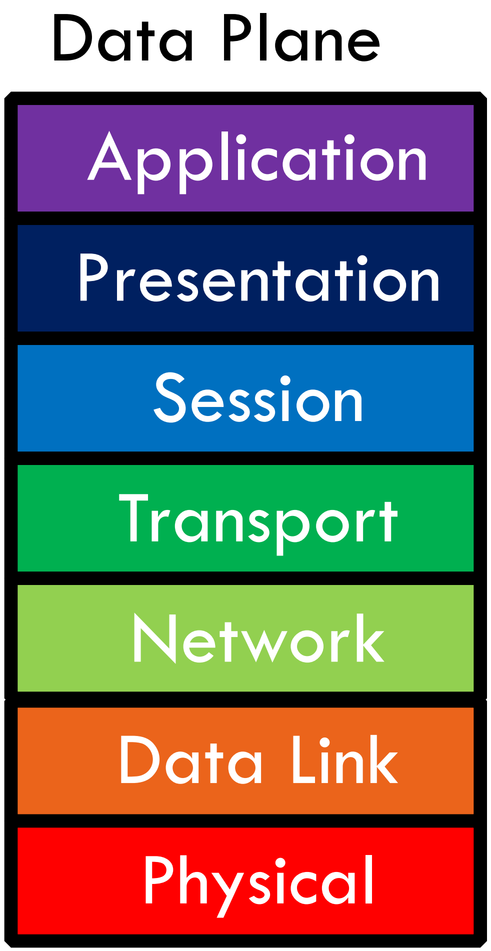
CS 3700
Networks and Distributed Systems

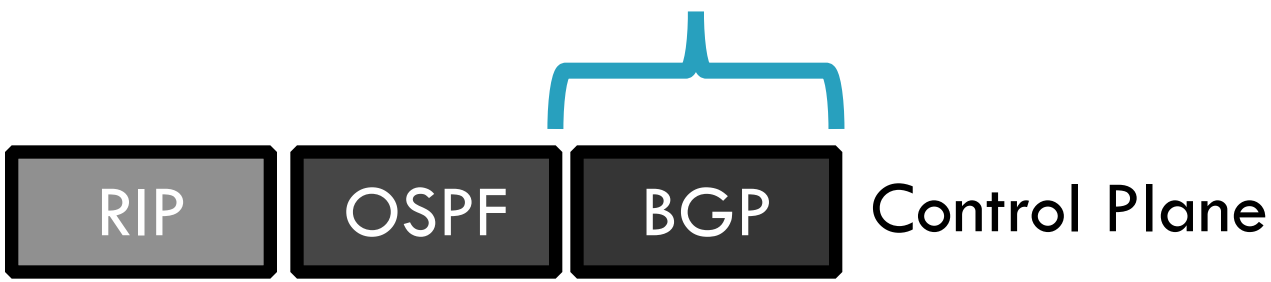
Lecture 8: Inter Domain Routing

Revised 2/4/2014

Network Layer, Control Plane

- Function:
  - Set up routes between networks
- Key challenges:
  - Implementing provider policies
  - Creating stable paths

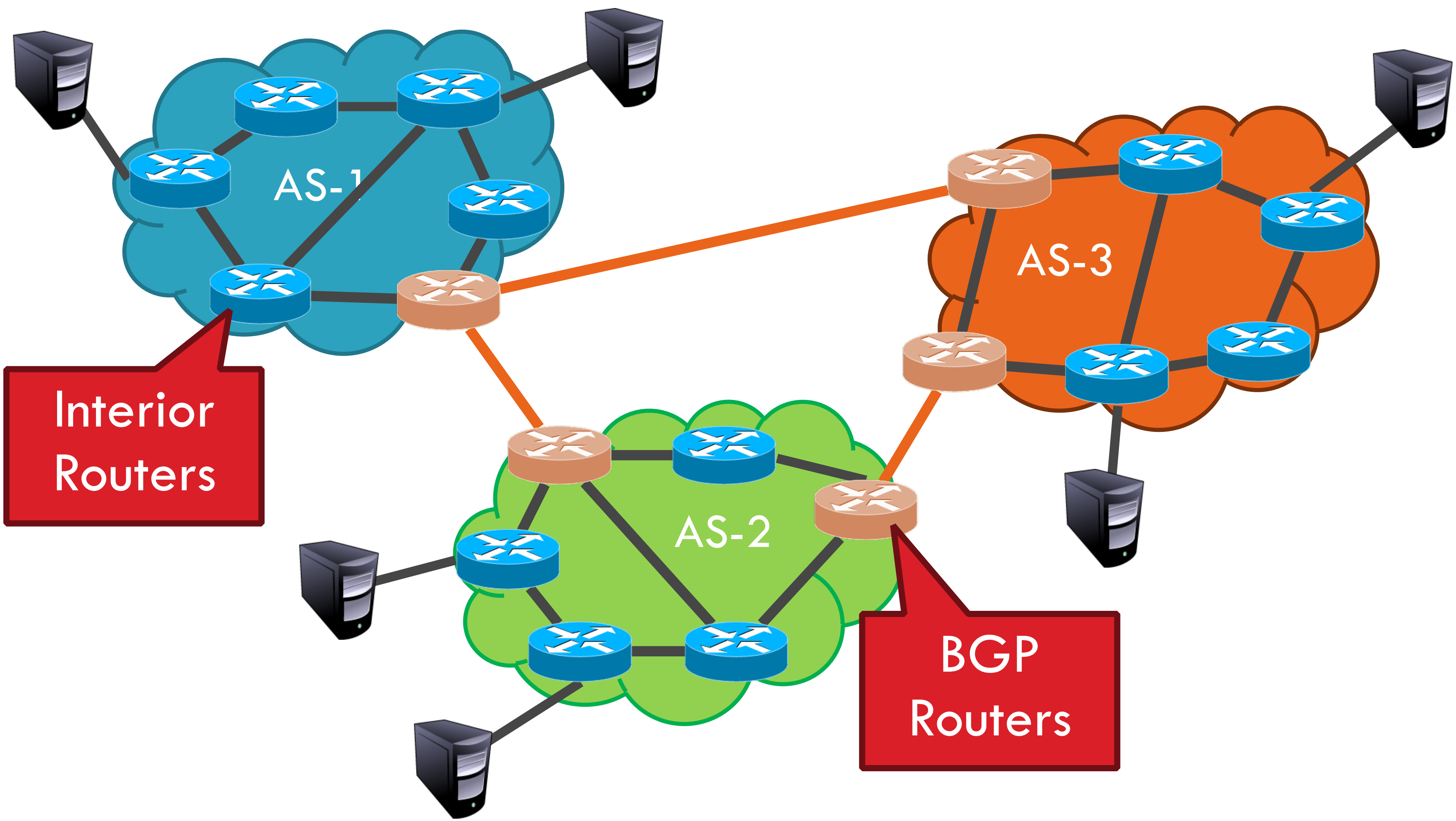




Outline

- BGP Basics
- Stable Paths Problem
- BGP in the Real World

ASs, Revisited



AS Numbers

- Each AS identified by an ASN number
  - 16-bit values (latest protocol supports 32-bit ones)
  - 64512 – 65535 are reserved
- Currently, there are > 20000 ASNs
  - AT&T: 5074, 6341, 7018, …
  - Sprint: 1239, 1240, 6211, 6242, …
  - Northeastern: 156
  - North America ASs → ftp://ftp.arin.net/info/asn.txt

Inter-Domain Routing

- Global connectivity is at stake!
  - Thus, all ASs must use the same protocol
  - Contrast with intra-domain routing

Inter-Domain Routing

- Global connectivity is at stake!
  - Thus, all ASs must use the same protocol
  - Contrast with intra-domain routing
- What are the requirements?
  - Scalability
  - Flexibility in choosing routes
    - Cost
    - Routing around failures

Inter-Domain Routing

- Global connectivity is at stake!
  - Thus, all ASs must use the same protocol
  - Contrast with intra-domain routing
- What are the requirements?
  - Scalability
  - Flexibility in choosing routes
    - Cost
    - Routing around failures
- Question: link state or distance vector?

Inter-Domain Routing

- Global connectivity is at stake!
  - Thus, all ASs must use the same protocol
  - Contrast with intra-domain routing
- What are the requirements?
  - Scalability
  - Flexibility in choosing routes
    - Cost
    - Routing around failures
- Question: link state or distance vector?
  - Trick question: BGP is a path vector protocol

BGP

- Border Gateway Protocol
  - De facto inter-domain protocol of the Internet
  - Policy based routing protocol
  - Uses a Bellman-Ford path vector protocol

BGP

- Border Gateway Protocol
  - De facto inter-domain protocol of the Internet
  - Policy based routing protocol
  - Uses a Bellman-Ford path vector protocol
- Relatively simple protocol, but…
  - Complex, manual configuration

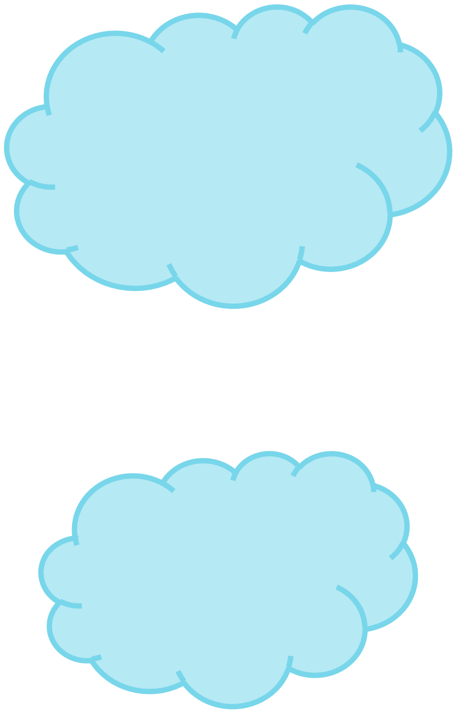
BGP

- Border Gateway Protocol
  - De facto inter-domain protocol of the Internet
  - Policy based routing protocol
  - Uses a Bellman-Ford path vector protocol
- Relatively simple protocol, but…
  - Complex, manual configuration
  - Entire world sees advertisements
    - Errors can screw up traffic globally

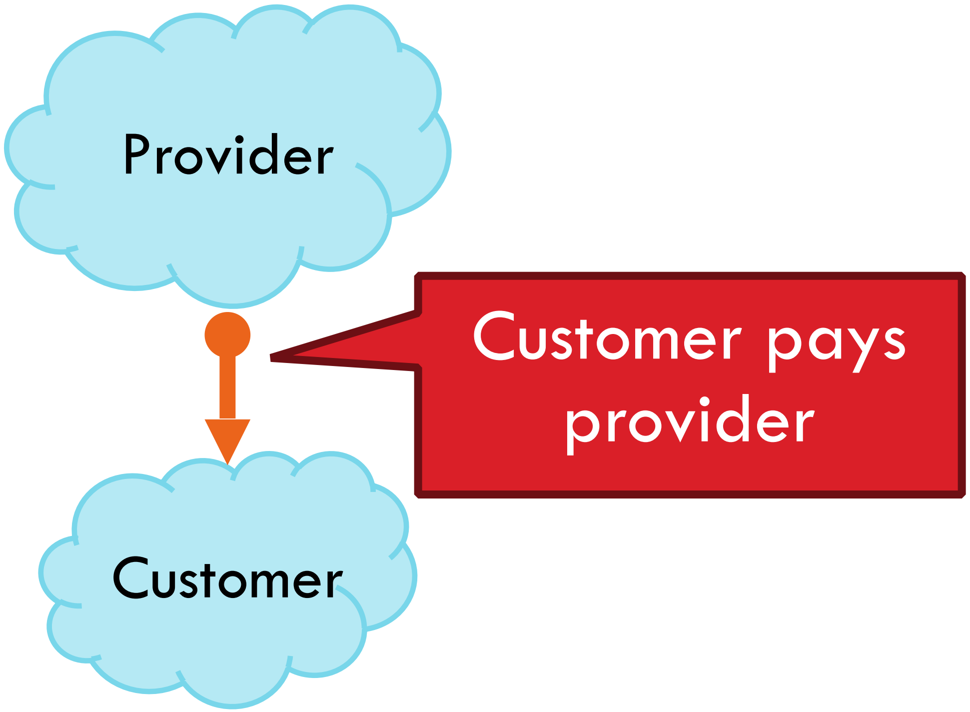
BGP

- Border Gateway Protocol
  - De facto inter-domain protocol of the Internet
  - Policy based routing protocol
  - Uses a Bellman-Ford path vector protocol
- Relatively simple protocol, but…
  - Complex, manual configuration
  - Entire world sees advertisements
    - Errors can screw up traffic globally
  - Policies driven by economics
    - How much $$$ does it cost to route along a given path?
    - Not by performance (e.g. shortest paths)

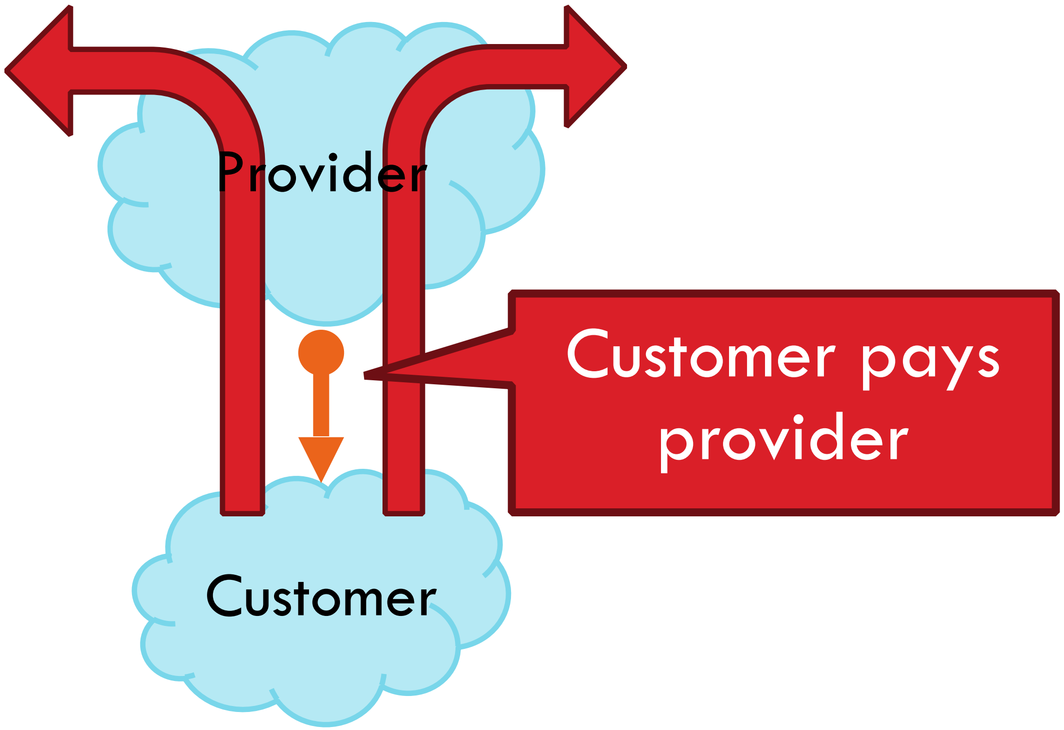
BGP Relationships



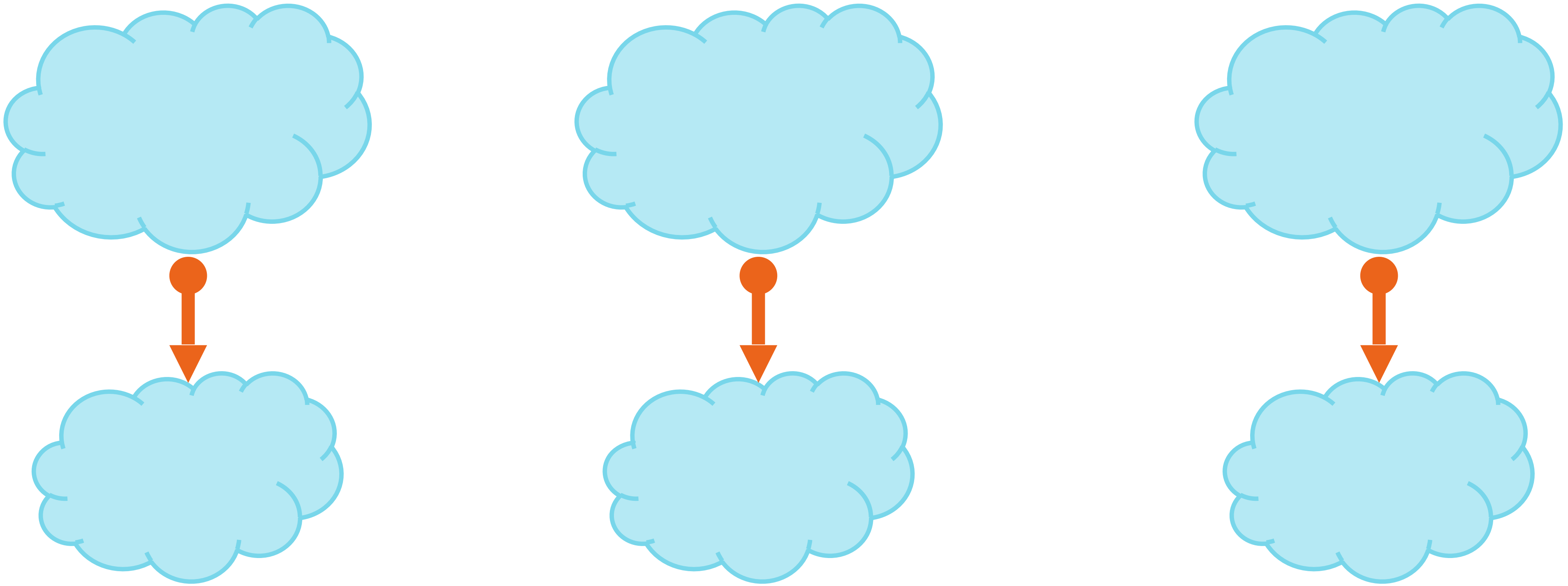
BGP Relationships



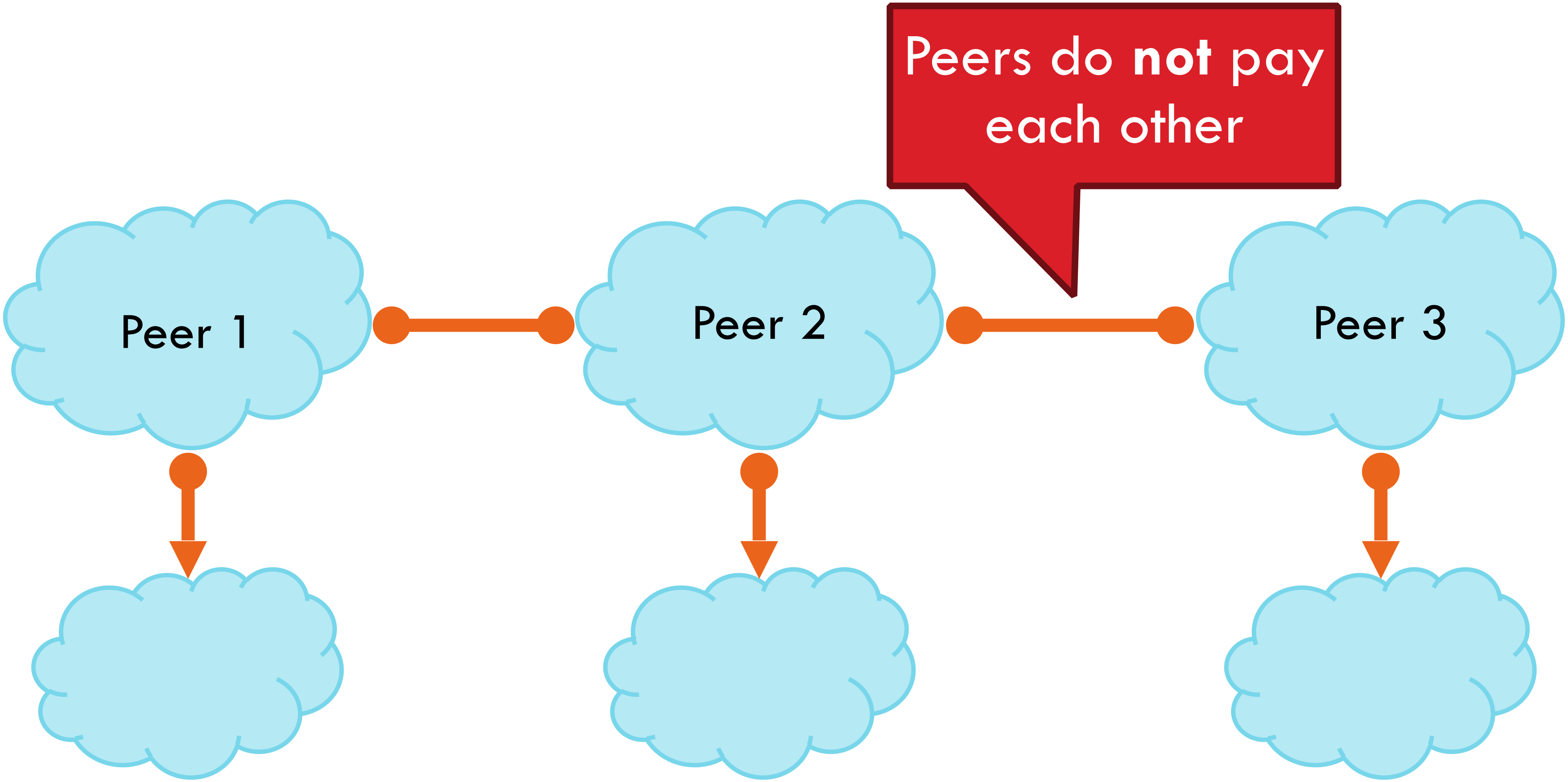
BGP Relationships



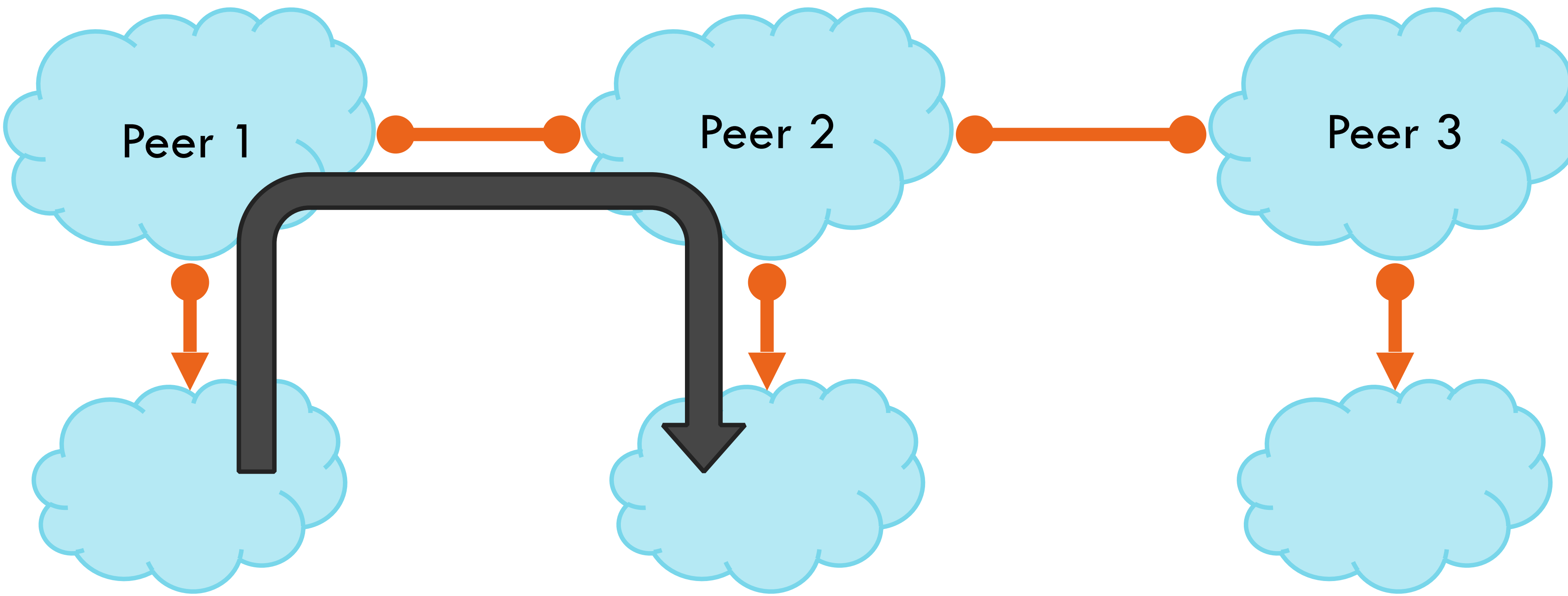
BGP Relationships



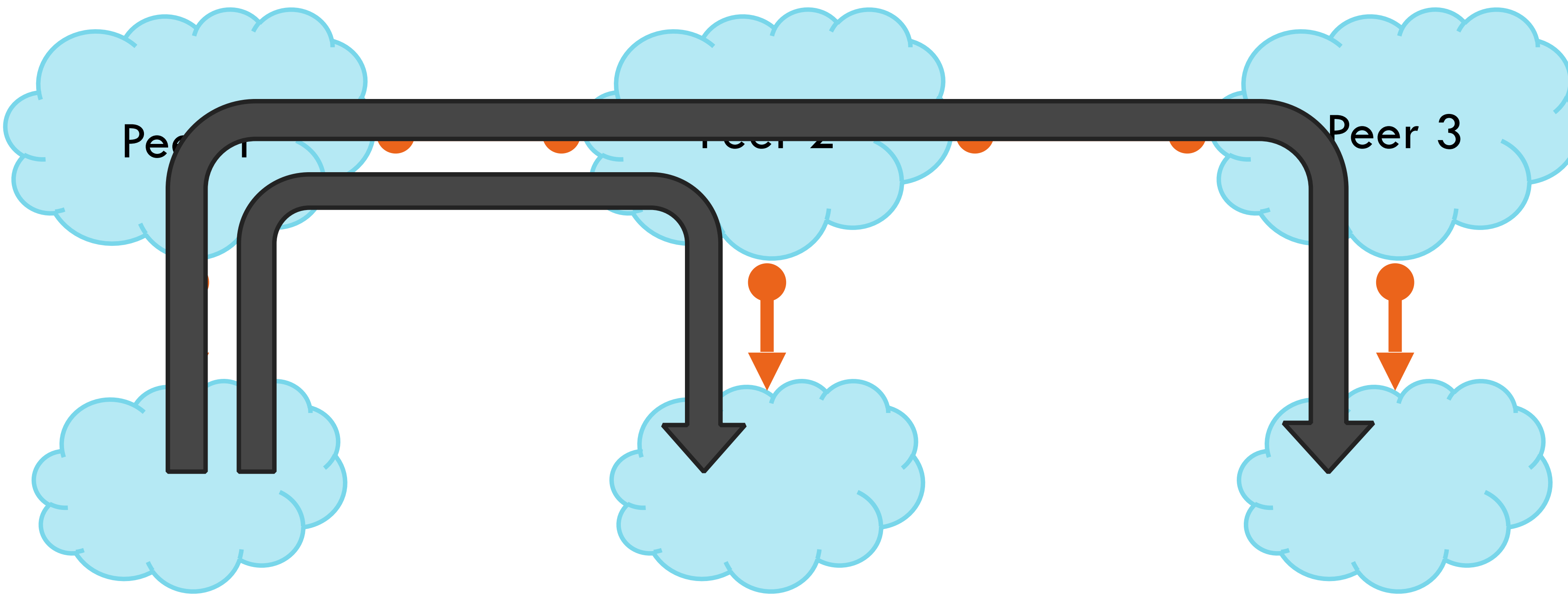
BGP Relationships



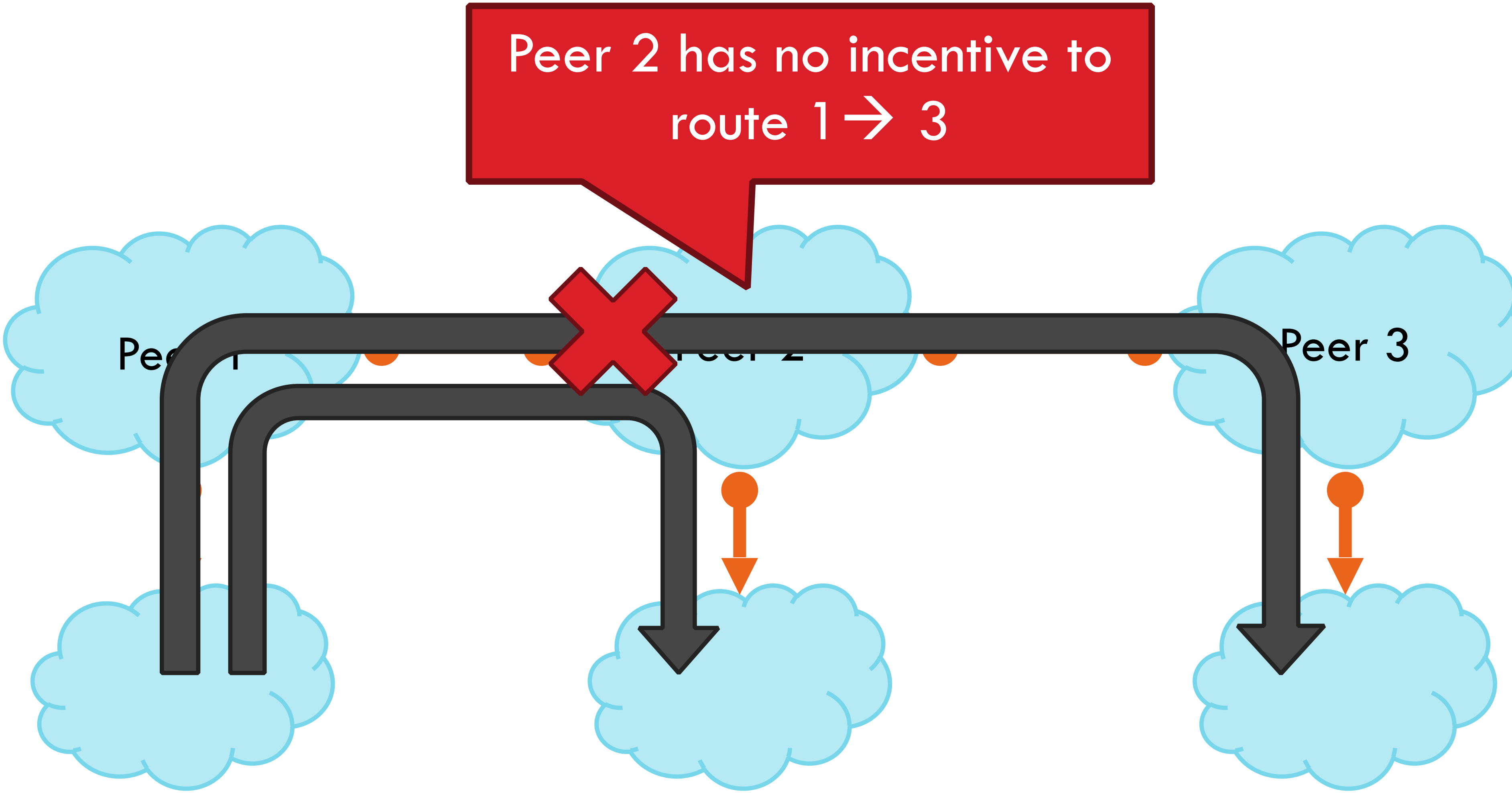
BGP Relationships



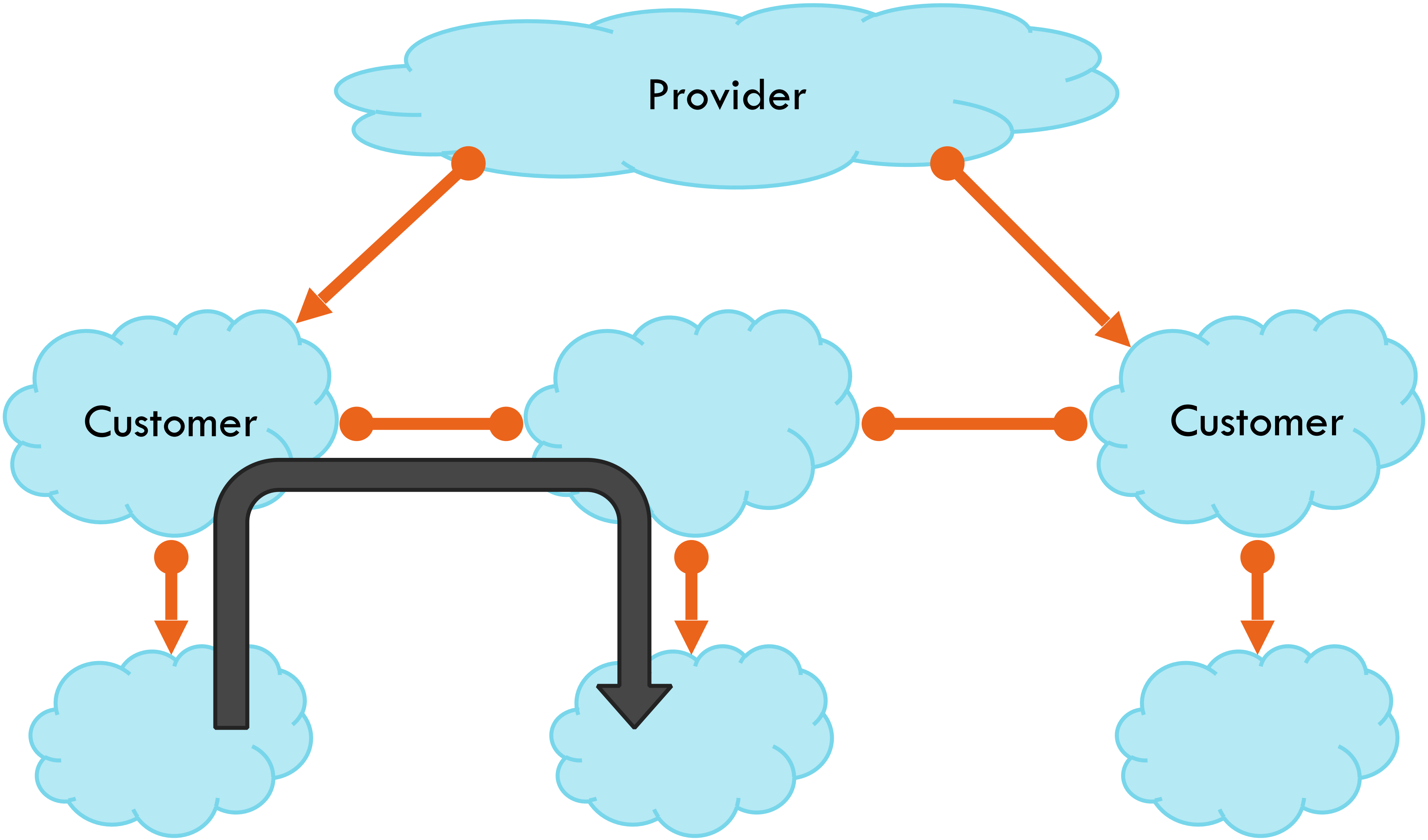
BGP Relationships



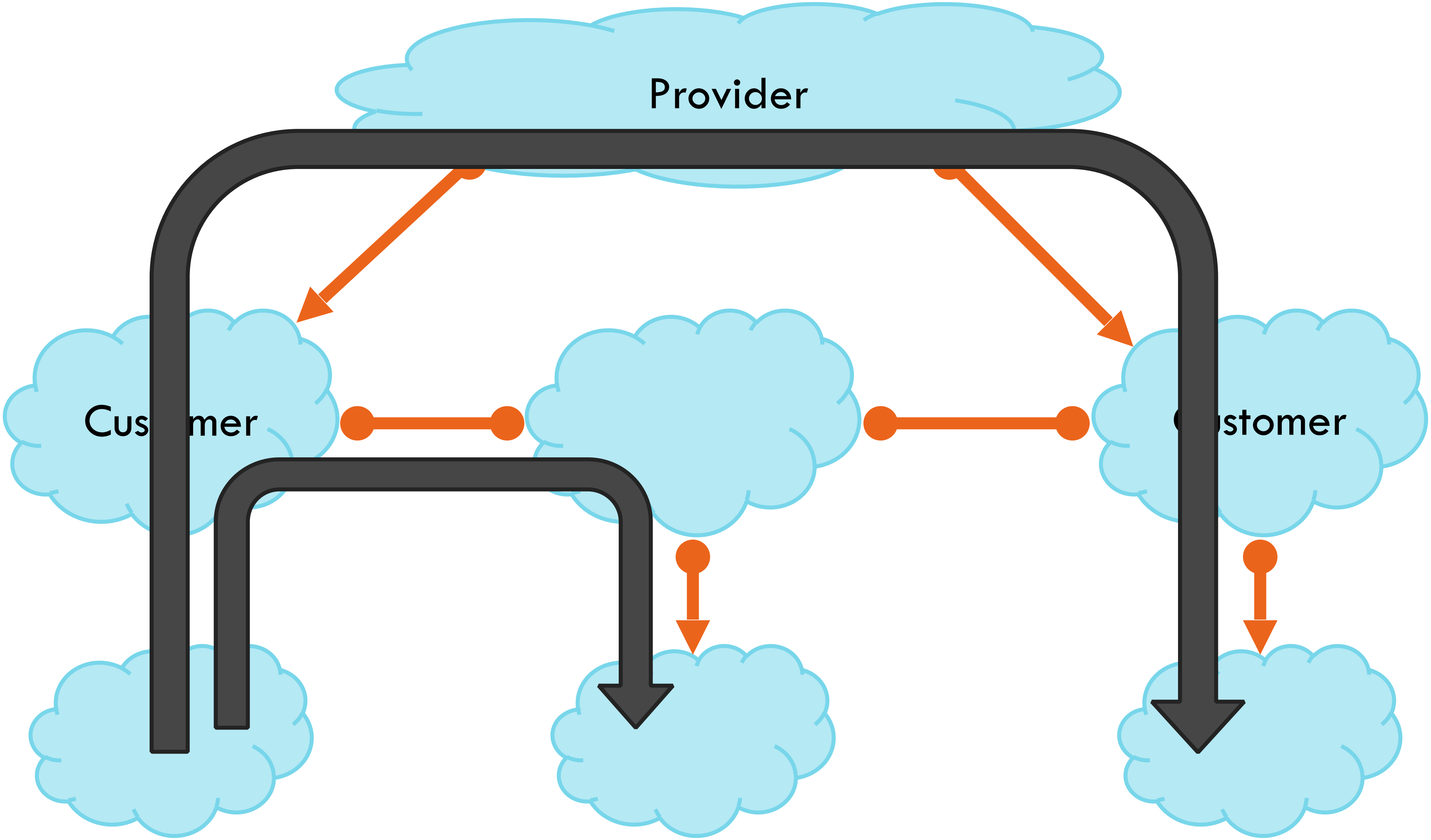
BGP Relationships



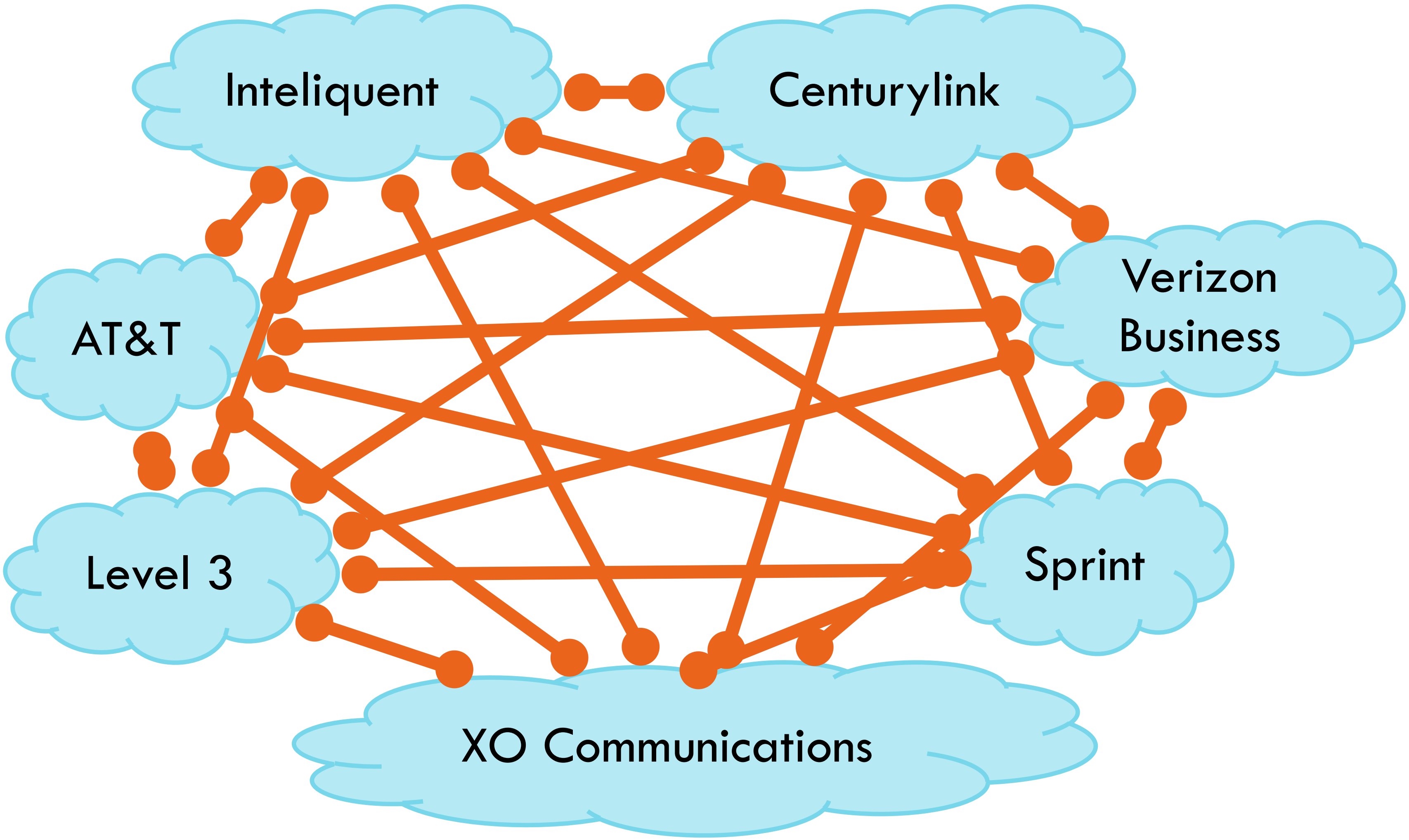
BGP Relationships



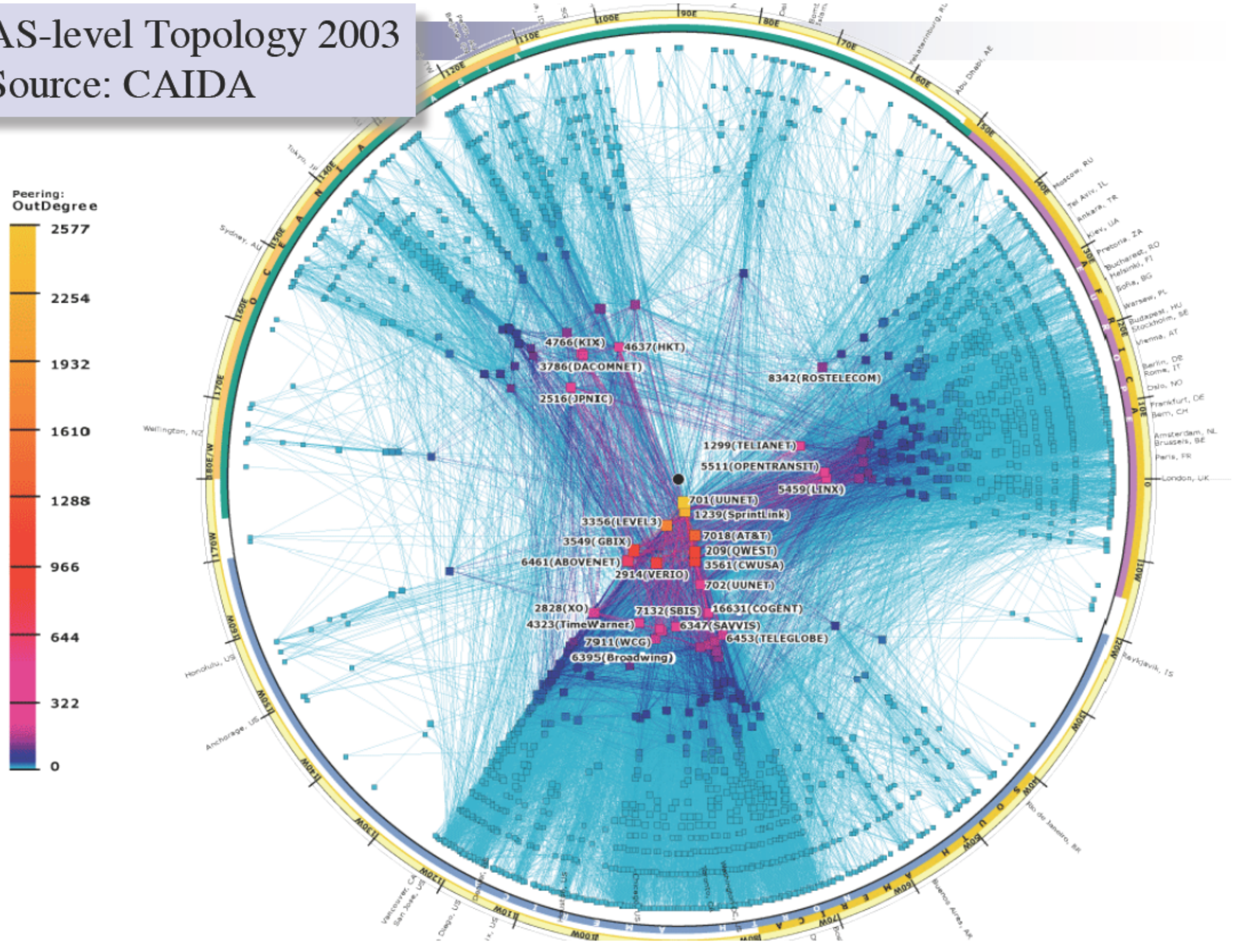
BGP Relationships



Tier-1 ISP Peering



AS-level Topology 2003
Source: CAIDA



Peering Wars

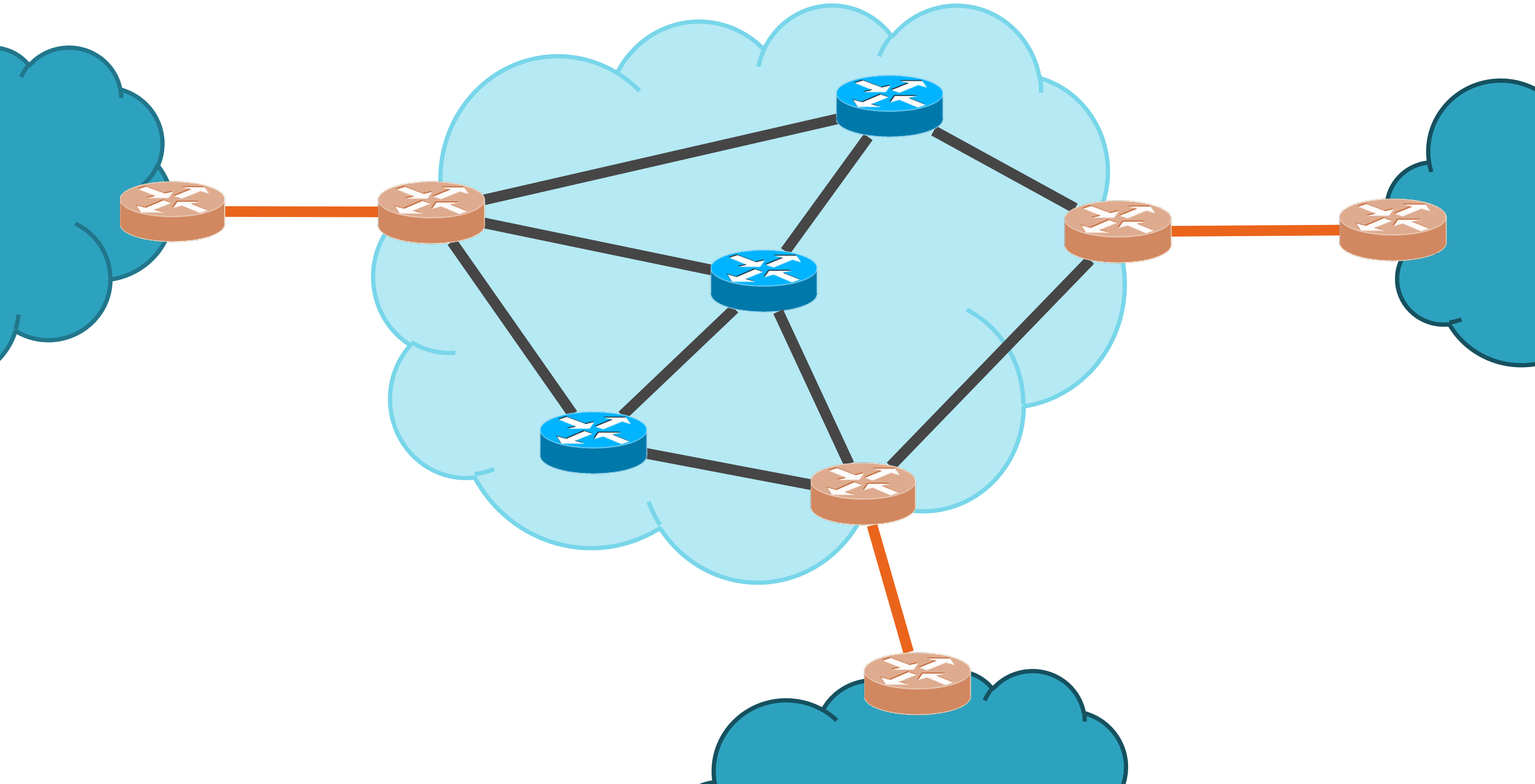
Peer	Don’t Peer
Reduce upstream costs	You would rather have customers
Improve end-to-end performance	Peers are often competitors
May be the only way to connect to parts of the Internet	Peering agreements require periodic renegotiation

