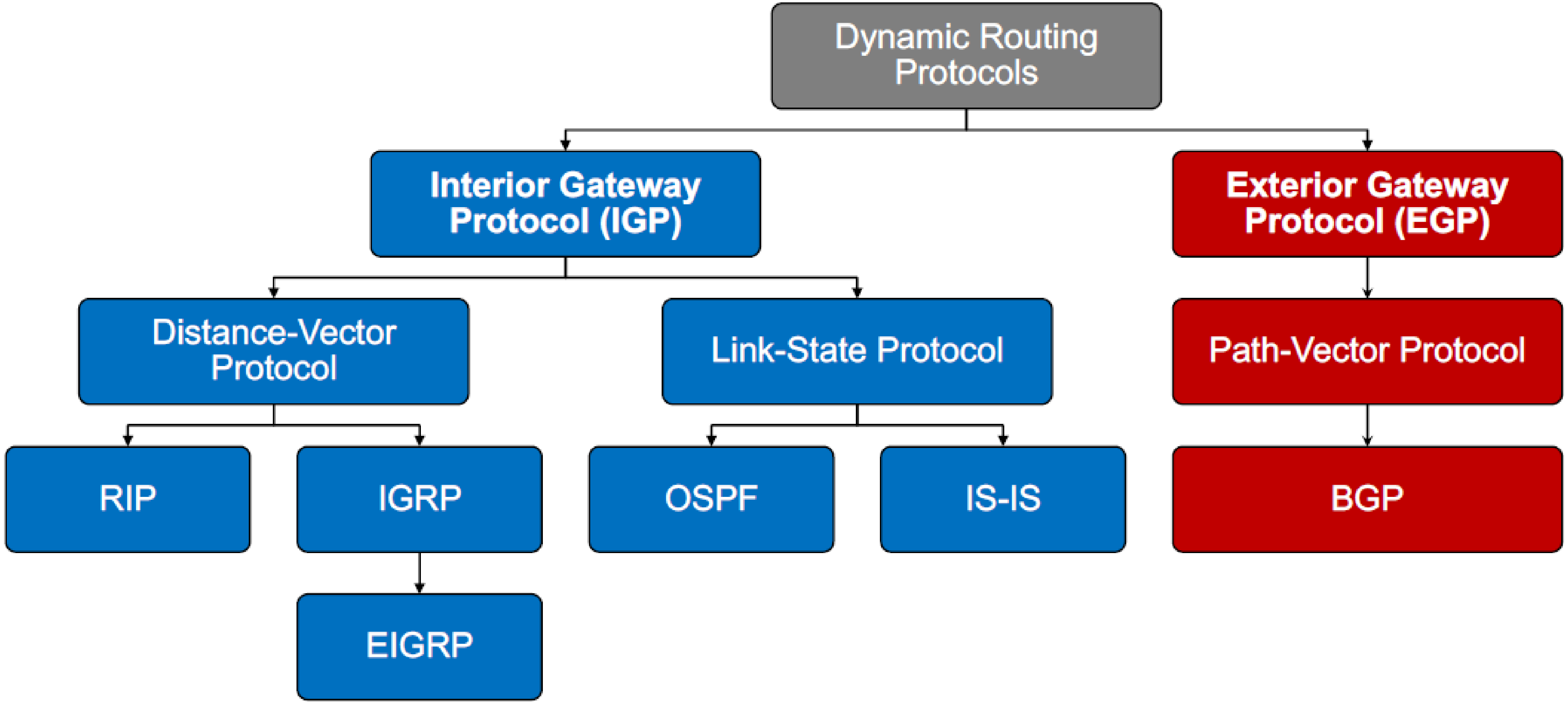
BGP

Mix of slides from: BRKENT-1179, APNIC, Kurose, Wu

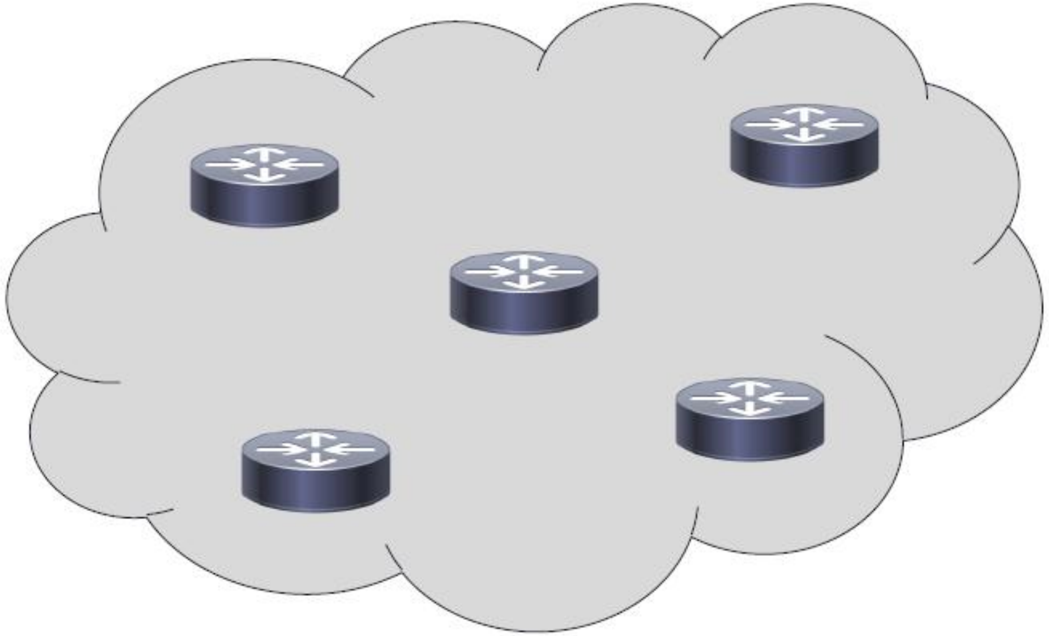
Dynamic Routing Protocol


Dynamic Routing Protocols
Interior Gateway Protocol (IGP)
Exterior Gateway Protocol (EGP)
Distance-Vector Protocol
Link-State Protocol
Path-Vector Protocol
RIP
IGRP
OSPF
IS-IS
BGP
EIGRP

Autonomous Systems & Peering

Autonomous System

- A group of one or more IP prefixes (lists of IP addresses accessible on a network) run by one or more network operators that maintain a single, clearly-defined routing policy.

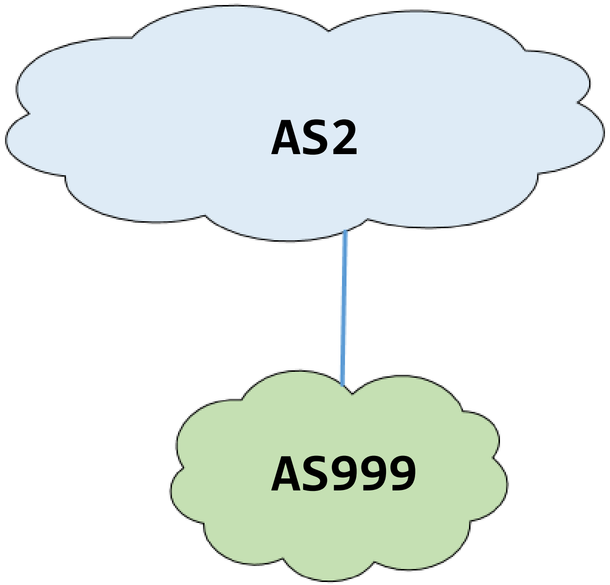


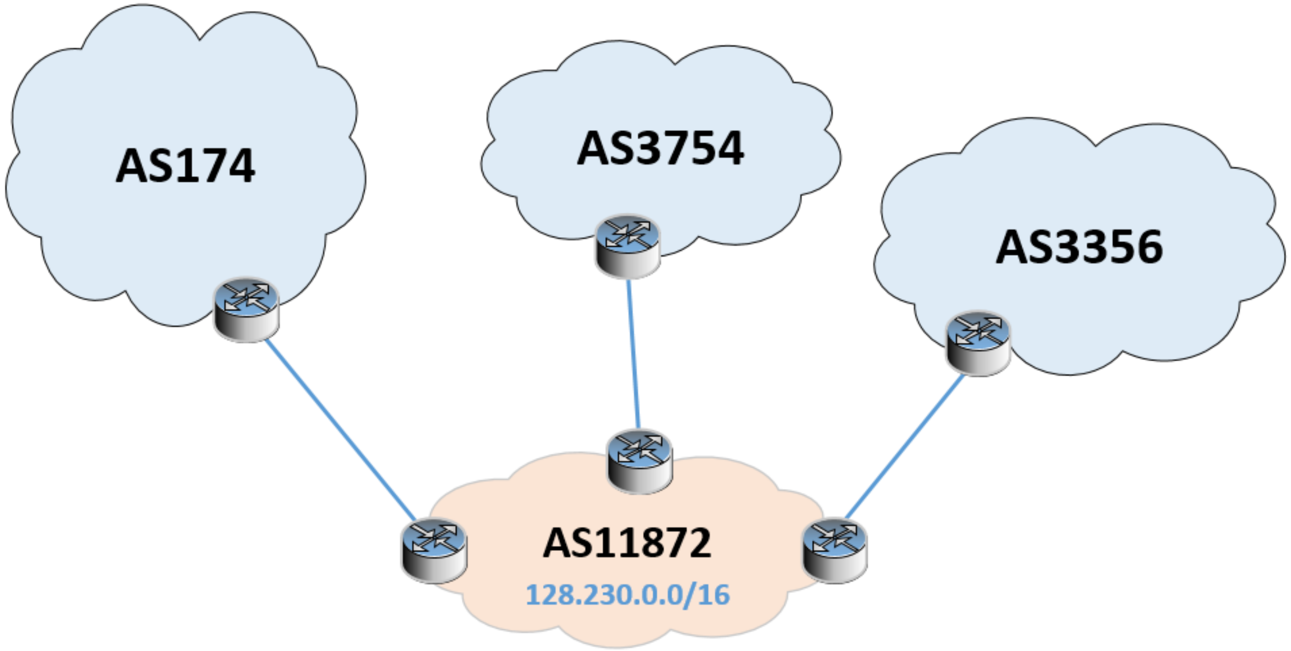
Autonomous System (AS)

- Types
  - Stub AS
  - Transit AS
- AS Number
  - Original scheme: 16 bits (BGP 2, RFC1105 (1989))
    - 0 to 65535 Private range 64512 to 65534
  - Extended number: 32 bits (BGP 4, RFC4893 (2007))
    - 0 to 4294967295 Additional private range 4200000000 to 4294967294

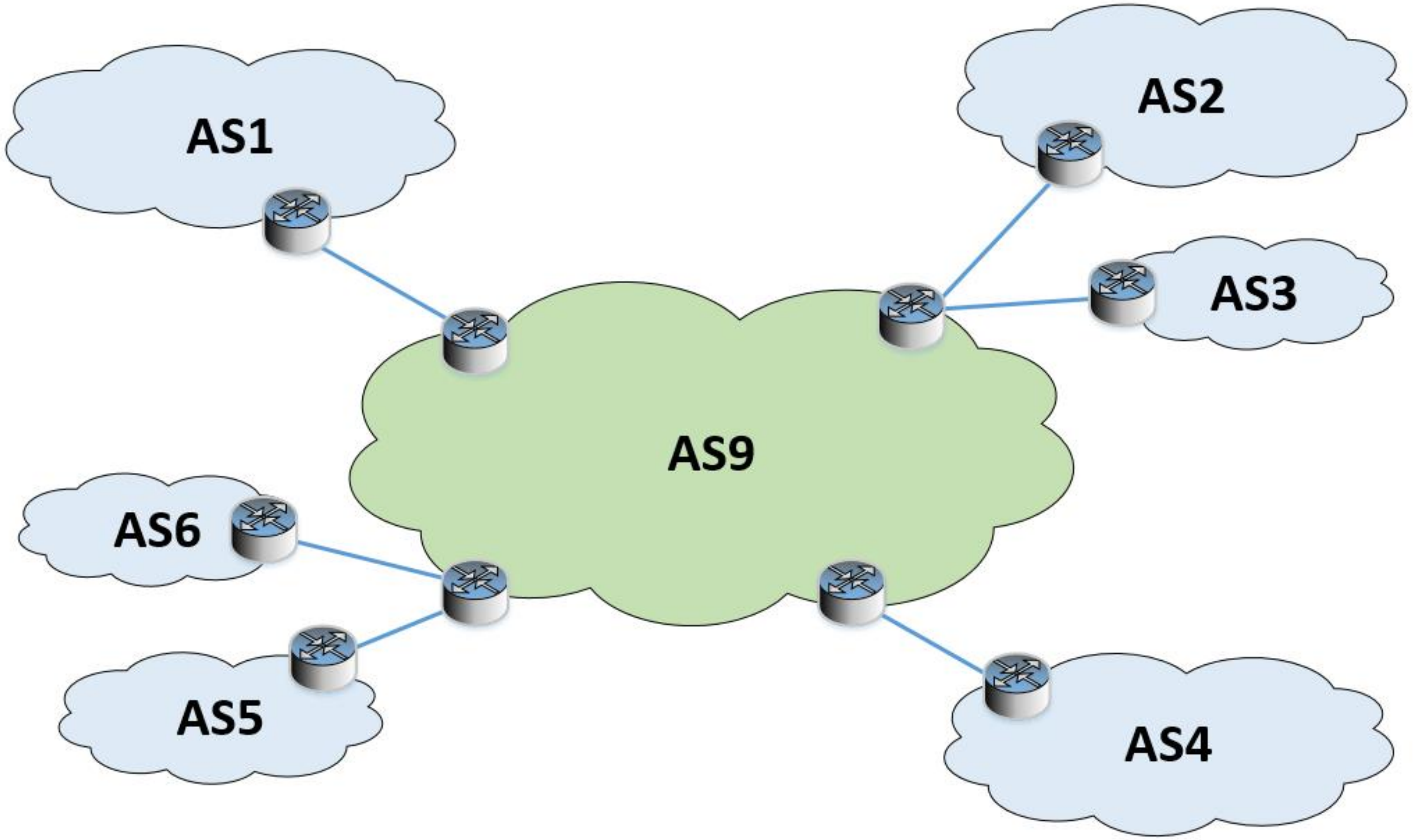
Stub AS

End customers; do not provide transit to others





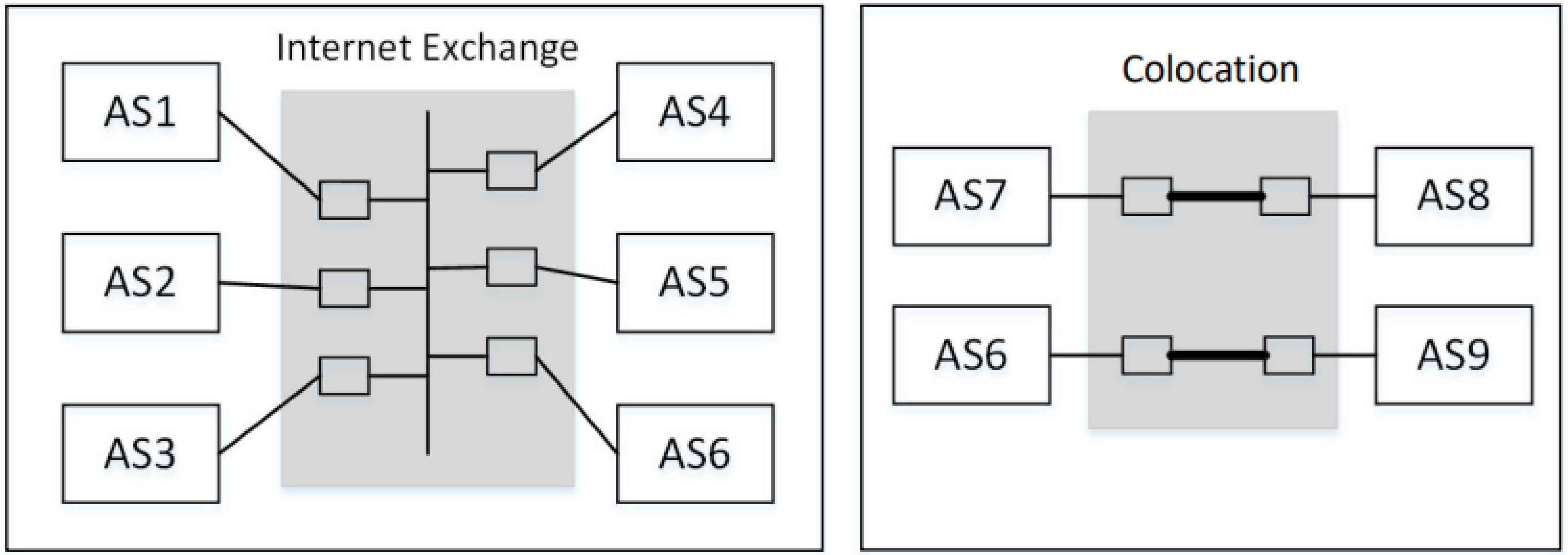
Transit AS



Provide transit to others

Peering

- Autonomous Systems peer with one another in an Internet Exchange or a data center

