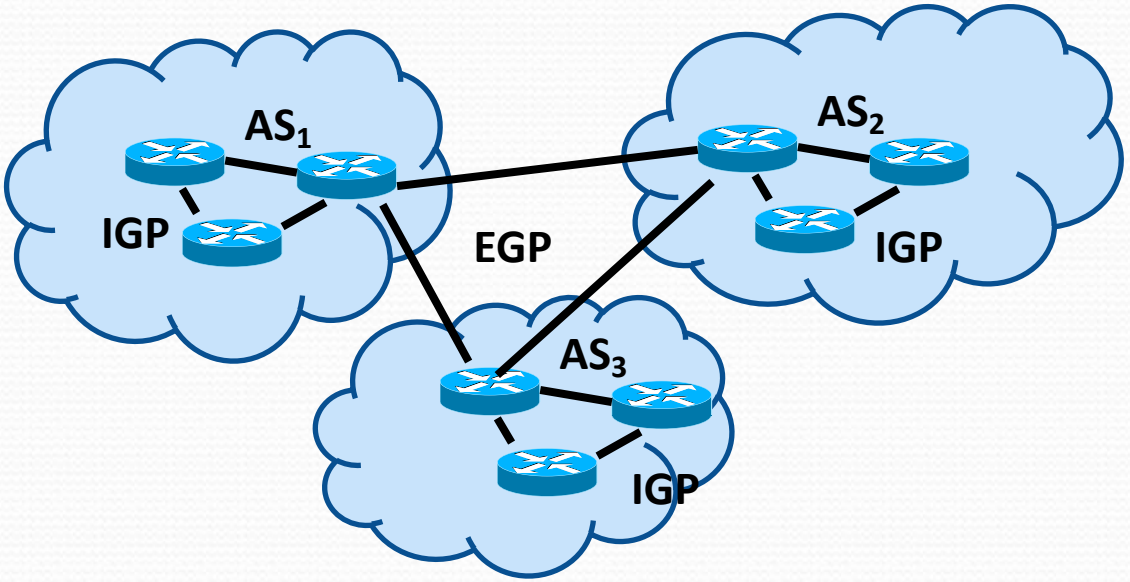
PI-Grau
(Internet Protocols)

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

- **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), called ASN (AS Numbers)
    - **0 ≤ ASN < 64512** are public AS numbers
    - **64512 ≤ ASN ≤ 65535** are private AS numbers (same idea as IP private addresses)
  - Extension to 32-bits (**ASN extedended**): **16-bits + 16-bit** (Old 16-bit ASN)
  - AS's exchange routes, IP subnets, using External Routing Protocols (EGP). Today, there only exists one EGP that is **BGPv4**

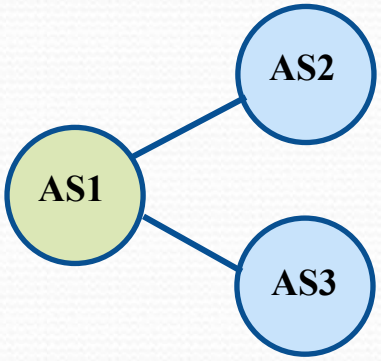


- **BGPv4**:
  - Is a routing protocol based on **policies**
  - It does **not** use **routing metrics** (hops, bandwidth, delay, ...)
  - Uses routing attributes that allow defining routing policies
  - BGP is encapsulated in TCP packets, thus, between two BGP routers should exist a TCP connection for 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

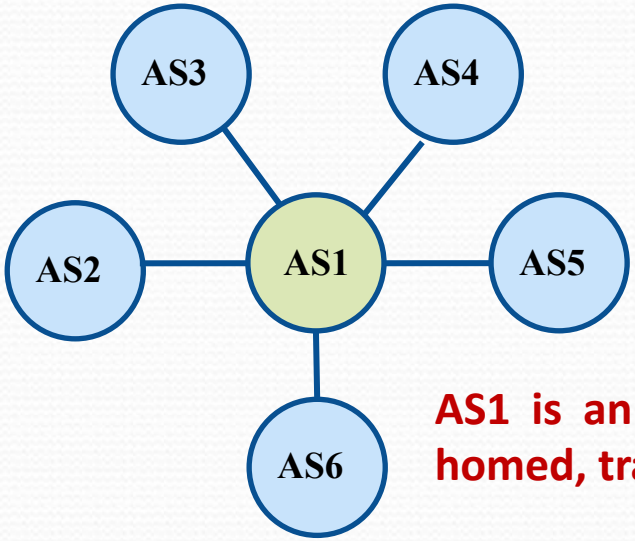
- **AS types of operation:**
  - **Stub AS or single-homed:** AS that reaches routes of other AS's using a **single connection point**
  - **Multi-homed AS:** AS that reaches routes of other AS's using more than one connection point but do not transit routes of other AS are called **Multi-homed non-transit**,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

AS1 is stub

AS1 is multi-homed (transit)



AS1 is multi-homed (non-transit)



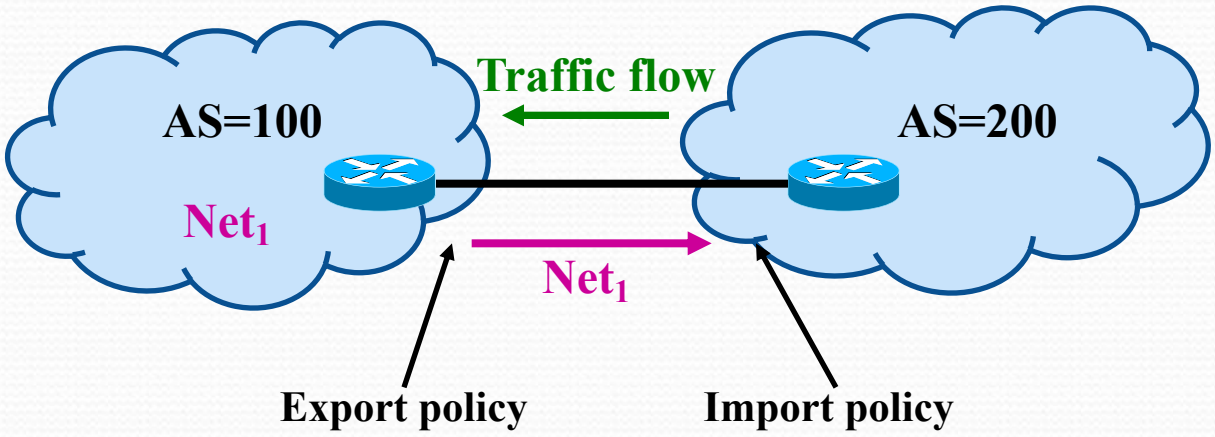
AS1 is an IPX (multi-homed, transit)

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

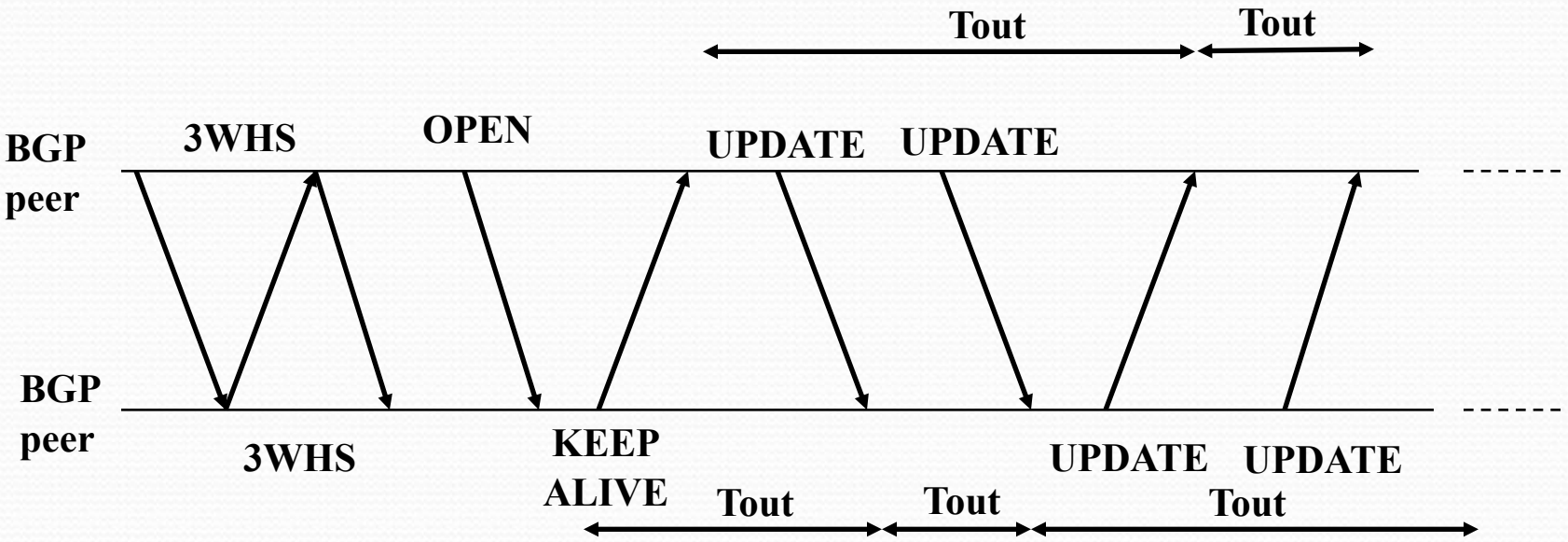
  http://www.ripe.net/ripe/docs/ripe-496

  - 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

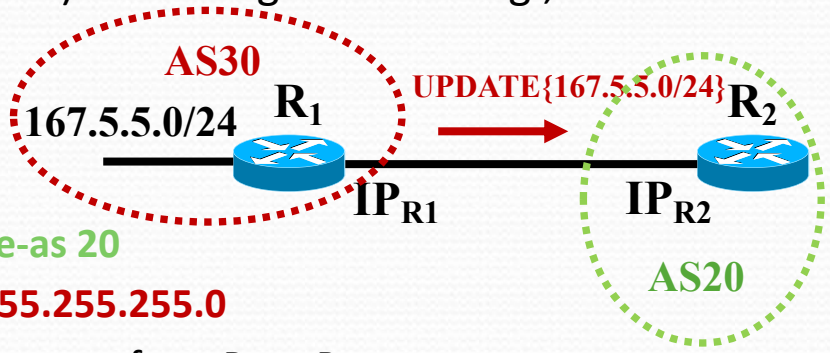
- **BGPv4 routing protocol:**
  - Announce routes, IP subnets, using administrative routing policies to other AS,
  - Traffic goes in opposite direction of the announcement of routes,
  - **Routing policy:** assume a subnet belongs to an AS, a routing policy means the decision of an AS to announce that 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 is alive
      - **UPDATE**: send routes and attributes
      - **NOTIFICATION**: error notification



- **BGPv4 routing operation:**
  - Neighbors open TCP connections (port 179) → e.g., command **neighbor.**
  - BGP exchange routes (UPDATE BGP packets) with neighbors → e.g., command **network,**
  - **The connection is unidirectional:**



