TDTS21 Advanced Networking

BGP and Inter-domain Routing

(It's all about the Money)

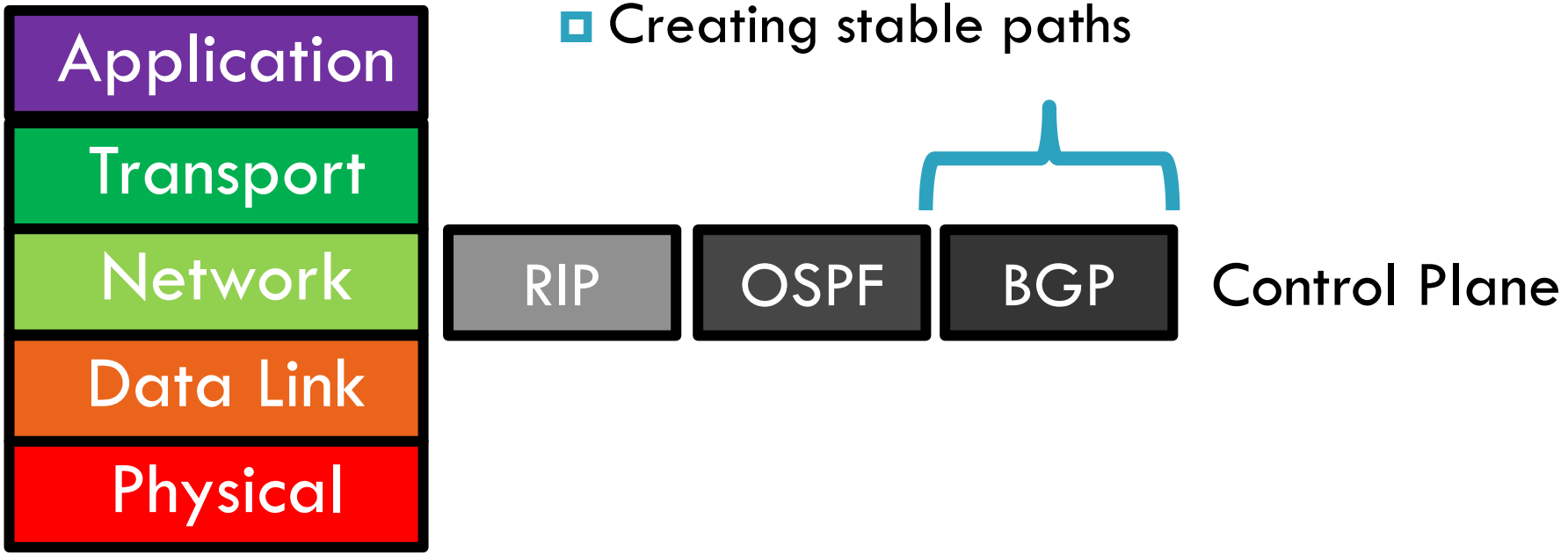
Based on slides from P. Gill, D. Choffnes, J. Rexford, and A. Feldman
Revised 2015, 2019, 2021 by N. Carlsson

Control plane vs. Data Plane

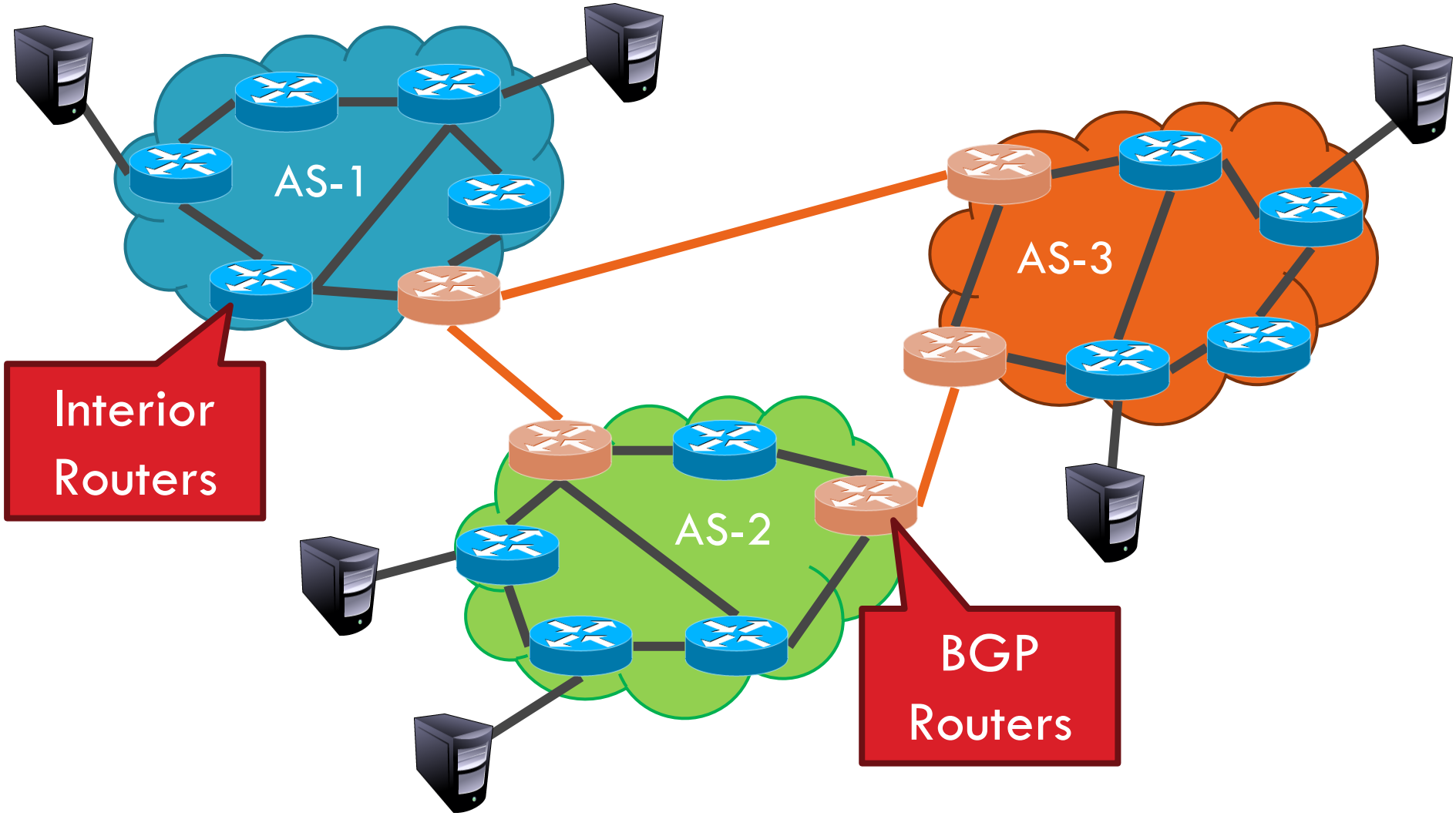
- Control:
  - Make sure that if there's a path available, data is forwarded over it
  - BGP sets up such paths at the AS-level
- Data:
  - For a destination, send packet to most-preferred next hop
  - Routers forward data along IP paths

Network Layer, Control Plane

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



ASs, Revisited



AS Numbers

- Each AS identified by an ASN number
  - 16-bit values (latest protocol supports 32-bit ones)
  - Some blocks (e.g., 64512 – 65535) are reserved
- Currently, there are ~ 100,000 ASNs
  - AT&T: 5074, 6341, 7018, …
  - Sprint: 1239, 1240, 6211, 6242, …
  - LIUNET: 2843 (prefix: 130.236.0.0/16)
  - Google 15169, 36561 (formerly YT), + others
  - Facebook 32934
  - North America ASs → ftp://ftp.arin.net/info/asn.txt

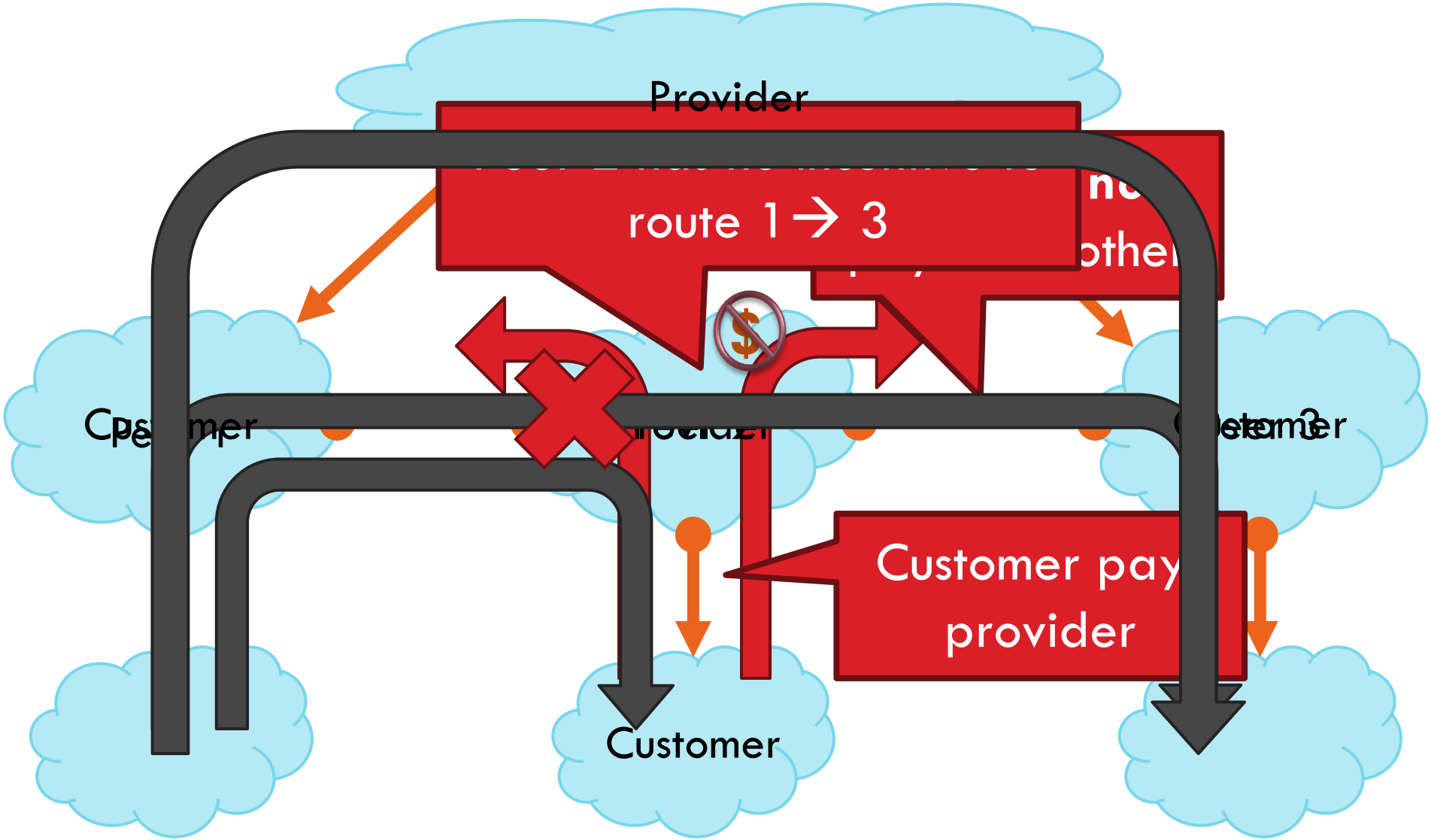
Inter-Domain Routing

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

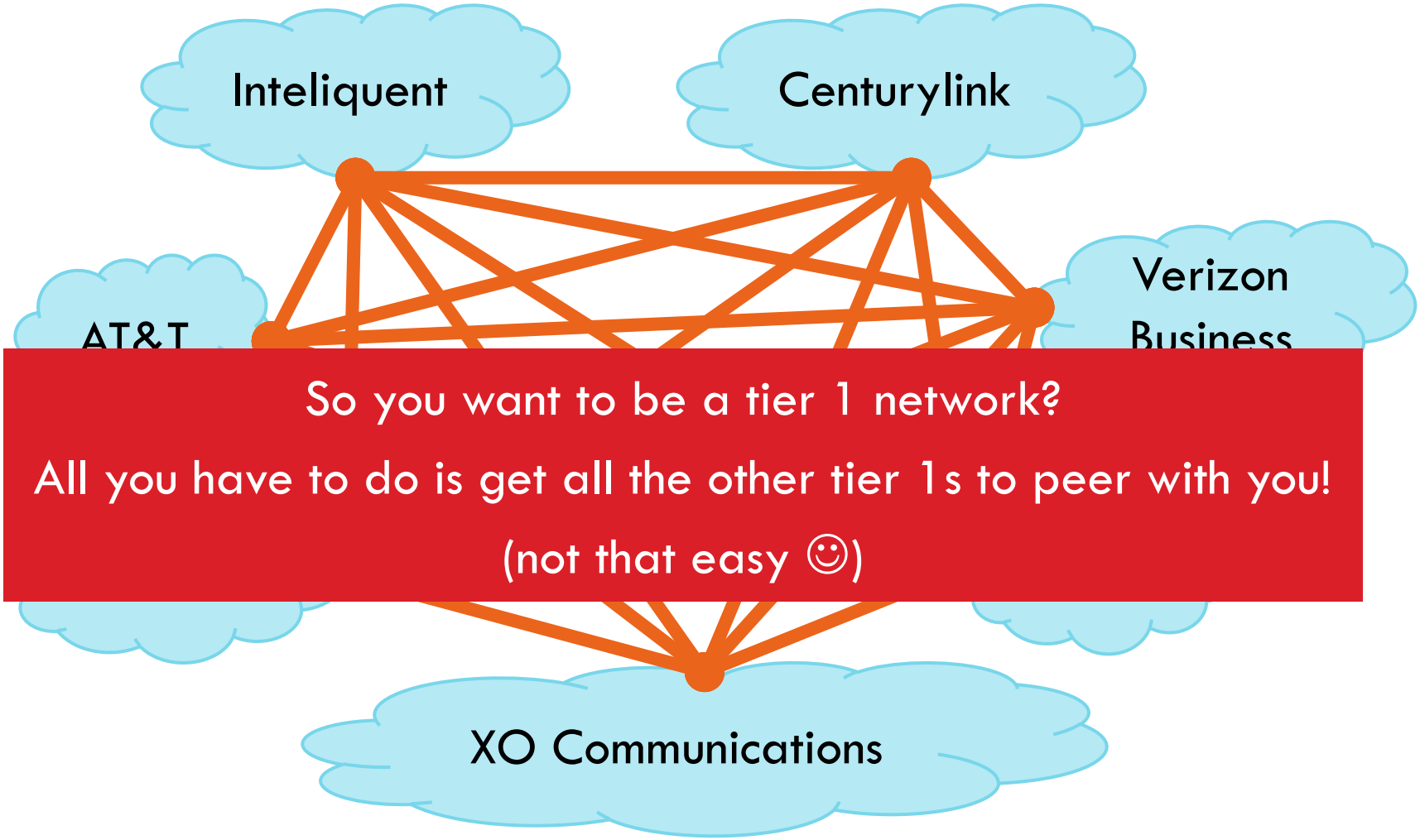
BGP

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

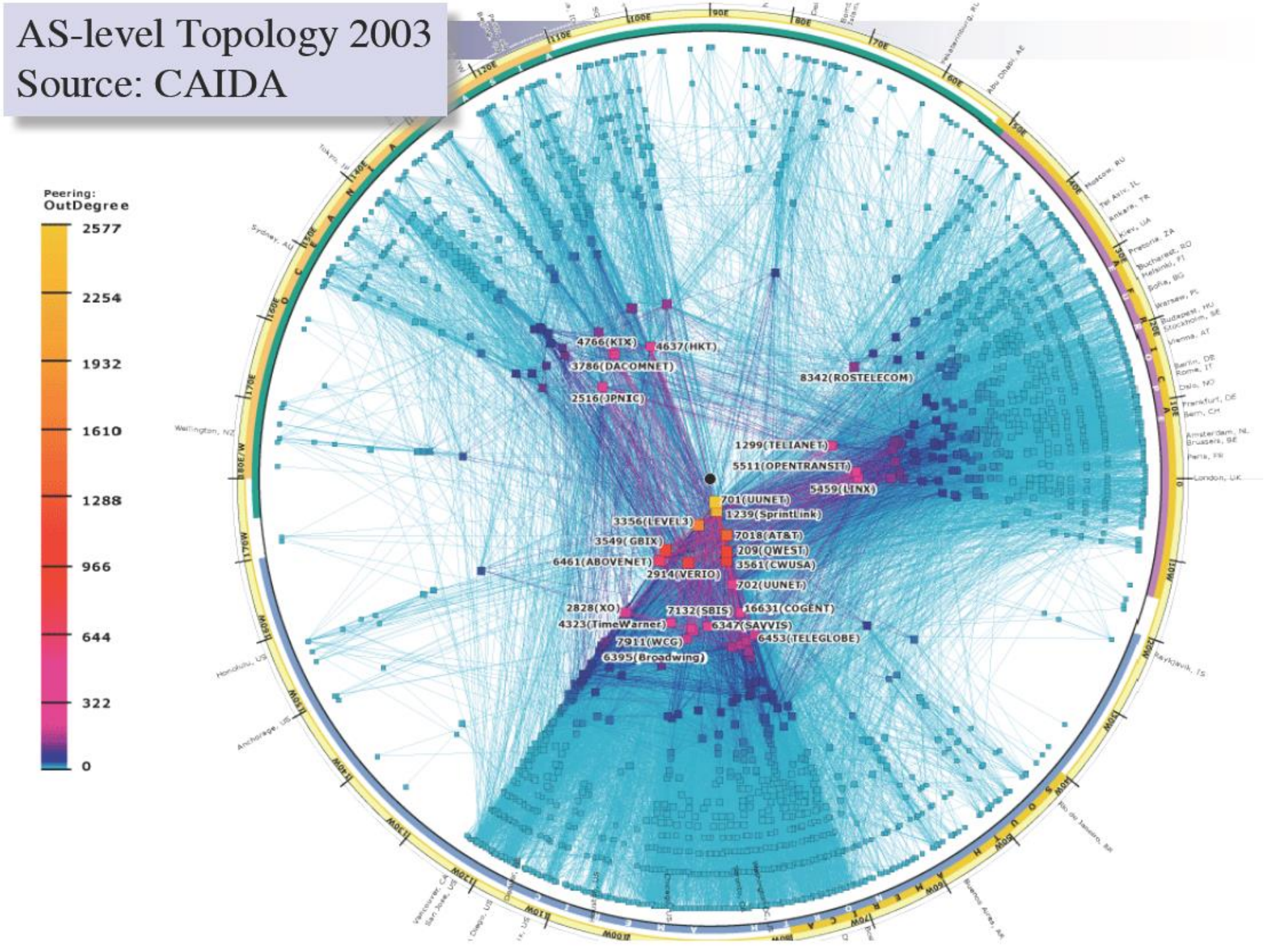
BGP Relationships



Tier-1 ISP Peering



AS-level Topology 2003
Source: CAIDA



Peering Wars

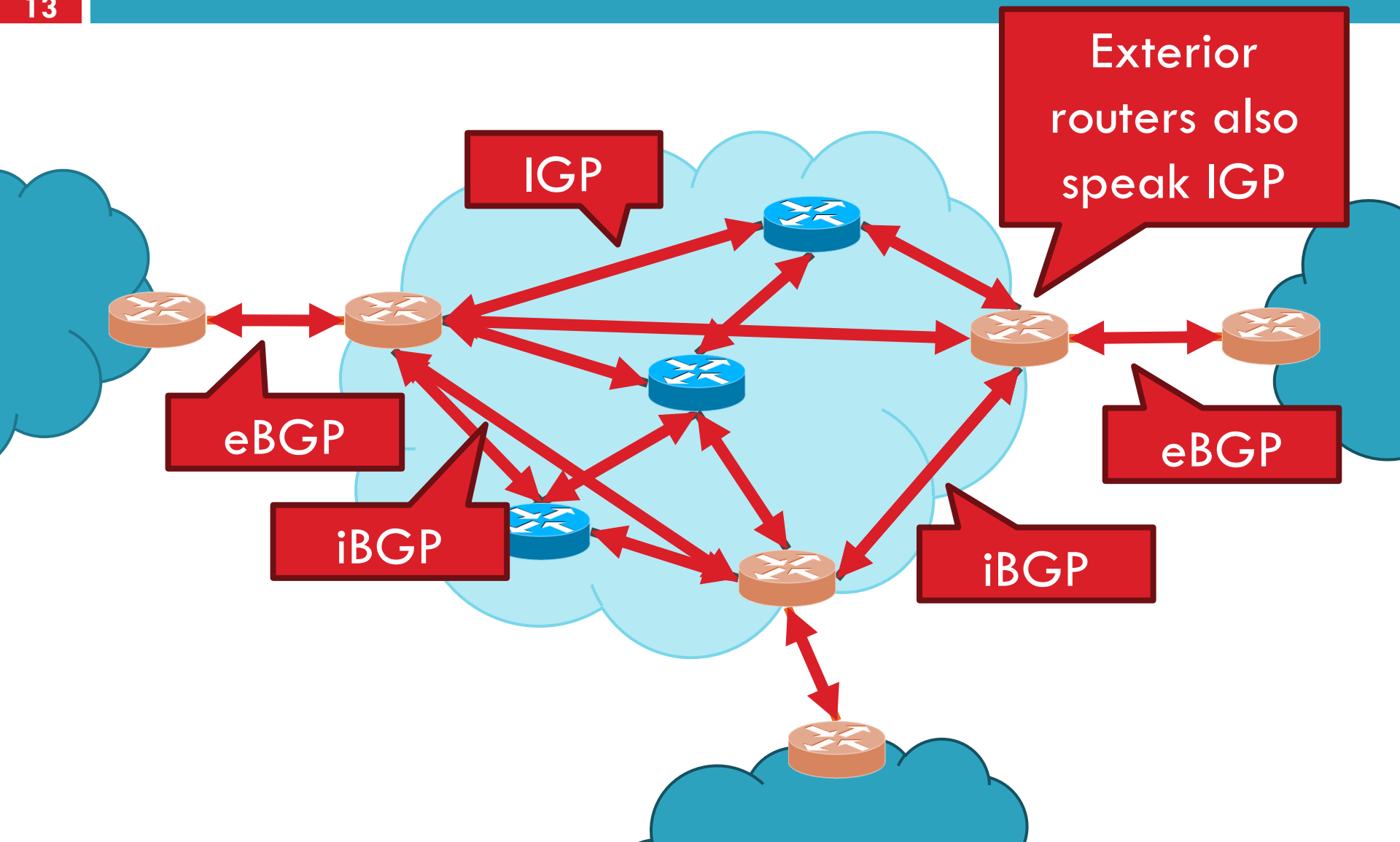
Peer	Don't Peer
☐ R	☐ You would rather have
☐ I	
☐ A	

Peering struggles in the ISP world are extremely contentious agreements are usually confidential

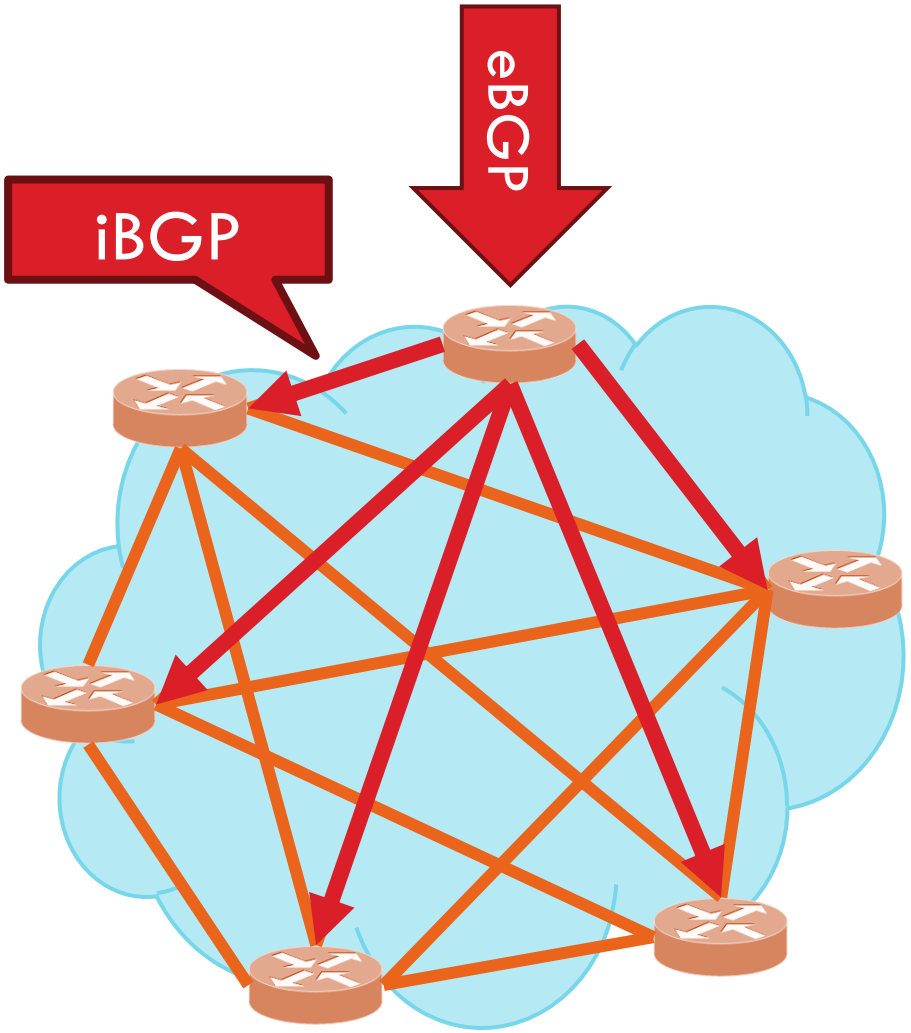
Example: If you are a customer of my peer why should I peer with you? You should pay me too!

Incentive to keep relationships private!

Two Types of BGP Neighbors

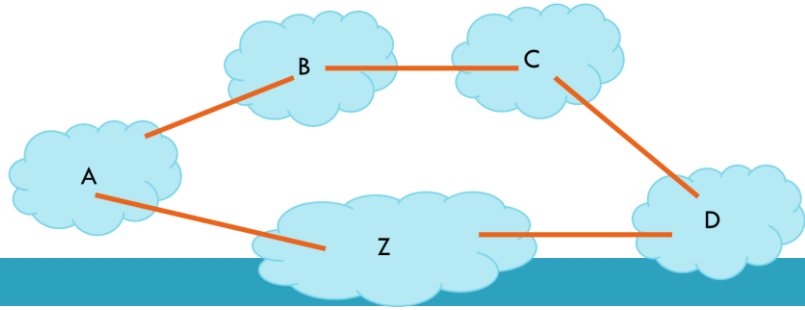


Full iBGP Meshes

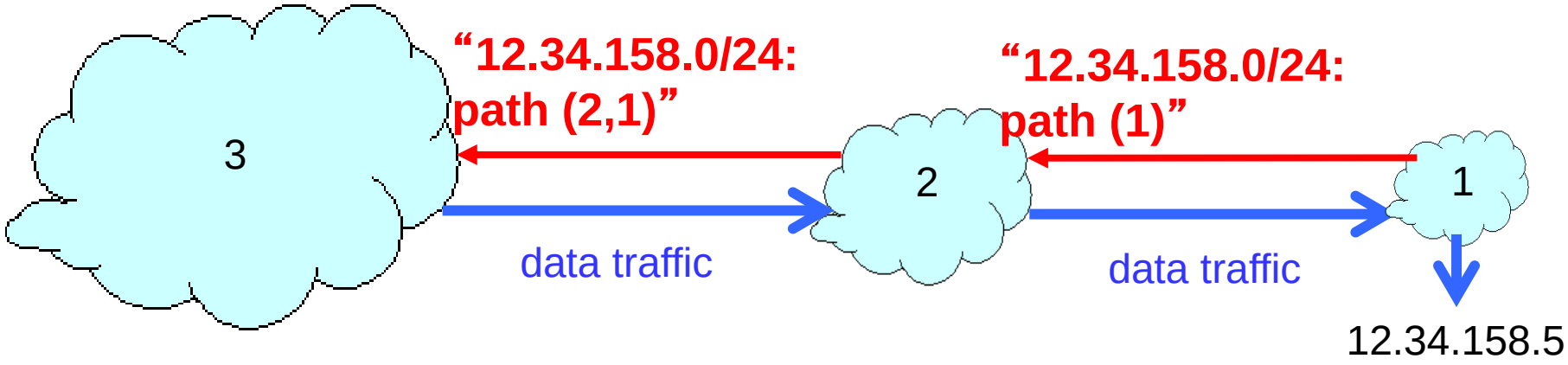


- Question: why do we need iBGP?
  - OSPF does not include BGP policy info
  - Prevents routing loops within the AS
- iBGP updates do not trigger announcements

Border Gateway Protocol

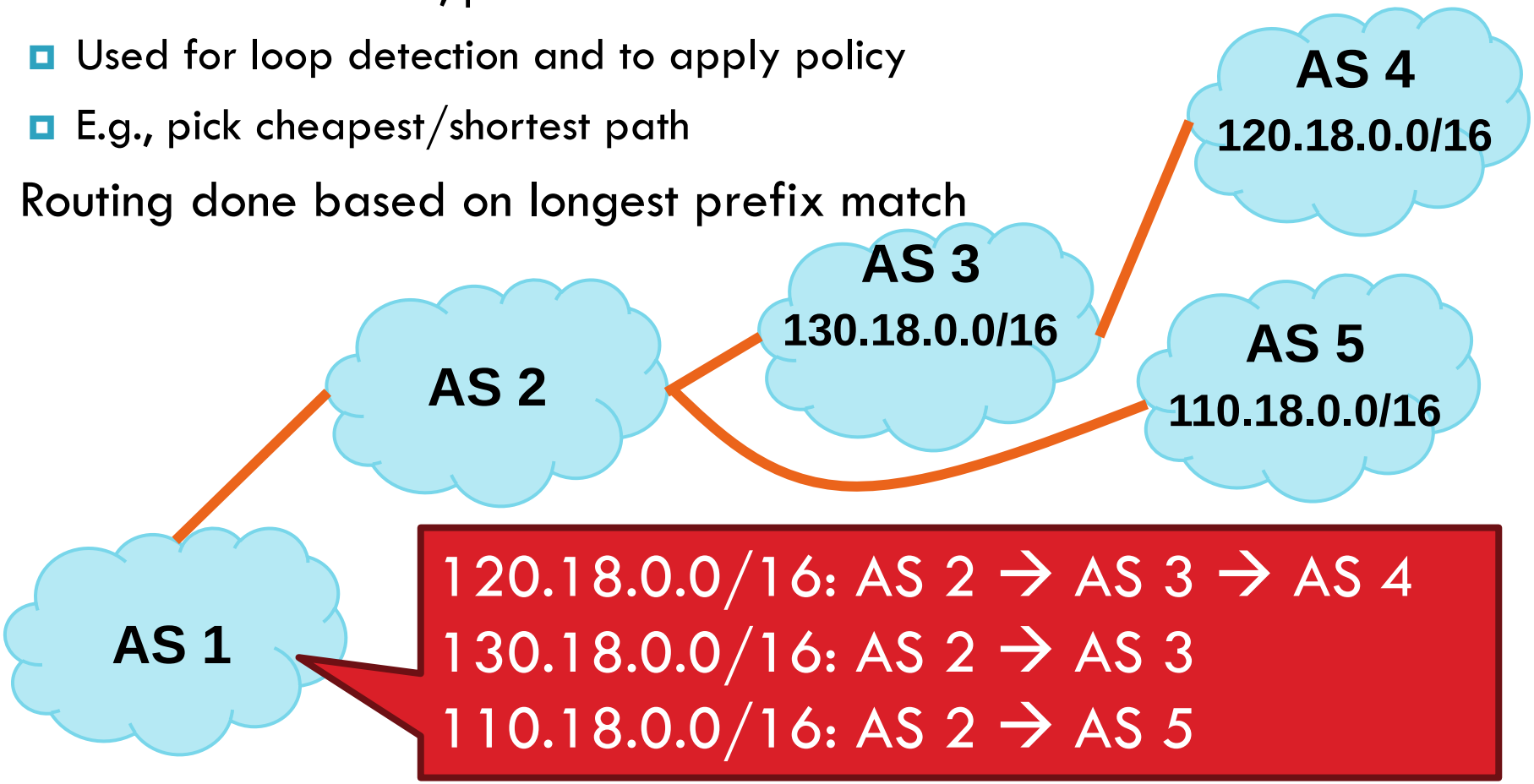


- ASes exchange info about who they can reach
  - IP prefix: block of destination IP addresses
  - AS path: sequence of ASes along the path
- Policies configured by the AS's operator
  - Path selection: which of the paths to use?
  - Path export: which neighbors to tell?

