Routing Convergence

Jennifer Rexford

Fall 2010 (TTh 1:30-2:50 in COS 302)

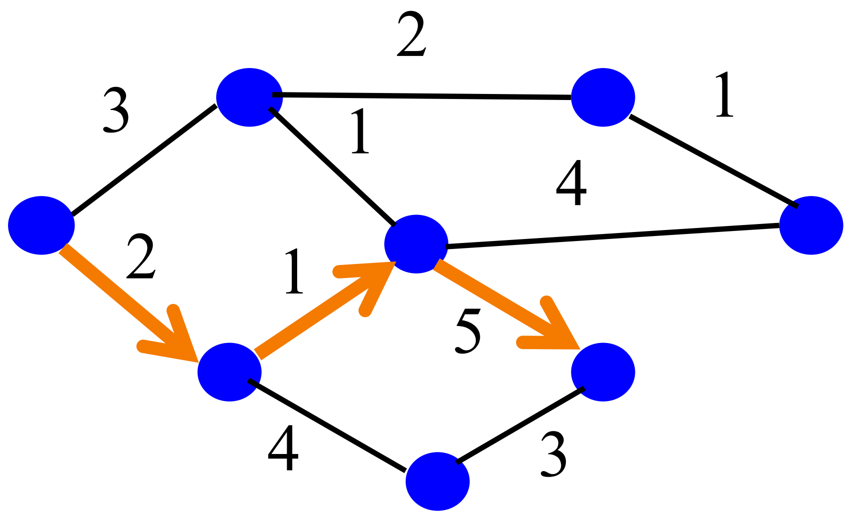
COS 561: Advanced Computer Networks

http://www.cs.princeton.edu/courses/archive/fall10/cos561/

Intradomain Routing

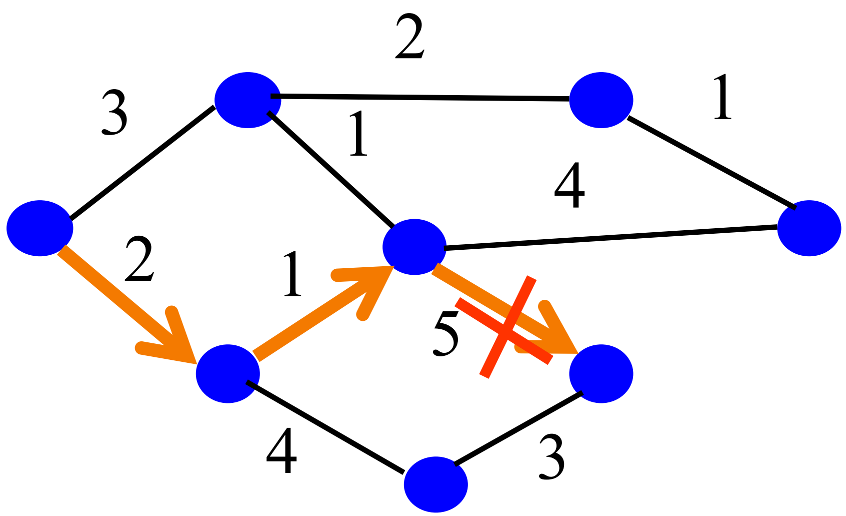
Convergence

- Getting consistent routing information to all nodes
  - E.g., all nodes having the same link-state database
- Consistent forwarding after convergence
  - All nodes have the same link-state database
  - All nodes forward packets on shortest paths
  - The next router on the path forwards to the next hop



