PATH VECTOR ROUTING AND THE BORDER GATEWAY PROTOCOL

READING: SECTIONS 4.3.3 PLUS OPTIONAL READING

COS 461: Computer Networks
Spring 2010 (MW 3:00-4:20 in COS 105)

Mike Freedman
http://www.cs.princeton.edu/courses/archive/spring10/cos461/

Goals of Today's Lecture

- Path-vector routing
  - Faster convergence than distance vector
  - More flexibility in selecting paths
- Interdomain routing
  - Autonomous Systems (AS)
  - Border Gateway Protocol (BGP)
- BGP convergence
  - Causes of BGP routing changes
  - Path exploration during convergence

Interdomain Routing and Autonomous Systems (ASes)

Interdomain Routing

- Internet is divided into Autonomous Systems
  - Distinct regions of administrative control
  - Routers/links managed by a single "institution"
  - Service provider, company, university, …
- Hierarchy of Autonomous Systems
  - Large, tier-1 provider with a nationwide backbone
  - Medium-sized regional provider with smaller backbone
  - Small network run by a single company or university
- Interaction between Autonomous Systems
  - Internal topology is not shared between ASes
  - … but, neighboring ASes interact to coordinate routing

Autonomous System Numbers

AS Numbers are 16 bit values.

Currently over 50,000 in use.

- **Level 3: 1**
- **MIT: 3**
- **Harvard: 11**
- **Yale: 29**
- **Princeton: 88**
- **AT&T: 7018, 6341, 5074, ...**
- **UUNET: 701, 702, 284, 12199, ...**
- **Sprint: 1239, 1240, 6211, 6242, ...**
- **...**

whois –h whois.arin.net as88

OrgName: Princeton University
OrgID: PRNU
Address: Office of Information Technology
Address: 87 Prospect Avenue
City: Princeton
StateProv: NJ
PostalCode: 08540
Country: US

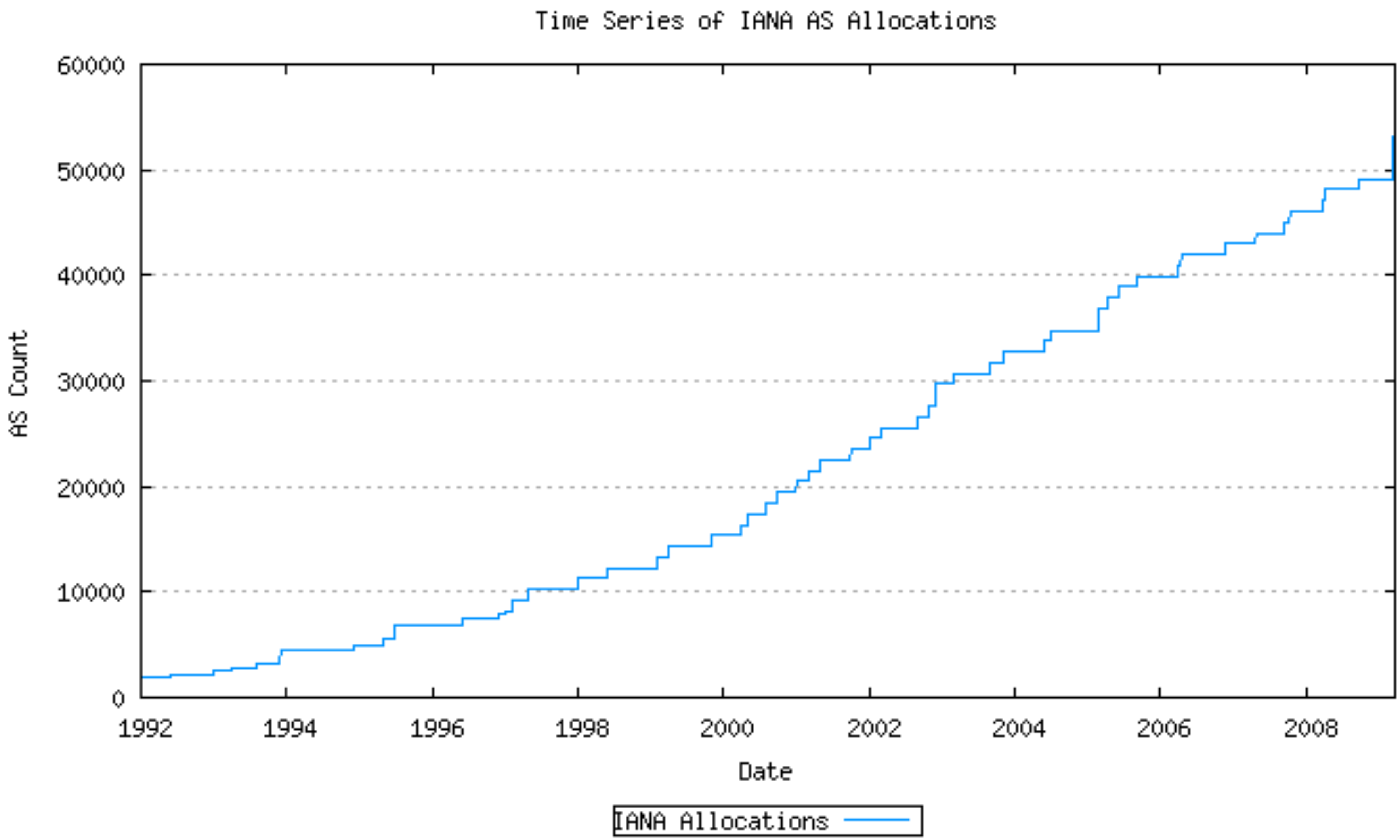
ASNumber: 88
ASName: PRINCETON-AS
ASHandle: AS88
Comment:
RegDate:
Updated: 2008-03-07

RTechHandle: PAO3-ARIN
RTechName: Olenick, Peter
RTechPhone: +1-609-258-6024
RTechEmail: polenick@princeton.edu
…

