E6998-02: Internet Routing

Lecture 17
Border Gateway Protocol, Part VI

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.

More BGP Extensions

- HELLO optional parameters:
  1. TCP MD5 Authentication (RFC2385).
  2. Capabilities negotiation (RFC2842).
     - TLVs indicating what optional capabilities the sender supports.
- If receiver does not support, closes connection with appropriate NOTIFICATION.

TCP MD5 Authentication

- TCP option type 19.
- 18 bytes long.
- 16 bytes of MD5 hash, including key, of TCP segment.

- Poor authentication.
- Should have used IPsec (of course).
- Does not make key management any easier.

Route Refresh Capability

- It's a request to the peer to send its Adj-RIB-Out.
- Used when the inbound policy of a peer changes.
  - All the routes that the peer had gotten (and potentially filtered or changed attributes thereof) have to be re-processed by the input policy engine.
- Alternative: close and reopen BGP session.
  - Causes lots of routes to flap.

- RFC 2918
- New BGP message (Type=5).

Outbound Route Filter Capability

- Request to the peer to send its inbound prefix filters.
- Rationale: why bother sending routes that will be filtered anyway?

- draft-ietf-idr-route-filter-06.txt

Graceful Restart Capability

- Indicates the ability to preserve BGP state across restarts.
- Minimizes distrubance.

- draft-ietf-idr-restart-05.txt

Dynamic Capability

- Capabilities are negotiated during OPEN.
- DC allows capabilities to be negotiated after OPEN.
- CAPABILITY message (Type=6)

- draft-ietf-idr-dynamic-cap-02.txt

Multiprotocol Extensions for BGP-4

- Negotiated capability.
- Extension to allow BGP-4 to carry routes for protocols other than unicast IPv4 (IPv6, multicast, *etc*.)

- Two new attributes:
  - MP_REACH_NLRI (Type=14)
    - Replaces NEXT_HOP attribute and NLRI field.
  - MP_UNREACH_NLRI (Type=15)
    - Replaces list of withdrawn routes.

- RFC2858 and draft-ietf-idr-rfc2858bis-02.txt

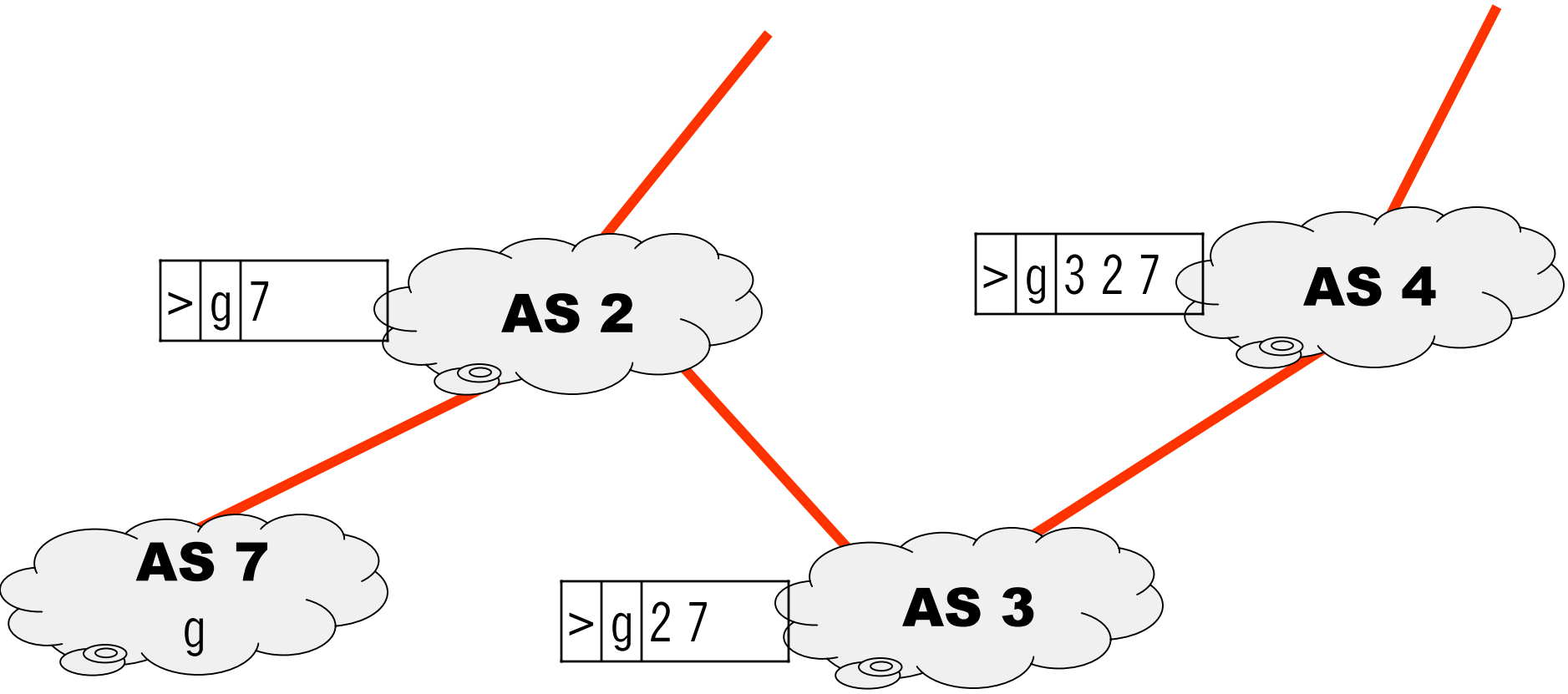
Dynamic Behavior of BGP

- The network is never in steady-state.
- Links break, routers crash, people make mistakes.
  - Routes get withdrawn.
  - New routes get advertised.
- How often do these happen?
- What is the effect on prefix reachability?
- Are they random or do they follow patterns?
- How disruptive are they?
- Can we/do we do anything to protect the network against them?

- Lots of recent and current research.

