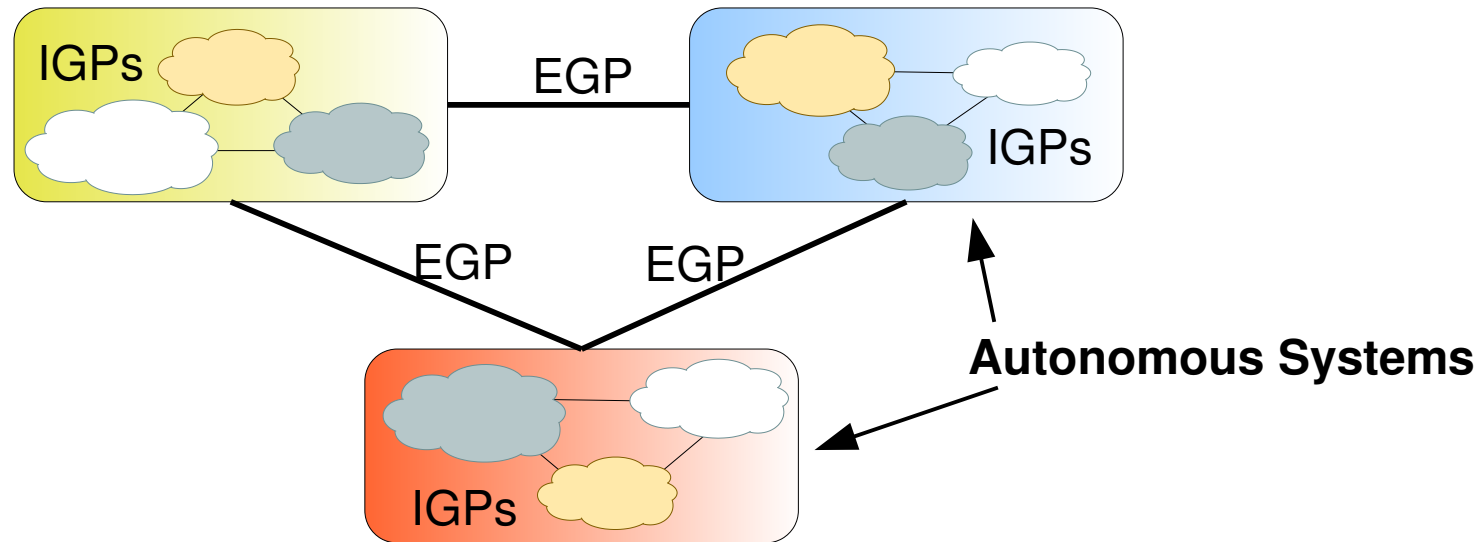


# External Routing (BGP and MP-BGP)

**Arquitectura de Redes Avanzadas**

# Border Gateway Protocol (BGP)



- Border Gateway Protocol Version 4 of the protocol (BGP4) was deployed in 1993 and currently is the protocol that assures Internet connectivity
- BGP is mainly used for routing between Autonomous Systems
- Autonomous System (AS) is a network under a single administration
  - One or more network operators with a common well defined global routing policy

# AS Numbers

- Allocated ID by InterNIC and is globally unique
- RFC 4271 defines an AS number as 2-bytes
  - Private AS Numbers = 64512 through 65535
  - Public AS Numbers = 1 through 64511
    - ➔ 39000+ have already been allocated
    - ➔ We will eventually run out of AS numbers
- Need to expand AS size from 2-bytes to 4-bytes
- RFC4893 defines BGP support for 4-bytes AS numbers
  - 4,294,967,295 AS numbers
  - As of January 1, 2009, all new Autonomous System numbers issued will be 4-byte by default, unless otherwise requested.
  - The full binary 4-byte AS number is split two words of 16 bits each
    - ➔ Notation:
    - ➔ <higher2bytes in decimal>.<lower2bytes in decimal>
    - ➔ Example1: AS 65546 is represented as “1.10”
    - ➔ Example2: AS 50000 is represented as “0.50000”
  - Cannot have a “flag day” solution



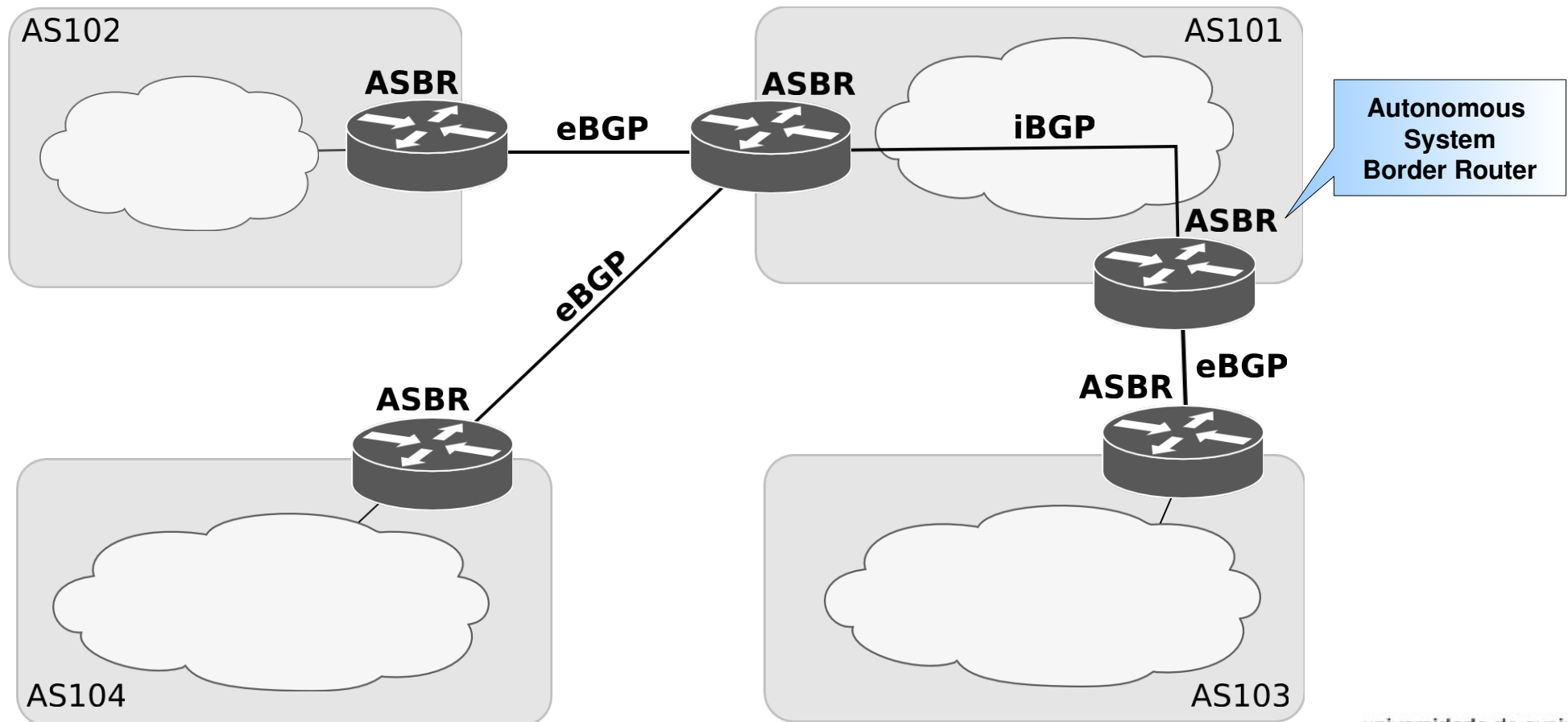
# BGP Neighbor Relationships

- Often called peering
  - Usually manually configured into routers by the administrator
- Each neighbor session runs over TCP (port 179)
  - Ensures reliable data delivery
- Peers exchange all their routes when the session is first established
- Updates are also sent when there is a topology change in the network or a change in routing policy
- BGP peers exchange session KEEPALIVE messages
  - To avoid extended periods of inactivity.
  - Low keepalive intervals can be set if a fast fail-over is required



# Internal BGP (iBGP) & External BGP (eBGP)

- Neighbor relations can be established between
  - Same AS routers (Internal BGP – iBGP).
  - Different AS routers (External BGP – eBGP).
- Routers that implement neighbor relations are called an Autonomous System Border Router (ASBR).



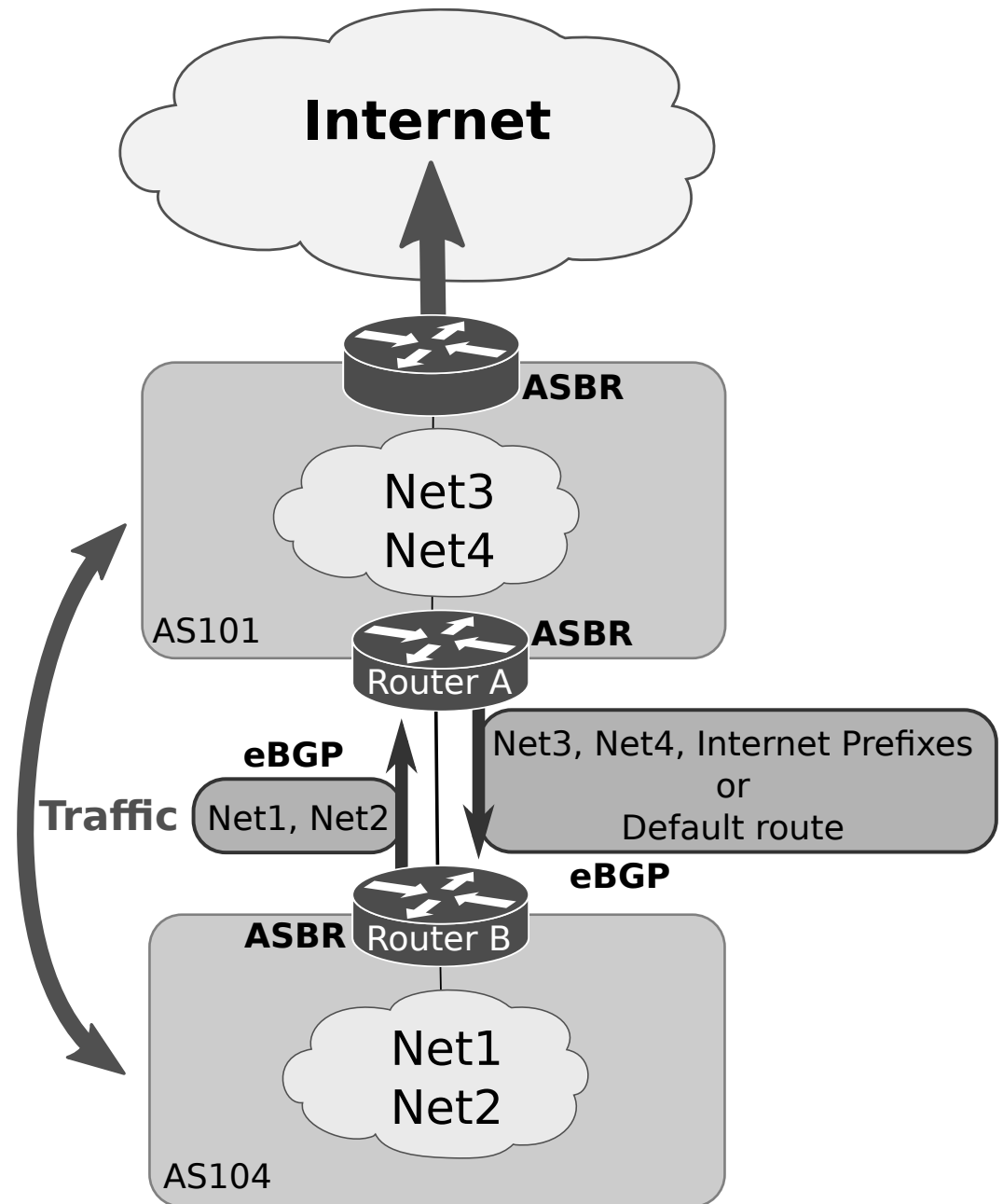
# External and Internal BGP

- External BGP (eBGP) is used between AS.
- Internal BGP (iBGP) is used within AS.
- A BGP router never forwards a path learned from one iBGP peer to another iBGP peer even if that path is the best path.
  - An exception is when a router is configured as route-reflector.
- A BGP forward the routes learned from one eBGP peer to both eBGP and iBGP peers.
  - Filters can be used to modify this behavior.
- iBGP routers in an AS **must maintain an iBGP session with all other iBGP routers** in the AS (iBGP Mesh).
  - To obtain complete routing information about external networks.
  - Most networks also use an IGP, such as OSPF.
  - Additional methods can be used to reduce iBGP Mesh complexity.
    - ➔ Route reflectors, private AS, ...



