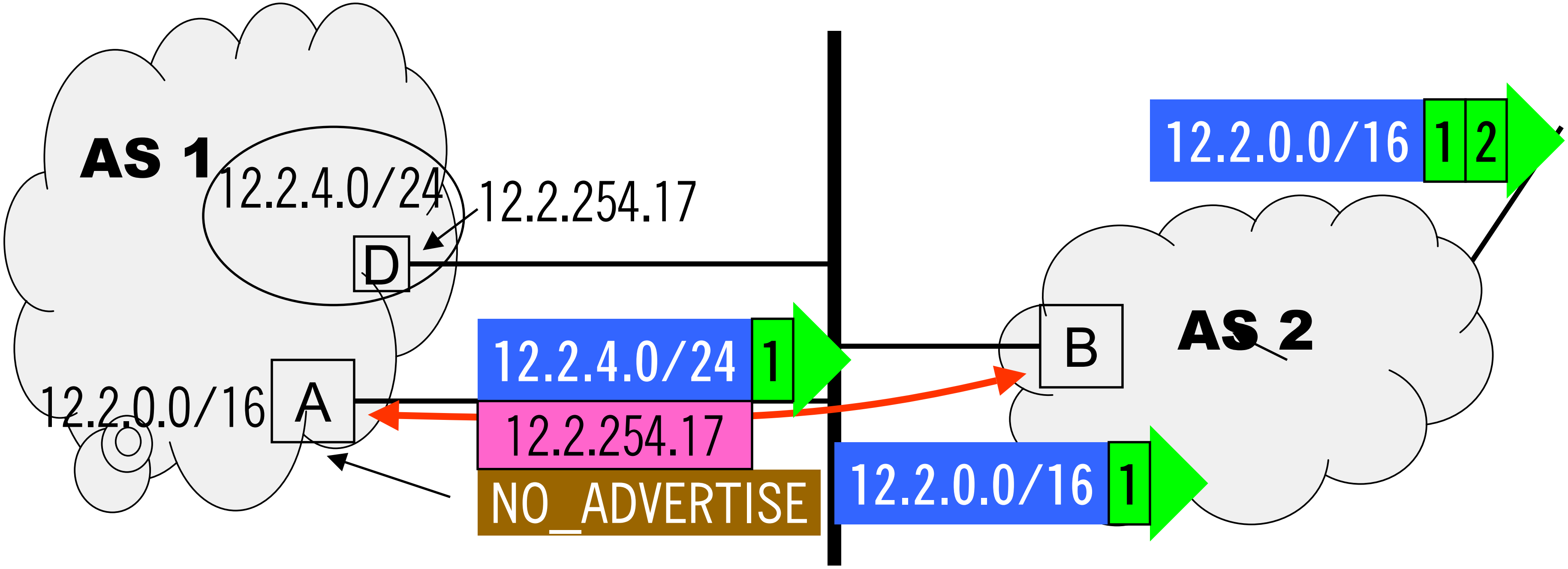
NO_EXPORT (0xFFFFFF01)

- Received routes with the NO_EXPORT community are not re-advertised beyond the receiving AS.



NO_ADVERTISE (0xFFFFFF02)

