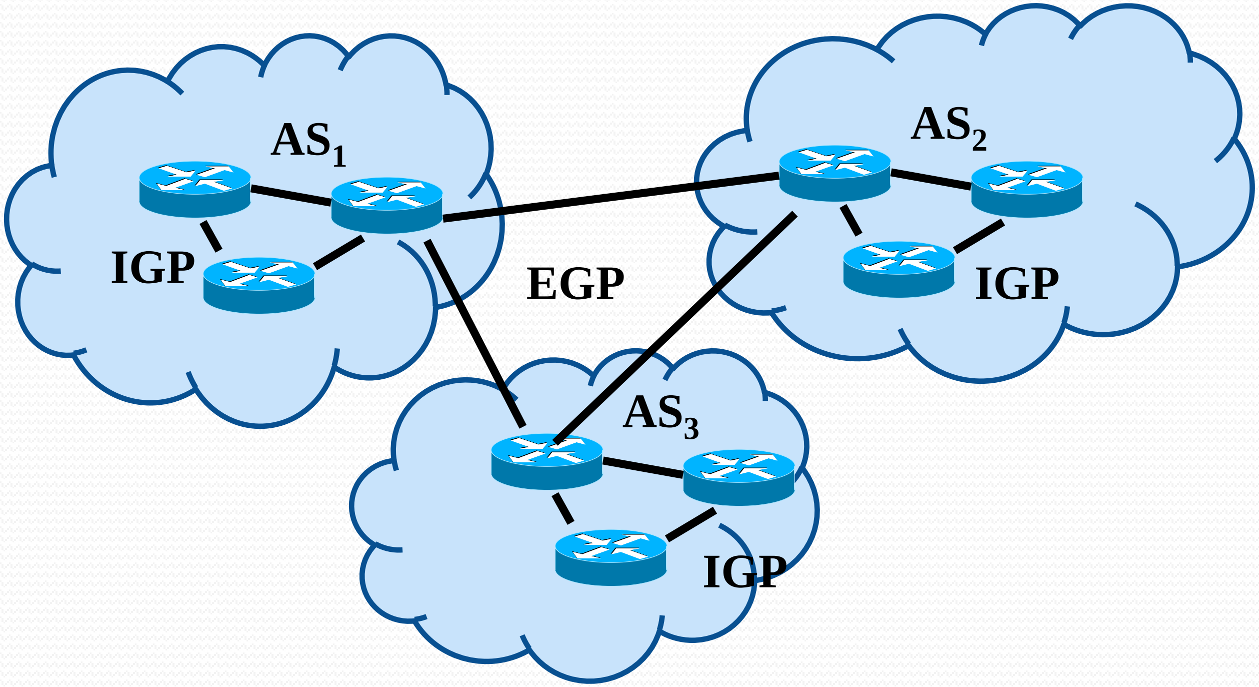
PI-Grau
(Internet Protocols)

José M. Barceló Ordinas
Departamento de Arquitectura de Computadores
(UPC)

- **Topic 4: Inter-domain Routing.**
  - Objectives
    - Introduce basic **inter-domain routing** concepts
    - Understand **BGP attributes**
    - Understand **Peer-to-peer relationships** among ISP
    - Learn **multi-homing** techniques

Topic 4: Inter-domain Routing

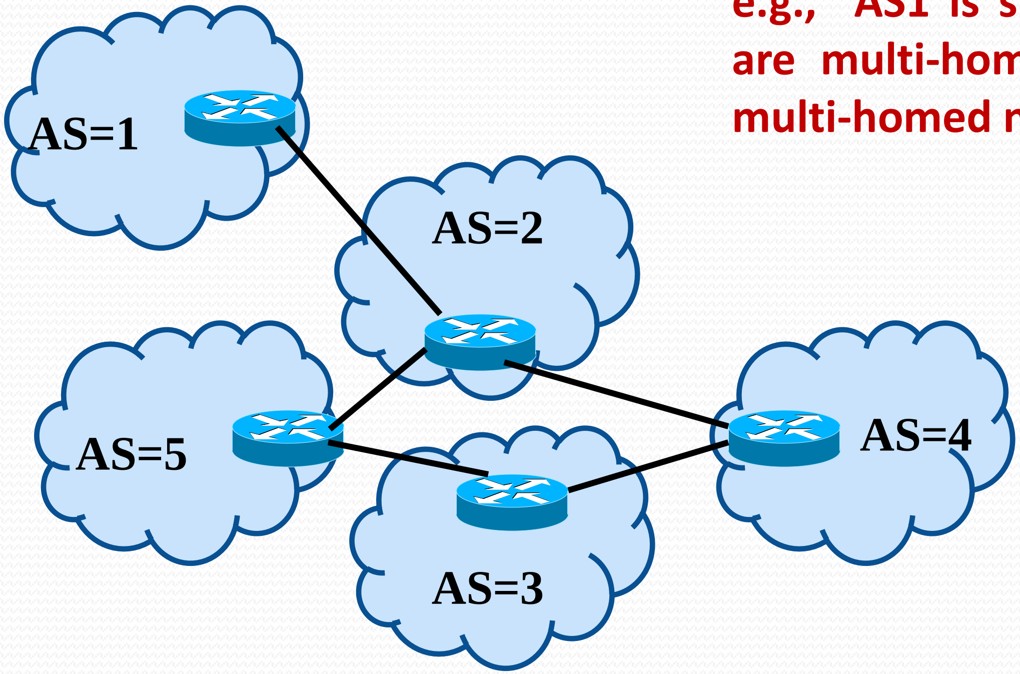
- **Autonomous Systems (AS):** set of routers with the same routing policy in a unique administrative domain
  - AS are identified with 16 bits (65535 AS’s)
    - **AS’s ≥ 64512** are private numbers (same as IP private addresses)
  - AS’s exchange routes, IP subnets, using External Routing Protocols (EGP) such as BGPv4



- **BGPv4**:
  - Is a routing protocol based on **policies**
  - It does **not** use **routing metrics** (hops, bandwidth, delay, ...)
  - Uses routing attributes that allow defining routing policies
  - BGP is encapsulated in TCP packets, thus, between two BGP routers should exist a TCP connection for each direction
- **ISP (Internet Service Provider)**
  - An ISP is an administrative entity that may have one or more AS numbers assigned depending of its architecture and geographical situation
  - In general an AS number may be assigned to an ISP or to a Corporative Network, thus, not all AS are ISP, however all ISPs have one or more AS number assigned

AS types of operation:

- **Stub AS or single-homed:** AS that reaches routes of other AS's using a **single connection point**
- **Multi-homed AS:** AS that reaches routes of other AS's using more than one connection point but do not transit routes of other AS are called **Multi-homed non-transit**, while if they transit routes are called **Multi-homed transit.**



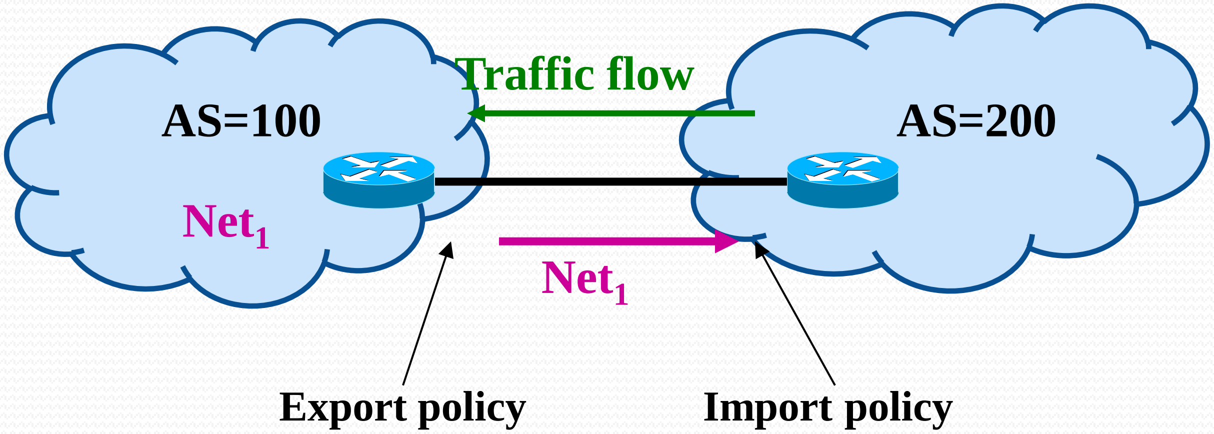
e.g., AS1 is stub, AS2, AS3 and AS4 are multi-homed transit and AS5 is multi-homed non-transit AS

- **RIPE-496: “Autonomous System (AS) Number Assignment Policies and Procedures”**

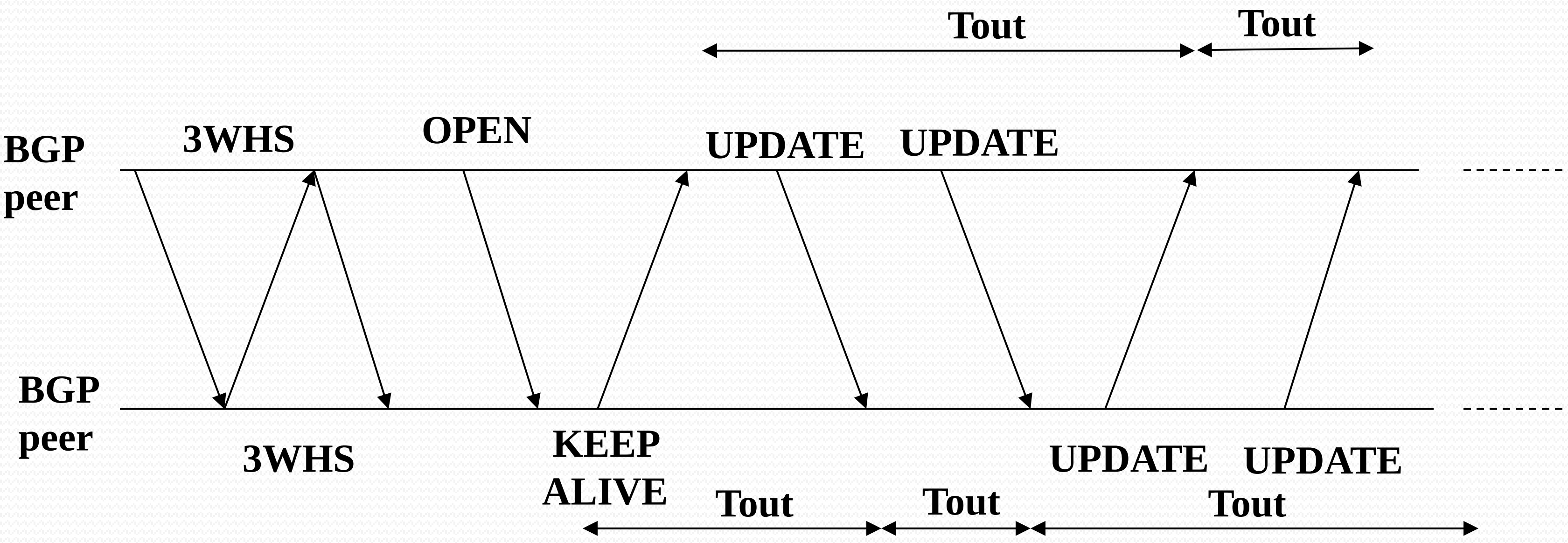
  http://www.ripe.net/ripe/docs/ripe-496

  - document indicating AS# assignment policies
  - “An Autonomous System (AS) is a group of IP networks run by one or more network operators with a single, clearly defined routing policy. When exchanging exterior routing information each AS is identified by a unique number”
  - If a Corporate Network is connected to a unique AS, does not need an AS number, however, if requires a different routing policy with respect its AS, it may require a AS#
- **RFC 1930,** "Guidelines for creation, selection, and registration of an Autonomous System (AS)“
  - It is obligatory that AS are multi-homed and that registers its routing policy in its RIR (Regional Internet Register) using RPSL (Routing Policy Specification Language)
  - Single-homed should use its Provider routing policy

- **BGPv4 routing protocol:**
  - Announce routes, IP subnets, using administrative routing policies to other AS
  - **Routing policy:** assume a subnet belongs to an AS, a routing policy means the decision of an AS to announce that route to other AS **("export policy")** and is the privilege of the other AS accept the route **("import policy")**
    - Combination of export and import policies define whether routes flows and this the direction in which information flows



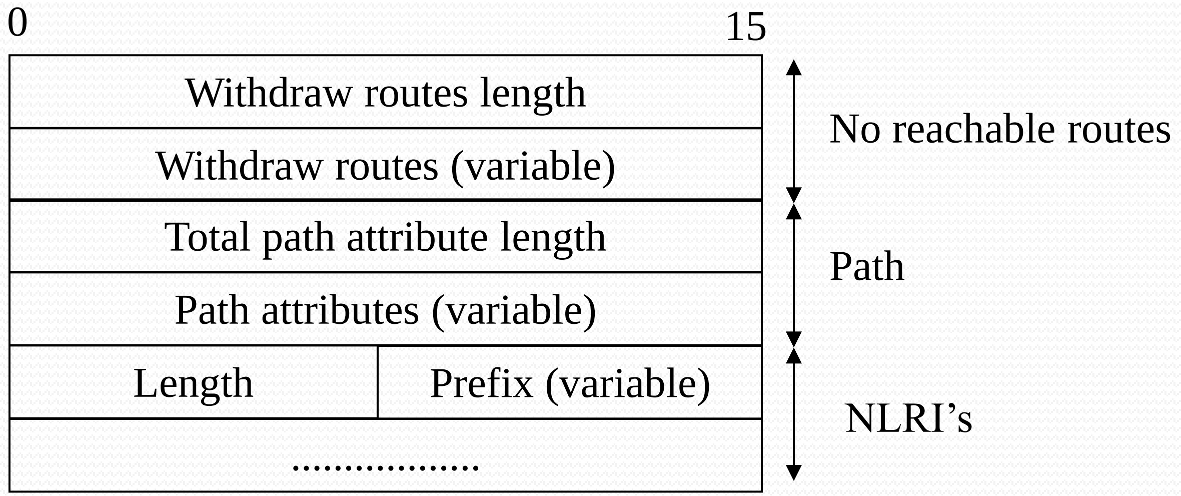
- **BGPv4 routing protocol:**
  - BGP packets:
    - BGP routers send packets encapsulated in TCP segments
    - The following types are defined
      - **OPEN**: create BGP connections
      - **KEEPALIVE**: test whether the TCP connection sis alive
      - **UPDATE**: send routes and attributes
      - **NOTIFICATION**: error notification



BGP Attributes:

- An UPDATE BGP message carry attributes indicating the routing policies that BGP should held
  - An attribute can be of the following types
    - **Well-known** attributes should be supported by all BGP implementations while **optional** attributes are not necessarily supported by BGP implementations
    - **Mandatory** attributes are always sent in UPDATE messages while **discretional** attributes are not sent in UPDATE messages (both combined only with well-known attributes)
    - **Transitive** and **non-transitive** attributes means that the routes transit or no-transit to other routers (both combined only with optional attributes)
    - **Complete** attribute is used if all the routers that transit an attribute implement the optional attributes and **Partial** attribute is used if only part of the routers that transit an attribute implement the optional attributes (both only used if optional transitive)
  - Not all combinations are possible, only the following ones:
    - **Well-known and mandatory**: AS-PATH, NEXT-HOP, ORIGIN
    - **Well-known and discretional**: LOCAL-PREFERENCE, ATOMIC AGGREGATE
    - **Optional and transitive** AGGREGATOR, COMMUNITY
    - **Optional and non-transitive**: MED (also called “metric”)

- **UPDATE messages**
  - **Withdraw routes length**: number of routes (in bytes) that the BGP has to withdraw. The field **withdraw routes** indicates the routes to be withdrawed (if 0 then there are no routes to withdraw).
  - **Total path attributes**: length of the vector of routes specified in the filed “path attributes”
  - **Path attributes**: list and description of the attributes
  - **NLRI** (Network Layer Reachable Information): routes to which the path attributes apply



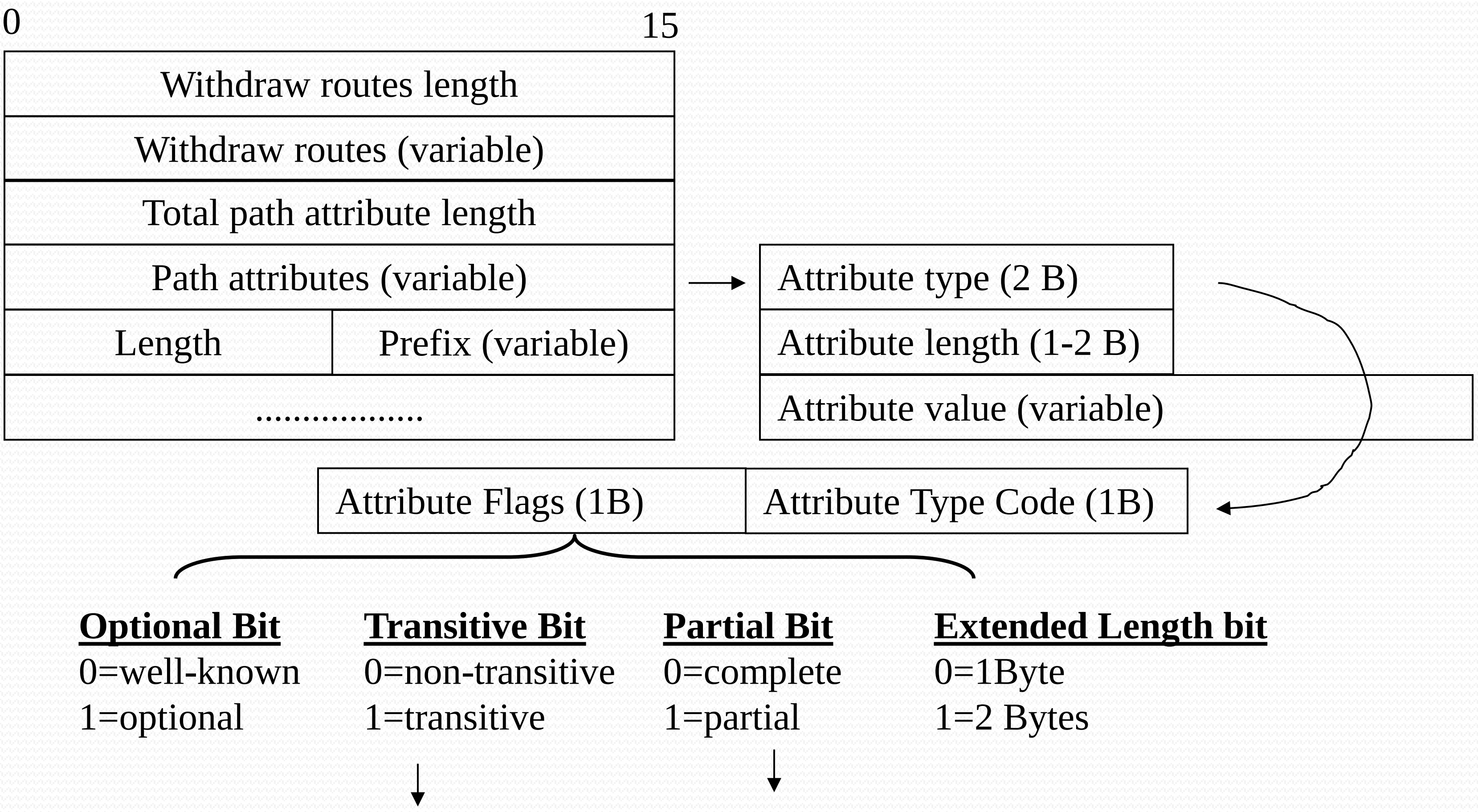
0	15
Withdraw routes length	
Withdraw routes (variable)	
Total path attribute length	
Path attributes (variable)	
Length	Prefix (variable)
....	

No reachable routes

Path

NLRI’s

- **UPDATE messages**



0 15

Withdraw routes length	
Withdraw routes (variable)	
Total path attribute length	
Path attributes (variable)	
Length	Prefix (variable)
....	