# Single-homed (or Stub) AS

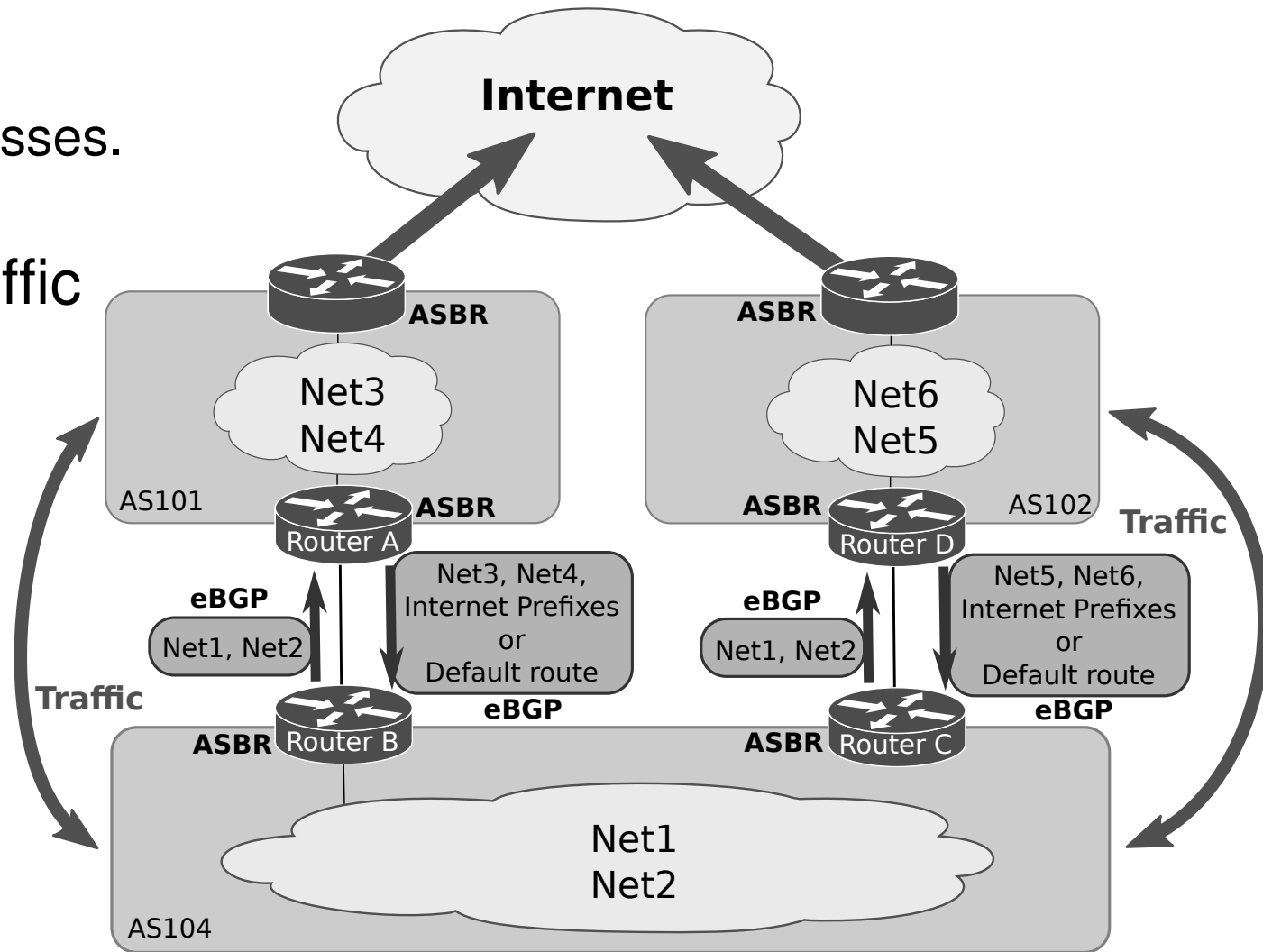
- AS has only one border router (ASBR)
  - Single Internet access.
  - Single ISP.





# Multi-homed Non-transit AS

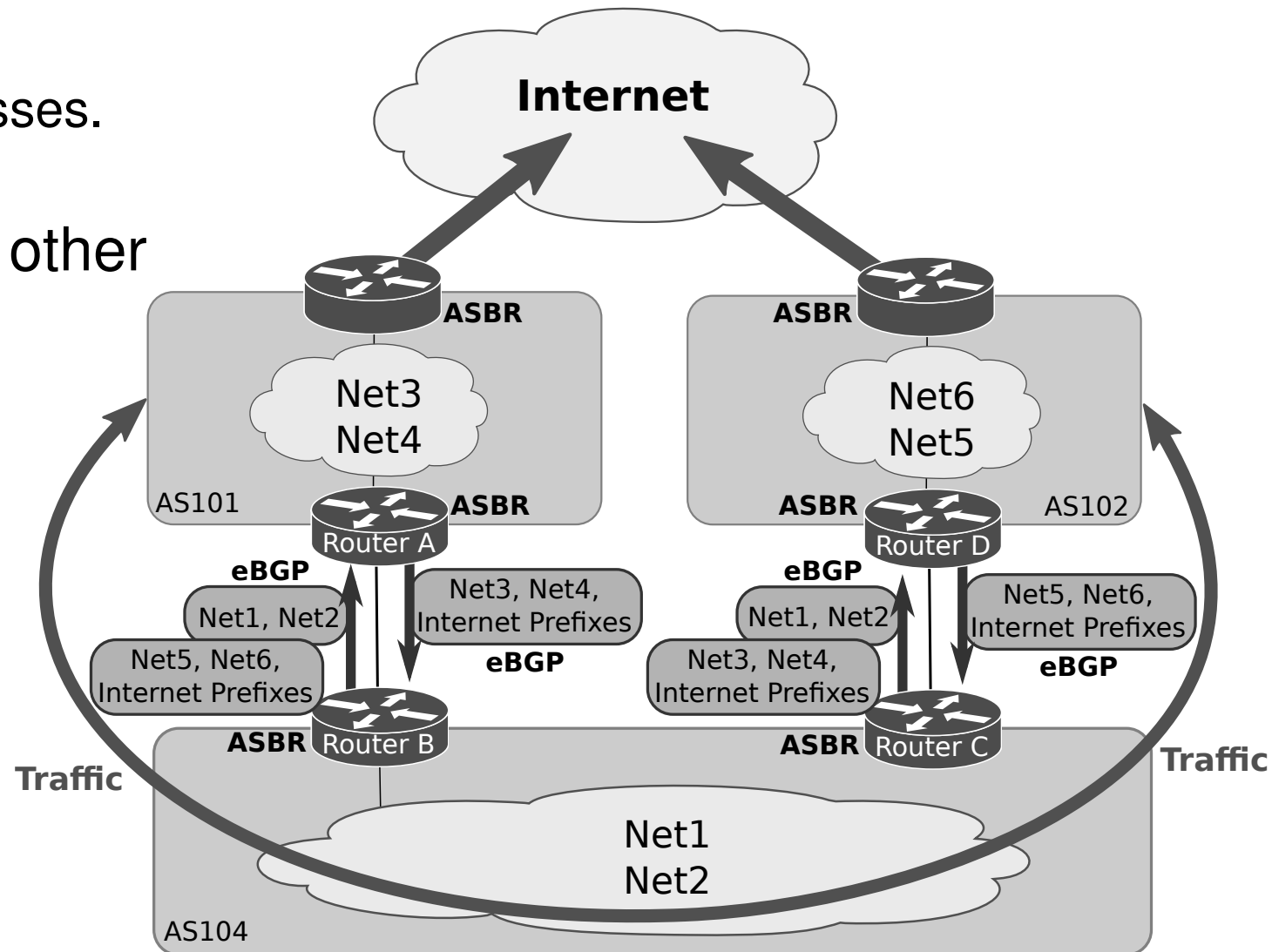
- AS has more than one border router (ASBR)
  - Multiple Internet accesses.
  - Multiple ISP.
- Does not transport traffic from other AS.





# Multi-homed Transit AS

- AS has more than one border router (ASBR).
  - Multiple Internet accesses.
  - Multiple ISP.
- Transports traffic from other AS.

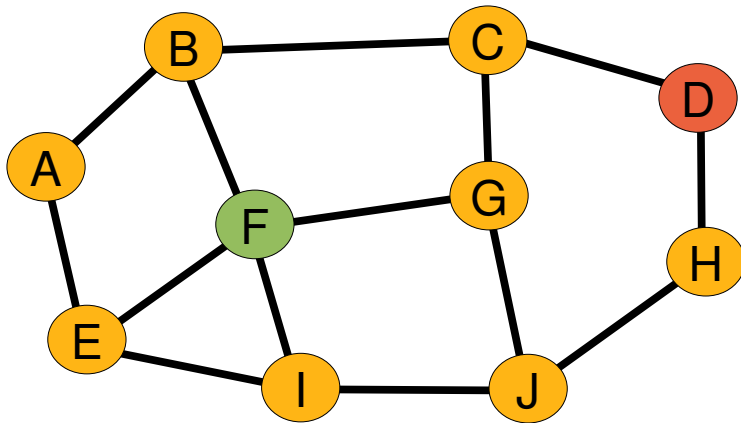


# Path-vector

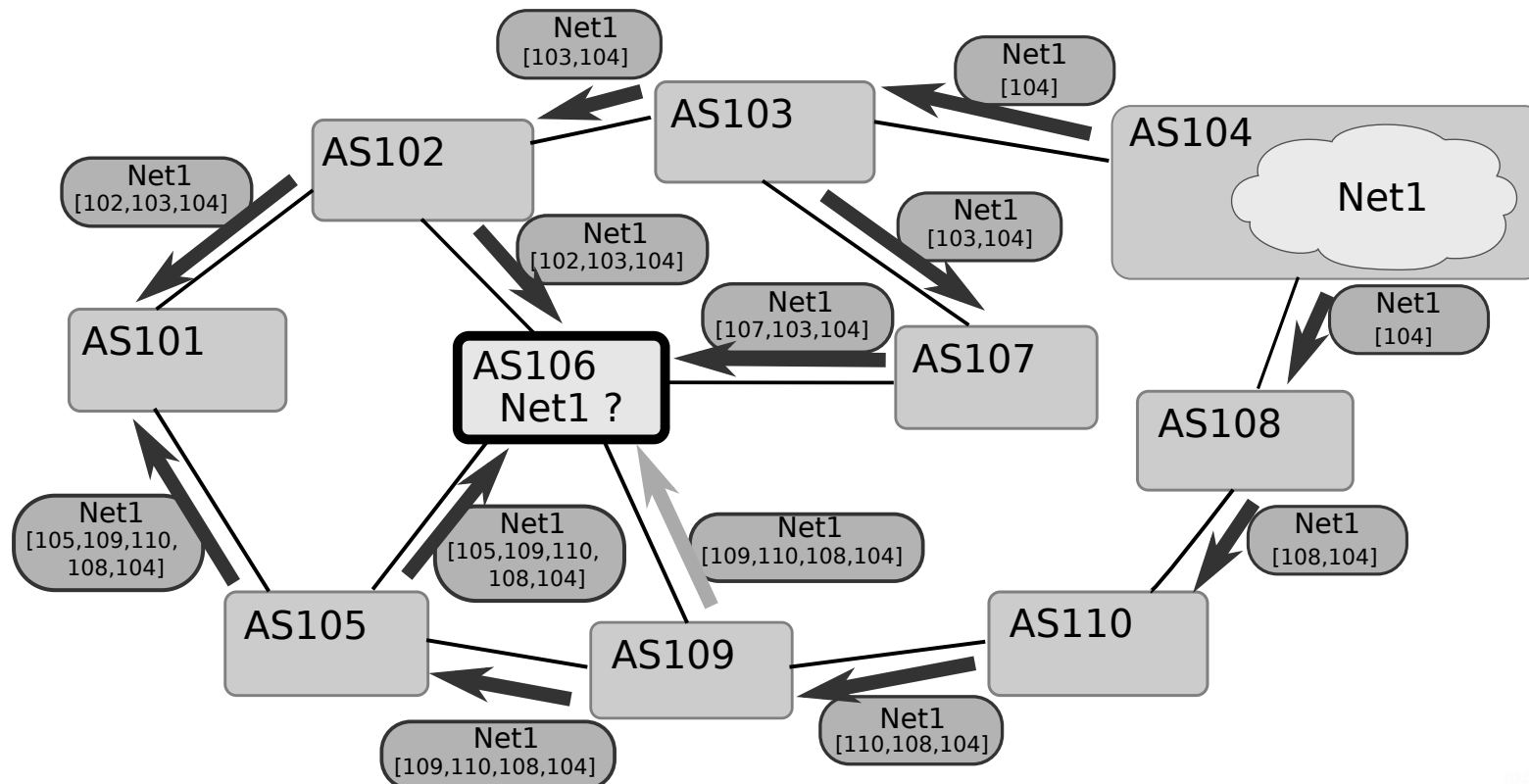
- BGP is a path-vector protocol
- Although it is essentially a distance-vector protocol that carries a list of the AS traversed by the route
  - Provides loop detection
- An EBGP speaker adds its own AS to this list before forwarding a route to another EBGP peer
- An IBGP speaker does not modify the list because it is sending the route to a peer within the same AS
  - AS list cannot be used to detect the IBGP routing loops



# Path vector

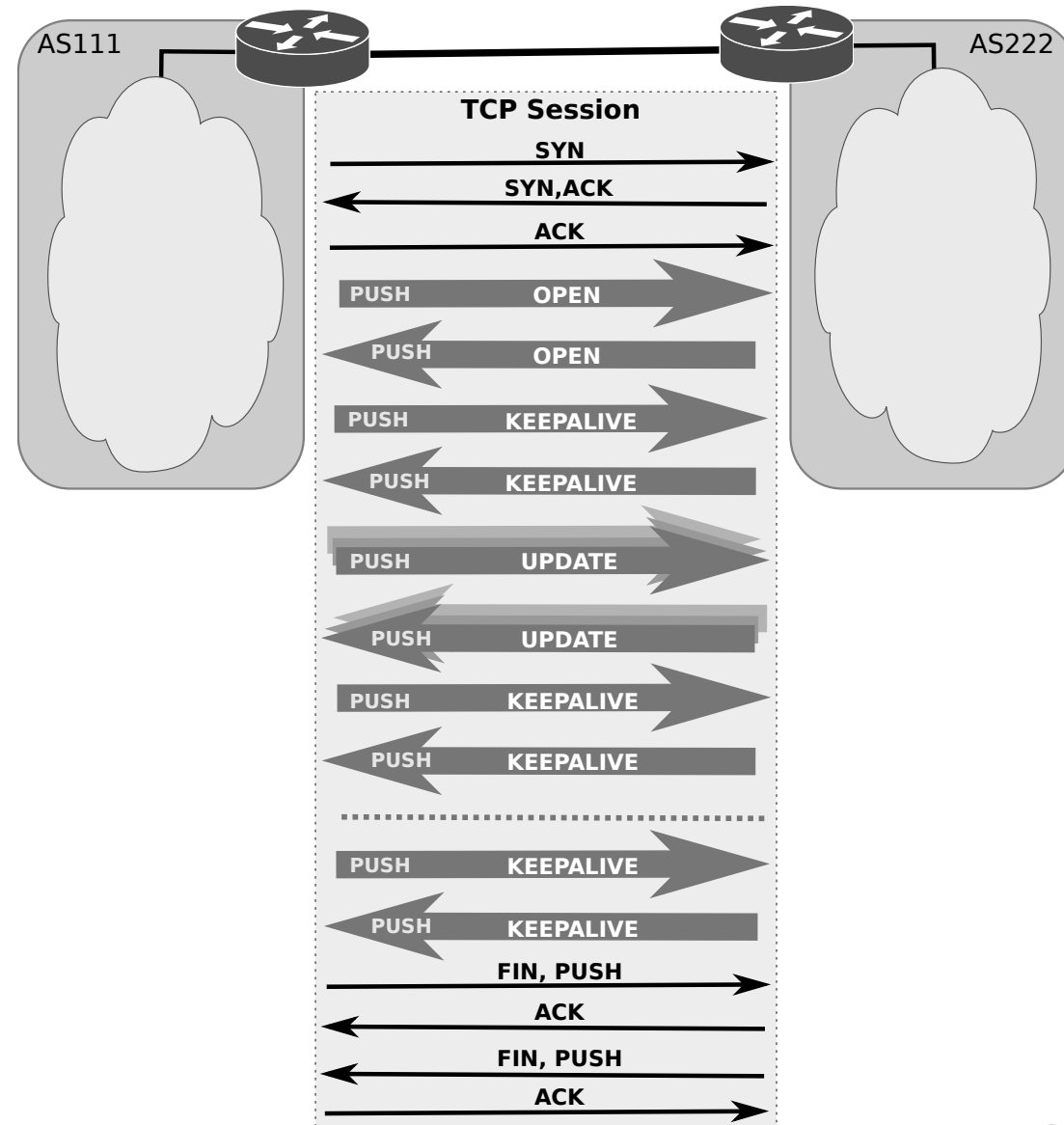


- F receives from its neighbors different paths to D:
  - De B: "I use BCD"
  - De G: "I use GCD"
  - De I: "I use IFGCD"
  - De E: "I use EFGCD"



