External Gateway Routing Protocols: BGP & MP-BGP

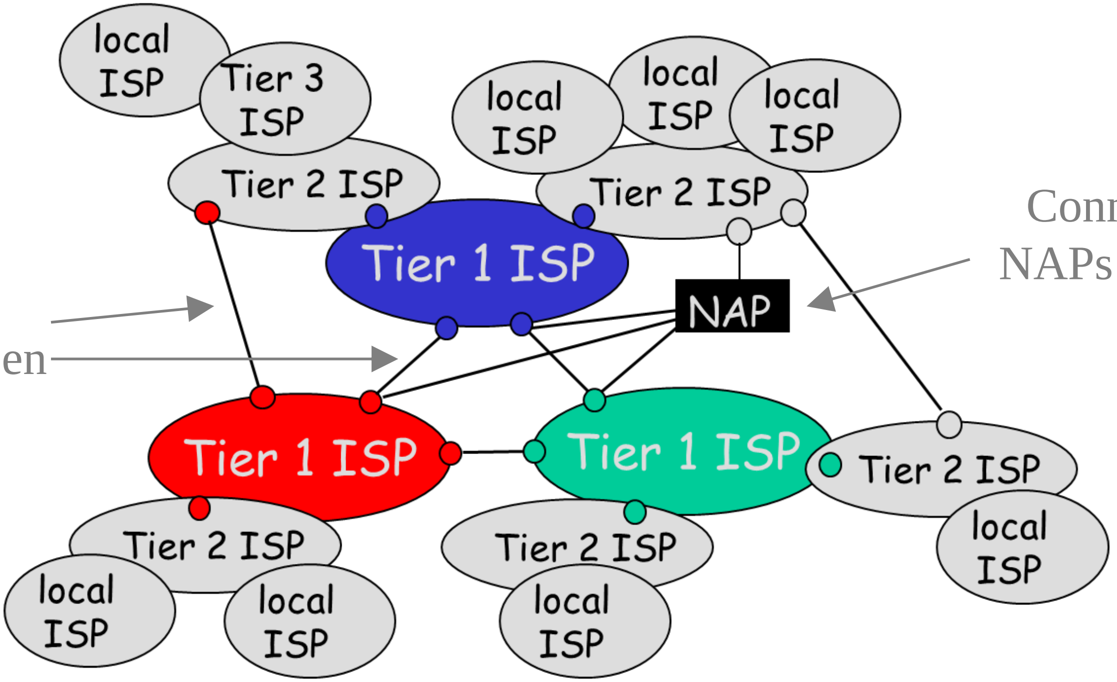
Redes de Comunicações II

Licenciatura em Engenharia de
Computadores e Informática

Prof. Amaro de Sousa (asou@ua.pt)
DETI-UA, 2024/2025

INTERNET: Network of Networks

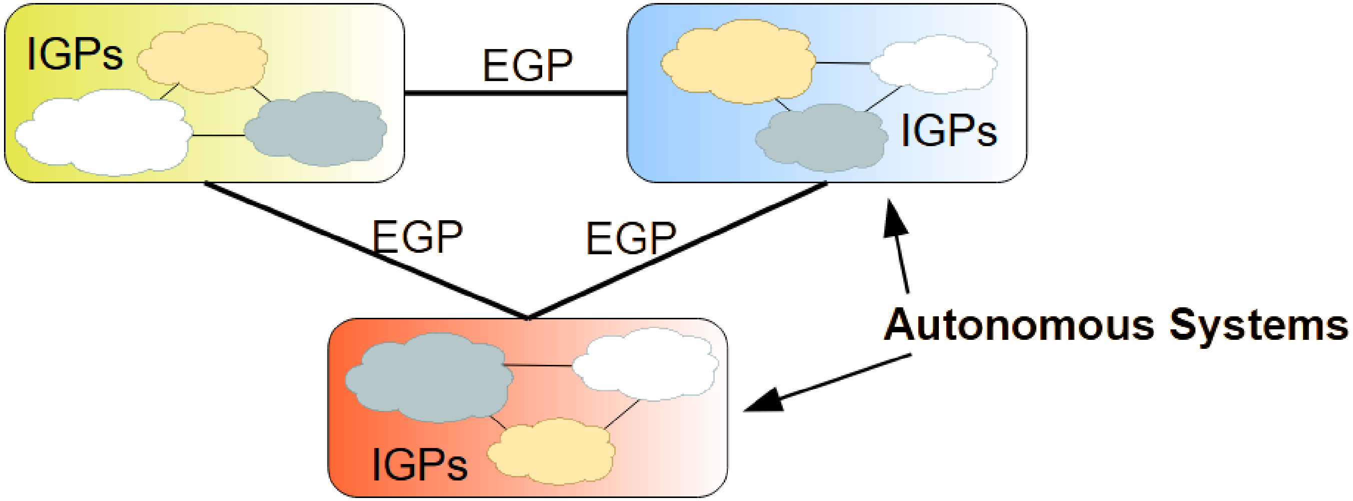
- Aproximately hierarchical
  - Tier 1 ISPs (world-wide coverage)
    - treat themselves as equal
  - Tier 2 ISPs (regional coverage)
    - connect to Tier 1 ISPs (and possibly between them)
    - pay to Tier 1 ISPs for global connectivity
  - Tier 3 and local ISPs (providing Internet access to end users)
    - connect to Tier 2 ISPs
    - pay to Tier 2 ISPs for global connectivity



Private peer connections between ISPs

Connections through NAPs (Network Access Points)

Border Gateway Protocol (BGP)



- The Border Gateway Protocol - version 4 (BGPv4) deployed in 1993 is currently the protocol that ensures Internet connectivity
- BGP is mainly used for routing between Autonomous Systems
- An Autonomous System (AS) is a set of networks under a single administration using IGP protocols

AS Identification

- An Autonomous System (AS) is identified by a number.
- AS numbers (ASNs) are assigned to Local Internet Registries (LIRs) and end-user organizations by their respective Regional Internet Registries (RIRs)
- RIRs, in turn, receive blocks of ASNs for reassignment from the Internet Assigned Numbers Authority (IANA).
- The IANA also maintains a registry of ASNs which are reserved for private use (and should therefore not be announced to the global Internet).

ASN (AS Number)

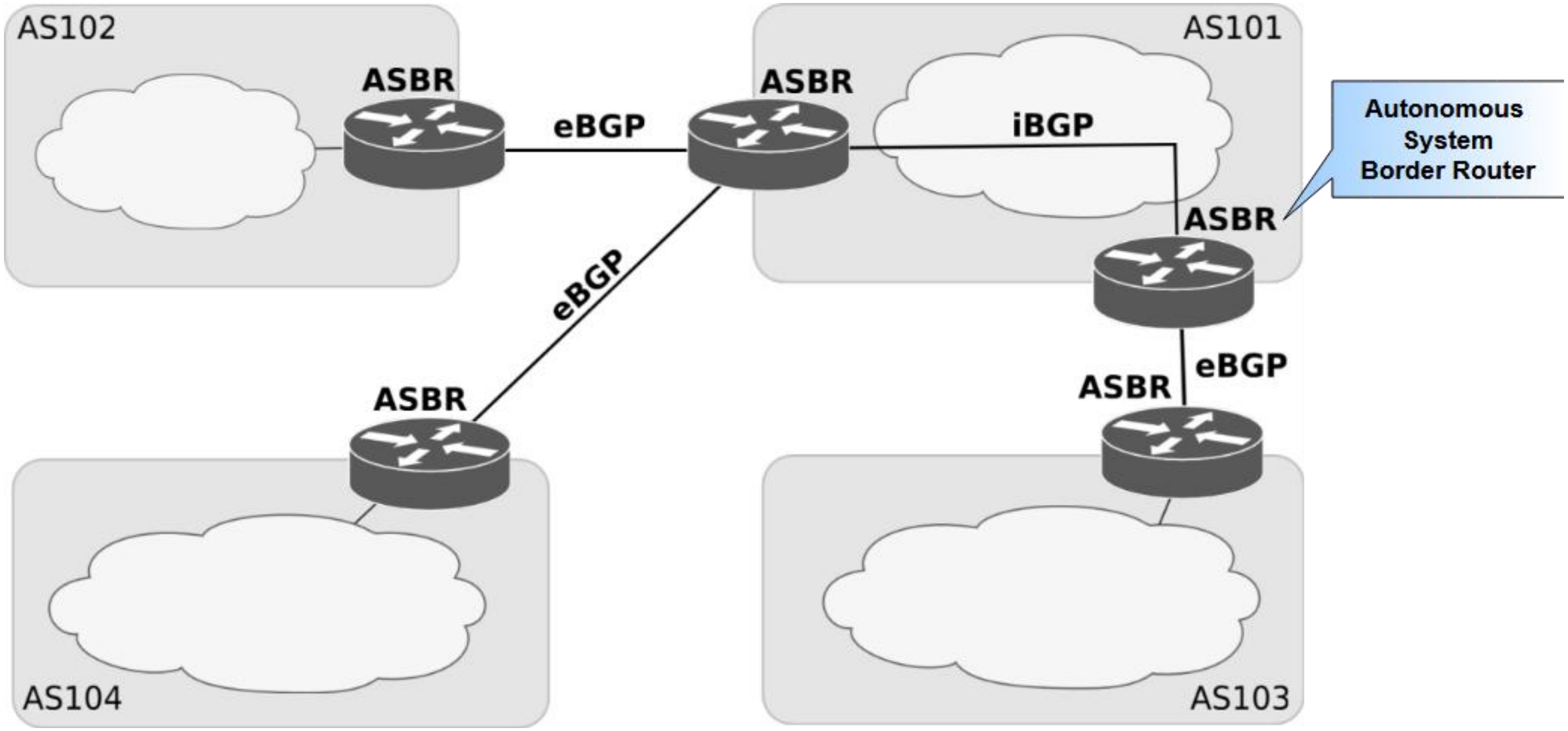
- RFC 4271 defines an ASN as a 2-bytes represented by a decimal
  - Private ASNs = 64512 through 65535
  - Public ASNs = 1 through 64511
  - 39000+ have already been allocated
  - Need to expand ASN size from 2-bytes to 4-bytes
- RFC 4893 defines BGP support for 4-bytes ASNs
  - Maximum ASN value is 4,294,967,295
  - Since 2009 that all new assigned ASNs are 4-byte by default, unless otherwise requested.
  - The 4-byte ASNs are represented by two decimals (2-bytes each). Notation:
    - <higher2bytes in decimal>.<lower2bytes in decimal>
    - Example1: AS 50000 is represented as “0.50000”
    - Example2: AS 65546 is represented as “1.10”