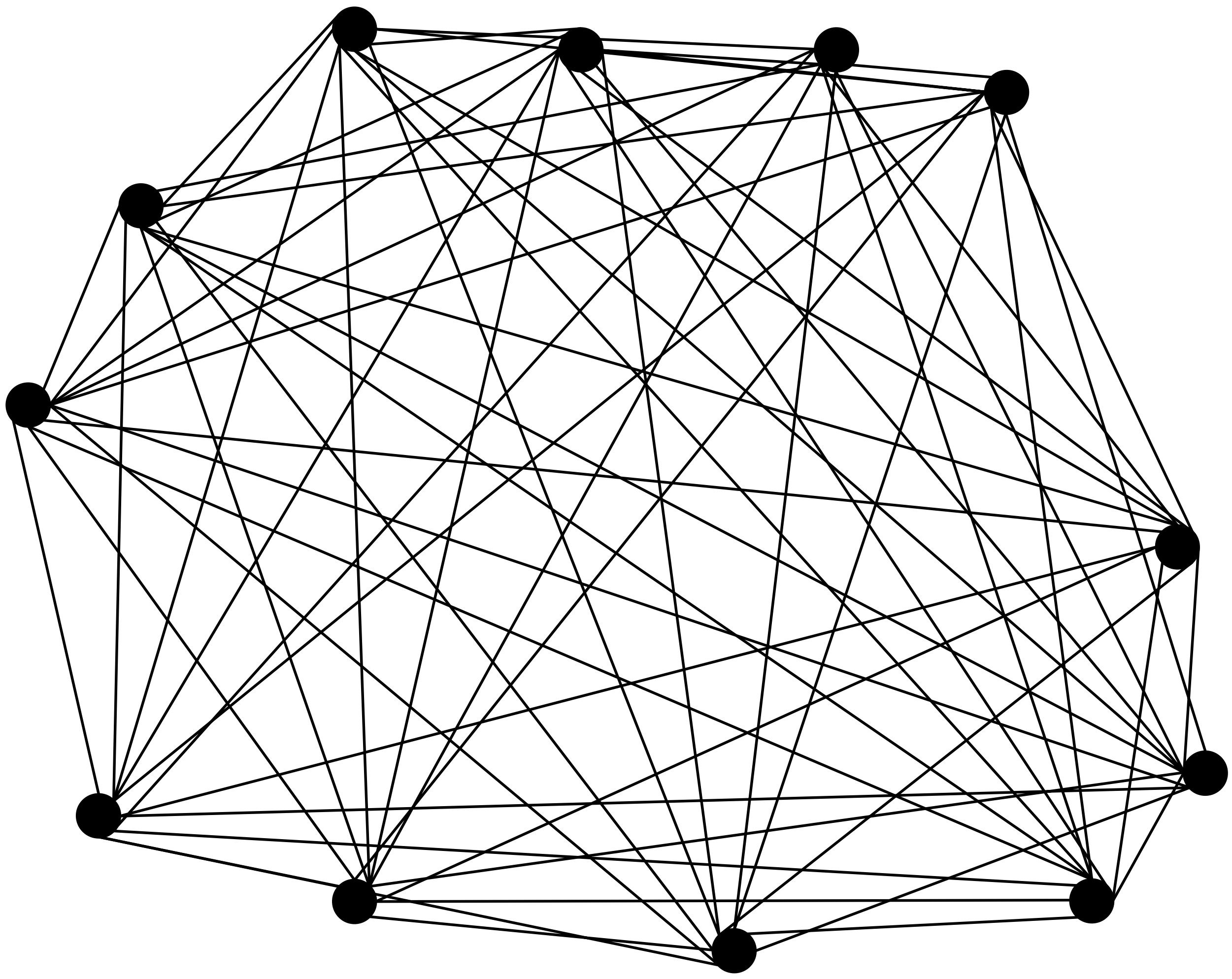
- Used in conjunction with the third-party NEXT_HOP.
- Most of AS1 is behind A.
- D does not speak BGP.
- AS1 advertises 12.2.4.0/24 with the NO_ADVERTISE.
- B uses D to forward packets to 12.2.4.0/24.
- This fine structure is not exported beyond AS2.