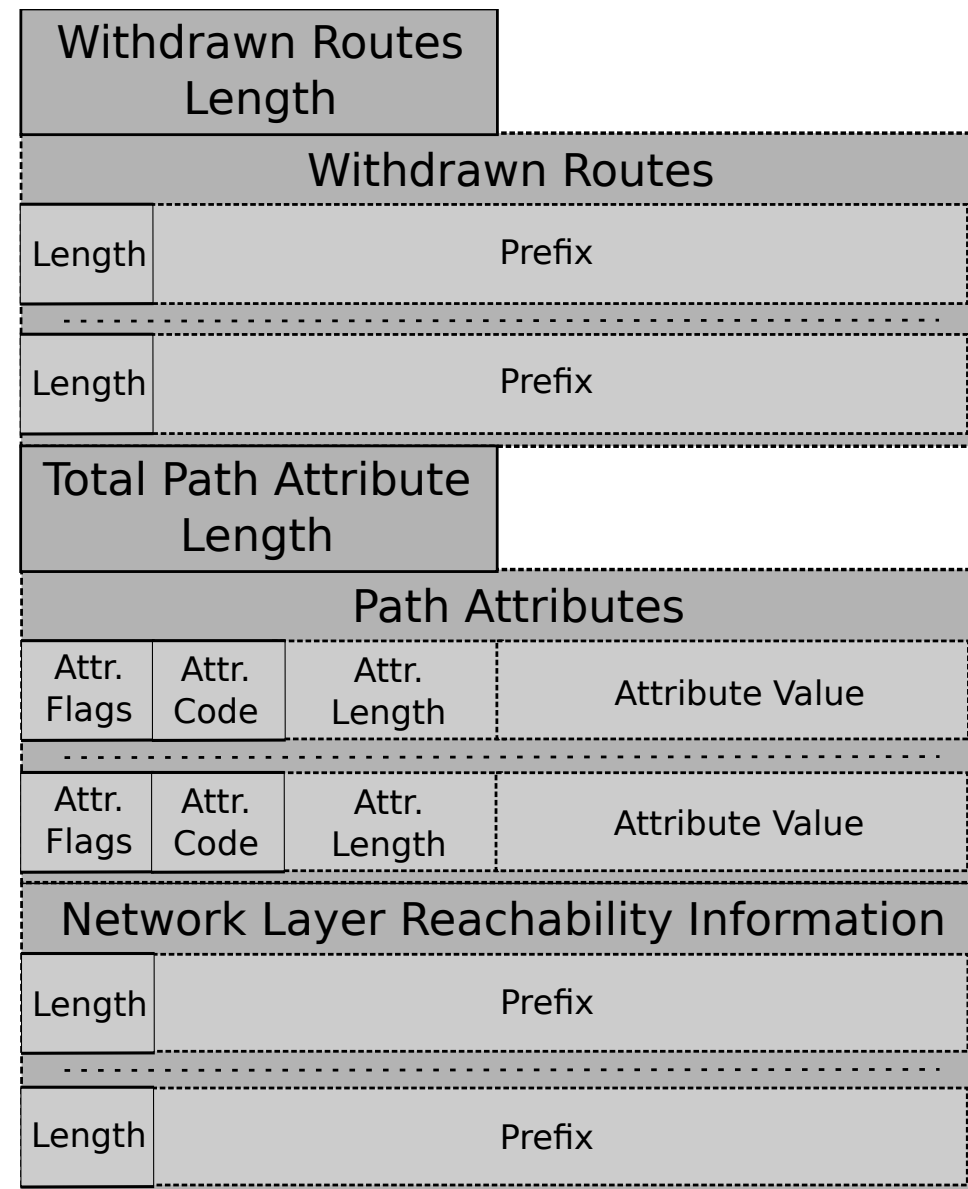
# BGP Messages

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

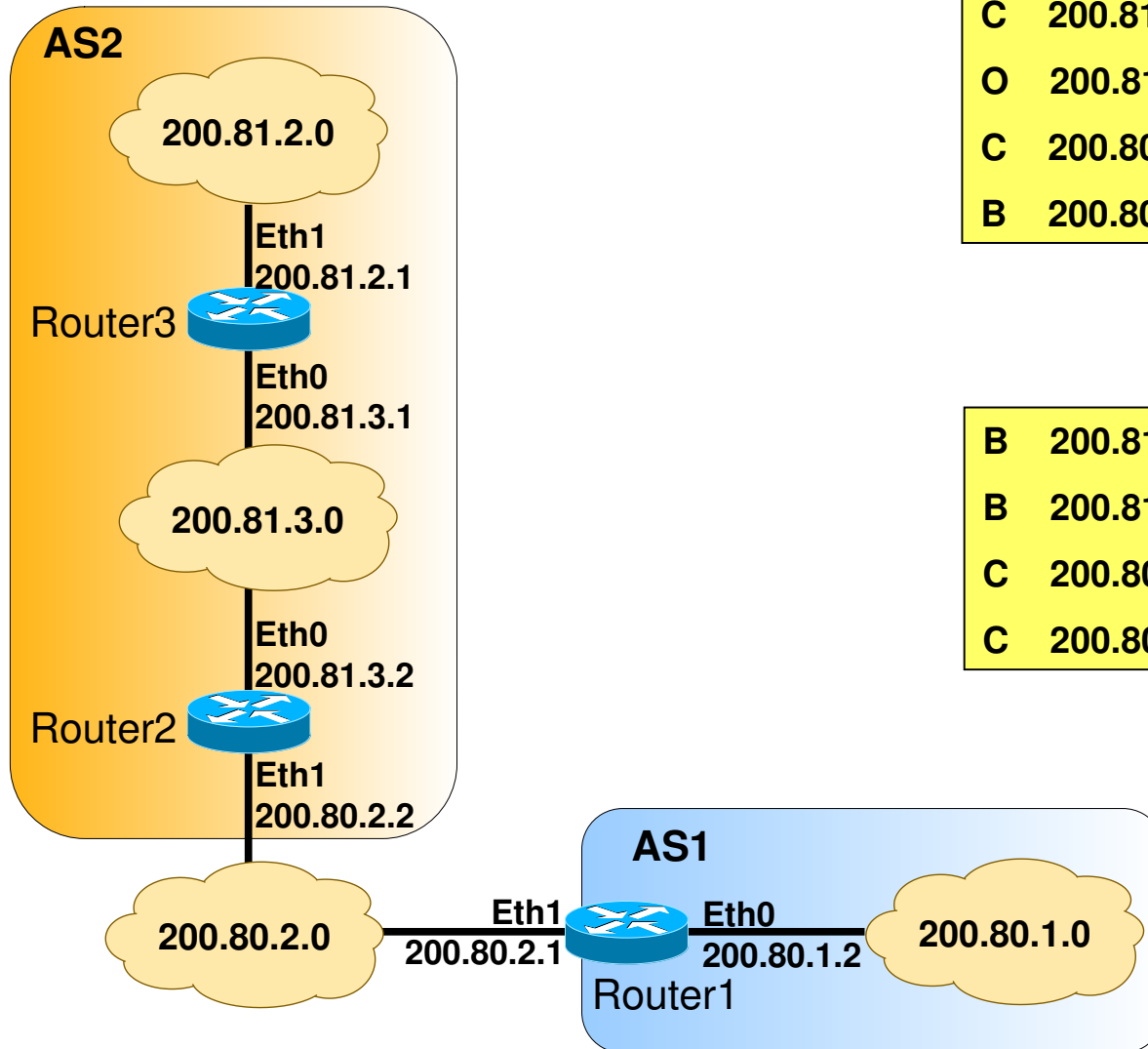


# Update Message

- Withdrawn routes – List of IP networks no longer accessible.
- Path attributes – parameters used to define routing and routing policies.
- Network layer reachability information – List of IP networks with connectivity.



# Example



C 200.81.3.0/24 is directly connected, Ethernet0  
O 200.81.2.0/24 [110/20] via 200.81.3.1, 00:01:12  
C 200.80.2.0/24 is directly connected, Ethernet1  
B 200.80.1.0/24 [20/0] via 200.80.2.1, 00:00:29

Router 2's routing table

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, Ethernet1  
C 200.80.1.0/24 is directly connected, Ethernet0

Router 1's routing table



# Example – BGP networks aggregation

## Before aggregation

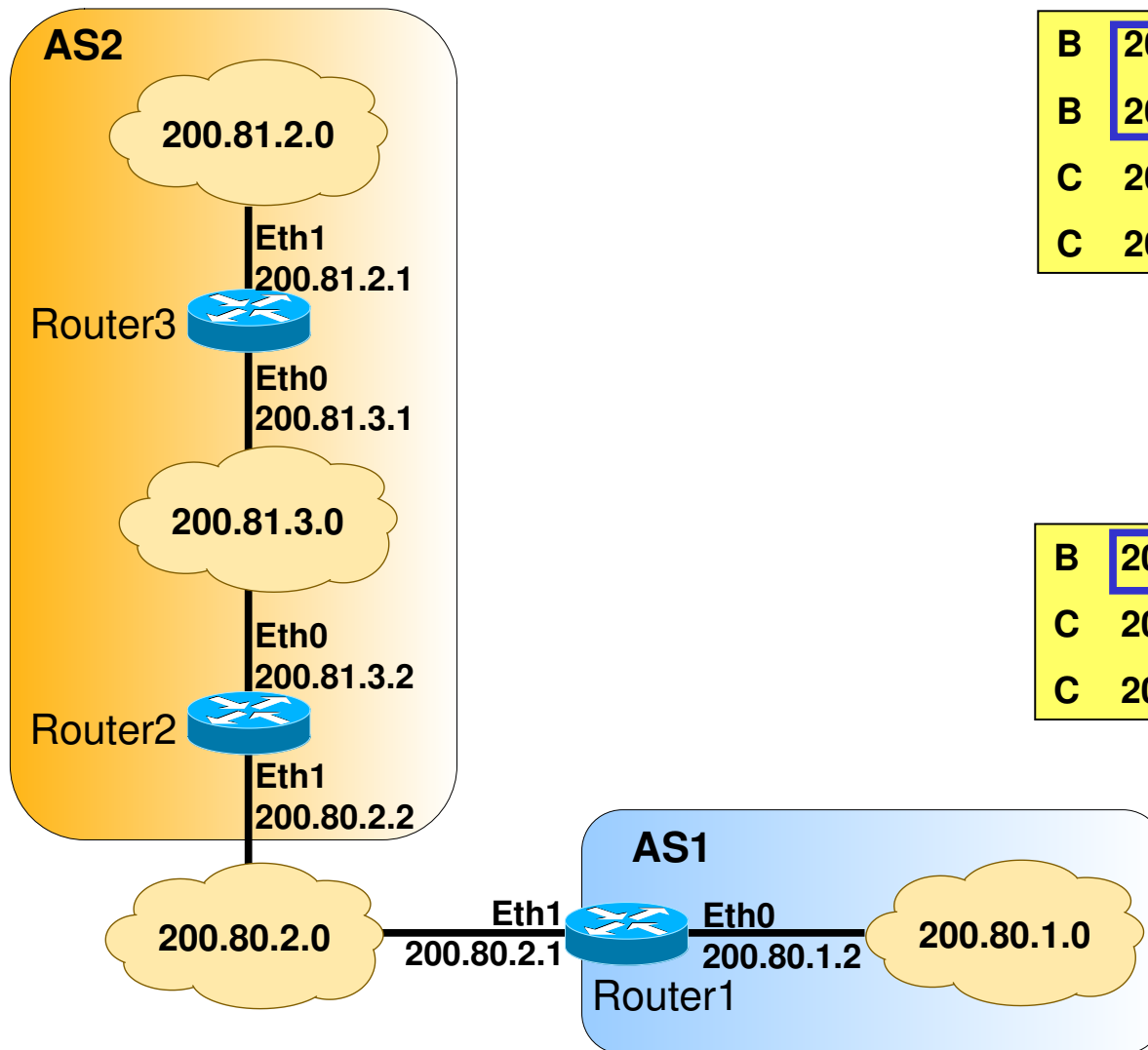
B 200.81.3.0/24 [20/0] via 200.80.2.2, 00:01:58  
B 200.81.2.0/24 [20/0] via 200.80.2.2, 00:01:57  
C 200.80.2.0/24 is directly connected, Ethernet1  
C 200.80.1.0/24 is directly connected, Ethernet0

Router 1

## After aggregation

B 200.81.2.0/23 [20/0] via 200.80.2.2, 00:01:06  
C 200.80.2.0/24 is directly connected, Ethernet1  
C 200.80.1.0/24 is directly connected, Ethernet0

Router 1





# BGP Attributes

- A BGP attribute, or path attribute, is a metric used to describe the characteristics of a BGP path.
- Attributes are contained in update messages passed between BGP peers to advertise routes. There are 4+1 categories of BGP attributes.
  - Well-known Mandatory (included in BGP updates)
    - ➔ AS-path, Next-hop, Origin.
  - 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)
    - ➔ Aggregator, Community, AS4\_Aggregator, AS4\_path.
  - Optional Non-transitive (may not be supported by all BGP implementations)
    - ➔ If the neighbor doesn't support that attribute it is deleted
    - ➔ Multi-exit-discriminator (MED).
  - Cisco-defined (local to router, not advertised)
    - ➔ Weight



# AS-PATH and ORIGIN Attributes

- 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.