Peering Wars

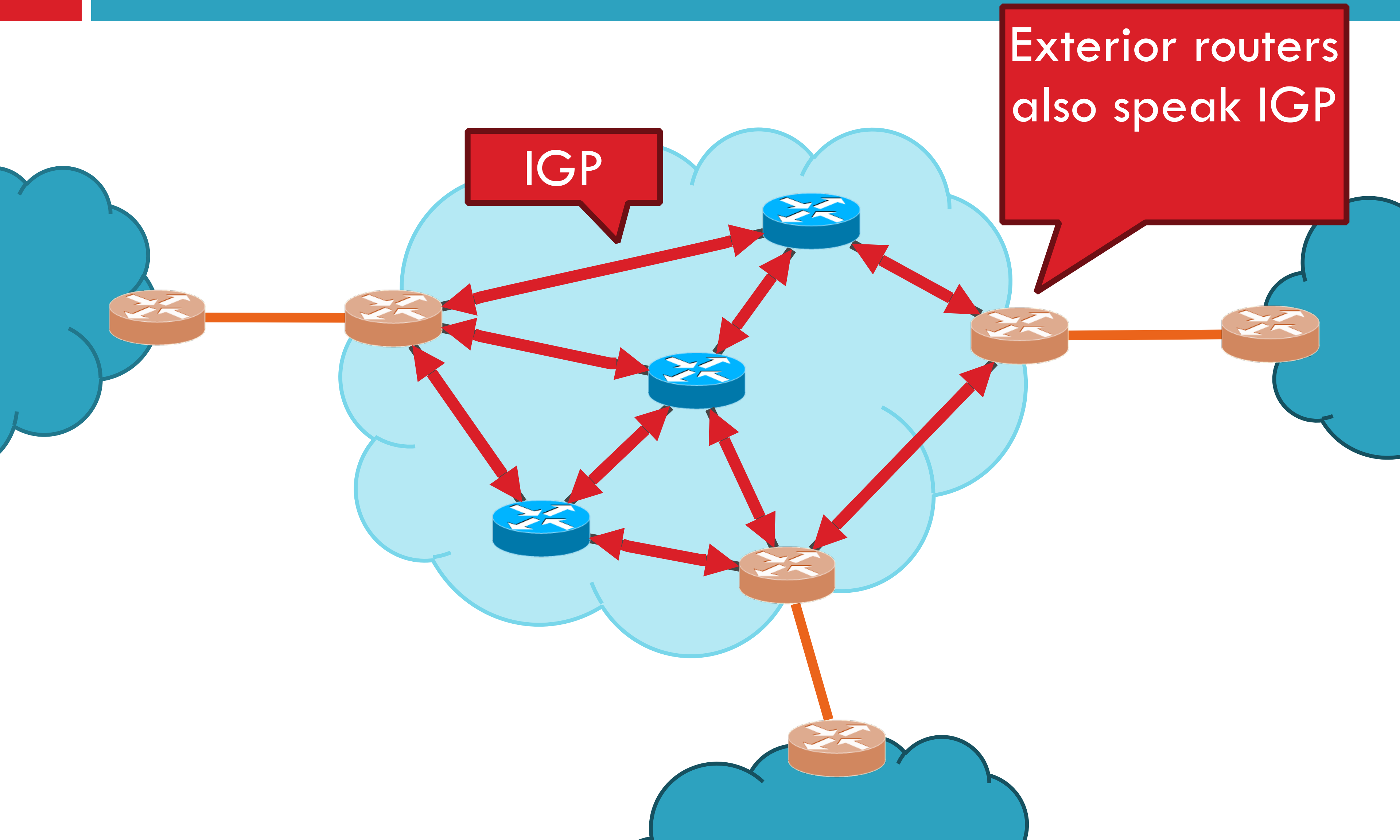
Peer	Don't Peer
Reduce upstream costs Improve end-to-end performance May be the only way to connect to parts of the Internet	You would rather have customers Peers are often competitors Peering agreements require periodic renegotiation

Peering struggles in the ISP world are extremely contentions, agreements are usually confidential

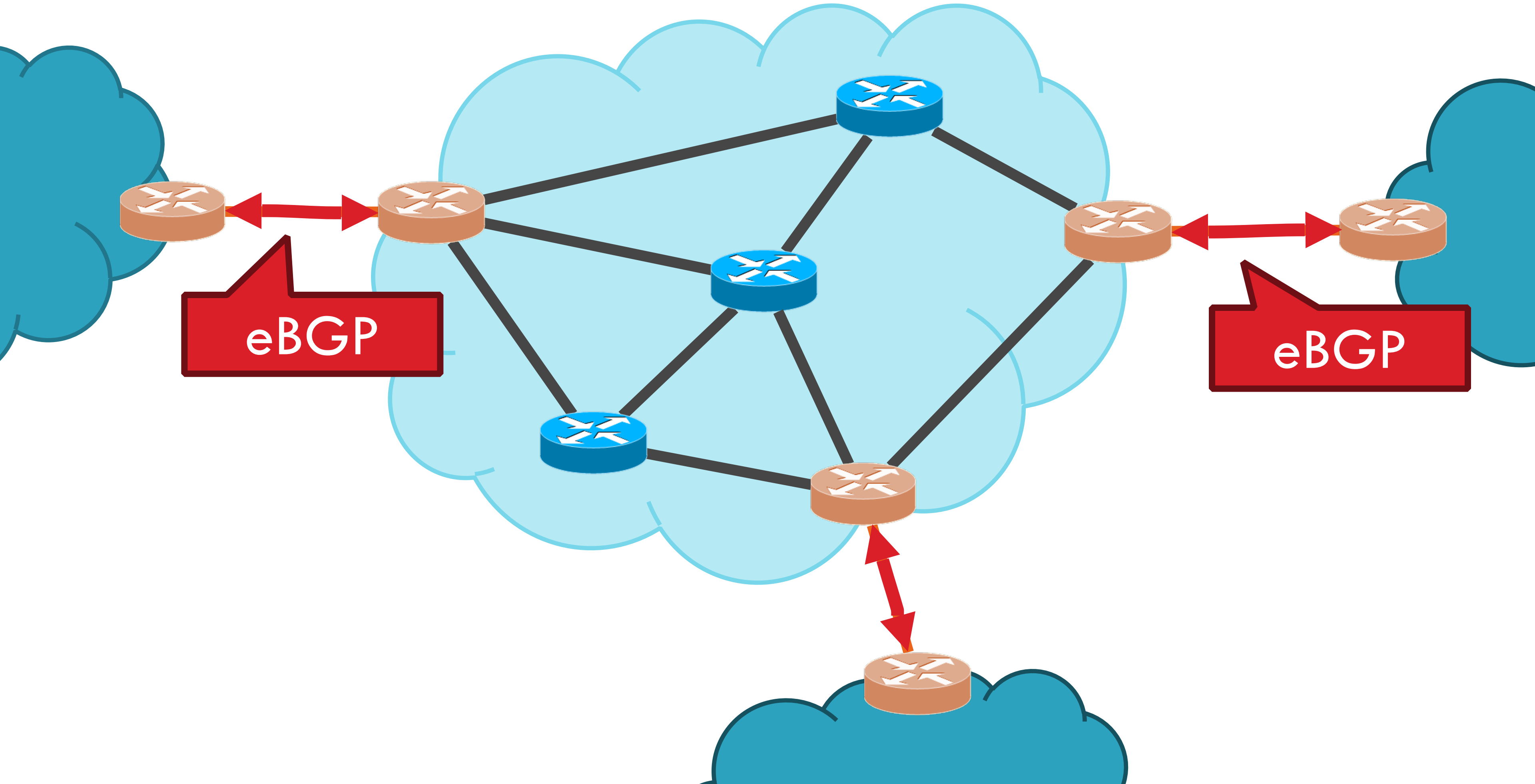
Two Types of BGP Neighbors



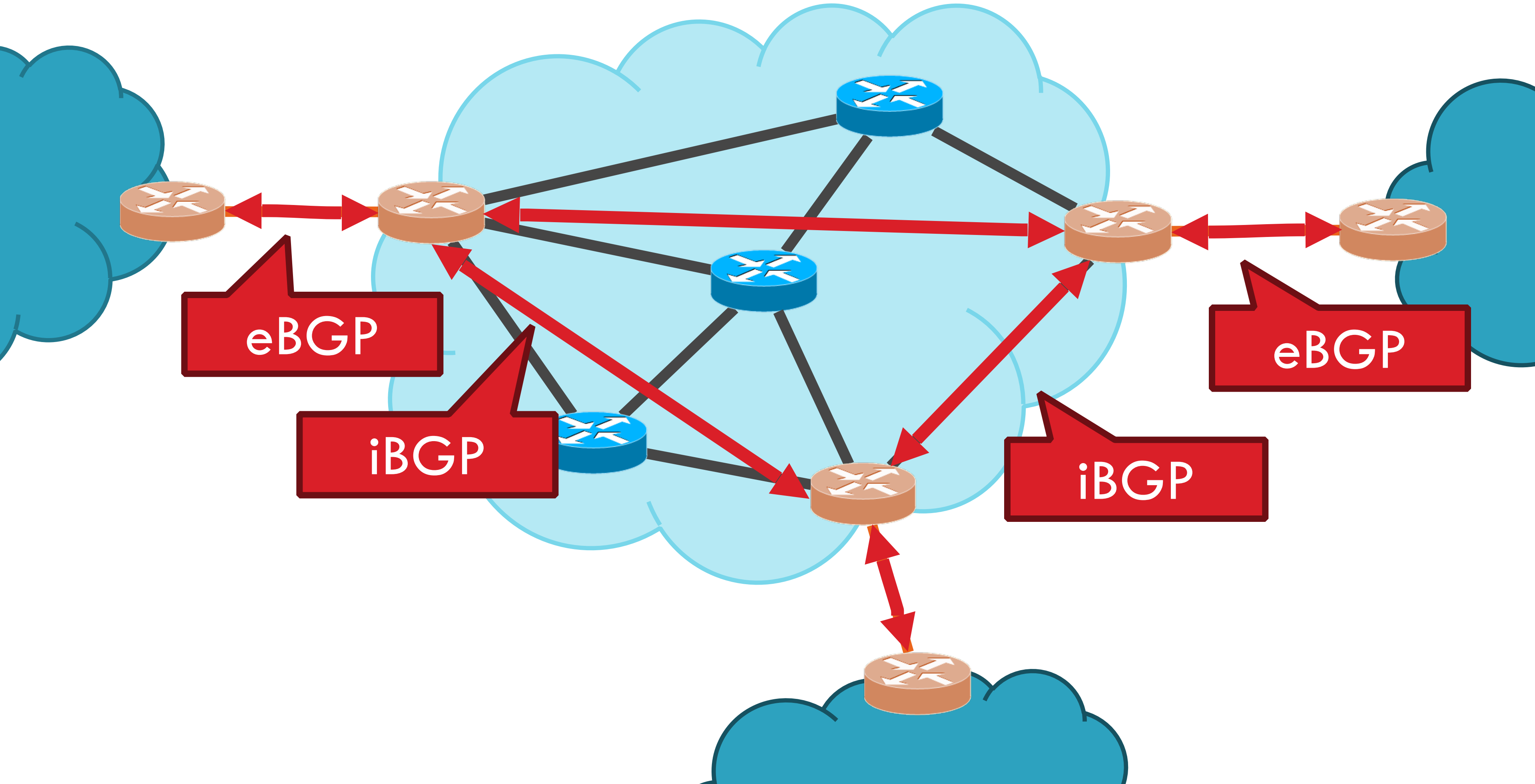
Two Types of BGP Neighbors



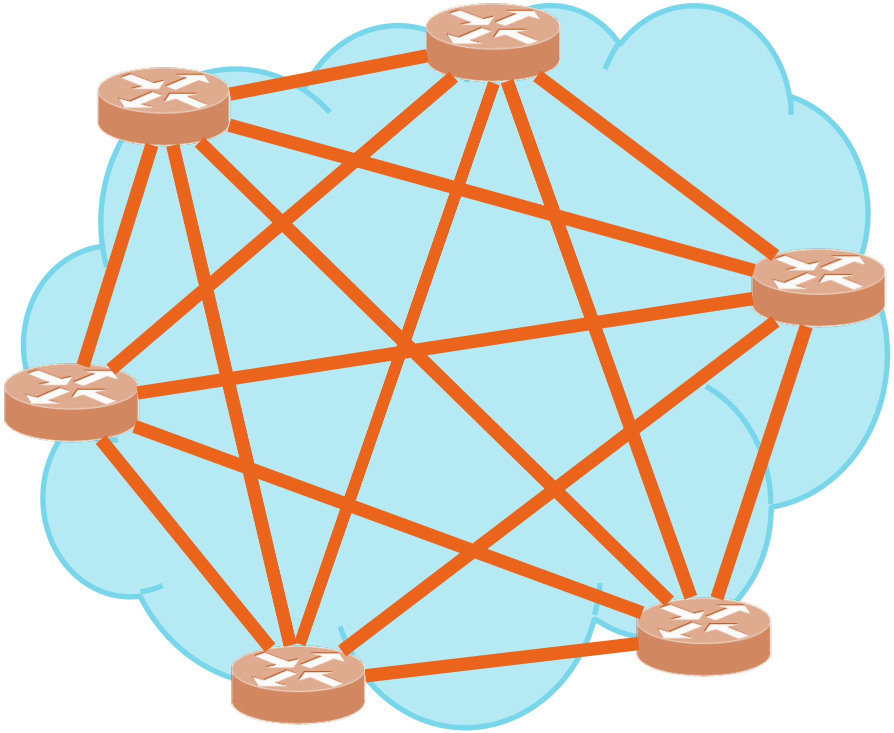
Two Types of BGP Neighbors



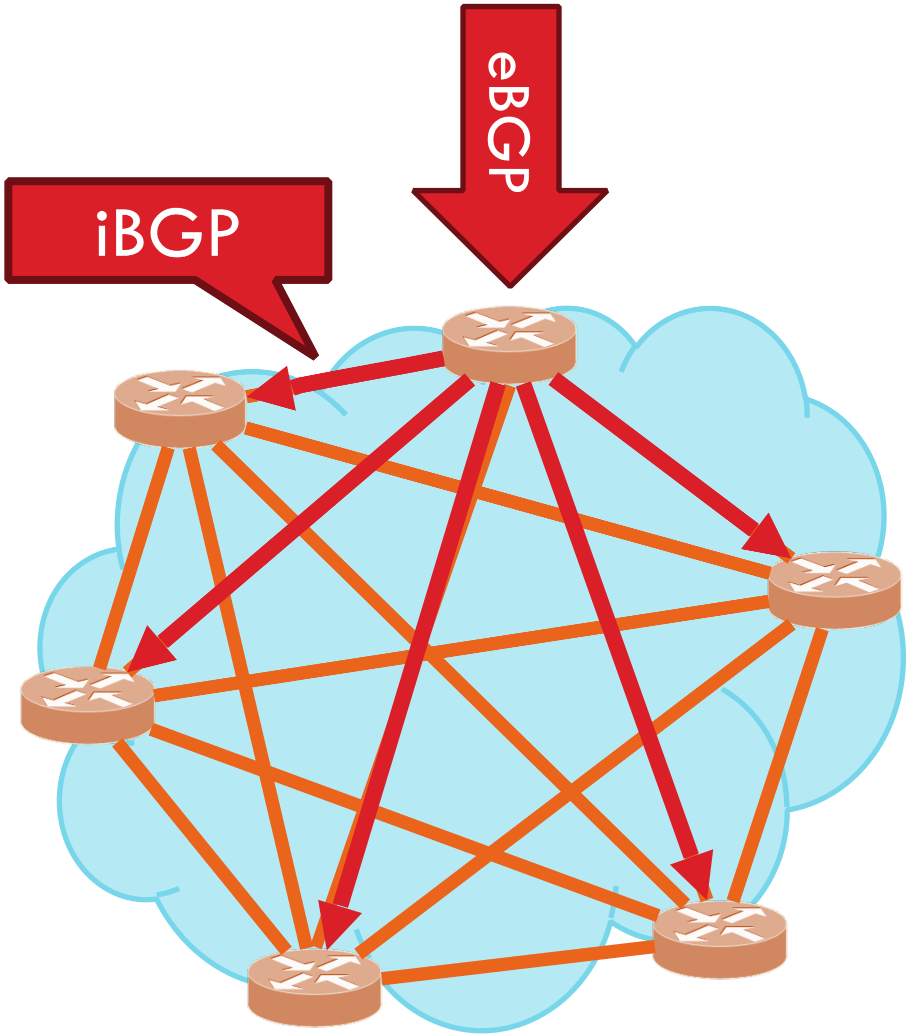
Two Types of BGP Neighbors



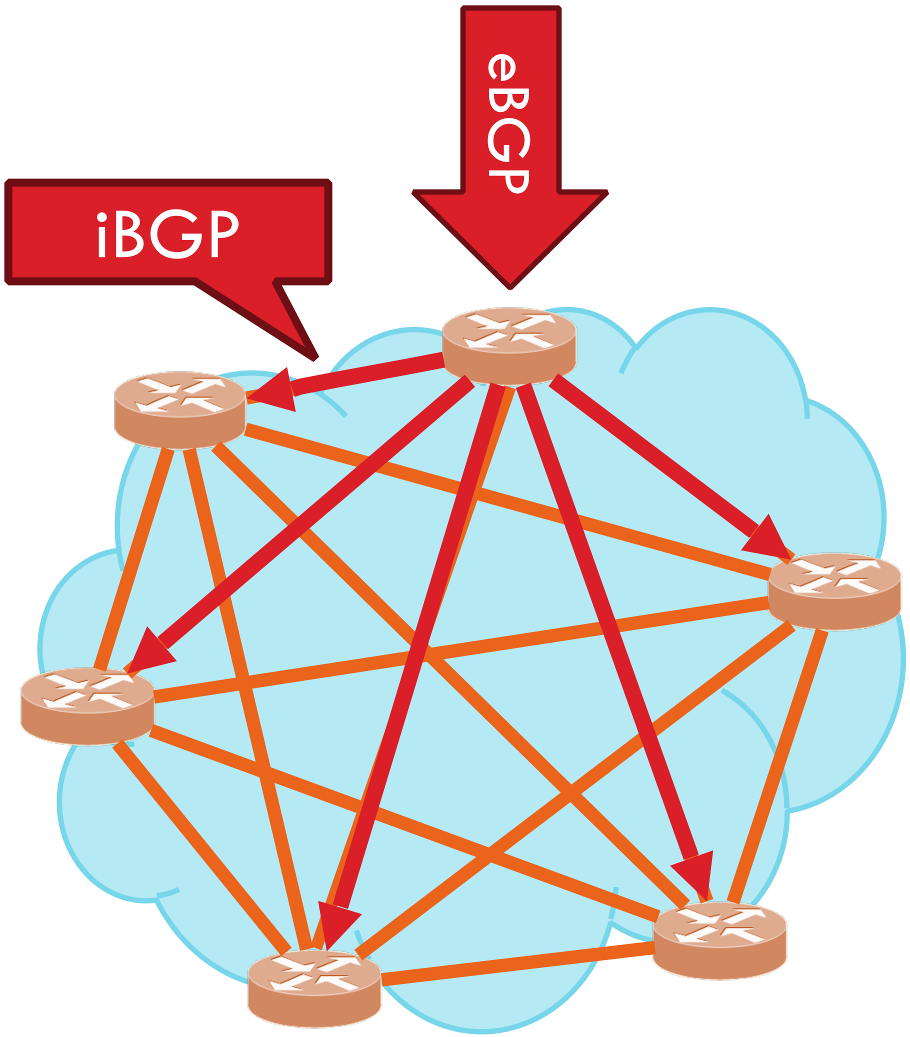
Full iBGP Meshes



Full iBGP Meshes

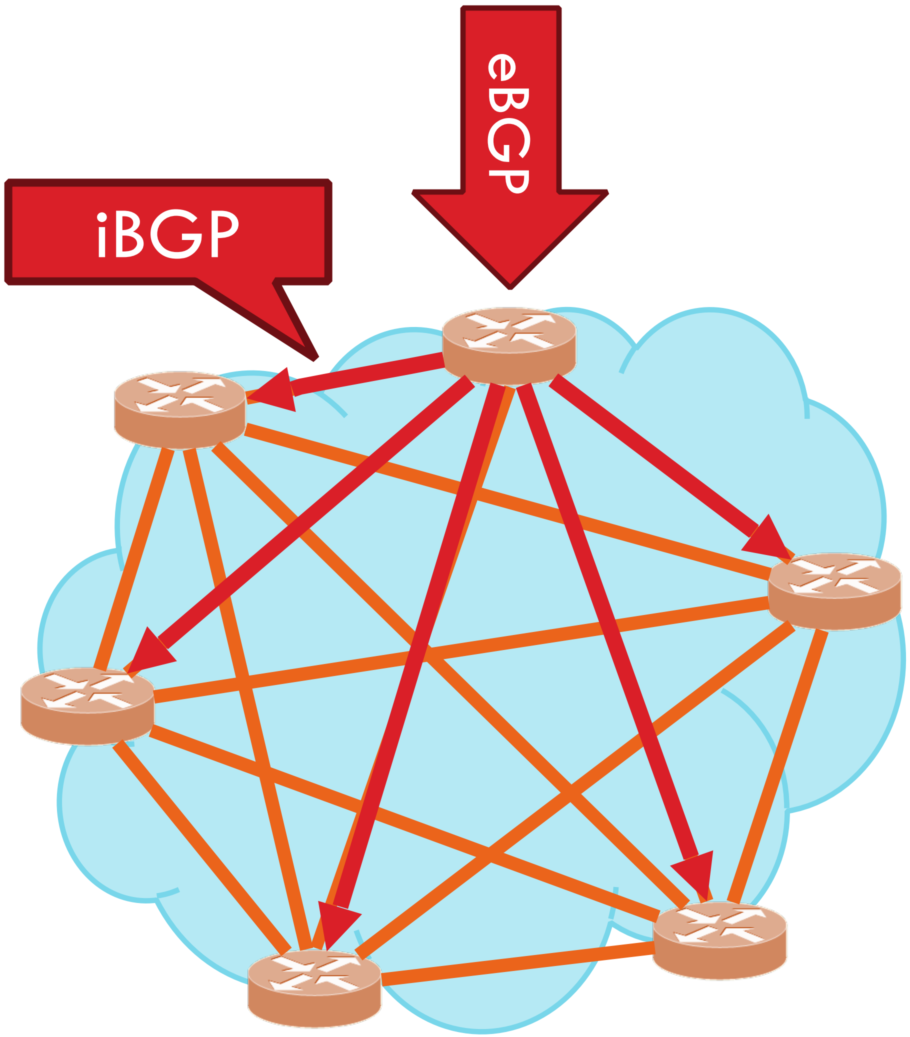


Full iBGP Meshes



- Question: why do we need iBGP?
  - OSPF does not include BGP policy info
  - Prevents routing loops within the AS

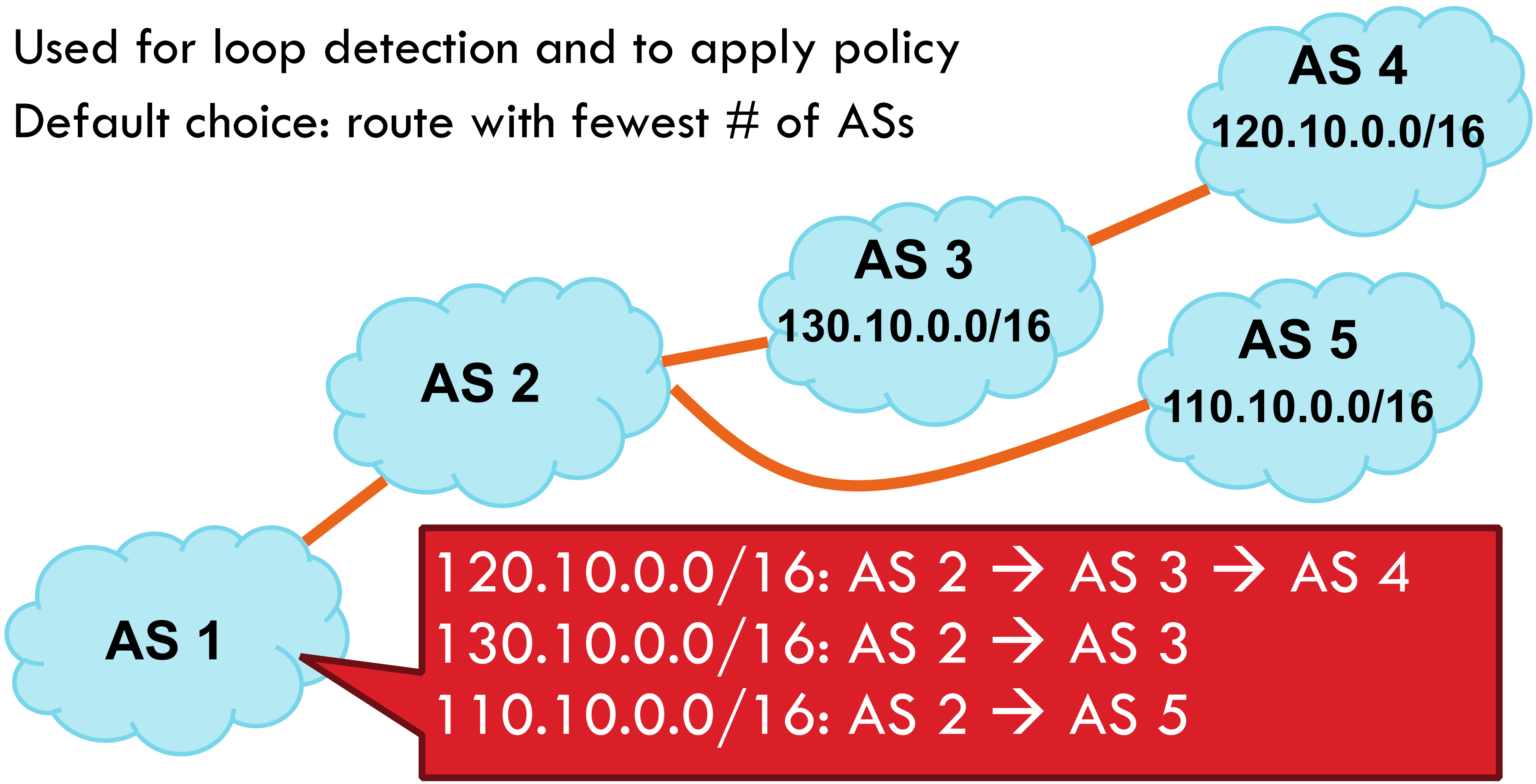
Full iBGP Meshes



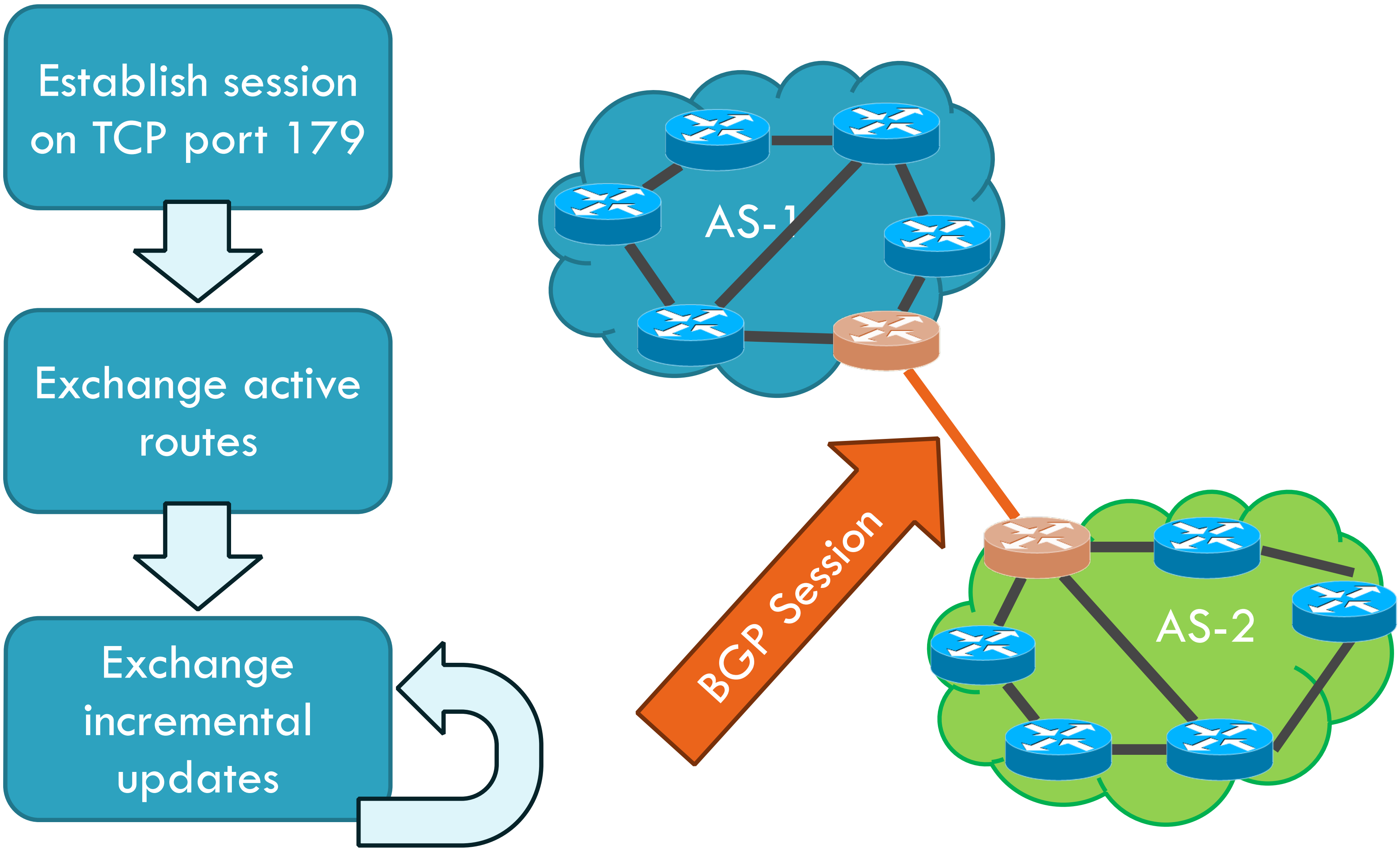
- Question: why do we need iBGP?
  - OSPF does not include BGP policy info
  - Prevents routing loops within the AS
- iBGP updates do not trigger announcements

Path Vector Protocol

- AS-path: sequence of ASs a route traverses
  - Like distance vector, plus additional information
- Used for loop detection and to apply policy
- Default choice: route with fewest # of ASs



BGP Operations (Simplified)



Four Types of BGP Messages

- Open: Establish a peering session.
- Keep Alive: Handshake at regular intervals.
- Notification: Shuts down a peering session.
- Update: Announce new routes or withdraw previously announced routes.

Four Types of BGP Messages

- Open: Establish a peering session.
- Keep Alive: Handshake at regular intervals.
- Notification: Shuts down a peering session.
- Update: Announce new routes or withdraw previously announced routes.

announcement = IP prefix + attributes values

BGP Attributes

- Attributes used to select "best" path
  - LocalPref
    - Local preference policy to choose most preferred route
    - Overrides default fewest AS behavior

BGP Attributes

- Attributes used to select “best” path
  - LocalPref
    - Local preference policy to choose most preferred route
    - Overrides default fewest AS behavior
  - Multi-exit Discriminator (MED)
    - Specifies path for external traffic destined for an internal network
    - Chooses peering point for your network

BGP Attributes

- Attributes used to select “best” path
  - LocalPref
    - Local preference policy to choose most preferred route
    - Overrides default fewest AS behavior
  - Multi-exit Discriminator (MED)
    - Specifies path for external traffic destined for an internal network
    - Chooses peering point for your network
  - Import Rules
    - What route advertisements do I accept?
  - Export Rules
    - Which routes do I forward to whom?

Route Selection Summary

Route Selection Summary

Highest Local Preference	**Enforce relationships**

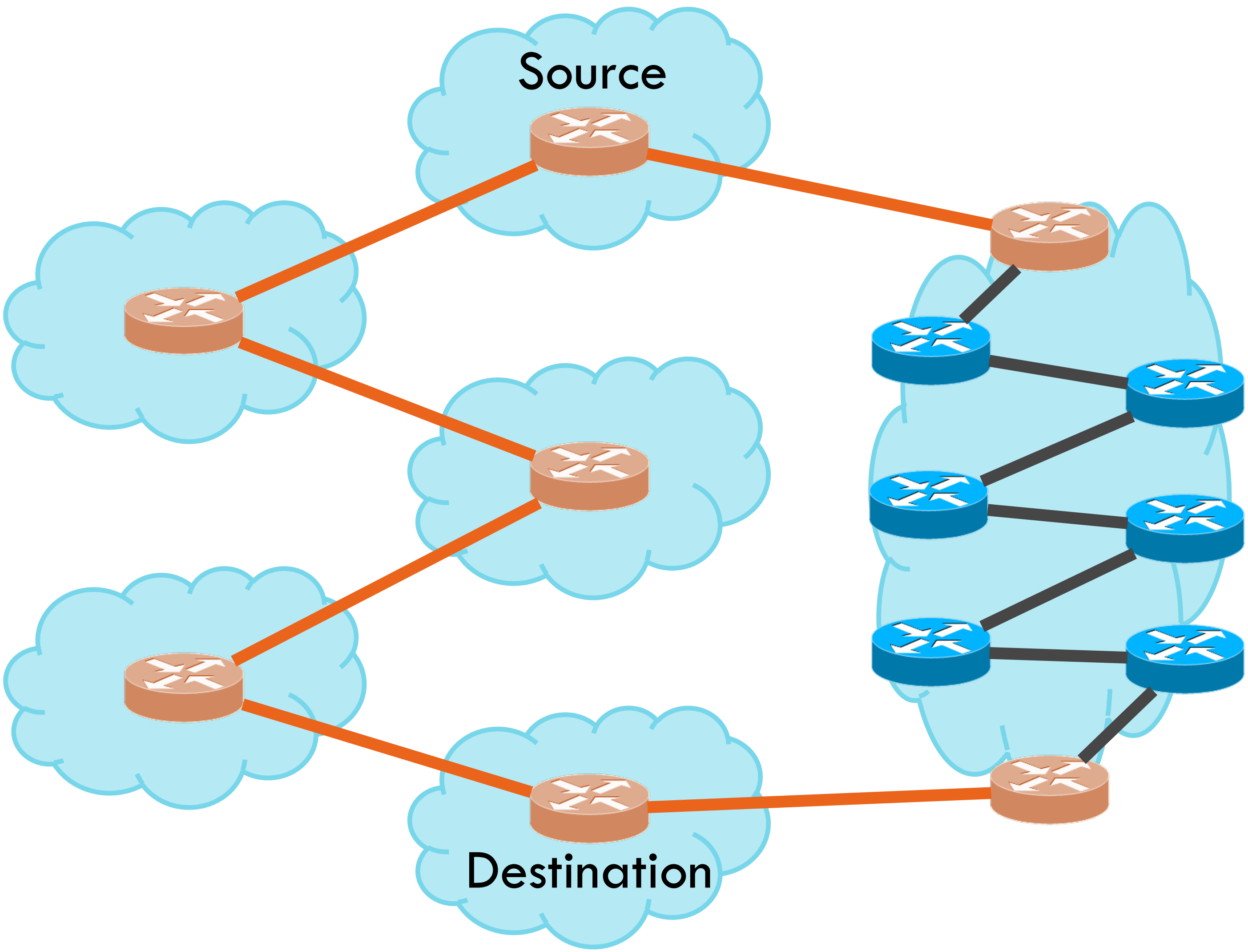
Route Selection Summary

Highest Local Preference	**Enforce relationships**
Shortest AS Path **Lowest MED** **Lowest IGP Cost to BGP Egress**	**Traffic engineering**

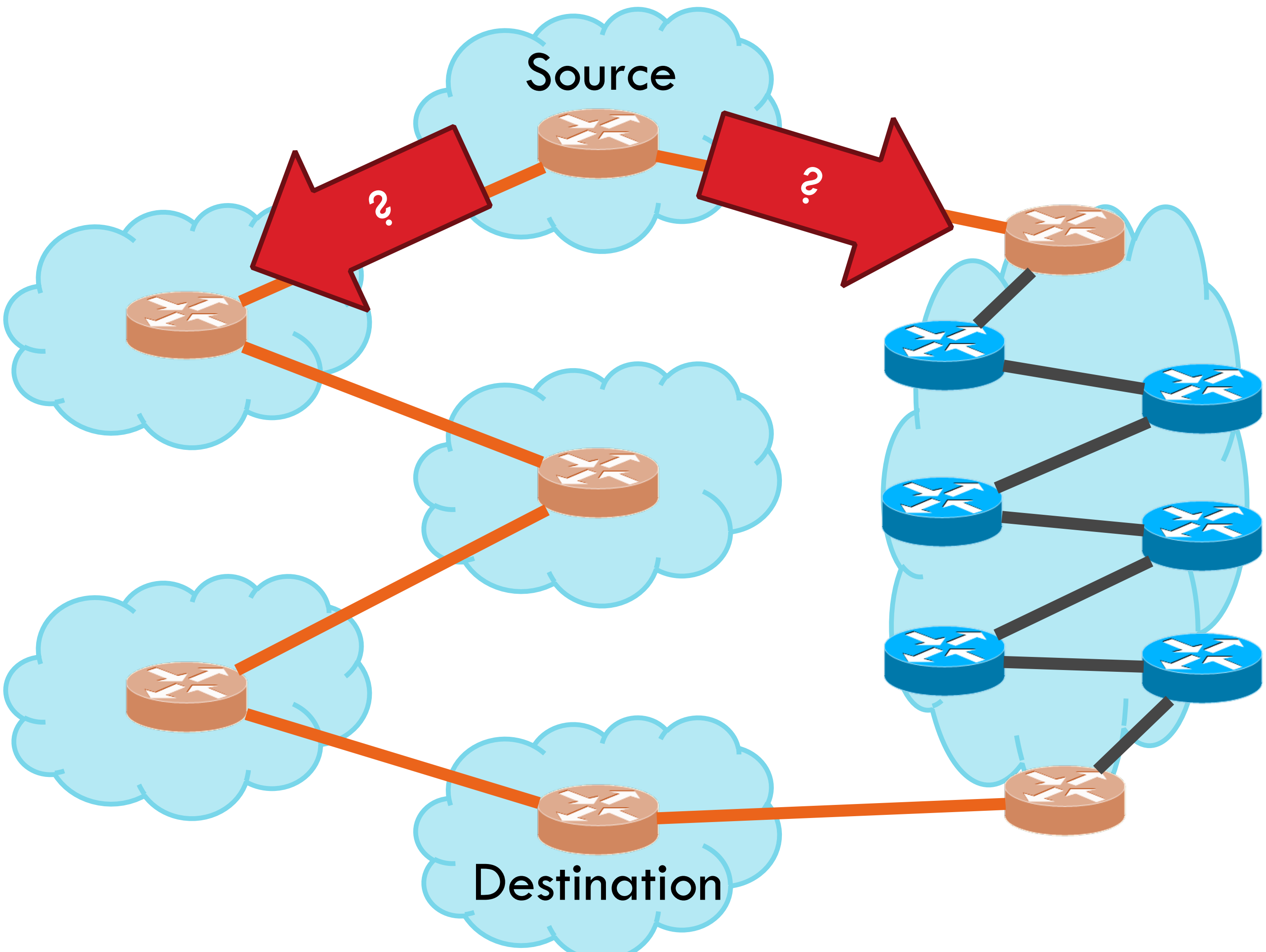
Route Selection Summary

Highest Local Preference	Enforce relationships
Shortest AS Path Lowest MED Lowest IGP Cost to BGP Egress	Traffic engineering
Lowest Router ID	When all else fails, break ties

