15-441/641: Computer Networks
BGP – Inter-domain Routing

15-441 Spring 2019
Profs Peter Steenkiste & **Justine Sherry**

Fall 2019
https://www.myheartisinthenetwork.com

Carnegie Mellon University

I’ve missed you!
What have you learned while I’ve been away?

Chat with a friend…

- What is the purpose of DHCP?
- What is the purpose of ARP?
- What are some benefits of DNS hierarchy?

Chat with a friend…

- What is the purpose of DHCP?
- What is the purpose of ARP?
- What are some benefits of DNS hierarchy?

Fun

Consider the following routing table:

Destination	Next Hop
192.1/16	1.2.3.4
192.1.0/23	1.2.3.5
192.1.4/24	1.2.3.6
192.1.1/24	1.2.3.7

Which next hop should the router use for a packet destined to 192.1.0.1?

- Routes:
  - 11000000.00000001.00000000.00000000
  - 11000000.00000001.00000000.00000000
  - 11000000.00000001.00000100.00000000
  - 11000000.00000001.00000001.00000000

- Packet:
  - 11000000.00000001.00000000.00000001

- Routes:
  - 11000000.00000001.00000000.00000000
  - 11000000.00000001.00000000.00000000
  -
  -

Pick the longer one

Don’t match

- Packet:
  - 11000000.00000001.00000000.00000001

EVEN MORE FUN

Pull out your laptop, if you have a Mac or Linux:
(Or if you have a Linux shell in Windows)

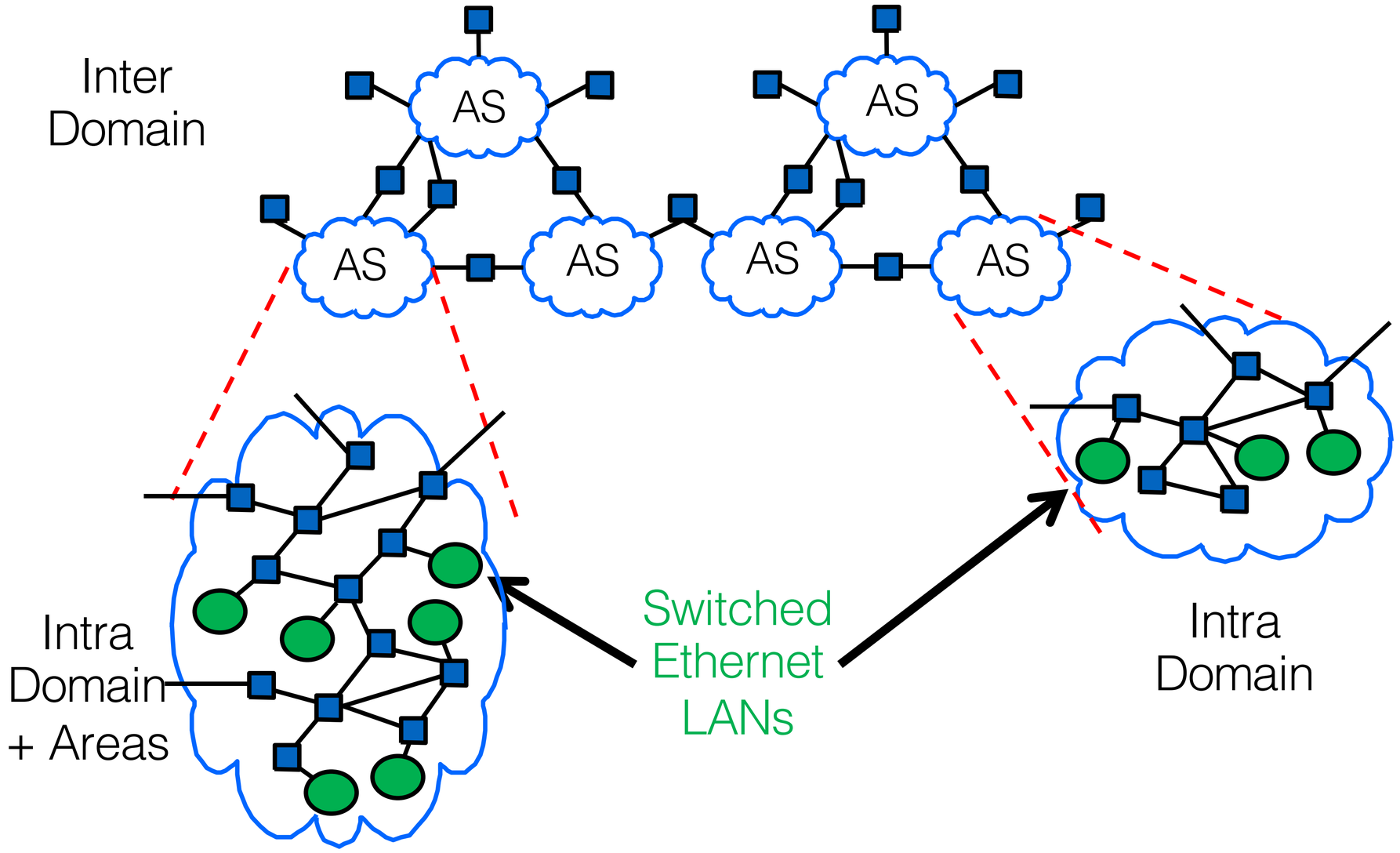
If you send a packet to facebook.com, what will the IP destination address be?
What will the Ethernet destination address be?

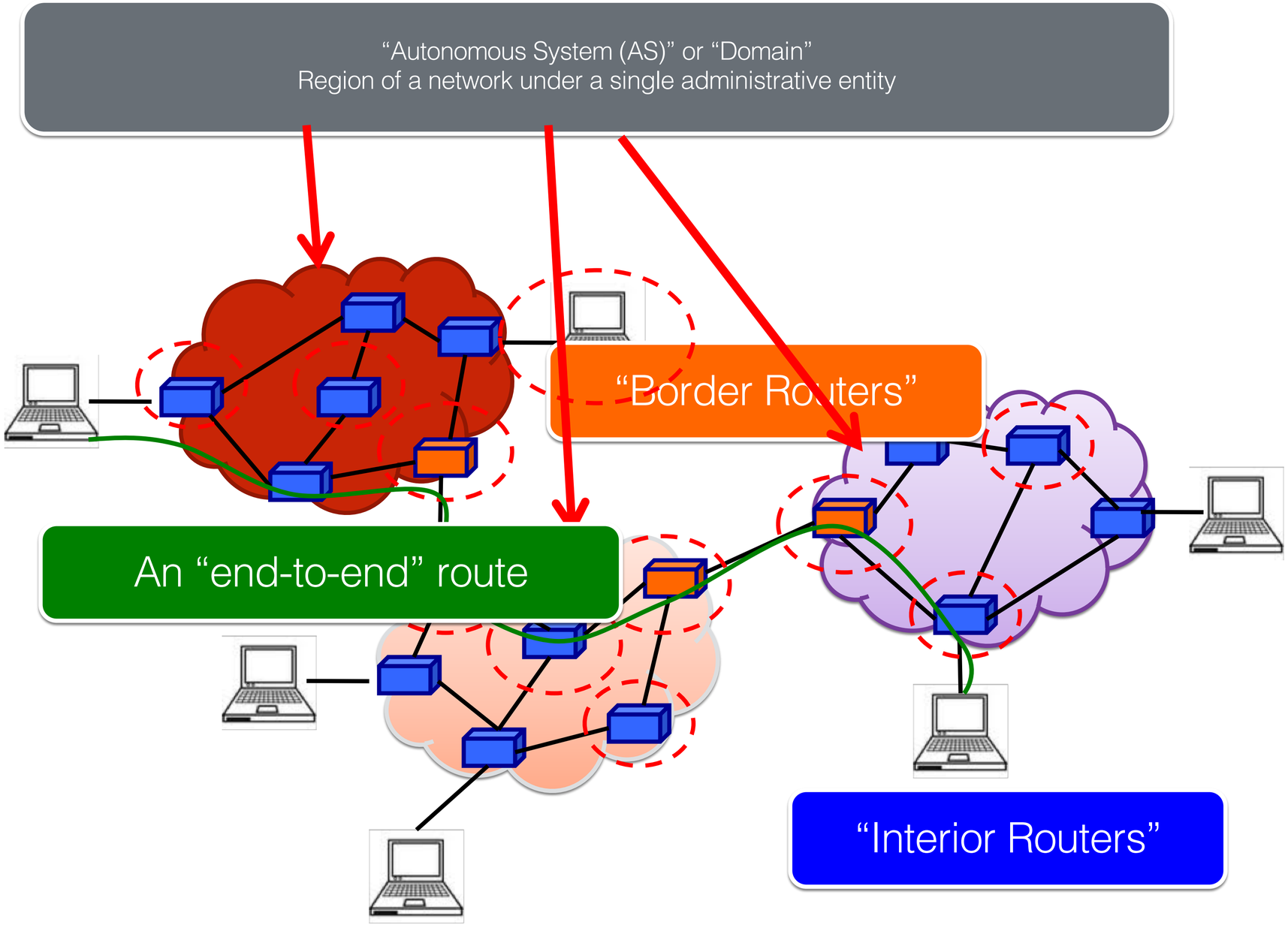
If you send a packet to nytimes.com, what will the IP destination address be?
What will the Ethernet destination address be?

Command line tools: dig, route

Okay great! On to our friend routing.

Inter and Intra-Domain Routing




"Autonomous System (AS)" or "Domain"
Region of a network under a single administrative entity
"Border Routers"
An "end-to-end" route
"Interior Routers"

Internet's Area Hierarchy

- What is an Autonomous System (AS)?
  - A set of routers under a single technical administration, using an *interior gateway protocol (IGP)* and common metrics to route packets within the AS and using an *exterior gateway protocol (EGP)* to route packets to other AS's
- Each AS assigned unique ID
  - Only transit domains really need it
- ASes peer with other ASes at network exchanges
  - "Gateway routers" forward packets across ASes

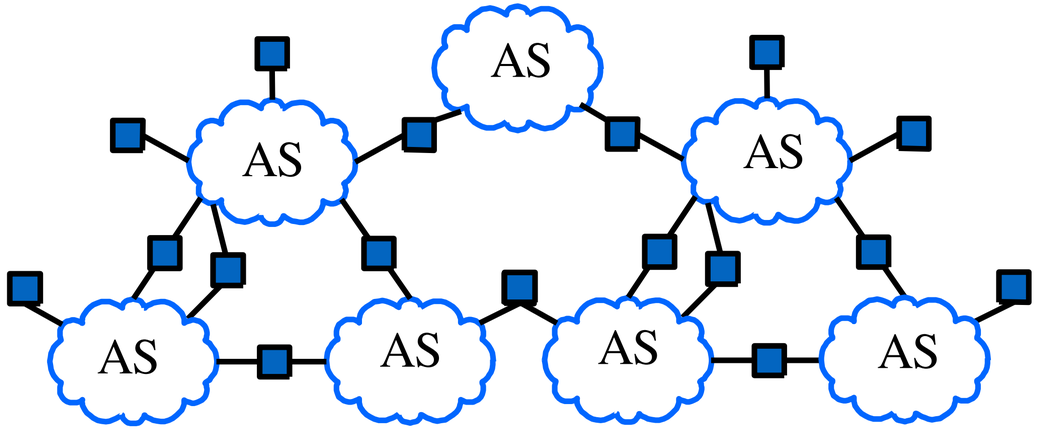
AS Numbers (ASNs)

ASNs are 16 bit values 64512 through 65535 are “private”

- Genuity: 1
- MIT: 3
- CMU: 9
- UC San Diego: 7377
- AT&T: 7018, 6341, 5074, …
- UUNET: 701, 702, 284, 12199, …
- Sprint: 1239, 1240, 6211, 6242, …
- …

ASNs represent units of routing policy

A Logical View of the Internet?



Algorithms we Know So Far

- Broadcast
- Distance Vector
- Link State
- Do you think they are a good choice for Internet, end to end routing?

Not so much

- Scale
  - Do we really want to run Distance Vector or Link State across all routers on the Internet?
- Administrative Control
  - Does an ISP really want to share all of its routes with the whole world?
  - Issues of autonomy, privacy, policy.

By now you should know the key ideas behind scaling

Addressing Goal: Scalable Routing

- State: Small forwarding tables at routers
  - Much less than the number of hosts

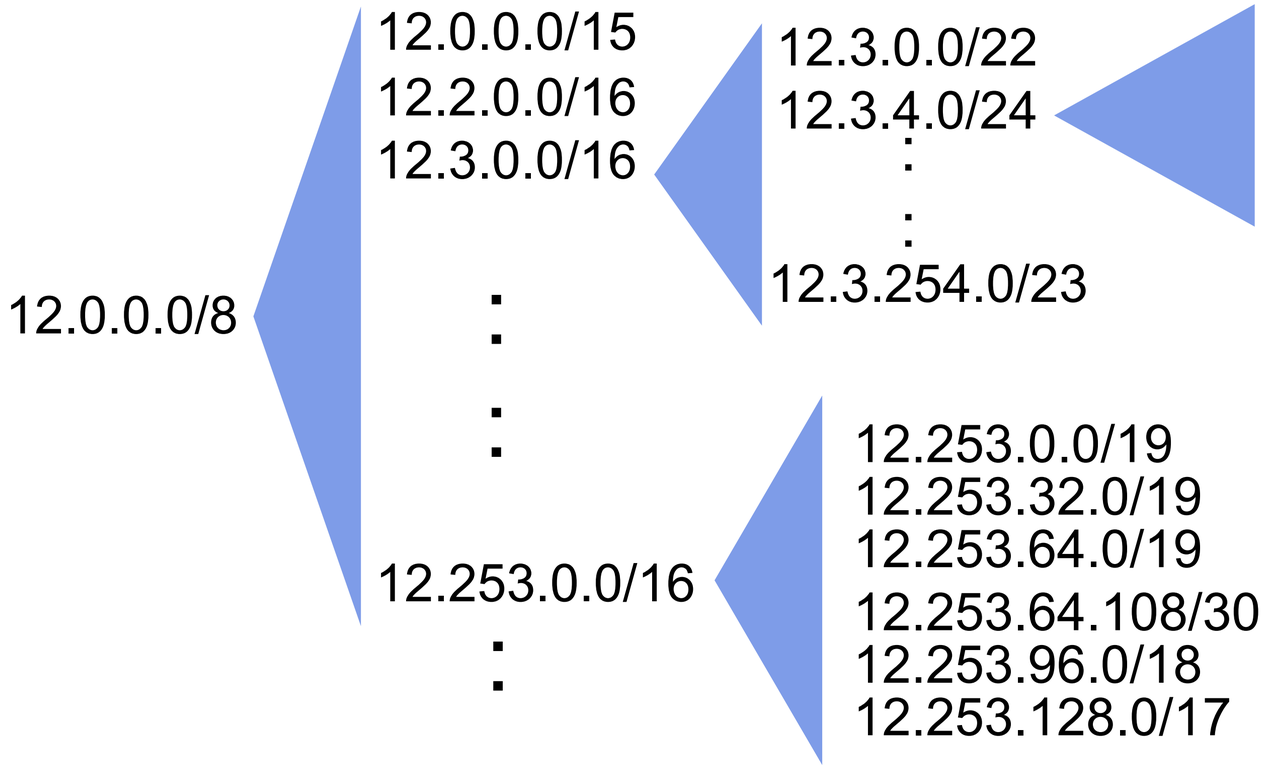
- Churn: Limited rate of change in routing tables
  - Traffic, inconsistencies, complexity

Ability to aggregate addresses is crucial for both

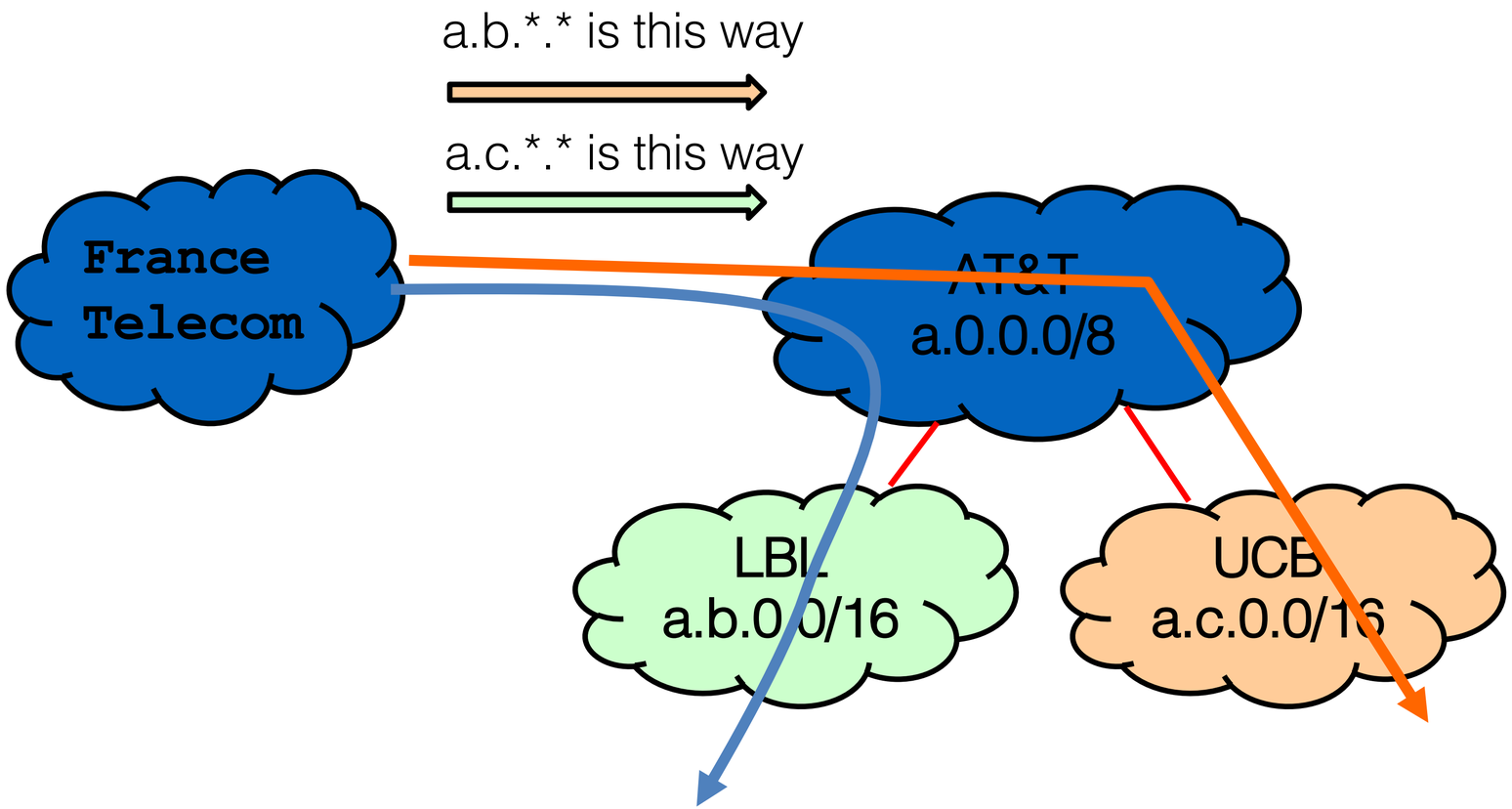
(one entry to *summarize* many addresses)

CIDR: Addresses allocated in contiguous prefix chunks

Recursively break down chunks as get closer to host

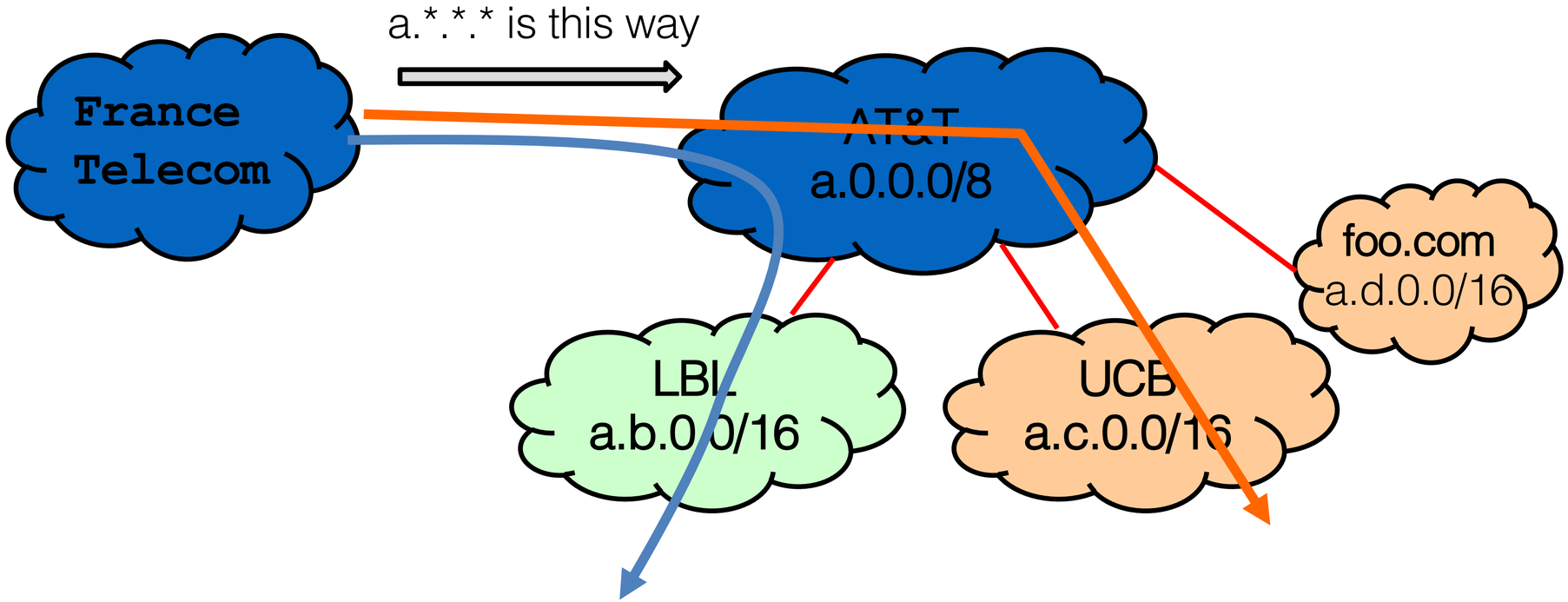


IP addressing → scalable routing?