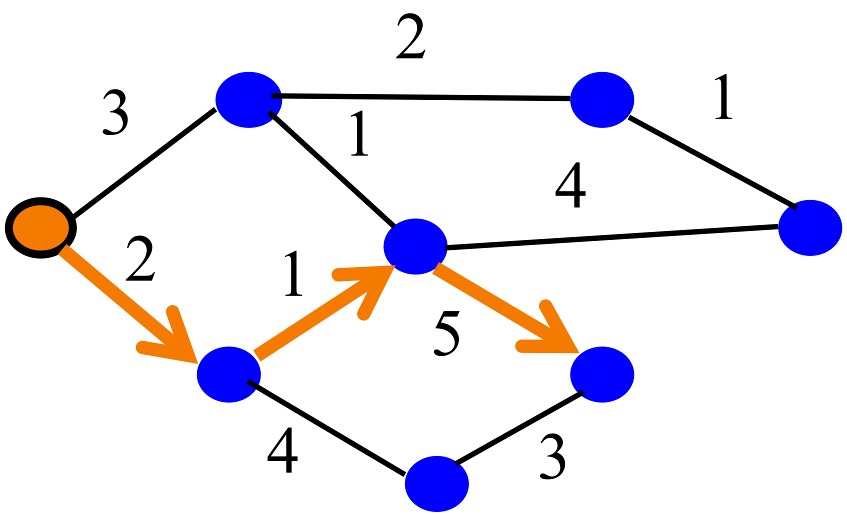
Transient Disruptions

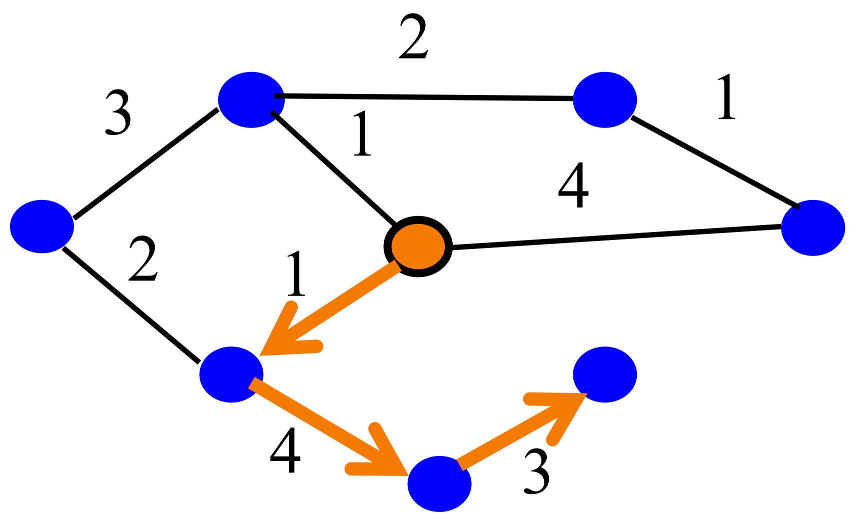
- Detection delay
  - A node does not detect a failed link immediately
  - … and forwards data packets into a “blackhole”
- Depends on timeout for detecting lost hellos
  - Or link media that can detect “loss of light”



Transient Disruptions

- Inconsistent link-state database
  - Some routers know about failure before others
  - The shortest paths are no longer consistent
  - Can cause transient forwarding loops





Convergence Delay

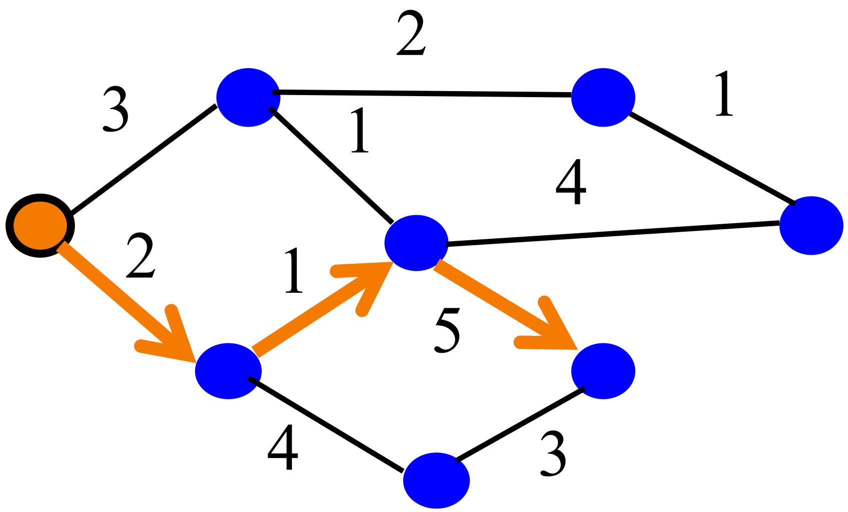
- Sources of convergence delay
  - Detection latency
  - Flooding of link-state information
  - Shortest-path computation
  - Creating the forwarding table
- Performance during convergence period
  - Lost packets due to blackholes and TTL expiry
  - Looping packets consuming resources
  - Out-of-order packets reaching the destination
- Very bad for VoIP, online gaming, and video

Reducing Convergence Delay

- Faster detection
  - Smaller hello timers
  - Link-layer technologies that can detect failures
- Faster flooding
  - Flooding immediately
  - Sending link-state packets with high-priority
- Faster computation
  - Faster processors on the routers
  - Incremental Dijkstra's algorithm
- Faster forwarding-table update
  - Data structures supporting incremental updates

Reducing Convergence Delay

- Weight tuning for planned maintenance
  - Gradually increase link weight
  - Before taking down the link
- MPLS fast-reroute
  - Backup paths to use when the primary path fails
  - Local protection to circumvent a failed link