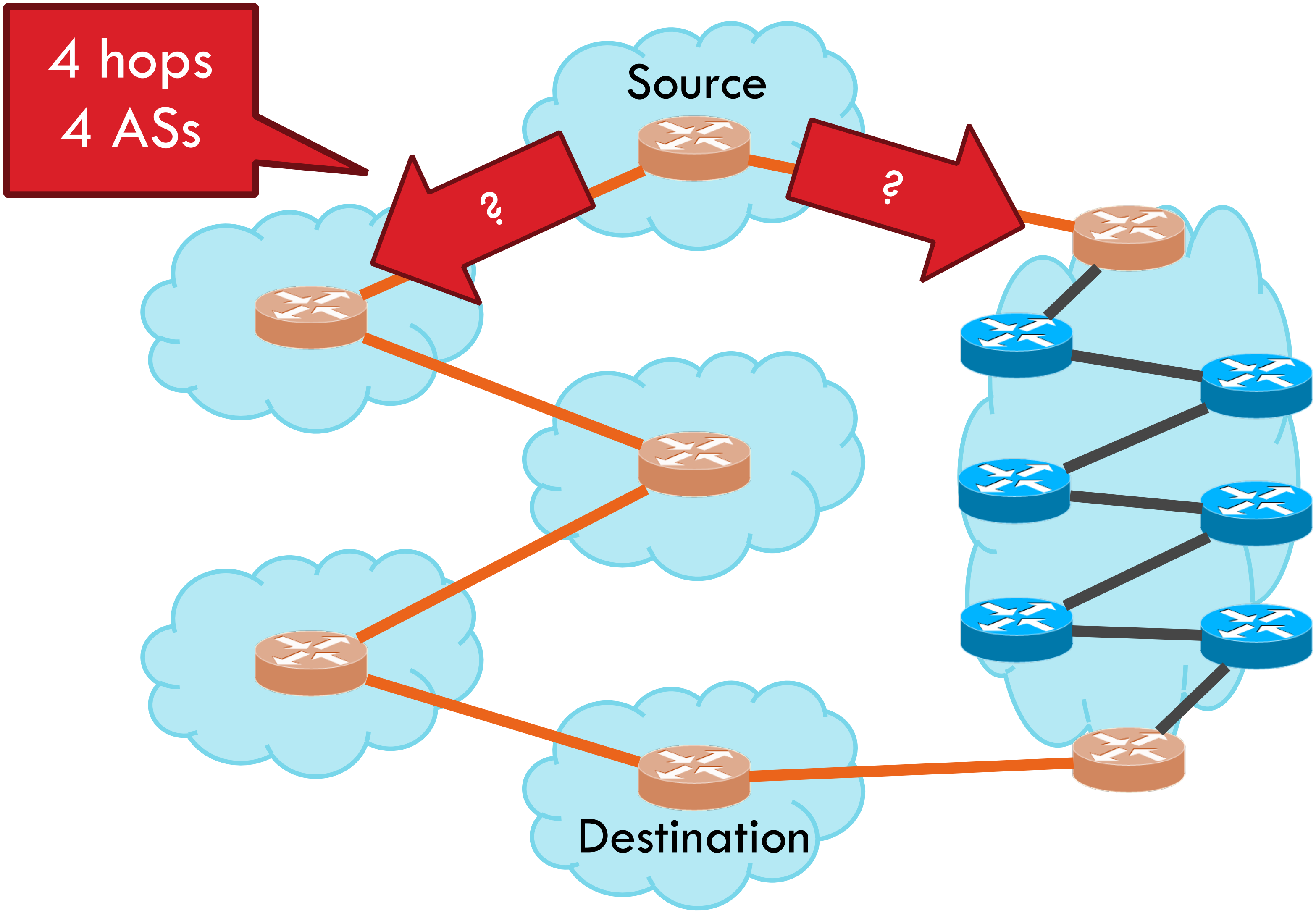
Shortest AS Path != Shortest Path



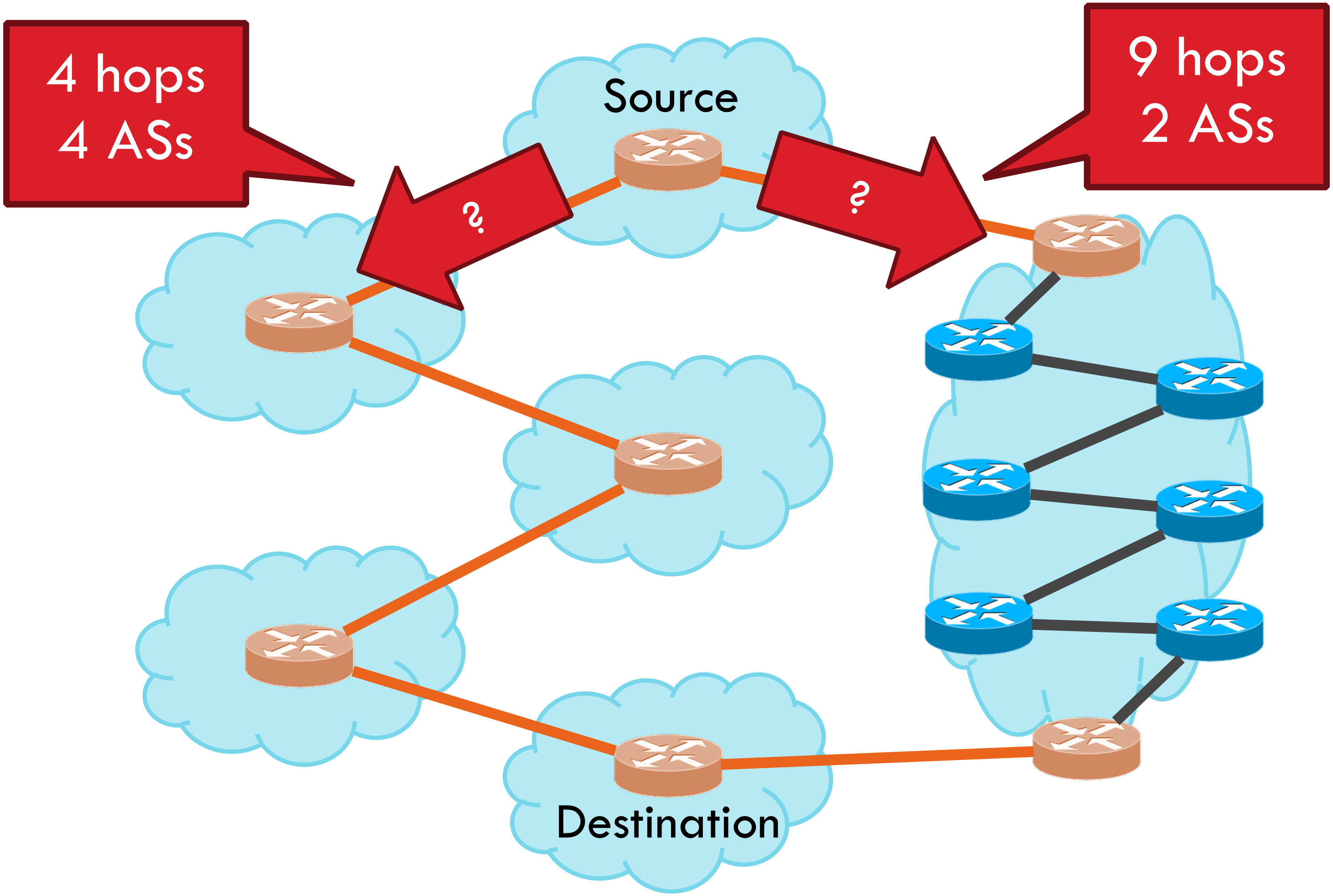
Shortest AS Path != Shortest Path



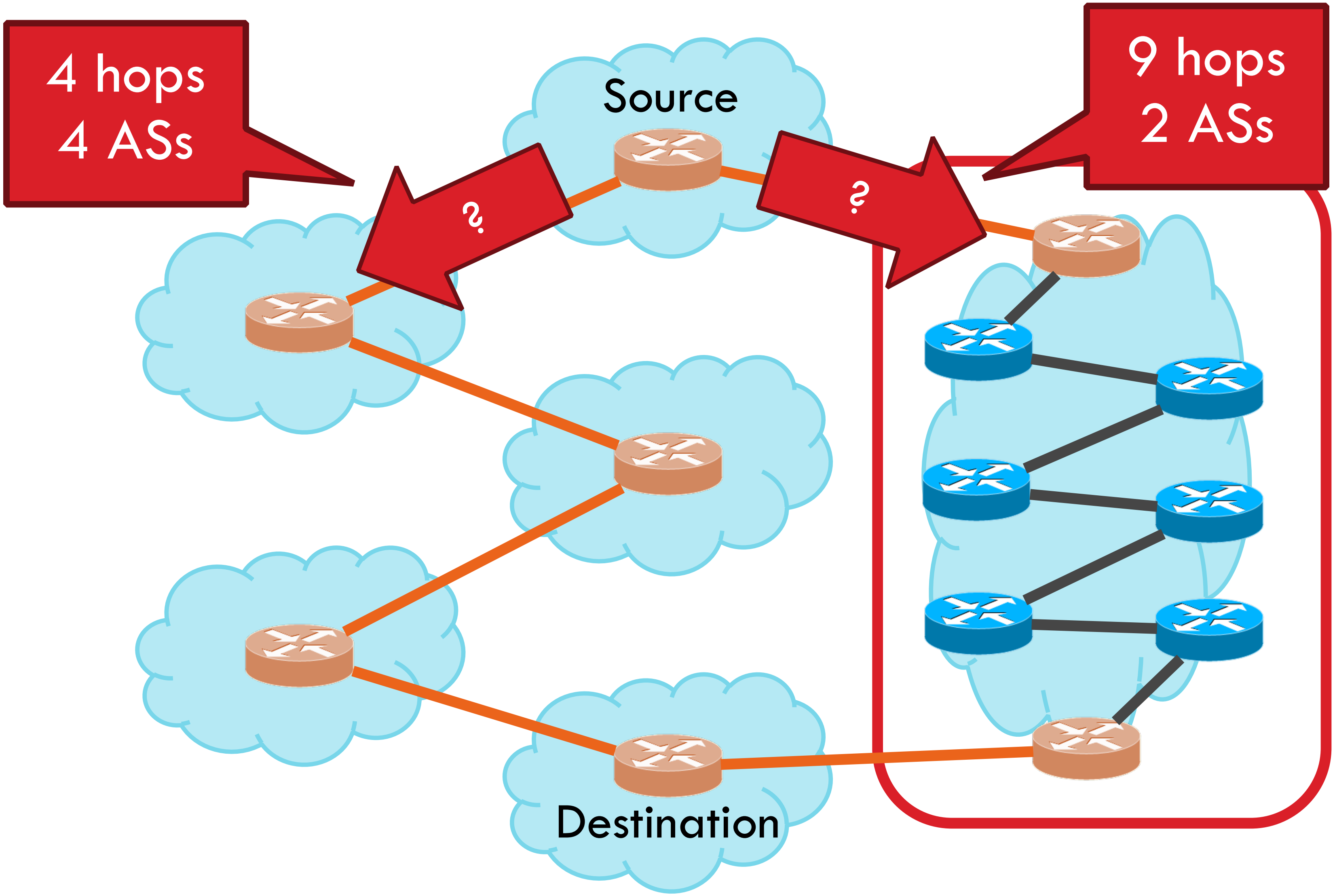
Shortest AS Path != Shortest Path



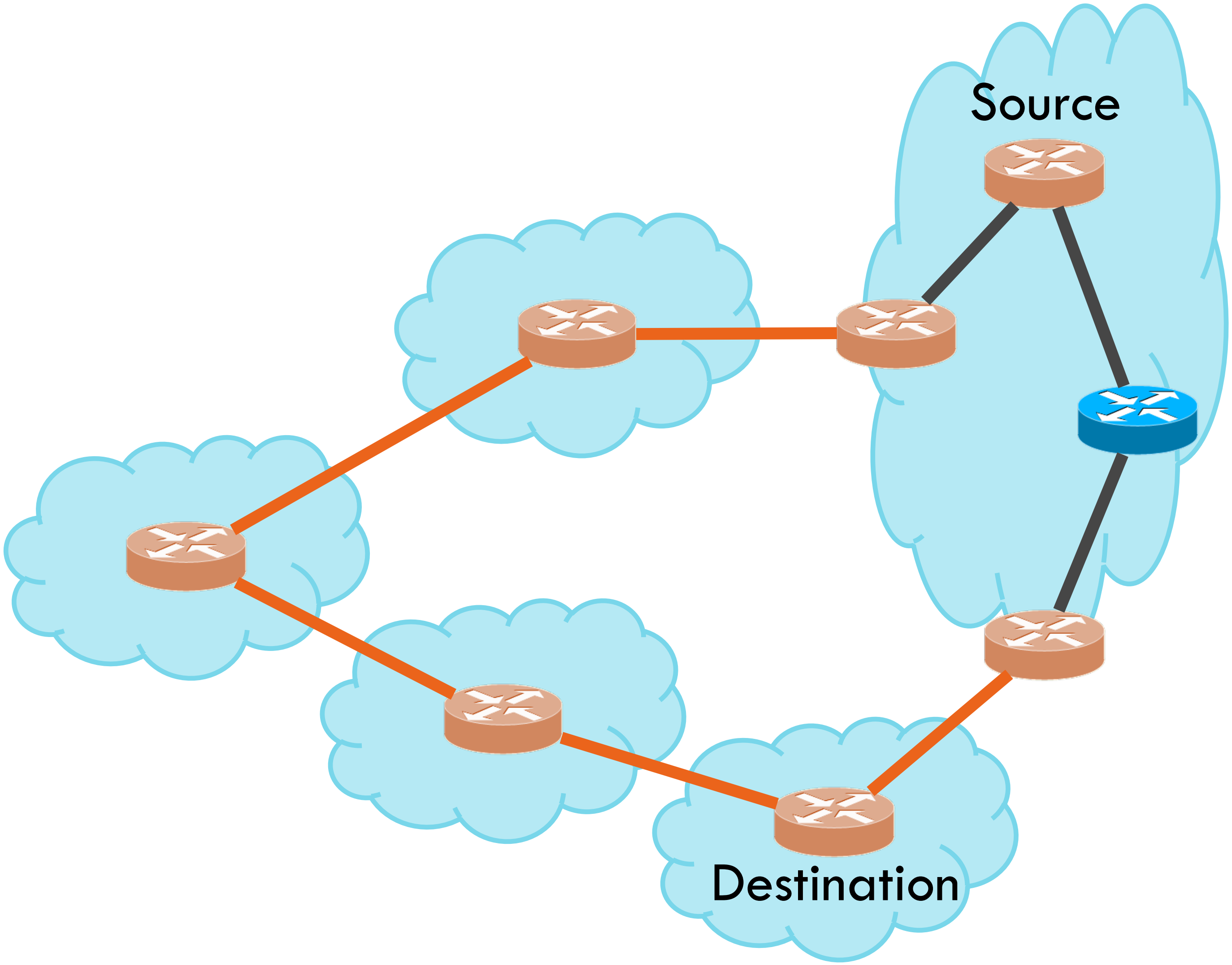
Shortest AS Path != Shortest Path



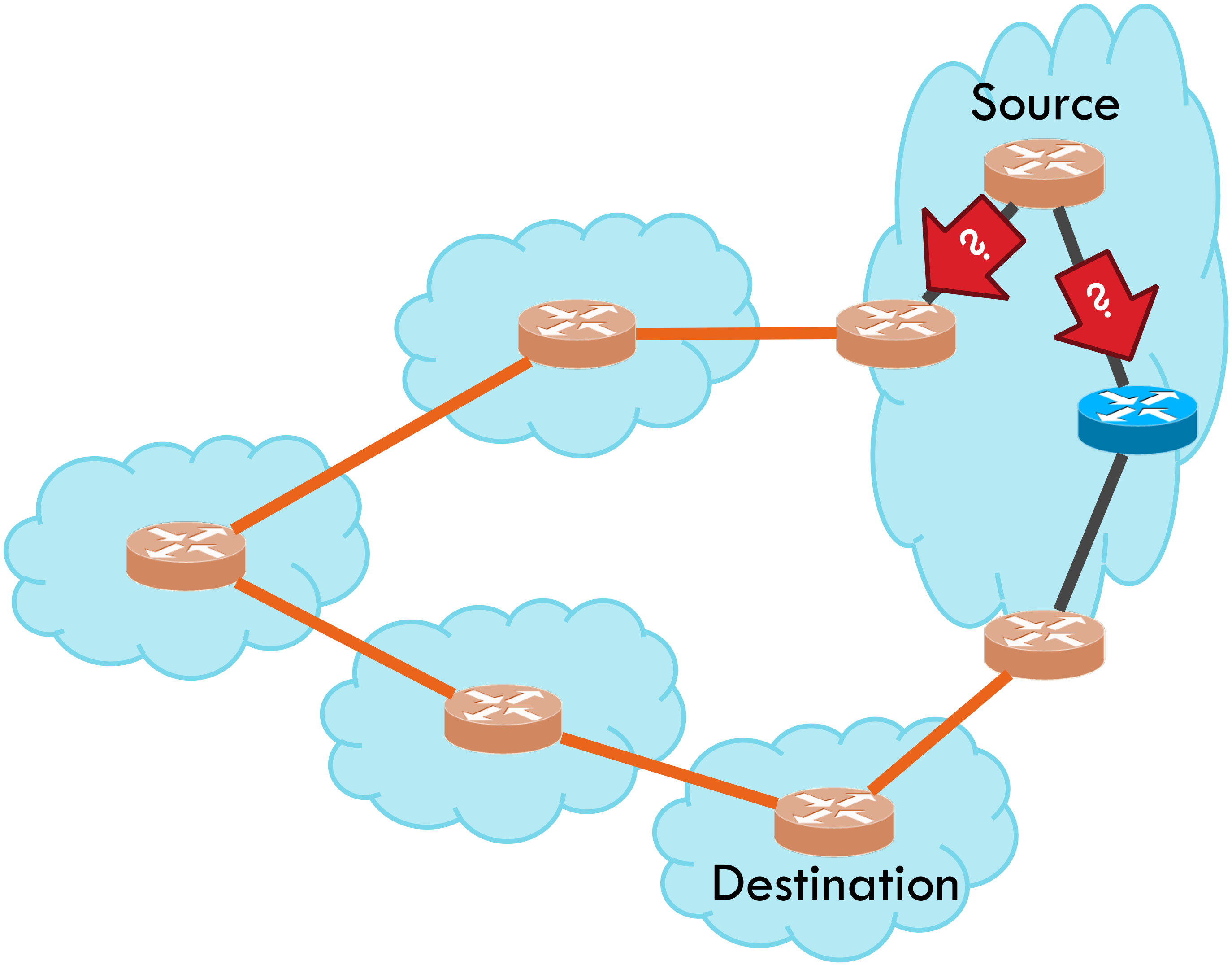
Shortest AS Path != Shortest Path



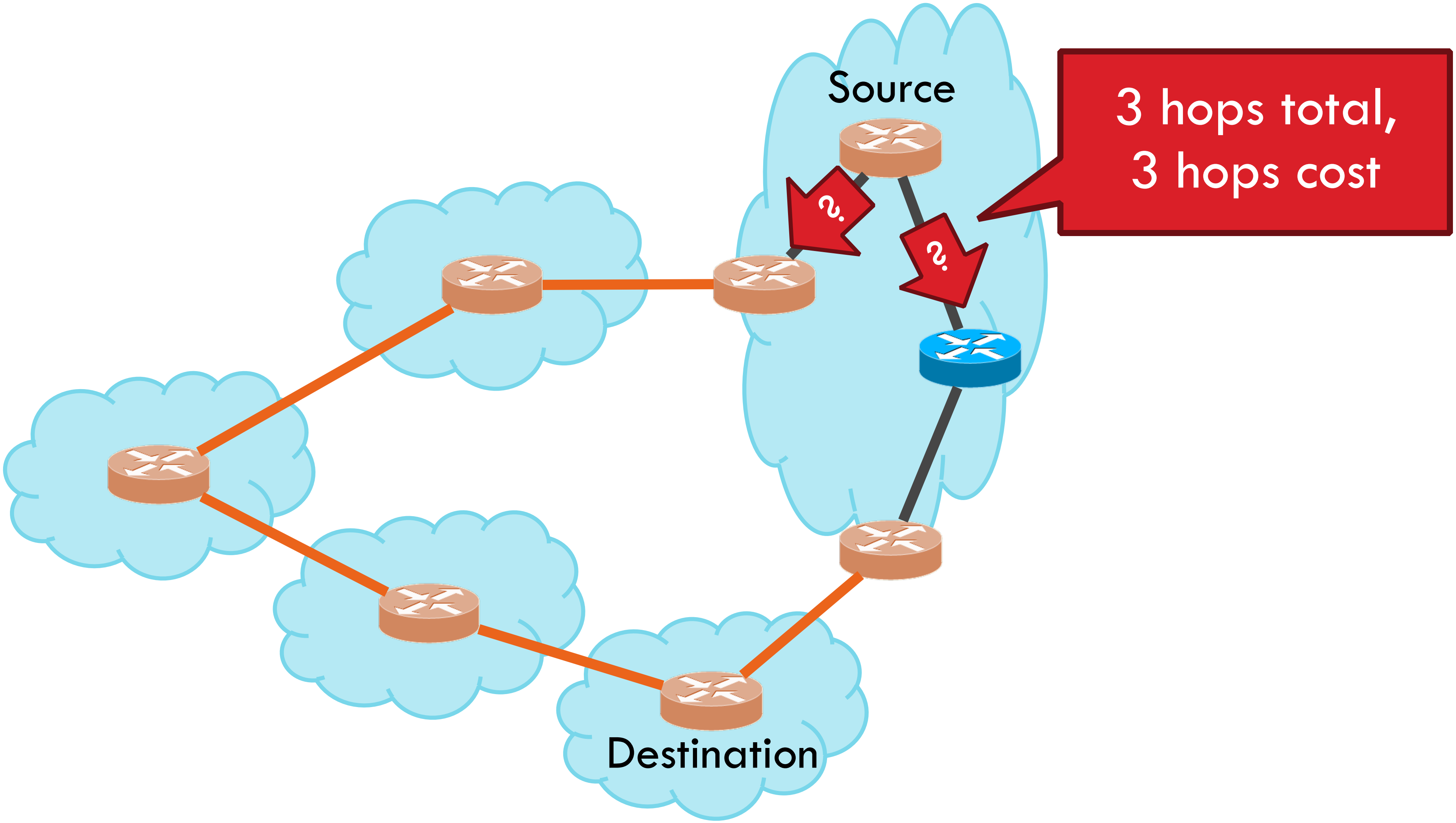
Hot Potato Routing



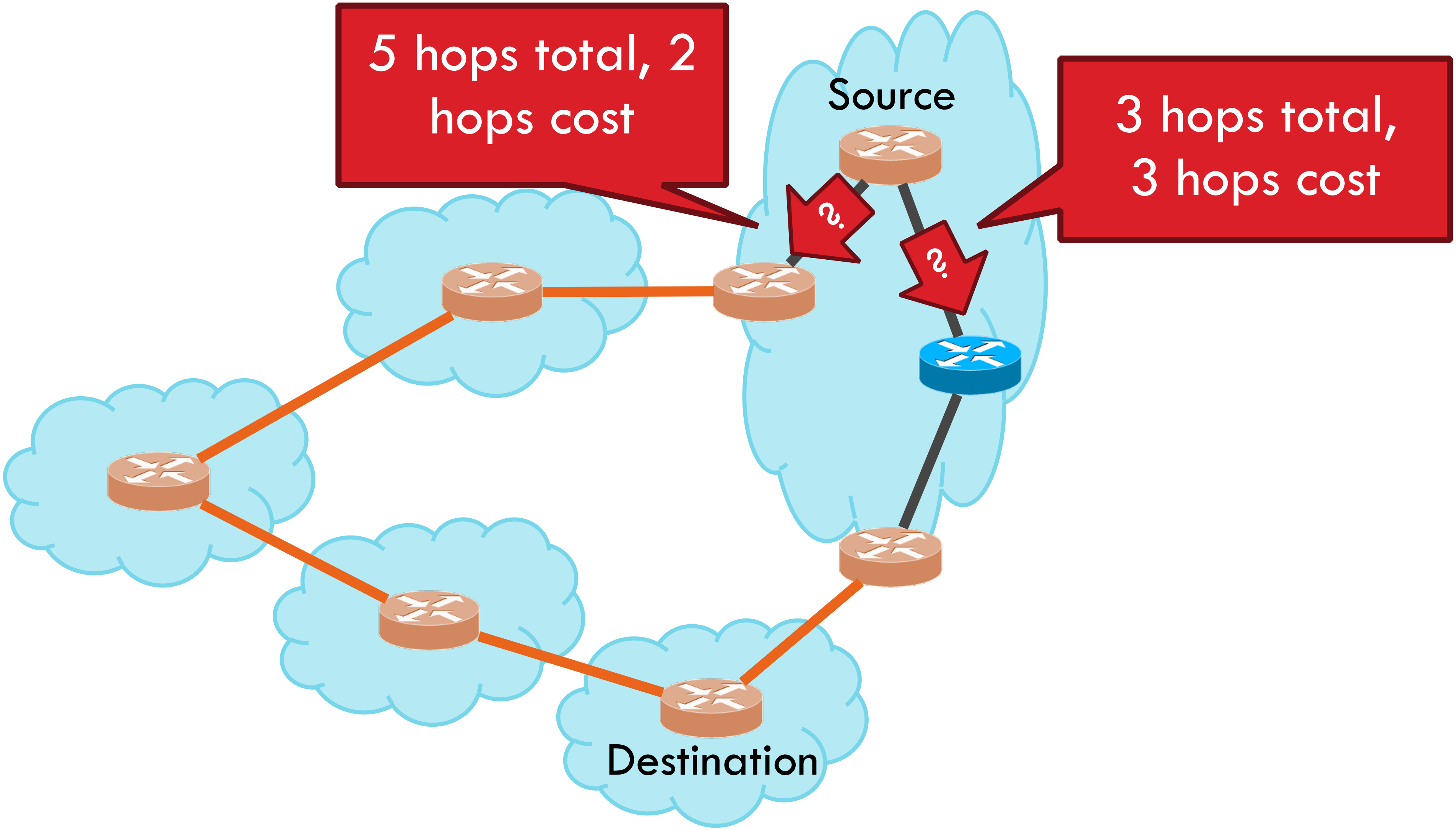
Hot Potato Routing



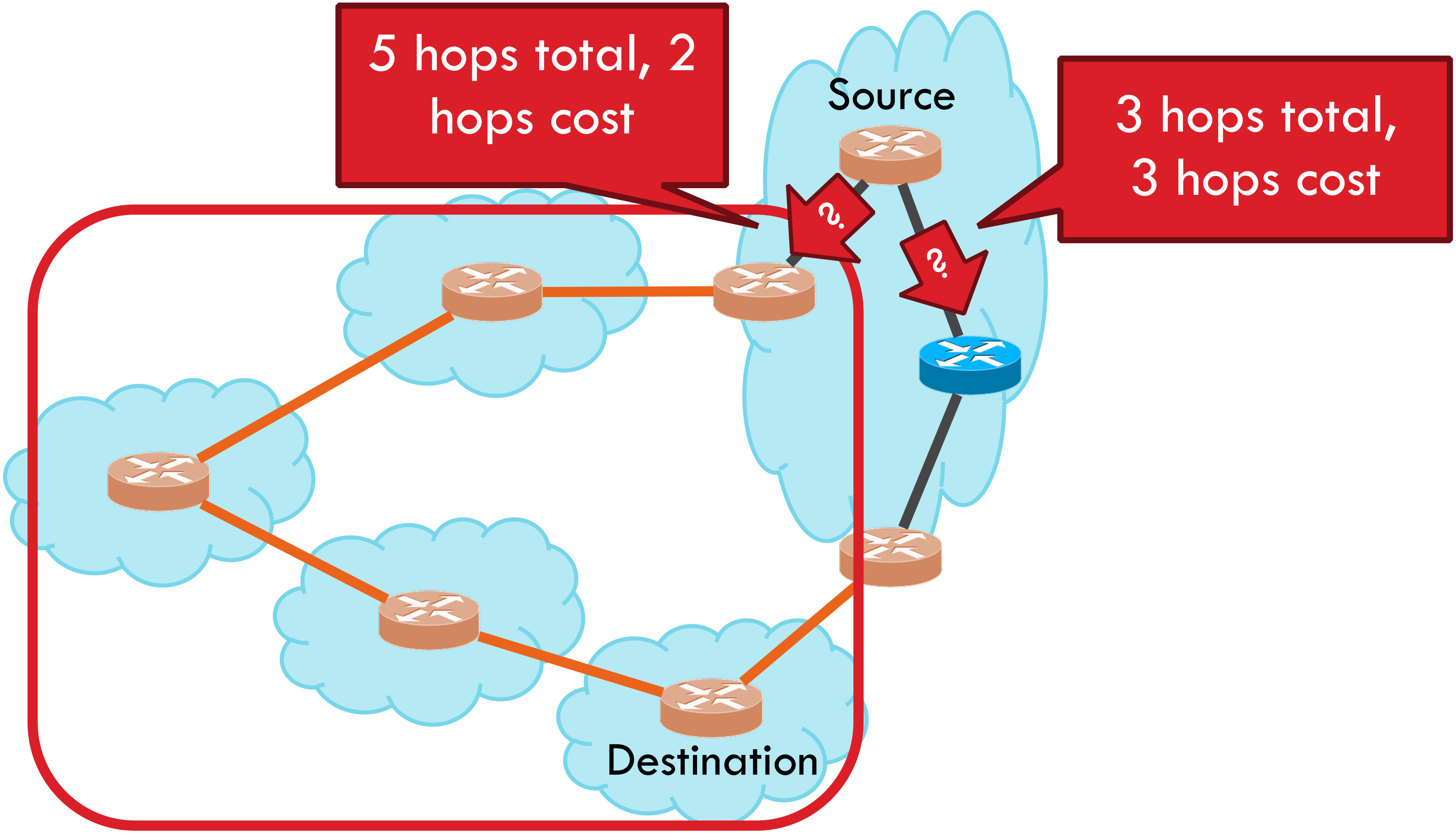
Hot Potato Routing



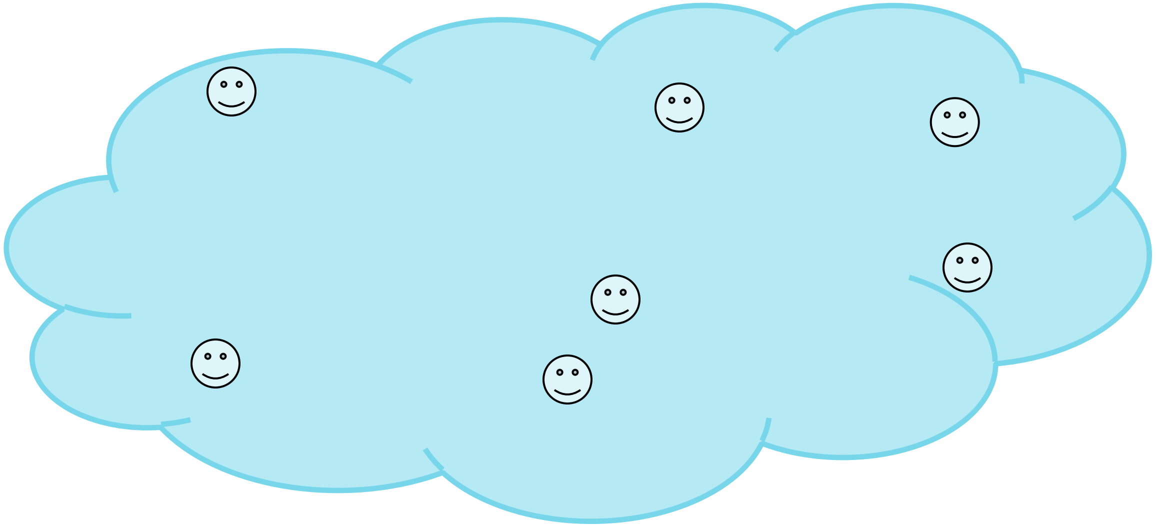
Hot Potato Routing



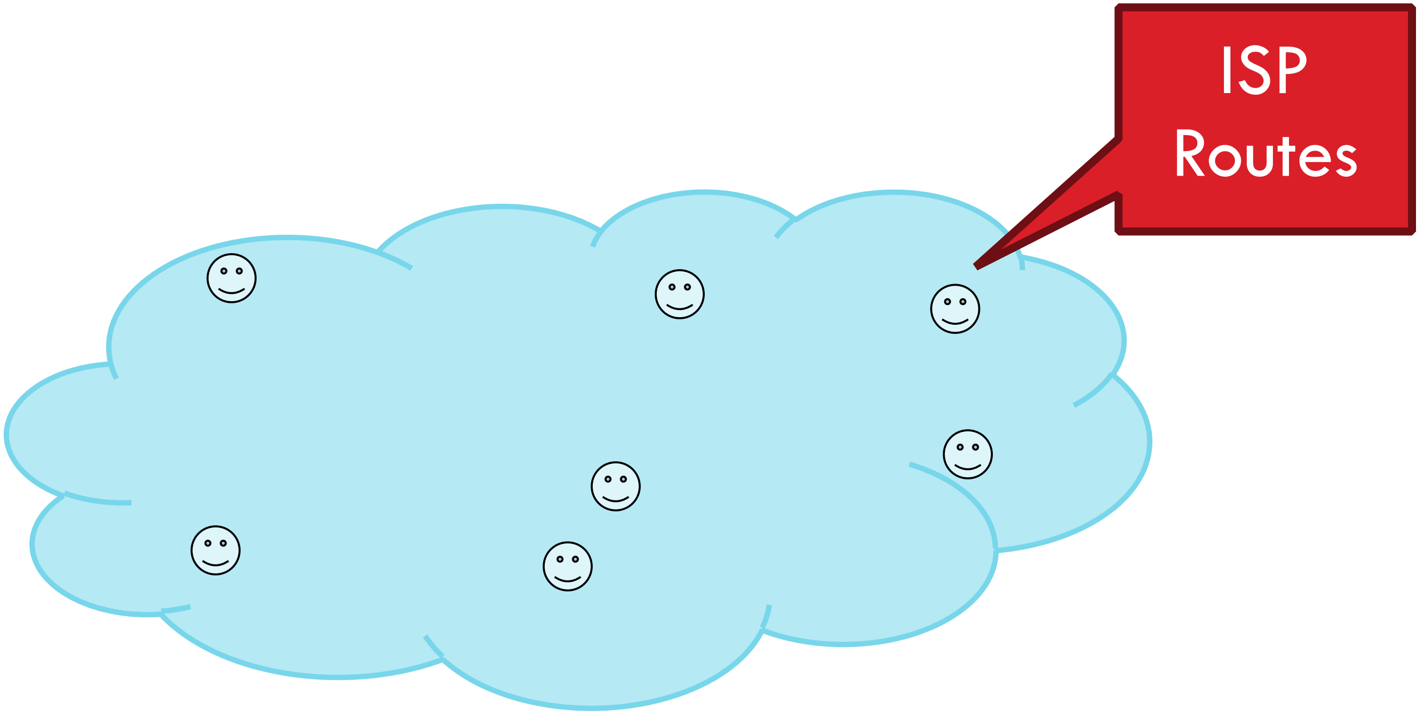
Hot Potato Routing



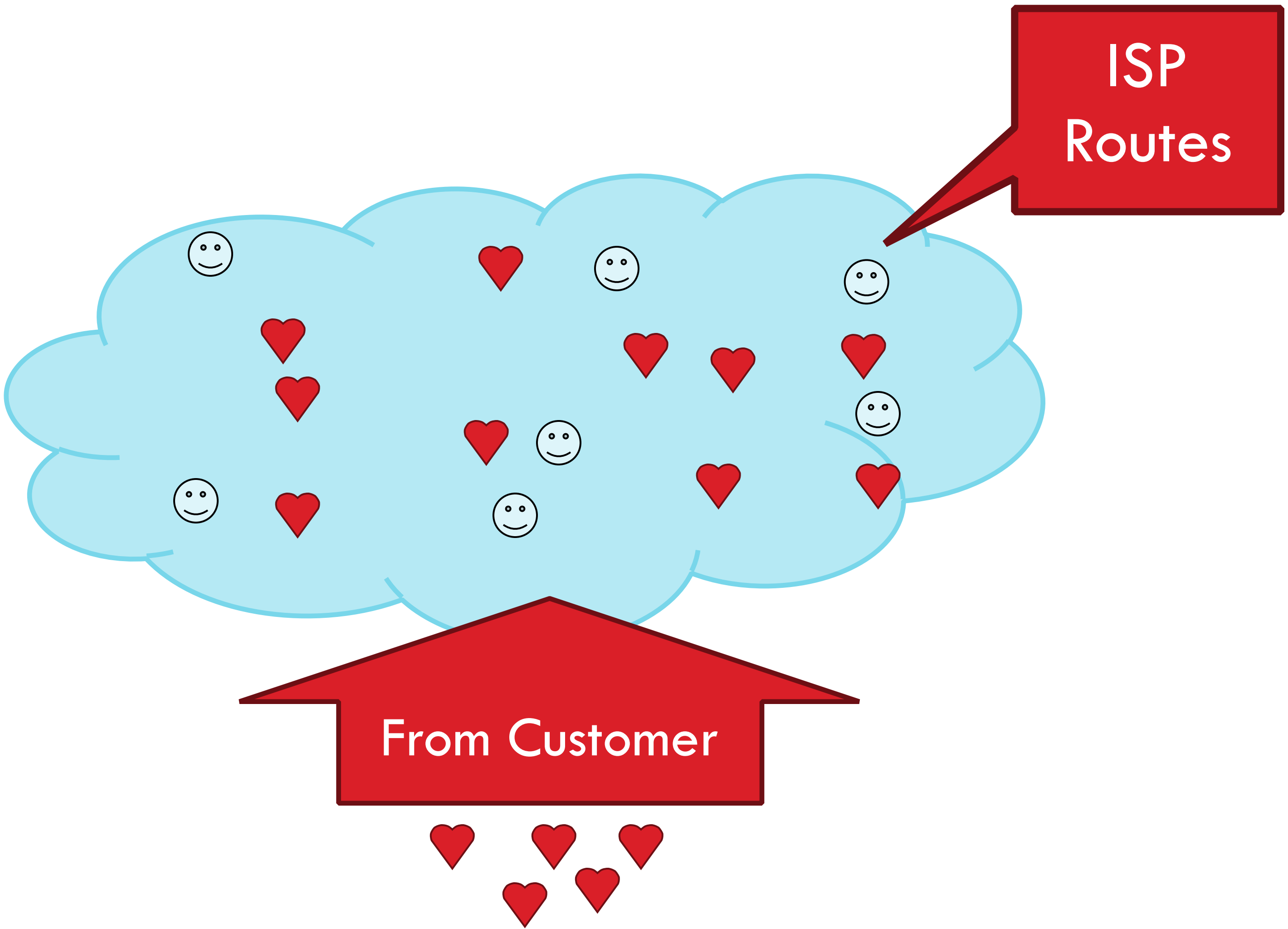
Importing Routes



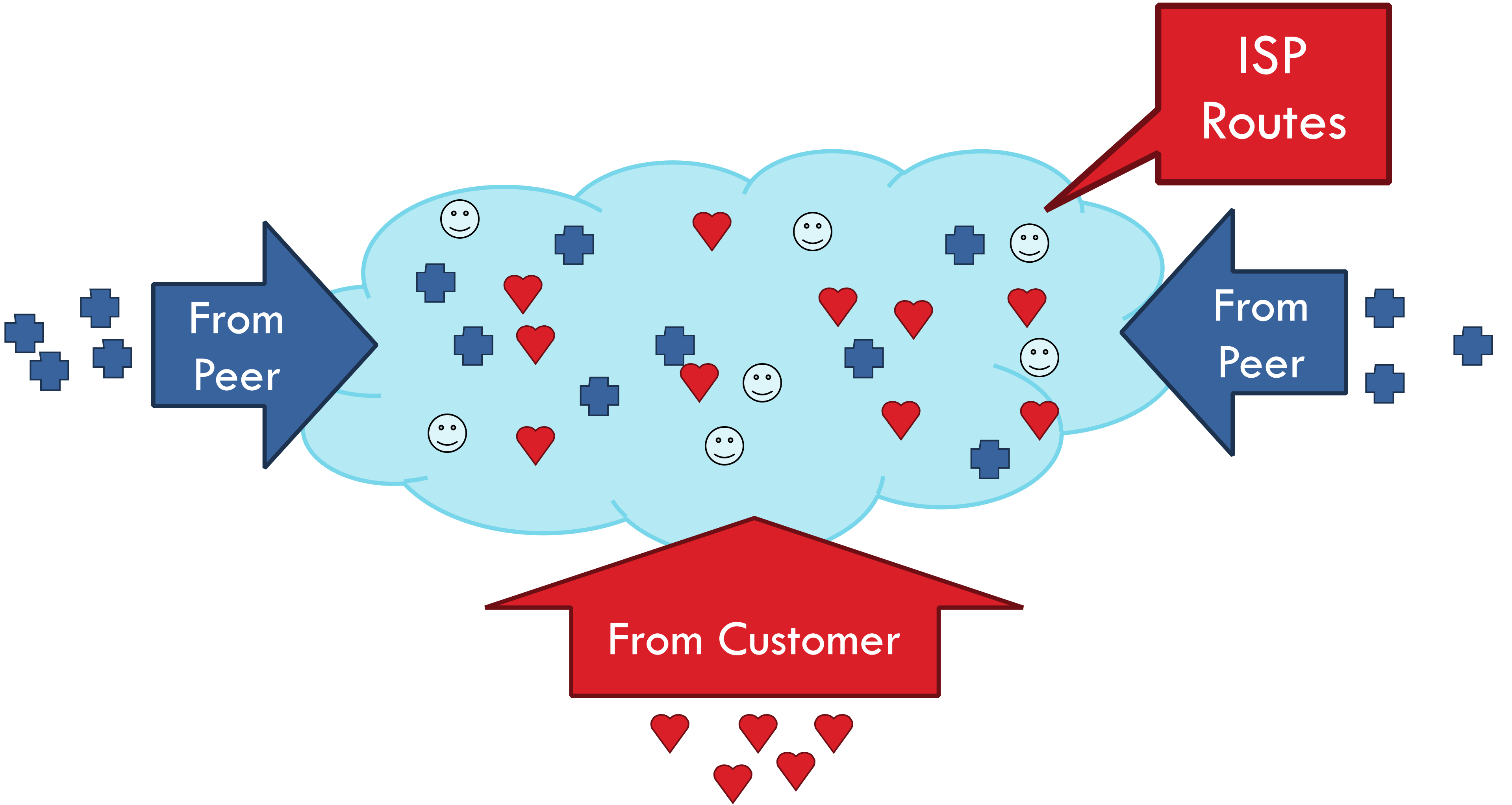
Importing Routes



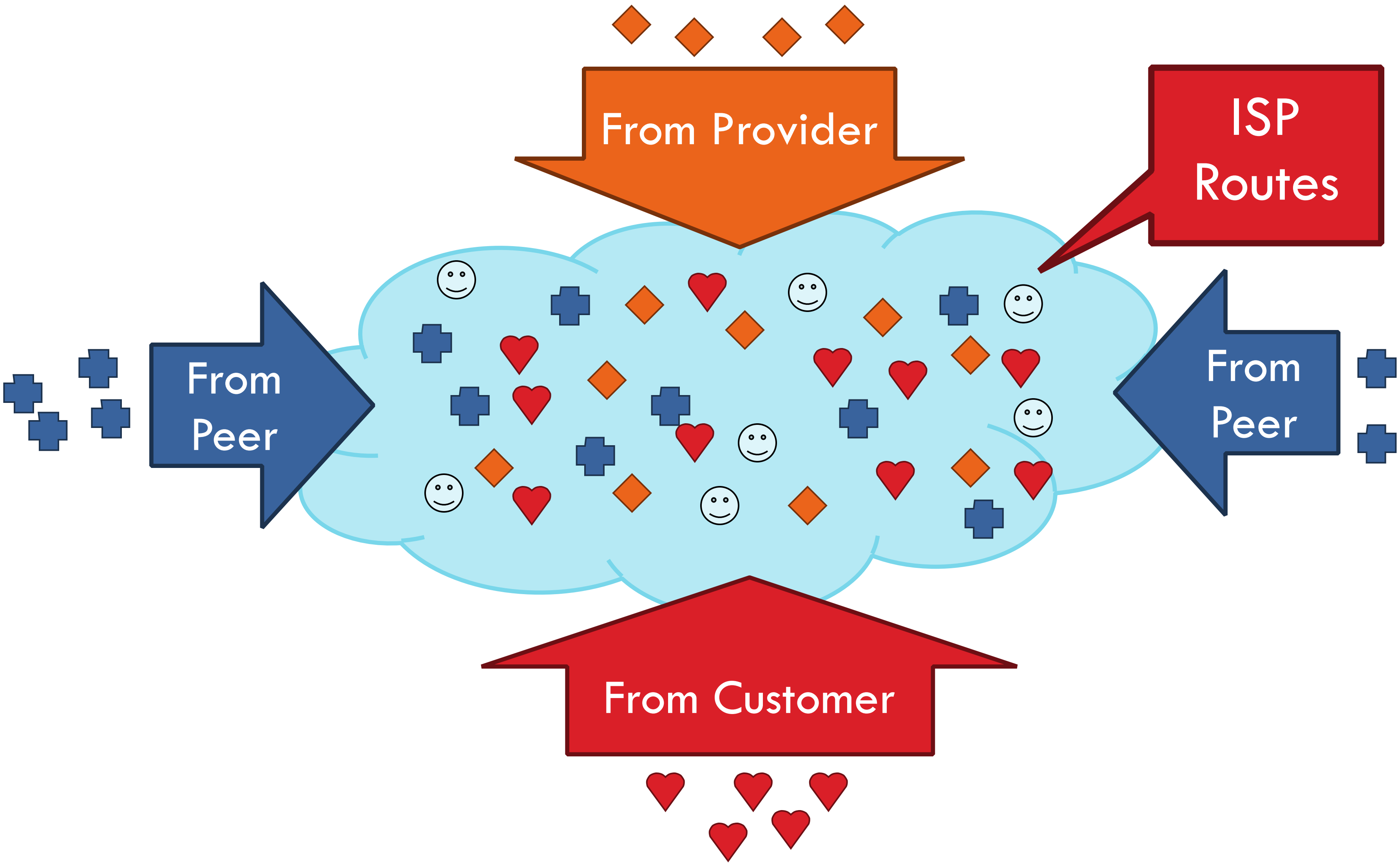
Importing Routes



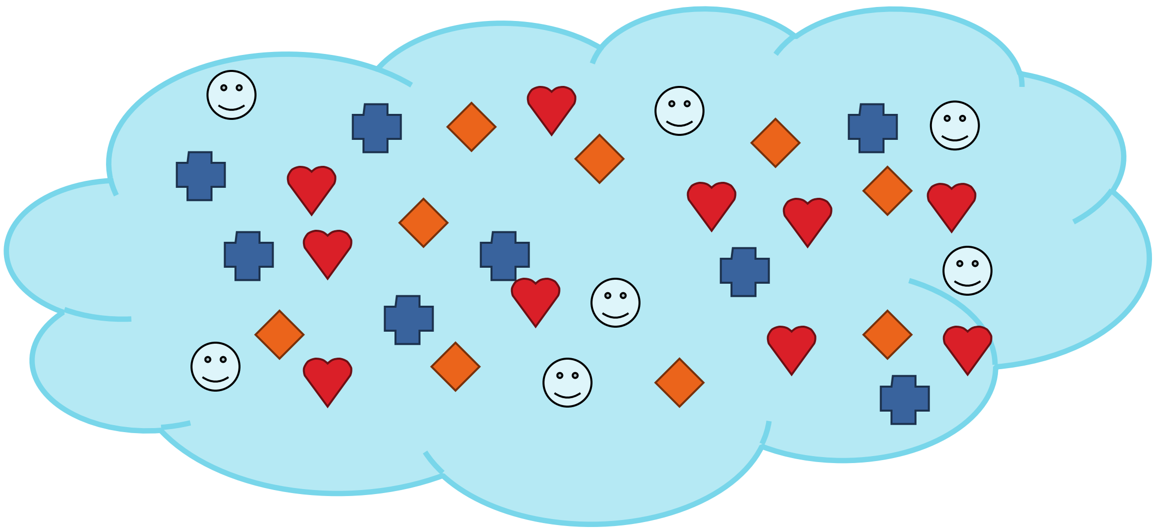
Importing Routes



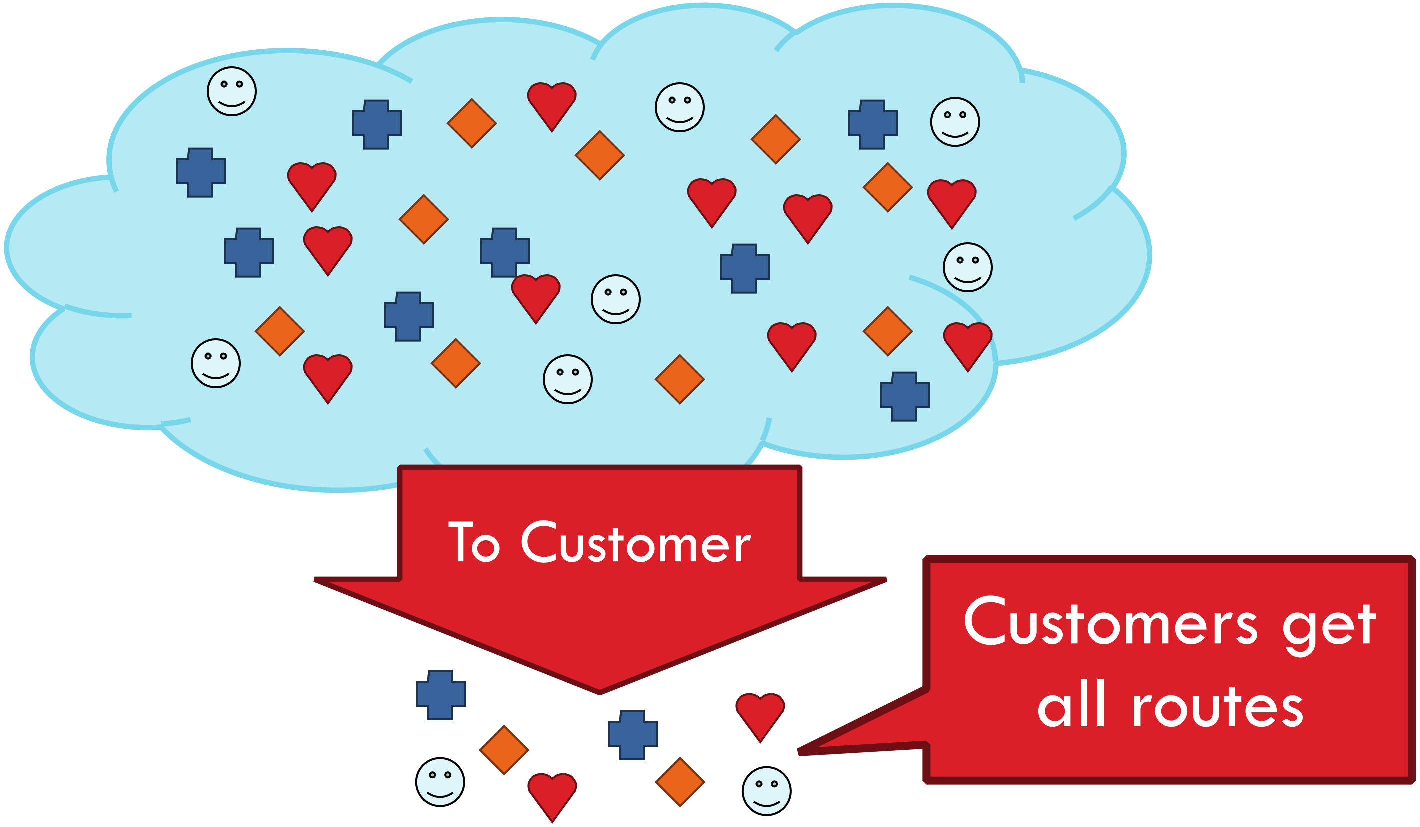
Importing Routes



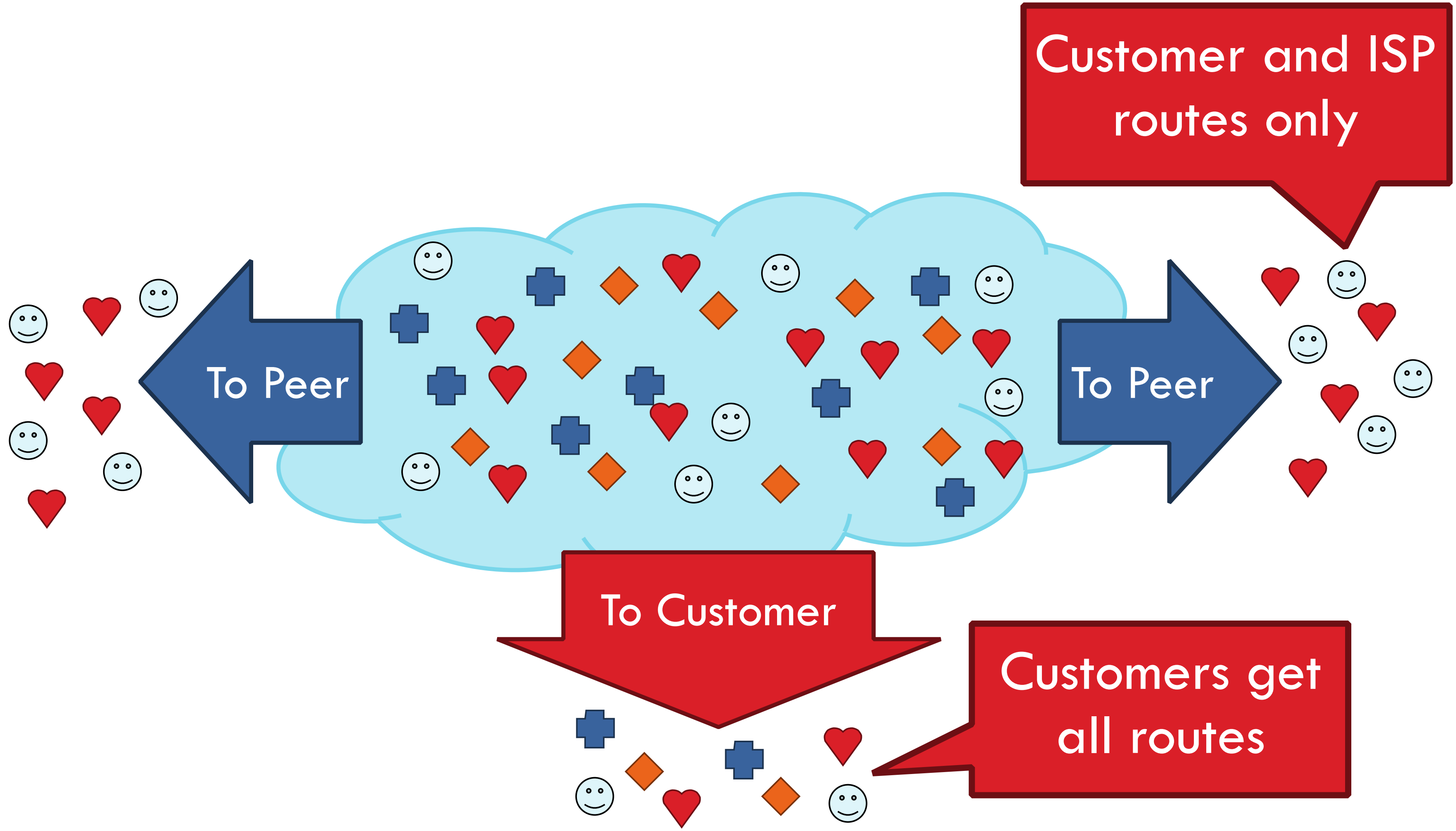
Exporting Routes



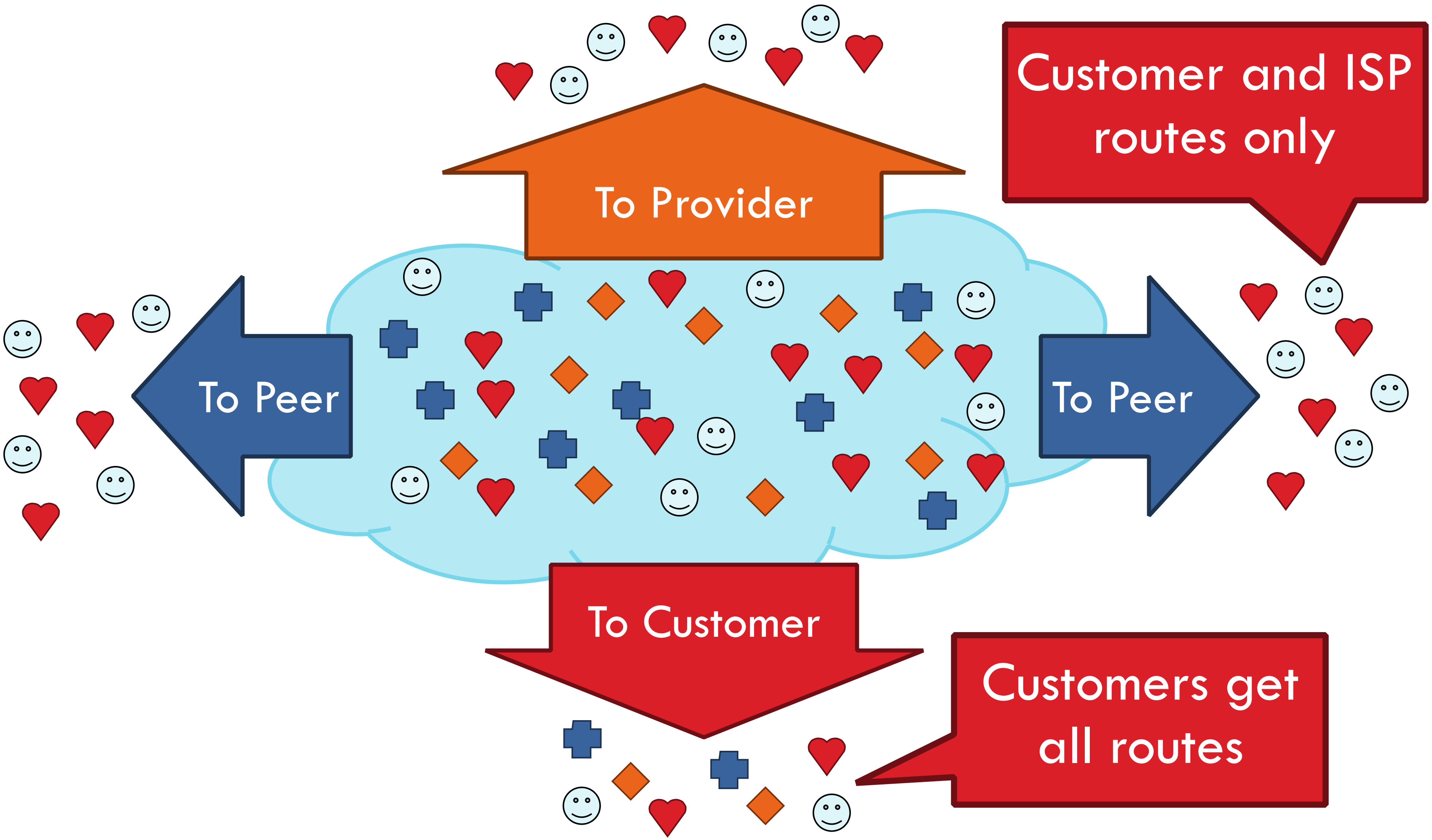
Exporting Routes



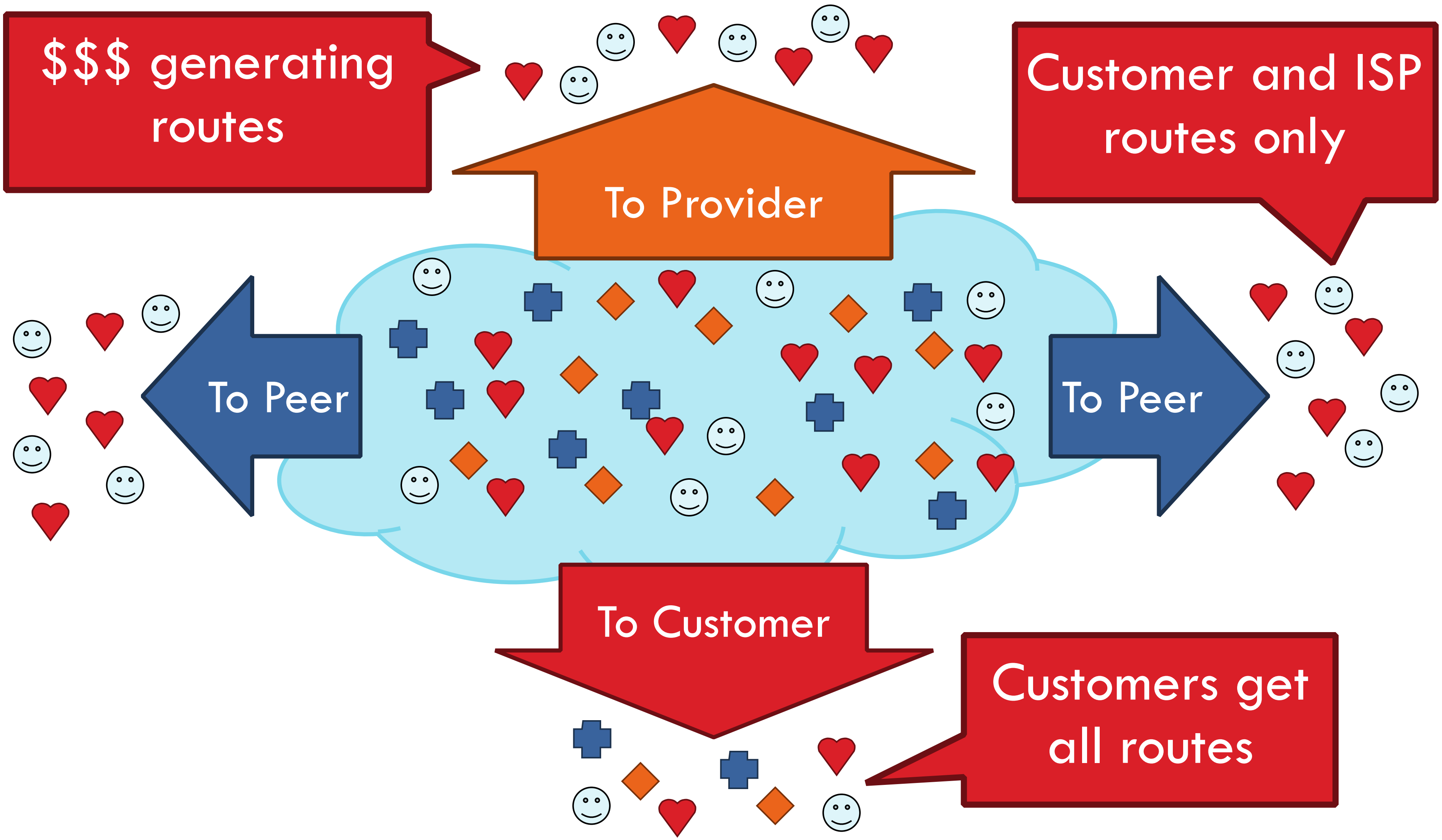
Exporting Routes



Exporting Routes



Exporting Routes

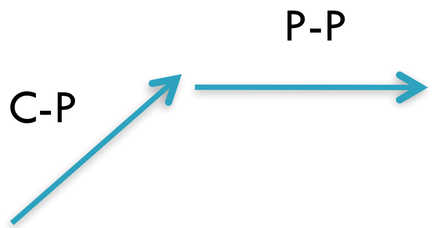


Modeling BGP

- AS relationships
  - Customer/provider
  - Peer
  - Sibling, IXP
- Gao-Rexford model
  - AS prefers to use customer path, then peer, then provider
    - Follow the money!
  - Valley-free routing
  - Hierarchical view of routing (incorrect but frequently used)