R_1(router)# router bgp 30

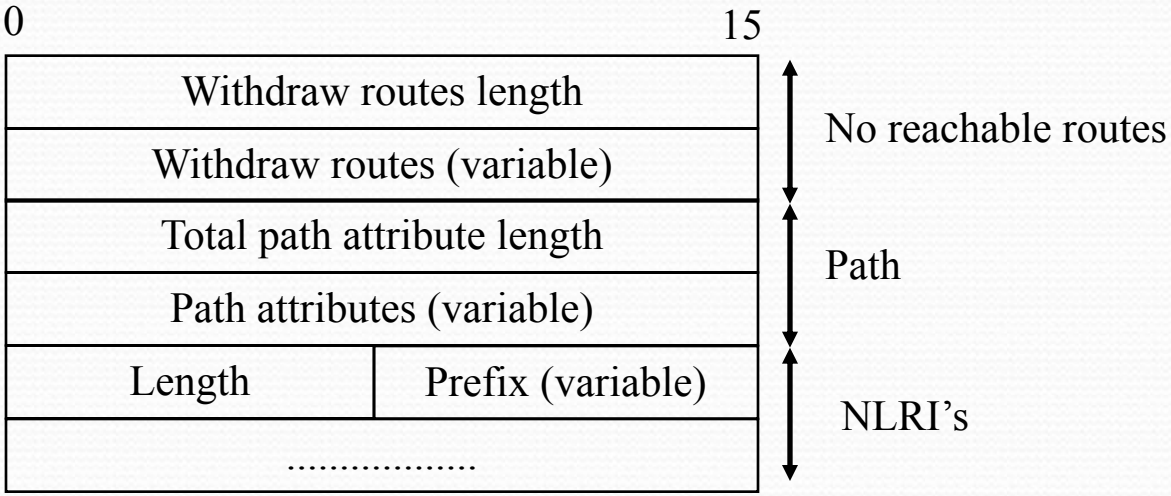
R_1(router)# neighbor IP_{R2} remote-as 20

R_1(router)# network 167.5.5.5 255.255.255.0

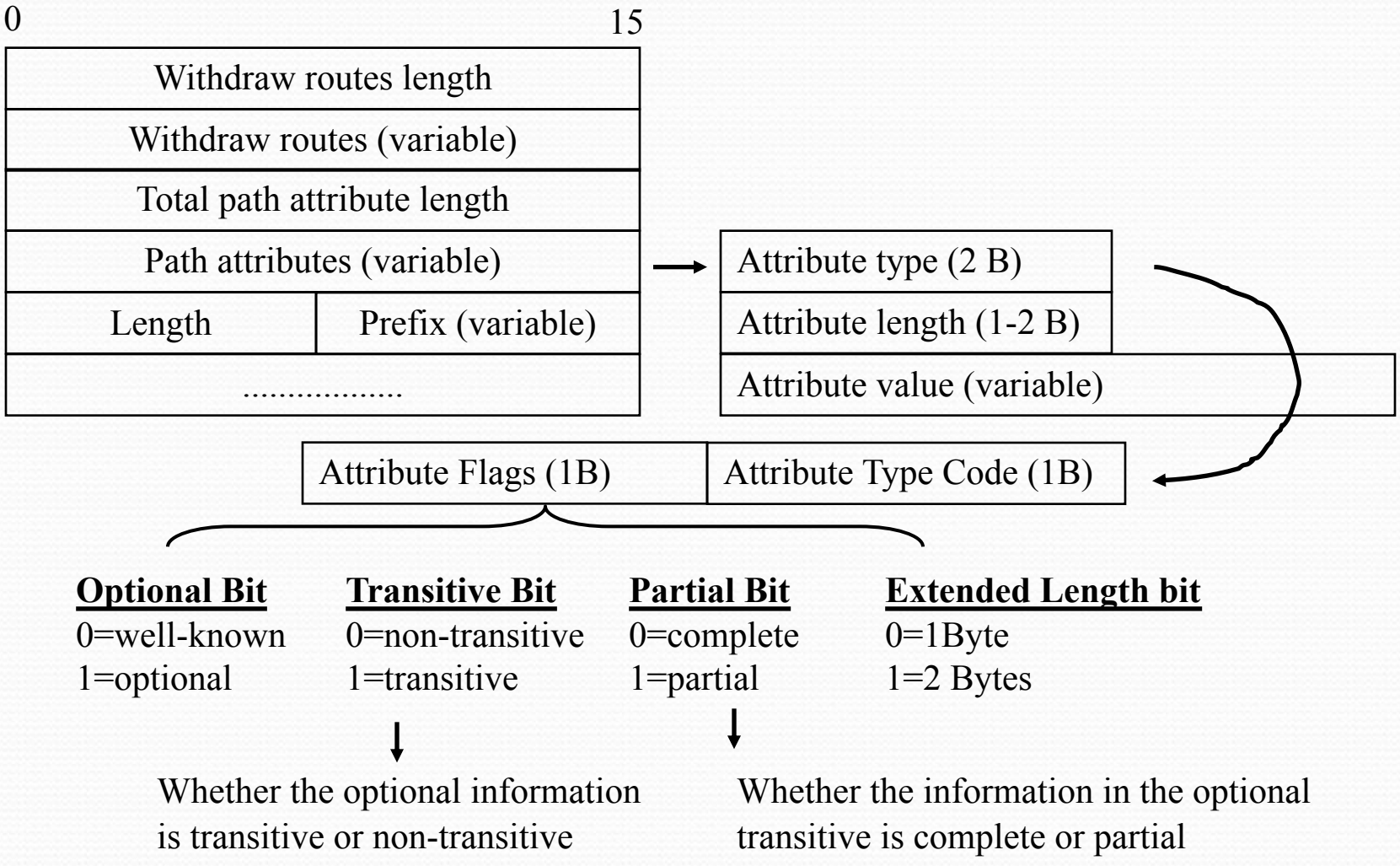
  - UPDATES in that connection for announcing routes from R_1 to R_2,
  - If R_2 wants to send UPDATES, then, R_2 has to open a neighbor connection with IP_{R1},
  - Route information is kept in **BGP tables (different from the routing table),**
  - The best path to each route (BGP routing decision process) is kept in the **routing table (RIB),**
    - BGP routing table grows (without filtering) as N*M, where N is the number of Internet routes and M is the number of AS neighbors,
  - BGP neighbors periodically exchange **keepalive messages** to maintain the TCP connection open (remember that TCP can close the connection after a timer, Tout, of several hours).

- **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 may or may not be 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



- **UPDATE messages**



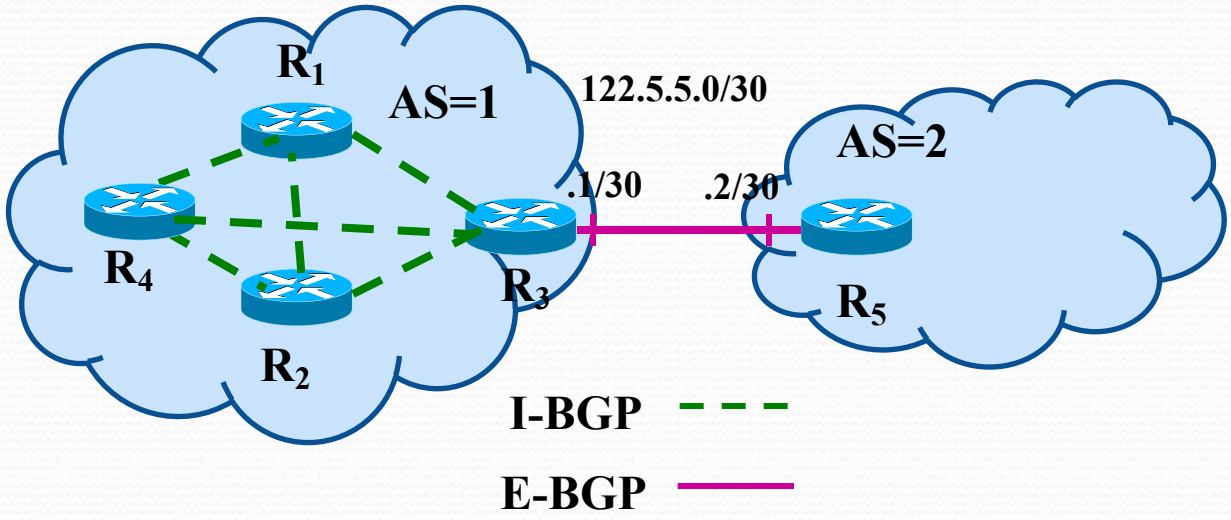
- **BGP Table**
  - Includes the following information: **subnet and mask**, **next-hop**, **MED** (metric), **Local_Pref**, **AS-path-vector** and **origin**
  - The decision process depends on:
    - manufacturer implementation, but basically all BGP tables are very similar,
    - maintain a DB for each active BGP session,
    - **a router only announces its "best route" in UPDATE BGP messages:**
      - Then a router has one entry per every BGP connection in its BGP table. If the router has N BGP connections, then it has N entries for each route except that a route has been filtered using ACL's, and then it has not been sent. In that case, the router has a partial Internet view since a router **only sees** what other routes decide to sent.
    - from the N entries for each route, the BGP router **uses a decision process** to select its best entry among the N possible entries for a route. Symbol **">"** indicates **the best entry** towards a route,
    - the best entry is copied in the Routing Table for being used in the routing process of IP packets.

- **BGP table**

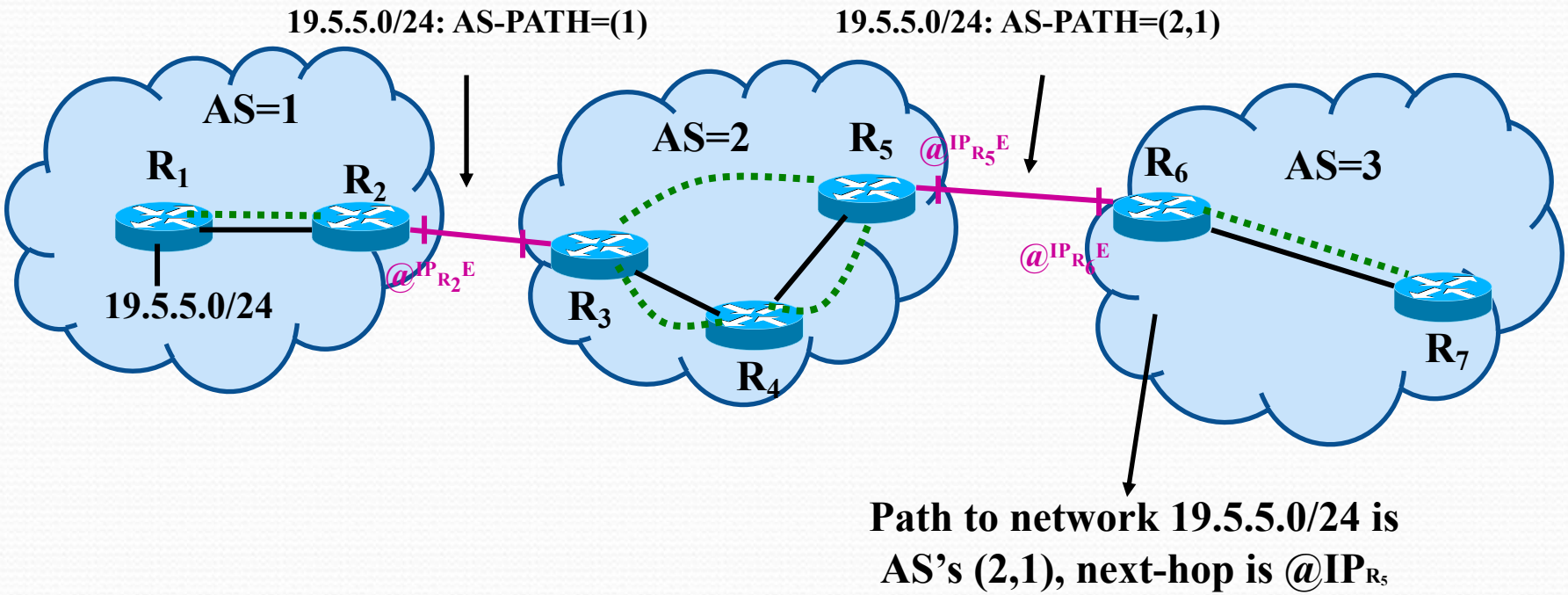
R2# show ip bgp

		Attributes			
Net/Mask	**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
……					
* 6.0.0.0/16	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