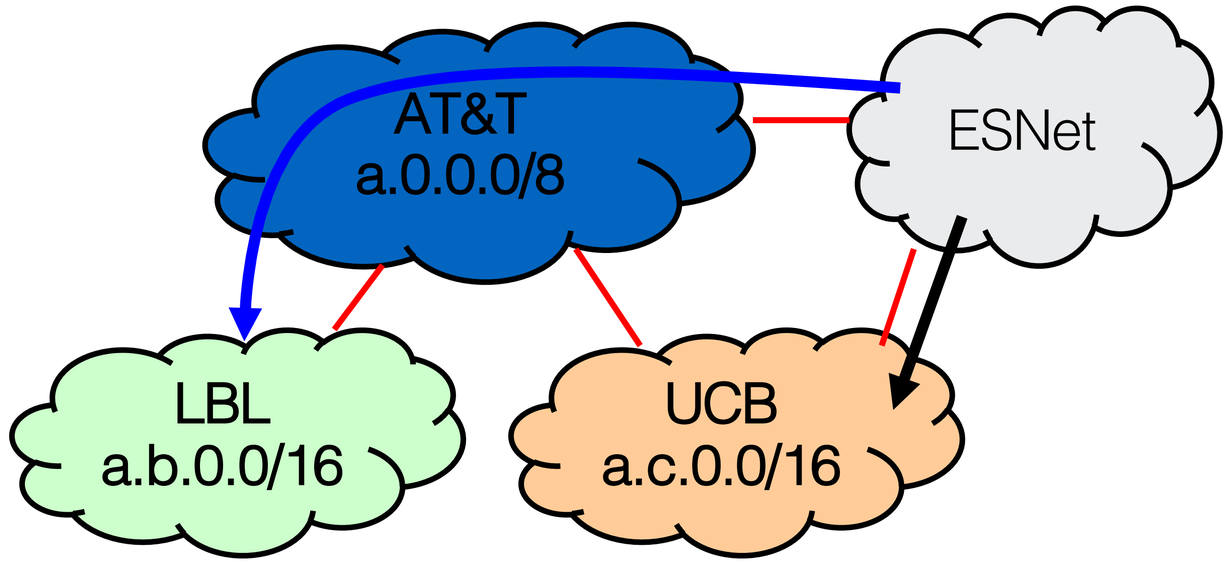
IP addressing → scalable routing?

Can add new hosts/networks without updating the routing entries at France Telecom



IP addressing → scalable routing?

ESNet must maintain routing entries for both a.*.*.* and a.c.*.*



This is important! Make sure you remember this from a few lectures ago!

Administrative structure shapes Interdomain routing

- ASes want freedom to pick routes based on **policy**
  - *“My traffic can’t be carried over my competitor’s network”*
  - *“I don’t want to carry A’s traffic through my network”*
  - Not expressible as Internet-wide “shortest path”!
- ASes want **autonomy**
  - Want to choose their own internal routing protocol
  - Want to choose their own policy
- ASes want **privacy**
  - choice of network topology, routing policies, *etc.*

Choice of Routing Algorithm

Link State (LS) *vs.* Distance Vector (DV)?

- LS offers no privacy -- global sharing of all network information (neighbors, policies)
- LS limits autonomy -- need agreement on metric, algorithm
- DV is a decent starting point
  - per-destination advertisement gives providers a hook for finer-grained control over whether/which routes to advertise
  - but DV wasn't designed to implement policy

The "Border Gateway Protocol" (BGP) extends distance-vector ideas to accommodate policy

BGP

- The role of policy
  - what we mean by it
  - why we need it
- Overall approach
  - four non-trivial changes to DV
  - how policy is implemented

Administrative structure shapes Interdomain routing

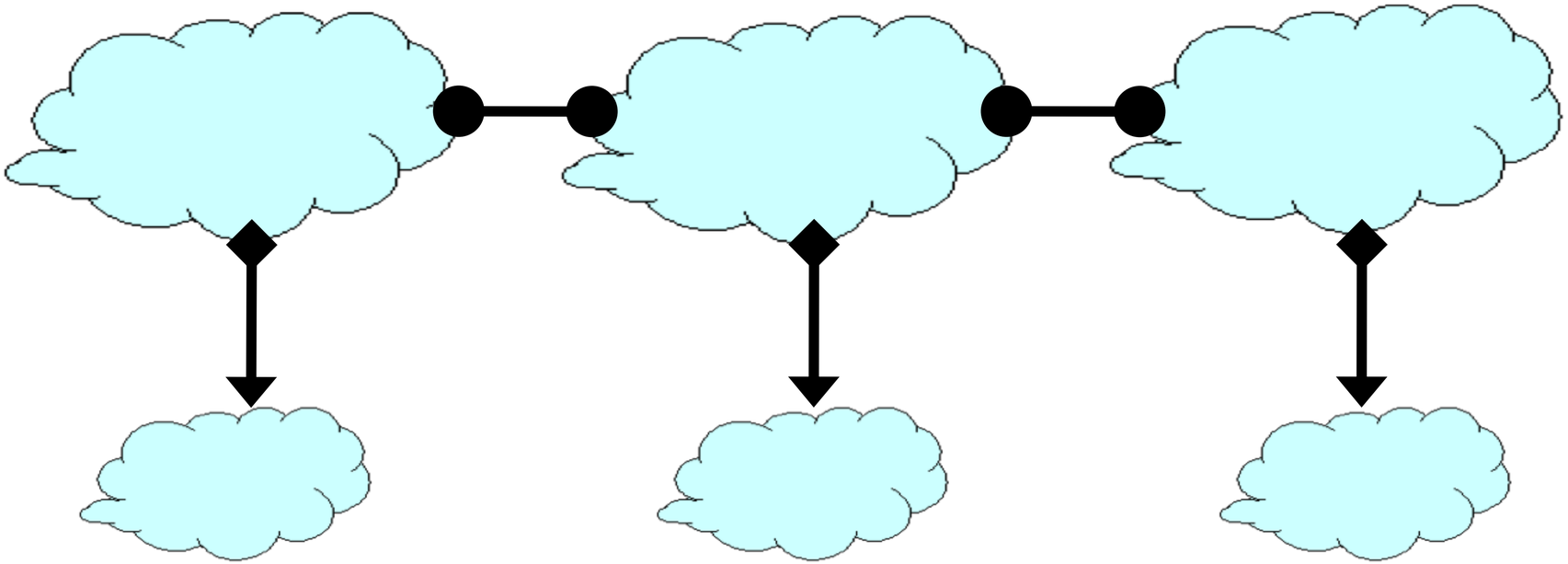
- ASes want freedom to pick routes based on policy
- ASes want autonomy
- ASes want privacy

Topology and policy is shaped by the business relationships between ASes

- Three basic kinds of relationships between ASes
  - AS A can be AS B's *customer*
  - AS A can be AS B's *provider*
  - AS A can be AS B's *peer*

- Business implications
  - Customer pays provider
  - Peers don't pay each other
    - Exchange roughly equal traffic

Business Relationships



Relations between ASes

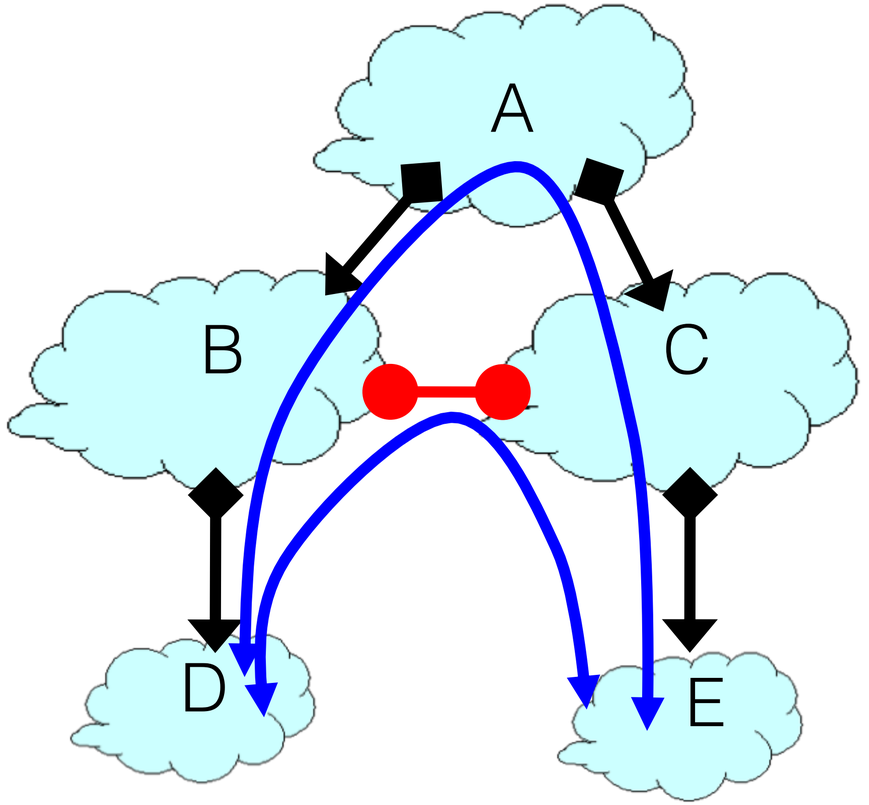
provider ⟶ customer

peer — peer

Business Implications

- **Customers pay provider**
- Peers don’t pay each other

Why peer?



E.g., D and E talk a lot

Peering saves B *and* C money

Relations between ASes

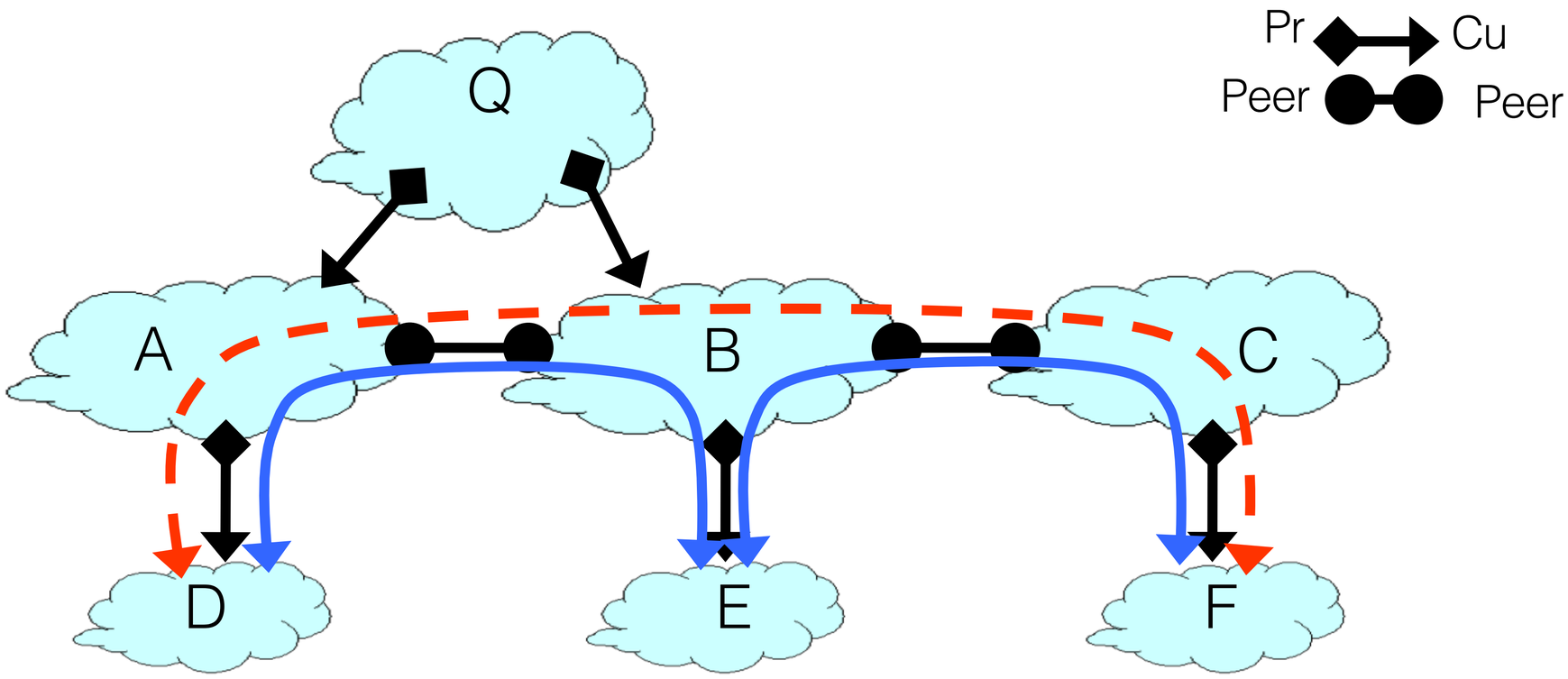
provider → customer

peer — peer

Business Implications

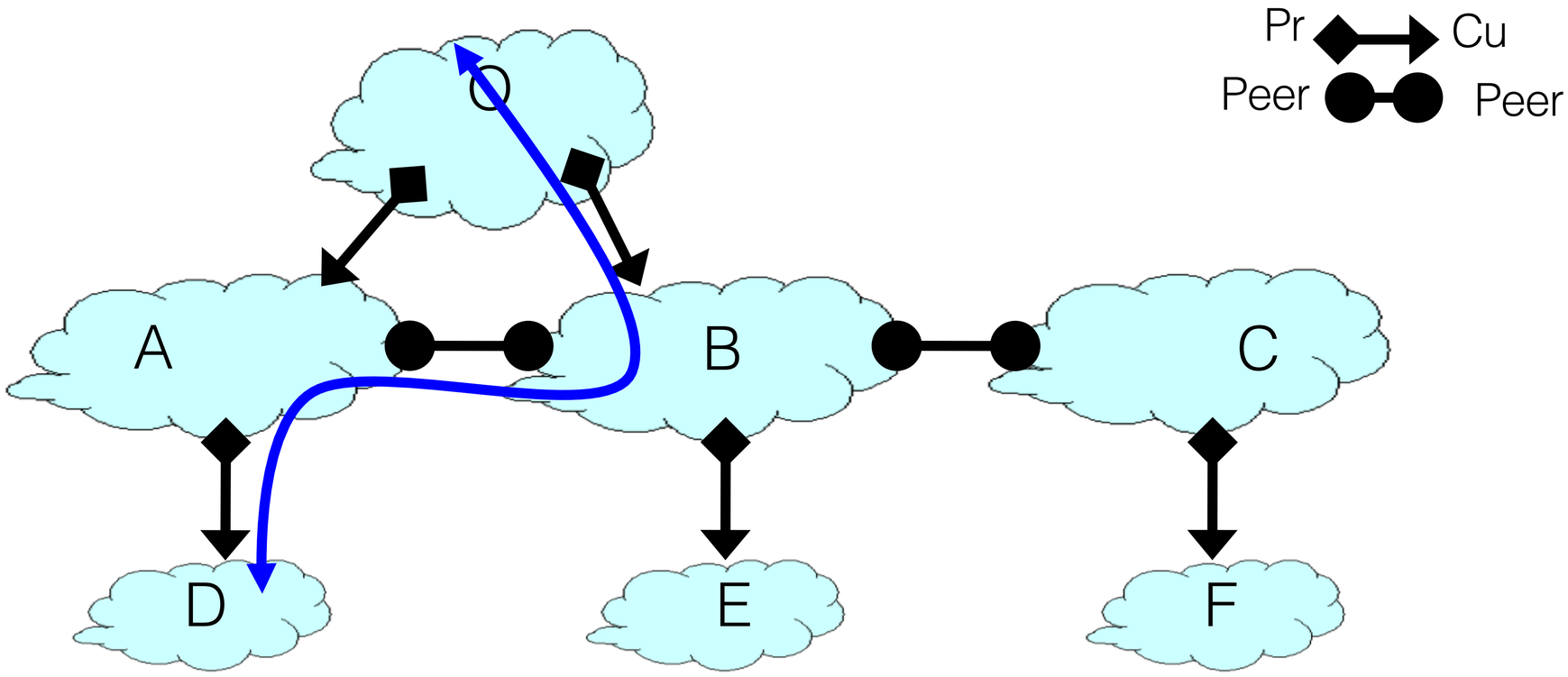
- **Customers pay provider**
- Peers don't pay each other

Routing Follows the Money!



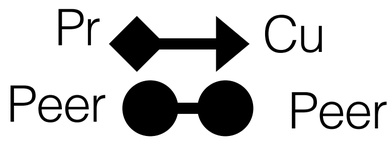
- ASes provide “transit” between their customers
- Peers do not provide transit between other peers

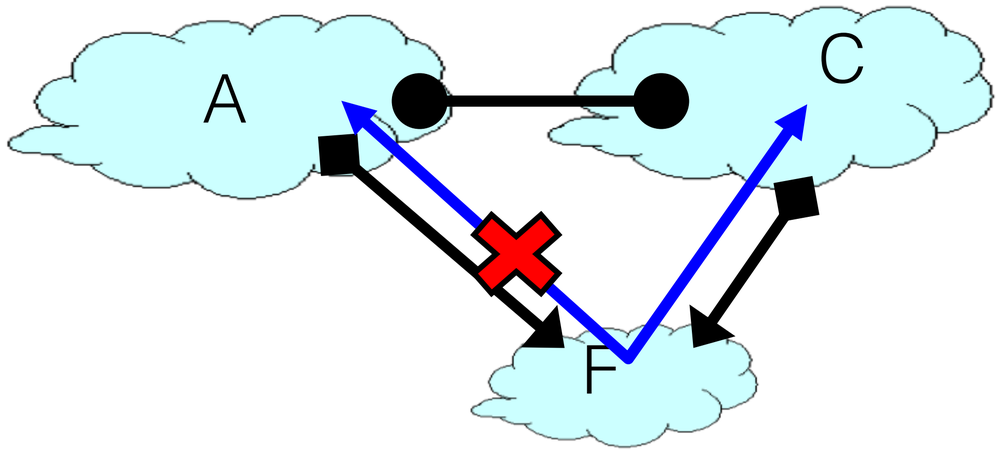
Routing Follows the Money!



- An AS only carries traffic to/from its own customers over a peering link

Routing Follows the Money!





- Routes are “valley free” (will return to this later)

In Short

- AS topology reflects business relationships between Ases
- Business relationships between ASes impact which routes are acceptable
- BGP Policy: Protocol design that allows ASes to control which routes are used

BGP

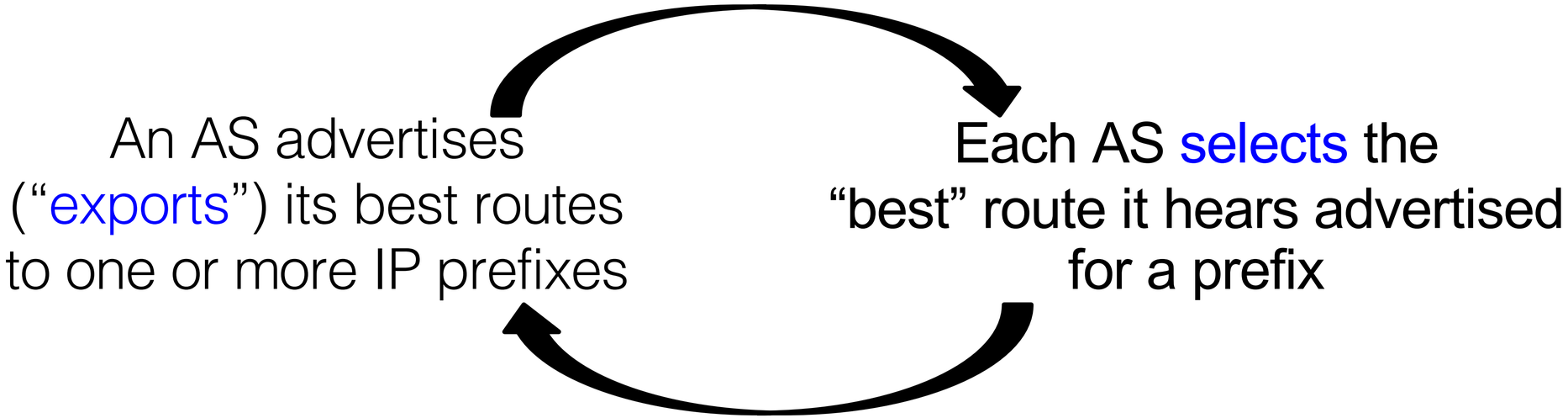
- The role of policy
  - what we mean by it
  - why we need it

- Overall approach
  - four non-trivial changes to DV
  - how policy is implemented

Interdomain Routing: Setup

- Destinations are IP prefixes (12.0.0.0/8)
- Nodes are Autonomous Systems (ASes)
  - Internals of each AS are hidden
- Links represent both physical links and business relationships
- BGP (Border Gateway Protocol) is the Interdomain routing protocol
  - Implemented by AS border routers

BGP: Basic Idea



You've heard this story before!

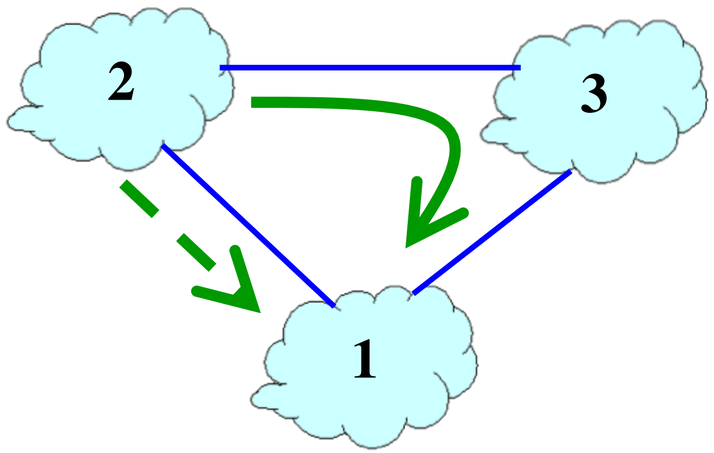
BGP inspired by Distance Vector

- Per-destination route advertisements
- No global sharing of network topology information
- Iterative and distributed convergence on paths
- With four crucial differences!