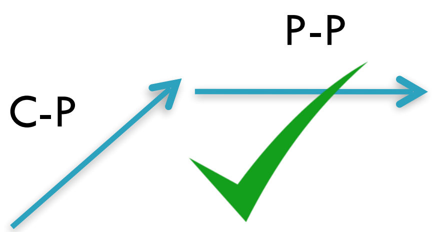
Modeling BGP

- AS relationships
  - Customer/provider
  - Peer
  - Sibling, IXP
- Gao-Rexford model
  - AS prefers to use customer path, then peer, then provider
    - Follow the money!
  - Valley-free routing
  - Hierarchical view of routing (incorrect but frequently used)



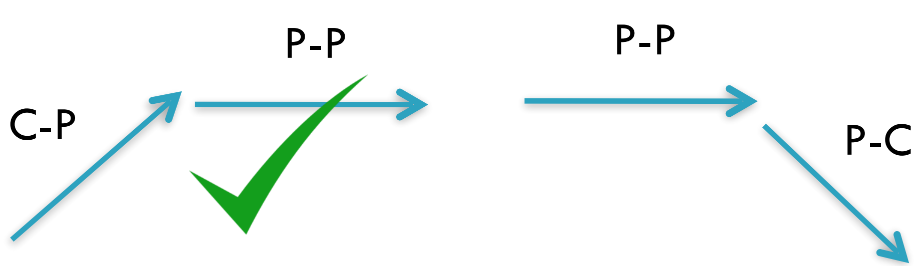
Modeling BGP

- AS relationships
  - Customer/provider
  - Peer
  - Sibling, IXP
- Gao-Rexford model
  - AS prefers to use customer path, then peer, then provider
    - Follow the money!
  - Valley-free routing
  - Hierarchical view of routing (incorrect but frequently used)



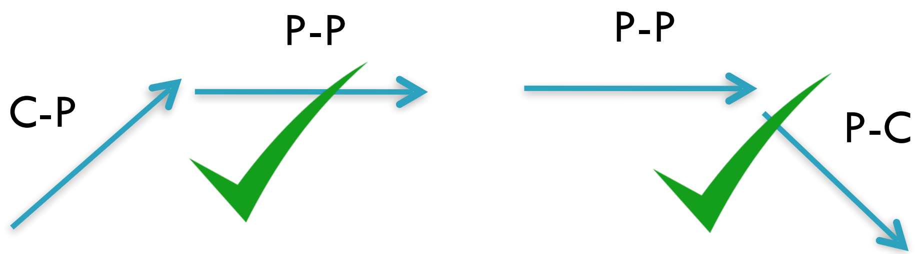
Modeling BGP

- AS relationships
  - Customer/provider
  - Peer
  - Sibling, IXP
- Gao-Rexford model
  - AS prefers to use customer path, then peer, then provider
    - Follow the money!
  - Valley-free routing
  - Hierarchical view of routing (incorrect but frequently used)



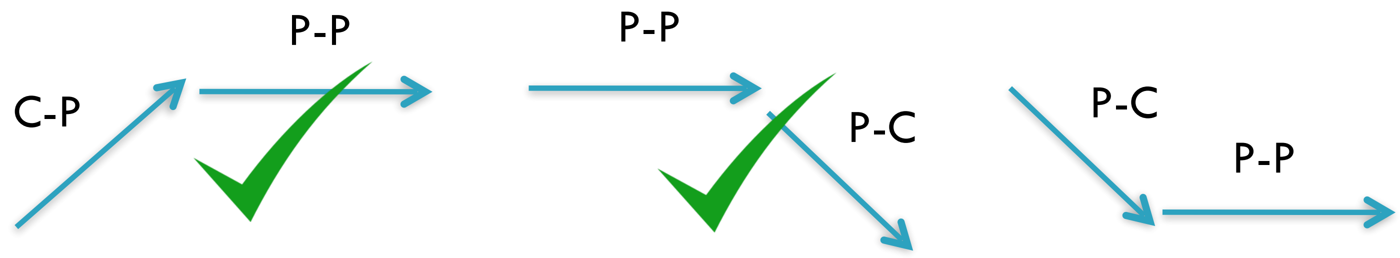
Modeling BGP

- AS relationships
  - Customer/provider
  - Peer
  - Sibling, IXP
- Gao-Rexford model
  - AS prefers to use customer path, then peer, then provider
    - Follow the money!
  - Valley-free routing
  - Hierarchical view of routing (incorrect but frequently used)



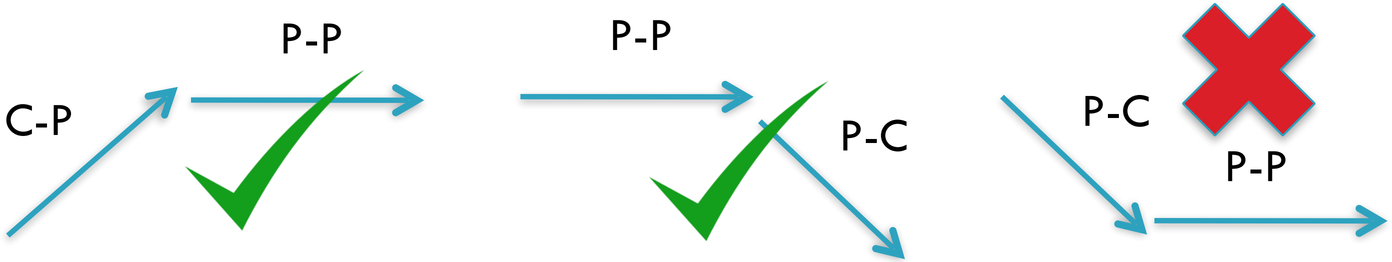
Modeling BGP

- AS relationships
  - Customer/provider
  - Peer
  - Sibling, IXP
- Gao-Rexford model
  - AS prefers to use customer path, then peer, then provider
    - Follow the money!
  - Valley-free routing
  - Hierarchical view of routing (incorrect but frequently used)



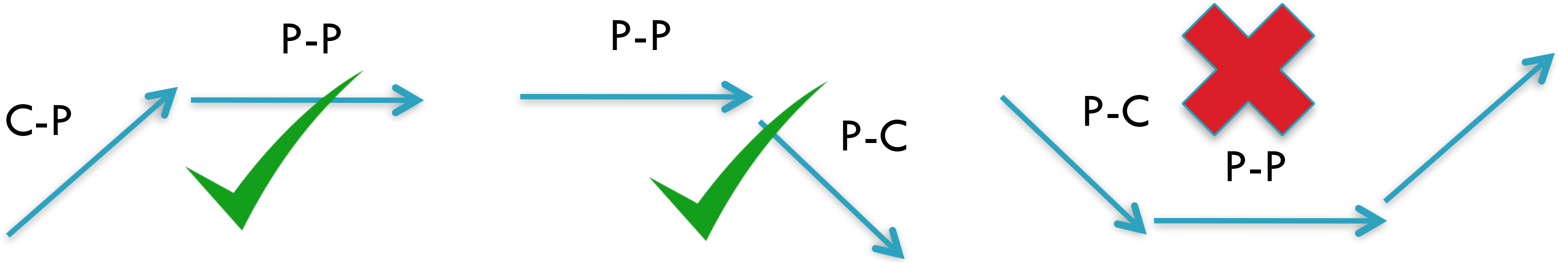
Modeling BGP

- AS relationships
  - Customer/provider
  - Peer
  - Sibling, IXP
- Gao-Rexford model
  - AS prefers to use customer path, then peer, then provider
    - Follow the money!
  - Valley-free routing
  - Hierarchical view of routing (incorrect but frequently used)



Modeling BGP

- AS relationships
  - Customer/provider
  - Peer
  - Sibling, IXP
- Gao-Rexford model
  - AS prefers to use customer path, then peer, then provider
    - Follow the money!
  - Valley-free routing
  - Hierarchical view of routing (incorrect but frequently used)



AS Relationships: It's Complicated

- GR Model is strictly hierarchical
  - Each AS pair has exactly one relationship
  - Each relationship is the same for all prefixes

AS Relationships: It's Complicated

- GR Model is strictly hierarchical
  - Each AS pair has exactly one relationship
  - Each relationship is the same for all prefixes
- In practice it's much more complicated
  - Rise of widespread peering
  - Regional, per-prefix peerings
  - Tier-1's being shoved out by "hypergiants"
  - IXPs dominating traffic volume

AS Relationships: It's Complicated

- GR Model is strictly hierarchical
  - Each AS pair has exactly one relationship
  - Each relationship is the same for all prefixes
- In practice it's much more complicated
  - Rise of widespread peering
  - Regional, per-prefix peerings
  - Tier-1's being shoved out by "hypergiants"
  - IXPs dominating traffic volume
- Modeling is very hard, very prone to error
  - Huge potential impact for understanding Internet behavior

Other BGP Attributes

- AS_SET
  - Instead of a single AS appearing at a slot, it's a set of Ases
  - Why?

Other BGP Attributes

- AS_SET
  - Instead of a single AS appearing at a slot, it’s a set of Ases
  - Why?
- Communities
  - Arbitrary number that is used by neighbors for routing decisions
    - Export this route only in Europe
    - Do not export to your peers
  - Usually stripped after first interdomain hop
  - Why?

Other BGP Attributes

- AS_SET
  - Instead of a single AS appearing at a slot, it's a set of Ases
  - Why?
- Communities
  - Arbitrary number that is used by neighbors for routing decisions
    - Export this route only in Europe
    - Do not export to your peers
  - Usually stripped after first interdomain hop
  - Why?
- Prepending
  - Lengthening the route by adding multiple instances of ASN
  - Why?

Outline

- BGP Basics
- Stable Paths Problem
- BGP in the Real World

What Problem is BGP Solving?

Underlying Problem	Distributed Solution
Shortest Paths	***RIP, OSPF, IS-IS, etc.***
???	***BGP***

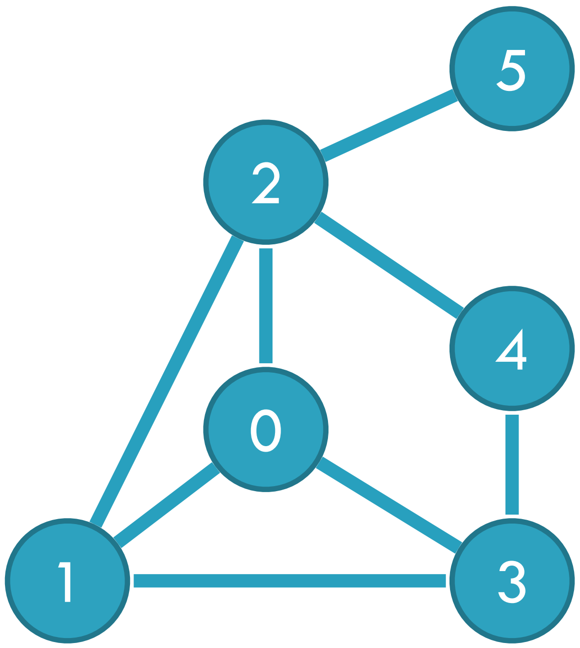
What Problem is BGP Solving?

Underlying Problem	Distributed Solution
Shortest Paths	***RIP, OSPF, IS-IS, etc.***
???	***BGP***

- Knowing ??? can:
  - Aid in the analysis of BGP policy
  - Aid in the design of BGP extensions
  - Help explain BGP routing anomalies
  - Give us a deeper understanding of the protocol

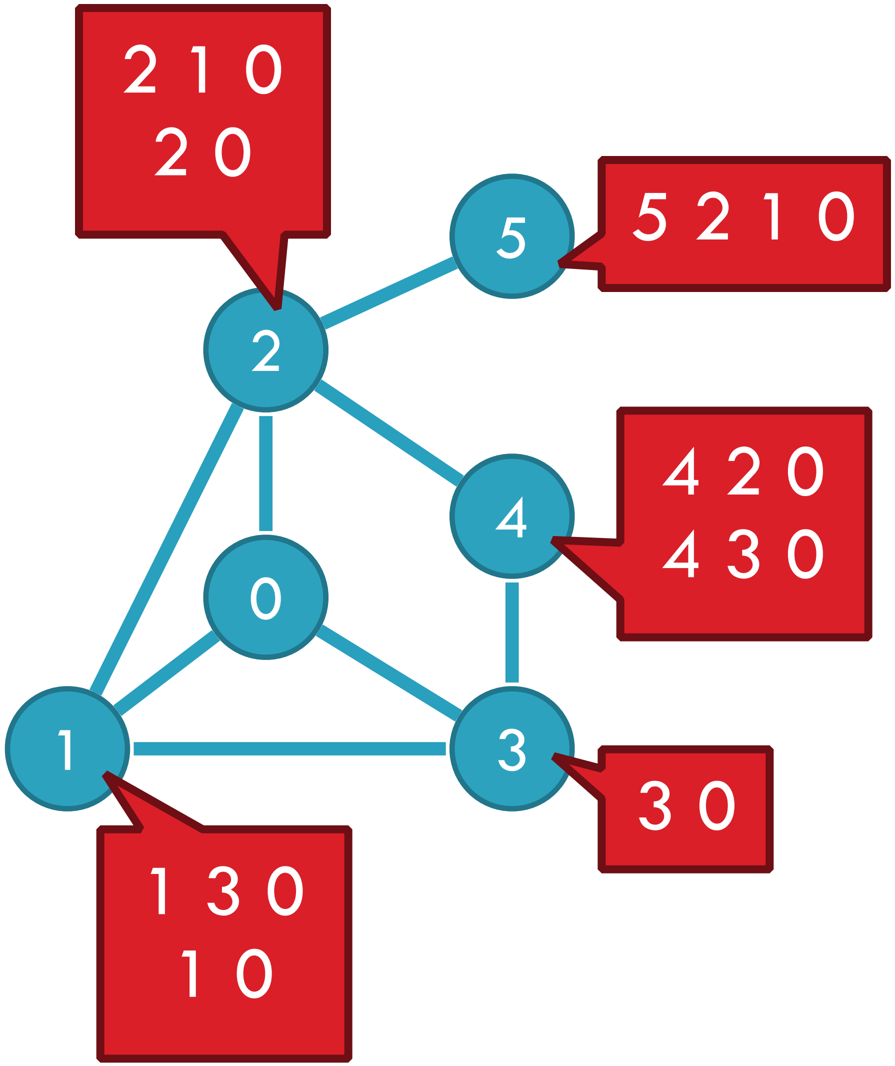
The Stable Paths Problem

- An instance of the SPP:
  - Graph of nodes and edges
  - Node 0, called the origin



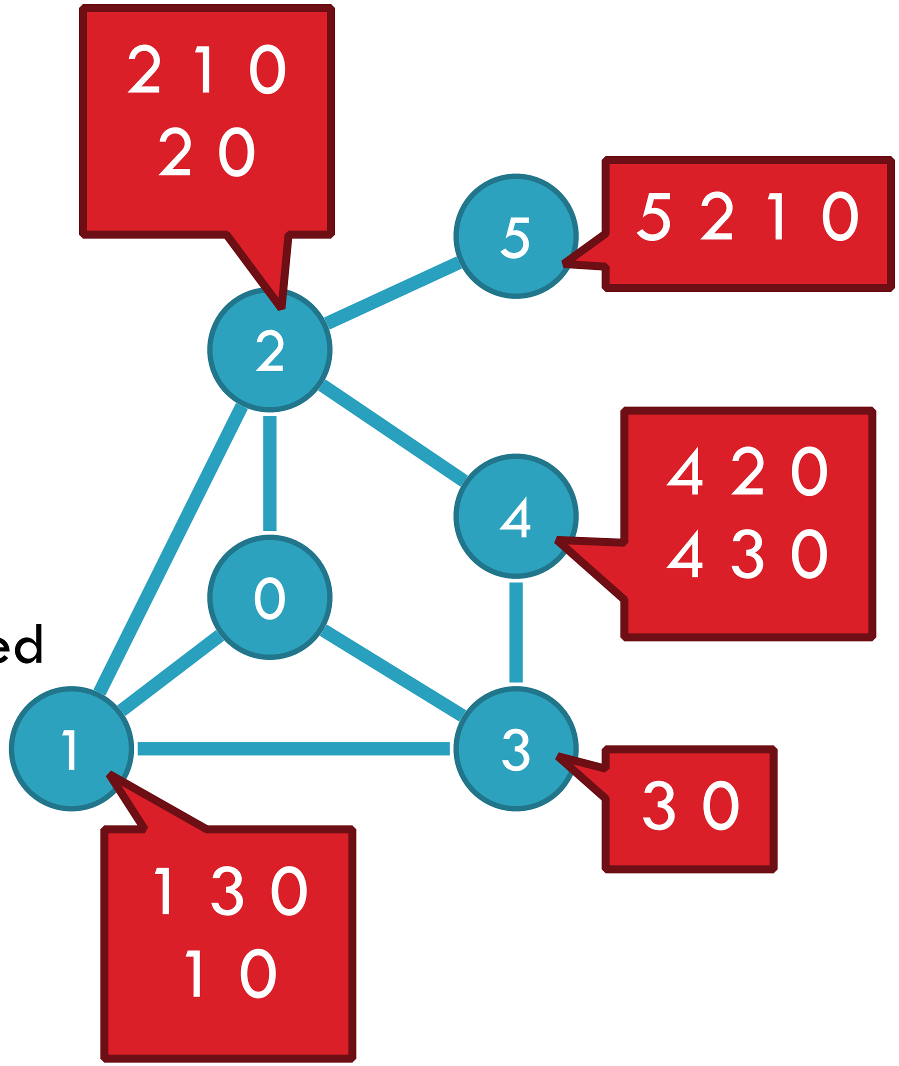
The Stable Paths Problem

- An instance of the SPP:
  - Graph of nodes and edges
  - Node 0, called the origin
  - A set of permitted paths from each node to the origin
    - Each set contains the null path



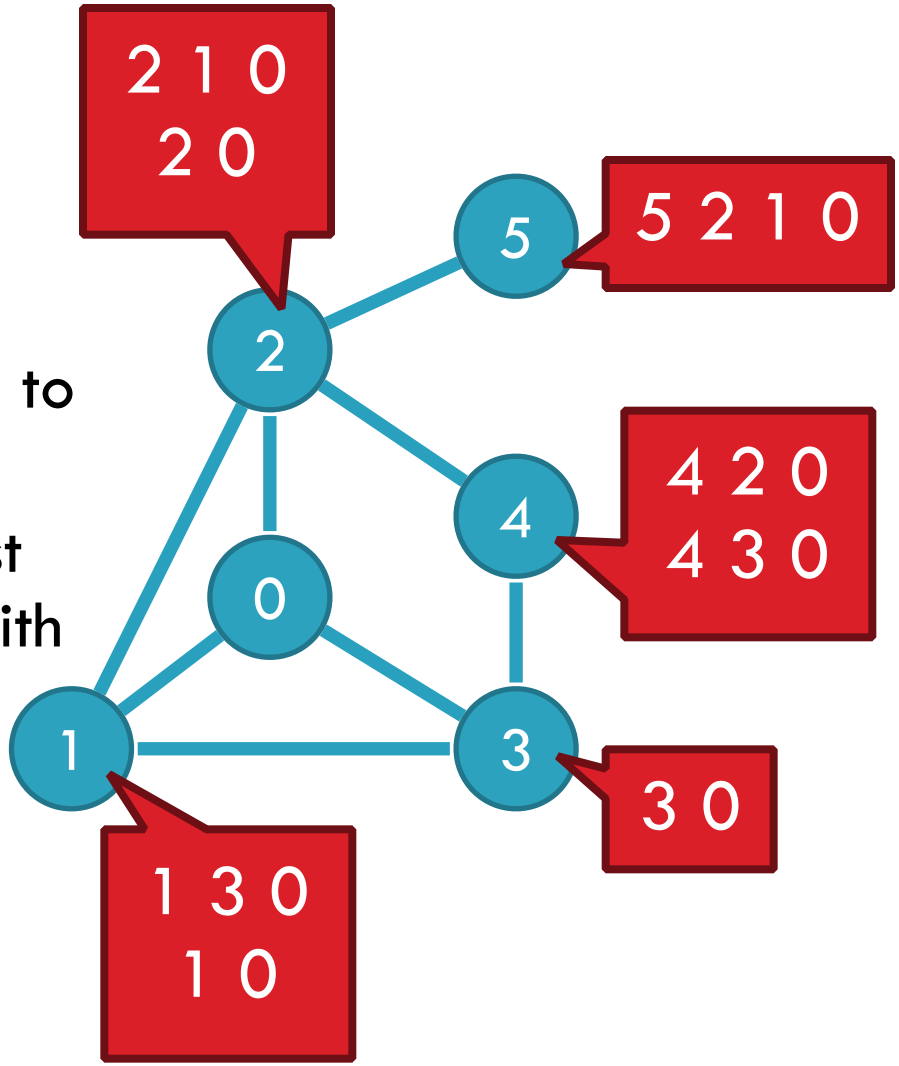
The Stable Paths Problem

- An instance of the SPP:
  - Graph of nodes and edges
  - Node 0, called the origin
  - A set of permitted paths from each node to the origin
    - Each set contains the null path
  - Each set of paths is ranked
    - Null path is always least preferred



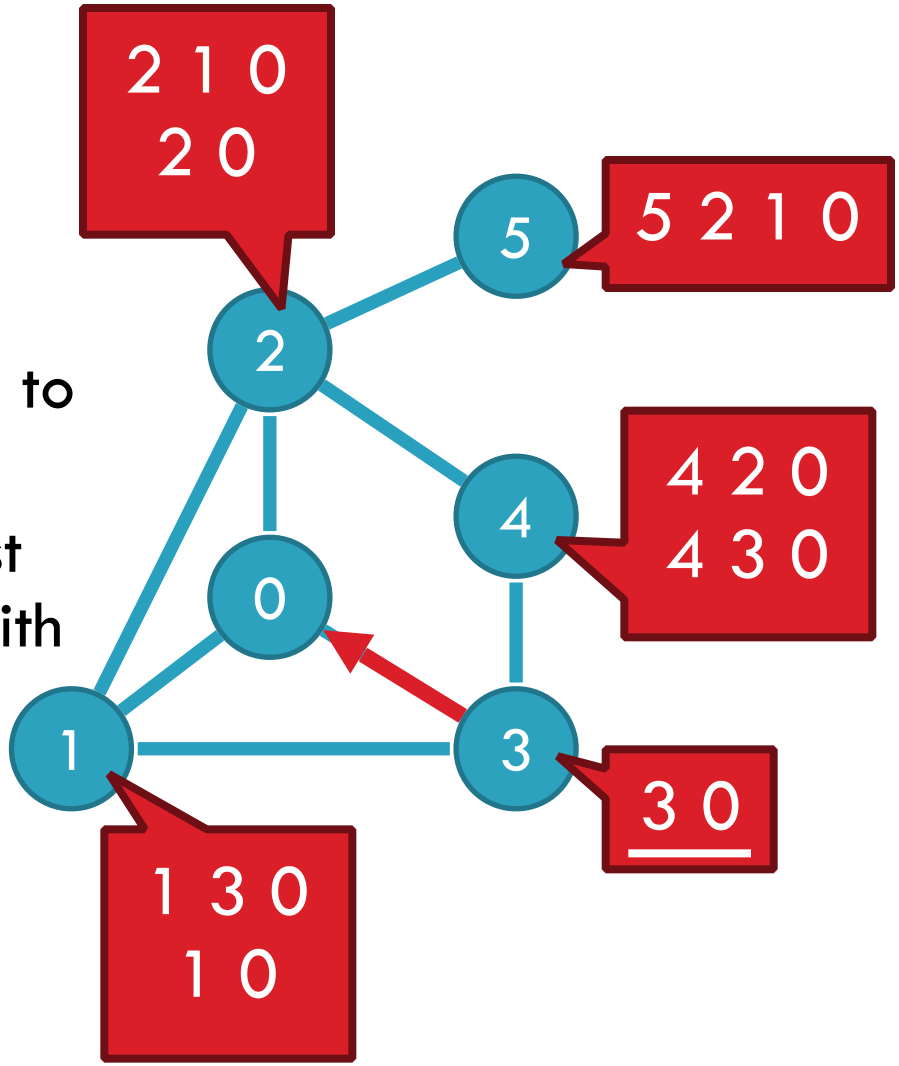
A Solution to the SPP

- A solution is an assignment of permitted paths to each node such that:
  - Node *u*'s path is either null or *uwP*, where path *uw* is assigned to node *w* and edge *u* → *w* exists
  - Each node is assigned the higest ranked path that is consistent with their neighbors



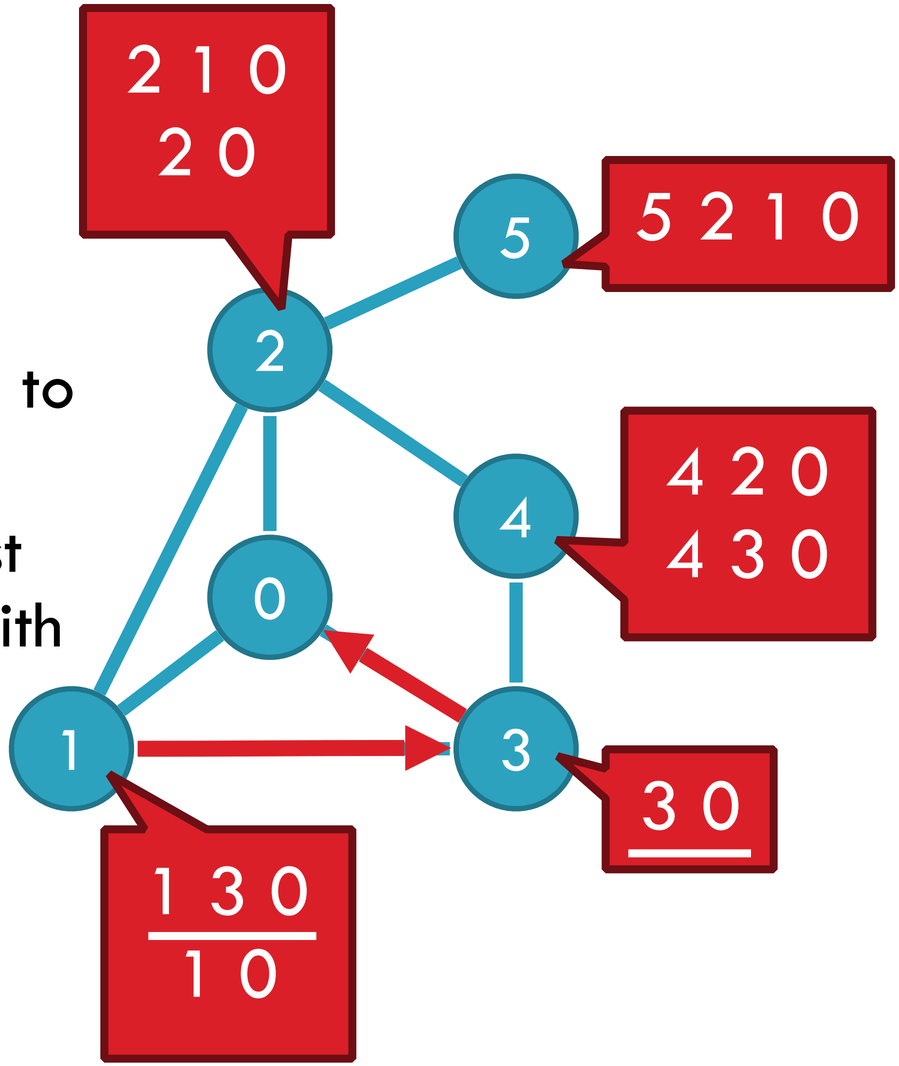
A Solution to the SPP

- A solution is an assignment of permitted paths to each node such that:
  - Node *u*'s path is either null or *uwP*, where path *uw* is assigned to node *w* and edge *u* → *w* exists
  - Each node is assigned the higest ranked path that is consistent with their neighbors



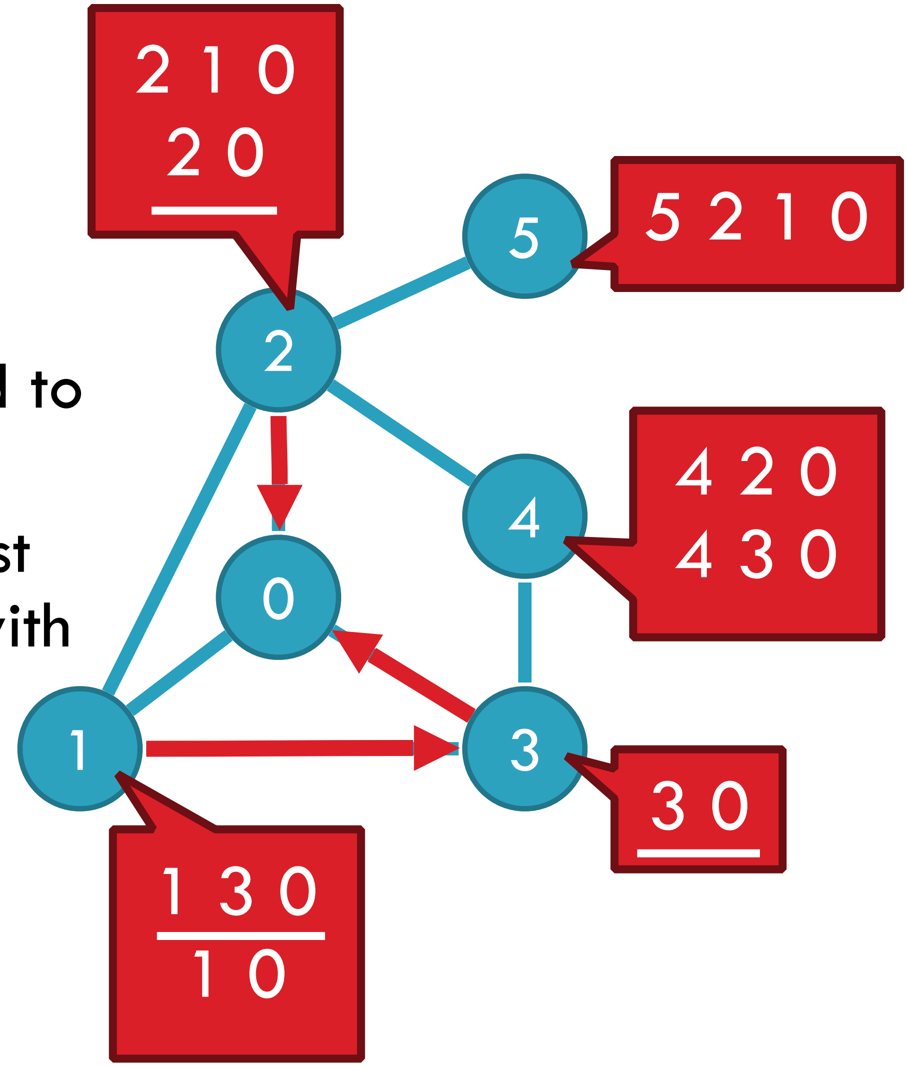
A Solution to the SPP

- A solution is an assignment of permitted paths to each node such that:
  - Node *u*'s path is either null or *uwP*, where path *uw* is assigned to node *w* and edge *u* → *w* exists
  - Each node is assigned the higest ranked path that is consistent with their neighbors



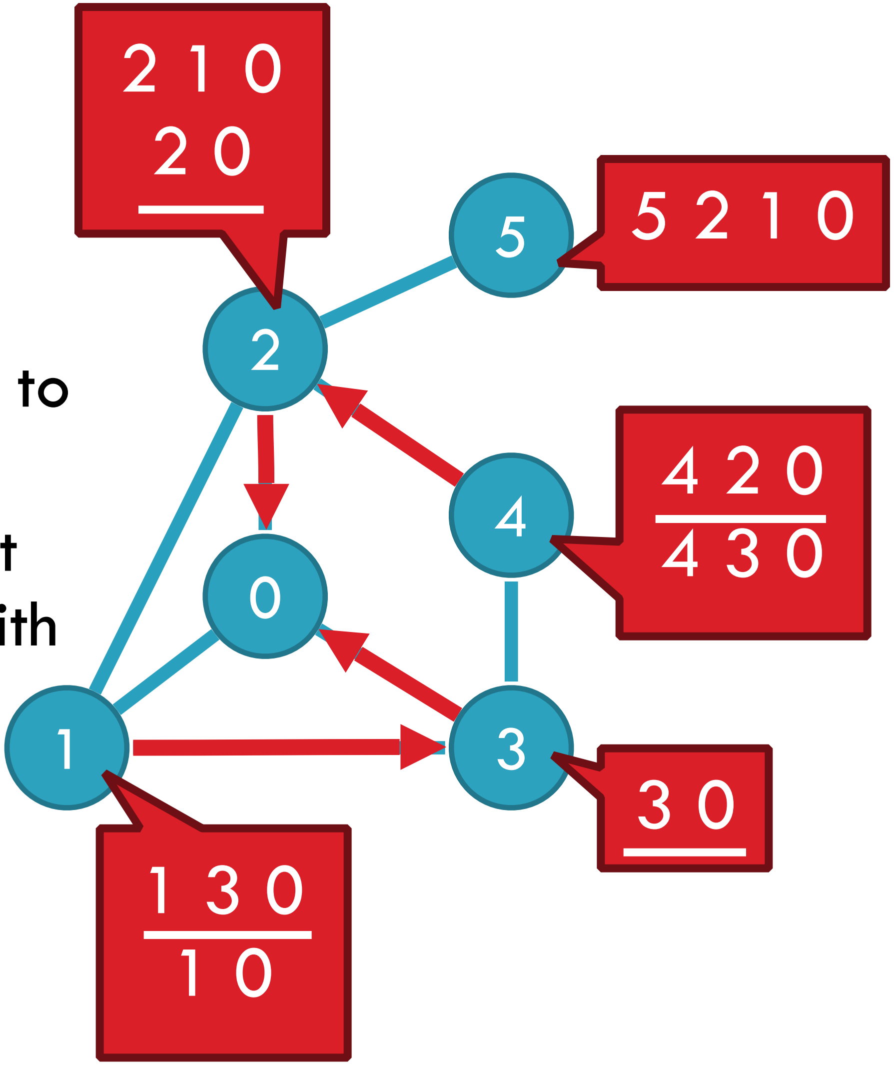
A Solution to the SPP

- A solution is an assignment of permitted paths to each node such that:
  - Node *u*'s path is either null or *uwP*, where path *uw* is assigned to node *w* and edge *u* → *w* exists
  - Each node is assigned the higest ranked path that is consistent with their neighbors



A Solution to the SPP

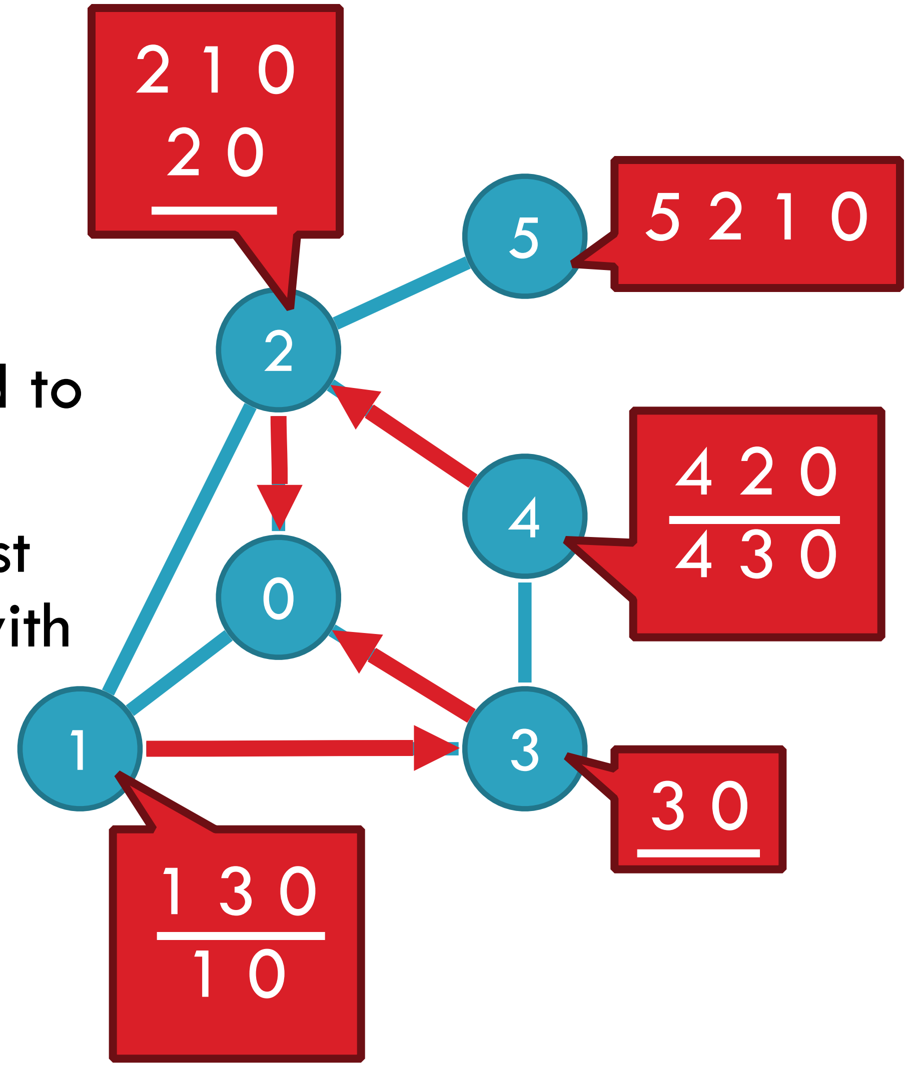
- A solution is an assignment of permitted paths to each node such that:
  - Node *u*'s path is either null or *uwP*, where path *uw* is assigned to node *w* and edge *u* → *w* exists
  - Each node is assigned the higest ranked path that is consistent with their neighbors



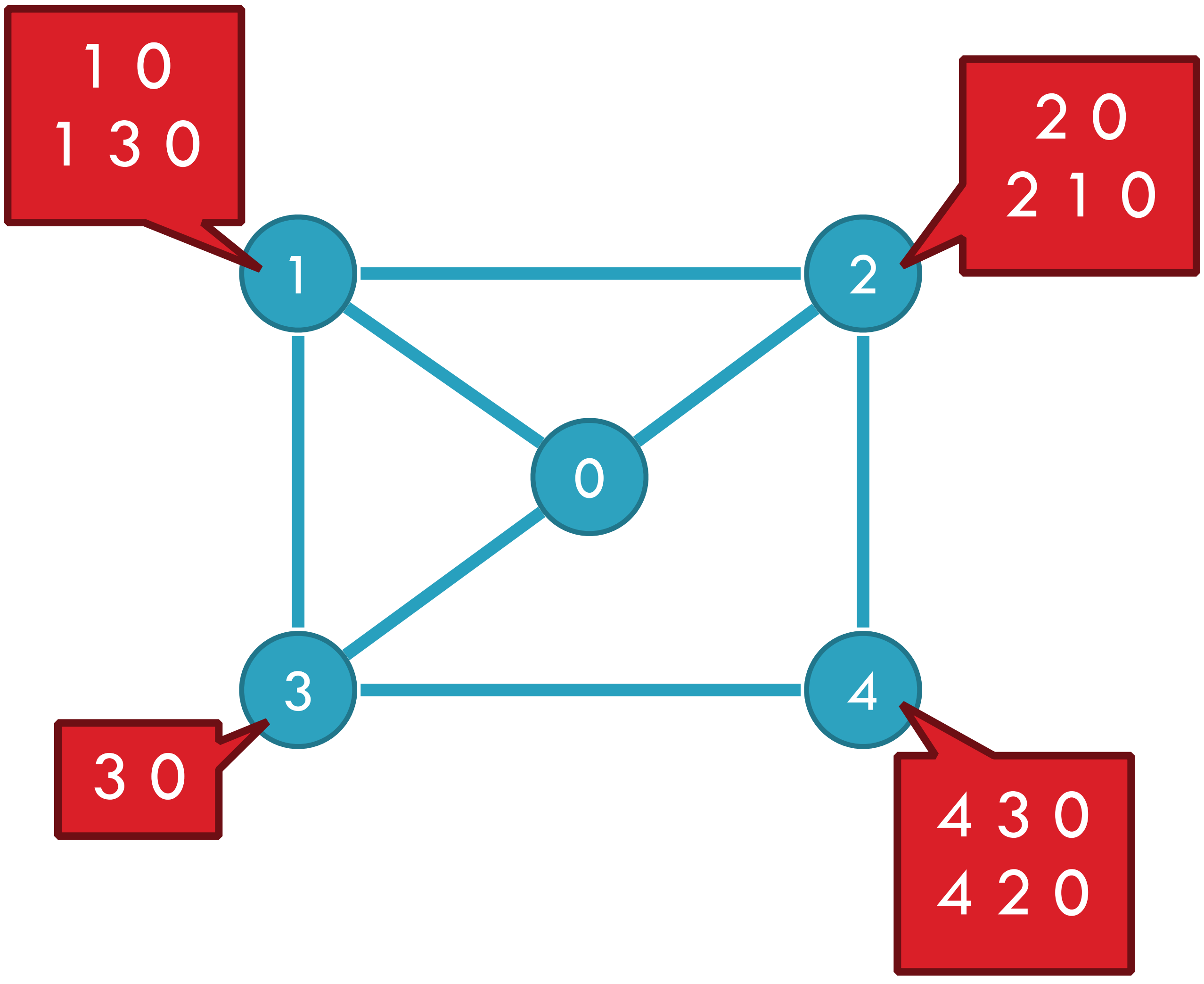
A Solution to the SPP

- A solution is an assignment of permitted paths to each node such that:

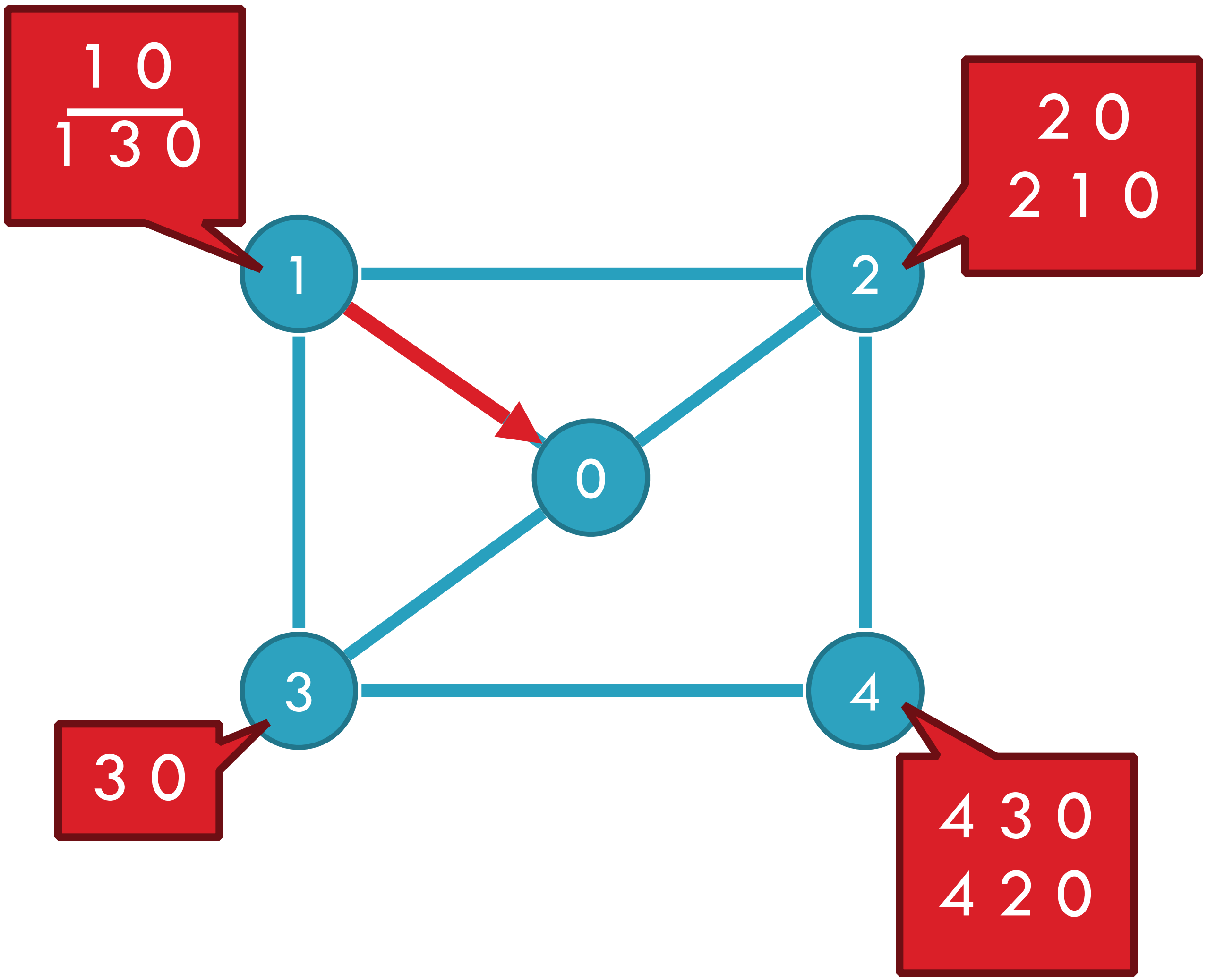
Solutions need not use the shortest paths, or form a spanning tree