Attribute type (2 B)
Attribute length (1-2 B)
Attribute value (variable)

Attribute Flags (1B)	Attribute Type Code (1B)

Optional Bit
0=well-known
1=optional

Transitive Bit
0=non-transitive
1=transitive

Partial Bit
0=complete
1=partial

Extended Length bit
0=1Byte
1=2 Bytes

Whether the optional information is transitive or non-transitive

Whether the information in the optional transitive is complete or partial

BGP routing table

- Includes the following information: **subnet and mask**, **next-hop**, **MED** (metric), **Local_Pref**, **AS-path-vector** and **origin**
- An AS may be connected to "N" AS's, thus will receive an UPDATE message with a possible path for a route for each E-BGP connection → "N" entries per route → **decision process** chooses the best entry as a function of attributes. The decision process depends on:
  - Manufacturer implementation, but basically all BGP routing tables are very similar
  - Maintain a DB for each active BGP session
  - Symbol **">"** indicates **the best entry** towards a route
  - A router only announces its "best route" in UPDATE BGP messages
    - Partial Internet view since a router **only sees** what other routes decide to sent

BGP routing table

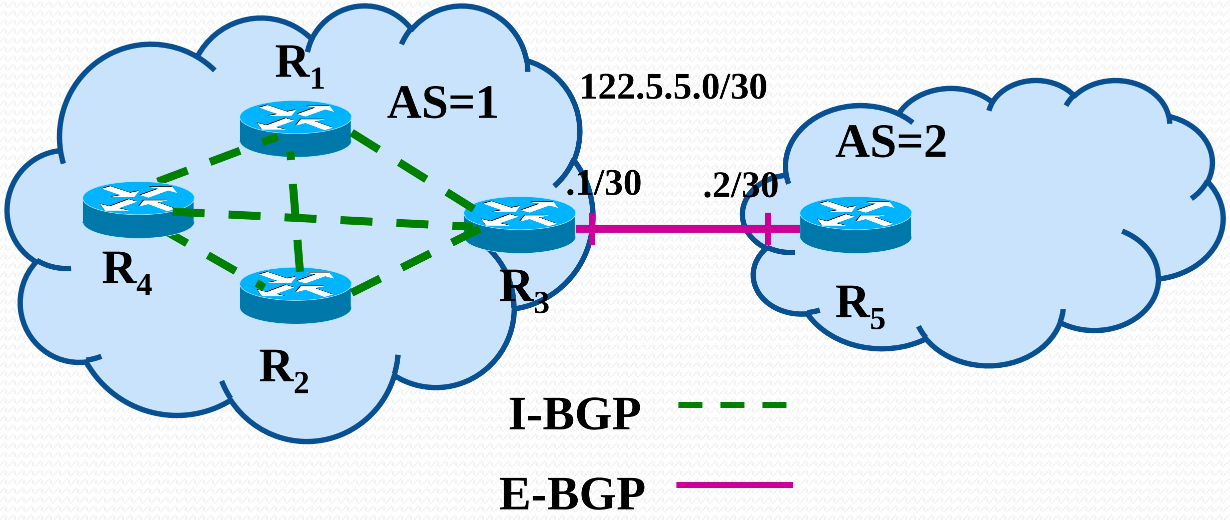
```
R2# show ip bgp
                                              Attributes

Network       Next Hop          Metric    LocPrf    AS-Path                        Origin
*  4.0.0.0    206.157.77.11       75        100      1673 1                           i
*>            12.127.0.249         0        200      7018 1                           i
*             204.70.4.89          0        100      3561 1                           i
*             204.42.253.253       0        200      267 1225 1239 1                  i
*             205.158.2.126        0        200      2828 4908 3561 1                 i

……………………………..

*  6.0.0.0    206.157.77.11       105       100      1673 1239 568 721 1455  i
*             12.127.0.249         0        100      7018 7170 1455                   i
*>            198.32.8.252         0        100      11537 7170 1455                  i
*             204.70.4.89          0        100      3561 568 721 1455                i
```

- **BGPv4 routing protocol:**
  - BGP Routers exchange routes. Each route has a list of attributes that allows other BGP routers to fix a policy with respect that route
  - BGP sessions: two BGP routers that open a TCP session on port 179 are called **neighbors** or **peers**
    - Two BGP routers belonging to the same AS use Internal BGP (I-BGP)
    - Two BGP routers belonging to different AS use External BGP (E-BGP)
      - **CAREFUL !!!** There is only one BGP protocol, however I-BGP and E-BGP operate differently

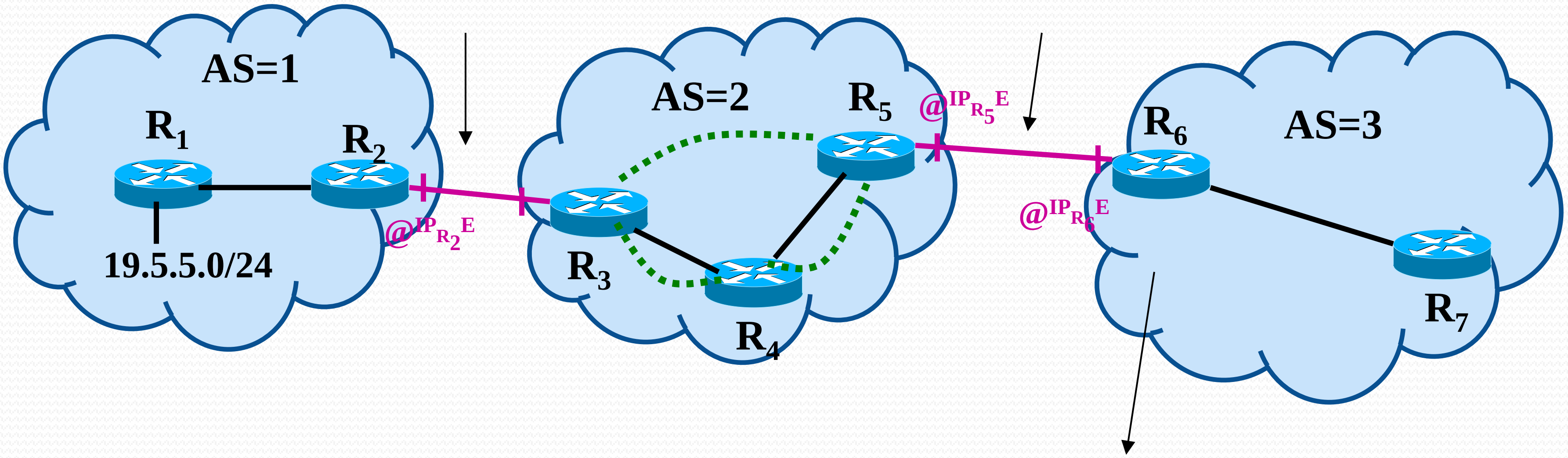


AS-PATH VECTOR (Well-known and Mandatory):

- Represents the path that a route follows from the origin AS → gives all AS traversed to reach that route
- AS-PATH = (AS_x, …., AS_{origin})
- Each AS adds its AS# to the AS-PATH vector

19.5.5.0/24: AS-PATH=(1)

19.5.5.0/24: AS-PATH=(2,1)

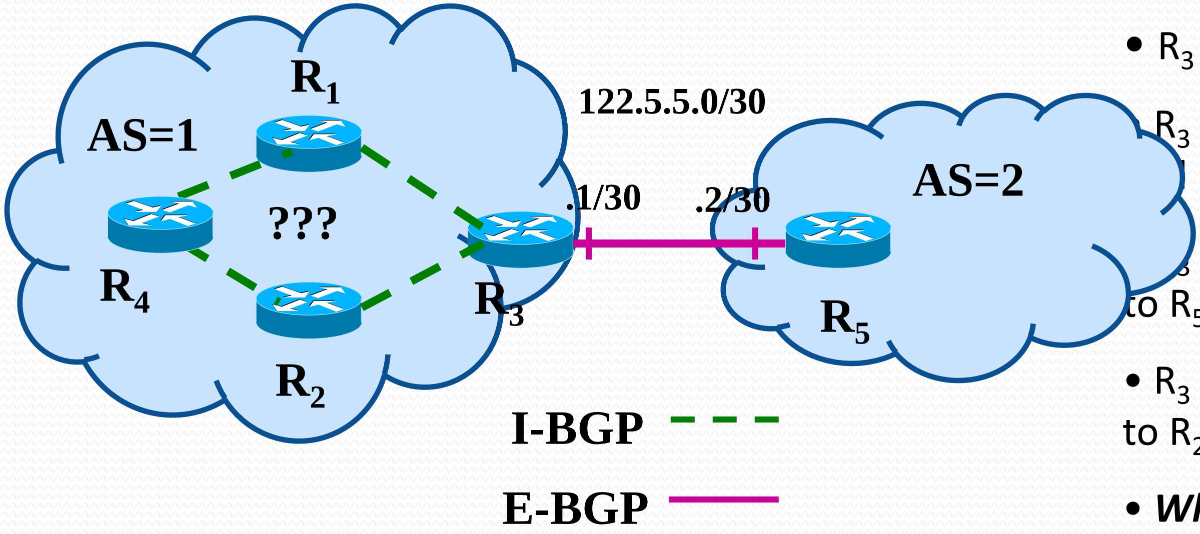


Path to network 19.5.5.0/24 is AS's (2,1), next-hop is $@IP_{R_5}$

I-BGP

- I-BGP is used to coordinate the routing policy inside the AS, furthermore is needed for allowing transit of external routes through the AS
  - Subnets learnt via E-BGP may be advertised via E-BGP and I-BGP
  - Subnets learnt via I-BGP only may be advertised via E-BGP
  - I-BGP routers **DO NOT** advertise routes learnt via I-BGP to other I-BGP neighbors

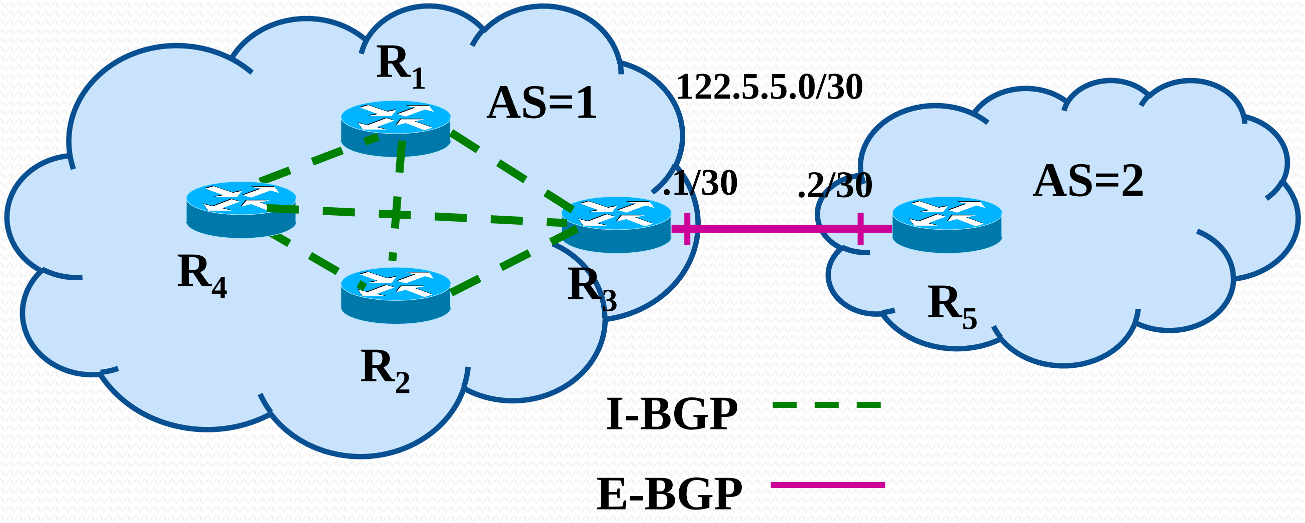
  → I-BGP routers should form a mesh I-BGP network → **problem of scalability that is solved via "BGP route reflectors" a "BGP confederations"**



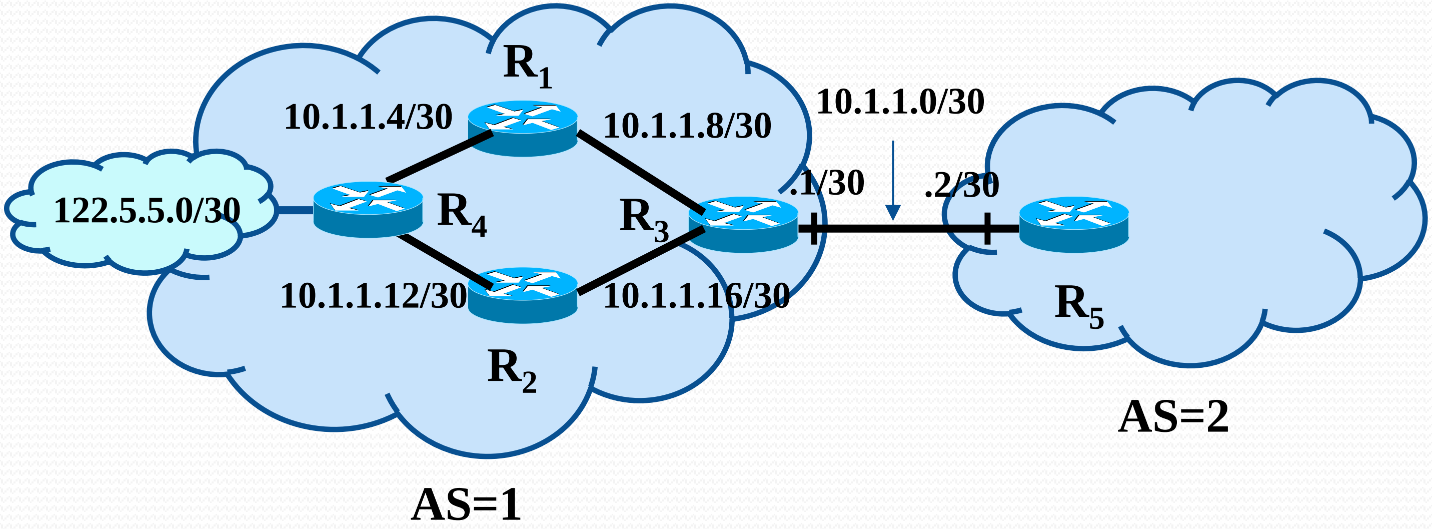
- R$_3$ learn routes from R$_5$, R$_1$ y R$_2$
- R$_3$ advertise routes learnt from R$_5$ to R$_1$ and R$_2$
- R$_3$ advertise routes learnt from R$_1$ and R$_2$ to R$_5$
- R$_3$ do not advertise routes learnt from R$_1$ to R$_2$ and learnt from R$_2$ to R$_1$
- ***What routes announce R$_1$?***

I-BGP

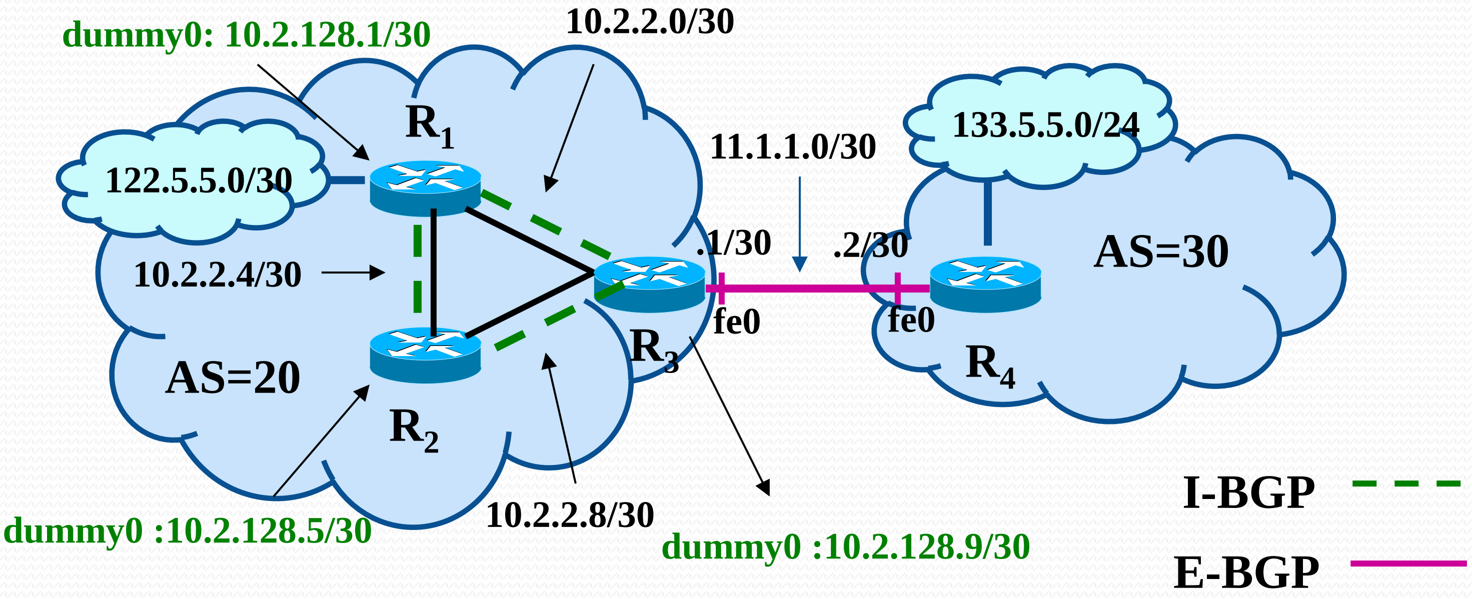
- Why I-BGP routers do not advertise routes learnt via I-BGP ?
  - BGP routers advertise an attribute called AS-path-vector that includes all the AS that the routes crosses.
  - **Objective**: **loop detection**. E.g.; vector AS=(1 2 5 6 2 8) crosses twice AS=2, → there is a loop
  - **Solution**: if a loop is detected do not advertise the route. *But then, what would happen if I-BGP announce routes learnt via I-BGP ?*



- BGP configuration with CISCO IoS

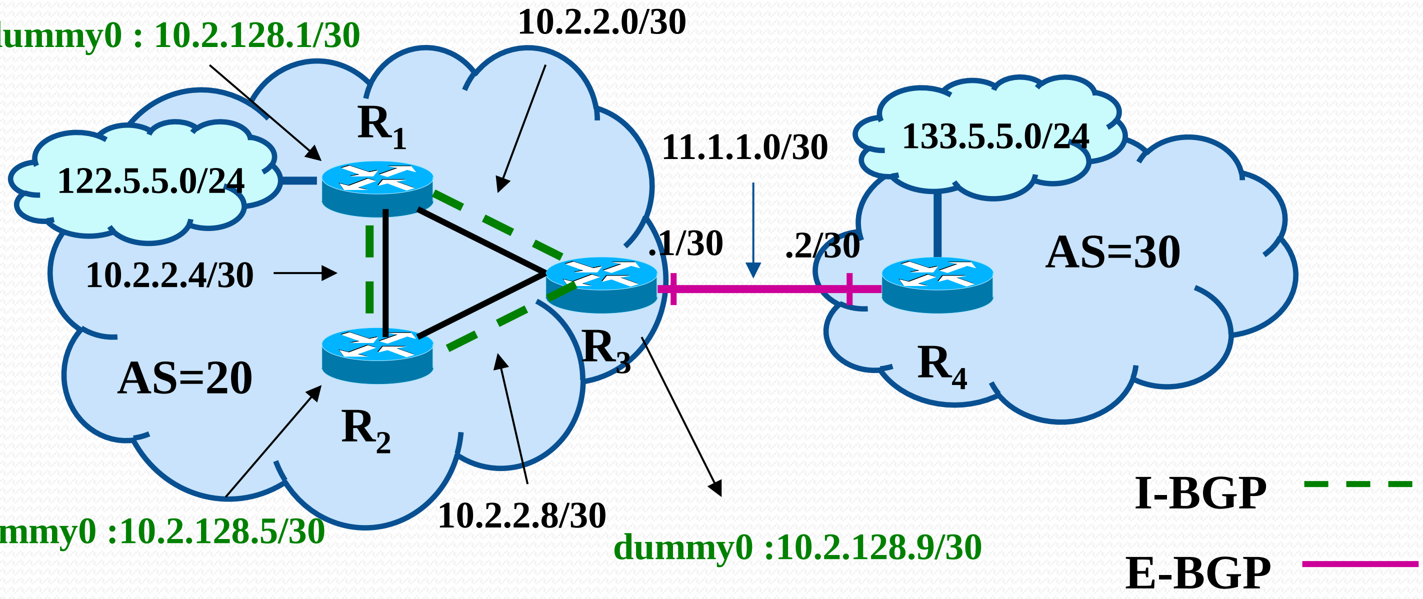


- **I-BGP (tricks):** I-BGP on loopback dummy interface
  - If the I-BGP session between R_1 and R_3 is built on IP addresses 11.2.2.1-11.2.2.2 and the link fails → the I-BGP session is lost
  - *Is it possible to reach R_3 from R_1?*
    - Use loopback interface with public/private IP addresses different from 127.0.0.0/8



- **BGP configuration with CISCO IoS**

Be carefull: AS's should be OSPF isolated (they belong to different domains).