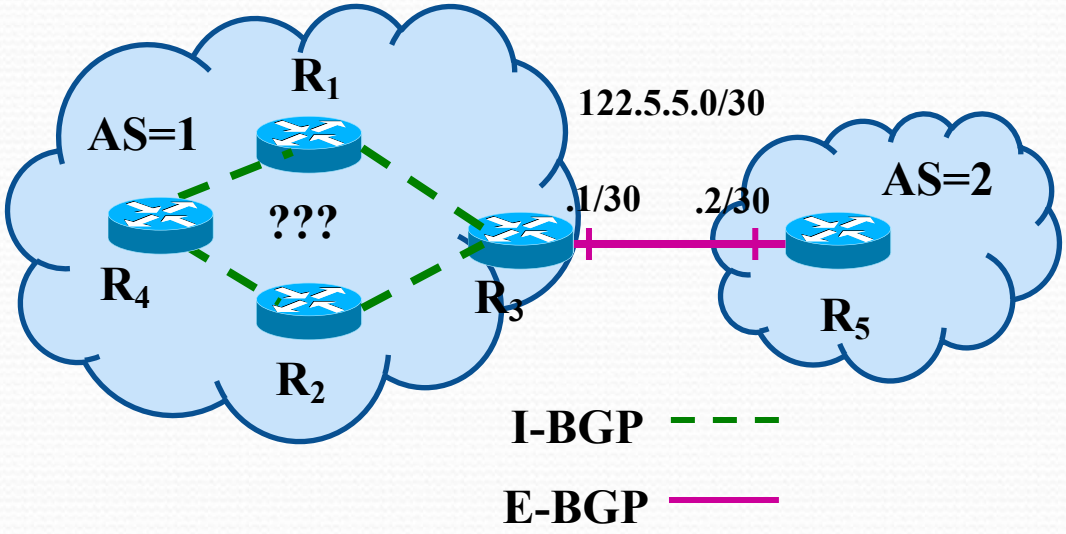
- **AS-PATH VECTOR (Well-known and Mandatory):**
  - Represents the path that a route follows from the origin AS → gives all AS traversed to reach that route,
  - Each AS adds its ASN to the AS-PATH vector when connected to E-BGP, never when connected to I-BGP routers,
  - **AS-PATH = (AS_x, …., AS_{origin})** → **AS_x** is the neighbor that sends the UPDATE, **AS_{origin}** is the owner of the route that is announce in the UPDATE, the rest are AS's that are crossed.



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:
  - Routes learnt via E-BGP may be advertised via E-BGP and I-BGP,
  - Routes 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,

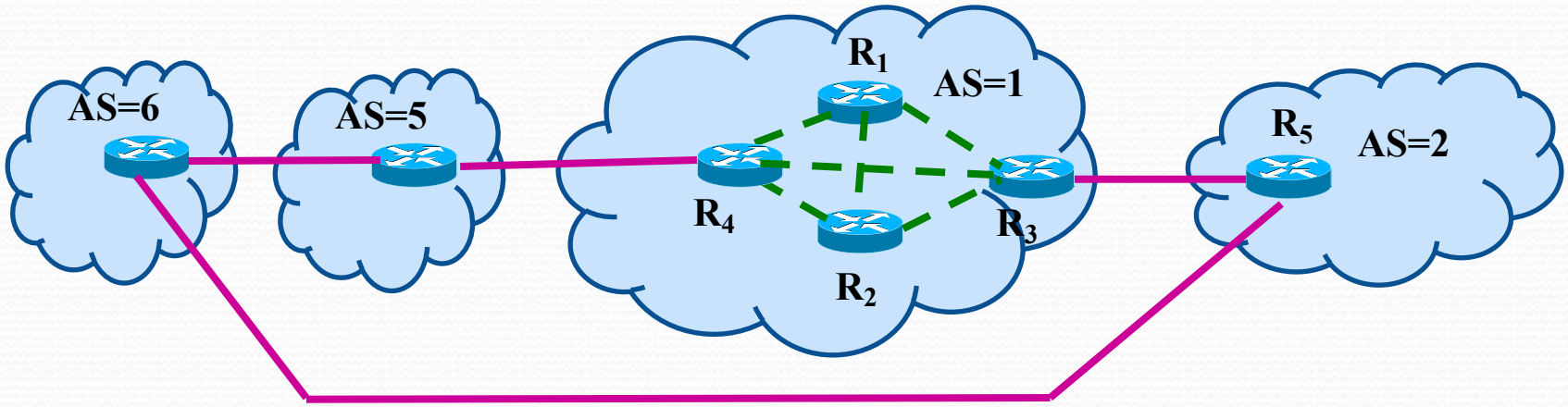
→ **I-BGP routers should form a mesh I-BGP network → problem of scalability that is solved via "BGP route reflectors" a "BGP confederations"**



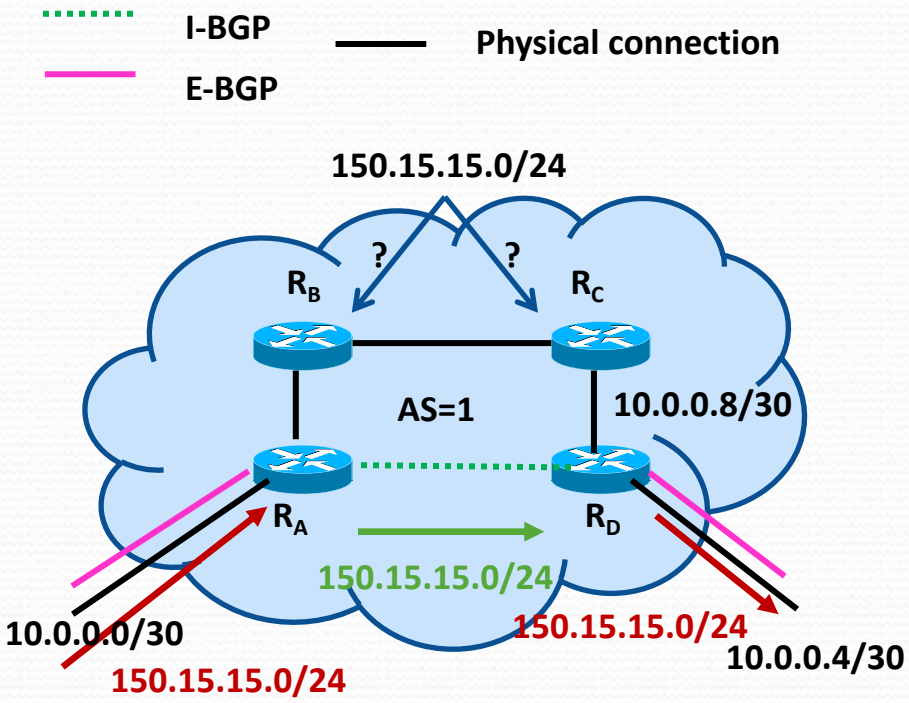
In this example. No I-BGP between R_4 and R_3:

- R_3 learn routes from R_5, R_1 y R_2 but not from R_4,
- 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?*** R_1 announces its own routes to R_3 and R_4, but it does not announce routes from R_3 to R_4 or from R_4 to R_3.

- **Why I-BGP routers do not advertise routes learnt via I-BGP ?**
  - BGP routers advertise the AS-path-vector attribute that includes all the AS that the routes crosses. **Objective**: **loop detection**, e.g.; AS2 receives and UPDATE with AS-PATH=(1 5 6 2). If it forwards the UPDATE then the UPDATE will cross twice AS=2 → there is a loop, then AS2 does not forward the UPDATE. **AS-PATH vectors allow the detection of loops.**
  - **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? e.g., AS1 has I-BGP connections:*
    - AS1 would add AS1 several times to the AS-PATH-vector. Then **i) I-BGP disables adding the ASN in the AS-PATH vector in I-BGP to avoid repetition of the same ASN in the AS-PATH**, e.g., avoid AS-PATH={2,1,1,1} because traverses 3 times BGP interior routers in AS=1,
    - But, then, it is unable to detect loops inside AS1. Then, **ii) I-BGP does not forward UPDATES learnt from another I-BGP router.**
  - **Consequence: I-BGP routers has to be FULL MESHED.**



- **BGP synchronization.** Synchronization implies **two issues**:
  1. 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 (not so good idea, BGP has better administrative distance than OSPF) or **create a full meshed I-BGP network** → all routers between two BGP routers have to be BGP,



Routing Table of R_C:

Net/Mask, gateway, interface

…..

150.15.15.0/24 is not in the routing table

BGP Table of R_D:

Net/Mask	Next-Hop	MED	LPref	AS-PATH	Orig
150.15.15.0/24	10.0.0.1	100	100	23 456	i

…..

Routing table of R_D:

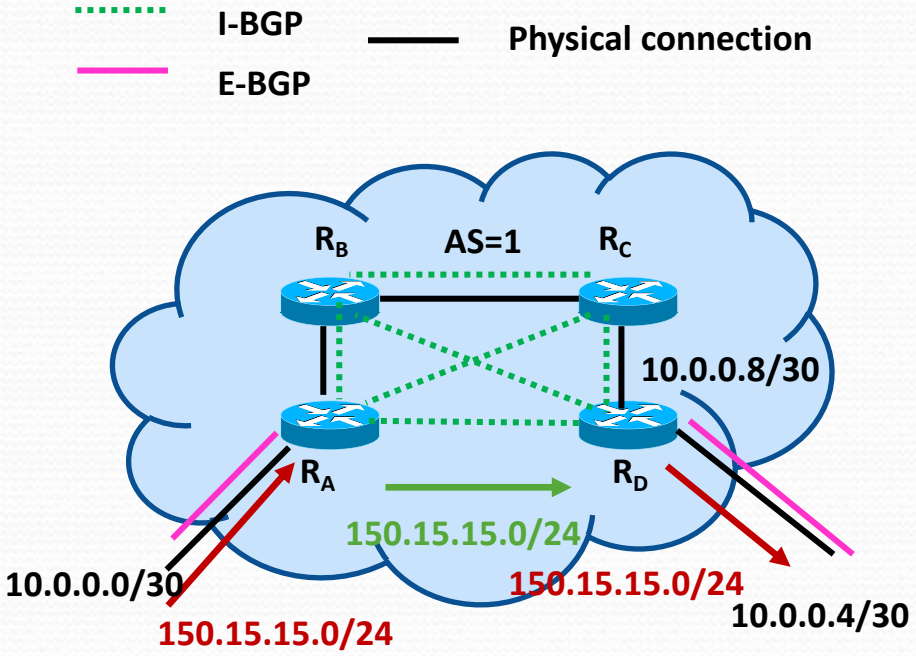
Protocol, Net/Mask, gateway, interface

BGP 150.15.15.0/24, via 10.0.0.1, interface ge0

OSPF 10.0.0.0/30, via 10.0.0.9, interface ge0

…..

- **BGP synchronization.** Synchronization implies **two issues**:
  2. **BGP Next Hop** address have to be known and learnt by OSPF → **OSPF synchronized with BGP**. That means that R_A learns via OSPF how to reach network 10.0.0.4/30 and R_D learns how to reach network 10.0.0.0/30 that are the boundary nets with other AS's.



BGP Table of R_D:

Net/Mask	Next-Hop	MED	LPref	AS-PATH	Orig
150.15.15.0/24	10.0.0.1	100	100	23 456	I
…..					