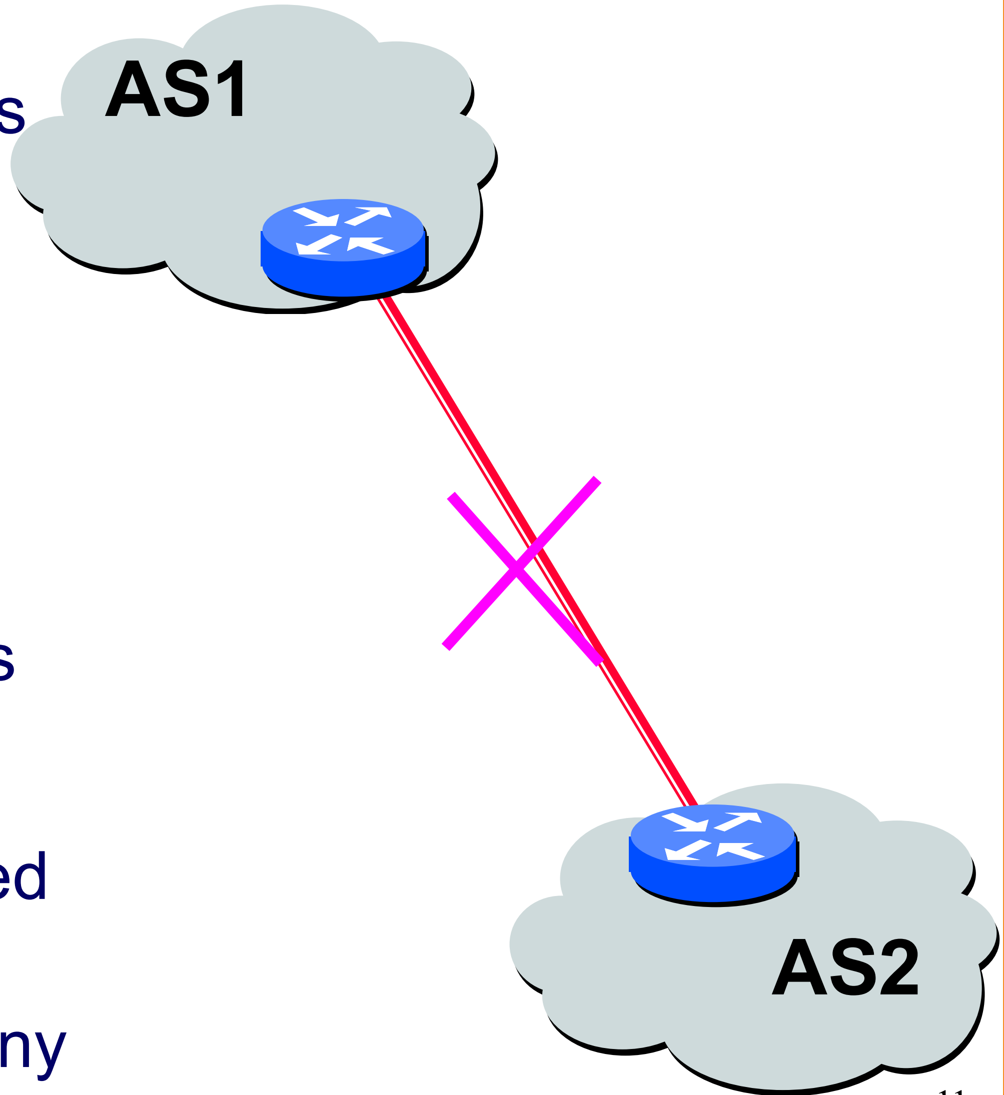
Interdomain Routing

Causes of BGP Routing Changes

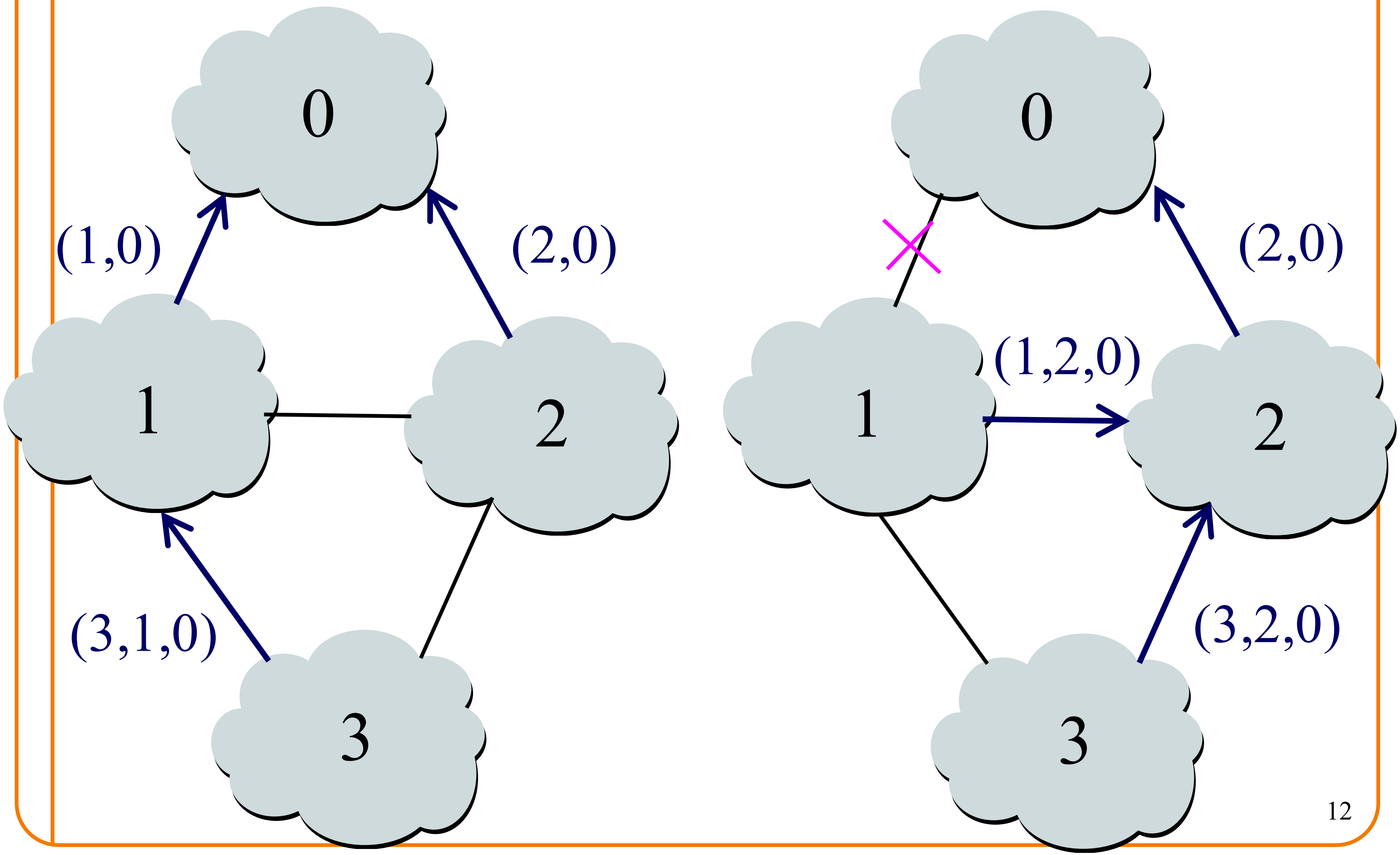
- Topology changes
  - Equipment going up or down
  - Deployment of new routers or sessions
- BGP session failures
  - Due to equipment failures, maintenance, etc.
  - Or, due to congestion on the physical path
- Changes in routing policy
  - Changes in preferences in the routes
  - Changes in whether the route is exported
- Persistent protocol oscillation
  - Conflicts between policies in different ASes