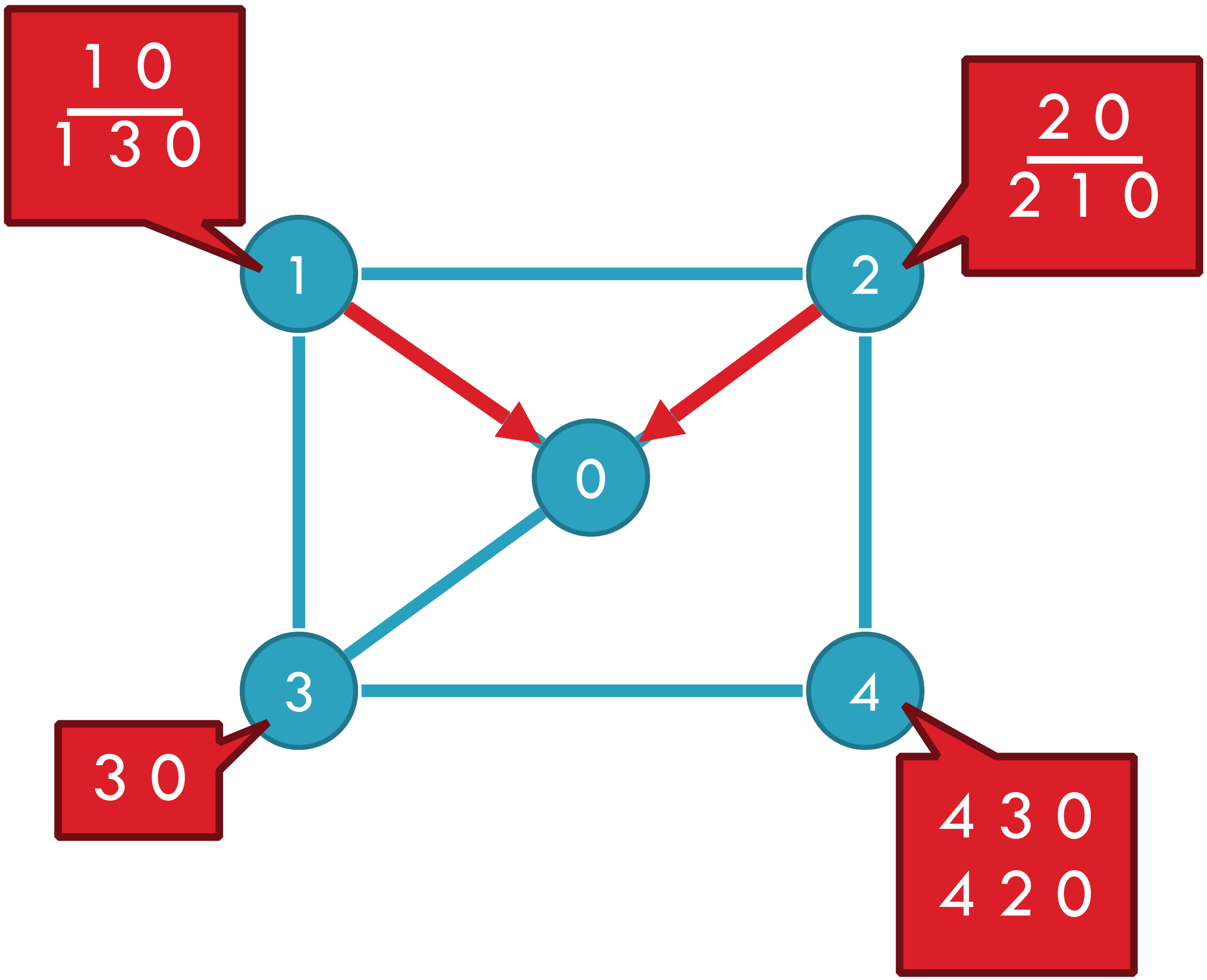
Simple SPP Example



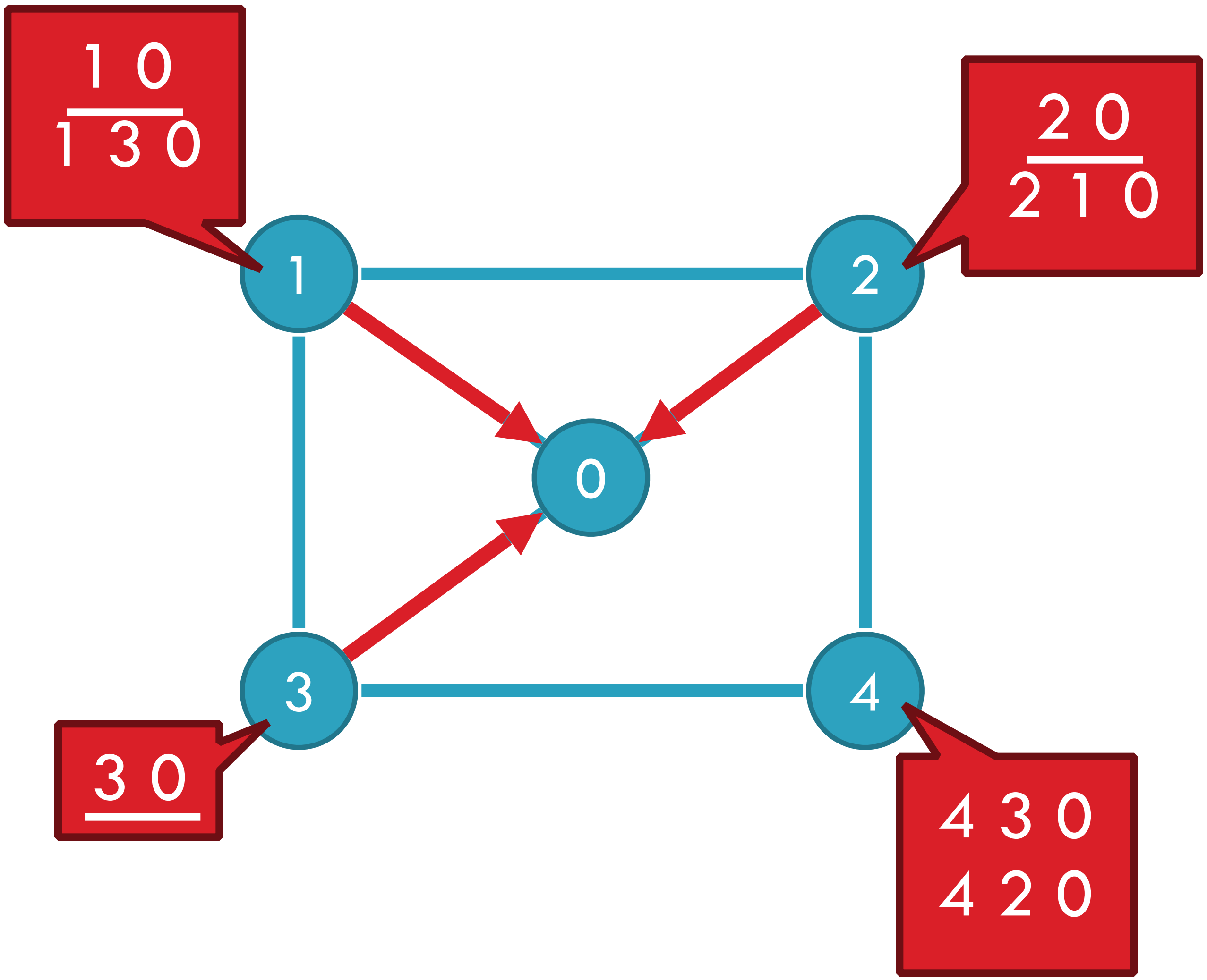
Simple SPP Example



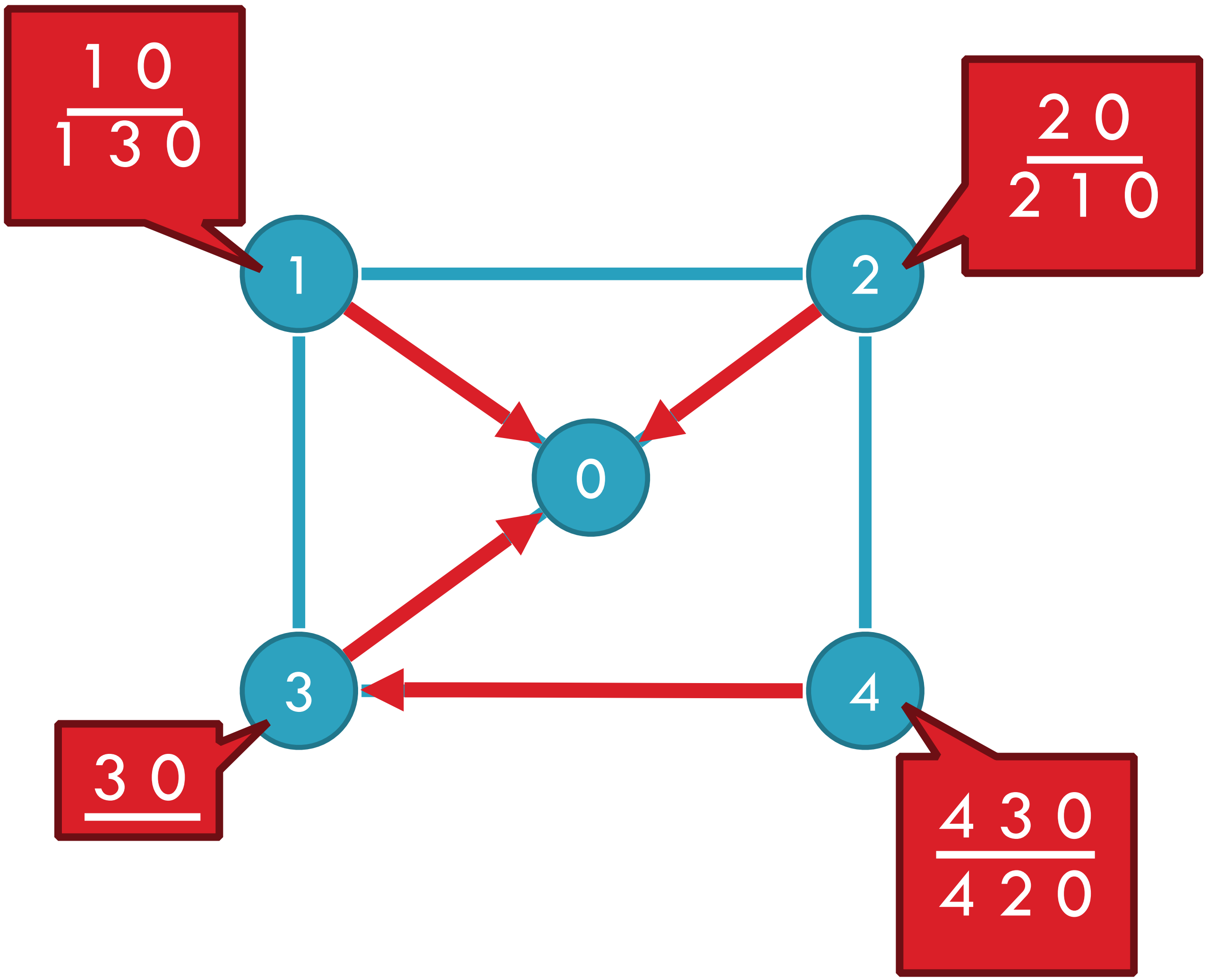
Simple SPP Example



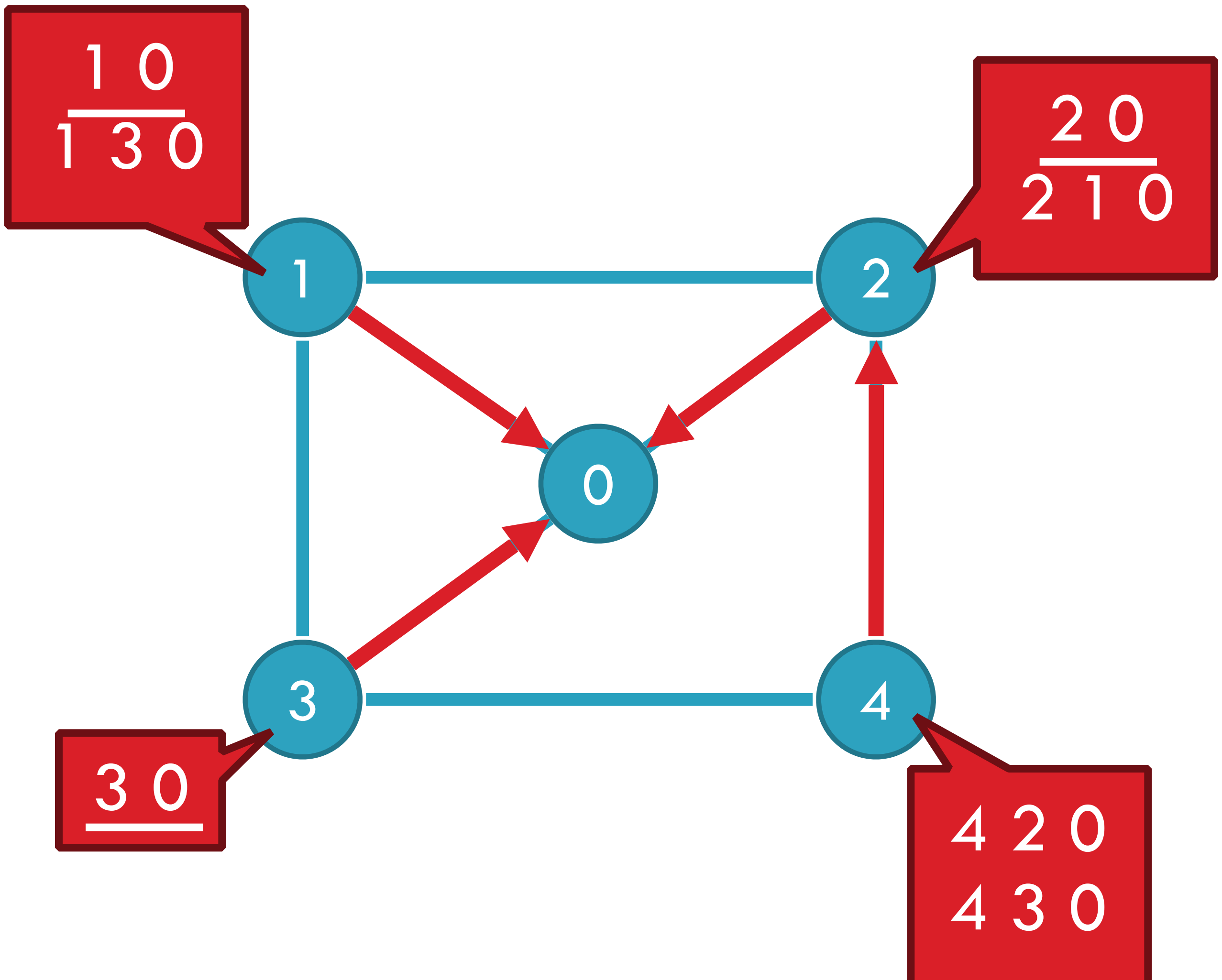
Simple SPP Example



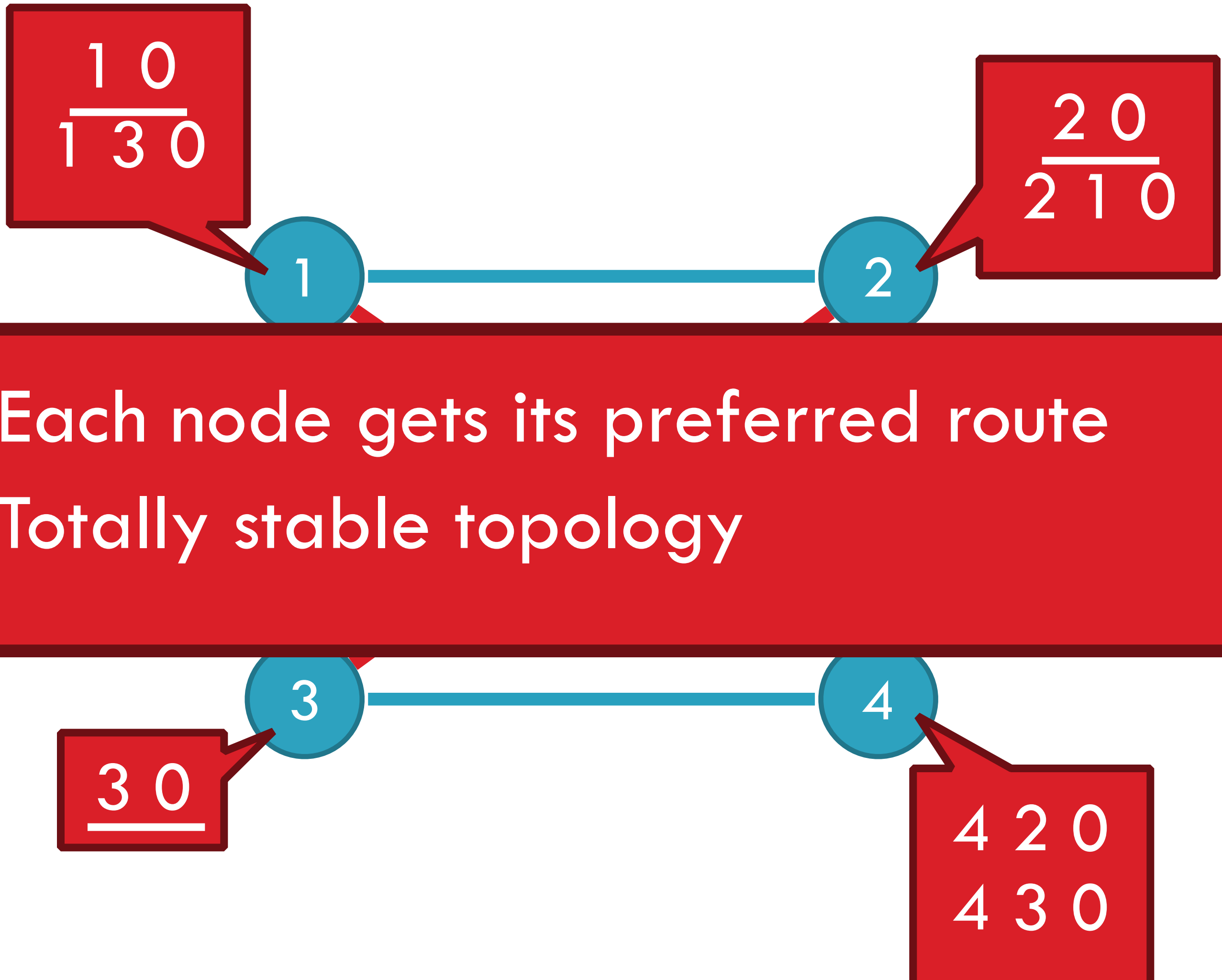
Simple SPP Example



Simple SPP Example

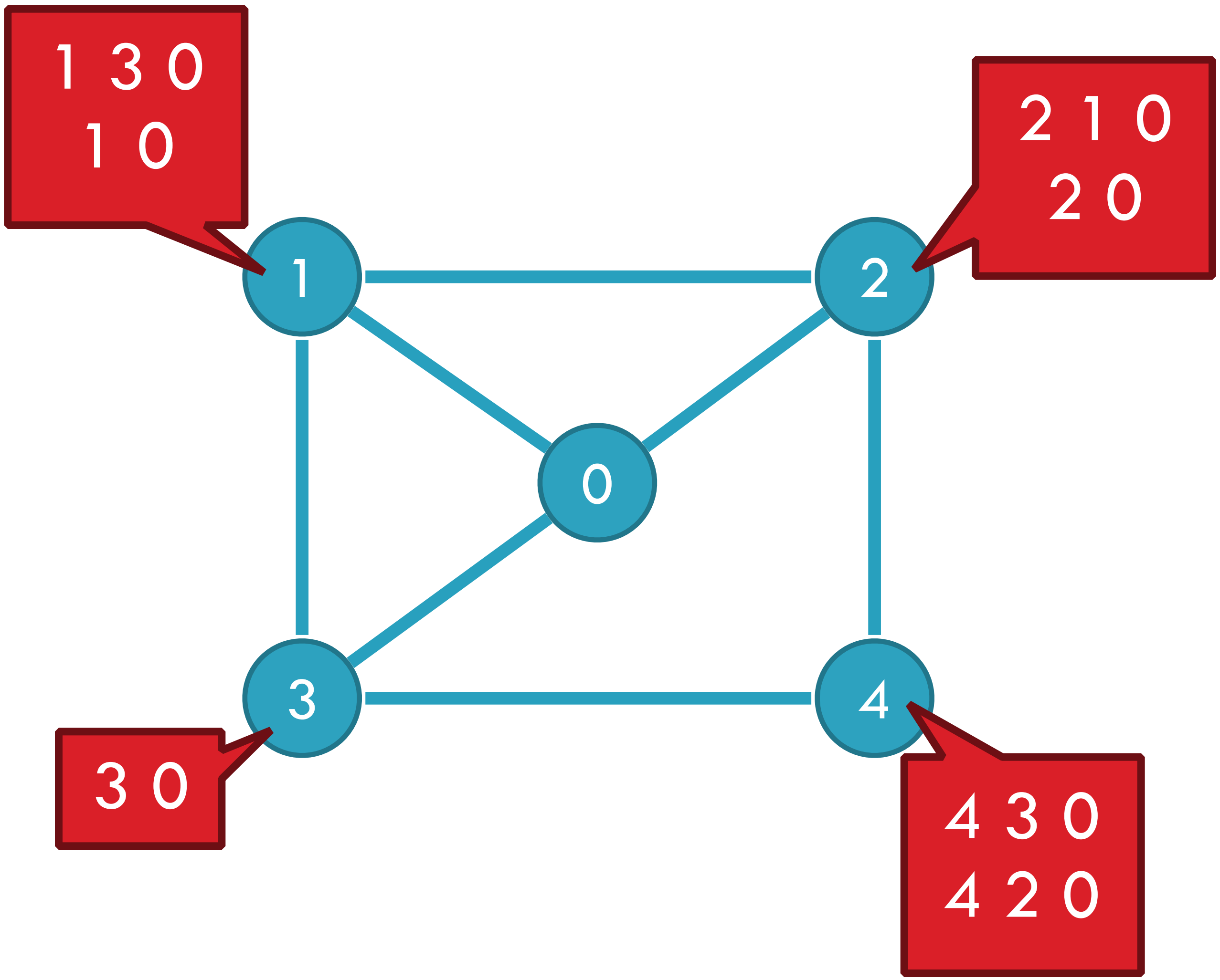


Simple SPP Example

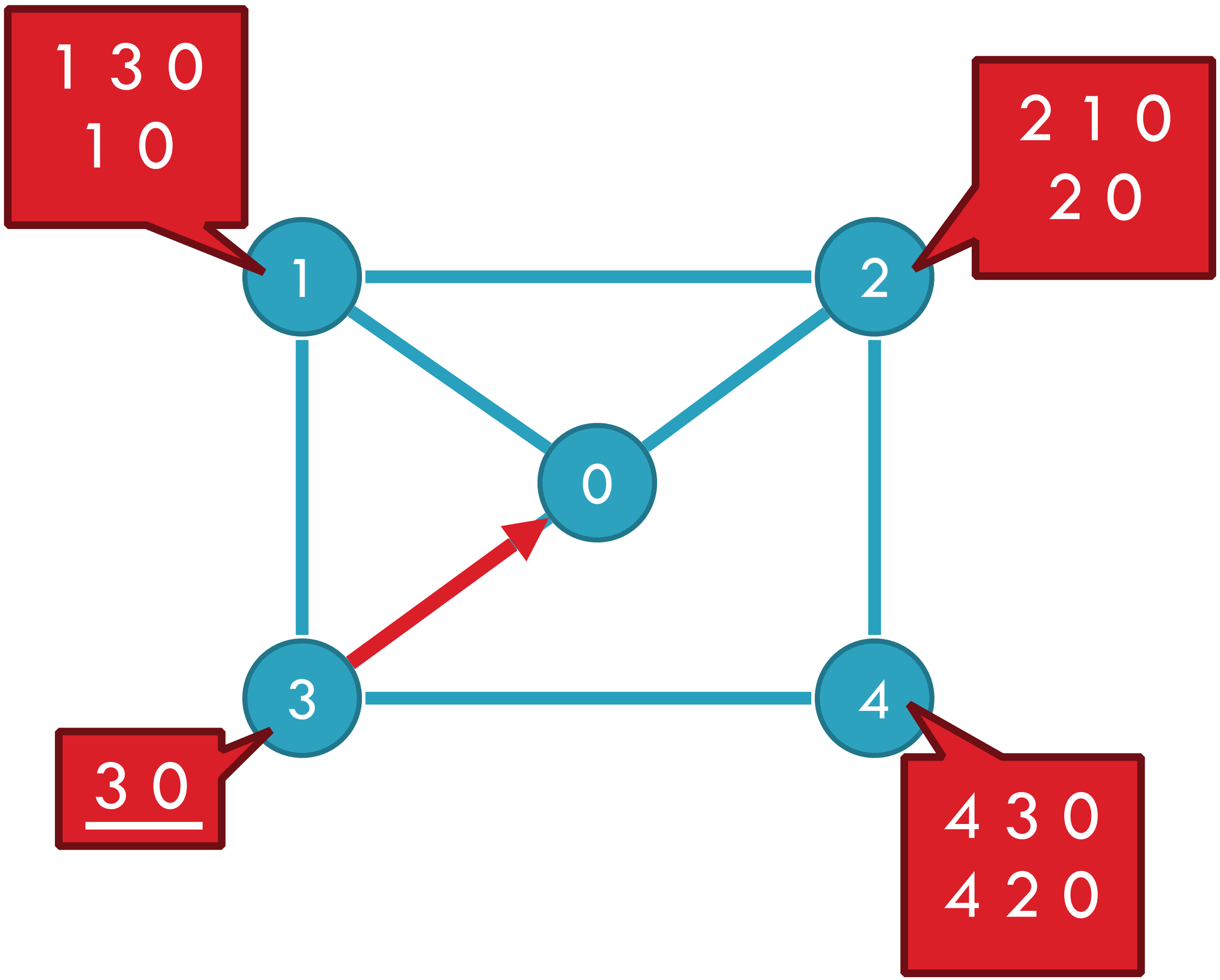


- Each node gets its preferred route
- Totally stable topology

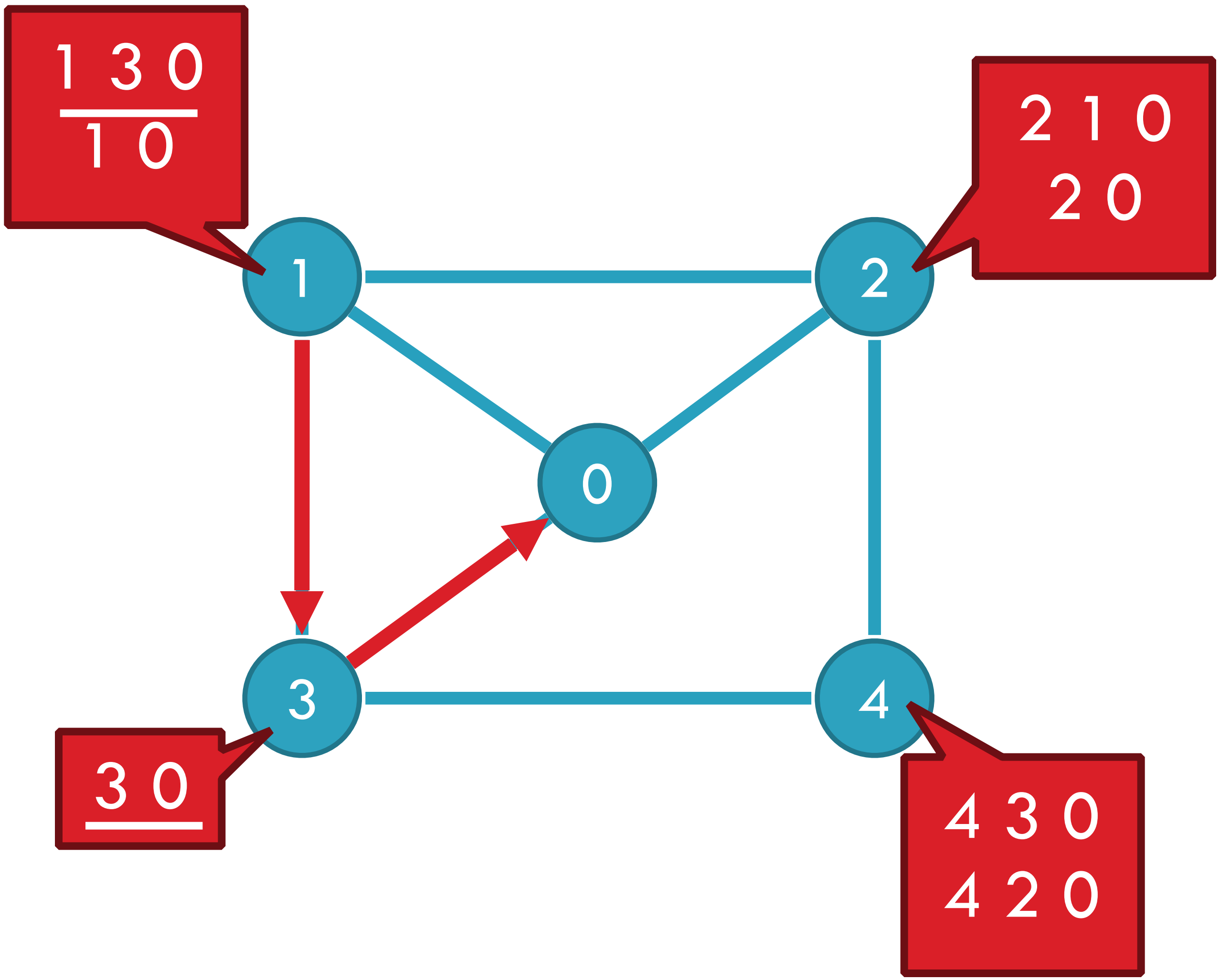
Good Gadget



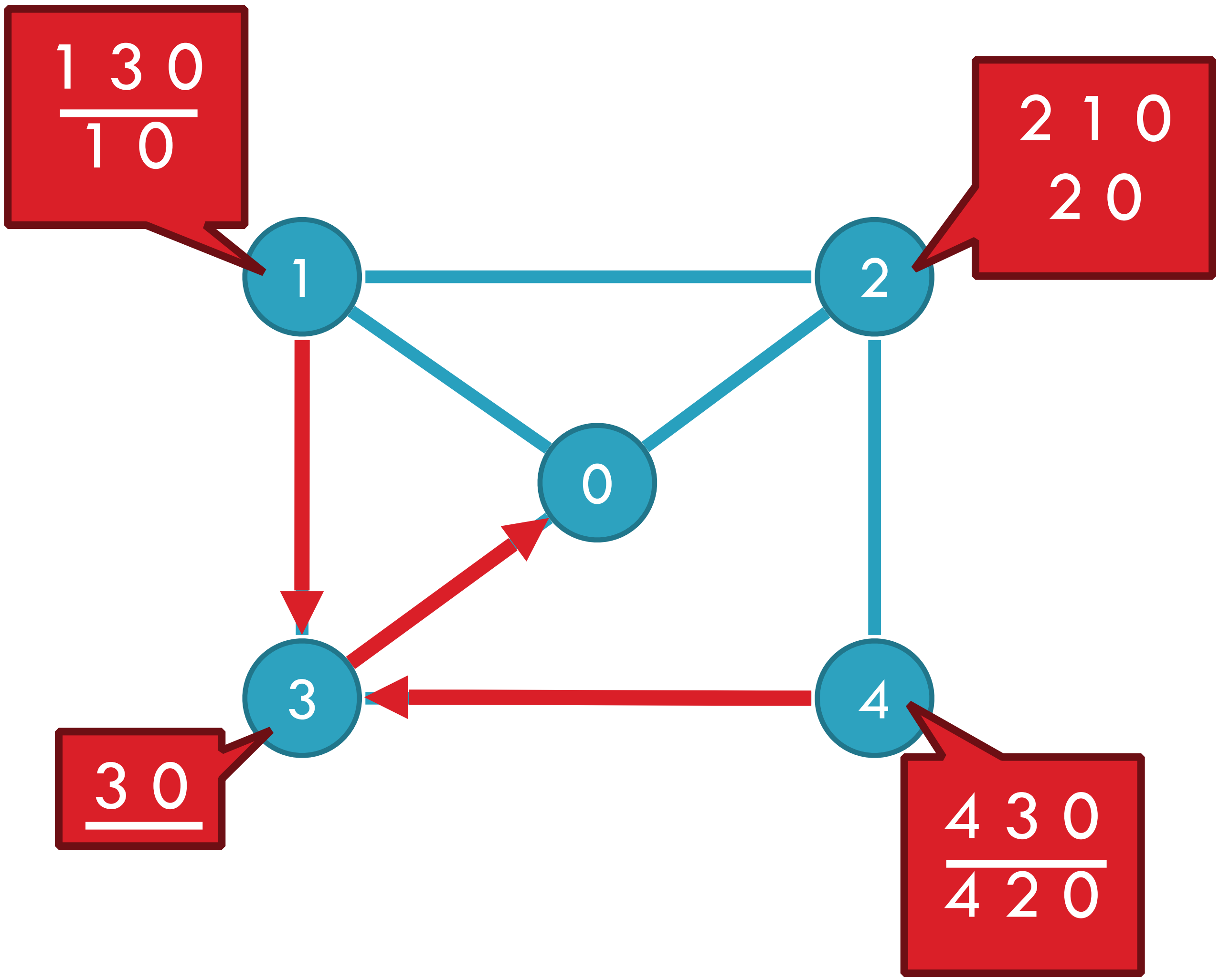
Good Gadget



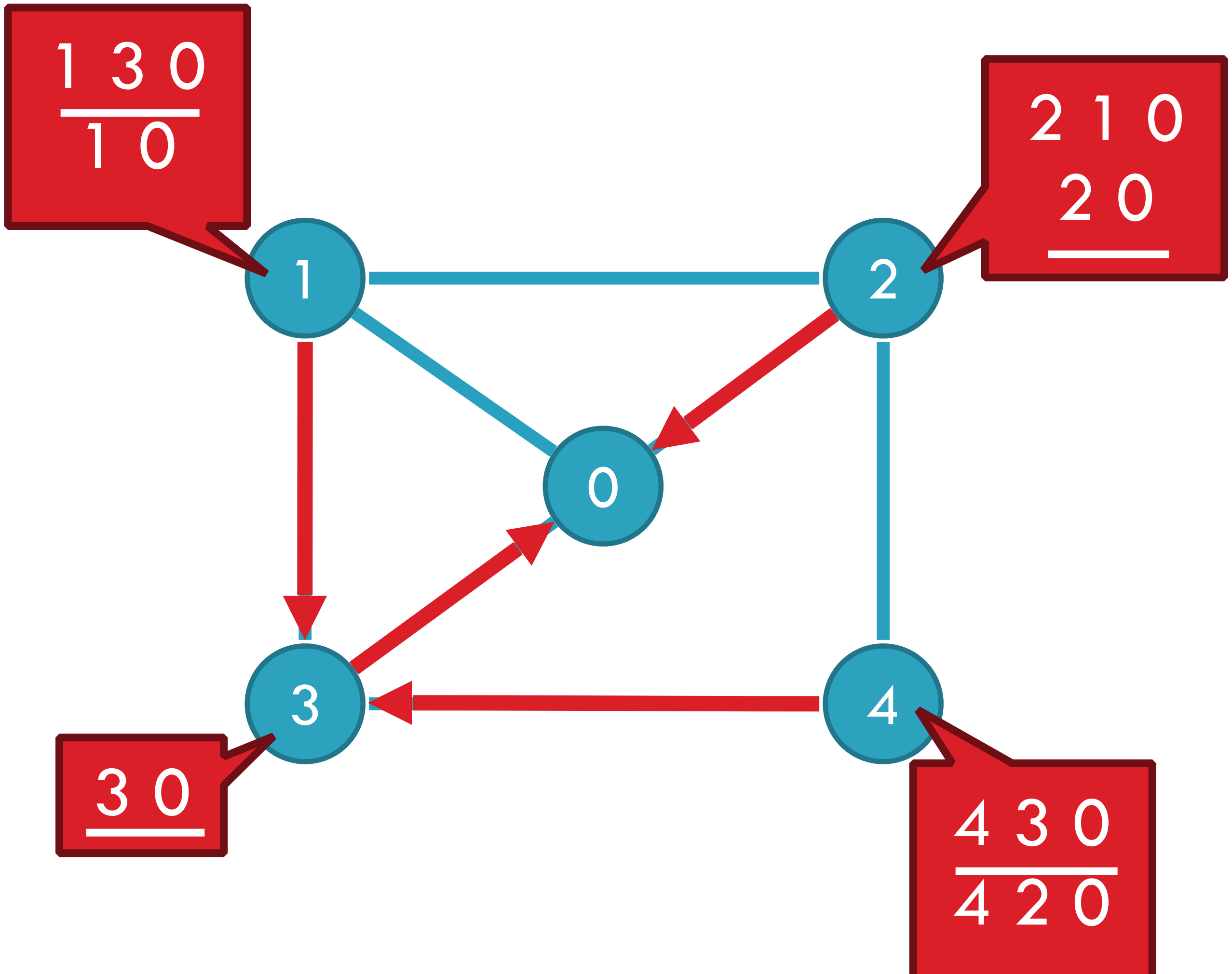
Good Gadget



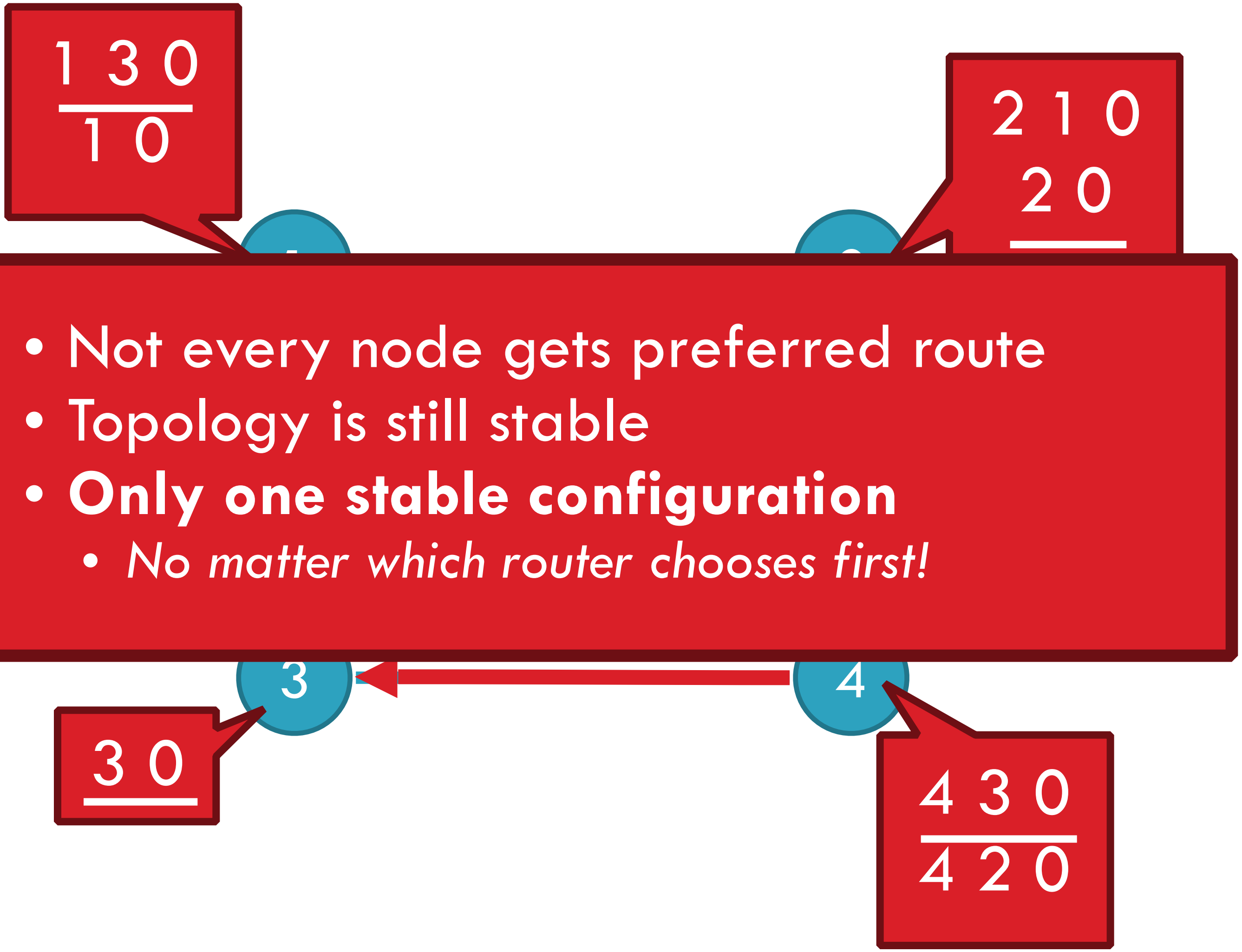
Good Gadget



Good Gadget

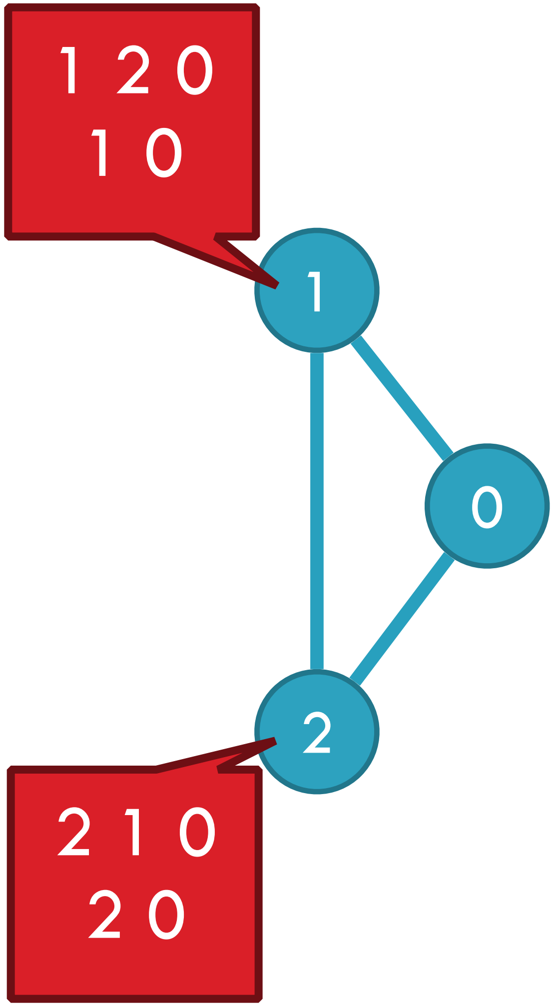


Good Gadget

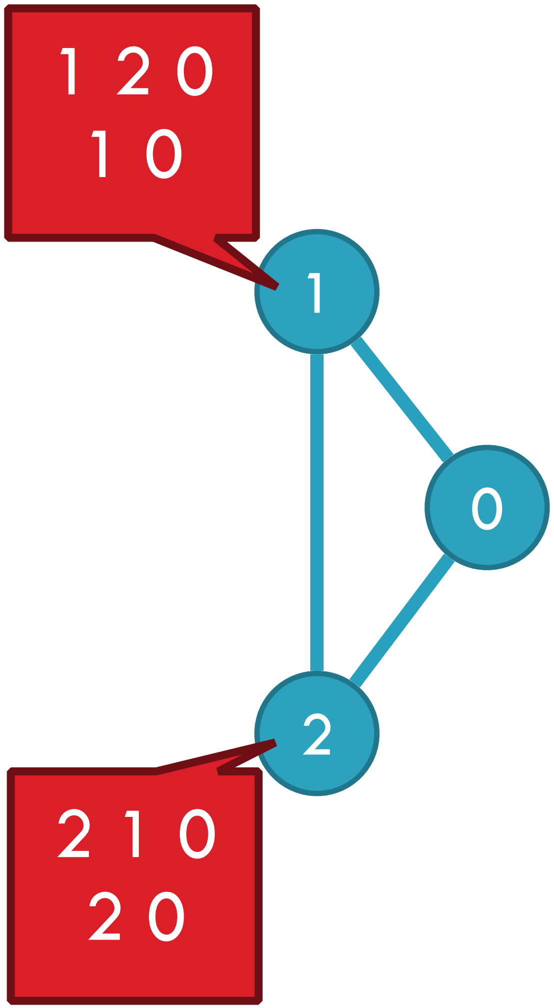


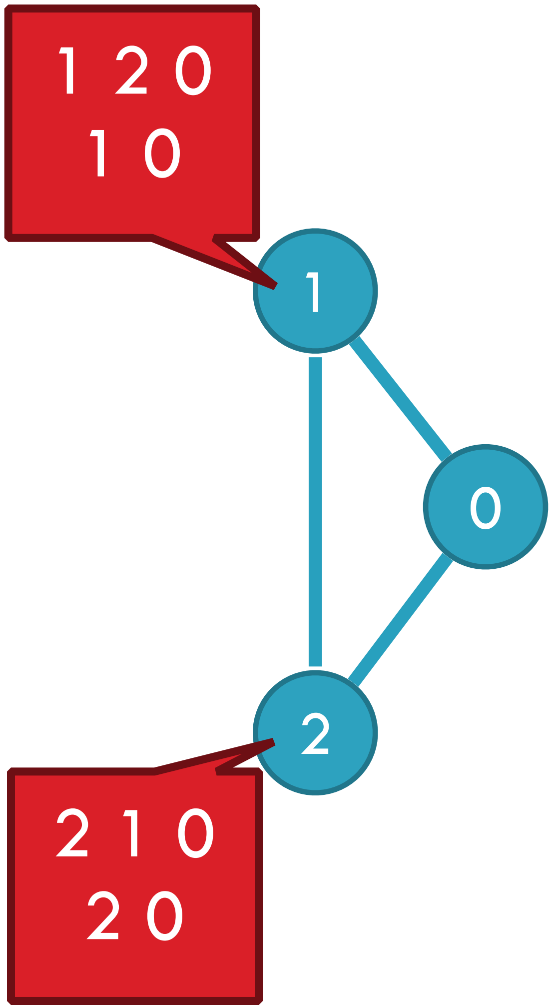
- Not every node gets preferred route
- Topology is still stable
- **Only one stable configuration**
  - *No matter which router chooses first!*