- ORIGIN

- Indicates how BGP learned about a particular route. Can take three possible values:
  - ➔ IGP (0) value is set if the route is interior to the originating AS, resulting from an explicit inclusion of a network within the BGP routing process by means of manual configuration.
  - ➔ INCOMPLETE (2) value is set if the route is learned by other means, namely, route redistribution from other routing processes into the BGP routing process.
  - ➔ EGP (1) is no longer used in modern networks.

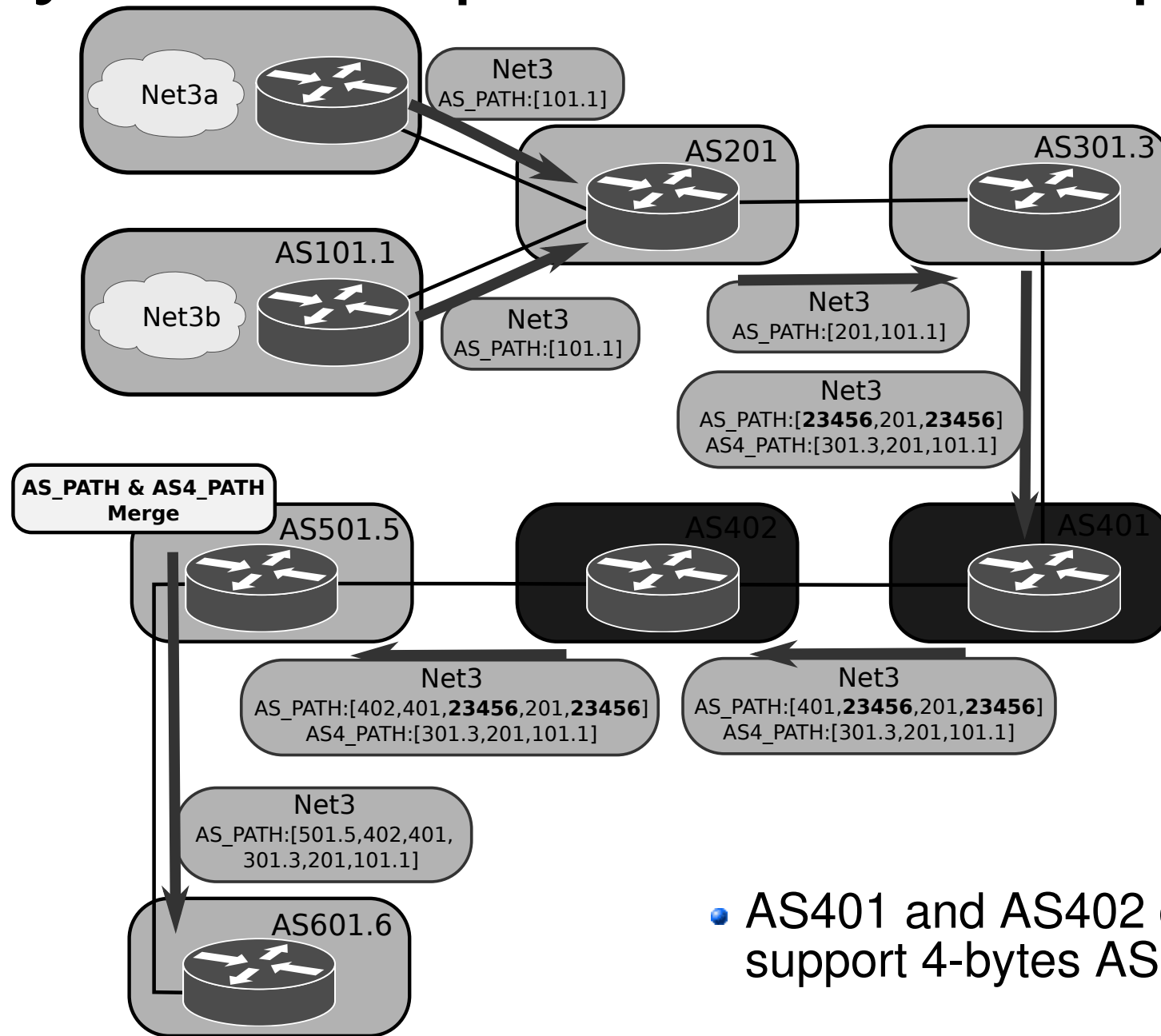


# AS4\_PATH & AS4\_AGGREGATOR

- AS4\_PATH attribute has the same semantics as the AS\_PATH attribute, except that it is optional transitive, and it carries 4-bytes AS numbers.
- AS4\_AGGREGATOR attribute has the same semantics as the AGGREGATOR attribute, except that it carries a 4-bytes AS number.
- 4-byte AS support is advertised via BGP capability negotiation
  - Speakers who support 4-byte AS are known as NEW BGP speakers
  - Those who do not are known as OLD BGP speakers
- New Reserved AS number
  - AS\_TRANS = AS 23456
    - ➔ 2-byte placeholder for a 4-byte AS number
    - ➔ Used for backward compatibility between OLD and NEW BGP speakers
- Receiving UPDATES from a NEW speaker
  - Decode each AS number as 4-bytes
  - AS\_PATH and AGGREGATOR are effected
- Receiving UPDATES from an OLD speaker
  - AS4\_AGGREGATOR will override AGGREGATOR
  - AS4\_PATH and AS\_PATH must be merged to form the correct as-path
- Merging AS4\_PATH and AS\_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 ]



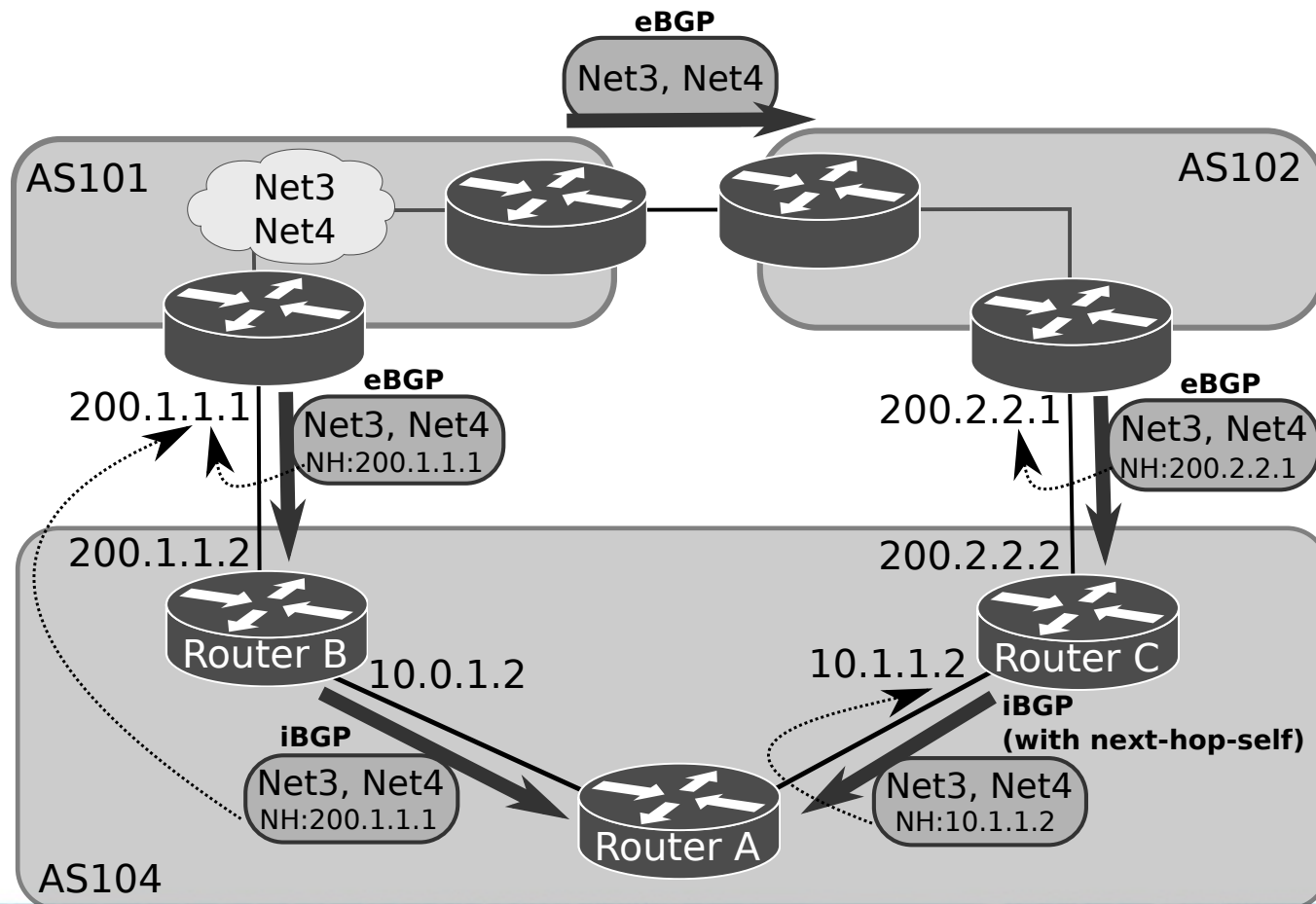
# 4-bytes AS Operational Example



- AS401 and AS402 do not support 4-bytes ASN.

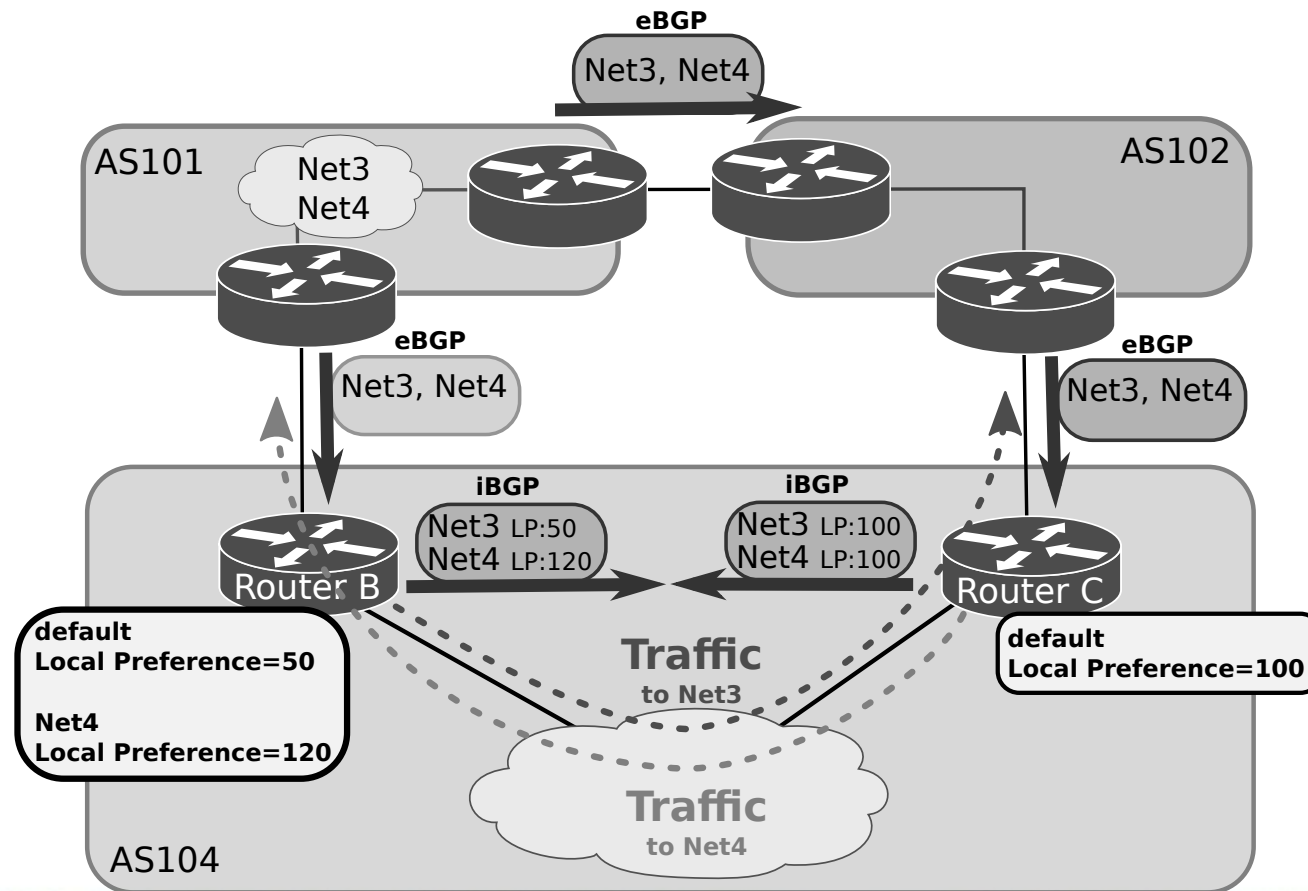
# Next-Hop Attribute

- The eBGP next-hop attribute is the IP address that is used to reach the advertising router
- For eBGP, the next-hop address is the IP address of the connection between the peers
- For iBGP, the eBGP next-hop address is carried into the local AS
  - By configuration the AS border router can be the next-hop to iBGP neighbors