BGP Session Failure

- BGP runs over TCP
  - BGP only sends updates when changes occur
  - TCP doesn't detect lost connectivity on its own
- Detecting a failure
  - Keep-alive: 60 seconds
  - Hold timer: 180 seconds
- Reacting to a failure
  - Discard all routes learned from the neighbor
  - Send new updates for any routes that change

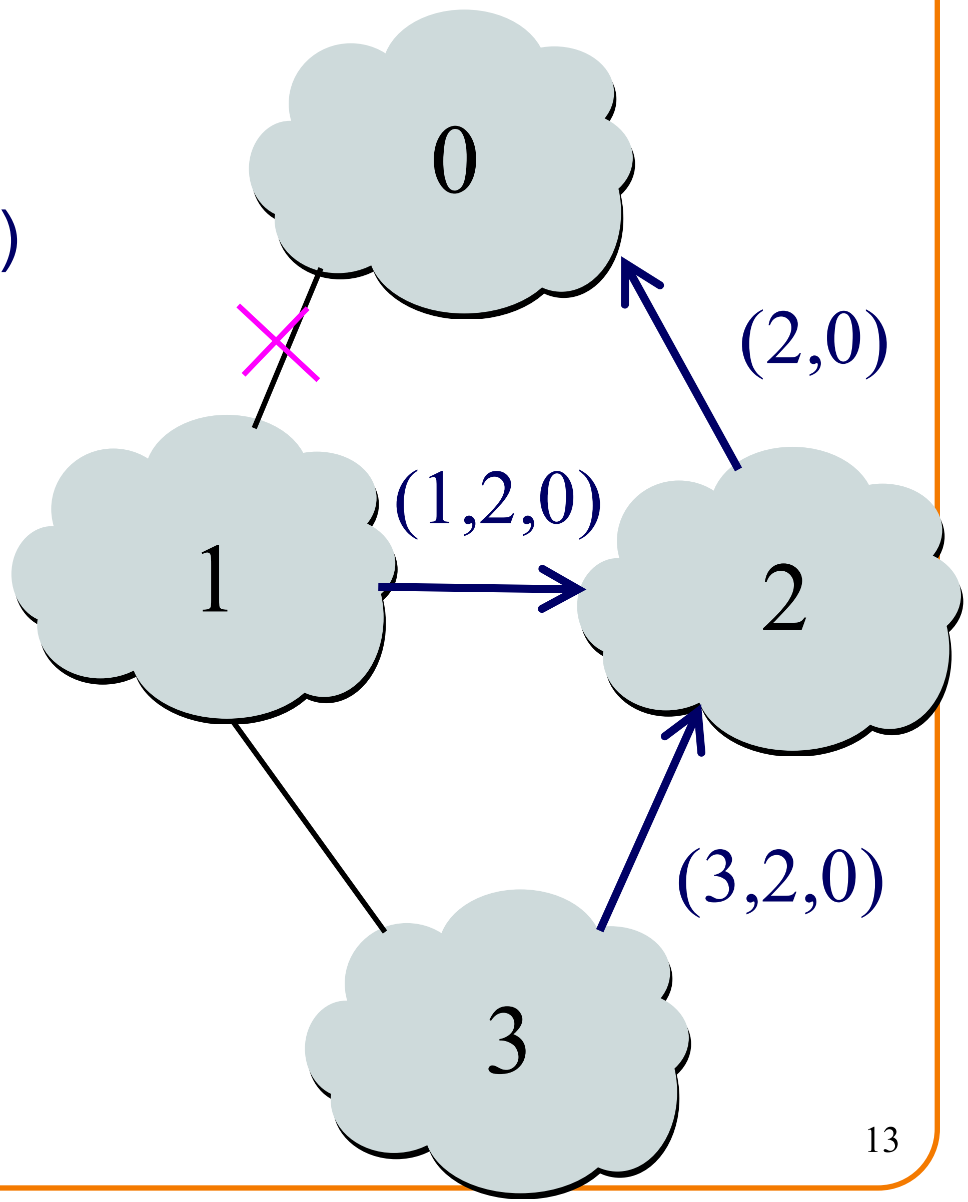


Routing Change: Before and After



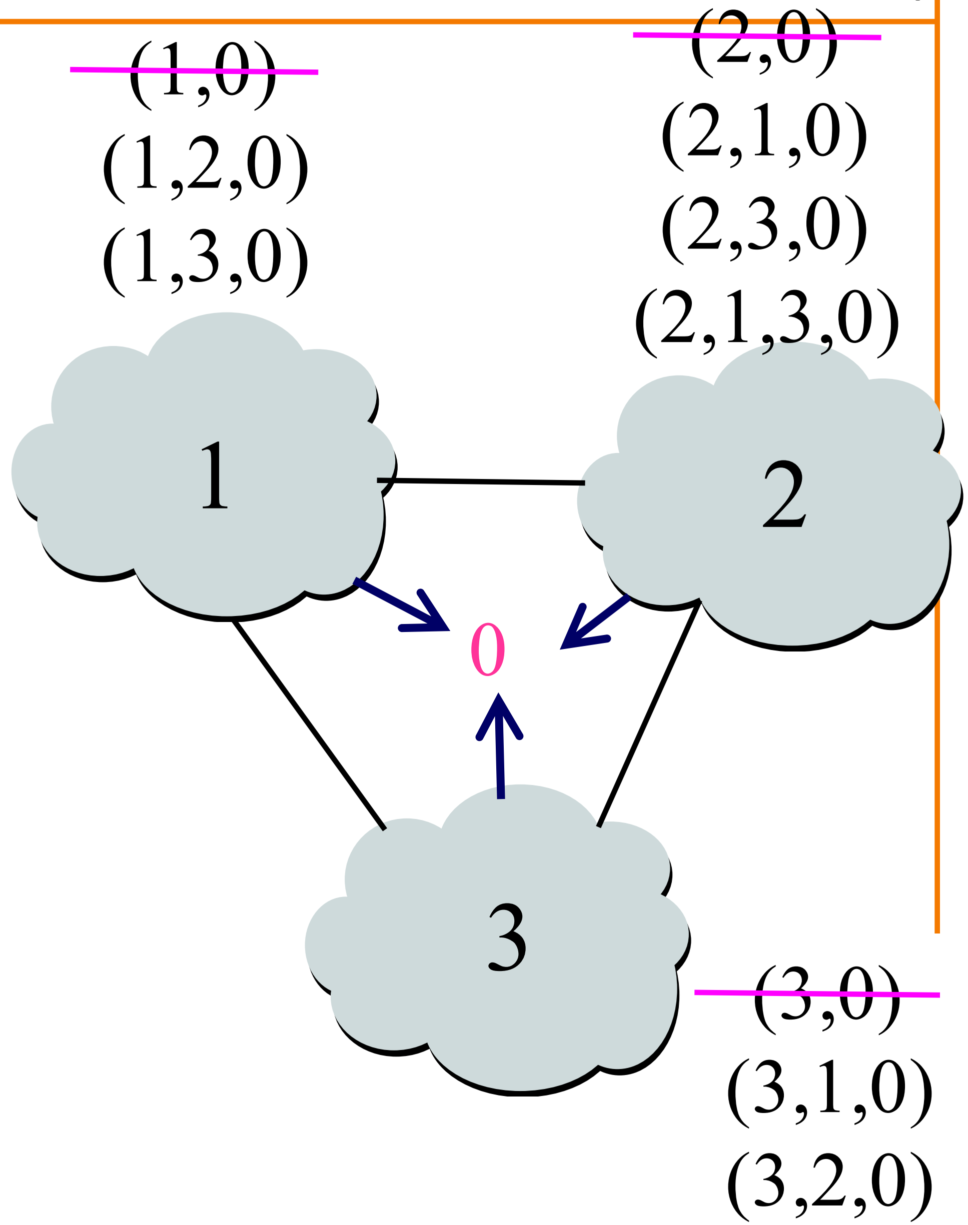
Routing Change: Path Exploration

- AS 1
  - Delete the route (1,0)
  - Switch to next route (1,2,0)
  - Send route (1,2,0) to AS 3
- AS 3
  - Sees (1,2,0) replace (1,0)
  - Compares to route (2,0)
  - Switches to using AS 2