BGP Peer Relationship

- A BGP peer relationship is a connection between two routers that use BGP to exchange routing information
  - Usually configured by management
- Each peer relationship runs over TCP (port 179)
  - Ensures the reliable data delivery property of TCP
- BGP peers exchange all their routes when the peer relationship is established
  - Updates are also sent to peers when there is a topology change in the network or a change in routing policy
- BGP peers exchange periodic KEEPALIVE messages
  - To avoid extended periods of inactivity
  - Low keepalive intervals can be set if fast fail-over is required

Internal BGP (iBGP) & External BGP (eBGP)

- Routers that implement peer relationships are called Autonomous System Border Routers (ASBRs).
- Peer relationships can be established between:
  - ASBRs of the same AS (Internal BGP – iBGP)
  - ASBRs of different ASs (External BGP – eBGP)

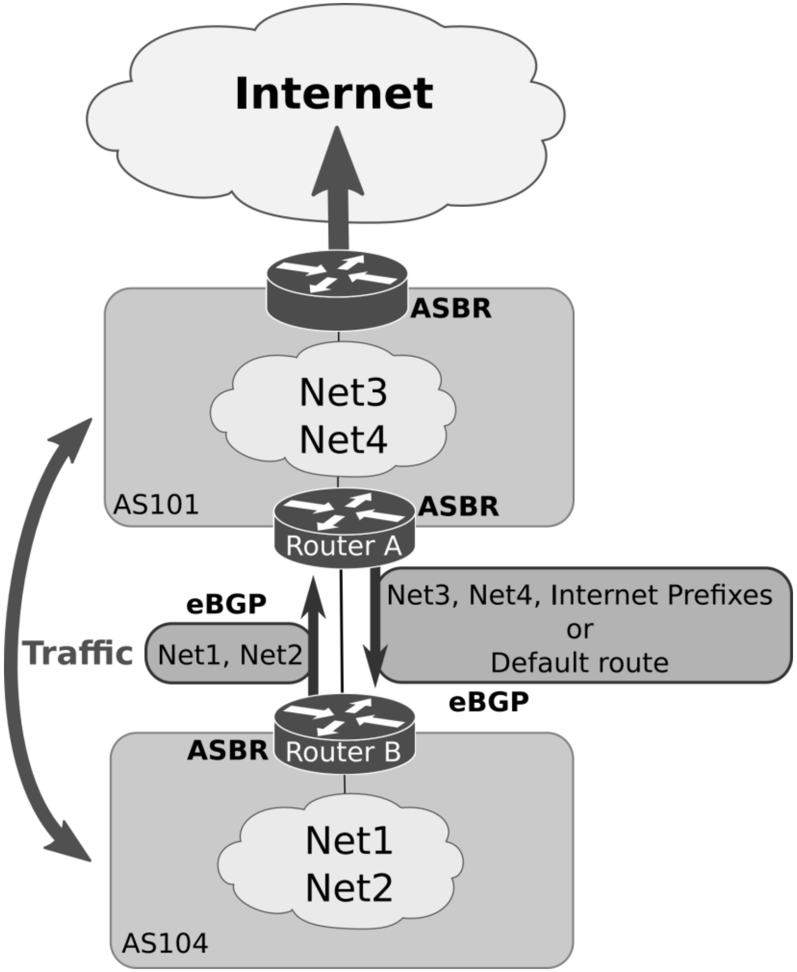


External and Internal BGP

- External BGP (eBGP) is used between ASs.
- Internal BGP (iBGP) is used within an AS.
- A BGP router forwards the routes learned from one eBGP peer to both the other eBGP and iBGP peers.
  - Filters can be used to modify this behaviour.
- A BGP router never forwards a route learned from one iBGP peer to another iBGP peer.
  - An exception is when a router is configured as route-reflector.
- **All ASBRs of an AS must maintain iBGP sessions with all other ASBRs in the same AS** (iBGP Mesh).
  - To obtain complete routing information from external routes.
  - Additional methods can be used to reduce iBGP Mesh complexity, like route reflectors and private ASs.