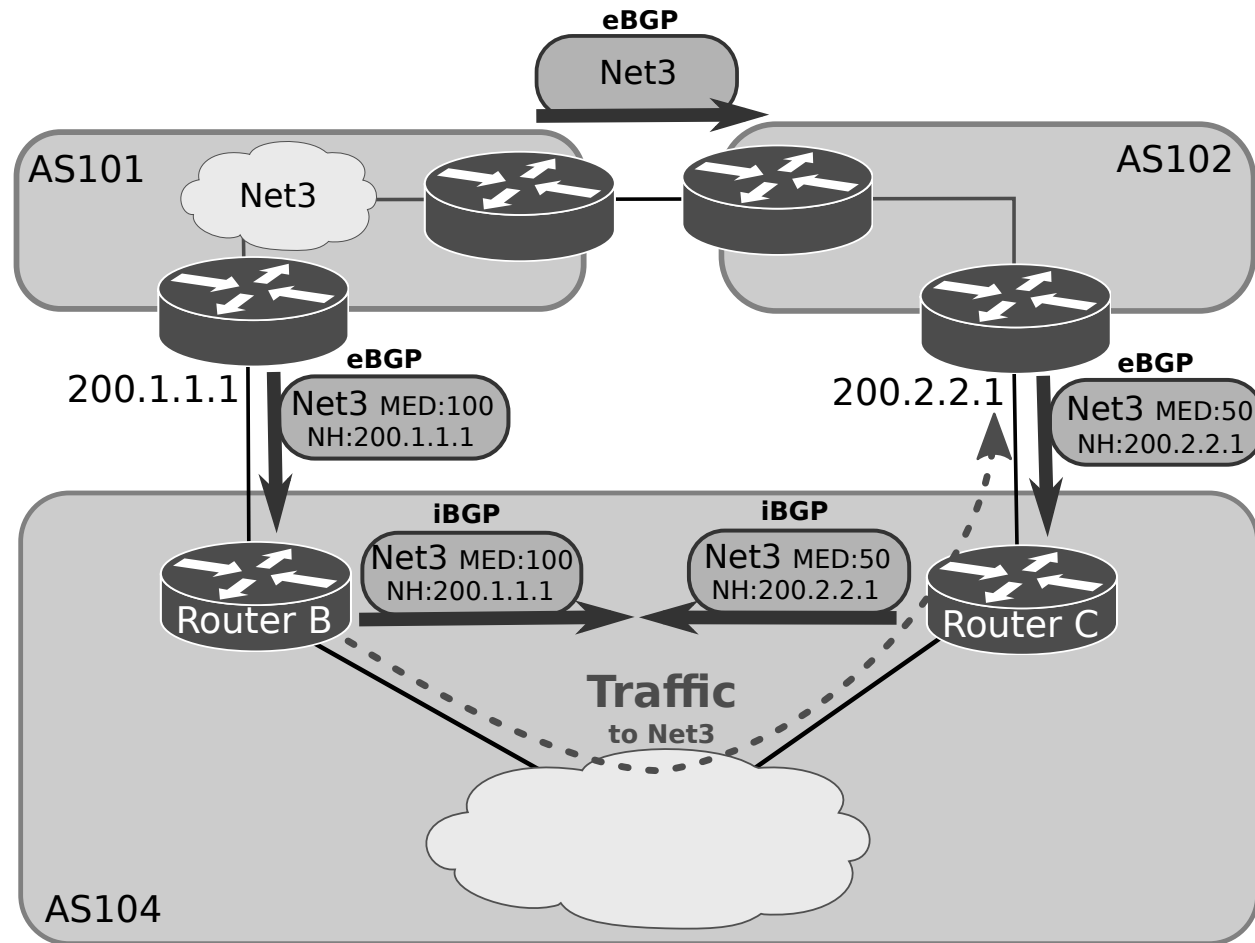
# Local Preference Attribute

- The local preference attribute is used to choose an exit point from the local autonomous system (AS).
  - **Higher value** is preferred.
- The local preference attribute is propagated throughout the local AS.
- Can be different, for different routes.



# Multi-Exit Discriminator Attribute (MED)

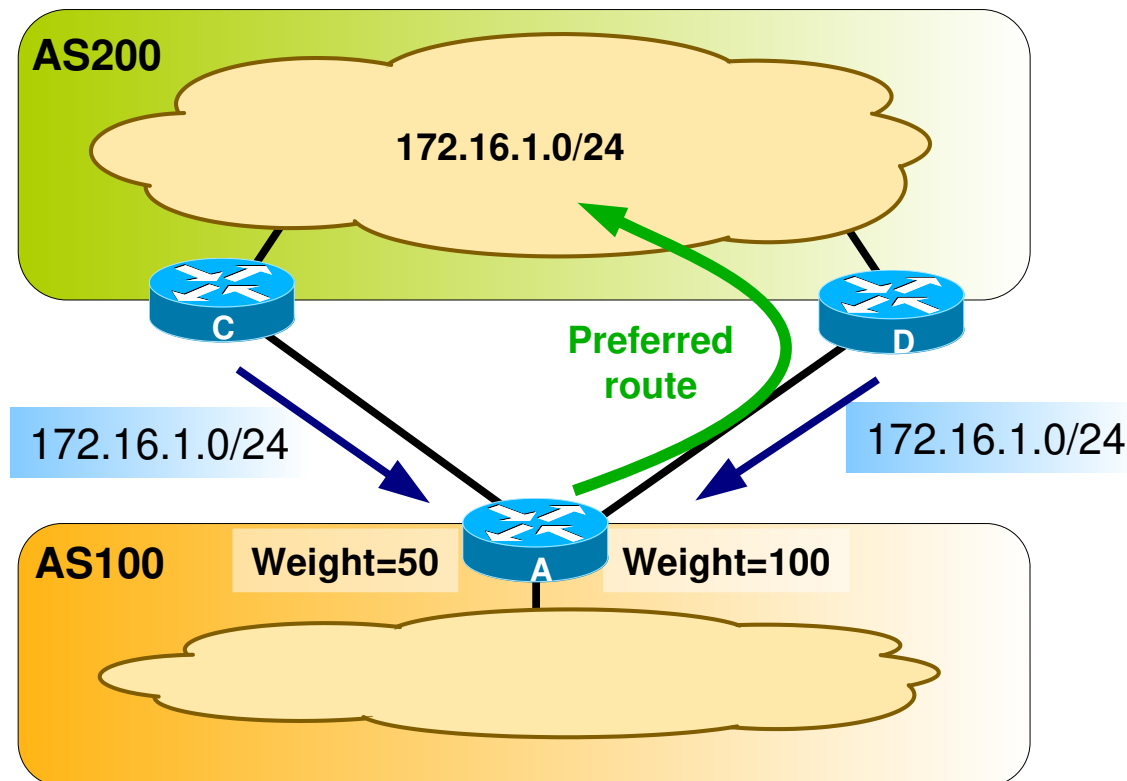
- The multi-exit discriminator (MED) or metric attribute is used as a suggestion to an external AS.
- The external AS that is receiving the MEDs may be using other BGP attributes for route selection.
- The **lower value** of the metric is preferred.
- MED is designed to influence incoming traffic.



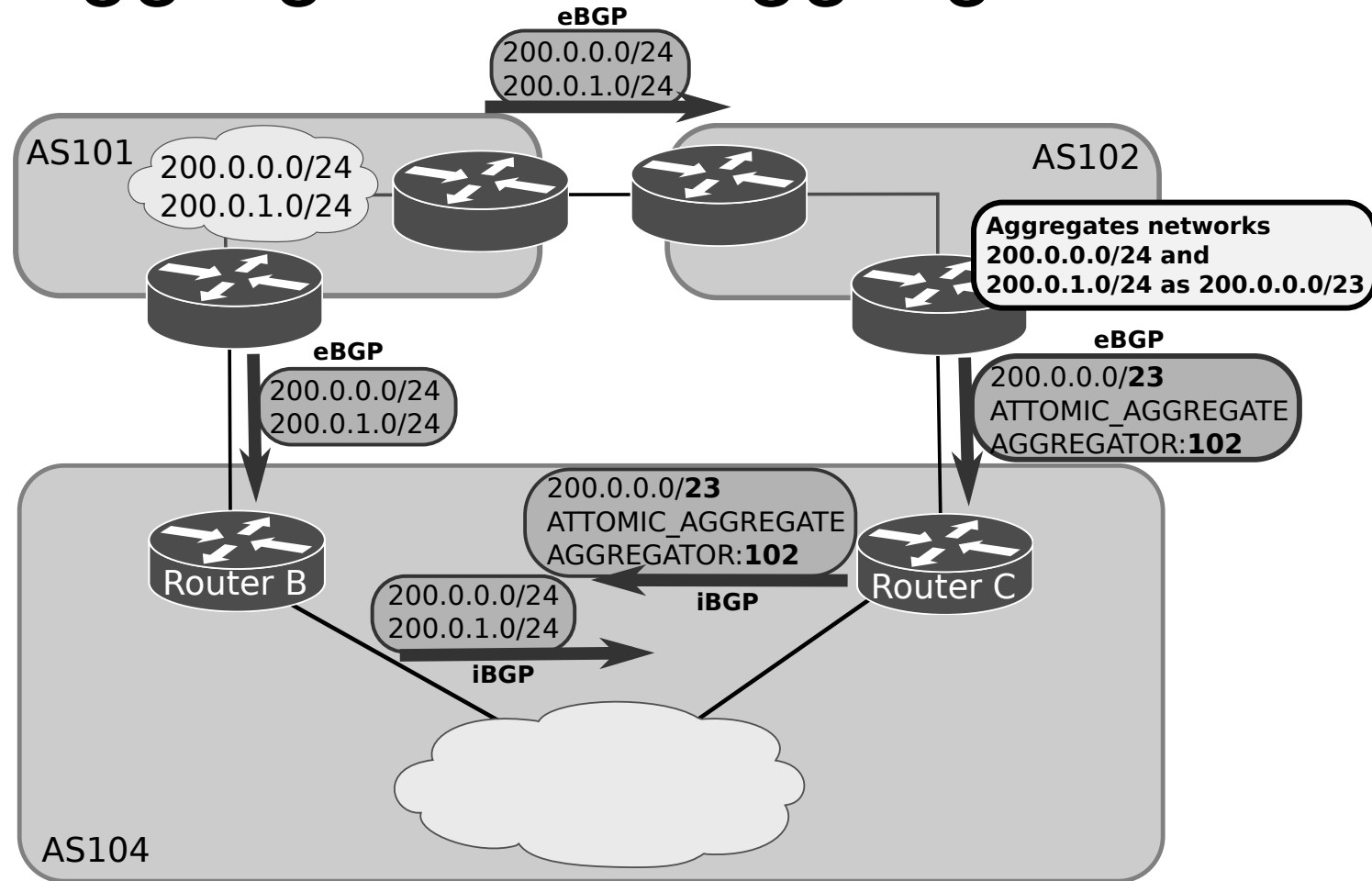


# Weight Attribute

- Weight is a Cisco-defined attribute that is local to a router.
- The weight attribute is not advertised to neighboring routers.
- If the router learns about more than one route to the same destination, the route with the **highest weight** will be preferred.



# Atomic Aggregate and Aggregator Attributes



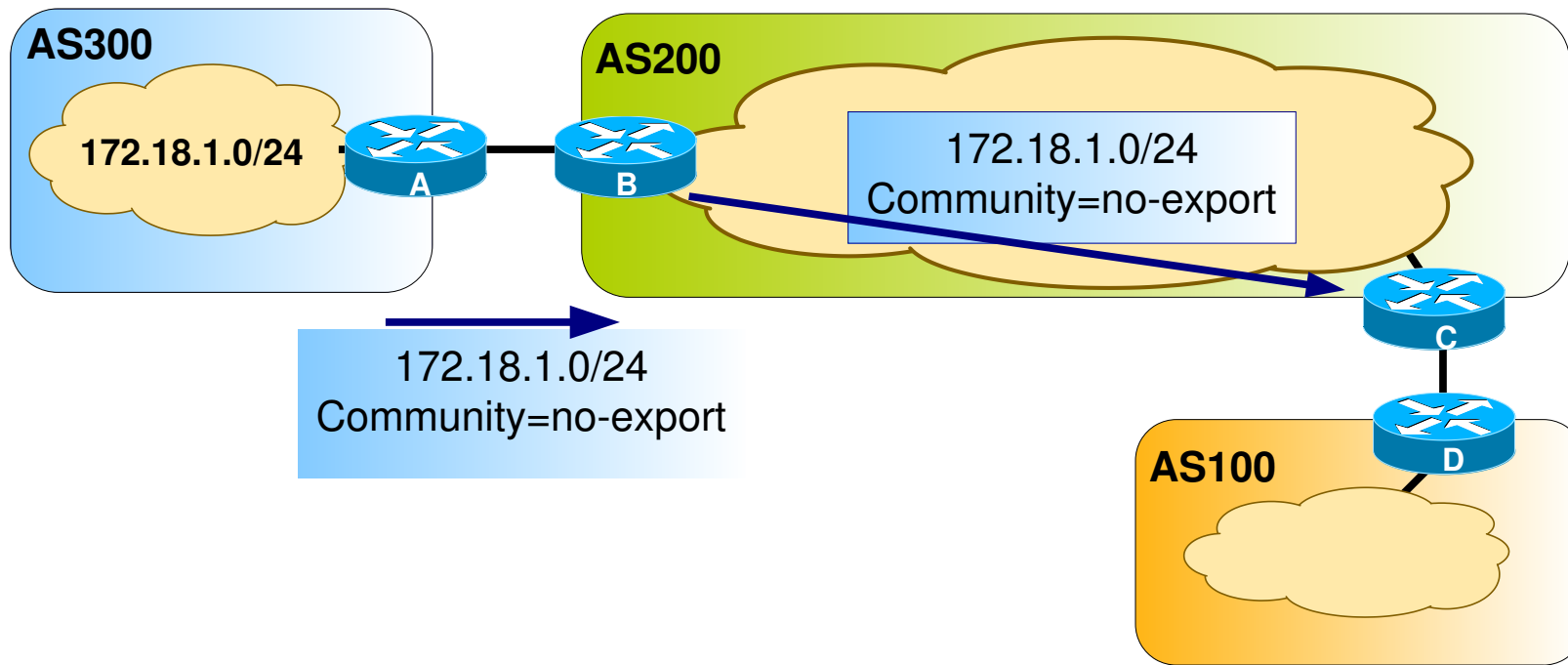
- Atomic Aggregate

- Is used to alert routers that specific routes have been aggregated into a less specific route.
- When aggregation like this occurs, more specific routes are lost.

- Aggregator

- Provides information about which AS performed the aggregation.
- And the IP address of the router that originated the aggregate.

# Community Attribute



- Used to group routes that share common properties so that policies can be applied at the group level
- Predefined community attributes are:
  - no-export - Do not advertise this route to EBGp peers
  - no-advertise - Do not advertise this route to any peer
  - internet - Advertise this route to the Internet community; all routers in the network belong to it
- General communities format is ASnumber:Cnumber
  - e.g. 300:1, 200:38, etc...