Differences between BGP and DV
(1) not picking shortest path routes

- BGP selects the best route based on policy, not shortest distance (least cost)

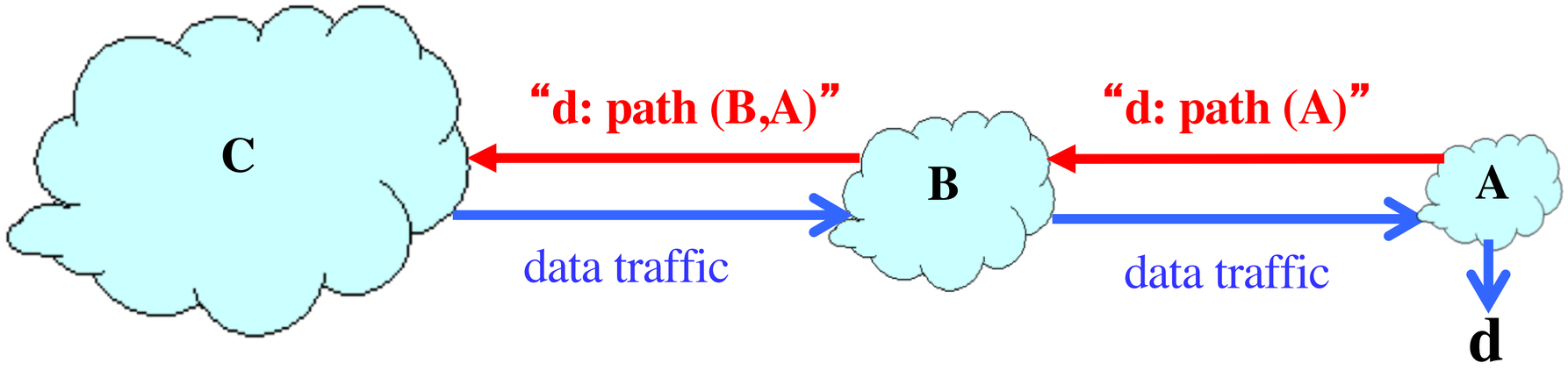
Node 2 may prefer "2, 3, 1" over "2, 1"



- How do we avoid loops?

Differences between BGP and DV
(2) path-vector routing

- Key idea: advertise the entire path
  - Distance vector: send *distance metric* per dest d
  - Path vector: send the *entire path* for each dest d



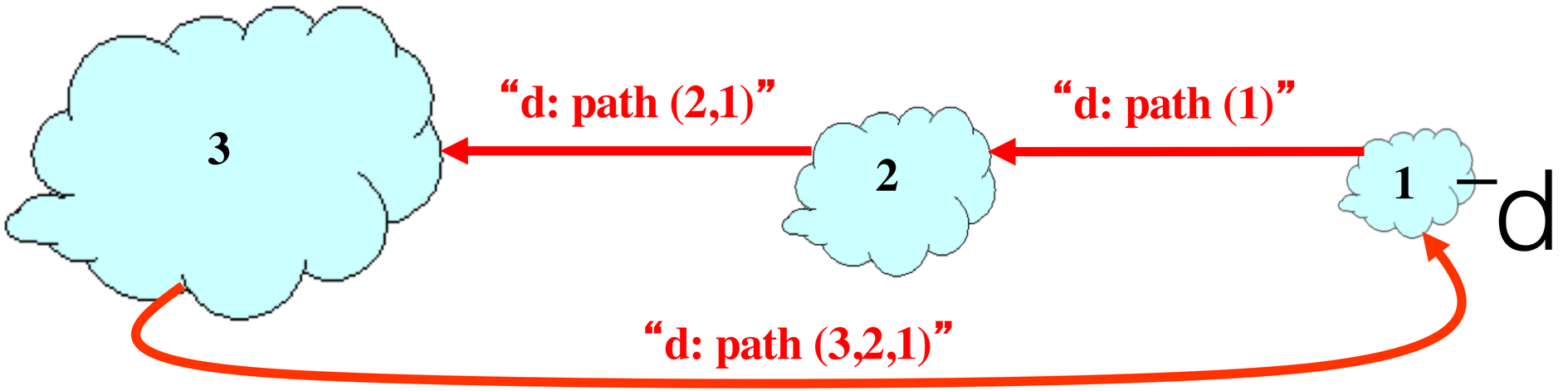
Differences between BGP and DV
(2) path-vector routing

- Key idea: advertise the entire path
  - Distance vector: send *distance metric* per dest d
  - Path vector: send the *entire path* for each dest d

- Benefits
  - loop avoidance is easy

Loop Detection w/ Path-Vector

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



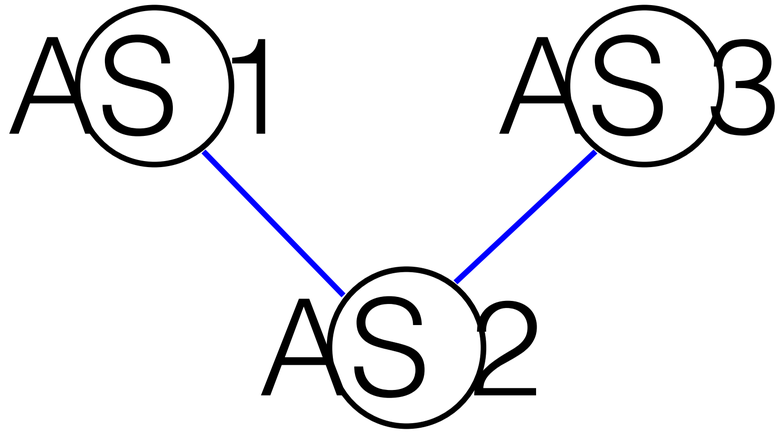
Differences between BGP and DV
(2) path-vector routing

- Key idea: advertise the entire path
  - Distance vector: send *distance metric* per dest d
  - Path vector: send the *entire path* for each dest d

- Benefits
  - loop avoidance is easy
  - flexible policies based on entire path

Differences between BGP and DV
(3) Selective route advertisement

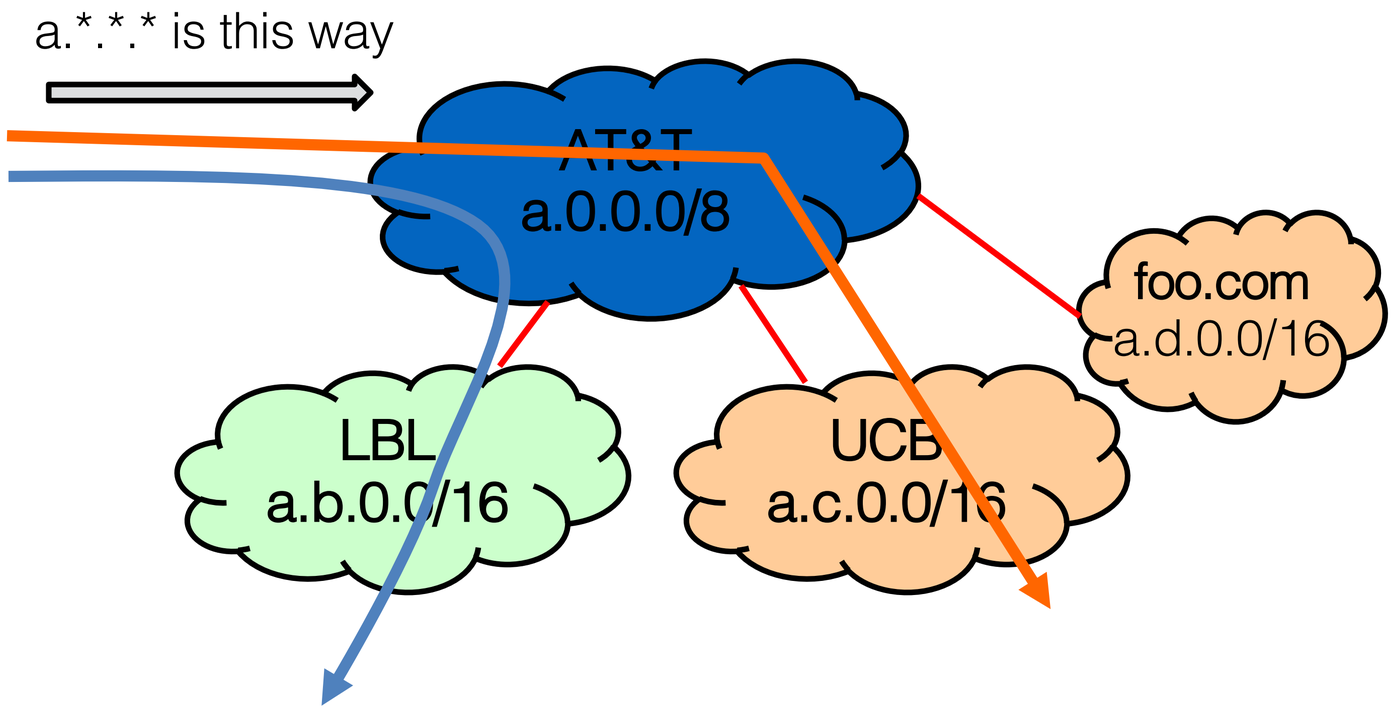
- For policy reasons, an AS may choose not to advertise a route to a destination
- **Hence, reachability is not guaranteed even if graph is connected**



Example: AS#2 does not want to carry traffic between AS#1 and AS#3

Differences between BGP and DV
(4) BGP may *aggregate* routes

- For scalability, BGP may aggregate routes for different prefixes

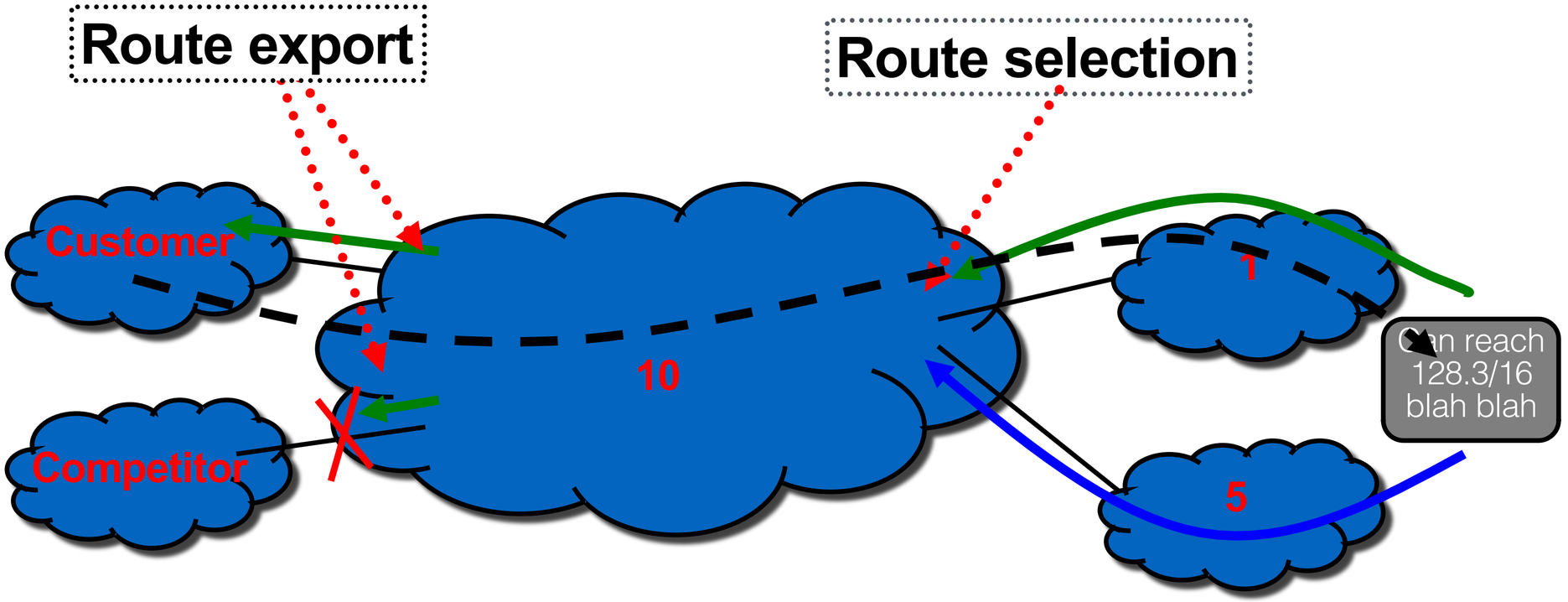


BGP

- The role of policy
  - what we mean by it
  - why we need it

- Overall approach
  - four non-trivial changes to DV
  - how policy is implemented

Policy imposed in how routes are selected and exported



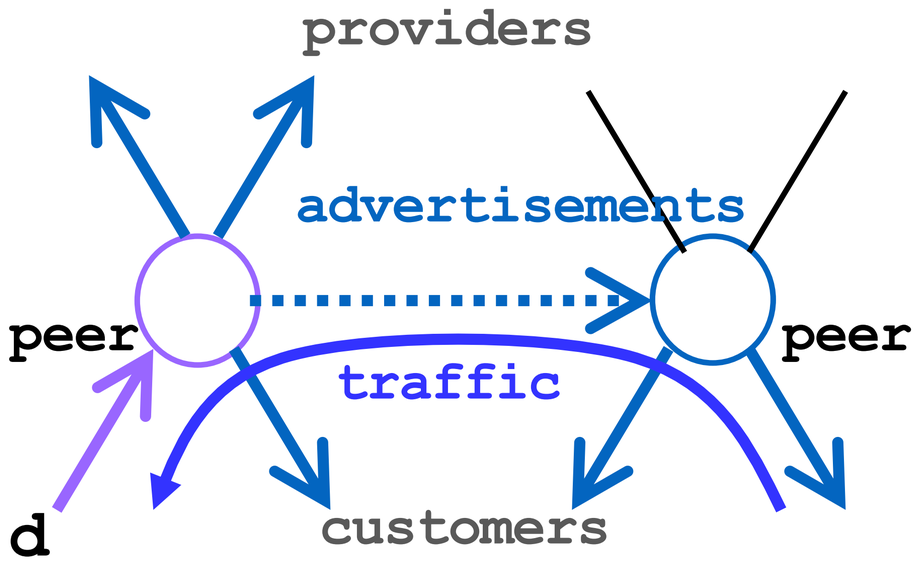
- **Selection**: Which path to use?
  - controls whether/how traffic leaves the network
- **Export**: Which path to advertise?
  - controls whether/how traffic enters the network

Typical Selection Policy

- In decreasing order of priority
  - make/save money (send to customer > peer > provider)
  - maximize performance (smallest AS path length)
  - minimize use of my network bandwidth (“hot potato”)
  - …
  - …
- BGP uses something called route “attributes” to implement the above (next lecture)

Typical Export: Peer-Peer Case

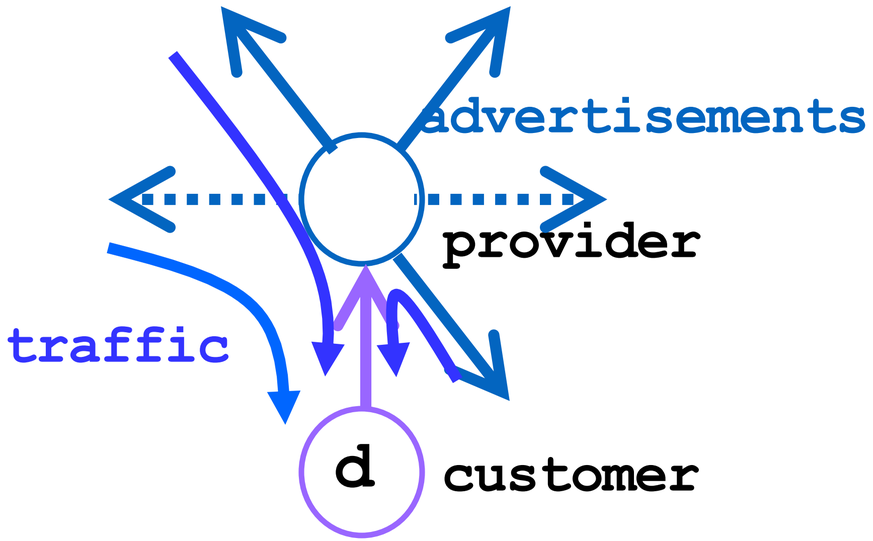
- Peers exchange traffic between their customers
  - AS exports only customer routes to a peer
  - AS exports a peer's routes only to its customers



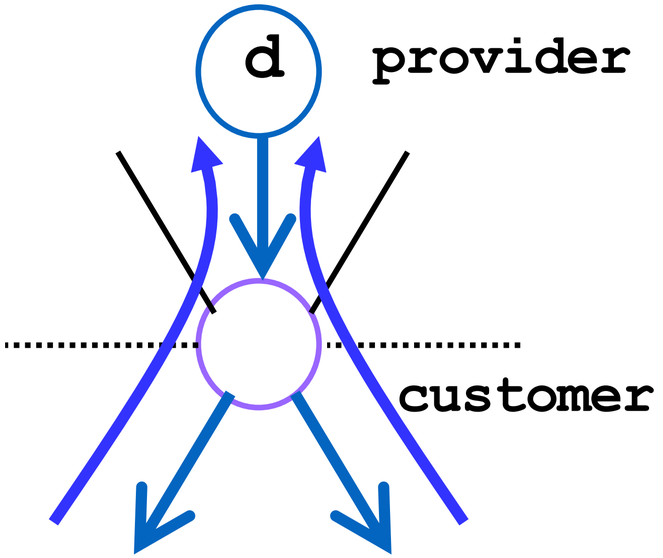
Typical Export: Customer-Provider

- Customer pays provider for access to Internet
  - Provider exports its customer routes to everybody
  - Customer exports provider routes only to its customers

Traffic to customer



Traffic from customer



Typical Export Policy

Destination prefix advertised by…	Export route to…
Customer	Everyone (providers, peers, other customers)
Peer	Customers
Provider	Customers

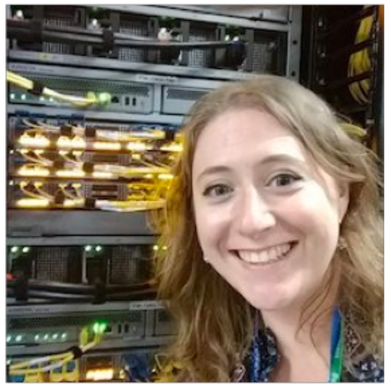
We'll refer to these as the "Gao-Rexford" rules (capture common -- but not required! -- practice!)

Jennifer Rexford

Department chair of CS at Princeton
ACM Fellow, SIGCOMM Achievement Award, National Academy of Engineers, Hopper Award

Known for:

- Gao Rexford Conditions
- Software Defined Networking Fundamentals
- Work prior to Princeton at AT&T bridging industry and research
- Original design of most networking slides used in every class in the country

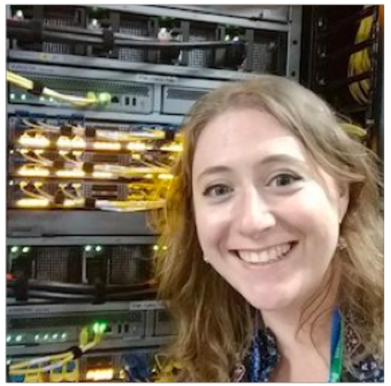


I stole slides from

Sylvia Ratnasamy
Who stole slides from

Scott Shenker
Who stole slides from

JEN REXFORD



I stole slides from

Sylvia Ratnasamy
Who stole slides from…

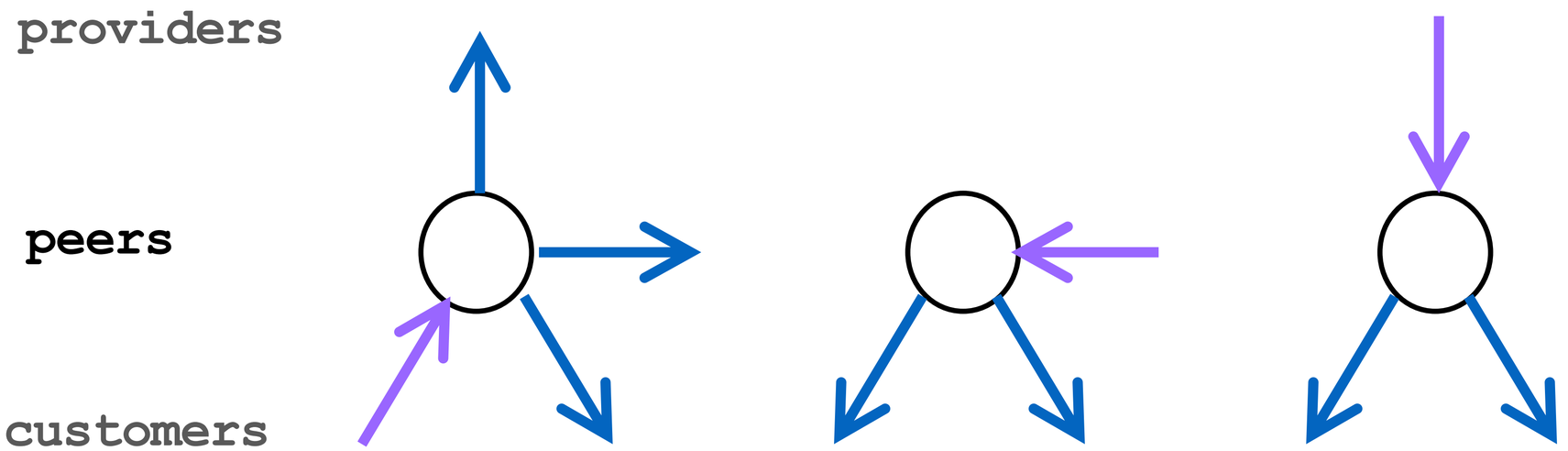
Scott Shenker
Who stole slides from..