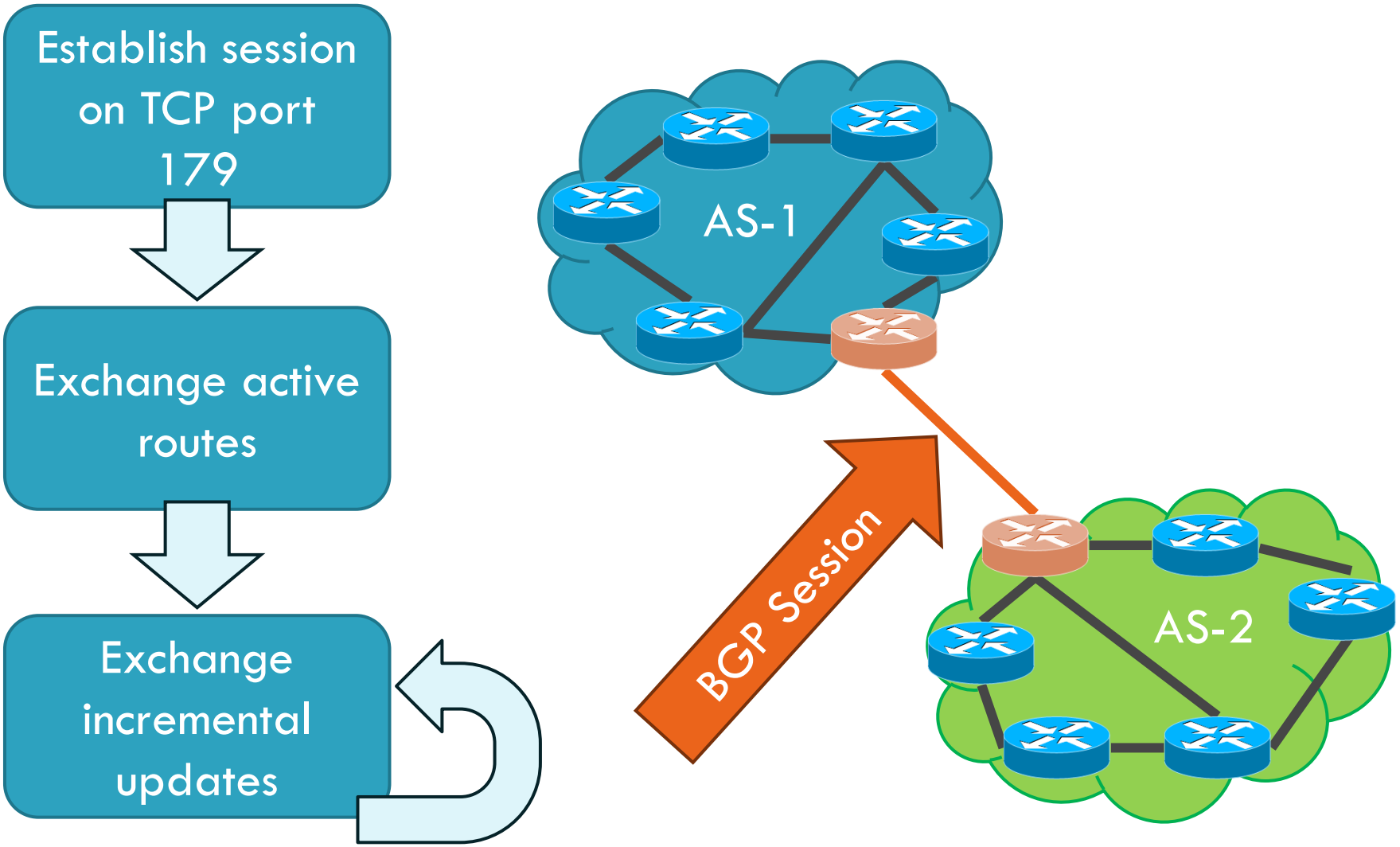


Path Vector Protocol

- AS-path: sequence of ASs a route traverses
  - Like distance vector, plus additional information
  - Used for loop detection and to apply policy
  - E.g., pick cheapest/shortest path
- Routing done based on longest prefix match



BGP Operations (Simplified)



Four Types of BGP Messages

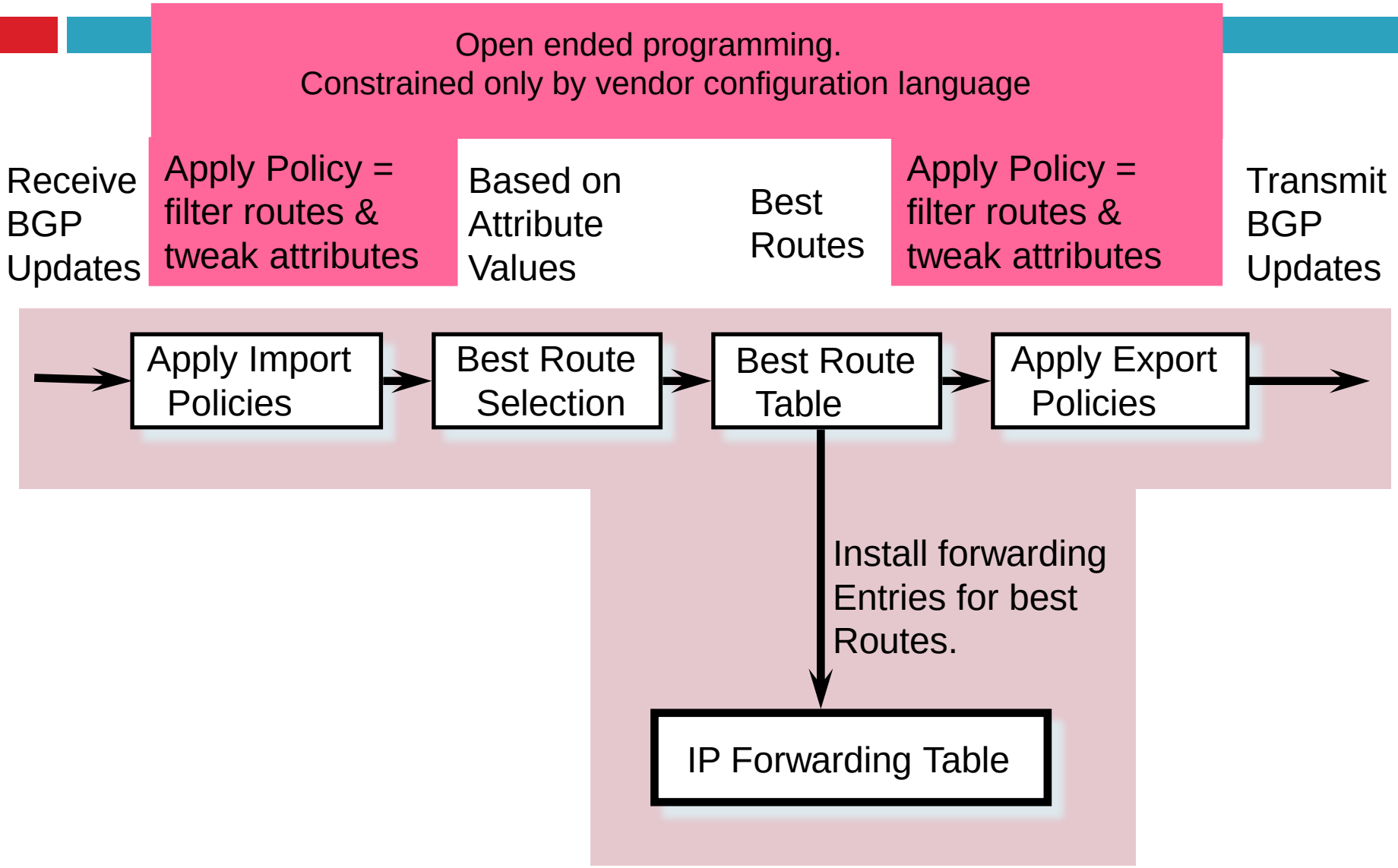
- Open: Establish a peering session.
- Keep Alive: Handshake at regular intervals.
- Notification: Shuts down a peering session.
- Update: Announce new routes or withdraw previously announced routes.

announcement = IP prefix + attributes values

Applying Policy to Routes

- Import policy
  - **Q: What route advertisements do I accept?**
  - Filter unwanted routes from neighbor
    - E.g. prefix that your customer doesn't own
  - Manipulate attributes to influence path selection
    - E.g., assign local preference to favored routes
- Export policy
  - **Q: Which routes do I forward to whom?**
  - Filter routes you don't want to tell your neighbor
    - E.g., don't tell a peer a route learned from other peer
  - Manipulate attributes to control what they see
    - E.g., make a path look artificially longer than it is

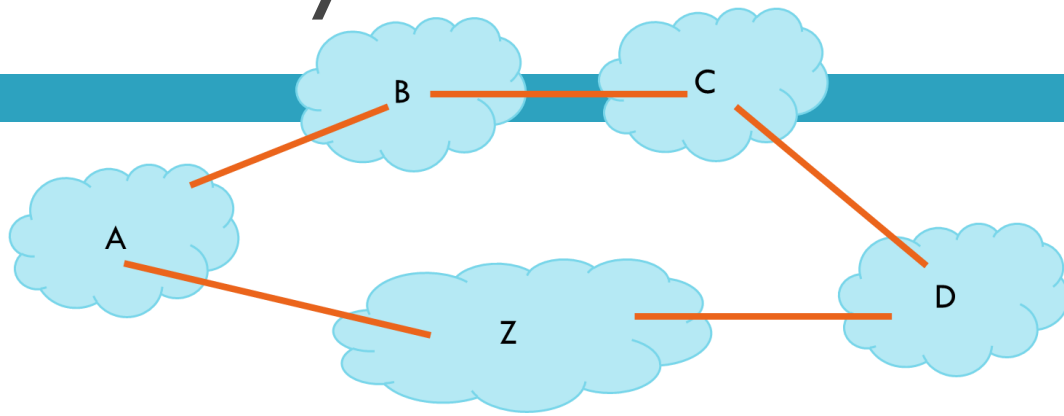
BGP Policy: Influencing Decisions



Routing Policies

- Economics
  - Enforce business relationships
  - Pick routes based on revenue and cost
  - Get traffic out of the network as early as possible
- Traffic engineering
  - Balance traffic over edge links
  - Select routes with good end-to-end performance
- Security and scalability
  - Filter routes that seem erroneous
  - Prevent the delivery of unwanted traffic
  - Limit the dissemination of small address blocks

Route Selection Summary



Highest Local Preference — Enforce relationships

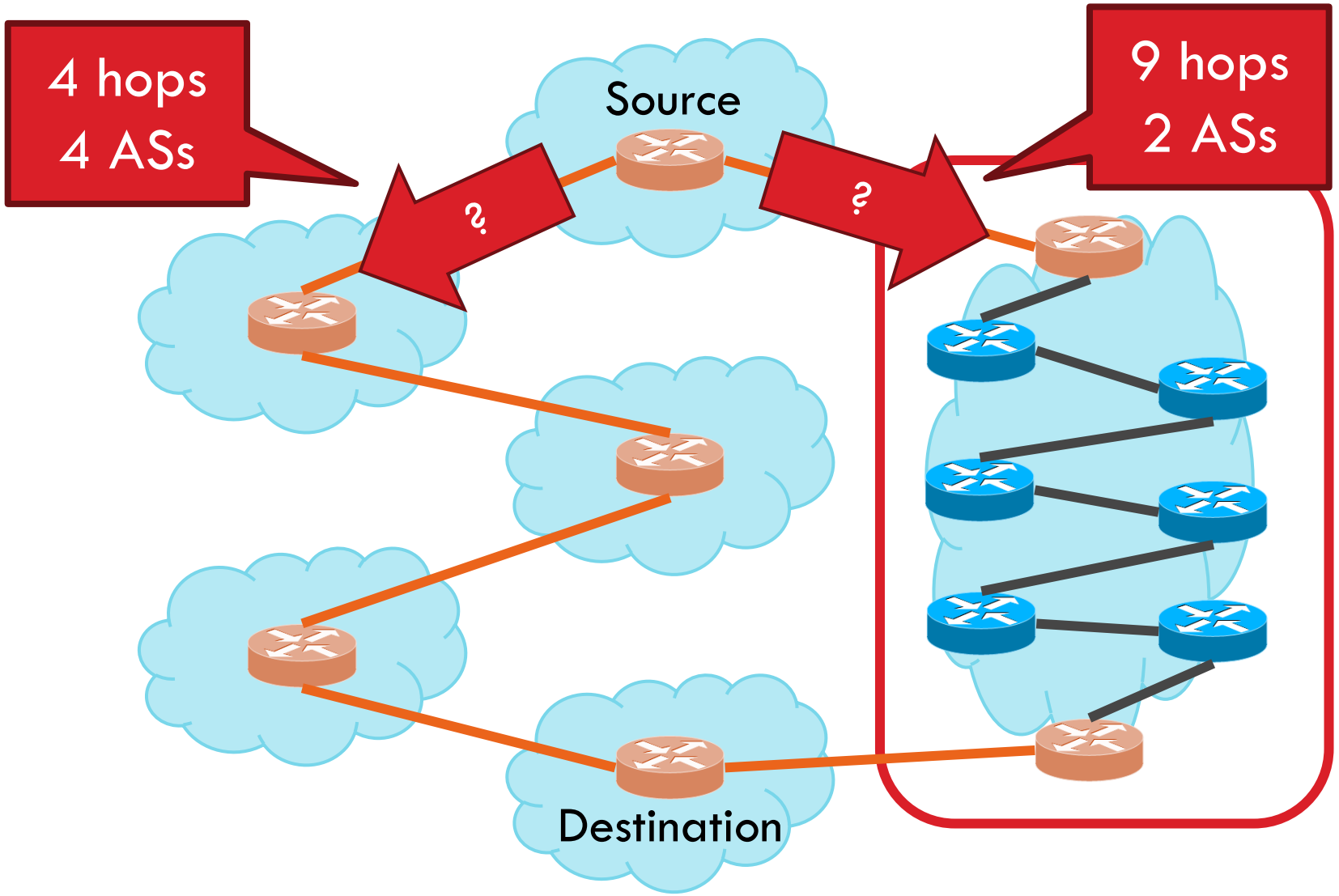
Shortest AS Path

Lowest MED

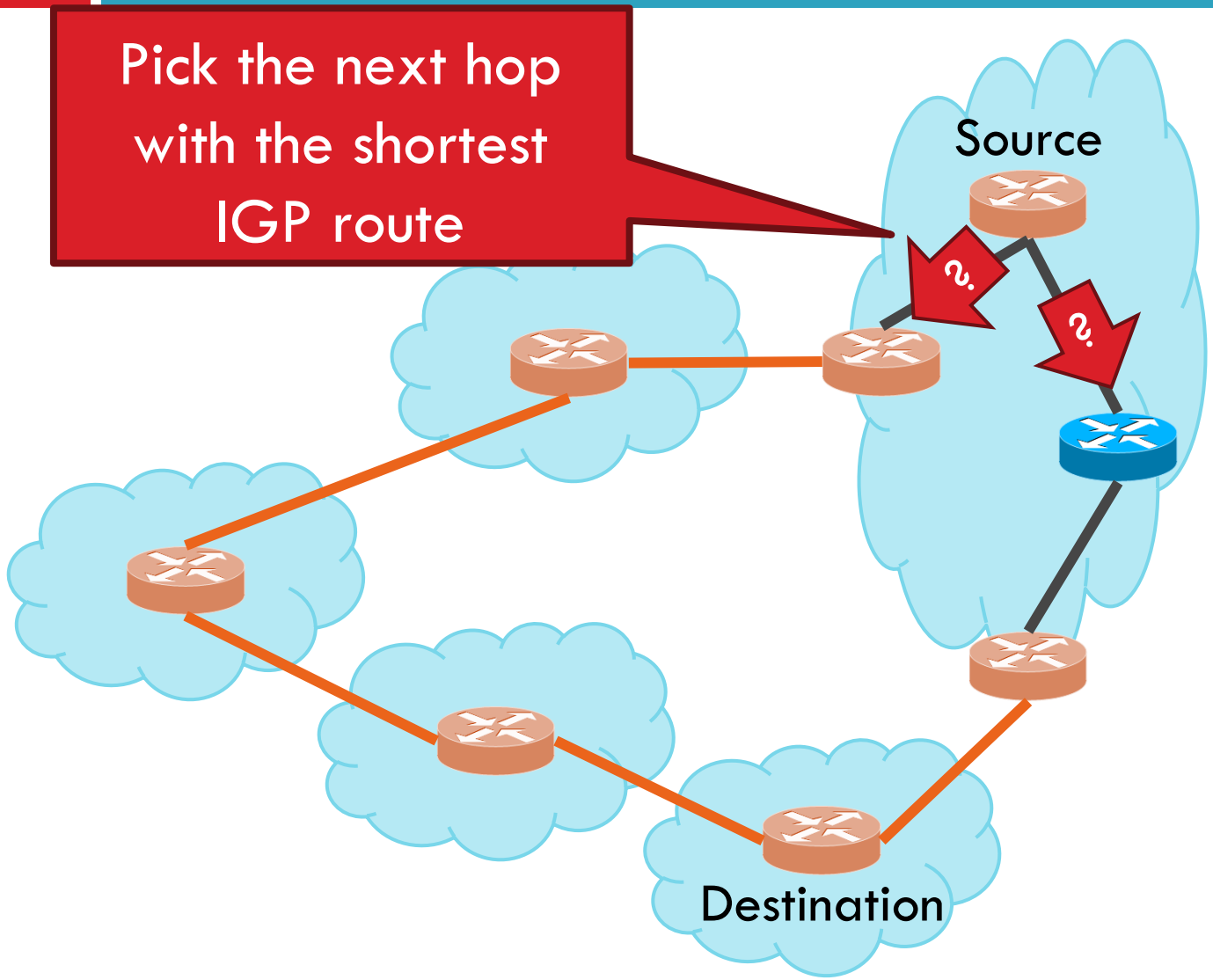
Lowest IGP Cost to BGP Egress — Traffic engineering

Lowest Router ID — When all else fails, break ties

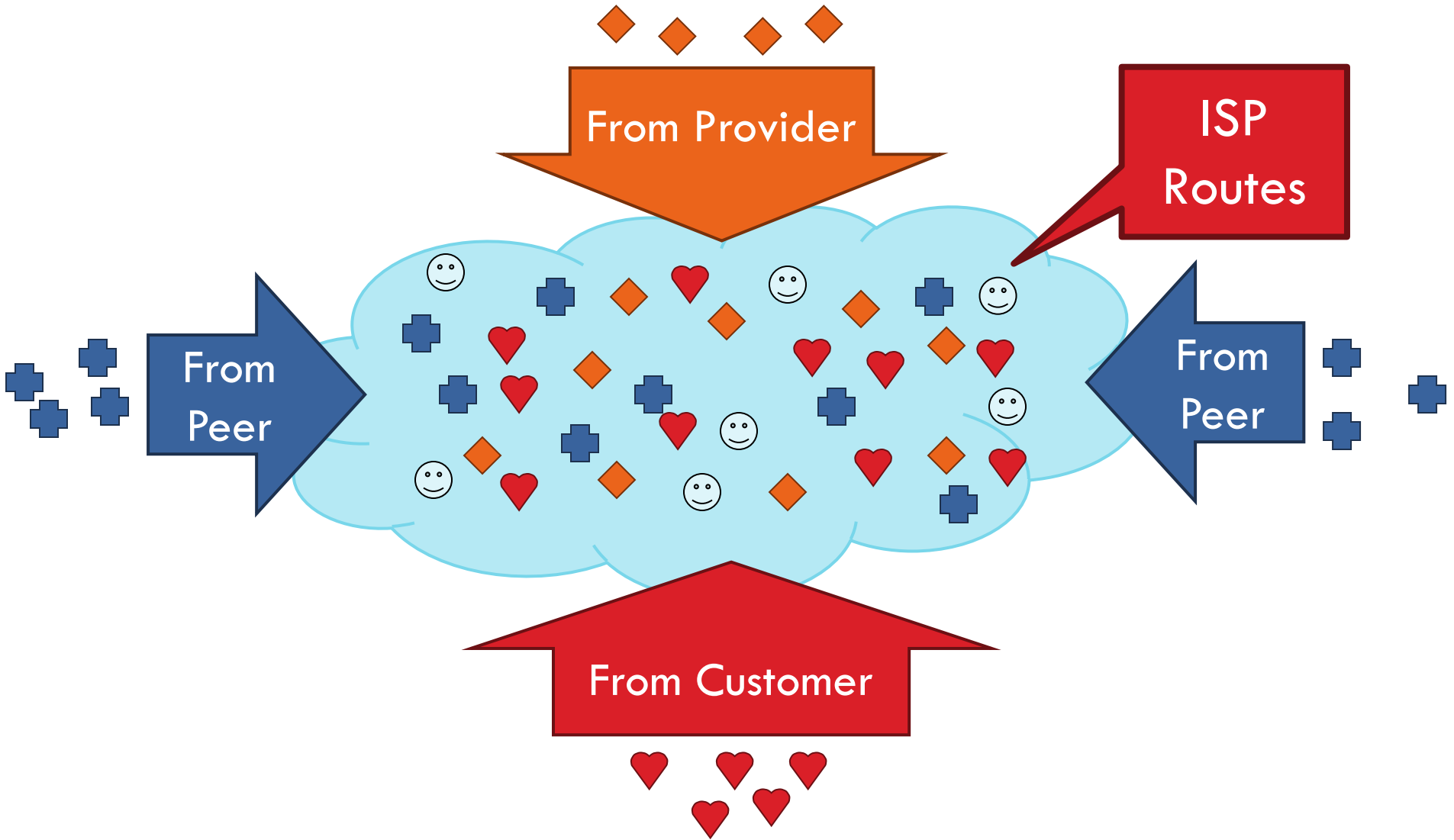
Shortest AS Path != Shortest Path



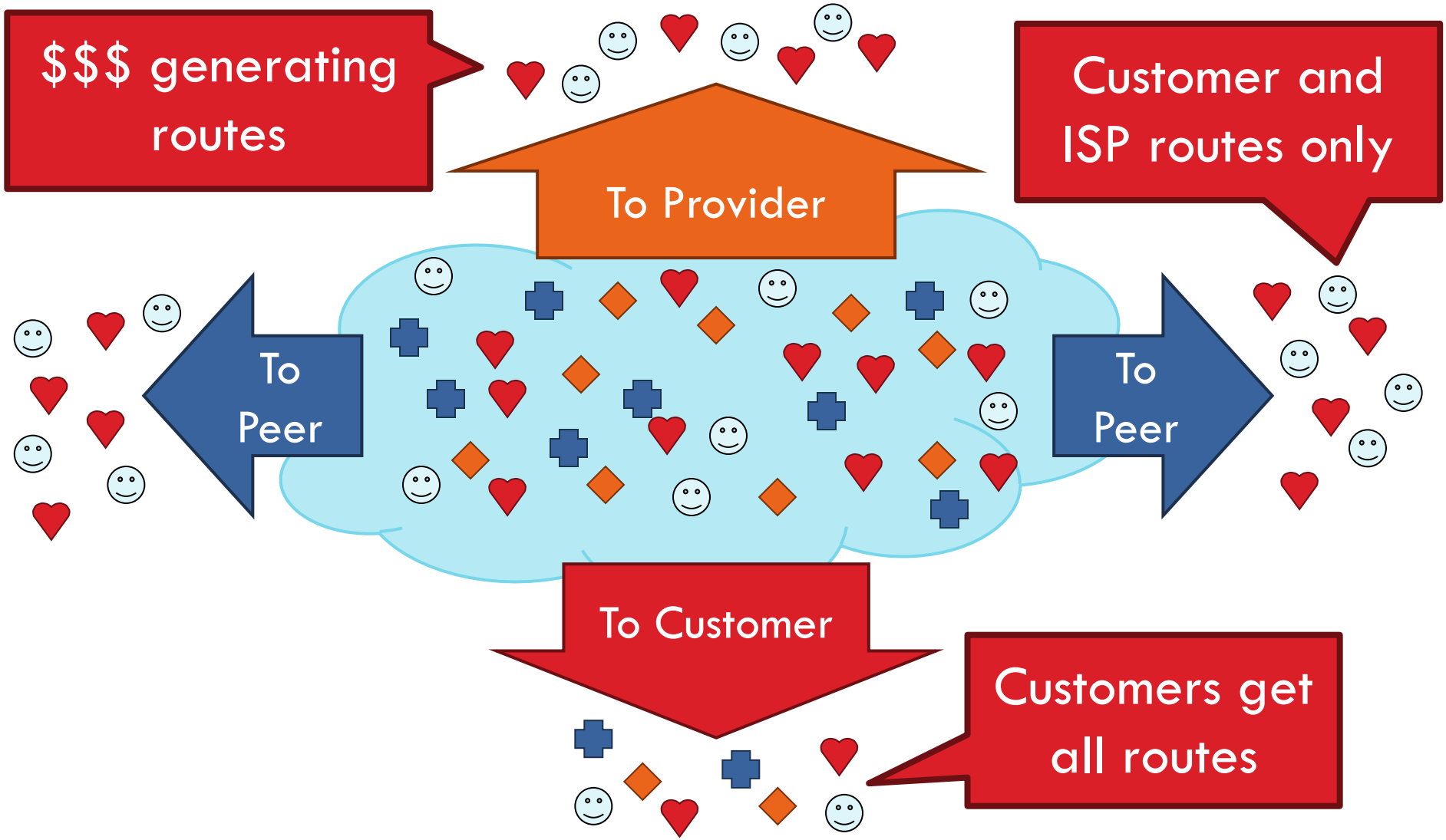
Hot Potato Routing



Importing Routes

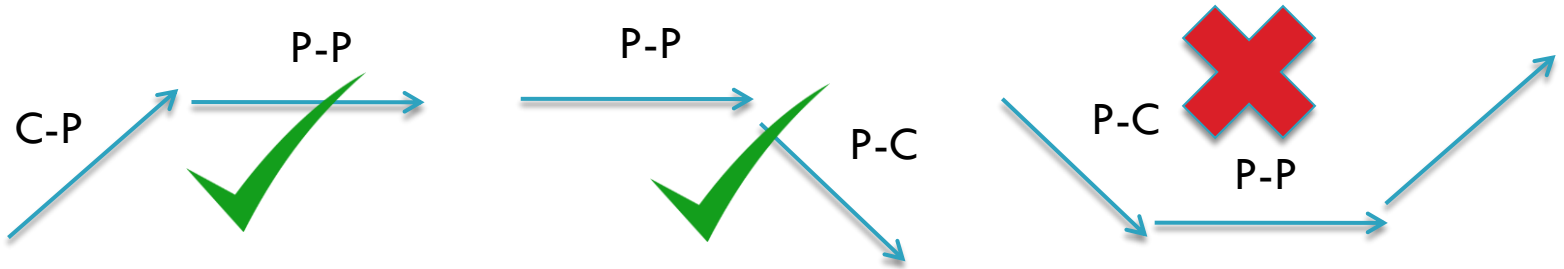


Exporting Routes



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)