# BGP Path Selection

- BGP may receive multiple advertisements for the same route from multiple sources.
- BGP selects only one path as the best path.
- BGP puts the selected path in the IP routing table and propagates the path to its neighbors. BGP uses the following criteria, in the order:
  - Largest weight (Cisco only)
  - Largest local preference
  - Path that was originated locally
  - Shortest path
  - Lowest origin type (IGP lower than EGP, EGP lower than incomplete)
  - Lowest MED attribute
  - Prefer the external path over the internal path
  - Closest IGP neighbor



# Multi-Protocol Border Gateway Protocol (MP-BGP)



# MP-BGP Description

- Extension to the BGP protocol
- Carries routing information about other protocols/families:
  - IPv6 Unicast
  - Multicast (IPv4 and IPv6)
  - 6PE - IPv6 over IPv4 MPLS backbone
  - Multi-Protocol Label Switching (MPLS) VPN (IPv4 and IPv6)
- Exchange of Multi-Protocol Reachability Information (NLRI)



# MP-BGP Attributes

- 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
- Attribute contains one or more triples
  - Address Family Information (AFI) with Sub-AFI
    - ➔ Identifies protocol information carried in the Network Layer Reachability Information
  - Next-hop information
    - ➔ Next-hop address must be of the same family
- Reachability information



# MP-BGP Negotiation Capabilities

- MP-BGP routers establish BGP sessions through the OPEN message
  - OPEN message contains optional parameters
  - If OPEN parameters are not recognized, BGP session is terminated
  - A new optional parameter: CAPABILITIES
- OPEN message with CAPABILITIES containing:
  - Multi-Protocol extensions (AFI/SAFI)
  - Route Refresh
  - Outbound Route Filtering

# MP-BGP New Features for IPv6

- IPv6 Unicast
  - MP-BGP enables the creation of IPv6 Inter-AS relations
- IPv6 Multicast
  - Unicast prefixes for Reverse Path Forwarding (RPF) checking
  - RPF information is disseminated between autonomous systems
  - Compatible with single domain Rendezvous Points or Protocol Independent Multicast-Source Specific Multicast (PIM-SSM)
  - Topology can be congruent or non-congruent with the unicast one
- IPv6 and label (6PE)
  - IPv6 packet is transported over an IPv4 MPLS backbone
- IPv6 VPN (6VPE)
  - Multiple IPv6 VPNs are created over an IPv4 MPLS backbone

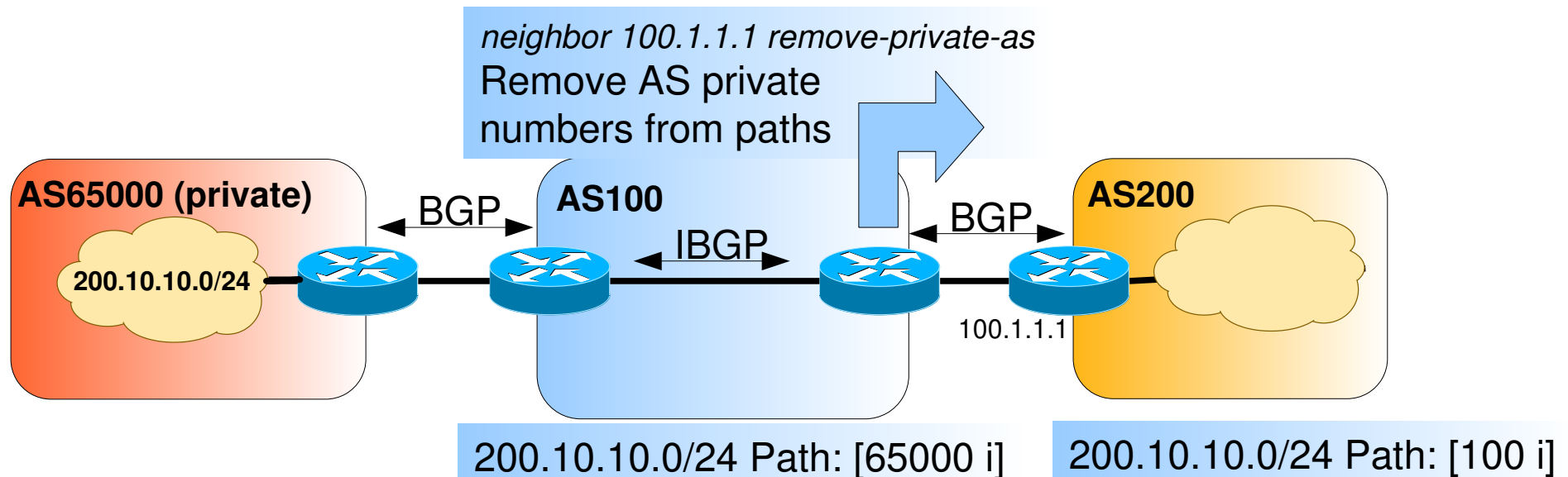


# Advanced BGP



# Private BGP AS

- Private autonomous system (AS) numbers range from 64512 to 65535
- When a customer network is large, the ISP may assign an AS number:
  - Permanently assigning a **Public** AS number in the range of 1 to 64511
    - ➔ Should have a unique AS number to propagate its BGP routes to Internet
    - ➔ Done when a customer network connects to two different ISPs, such as multihoming
  - Assigning a **Private** AS number in the range of 64512 to 65535.
    - ➔ It is not recommended that you use a private AS number when planning to connect to multiple ISPs in the future



# BGP AS Routing Policies

```
aut-num: AS15525
as-name: PTPRIMENET
descr: PT Prime Autonomous System
descr: Corporate Data Communications Services
descr: Portugal
import: from AS1930 action pref=100;
       accept AS-RCCN # RCCN
import: from AS3243 action pref=200;
       accept AS-TELEPAC # Telepac
import: from AS5516 action pref=100;
       accept AS5516 # INESC
import: from AS5533 action pref=100;
       accept AS-VIAPT # Via NetWorks Portugal
import: from AS8657 action pref=300;
       accept ANY # CPRM
import: from AS12305 action pref=100;
       accept AS12305 # Nortenet
import: from AS1897 action pref=100;
       accept AS1897 AS9190 AS13134 AS15931 # KPN Qwest
import: from AS13156 action pref=100;
       accept AS13156 # Cabovisao
import: from AS8824 action pref=100;
       accept AS8824 AS15919 # Eastecnica
```

```
export: to AS1897 announce RS-PTPRIME # KPNQwest
export: to AS1930 announce RS-PTPRIME # RCCN
export: to AS3243 announce RS-PTPRIME # Telepac
export: to AS5516 announce {0.0.0.0/0} # INESC
export: to AS5533 announce RS-PTPRIME # Via NetWorks
       Portugal
export: to AS8657 announce RS-PTPRIME # CPRM
export: to AS8824 announce RS-PTPRIME # Eastecnica
export: to AS8826 announce {0.0.0.0/0} # Siemens
export: to AS9186 announce RS-PTPRIME # ONI
export: to AS12305 announce RS-PTPRIME # Nortenet
export: to AS12353 announce RS-PTPRIME # Vodafone Portugal
export: to AS13156 announce RS-PTPRIME # Cabovisao
export: to AS13910 announce ANY # register.com
export: to AS15931 announce ANY # YASP Hiperbit
export: to AS24698 announce RS-PTPRIME # Optimus
export: to AS25005 announce ANY # Finibanco
export: to AS25253 announce {0.0.0.0/0} # CGDNet
export: to AS28672 announce ANY # BPN
export: to AS31401 announce {0.0.0.0/0} # SICAMSERV
export: to AS39088 announce {0.0.0.0/0} # Santander-Totta
export: to AS41345 announce RS-PTPRIME # Visabeira
export: to AS43064 announce RS-PTPRIME # Teixeira Duarte
export: to AS43643 announce ANY # TAP
```

From RIPE database  
<http://www.db.ripe.net>

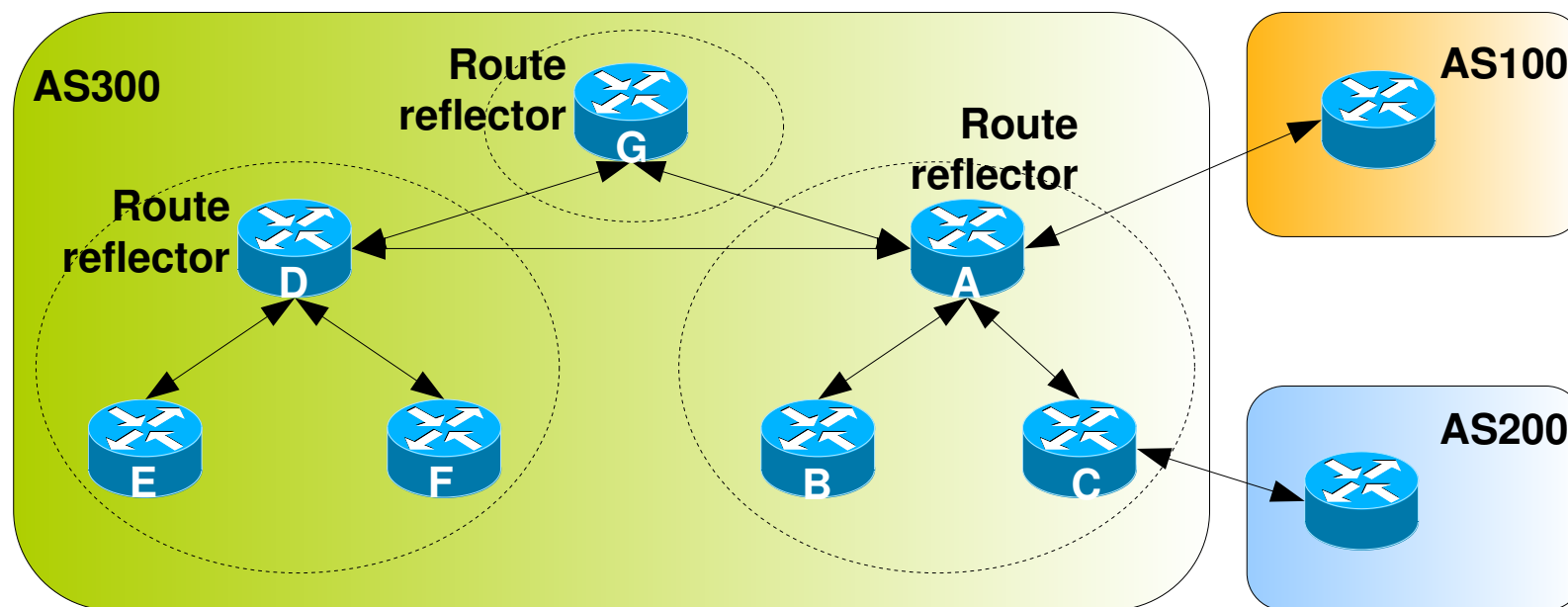


# BGP Synchronization

- Synchronization states that, if your AS passes traffic from another AS to a third AS, BGP should not advertise a route before all the routers in your AS have learned about the route via IGP.
- BGP waits until IGP has propagated the route within the AS. Then, BGP advertises the route to external peers.



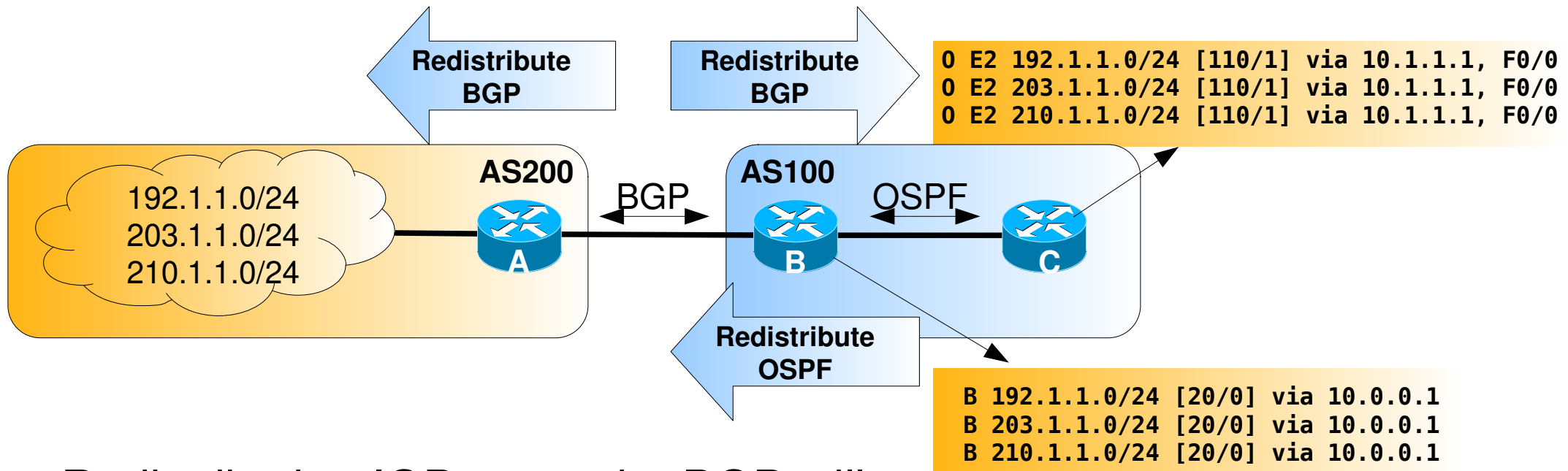
# BGP Route Reflectors



- Without a route reflector, the network requires a full iBGP mesh within AS300.
- The route reflector and its clients are called a cluster.
  - Router A is configured as a route reflector, iBGP peering between Routers B and C (and others) is not required.
  - Router D is configured as a route reflector, iBGP peering between Routers E and F (and others) is not required.
- Full iBGP mesh between route reflector Routers.