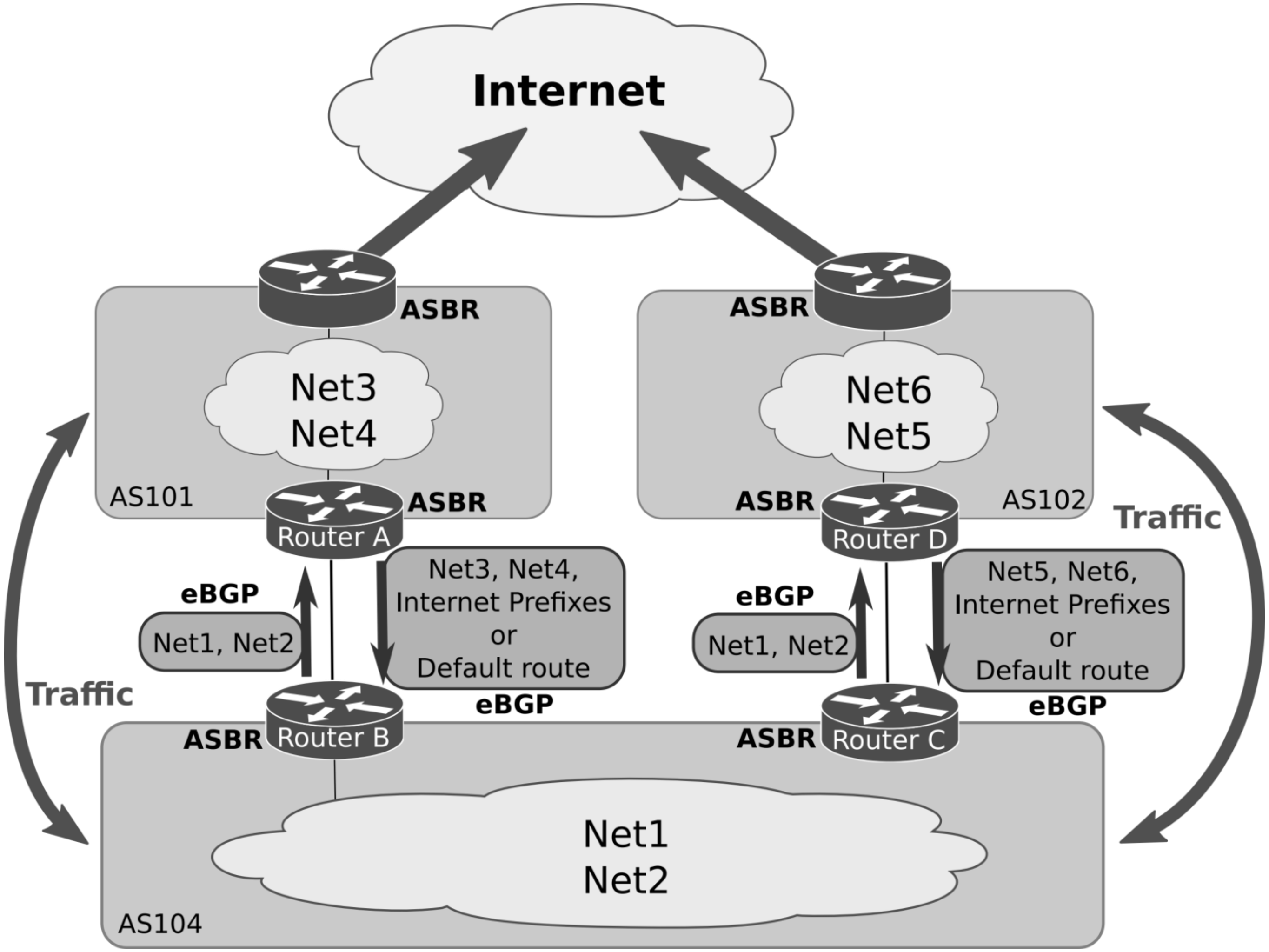
Single-homed (or Stub) AS

- AS has only one ASBR
  - Single Internet access from a single ISP



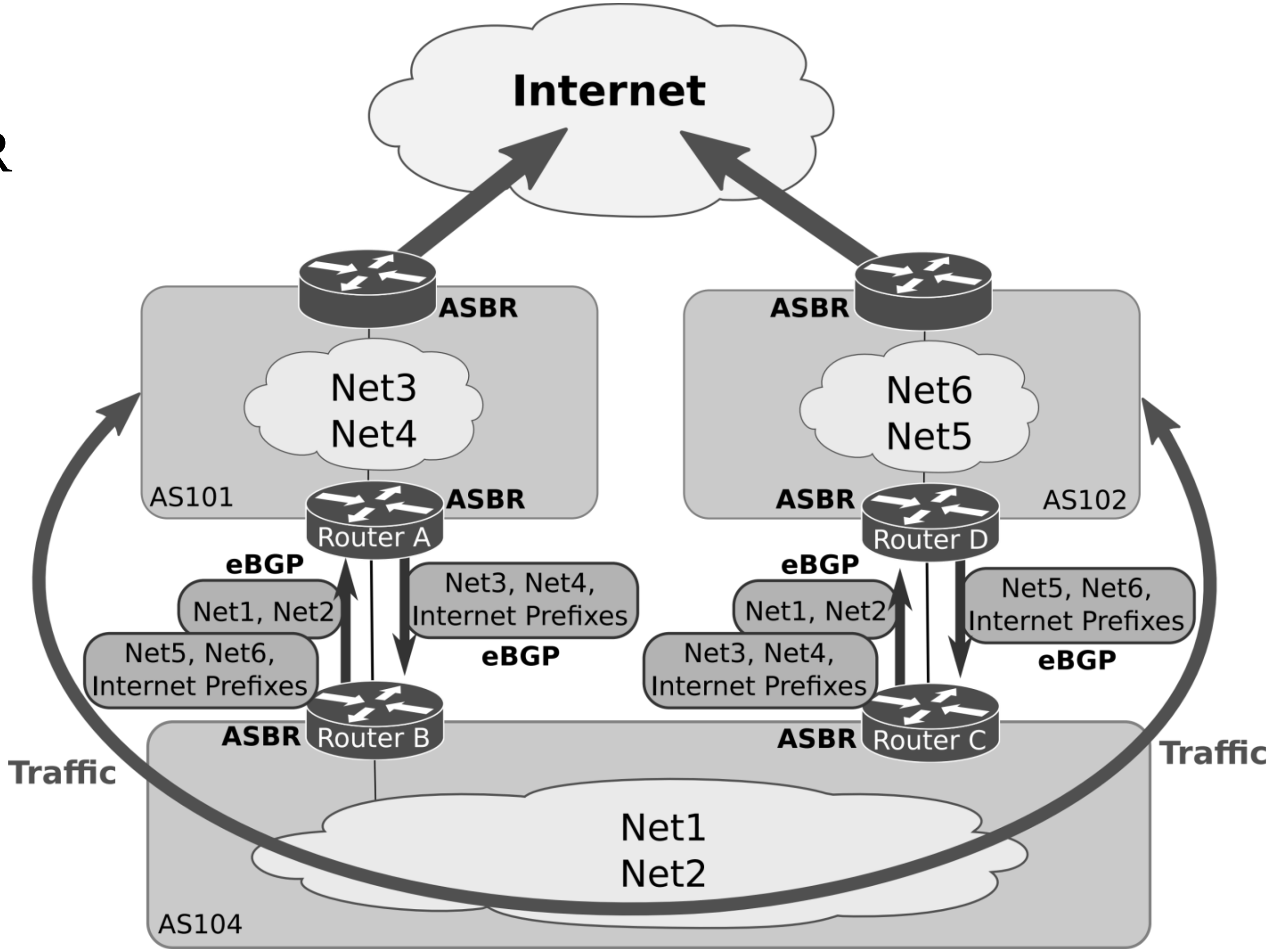
Multi-homed Non-transit AS

- AS has more than one ASBR
  - Multiple Internet accesses from multiple ISPs
- The AS does not support traffic between other ASs.



Multi-homed Transit AS

- AS has more than one ASBR
  - Multiple Internet accesses from multiple ISPs
- The AS supports traffic between the other ASs.

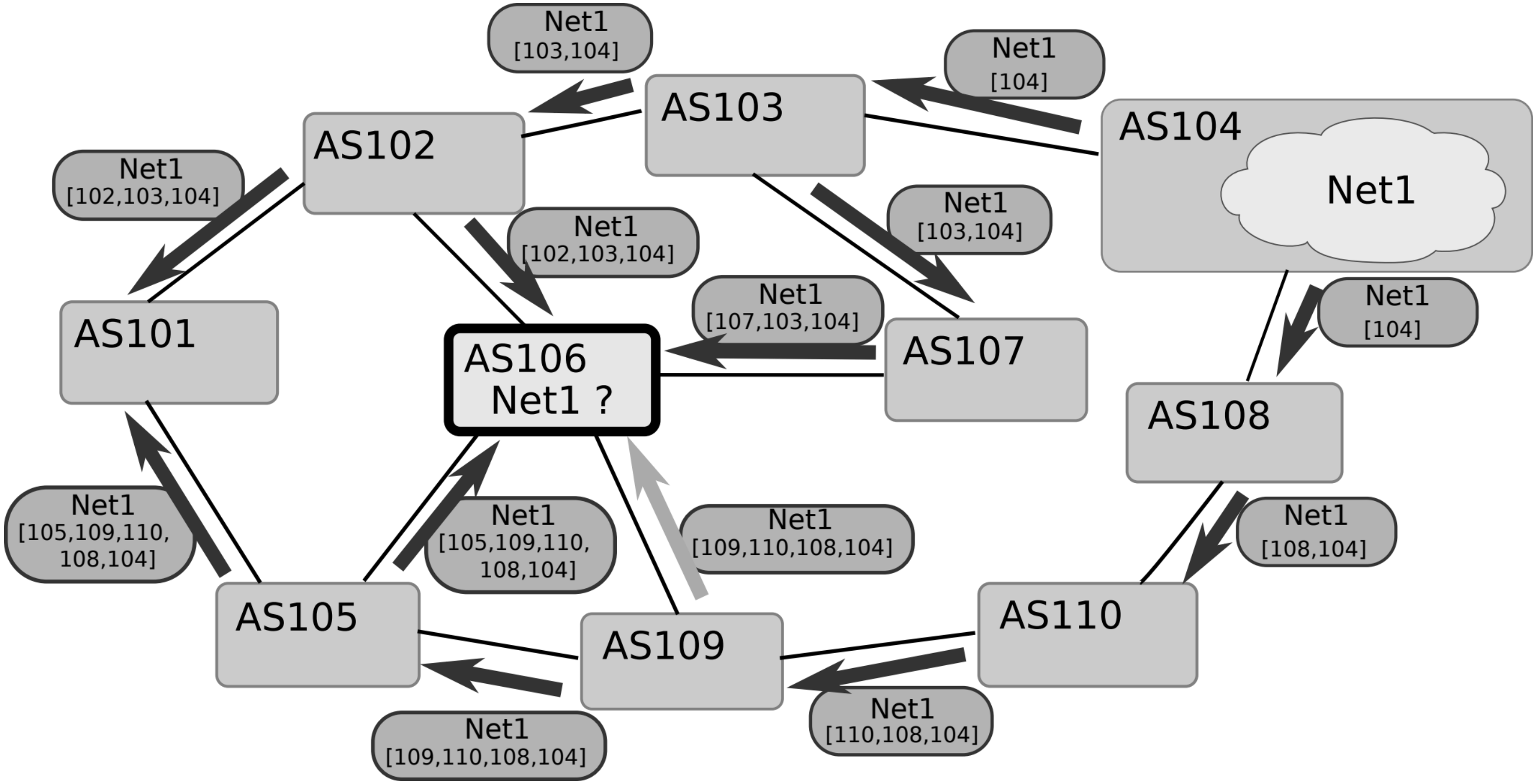


Path-Vector Protocol

- BGP is a path-vector protocol
- It works similarly as a distance-vector protocol, but each announced route contains the sequence of traversed ASs (identified by their ASNs)
  - Provides loop detection and policy-based routing selection
  - Illustration in the next slide
- An eBGP sending peer adds its own ASN to the AS sequence before forwarding the route announcement to its peer
- An iBGP sending peer does not modify the sequence because it is sending the route announcement to a peer within the same AS
  - The AS sequence list cannot be used to detect the iBGP routing loops