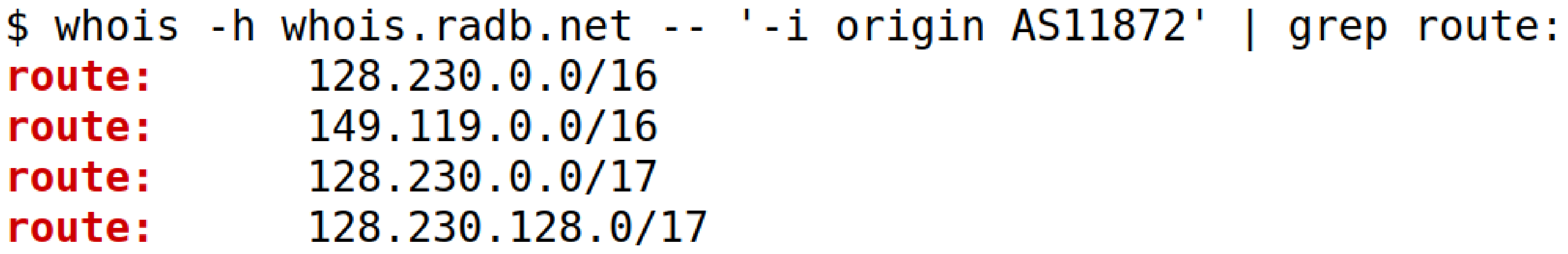


(A) Public peering

(B) Private peering

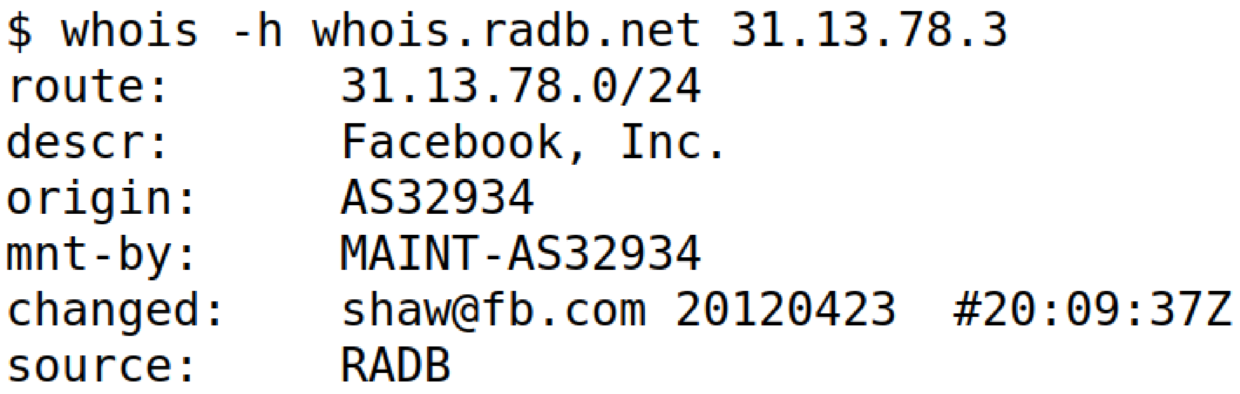
ASN and IP

- Find ASN from IP Address



```
$ whois -h whois.radb.net -- '-i origin AS11872' | grep route:
route:      128.230.0.0/16
route:      149.119.0.0/16
route:      128.230.0.0/17
route:      128.230.128.0/17
```

- Find IP prefix from ASN



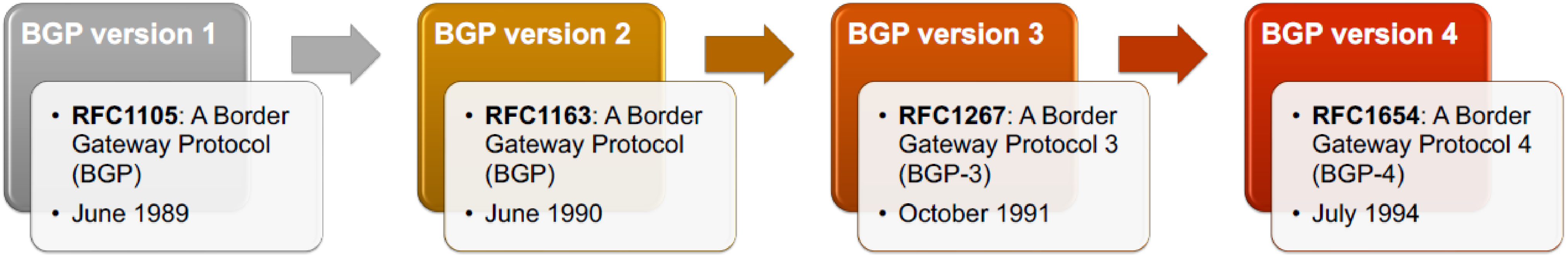
```
$ whois -h whois.radb.net 31.13.78.3
route:      31.13.78.0/24
descr:      Facebook, Inc.
origin:     AS32934
mnt-by:     MAINT-AS32934
changed:    shaw@fb.com 20120423  #20:09:37Z
source:     RADB
```

Border Gateway Protocol

- Border Gateway Protocol Large scale, robust and stable routing protocol designed to operate between autonomous systems
- Based on TCP, listens on port 179
- Fundamentally a distance vector protocol
- Does not have the concept of a simple metric
- Strong control over advertised routes and their attributes

BGP Versions

- BGP was first described in 1989 and has been in use on the Internet since 1994
- There are four versions of BGP:



BGP version 1

- **RFC1105**: A Border Gateway Protocol (BGP)
- June 1989

BGP version 2

- **RFC1163**: A Border Gateway Protocol (BGP)
- June 1990

BGP version 3

- **RFC1267**: A Border Gateway Protocol 3 (BGP-3)
- October 1991

BGP version 4

- **RFC1654**: A Border Gateway Protocol 4 (BGP-4)
- July 1994

BGP Stability Considerations

- Network events burst.
- Fast reaction: better convergence, more churn.
- Delayed reaction: more stability, slower convergence.

- BGP prioritizes stability.
- Delays updates to reduce churn, combines changes.
- Sends only incremental updates.

- "Churn" reflects the percentage of customers who discontinue their use of a business's products or services over a certain period of time.

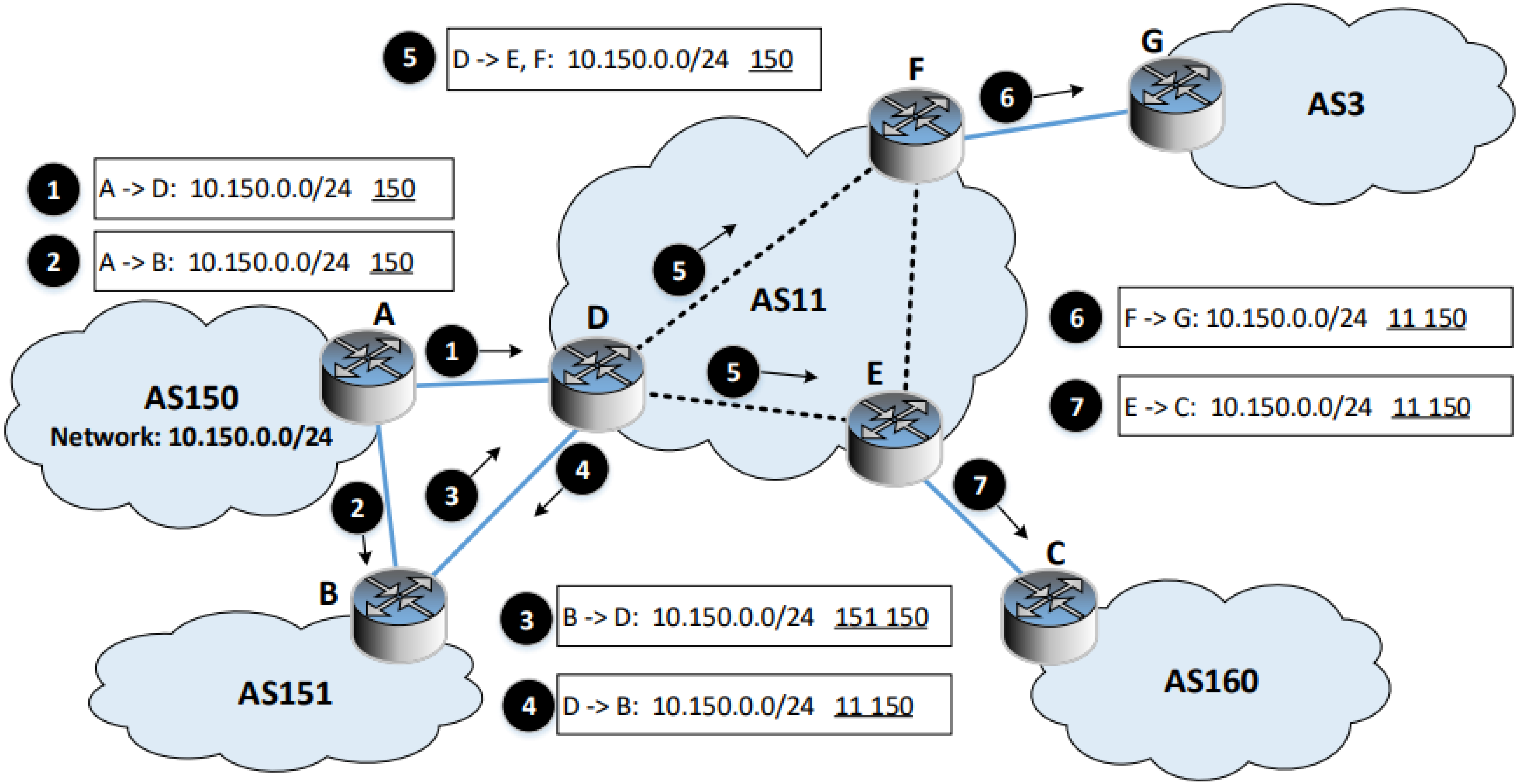
How BGP Works

Internet inter-AS routing: BGP

- BGP provides each AS a means to:
  - eBGP (External BGP): for BGPs from different AS
    - obtain subnet reachability information from neighboring ASes
  - iBGP (Internal BGP): for BGPs from the same AS
    - propagate reachability information to all AS-internal routers.

  - determine "good" routes to other networks based on reachability information and *policy*

IP Prefix Announcement



1 A -> D: 10.150.0.0/24 150

2 A -> B: 10.150.0.0/24 150

3 B -> D: 10.150.0.0/24 151 150

4 D -> B: 10.150.0.0/24 11 150

5 D -> E, F: 10.150.0.0/24 150

6 F -> G: 10.150.0.0/24 11 150

7 E -> C: 10.150.0.0/24 11 150

BGP Message Types

- BGP runs on TCP:
    - Byte stream-oriented
    - Unicast only
    - Connection-oriented and reliable
    - Provides flow and congestion control

BGP Message Types

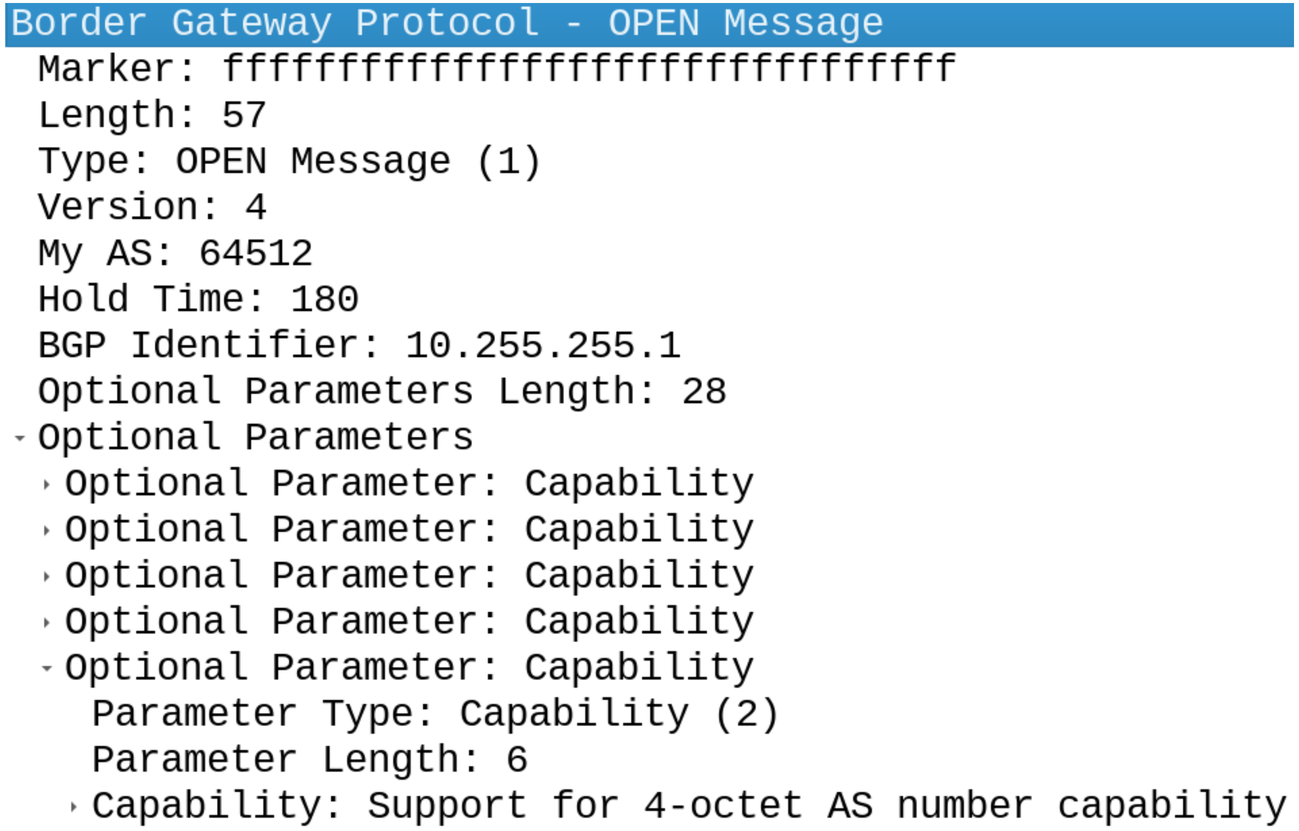
- BGPv4 uses (only) 5 message types
  - OPEN
  - UPDATE
  - NOTIFICATION
  - KEEPALIVE
  - ROUTE-REFRESH

BGP OPEN Message

- BGP speakers use OPEN to advertise configuration and capabilities after TCP session starts:
  - Version
  - Autonomous System Number
  - Hold Time (advertise/negotiation)
  - BGP Router ID
  - Optional Capabilities (advertise/negotiation)

- Incompatible configurations terminate peering and close TCP session.