SPP May Have Multiple Solutions

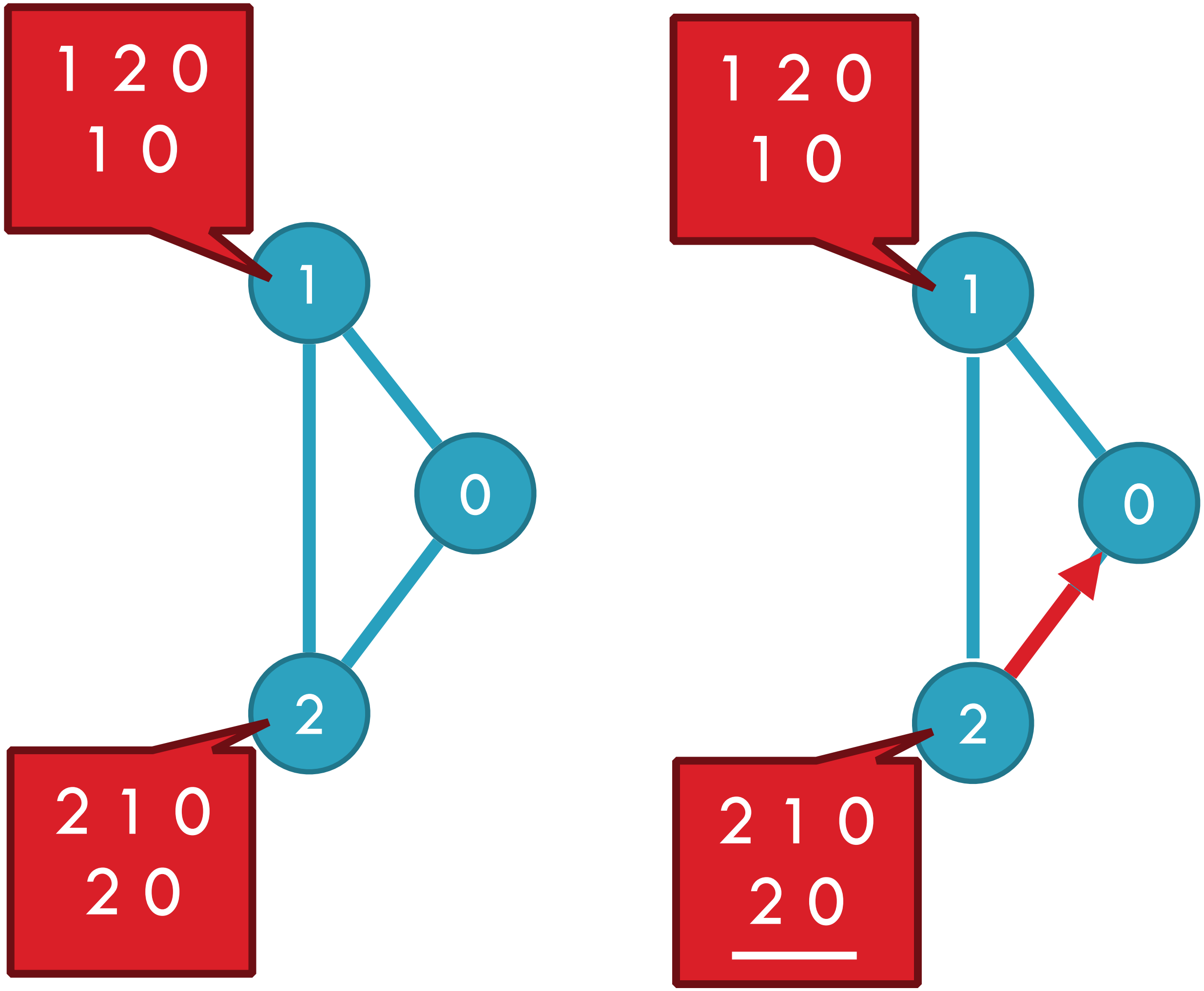


SPP May Have Multiple Solutions

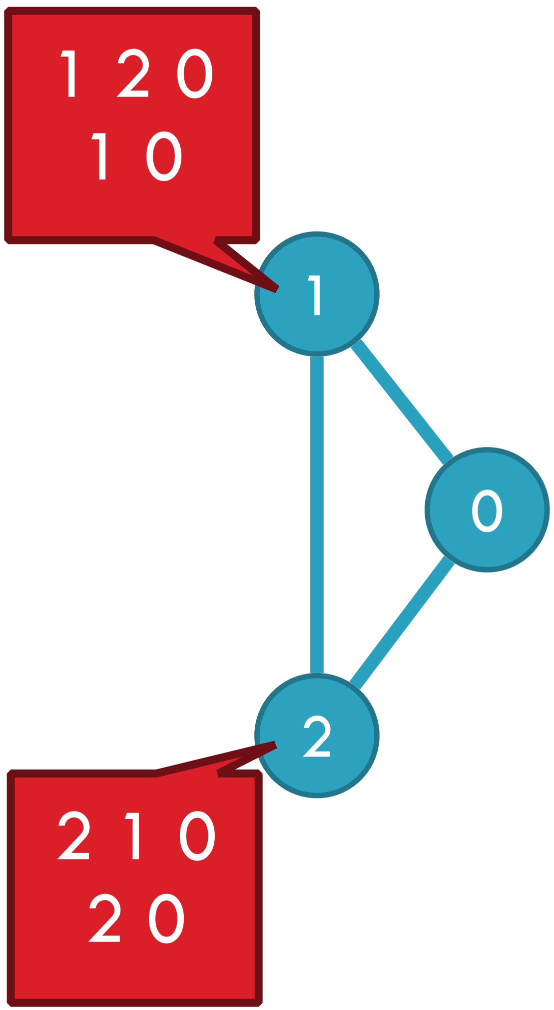


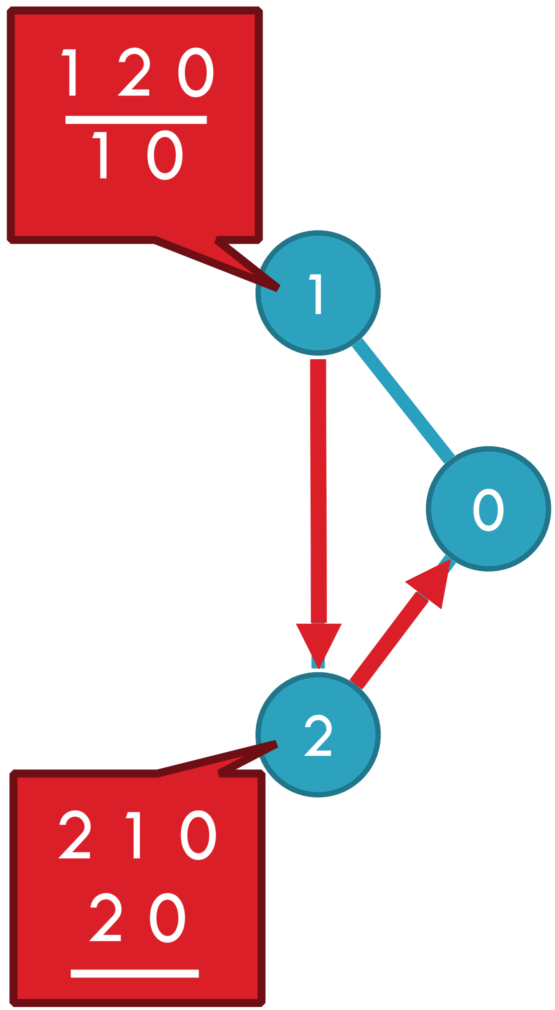


SPP May Have Multiple Solutions

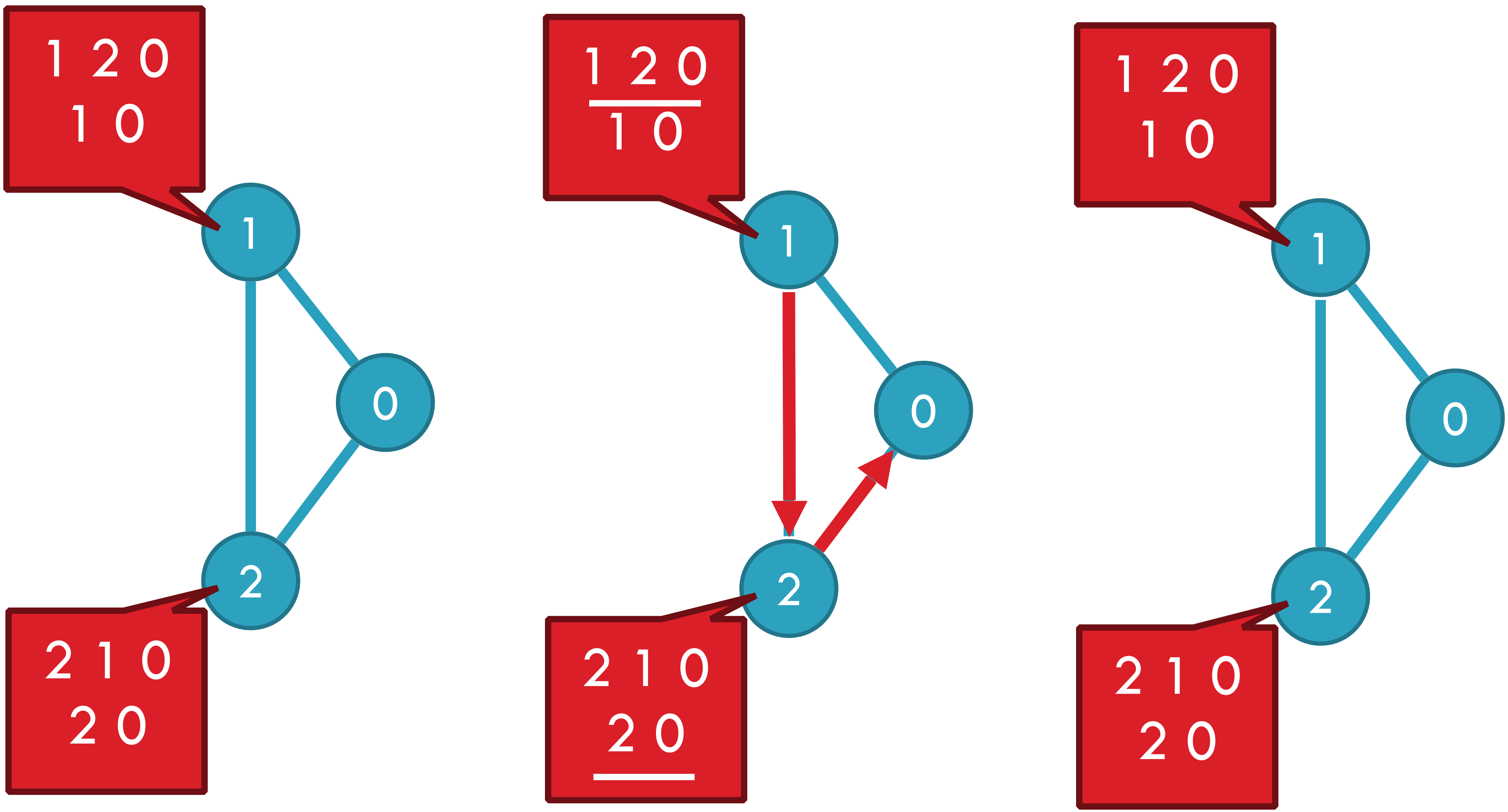


SPP May Have Multiple Solutions

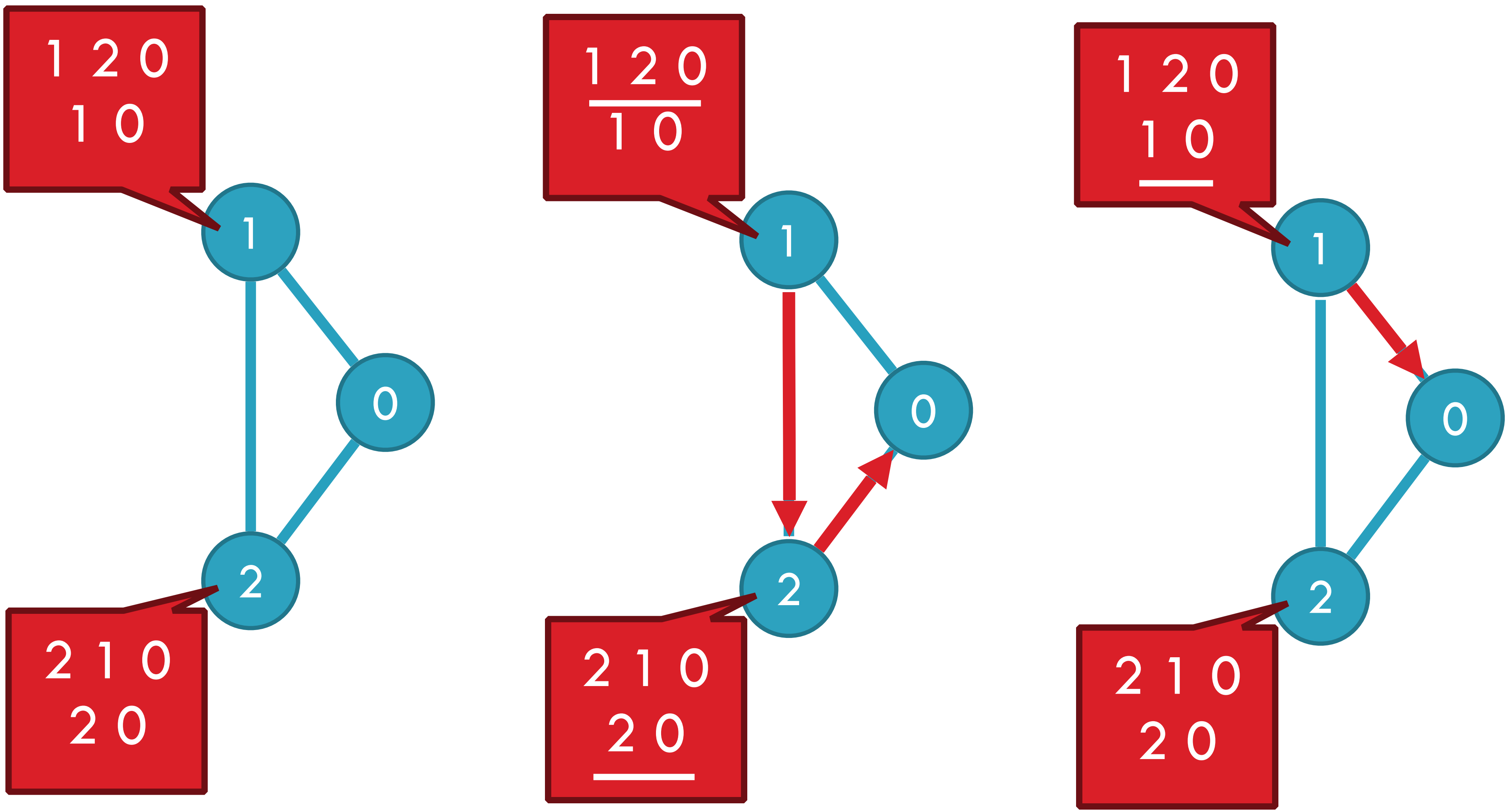




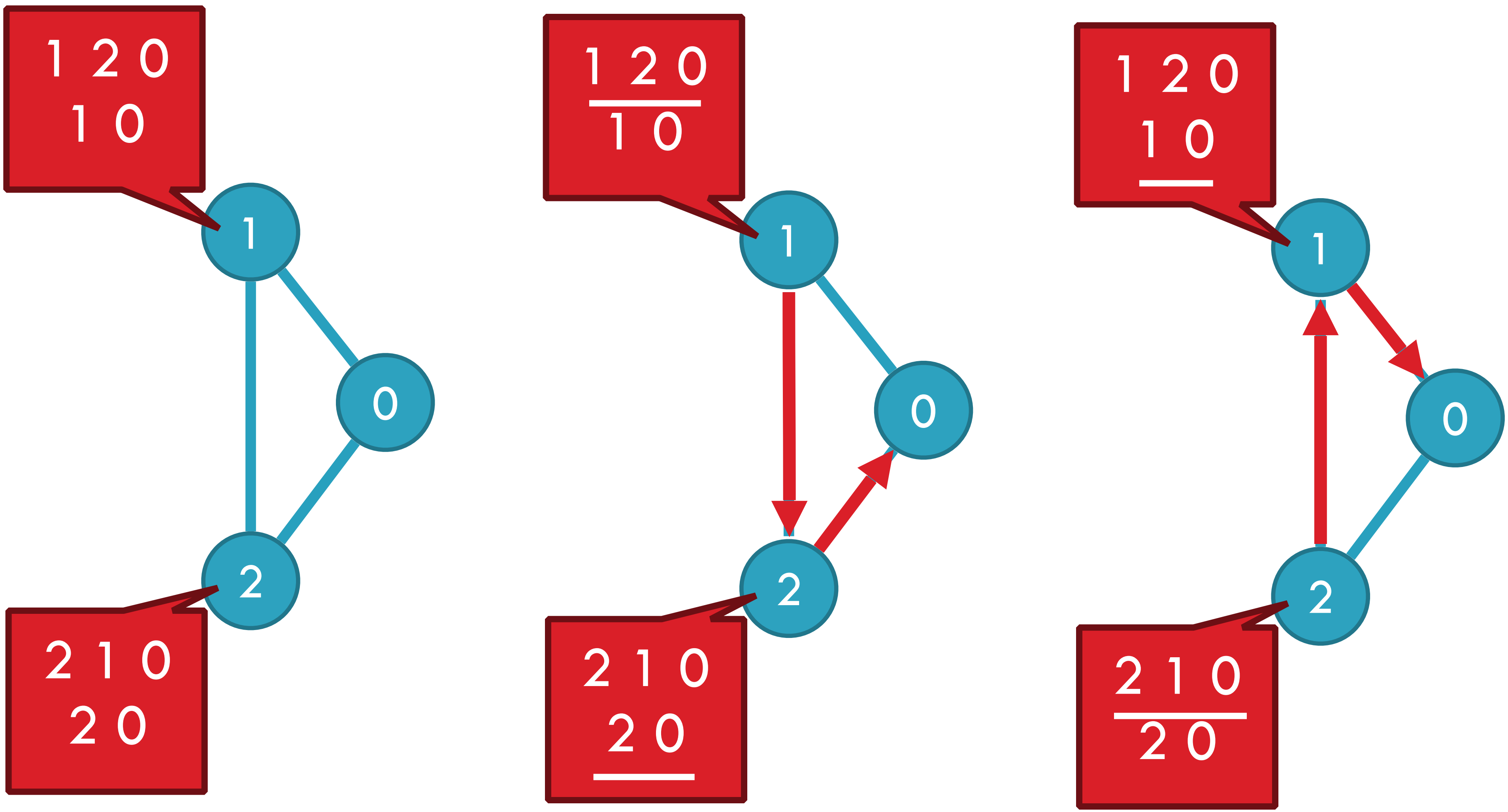
SPP May Have Multiple Solutions



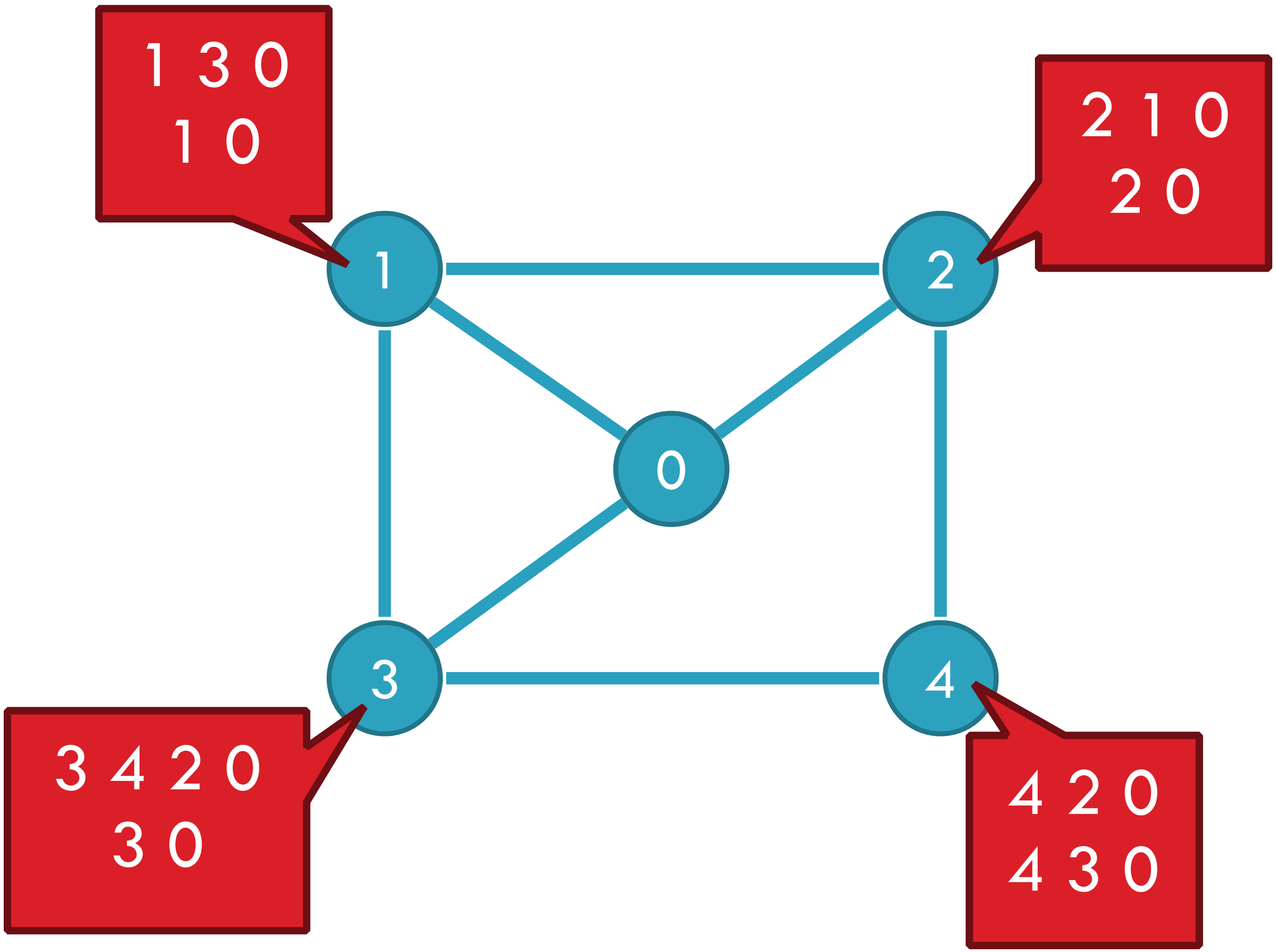
SPP May Have Multiple Solutions



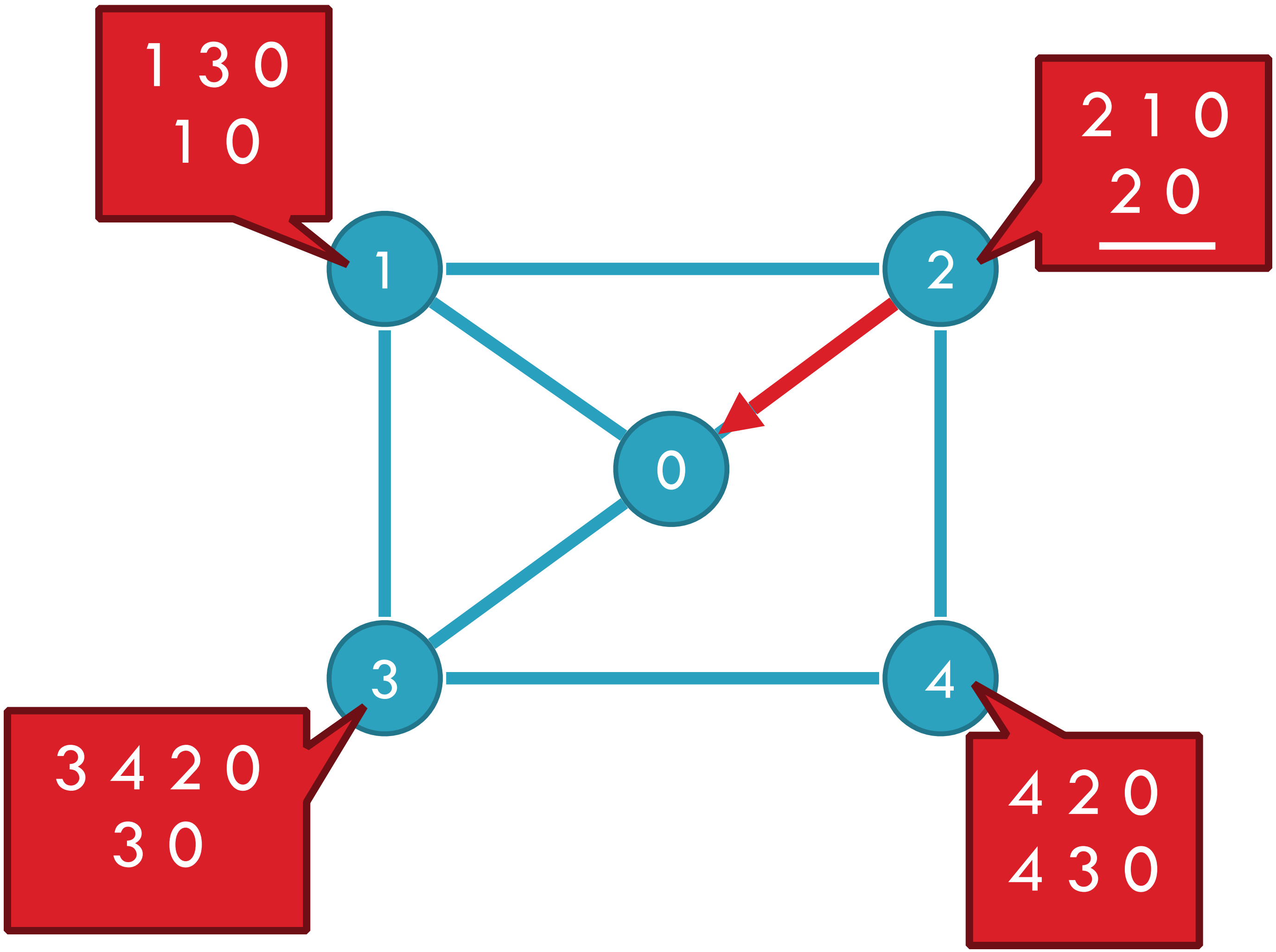
SPP May Have Multiple Solutions



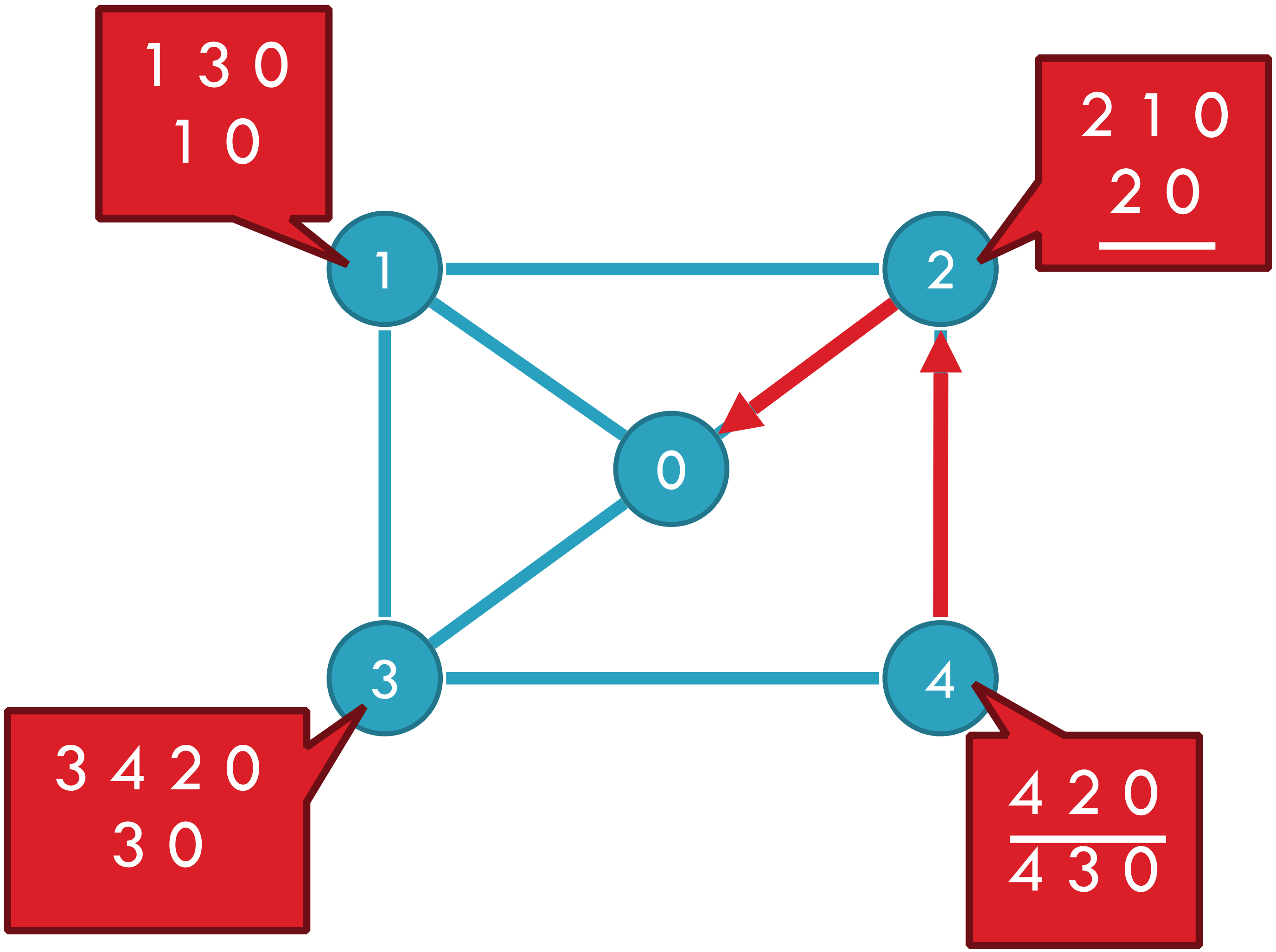
Bad Gadget



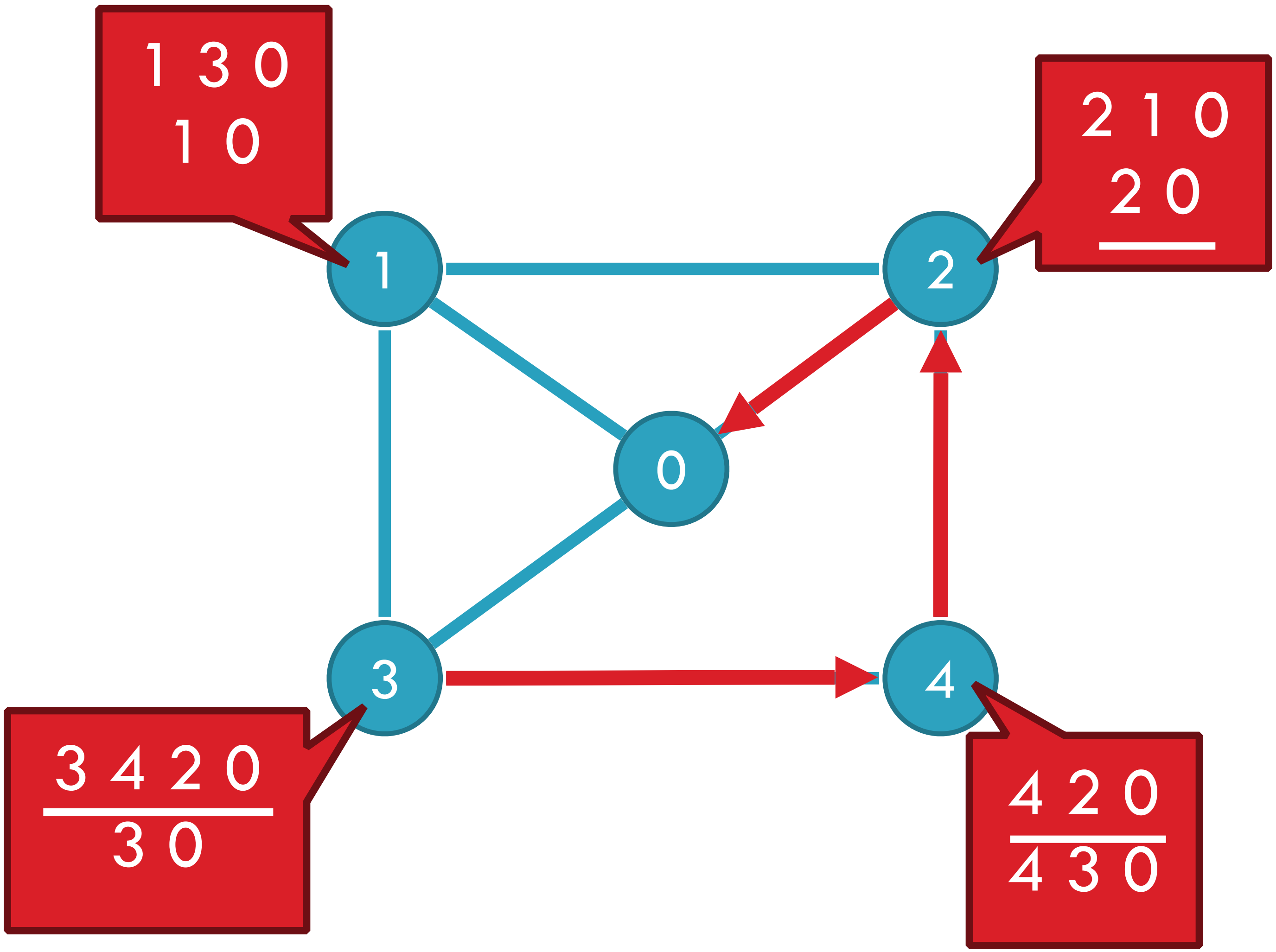
Bad Gadget



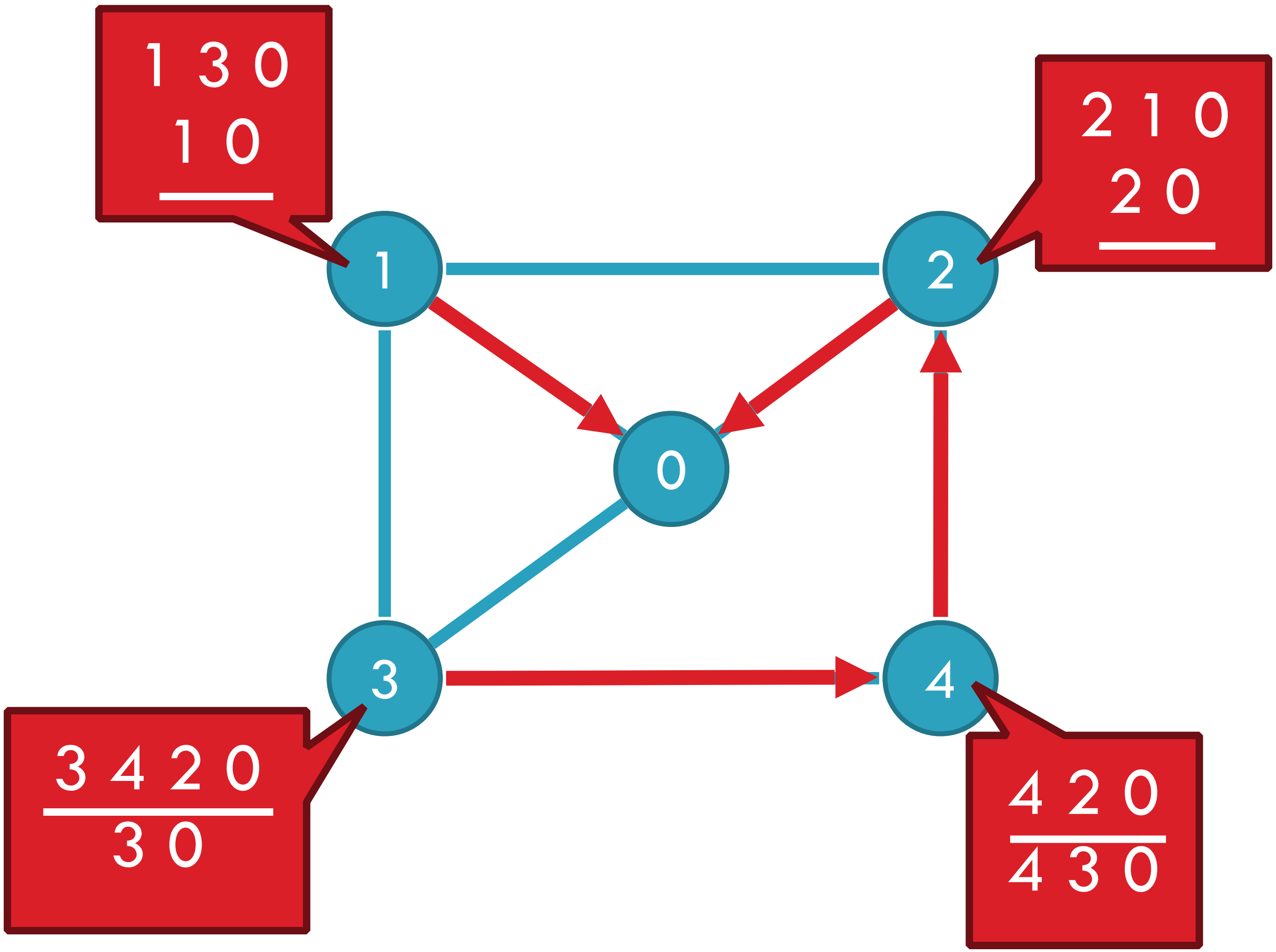
Bad Gadget



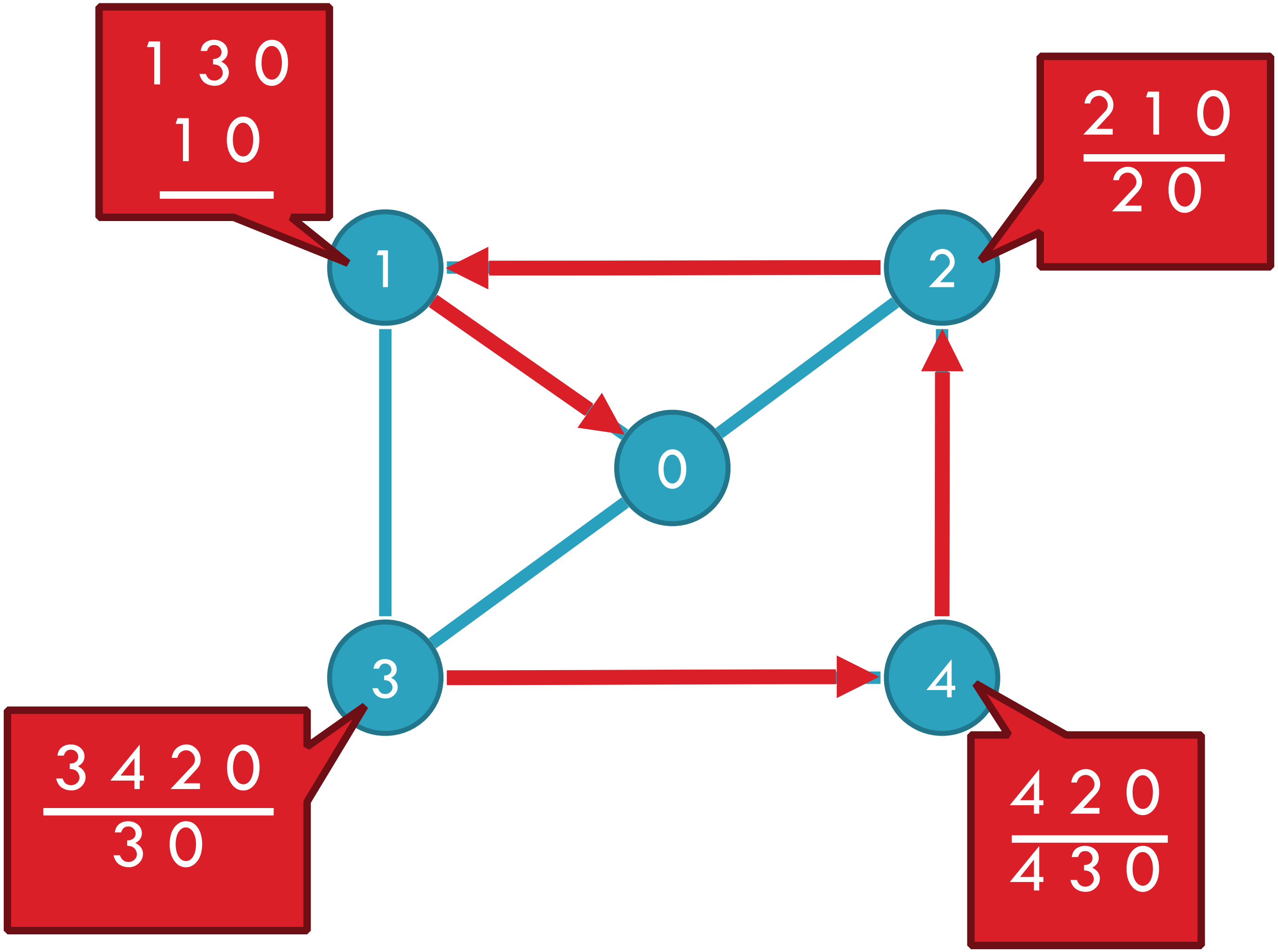
Bad Gadget



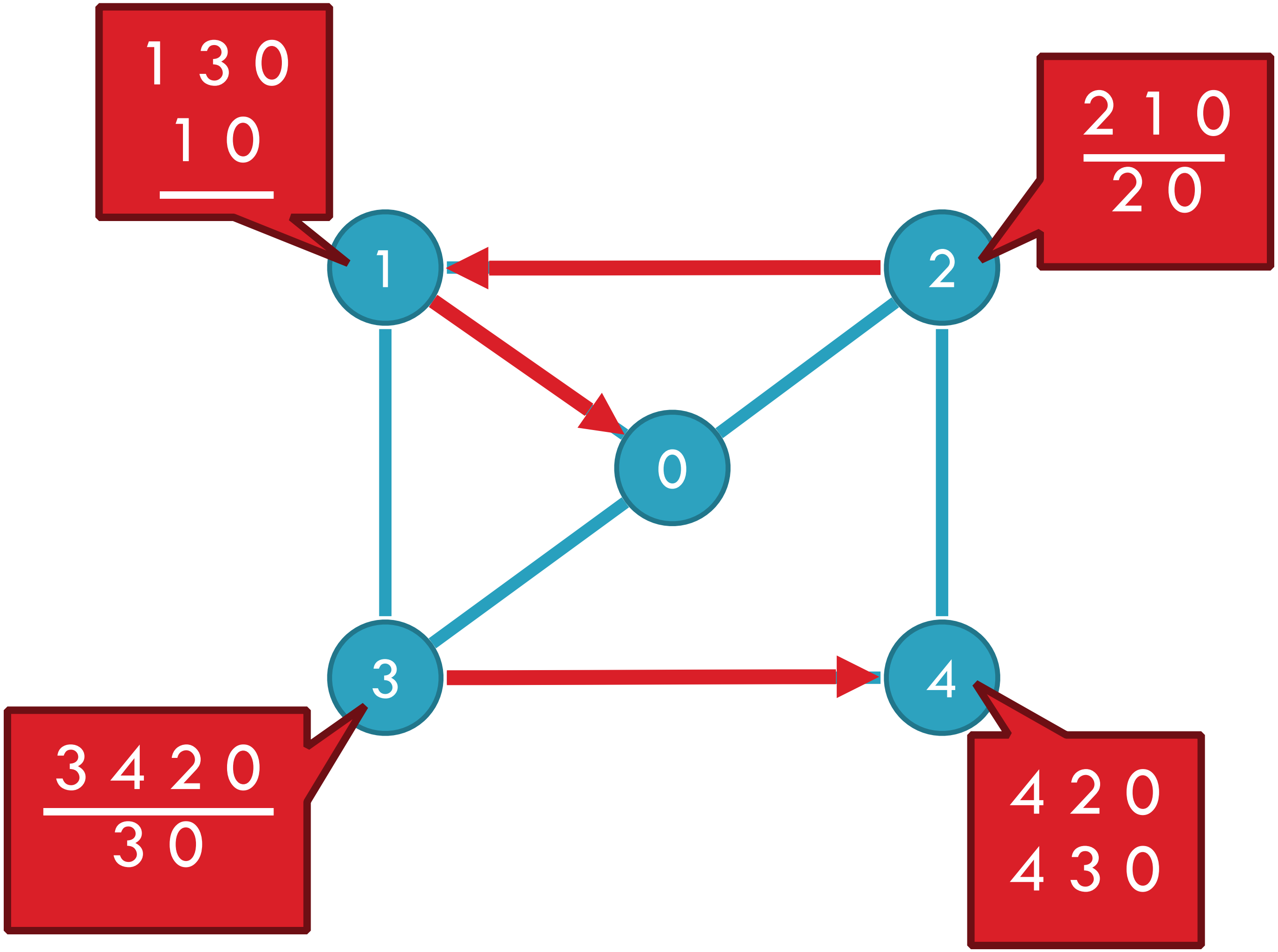
Bad Gadget



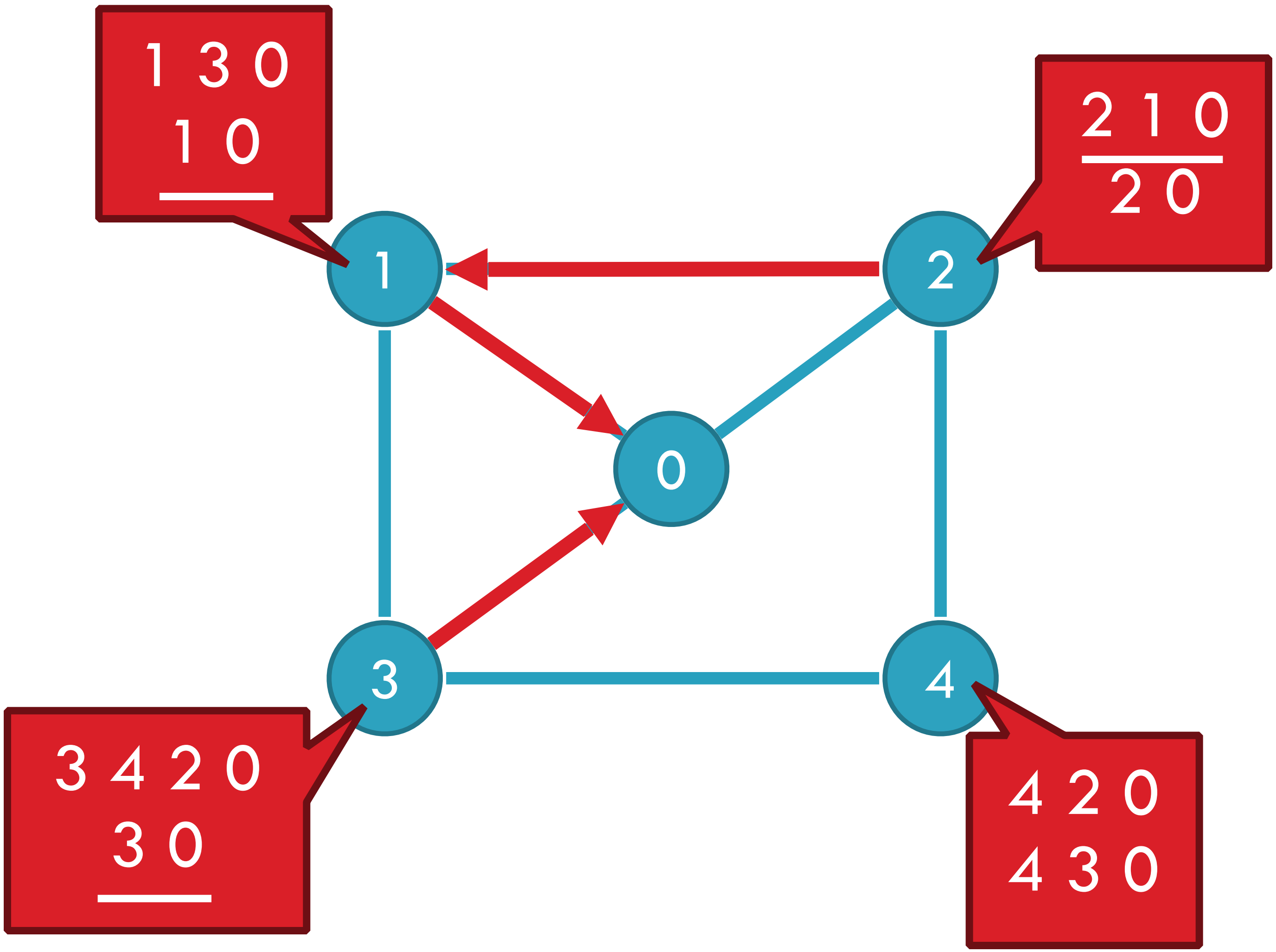
Bad Gadget



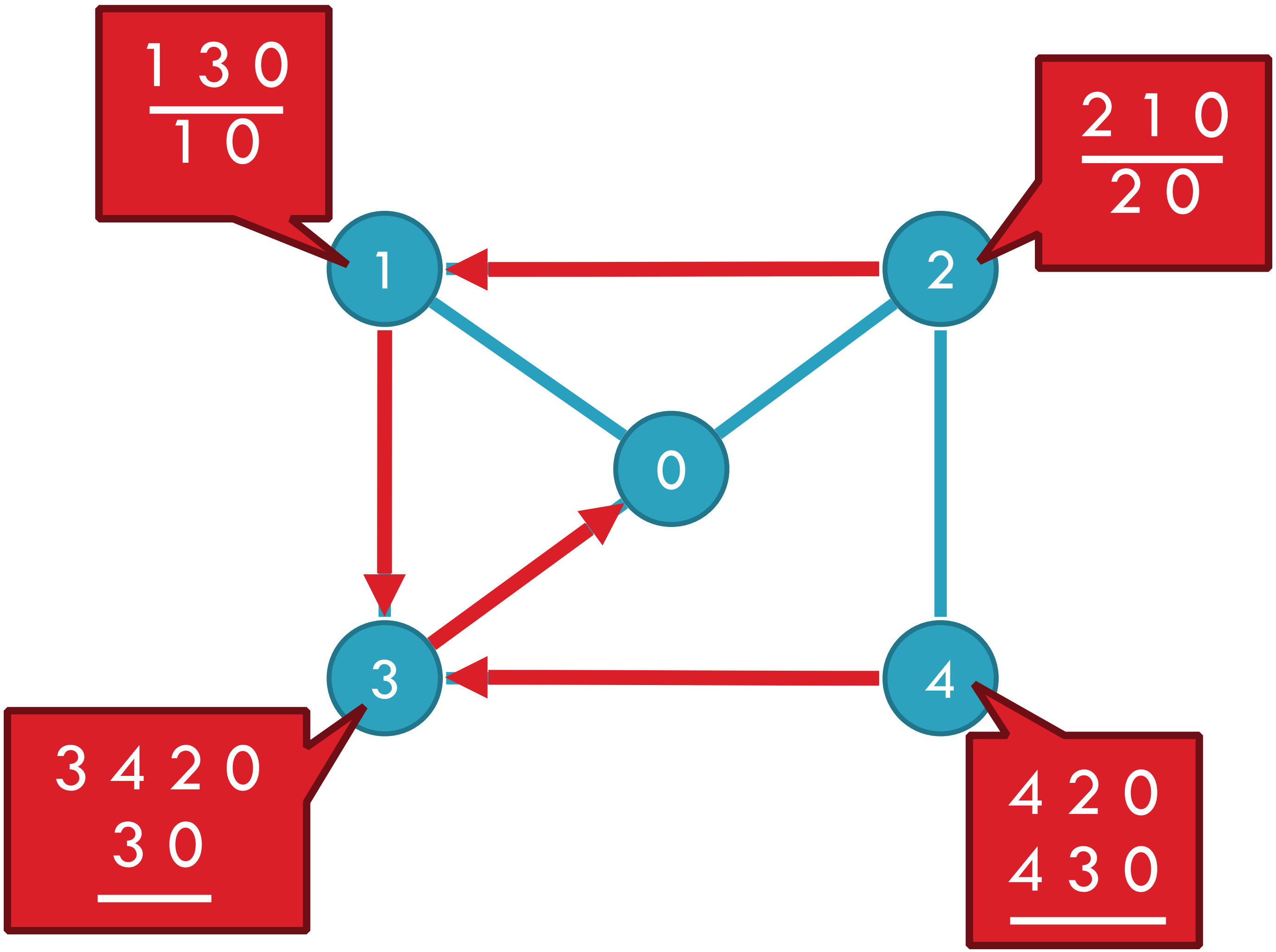
Bad Gadget



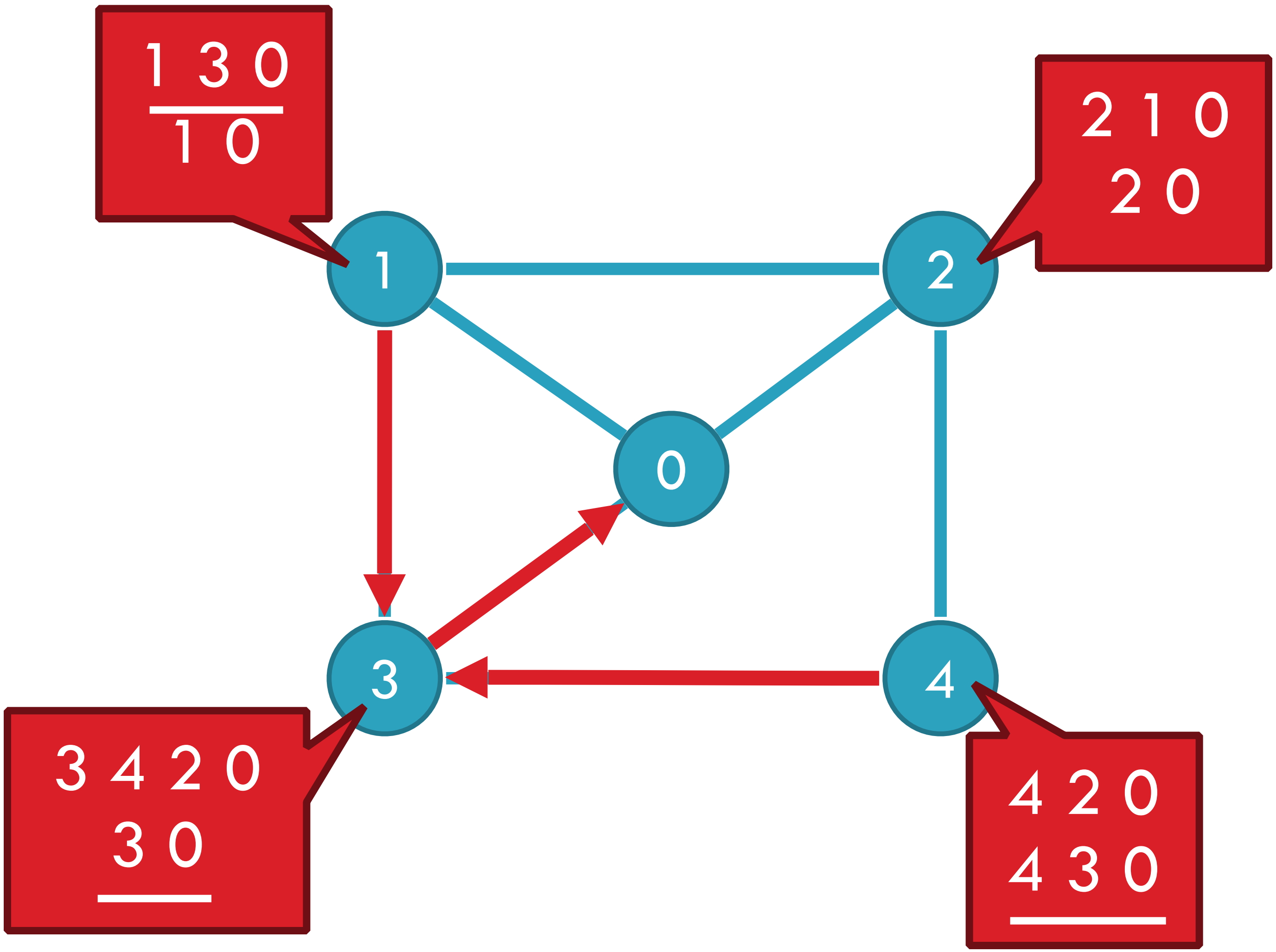
Bad Gadget



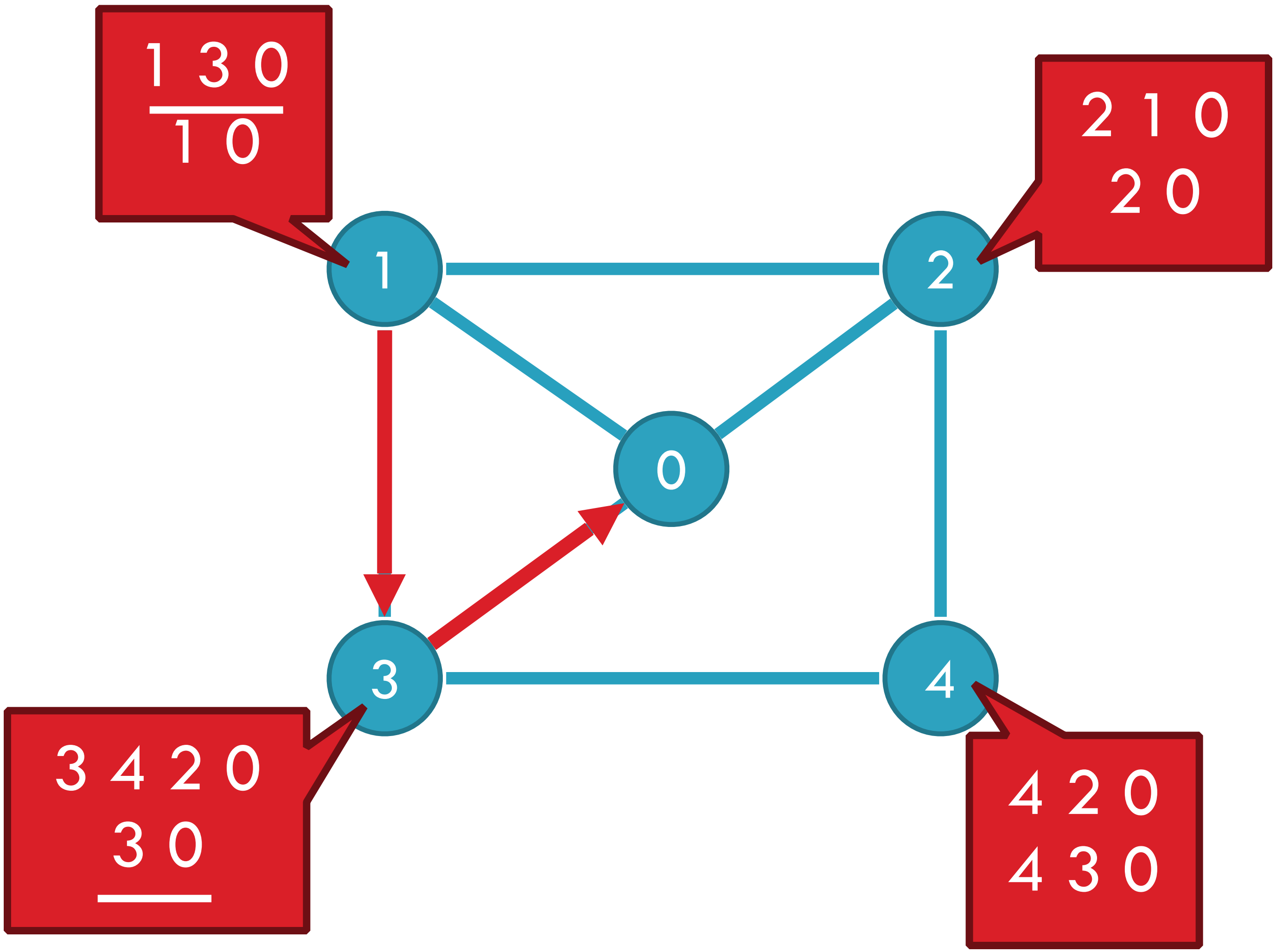
Bad Gadget



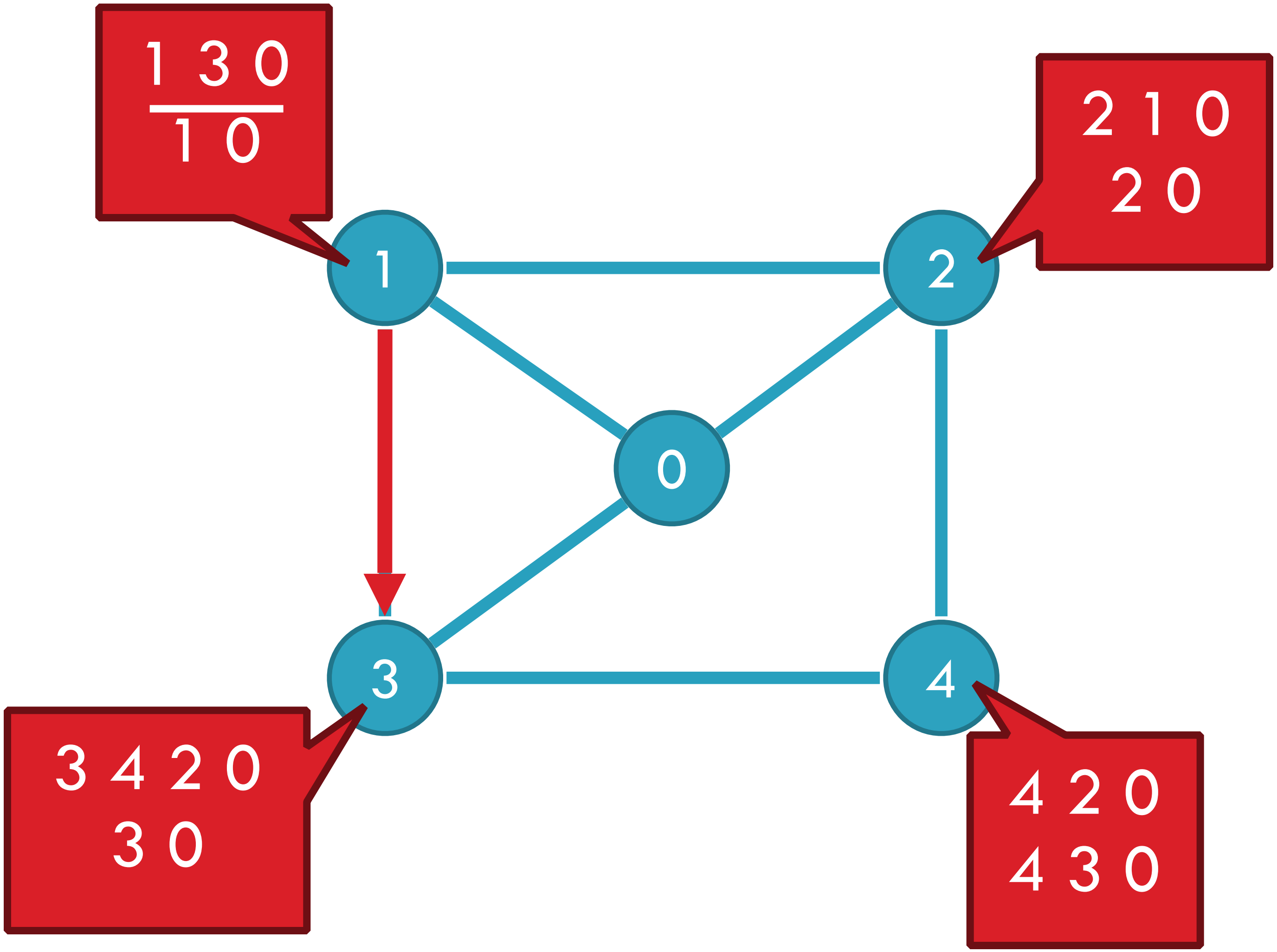
Bad Gadget



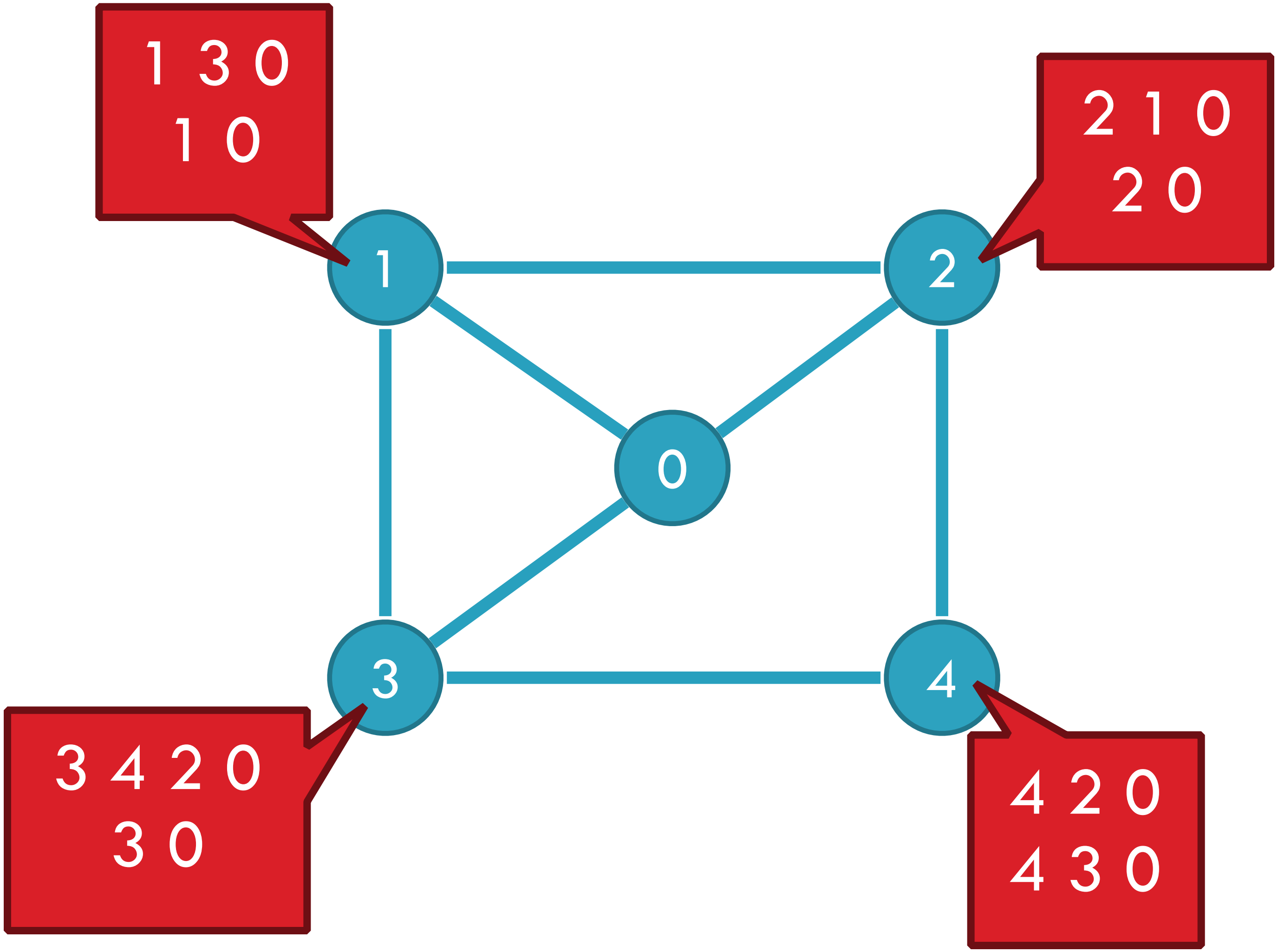
Bad Gadget



Bad Gadget



Bad Gadget



Bad Gadget

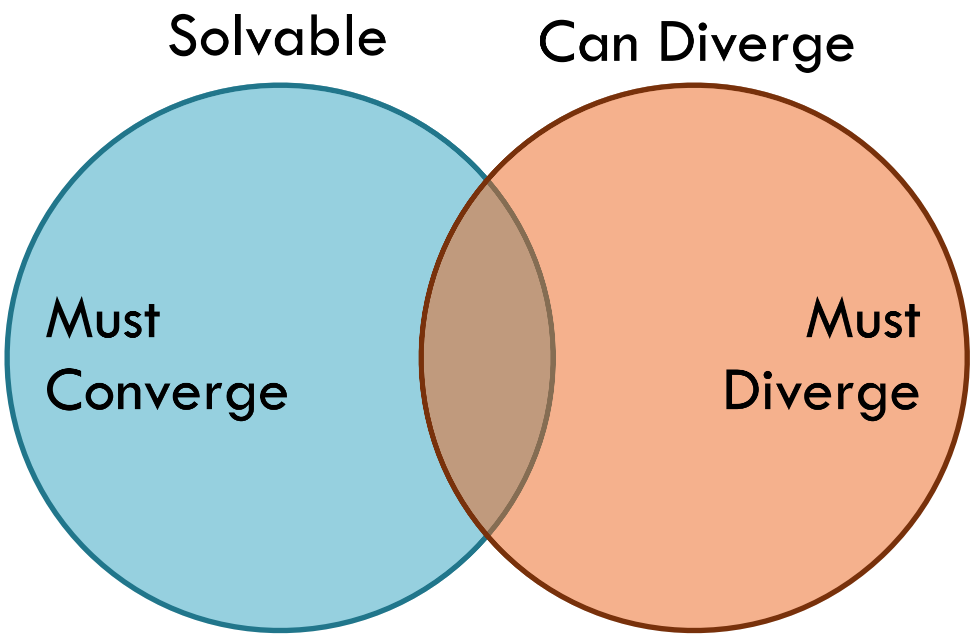