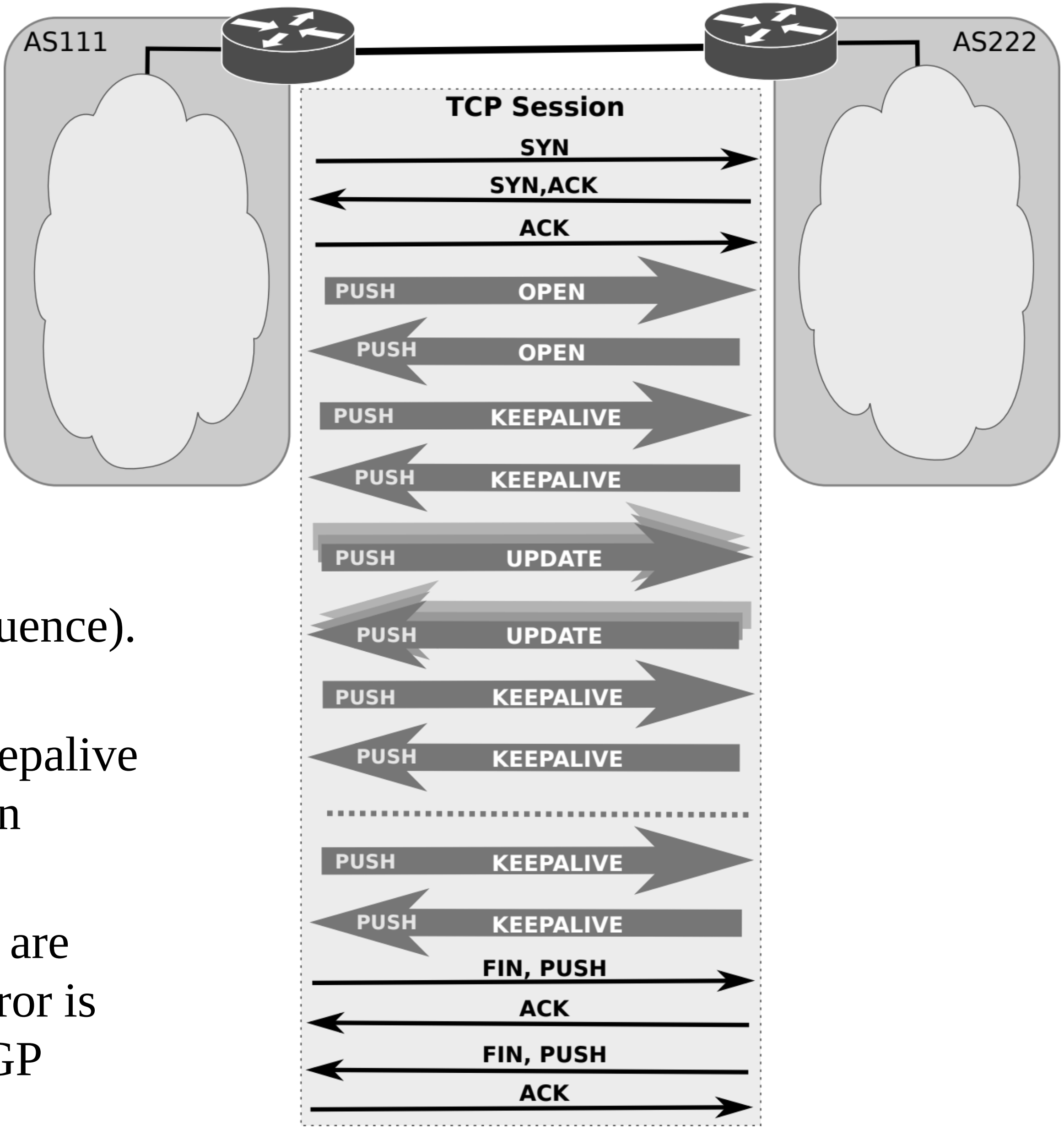
Path-Vector Protocol Illustration

- AS106 receives from its eBGP peers a route announcement of **Net1** (that belongs to AS104) with their routing sequence towards AS104
- Based on its routing configuration, AS106 selects the eBGP peer to forward the traffic towards **Net1**



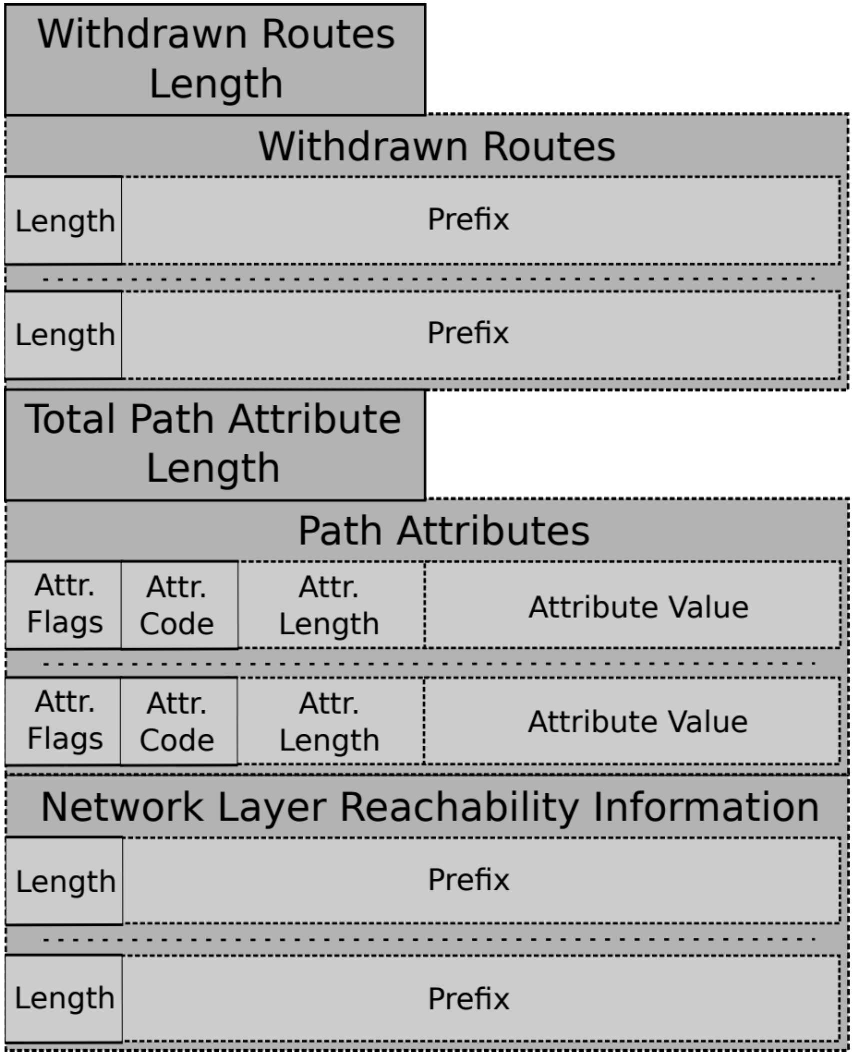
BGP Messages

- OPEN messages are used to establish the BGP session.
- UPDATE messages are used to send IP prefixes, along with their associated BGP attributes (such as the AS_PATH with the AS sequence).
- KEEPALIVE messages are exchanged whenever the keepalive period is reached, without an update being exchanged.
- NOTIFICATION messages are sent whenever a protocol error is detected, after which the BGP session is closed.



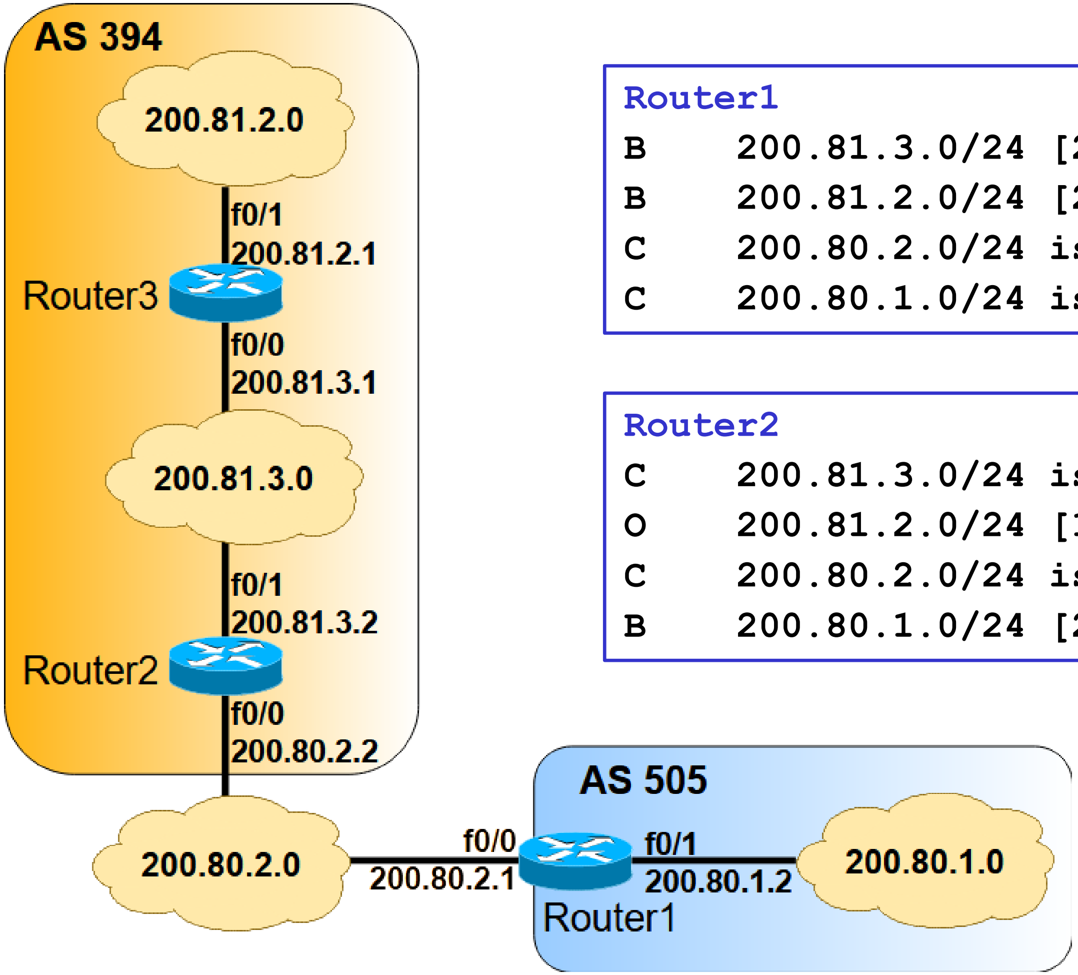
BGP UPDATE Message

- Withdrawn Routes
  - list of IP prefixes no longer accessible.
- Path Attributes
  - parameters used to compute the best routing path.
- Network Layer Reachability Information
  - list of known IP prefixes characterized by the parameters specified in the Path Attributes.