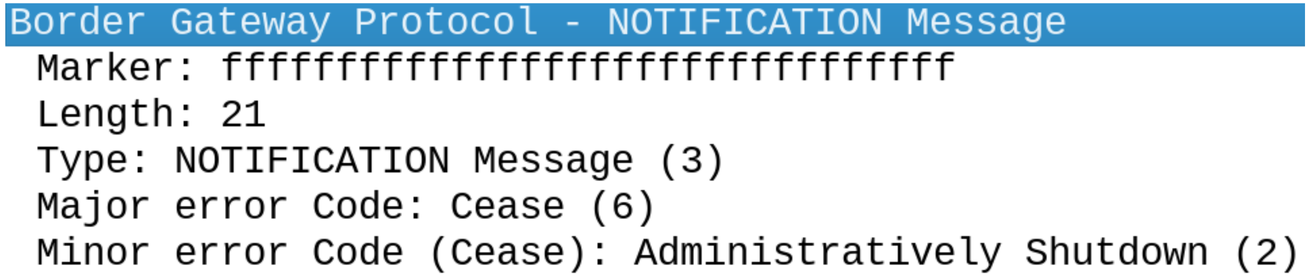
BGP OPEN Message



```
Border Gateway Protocol - OPEN Message
  Marker: ffffffffffffffffffffffffffffffff
  Length: 57
  Type: OPEN Message (1)
  Version: 4
  My AS: 64512
  Hold Time: 180
  BGP Identifier: 10.255.255.1
  Optional Parameters Length: 28
▾ Optional Parameters
  ▸ Optional Parameter: Capability
  ▸ Optional Parameter: Capability
  ▸ Optional Parameter: Capability
  ▸ Optional Parameter: Capability
  ▾ Optional Parameter: Capability
      Parameter Type: Capability (2)
      Parameter Length: 6
    ▸ Capability: Support for 4-octet AS number capability
```

BGP NOTIFICATION Message

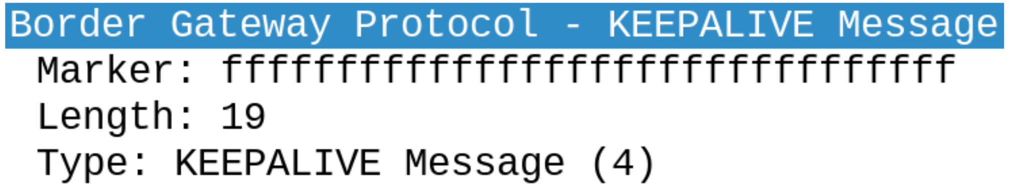
- NOTIFICATION message sent for unrecoverable conditions to terminate peering.
- Sender closes session after NOTIFICATION.
- Contents useful for diagnostics.



```
Border Gateway Protocol - NOTIFICATION Message
 Marker: ffffffffffffffffffffffffffffffff
 Length: 21
 Type: NOTIFICATION Message (3)
 Major error Code: Cease (6)
 Minor error Code (Cease): Administratively Shutdown (2)
```

BGP KEEPALIVE Message

- BGP uses KEEPALIVE instead of TCP keepalives to show liveliness.
- KEEPALIVE is sent: Immediately after an agreeable OPEN message.
- Periodically, default is one-third of Hold Time.

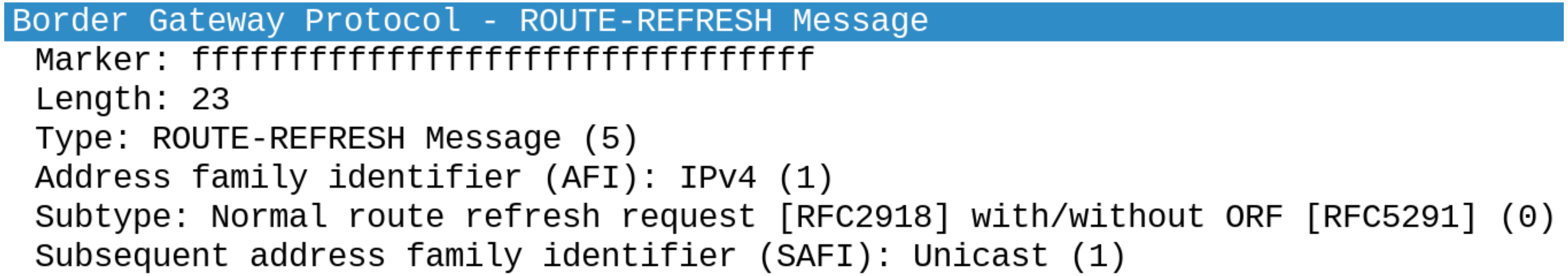


```
Border Gateway Protocol - KEEPALIVE Message
 Marker: ffffffffffffffffffffffffffffffff
 Length: 19
 Type: KEEPALIVE Message (4)
```

BGP ROUTE-REFRESH Message

- Original BGP lacked a way to request prefix resend. Needed for inbound route policy changes.
- Vendors used "Soft Reconfiguration" to store unfiltered routes.

- RFC 2918 introduced ROUTE-REFRESH to request route resend for any address family

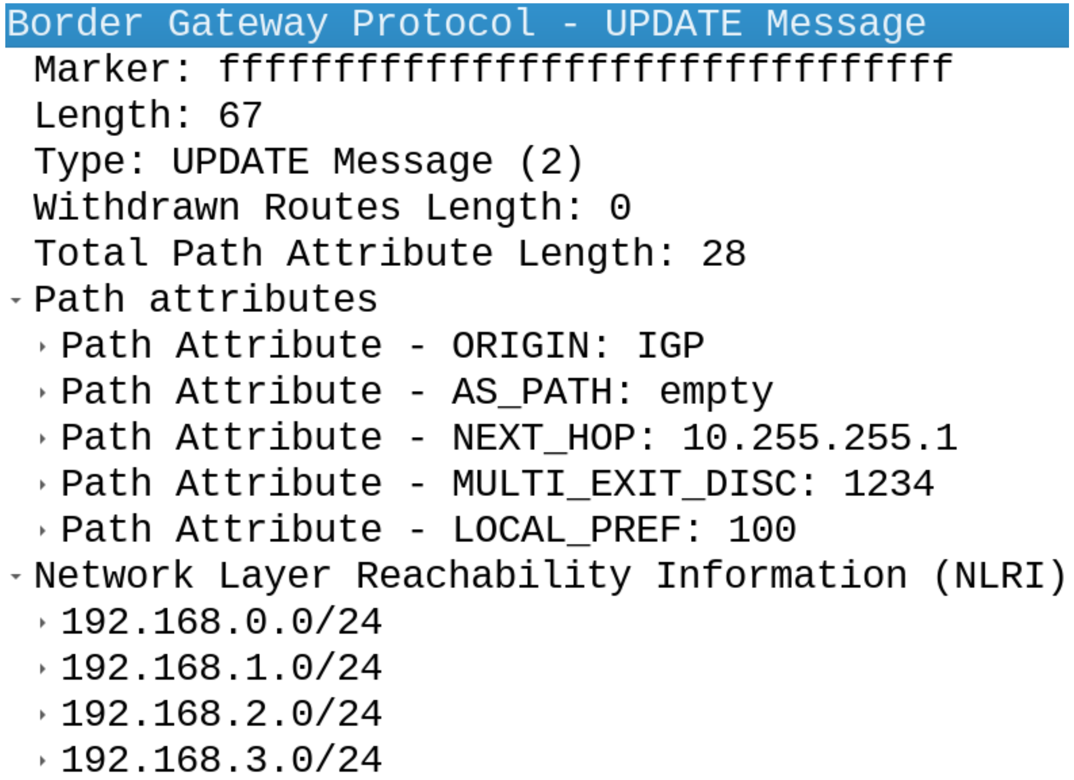


```
Border Gateway Protocol - ROUTE-REFRESH Message
 Marker: ffffffffffffffffffffffffffffffff
 Length: 23
 Type: ROUTE-REFRESH Message (5)
 Address family identifier (AFI): IPv4 (1)
 Subtype: Normal route refresh request [RFC2918] with/without ORF [RFC5291] (0)
 Subsequent address family identifier (SAFI): Unicast (1)
```

BGP UPDATE Message

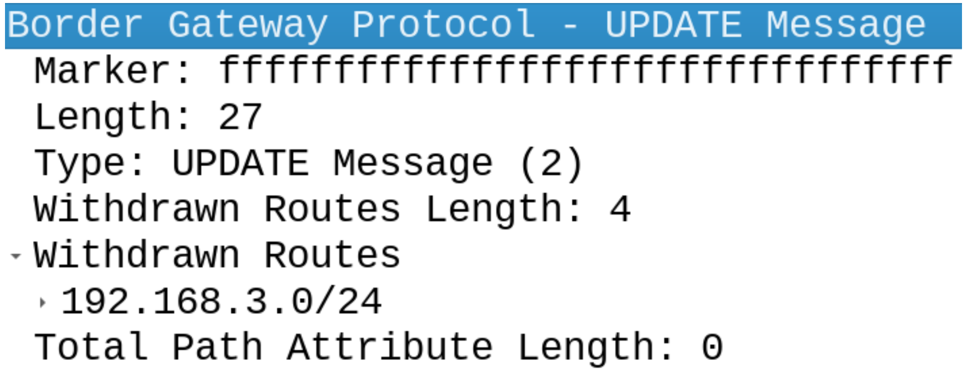
- UPDATE message is BGP's workhorse.
- Advertises reachable (Network Layer Reachability Information) NLRIs with attributes.
- Withdraws unreachable NLRIs.
- Designed for maximum efficiency:
- Path attributes included once, followed by NLRIs sharing them.
- Each NLRI contains only the network prefix (with padding if needed).

BGP Update Message – New/Updated Routes



```
Border Gateway Protocol - UPDATE Message
  Marker: ffffffffffffffffffffffffffffffff
  Length: 67
  Type: UPDATE Message (2)
  Withdrawn Routes Length: 0
  Total Path Attribute Length: 28
▾ Path attributes
  ▸ Path Attribute - ORIGIN: IGP
  ▸ Path Attribute - AS_PATH: empty
  ▸ Path Attribute - NEXT_HOP: 10.255.255.1
  ▸ Path Attribute - MULTI_EXIT_DISC: 1234
  ▸ Path Attribute - LOCAL_PREF: 100
▾ Network Layer Reachability Information (NLRI)
  ▸ 192.168.0.0/24
  ▸ 192.168.1.0/24
  ▸ 192.168.2.0/24
  ▸ 192.168.3.0/24
```

BGP Update Message – Withdrawn Routes



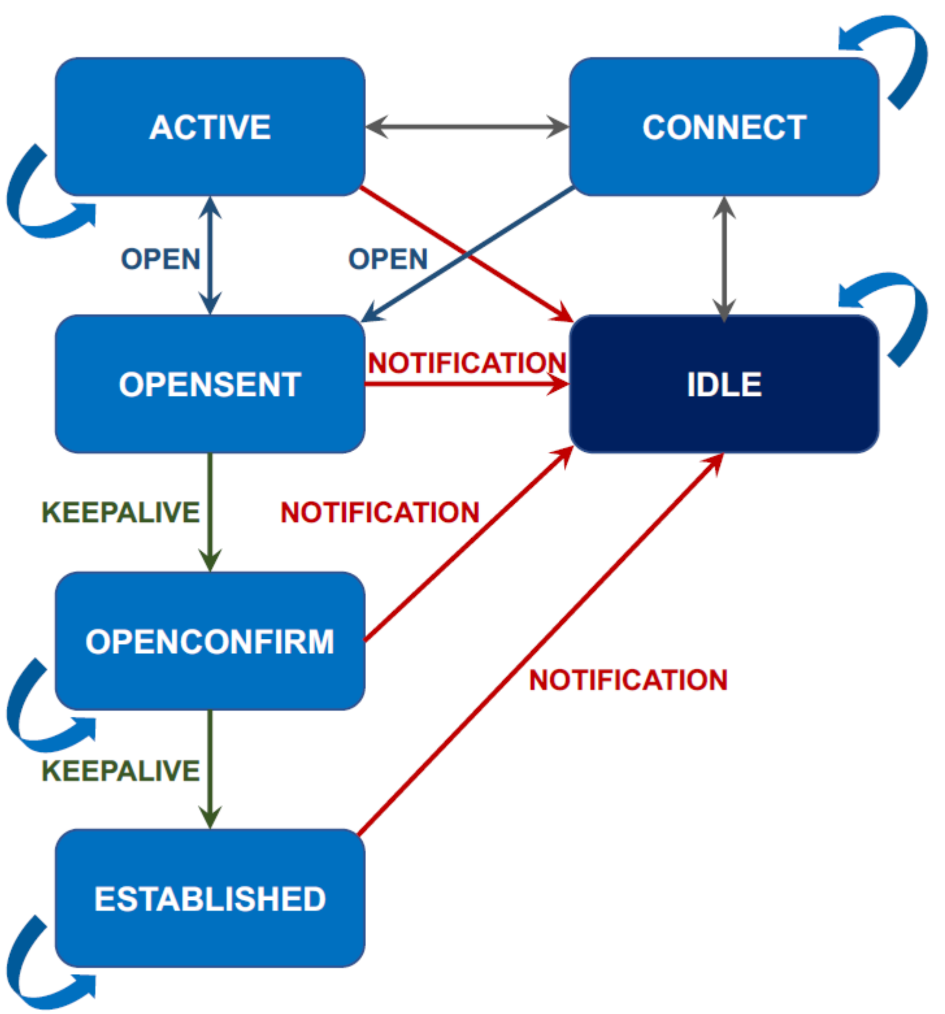
```
Border Gateway Protocol - UPDATE Message
  Marker: ffffffffffffffffffffffffffffffff
  Length: 27
  Type: UPDATE Message (2)
  Withdrawn Routes Length: 4
▾ Withdrawn Routes
  ▸ 192.168.3.0/24
  Total Path Attribute Length: 0
```

BGP Finite State Machine (FSM)

- BGP peer undergoes several state changes in its life cycle
    - IDLE
    - CONNECT
    - ACTIVE
    - OPENSENT
    - OPENCONFIRM
    - ESTABLISHED
- During each state, peers must send and receive messages, process message data, and initialize resources before proceeding to the next state.