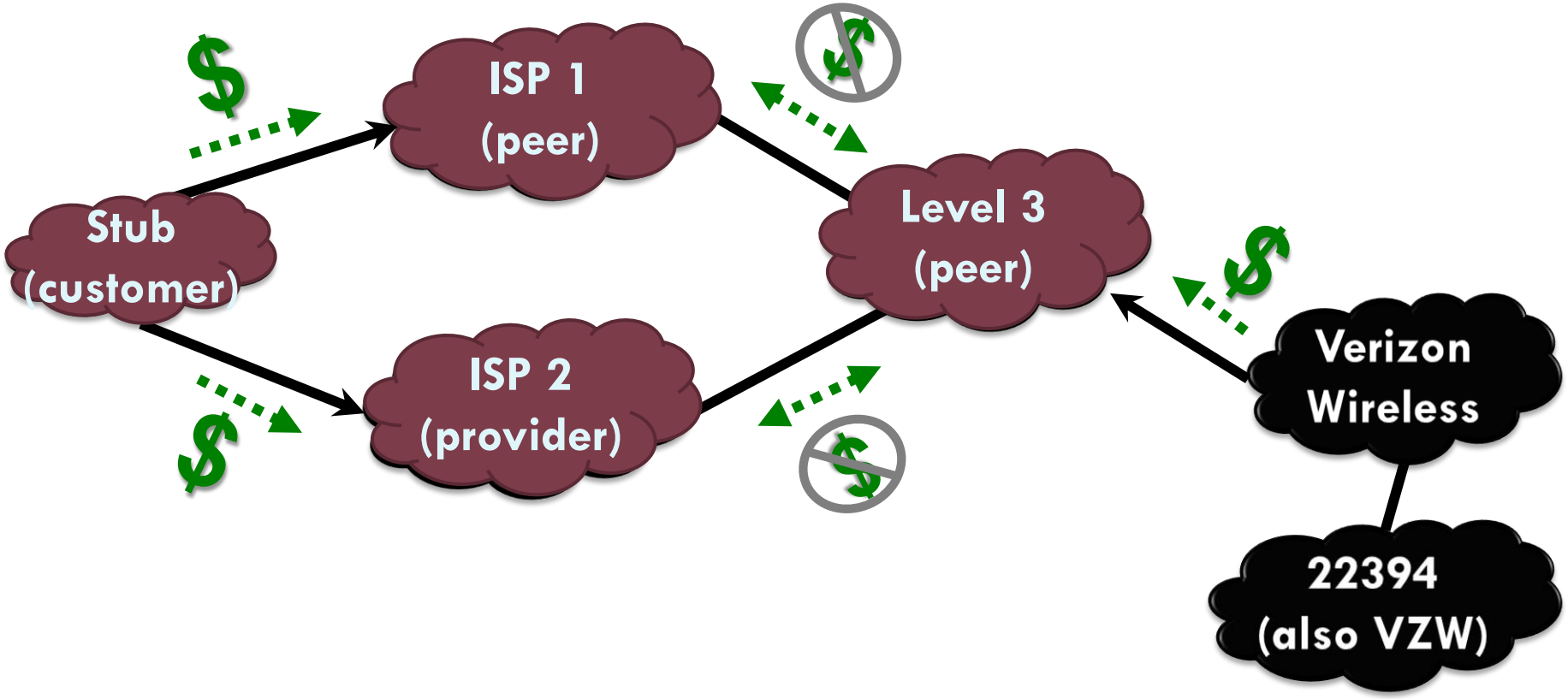


AS Relationships: It's Complicated

- 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

BGP: The Internet's Routing Protocol

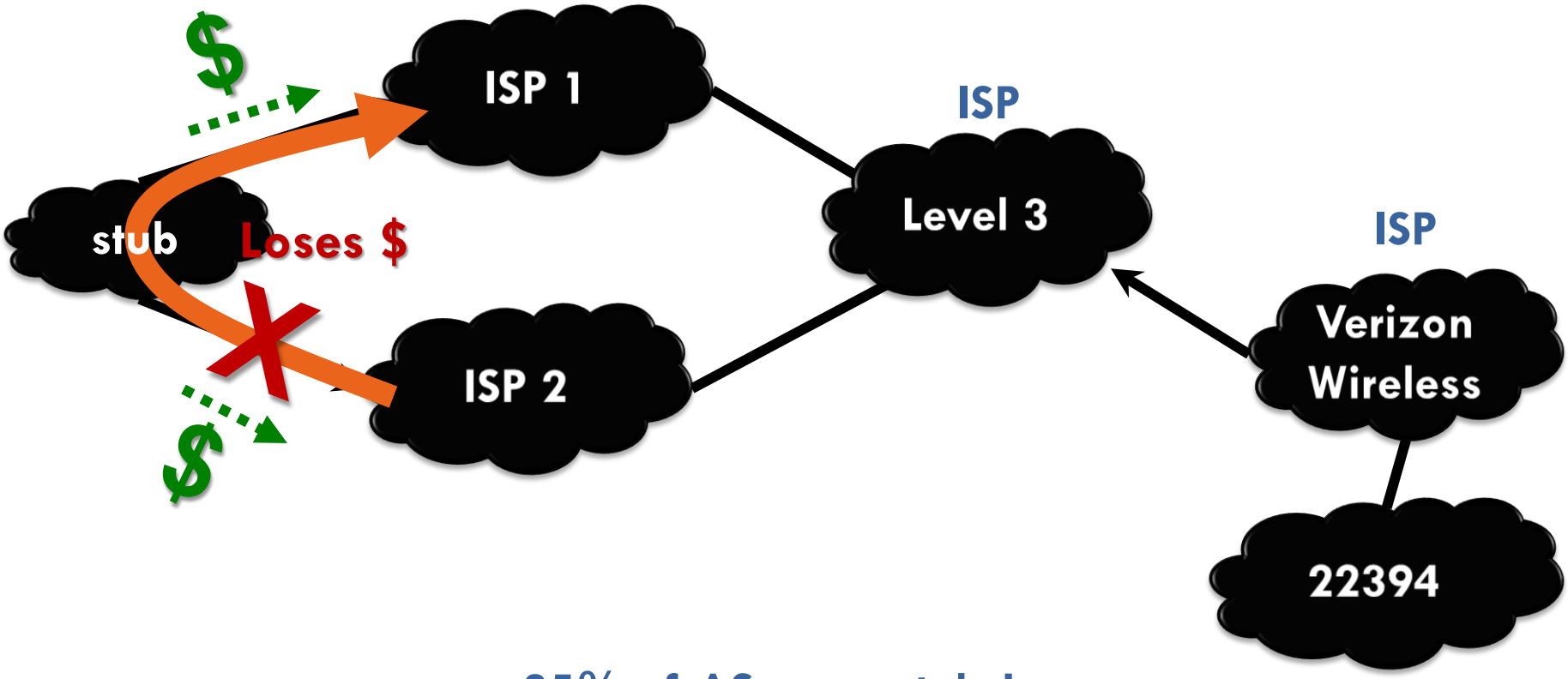
A simple model of AS-level business relationships.



BGP: The Internet's Routing Protocol (2)

A stub is an AS with no customers that never transits traffic.

(Transit = carry traffic from one neighbor to another)

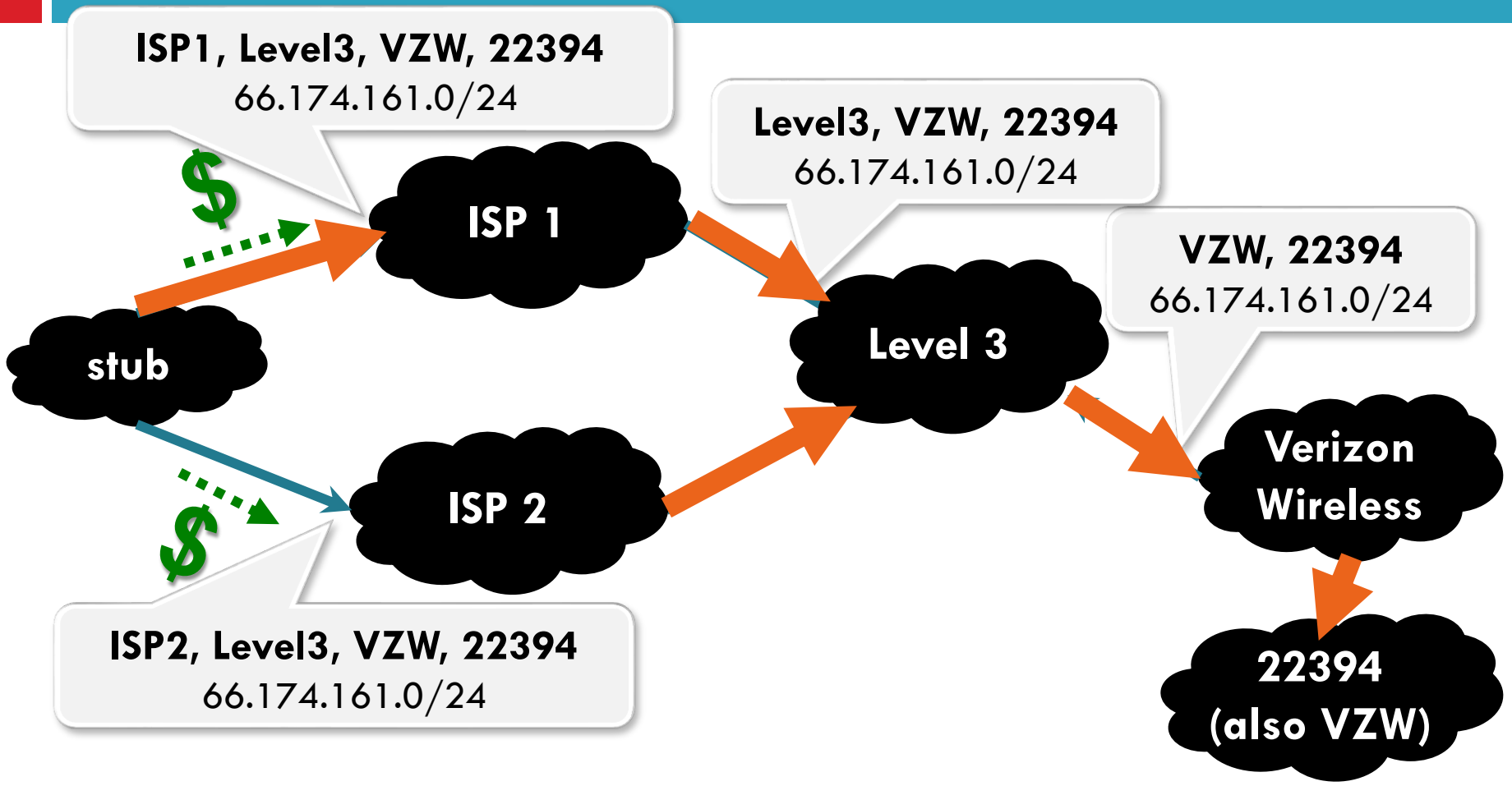


85% of ASes are stubs!

We call the rest (15%) ISPs.

BGP: The Internet's Routing Protocol (3)

BGP sets up paths from ASes to destination IP prefixes.



A model of BGP routing policies:

Prefer cheaper paths. Then, prefer shorter paths.

Standard model of Internet routing

- Proposed by Gao & Rexford 20 years ago
- Based on practices employed by a large ISP
- Provide an intuitive model of path selection and export policy

Standard model of Internet routing

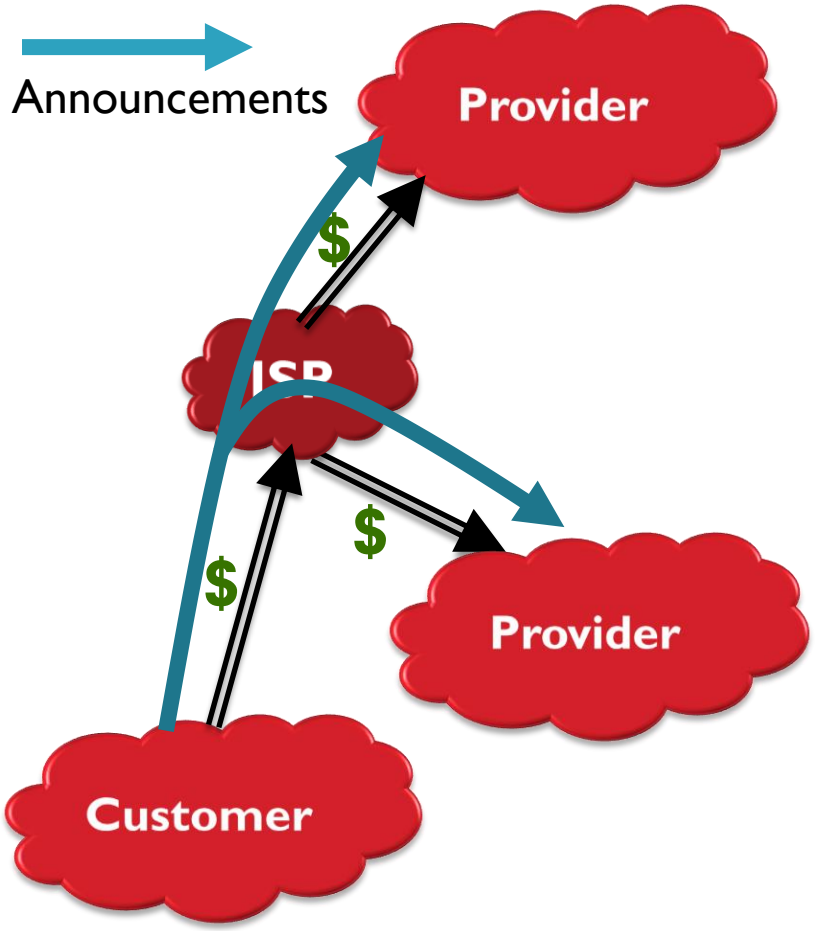
- Proposed by Gao & Rexford 20 years ago

Path Selection:

1. LocalPref: Prefer customer paths over peer paths over provider paths
2. Prefer shorter paths
3. Arbitrary tiebreak

Standard model of Internet routing

- Proposed by Gao & Rexford 20 years ago



Path Selection:

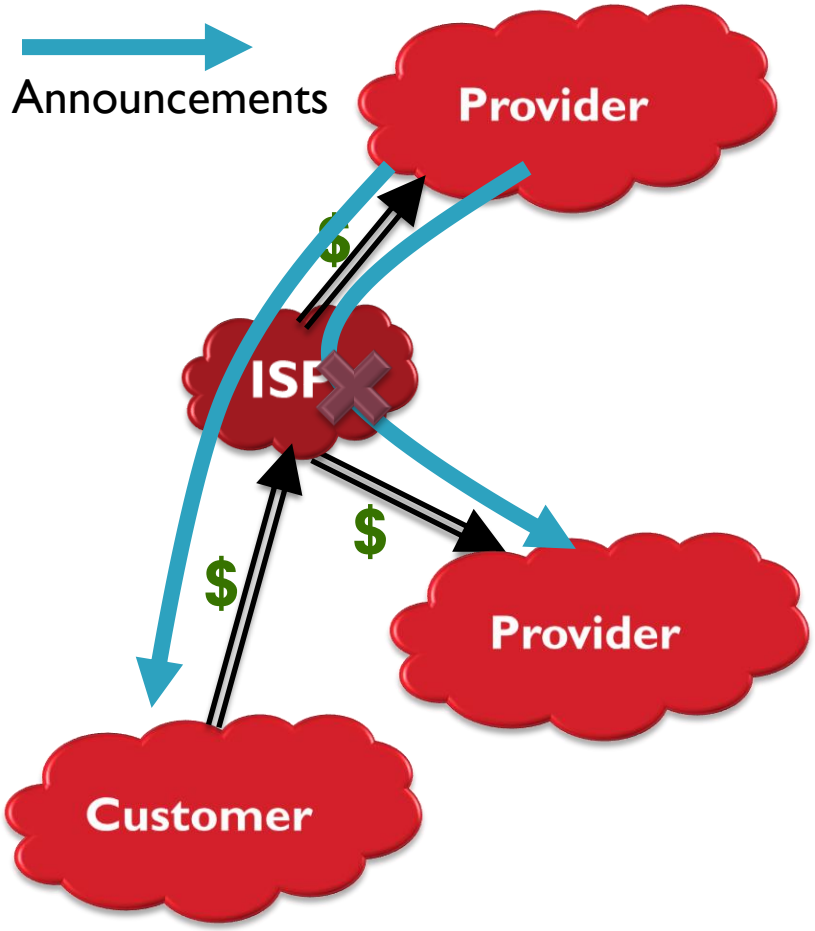
1. LocalPref: Prefer customer paths over peer paths over provider paths
2. Prefer shorter paths
3. Arbitrary tiebreak

Export Policy:

1. Export customer path to all neighbors.
2. Export peer/provider path to all customers.

Standard model of Internet routing

- Proposed by Gao & Rexford 20 years ago



Path Selection:

1. LocalPref: Prefer customer paths over peer paths over provider paths
2. Prefer shorter paths
3. Arbitrary tiebreak

Export Policy:

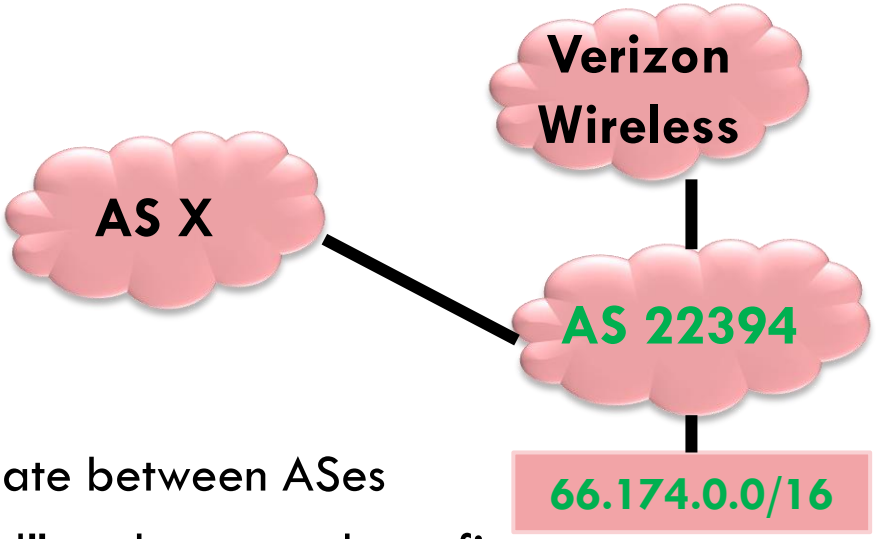
1. Export customer path to all neighbors.
2. Export peer/provider path to all customers.

BGP-related Hijacks

Normal operation

- Origin AS announces prefix
- Route announcements propagate between ASes
- Helps ASes learn about "good" paths to reach prefix

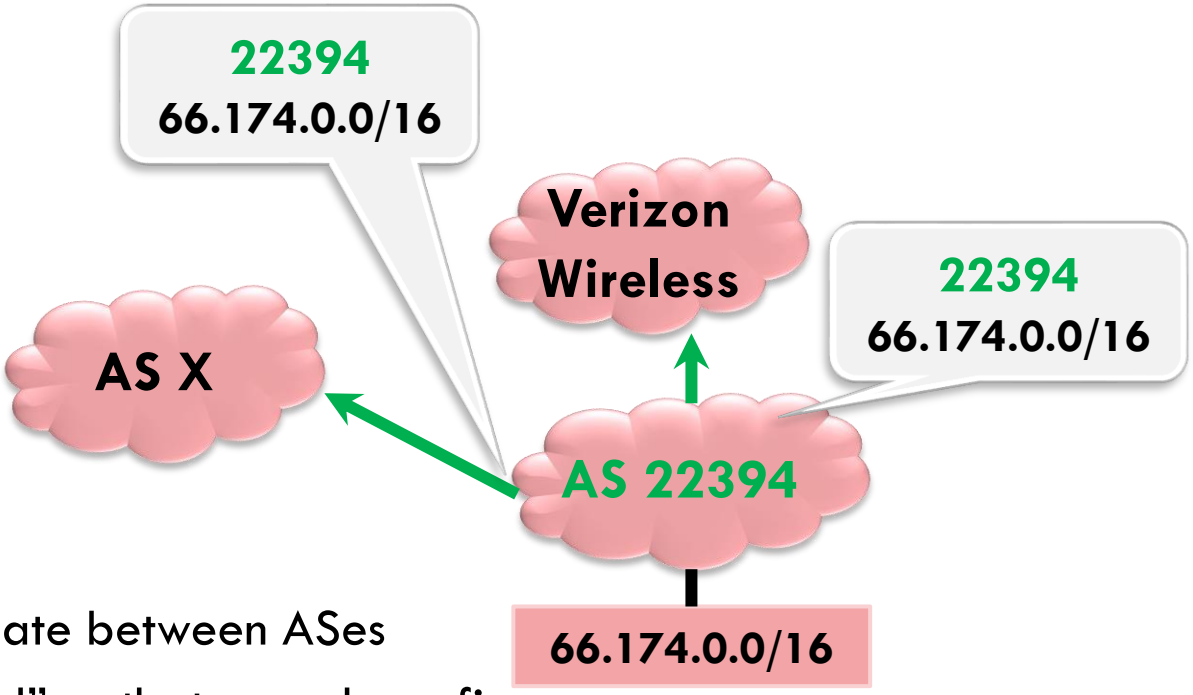
BGP-related Hijacks



Normal operation

- Origin AS announces prefix
- Route announcements propagate between ASes
- Helps ASes learn about "good" paths to reach prefix

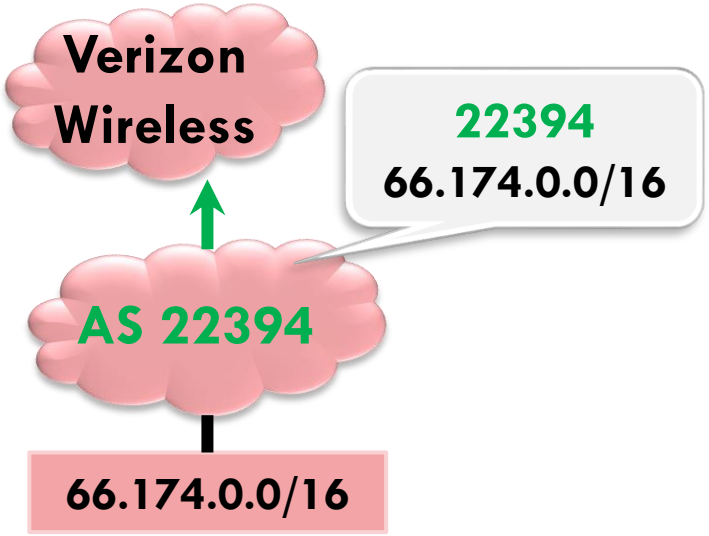
BGP-related Hijacks



Normal operation

- Origin AS announces prefix
- Route announcements propagate between ASes
- Helps ASes learn about "good" paths to reach prefix

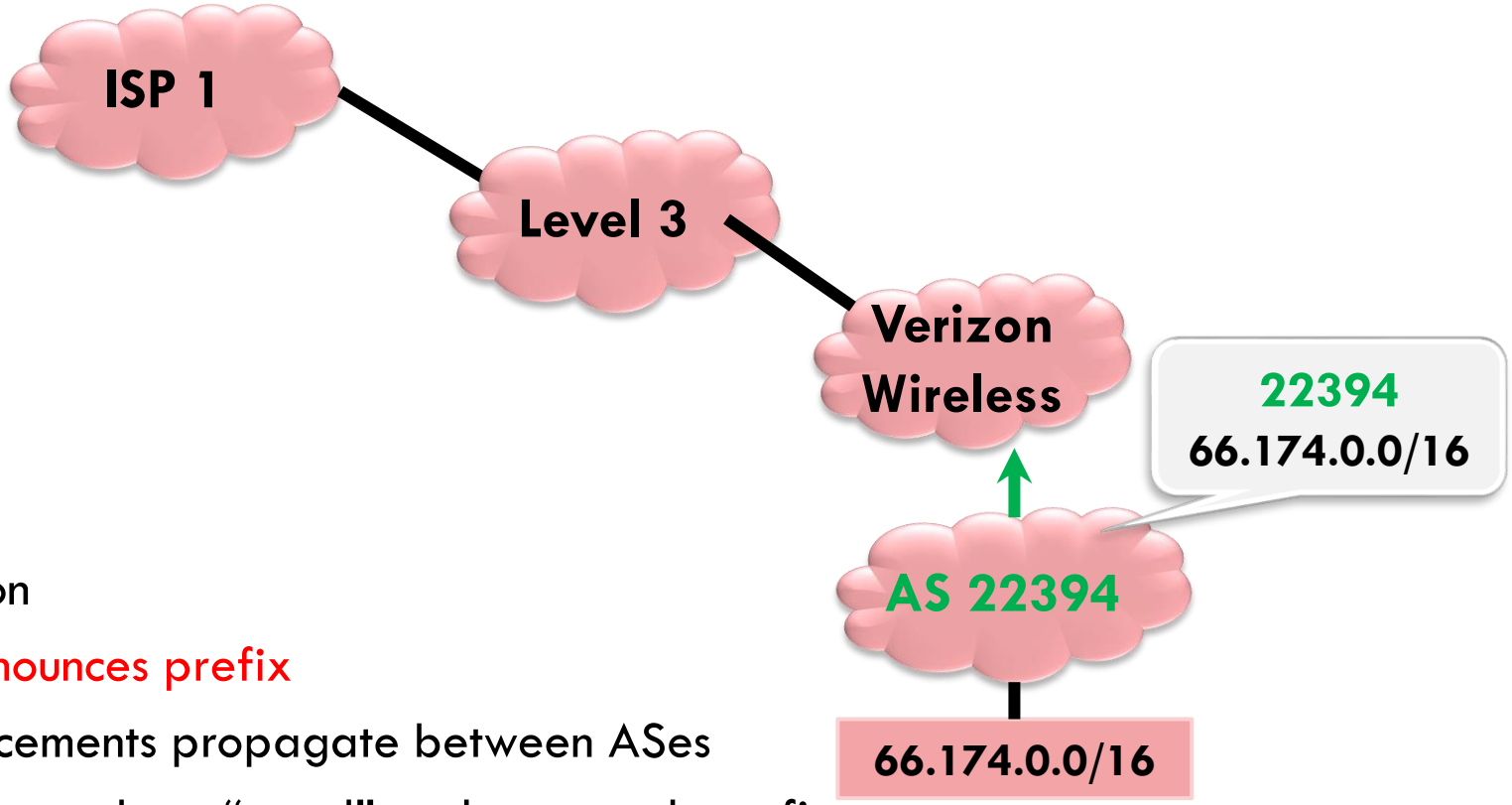
BGP-related Hijacks



Normal operation

- Origin AS announces prefix
- Route announcements propagate between ASes
- Helps ASes learn about "good" paths to reach prefix

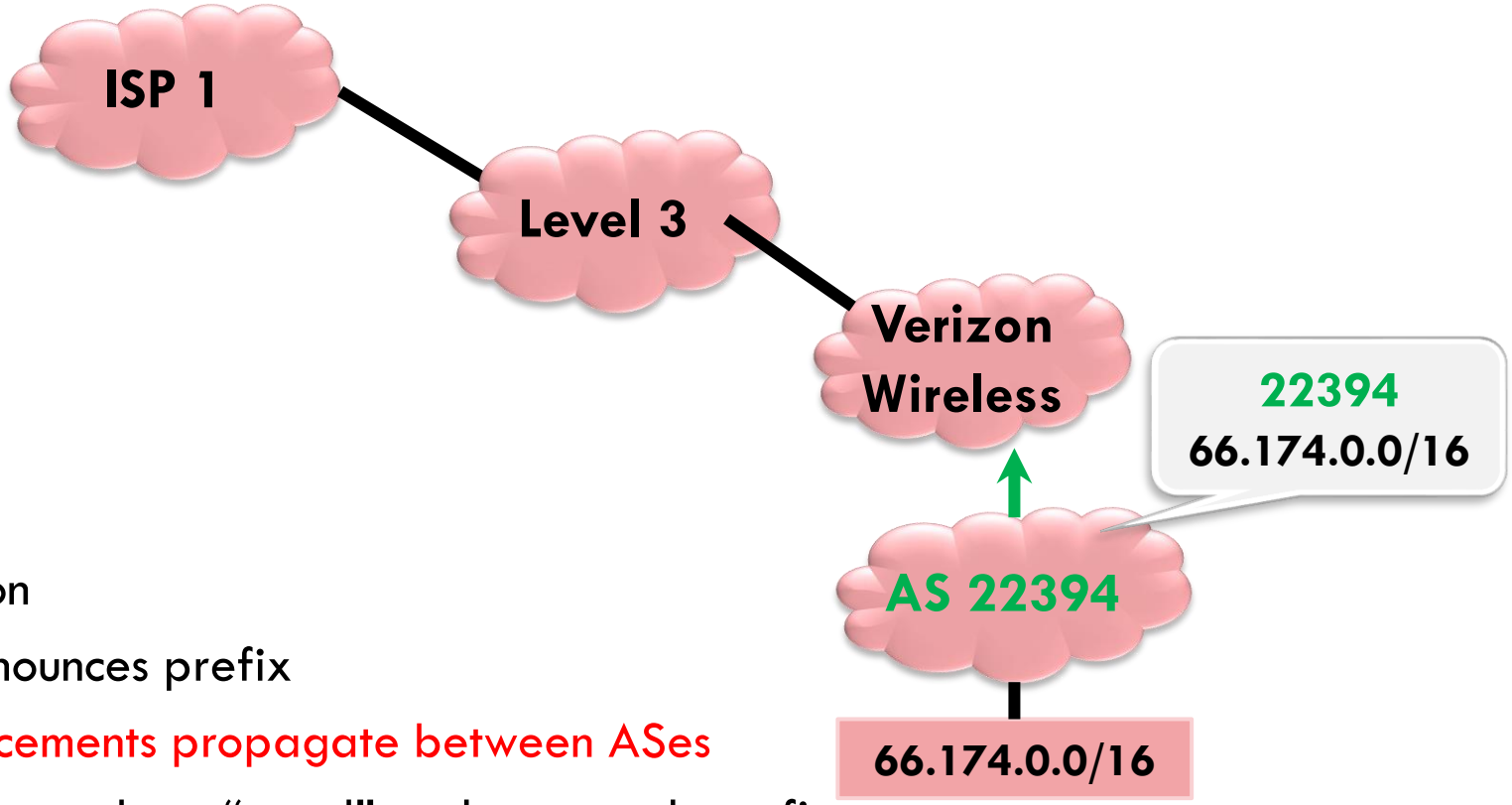
BGP-related Hijacks



Normal operation

- Origin AS announces prefix
- Route announcements propagate between ASes
- Helps ASes learn about "good" paths to reach prefix