Example

- Router1 is the ASBR of AS 505
- Router2 is the ASBR of AS 394
- eBGP session configured between Router1 and Router2
- AS 394 internal routing is configured with OSPF

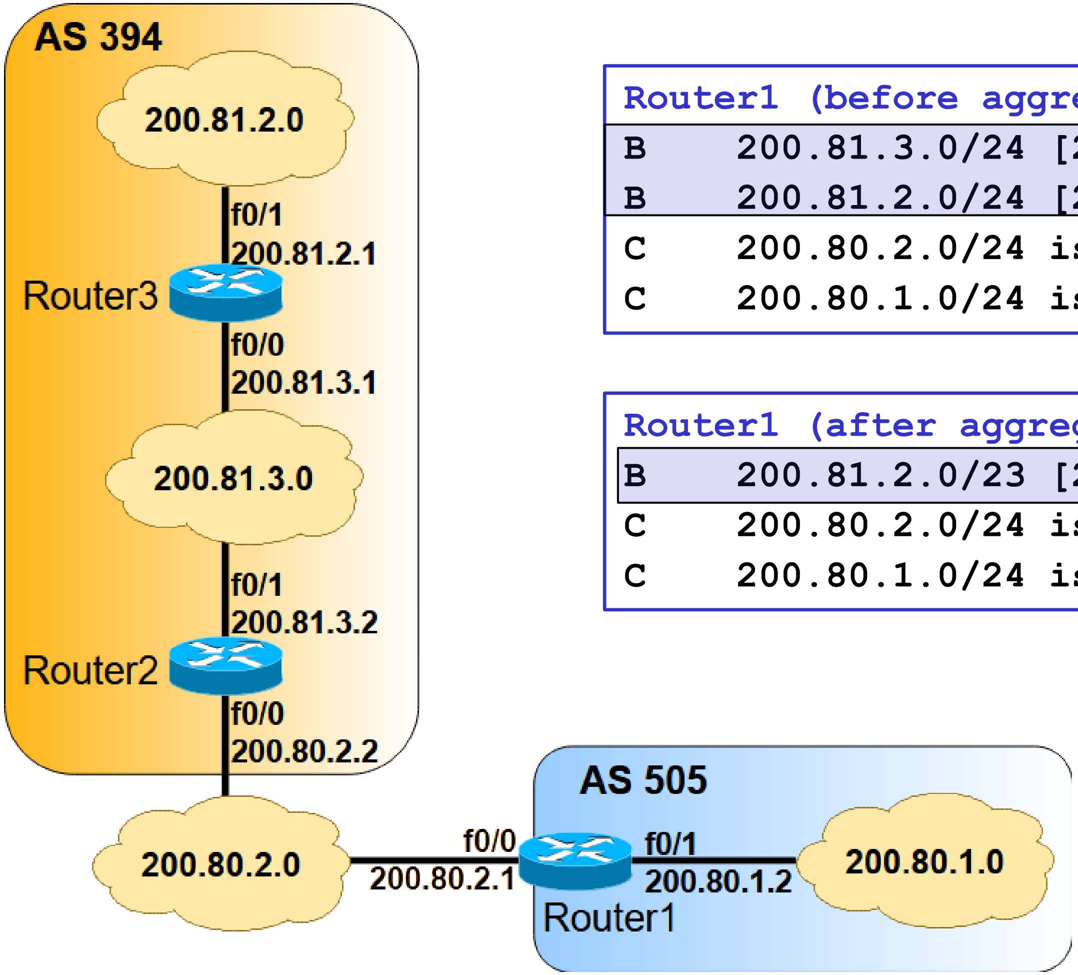


```
Router1
B     200.81.3.0/24 [20/0] via 200.80.2.2, 00:01:58
B     200.81.2.0/24 [20/0] via 200.80.2.2, 00:01:57
C     200.80.2.0/24 is directly connected, f0/0
C     200.80.1.0/24 is directly connected, f0/1
```

```
Router2
C     200.81.3.0/24 is directly connected, f0/1
O     200.81.2.0/24 [110/2] via 200.81.3.1, 00:01:12
C     200.80.2.0/24 is directly connected, f0/0
B     200.80.1.0/24 [20/0] via 200.80.2.1, 00:00:29
```

Example with an Aggregated Prefix

- 200.81.2.0/24 and 200.81.3.0/24 are represented by the aggregated prefix 200.81.2.0/23
- Prefix 200.81.2.0/23 is announced by Router2 to its eBGP peer Router1



```
Router1 (before aggregation)
B     200.81.3.0/24 [20/0] via 200.80.2.2, 00:01:58
B     200.81.2.0/24 [20/0] via 200.80.2.2, 00:01:57
C     200.80.2.0/24 is directly connected, f0/0
C     200.80.1.0/24 is directly connected, f0/1
```

```
Router1 (after aggregation)
B     200.81.2.0/23 [20/0] via 200.80.2.2, 00:01:03
C     200.80.2.0/24 is directly connected, f0/0
C     200.80.1.0/24 is directly connected, f0/1
```

BGP Attributes

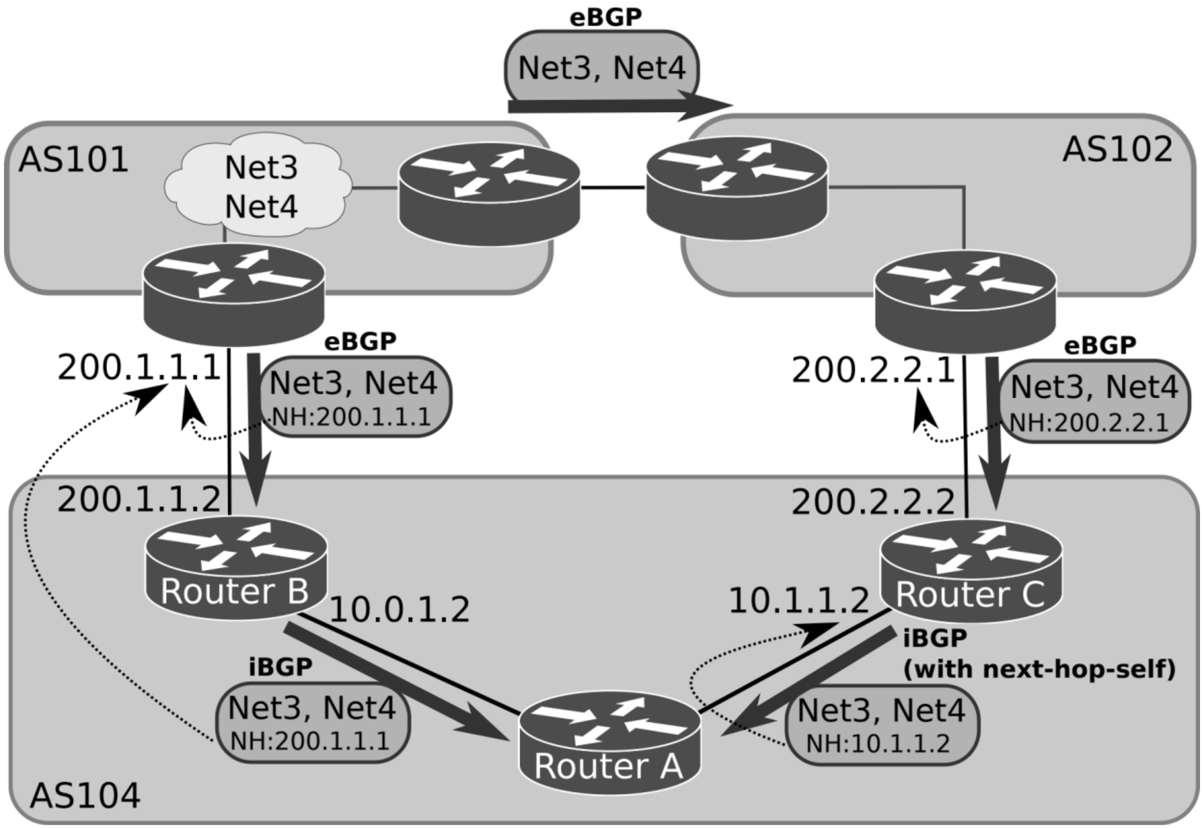
- BGP attributes are metrics that describe the characteristics of BGP routes and are used to select routing paths.
- Attributes are included in the BGP UPDATE messages when announcing IP prefixes. There are different attribute categories:
  - Well-known Mandatory (always included in BGP UPDATEs)
    - AS_PATH, ORIGIN, Next-Hop
  - Well-known Discretionary (may or may not be included in BGP UPDATEs)
    - Local Preference, Atomic Aggregate
  - Optional Transitive (may not be supported by all BGP implementations)
    - If not supported, BGP routers forward them unchanged.
    - Aggregator, Community, AS4_Aggregator, AS4_PATH
  - Optional Non-transitive (may not be supported by all BGP implementations)
    - Multi-exit-discriminator (MED), Originator, Cluster-ID
  - Cisco-defined (local to router, not included in BGP UPDATEs)
    - Weight

BGP Well-known Mandatory Attributes: AS_PATH and ORIGIN

- AS_PATH
  - When a route advertisement passes through an autonomous system, the AS number is added to an ordered list of AS numbers that the route advertisement has traversed (the basis of the path-vector property).
- ORIGIN
  - Indicates how BGP has learned the route in the originating AS. It can take one of three possible values:
    - IGP (value 0) is used if the route is interior to the originating AS, resulting from an explicit inclusion of the network prefix within the BGP routing process.
    - EGP (value 1) is used if the route is learned by other EGP protocol (no longer used in modern networks).
    - INCOMPLETE (value 2) is used if the route is learned by other means, namely, route redistribution from other routing protocols into the BGP routing process.

BGP Well-known Mandatory Attribute: Next-Hop

- The eBGP Next-Hop attribute is the IP address used to reach the advertising router, i.e., the IP address of the eBGP peer.
  - By default, the IP networks of eBGP sessions do not belong to any AS.
  - So, the IP address of the eBGP Next-Hop attribute is not known inside an AS.
- In the BGP router, its IP address inside the AS must be configured as the Next-Hop attribute to its iBGP neighbours.