AS Number Trivia

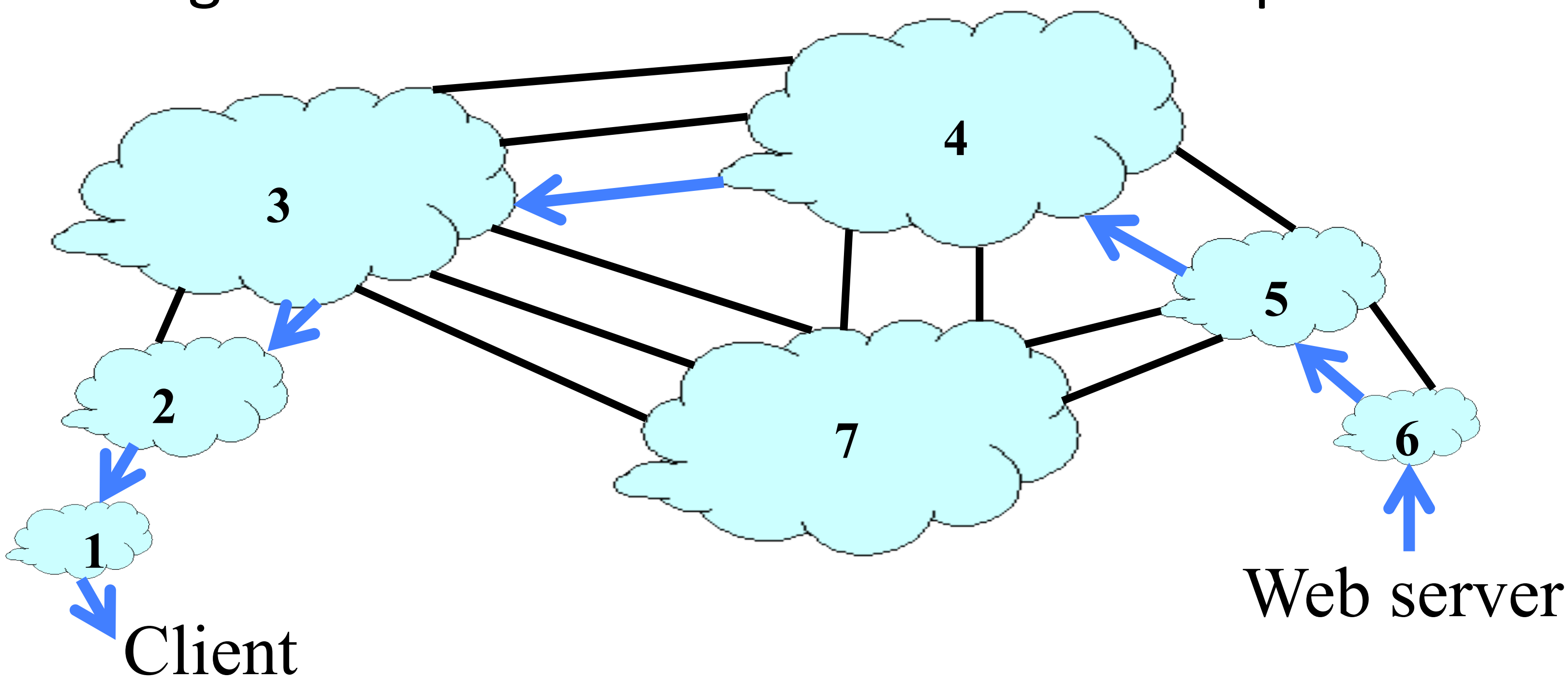
- AS number is a 16-bit quantity
  - So, 65,536 unique AS numbers
- Some are reserved (e.g., for private AS numbers)
  - So, only 64,510 are available for public use
- Managed by Internet Assigned Numbers Authority
  - Gives blocks of 1024 to Regional Internet Registries
  - IANA has allocated 39,934 AS numbers to RIRs (Jan'06)
- RIRs assign AS numbers to institutions
  - RIRs have assigned 34,827 (Jan'06)
  - Only 21,191 are visible in interdomain routing (Jan'06)
- Recently started assigning 32-bit AS #s (2007)

Growth of AS numbers



Interdomain Routing

- AS-level topology
  - Destinations are IP prefixes (e.g., 12.0.0.0/8)
  - Nodes are Autonomous Systems (ASes)
  - Edges are links and business relationships



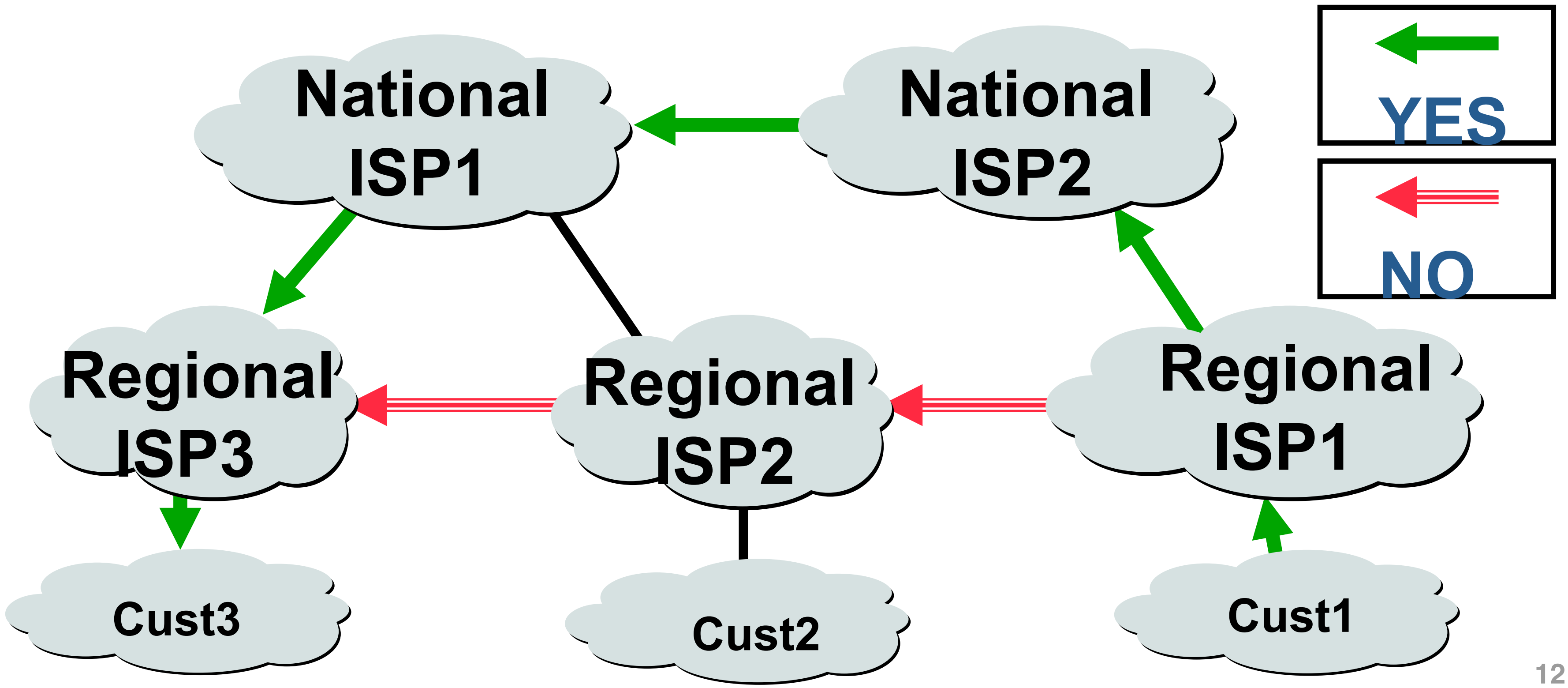
Challenges for Interdomain Routing

- Scale
  - Prefixes: 200,000, and growing
  - ASes: 20,000+ visible ones, and 60K allocated
  - Routers: at least in the millions…
- Privacy
  - ASes don't want to divulge internal topologies
  - … or 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

Path-Vector Routing

Shortest-Path Routing is Restrictive

- All traffic must travel on shortest paths
- All nodes need common notion of link costs
- Incompatible with commercial relationships



Link-State Routing is Problematic

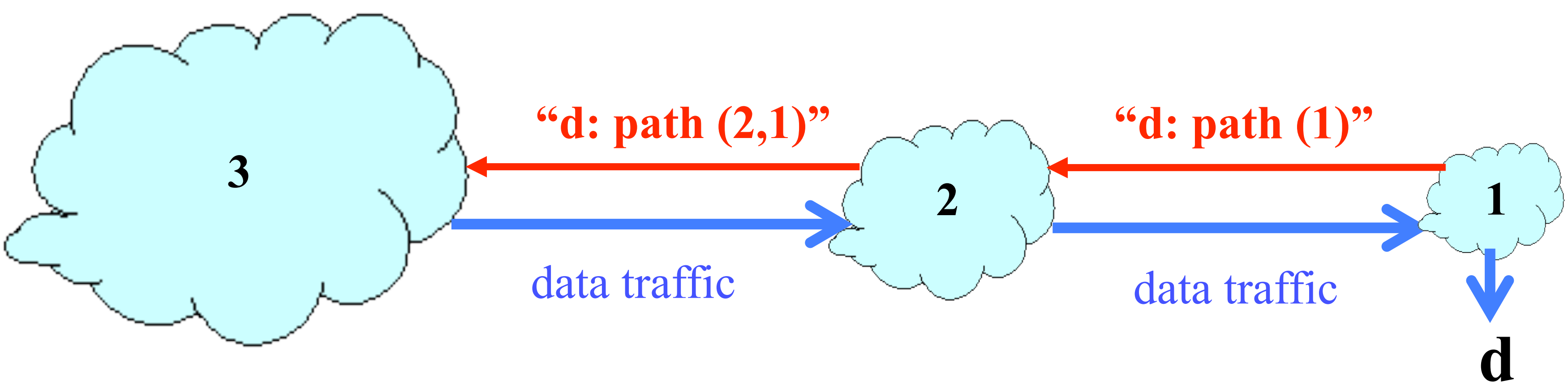
- Topology information is flooded
  - High bandwidth and storage overhead
  - Forces nodes to divulge sensitive information
- Entire path computed locally per node
  - High processing overhead in a large network
- Minimizes some notion of total distance
  - Works only if policy is shared and uniform
- Typically used only inside an AS
  - E.g., OSPF and IS-IS