I-BGP Scaling

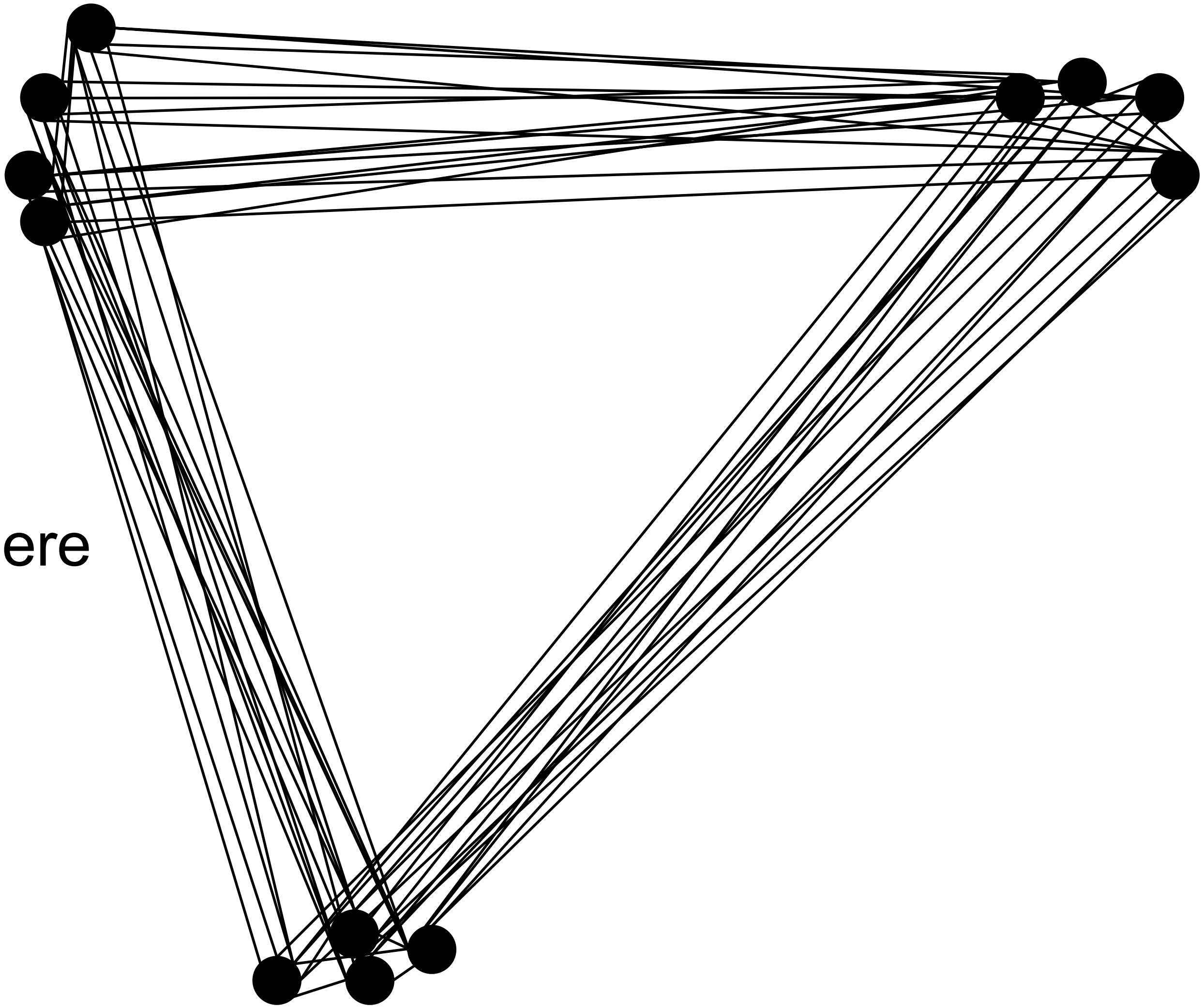
- I-BGP peering sessions can be wasteful of resources.

  (Lines represent I-BGP sessions, NOT physical links!)



I-BGP Scaling

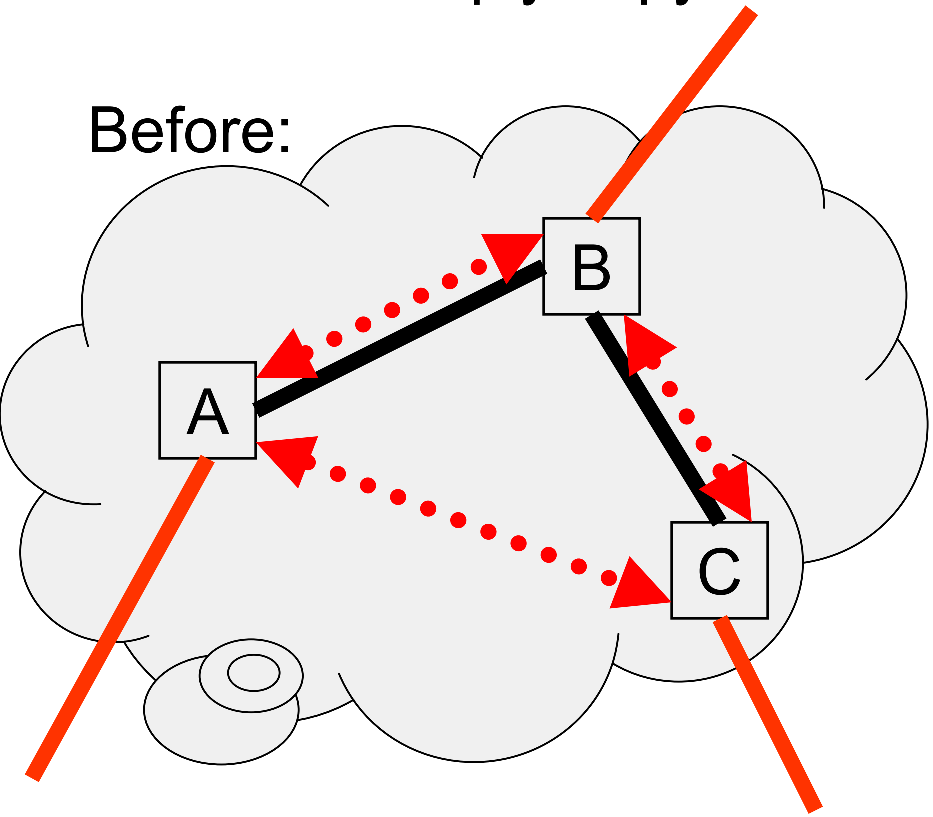
- Really wasteful!
  - CPU
  - Memory
  - Link capacity
- Poor scaling.
- Replicated traffic.
  - Chances are there is only one link between each group of four routers in the picture!