Routing table of R_D:

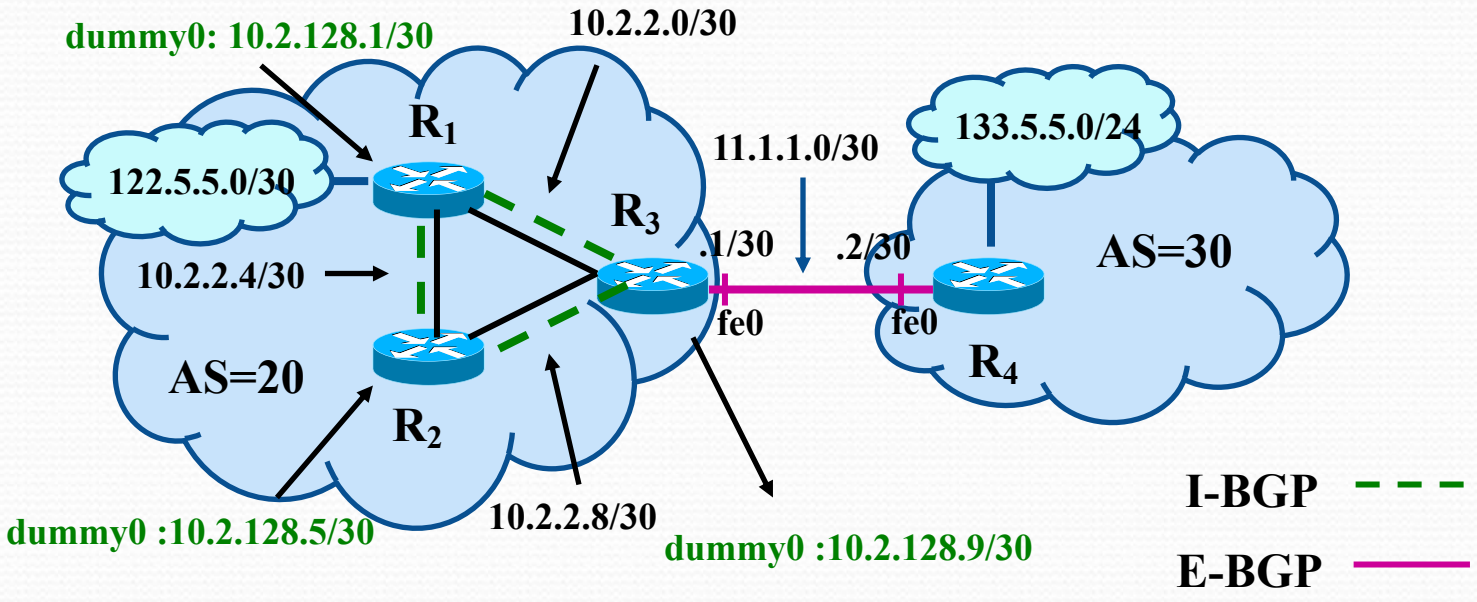
Protocol, Net/Mask, gateway, interface

BGP 150.15.15.0/24, via 10.0.0.1, interface ge0

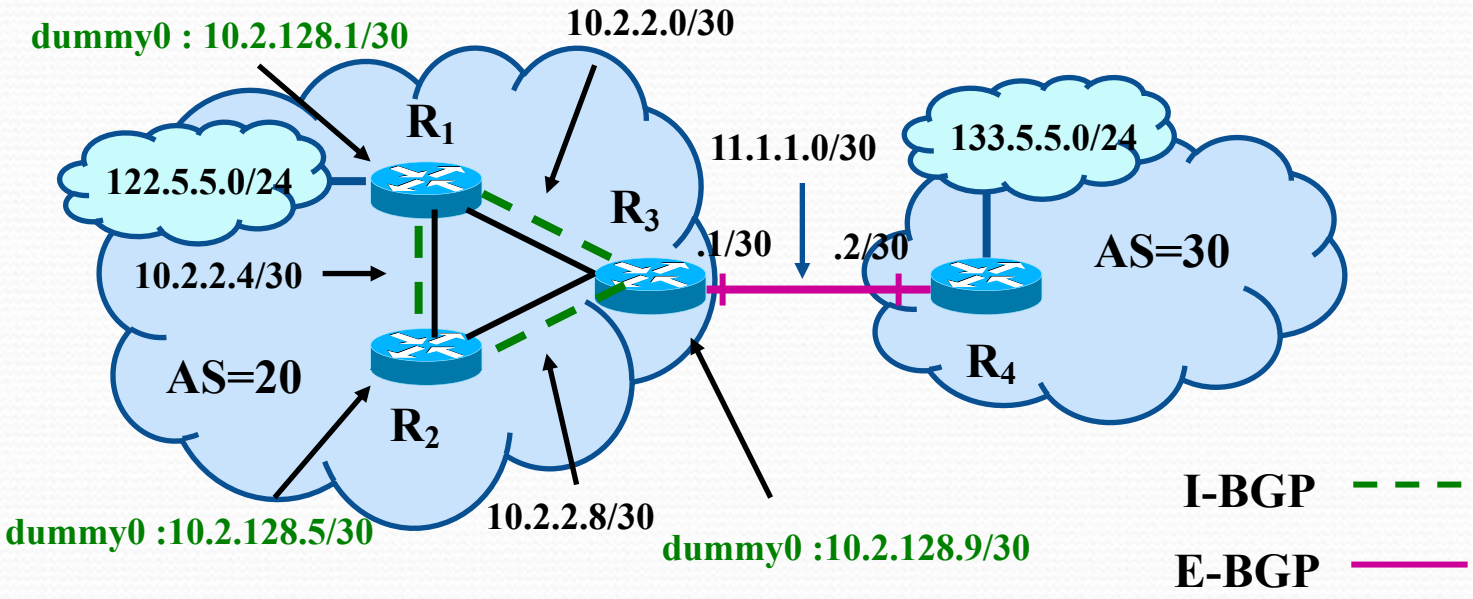
…..

Net 10.0.0.0/30 has not been announced by OSPF and R_D does not know how to reach it, it is to say, this net is not in the R_D routing table.

- **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 10.2.2.1-10.2.2.2 and the link fails → **the I-BGP session is lost**
  - *Is it possible to reach R_3 from R_1?* Yes, R_3 is reach directly (R_1-R_3) or via R_2 (R_1-R_2-R_3):
    - Use a loopback interface: loopback of host is 127.0.0.0/8, but loopback (also called sometimes dummy) is a public/private IP addresses, e.g. R_1 has as loopback address 10.2.128.1/30, **that also has to be announced via OSPF** in order that R_3 is able to reach it.
    - **Take care !!!** When "update-source" command is used, the BGP IP@ should paired. It is to say, if R_1 uses 10.2.128.1 as "update-source" address to R_3, then R_3 should bgp with the command neighbor to 10.2.128.1 (source address of R_1).



- **BGP configuration with CISCO IoS**
  - Be carefull: AS's should be OSPF isolated (they belong to different domains), thus, you have to use "passive-interface" OSPF command between R_3-R_4 and viceversa,
  - The command "update-source" and loopback addresses between BGP connections only is used between I-BGP connections. It has no sense between E-BGP, if there are no OSPF alternatives between E-BGP routers.



- **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

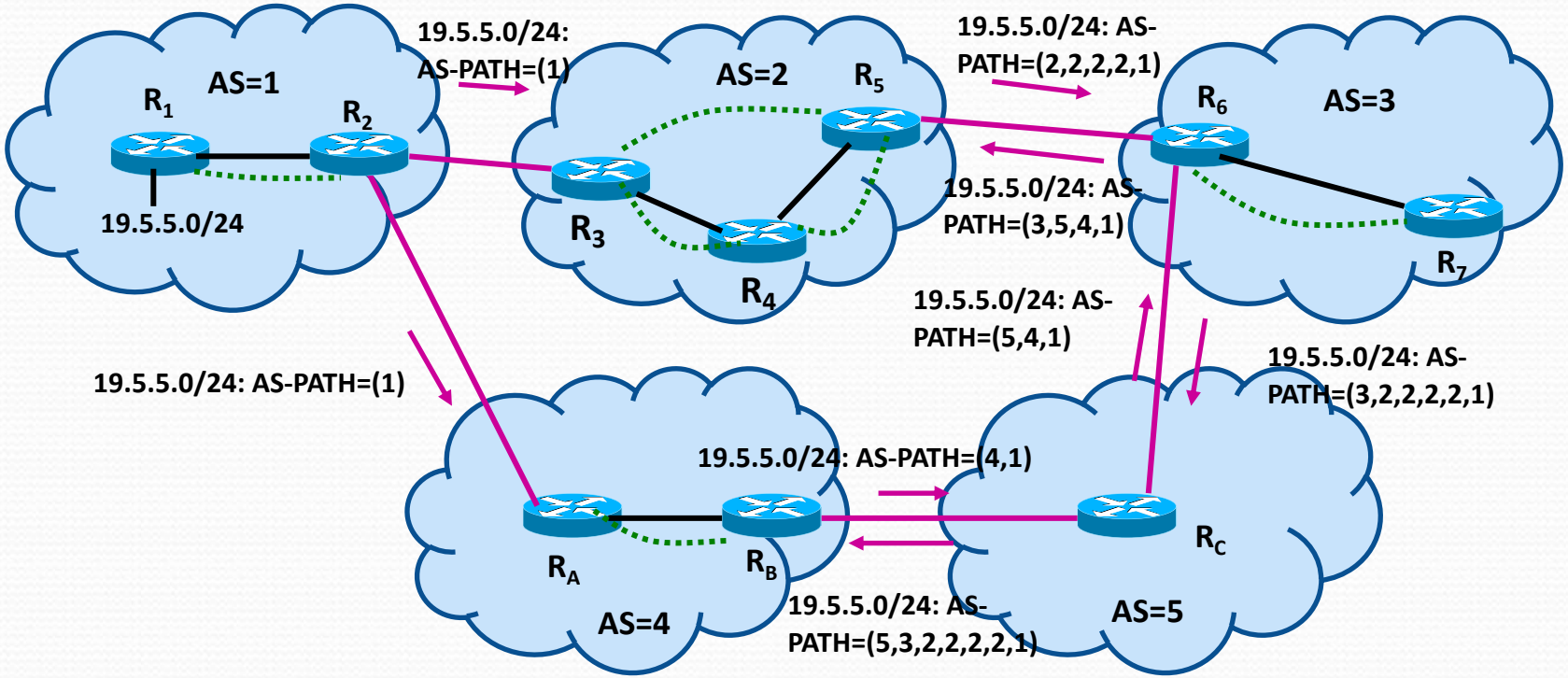
R4(conf-r)# **neighbor** 11.1.1.1 **remote-as** 20

R1(conf-r)# **network** 133.5.5.0 255.255.255.0

- **Manipulating the AS-PATH vector attribute:**
  - One of the decision process steps is the **shortest path** (in AS hops),
  - Manipulate the AS-PATH: **"prepending" (increase) the AS-PATH vector.**