Peter Steenkiste,
who stole slides from…

Srini Seshan,
who stole slides from…

JEN REXFORD

Gao-Rexford



With Gao-Rexford, the customer-provider graph is a DAG (directed acyclic graph) and routes are “valley free”
What does “Valley Free” mean here?

Activity

- X is a small university network with two providers, A and B.
- A's provider is C.
- B's provider is D.
- C's provider is Z.
- D's provider is Z.

- What AS path does traffic take from A to B?
  - Why?

Activity

- A’s provider is Z. A peers with B.
- B’s provider is Z. B peers with A and C.
- C’s provider is Y. C peers with B.
- Z’s provider is X.
- Y’s provider is X.

- What AS path does traffic take from A to C?
  - Why?

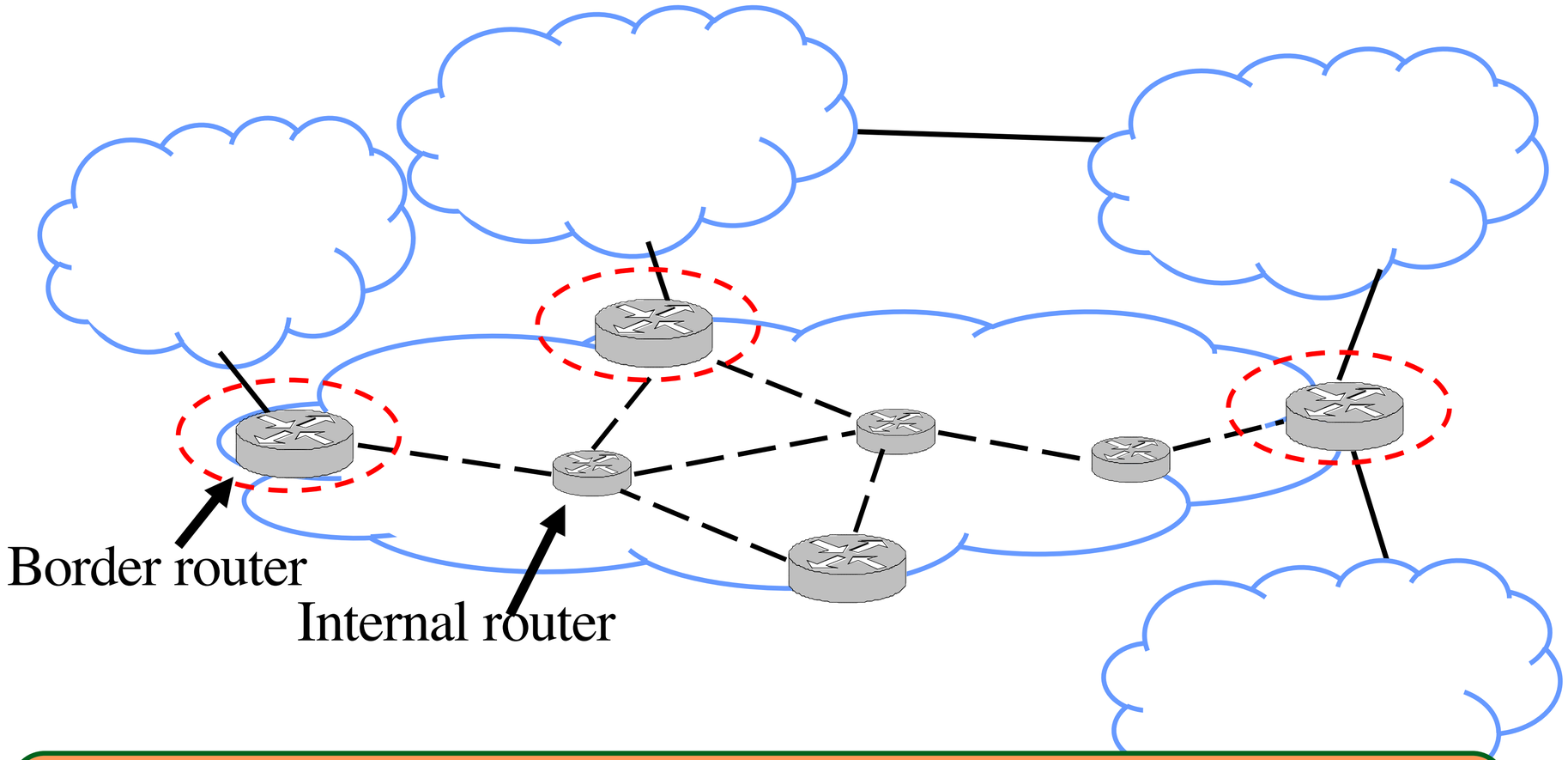
BGP

- BGP policy
  - typical policies, how they're implemented

- BGP protocol details
  - stay awake as long as you can…

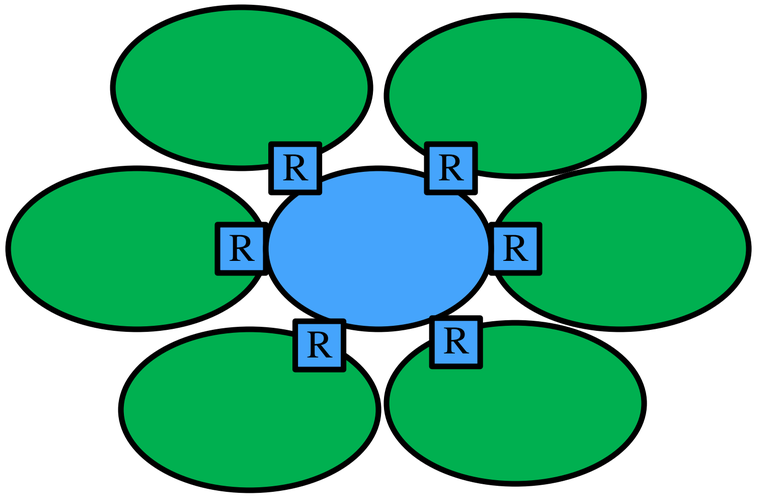
- BGP issues

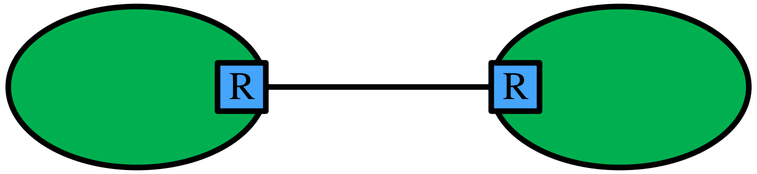
Who speaks BGP?



Border routers at an Autonomous System

How Do ISPs Peer?



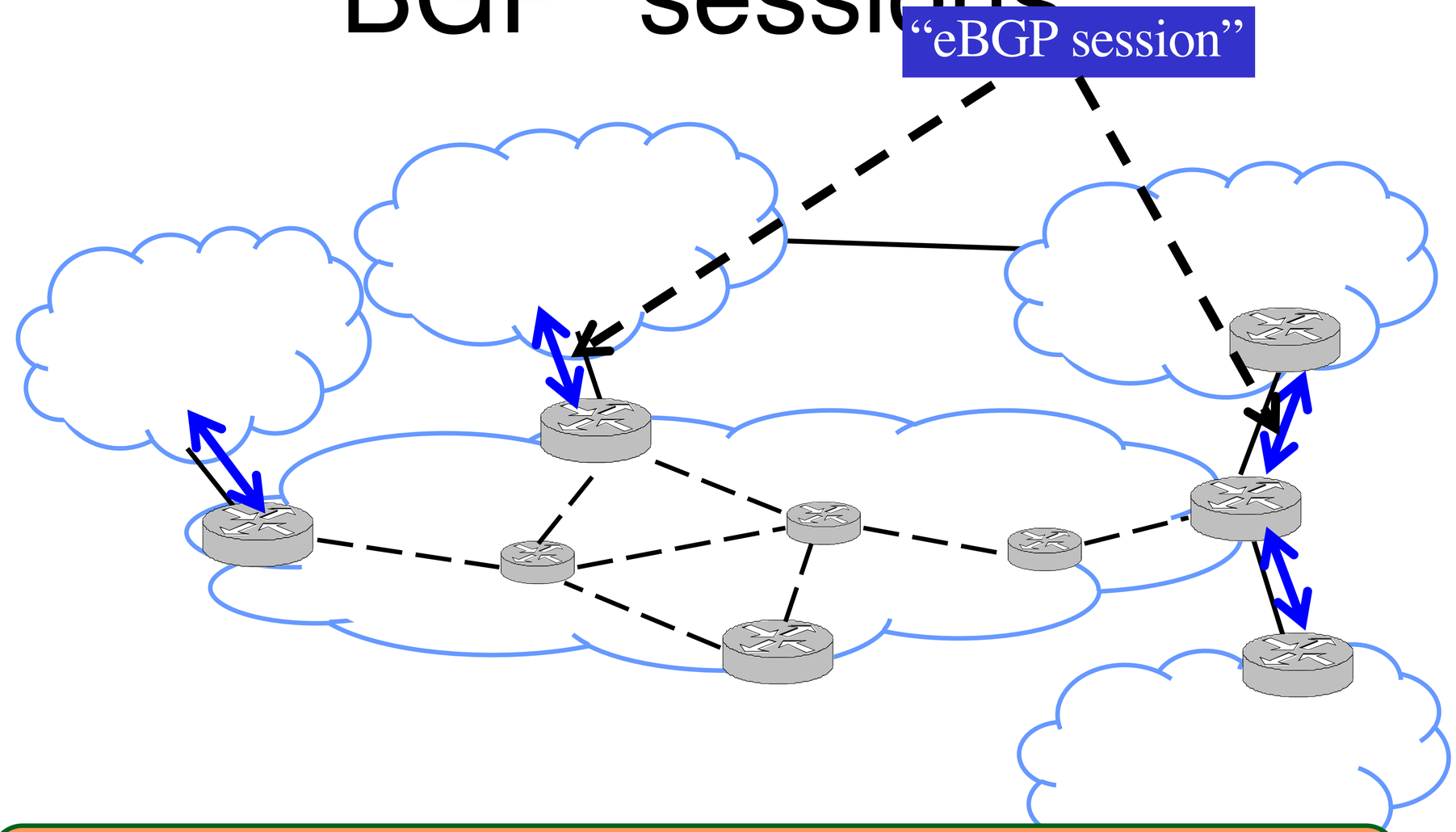


- Public peering: use network to connect large number of ISPs in Internet eXchange Point (IXP)
  - Managed by IXP operator
  - Layer 2 private network
  - Efficient: can have 100s of ISPs
  - Has led to increase in peering
- Private peering: directly connect ISP border routers
  - Set up as private connection
  - Typically done in an Internet eXchange Point (IXP)

What does “speak BGP” mean?

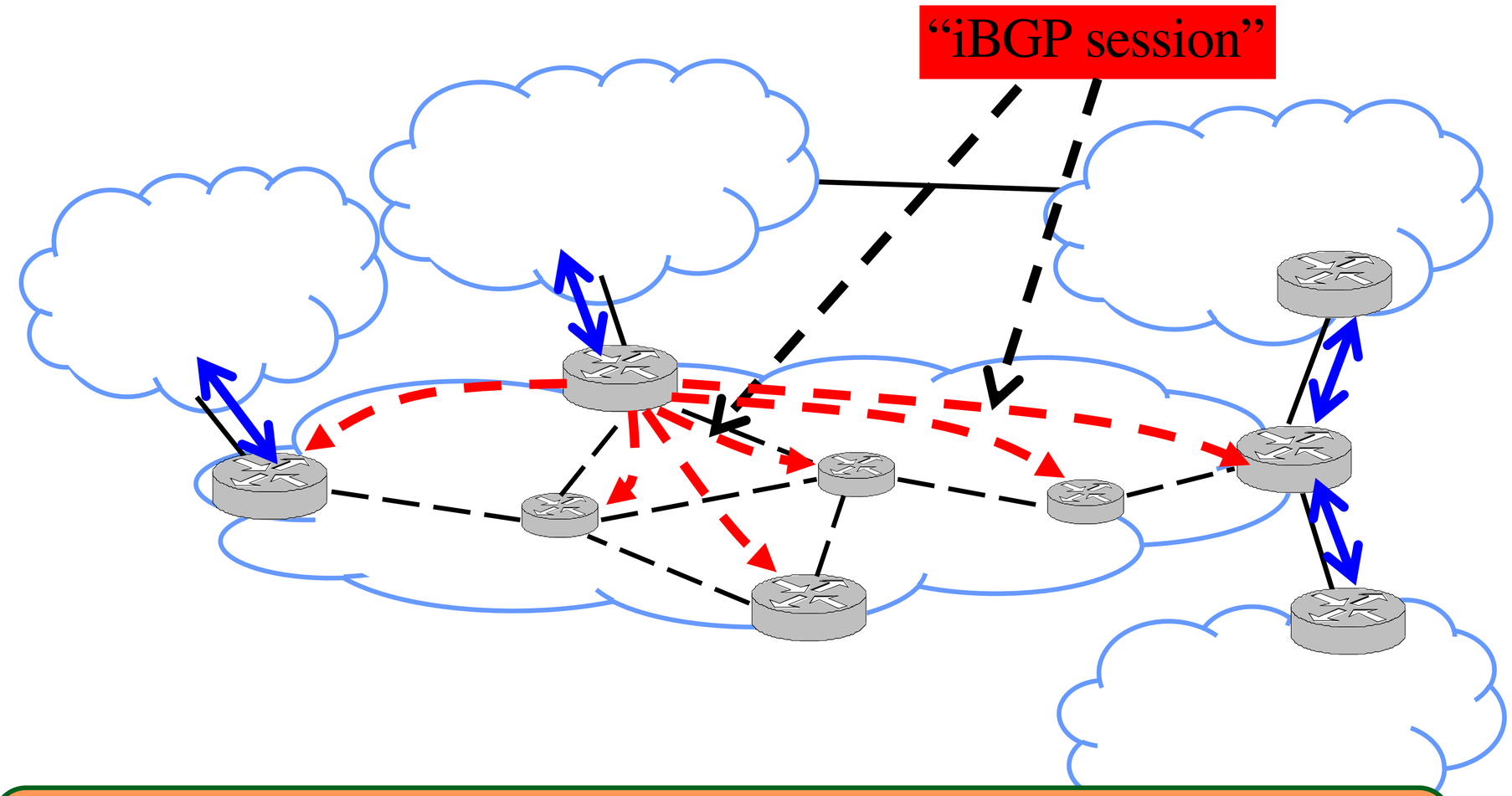
- Implement the standardized BGP protocol
  - read more here: http://tools.ietf.org/html/rfc4271
- Specifies what messages to exchange with other BGP “speakers”
  - message types: e.g., route advertisements
  - message syntax: e.g., first X bytes for dest prefix; next Y for AS path, *etc.*
- And how to process these messages
  - e.g., *“when you receive a message of type X, apply this selection rule, then…”*
  - as per BGP state machine in the protocol spec + policy decisions, etc.

BGP “sessions”



A border router speaks BGP with
border routers in other ASes

BGP “sessions”



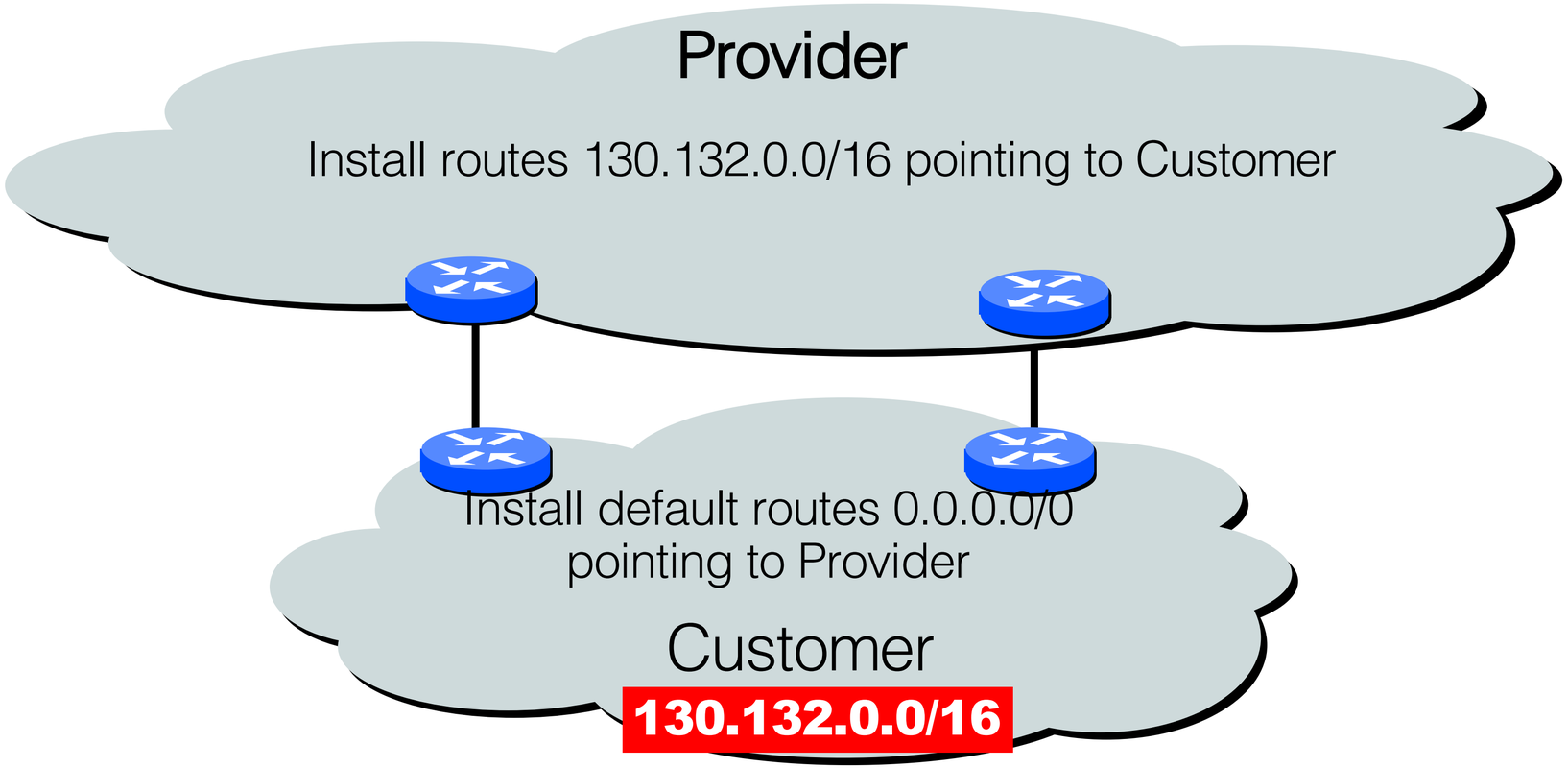
A border router speaks BGP with other (interior and border) routers in its own AS

eBGP, iBGP, IGP

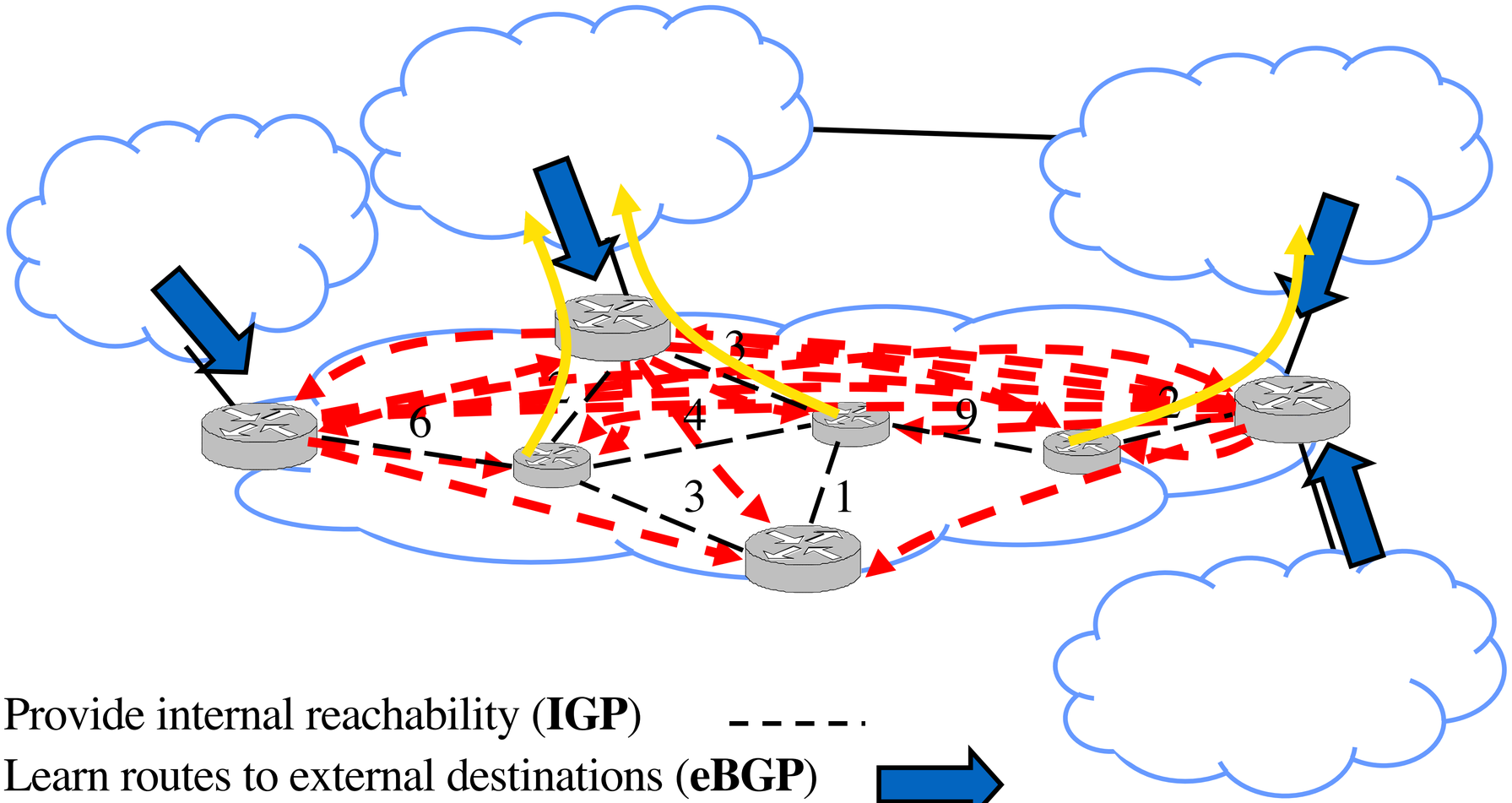
- eBGP: BGP sessions between border routers in different ASes
  - Learn routes to external destinations
- iBGP: BGP sessions between border routers and other routers within the same AS
  - distribute externally learned routes internally
  - assume a full all-to-all mesh of iBGP sessions
- IGP: “Interior Gateway Protocol” = Intradomain routing protocol
  - provide internal reachability
  - e.g., OSPF, RIP