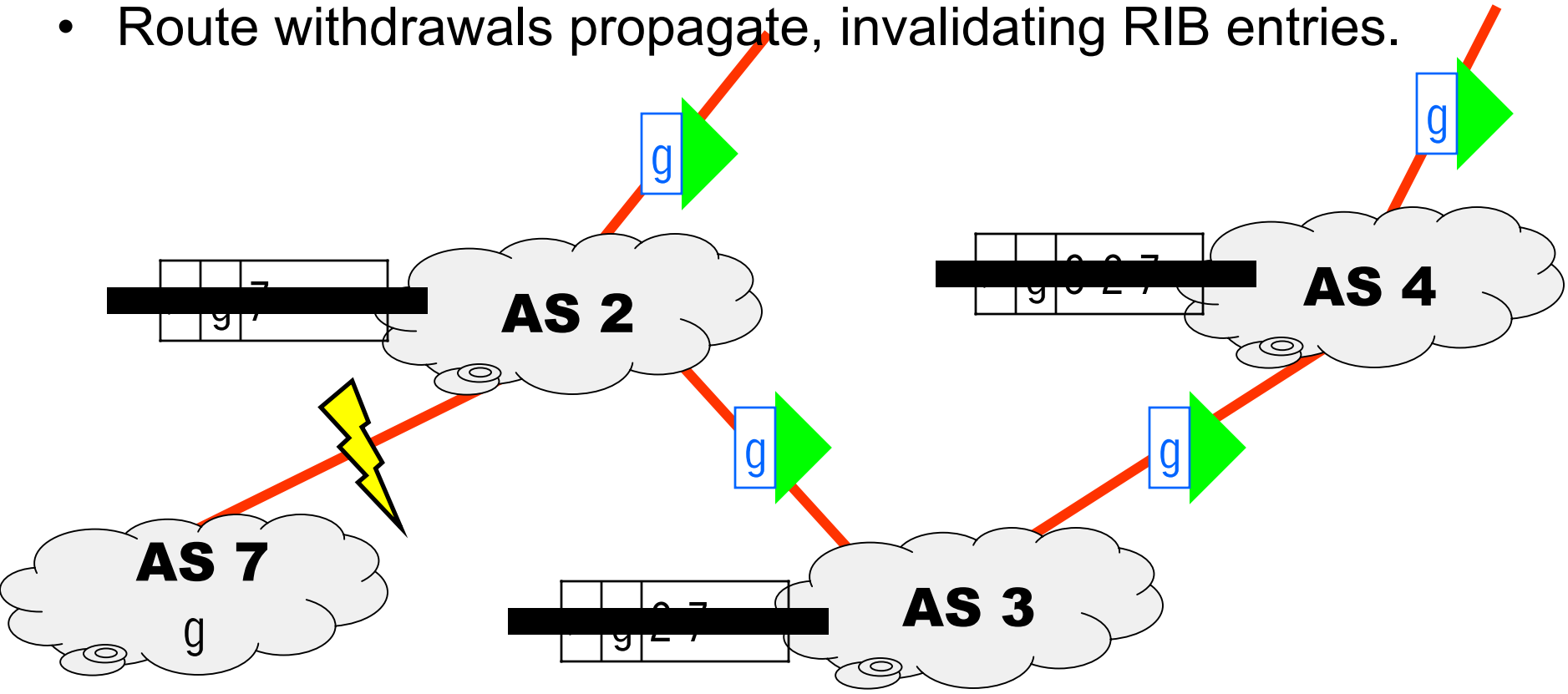
Link Failure (Single-homed system)

- AS 7 (prefix: g) is single-homed.



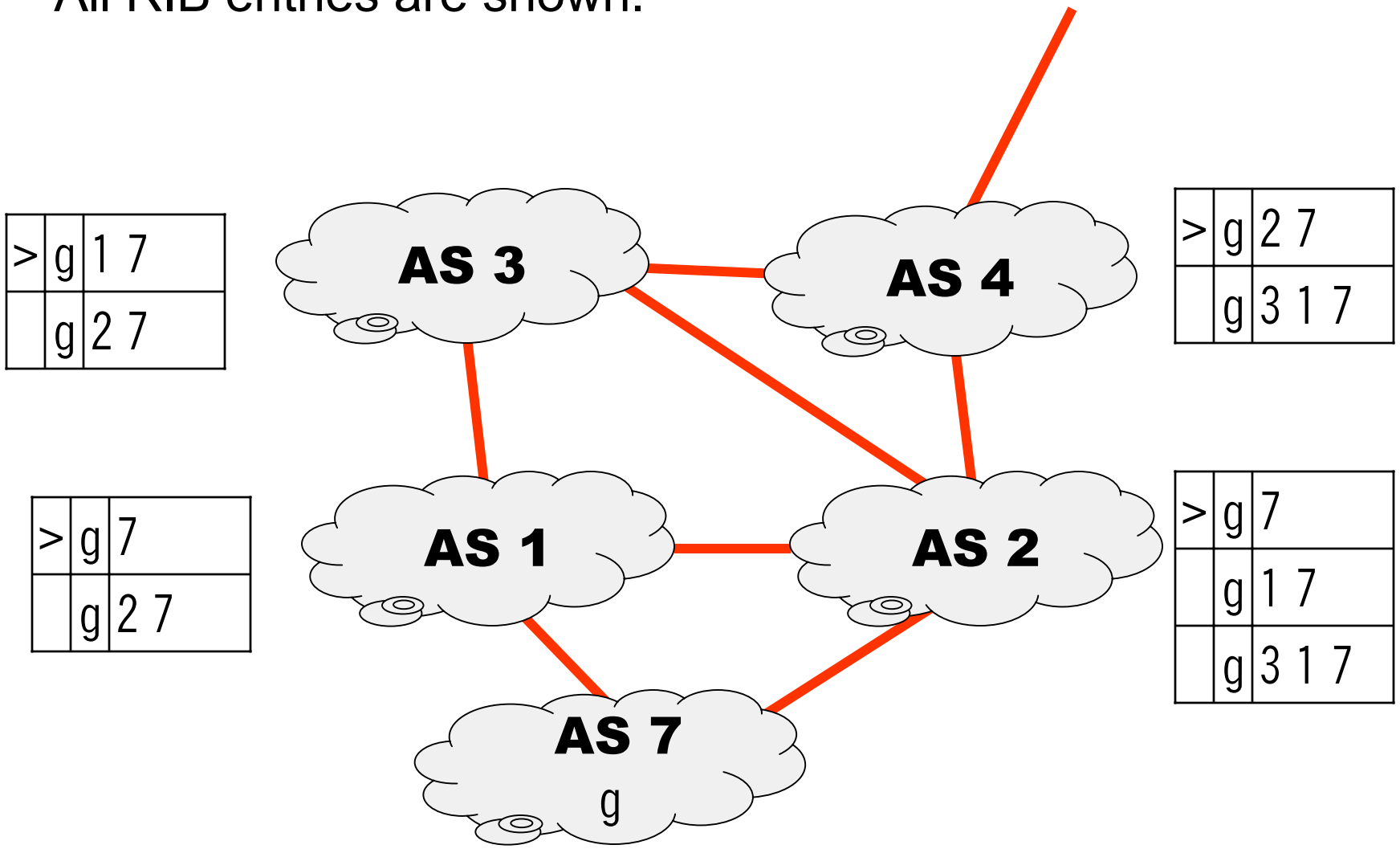
Link Failure (Single-homed system)

- Link between AS2 and AS7 fails.
- AS2 removes g from its RIB (both its Adj-RIB-1 and its Loc-RIB).
- AS2 withdraws route to g.
- Route withdrawals propagate, invalidating RIB entries.