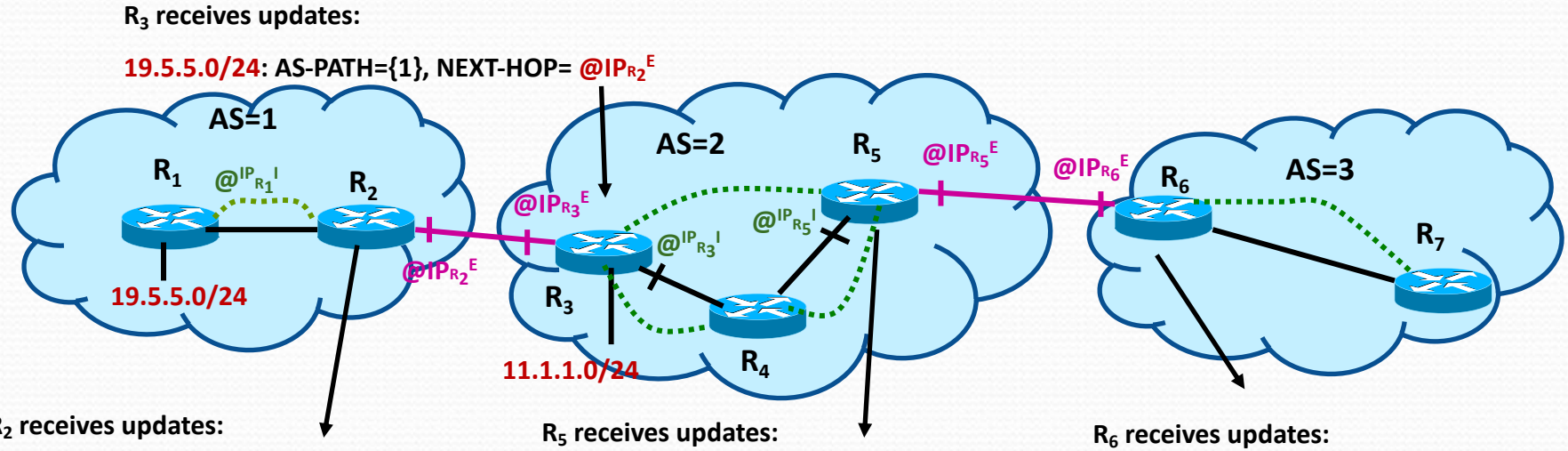
BGP Table of R_6:

	Net/Mask	Next-Hop	MED	LPref	AS-PATH	Orig
*>	19.5.5.0/24	IP@-R_C	100	100	5,4,1	i
	19.5.5.0/24	IP@-R_5	100	100	2,2,2,2,1	i

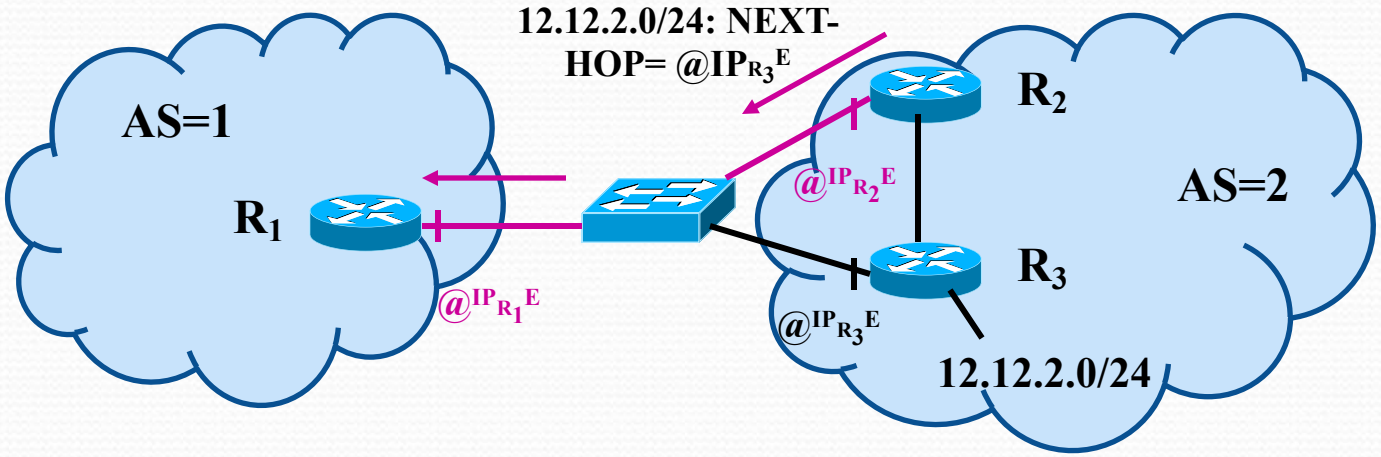


NEXT-HOP (Well-known and Mandatory):

- **For an E-BGP session** is the @IP of the BGP router that forwards the route (E-BGP neighbor),
- **For an I-BGP session:**
  1. **The Net is generated in an internal BGP router of the same AS:** is the @IP of the I-BGP router that generated (owner) the route,
  2. **The Net comes from other AS, e.g., from an external BGP router:** is the @IP of the external E-BGP router that forwarded the route from the other AS,
- 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$_{R_2}^{E}$**



- **Manipulating 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,
  - when it has to announce network 12.12.2.0/24, @IP_{R3} is best next-hop than @IP_{R2} to reach R_1 since there is a switch. However, there is no BGP connection between R_3-R_1 (maybe R_3 is not a BGP router)
  - R2 can announce route 12.12.2.0/24 in behalf of R_3, but instead of using its @IP_{R2} as nexh-hop, it manipulates it to be @IP_{R3}.



- **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

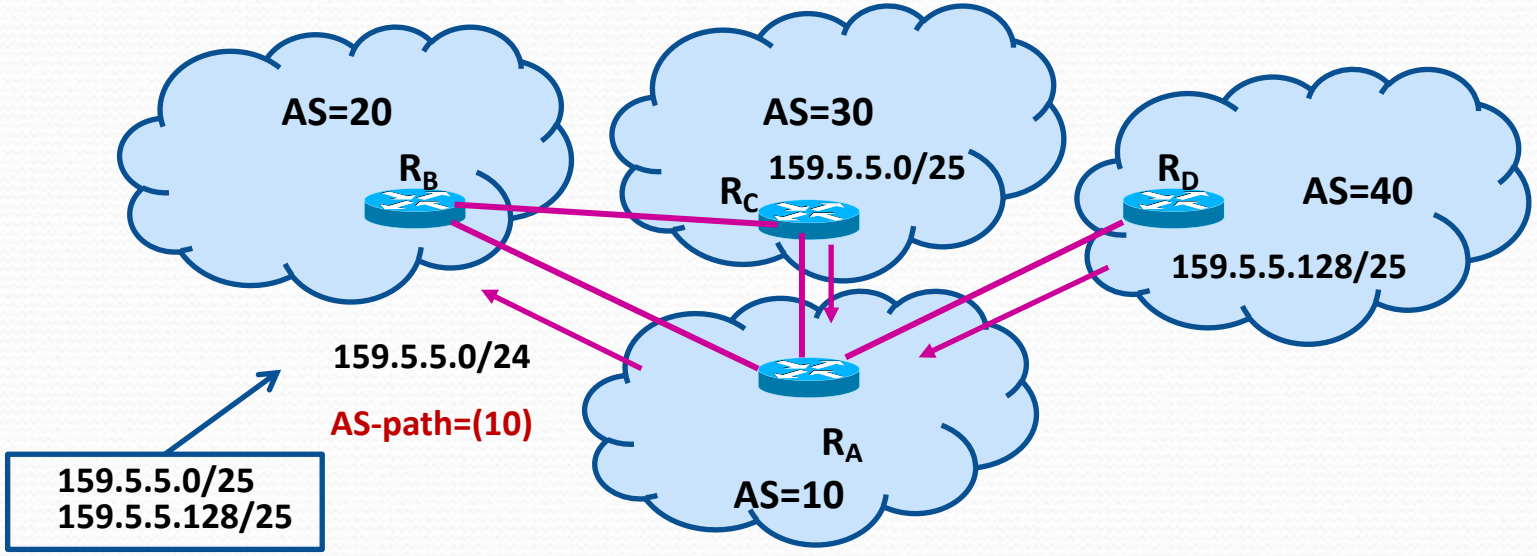
- **AGGREGATOR (Optional and transitive)**
  - A BGPv4 UPDATE message sends a subnet/mask that may be aggregated,
  - since there is no information on AS30 and AS40, **it is not possible to detect loops on these AS's**, e.g., when R_B send the UPDATE to R_C, it will be **159.5.5.0/24 AS-PATH=(20,10)**
  - it has not influence in the path selection.

Code in router R_A:

```
router bgp 10
   neighbor  IP@-R_B remote-as 20
   aggregate-address 159.5.5.0   255.255.255.0   summary-only
```

BGP Table of R_B:

	Net/Mask	Next-Hop	MED	LPref	AS-PATH	Orig
*>	159.5.5.0/24	IP@-R_A	100	100	10	i



AGGREGATOR (Optional and transitive)

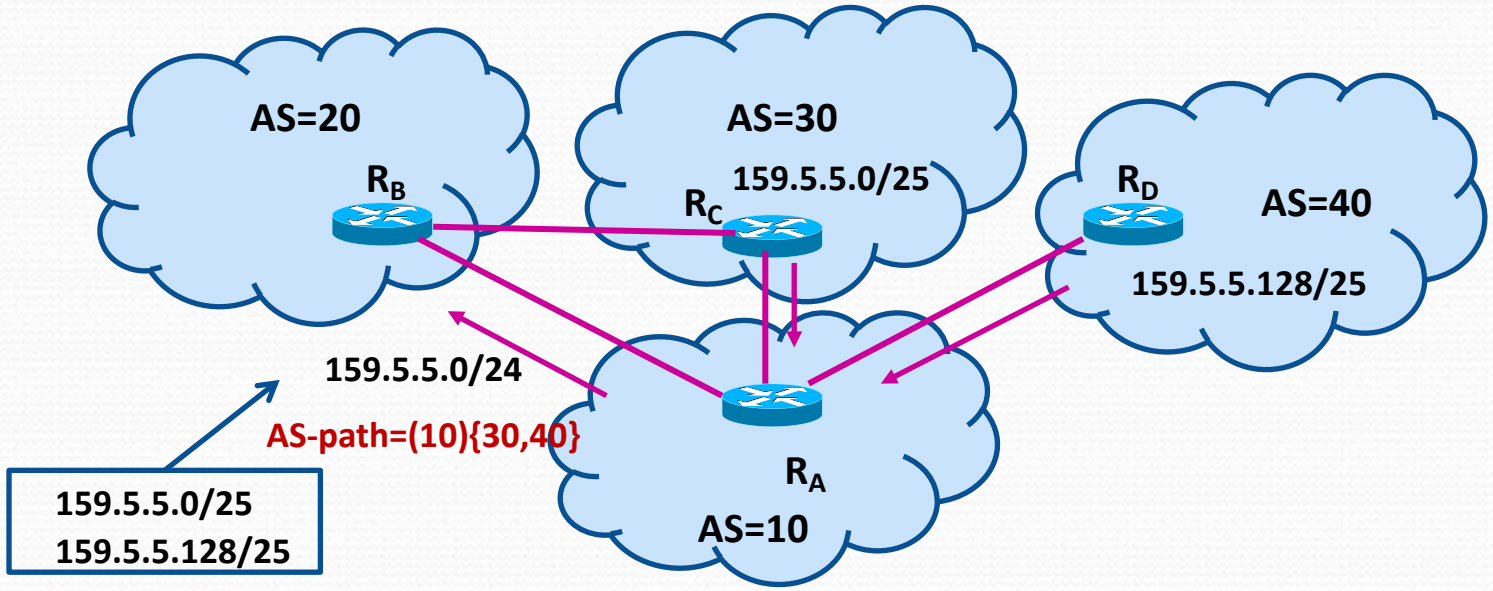
- A BGPv4 UPDATE message sends a subnet/mask that may be aggregated,
- The BGP router that aggregates can indicate in the AS-PATH vector the partition of the subnet aggregated (**AS-SET option**), **this helps the detection of loops,** e.g., when R_B send the UPDATE to R_C, it will be 159.5.5.0/24 AS-PATH=(20,10 {30,40})
- It has not influence in the path selection

Code in router R_A:

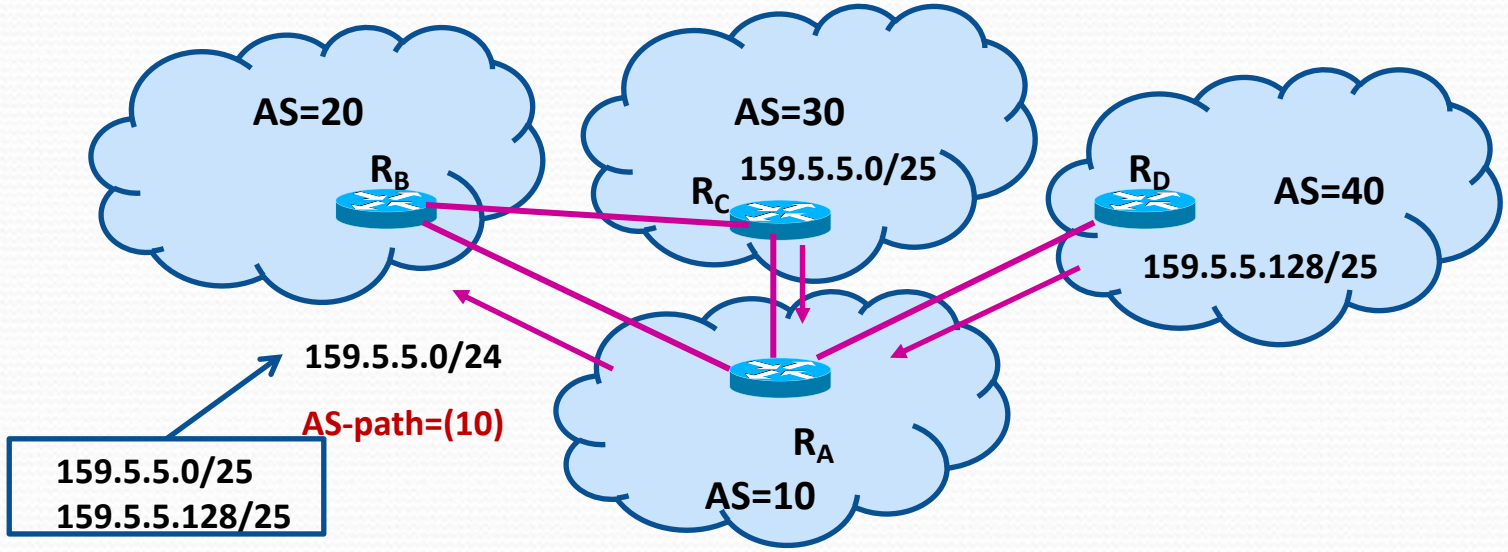
```
router bgp 10
  neighbor IP@-RB remote-as 20
  aggregate-address 159.5.5.0 255.255.255.0 summary-only as-set
```

BGP Table of R_B:

	Net/Mask	Next-Hop	MED	LPref	AS-PATH	Orig
*>	159.5.5.0/24	IP@-R_A	100	100	10 {30,40}	i



- **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 → it is mandatory that the Atomic Aggregate is active
  - **BE CAREFUL**:
    1. CISCO IOS by default does not use as-set, however JUNOS does the opposite. Thus, check manuals to see how to proceed when aggregating routes.
    2. Moreover, when aggregating may be interesting to modify other attributes, so, in general, aggregations is used in conjunction with communities.

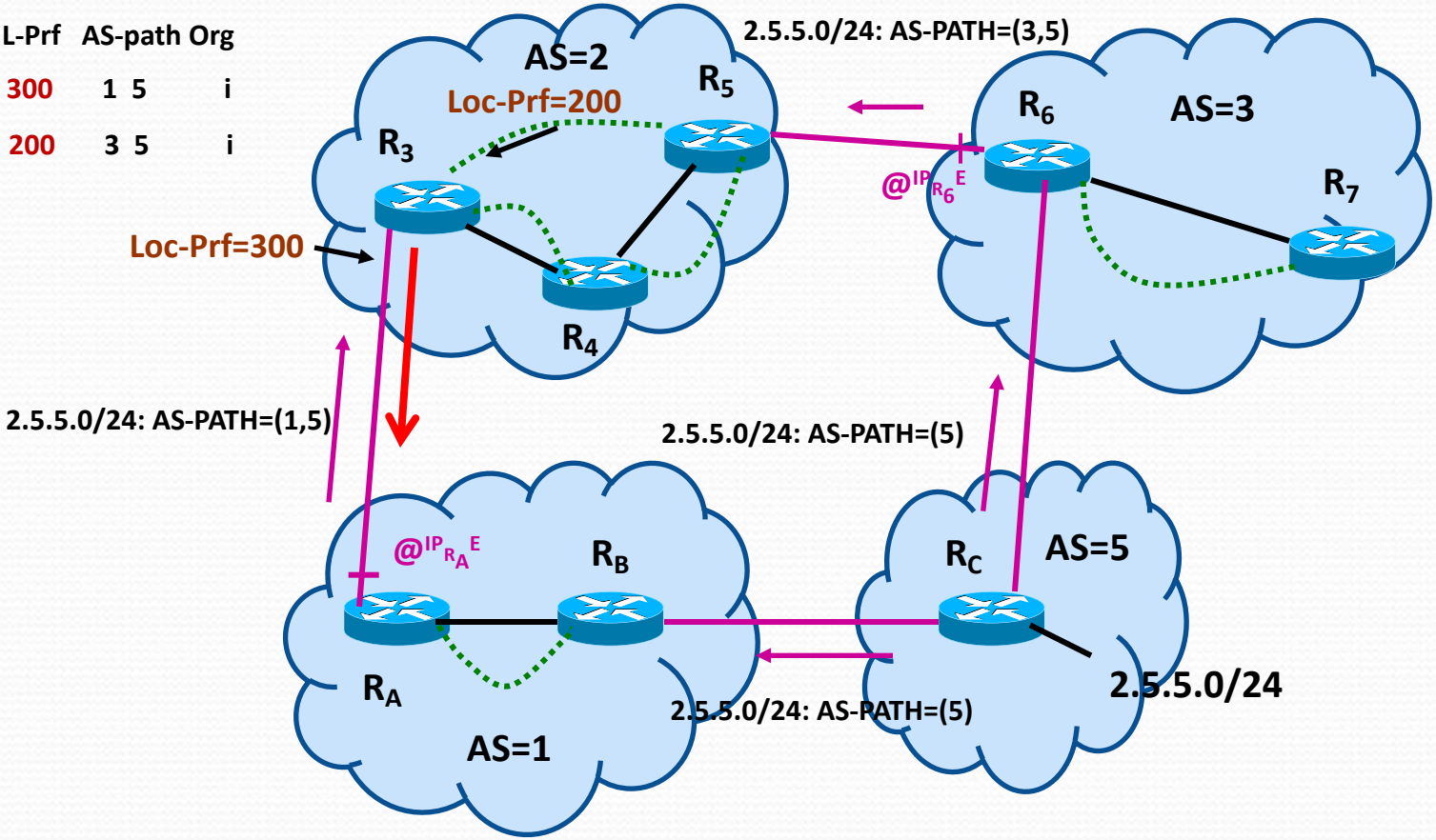


LOCAL-PREFERENCE (Well-Known and discretional):

- Attribute that indicates the preferred output link
- It is **ALWAYS** sent in UPDATES in I-BGP, but **NEVER** in UPDATES using E-BGP
- Highest values of Loc-Prf have higher preference (default value=100)

BGP table of R_3:

Net/Mask	N-hop	MED	L-Prf	AS-path	Org
*>2.5.5.0/24	$IP_{R_A}{}^E$	100	300	1 5	i
2.5.5.0/24	$IP_{R_6}{}^E$	100	200	3 5	i



- **BGP configuration with CISCO IoS: manipulate attributes**
  - **Route Maps: tool to create conditionals in CISCO IOS**

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

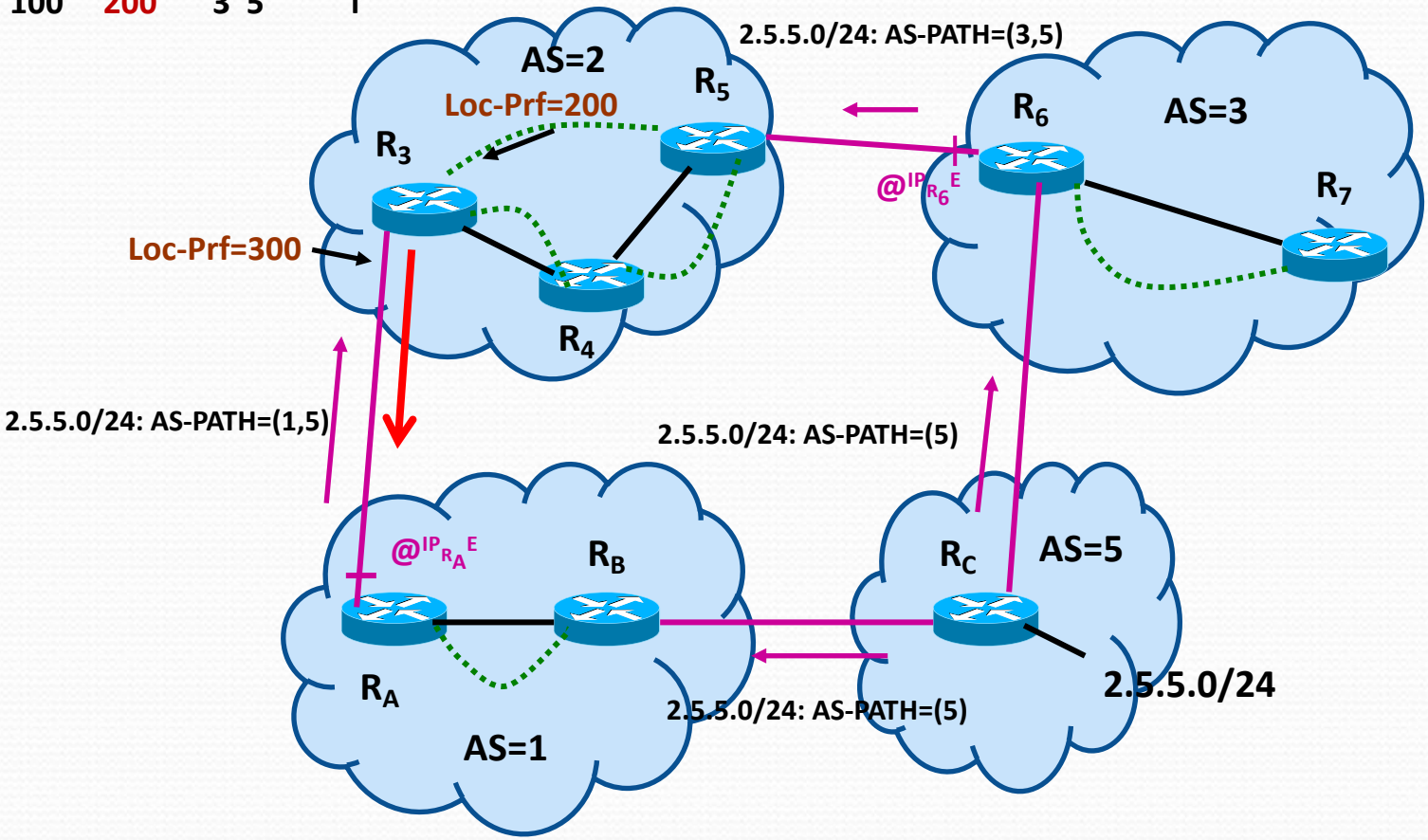
If condición1 **then** acción1
elseif condición2 **then** acción2
elseif condición3 **then** acción3
else acción4

route-map MYMAP **permit** 10
match condición 1
set acción 1
route-map MYMAP **permit** 20
match condición 2
set acción 2
route-map MYMAP **permit** 30
match condición 3
set acción 3
route-map MYMAP **permit** 40
set acción 4
route-map MYMAP **permit** 50