BGP Optional Transitive Attribute: AS4_PATH

- AS4_PATH attribute has the same semantics as the AS_PATH attribute, except that it is Optional Transitive, and it carries 4-bytes ASNs.
  - 4-byte AS support is advertised via BGP capability negotiation in the initial exchanged OPEN messages
  - Peers supporting 4-byte AS are known as NEW BGP peers
  - Peers not supporting 4-byte AS are known as OLD BGP peers
- New Reserved ASN: AS_TRANS = 23456
  - 2-byte placeholder for a 4-byte ASN in the AS_PATH attribute
- A NEW BGP router receives UPDATEs from a NEW BGP peer
  - It decodes each ASN as 4-bytes in the AS_PATH attribute
- A NEW BGP router receives UPDATEs from an OLD BGP peer
  - AS_PATH and AS4_PATH must be merged to form the correct AS_PATH
- Merging AS_PATH and AS4_PATH
  - AS_PATH : { 275 , 250 , 225 , 23456 , 23456 , 200 , 23456 , 175 }
  - AS4_PATH : { 100.1 , 100.2 , 200 , 100.3 , 175 }
  - Merged AS-PATH : { 275 , 250 , 225 , 100.1 , 100.2 , 200 , 100.3 , 175 }

Illustration of AS_PATH and AS4_PATH Attributes

- Since AS4_PATH is an Optional Transitive attribute, BGP routers not supporting it (i.e., the OLD BGP peers) forward them unchanged (as in AS 13 and AS 14 of the figure)

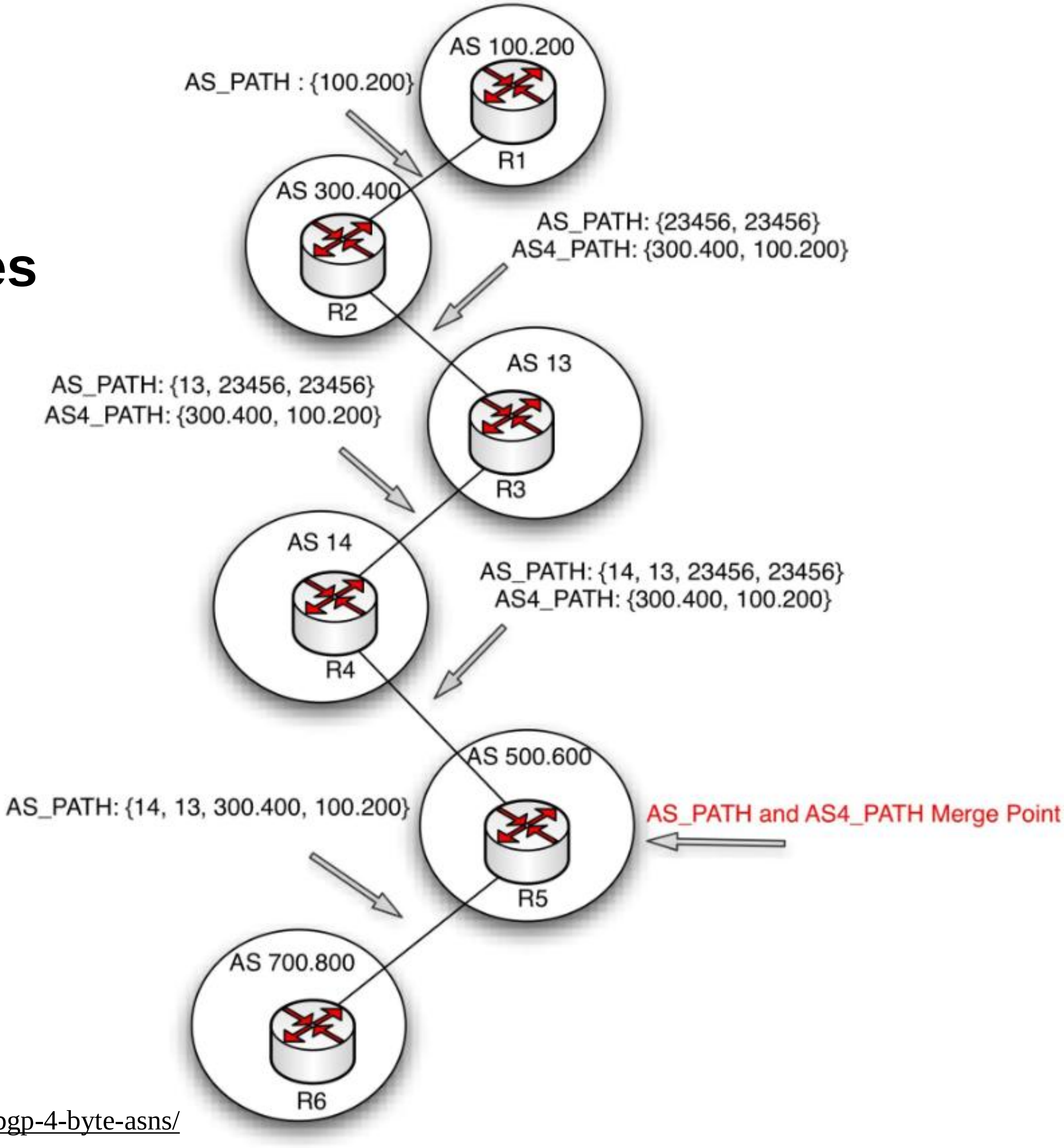
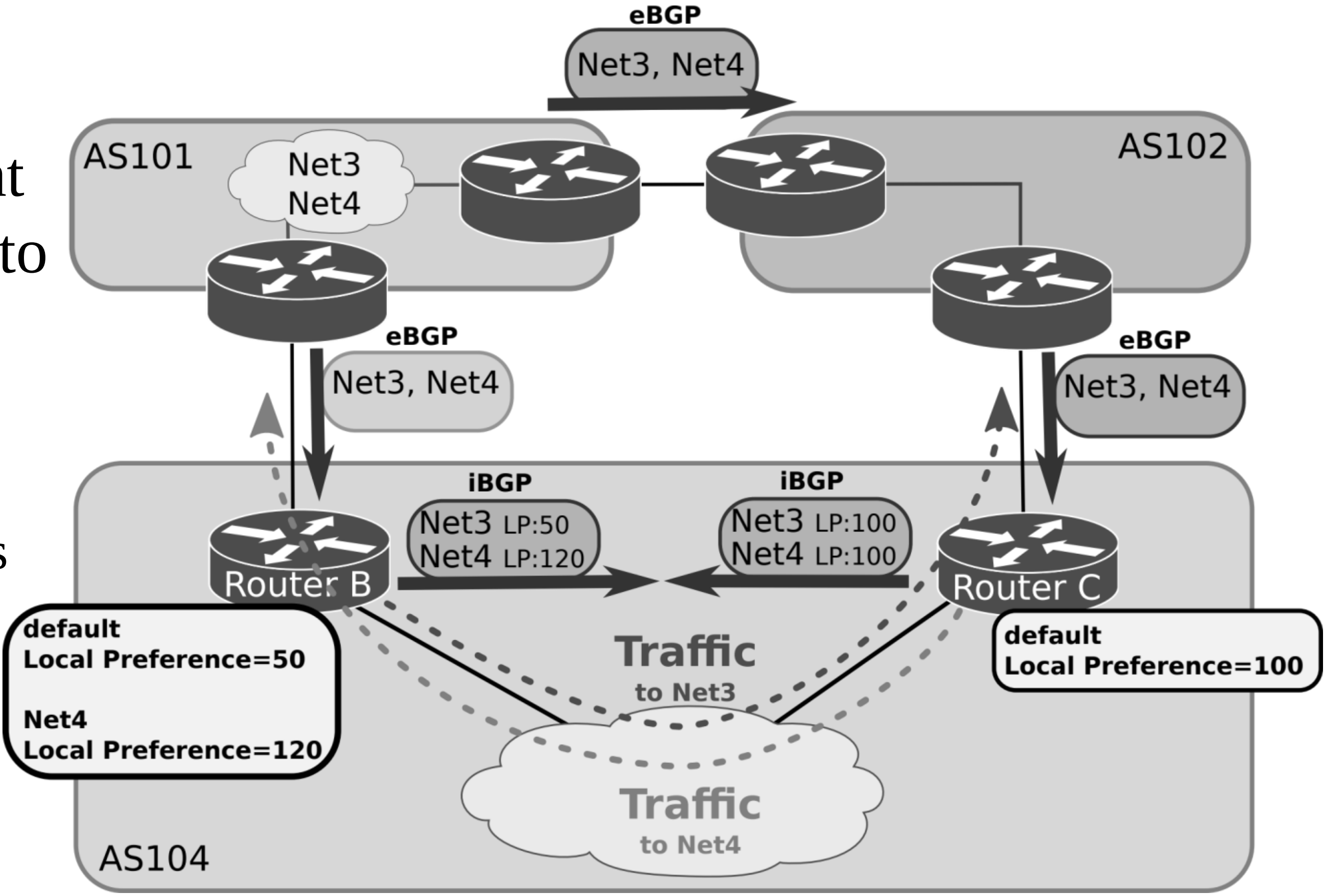

AS 100.200
R1
AS_PATH : {100.200}
AS 300.400
R2
AS_PATH: {23456, 23456}
AS4_PATH: {300.400, 100.200}
AS 13
R3
AS_PATH: {13, 23456, 23456}
AS4_PATH: {300.400, 100.200}
AS 14
R4
AS_PATH: {14, 13, 23456, 23456}
AS4_PATH: {300.400, 100.200}
AS 500.600
R5
AS_PATH and AS4_PATH Merge Point
AS_PATH: {14, 13, 300.400, 100.200}
AS 700.800
R6