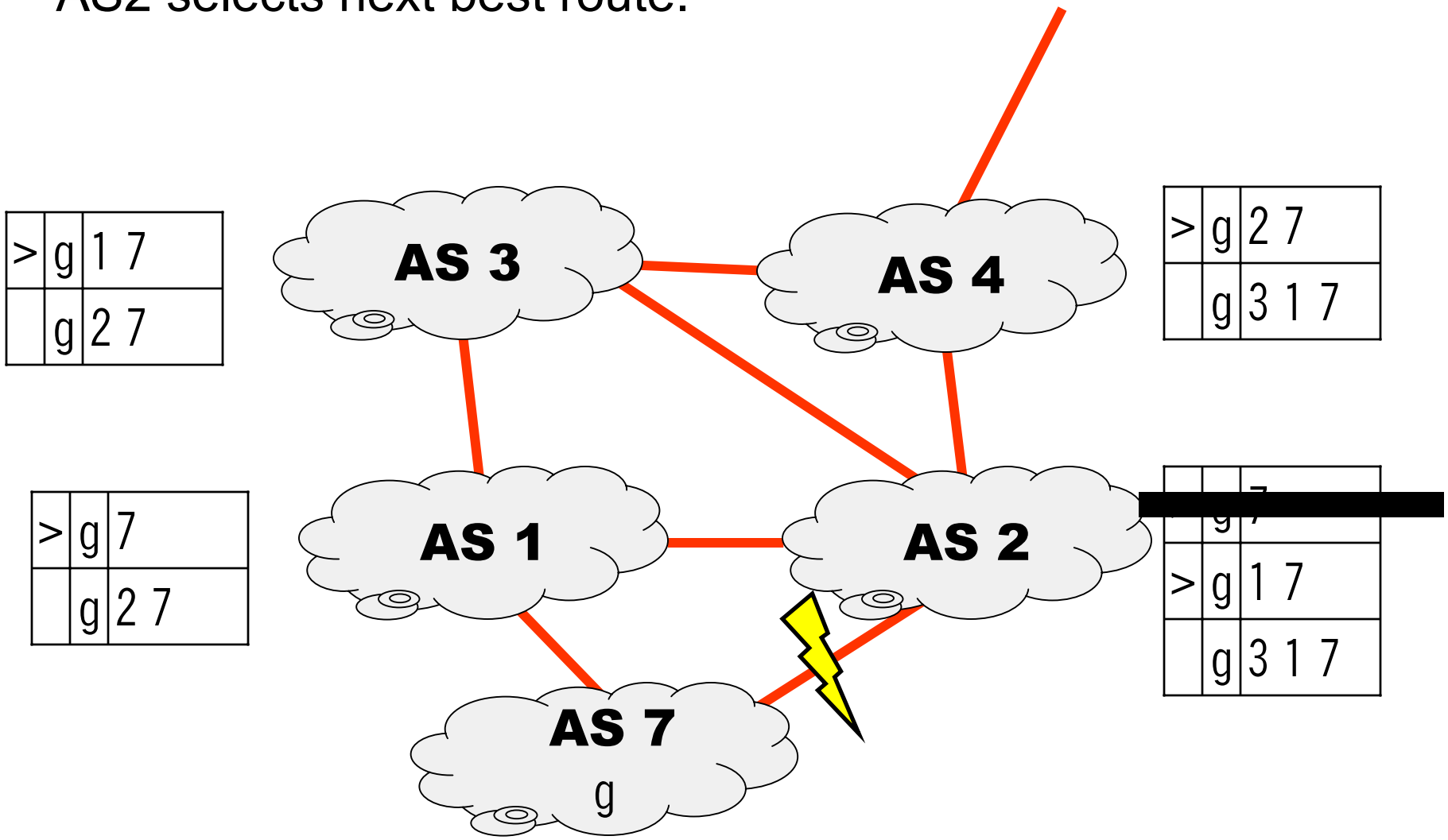
Link Failure (Multihomed system)

- AS 7 (prefix: g) is dual-homed.
- All RIB entries are shown.



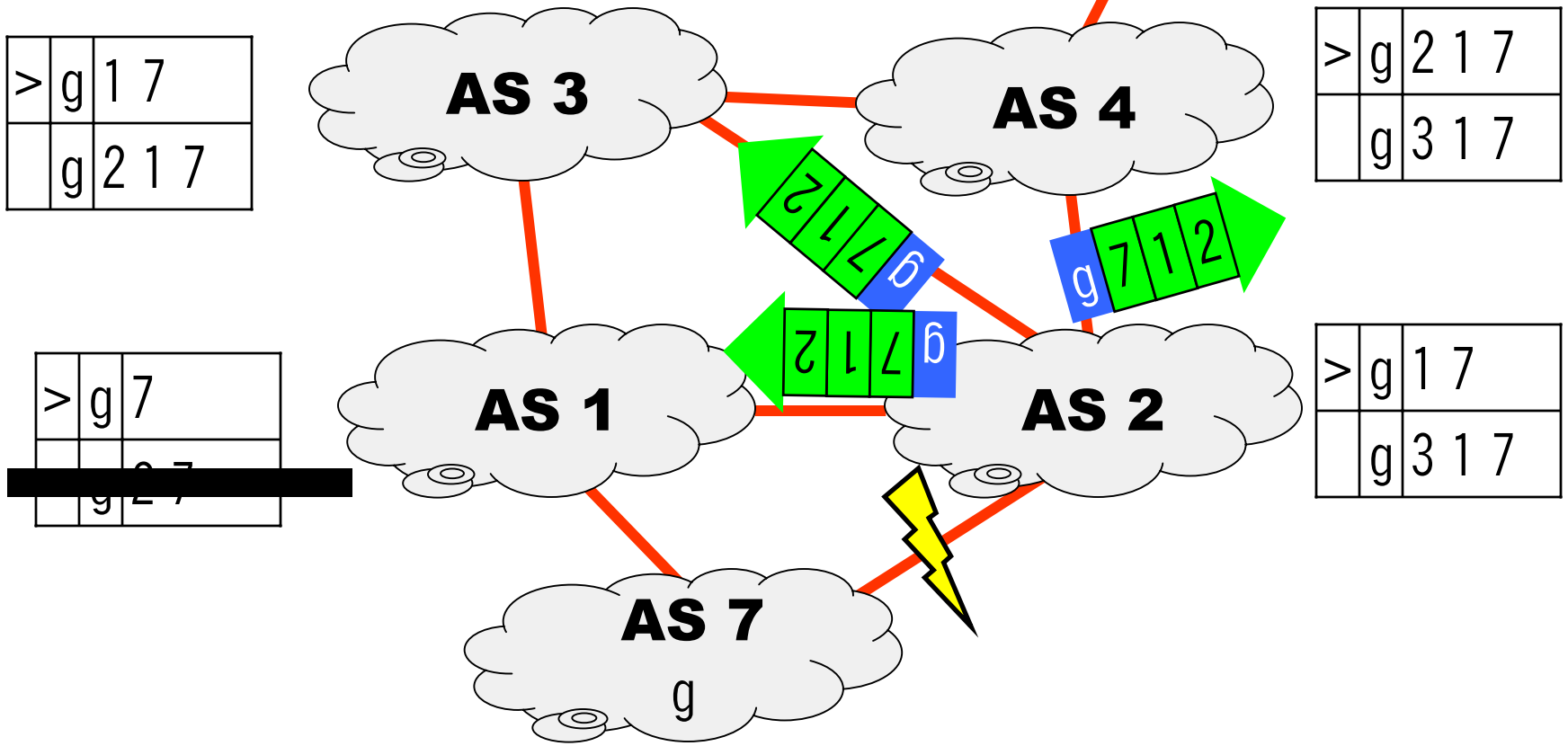
Link Failure (Multihomed system)

- Link 2-7 fails.
- AS2 selects next best route.