- **BGP configuration with CISCO IOS: add Local-Pref**

BGP table of R_3:

Net/Mask	N-hop	MED	L-Prf	AS-path	Org
*>2.5.5.0/24	$IP_{R_A}{}^E$	100	300	1 5	i
2.5.5.0/24	$IP_{R_6}{}^E$	100	200	3 5	i



BGP configuration with CISCO IOS: add Local-Pref

!!!! 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** rr5 **permit** 10

R3(conf)# **match ip address** 1

R3(conf)# **set local-pref= 200**

R3(conf)# **route-map** rr5 **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

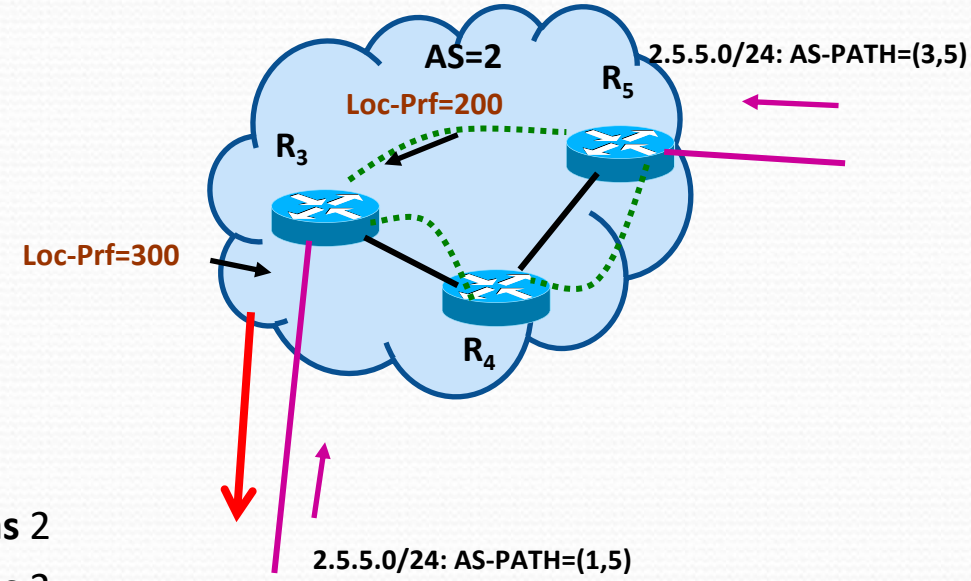
R3(conf-r)# **neighbor** IP@R4 **remote-as** 2

R3(conf-r)# **neighbor** IP@R5 **remote-as** 2

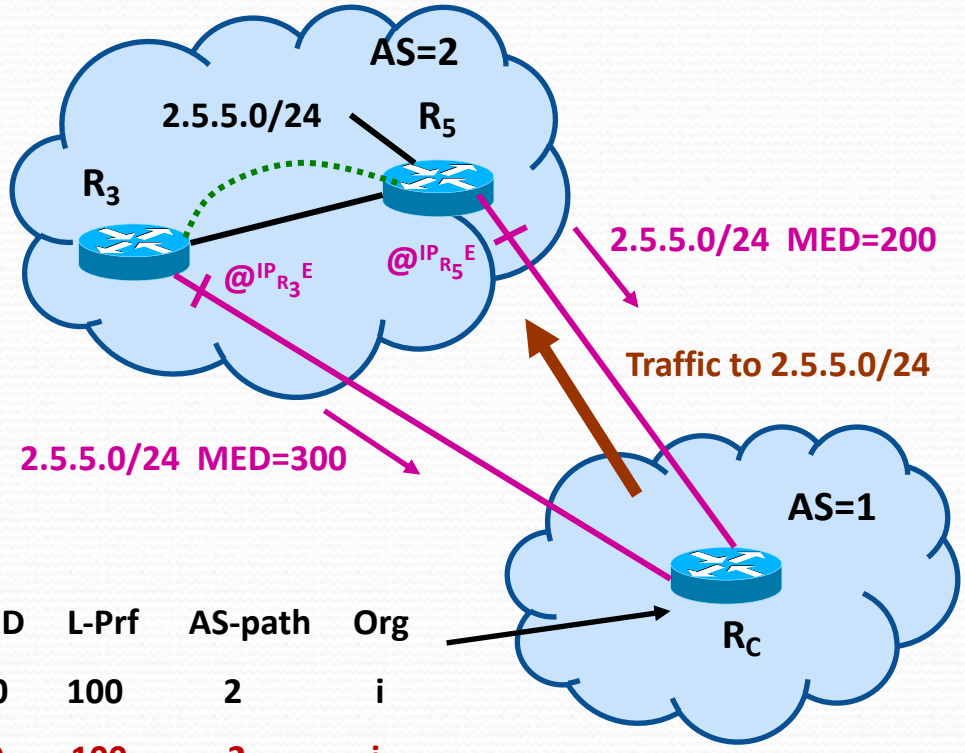
R3(conf-r)# **neighbor** IP@RA **remote-as** 1

R3(conf-r)# **neighbor** IP@R5 **route-map** rr5 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



BGP table of R_C:

Net/Mask	Next-hop	MED	L-Prf	AS-path	Org
2.5.5.0/24	$IP_{R_3}{}^E$	300	100	2	i
*> 2.5.5.0/24	$IP_{R_5}{}^E$	200	100	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= 200**

R5(conf)# **route-map** rrC **permit** 20

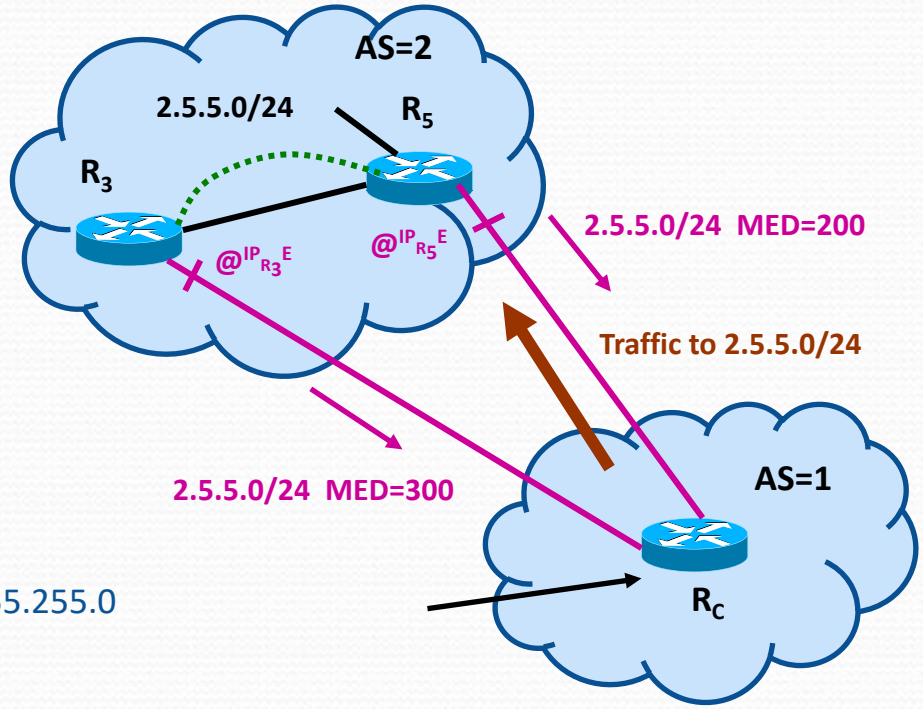
R5(conf)# **router** bgp 2

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

R5(conf-r)# **network** 2.5.5.0 255.255.255.0



!!!! 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= 300**

R3(conf)# **route-map** rrC **permit** 20

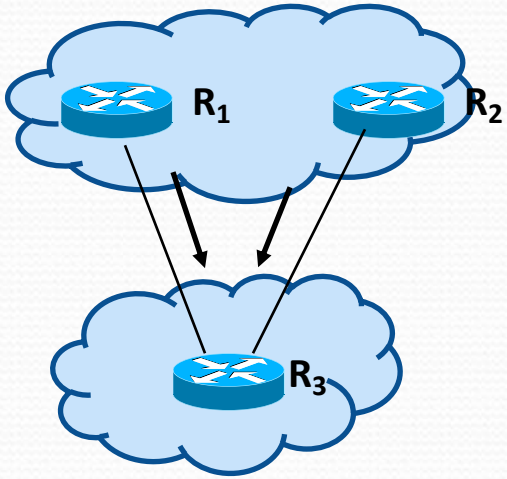
R3(conf)# **router** bgp 2

R3(conf-r)# **neighbor** IP@R5 **remote-as** 2

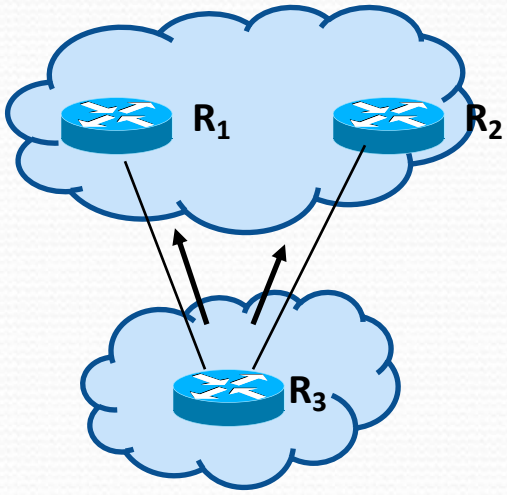
R3(conf-r)# **neighbor** IP@RC **remote-as** 1

R3(conf-r)# **neighbor** IP@RC **route-map** rrC out

- **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**

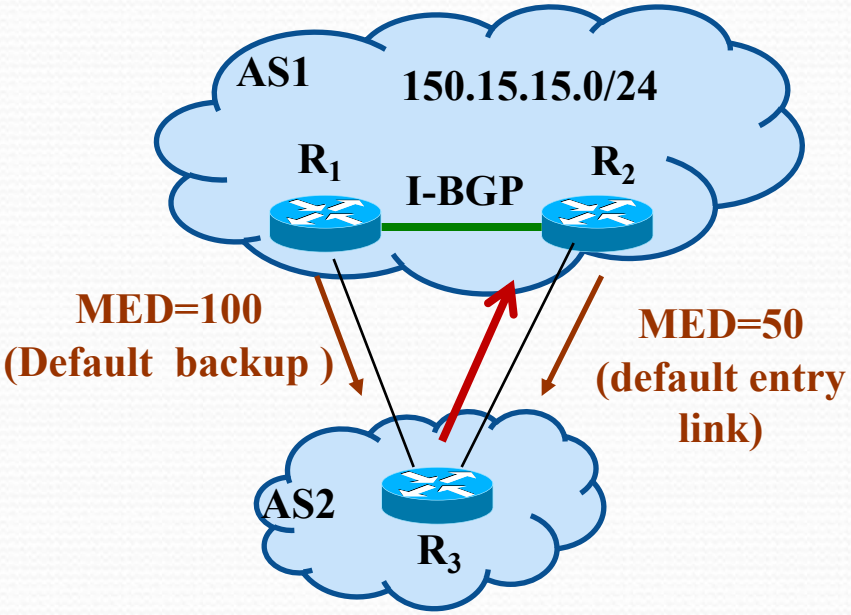


Inbound traffic control: I choose which is the entry link

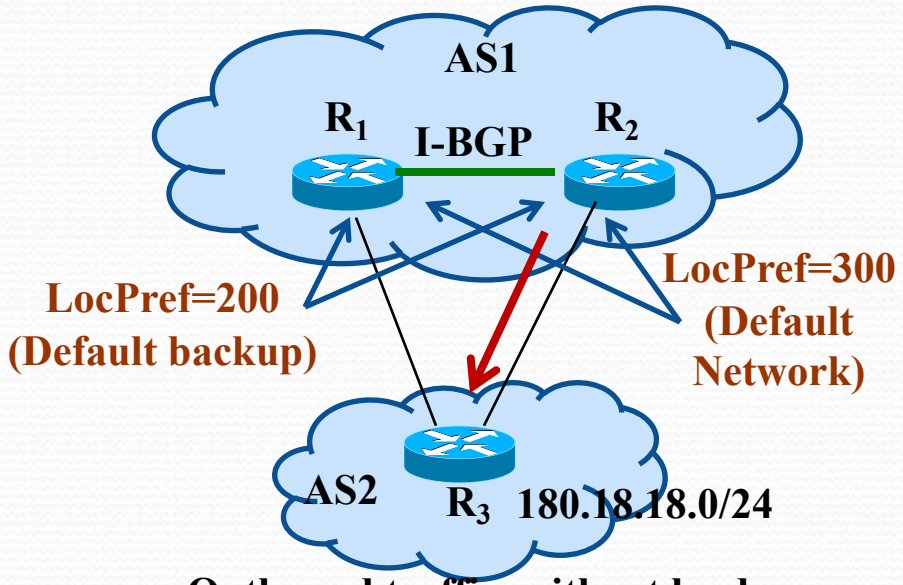


Outbound traffic control: I choose the output link

- **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)