BGP configuration with CISCO IoS

```
!!!! Configure OSPF in Router R1
R1(conf)# router ospf 1
R1(conf-r)# network 10.2.2.0   255.255.255.254    area 0
R1(conf-r)# network 10.2.2.4   255.255.255.254    area 0
R1(conf-r)# network 10.2.128.0   255.255.255.254    area 0
R1(conf-r)# network 122.5.5.0   255.255.255.0    area 0

!!!! Configure BGP in Router R1
R1(conf)# router bgp 20
R1(conf-r)# neighbor 10.2.128.5 remote-as 20
R1(conf-r)# neighbor 10.2.128.5 update-source dummy0
R1(conf-r)# neighbor 10.2.128.9 remote-as 20
R1(conf-r)# neighbor 10.2.128.9 update-source dummy0
R1(conf-r)# network 122.5.5.0   255.255.255.0
```

- **BGP configuration with CISCO IoS**

```
!!!! Configure OSPF in Router R2
R2(conf)# router ospf 1
R2(conf-r)# network 10.2.2.4   255.255.255.254    area 0
R2(conf-r)# network 10.2.2.8   255.255.255.254    area 0
R2(conf-r)# network 10.2.128.4   255.255.255.254    area 0

!!!! Configure BGP in Router R2
R2(conf)# router bgp 20
R2(conf-r)# neighbor 10.2.128.1 remote-as 20
R2(conf-r)# neighbor 10.2.128.1 update-source dummy0
R2(conf-r)# neighbor 10.2.128.9 remote-as 20
R2(conf-r)# neighbor 10.2.128.9 update-source dummy0
```

BGP configuration with CISCO IoS

```
!!!! Configure OSPF in Router R3
R3(conf)# router ospf 1
R3(conf-r)# network 10.2.2.0   255.255.255.254    area 0
R3(conf-r)# network 10.2.2.8   255.255.255.254    area 0
R3(conf-r)# network 10.2.128.8   255.255.255.254    area 0
R3(conf-r)# network 11.1.1.0   255.255.255.254    area 0
R3(conf-r)# passive-interface fe0        ← domain isolation

!!!! Configure BGP in Router R3
R3(conf)# router bgp 20
R3(conf-r)# neighbor 10.2.128.1 remote-as 20
R3(conf-r)# neighbor 10.2.128.1 update-source dummy0
R3(conf-r)# neighbor 10.2.128.5 remote-as 20
R3(conf-r)# neighbor 10.2.128.5 update-source dummy0
R3(conf-r)# neighbor 11.1.1.2 remote-as 30
```

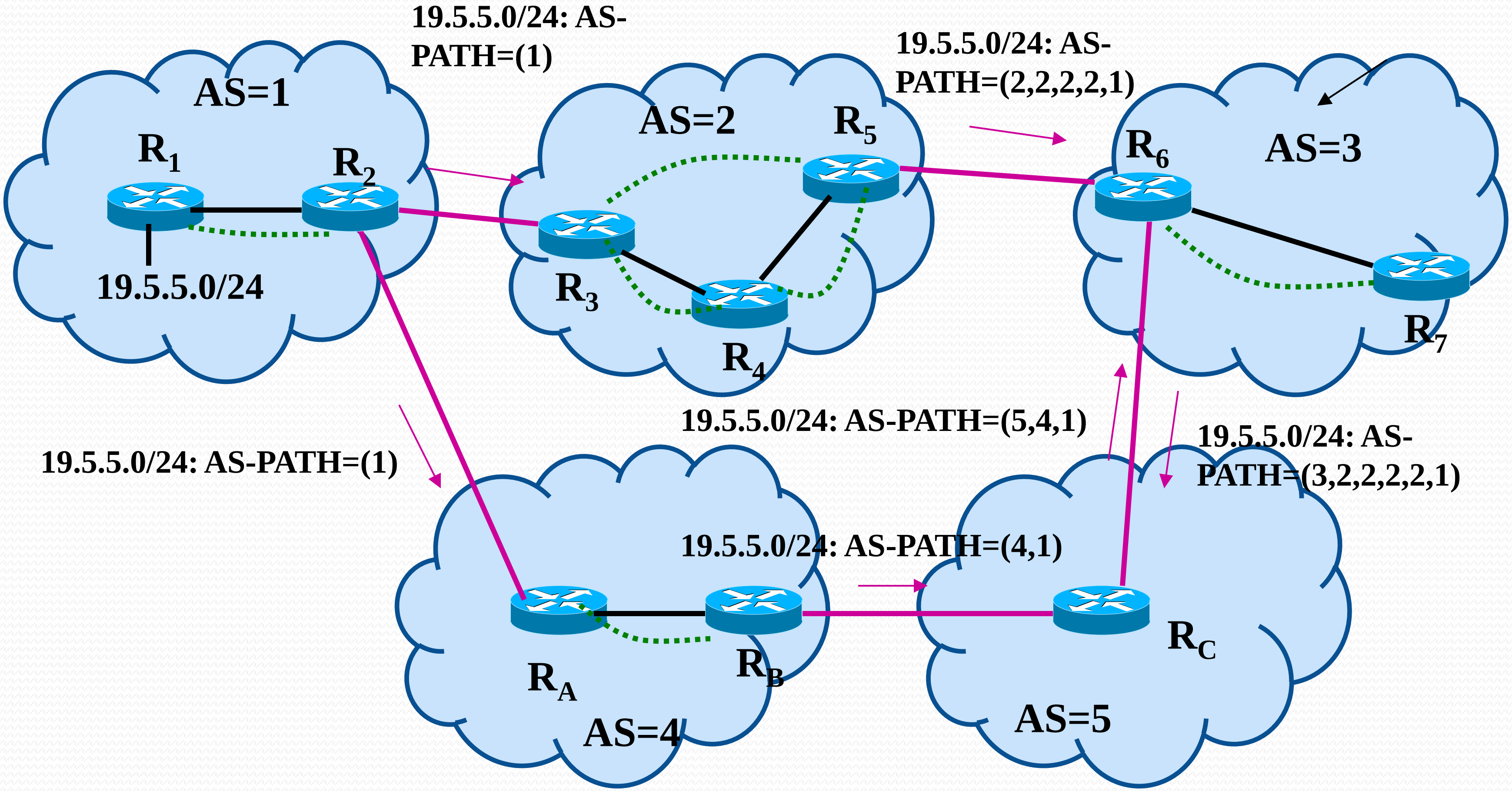
- **BGP configuration with CISCO IoS**

```
!!!! Configure OSPF in Router R4
R4(conf)# router ospf 1
R4(conf-r)# network 133.5.5.0   255.255.255.0    area 0
R4(conf-r)# network 11.1.1.0   255.255.255.254    area 0
R4(conf-r)# passive-interface fe0          ←  domain isolation

!!!! Configure BGP in Router R4
R4(conf)# router bgp 30
R4(conf-r)# neighbor 11.1.1.1 remote-as 20
R1(conf-r)# network 133.5.5.0   255.255.255.0
```

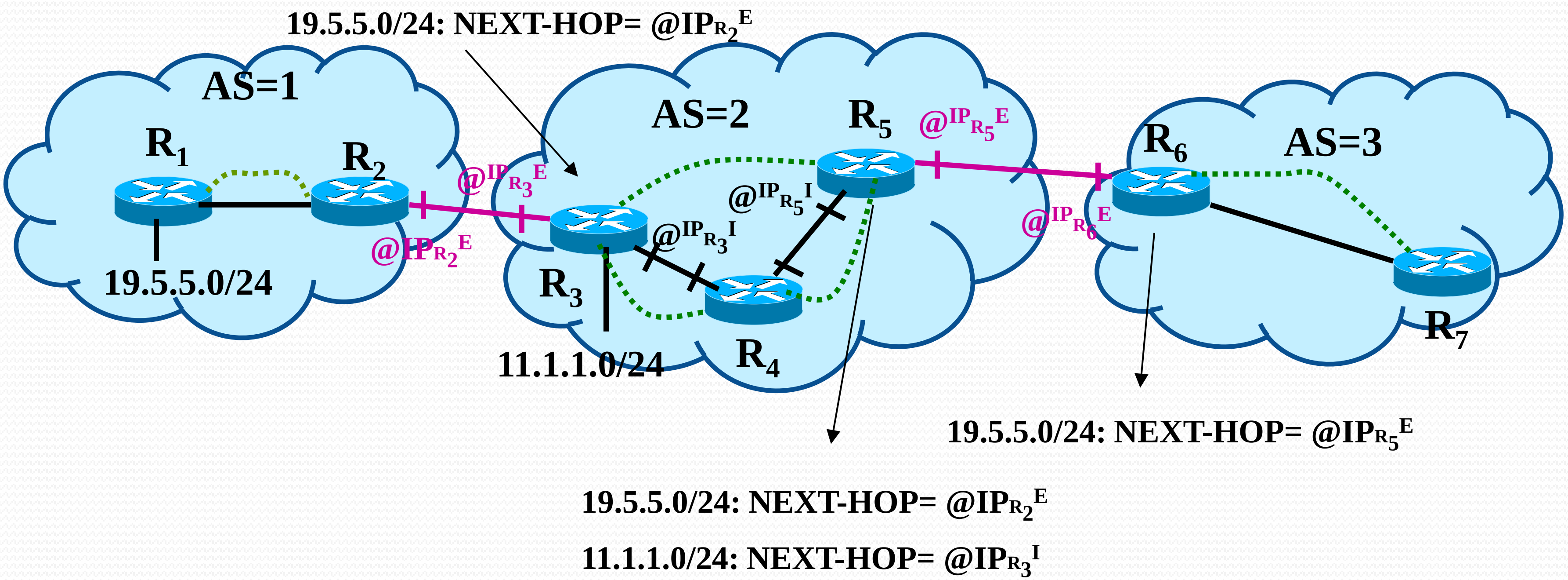
Manipulating the AS-PATH vector attribute:

- BGP **always** prefers the shortest path (in AS hops)
- Increase "prepending" the AS-PATH


19.5.5.0/24: AS-PATH=(1)
19.5.5.0/24: AS-PATH=(2,2,2,2,1)
AS=1
R1
R2
19.5.5.0/24
AS=2
R5
R3
R4
AS=3
R6
R7
19.5.5.0/24: AS-PATH=(5,4,1)
19.5.5.0/24: AS-PATH=(1)
19.5.5.0/24: AS-PATH=(3,2,2,2,2,1)
19.5.5.0/24: AS-PATH=(4,1)
RA
RB
AS=4
RC
AS=5

NEXT-HOP (Well-known and Mandatory):

- For a E-BGP session is the @IP of the BGP router that advertises the route
- For a I-BGP session is the @IP of the BGP router that advertises the route
  - If R_3 advertises 11.1.1.0/24 to R_5 it is necessary that R_5 knows how to arrive to **@IP$_{R_3}^{I}$**
- I-BGP sessions **do no change** the next-hop advertised by an E-BGP UPDATE message
  - If R_3 advertises 19.5.5.0/24 to R_5 it is necessary that R_5 knows how to arrive to **@IP$_{R_2}^{E}$**



19.5.5.0/24: NEXT-HOP= @IP$_{R_2}^{E}$

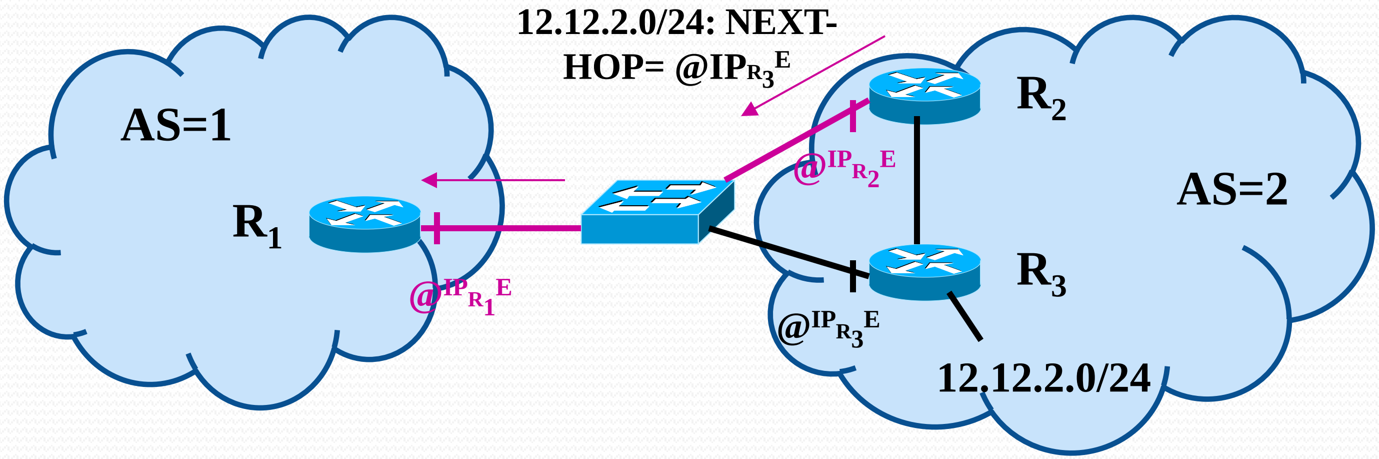
19.5.5.0/24: NEXT-HOP= @IP$_{R_5}^{E}$

19.5.5.0/24: NEXT-HOP= @IP$_{R_2}^{E}$

11.1.1.0/24: NEXT-HOP= @IP$_{R_3}^{I}$

NEXT-HOP in BMA networks:

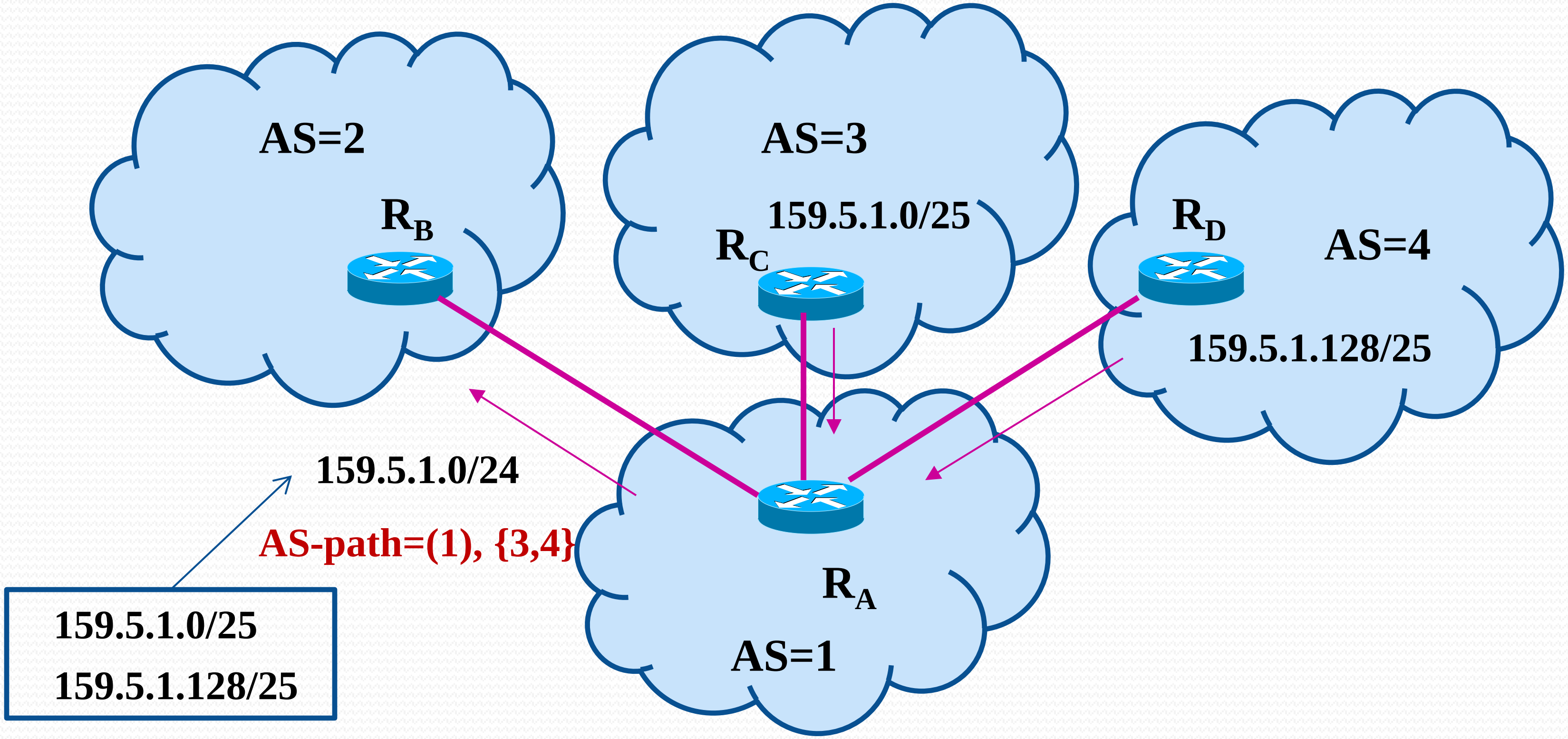
- R_1 and R_2 maintain a E-BGP connection. R_2 announces its networks with @IP$_{R2}$ as next-hop to R_1
- However when it has to announce network 12.12.2.0/24, @IP$_{R3}$ is best next-hop than @IP$_{R2}$ to reach R_1, so announces @IP$_{R3}$ instead of @IP$_{R2}$



- **ORIGIN (Well-known and Mandatory):**
  - Indicates who originated a route
    - **IGP**: the route was originated by an internal mechanism (in CISCO routers BGP network advertising is activated using the command “network”) and is indicated with the character “i” in the BGP routing table
    - **EGP**: the route was originated by the EGP protocol from am external AS and is indicated with the character “e” in the BGP routing table (EGP is obsolete and is not currently used)
    - **Incomplete**: unknown origin (e.g.; redistributed in BGP from internal IGP protocols such as RIP, OSPF, IS-IS) and is indicated with the character “?” in the BGP routing table

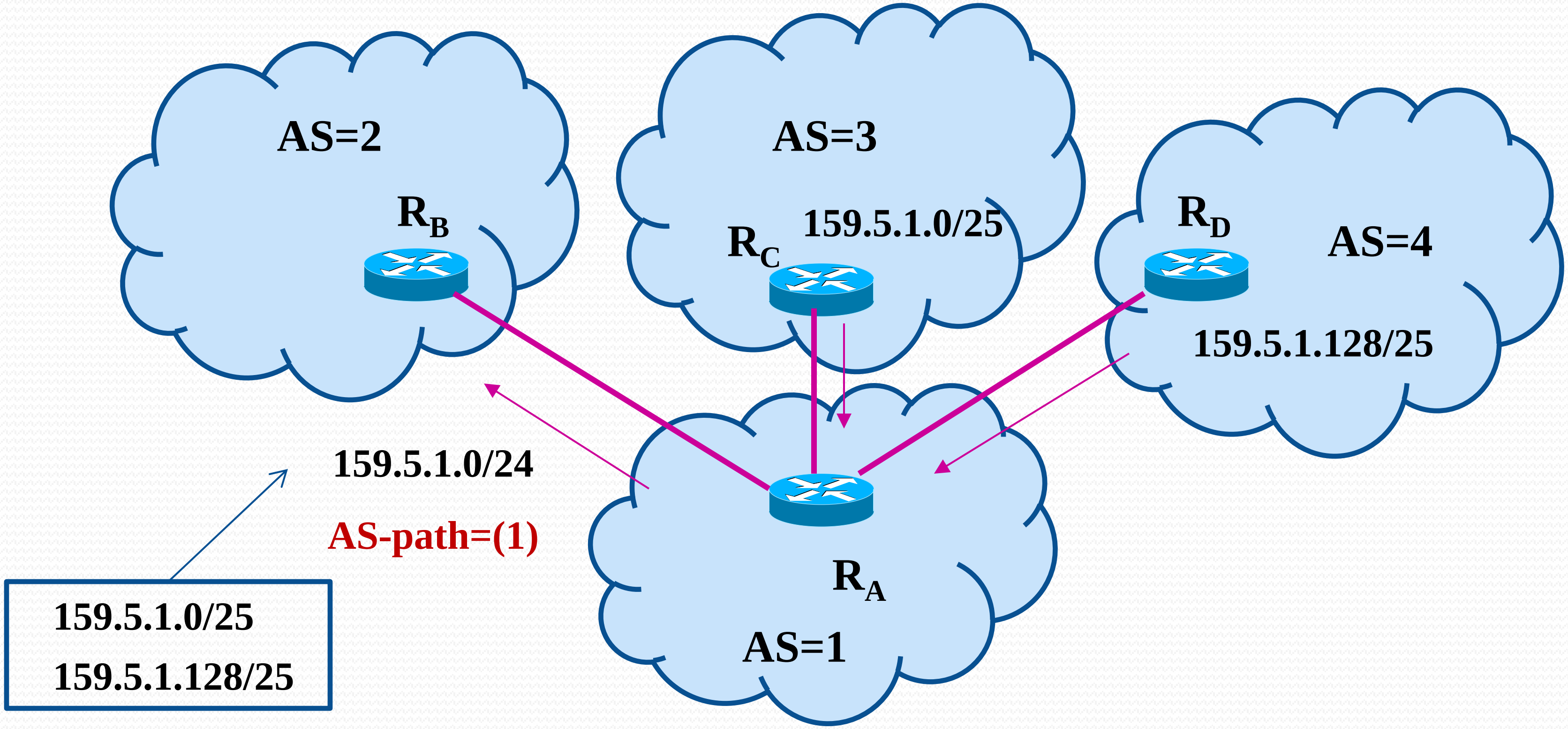
AGGREGATOR (Optional and transitive)