Distance Vector is on the Right Track

- Advantages
  - Hides details of the network topology
  - Nodes determine only "next hop" toward the dest
- Disadvantages
  - Minimizes some notion of total distance, which is difficult in an interdomain setting
  - Slow convergence due to the counting-to-infinity problem ("bad news travels slowly")
- Idea: extend the notion of a distance vector
  - To make it easier to detect loops

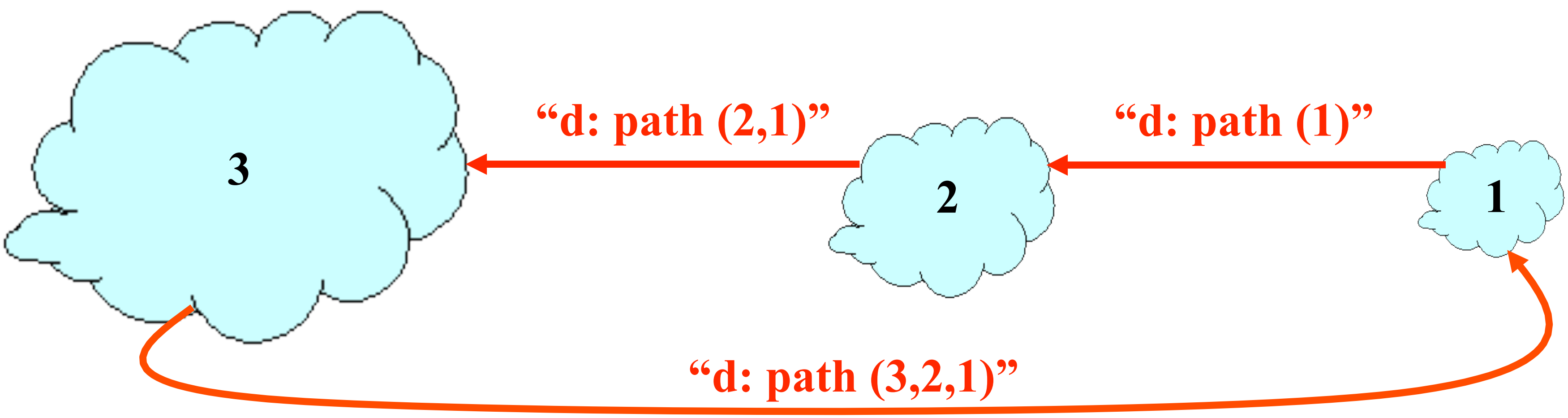
Path-Vector Routing

- Extension of distance-vector routing
  - Support flexible routing policies
  - Avoid count-to-infinity problem
- Key idea: advertise the entire path
  - Distance vector: send *distance metric* per dest d
  - Path vector: send the *entire path* for each dest d



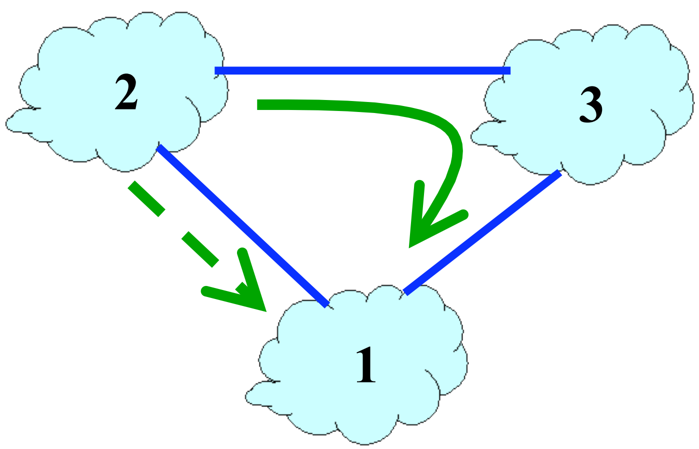
Faster Loop Detection

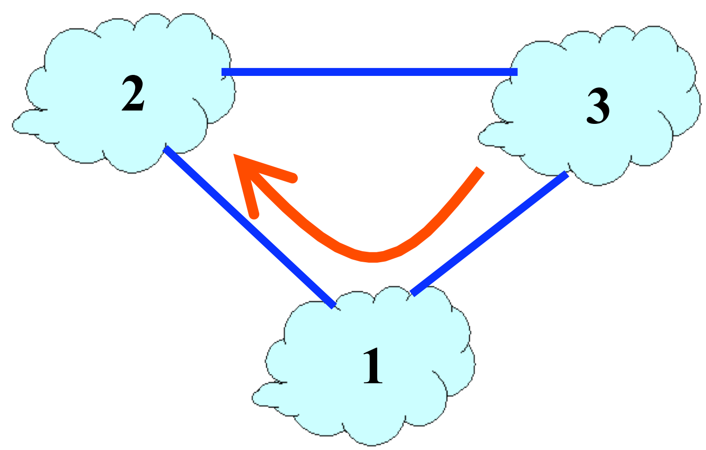
- Node can easily detect a loop
  - Look for its own node identifier in the path
  - E.g., node 1 sees itself in the path "3, 2, 1"
- Node can simply discard paths with loops
  - E.g., node 1 simply discards the advertisement



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 let node 3 hear the path “1, 2”





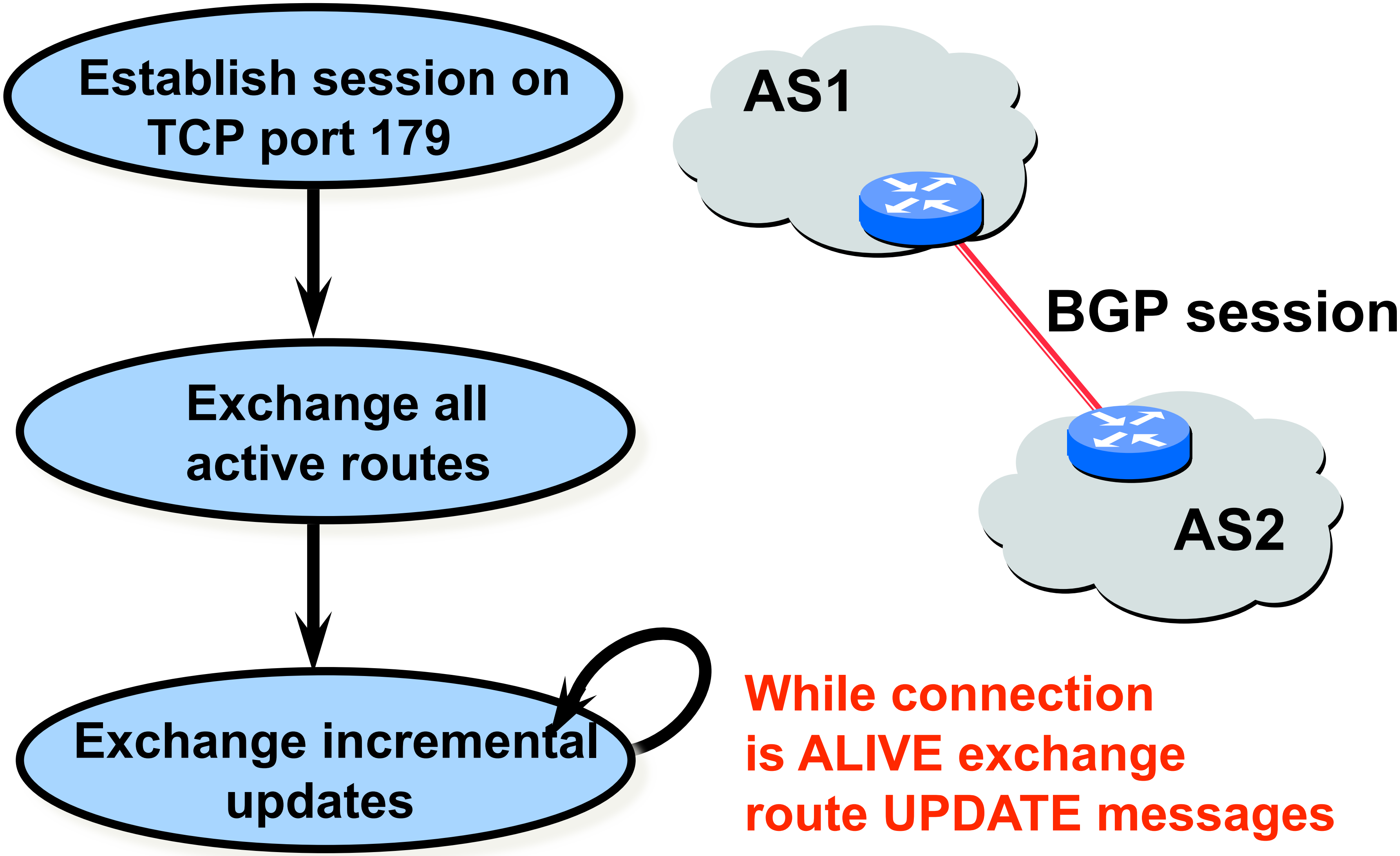
Border Gateway Protocol (BGP)