Figure from: https://networkn3rd.wordpress.com/2013/02/01/bgp-4-byte-asns/

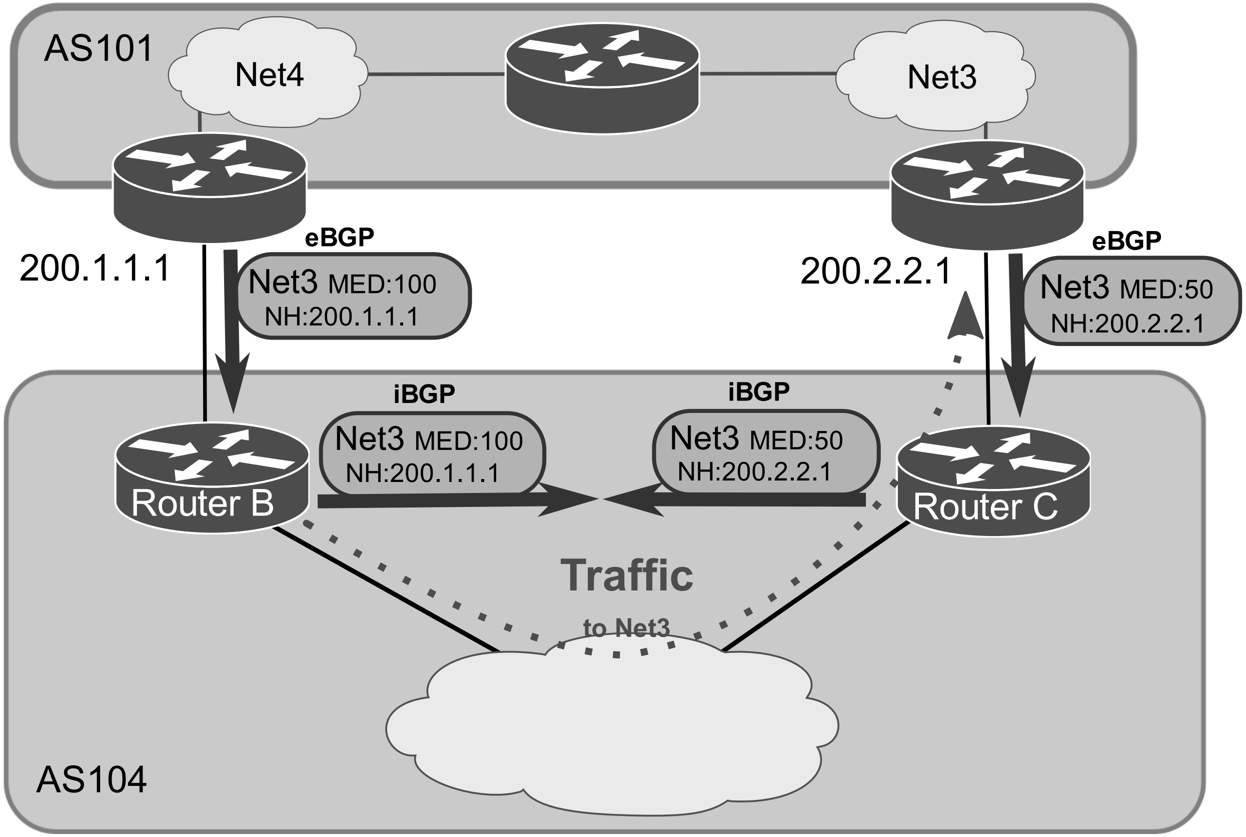
BGP Well-known Discretionary Attribute: Local Preference

- Local Preference (LP) is used to select the exit point from the local AS to outside routes.
  - **Higher value** is preferred.
  - The LP attribute is propagated throughout the local AS.
  - The LP value can be different, for different routes.



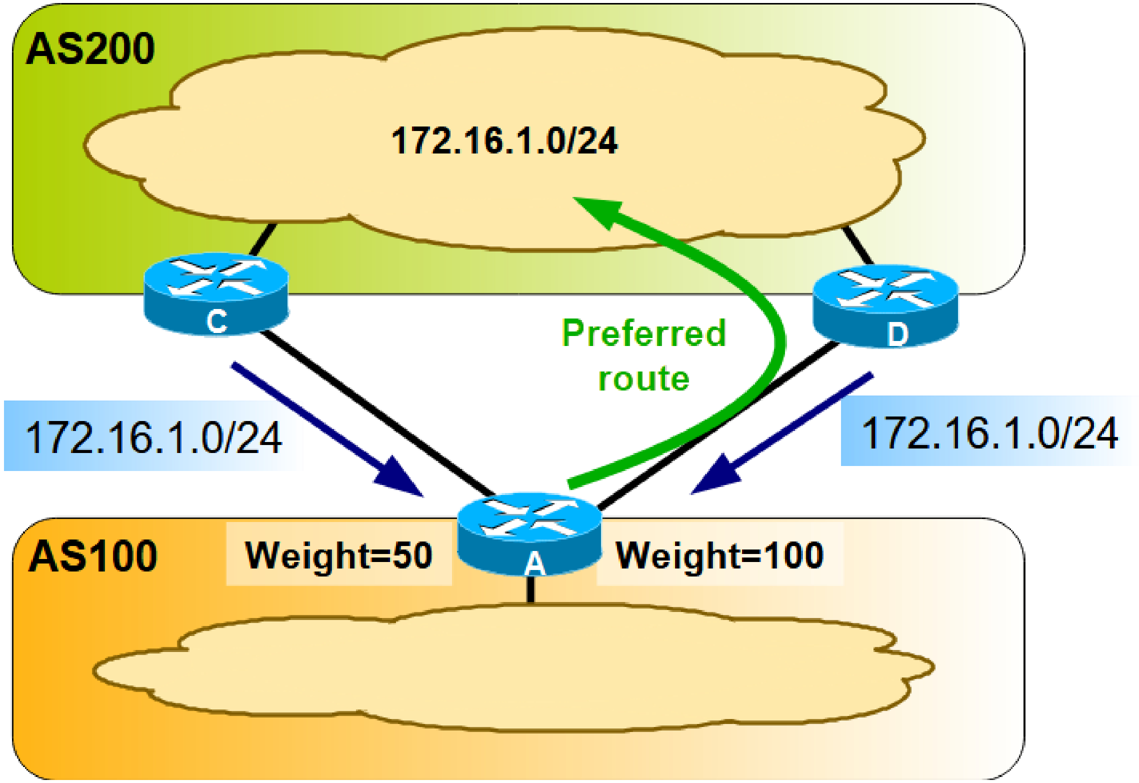
BGP Optional Non-transitive Attribute: Multi-Exit Discriminator (MED)

- MED is used as a suggestion to a peer AS (aiming to influence the incoming traffic).
  - **Lower value** is preferred.
  - The external AS receiving the MED values may be using other BGP attributes for route selection.



CISCO Weight Attribute

- Weight is a Cisco-defined local attribute.
  - The weight attribute is not advertised to BGP peers.
  - It is useful when a BGP router has multiple eBGP peers.
- If the router learns more than one route to the same destination prefix, the route with **the highest weight** is preferred.