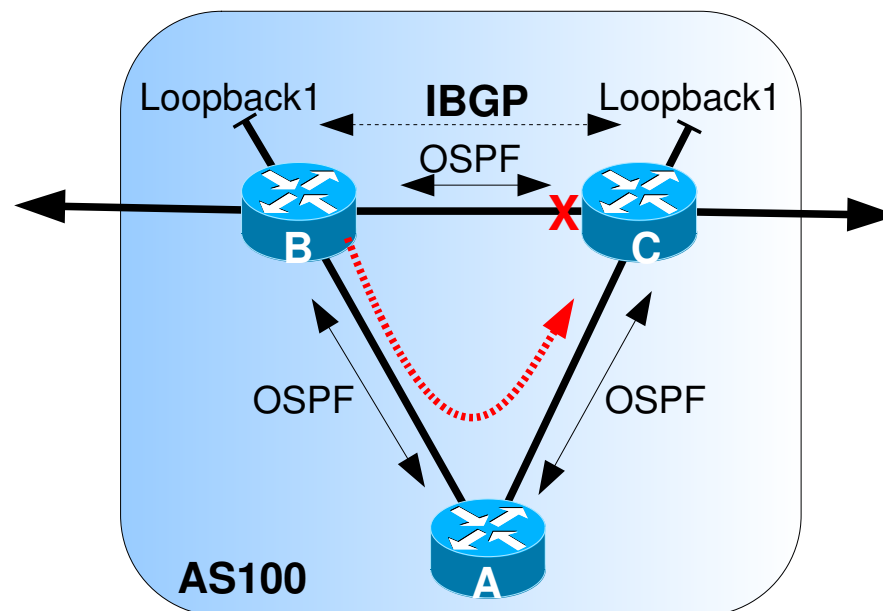
# Routes Redistribution



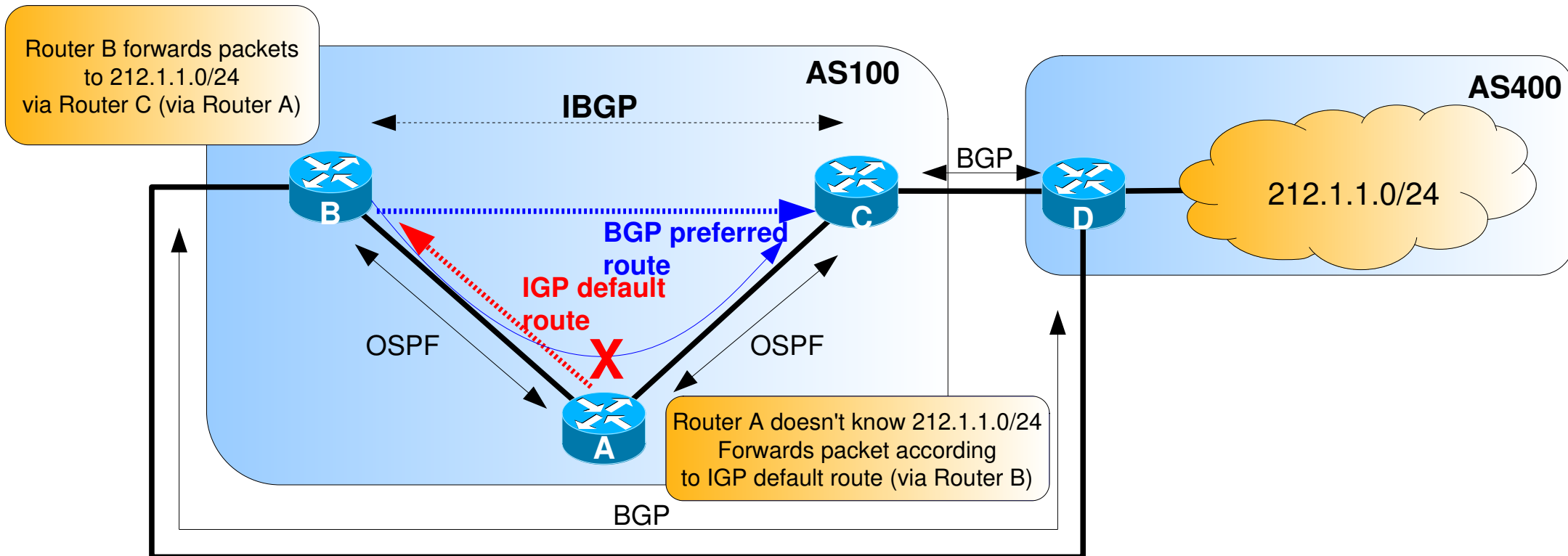
- Redistributing IGP routes by BGP will:
  - Simplify BGP configuration (advantage)
  - And BGP will announce only internal networks with connectivity (advantage)
- Redistributing BGP routes by IGP protocols will:
  - Make internal routes know all external routes (disadvantage/advantage?)
  - Increase routing tables size in internal routers (disadvantage)
    - Decrease routing time, imposes memory requirements, ...
  - Avoid the usage of internal default routes (disadvantage/advantage?)

# BGP Neighborhood Resilience

- BGP neighbor relations between physical interfaces are dependent on interface stability/status
- (Virtual) neighbor relations using Loopback interfaces/addresses
  - Loopback interfaces are virtual and software based
    - ➔ If the router is active Loopback interfaces are always active
  - Neighbor relation is active while a path exists between the virtual networks
    - ➔ (Alternative) Routing provided by IGPs

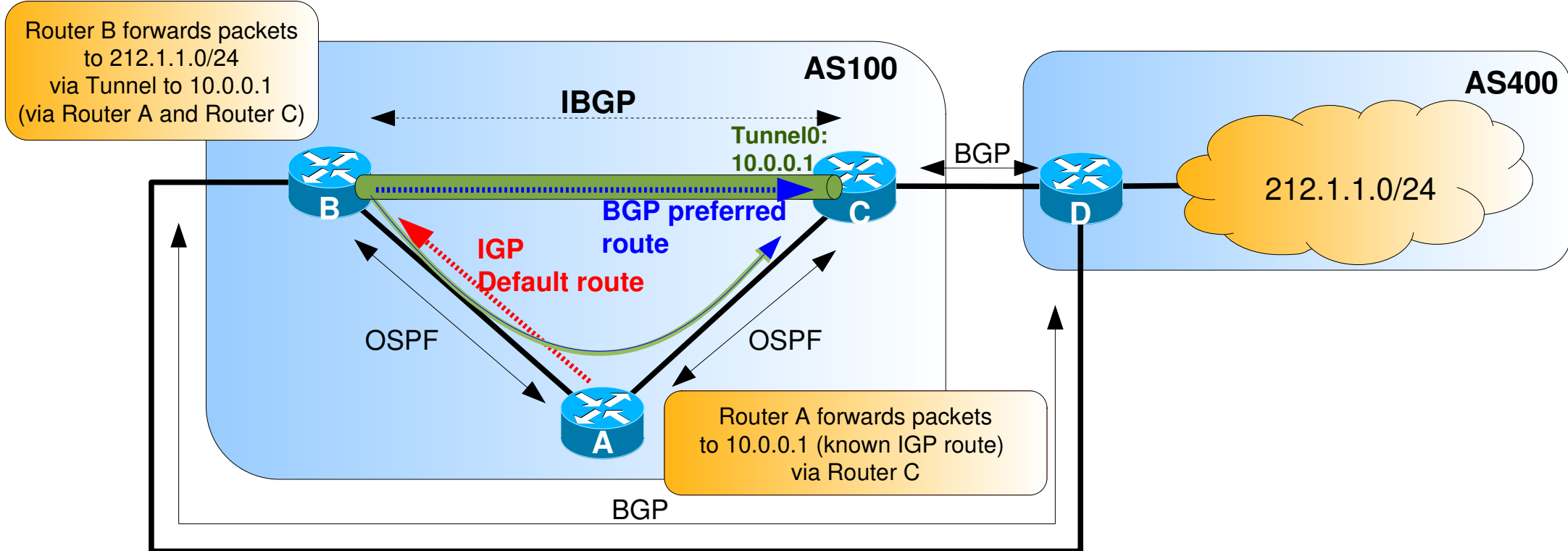


# BGP and IGP conflicts



- Routing conflicts may arise with
  - Internal routers without BGP
  - No redistribution of BGP routes by IGP
  - IGP default routes
  - BGP preferred routes (with no agreement with IGP default routes)
- Solutions
  - Adjust IGP default routes
  - Adjust BGP preferred routes (e.g. with local preference)
  - BGP neighborhood and Internal routing via IP-IP tunnels

# BGP over Tunnels (over IGP)



- IP-IP tunnels to solve BGP/IGP routing conflicts

- Tunnels manually configured
  - Between physical or Loopback interfaces
- BGP neighborhood via Tunnel
- BGP routes learned via Tunnel (next hop is remote Tunnel end-point)
- Tunnel “network” distributed internally via IGP

- In Router A, to any packet destined to an outside network it's forwarded via Tunnel

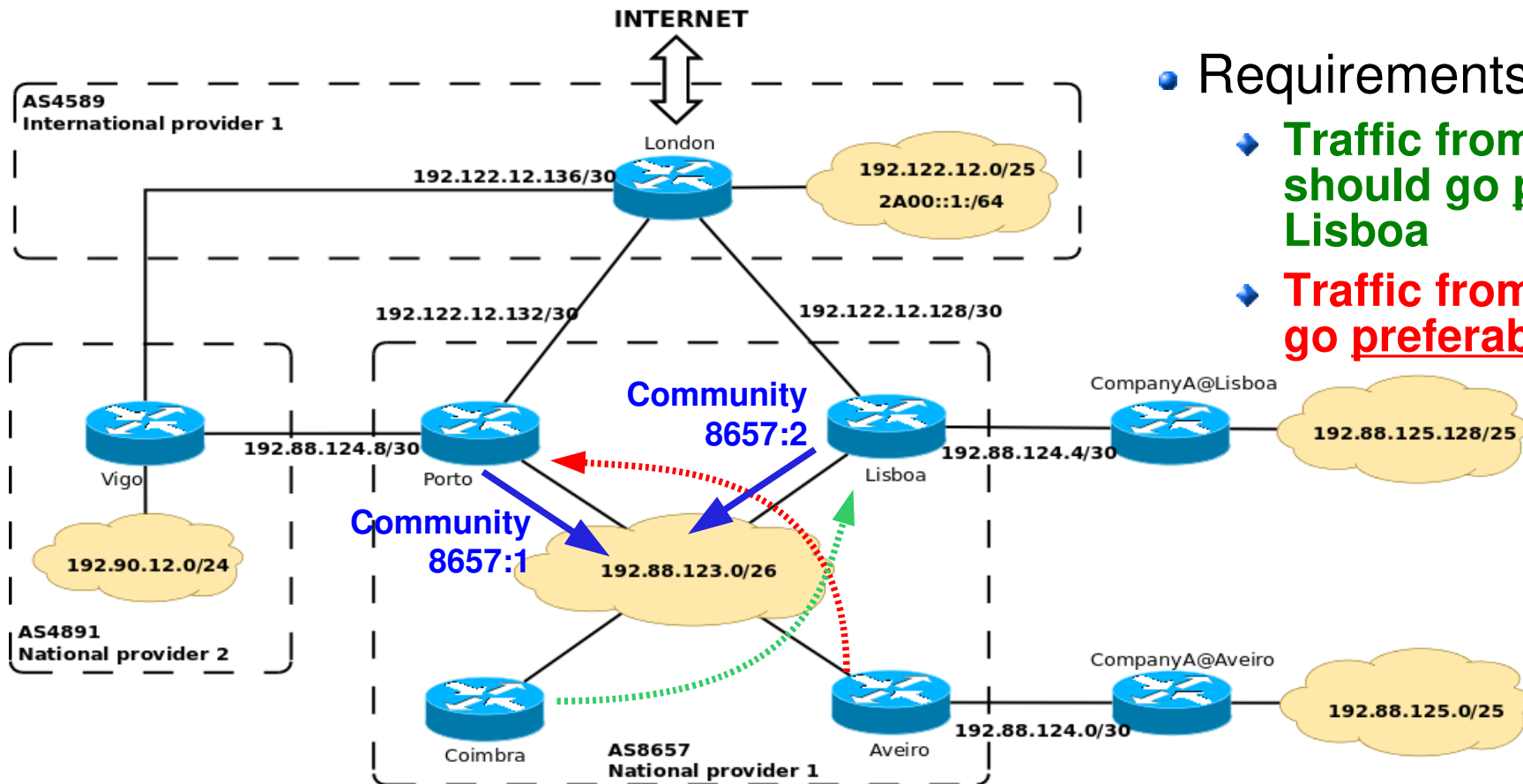
- A new IP header is added, new IP destination address is the remote Tunnel end-point
- Internally, packet is routed according to the new IP header (Tunnel end-points IP addresses)

# BGP Filtering and Route Maps

- Sending and receiving BGP updates can be controlled by using a number of different filtering methods.
- BGP updates can be filtered based on:
  - Route information,
  - Path information,
  - Communities.
- Route maps are used with BGP to
  - Control and modify routing information.
  - Define the conditions by which routes are redistributed between routing domains.