- A BGPv4 UPDATE message sends a subnet/mask that may be aggregated
- The BGP router that aggregates can indicate in the AS-PATH vector the partition of the subnet aggregated (**AS-SET option**),
- It has not influence in the path selection



ATOMIC AGGREGATE (Well-Known and discretional)

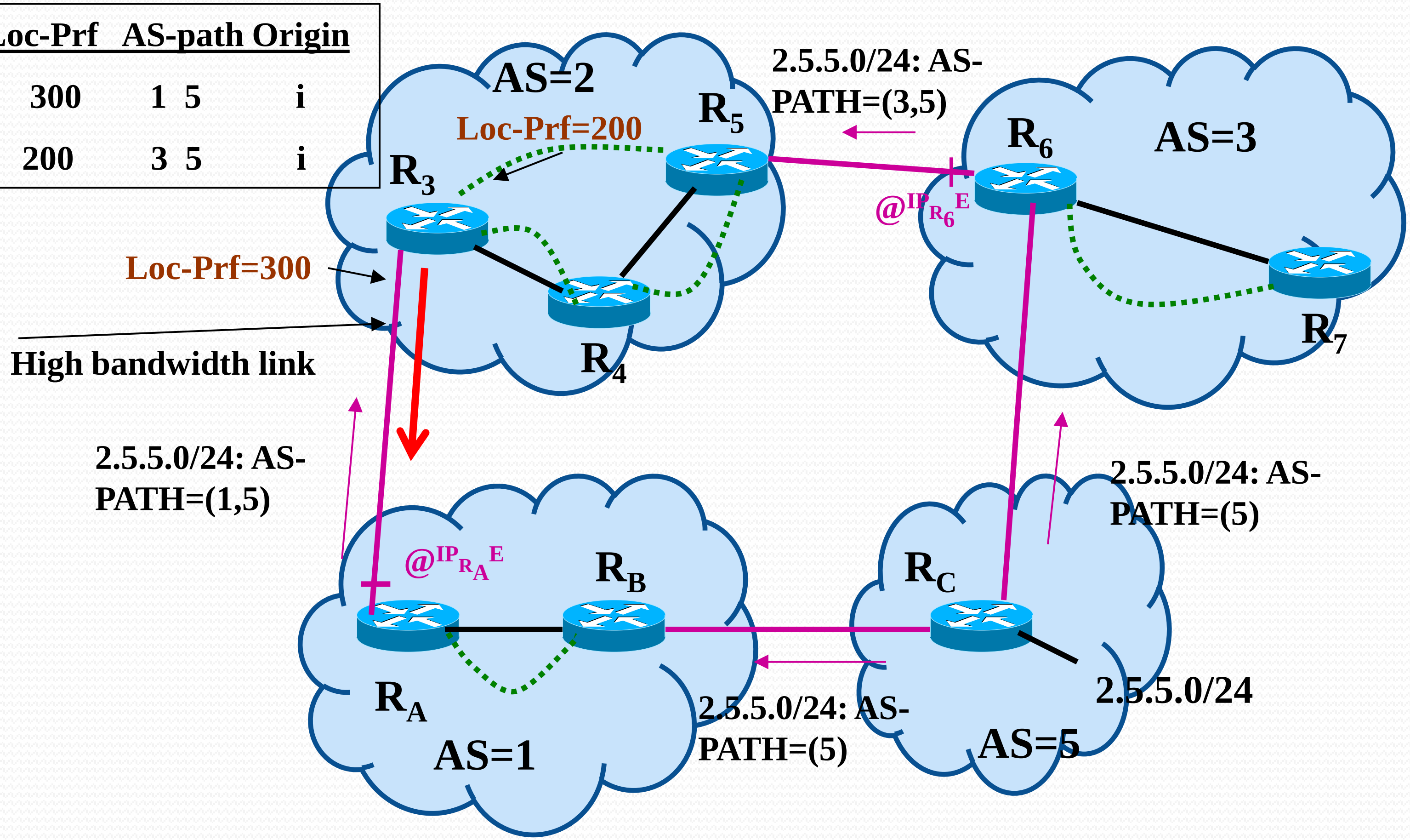
- The purpose of the attribute is to alert BGP speakers along the path that some information have been lost due to the route aggregation process and that the aggregate path might not be the best path to the destination.
- If when aggregating, the AS-SET has not been activated, then the AS-PATH vector can loss information of the original PATHs previous to aggregating → <u>it is mandatory that the Atomic Aggregate is active</u>



LOCAL-PREFERENCE (Well-Known and discretional):

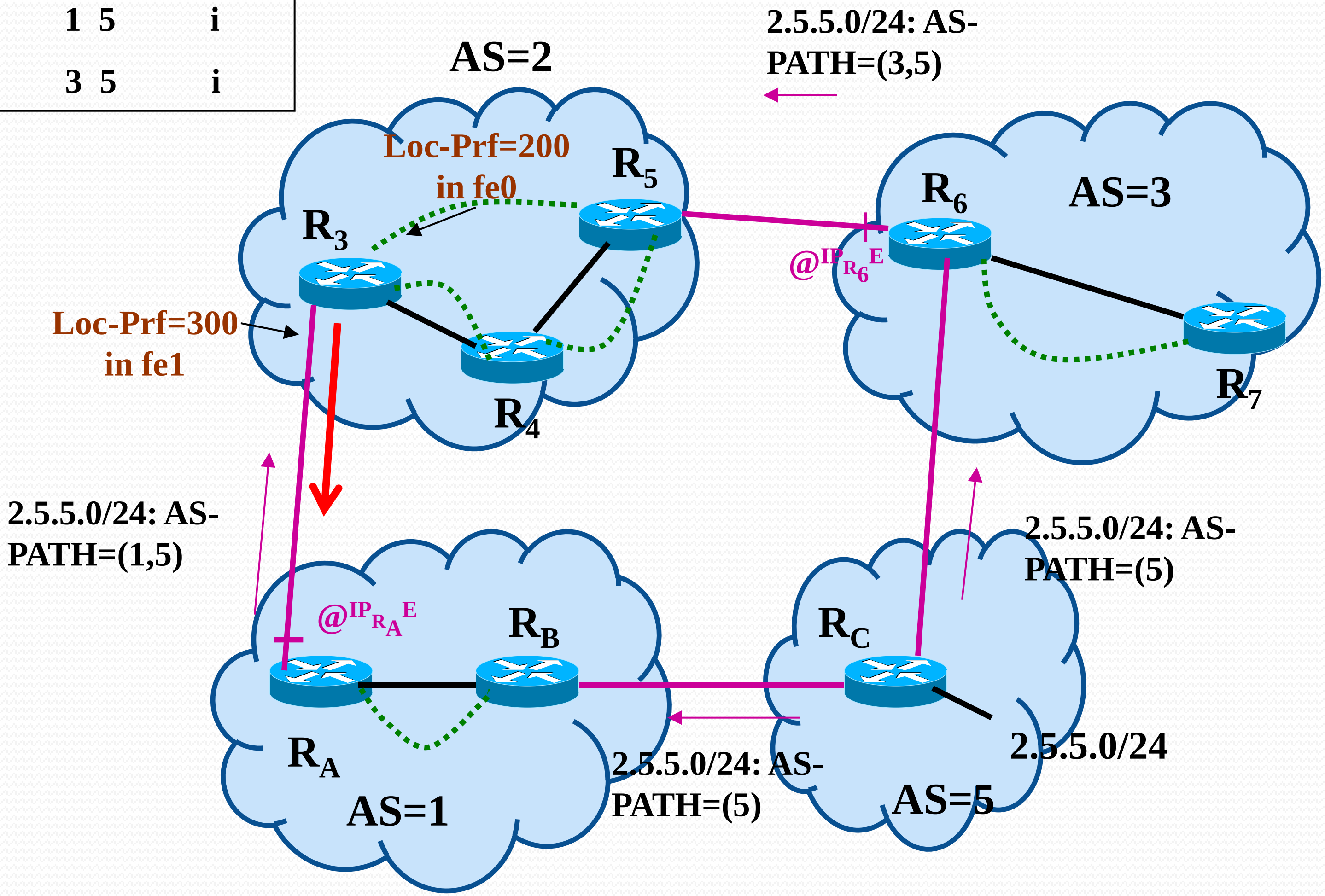
- Attribute that indicates the preferred output link
- Highest values of Loc-Prf have higher preference (default value=100)

Net	Next-hop	Loc-Prf	AS-path	Origin
>2.5.5.0/24	$@^{IP}{}_{R_A}{}^{E}$	300	1 5	i
2.5.5.0/24	$@^{IP}{}_{R_6}{}^{E}$	200	3 5	i



- **Local Preference with CISCO IoS**

Net	Next-hop	Loc-Prf	AS-path	Origin
>2.5.5.0/24	$@^{IP}{}_{R_A}{}^{E}$	300	1 5	i
2.5.5.0/24	$@^{IP}{}_{R_6}{}^{E}$	200	3 5	i



- ## BGP configuration with CISCO IoS
  - ### Route Maps:

route-map map-tag [**permit | deny**] [seq-number]

match: comando que especifica el criterio que debe ser comprobado

set: comando que indica la acción a ejecutar si el match aplica

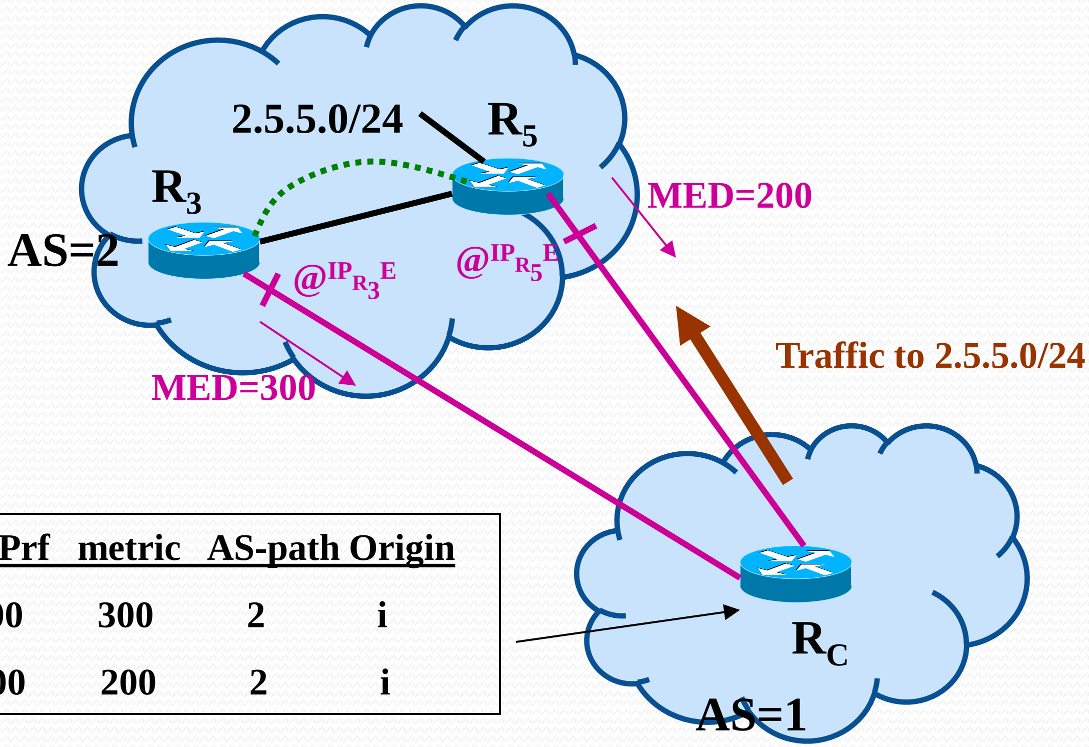
```
If condición then acción
    elseif condición then acción
    else acción
```

- **BGP configuration with CISCO IoS**

```
!!!! Create the route-map in Router R3
R3(conf)# ip access-list 1 permit 2.5.5.0   255.255.255.0
R3(conf)# route-map rr4 permit 10
R3(conf)# match ip address 1
R3(conf)# set  local-pref= 200
R3(conf)# route-map rr4 permit 20
R3(conf)# route-map rrA permit 10
R3(conf)# match ip address 1
R3(conf)# set  local-pref= 300
R3(conf)# route-map rrA permit 20
!!!! Configure BGP in Router R1
R3(conf)# router bgp 2
R3(conf-r)# neighbor IP@R4 remote-as 2
R3(conf-r)# neighbor IP@RA remote-as 1
R3(conf-r)# neighbor IP@R4 route-map rr4 in
R3(conf-r)# neighbor IP@RA route-map rrA in
```

MED, Multi-Exit-Discriminator (optional and non-transitive)

- Also called “metric”, indicates to the neighbors what is the preferred entry link
- The **lowest value** is the preferred value



Net	Next-hop	Loc-Prf	metric	AS-path	Origin
2.5.5.0/24	$@^{IP}{}_{R_3}{}^{E}$	100	300	2	i
>2.5.5.0/24	$@^{IP}{}_{R_5}{}^{E}$	100	200	2	i

Topic 4: Inter-domain Routing

```
!!!! Create the route-map in Router R5
R5(conf)# ip access-list 1 permit 2.5.5.0   255.255.255.0
R5(conf)# route-map rrC permit 10
R5(conf)# match ip address 1
R5(conf)# set med= 300
R5(conf)# route-map rrC permit 20
R5(conf)# router bgp 2
R5(conf-r)# neighbor IP@R3 remote-as 2
R5(conf-r)# neighbor IP@RC remote-as 1
R5(conf-r)# neighbor IP@RC route-map rrC out

!!!! Create the route-map in Router R3
R3(conf)# ip access-list 1 permit 2.5.5.0   255.255.255.0
R3(conf)# route-map rrC permit 10
R3(conf)# match ip address 1
R3(conf)# set med= 200
R3(conf)# route-map rrC permit 20
R3(conf)# router bgp 2
R3(conf-r)# neighbor IP@R5 remote-as 2
R3(conf-r)# neighbor IP@RC remote-as 1
R3(conf-r)# neighbor IP@RA route-map rrC out
```

Community (Optional and transitive)

- Offers the possibility to associate a identifier with a route
- Allows that this route receives the same policy by all AS associated to that policy

- Coded with 32 bits (4 Bytes)
  - Two first Bytes are the AS# that creates the community
  - Last two bytes are defined by the AS

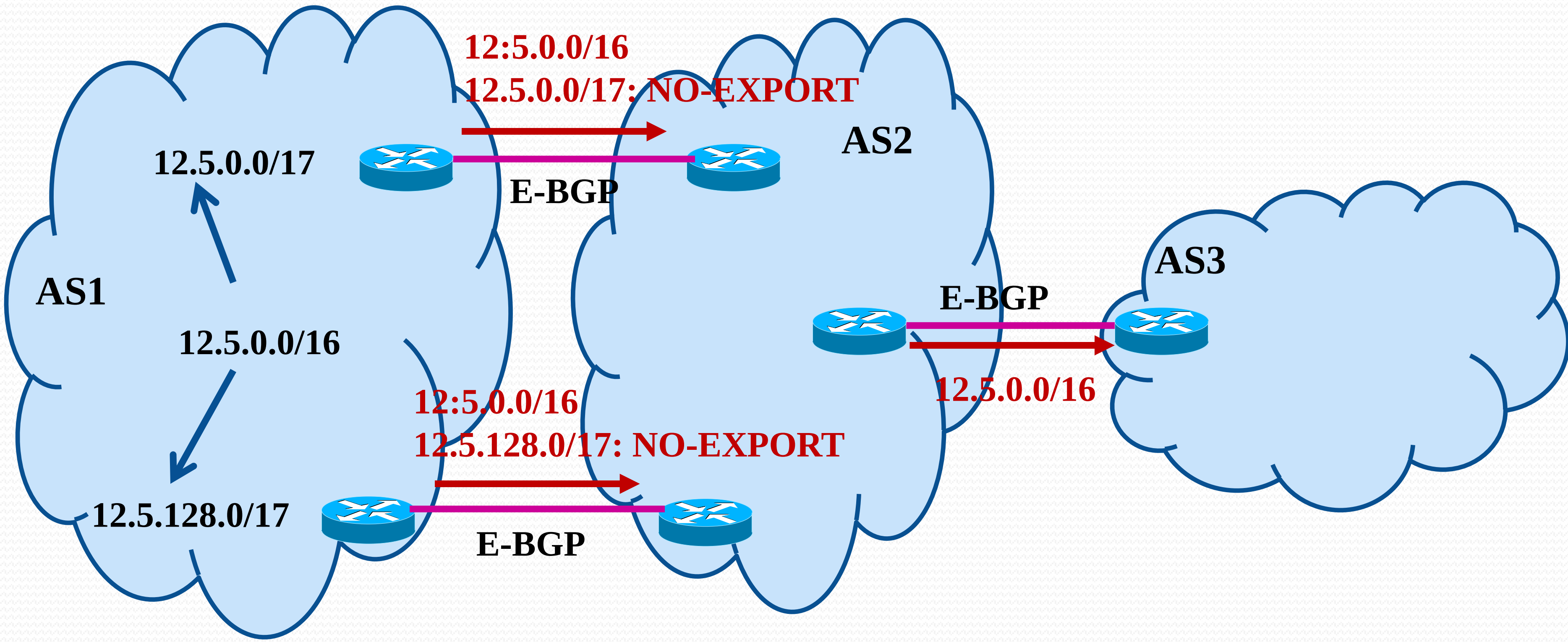
AS:value (decimal)

- Communities reserved: **0x00000000 to 0x0000ffff** (**0:value**) and **0xffff0000 to 0xffffffff** (**65535:value**)

- The rest may be freely used: from **1:0 to 65534:65535**

- ## Standard Communities:
  - Three standard communities (RFC 1997)
    - **NO_EXPORT (0xFFFFFF01):** all the received routes with this attribute SHOULD NOT be advertised out of the AS
    - **NO_ADVERTISE (0xFFFFFF02):** all the received routes with this attribute SHOULD NOT be advertised to other BGP neighbors (inside the same AS)
    - **NO_EXPORT_SUBCONFED (0xFFFFFF03):** all the received routes with this attribute SHOULD NOT be advertised to external BGP routers (from other confederation)

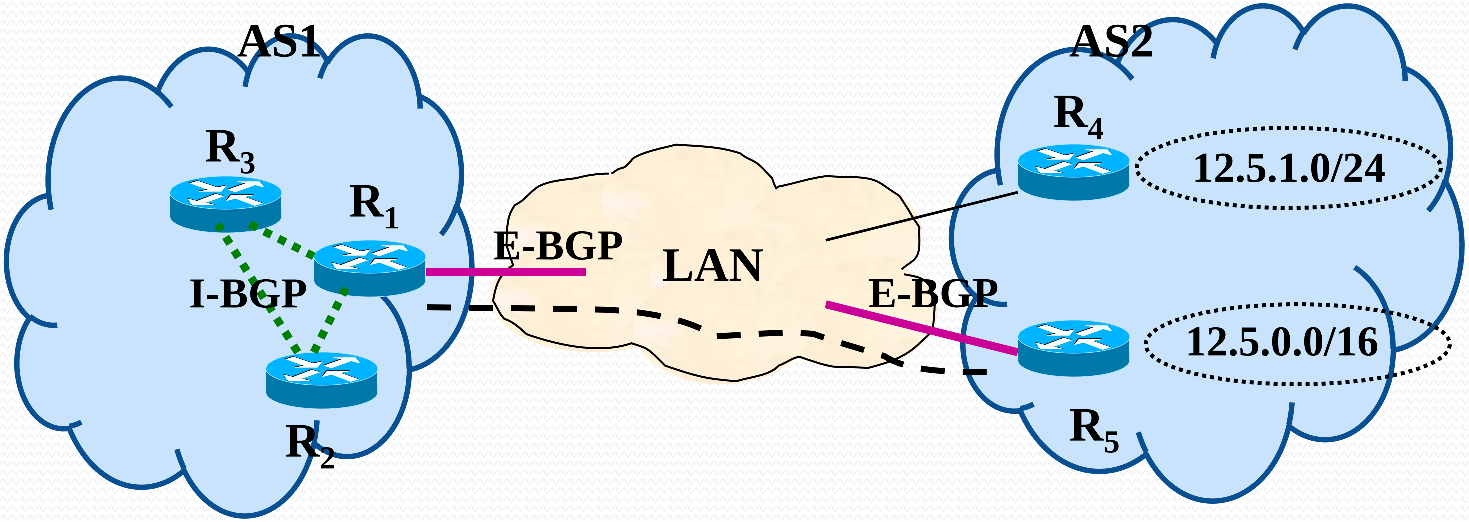
- **Standard Communities:**
  - **Example: NO-EXPORT community**
    - AS1 wants to perform load balancing with AS2 with subnets 12.5.0.0/17 and 12.5.128.0/17
    - AS3 does not need to receive the 2 routes (it is enough /16). Thus, AS1 exports to AS2 the /17 with the NO-EXPORT community and the /16 without community.