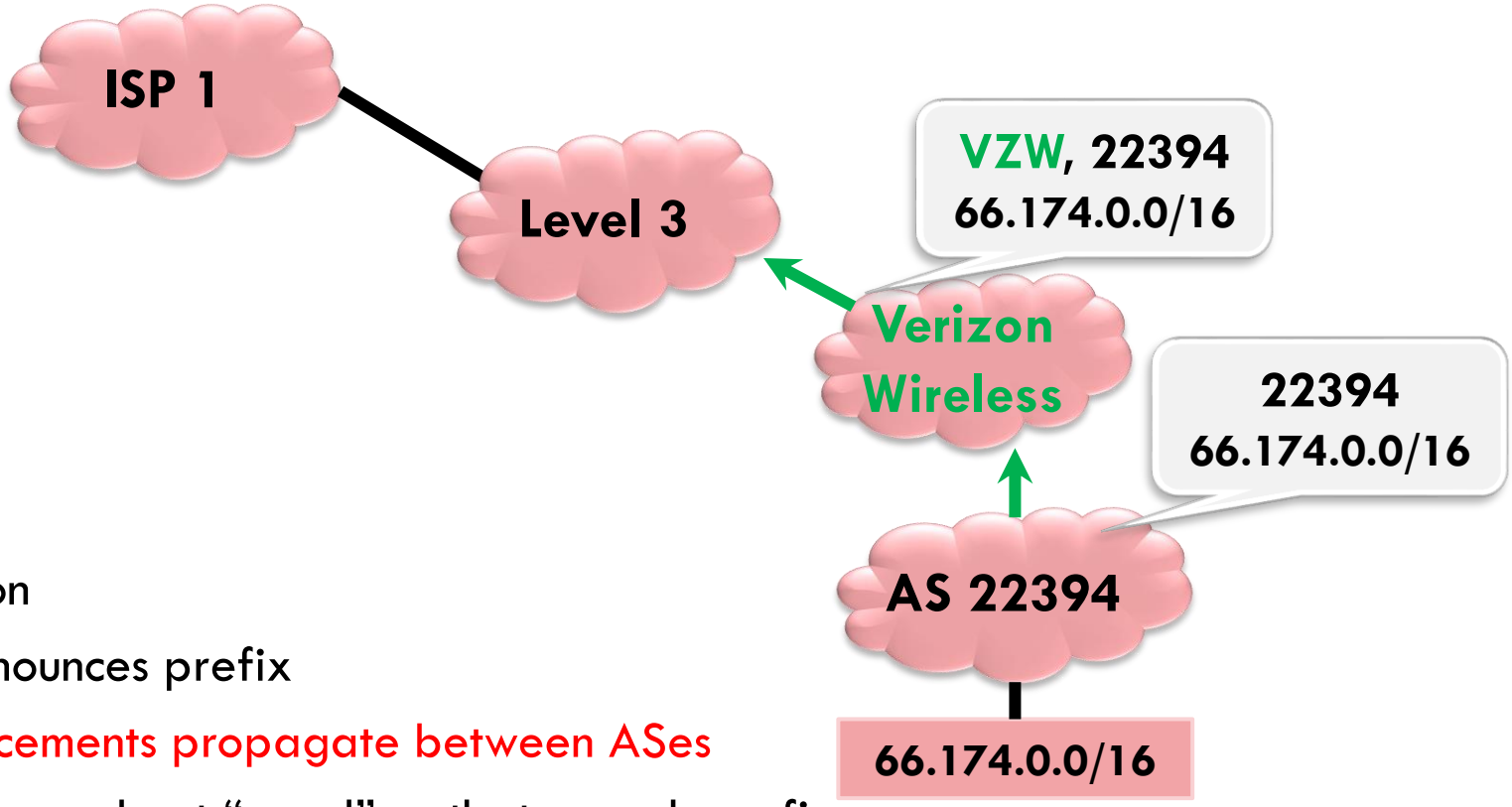
BGP-related Hijacks



Normal operation

- Origin AS announces prefix
- Route announcements propagate between ASes
- Helps ASes learn about "good" paths to reach prefix

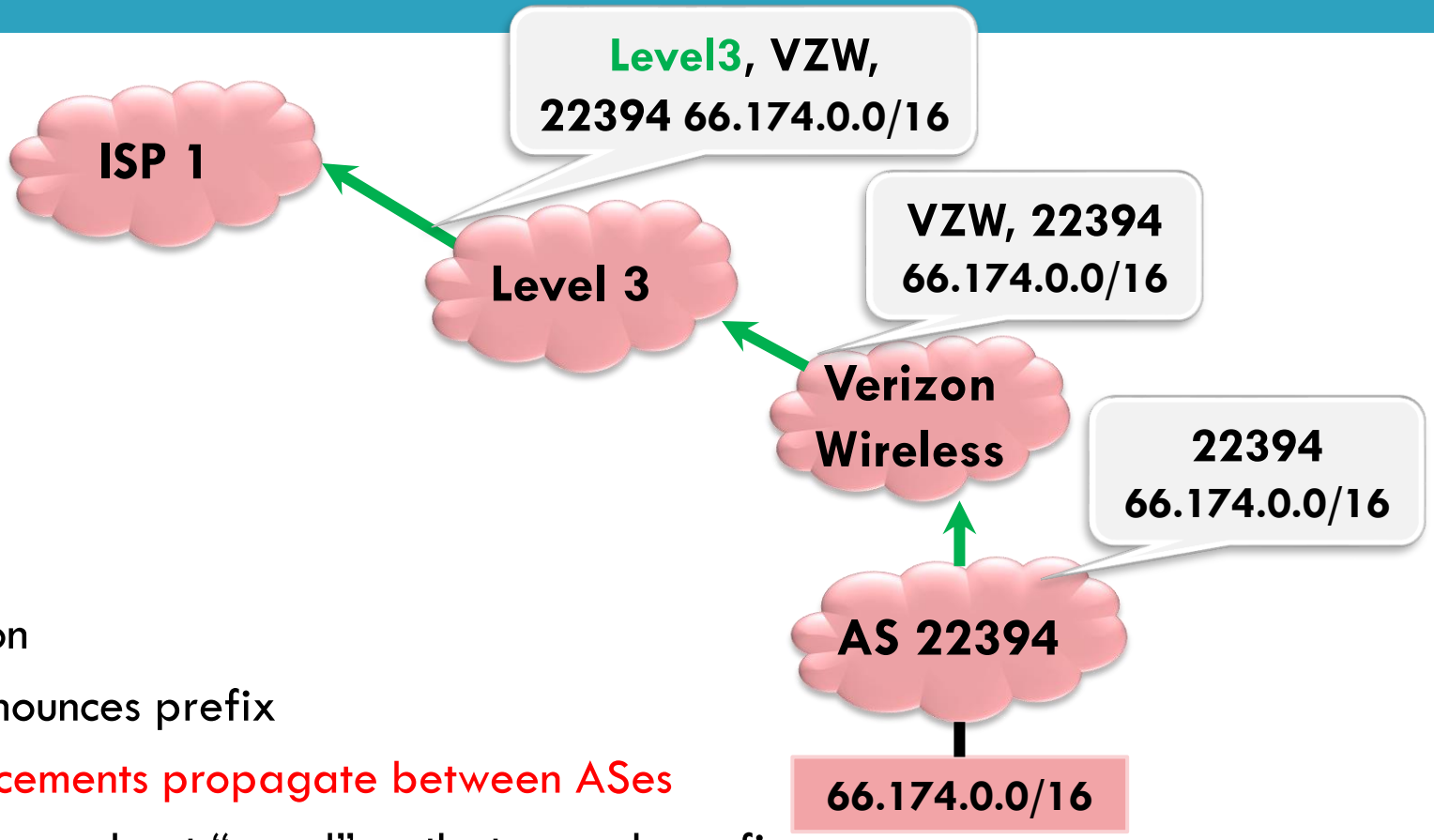
BGP-related Hijacks



Normal operation

- Origin AS announces prefix
- Route announcements propagate between ASes
- Helps ASes learn about "good" paths to reach prefix

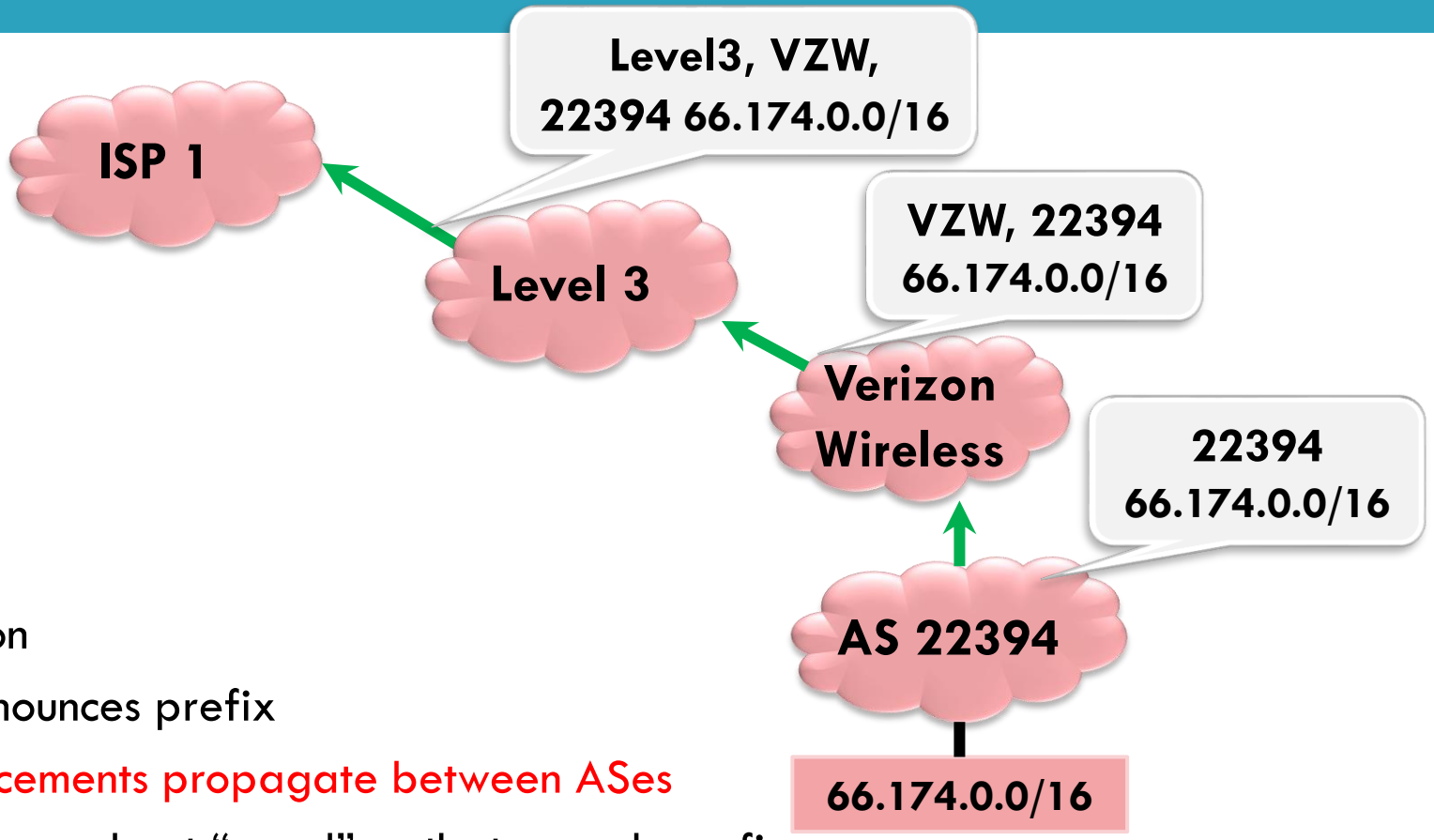
BGP-related Hijacks



Normal operation

- Origin AS announces prefix
- Route announcements propagate between ASes
- Helps ASes learn about "good" paths to reach prefix

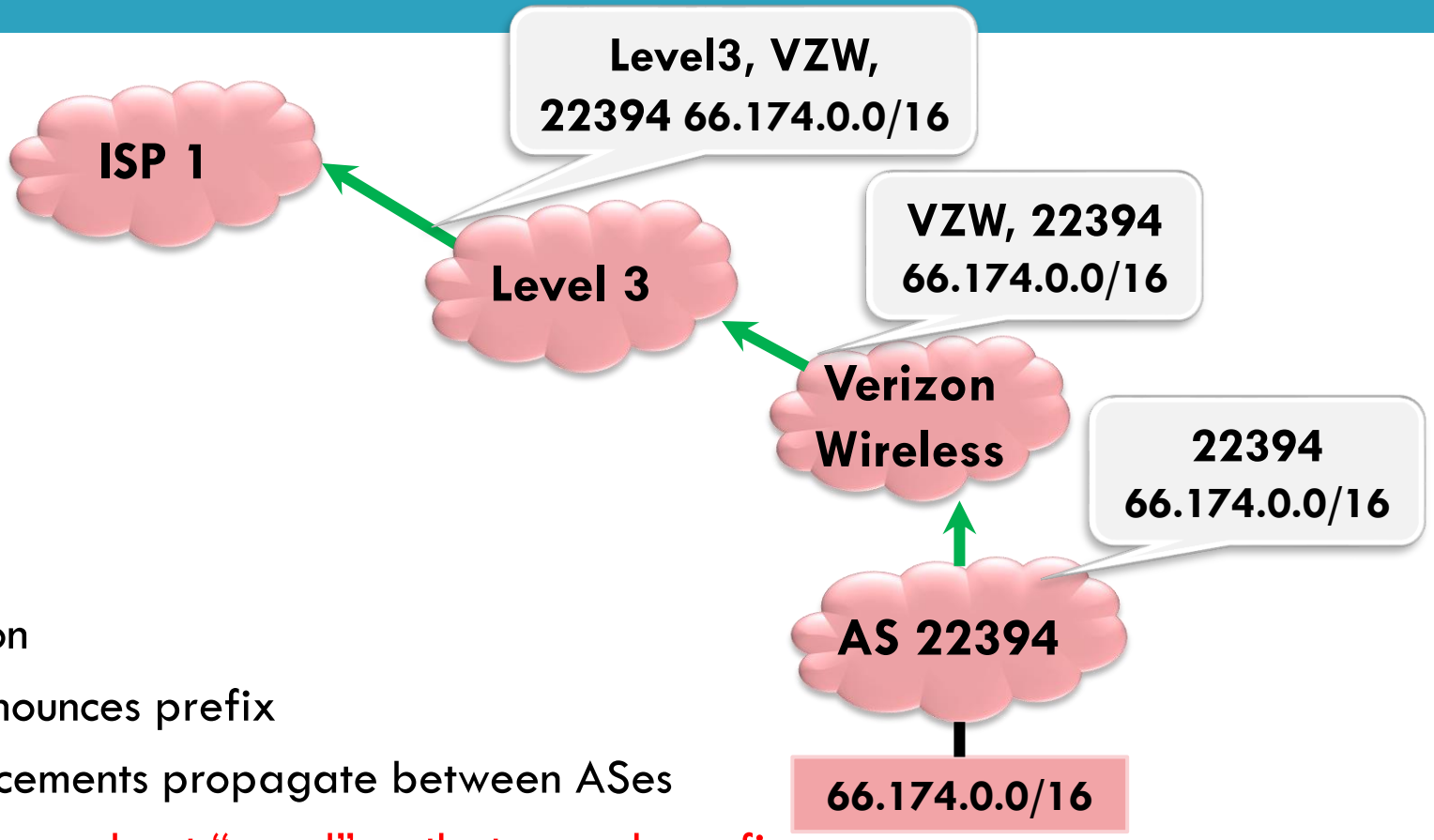
BGP-related Hijacks



Normal operation

- Origin AS announces prefix
- Route announcements propagate between ASes
- Helps ASes learn about "good" paths to reach prefix

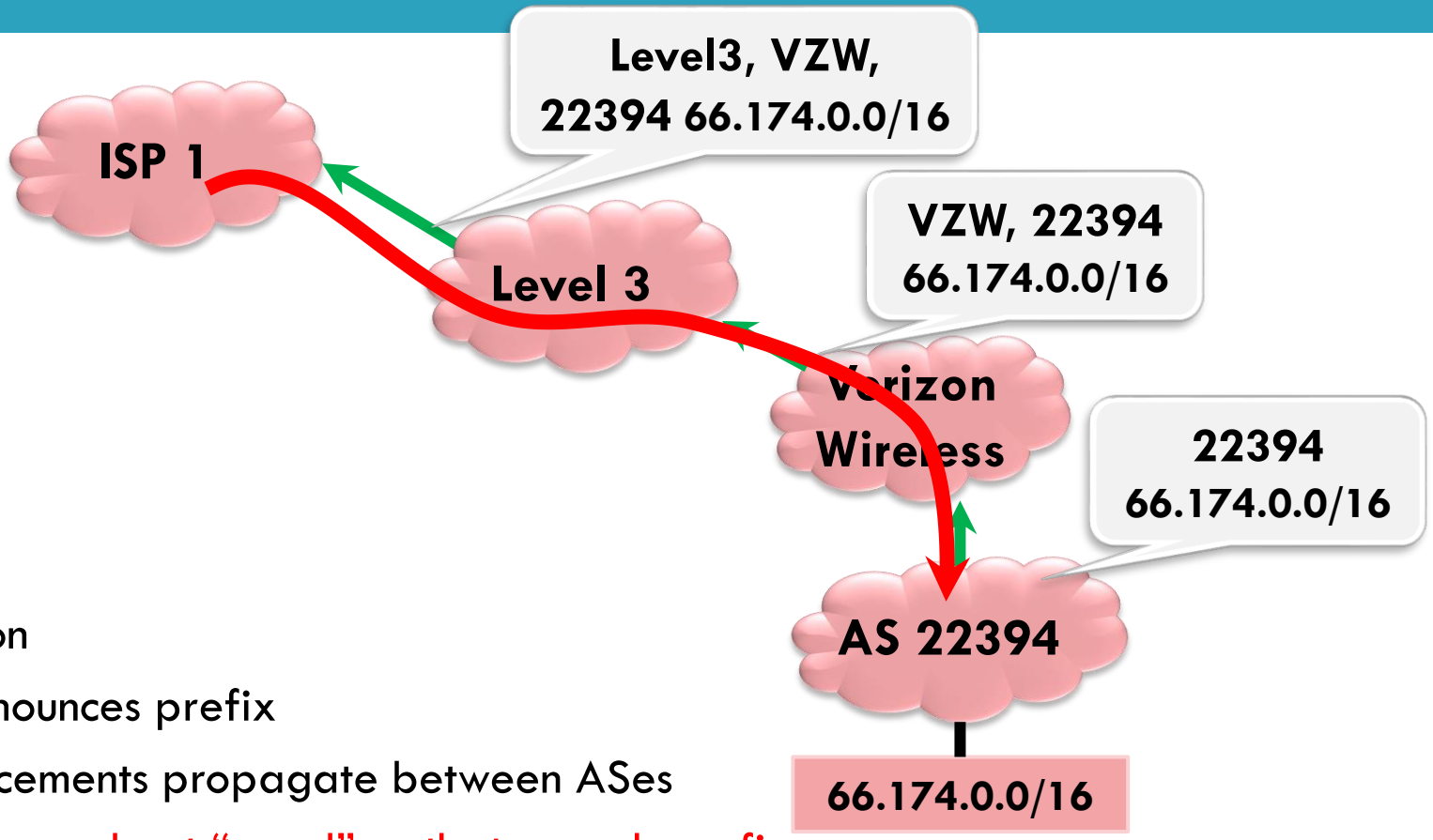
BGP-related Hijacks



Normal operation

- Origin AS announces prefix
- Route announcements propagate between ASes
- Helps ASes learn about "good" paths to reach prefix

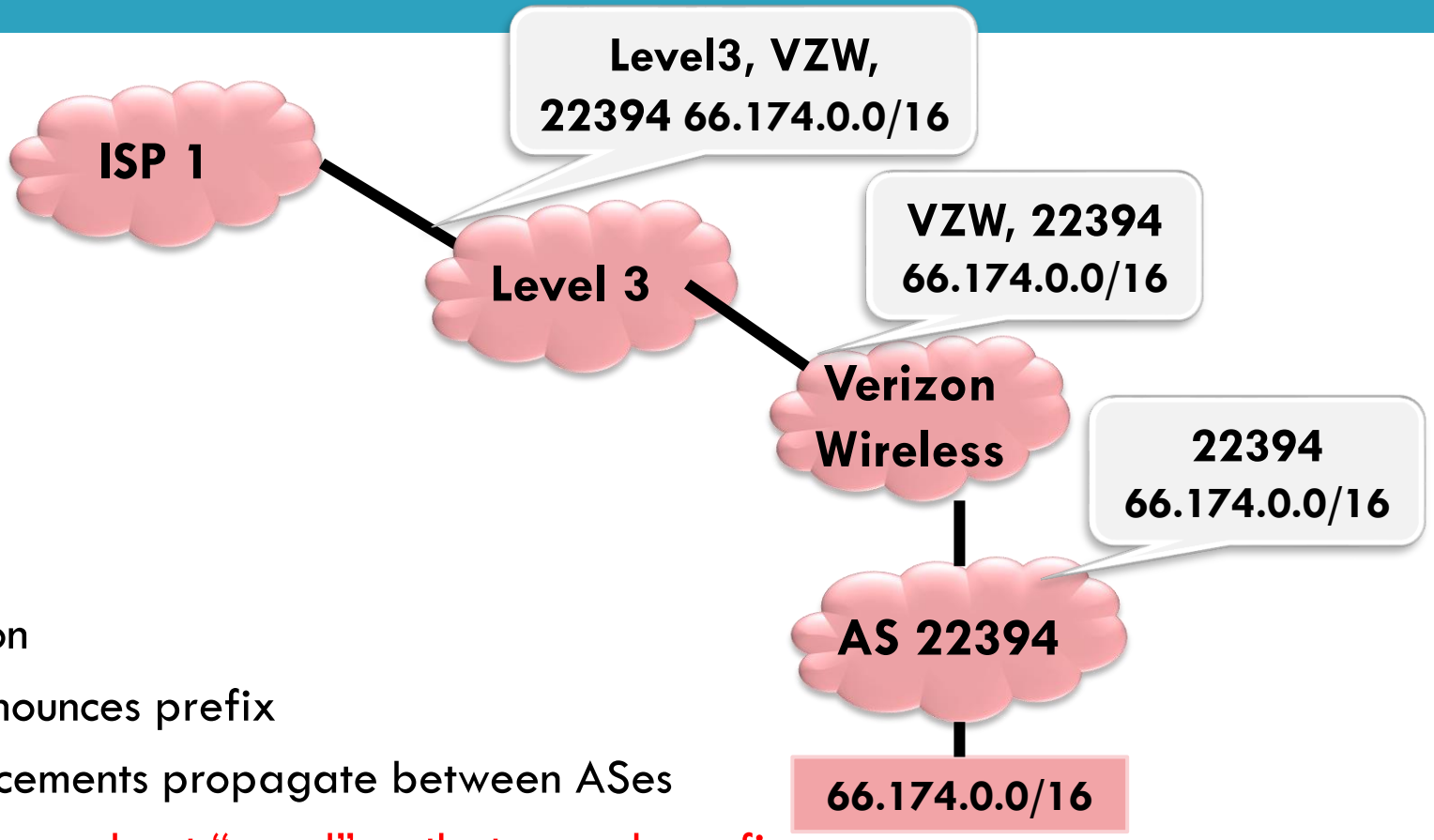
BGP-related Hijacks



Normal operation

- Origin AS announces prefix
- Route announcements propagate between ASes
- Helps ASes learn about "good" paths to reach prefix

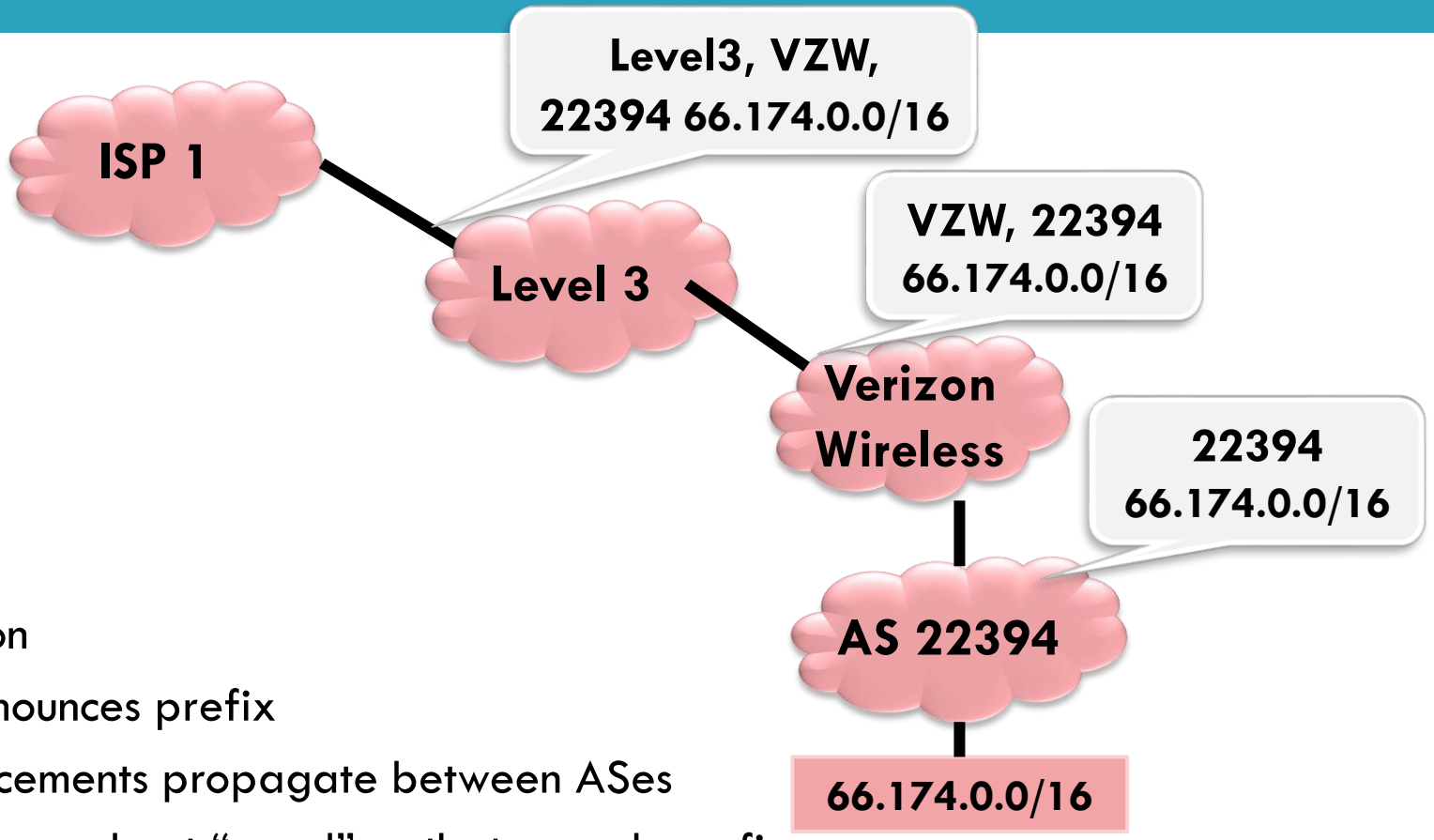
BGP-related Hijacks



Normal operation

- Origin AS announces prefix
- Route announcements propagate between ASes
- Helps ASes learn about "good" paths to reach prefix