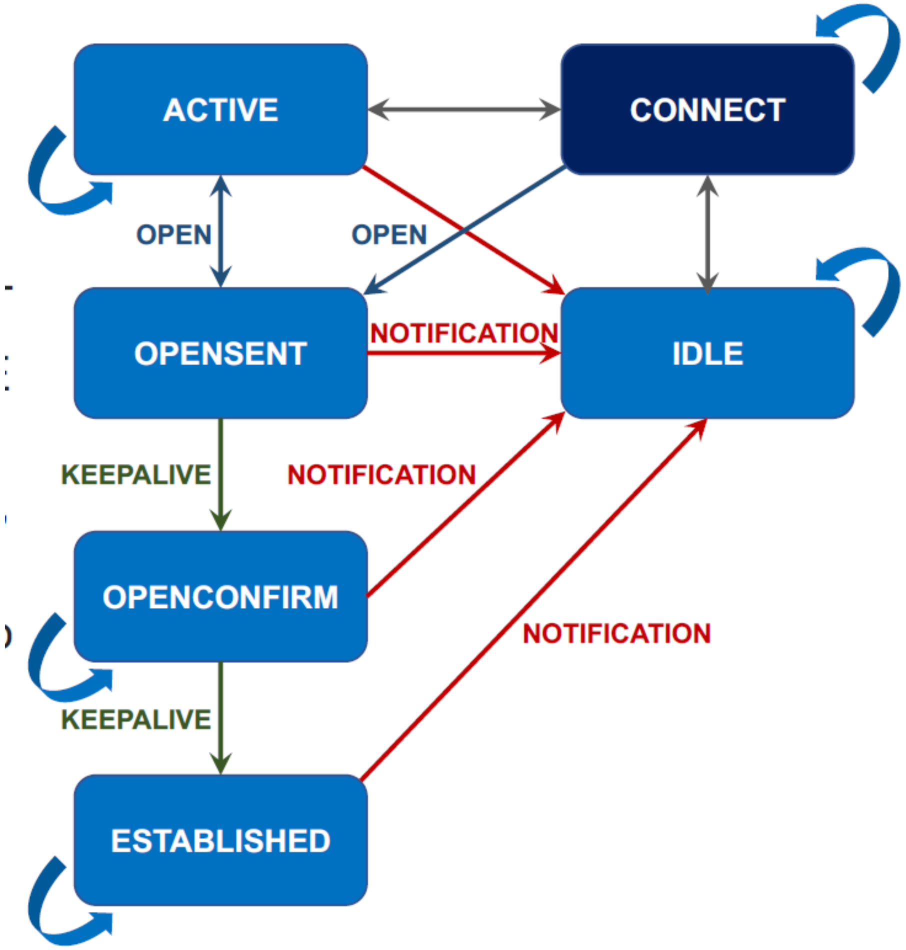
BGP Finite State Machine (FSM)

- **IDLE**
  - Initializes resources
  - Resets ConnectRetryTimer
  - RFC4271 suggested 120 seconds as the default value for ConnectRetryTime
  - Initiates a TCP connection with its configured BGP peer, and listens for a TCP connection from its peer
- If no error, changes peer’s state to CONNECT
- If error occurred, peer remains in IDLE state



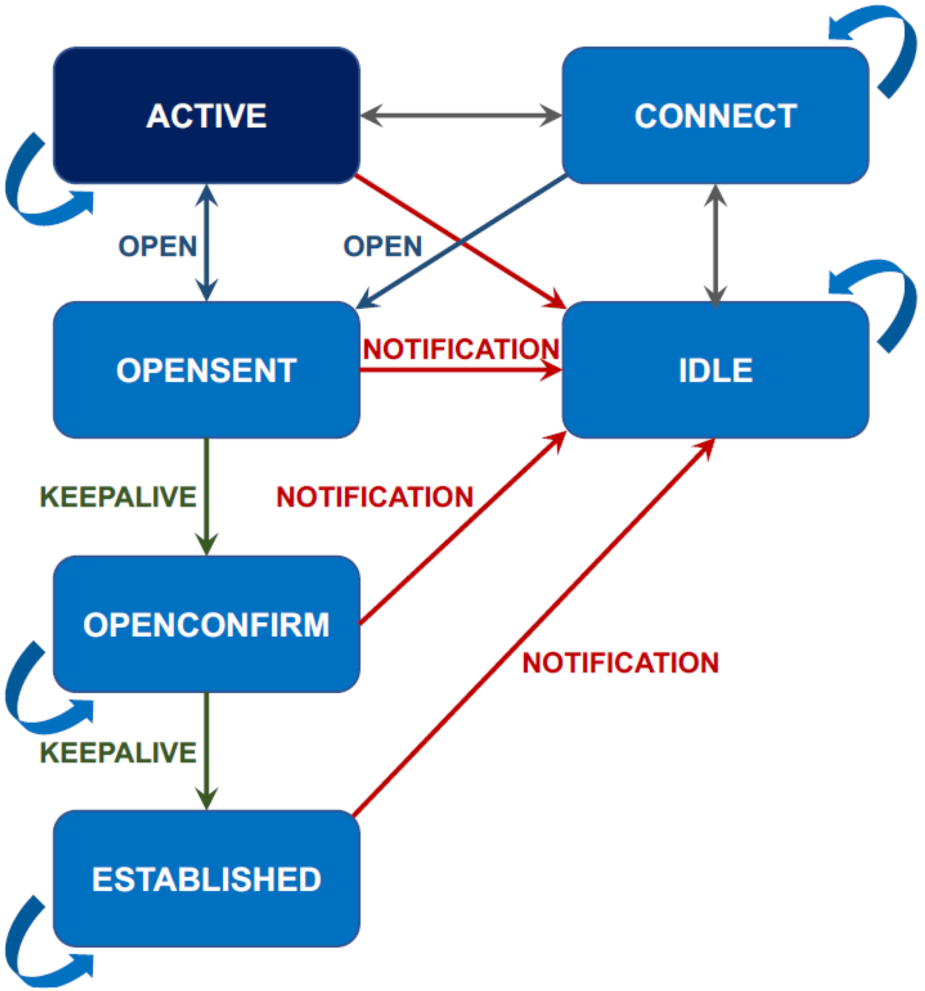
CONNECT

- Waits for successful TCP session
- Sends OPEN to peer
  - If no error, changes peer's state to OPENSENT
  - If error occurs, changes peer's state to ACTIVE
  - If ConnectRetryTimer expired, keeps peer in CONNECT state and resets ConnectRetryTimer, then tries a new TCP three-way handshake
- If something else happens, moves peer back to IDLE state



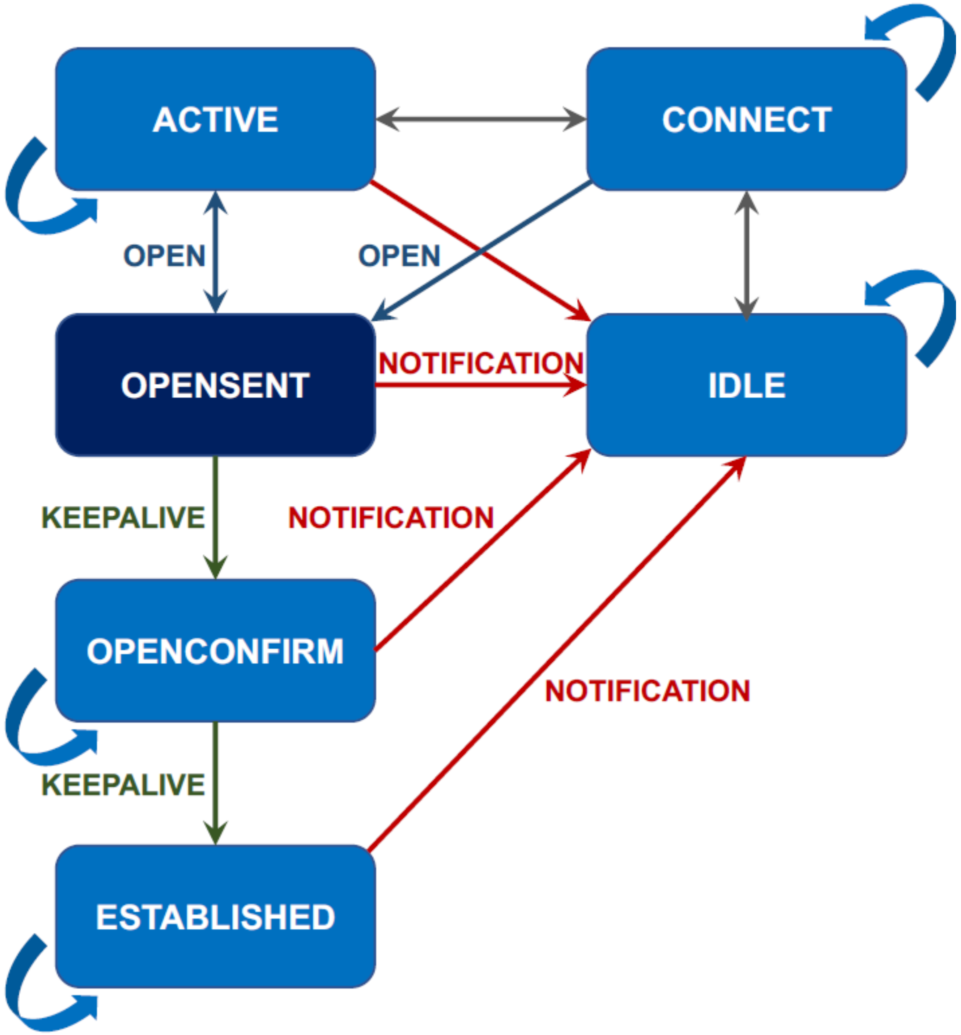
ACTIVE

- Unable to establish a successful TCP session
- Tries to restart another TCP session withthe peer
    - If successful, sends an OPEN to the peer and changes peer’s state to OPENSENT
    - If unsuccessful, changes peer’s state to IDLE
    - If ConnectRetryTimer expires, moves peer back to CONNECT state



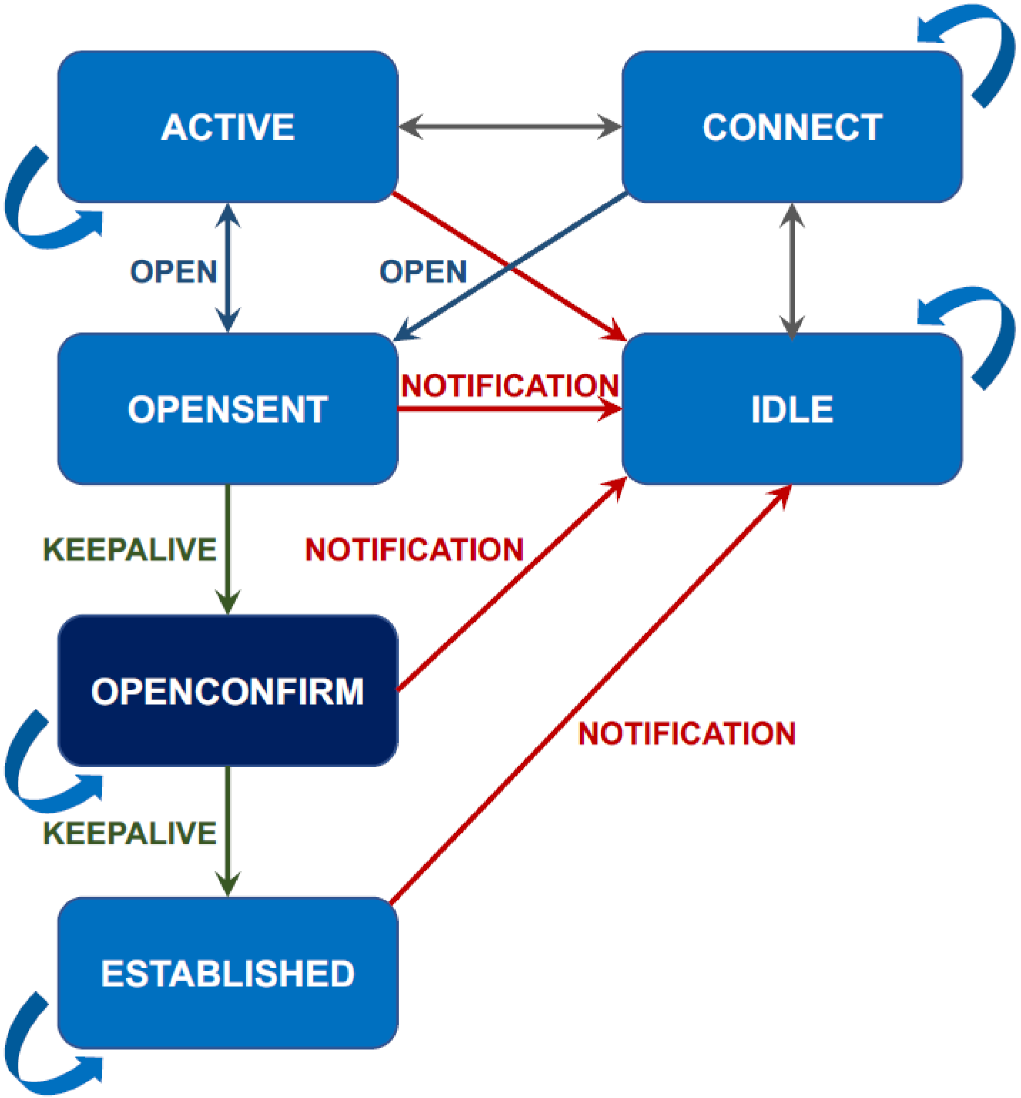
OPENSENT

- OPEN has been sent to peer
- Waits for OPEN from peer
- Checks validity of the received OPEN
  - If there is no error, sends KEEPALIVE message and changes peer’s state to OPENCONFIRM
  - If error occurs due to mismatched OPEN between peers, sends NOTIFICATION and change peer’s state to IDLE
- In case TCP session fails, moves peer back to ACTIVE state



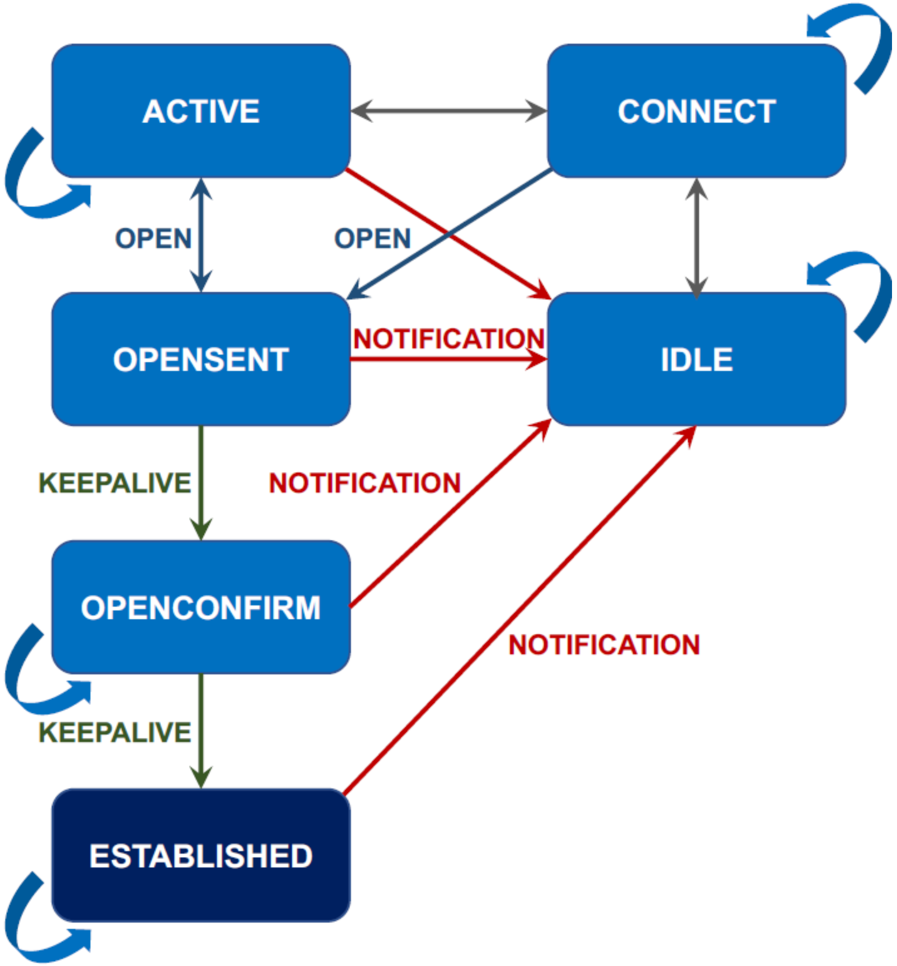
OPENCONFIRM

- Waits for a KEEPALIVE or NOTIFICATION from the peer
  - Upon receipt of peer’s KEEPALIVE, changes peer’s state to ESTABLISHED
  - If the HoldTimer expires or NOTIFICATION is received, changes peer’s state to IDLE



ESTABLISHED

- BGP peer adjacency is complete
- UPDATE is used for exchanging reachability information
  - Initial full routing table exchange
  - Incremental updates for later changes
- In case NOTIFICATION is received, changes peer's state back to IDLE



BGP Attributes

- An attribute is an additional piece of information accompanying an advertised NLRI

- BGP uses attributes in multiple ways
  - Prevents routing loops
  - Performs best path selection
  - Filters or sorts routes
  - … and many more

- Basic BGP specification recognizes only a handful of attributes
- Several new have been added over time for various applications and uses