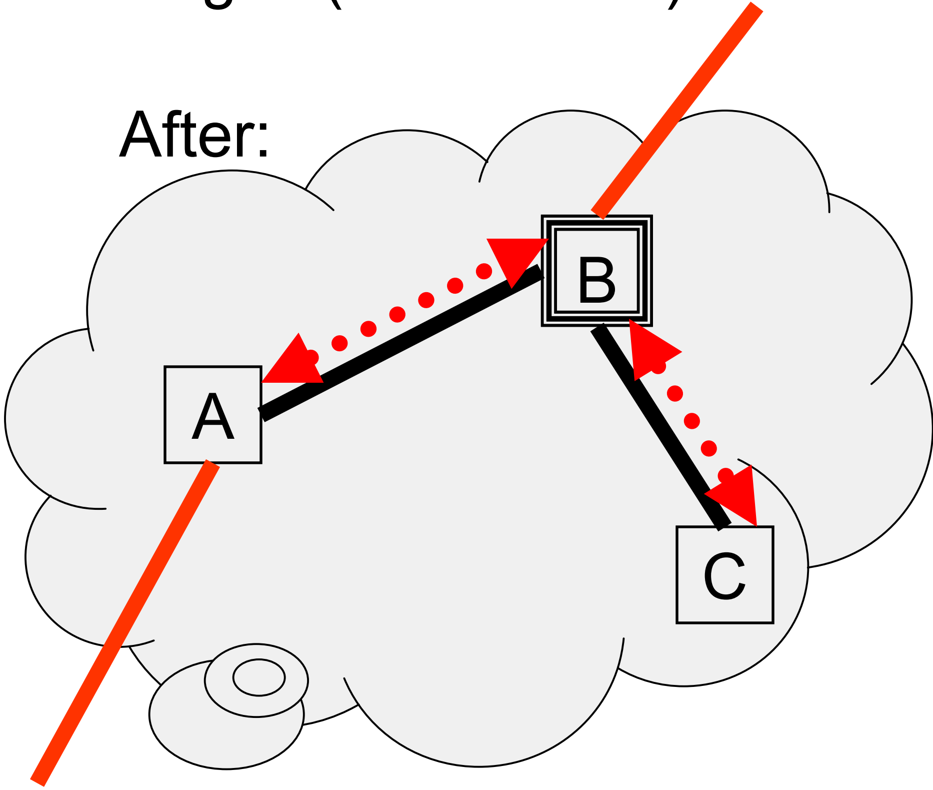
Route Reflection

- Relax the rule about not re-advertising I-BGP-learned routes.
  - – Add hierarchy to I-BGP.
- Reduces # of sessions.
- RR can simply copy UPDATE messages (saves CPU).