Standard Communities:

Example: NO-ADVERTISE community

- AS1 and AS2 have a E-BGP connection between R_1 and R_5, R_4 does not understand BGP
- R_4 uses /24 and R_5 the rest of the /16 block, thus R_5 would like that R_1 send packets destined to /24 directly to R_4 and not to R_5
- R_5 uses NEXT-HOP so R_1 sends packets to R_4, but R_1 should be the only router that understand this policy in AS1, the other AS1 BGP routers does not need to know the policy



BGP routing table (Re-visited)

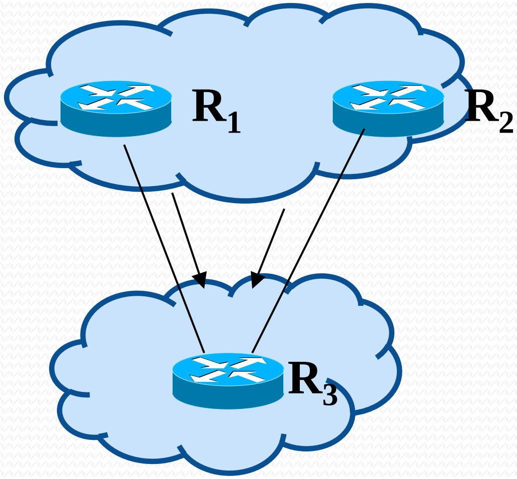
```
R2# show ip bgp
                                               Attributes
                   ________________________________________________________________
Network    Next Hop          Metric   LocPrf   AS-Path                        Origin
*  4.0.0.0 206.157.77.11       75      100      1673 1                          i
*>         12.127.0.249         0      200      7018 1                          i
*          204.70.4.89          0      100      3561 1                          i
*          204.42.253.253       0      200      267 1225 1239 1                 i
*          205.158.2.126        0      200      2828 4908 3561 1                i

…………………………..
*  6.0.0.0 206.157.77.11      105      100      1673 1239 568 721 1455          i
*          12.127.0.249         0      100      7018 7170 1455                  i
*>         198.32.8.252         0      100      11537 7170 1455                 i
*          204.70.4.89          0      100      3561 568 721 1455               i
```

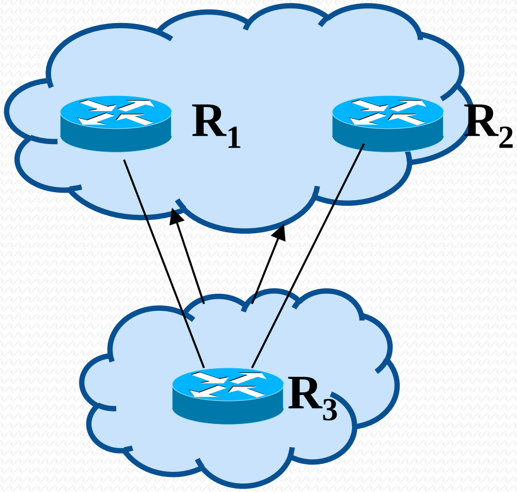
BGP Routing Table: Decision Process

- Depends on implementation. E.g.; in a CISCO router
  1. For internal paths, synchronization ON, otherwise→ reject the route
  2. If the "next-hop" is not reachable → reject the route
  3. Prefer route with maximum "weight" (CISCO attribute)
  4. Multiple routes with the same "weight", choose the highest Loc-Prf
  5. Multiple routes with the same Loc-Pref, choose the lowest AS-path
  6. Multiple routes with the same AS-path, choose the lowest "origin" (IGP<EGP<Incomplete)
  7. Multiple routes with the same "origin", choose the lowest MED
  8. ....
  9. Choose the route of the BGP router with lowest Router-ID and if there is more than one route from the same router, choose that one with lowest interface IP@

- **Multi-homing**:
  - **Single-homed AS**: a customer only has one connection with other ISP
  - **Multi-homed AS:** a customer has more than one connection with one or more ISP
    - Multi-homing increases access reliability since a link fails the customer has a back-up line
    - **Load balancing:** balance traffic among links allowing **Inbound traffic control** and **Outbound traffic control**

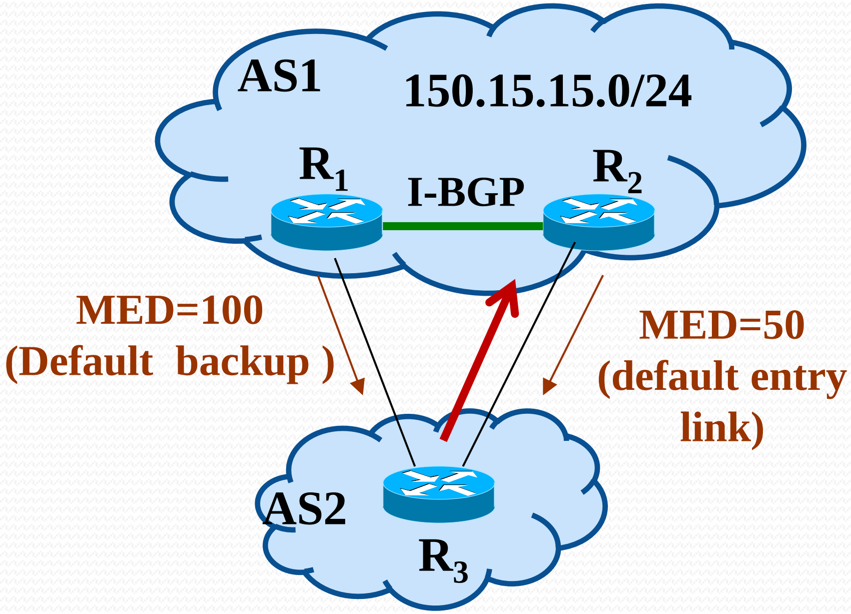


Inbound traffic control: I choose which is the entry link

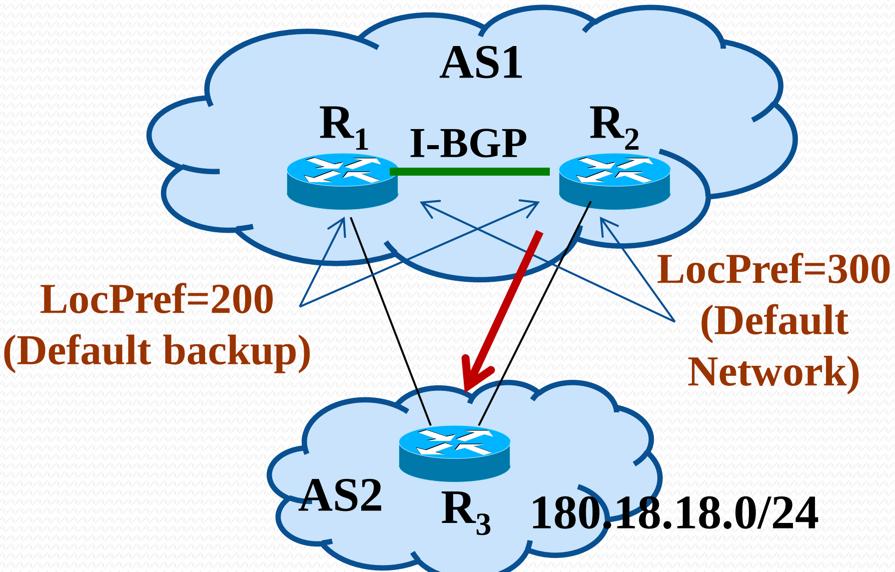


Outbound traffic control: I choose the output link

- **Some examples: Multi-homed AS to the same provider**
  - AS1 use BGP and export routes with different MED attribute in order to force AS2 to use the entry link (R_2 over R_1)
  - AS1 use BGP and import routes related to Local-Pref attribute in order to select the output link (R_2 over R_1)



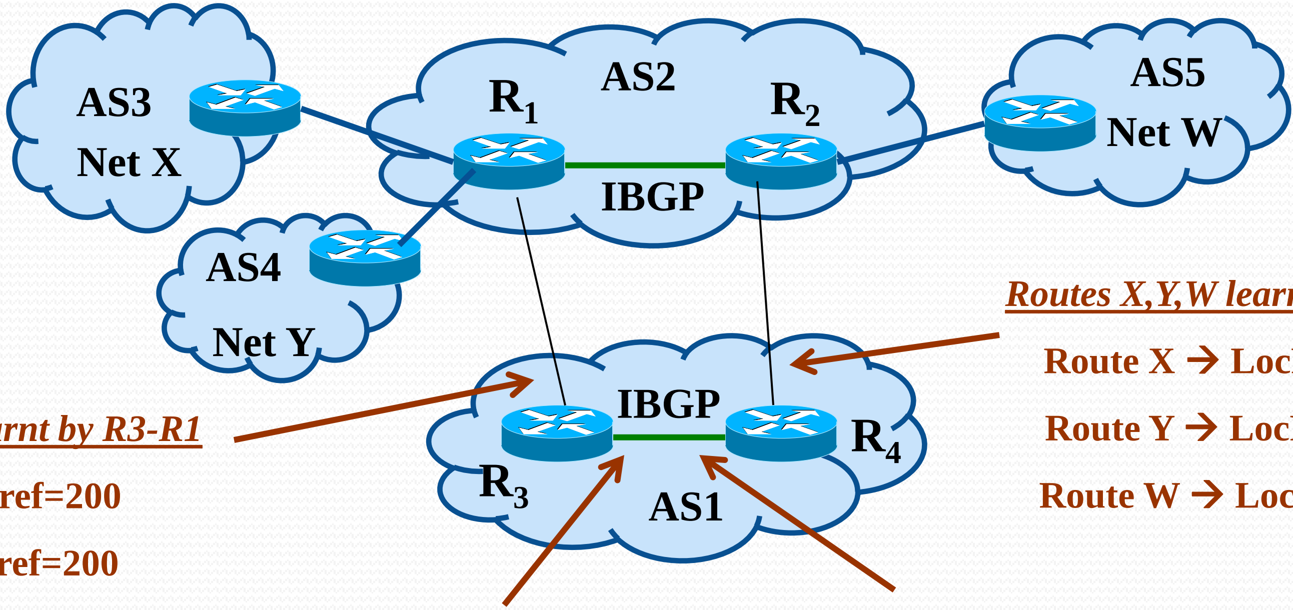
Inbound traffic control without load balancing (entry link is R2-R3)



Outbound traffic without load balance (out link is R2-R3)

Some examples: Load balancing: Outbound traffic control

- Use LocPref to choose outer links: Traffic towards customers AS3 and AS4 get out using R3 and traffic towards C4 leave using R4



Routes X,Y, W learnt by R3-R1

Route X → LocPref=200

Route Y → LocPref=200

Route W → LocPref=100

Routes X,Y,W learnt by R4-R2

Route X → LocPref=100

Route Y → LocPref=100

Route W → LocPref=200

Routes X,Y, W learnt by R3-R4

Route X → LocPref=100

Route Y → LocPref=100

Route W → LocPref=200

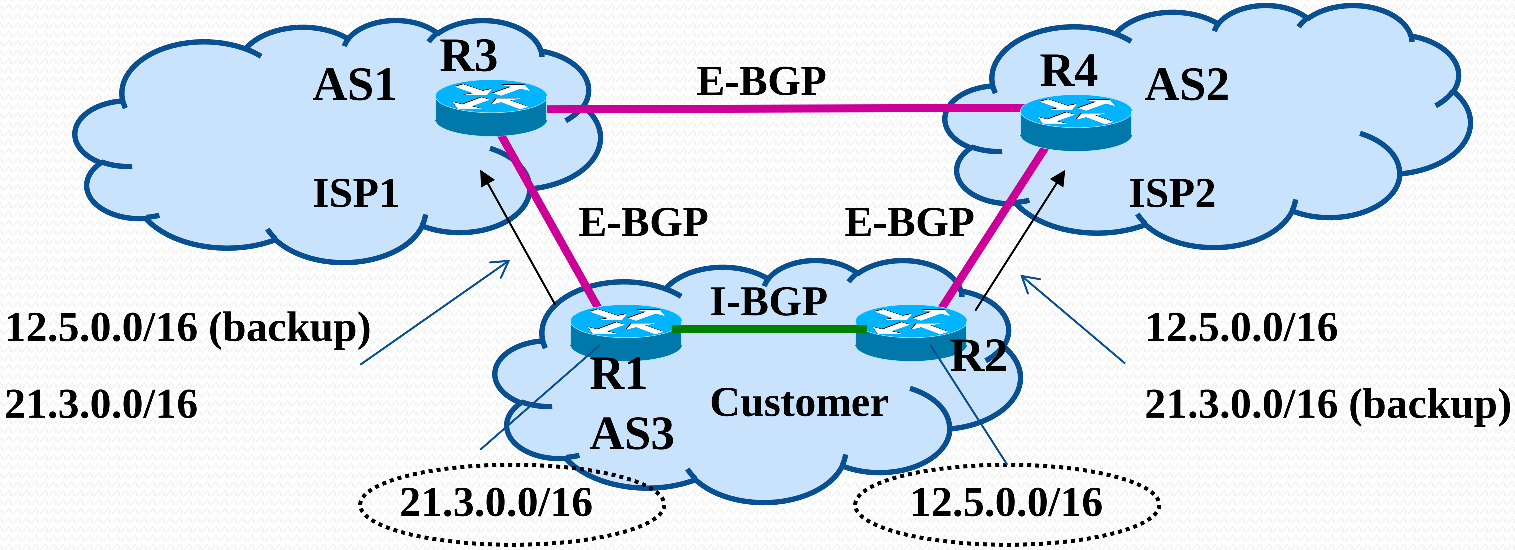
Routes X,Y, W learnt by R4-R3

Route X → LocPref=200

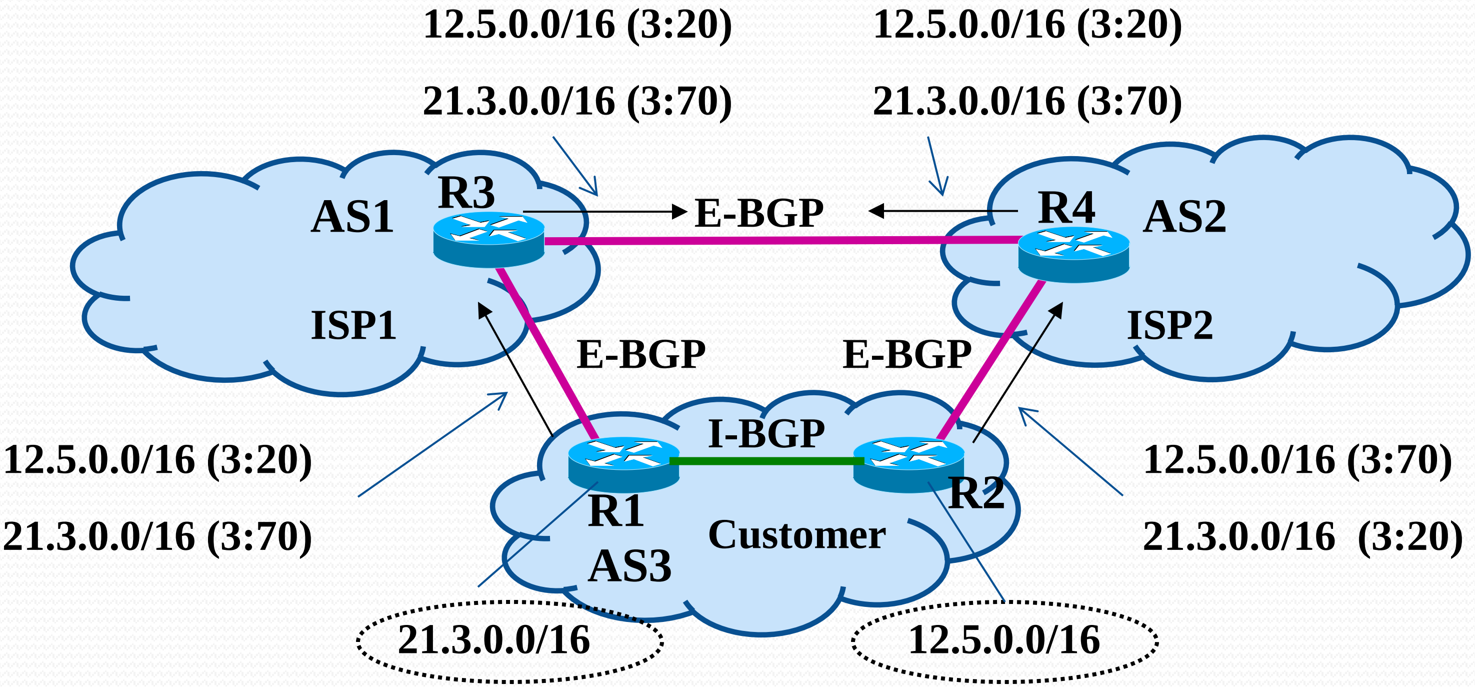
Route Y → LocPref=200

Route W → LocPref=100

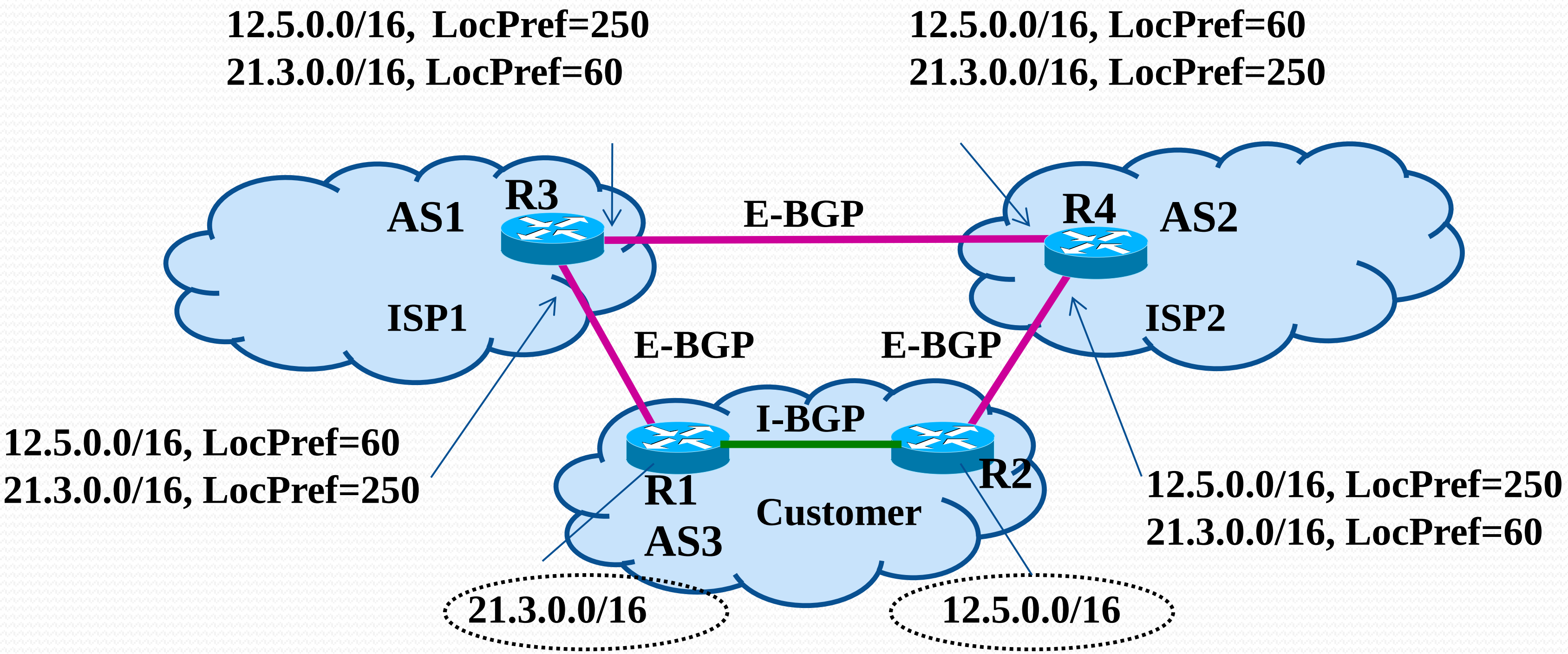
- **Communities (revisited): automatic back-up routes in multi-homing**
  - AS3 wants multi-homing with ISP1 and ISP2. Traffic towards network 12.5.0.0/16 enters via ISP2 and traffic towards network 21.3.0.0/16 enters via ISP1, but both connections act as backup with respect the other network (they can not use MED since it is not transitive)