BGP Attribute Types

- **Well-known: Every BGP implementation must support it**
  - Well-known mandatory: Must always be included with a NLRI
  - Well-known discretionary: May be included with a NLRI as needed

- **Optional:** BGP implementations do not need to support it
  - Optional transitive: When advertising a learned NLRI, keep the attribute with the NLRI even if not recognized
  - Optional non-transitive: When advertising a learned NLRI, remove the attribute from the NLRI if not recognized
- Note: All well-known attributes are transitive

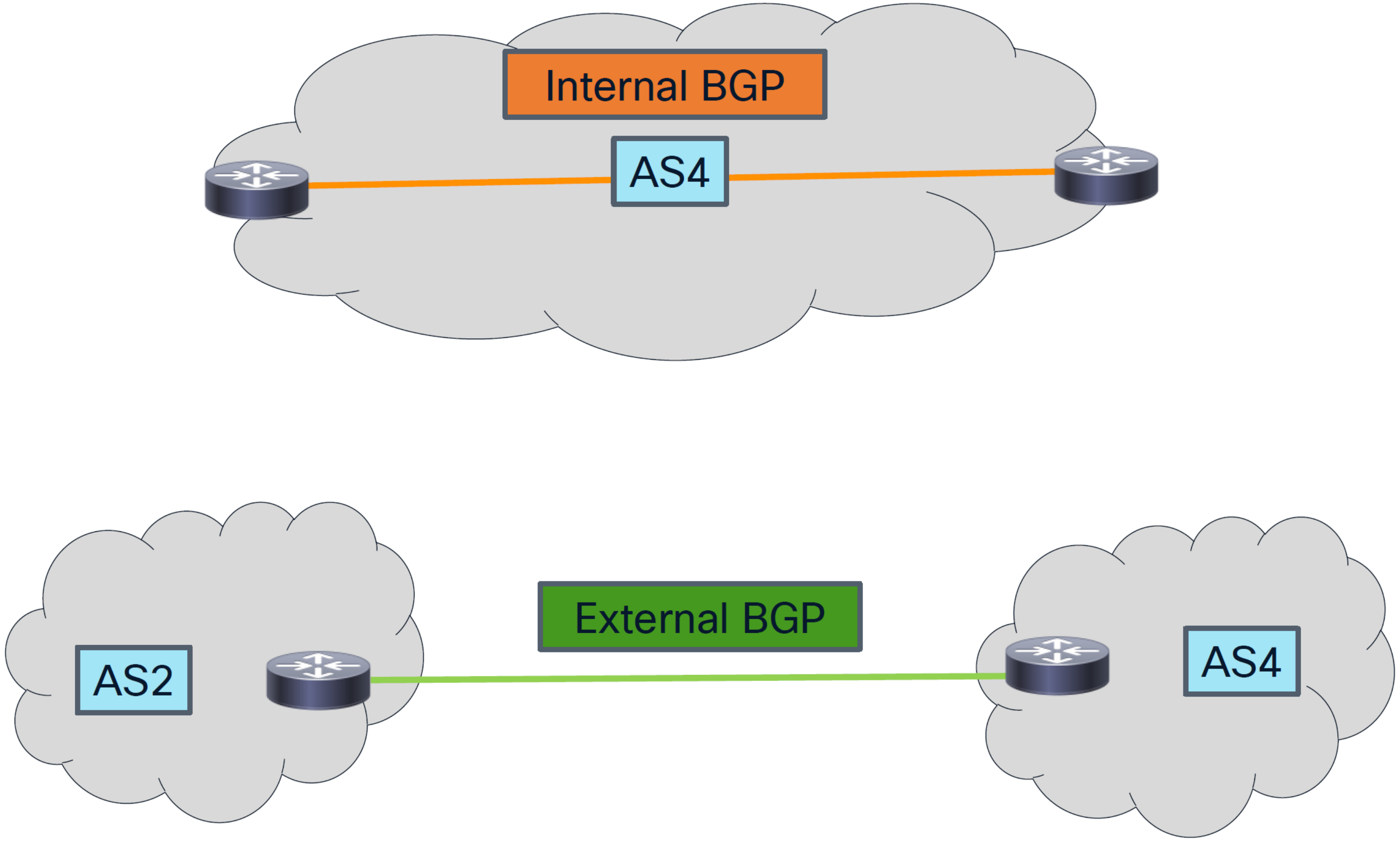
BGP Attributes

- Well-known mandatory
  - AS_PATH
  - NEXT_HOP
  - ORIGIN
- Well-known discretionary
  - LOCAL_PREF
  - ATOMIC_AGGREGATE

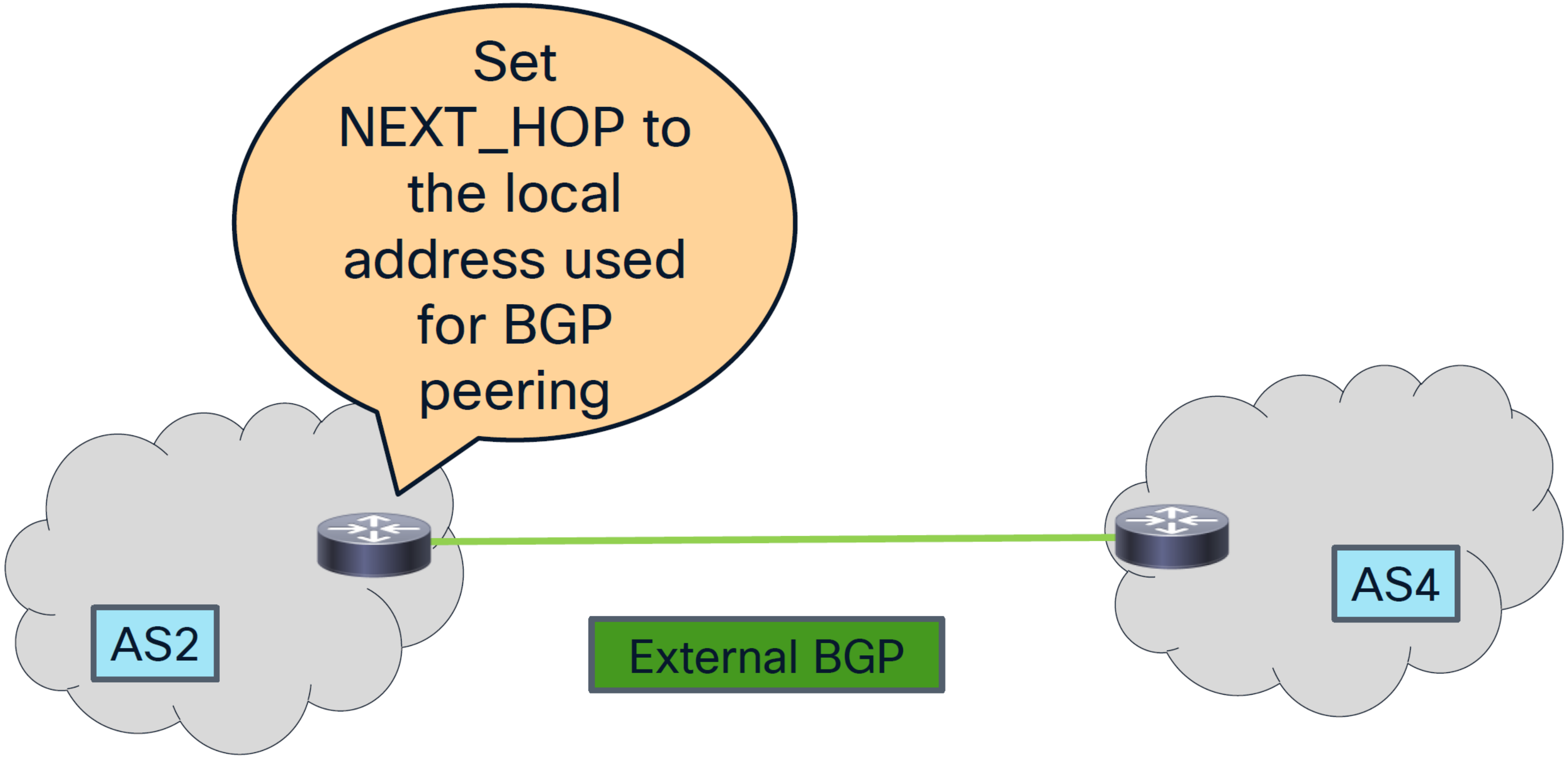
- Optional transitive
  - AGGREGATOR
  - COMMUNITIES
  - EXTENDED_COMMUNITIES
- Optional non-transitive
  - MULTI_EXIT_DISC
  - CLUSTER_LIST