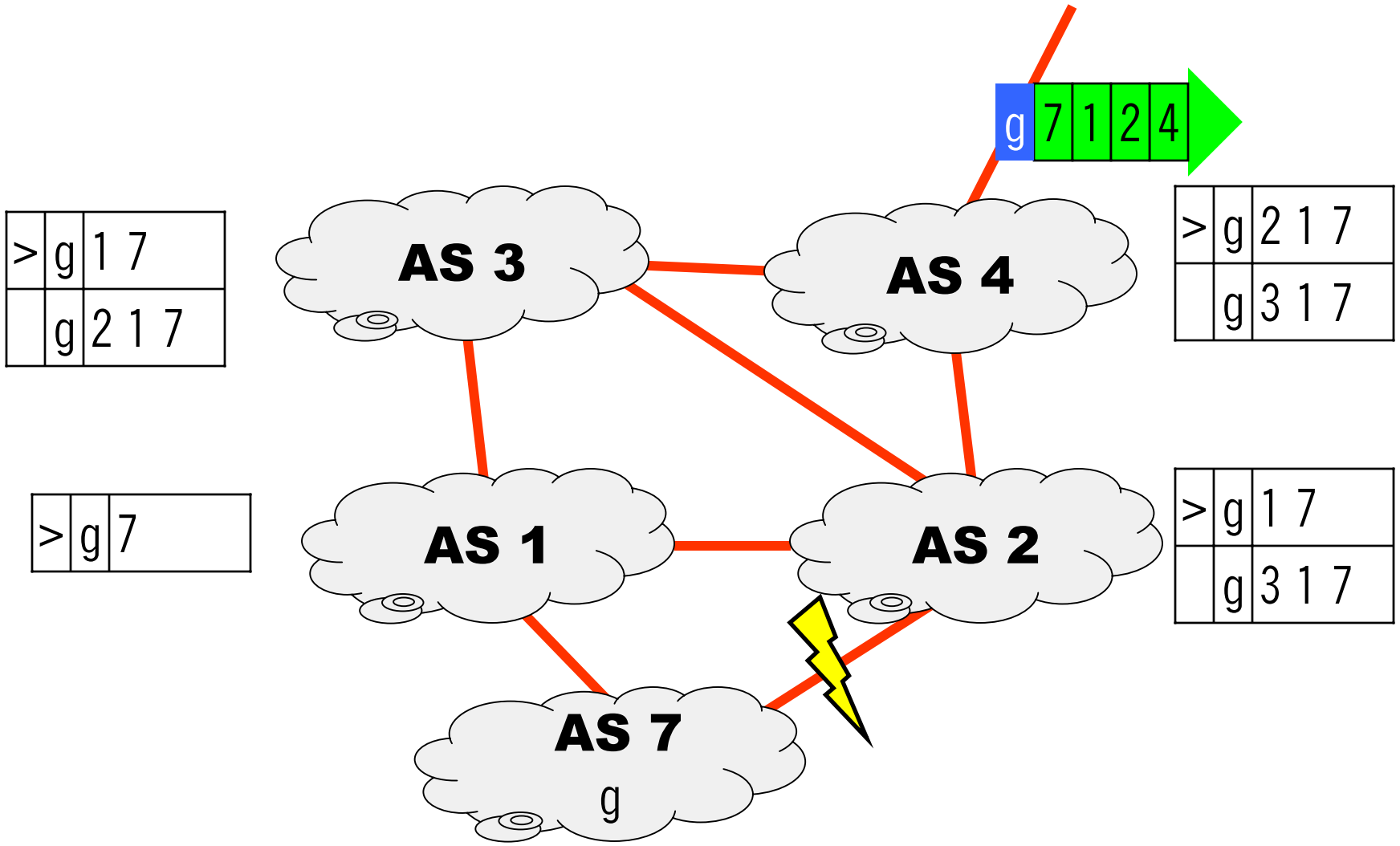
Link Failure (Multihomed system)

- New route is advertised.
- AS1 removes route to g (it's in the AS_PATH).
- AS3 puts replaces route from AS2, but prefers route via AS1 (shorter).
- AS4 has to choose between 7-1-2 and 7-1-3.



Link Failure (Multihomed system)

- AS4 sticks decides to stick with AS2 (higher LOCAL_PREF).
- Has to advertise new route, since AS_PATH changed.



Route Flapping

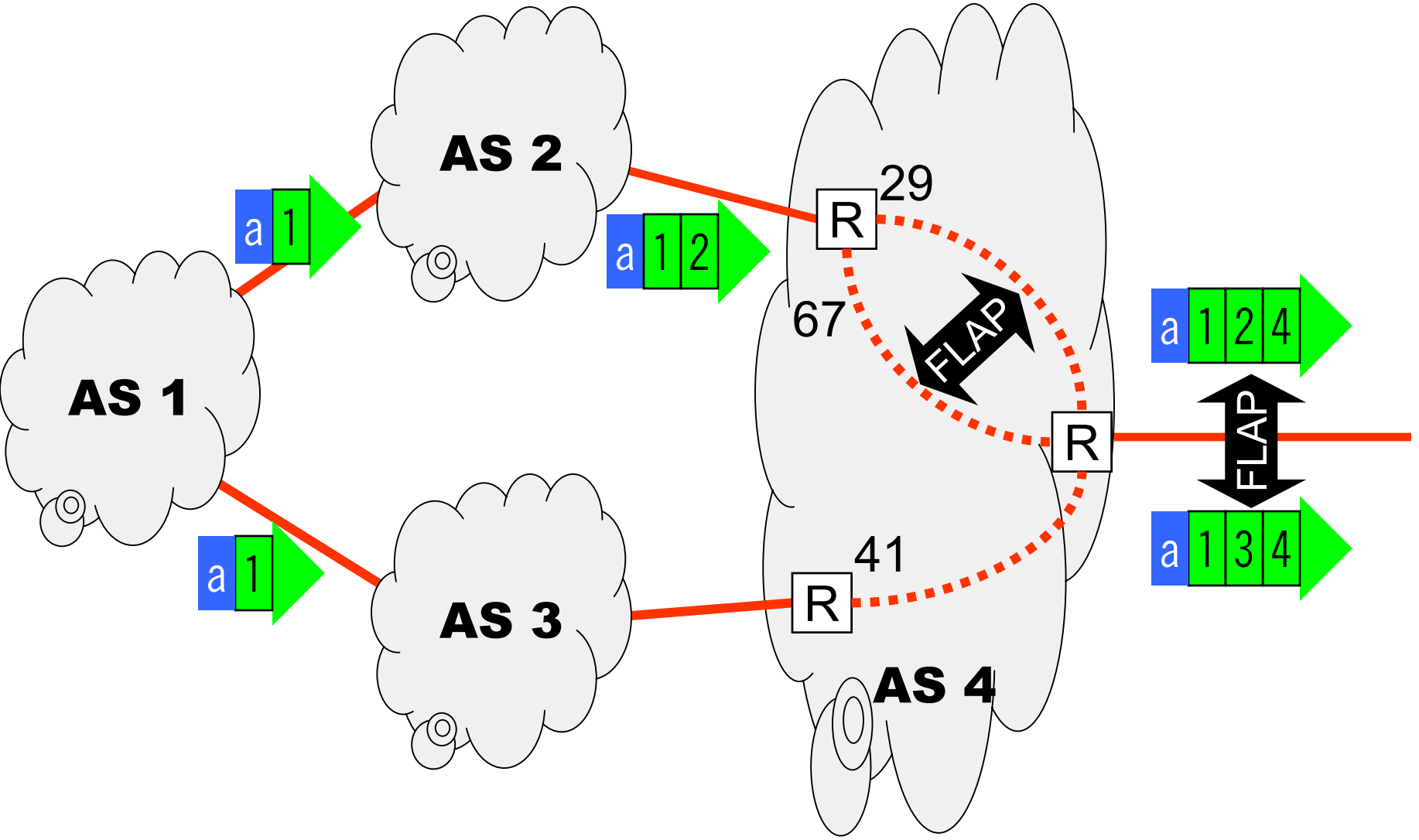
- Routing instability.
- Route disappears, appears again, disappears again…
  - Withdrawal, announcement, withdrawal, announcement…
- Visible to the entire Internet.
  - Wastes resources, triggers more instability.
- Some causes of *Route Flapping*:
  - Flaky inter-AS links.
  - Flaky or insufficient hardware.
  - Link congestion.
  - IGP instability.
  - Operator error.