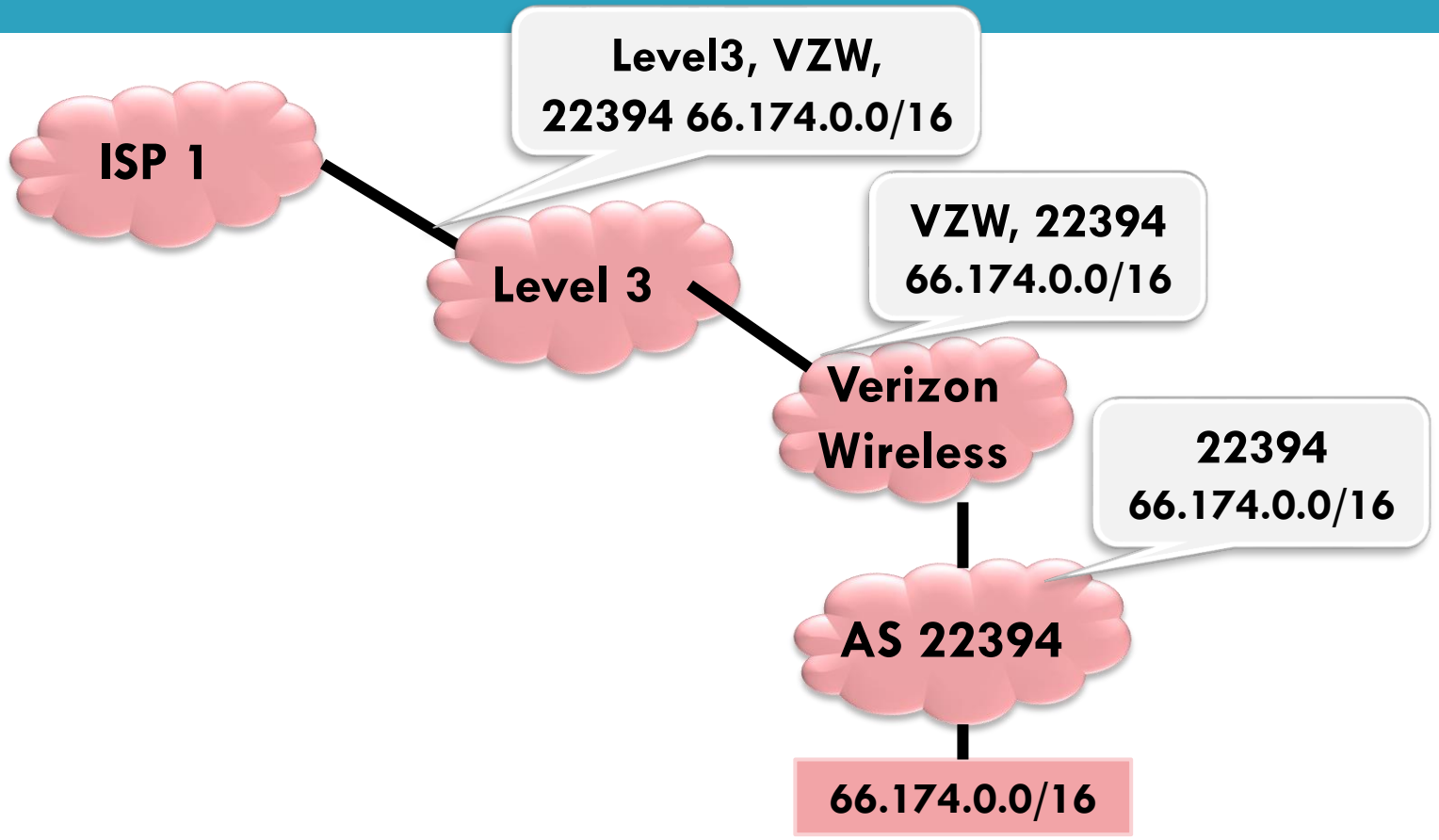
BGP-related Hijacks



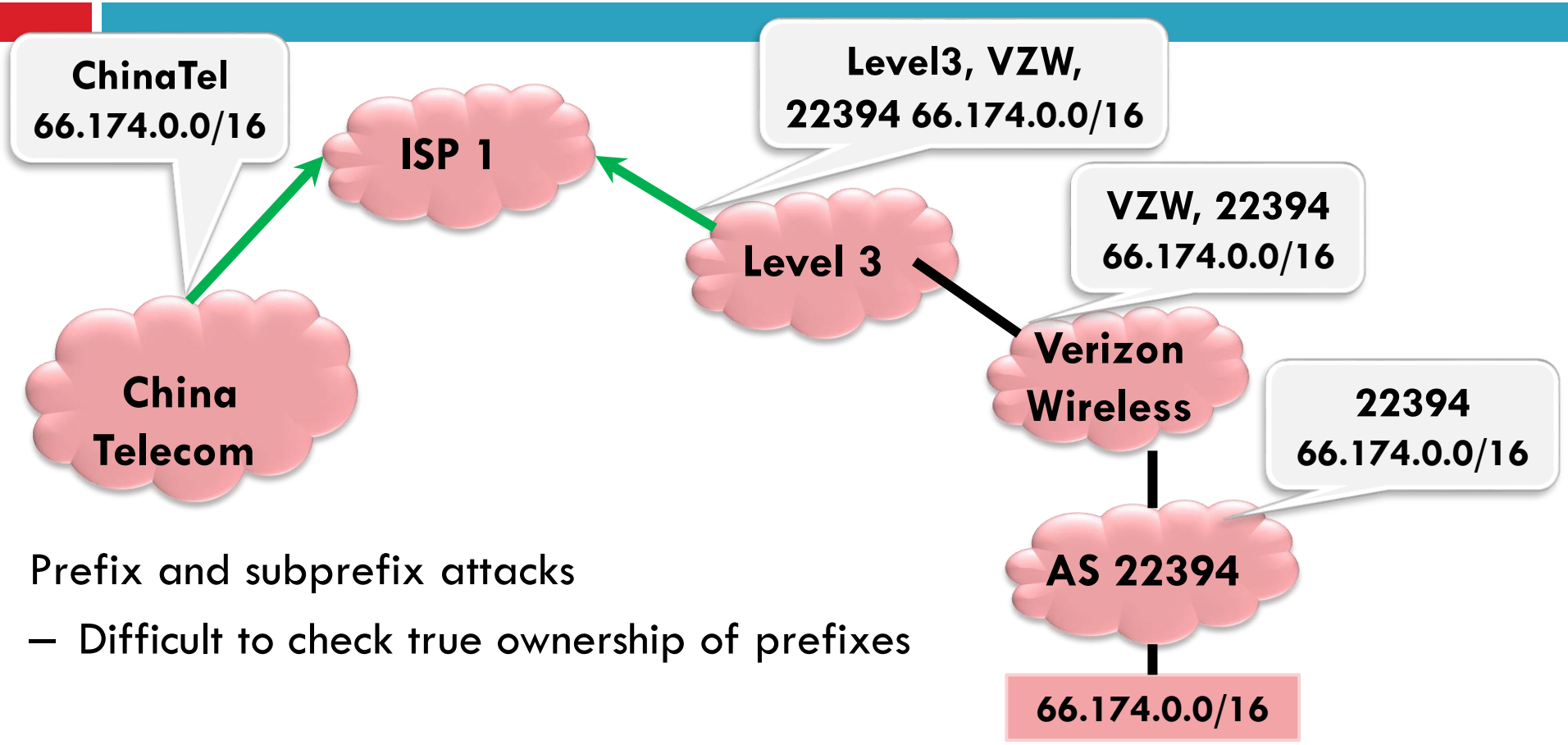
Normal operation

- Origin AS announces prefix
- Route announcements propagate between ASes
- Helps ASes learn about "good" paths to reach prefix

BGP-related Hijacks



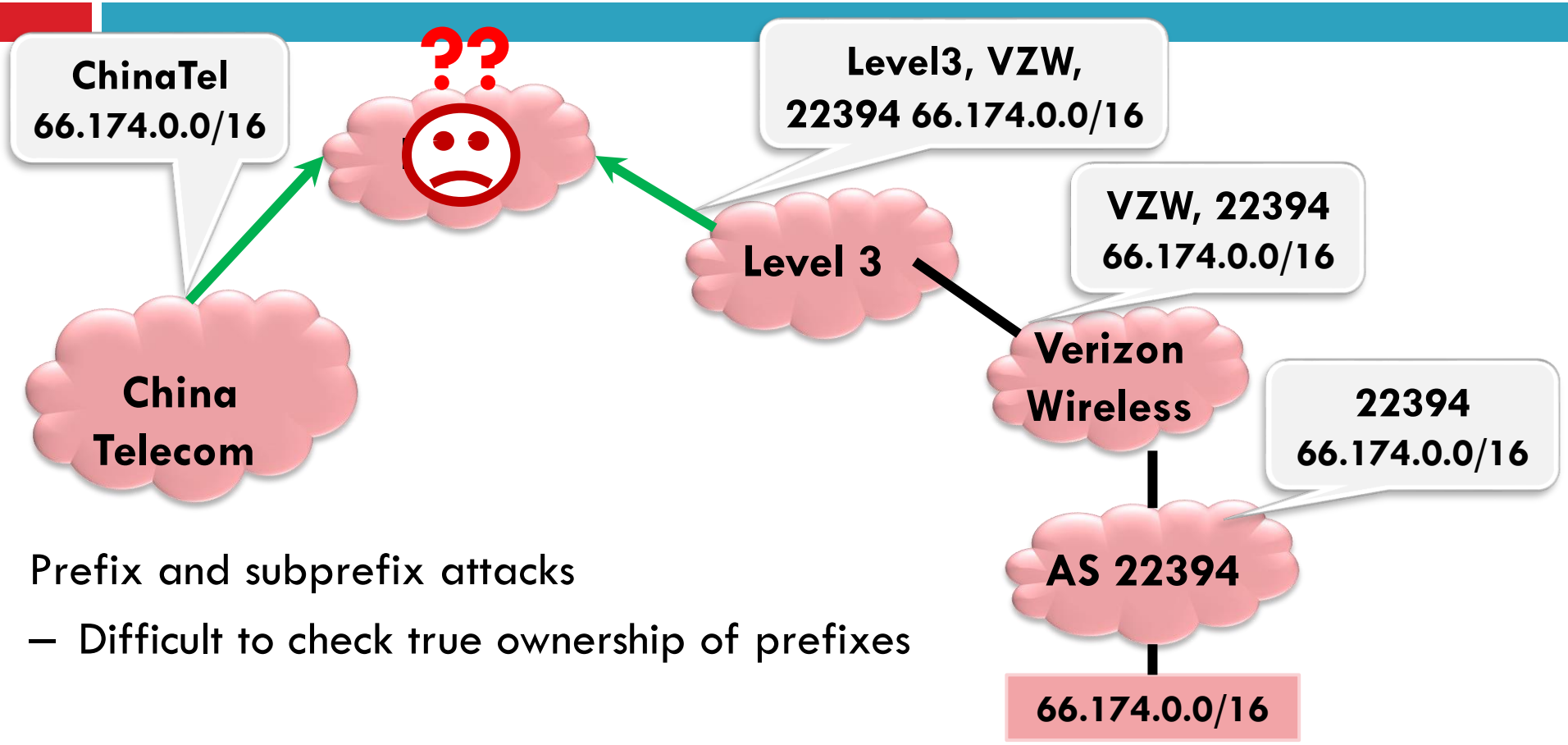
BGP-related Hijacks



Prefix and subprefix attacks

– Difficult to check true ownership of prefixes

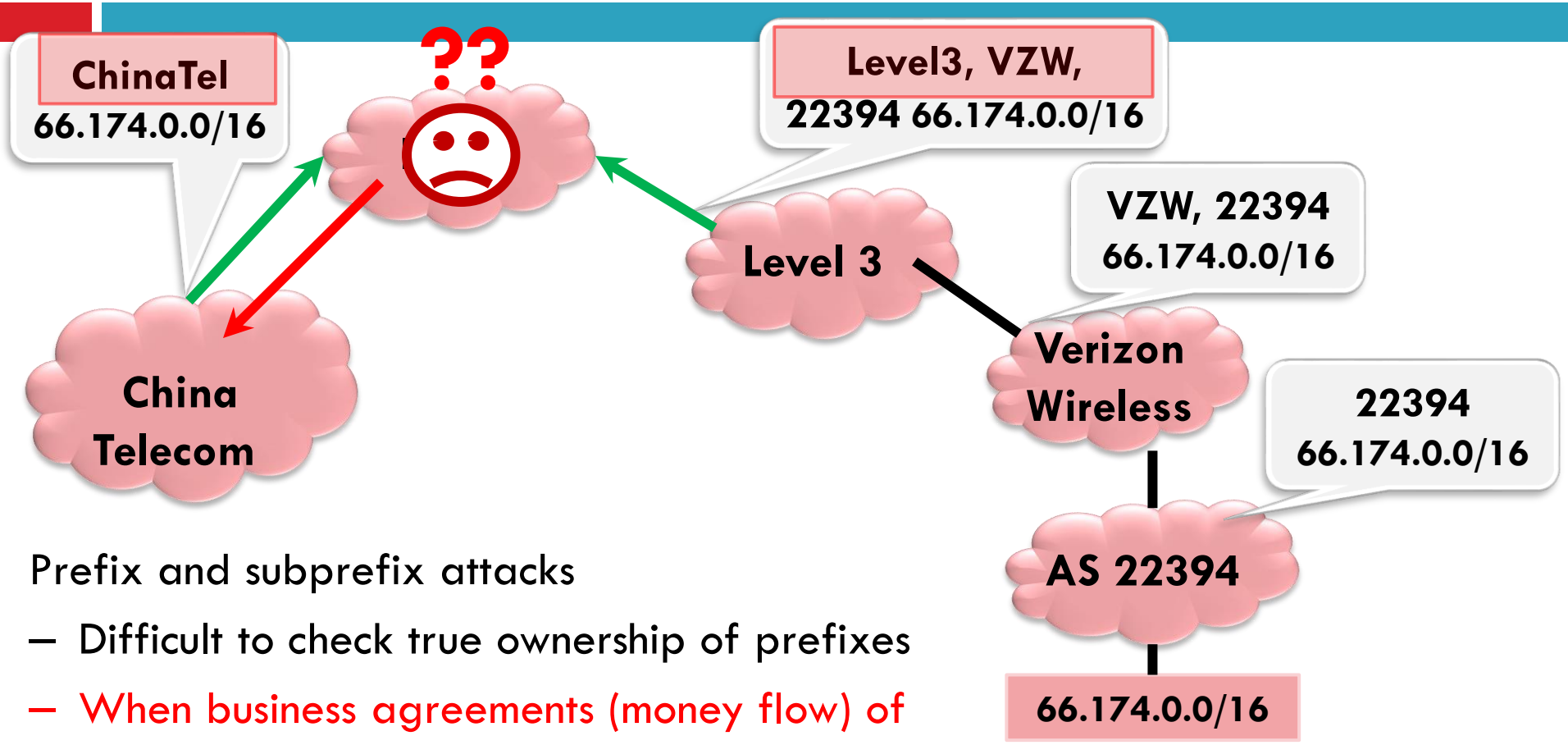
BGP-related Hijacks



Prefix and subprefix attacks

– Difficult to check true ownership of prefixes

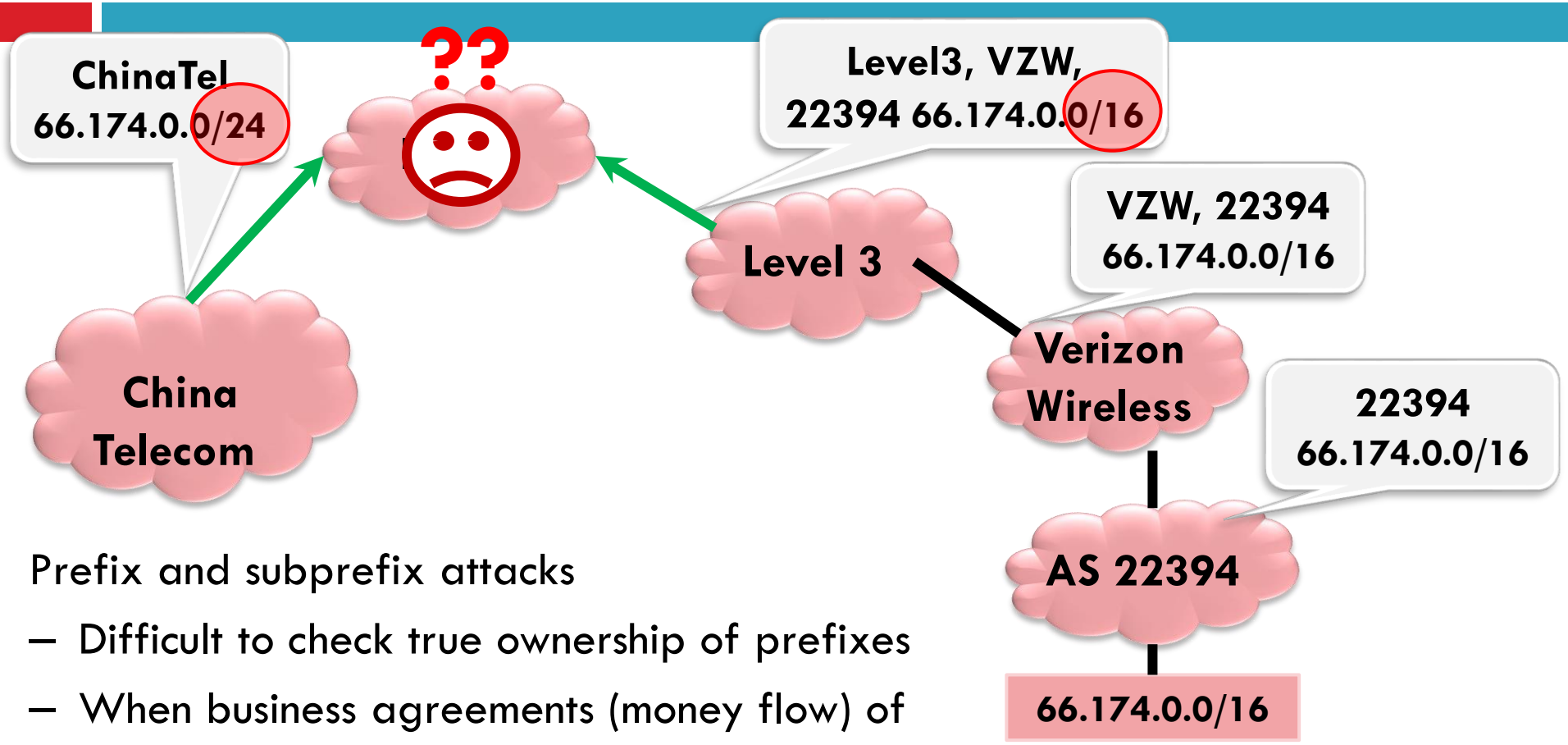
BGP-related Hijacks



Prefix and subprefix attacks

- Difficult to check true ownership of prefixes
- When business agreements (money flow) of same type, typically pick "shorter" path

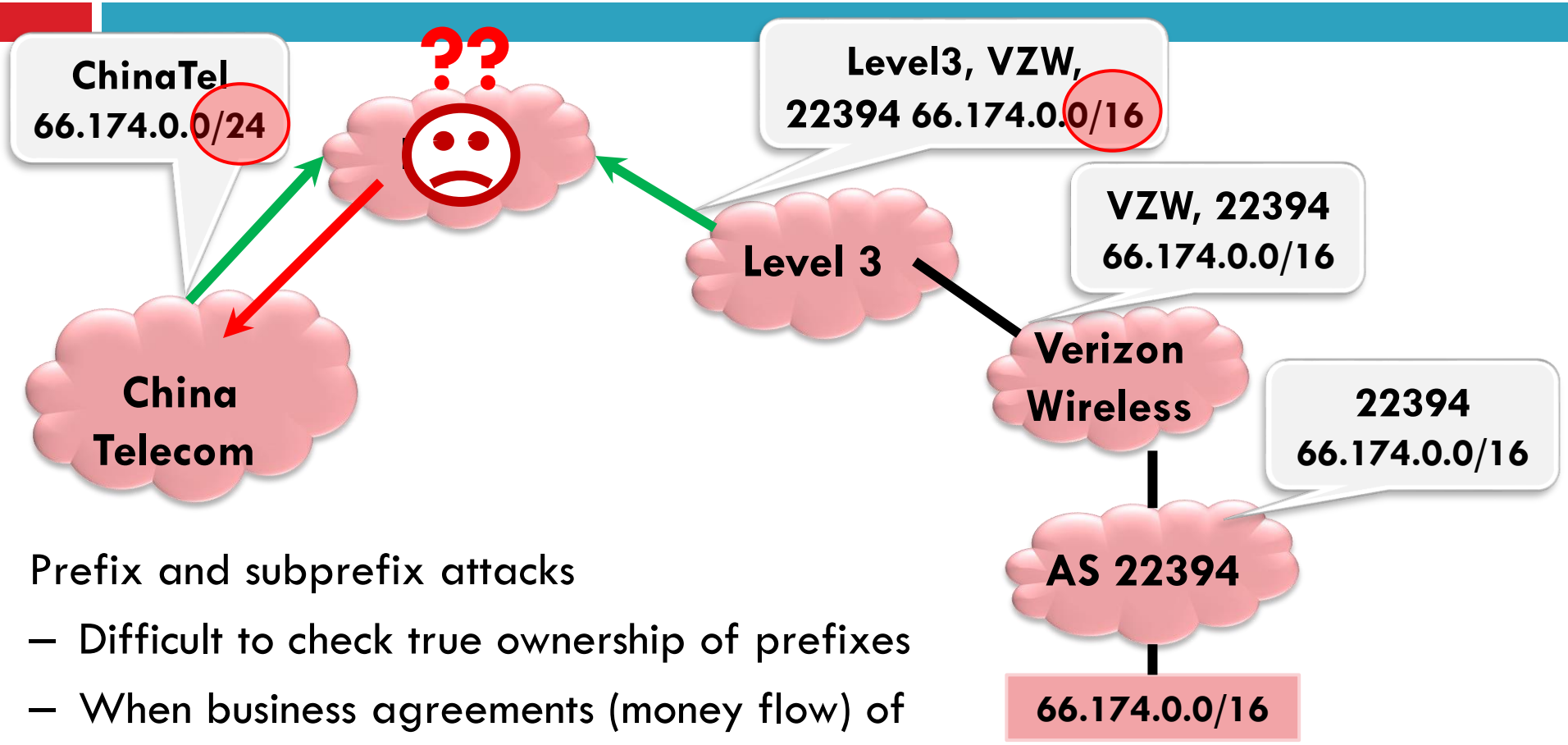
BGP-related Hijacks



Prefix and subprefix attacks

- Difficult to check true ownership of prefixes
- When business agreements (money flow) of same type, typically pick "shorter" path
- Or more specific prefix (subprefix attack)

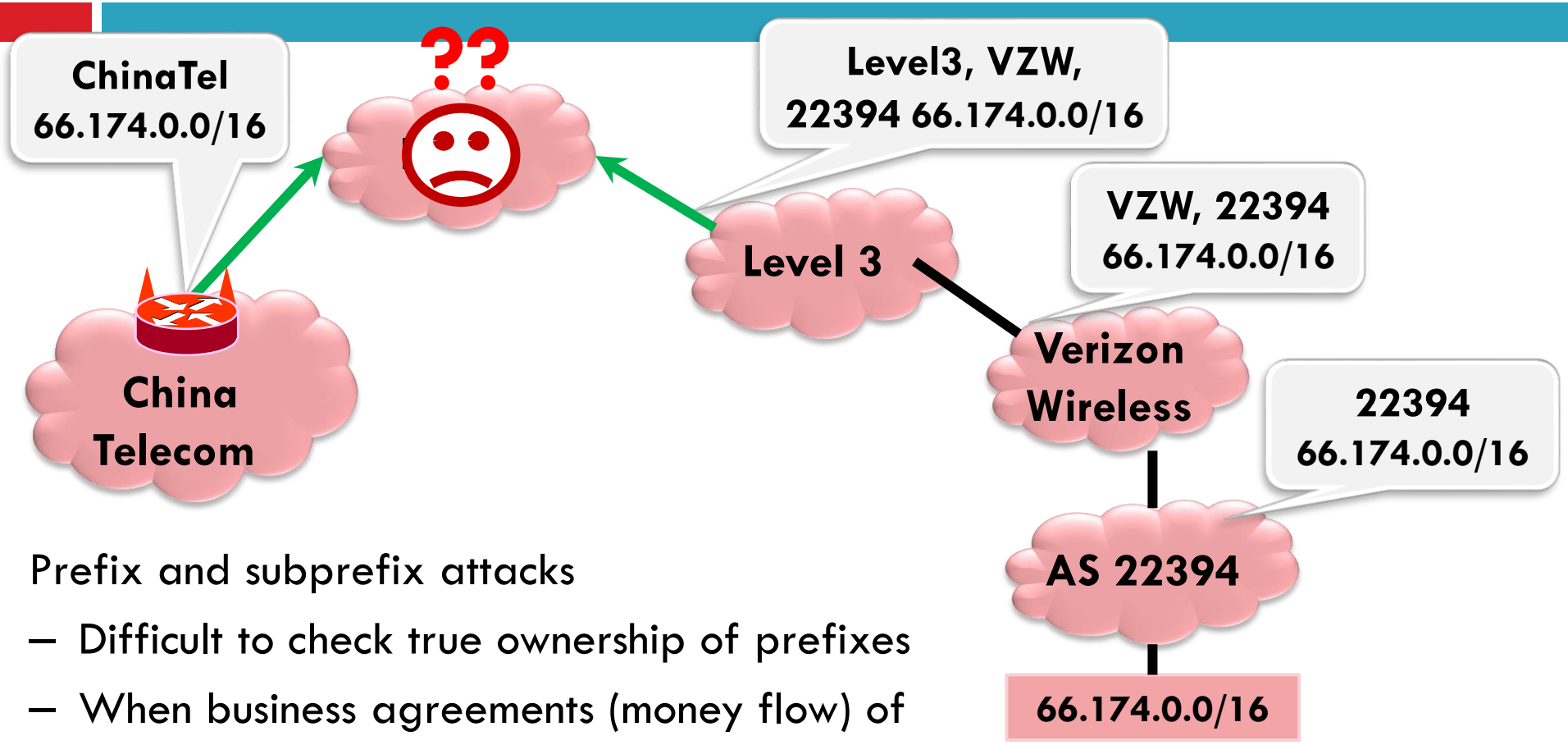
BGP-related Hijacks



Prefix and subprefix attacks

- Difficult to check true ownership of prefixes
- When business agreements (money flow) of same type, typically pick "shorter" path
- Or more specific prefix (subprefix attack)

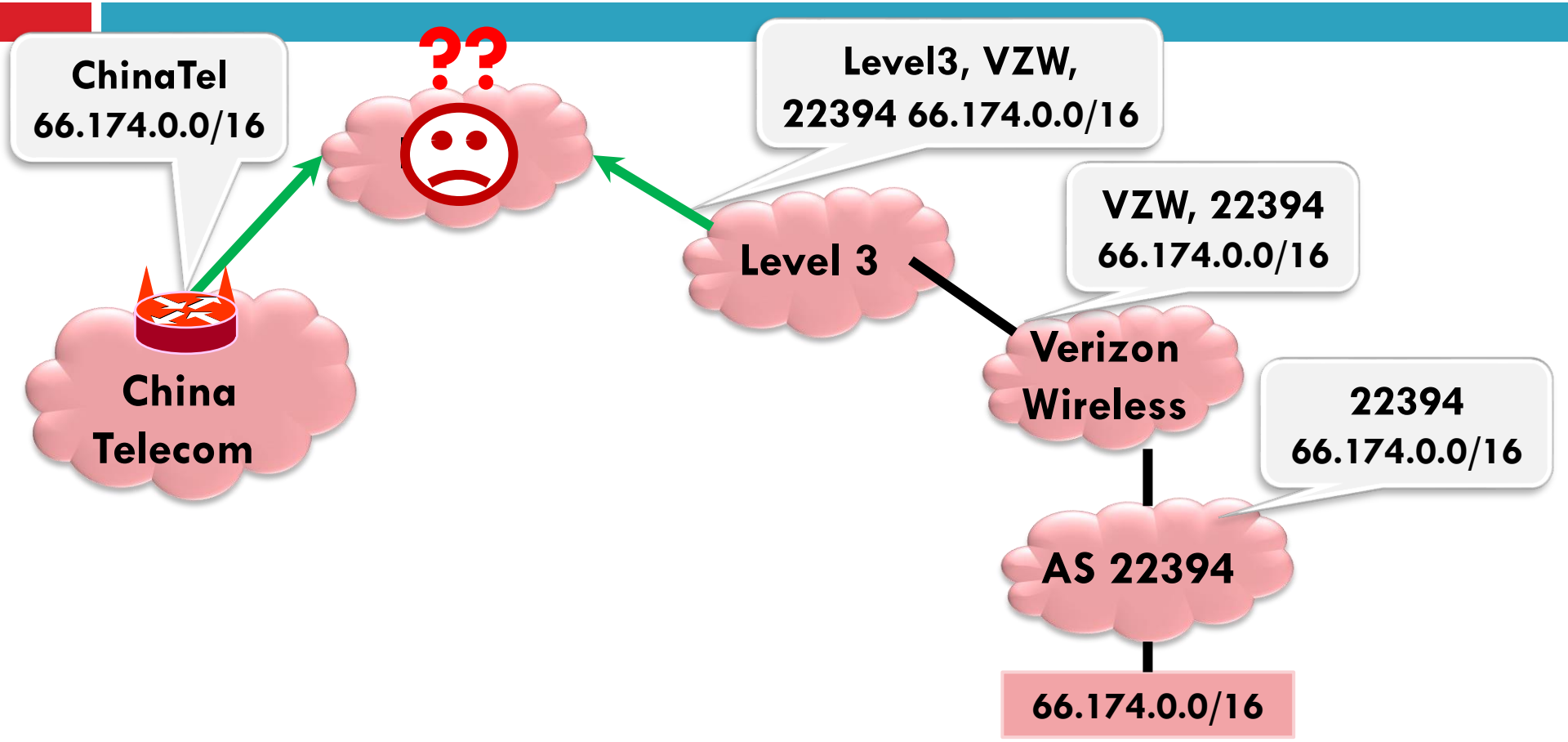
BGP-related Hijacks



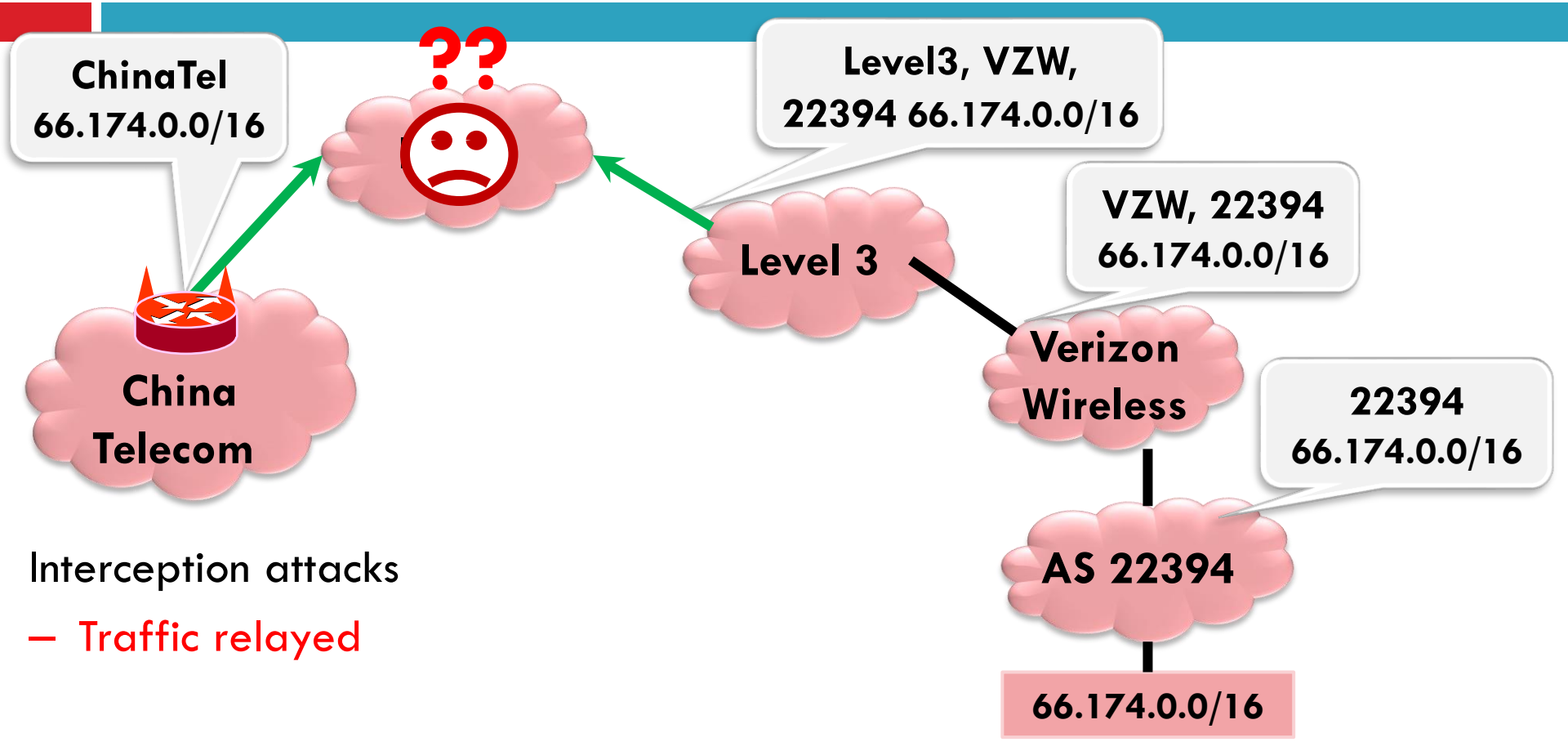
Prefix and subprefix attacks

- Difficult to check true ownership of prefixes
- When business agreements (money flow) of same type, typically pick "shorter" path
- Or more specific prefix (subprefix attack)
- Apr. 2010: ChinaTel announces 50K prefixes