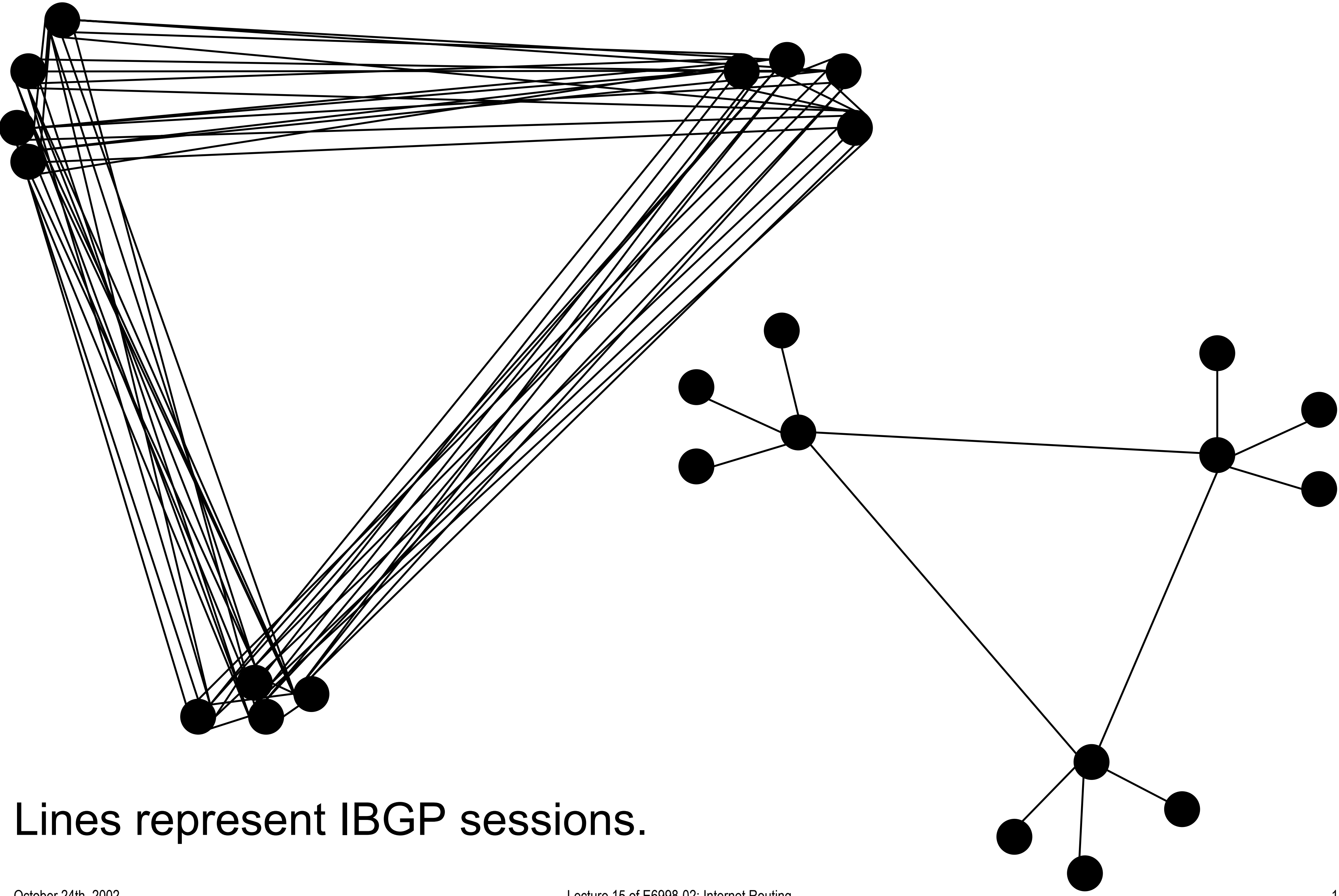
Before:



After:



Before/After



Lines represent IBGP sessions.

Route Reflection, cont'd

- I-BGP peers of a Route Reflector:
  - *Clients*
  - *Non-clients*
- A RR and its clients form a *Cluster*.
- Non-clients still form a full I-BGP mesh with each other.
- Clients only talk to their RR
  - And external peers, of course.
- Clients are normal I-BGP peers.
  - All they know is that they have been configured to peer with the RR.
- Which routers become RR depends on the topology.
  - Ditto for clusters.

Route-Reflector Route Selection

- RR receiving multiple routes to same destination runs regular BGP route selection procedure.