Internal vs External BGP

Internal vs External BGP

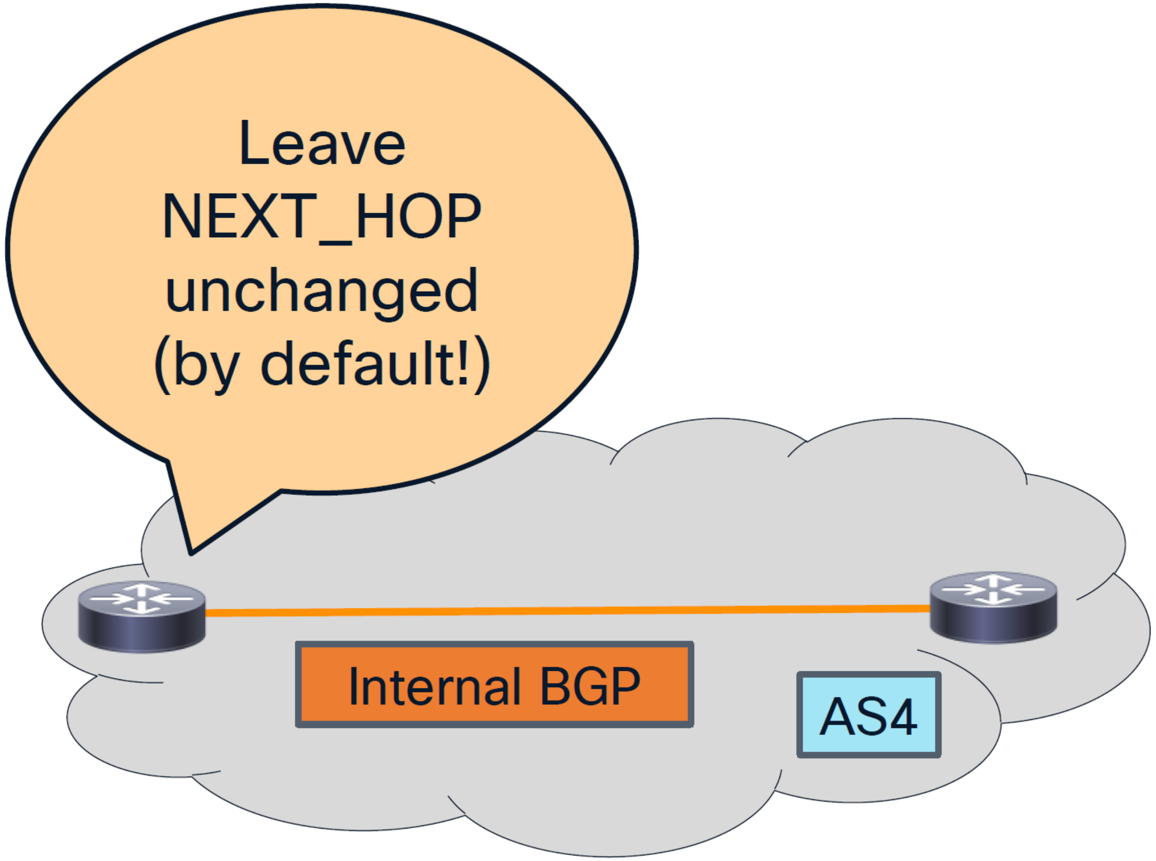


NEXT_HOP in eBGP

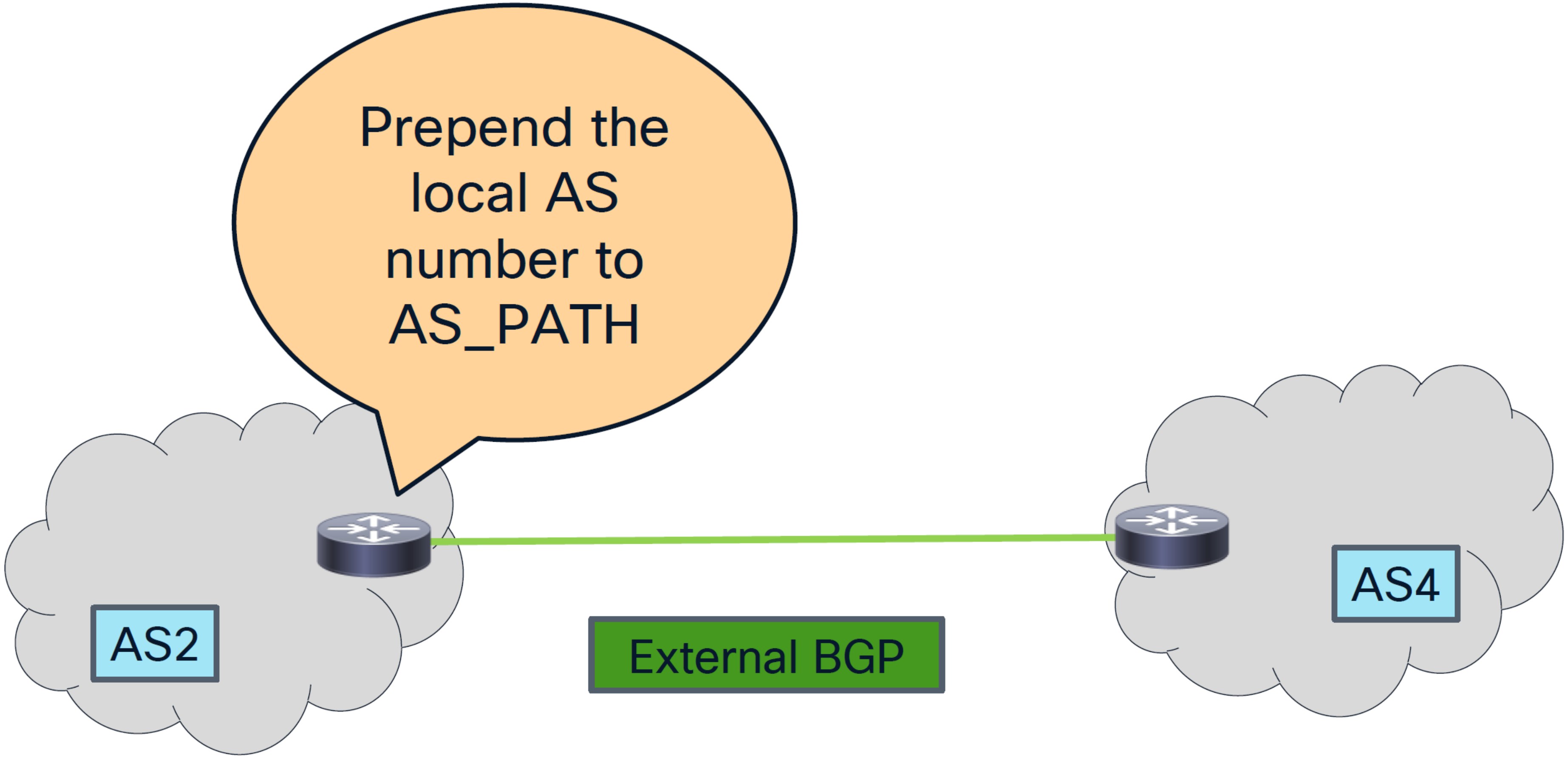


Set NEXT_HOP to the local address used for BGP peering

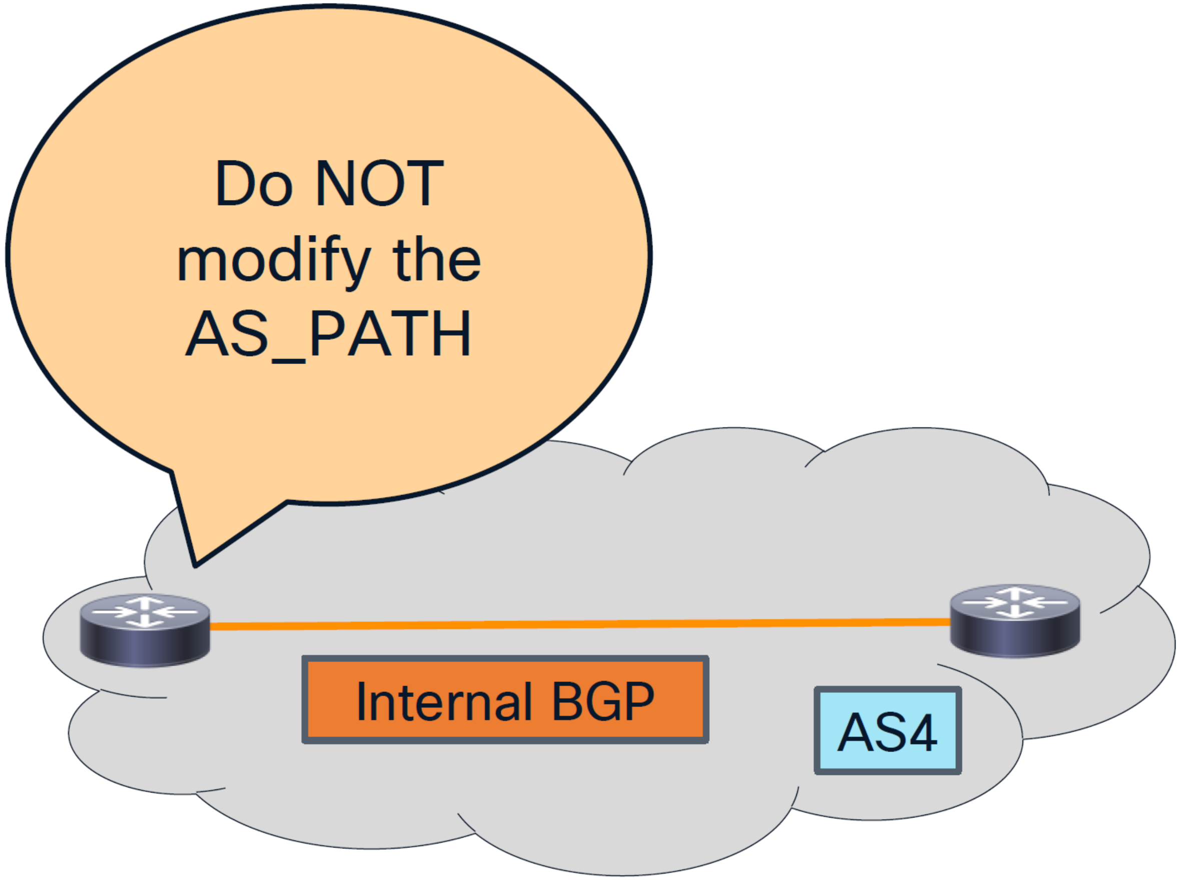
NEXT_HOP in iBGP



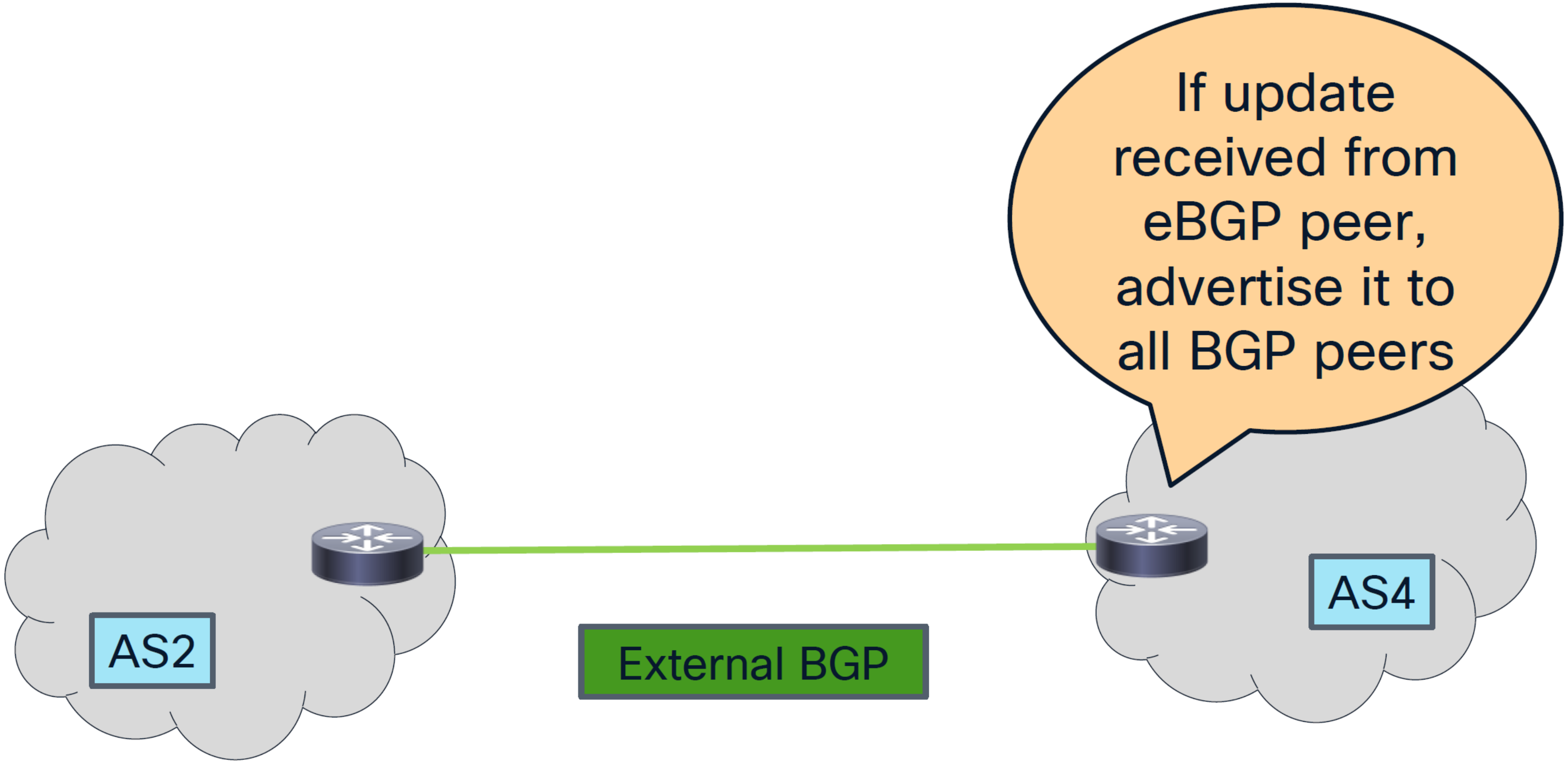
AS_PATH in eBGP



AS_PATH in iBGP

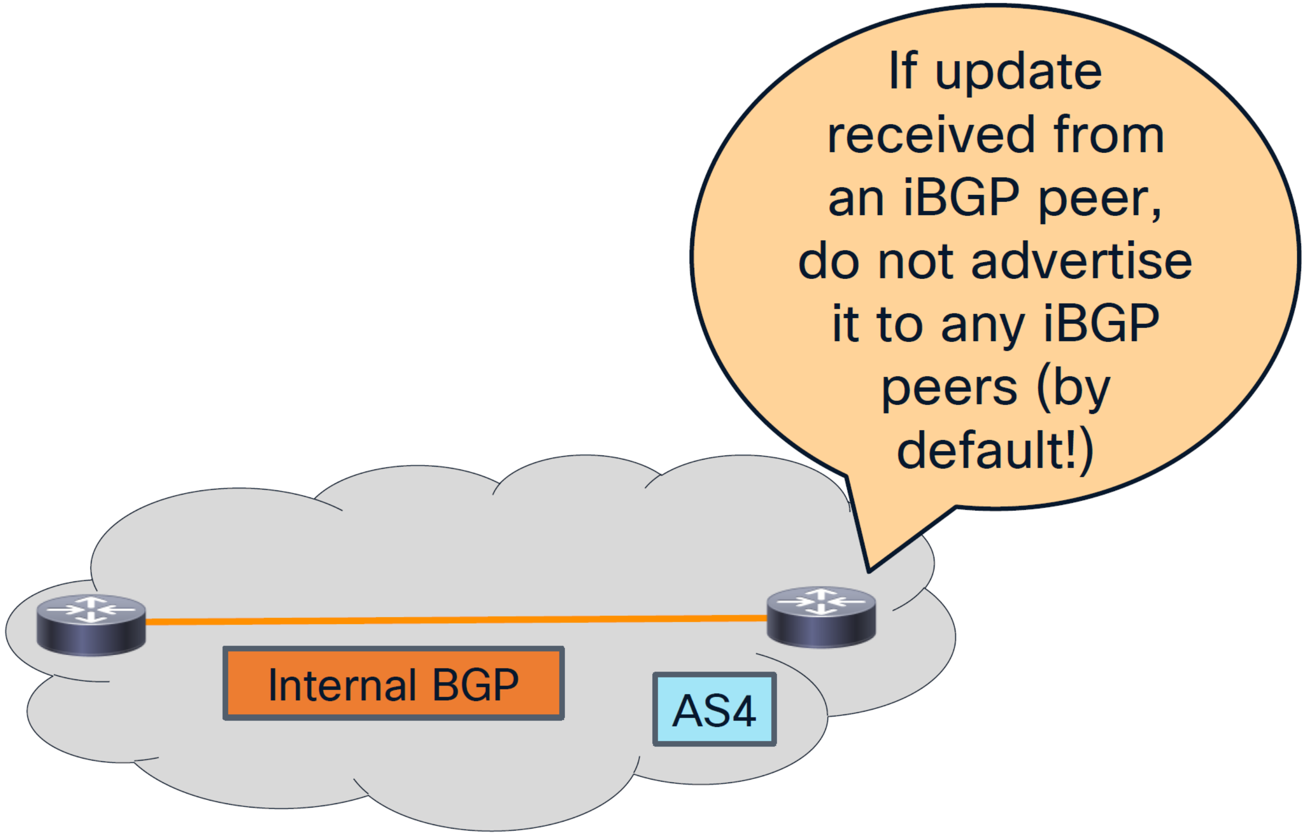


Updates in eBGP



If update received from eBGP peer, advertise it to all BGP peers

Updates in iBGP



If update received from an iBGP peer, do not advertise it to any iBGP peers (by default!)

BGP Operations

BGP Transport

- BGP operates by exchanging Network Layer Reachability Information (NLRI).
  - NLRI includes a set of BGP path attributes and one or more prefixes which those attributes are associated
  - NLRI is encapsulated inside the BGP UPDATE message

- Does not have own transport protocol.
- Utilizes TCP and runs on TCP port 179.
- BGP messages are exchanged over the TCP session.

BGP Capabilities

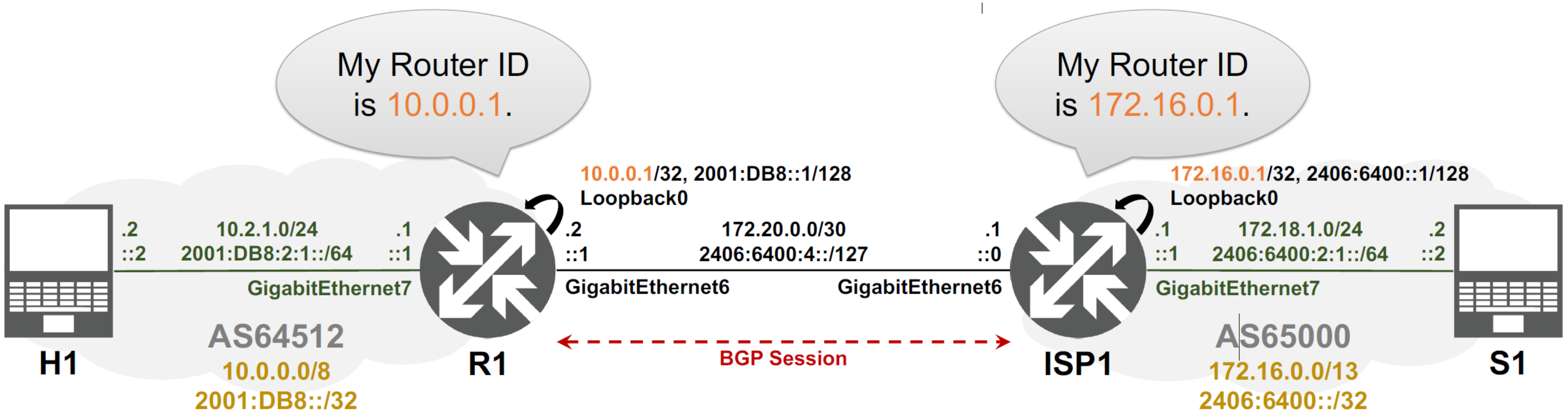
- Capability codes indicate whether a BGP router is able to accommodate particular capabilities.
    - Advertised in OPEN message
- If received capability is not supported by remote peer, it sends back a NOTIFICATION message.
- BGP routers attempt to peer without the unsupported capability.
- Commonly implemented capabilities:
    - Route Refresh
    - Multi-protocol Extension
    - Support for 4-octet AS Number

BGP Router ID

- The BGP router identifier (ID) is a 4-byte field that is set to the highest IP address on the router.
  - (Cisco) Loopback interface addresses are considered before physical interface addresses because loopback interfaces are more stable than physical interfaces.
  - The BGP router ID is used in the BGP algorithm for determining the best path to a destination where the preference is for the BGP router with the lowest router ID.
  - It is possible to manually configure the BGP router ID using the bgp router-id command to influence the best path algorithm.