Border Gateway Protocol

- Interdomain routing protocol for the Internet
  - Prefix-based path-vector protocol
  - Policy-based routing based on AS Paths
  - Evolved during the past 18 years

- **1989 : BGP-1 [RFC 1105], replacement for EGP**
- **1990 : BGP-2 [RFC 1163]**
- **1991 : BGP-3 [RFC 1267]**
- **1995 : BGP-4 [RFC 1771], support for CIDR**
- **2006 : BGP-4 [RFC 4271], update**

BGP Operations



Incremental Protocol

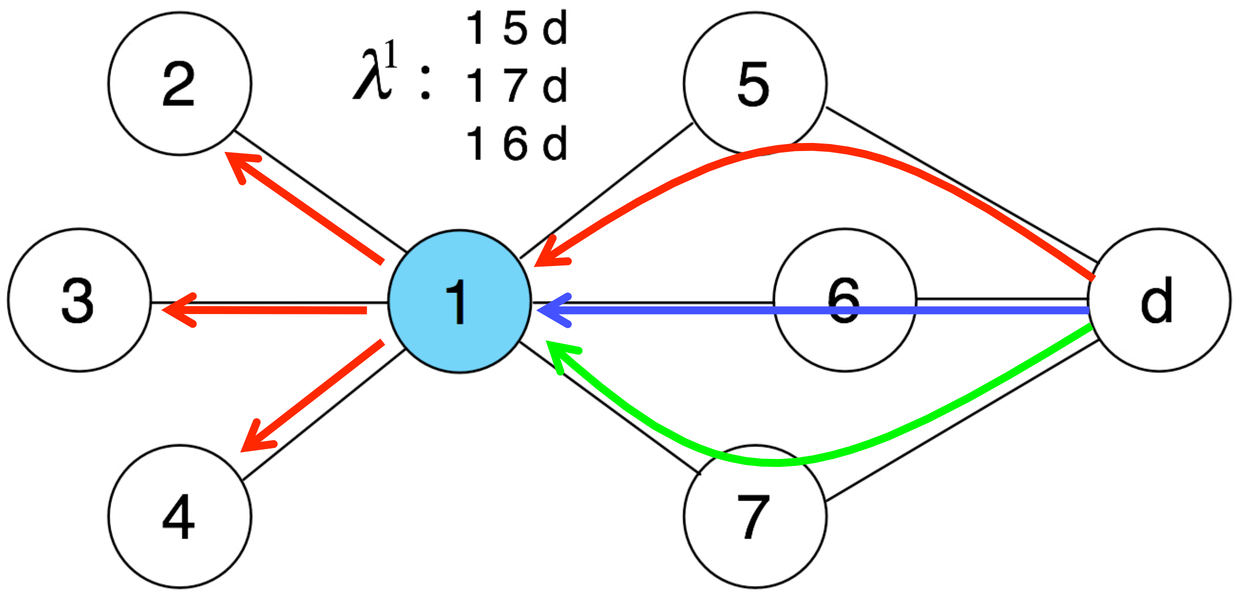
- A node learns multiple paths to destination
  - Stores all of the routes in a routing table
  - Applies policy to select a single active route
  - … and may advertise the route to its neighbors
- Incremental updates
  - Announcement
    - Upon selecting a new active route, add node id to path
    - … and (optionally) advertise to each neighbor
  - Withdrawal
    - If the active route is no longer available
    - … send a withdrawal message to the neighbors

Incremental Protocol

- A node learns multiple paths to destination
  - Stores all of the routes in a routing table
  - Applies policy to select a single active route
  - … and may advertise the route to its neighbors
- Incremental updates
  - Announcement
    - Upon selecting a new active route, add node id to path
    - … and (optionally) advertise to each neighbor
  - Withdrawal
    - If the active route is no longer available
    - … send a withdrawal message to the neighbors

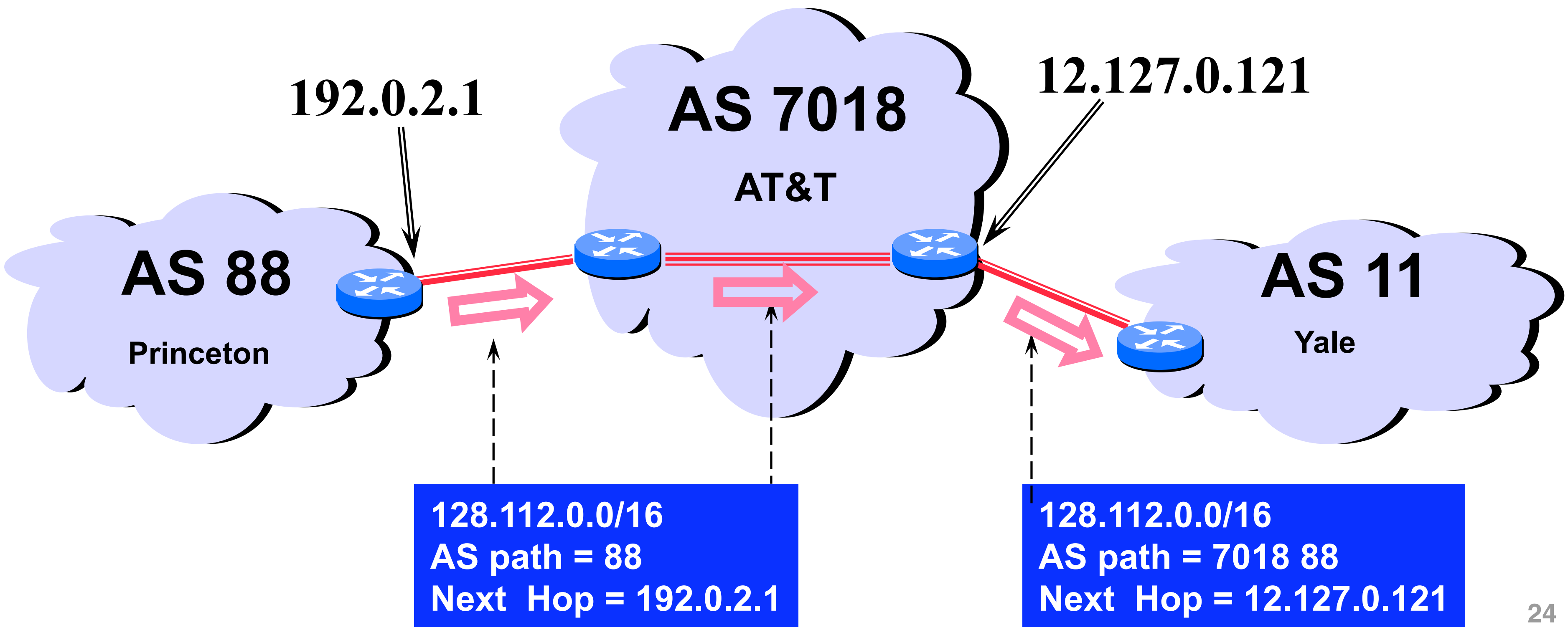
Export active routes

- In conventional path vector routing, a node has one *ranking* function, which reflects its routing policy