Routing Change: Path Exploration

- Initial situation
  - Destination 0 is alive
  - All ASes use direct path
- When destination dies
  - All ASes lose direct path
  - All switch to longer paths
  - Eventually withdrawn
- E.g., AS 2
  - (2,0) → (2,1,0)
  - (2,1,0) → (2,3,0)
  - (2,3,0) → (2,1,3,0)
  - (2,1,3,0) → null



BGP Converges Slowly

- Path vector avoids count-to-infinity
  - But, ASes still must explore many alternate paths
  - … to find the highest-ranked path that is still available
- Fortunately, in practice
  - Most popular destinations have very stable BGP routes
  - And most instability lies in a few unpopular destinations
- Still, lower BGP convergence delay is a goal
  - Can be tens of seconds to a few minutes
  - High for important interactive applications
  - … or even conventional application, like Web browsing

Beyond Faster Convergence

Research Ideas

- Modeling of routing convergence
  - Bounds for BGP as function of topology and policy
  - Impact on timer configurations on convergence
- Much smaller timers for faster convergence
  - Understand trade-off between convergence time and protocol overhead
- Distributed coordination of routing changes
  - Avoid loops and blackholes during convergence
- Failure carrying packets
  - Faster detection by piggybacking on data packets

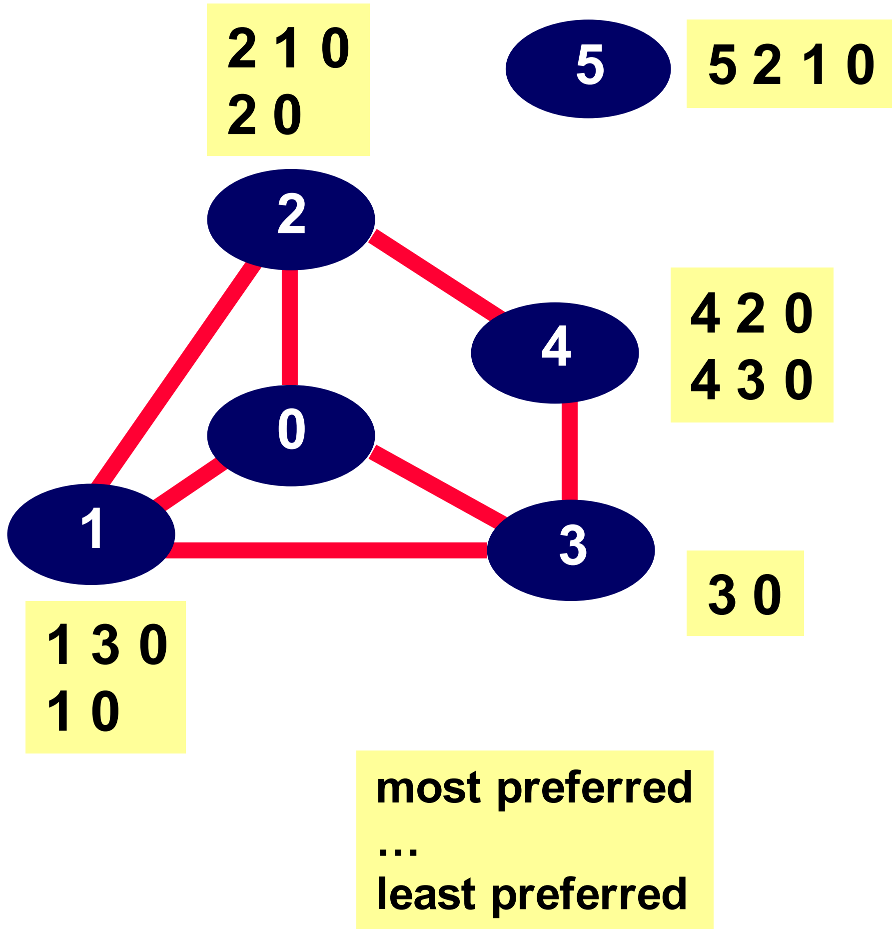
Research Ideas

- Temporary backup routes in BGP
  - Have an alternate path through another AS
- Multipath routing to survive failures
  - Multiple paths, possibly computed in advance
  - Load balancing over the currently-working paths
- Error-correction codes on multiple paths
  - Spreading redundant traffic over multiple paths
  - Reconstructing the traffic at the receiver