Router ID

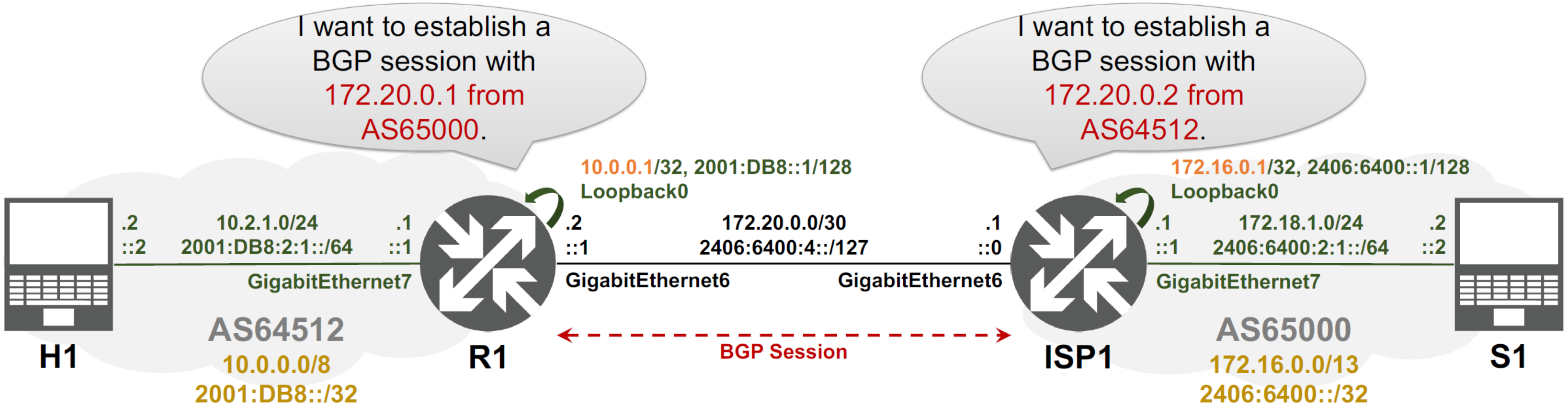
- Cisco IOS: Highest Loopback IPv4 prefers than Highest active interface IPv4
- Juniper Junos OS: Lowest Loopback IPv4 prefers than Lowest physical interface IPv4
- MikroTik RouterOS: Lowest active interface IPv4



APNIC

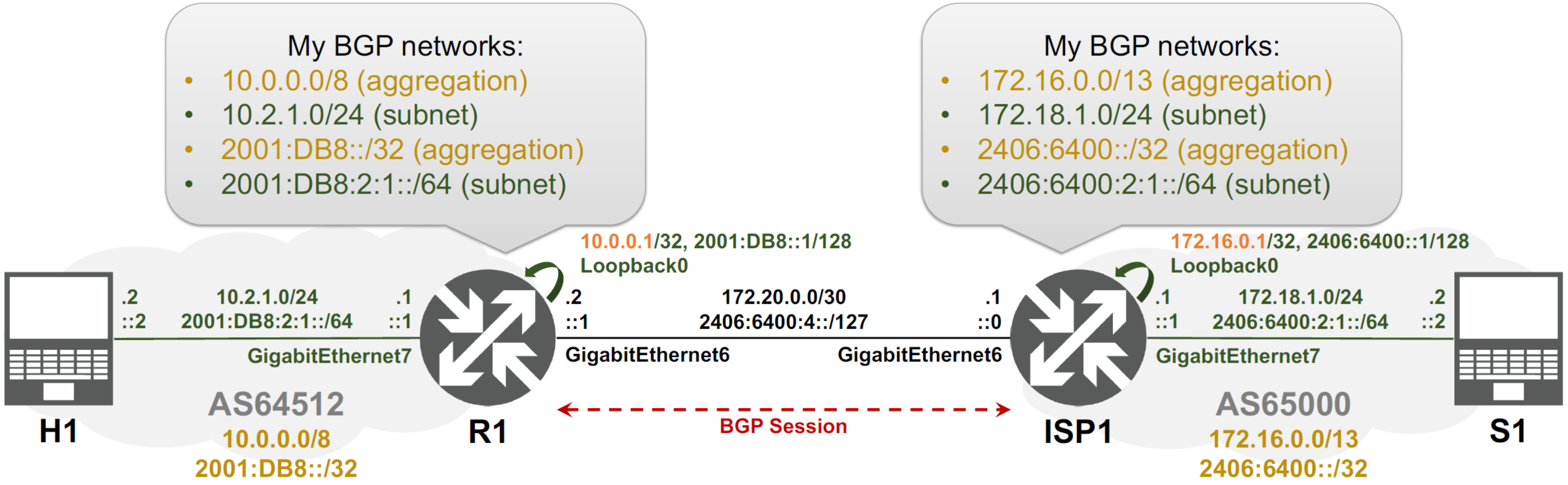
BGP Peer

- BGP does not perform auto-discovery for peers (neighbors).
- BGP peers are manually configured.
  - Local peer address and ASN
  - Remote peer address and ASN
  - Authentication (if necessary)



BGP Network

- Indicates a BGP prefix that should be originated by the router.
- By default, the prefix is advertised only if corresponding route is present in the routing table



My BGP networks:

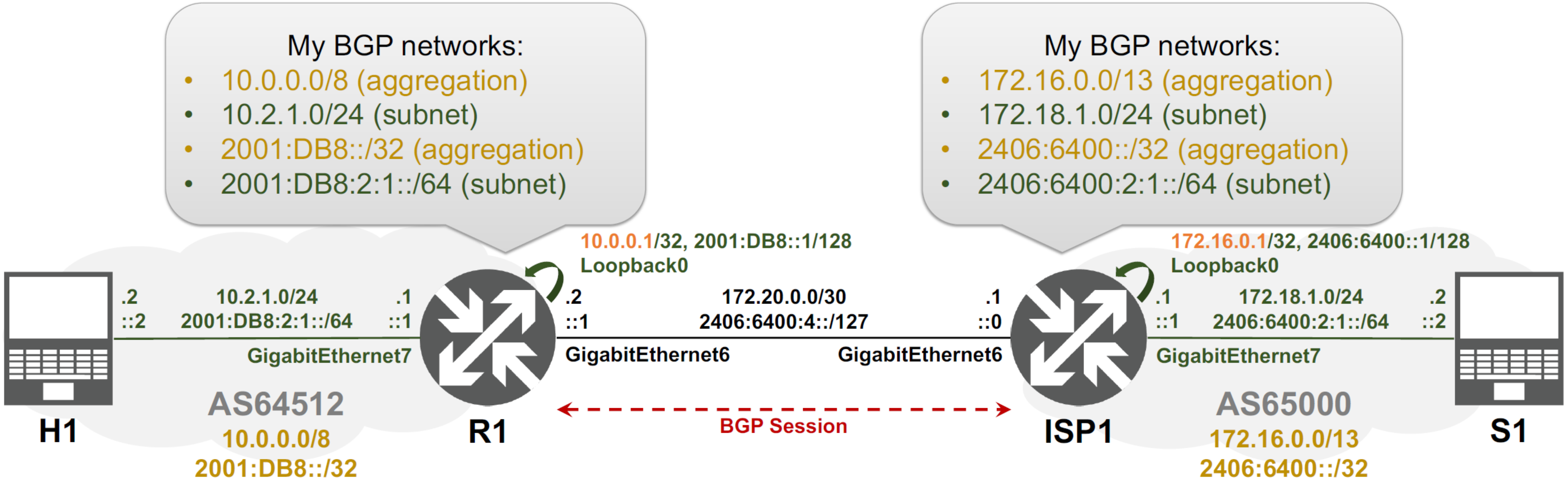
- 10.0.0.0/8 (aggregation)
- 10.2.1.0/24 (subnet)
- 2001:DB8::/32 (aggregation)
- 2001:DB8:2:1::/64 (subnet)

My BGP networks:

- 172.16.0.0/13 (aggregation)
- 172.18.1.0/24 (subnet)
- 2406:6400::/32 (aggregation)
- 2406:6400:2:1::/64 (subnet)

BGP Network

- Prefixes are usually subnet or aggregate routes instead of
- individual host routes.
- IPv4 prefixes longer than /24 and IPv6 prefixes longer than /48 won’t generally be accepted on the Internet.



My BGP networks:

- 10.0.0.0/8 (aggregation)
- 10.2.1.0/24 (subnet)
- 2001:DB8::/32 (aggregation)
- 2001:DB8:2:1::/64 (subnet)

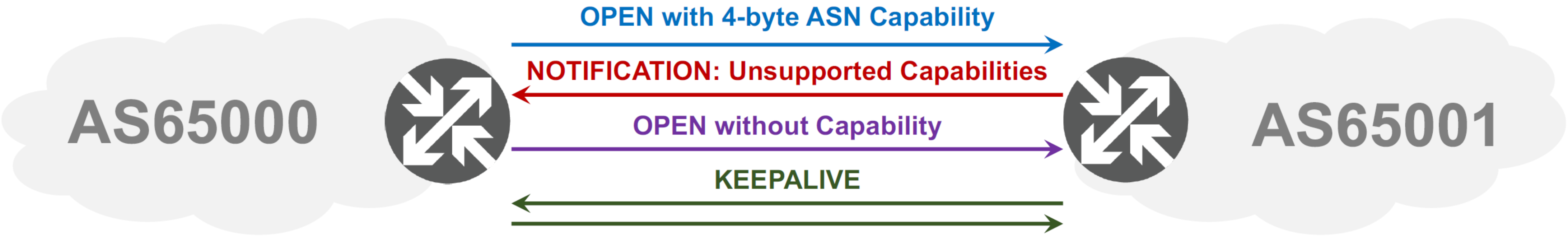
My BGP networks:

- 172.16.0.0/13 (aggregation)
- 172.18.1.0/24 (subnet)
- 2406:6400::/32 (aggregation)
- 2406:6400:2:1::/64 (subnet)

BGP Best Path

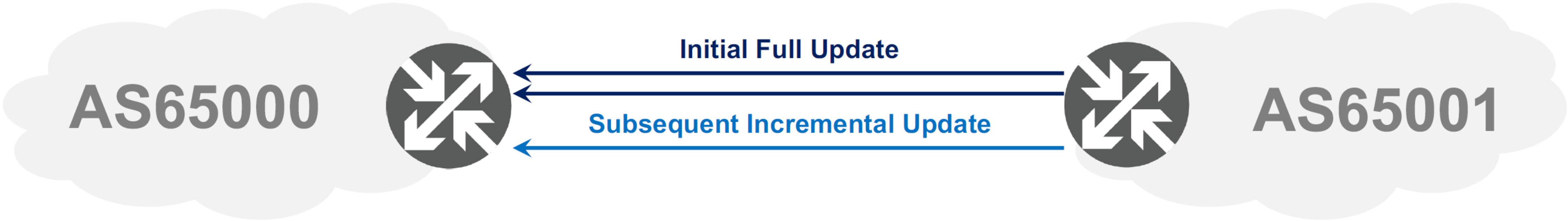
- Best path is the path that BGP selected to use in RIB.
- BGP uses path attributes to determine best path.
    - Administrator influence the selection process by routing policy
    - Best paths might not be the shortest path, but the most suitable path based on the routing policy
- By default, BGP installs single best path for each destination.
- BGP propagates only the best path to the peers.
- BGP Multipath is a feature that allows BGP to install multiple best paths when they have the same metrics.
    - For load sharing over multiple next hops

BGP Session Establishment


AS65000
OPEN with 4-byte ASN Capability
NOTIFICATION: Unsupported Capabilities
OPEN without Capability
KEEPALIVE
AS65001

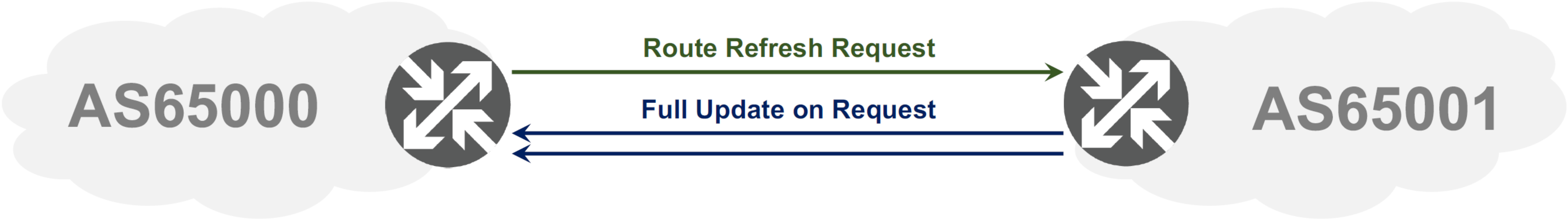
- Sends OPEN to peer after TCP three-way handshake.
- Peer replies NOTIFICATION if capabilities unsupported.
- Resends OPEN without unsupported capabilities.
- Peer replies KEEPALIVE if OPEN is acceptable.
- KEEPALIVE is sent periodically for maintaining the session.

BGP Updates



- Initial full update upon BGP session establishment.
- Subsequent incremental updates after initial full update.
  - When new prefixes are being advertised
  - When existing prefixes are being updated
  - When existing prefixes are being withdrawn

Route Refresh Capability

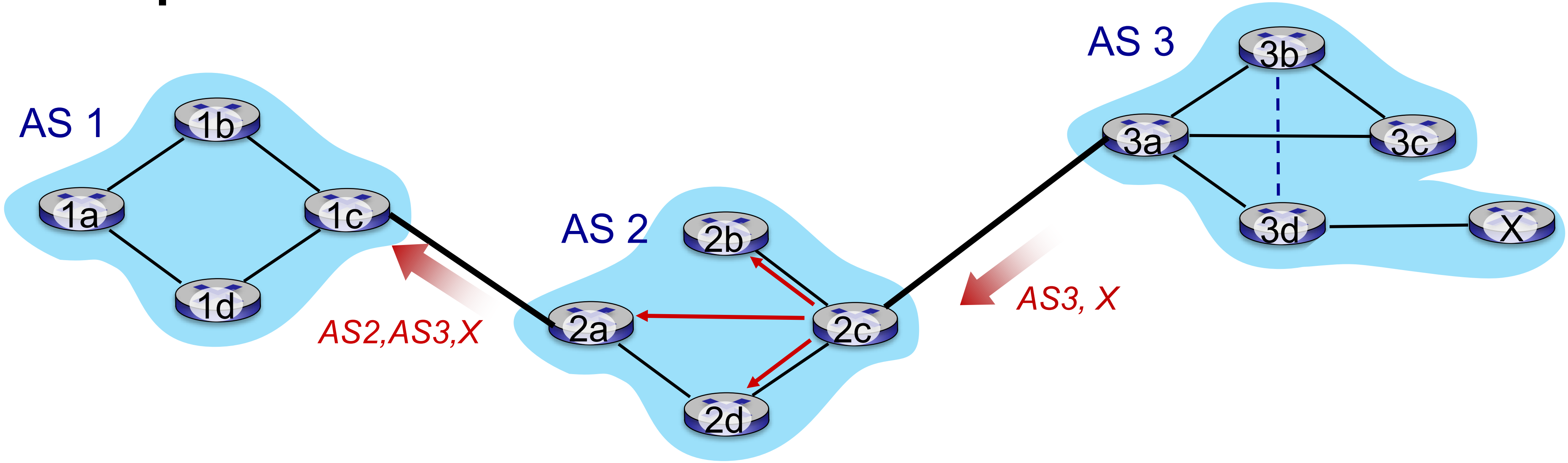


- Requests peer to resend full BGP update.
  - RFC2918: Route Refresh Capability for BGP-4

Path attributes and BGP routes

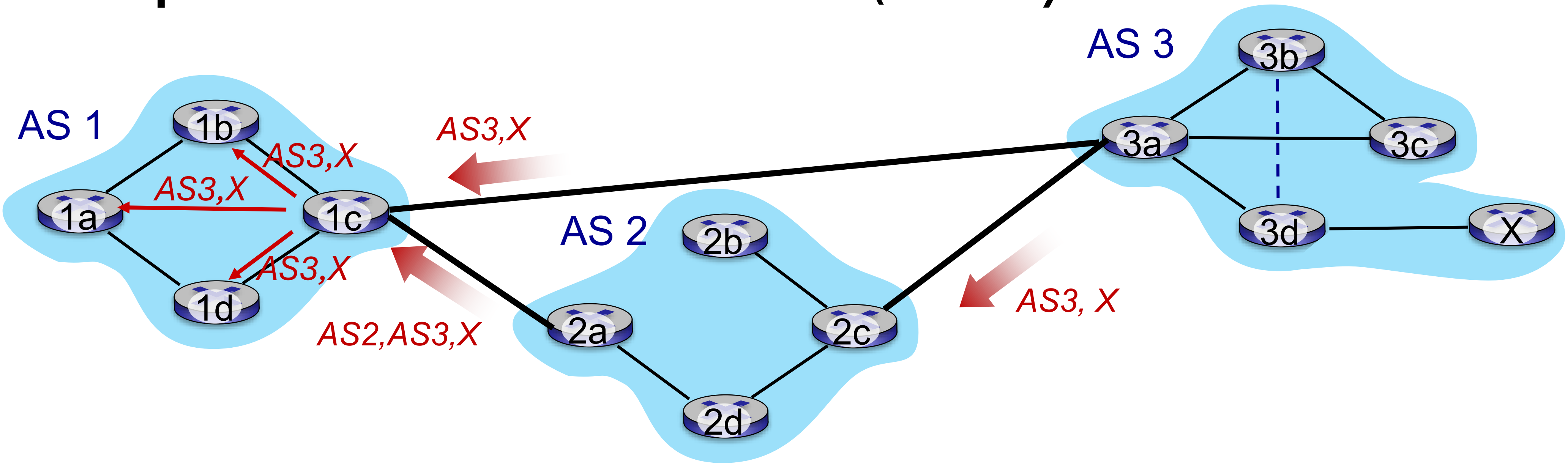
- BGP advertised route: prefix + attributes
  - prefix: destination being advertised
  - two important attributes:
    - AS-PATH: list of ASes through which prefix advertisement has passed
    - NEXT-HOP: indicates specific internal-AS router to next-hop AS
- policy-based routing:
  - gateway receiving route advertisement uses *import policy* to accept/decline path (e.g., never route through AS Y).
  - AS policy also determines whether to *advertise* path to other other neighboring ASes