# BGP Case Studies



## Requirements

- ♦ Traffic from Coimbra should go preferably by Lisboa
- ♦ Traffic from Aveiro should go preferably by Porto

### @Porto

- ♦ Route-map applied to all BGP announced external routes/nets
- ♦ Adds BGP attribute: **Community 8657:1**

### @Lisboa

- ♦ Route-map applied to all BGP announced external routes/nets
- ♦ Adds BGP attribute: **Community 8657:2**

### @Aveiro

- ♦ Route-map applied to all BGP received routes/nets
- ♦ If **Community 8657:1** → **Local-preference 200**
- ♦ If **Community 8657:2** → **Local-preference 100**

### @Coimbra

- ♦ Route-map applied to all BGP received routes/nets
- ♦ If **Community 8657:1** → **Local-preference 100**
- ♦ If **Community 8657:2** → **Local-preference 200**





# BGP Community Attribute (real data)

## TeliaNet Global Network

```
remarks: BGP COMMUNITY SUPPORT FOR AS1299 TRANSIT CUSTOMERS:
remarks:
remarks: Community Action
remarks: -----
remarks: 1299:50 Set local pref 50 within AS1299 (lowest possible)
remarks: 1299:150 Set local pref 150 within AS1299 (equal to peer, backup)
remarks:
remarks: European peers/ix-points      US peers/ix-points      Asia peers/ix-points
remarks: Community Action              Community Action        Community Action
remarks: -----
remarks: 1299:200x All peers Europe incl: 1299:500x All peers US incl: 1299:700x All peers Asia incl:
...
remarks: 1299:250x Sprint/1239          1299:550x Sprint/1239      -
remarks: 1299:251x Savvis/3561          1299:551x Savvis/3561      -
remarks: 1299:252x Verio/2914            1299:552x Verio/2914      -
remarks: 1299:253x Abovenet/6461         1299:553x Abovenet/6461    -
remarks: 1299:254x FT/5511                1299:554x FT/5511         1299:754x FT/5511
remarks: 1299:255x GBLX/3549              1299:555x GBLX/3549        1299:755x GBLX/3549
remarks: 1299:256x Level3/3356            1299:556x Level3/3356      -
remarks: 1299:257x UUnet/702              1299:557x UUnet/701        -
remarks: 1299:558x AT&T/7018              1299:758x AT&T/2687
remarks: 1299:259x Telefonica/12956        1299:559x Telefonica/12956  -
remarks: 1299:260x BT/Concert/5400        -
remarks: 1299:261x Qwest/209              1299:561x Qwest/209        -
remarks: 1299:263x Teleglobe/6453          1299:563x Teleglobe/6453   -
remarks: 1299:264x DTAG/3320              1299:564x DTAG/3320        -
remarks: 1299:268x AOL/1668                1299:568x AOL/1668         -
remarks: 1299:269x Tiscali/3257            1299:569x Tiscali/3257     1299:769x Tiscali/3257
remarks: 1299:270x UPC/6830                -
remarks: 1299:273x Cogent/174              1299:573x Cogent/174        -
remarks: 1299:274x Telecom Italia/6762      1299:574x Telecom Italia/6762 1299:774x Telecom Italia/6762
remarks: 1299:275x Tele2/1257              -
...
remarks: 1299:284x Cable & Wireless DE/1273 1299:584x Cable & Wireless DE/1273 -
remarks: 1299:286x KPN/286                  -
remarks: 1299:287x China Netcom/4837        1299:587x C
remarks: 1299:288x China Telecom/4134        1299:588x C
```

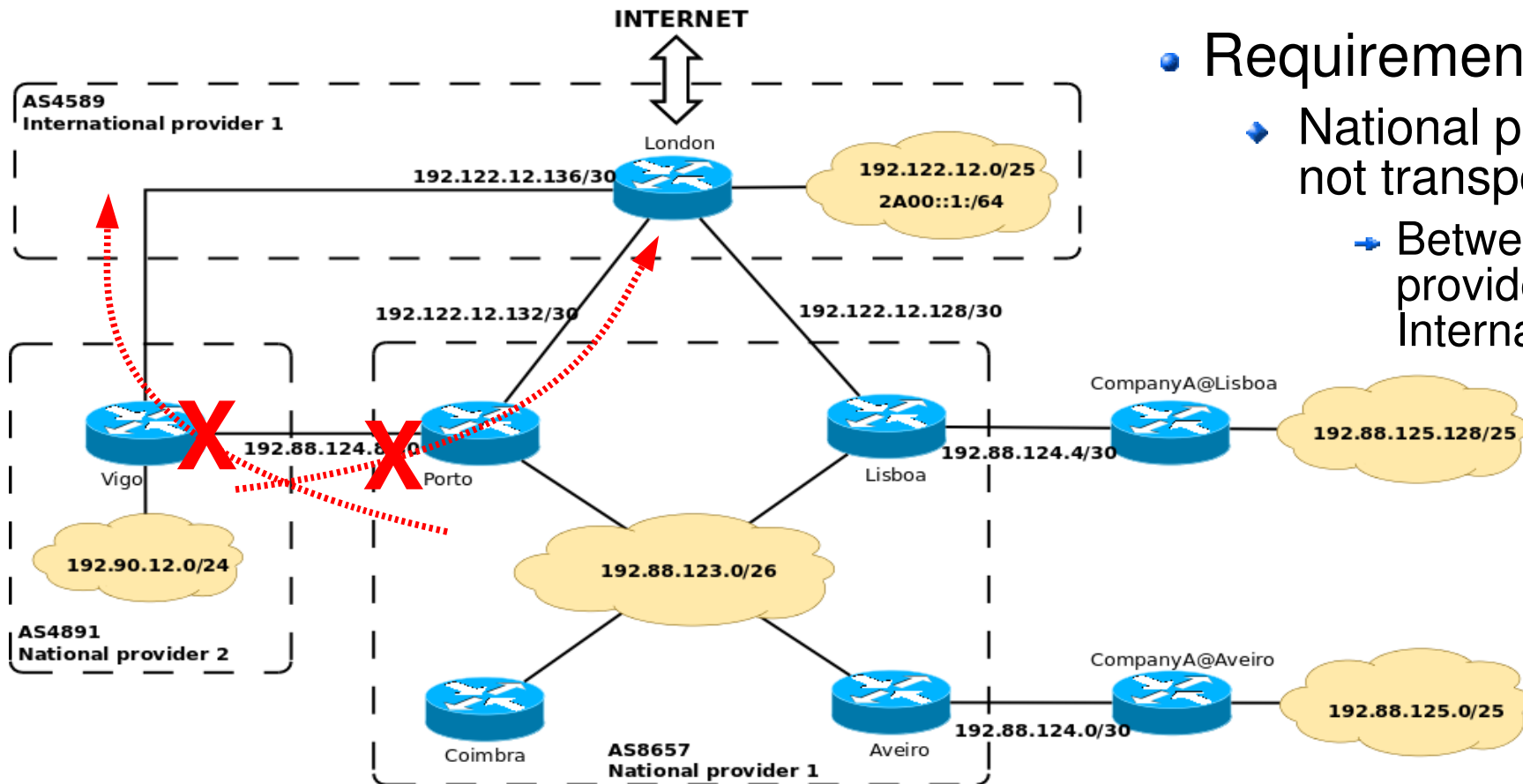
From RIPE database

<https://apps.db.ripe.net/>

e.g., <https://apps.db.ripe.net/db-web-ui/#/query?bflag=false&dflag=false&rflag=true&searchtext=as1299>



# BGP Case Studies



## Requirements

- ◆ National providers should not transport traffic
  - ➔ Between other national providers and the International provider

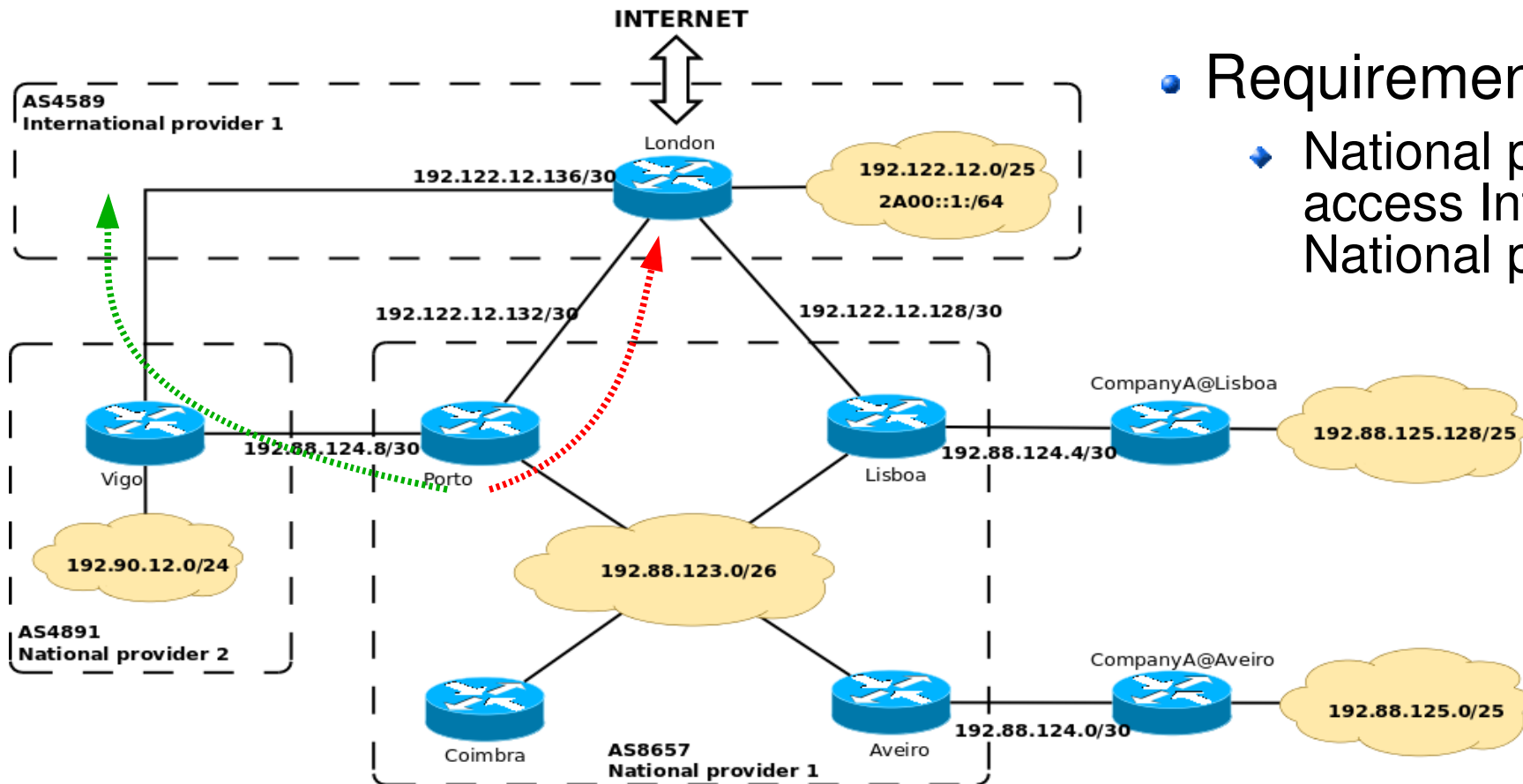
### @Porto, @Lisboa

- ◆ Route-map applied to all external BGP announcements
- ◆ Announce only internal routes/nets
  - ➔ Empty path “^\$”

### @Vigo

- ◆ Route-map applied to all external BGP announcements
- ◆ Announce only internal routes/nets
  - ➔ Empty path “^\$”

# BGP Case Studies



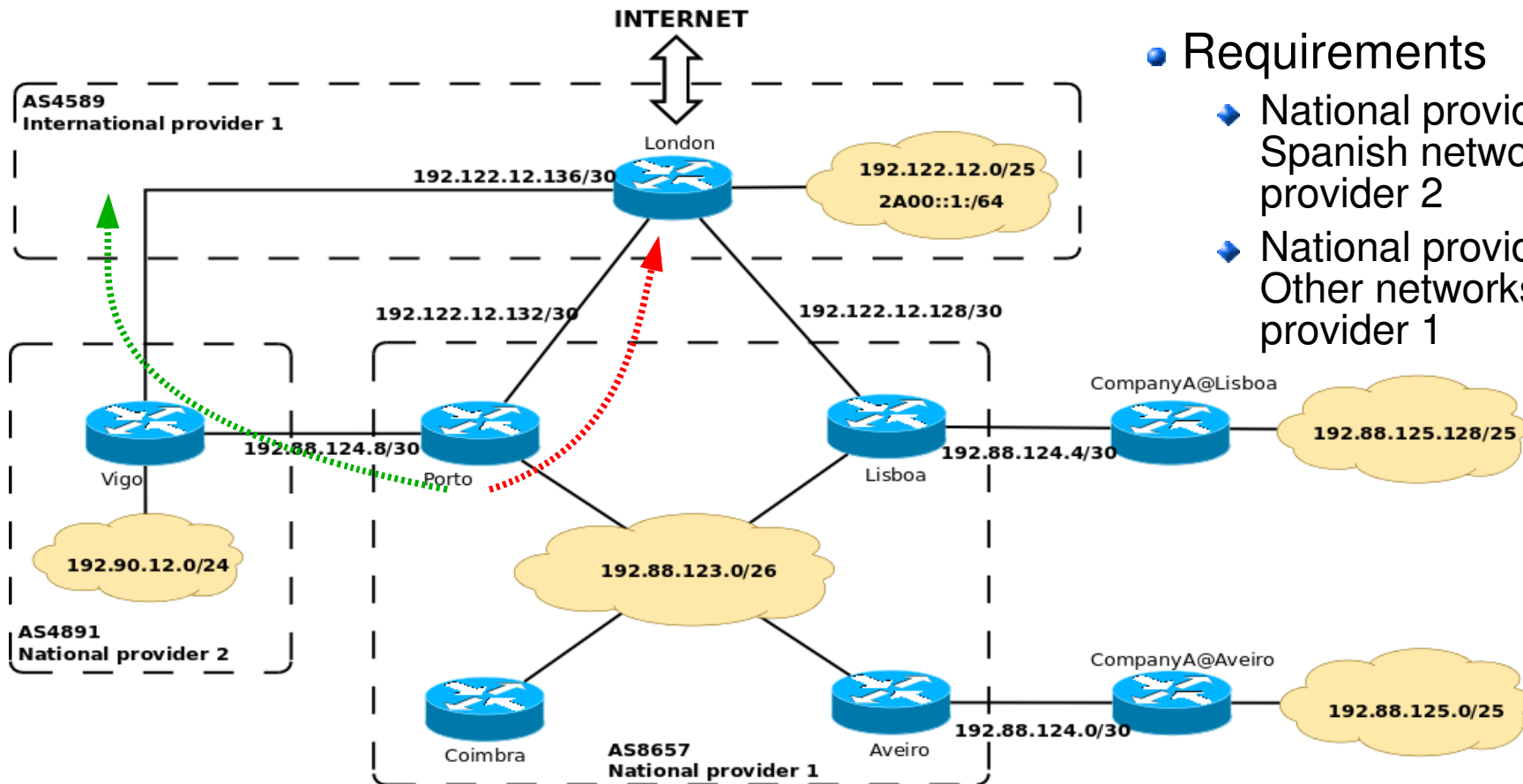
- Requirements

- National provider 1 should access Internet using National provider 2

- @Porto, @Lisboa

- Route-map applied to all BGP announcements received
- If Path contains "4891" → **Local-preference 200**
- If Path does not contain "4891" → **Local-preference 100**

# BGP Case Studies



## Requirements

- ◆ National provider 1 should access Spanish networks using National provider 2
- ◆ National provider 1 should access Other networks using International provider 1

## @Porto, @Lisboa

- ◆ Route-map applied to all BGP announcements received
  - E.g. known Spanish operators AS: 4891, 7654, 9876 and 3352
- ◆ If Path starts (from right to left) with “4891\$ or 7654\$ or 9876\$ or 3352\$” and ends in “^4891” → **Local-preference 200**
- ◆ If Path does not start with “4891\$ or 7654\$ or 9876\$ or 3352\$” and ends in “^4891” → **Local-preference 50**
- ◆ Assuming default Local-preference 100.

# BGP Case Studies

- Requirements

- AS104 wants to avoid paths to Net3 that use “slow” links.