BGP Instability

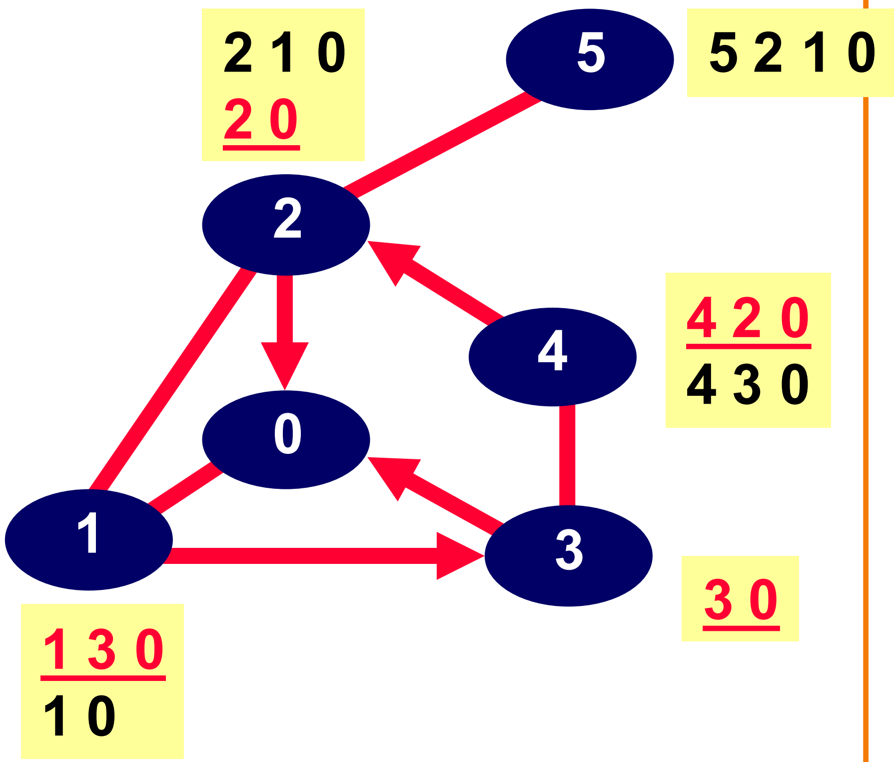
Stable Paths Problem (SPP) Instance

- Node
  - BGP-speaking router
  - Node 0 is destination
- Edge
  - BGP adjacency
- Permitted paths
  - Set of routes to 0 at each node
  - Ranking of the paths