BGP path advertisement



- AS2 router 2c receives path advertisement AS3,X (via eBGP) from AS3 router 3a
- based on AS2 policy, AS2 router 2c accepts path AS3,X, propagates (via iBGP) to all AS2 routers
- based on AS2 policy, AS2 router 2a advertises (via eBGP) path AS2, AS3, X to AS1 router 1c

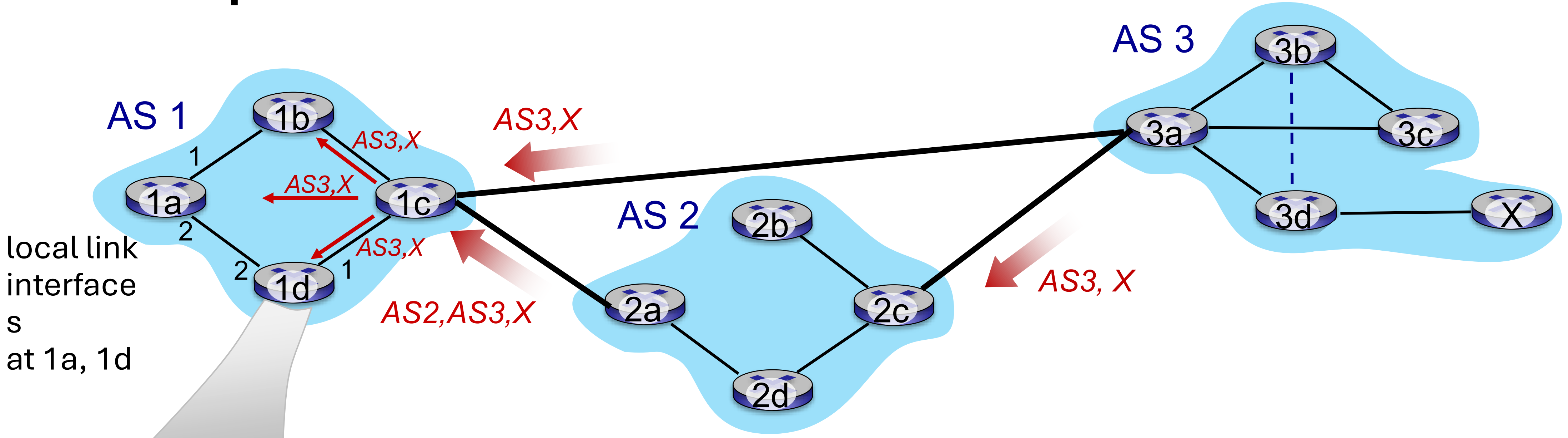
BGP path advertisement (more)

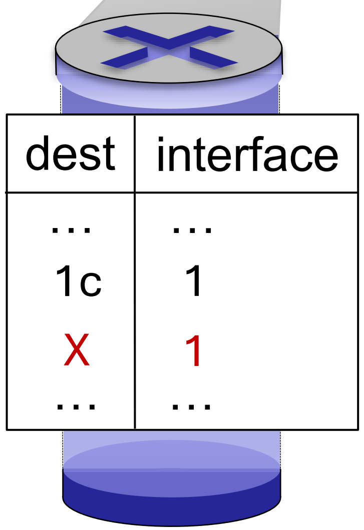


gateway router may learn about multiple paths to destination:

- AS1 gateway router 1c learns path *AS2,AS3,X* from 2a
- AS1 gateway router 1c learns path *AS3,X* from 3a
- based on *policy,* AS1 gateway router 1c chooses path *AS3,X* and advertises path within AS1 via iBGP

BGP path advertisement

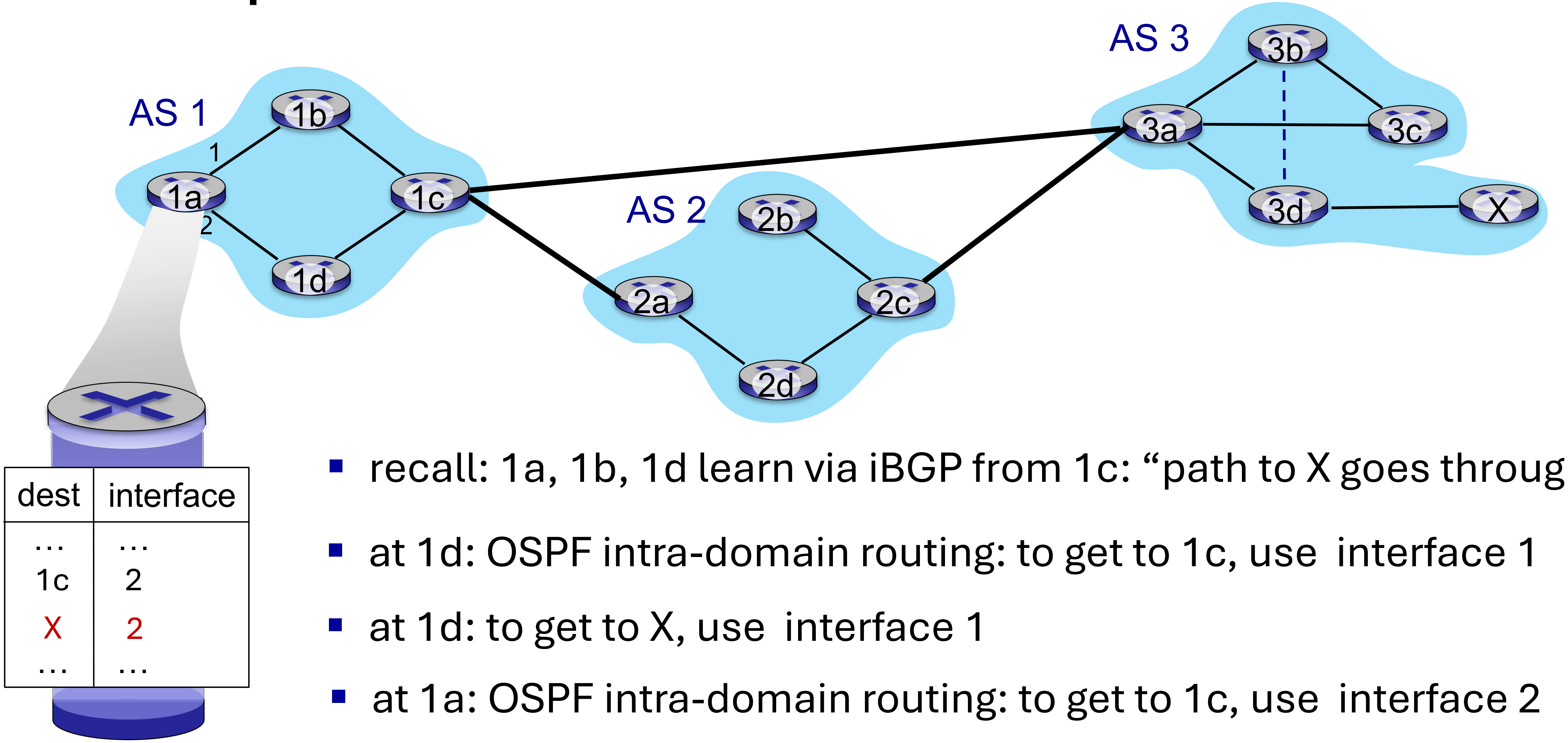




dest	interface
…	…
1c	1
X	1
…	…

- recall: 1a, 1b, 1d learn via iBGP from 1c: “path to X goes through 1c”
- at 1d: OSPF intra-domain routing: to get to 1c, use interface 1
- at 1d: to get to X, use interface 1

BGP path advertisement



dest	interface
…	…
1c	2
X	2
…	…

- recall: 1a, 1b, 1d learn via iBGP from 1c: "path to X goes through 1c"
- at 1d: OSPF intra-domain routing: to get to 1c, use interface 1
- at 1d: to get to X, use interface 1
- at 1a: OSPF intra-domain routing: to get to 1c, use interface 2
- at 1a: to get to X, use interface 2

Why different Intra-, Inter-AS routing ?

policy:

- inter-AS: admin wants control over how its traffic routed, who routes through its network
- intra-AS: single admin, so policy less of an issue

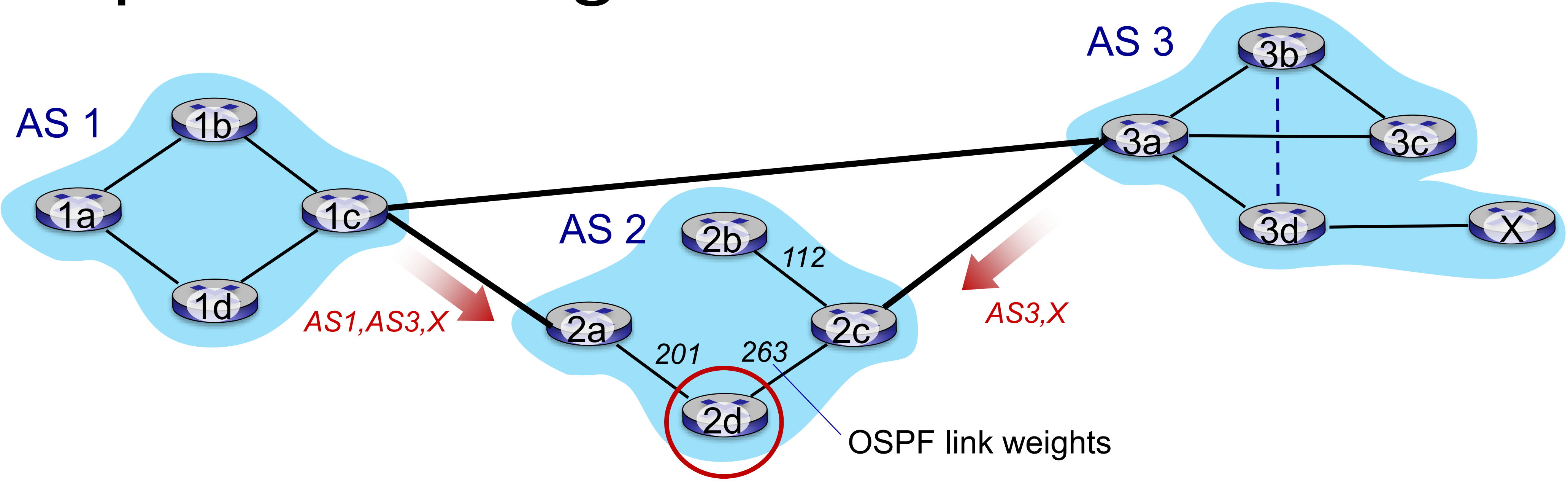
scale:

- hierarchical routing saves table size, reduced update traffic

performance:

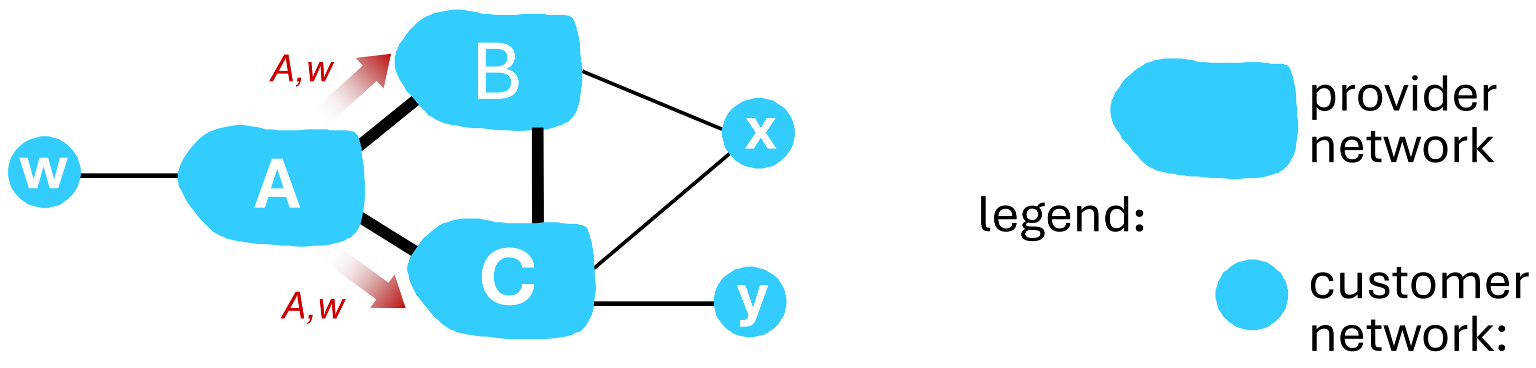
- intra-AS: can focus on performance
- inter-AS: policy dominates over performance

Hot potato routing



- 2d learns (via iBGP) it can route to X via 2a or 2c
- hot potato routing: choose local gateway that has least *intra-domain* cost (e.g., 2d chooses 2a, even though more AS hops to *X*): don't worry about inter-domain cost!

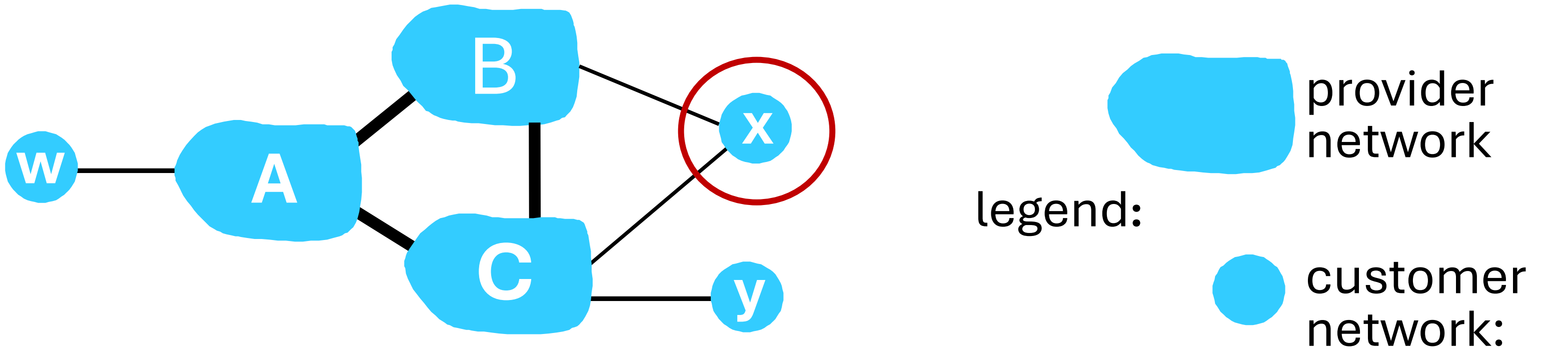
BGP: achieving policy via advertisements



ISP only wants to route traffic to/from its customer networks (does not want to carry transit traffic between other ISPs – a typical "real world" policy)

- A advertises path Aw to B and to C
- B *chooses not to advertise* BAw to C!
  - B gets no "revenue" for routing CBAw, since none of C, A, w are B's customers
  - C does *not* learn about CBAw path
- C will route CAw (not using B) to get to w

BGP: achieving policy via advertisements (more)



ISP only wants to route traffic to/from its customer networks (does not want to carry transit traffic between other ISPs – a typical "real world" policy)

- A,B,C are provider networks
- x,w,y are customer (of provider networks)
- x is dual-homed: attached to two networks
- *policy to enforce:* x does not want to route from B to C via x
    - .. so x will not advertise to B a route to C

BGP route selection

- router may learn about more than one route to destination AS, selects route based on:
    1. local preference value attribute: policy decision
    2. shortest AS-PATH
    3. closest NEXT-HOP router: hot potato routing
    4. additional criteria