- **Communities: automatic back-up routes in multi-homing**
  - AS1 and AS2 react to community 3:20 activating LocalPref=60
  - AS1 and AS2 react to community 3:70 activating LocalPref=250

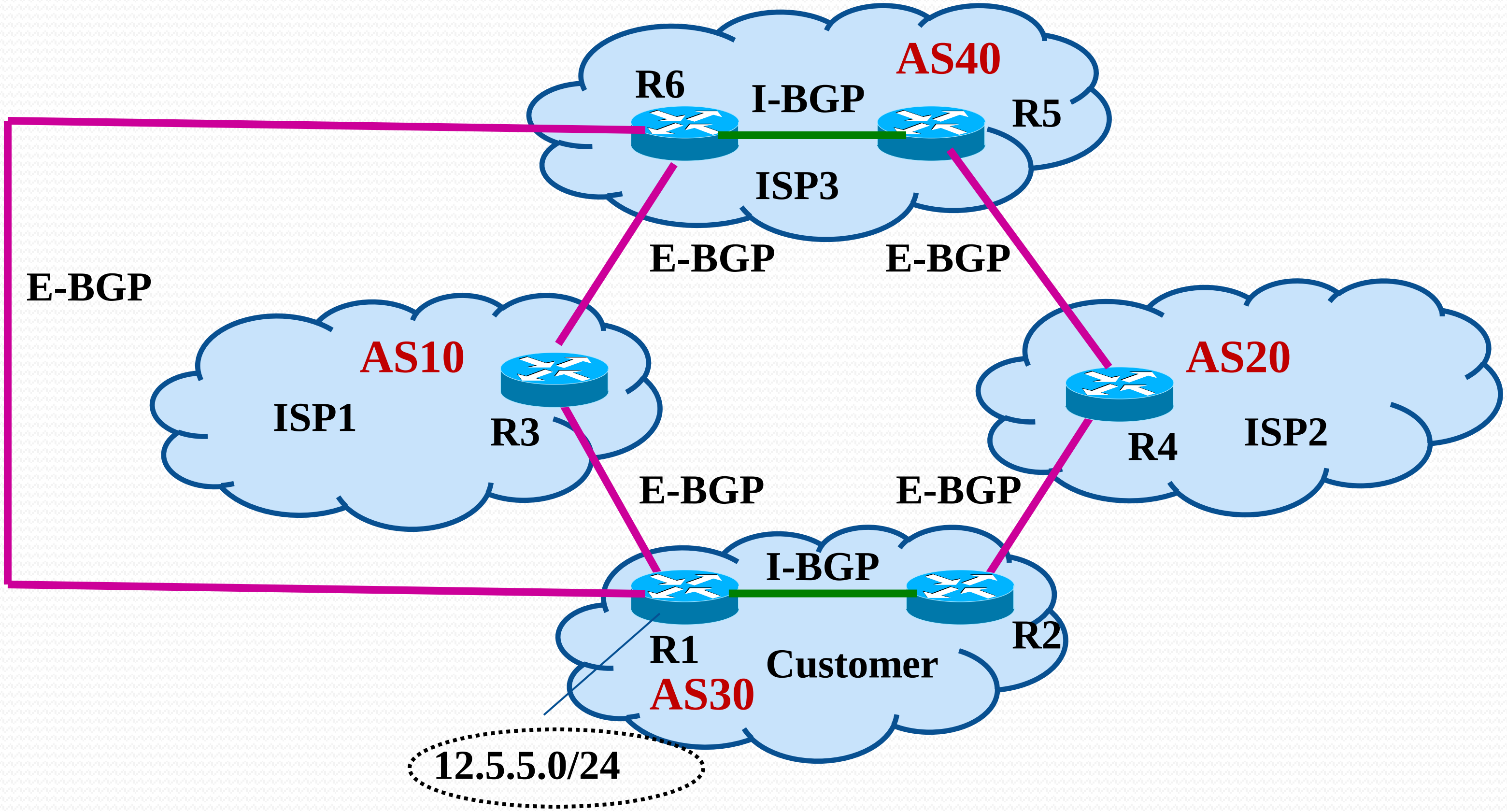


- **Communities: automatic back-up routes in multi-homing**
  - AS1 and AS2 react to community 3:20 activating LocalPref=60
  - AS1 and AS2 react to community 3:70 activating LocalPref=250



BGP Communities CISCO IoS

- Access to network 12.5.5.0/24 is done via i) R6-R1, if not possible, via ii) R6-R3-R1, elsewhere, iii) via R5-R4-R2



Topic 4: Inter-domain Routing

R1# .

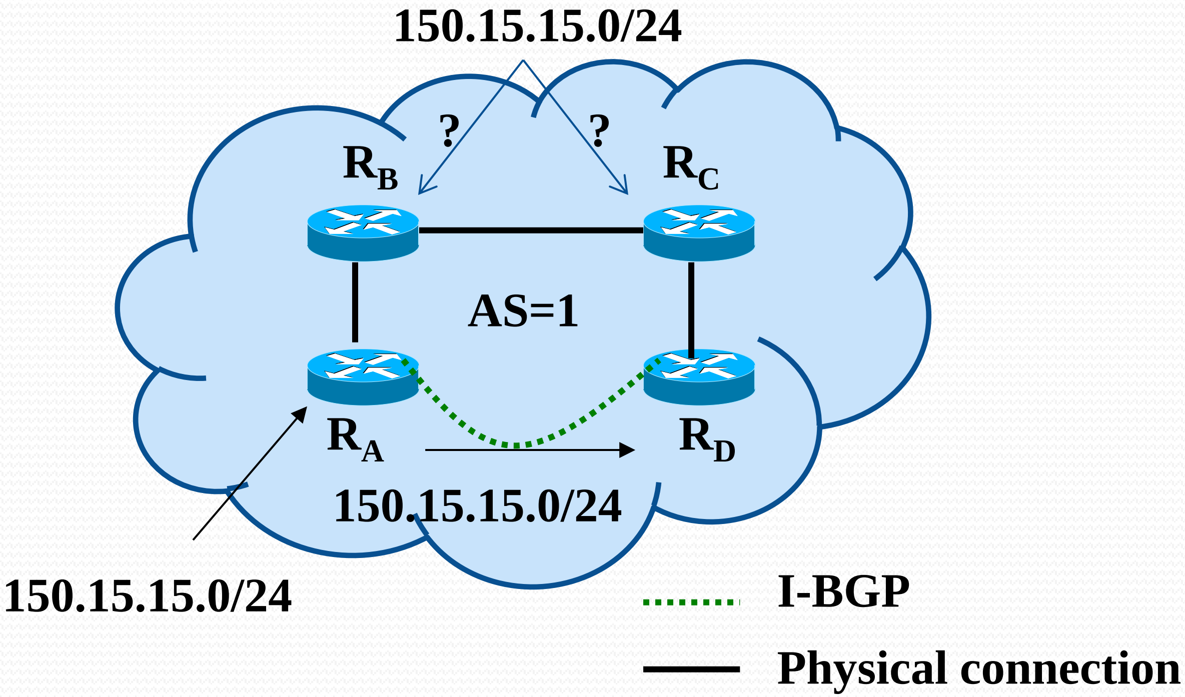
```
router bgp 30
neighbor 1.1.1.1 remote-as 40
neighbor 2.2.2.2  route-as 10
neighbor 3.3.3.3  route-as 30
network  12.5.5.0/24
neighbor 1.1.1.1 send-community
neighbor 2.2.2.2 send-community
neighbor 1.1.1.1  route-map Peer-R6 out
neighbor 2.2.2.2  route-map Peer-R3 out
!
route-map Peer-R6 permit 10
match ip address 1
set community 30:20
route-map Peer-R6 permit 20
!
route-map Peer-R3 permit 10
match ip address 1
set community 30:10
route-map Peer-R3 permit 20
!
ip access-list 1 permit  12.5.5.0  0.0.0.255
```

R6# .

```
router bgp 40
neighbor 1.1.1.2 remote-as 30
neighbor 5.5.5.2  route-as 10
neighbor 6.6.6.2  route-as 40
neighbor 1.1.1.2  route-map Peer-R1 in
neighbor 5.5.5.2  route-map Peer-R3 in
neighbor 6.6.6.2  route-map Peer-R5 in
!
route-map Peer-R5 permit 10
match ip address 1
set Local-Preference=100
route-map Peer-R5 permit 20
!
route-map Peer-R3 permit 10
match community 1
set Local-Preference=200
route-map Peer-R3 permit 20
!
route-map Peer-R1 permit 10
match community 2
set Local-Preference=300
route-map Peer-R1 permit 20
!
ip access-list 1 permit  12.5.5.0  0.0.0.2552
ip community-list 1 permit  30:10
ip community-list 2 permit  30:20
```

BGP synchronization

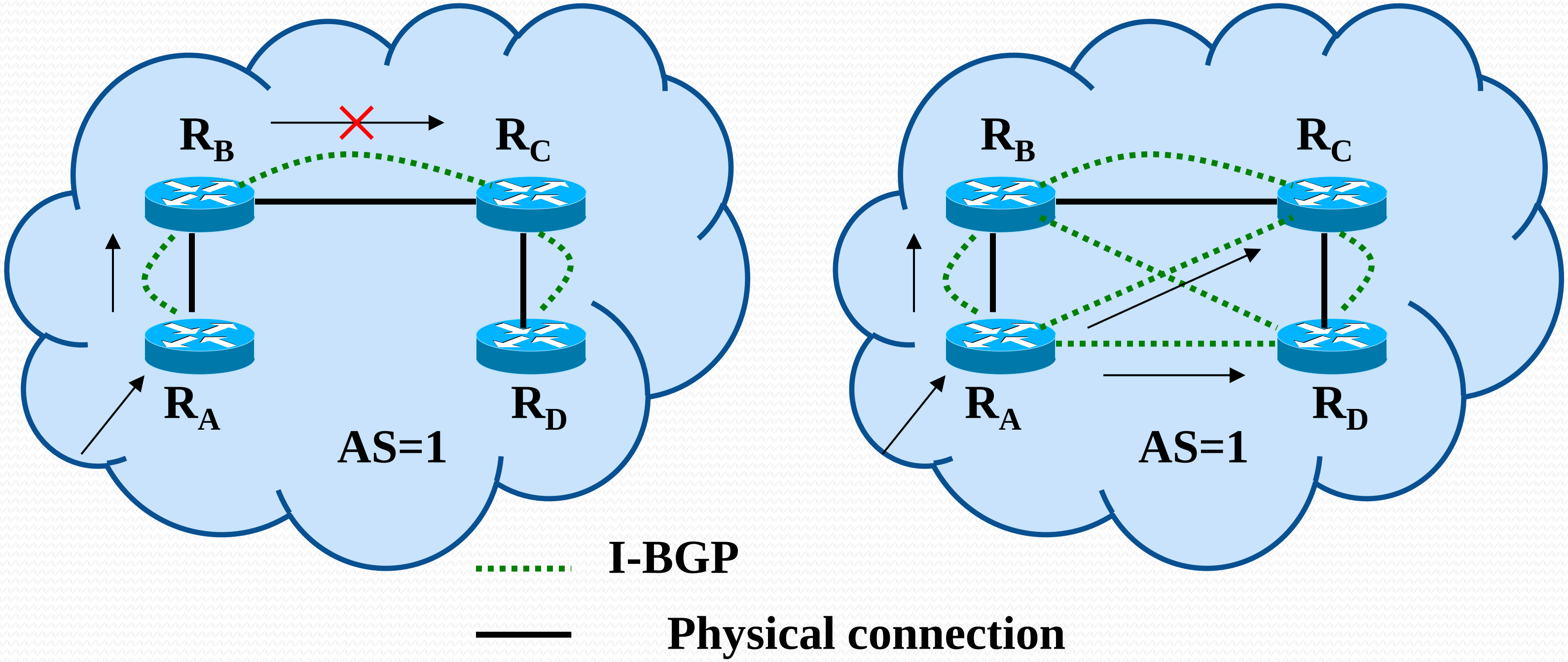
- Routes received by R_A are sent via I-BGP to R_D. However, routers R_B and R_C are not aware of route 150.15.15.0/24
- Then, routers R_B and R_C are not synchronized (they do not know how to get to route 150.15.15.0/24) → redistribute BGP in IGP protocols or create a full meshed I-BGP network



BGP scalability

- Split-horizon:
  - A route learnt by I-BGP is not propagated to I-BGP neighbor routers
  - A I-BGP **“full-mesh”** network is needed → if N routers

N (N-1)/2 I-BGP connections

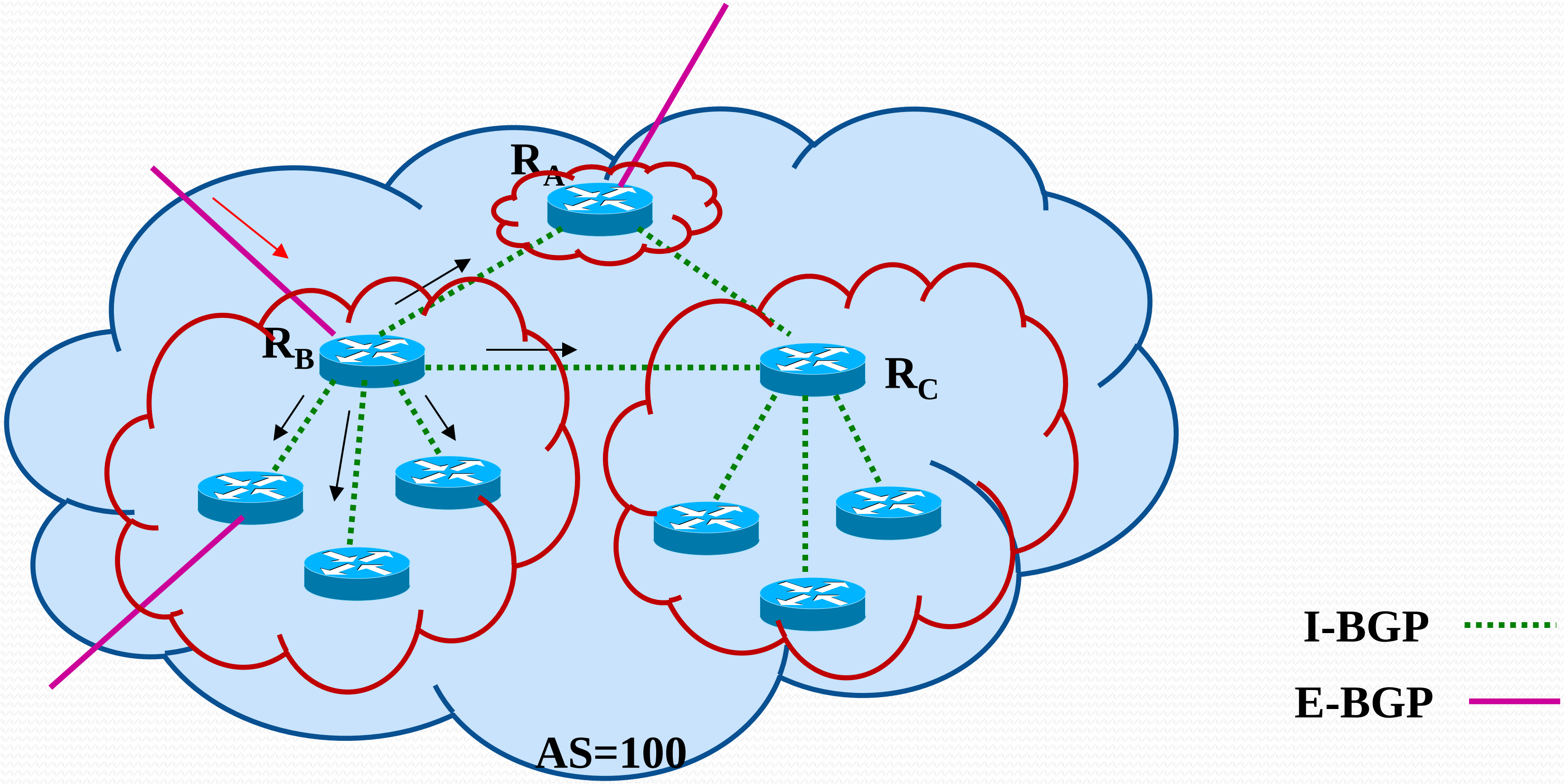


BGP scalability: Route Reflectors and Confederations

- **Route Reflectors:** split-horizon rule is modified in order the route reflector may propagate routes learnt by I-BGP connections under certain conditions reducing the number of I-BGP sessions in the AS

- **Clusters** are used to define the network
  - A **route reflector** acts as **cluster-head**
  - Each route reflector maintain I-BGP sessions with its **customers** (routers that belong to the cluster)
  - The route reflectors should form a mesh network between them, but customers do not need to form a mesh

- **Route Reflectors**
  - Each route reflector maintain ONE I-BGP sessions with each of its **customers** (routers that belong to the cluster)
  - The route reflectors should form a mesh network between them, but customers do not need to form a mesh



Route Reflectors

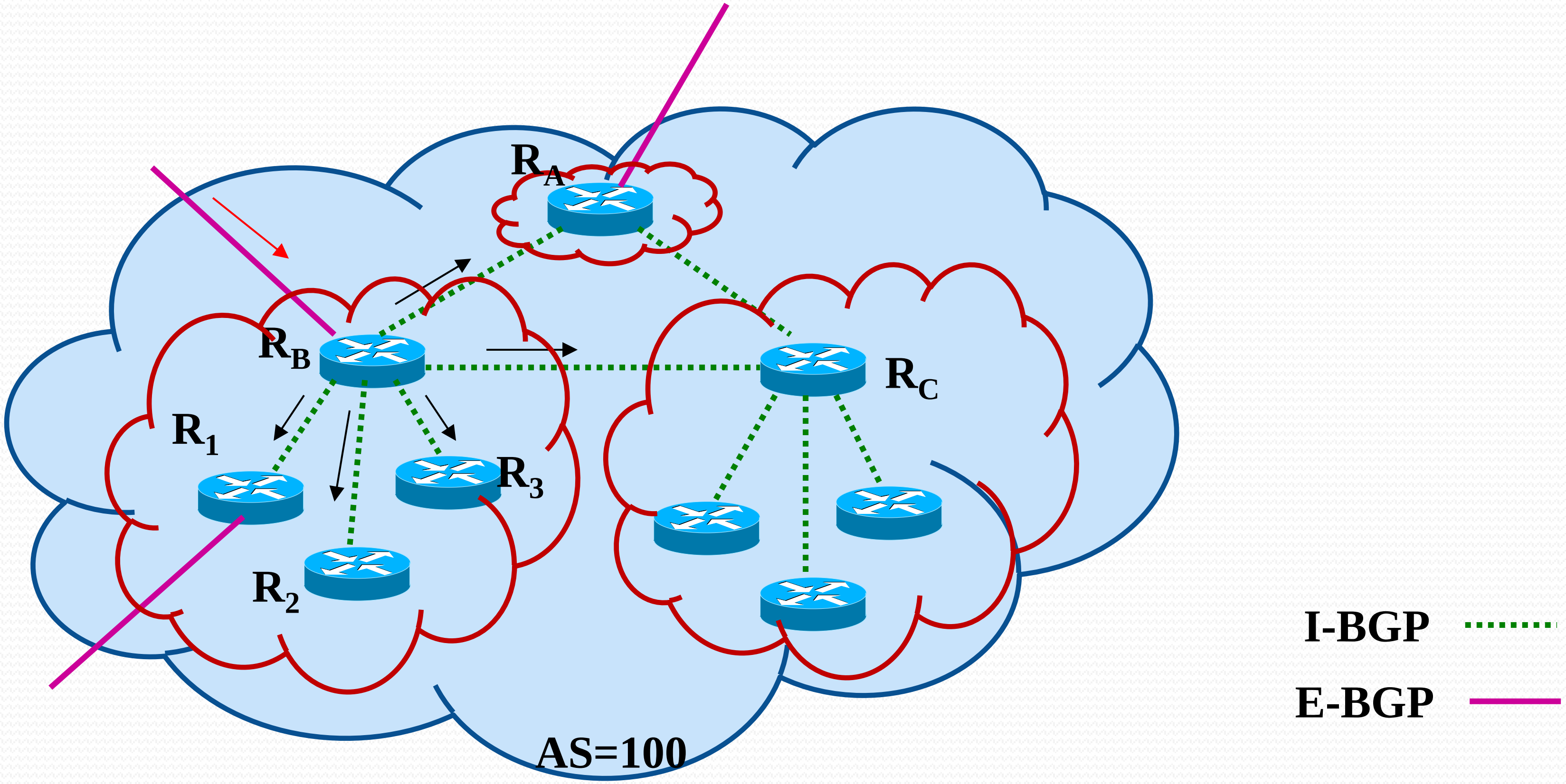
En concreto el router RR sigue estas reglas al recibir un mensaje BGP:

- Si el mensaje BGP proviene de un vecino no cliente (por ejemplo otro RR), entonces el RR la refleja a todos sus clientes dentro de su cluster.