1 3 0

- That was only one round of oscillation!
- This keeps going, infinitely
- Problem stems from:
  - Local (not global) decisions
  - Ability of one node to improve its path selection

4 3 0

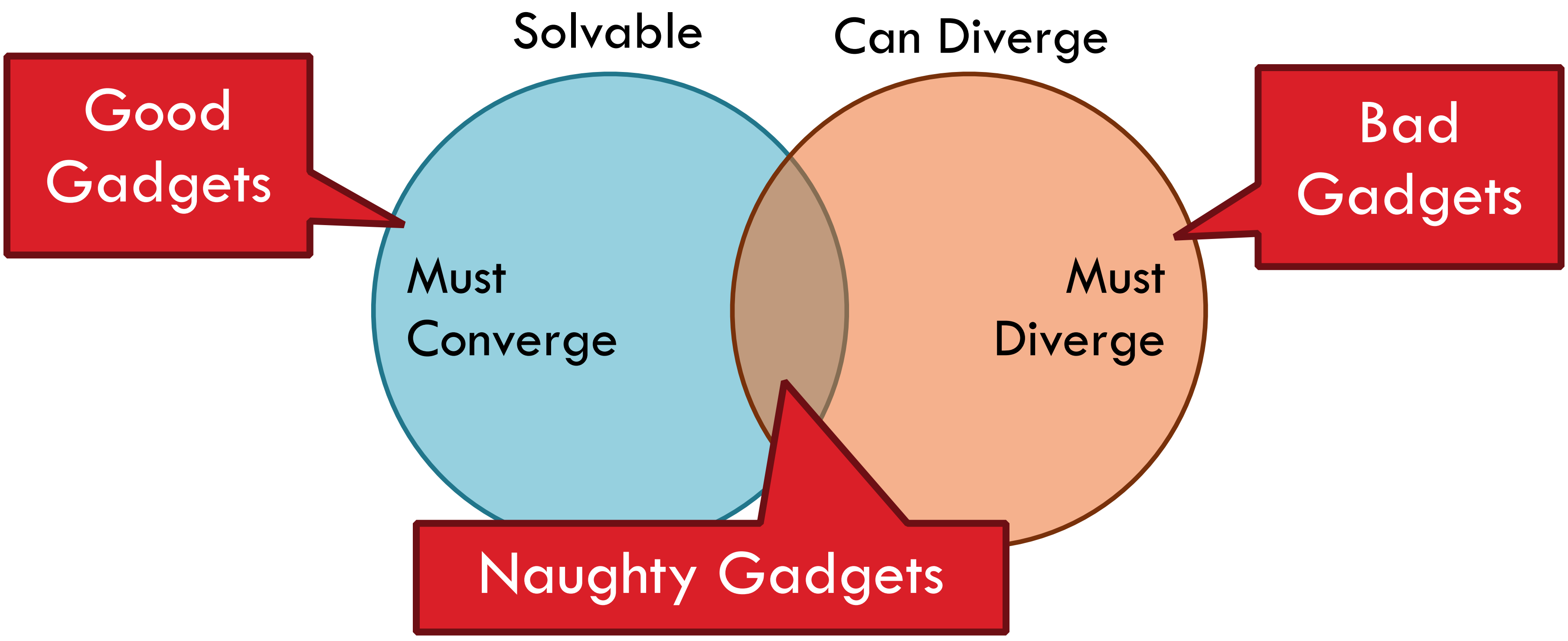
SPP Explains BGP Divergence

- BGP is not guaranteed to converge to stable routing
  - Policy inconsistencies may lead to “livelock”
  - Protocol oscillation

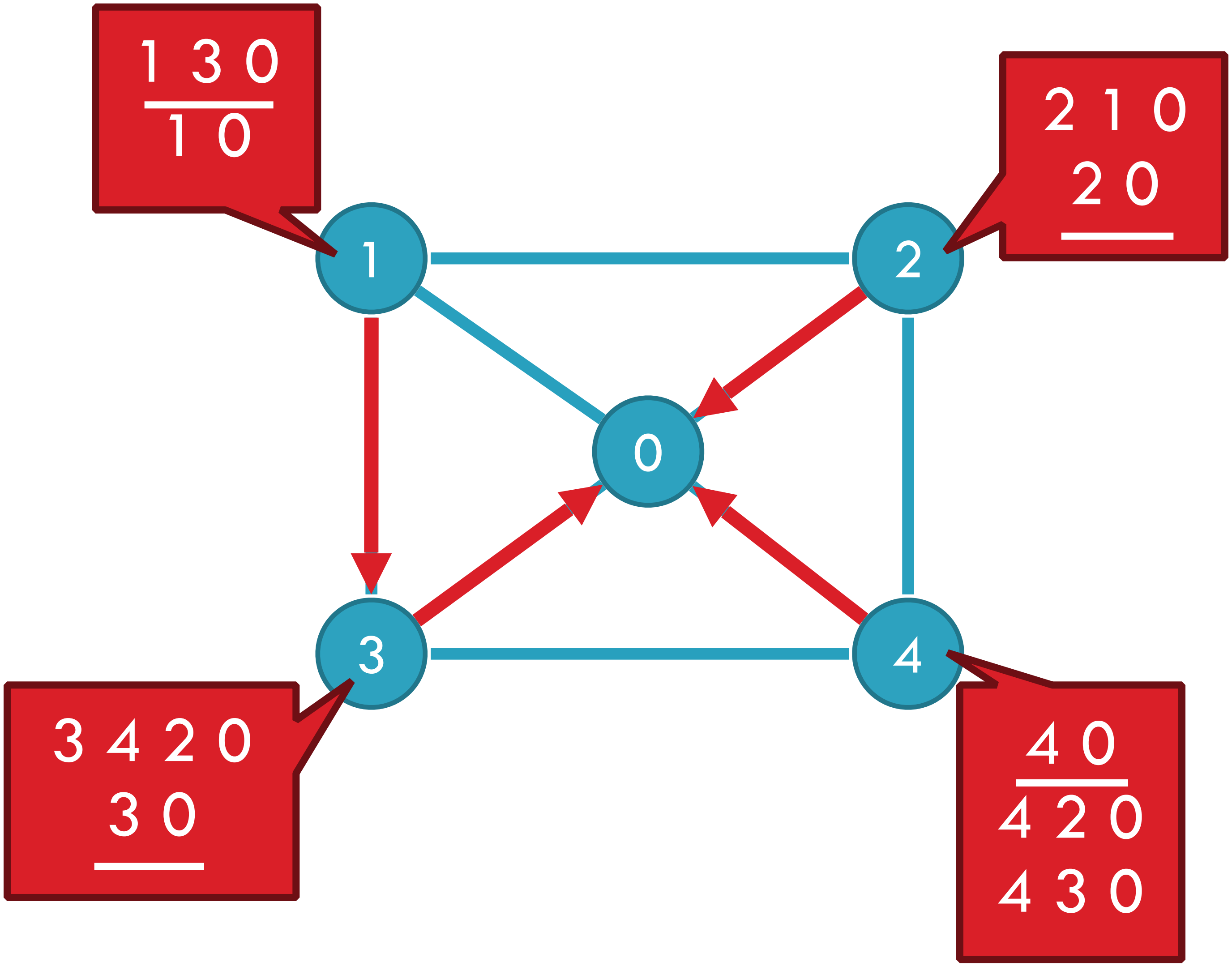


SPP Explains BGP Divergence

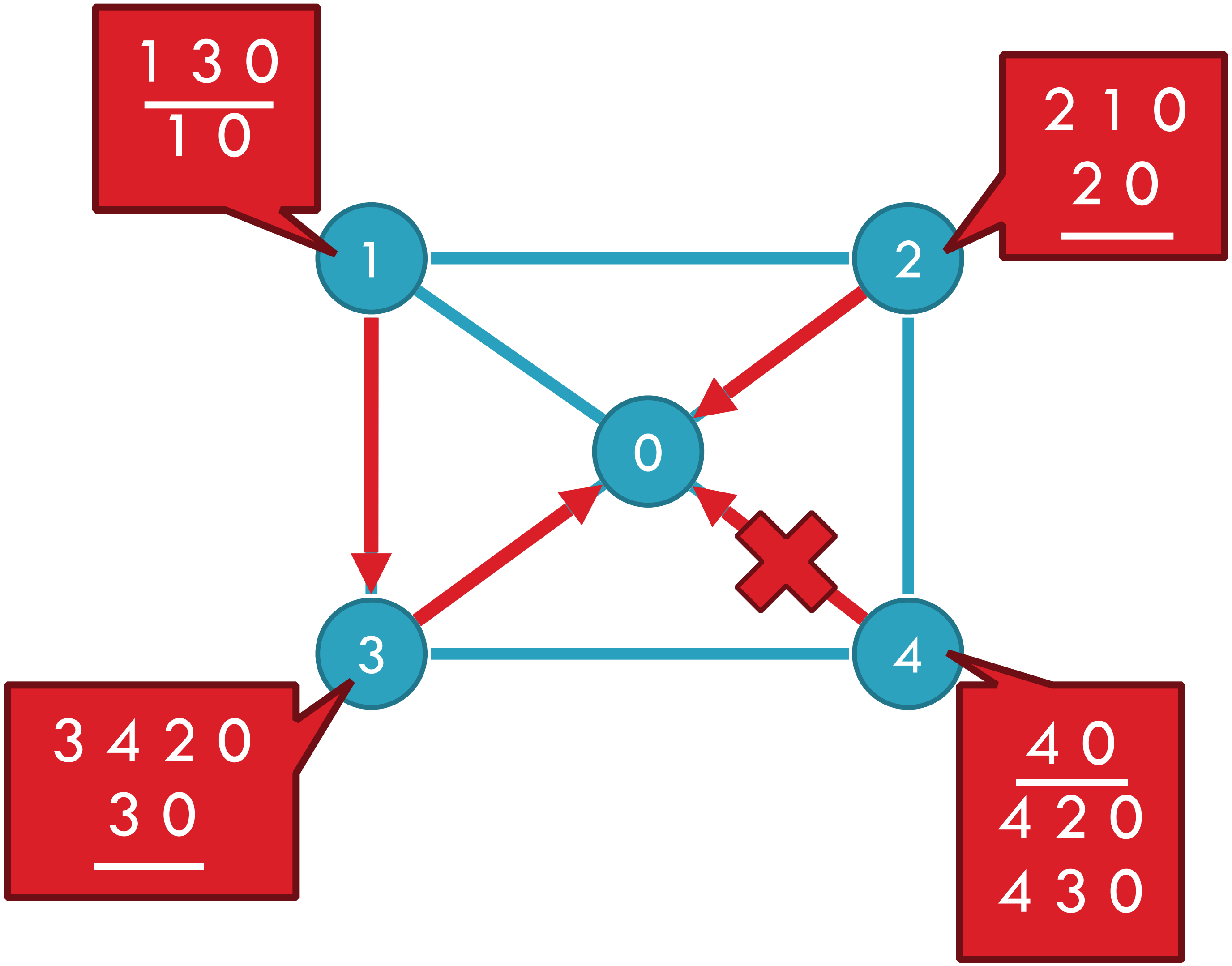
- BGP is not guaranteed to converge to stable routing
  - Policy inconsistencies may lead to "livelock"
  - Protocol oscillation



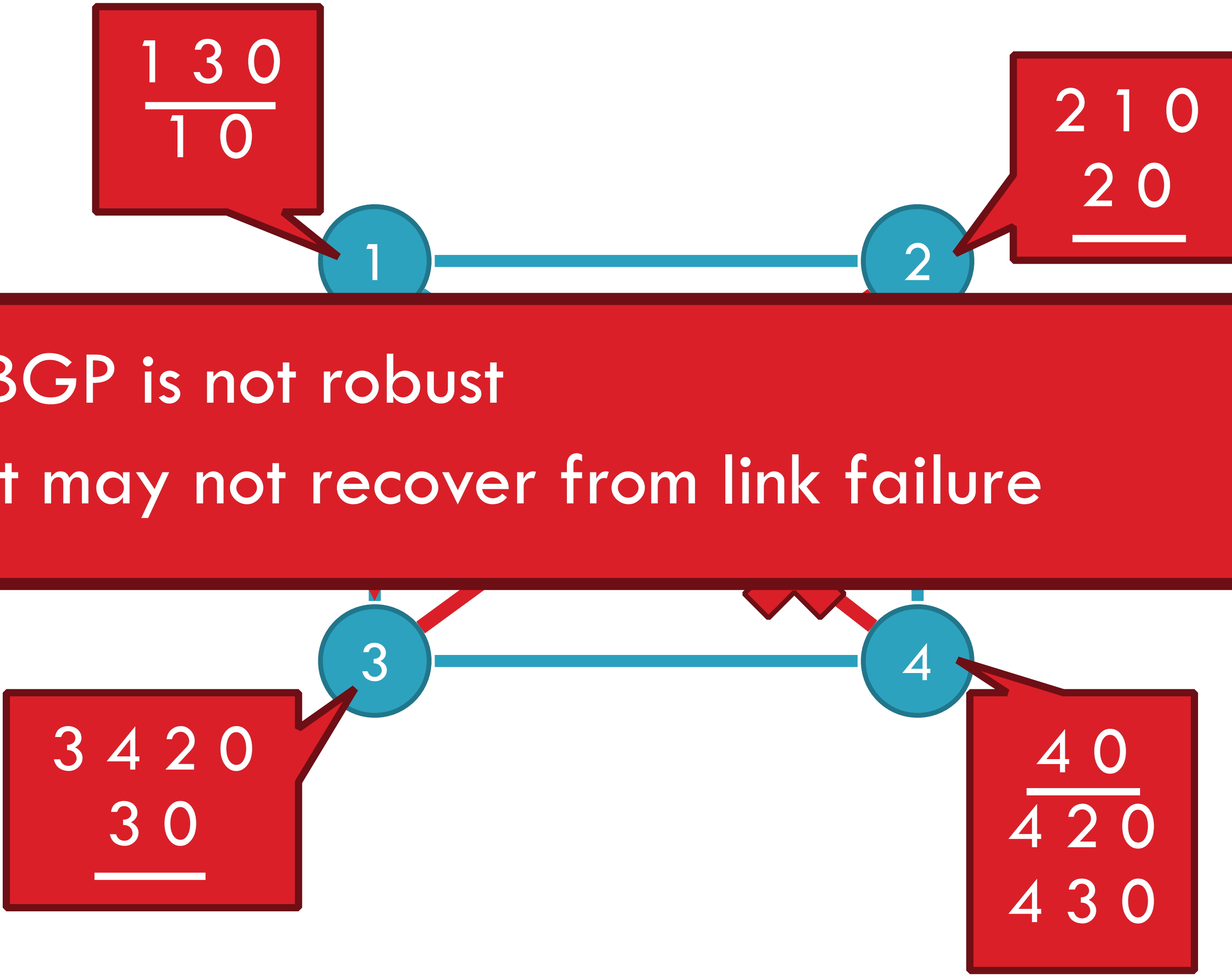
Beware of Backup Policies



Beware of Backup Policies

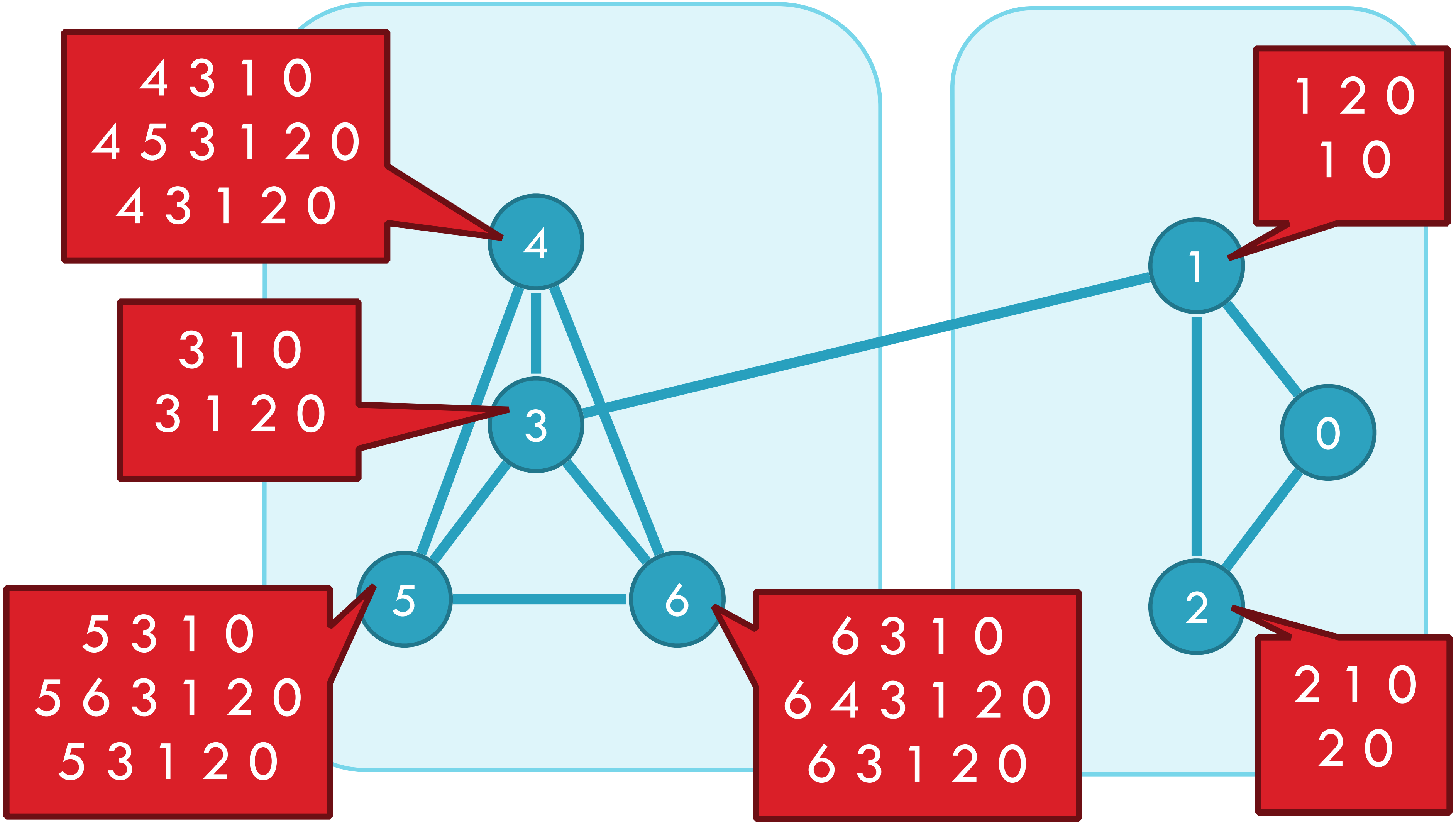


Beware of Backup Policies

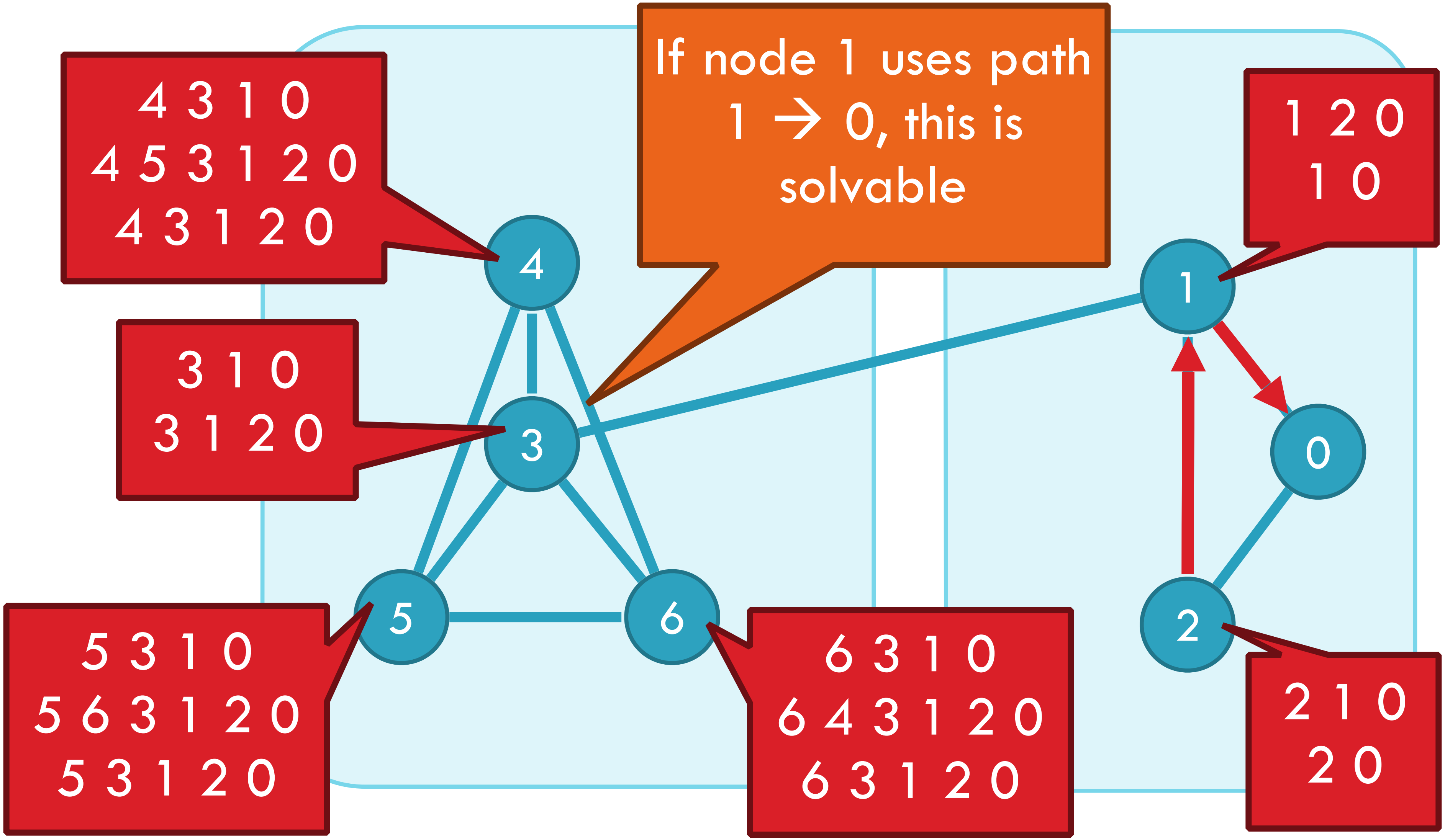


- BGP is not robust
- It may not recover from link failure

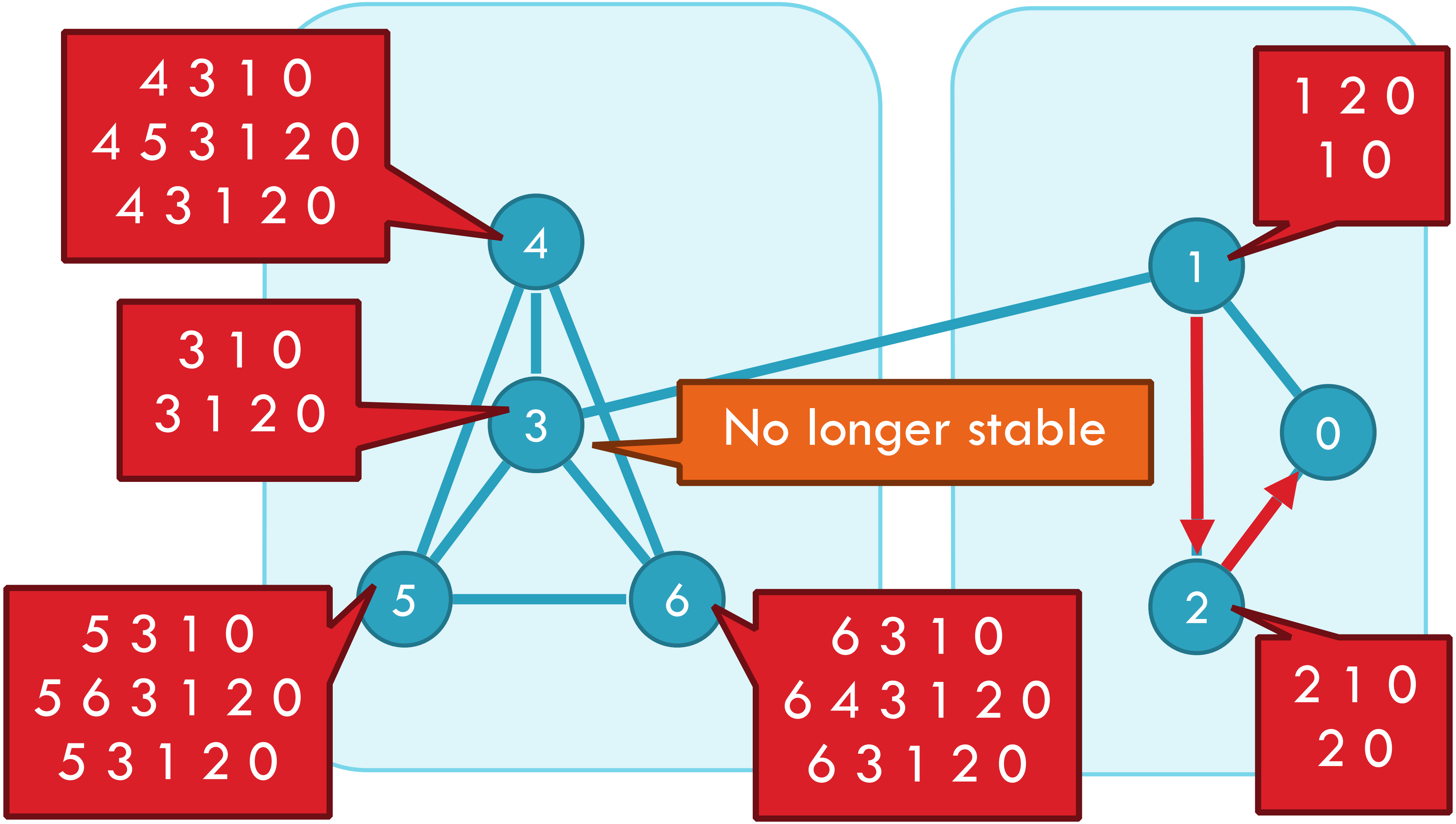
BGP is Precarious



BGP is Precarious



BGP is Precarious



Can BGP Be Fixed?

- Unfortunately, SPP is NP-complete

Can BGP Be Fixed?

- Unfortunately, SPP is NP-complete

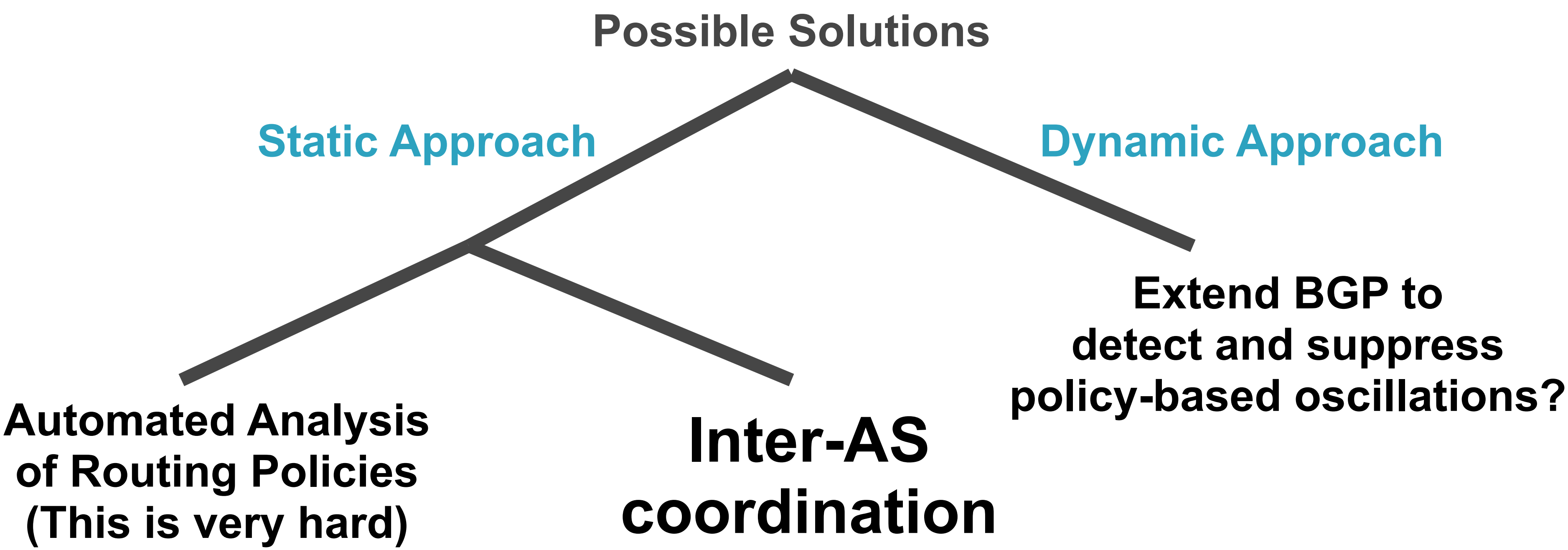
Possible Solutions

Dynamic Approach

Extend BGP to detect and suppress policy-based oscillations?