BGP-related Hijacks



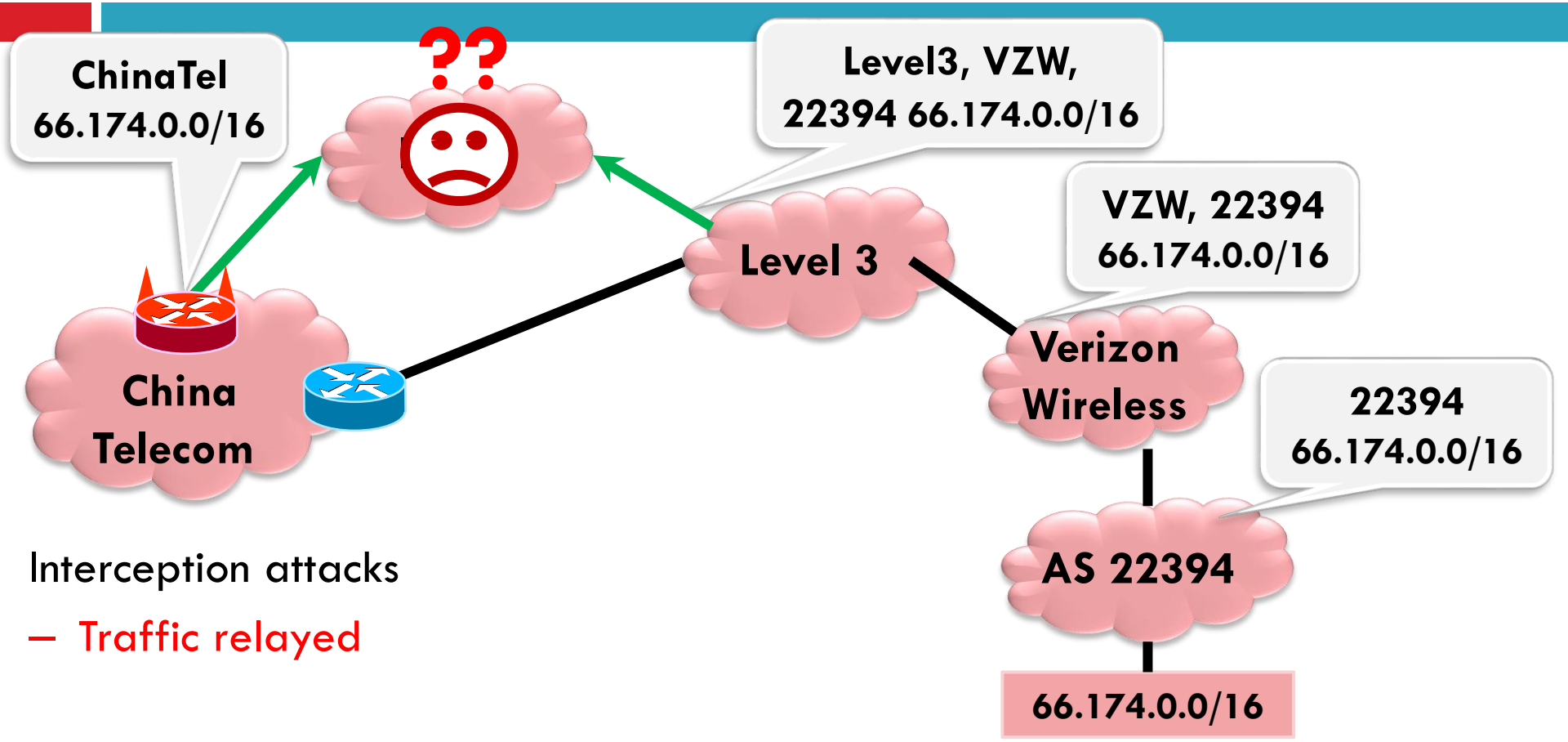
BGP-related Hijacks



Interception attacks

– Traffic relayed

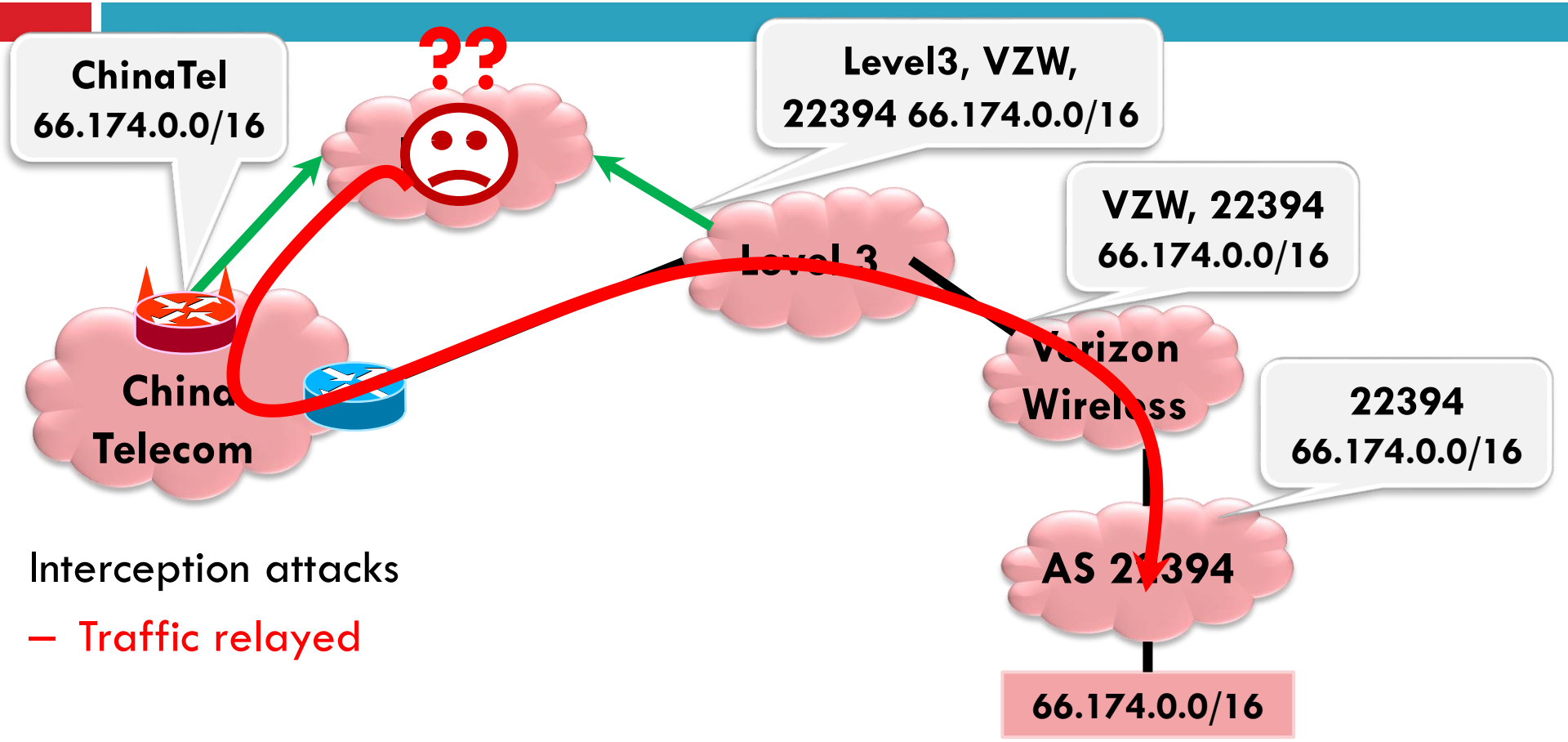
BGP-related Hijacks



Interception attacks

– Traffic relayed

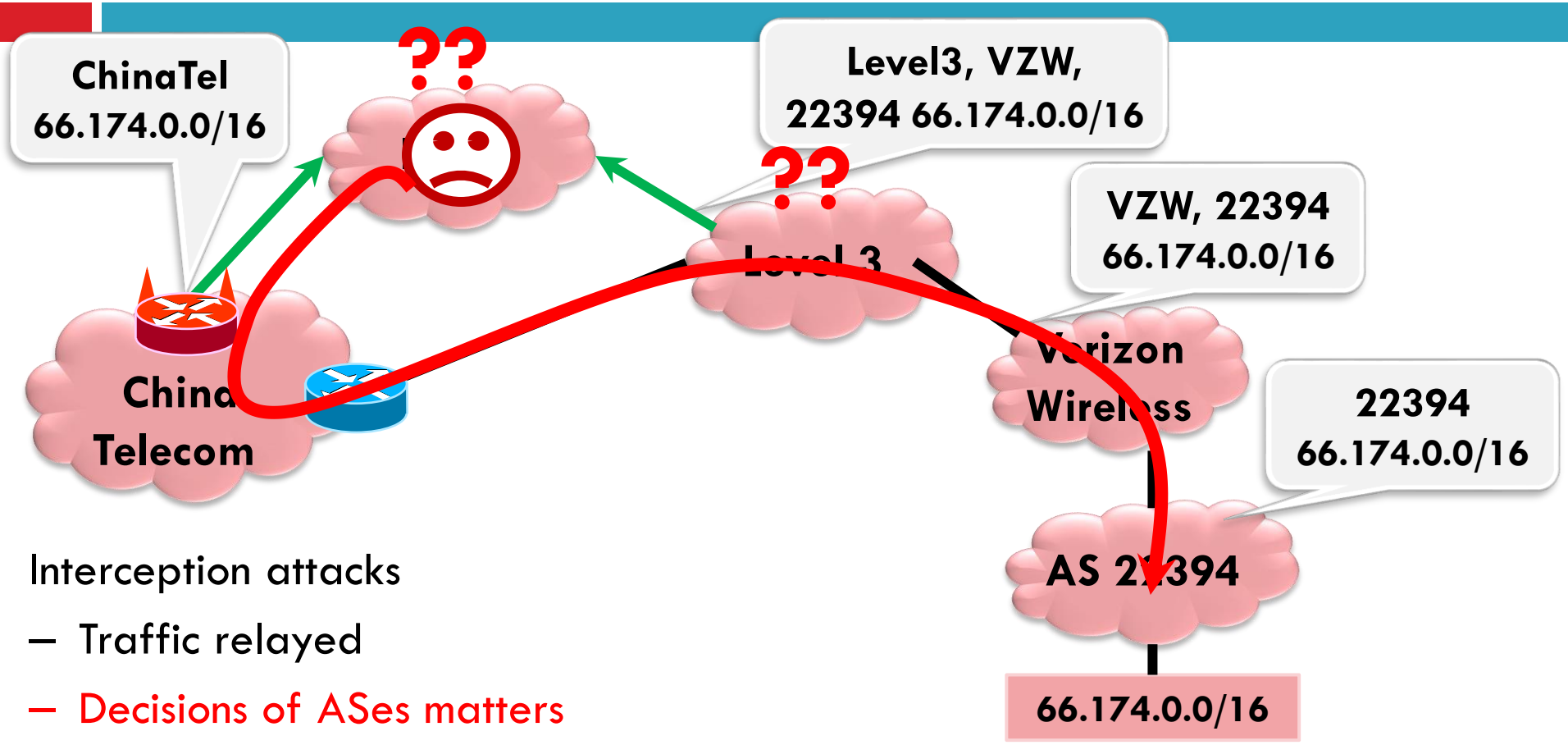
BGP-related Hijacks



Interception attacks

– Traffic relayed

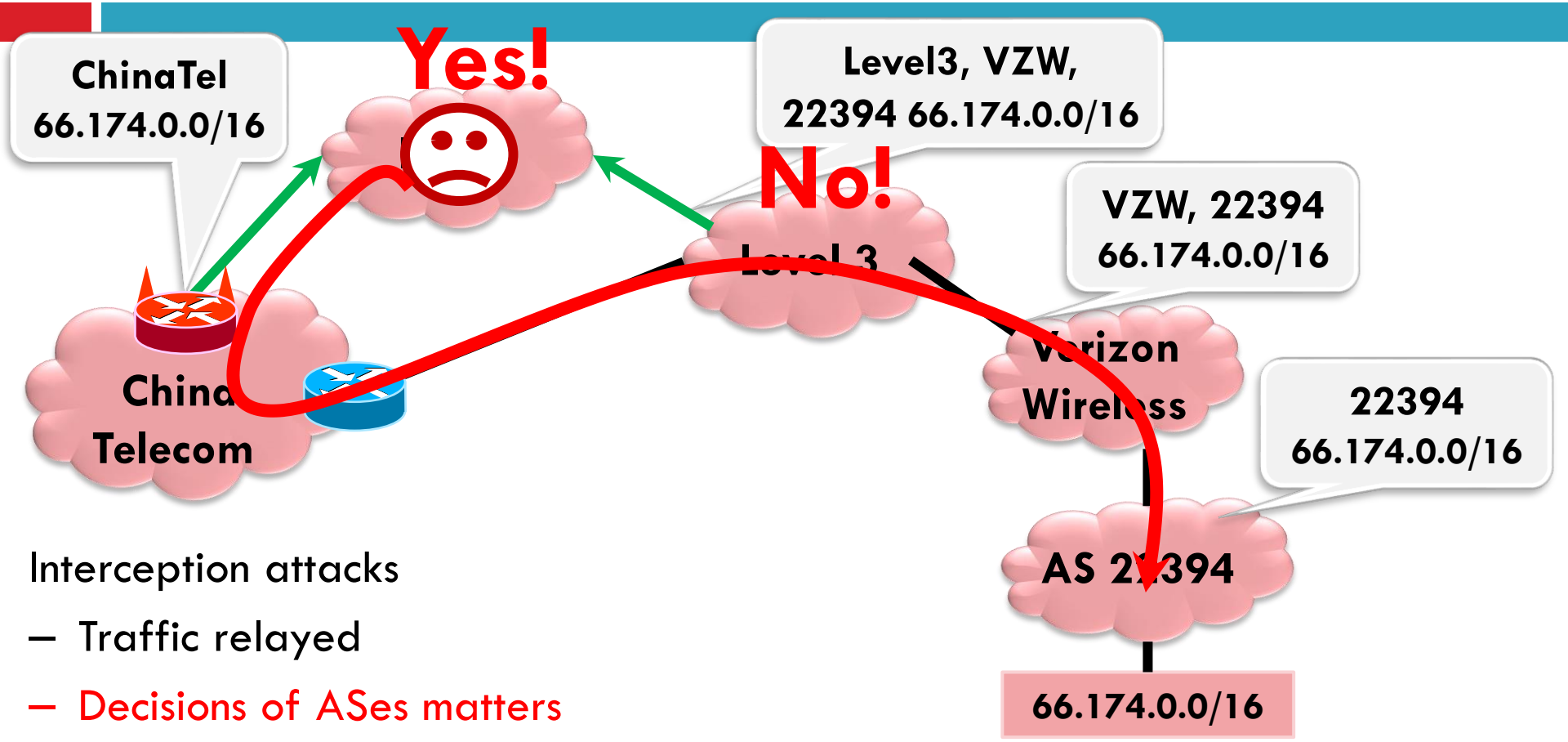
BGP-related Hijacks



Interception attacks

- Traffic relayed
- Decisions of ASes matters
  - E.g., selection of ChinaTel path
- Collaboration important

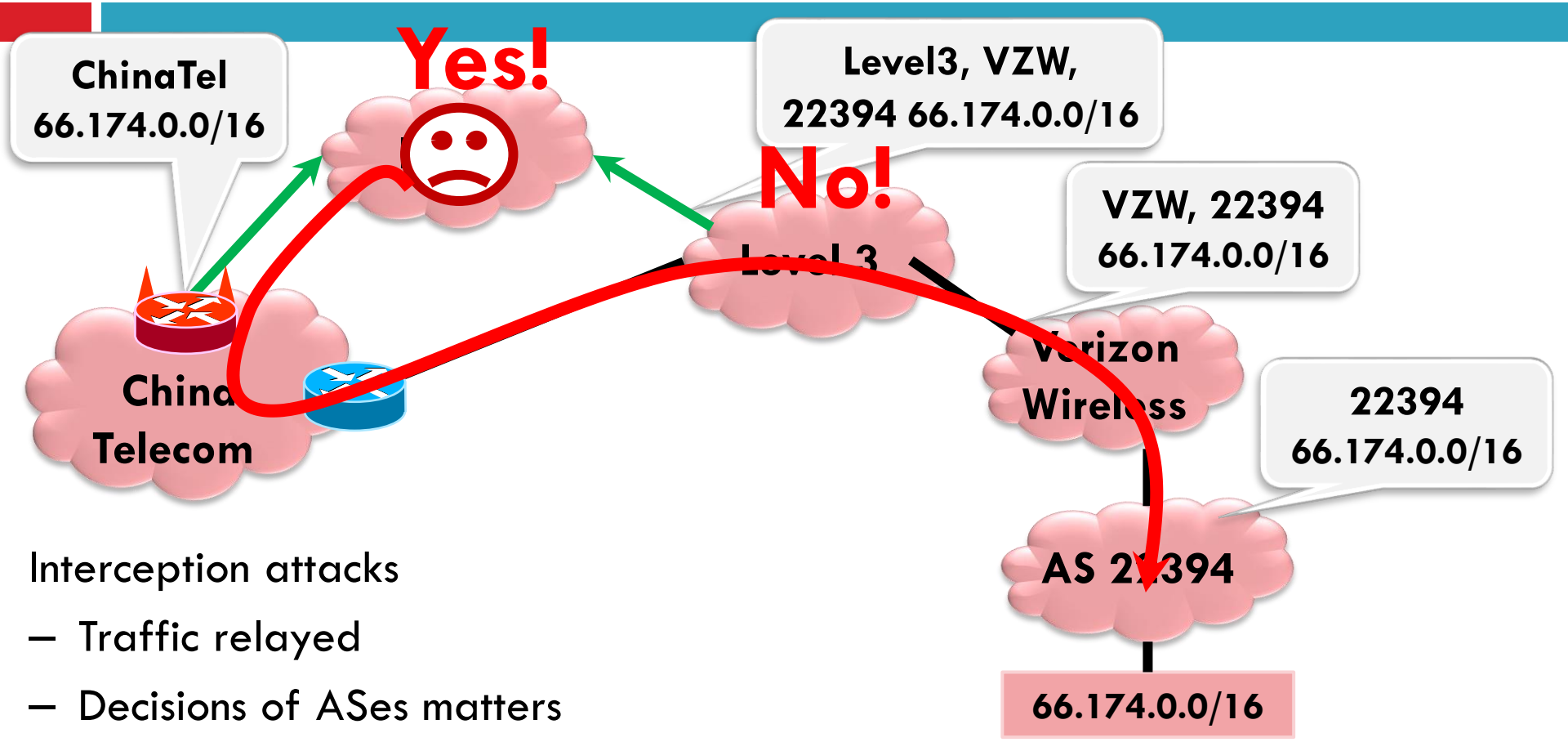
BGP-related Hijacks



Interception attacks

- Traffic relayed
- Decisions of ASes matters
  - E.g., selection of ChinaTel path
- Collaboration important

BGP-related Hijacks



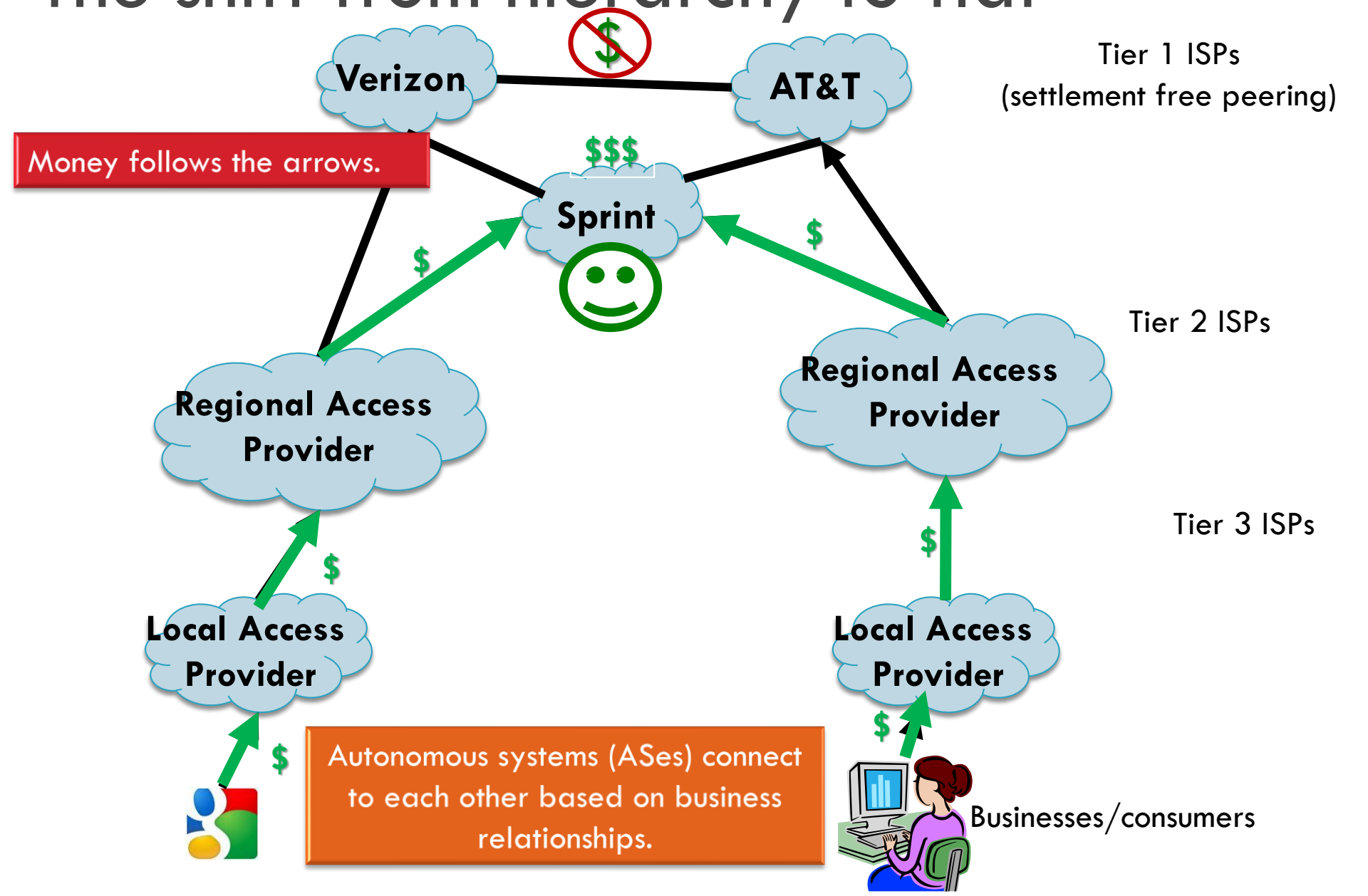
Interception attacks

- Traffic relayed
- Decisions of ASes matters
  - E.g., selection of ChinaTel path
- Collaboration important

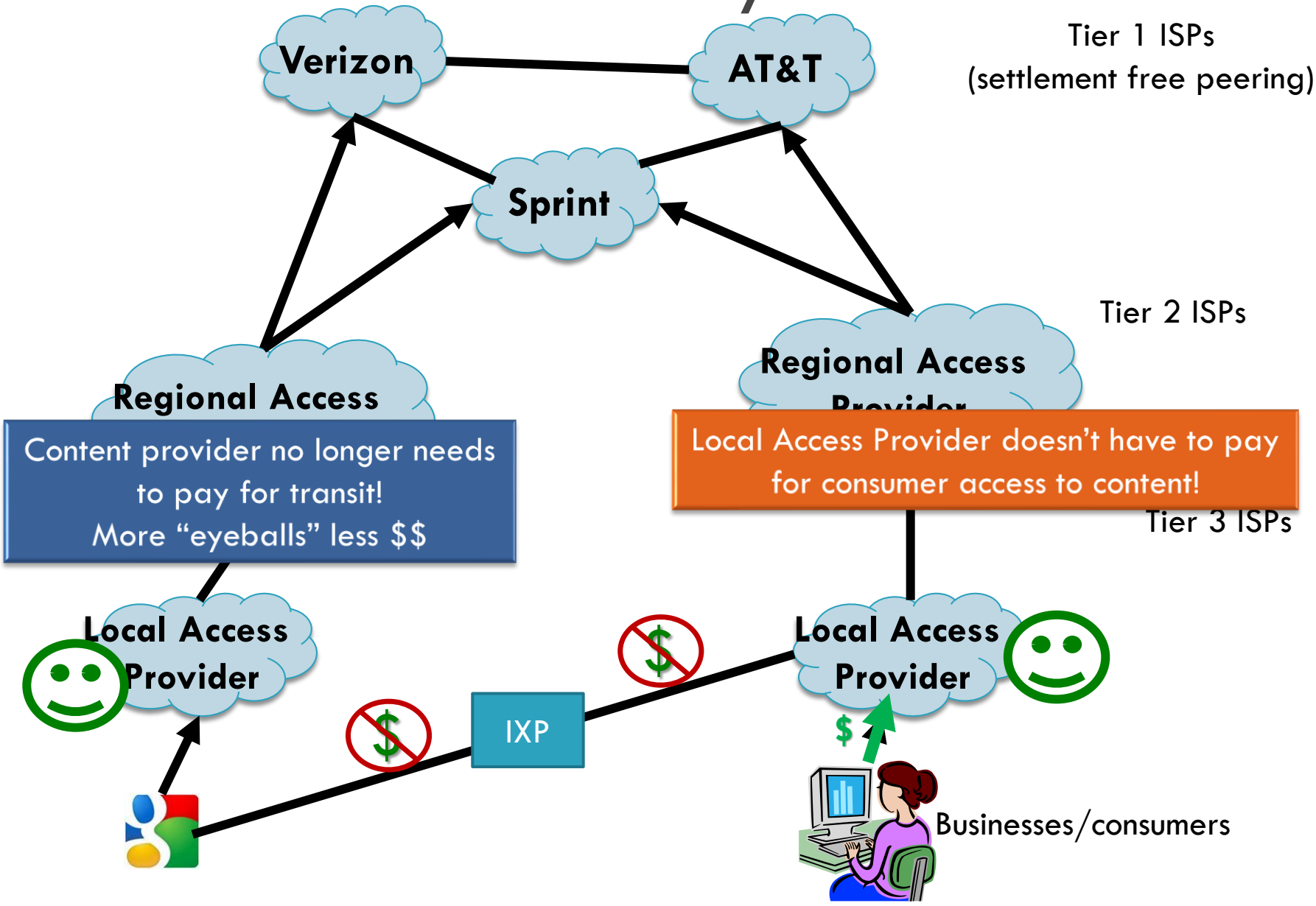
Example attacks

- "Characterizing Large-scale Routing Anomalies: A Case Study of the China Telecom Incident", Hiran et al., Proc. PAM 2013

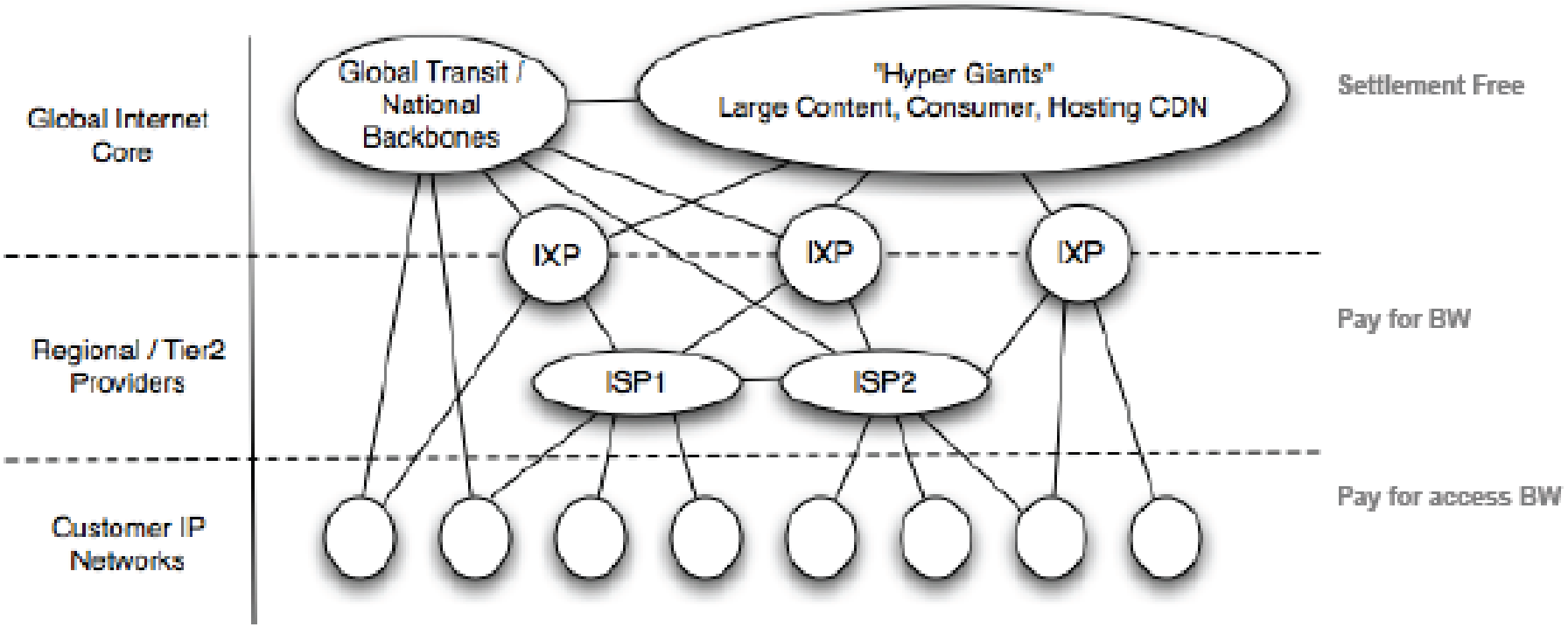
The shift from hierarchy to flat



The shift from hierarchy to flat



A new Internet model



- Flatter and much more densely interconnected Internet
- Disintermediation between content and "eyeball" networks
- New commercial models between content, consumer and transit

How do ASes connect?