- Si el mensaje BGP proviene de un cliente, el RR la refleja a todos los vecinos clientes y no clientes.

- Si el mensaje BGP se aprende de un vecino eBGP, éste se envía a todos los vecinos clientes y no clientes.

- Route Reflectors CISCO IoS



Route Reflectors CISCO IoS

```
!!!! Create RR in RB
RB(conf)# router bgp 100
RB(conf-r)# neighbor IP@RA remote-as 100            ← neighboring with RR A
RB(conf-r)# neighbor IP@RC remote-as 100            ← neighboring with RR C
!
RB(conf-r)# neighbor IP@R1 remote-as 100
RB(conf-r)# neighbor IP@R1 route-reflector-client     ← customer of RB
!
RB(conf-r)# neighbor IP@R2 remote-as 100
RB(conf-r)# neighbor IP@R2 route-reflector-client    ← customer of RB
!
RB(conf-r)# neighbor IP@R3 remote-as 100
RB(conf-r)# neighbor IP@R3 route-reflector-client    ← customer of RB
```

Route Reflectors

- BGP session savings:
  - N routers in the domain
  - N_R Route Reflectors and NR_i ($i=1,2,\ldots, N_R$) customers per Route Reflector:

$$N = \sum_{i=1}^{N_R} NR_i + N_R$$

  - Then, the number of I-BGP sessions that have to be configured is:

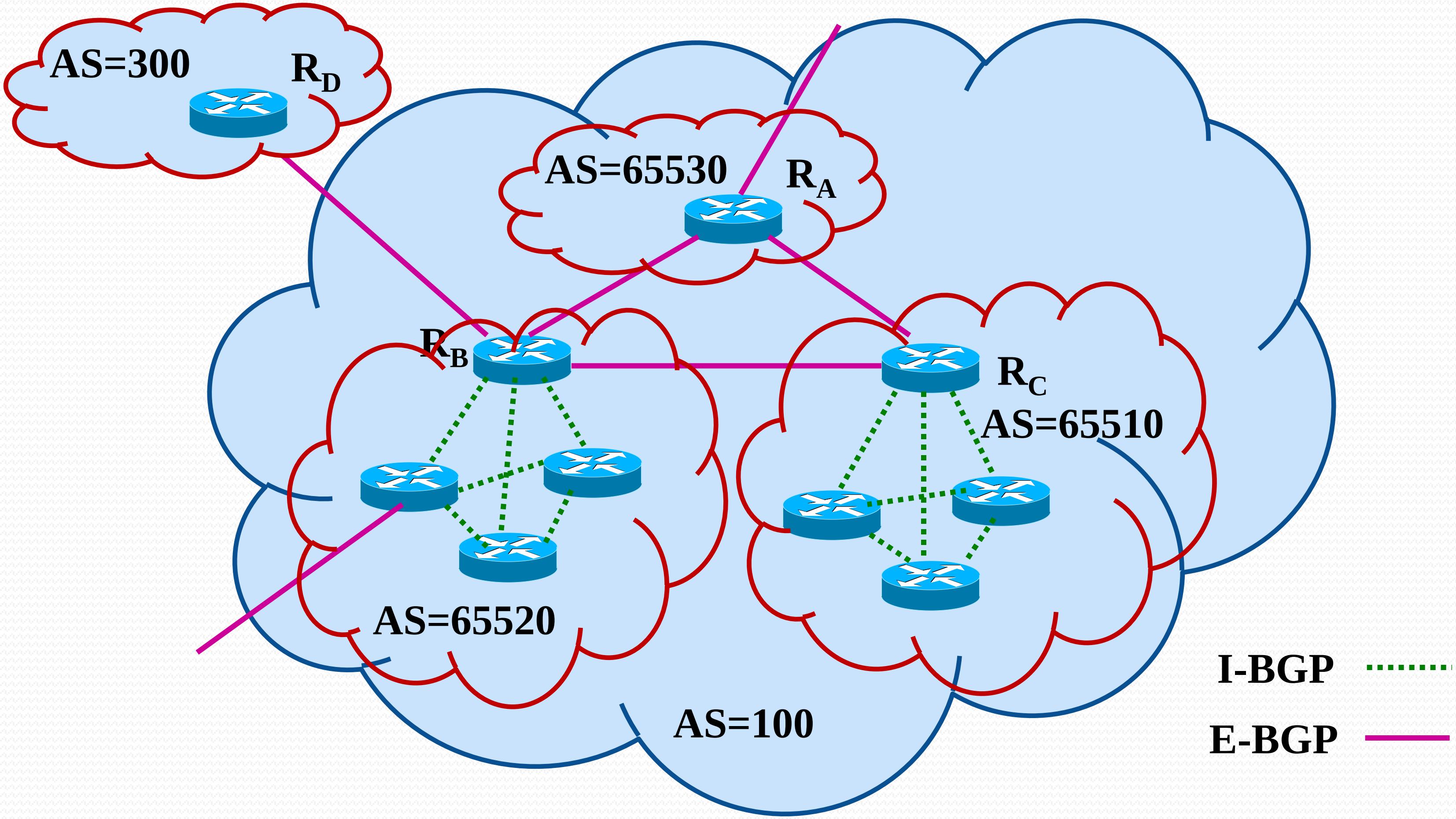
$$I-BGP = \sum_{i=1}^{N_R} NR_i + \frac{N_R(N_R-1)}{2}$$

  - For example, in the previous figure: N=9, N_R=3, NR_1=0, NR_2=3, NR_3=3
    - Without Route Reflectors: I-BGP=N*(N-1)/2= 9*8/2= **36 I-BGP sessions**
    - With Route Reflectors: I-BGP=0+3+3+(3*2/2)= **9 I-BGP sessions**

→ a saving of (36-9)/36=75% of I-BGP sessions

- **BGP scalability**
  - **Confederation** is another solution to reduce the number of I-BGP sessions
    - Create mini-AS using private AS numbers inside the AS
    - Each mini-AS should form a mesh network
    - Each mini-AS needs E-BGP sessions with other mini-AS
    - From the external point of view they are seen as a unique public AS

- **BGP Confederations**
  - Each private AS is full-meshed
  - Private AS are connected via E-BGP



Confederations

- BGP session savings:
  - N routers in the domain
  - N_C confederations and NC_i (i=1,2,…, N_C) rotuers per confederation:

$$N = \sum_{i=1}^{N_C} NC_i$$

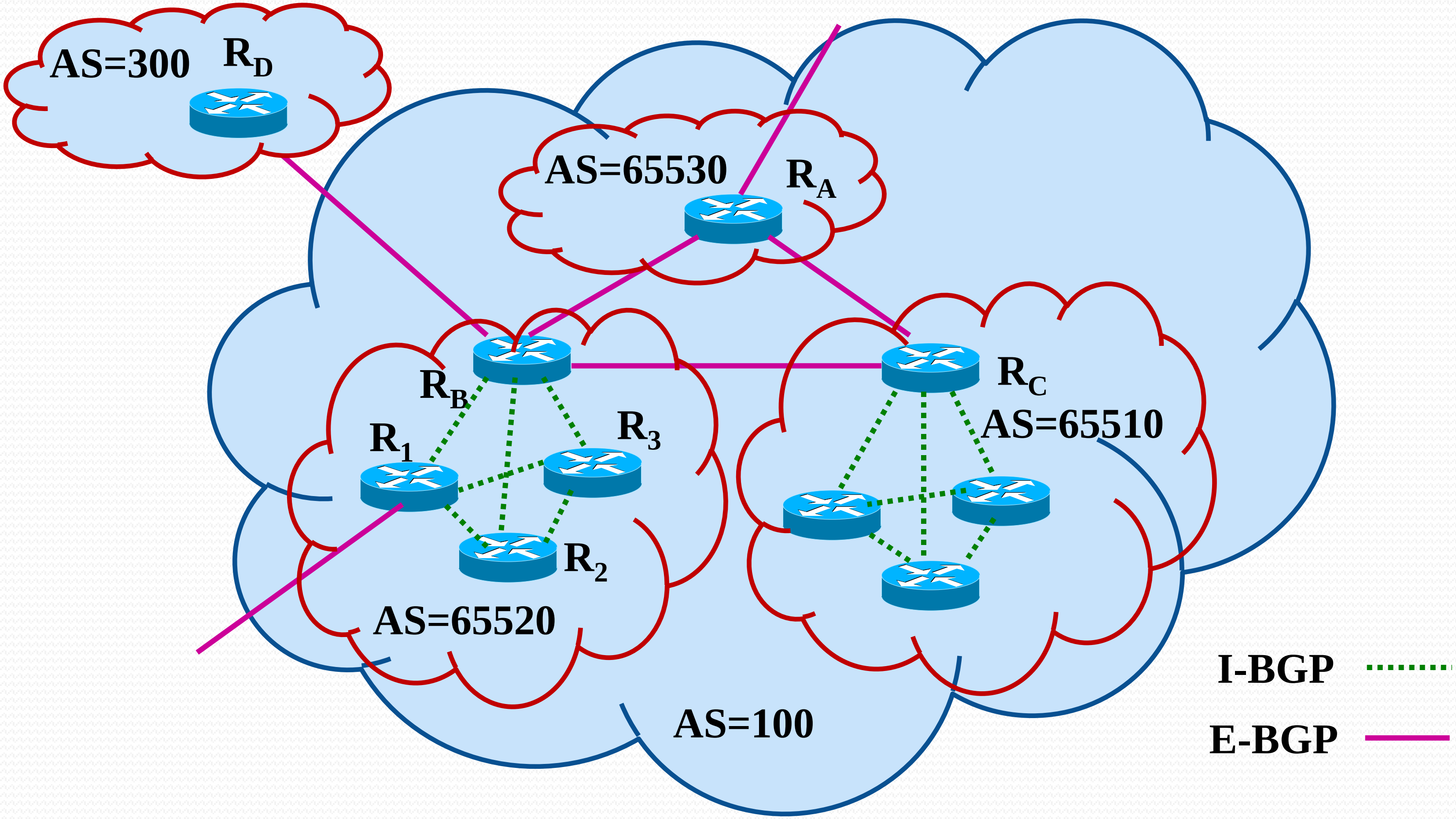
  - Then, the number of BGP sessions that have to be configured is:

$$BGP = I\text{-}BGP + E\text{-}BGP = \sum_{i=1}^{N_C} \frac{NC_i * (NC_i - 1)}{2} + \min(E - BGP)$$

  - For example, in the previous figure: N=9, N_C=3, NR_1=1, NR_2=4, NR_3=4
    - Without confederations: I-BGP=N*(N-1)/2= 9*8/2= **36 I-BGP sessions**
    - With confederations: I-BGP=0+4*3/2+4*3/2= **12 I-BGP sessions** and **min(E-BGP)=2 E-BGP sessions → BGP=12+2=14 BGP sessions**

→ a saving of (36-12)/36=66.6% of BGP sessions

Confederations in CISCO IoS



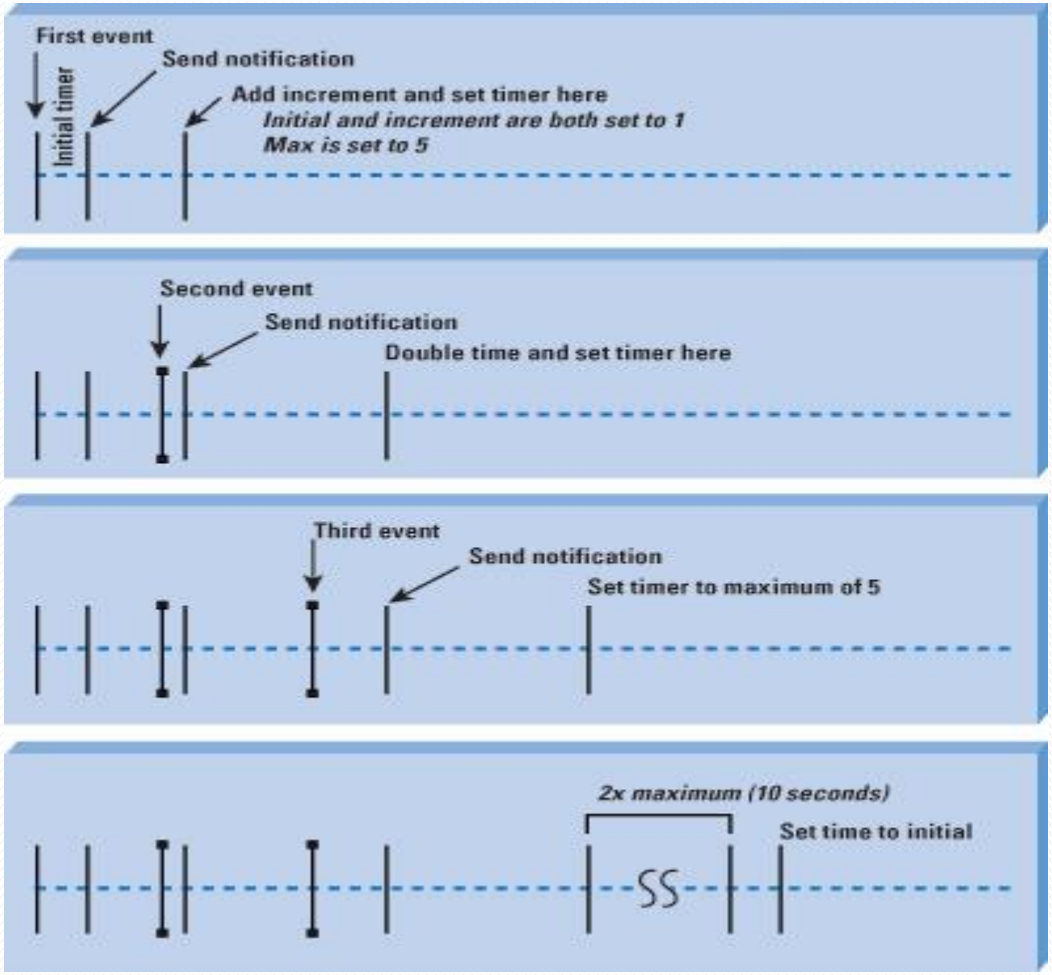
Route Reflectors CISCO IoS

```
!!!! Create Confederation in RB
!
RB(conf)# router bgp 65520
RB(conf-r)# bgp confederation identifier 100        ← defines the public AS#
RB(conf-r)# bgp confederation peers 65510           ← defines the private AS#
RB(conf-r)# bgp confederation peers 65530           ← defines the private AS#
!
RB(conf-r)# neighbor IP@R1 remote-as 65520          ← same AS confederation
RB(conf-r)# neighbor IP@R2 remote-as 65520          ← same AS confederation
RB(conf-r)# neighbor IP@R3 remote-as 65520          ← same AS confederation
RB(conf-r)# neighbor IP@RA remote-as 65530          ← other AS confederation
RB(conf-r)# neighbor IP@RC remote-as 65510          ← other AS confederation
```

- **BGP convergence:**
  - **Flapping:** a link changes constantly from one state to other (up and down), provoking updating of messages and thus low network convergence, loops and network failures, “**meltdown**”
  - Solutions:
    - L2 has to wait before announcing that the link has failed L3 (“debouncing the interface”)
    - Wait (L3) before sending routing messages
    - Wait before eliminating routes in the routing table
    - Wait before react to topological changes

- **BGP convergence:**
- Reduce (**slow-down techniques**) the frequency at which update routing messages are sent to other BGP routers
  - The more changes the more "slow" frequency
  - Speed up if events occur from time to time
  - Slow-down techniques have as objective to minimize instabilities (meltdown) produced by "route flapping"
    - **Exponential back-off**: slow down message reporting
    - **Dampening**: do not report an event if this one occurs frequently

- BGP convergence: Exponential Back-off



Tout=1 s

→ Tout=2 s

→ Tout=4 s

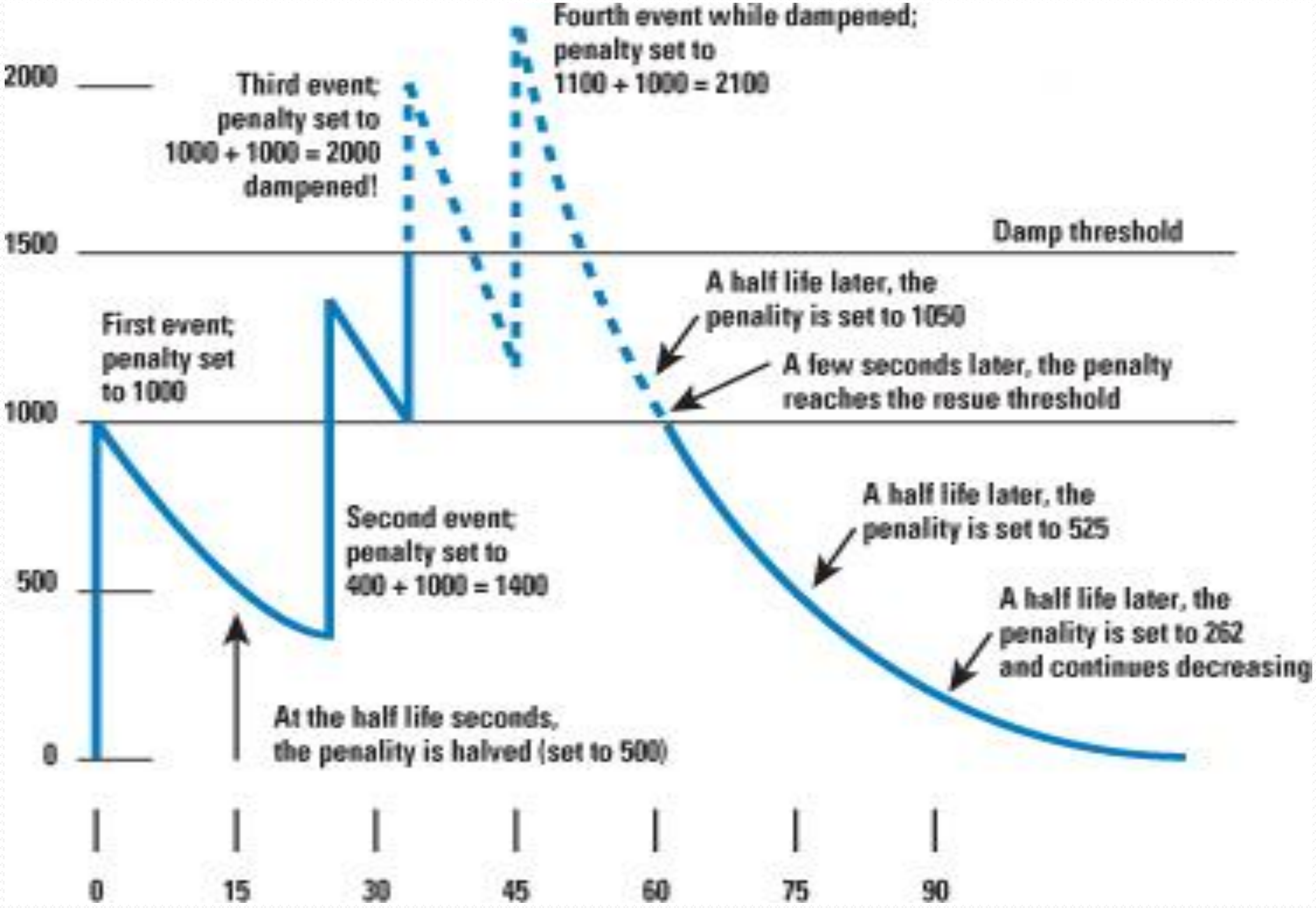
→ Tout=5 s

Reset to Tout=1 s
if 10 seconds
without events

BGP convergence: Dampening

- Each time an event occurs, a counter is incremented by a **penalty value**
- After a time without occurring the event, the counter is decremented
- If the counter reaches the **“damp threshold“** the event enters in the **“DAMPENED” state**
  - The link and route pass to the down state
- If the counter reaches the “reuse threshold”
  - The link and route pass to the up state

- **BGP convergence: Dampening**



BGP convergence: Dampening

- Example:
  - Penalty :1000
  - Suppress Limit: 2000
  - Reuse Limit: 750
  - Half-Life: 15 Minutes
  - Maximum Suppress-Limit: 60 Minutes

Once a route has been dampened, the penalty must be reduced to a value lower than the reuse limit in order to be advertised once again. The half-life timer does this automatically. After a penalty has been assigned and the prefix has become stable again, the half-life timer starts. When the half-life time has been reached, the penalty will be reduced by half (it decreases exponentially every fifteen minutes).