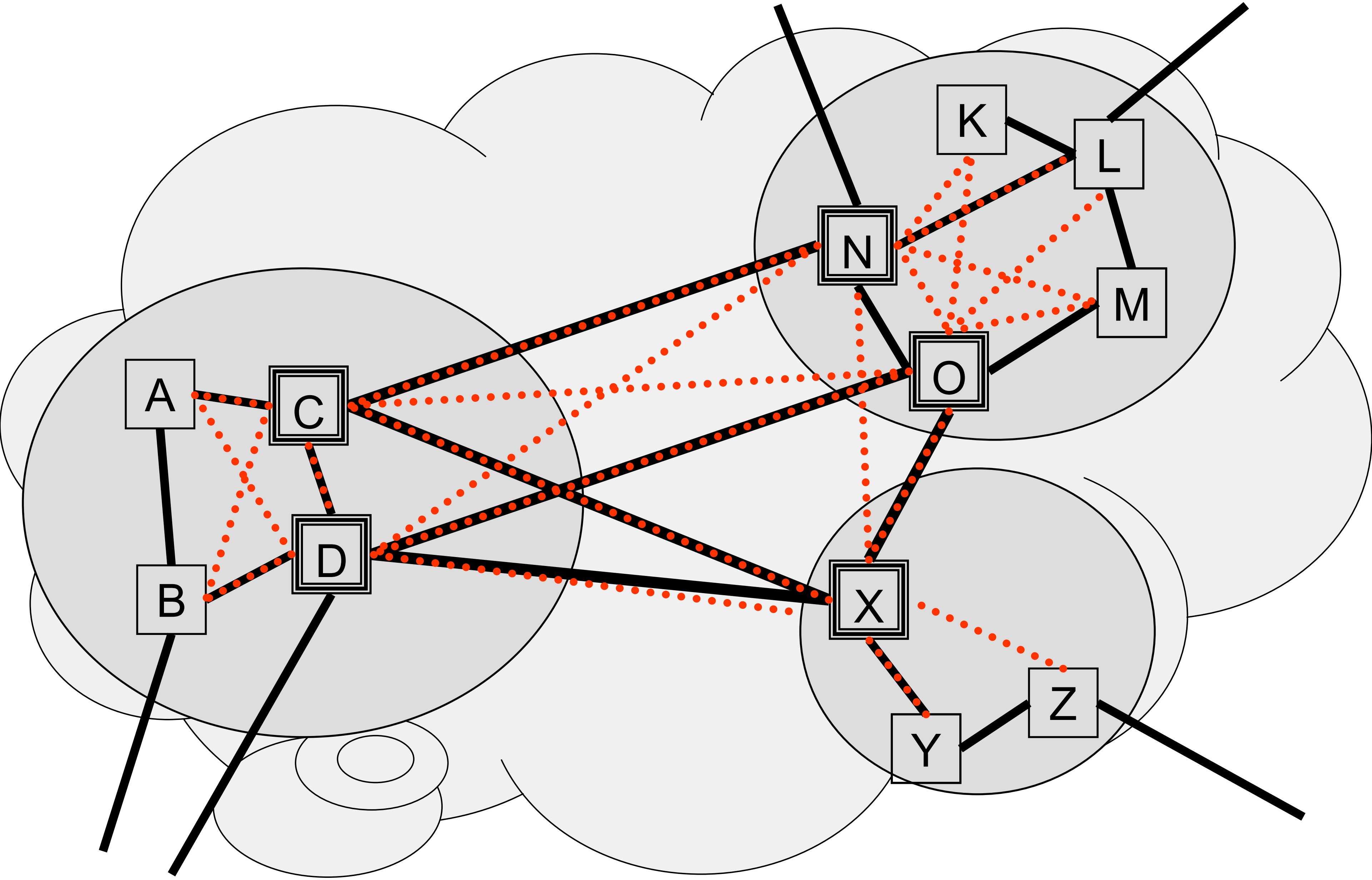
Received from:	**Reflect to:**
nonclient peer (RR or otherwise)	clients only
client	all other clients* all nonclient peers
EBGP	all clients all nonclient peers

*Except when clients are fully-meshed.

Redundancy in RR

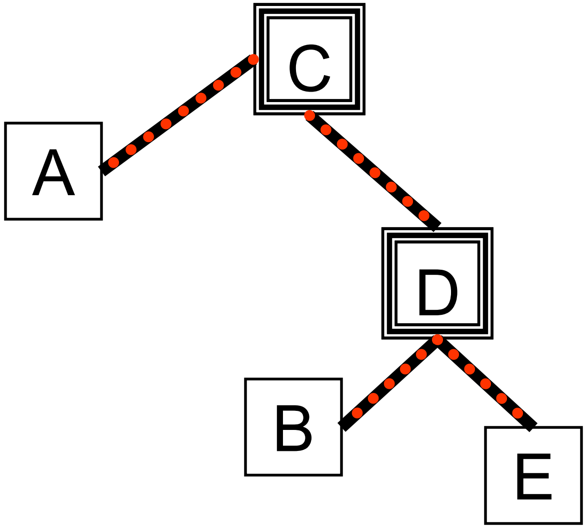
- If a route reflector goes down, I-BGP setup gets partitioned.
  - Not good!
- Redundancy.
- Each cluster gets at least two RRs.
  - Each client in the cluster talks to both RRs.
  - Yes, they get duplicate UPDATEs.
- RRs fully meshed.
- Clients can also be fully meshed inside a cluster.
  - RR must be configured not to readvertise to its own clients.
- Topology considerations.
  - I-BGP sessions should (if possible) flow over distinct links.

RR with Redundancy



Nested RR Configurations

- A client does not know it is a client!
  - A RR can be client of another RR.



- D is C’s client, but B&E’s RR.

RR and Attributes

- RR preserve BGP attributes.
- Necessary to avoid loops due to interactions with the IGP.
- NEXT_HOP in particular.