- Point of Presence (PoP)
  - Usually a room or a building (windowless)
  - One router from one AS is physically connected to the other
  - Often in big cities
  - Establishing a new connection at PoPs can be expensive
- Internet eXchange Points
  - Facilities dedicated to providing presence and connectivity for large numbers of ASes
  - Many fewer IXPs than PoPs
  - Economies of scale

IXPs Definition

- Industry definition (according to Euro-IX)

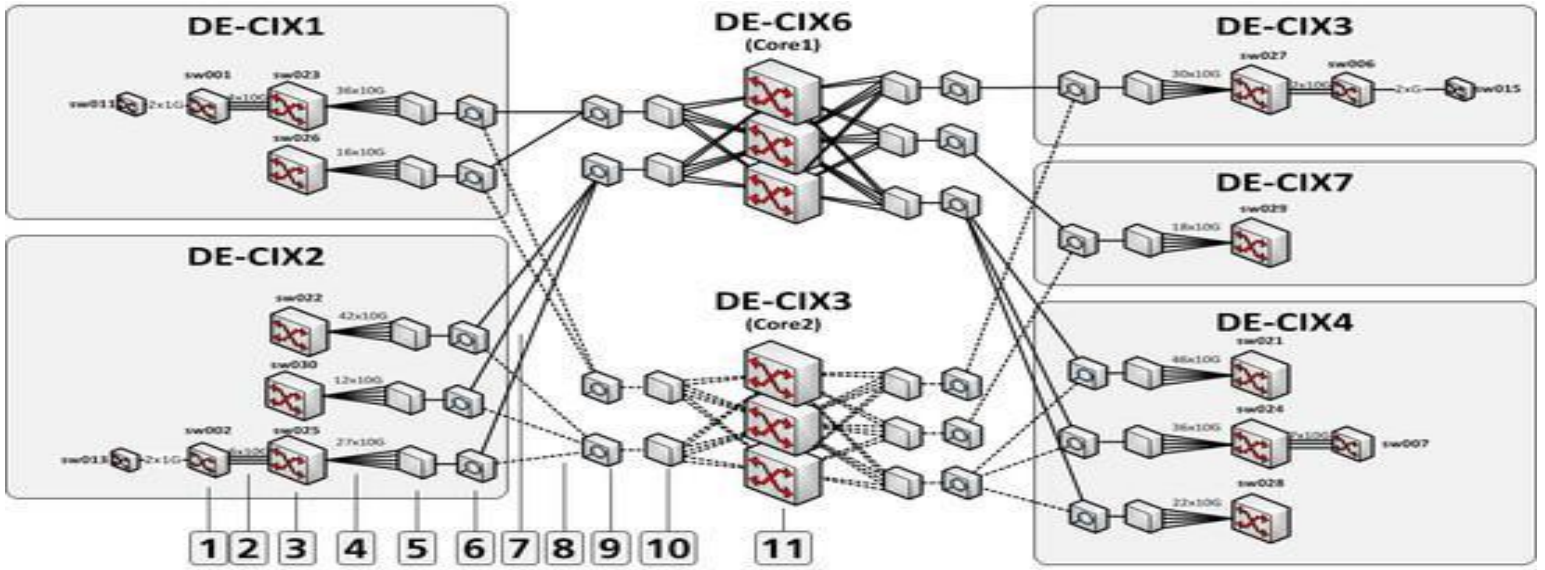
A physical network infrastructure operated by a single entity with the purpose to **facilitate** the **exchange** of Internet traffic between **Autonomous Systems**

The number of Autonomous Systems connected should be at least three and there **must** be a **clear** and **open policy** for others to **join**.

https://www.euro-ix.net/what-is-an-ixp

Internet eXchange Points

Inside an IXP



1 Force10 Terascale E1200
2 Multiple 10G-Connections
3 Force10 Exascale E1200i
4 Multiple 10G-Connections
5 DWDM MUX 32 Channel
6 Lynx LightLeader Master Unit
7 Dark Fiber Working Line
8 Dark Fiber Protection Line
9 Lynx LightLeader Slave Unit
10 DWDM MUX 32 Channel
11 2xBrocade MLX32 and 1xForce10 Exascale 1200i per Core

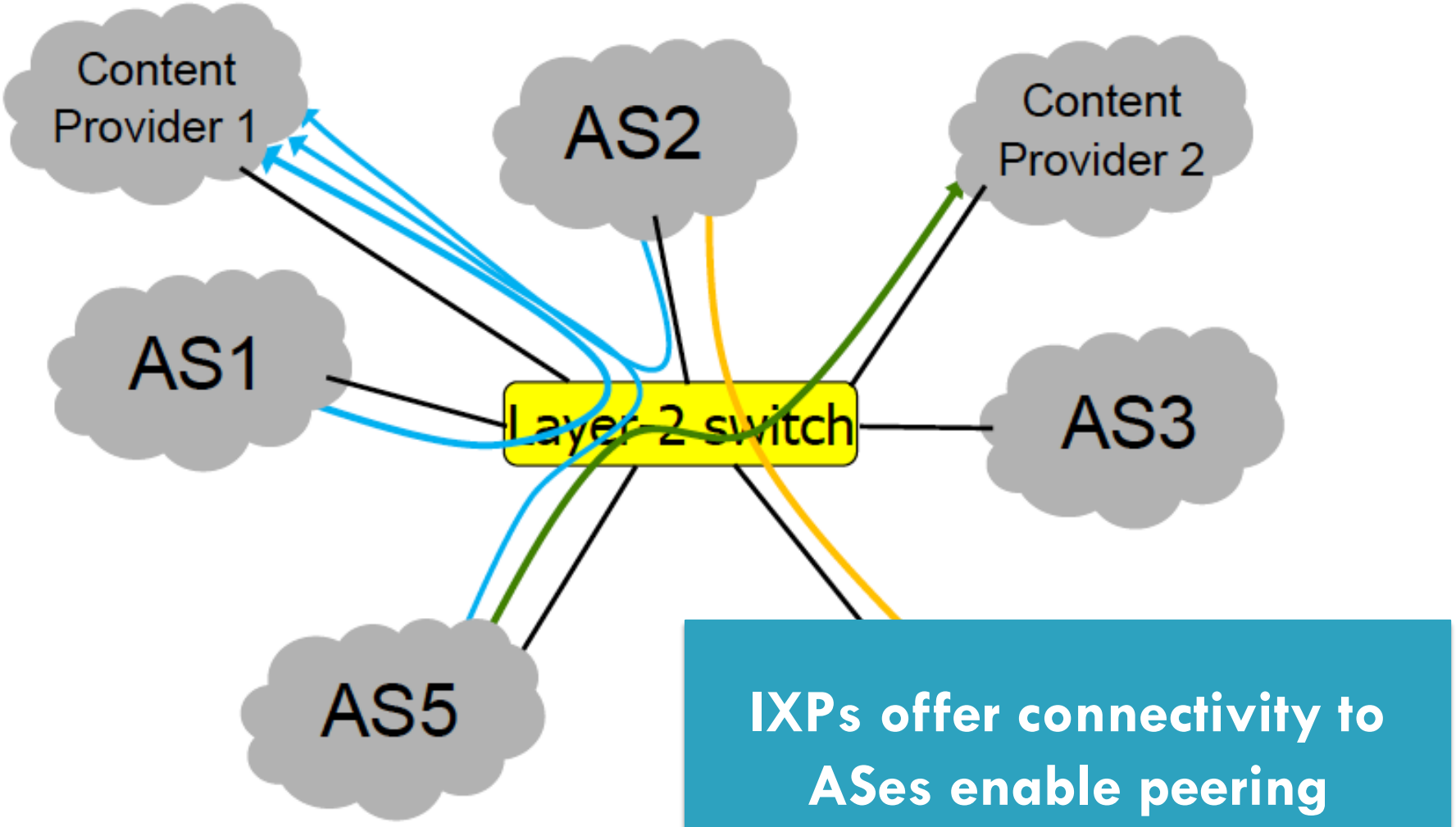
Robust infrastructure with redundency

http://www.de-cix.net/about/topology/

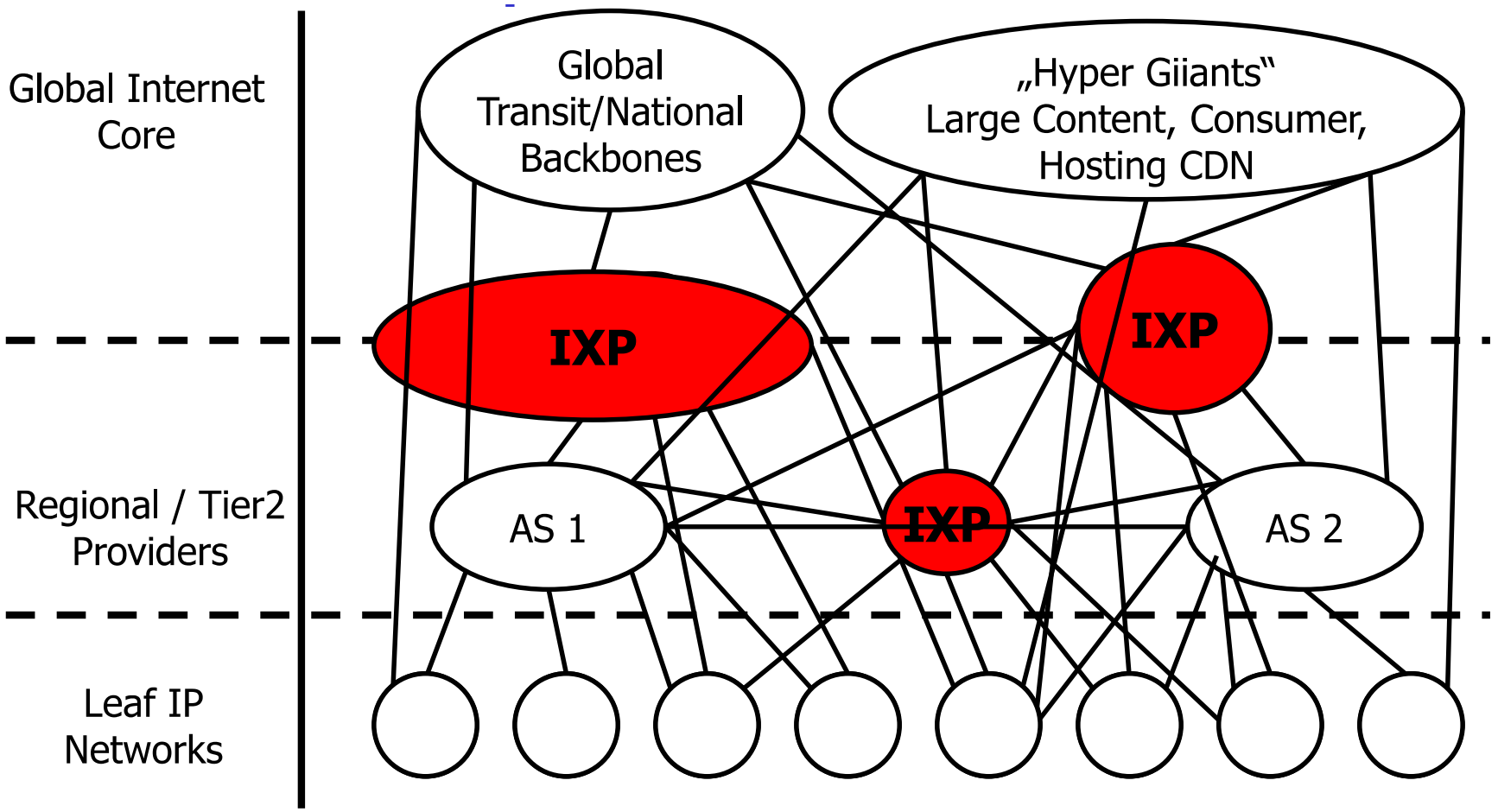
IXPs worldwide

https://prefix.pch.net/applications/ixpdir/

Structure


Content Provider 1
AS2
Content Provider 2
AS1
Layer-2 switch
AS3
AS5
IXPs offer connectivity to ASes enable peering

Revised model 2012+



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

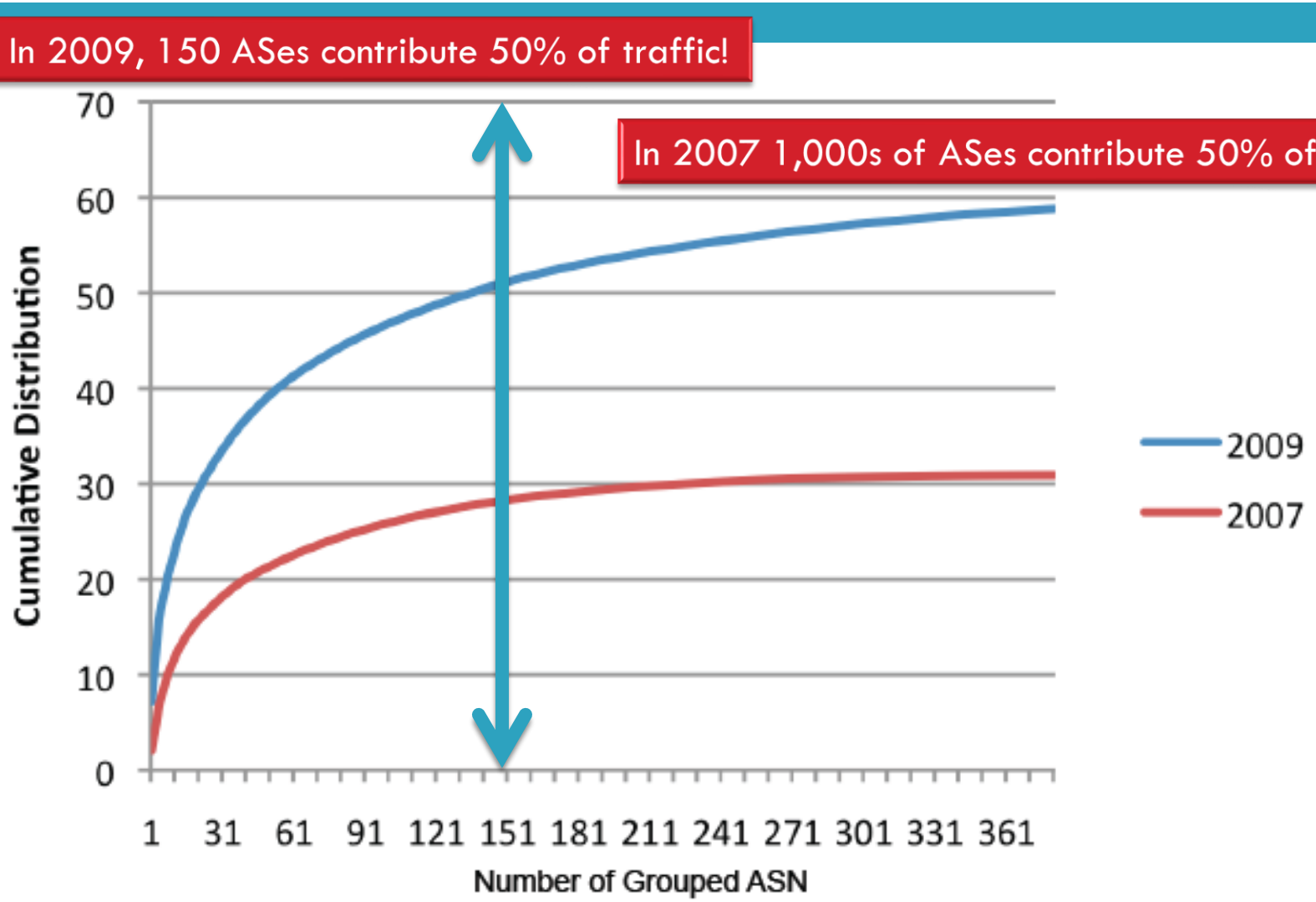
Why are these still issues?

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

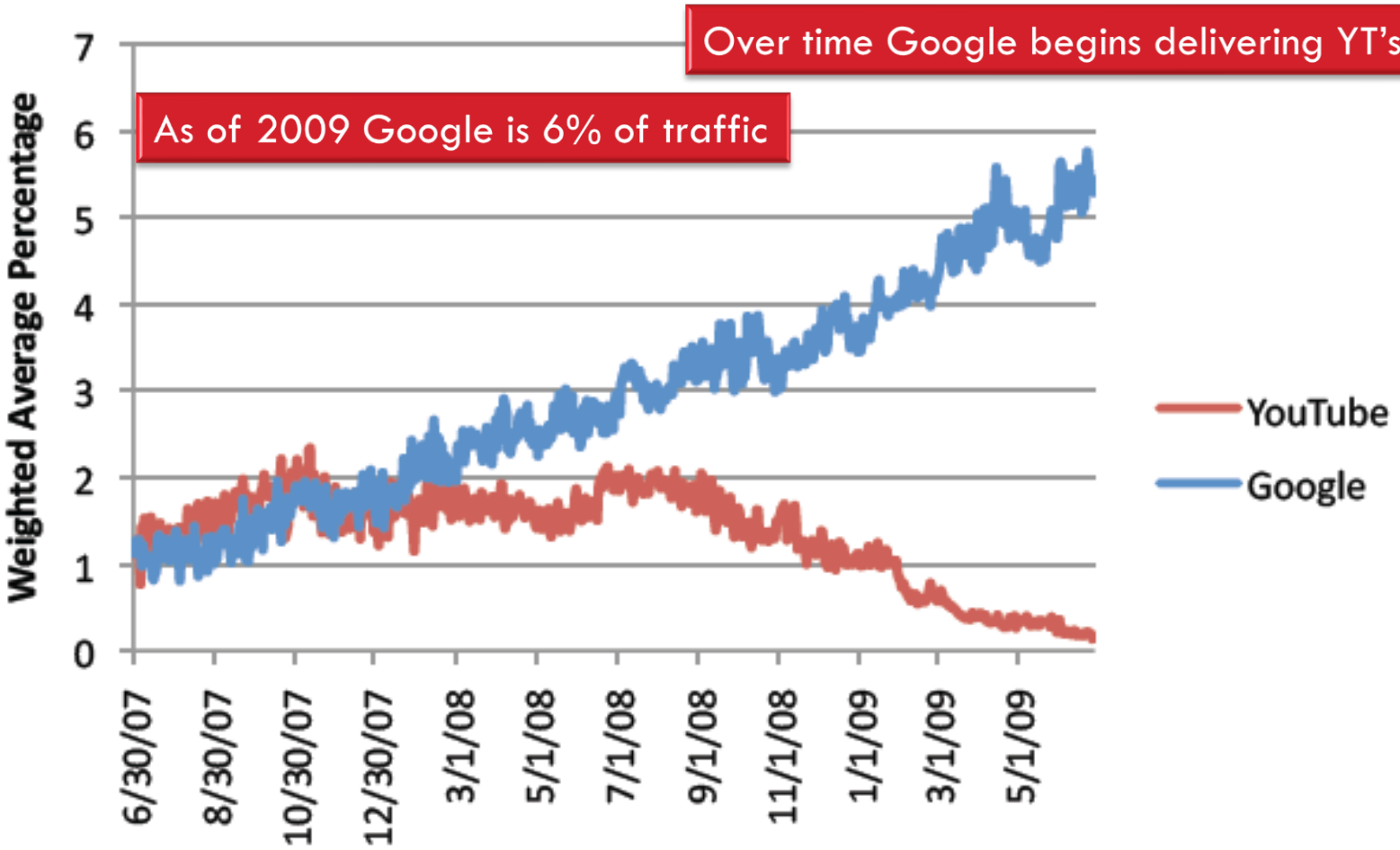
Very similar issues to IPv6 deployment

More slides …

Consolidation of Content



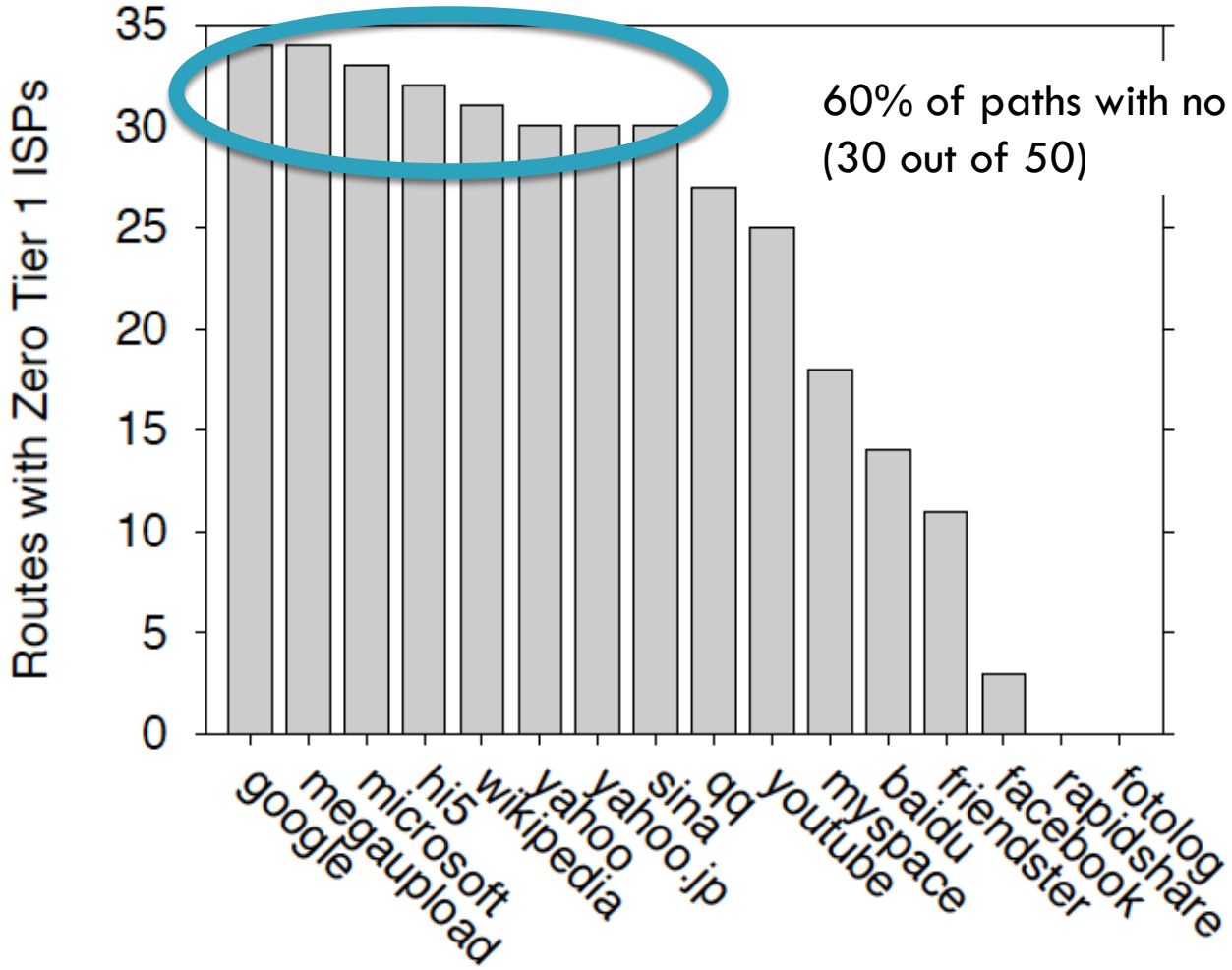
Case Study: Google



Graph of weighted averaged grouped ASNs

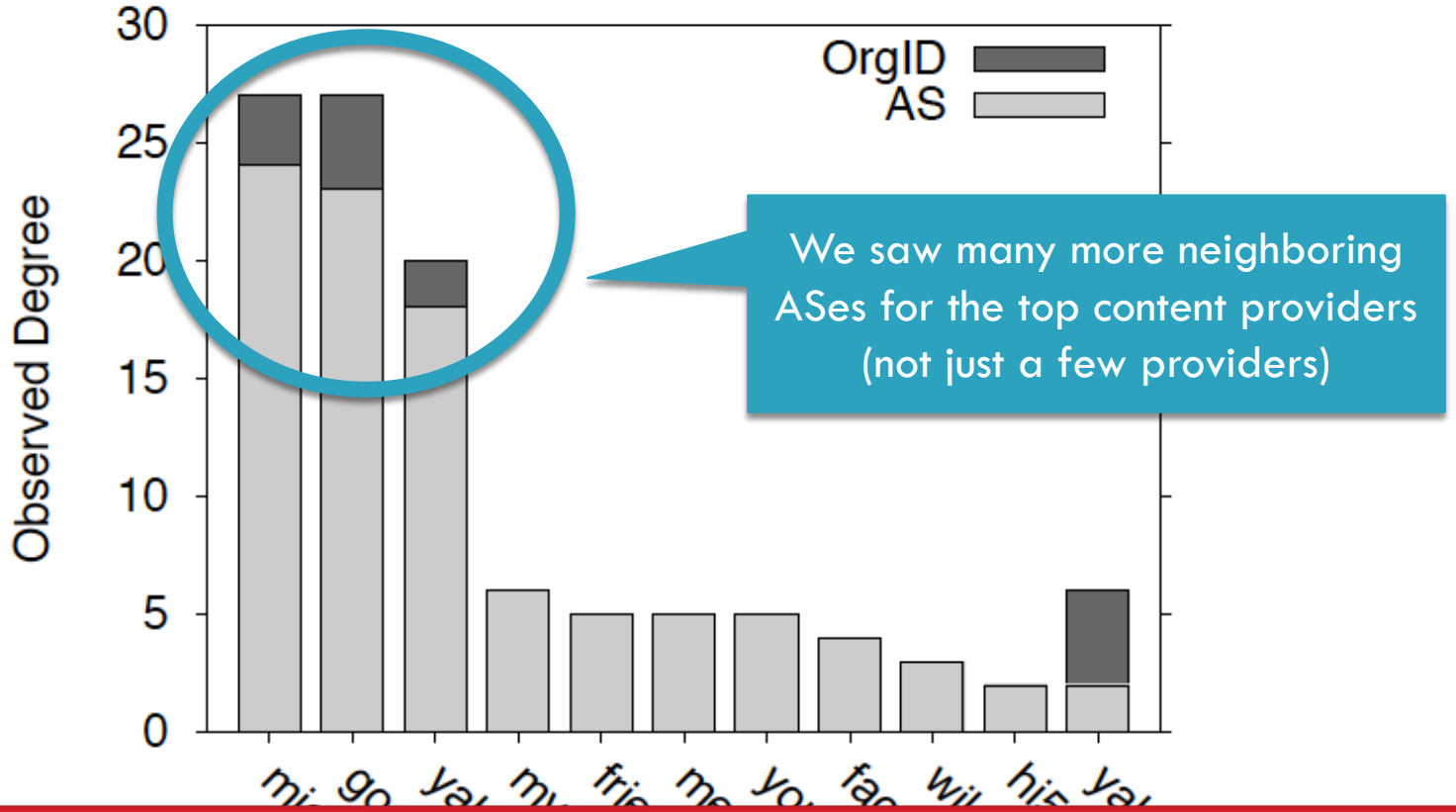
Flattening: Paths with no Tier 1s

The Flattening Internet Topology: Natural Evolution, Unsightly Barnacles or Contrived Collapse?, Proc. PAM 2008



Relative degree of top content providers

The Flattening Internet Topology: Natural Evolution, Unsightly Barnacles or Contrived Collapse?, Proc. PAM 2008



These numbers are actually way lower than the true degree of these ASes

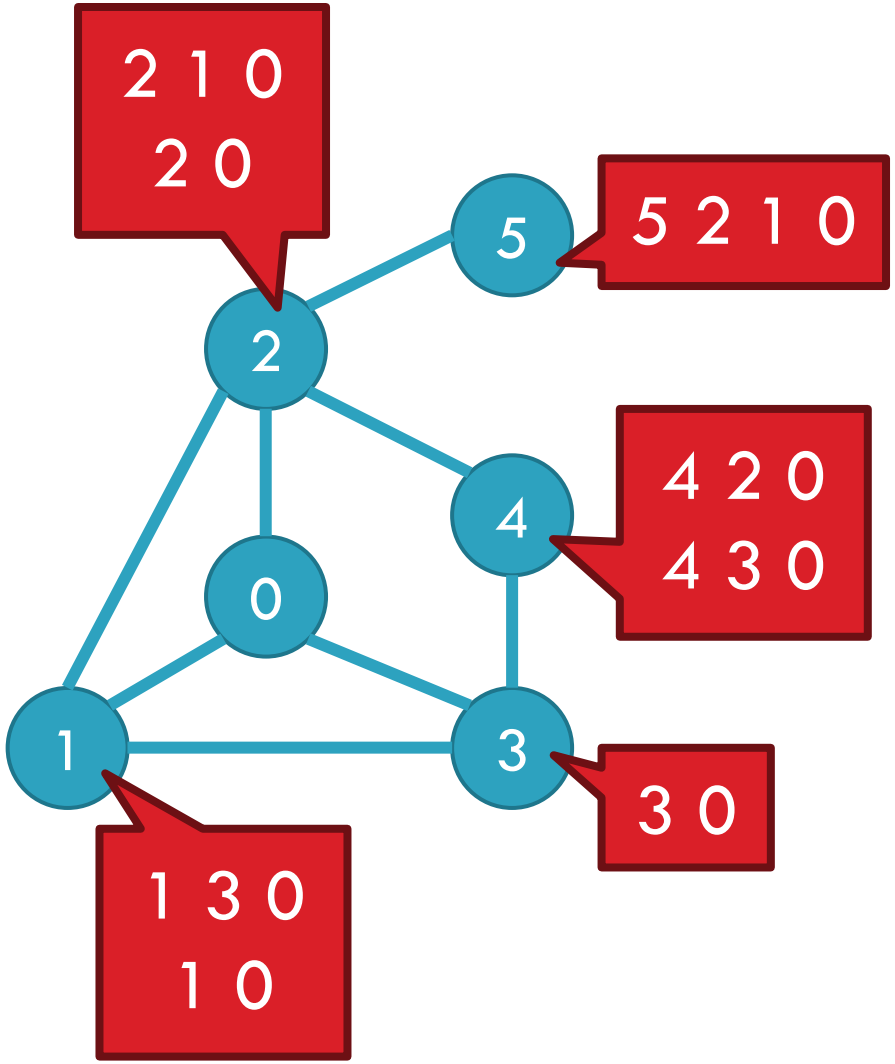
What Problem is BGP Solving?