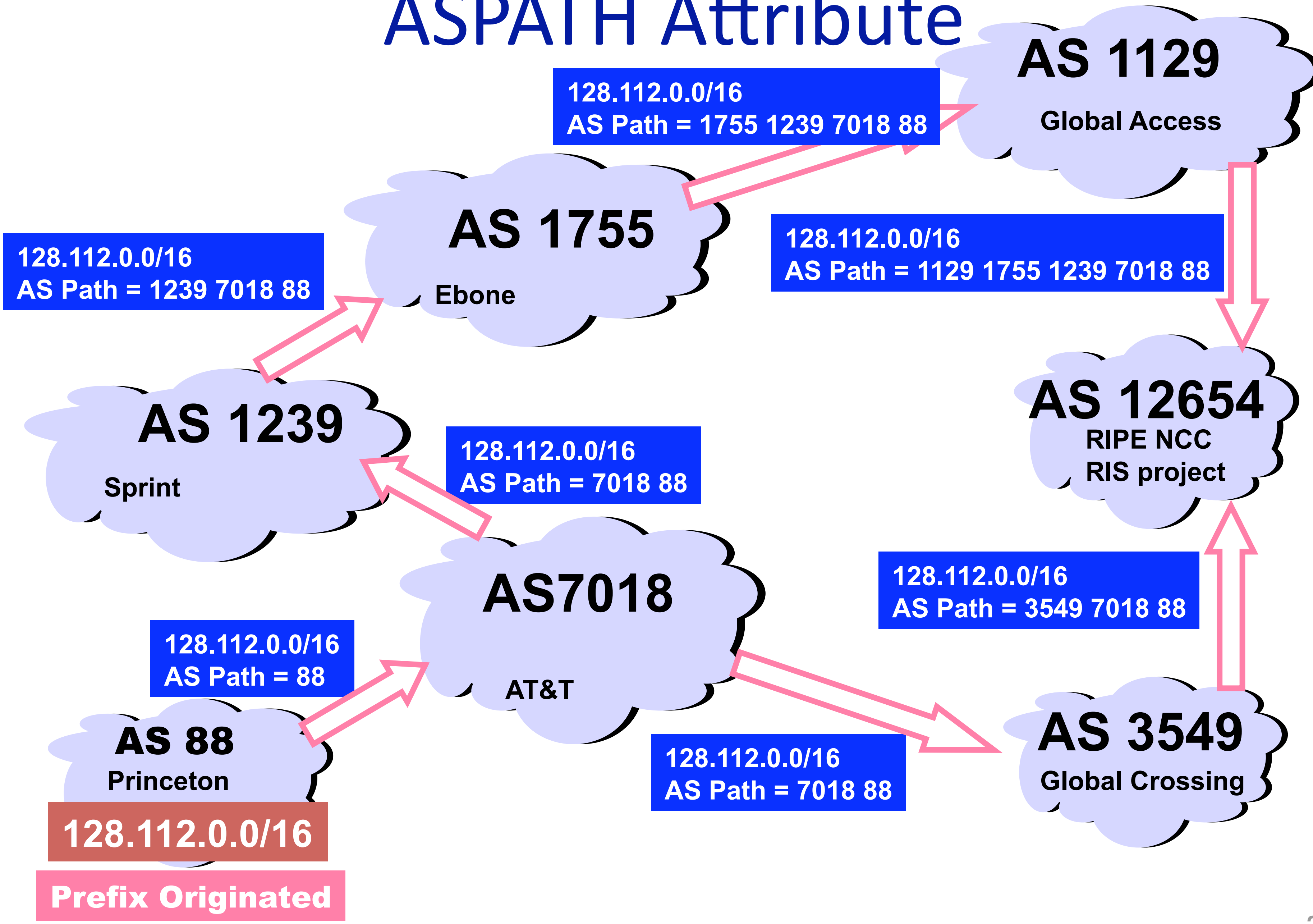


BGP Route

- Destination prefix (e.g., 128.112.0.0/16)
- Route attributes, including
  - AS path (e.g., "7018 88")
  - Next-hop IP address (e.g., 12.127.0.121)