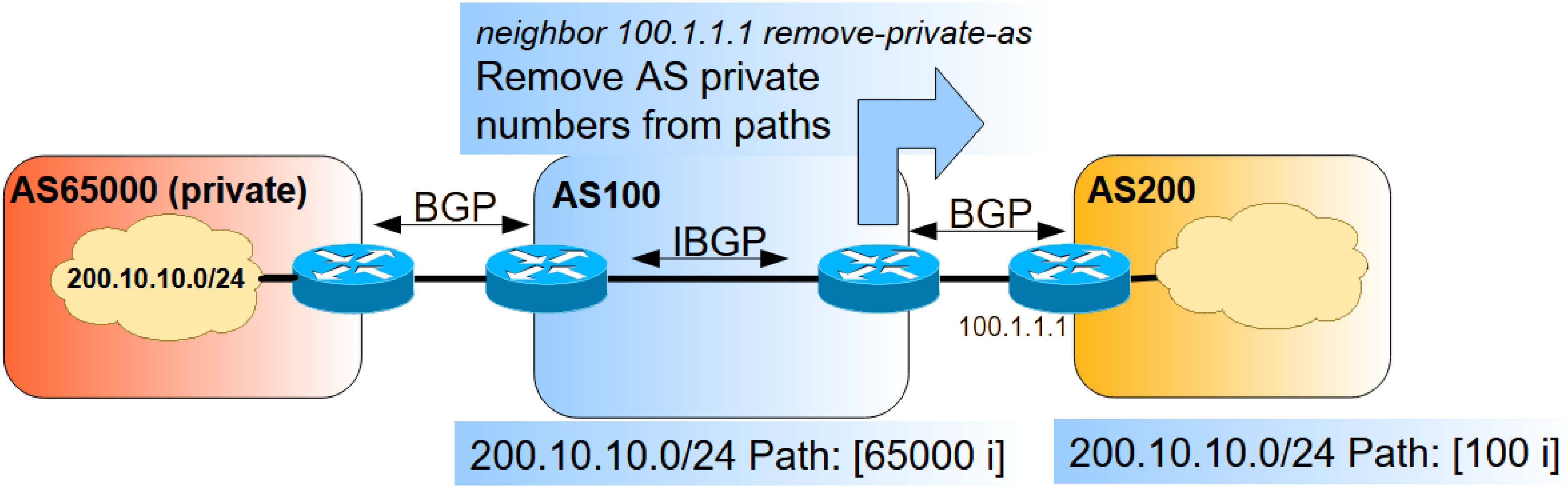


BGP Path Selection

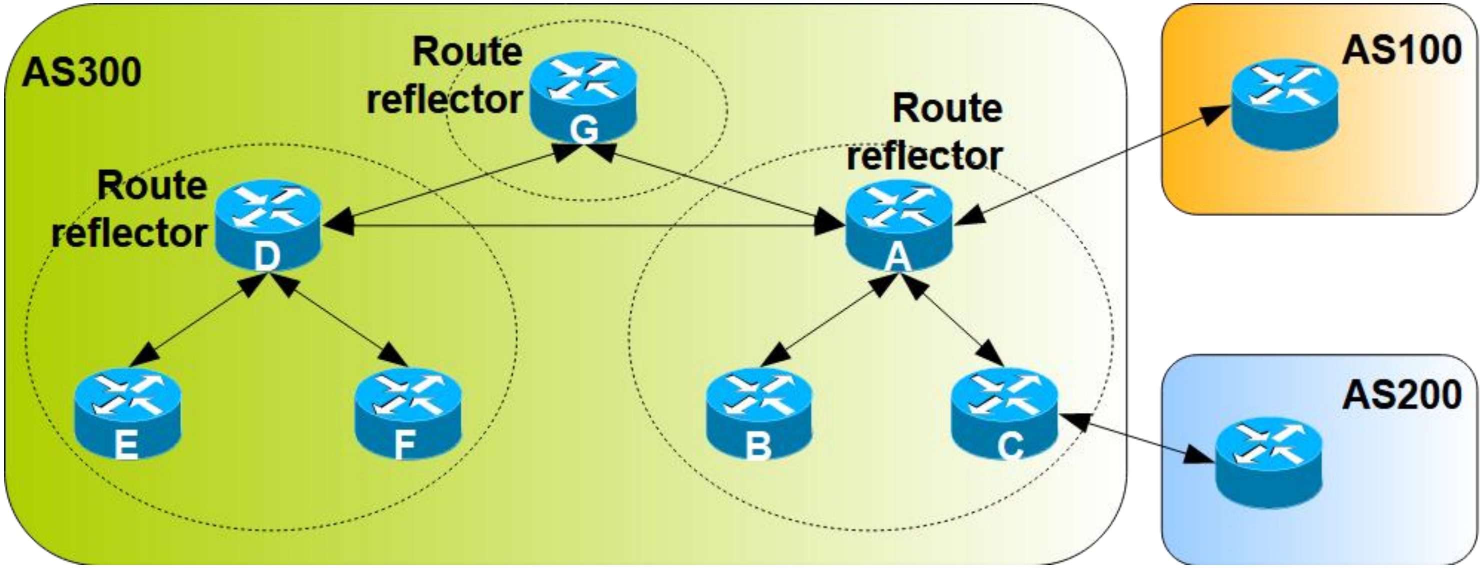
- BGP may receive advertisements for the same network prefix from multiple sources and selects only one path.
- BGP sets the selected path in the IP routing table of the router and propagates the path to its BGP neighbours.
- The path selection uses the following criteria, in the order:
  - Largest weight value (Cisco only)
  - Largest Local Preference value
  - Path that was originated locally (i.e., belongs to the local AS)
  - Shortest path (i.e., with a shorter AS_PATH attribute length)
  - Lowest ORIGIN value
  - Lowest MED value
  - External path (through eBGP) over internal path (through iBGP)
  - Lowest IGP Metric (the path with the lowest IGP cost to the Next Hop)
  - Lowest BGP Router ID

BGP Private Autonomous System (AS)

- When a customer network is large, the network might be set as an Autonomous System with a given ASN:
  - Assigning a **Public** ASN in the range of 1 to 64511
    - Required when the customer network connects to different ISPs
  - Assigning a **Private** ASN in the range of 64512 to 65535.
    - Possible when the costumer networks connects to a single ISP (not recommended if the costumer plans to connect to other ISPs in the future)

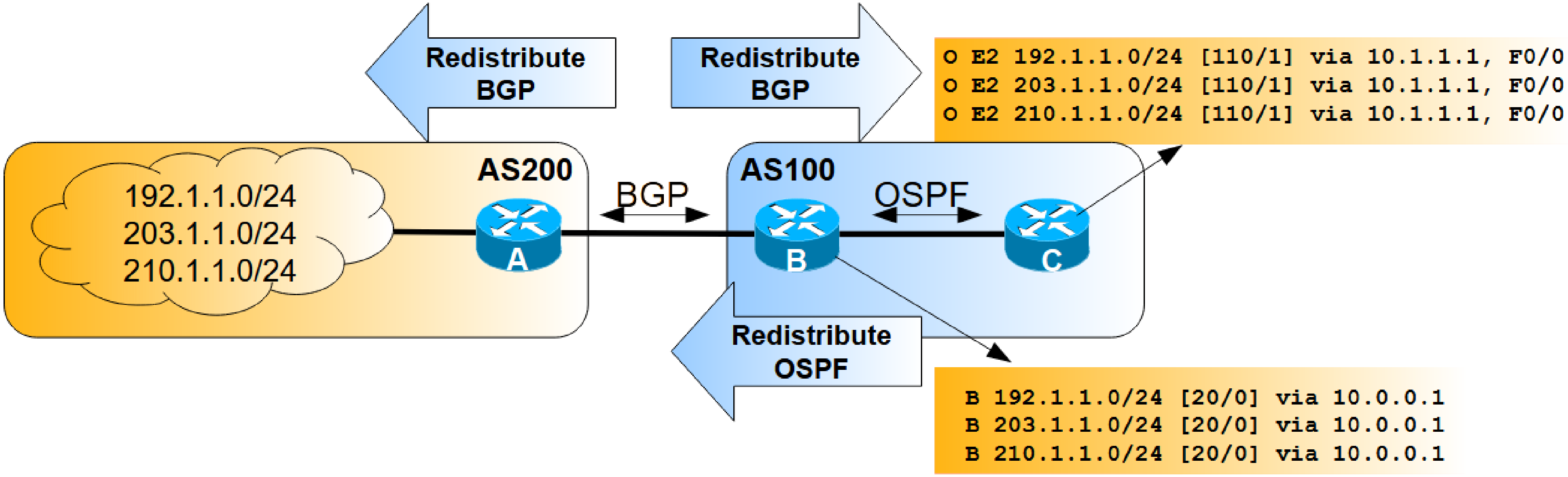

neighbor 100.1.1.1 remove-private-as
Remove AS private numbers from paths
AS65000 (private)
200.10.10.0/24
BGP
AS100
IBGP
BGP
AS200
100.1.1.1
200.10.10.0/24 Path: [65000 i]
200.10.10.0/24 Path: [100 i]

BGP Route Reflectors



- Without a route reflector, the network requires a full iBGP mesh within AS300 in the above example.
- The route reflector and its clients are called a cluster
  - Router A is configured as a route reflector with clients B and C, iBGP peering between Routers B and C is not required.
  - Router D is configured as a route reflector with clients E and F, iBGP peering between Routers E and F is not required.
- Route reflector routers are connected in a full iBGP mesh.

Routes Redistribution



```
O E2 192.1.1.0/24 [110/1] via 10.1.1.1, F0/0
O E2 203.1.1.0/24 [110/1] via 10.1.1.1, F0/0
O E2 210.1.1.0/24 [110/1] via 10.1.1.1, F0/0
```

```
B 192.1.1.0/24 [20/0] via 10.0.0.1
B 203.1.1.0/24 [20/0] via 10.0.0.1
B 210.1.1.0/24 [20/0] via 10.0.0.1
```

- Redistributing IGP routes by BGP will:
  - Simplify BGP configuration (advantage)
  - BGP announces only internal networks with connectivity (advantage)
  - Prevent announcement of aggregated prefixes (disadvantage)
- Redistributing BGP routes by IGP protocols will:
  - Make internal routes know all external routes (advantage)
  - Increase routing tables size in internal routers (disadvantage)

BGP Filtering

- By default, BGP processes announce every network path received from other BGP peers or redistributed by IGP.
- Sending and accepting BGP updates can be controlled by using different filtering methods.
  - Route-maps, prefix-lists, distribution-lists.
- BGP updates can be filtered based on:
  - Route information, Path information (AS_PATH attribute), etc…
- Most common actions:
  - Configure Local Preference values to define the preferred paths.
  - Block/Deny paths to prevent undesired routing paths of being selected.
- Best basic practices:
  - Block all IPv4 private networks,
  - Announce default routes only to AS peers with a traffic transport contract.
  - Accept default routes only from AS peers providing a traffic transport service.

Multi-Protocol Border Gateway Protocol (MP-BGP)

- BGP carries only routing information about IPv4 unicast
- MP-BGP is an extension to the BGP protocol:
  - It behaves as BGP for IPv4 unicast and, therefore, a BGP peer relationship is possible between a BGP router and a MP-BGP router.
- MP-BGP carries routing information about different protocols/families:
  - IPv4 and IPv6 (both Unicast and Multicast)
  - Multi-Protocol Label Switching (MPLS) VPNs (IPv4 and IPv6)
  - 6PE: IPv6 packets transported over an IPv4 MPLS backbone
  - 6VPE: Multiple IPv6 VPNs created over an IPv4 MPLS backbone
  - Ethernet VPN (EVPN): a protocol to distribute IP and MAC addresses
    - it is a standards-based solution for VXLANs used in Data Centres
- MP-BGP exchanges Multi-Protocol NLRI (Network Layer Reachability Information)

MP-BGP New Attributes

- MP-BGP introduces new non-transitive and optional attributes:
  - MP_REACH_NLRI
    - Carry the set of reachable destinations together with the next- hop information to be used for forwarding to these destinations
  - MP_UNREACH_NLRI
    - Carry the set of unreachable destinations
- The new attributes contain one or more triples
  - Address Family Identifier (AFI) with SAFI (Subsequent AFI)
    - AFI is IPv4 or IPv6, for example
    - In the case IPv4 or IPv6, SAFI is unicast or multicast
    - AFI (and SAFI) provide the protocol information associated with the content of the NLRI (Network Layer Reachability Information)
  - Next-Hop
    - Next-hop address must be of the family identified by the AFI

MP-BGP Negotiation Capabilities

- Like BGP, MP-BGP routers establish BGP peer relationships through the exchange of OPEN messages
  - MP-BGP introduces a new optional parameter: CAPABILITIES
- Some important CAPABILITIES parameters:
  - Multi-Protocol extensions (AFI/SAFI): identify the supported AFI/SAFI protocol/families
  - Route Refresh: allows a BGP router to request the resend of the reachability information from the BGP peer router
  - Outbound Route Filtering (ORF): allows a BGP router to send (to the BGP peer router) prefix information that will be denied if announced in the inbound direction
- MP-BGP negotiation is the process of determining the supported capabilities in a BGP peer relationship:
  - process done by identifying the CAPABILITIES parameters announced in the OPEN messages that are common to both BGP peers.