Link Instability

- The first three are examples of link instability.
  - Link itself fails.
  - Router/router interface fails.
  - Messages can't get through.
- When a link goes down, routers withdraw routes associated with this link.
  - Customer-ISP.
  - ISP-ISP.
- Announcements travel throughout the default-free zone.
- Aggregation may mask downstream flapping.
  - Does not work for multihoming

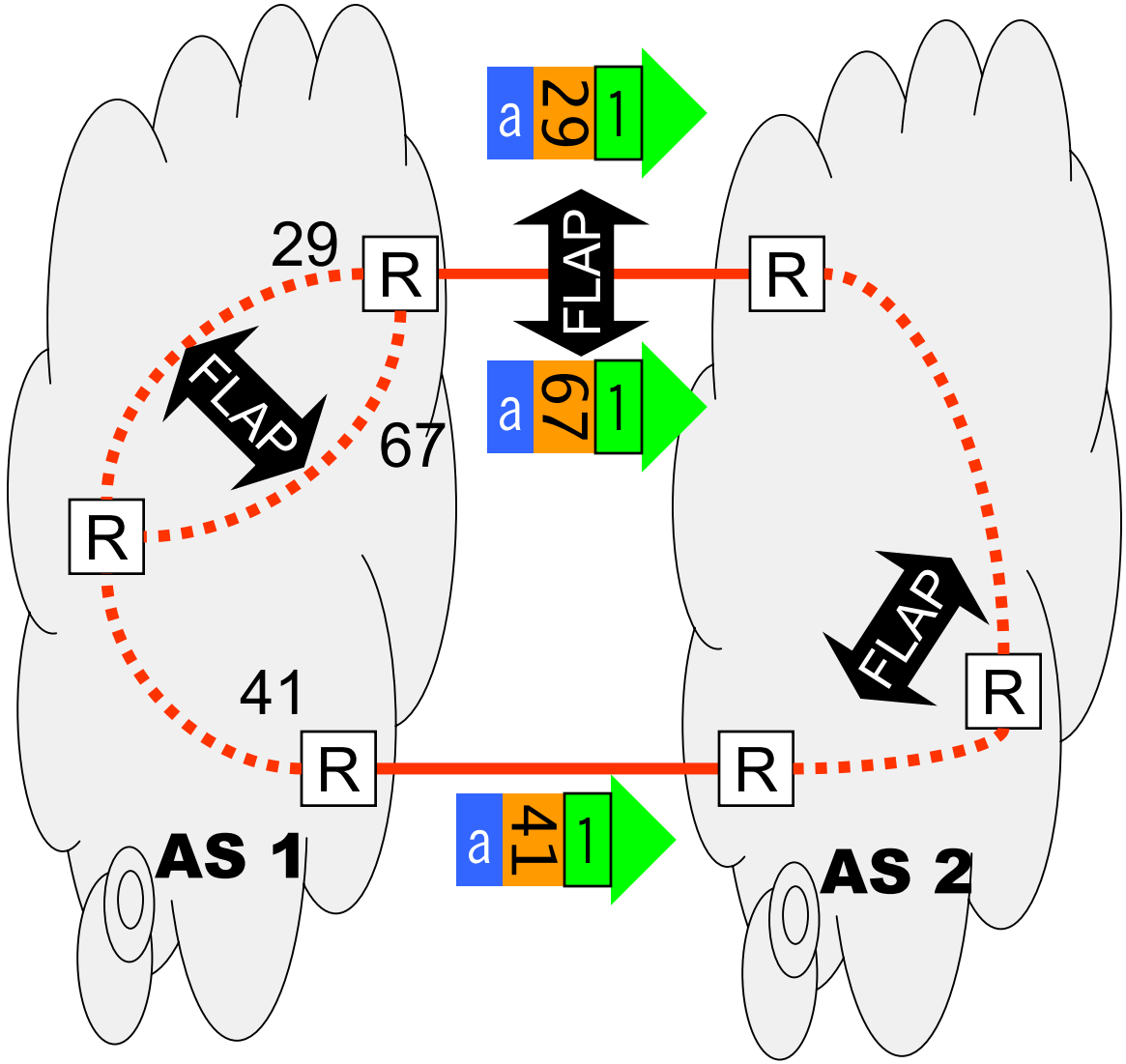
IGP Instability

- IGP route-preference rule exports instability.



IGP Instability

- MEDs can export internal instability.



Route Flap Dampening

- RFC2439
- Router detects route flapping.
- *Penalty*:
  - Increased each time a route flaps.
  - Decreased over time.
- If penalty threshold exceeded (*suppress limit*), route is suppressed.
- Until penalty drops below a certain level (*reuse limit*).

- There is evidence that it may be harmful.
  - BGP explores alternate paths when a route is withdrawn.
  - Dampening merely makes the exploration run in slow motion.
  - Too aggressive.