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
  - A set of permitted paths from each node to the origin
  - Each set of paths is ranked

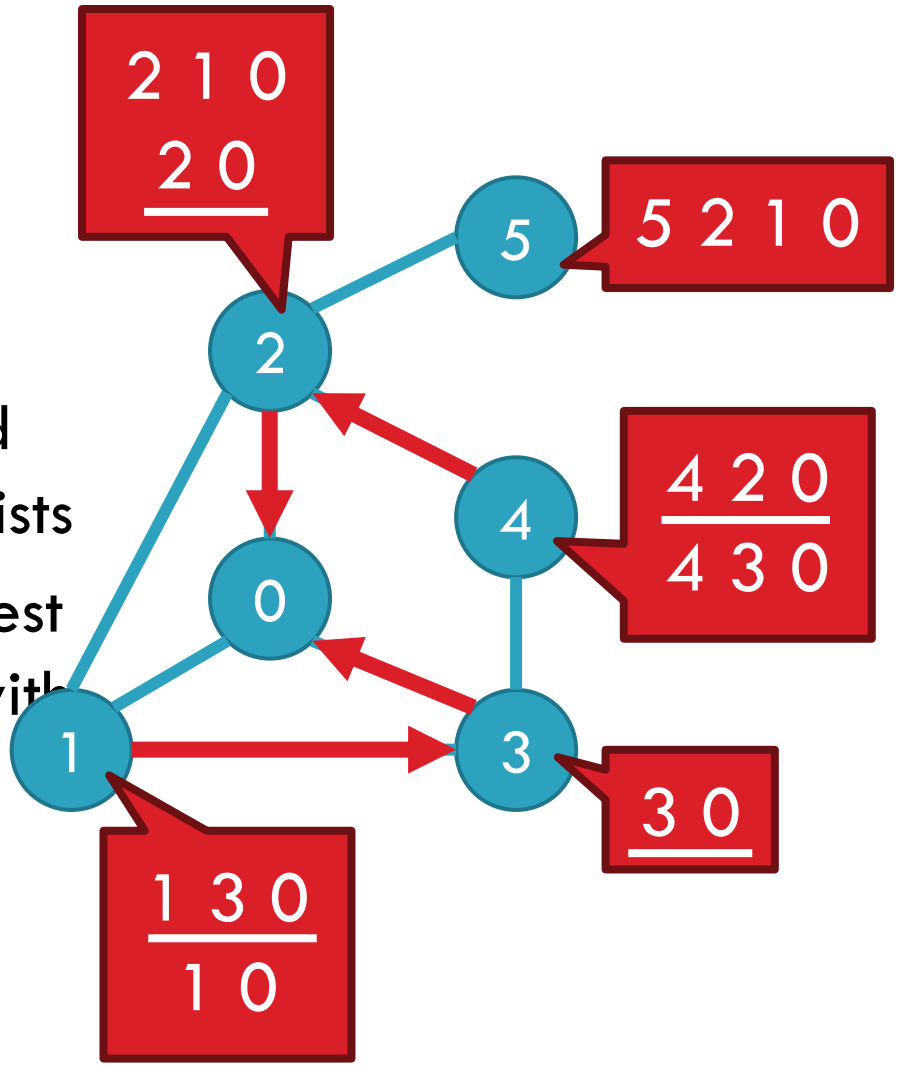


A Solution to the SPP

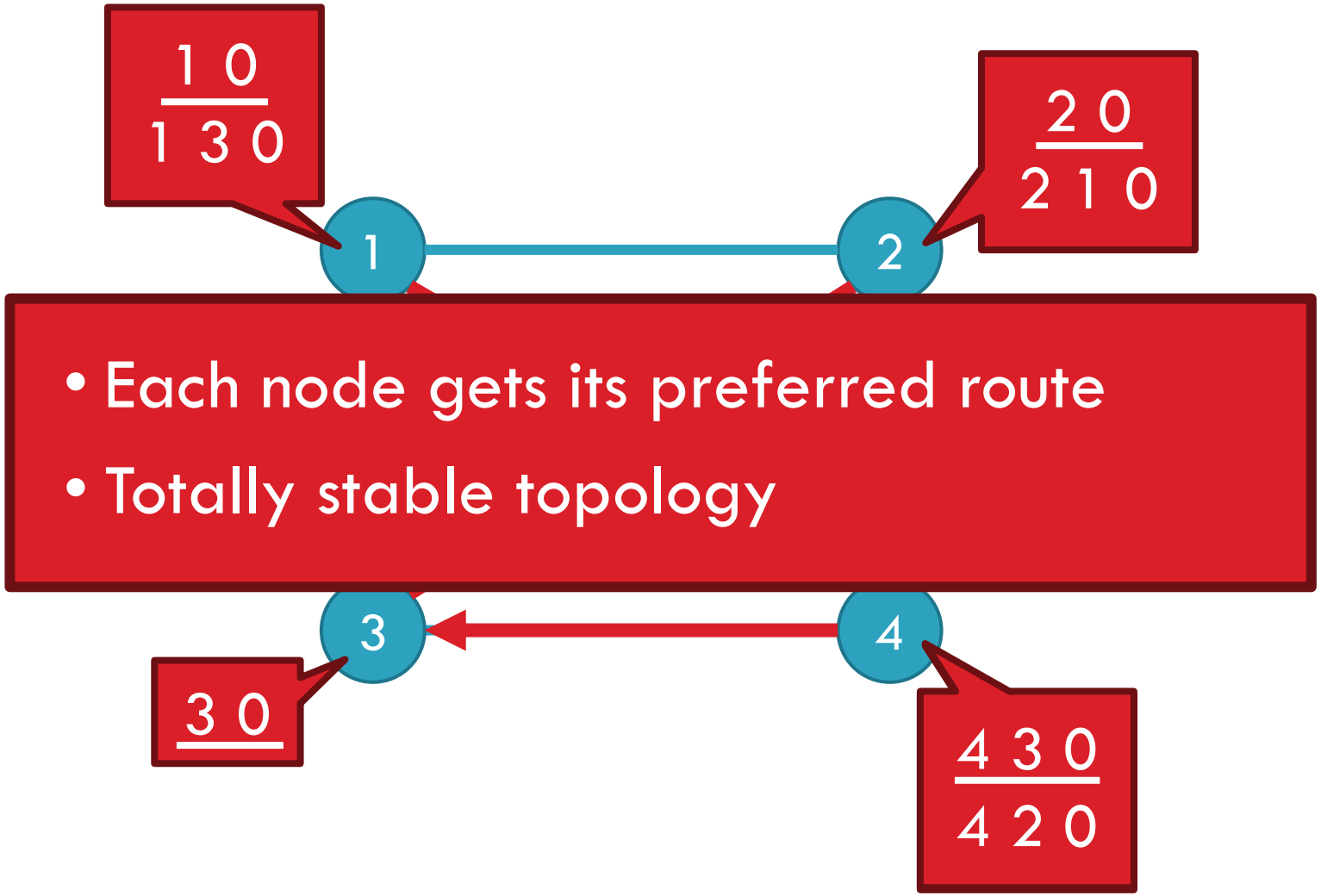
- A solution is an assignment of permitted paths to each node

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

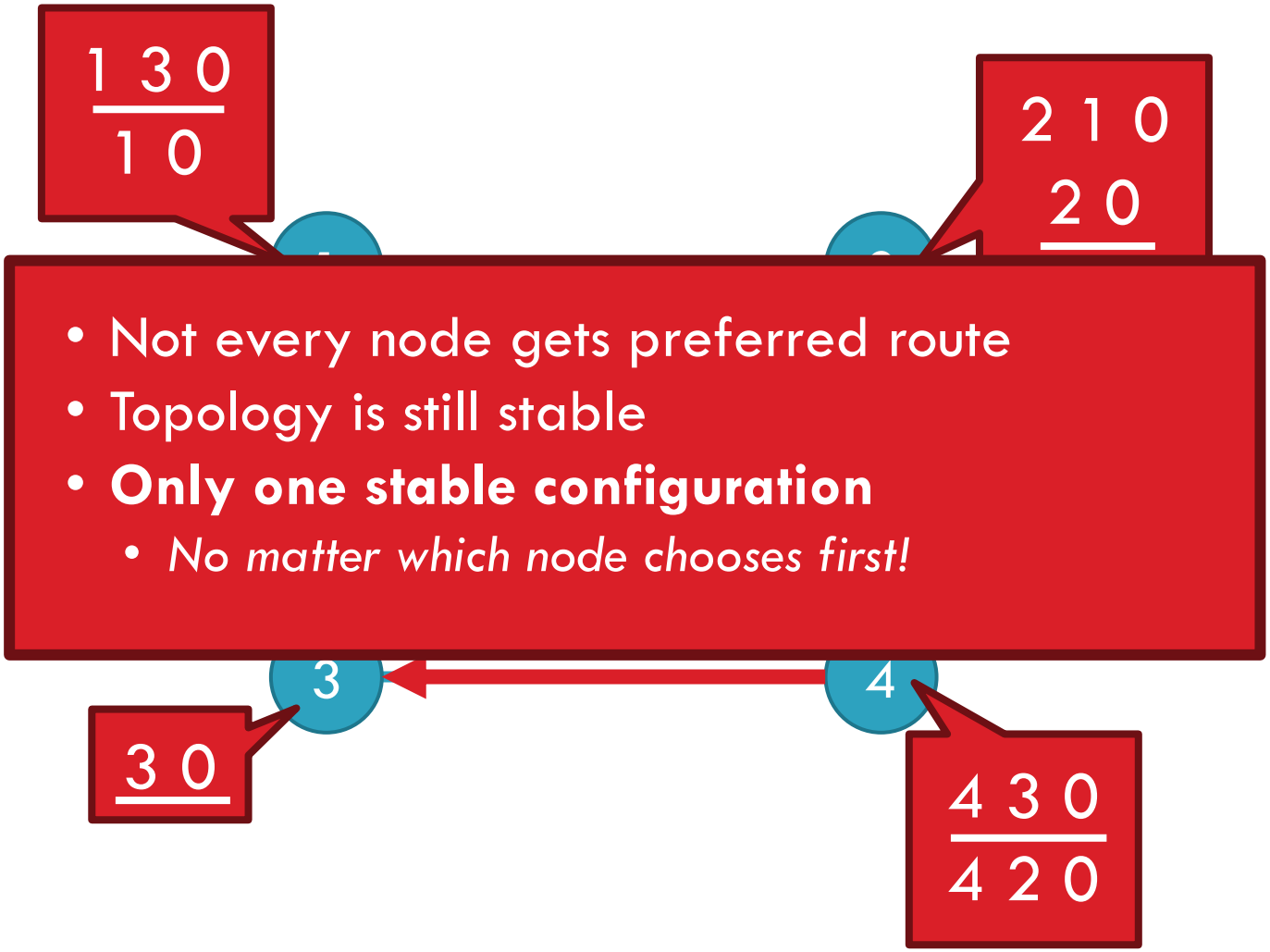
their neighbors



Simple SPP Example

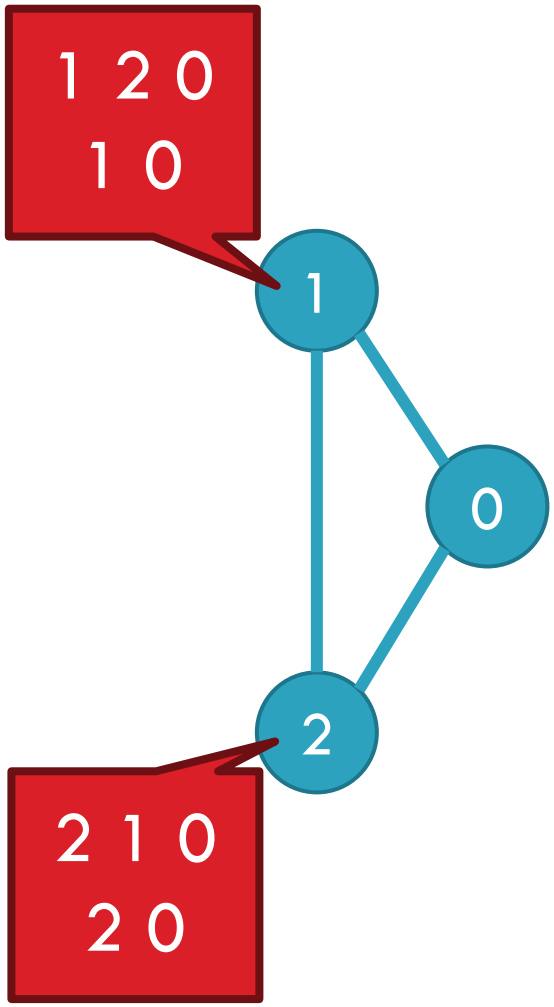


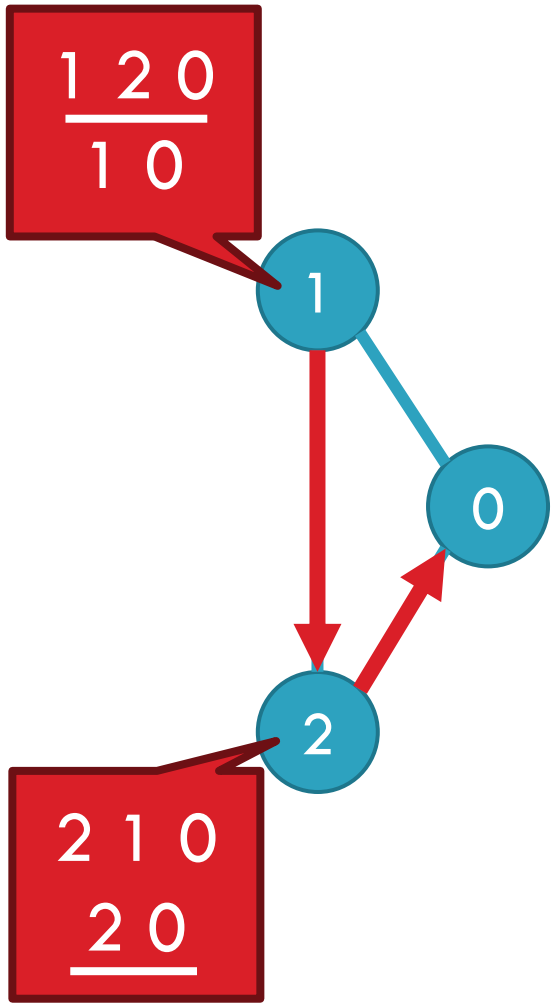
Good Gadget

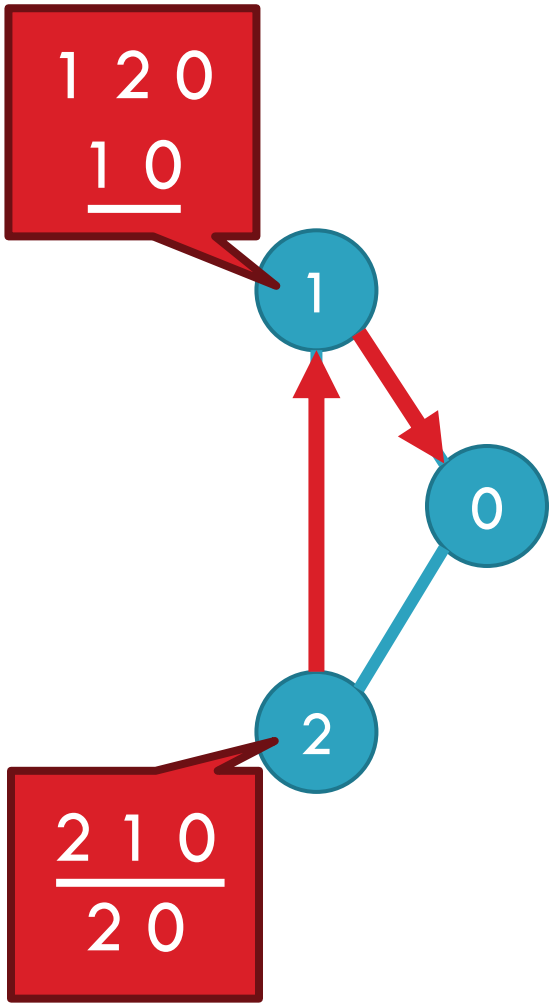


- Not every node gets preferred route
- Topology is still stable
- **Only one stable configuration**
  - *No matter which node chooses first!*

SPP May Have Multiple Solutions







Bad Gadget

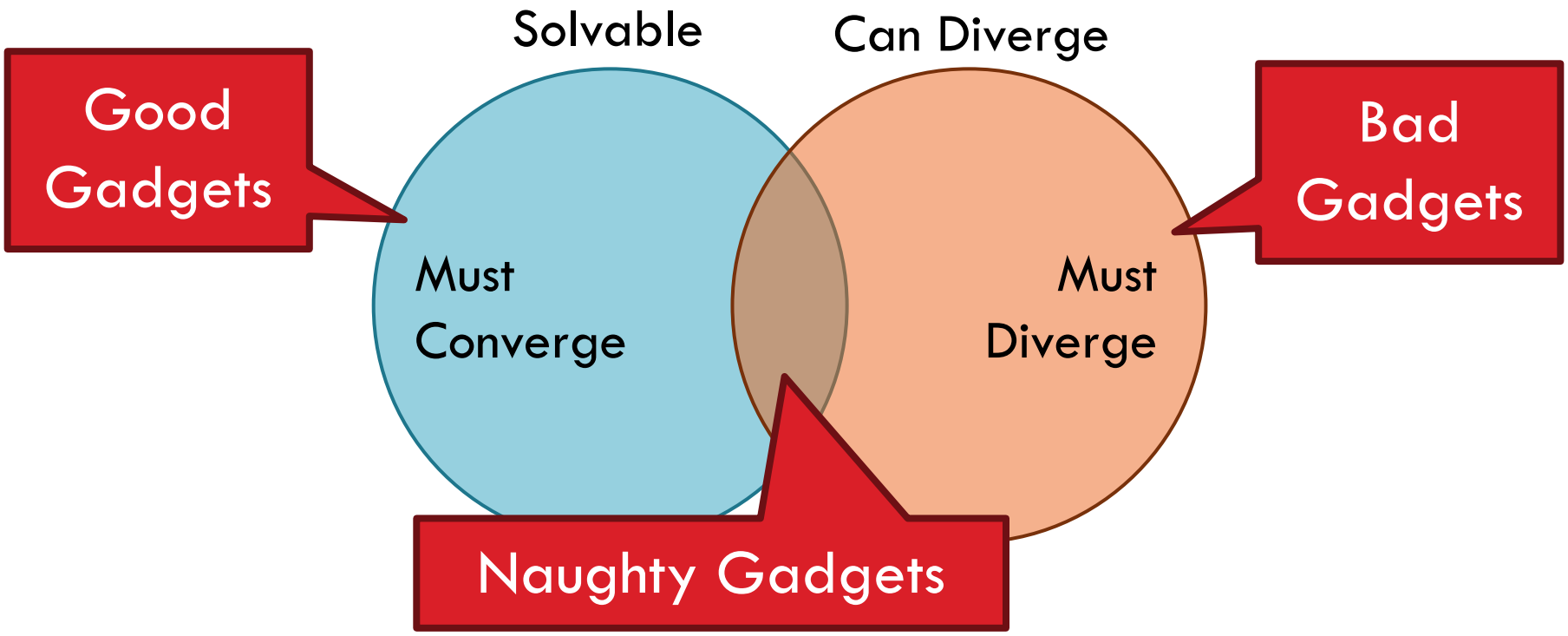
1 3 0

- 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

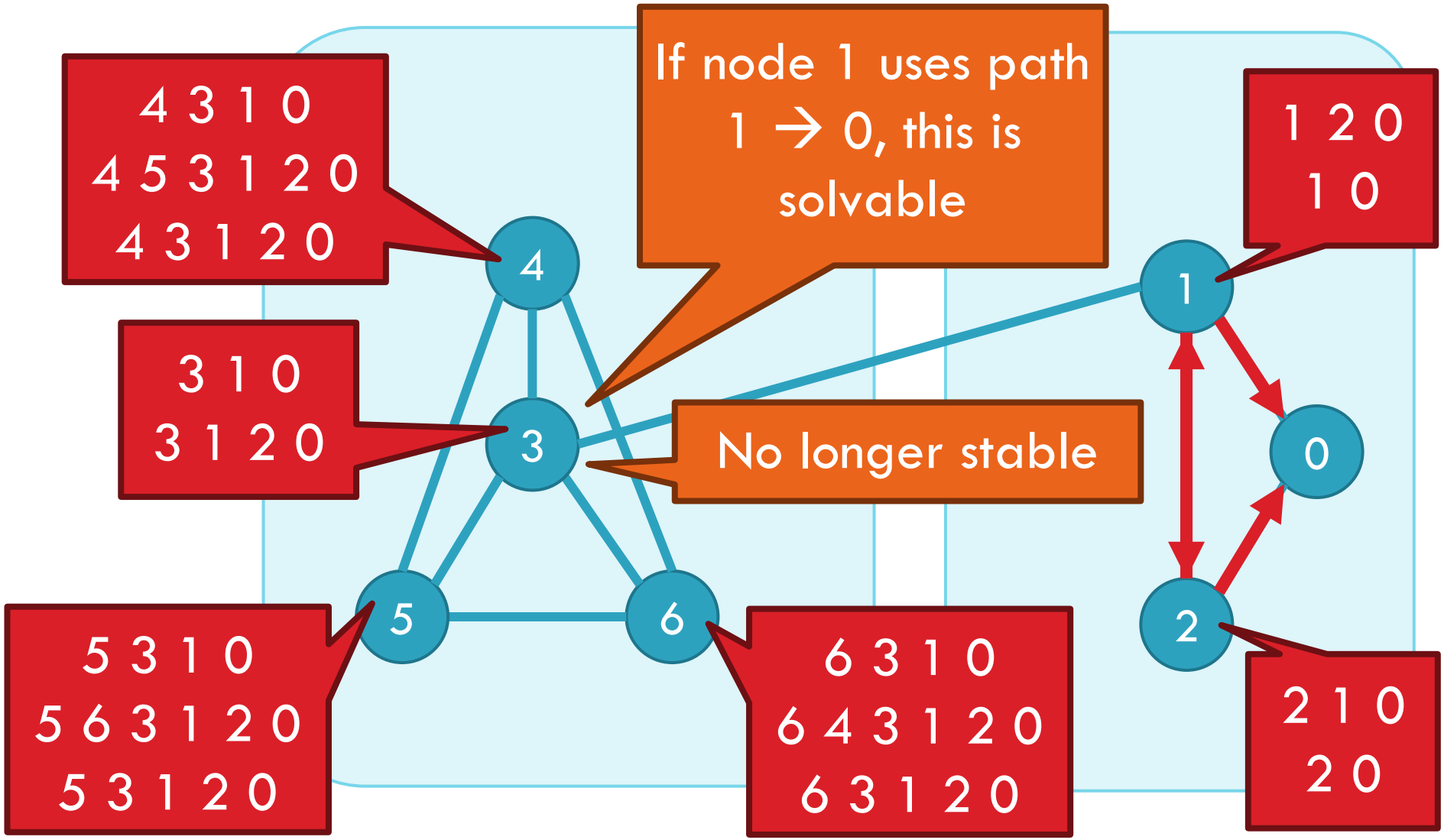
4 3 0

SPP Explains BGP Divergence

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

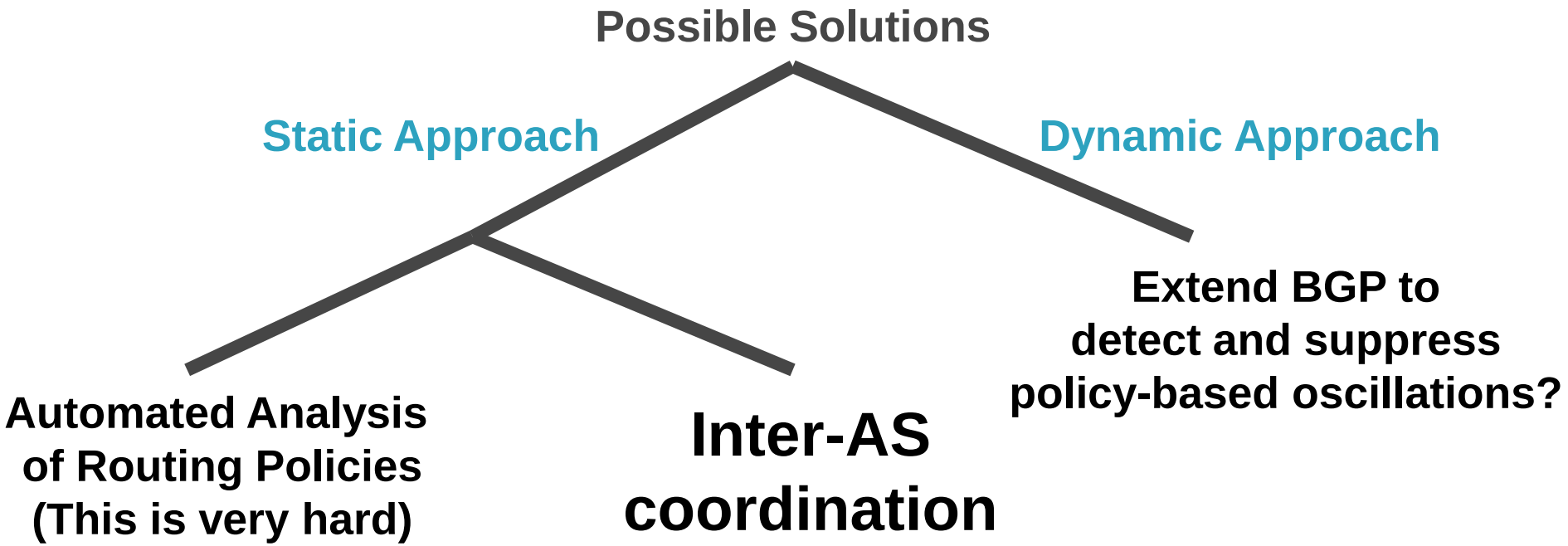


BGP is Precarious

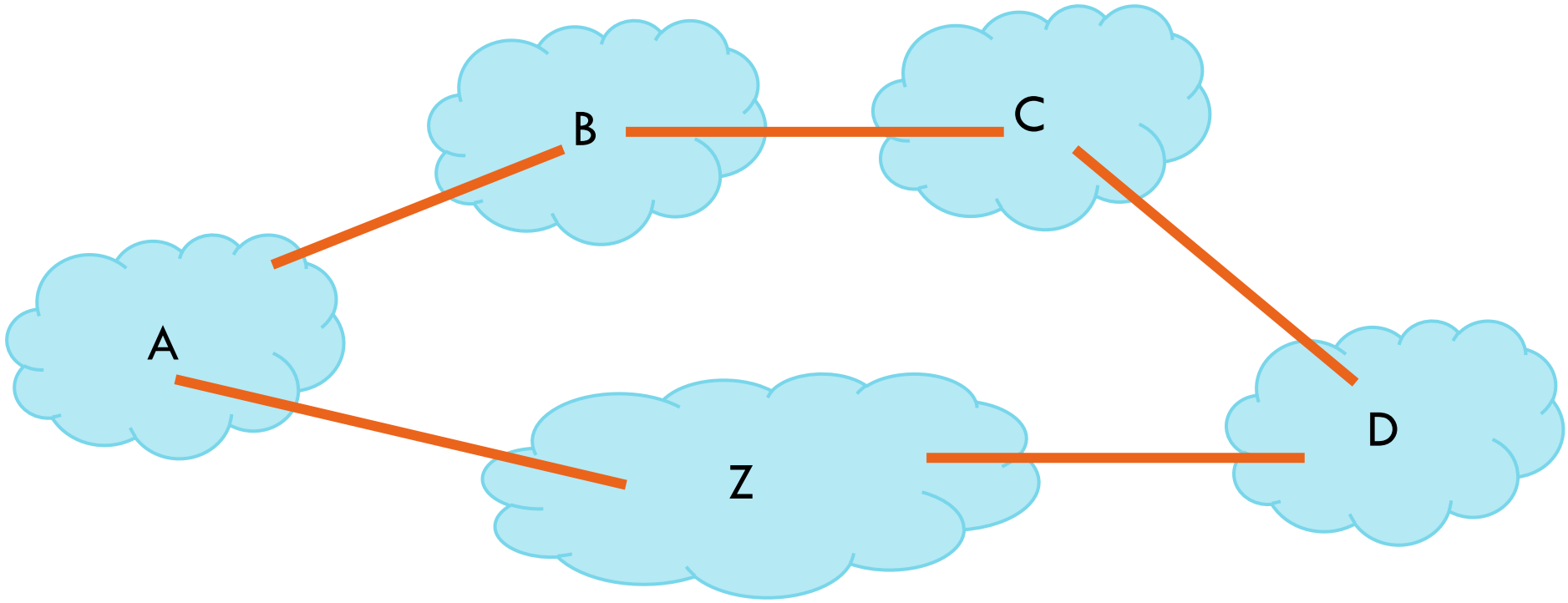


Can BGP Be Fixed?

- Unfortunately, SPP is NP-complete



These approaches are complementary


B
C
A
D
Z