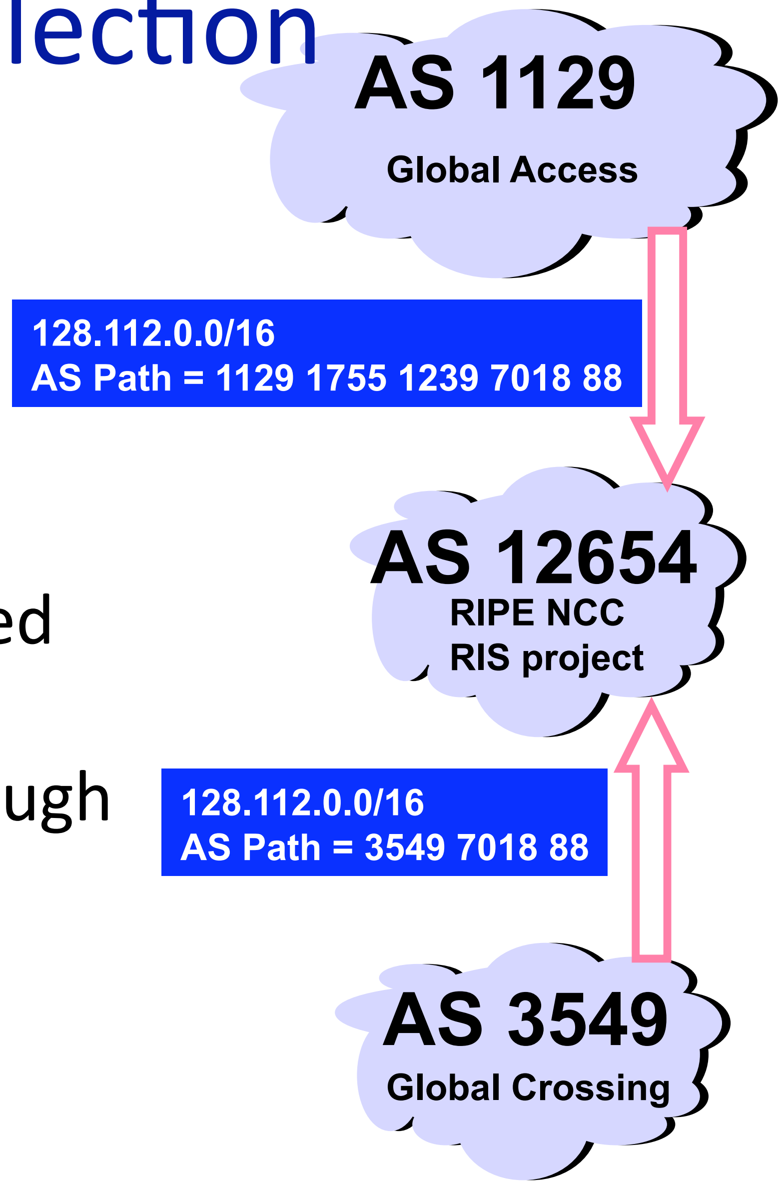


ASPATH Attribute



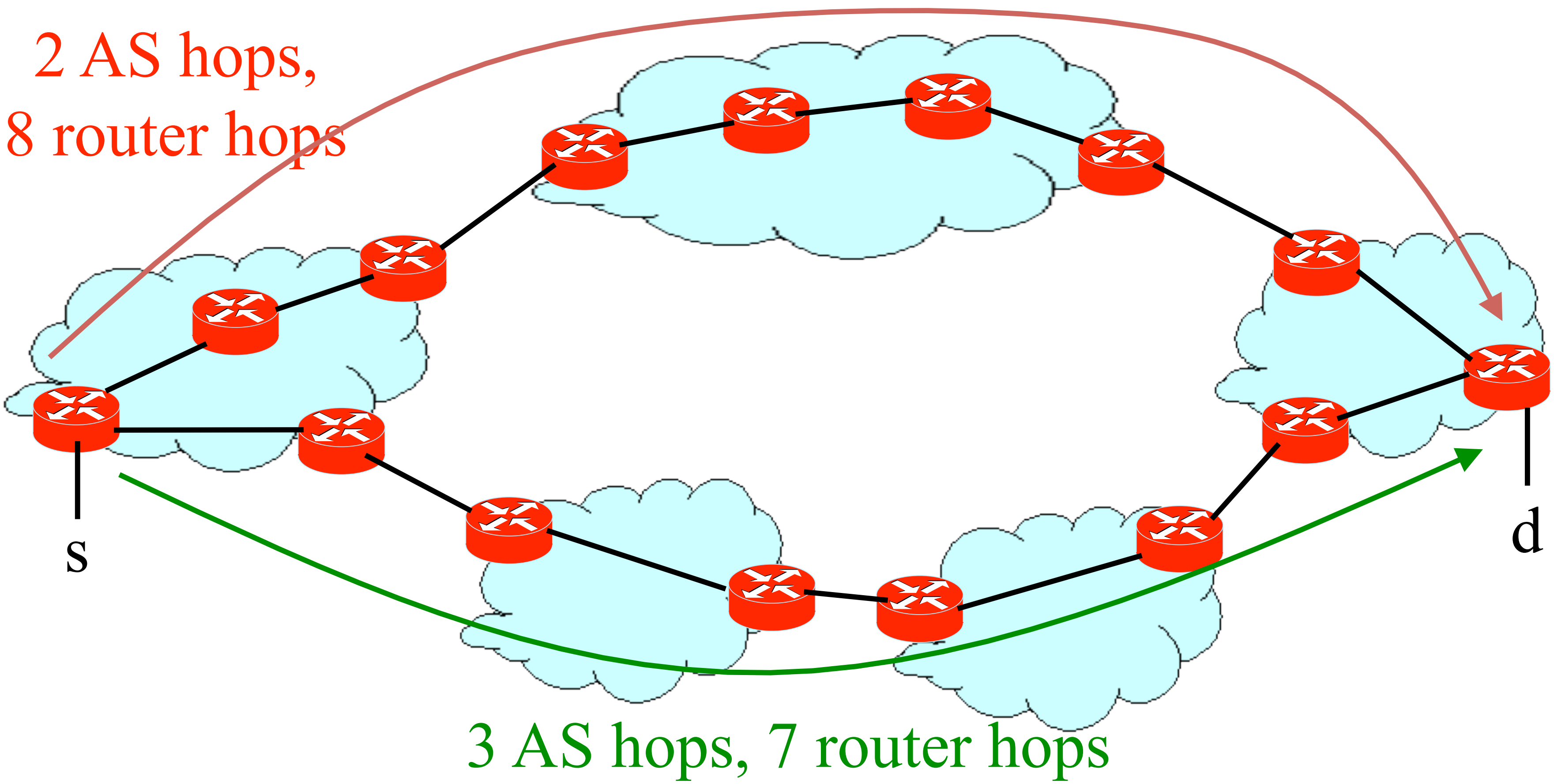
BGP Path Selection

- Simplest case
  - Shortest AS path
  - Arbitrary tie break
- Example
  - Three-hop AS path preferred over a five-hop AS path
  - AS 12654 prefers path through Global Crossing
- But, BGP is not limited to shortest-path routing
  - Policy-based routing



AS Path Length != Router Hops

- AS path may be longer than shortest AS path
- Router path may be longer than shortest path