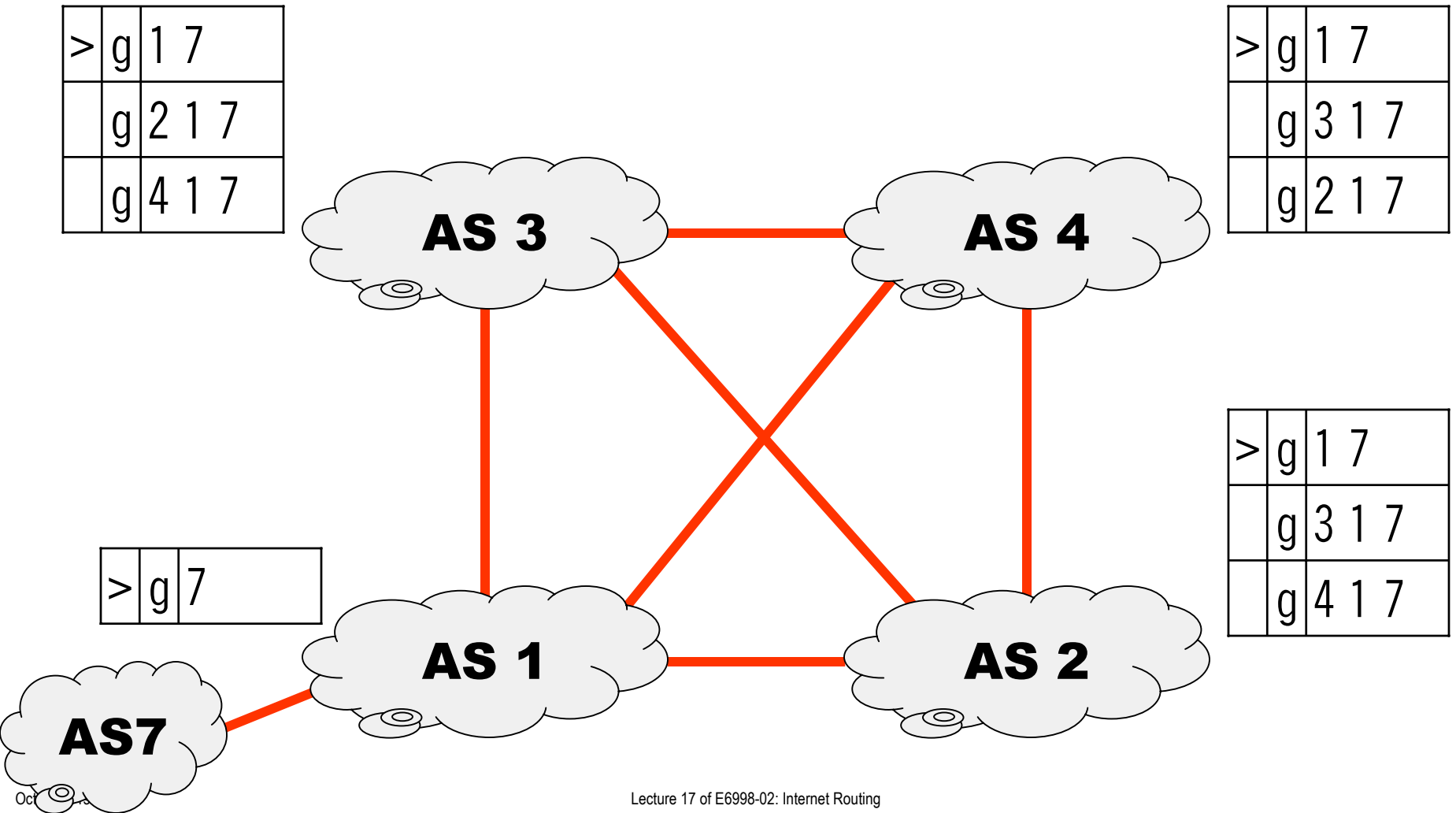
Convergence

- Link-State algorithms avoid loops by running the same computation (Dijkstra SPF) on the same data.
- Distance-Vector (Bellman-Ford-like) algorithms (e.g., RIP) avoid loops by selecting routes with a lower metric.
- Path-Vector algorithms (e.g., BGP) avoid loops by detecting self in path.

- LS converges as soon as new LSAs flooded.
- DV counts to infinity.
  - Split horizon/poison reverse/triggered updates just make the counting-to-infinity faster.
- How about BGP?

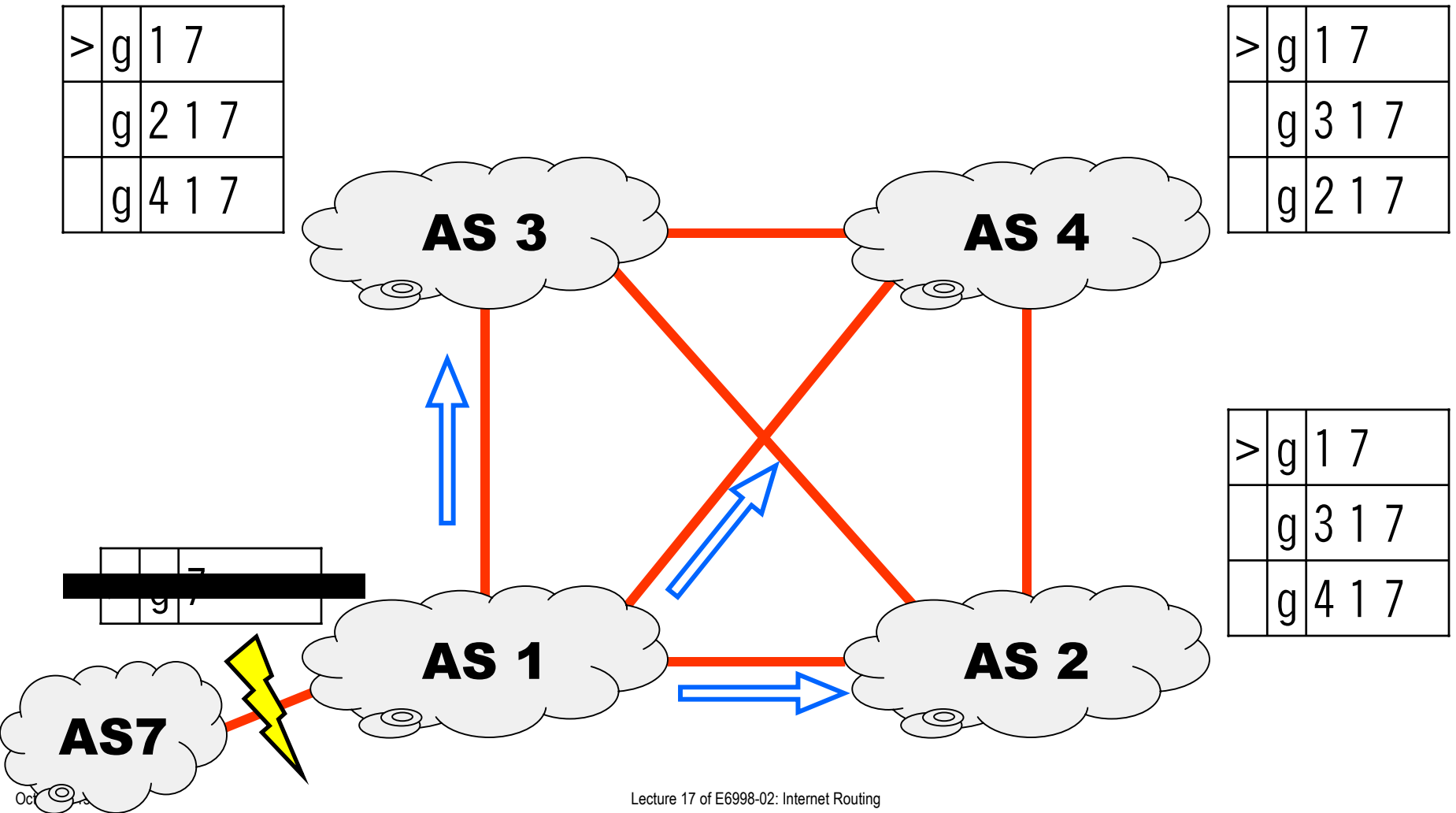
BGP Explores All Paths!