Some Border Routers Don’t Need BGP

- Customer that connects to a single upstream ISP
  - The ISP can advertise prefixes into BGP on behalf of customer
  - … and the customer can simply default-route to the ISP



Putting the pieces together

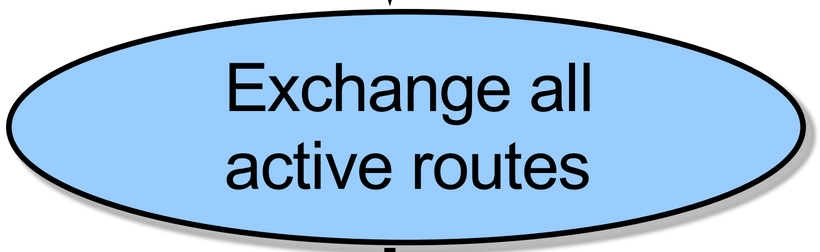


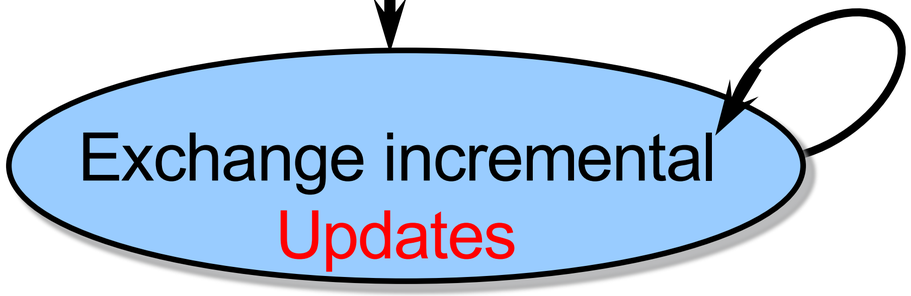
1. Provide internal reachability (**IGP**)
2. Learn routes to external destinations (**eBGP**)
3. Distribute externally learned routes internally (**iBGP**)
4. Travel shortest path to egress (IGP)

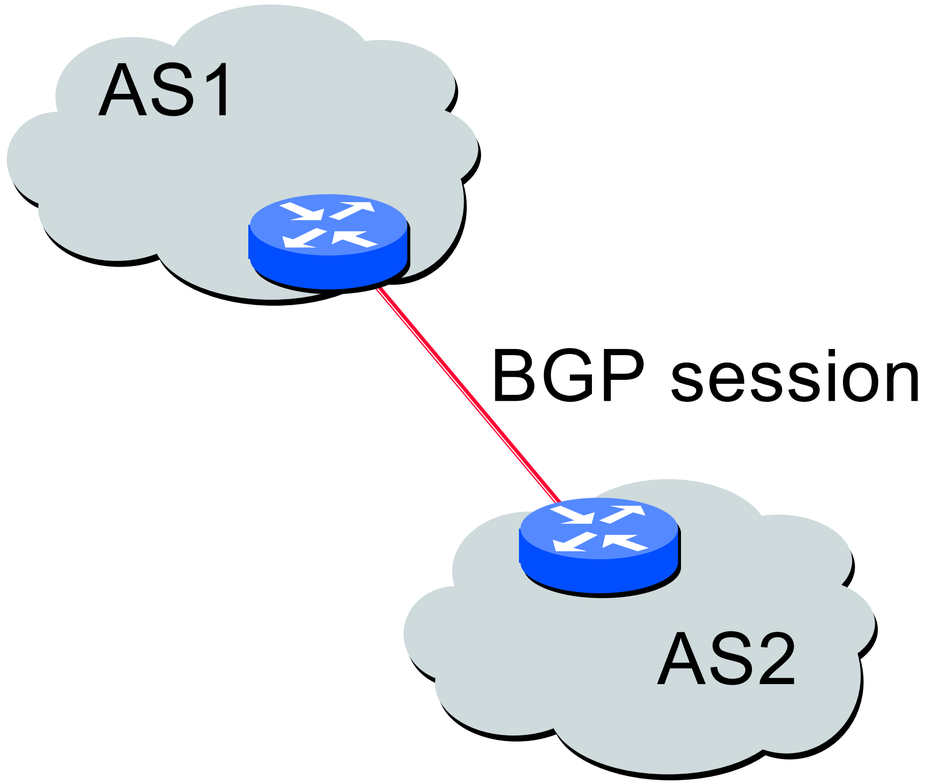
Basic Messages in BGP

- **Open**
  - Establishes BGP session
  - BGP uses TCP *[will make sense in 1-2weeks]*
- **Notification**
  - Report unusual conditions
- **Update**
  - Inform neighbor of new routes
  - Inform neighbor of old routes that become inactive
- **Keepalive**
  - Inform neighbor that connection is still viable

BGP Operations







While connection is ALIVE exchange

Route Updates

- Format *<IP prefix: route attributes>*
  - attributes describe properties of the route

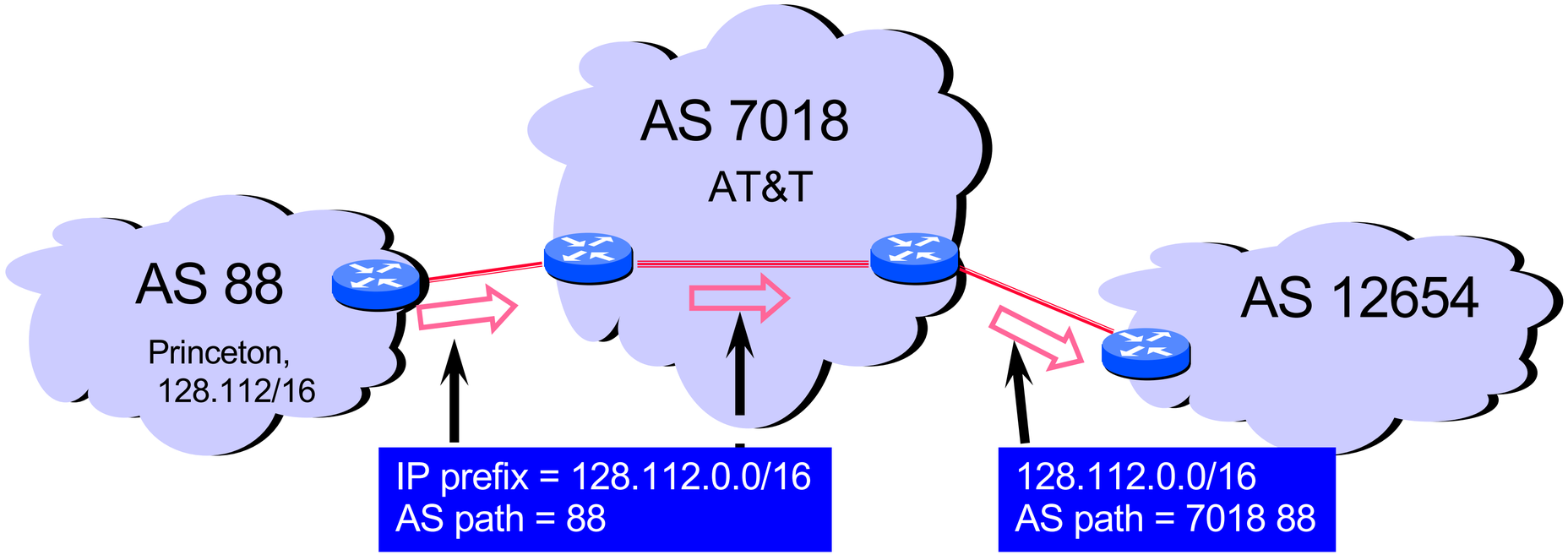
- Two kinds of updates
  - announcements: new routes or changes to existing routes
  - withdrawal: remove routes that no longer exist

Route Attributes

- Routes are described using attributes
  - Used in route selection/export decisions
- Some attributes are local
  - i.e., private within an AS, not included in announcements
  - e.g., `LOCAL PREF`, `ORIGIN`
- Some attributes are propagated with eBGP route announcements
  - e.g., `NEXT HOP`, `AS PATH`, `MED`, *etc.*
- There are many standardized attributes in BGP
  - We will discuss a few

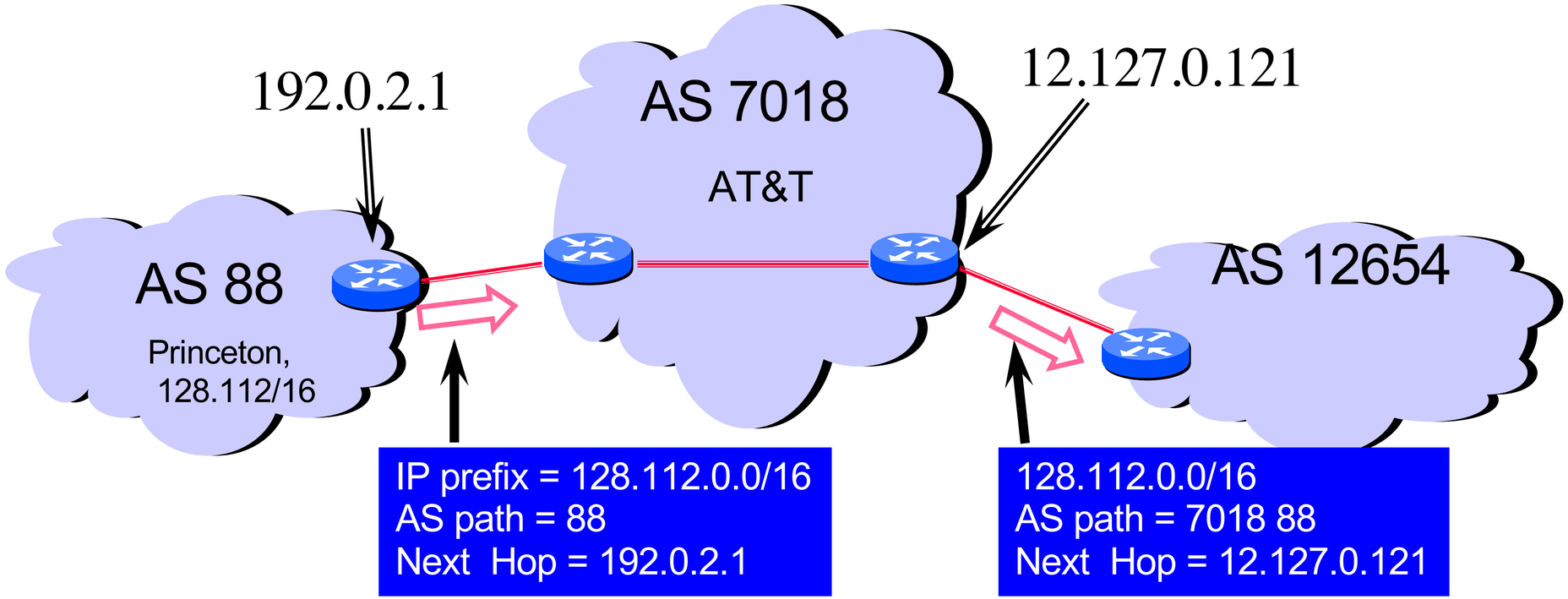
Attributes (1): ASPATH

- Carried in route announcements
- Vector that lists all the ASes a route announcement has traversed (in reverse order)
- e.g., “7018 88”



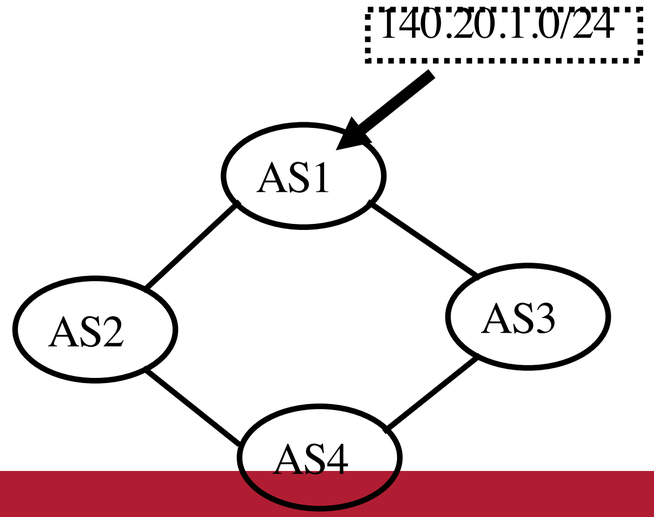
Attributes (2): NEXT HOP

- Carried in a route update message
- IP address of next hop router on path to destination
- Updated as the announcement leaves AS



Attributes (3): LOCAL PREF

- “Local Preference”
- Used to choose between different AS paths
- The higher the value the more preferred
- Local to an AS; carried only in iBGP messages
- Ensures consistent route selection across an AS

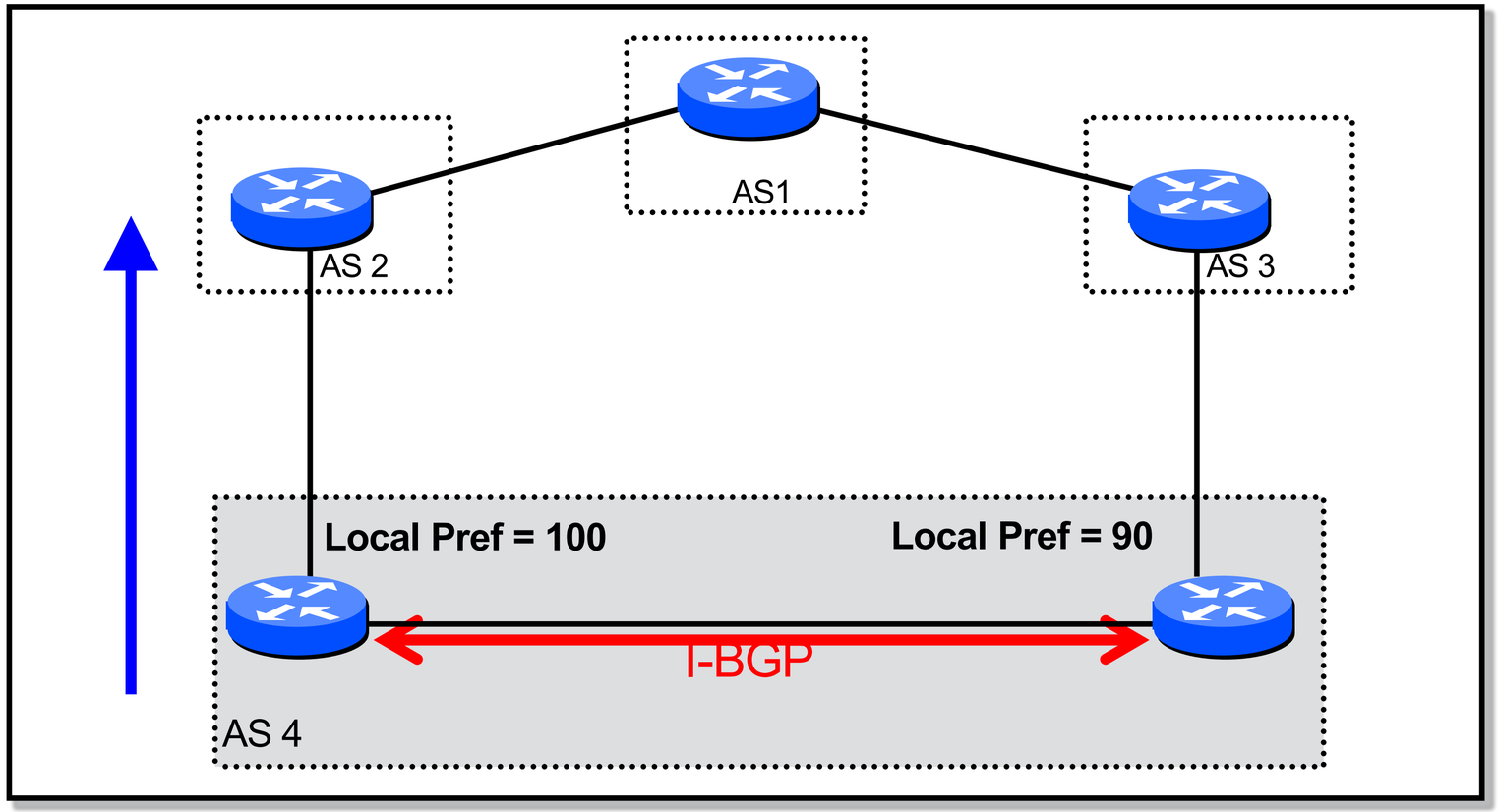


BGP table at AS4:

Destination	AS Path	Local Pref
140.20.1.0/24	**AS3 AS1**	**300**
140.20.1.0/24	**AS2 AS1**	**100**

Example: iBGP and LOCAL PREF

- Both routers prefer the path through AS 100 on the left

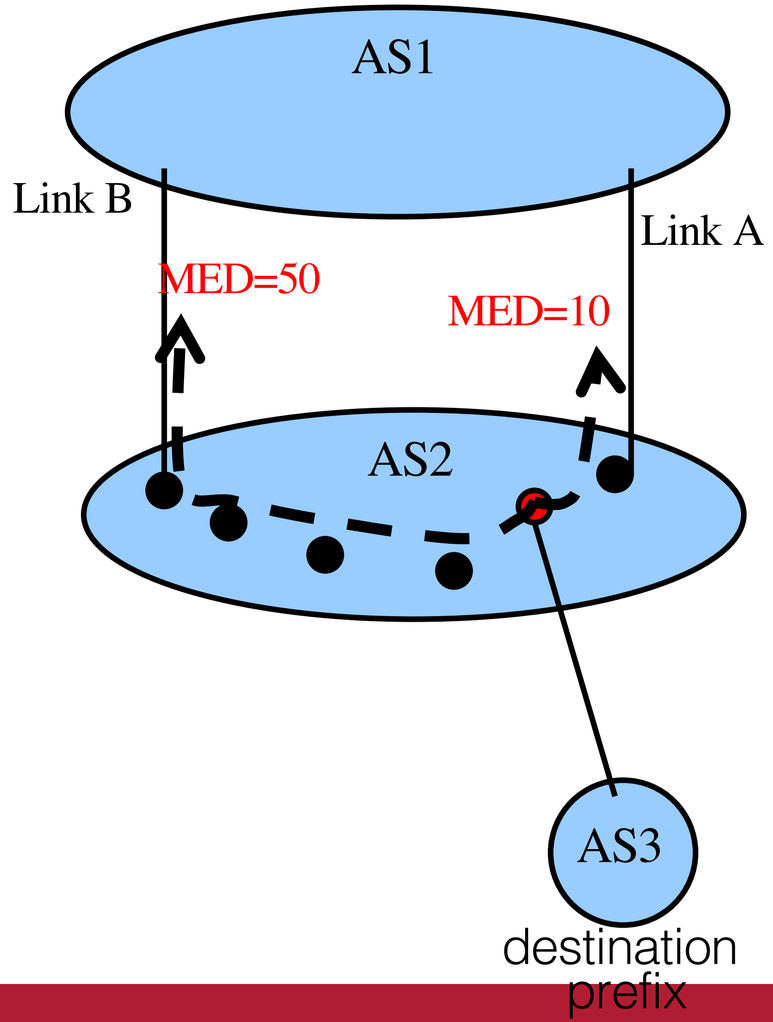


Attributes (4): ORIGIN

- Records who originated the announcement
- Local to an AS
- Options:
  - “e” : from eBGP
  - “i” : from iBGP
  - “?” : Incomplete; often used for static routes
- Typically: e > i > ?

Attributes (5) : MED

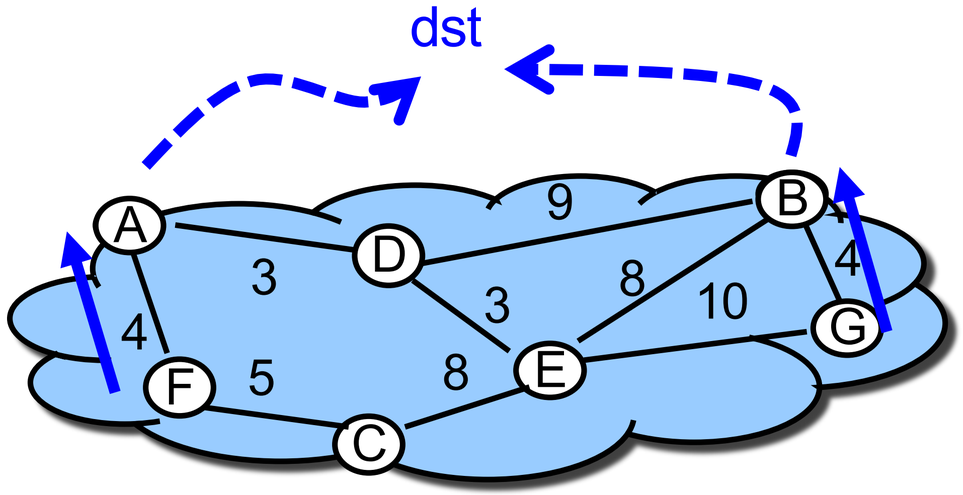
- “Multi-Exit Discriminator”
- Used when ASes are interconnected via 2 or more links to specify how close a prefix is to the link it is announced on
- Lower is better
- AS announcing prefix sets MED (AS2 in picture)
- AS receiving prefix (optionally!) uses MED to select link (AS1 in pic.)