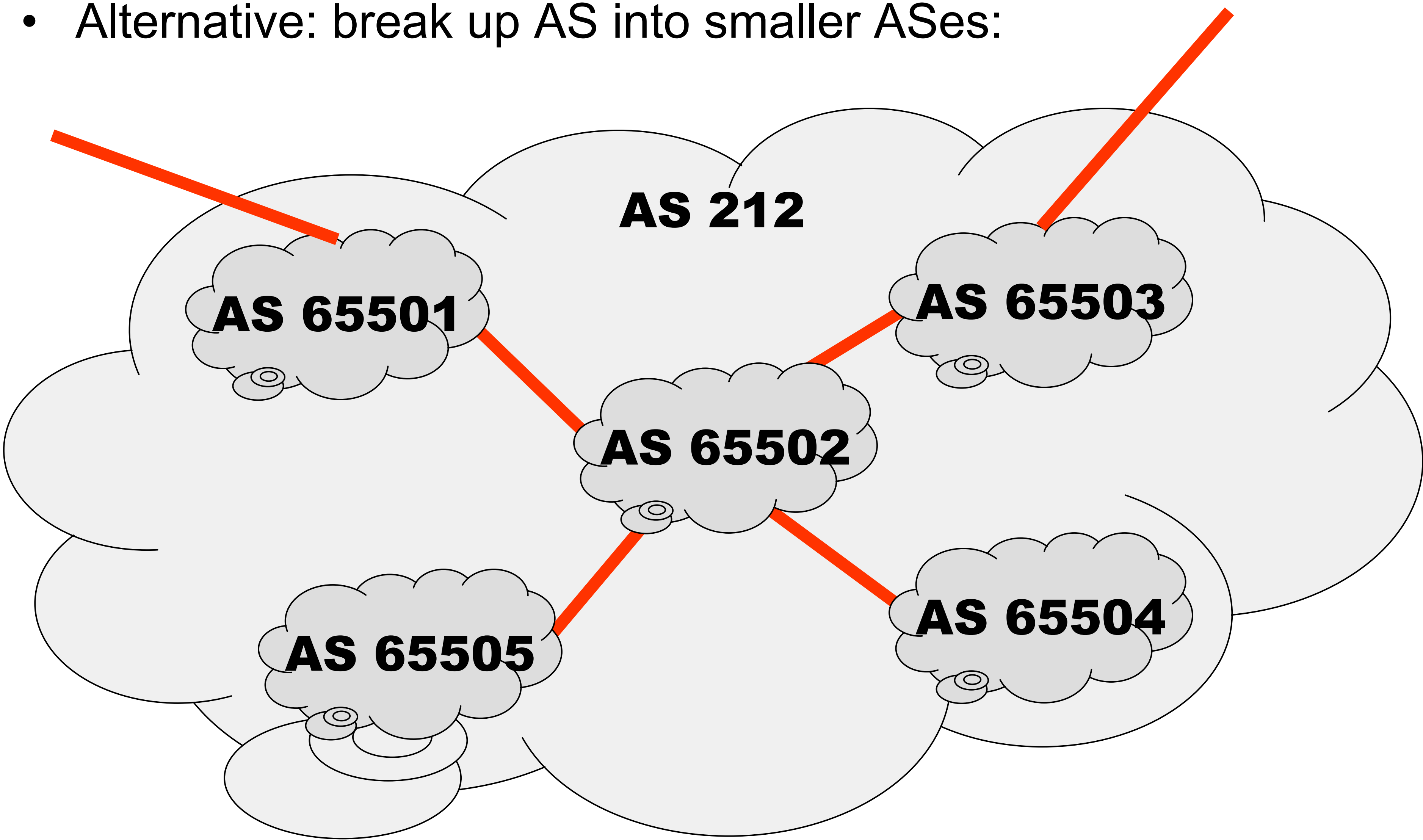
- Fewer actual paths are possible.
- Bizarre interactions can occur.
- RR/Clustering should follow topology.

Avoiding Loops

- Relaxation of the I-BGP re-advertising rule can lead to loops.
  - In cases of misconfiguration.
- ORIGINATOR_ID
  - Optional, non-transitive (type code 9).
  - Router ID of router that injected the route.
  - Added by the RR.
- CLUSTER_LIST
  - Optional, non-transitive (type code 10).
  - List of clusters that an UPDATE has traversed.
    - CLUSTER_ID should be the same in RRs of the same cluster.
  - Also added by the RR.
  - Remind you of anything?