- See Labovitz *et al.*, SIGCOMM 2000.

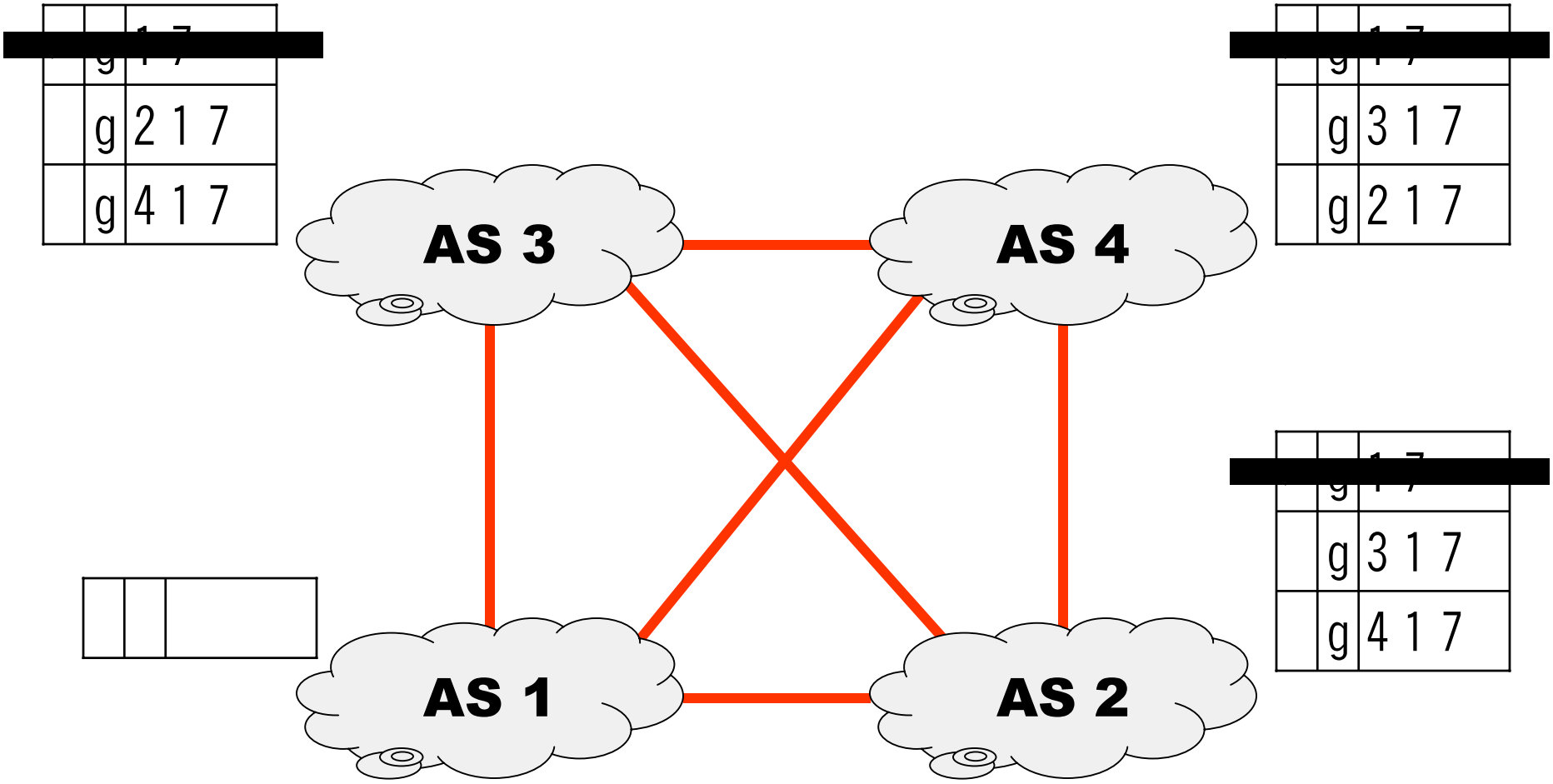


BGP Explores All Paths!