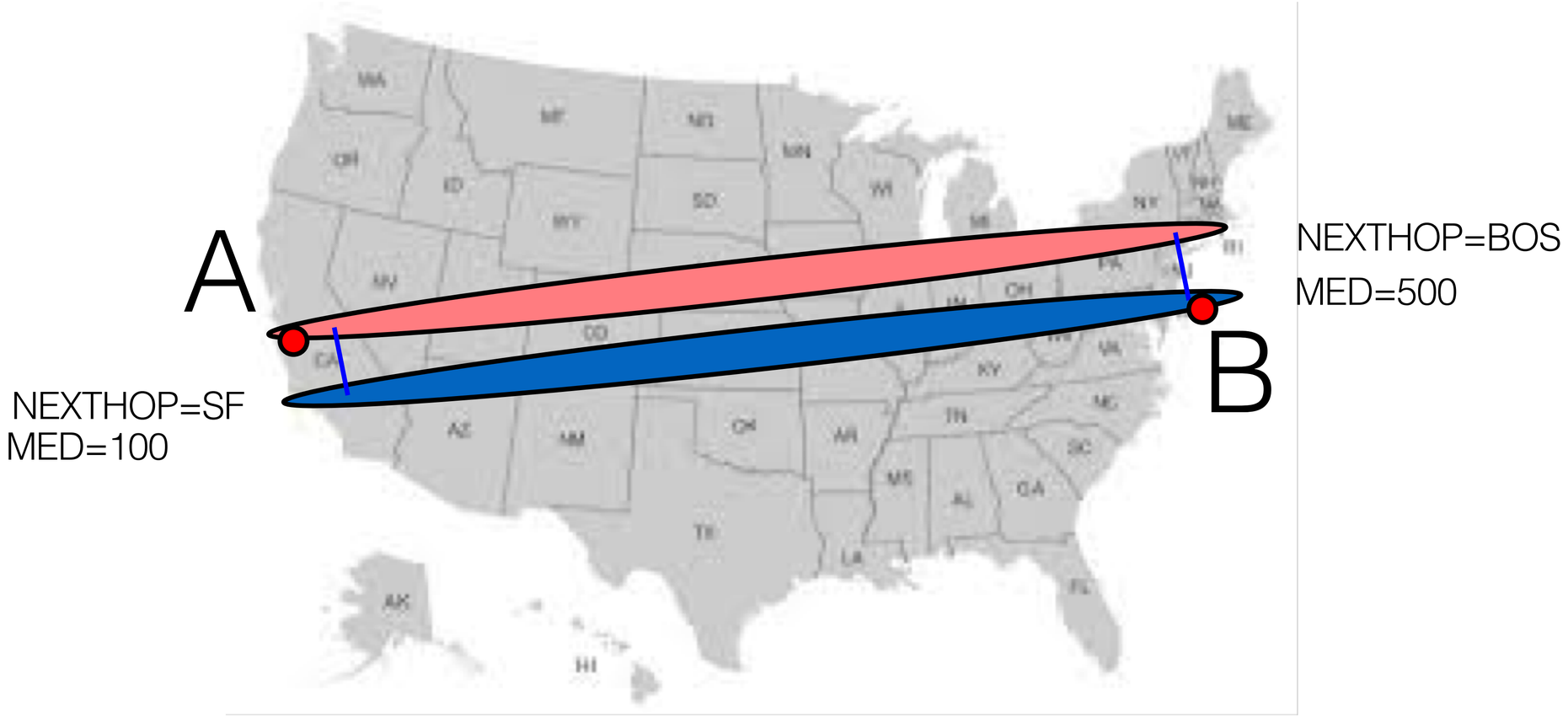
Attributes (6): IGP cost

- Used for hot-potato routing
  - Each router selects the closest egress point based on the path cost in intra-domain protocol

hot potato



IGP may conflict with MED

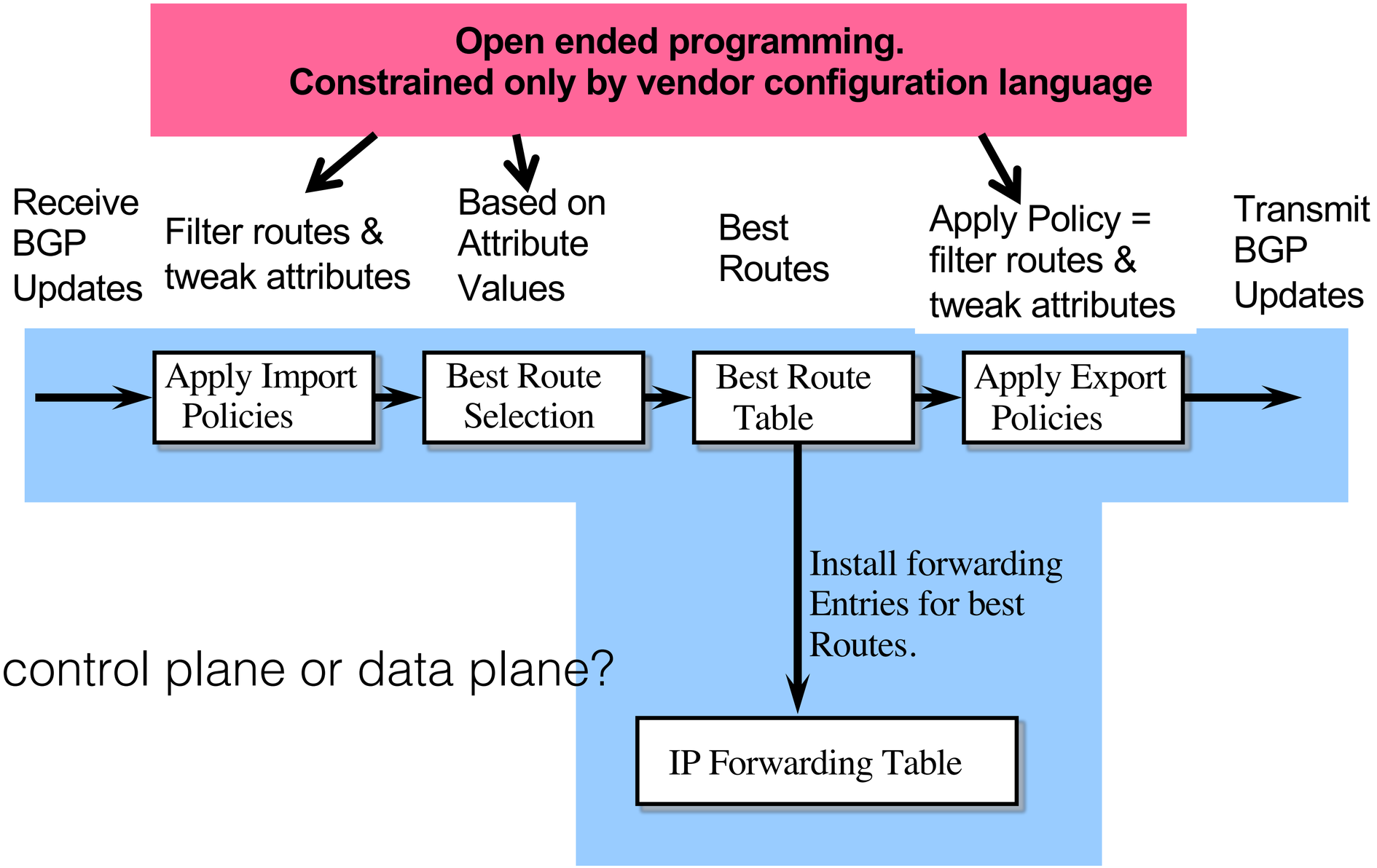


Using Attributes

- Rules for route selection in priority order

Priority	Rule	Remarks
1	LOCAL PREF	Pick highest LOCAL PREF
2	ASPATH	Pick shortest ASPATH length
3	MED	Lowest MED preferred
4	iBGP path	Lowest IGP cost to next hop (egress router)
5	Router ID	Smallest router ID (IP address) as tie-breaker

BGP UPDATE Processing



Is this control plane or data plane?

Issues with BGP

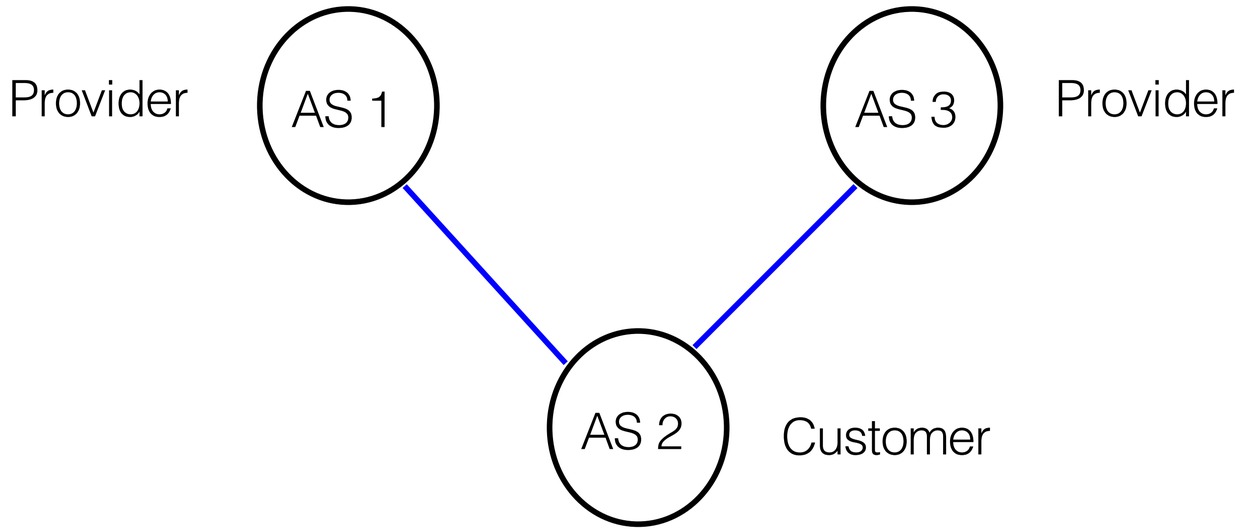
- Reachability
- Security
- Convergence
- Performance

Thoughts on why these might be difficult?

Reachability

- In normal routing, if graph is connected then reachability is assured

- With policy routing, this does not always hold



Security

- An AS can claim to serve a prefix that they actually don't have a route to (blackholing traffic)
  - Problem not specific to policy or path vector
  - Important because of AS autonomy
  - *Fixable: make ASes "prove" they have a path*

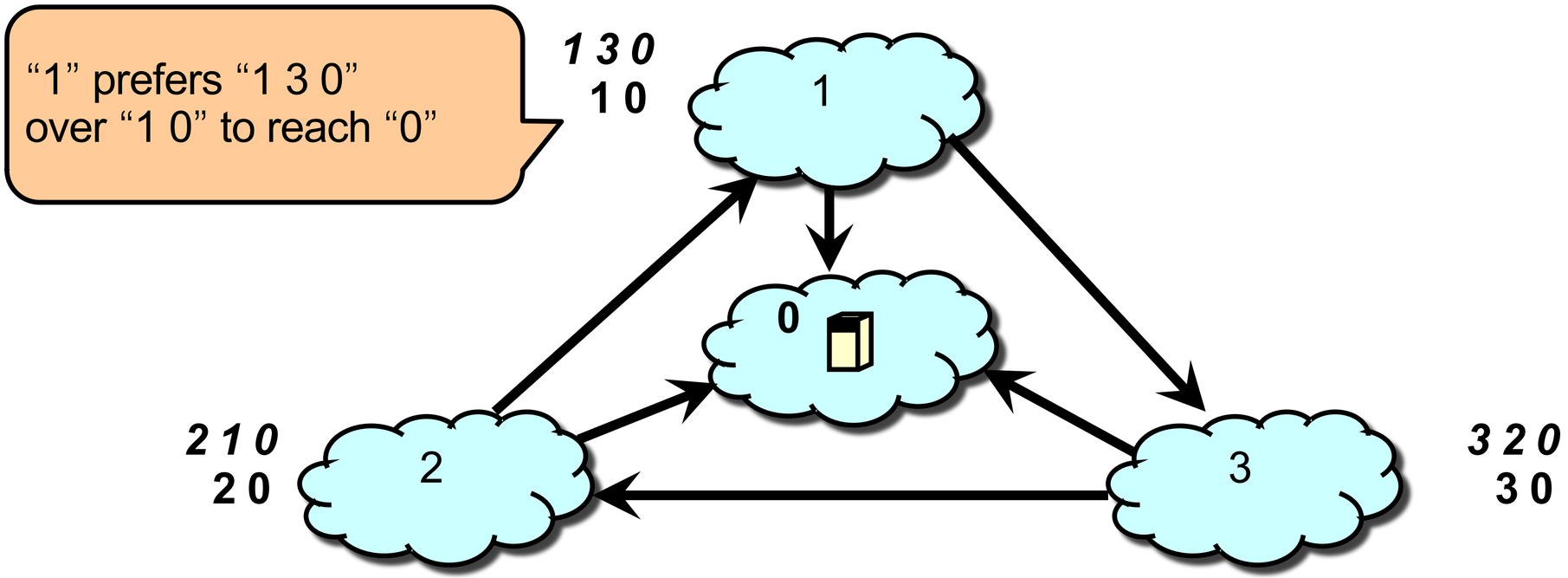
- Note: AS can also have incentive to forward packets along a route different from what is advertised
  - Tell customers about fictitious short path…
  - Much harder to fix!

Convergence

- Result: If all AS policies follow “Gao-Rexford” rules, BGP is guaranteed to converge (safety)

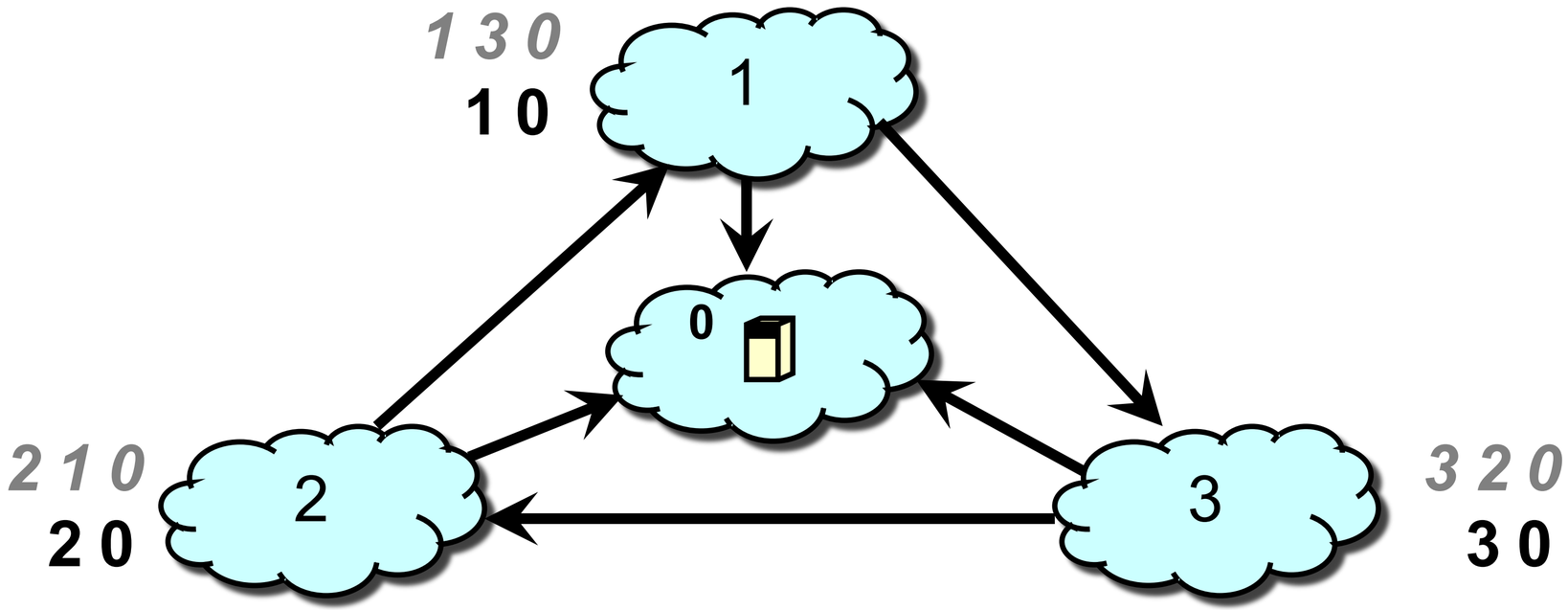
- For arbitrary policies, BGP may fail to converge!