For example, if the penalty was 3000, then fifteen minutes later, the half-life will have reduced the penalty to 1500. Another 15minutes will reduce the penalty to 750, and so on. Once the penalty goes below half of the re-use limit (375 in this case), the penalty is completely removed.

The maximum suppress-limit is used to ensure the prefix doesn't get dampened indefinitely. Using the default values above, a prefix would become un-suppressed after 60 minutes regardless of penalty.

BGP convergence: Dampening

There is also a hidden value called the max penalty; which gets calculated behind the scenes. It is used to ensure you haven't entered dampening values that aren't going to work. Lets look at an example:

- Penalty :1000
- Suppress Limit: 10000
- Reuse Limit: 1500
- Half-Life: 30 Minutes
- Maximum Suppress-Limit: 60 Minutes

To work out the maximum penalty that can possibly be assigned to a prefix you can use the formula below:

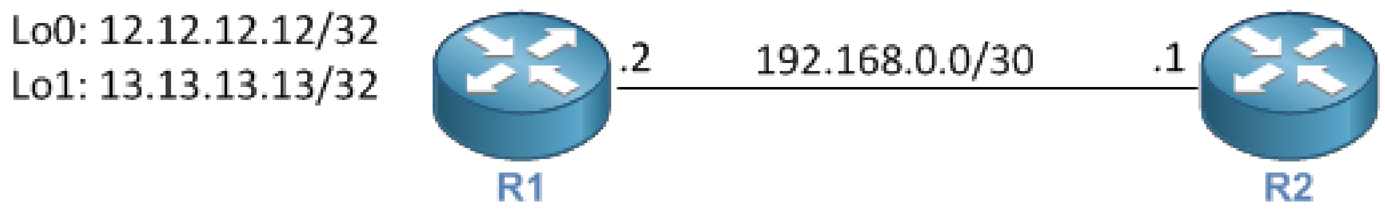
$$\textbf{max-penalty} = \textbf{reuse-limit} * 2^{\textbf{(max-suppress-time/half-life)}}$$

Take the values above, and: max-penalty = $1500*2^{(60/30)}$ = **6000**

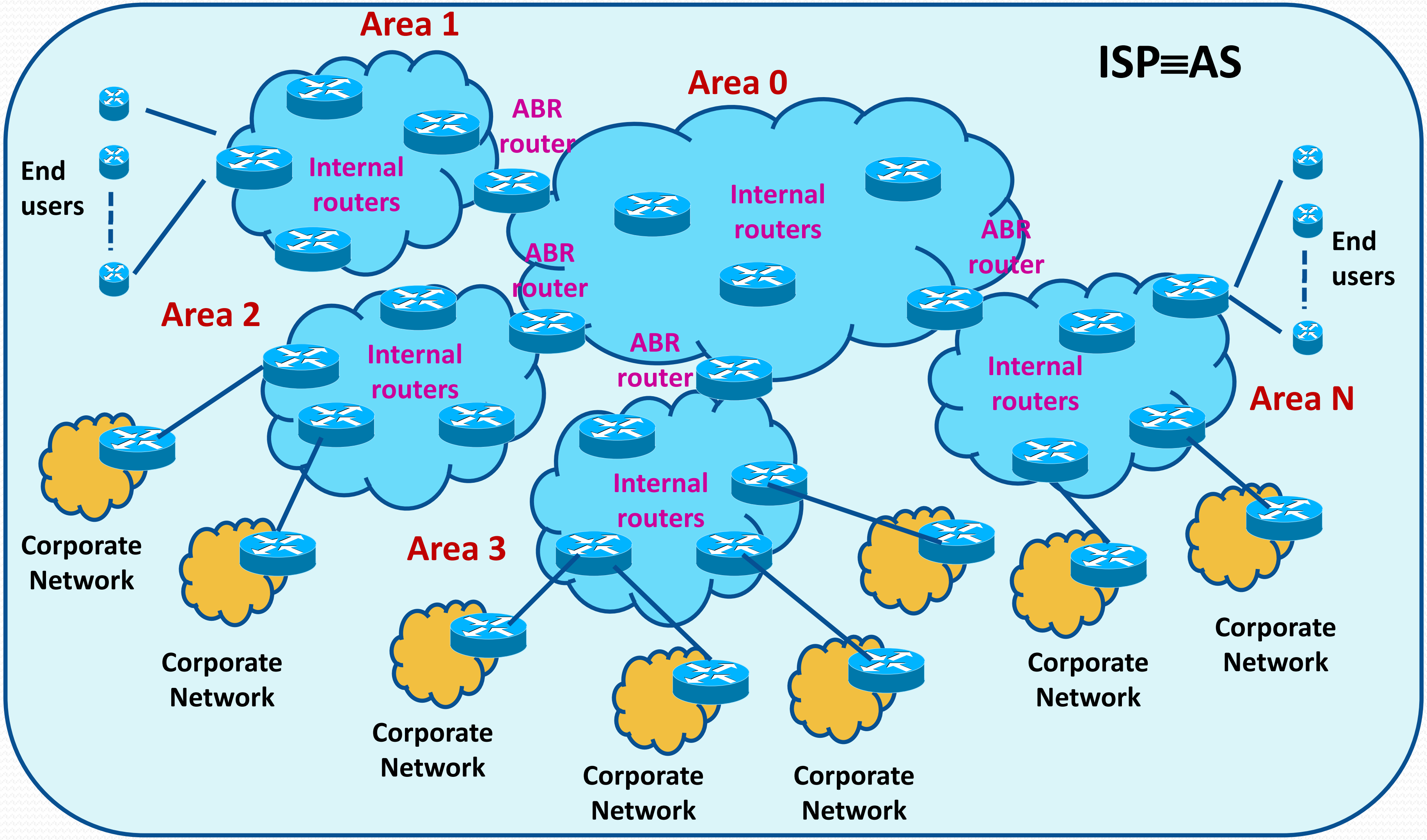
So a route flap causes a penalty of 1000. When the penalty reaches 10,000, the prefix gets dampened. However, the maximum penalty that can be assigned is 6000. This means we will never incur a penalty significant enough to dampen the prefix. When deploying bgp dampening, you should run your values through the formula above to ensure you can actually dampen prefixes.

BGP convergence: Dampening

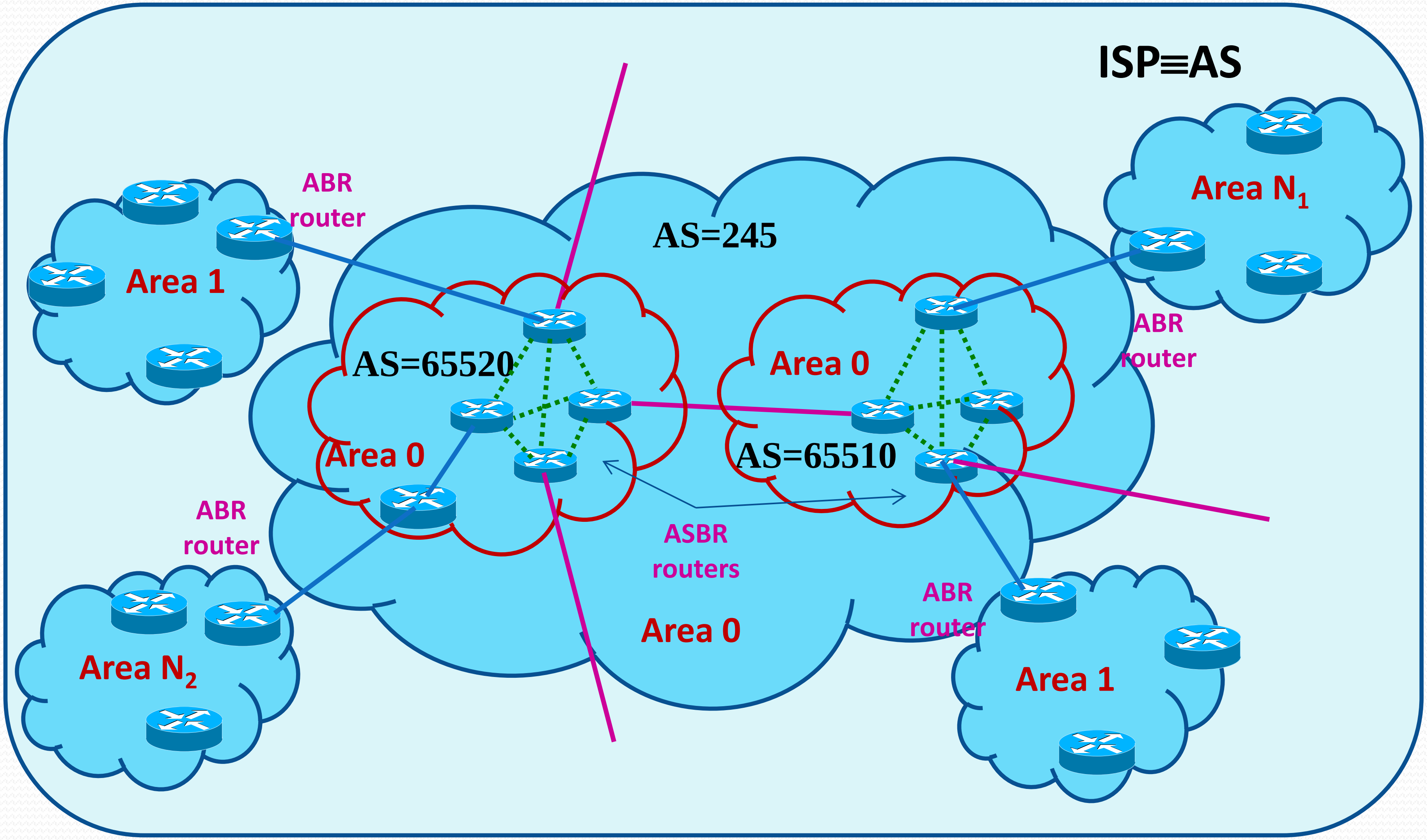
```
!!!! Activating dampening in route 12.12.12.12/32 at router R1
R1(conf)# ip access-list 1 permit 12.12.12.12   255.255.255.255
R1(conf)# route-map damp-R1 permit 10
R1(conf)# match ip address 1
R1(conf)# set dampening  5 1900 2000 10
!!!     5=half-life, 1900=reuse-limit, 2000=suppress-limit, 10=max-suppress-limit
R1(conf)# route-map damp-R1 permit 20
!
R1(conf)# router bgp 200
R1(conf-r)# neighbor IP@R2 remote-as 200
R1(conf-r)# bgp dampening route-map damp-R1
```



- ISP Design: Intra/Inter domain design

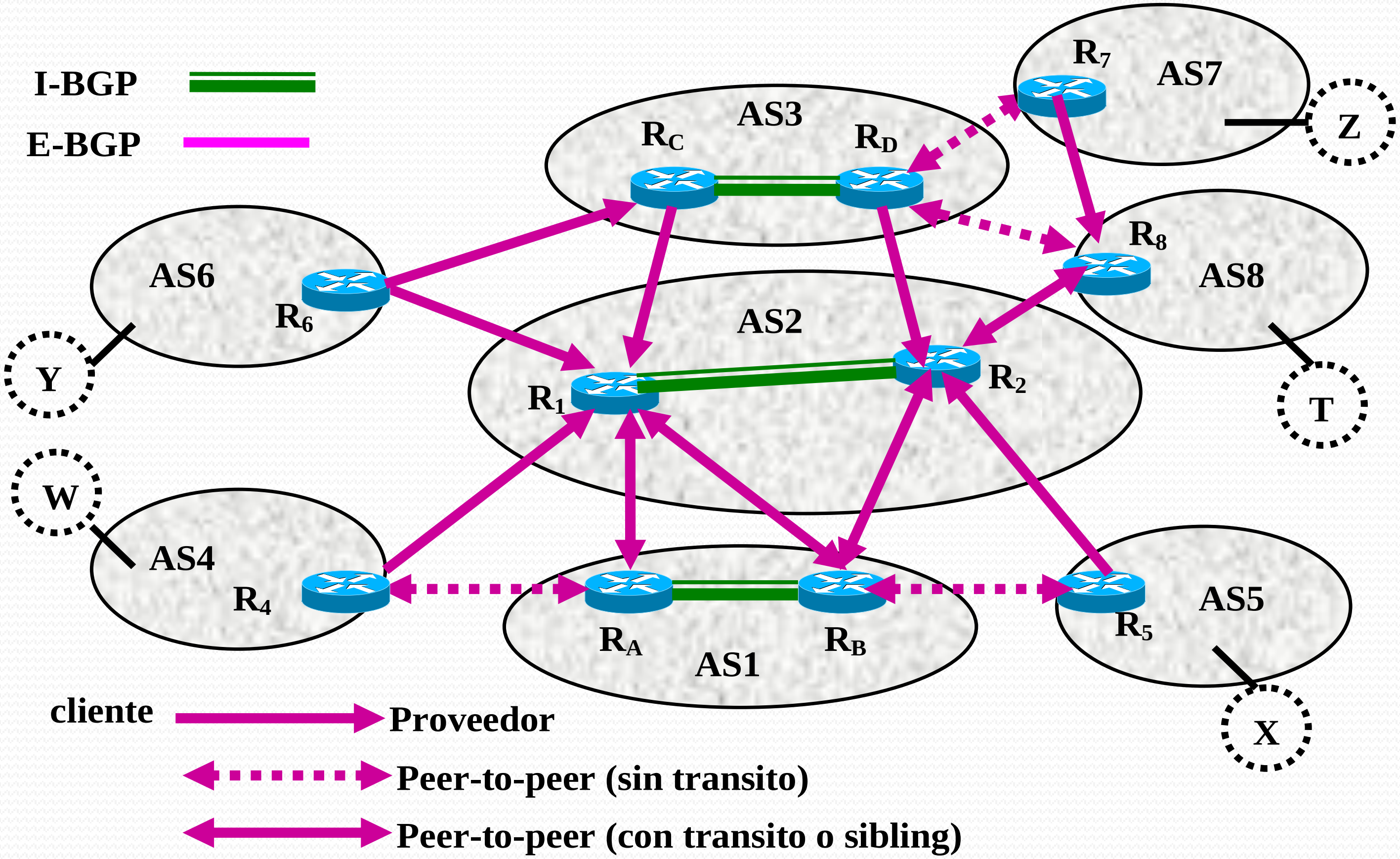


- ISP Design: Intra/Inter domain design

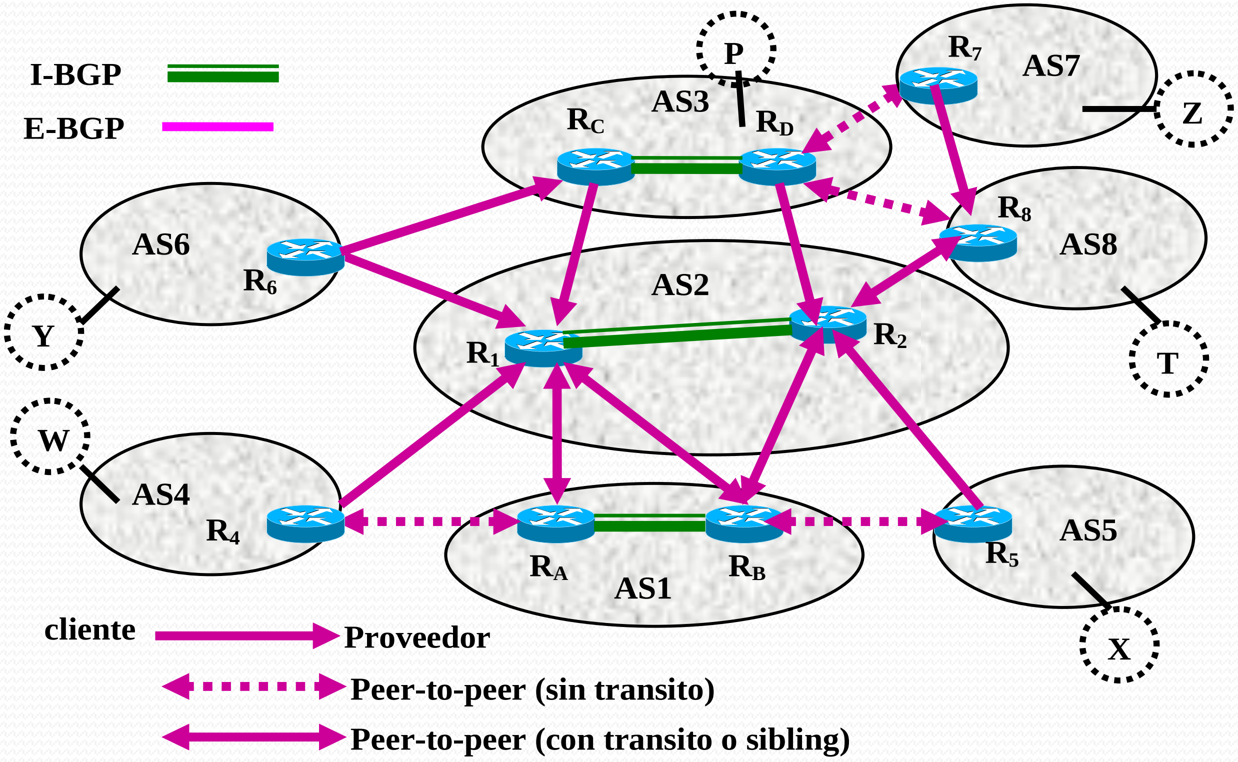


- **Exercise - Scalability:**
  - Assume an ISP of 200 BGP routers.
    a) Calculate the total number of i-BGP connections
    b) If we define 5 Route Reflectors with 39 customers each, calculate the number of i-BGP connections and the savings (in %)
    c) If we define 5 Confederations with 40 routers each, calculate the number of i-BGP + e-BGP connections and the savings (in%)
    d) If we define 5 Confederations with 40 routers each, and create 2 Route Reflectors with 19 routers per confederation, calculate the number of i-BGP + e-BGP connections and the savings (in%)

- **Exercise – Peering. BGP Routing table:**
  - Write the AS-Path between networks: a) X → Y, b) W → X, c) W→ Z, d) Y → Z



- **Exercise – Peering. UPDATES:**
  - Write which networks are announced from: a) AS2 to AS8, b) AS2 to AS8, c) AS8 to AS7, d) AS3 to AS8, e) AS8 to AS3


I-BGP
E-BGP
P
AS3
RC
RD
R7
AS7
Z
R8
AS8
AS6
R6
AS2
Y
R1
R2
T
W
AS4
R4
RA
RB
AS1
R5
AS5
X
cliente
Proveedor
Peer-to-peer (sin transito)
Peer-to-peer (con transito o sibling)