Confederations

- RR enforces hierarchy.
- Alternative: break up AS into smaller ASes:

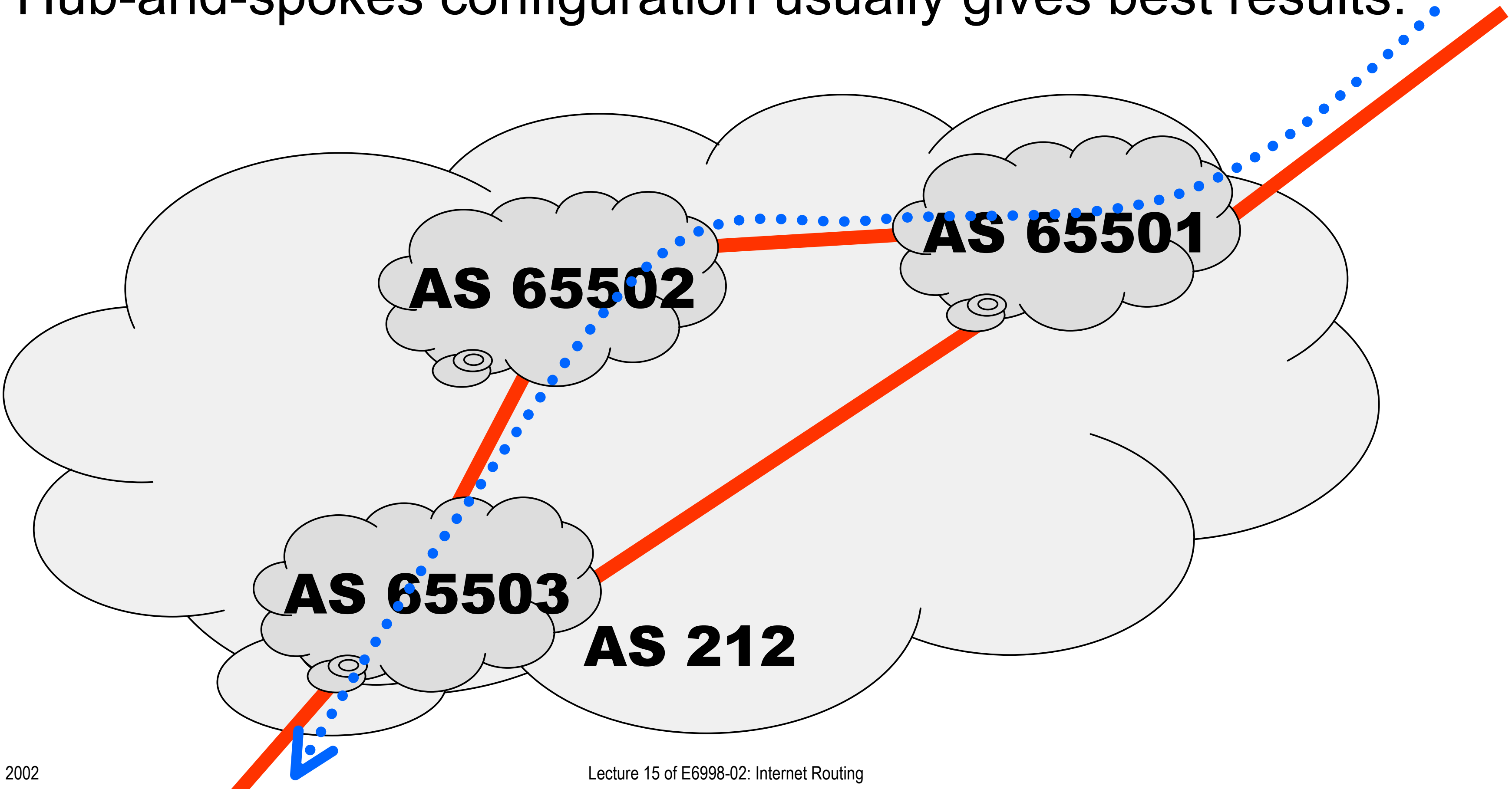


Confederations, cont'd

- Entire AS runs a single IGP.
  - Areas may or may not overlap with sub-ASes.
- Routers inside each sub-AS run normal I-BGP.
- BGP sessions between border routers of sub-ASes in the same confederation: EIBGP (what else!)
- Like E-BGP but with some changes.
  - LOCAL_PREF and MED are carried along.
  - NEXT_HOP is set by the first router, then carried along.
  - New AS_PATH segments:
    - AS_CONFED_SET (type 3).
    - AS_CONFED_SEQUENCE (type 4).
    - Stripped when going over a (real) EBGP session.
  - NO_EXPORT_SUBCONFED community.
- Route selection process is the same as with "regular" BGP.
  - Change: Prefer EBGP over EIBGP over IBGP.

Confederation Topology Considerations

- AS_PATH length stays constant (sub-AS components don't count).
  - Packets may take suboptimal path:
- Confederations should follow physical topology.
- Hub-and-spokes configuration usually gives best results.



RR *vs.* Confederations

- Experience varies.
- In RR, only the reflectors have to support the extension.
  - Not so in Confederations.
- Sub-ASs in a confederation can run individual IGPs.
- You can actually do RR inside a confederation.