Example of Policy Oscillation



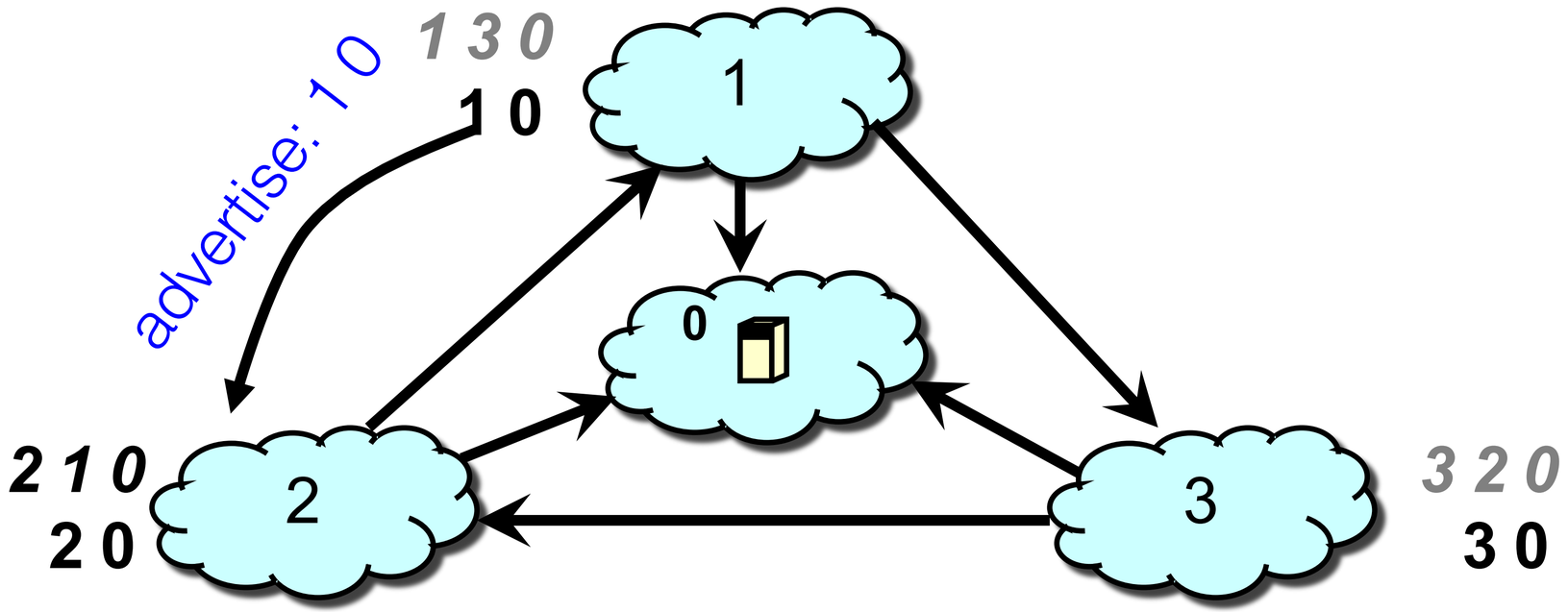
Step-by-Step of Policy Oscillation

Initially: nodes 1, 2, 3 know only shortest path to 0

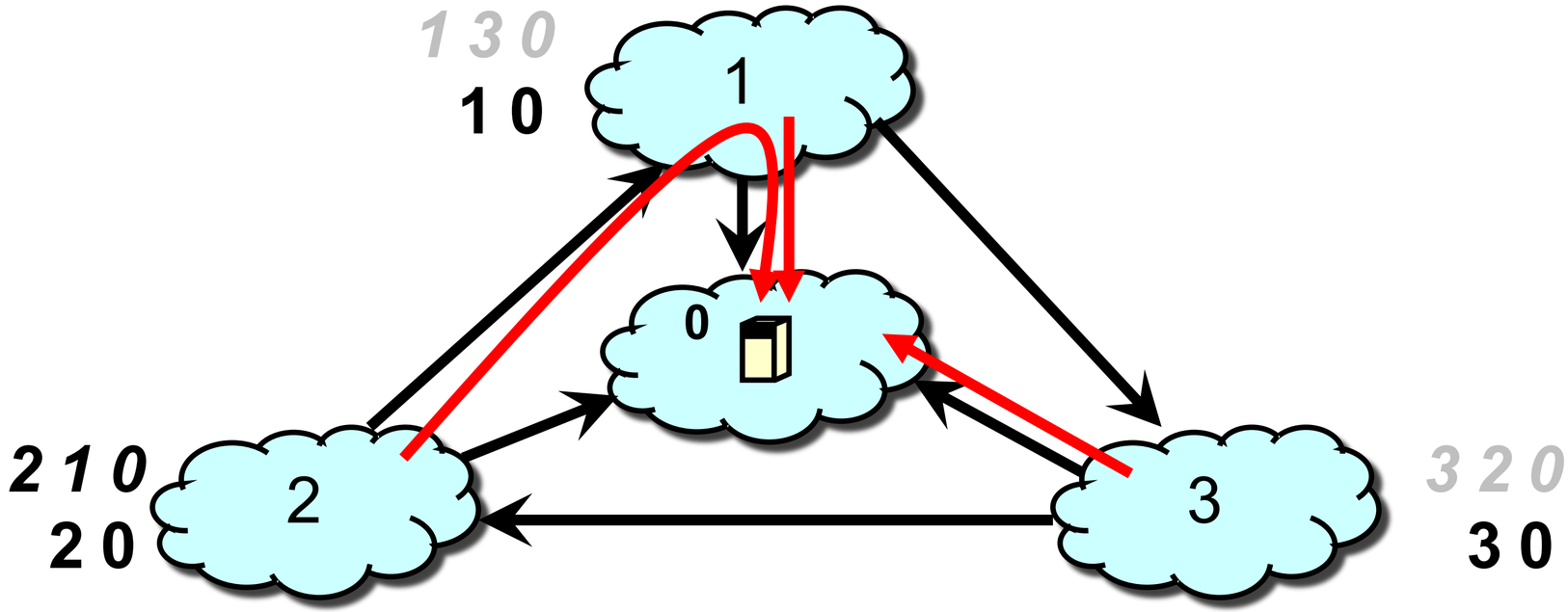


Step-by-Step of Policy Oscillation

1 **advertises** its path 1 0 to 2

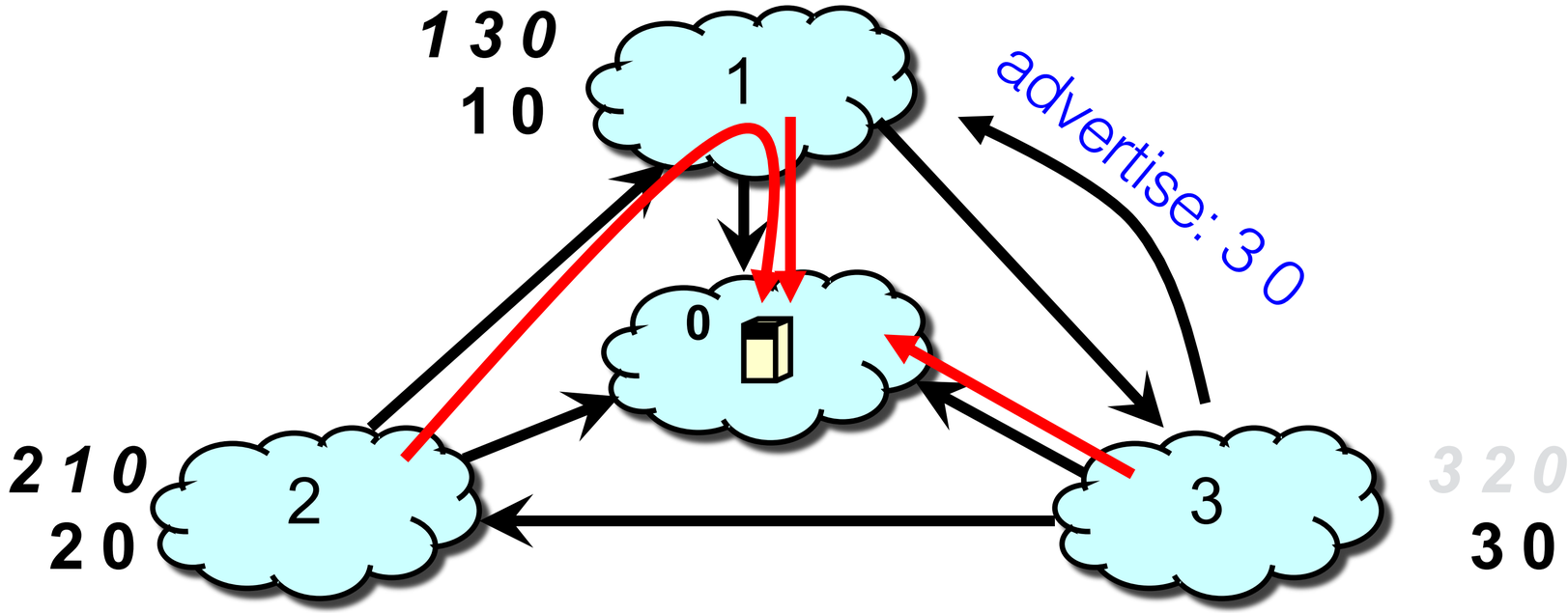


Step-by-Step of Policy Oscillation

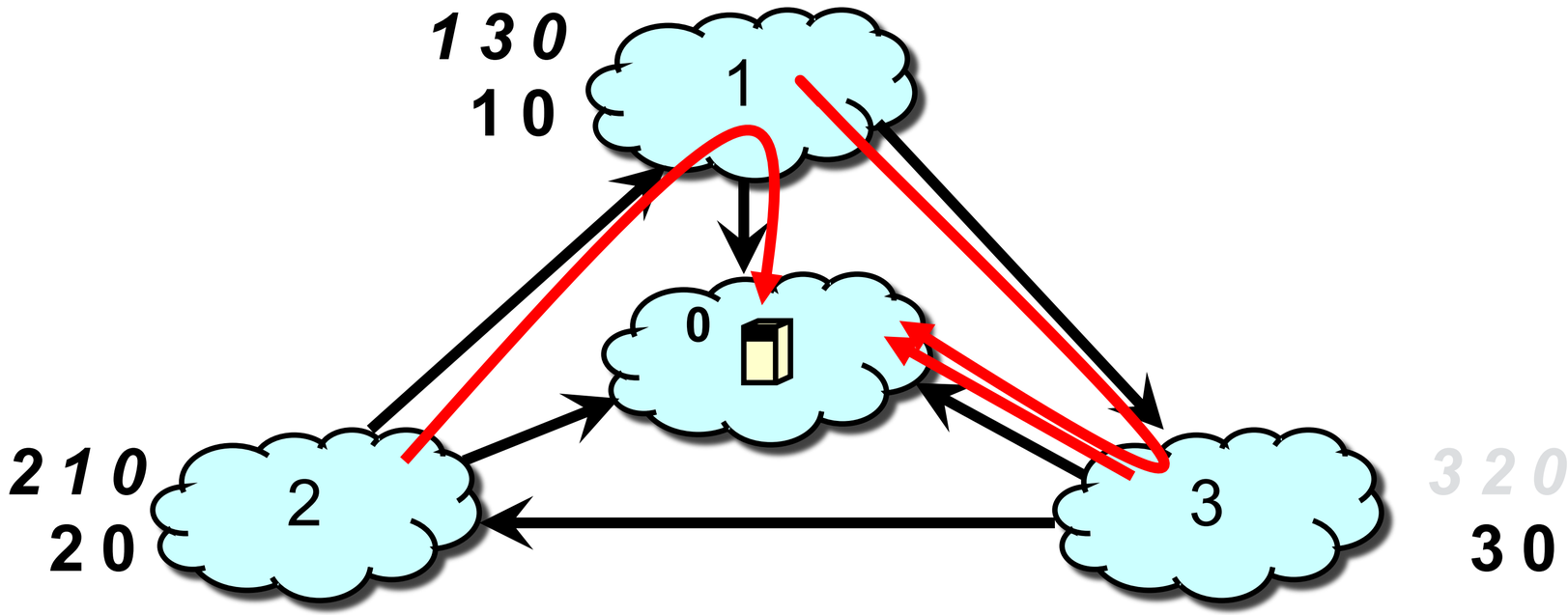


Step-by-Step of Policy Oscillation

3 **advertises** its path 3 0 to 1

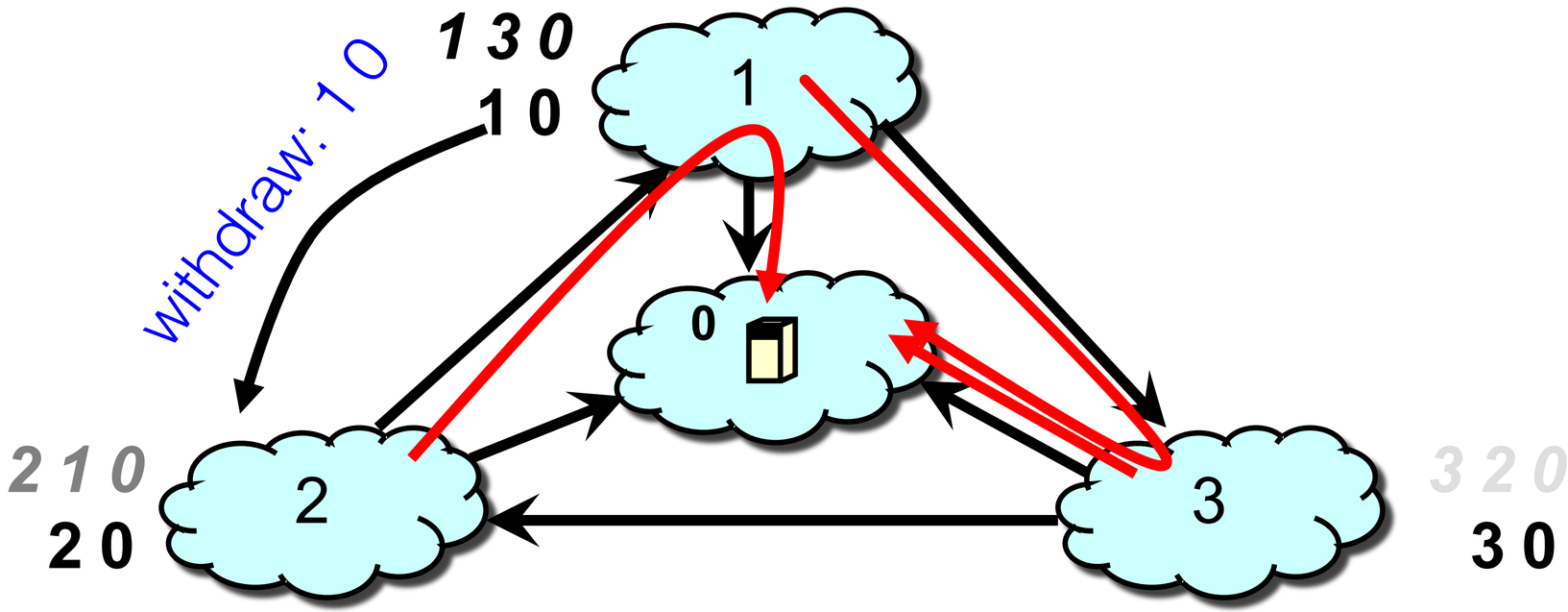


Step-by-Step of Policy Oscillation

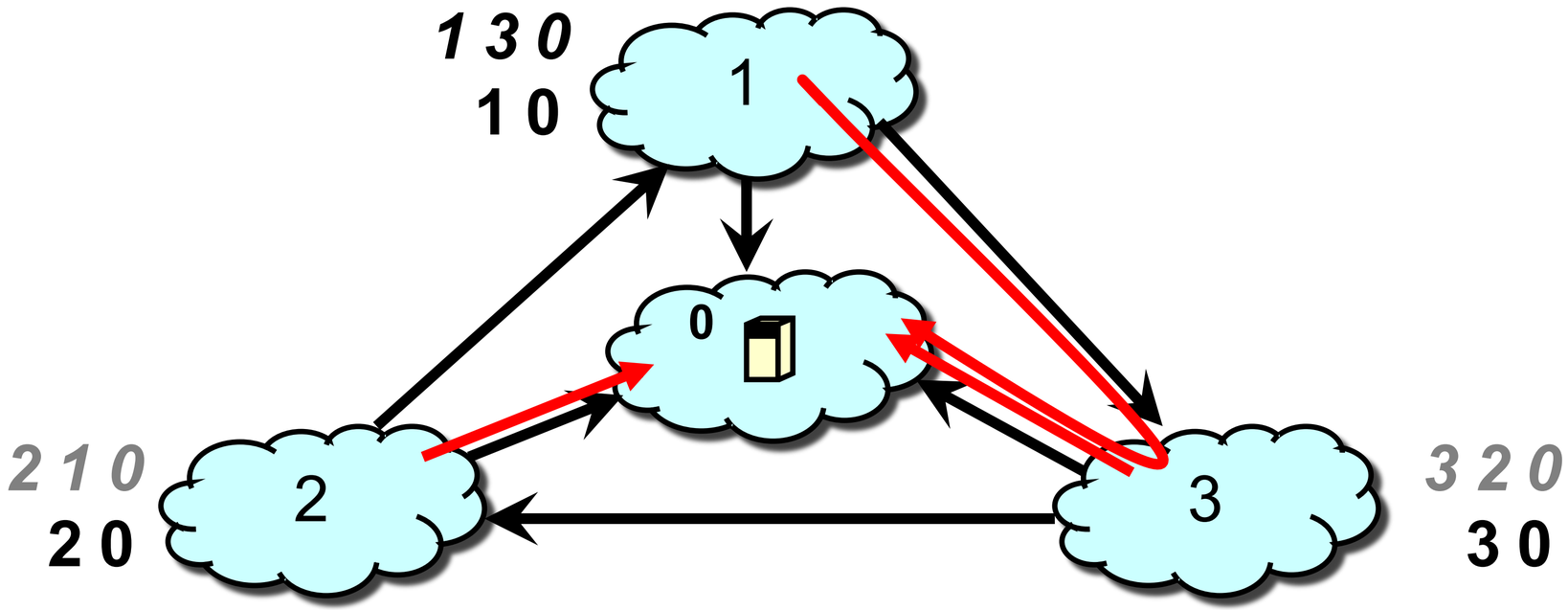


Step-by-Step of Policy Oscillation

1 **withdraws** its path 1 0 from 2

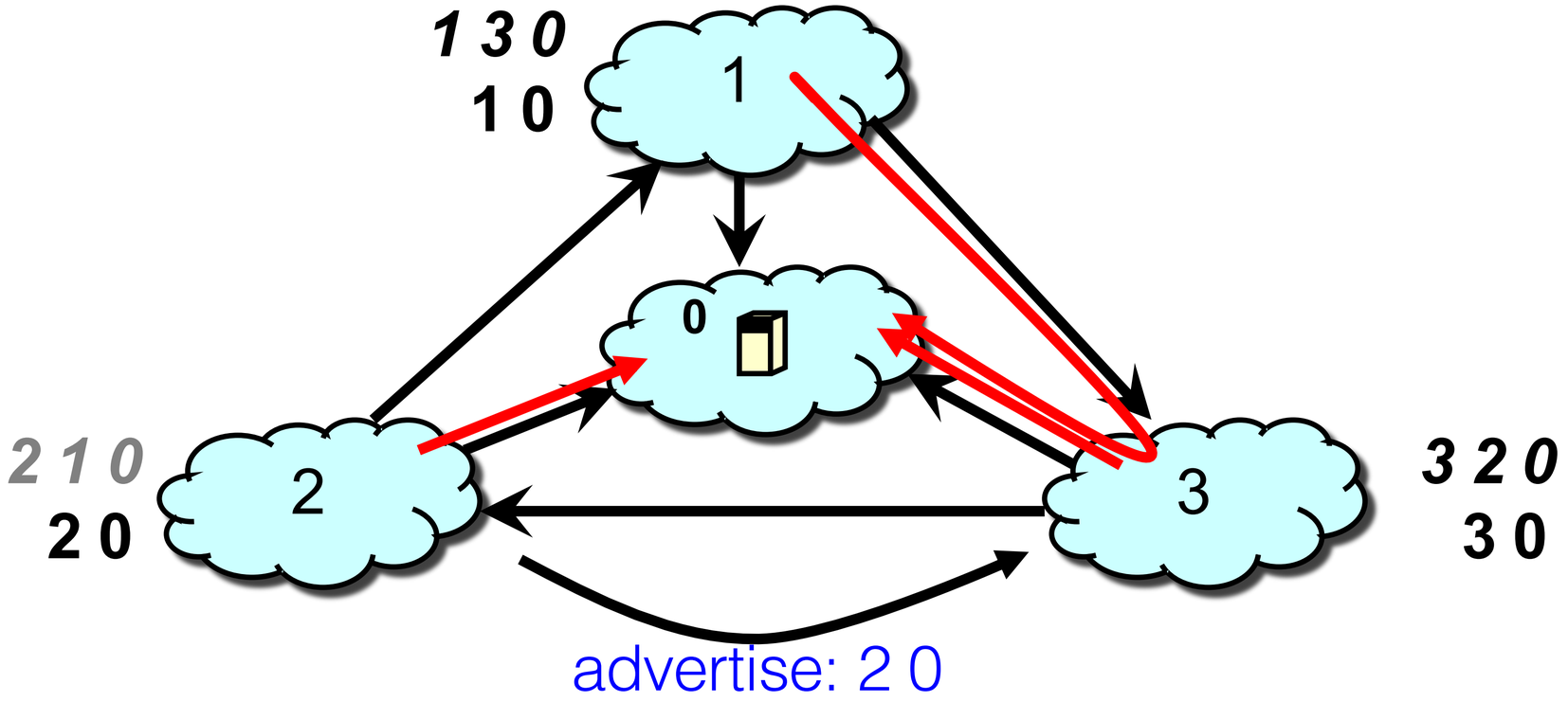


Step-by-Step of Policy Oscillation

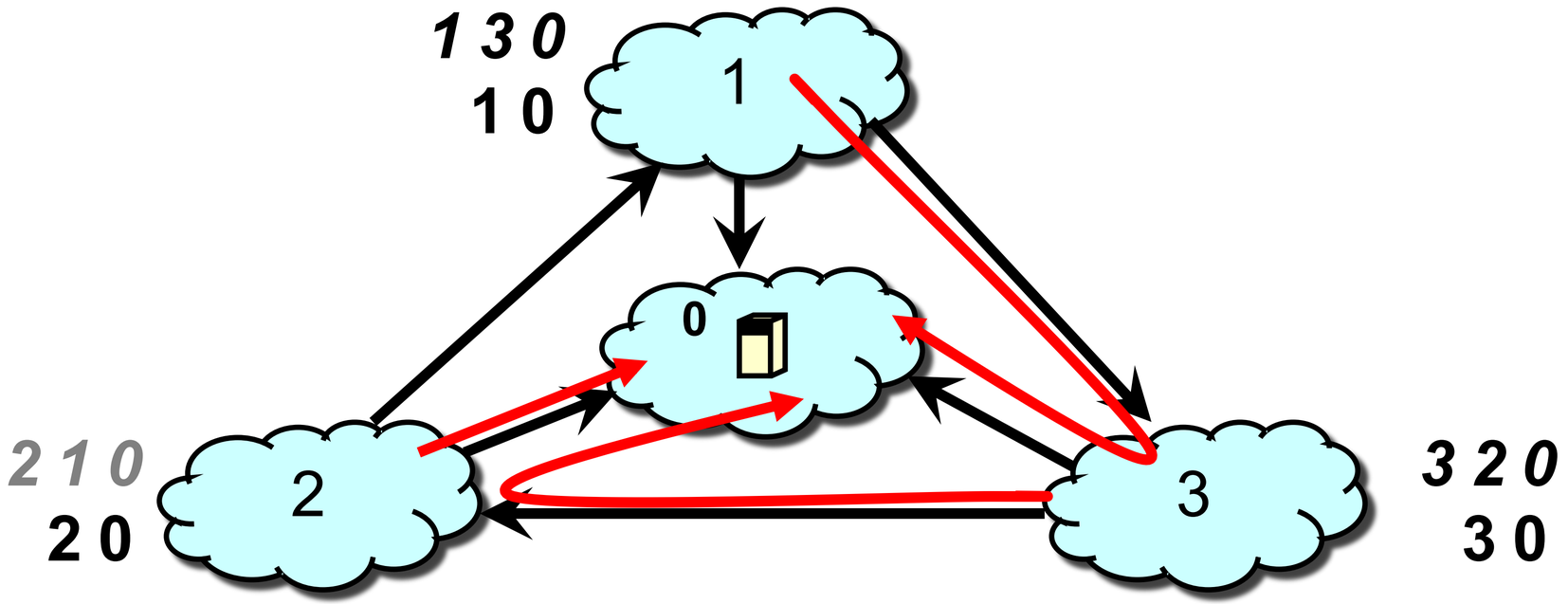


Step-by-Step of Policy Oscillation

2 **advertises** its path 2 0 to 3

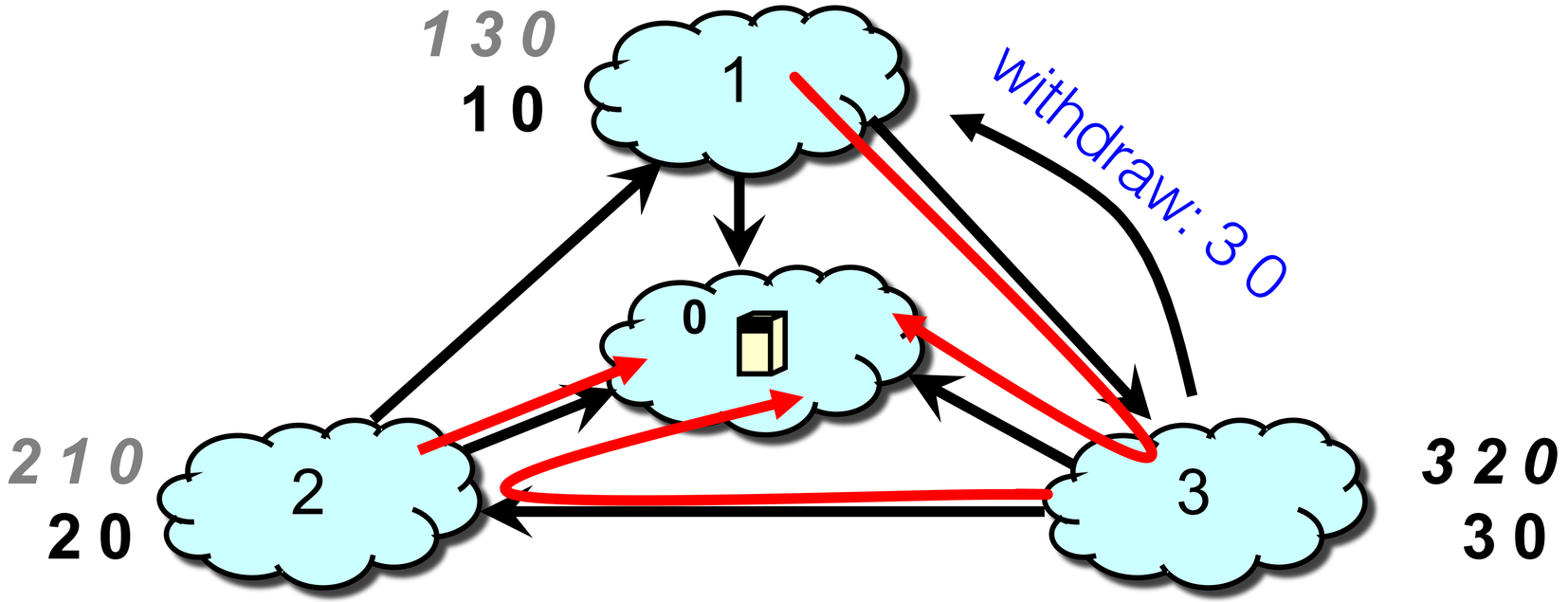


Step-by-Step of Policy Oscillation

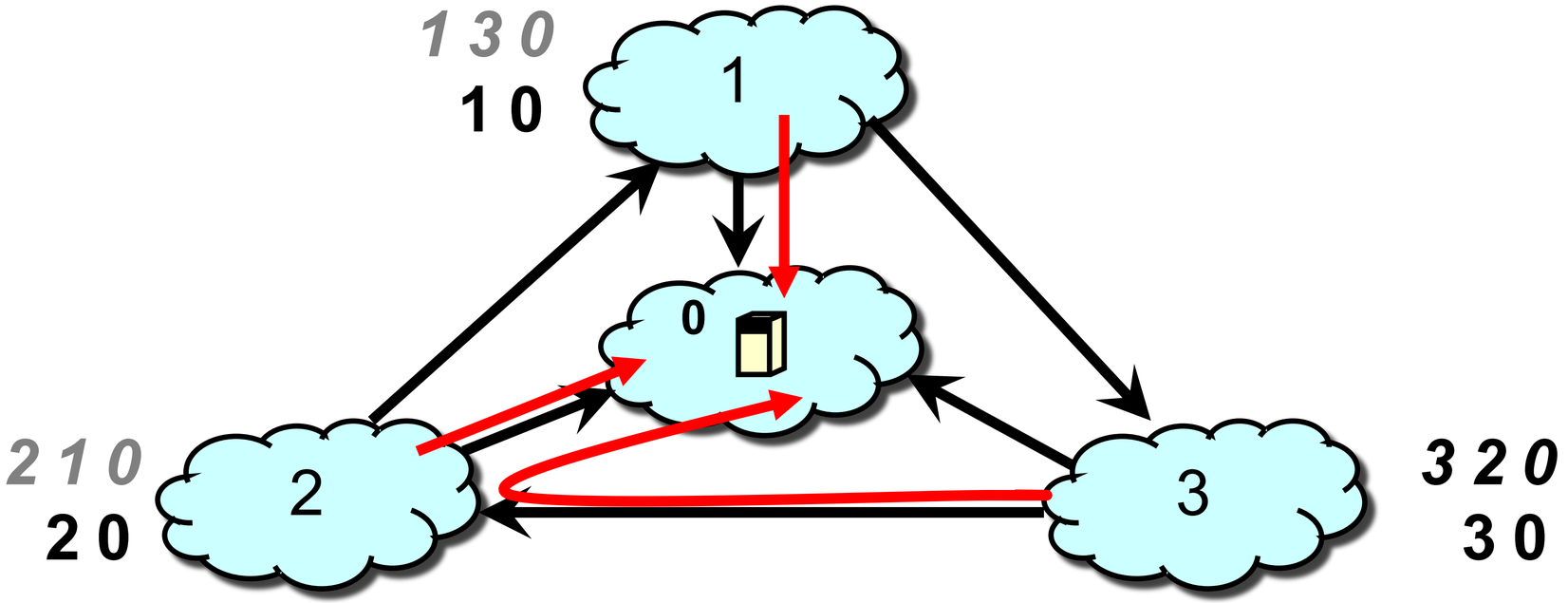


Step-by-Step of Policy Oscillation

3 **withdraws** its path 3 0 from 1

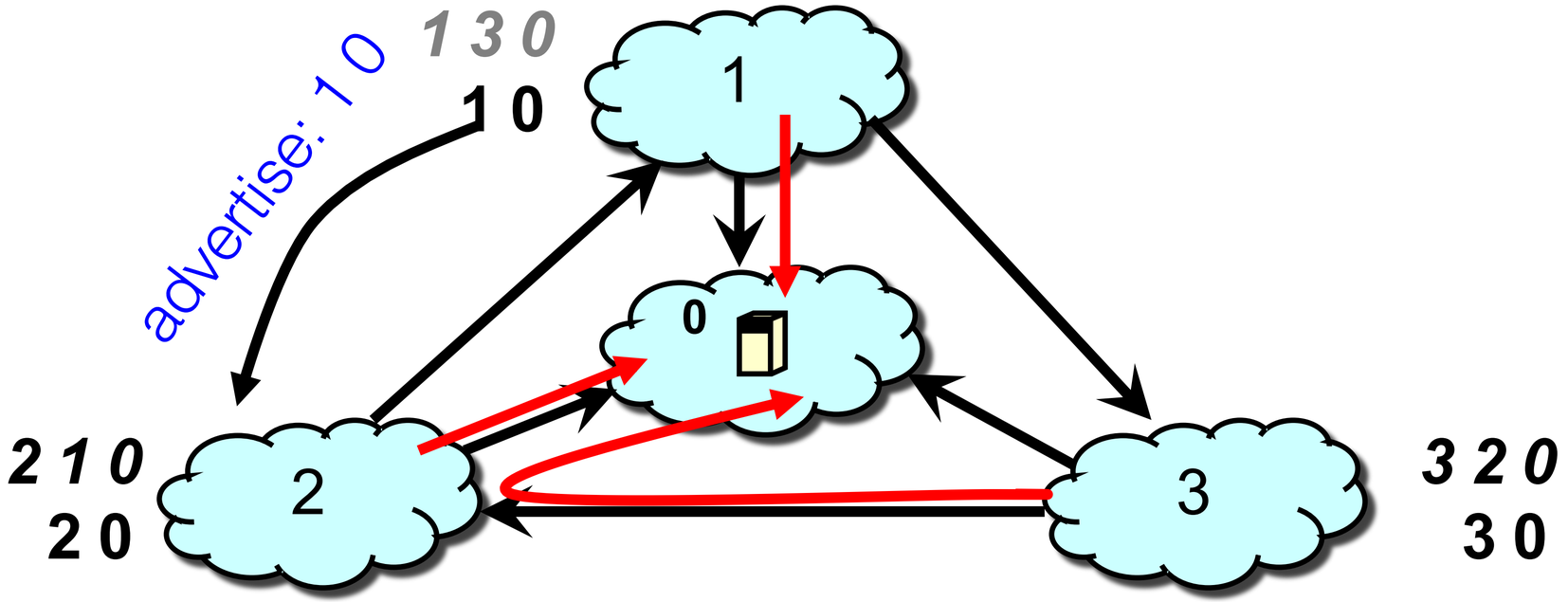


Step-by-Step of Policy Oscillation

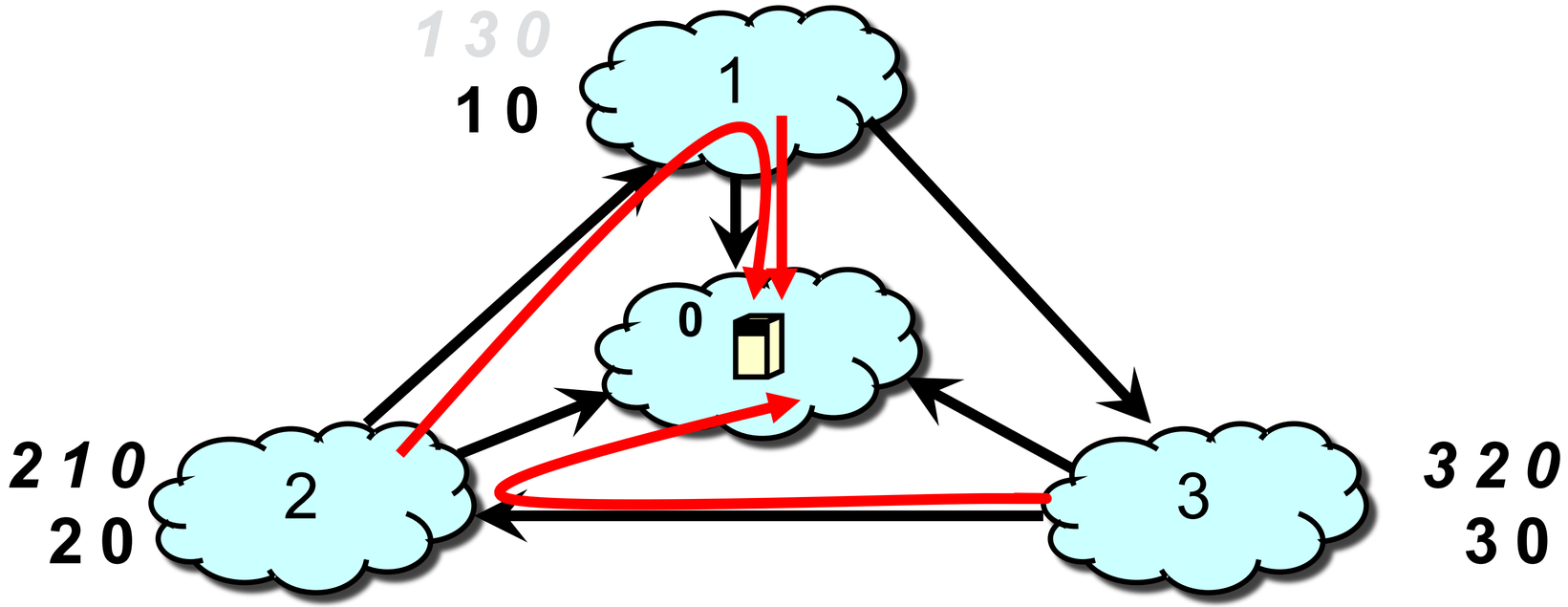


Step-by-Step of Policy Oscillation

1 **advertises** its path 1 0 to 2

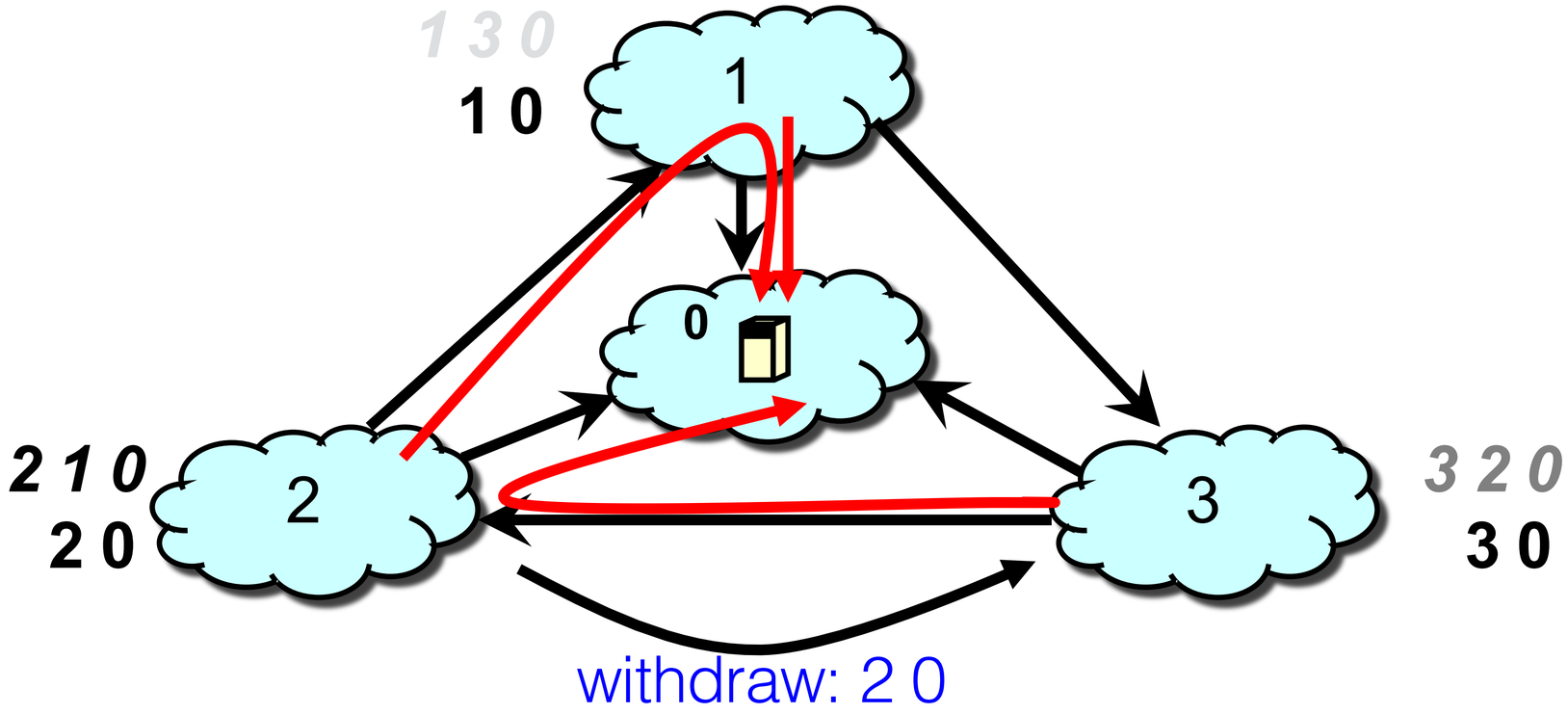


Step-by-Step of Policy Oscillation

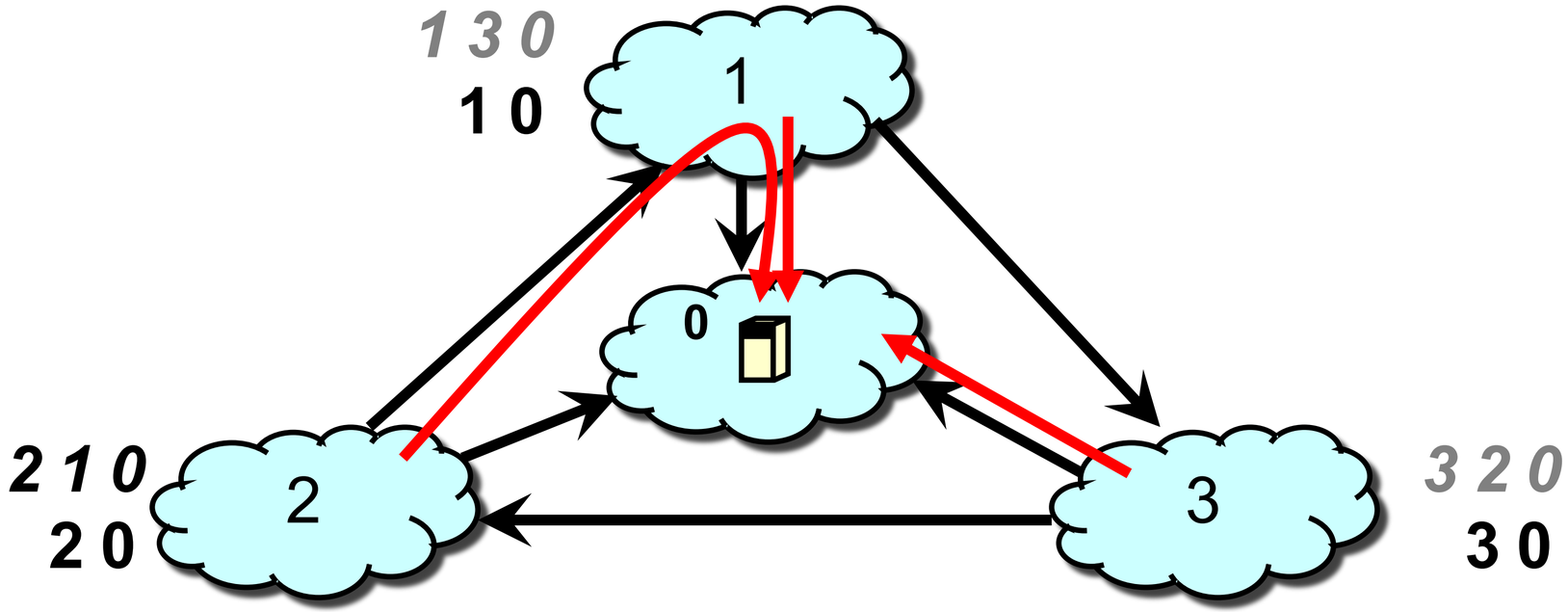


Step-by-Step of Policy Oscillation

2 **withdraws** its path 2 0 from 3



Step-by-Step of Policy Oscillation



We are back to where we started!

Convergence

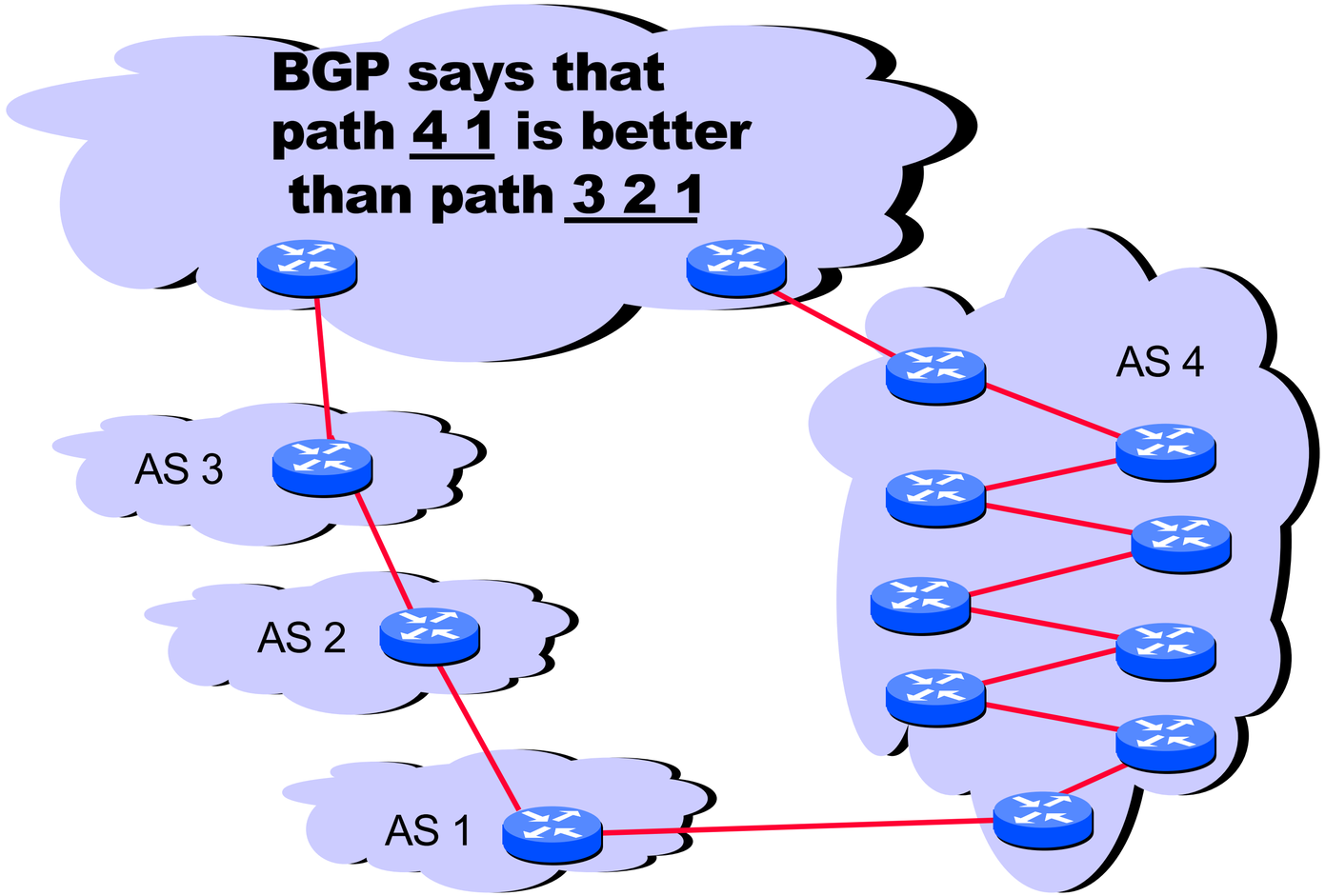
- Result: If all AS policies follow “Gao-Rexford” rules, BGP is guaranteed to converge (safety)
- For arbitrary policies, BGP may fail to converge!