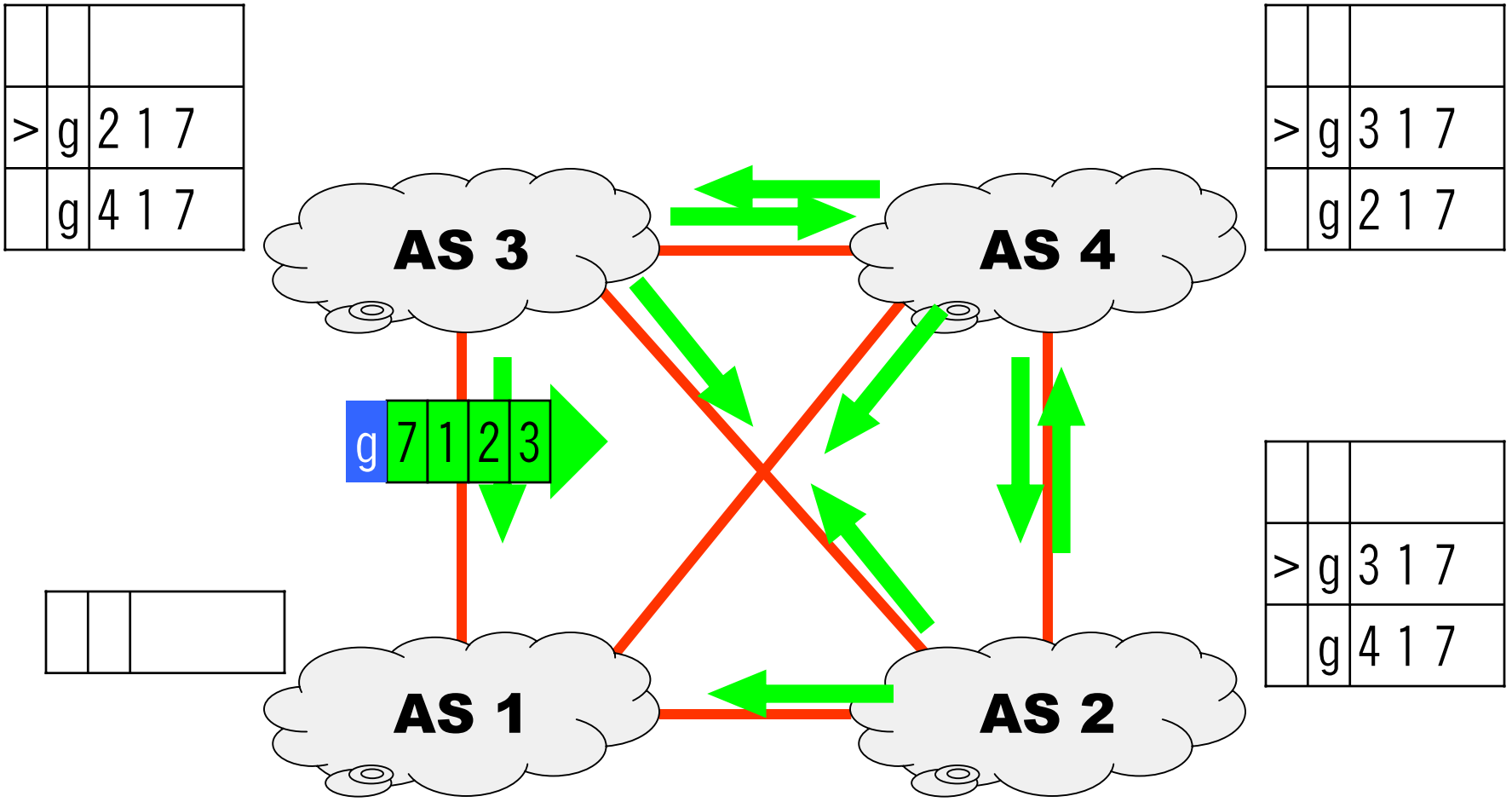
- Link 7-1 goes down.
- AS1 withdraws the route to prefix g.



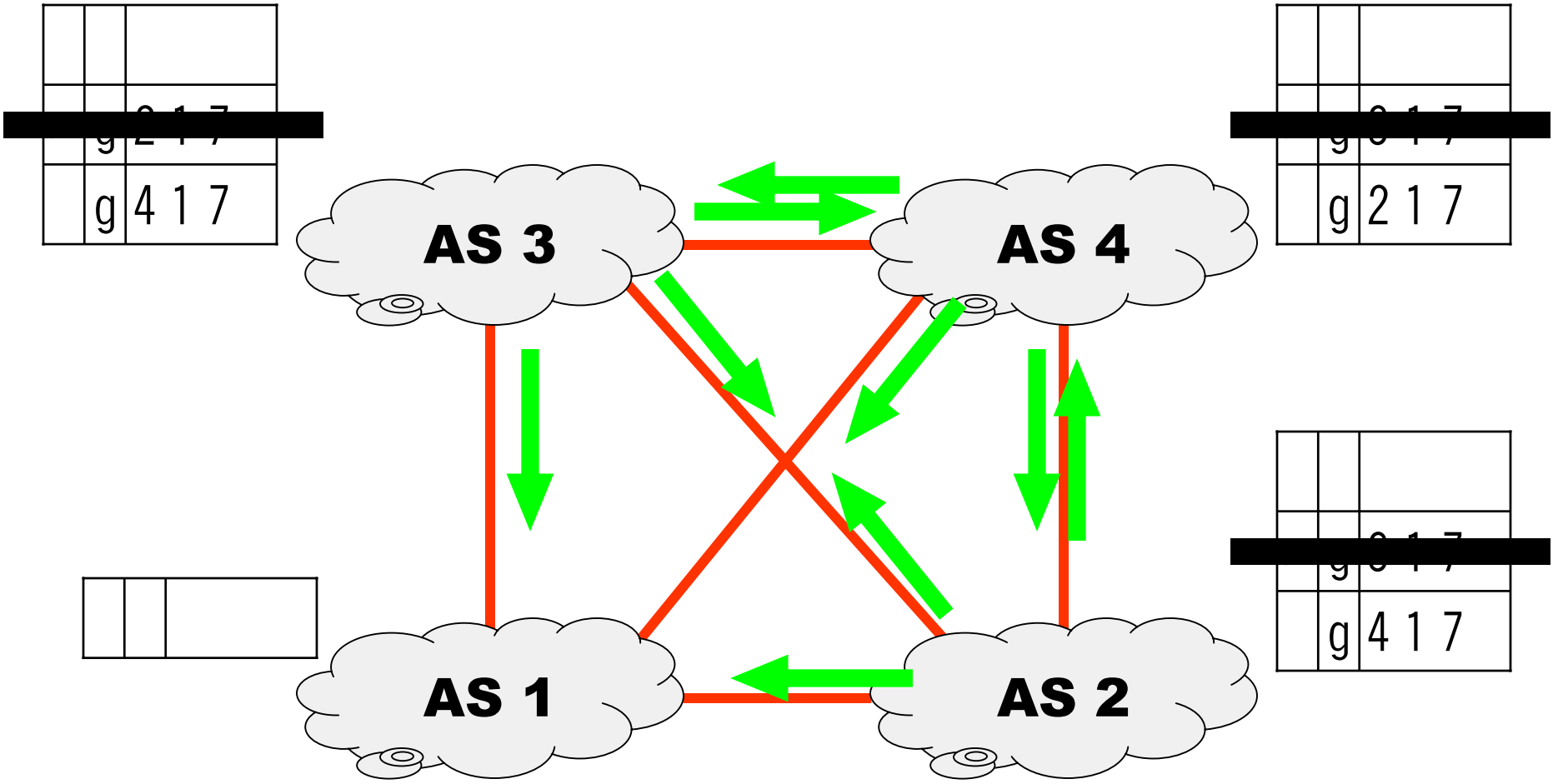
- AS 2, 3, 4 remove [1 7] route.



- Select their next best route.
- Advertise it.



- AS1 ignores the routes it gets (self in AS_PATH).
- (e.g.) AS2 gets [3 2 1 7] from AS3; treats it as implicit withdrawal of [3 1 7], then rejects it (self in AS_PATH).
- Process repeats one more time, then all ASes lose their routes to g.



BGP Explores n! Paths (cont'd)

- Problem was exacerbated by MinRouteAdvertisementInterval.
- Routers would wait 30 seconds before sending next set of updates.
- Common perception at the time was "BGP converges within 30 seconds".
- There were paths that took over 15 minutes to converge.

- This sort of behavior creates routing traffic without always benefiting connectivity.

- Lots of other sources of instability.

BGP Conclusion

- Protocol (deceptively) simple.
- Lots of accumulated current practices.
- It mostly works.
- But for how much longer?