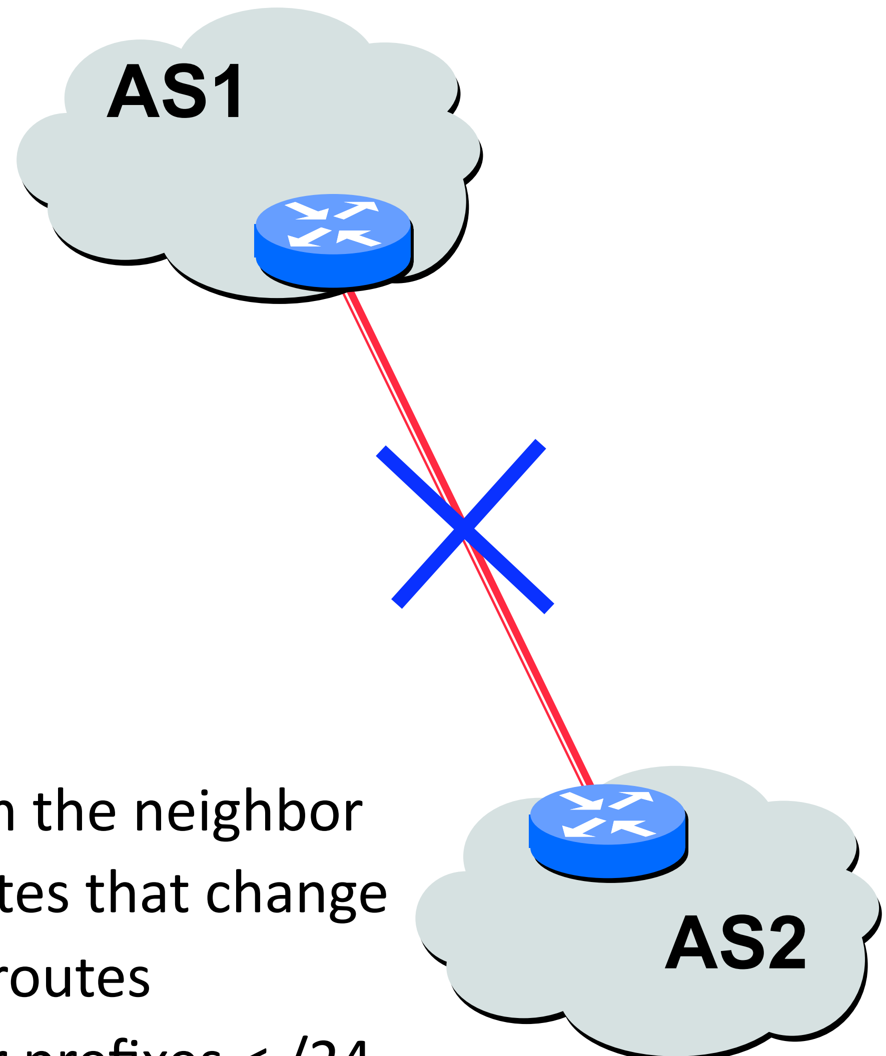


BGP Convergence

Causes of BGP Routing Changes

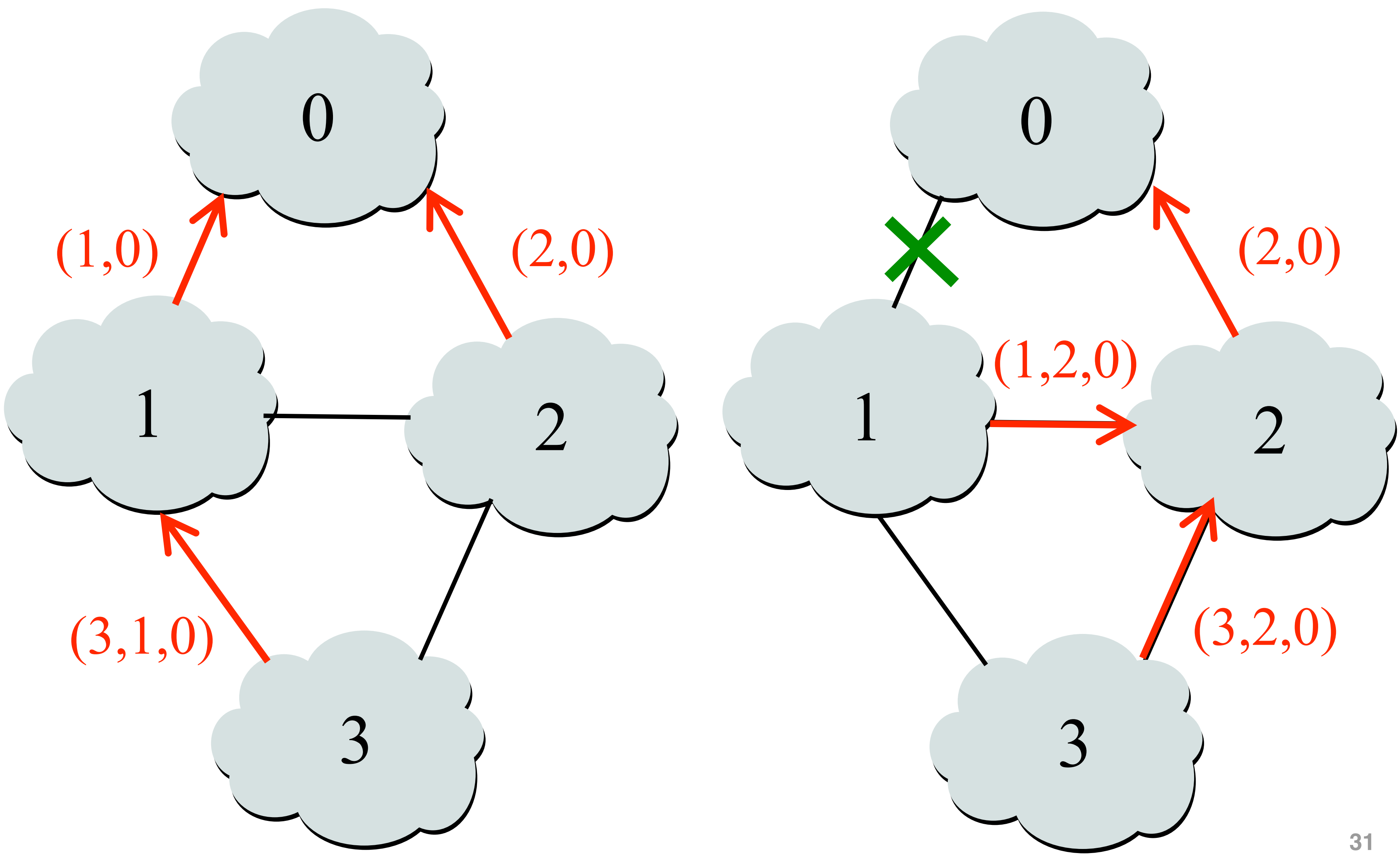
- Topology changes
  - Equipment going up or down
  - Deployment of new routers or sessions
- BGP session failures
  - Due to equipment failures, maintenance, etc.
  - Or, due to congestion on the physical path
- Changes in routing policy
  - Changes in preferences in the routes
  - Changes in whether 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
  - Overhead increases with # of routes
    - Why many tier-1 ASes filter prefixes < /24