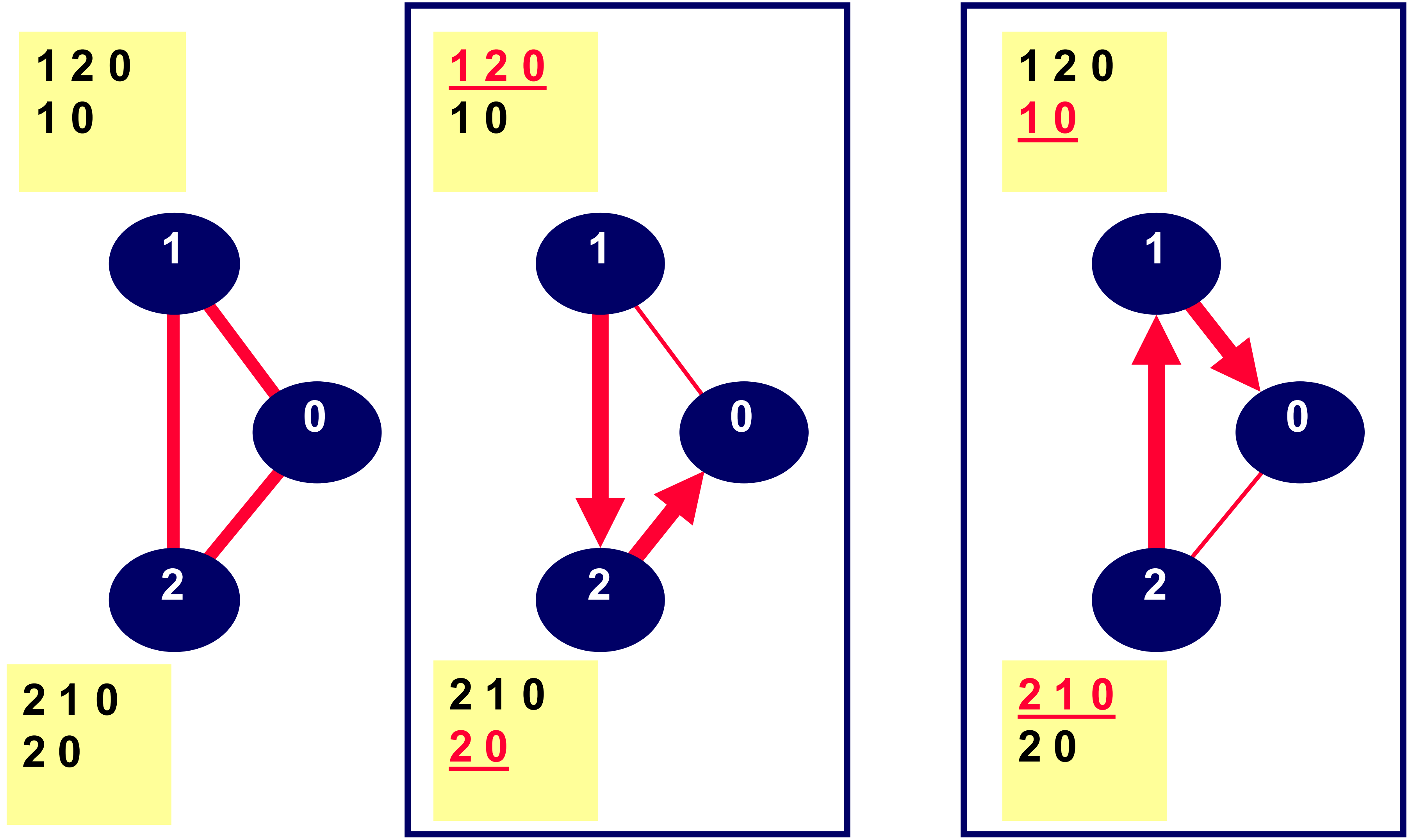
A Solution to a Stable Paths Problem

- Solution
  - Path assignment per node
  - Can be the “null” path
- If node u has path uwP
  - {u,w} is an edge in the graph
  - Node w is assigned path wP
- Each node is assigned
  - The highest ranked path consistent with assignment of its neighbors



A solution need not represent a shortest path tree, or a spanning tree.

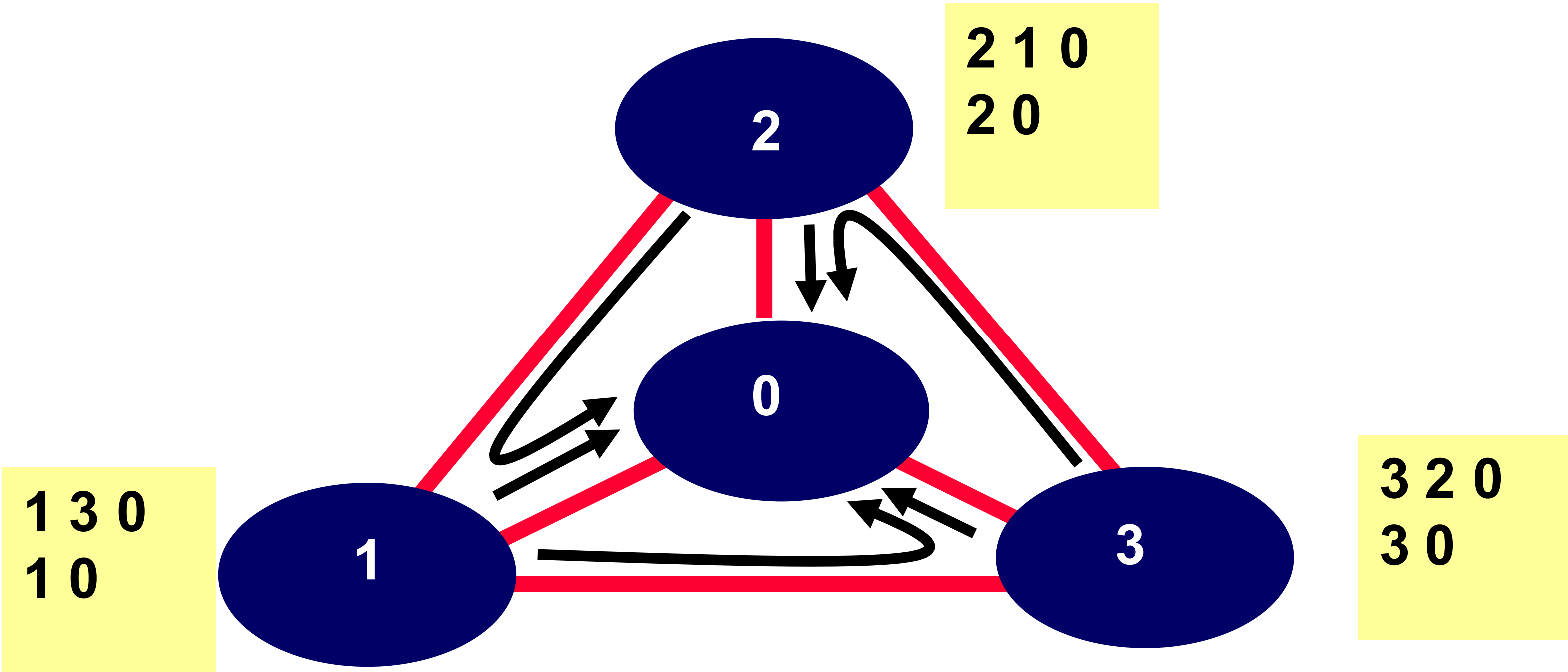
An SPP May Have Multiple Solutions



First solution

Second solution

An SPP May Have No Solution



Avoiding BGP Instability

- Detecting conflicting policies
  - Computationally expensive
  - Requires too much cooperation
- Detecting oscillations
  - Observing the repetitive BGP routing messages
- Restricted routing policies and topologies
  - Policies based on business relationships
    - Prefer paths through customers
    - Don't provide transit service to peers and providers
  - No cycles of provider-customer relationship
- Getting rid of BGP ☺