- Should this trouble us?

Performance Nonissues

- Internal routing (non)
  - Domains typically use “hot potato” routing
  - Not always optimal, but economically expedient
- Policy not about performance (non)
  - So policy-chosen paths aren’t shortest
- Choosing among policy-compliant paths (non)
  - Fewest AS hops has little to do with actual delay
  - 20% of paths inflated by at least 5 router hops

Performance (example)

- AS path length can be misleading
  - An AS may have many router-level hops



Real Performance Issue: Slow convergence

- BGP outages are biggest source of Internet problems
- Labovitz *et al. SIGCOMM'97*
  - 10% of routes available less than 95% of time
  - Less than 35% of routes available 99.99% of the time
- Labovitz *et al. SIGCOMM 2000*
  - 40% of path outages take 30+ minutes to repair
- But most popular paths are very stable

BGP Misconfigurations

- BGP protocol is both bloated and underspecified
  - lots of leeway in how to set and interpret attribute values, route selection rules, *etc.*
  - necessary to allow autonomy, diverse policies
  - but also gives operators plenty of rope

- Much of this configuration is manual and *ad hoc*

- And the core abstraction is fundamentally flawed
  - per-router configuration to effect AS-wide policy
  - now strong industry interest in changing this! [later: SDN]

Important Concepts

- Wide area Internet structure and routing driven by economic considerations
  - Customer, providers and peers
- BGP designed to:
  - Provide hierarchy that allows scalability
  - Allow enforcement of policies related to structure
- Mechanisms
  - Path vector – scalable, hides structure from neighbors, detects loops quickly