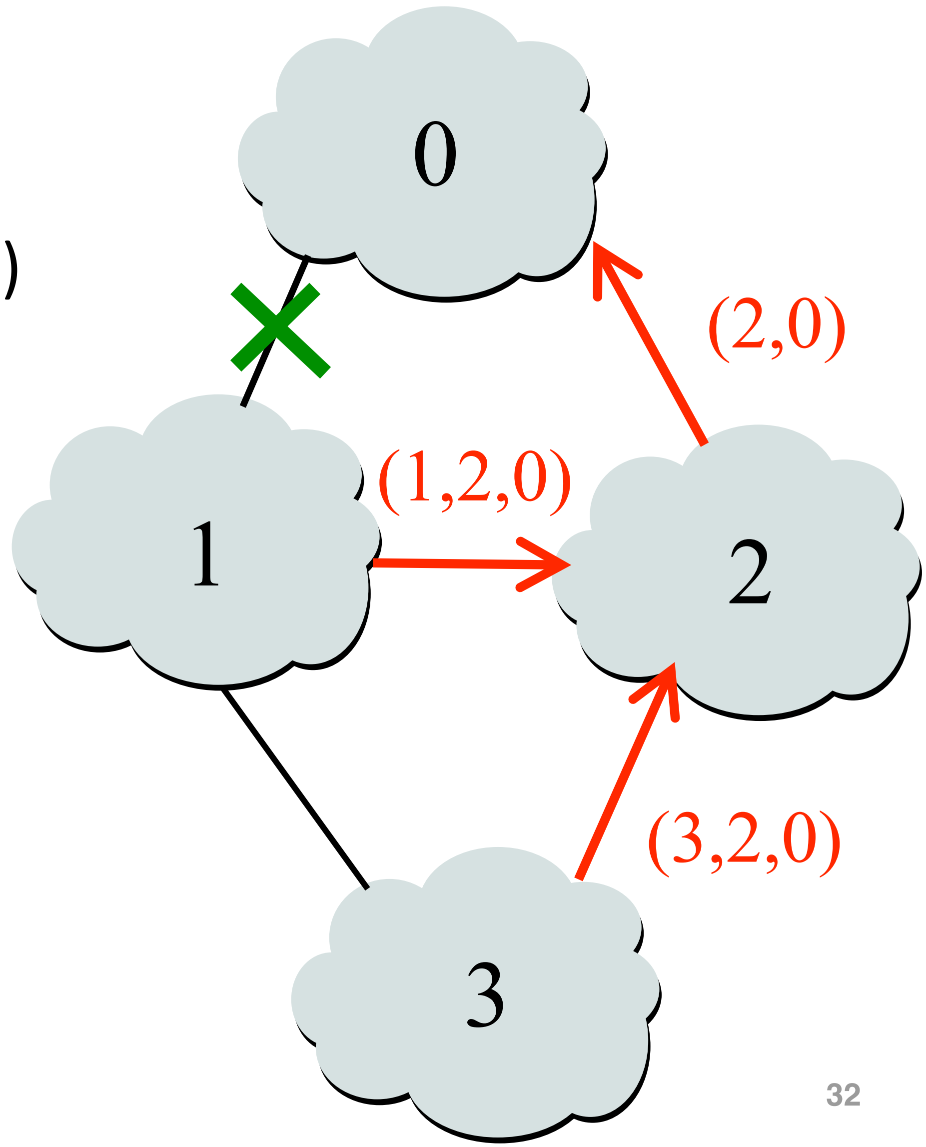
Routing Change: Before and After



Routing Change: Path Exploration

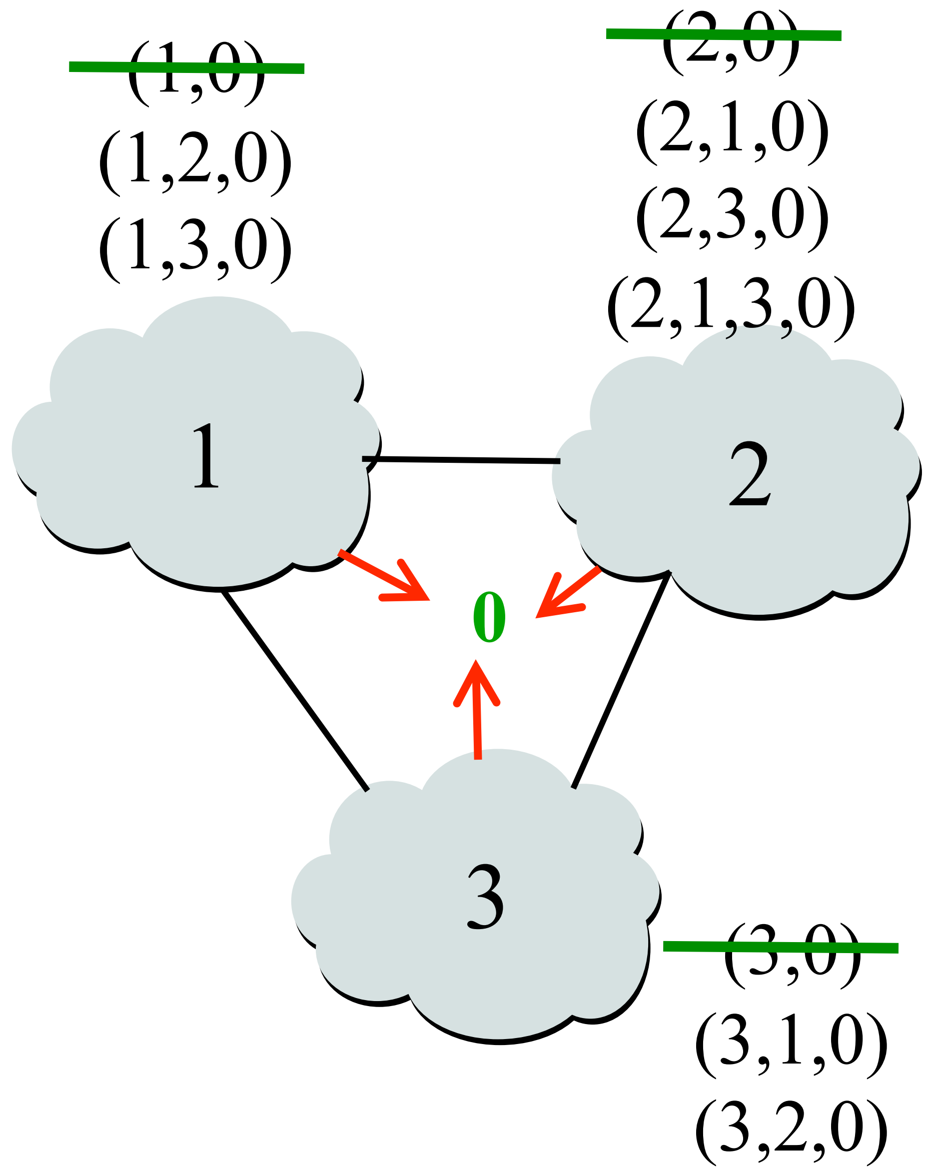
- AS 1
  - Delete the route (1,0)
  - Switch to next route (1,2,0)
  - Send route (1,2,0) to AS 3

- AS 3
  - Sees (1,2,0) replace (1,0)
  - Compares to route (2,0)
  - Switches to using AS 2



Routing Change: Path Exploration

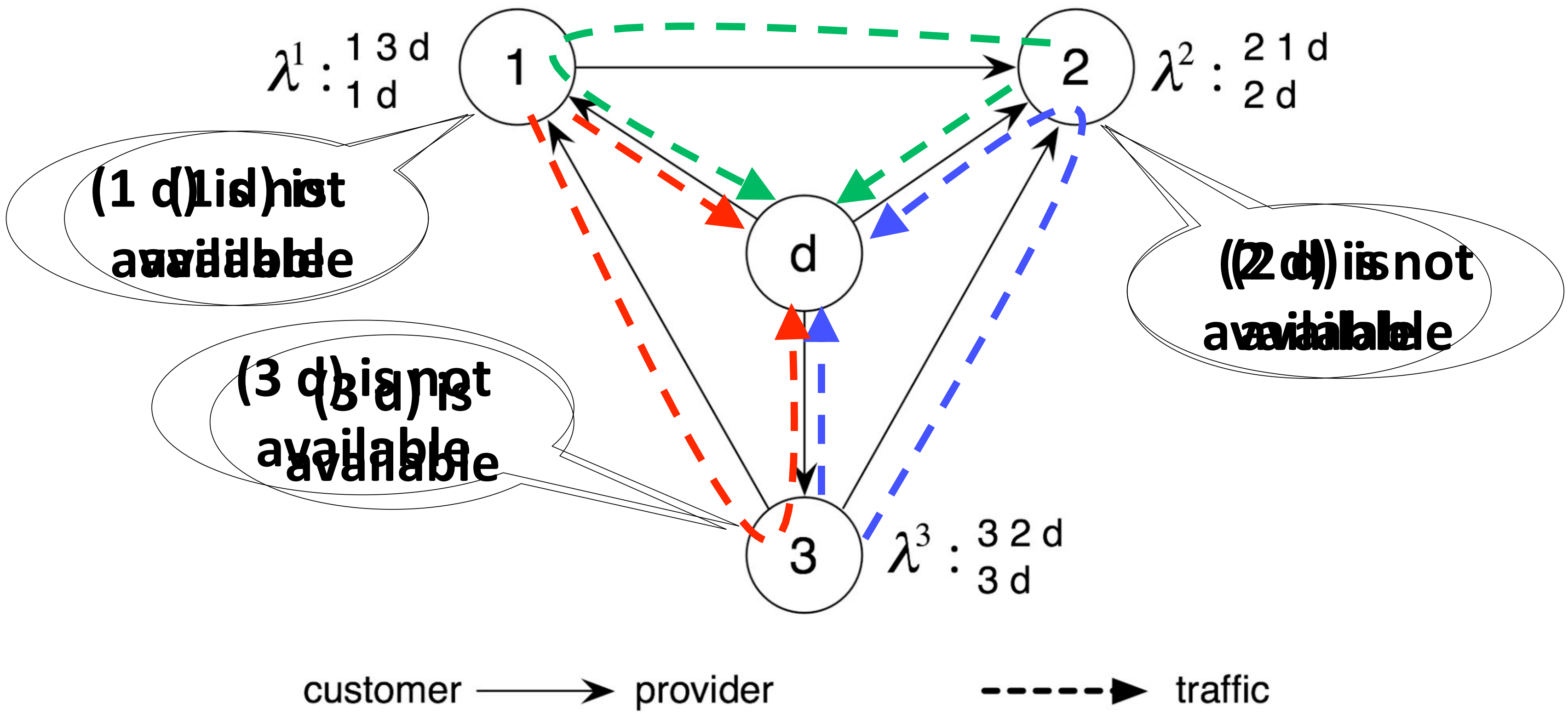
- Initial situation
  - Destination 0 is alive
  - All ASes use direct path
- When destination dies
  - All ASes lose direct path
  - All switch to longer paths
  - Eventually withdrawn
- E.g., AS 2
  - (2,0) → (2,1,0)
  - (2,1,0) → (2,3,0)
  - (2,3,0) → (2,1,3,0)
  - (2,1,3,0) → null



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 tens of minutes
  - High for important interactive applications
  - … or even conventional application, like Web browsing

BGP Not Guaranteed to Converge



Example known as a “dispute wheel”

Conclusions

- BGP is solving a hard problem
  - Routing protocol operating at a global scale
  - With tens of thousands of independent networks
  - That each have their own policy goals
  - And all want fast convergence
- Key features of BGP
  - Prefix-based path-vector protocol
  - Incremental updates (announcements and withdrawals)
- Next lecture: Tricks for setting policy!