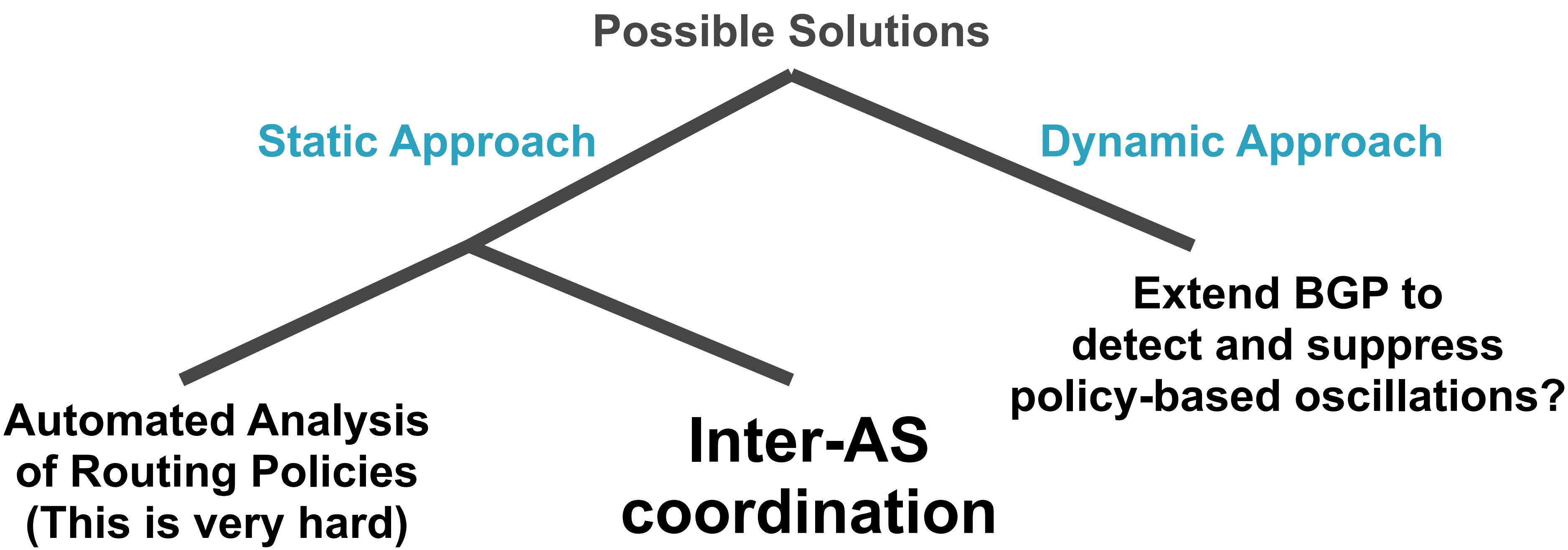
Can BGP Be Fixed?

- Unfortunately, SPP is NP-complete



Can BGP Be Fixed?

- Unfortunately, SPP is NP-complete



These approaches are complementary

Outline

- BGP Basics
- Stable Paths Problem
- BGP in the Real World

Motivation

- Routing reliability/fault-tolerance on small time scales (minutes) not previously a priority
- Transaction oriented and interactive applications (e.g. Internet Telephony) will require higher levels of end-to-end network reliability
- How well does the Internet routing infrastructure tolerate faults?

Conventional Wisdom

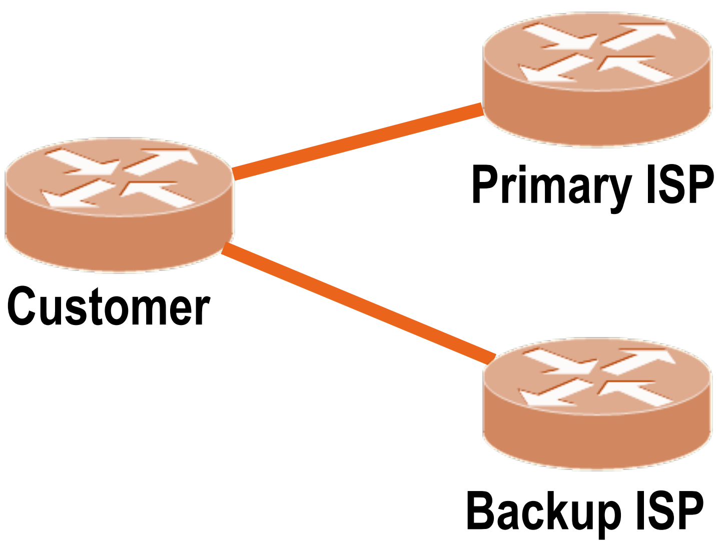
- Internet routing is robust under faults
  - Supports path re-routing
  - Path restoration on the order of seconds
- BGP has good convergence properties
  - Does not exhibit looping/bouncing problems of RIP
- Internet fail-over will improve with faster routers and faster links
- More redundant connections (multi-homing) will always improve fault-tolerance

Delayed Routing Convergence

- Conventional wisdom about routing convergence is not accurate
  - Measurement of BGP convergence in the Internet
  - Analysis/intuition behind delayed BGP routing convergence
  - Modifications to BGP implementations which would improve convergence times

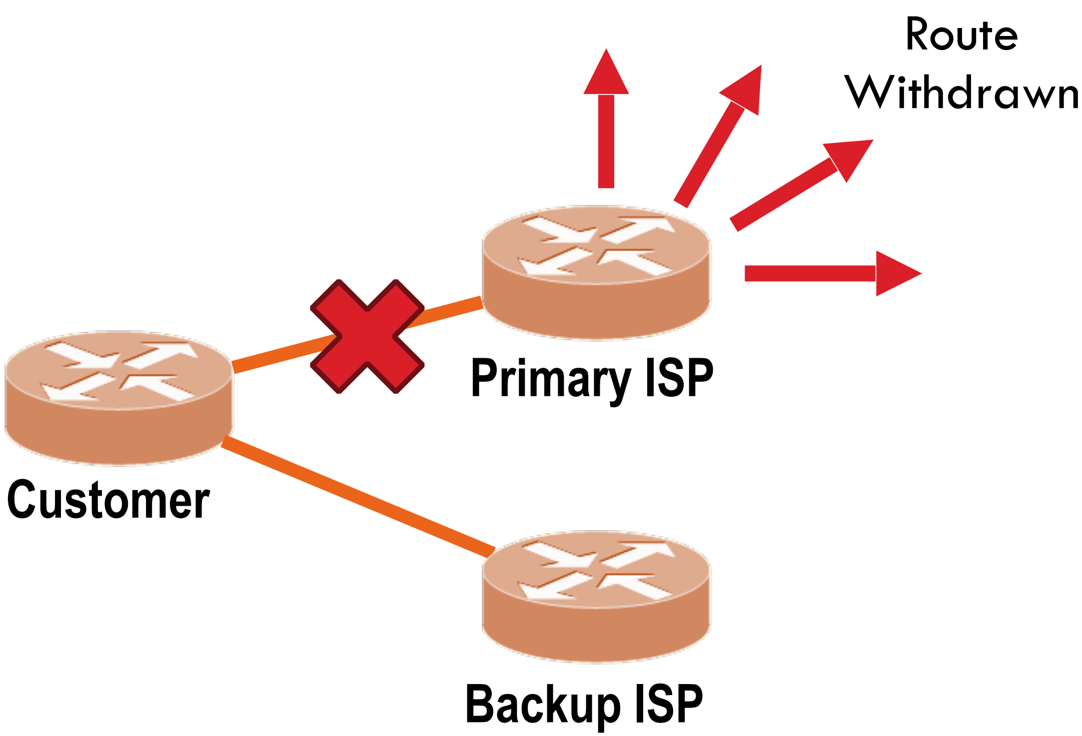
Open Question

- After a fault in a path to multi-homed site, how long does it take for majority of Internet routers to fail-over to secondary path?



Open Question

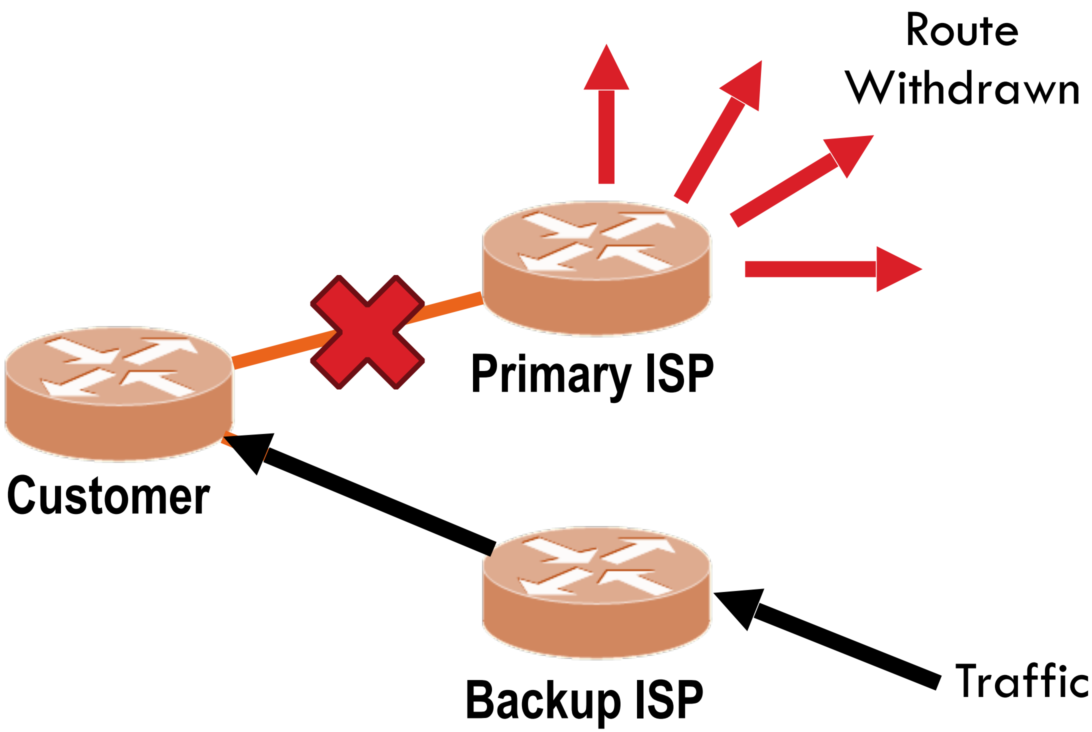
- After a fault in a path to multi-homed site, how long does it take for majority of Internet routers to fail-over to secondary path?



- Routing table convergence

Open Question

- After a fault in a path to multi-homed site, how long does it take for majority of Internet routers to fail-over to secondary path?



- Routing table convergence
- Stable end-to-end paths

Bad News

- With unconstrained policies:
  - Divergence
  - Possible create unsatisfiable policies
  - NP-complete to identify these policies
  - Happening today?

Bad News

- With unconstrained policies:
  - Divergence
  - Possible create unsatisfiable policies
  - NP-complete to identify these policies
  - Happening today?
- With constrained policies (e.g. shortest path first)
  - Transient oscillations
  - BGP usually converges
  - It may take a very long time…

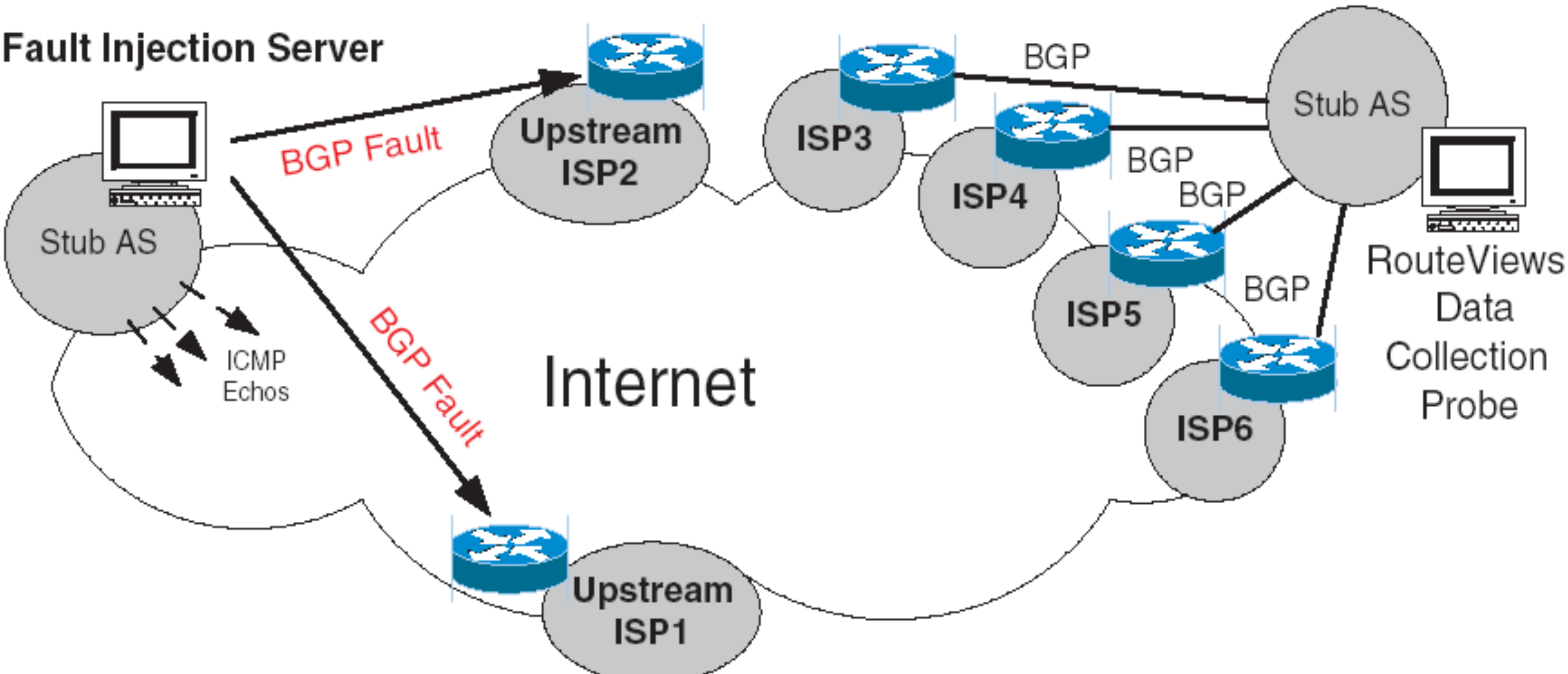
Bad News

- With unconstrained policies:
  - Divergence
  - Possible create unsatisfiable policies
  - NP-complete to identify these policies
  - Happening today?
- With constrained policies (e.g. shortest path first)
  - Transient oscillations
  - BGP usually converges
  - It may take a very long time…
- BGP Beacons: focuses on constrained policies

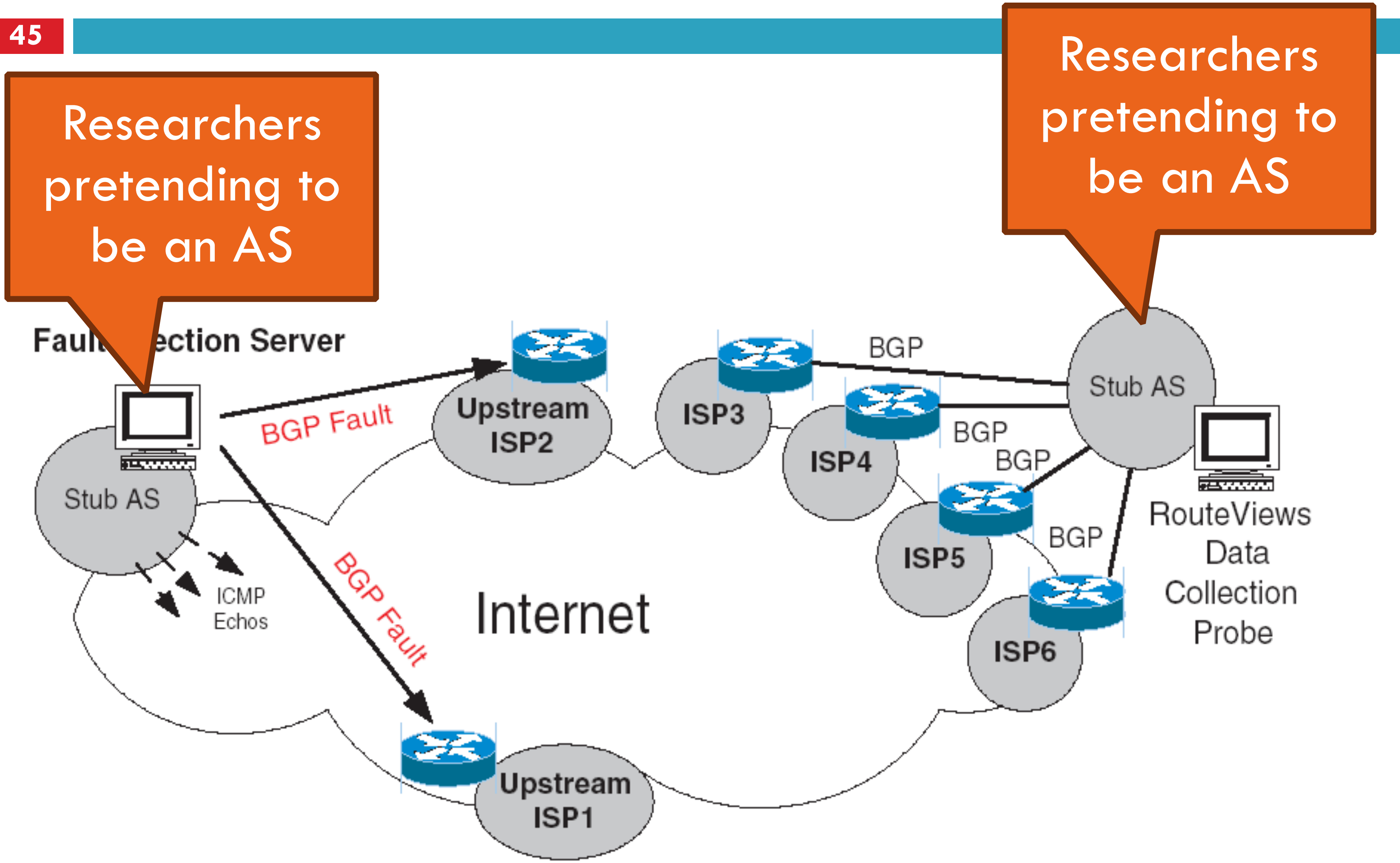
16 Month Study of Convergence

- Instrument the Internet
  - Inject BGP faults (announcements/withdrawals) of varied prefix and AS path length into topologically and geographically diverse ISP peering sessions
  - Monitor impact faults through
    - Recording BGP peering sessions with 20 tier1/tier2 ISPs
    - Active ICMP measurements (512 byte/second to 100 random web sites)
  - Wait two years (and 250,000 faults)

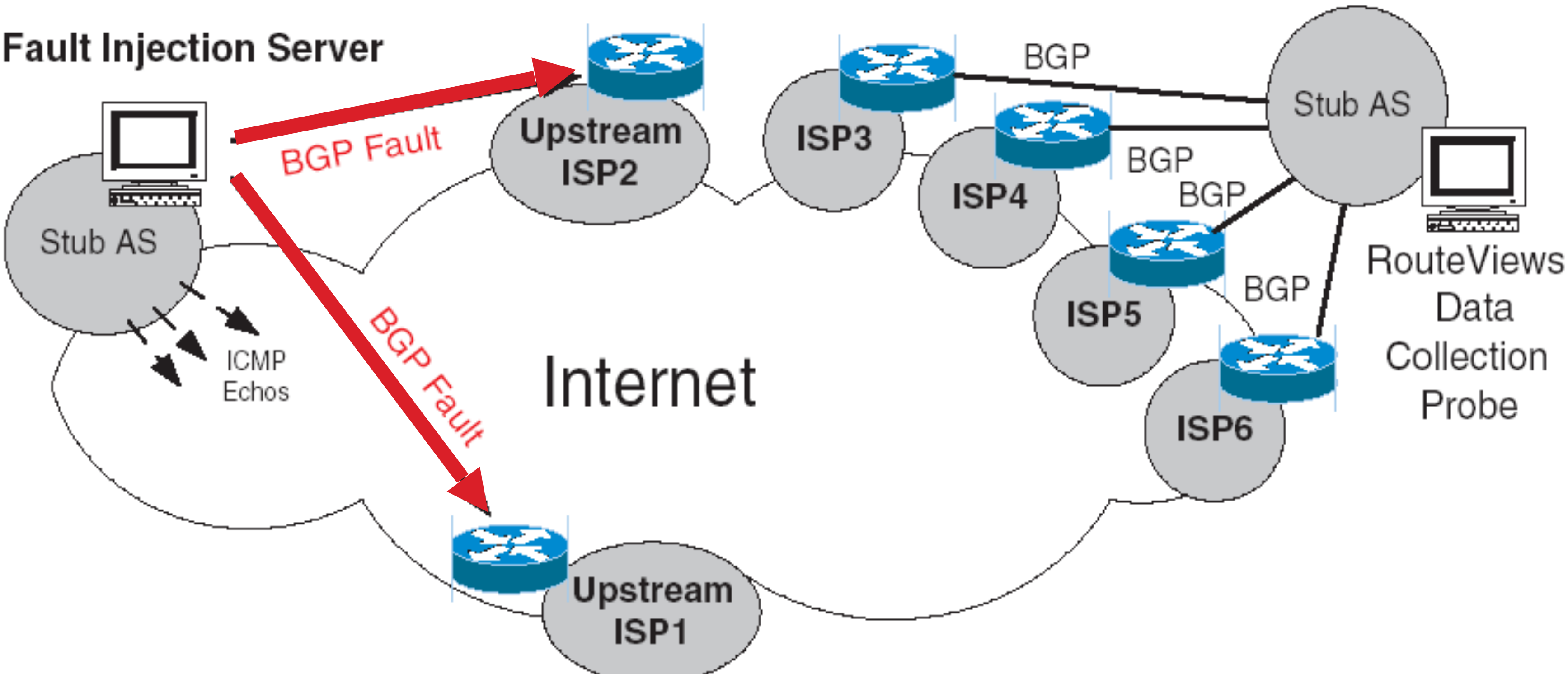
Measurement Architecture



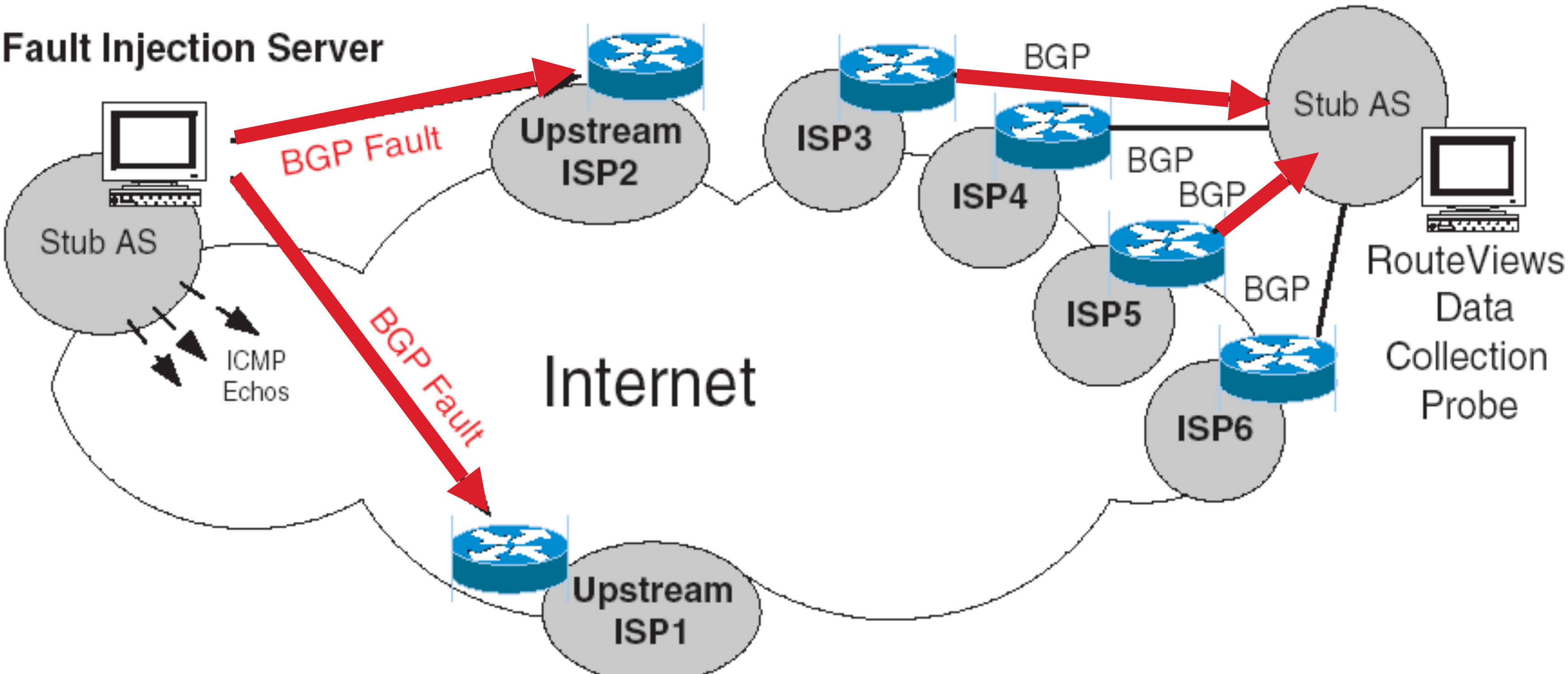
Measurement Architecture



Measurement Architecture



Measurement Architecture



Announcement Scenarios

- Tup – a new route is advertised
- Tdown – A route is withdrawn
  - i.e. single-homed failure
- Tshort – Advertise a shorter/better AS path
  - i.e. primary path repaired
- Tlong – Advertise a longer/worse AS path
  - i.e. primary path fails

Major Convergence Results

- Routing convergence requires an order of magnitude longer than expected
  - 10s of minutes
- Routes converge more quickly following Tup/Repair than Tdown/Failure events
  - Bad news travels more slowly
- Withdrawals (Tdown) generate several more announcements than new routes (Tup)

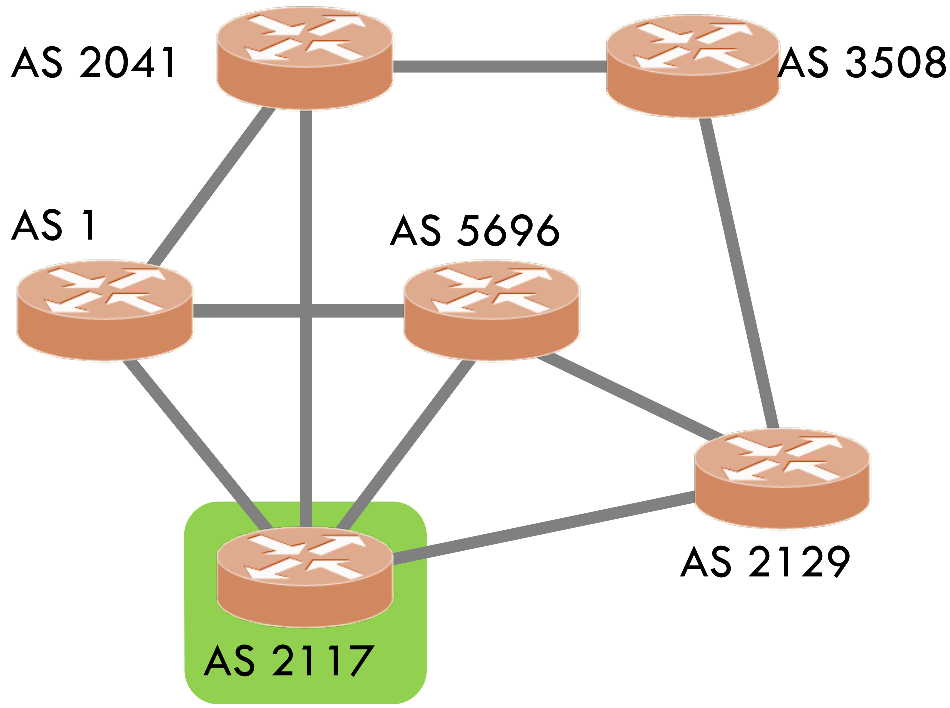
Example

TIME	BGP Message/Event
10:40:30	Route Fails/Withdrawn by AS2129
10:41:08	2117 announce 5696 2129
10:41:32	2117 announce 1 5696 2129
10:41:50	2117 announce 2041 3508 3508 4540 7037 1239 5696 2129
10:42:17	2117 announce 1 2041 3508 3508 4540 7037 1239 5696 2129
10:43:05	2117announce 2041 3508 3508 4540 7037 1239 6113 5696 2129
10:43:35	2117 announce 1 2041 3508 3508 4540 7037 1239 6113 5696 2129
10:43:59	2117 sends withdraw

- BGP log of updates from AS2117 for route via AS2129
- One withdrawal triggers 6 announcements and one withdrawal from 2117
- Increasing AS path length until final withdrawal

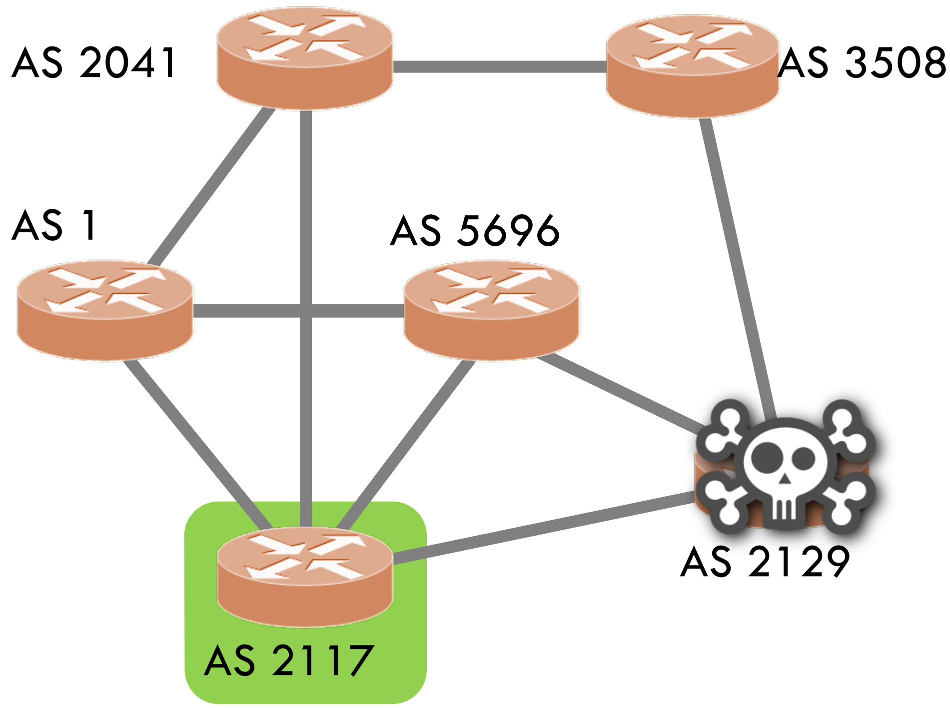
Why So Many Announcements?

Events from AS 2177



Why So Many Announcements?

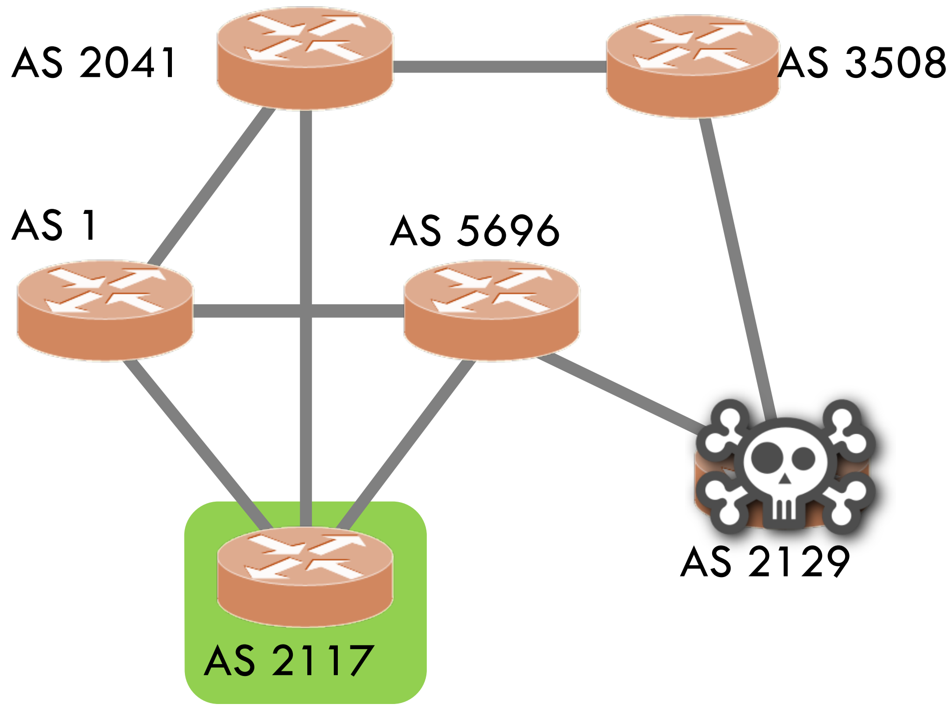
Events from AS 2177



Why So Many Announcements?

Events from AS 2177

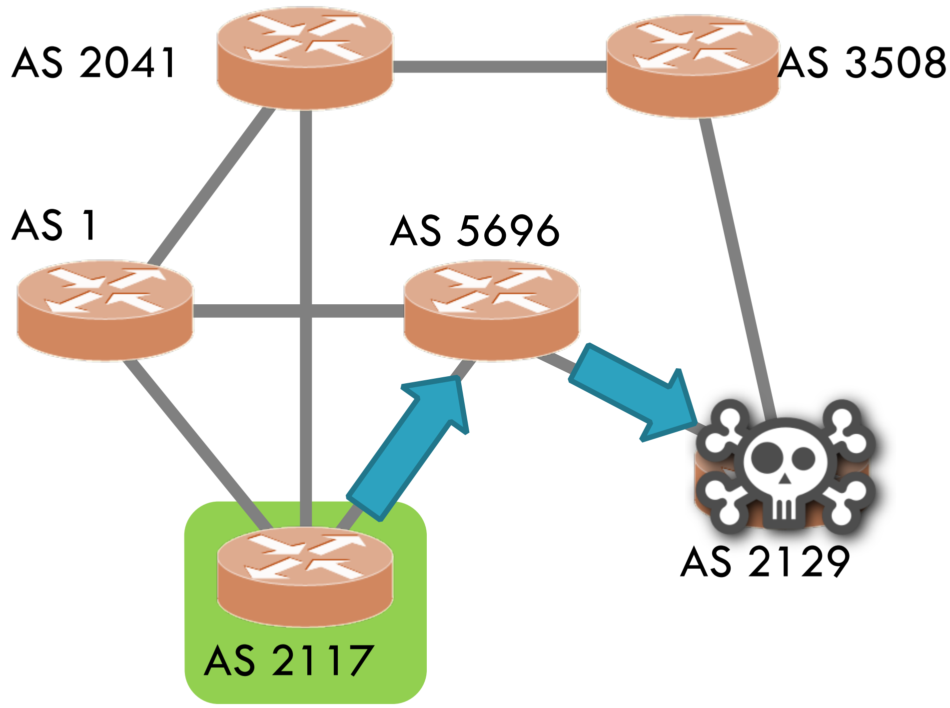
1. Route Fails: AS 2129



Why So Many Announcements?

Events from AS 2177

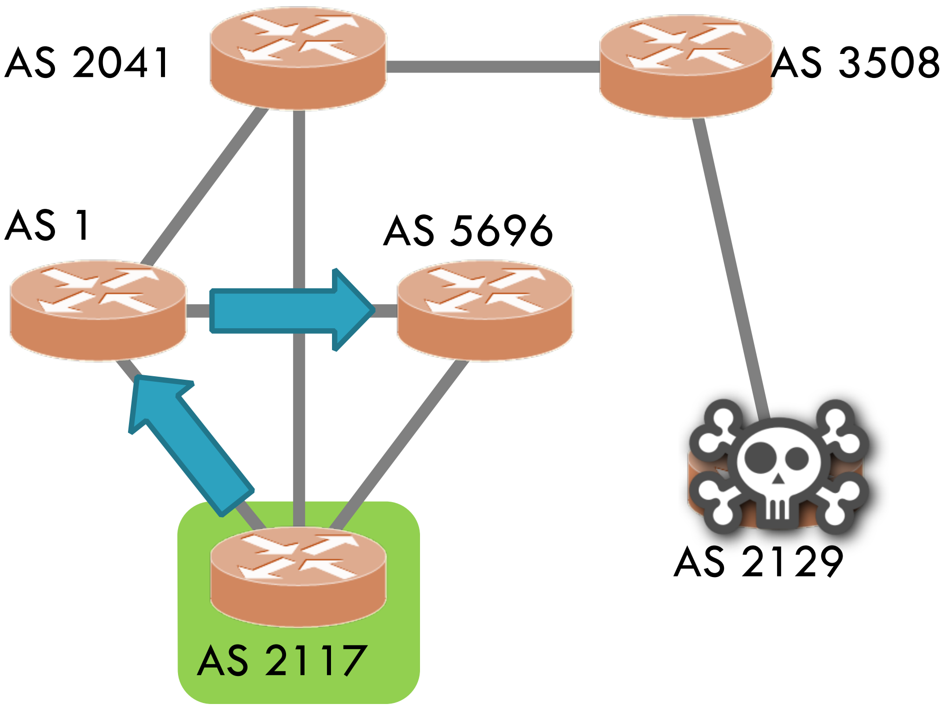
1. Route Fails: AS 2129
2. Announce: 5696 2129



Why So Many Announcements?

Events from AS 2177

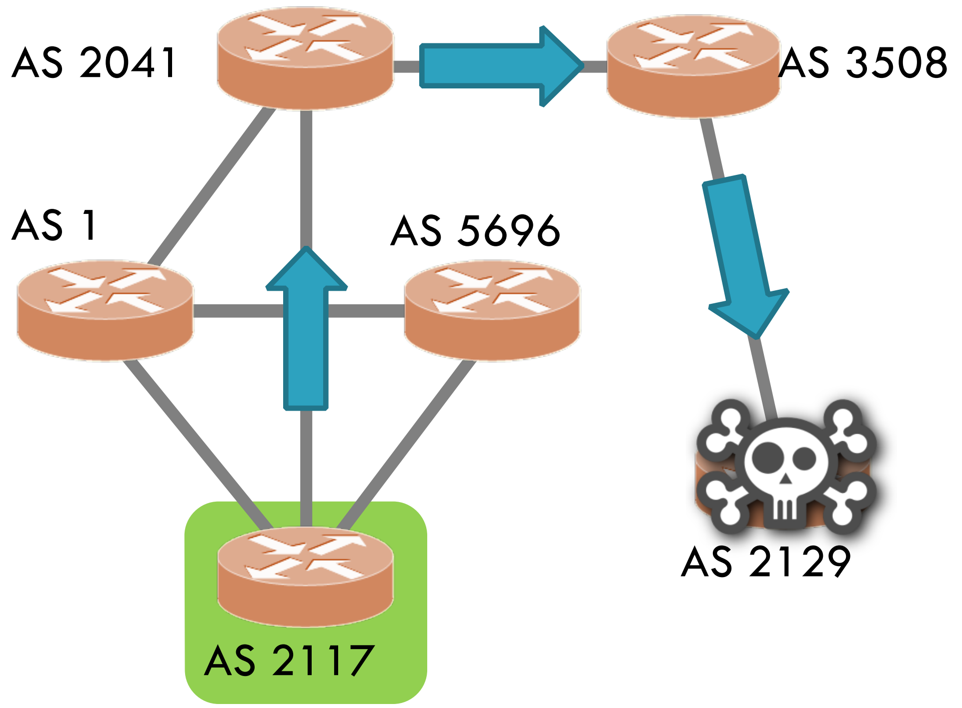
1. Route Fails: AS 2129
2. Announce: 5696 2129
3. Announce: 1 5696 2129



Why So Many Announcements?

Events from AS 2177

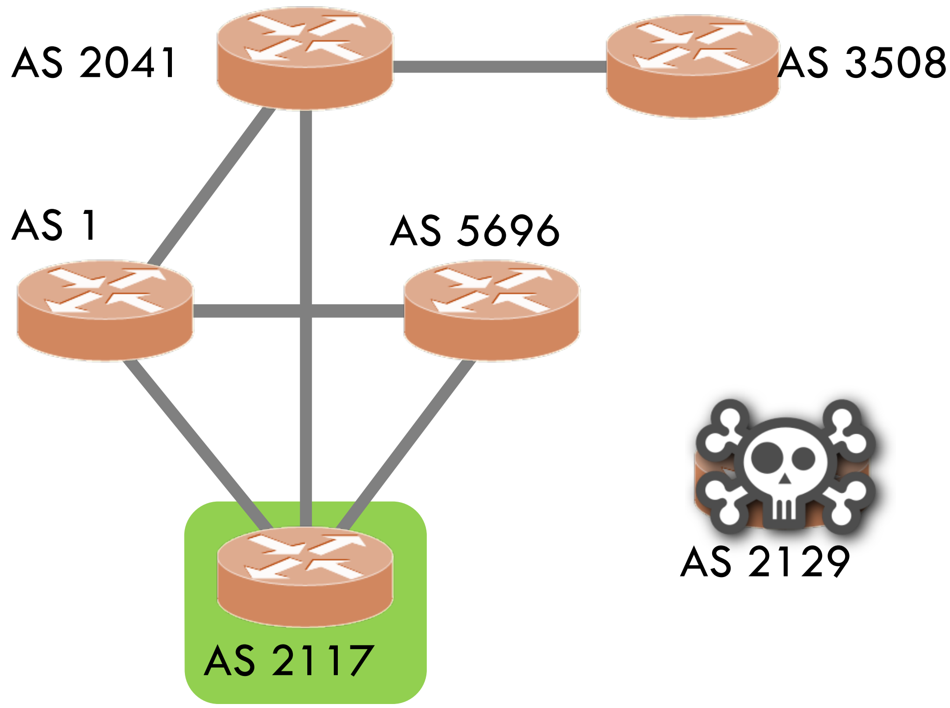
1. Route Fails: AS 2129
2. Announce: 5696 2129
3. Announce: 1 5696 2129
4. Announce: 2041 3508 2129



Why So Many Announcements?

Events from AS 2177

1. Route Fails: AS 2129
2. Announce: 5696 2129
3. Announce: 1 5696 2129
4. Announce: 2041 3508 2129
5. Announce: 1 2041 3508 2129
6. Route Withdrawn: 2129



How Many Announcements Does it Take For an AS to Withdraw a Route?

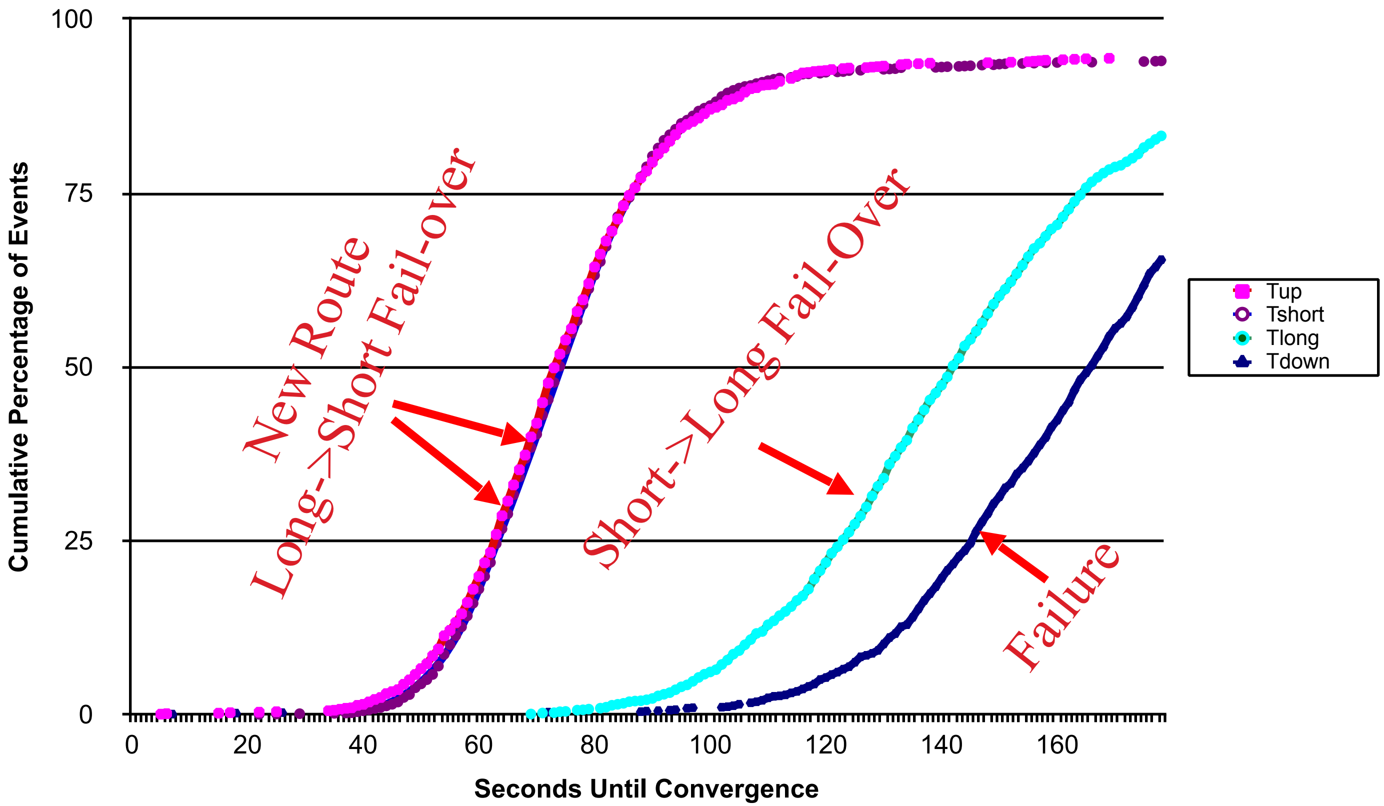
7/5 19:33:25	Route **R** is withdrawn	
7/5 19:34:15	AS6543 announce **R**	6543 66665 8918 1 5696 999
7/5 19:35:00	AS6543 announce **R**	6543 66665 8918 67455 6461 5696 999
7/5 19:35:37	AS6543 announce **R**	6543 66665 4332 6461 5696 999
7/5 19:35:39	AS6543 announce **R**	6543 66665 5378 6660 67455 6461 5696 999
7/5 19:35:39	AS6543 announce **R**	6543 66665 65 6461 5696 999
7/5 19:35:52	AS6543 announce **R**	6543 66665 6461 5696 999
7/5 19:36:00	AS6543 announce **R**	6543 66665 5378 6765 6660 67455 6461 5696 999
		…
7/5 19:38:22	AS6543 withdraw **R**	

How Many Announcements Does it Take For an AS to Withdraw a Route?

7/5 19:33:25	Route **R** is **withdrawn**	
7/5 19:34:15	AS6543 **announce** **R**	6543 66665 8918 1 5696 999
7/5 19:35:00	AS6543 **announce** **R**	6543 66665 8918 67455 6461 5696 999
7/5 19:35:37	AS6543 **announce** **R**	6543 66665 4332 6461 5696 999
7/5 19:35:39	AS6543 **announce** **R**	6543 66665 5378 6660 67455 6461 5696 999
7/5 19:35:39	AS6543 **announce** **R**	6543 66665 65 6461 5696 999
7/5 19:35:52	AS6543 **announce** **R**	6543 66665 6461 5696 999
7/5 19:36:00	AS6543 **announce** **R**	6543 66665 5378 6765 6660 67455 6461 5696 999
		...
7/5 19:38:22	AS6543 **withdraw** **R**	

Answer: ***up to 19***

BGP Routing Table Convergence Times



- Less than half of Tdown events converge within two minutes
- Tup/Tshort and Tdown/Tlong form equivalence classes
- Long tailed distribution (up to 15 minutes)

Failures, Fail-overs and Repairs

- Bad news does not travel fast…
- Repairs (Tup) exhibit similar convergence as long-short AS path fail-over
- Failures (Tdown) and short-long fail-overs (e.g. primary to secondary path) also similar
  - Slower than Tup (e.g. a repair)
  - 80% take longer than two minutes
  - Fail-over times degrade the greater the degree of multi-homing

Intuition for Delayed Convergence

- There exists possible ordering of messages such that BGP will explore ALL possible AS paths of ALL possible lengths
- BGP is O(N!), where N number of default-free BGP routers in a complete graph with default policy

Impact of Delayed Convergence

- Why do we care about routing table convergence?
  - It impacts end-to-end connectivity for Internet paths
- ICMP experiment results
  - Loss of connectivity, packet loss, latency, and packet re-ordering for an average of 3-5 minutes after a fault
- Why?
  - Routers drop packets when next hop is unknown
  - Path switching spikes latency/delay
  - Multi-pathing causes reordering

In real life …

- Discussed worst case BGP behavior
- In practice, BGP policy prevents worst case from happening
- BGP timers also provide synchronization and limits possible orderings of messages

Inter-Domain Routing Summary

- BGP4 is the only inter-domain routing protocol currently in use world-wide
- Issues?
  - Lack of security
  - Ease of misconfiguration
  - Poorly understood interaction between local policies
  - Poor convergence
  - Lack of appropriate information hiding
  - Non-determinism
  - Poor overload behavior

Lots of research into how to fix this

- Security
  - BGPSEC, RPKI
- Misconfigurations, inflexible policy
  - SDN
- Policy Interactions
  - PoiRoot (root cause analysis)
- Convergence
  - Consensus Routing
- Inconsistent behavior
  - LIFEGUARD, among others

Why are these still issues?

- Backward compatibility
- Buy-in / incentives for operators
- Stubbornness

Why are these still issues?

- Backward compatibility
- Buy-in / incentives for operators
- Stubbornness

Very similar issues to IPv6 deployment