- **BGP 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 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@

- **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

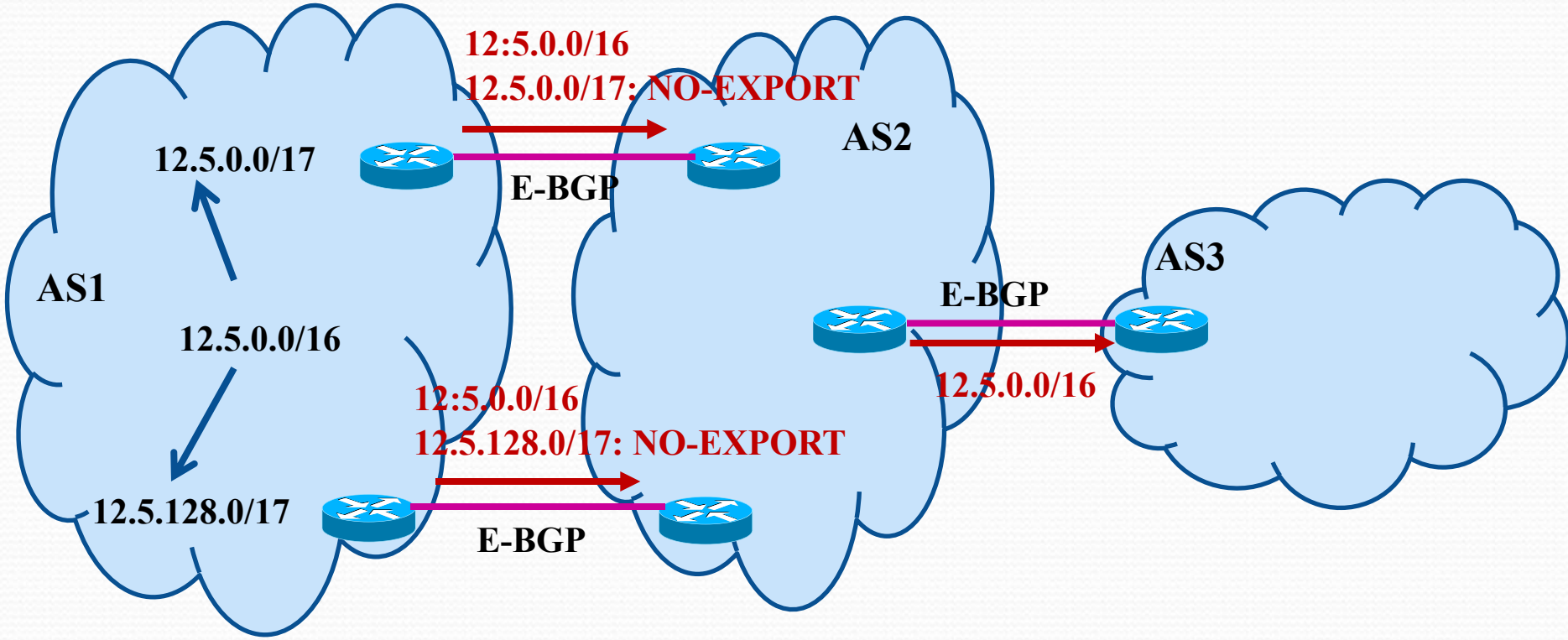
      AS:value (decimal)
  - Communities reserved: **0x00000000 to 0x0000ffff** (**0:value**) and **0xffff0000 to 0xffffffff** (**65535:value**)
  - The rest may be freely used: from **1:0 to 65534:65535**

Use of Communities

- It is a “signal” or “flag” to indicate to other BGP routers that they have to execute some action previously agreed between a set of AS’s,
- Assign prefixes to pre-defined groups
  - e.g. standard NO-EXPORT or NO-ADVERTISE
- Control how prefixes are advertised by peers
  - Control your neighbors LOCAL-PREFERENCE for the specific prefix
  - Signal neighbor to prepend multiple ASNs to AS_PATH
  - Blackhole all traffic to specific prefix

- **Standard Communities:**
  - Three standard communities (RFC 1997)
    - **NO_EXPORT (65535:65281):** all the received routes with this attribute SHOULD NOT be advertised out of the AS
    - **NO_ADVERTISE (65535:65281):** all the received routes with this attribute SHOULD NOT be advertised to other BGP neighbors (inside the same AS)
    - **NO_EXPORT_SUBCONFED (65535:65281):** all the received routes with this attribute SHOULD NOT be advertised to external BGP routers (from other confederation)

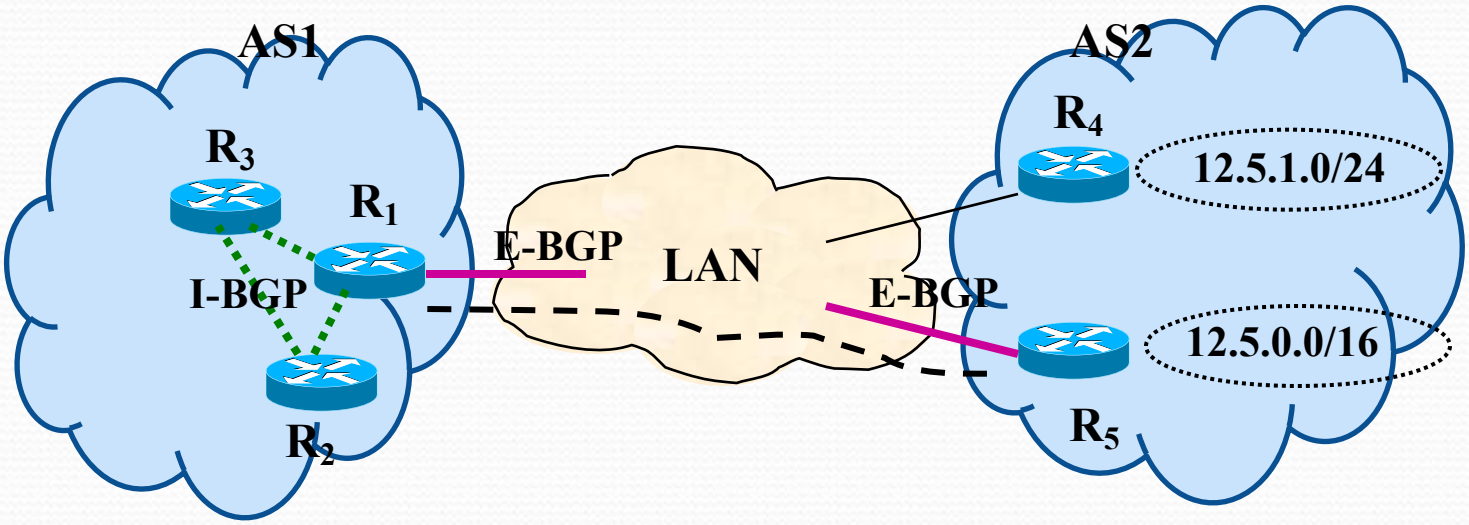
- **Standard Communities:**
  - **Example: NO-EXPORT community**
    - AS1 wants to perform load balancing with AS2 with subnets 12.5.0.0/17 and 12.5.128.0/17
    - AS3 does not need to receive the 2 routes (it is enough /16). Thus, AS1 exports to AS2 the /17 with the NO-EXPORT community and the /16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



Standard Communities:

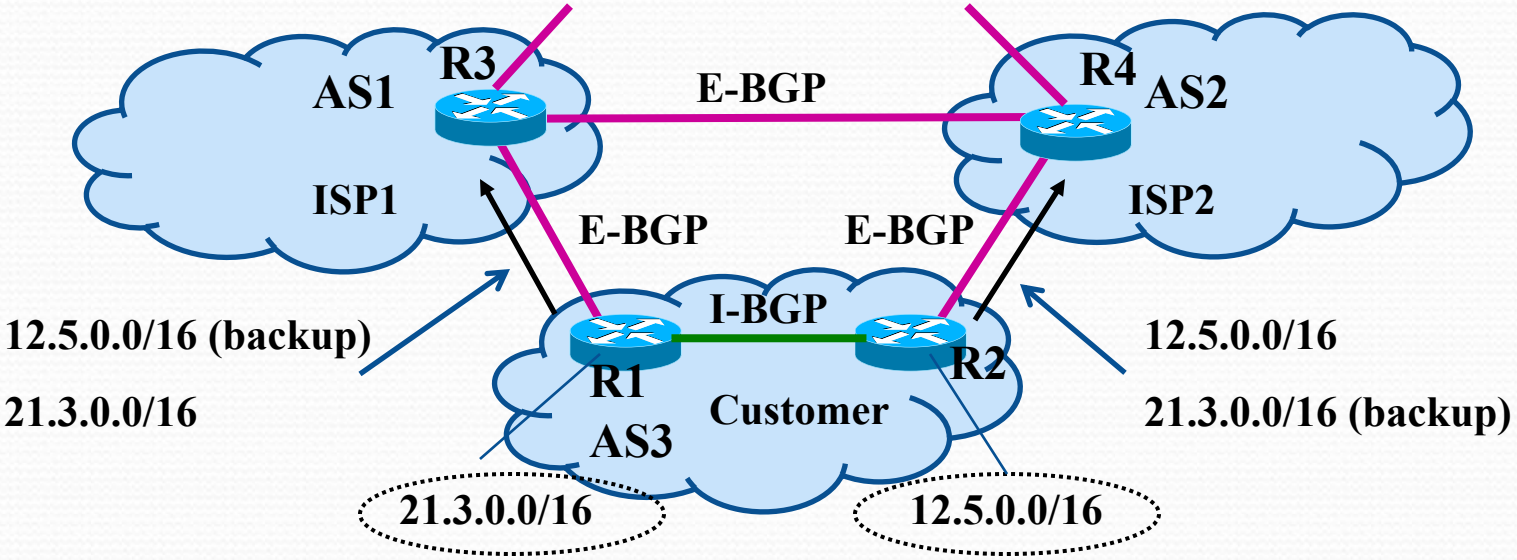
Example: NO-ADVERTISE community

- 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

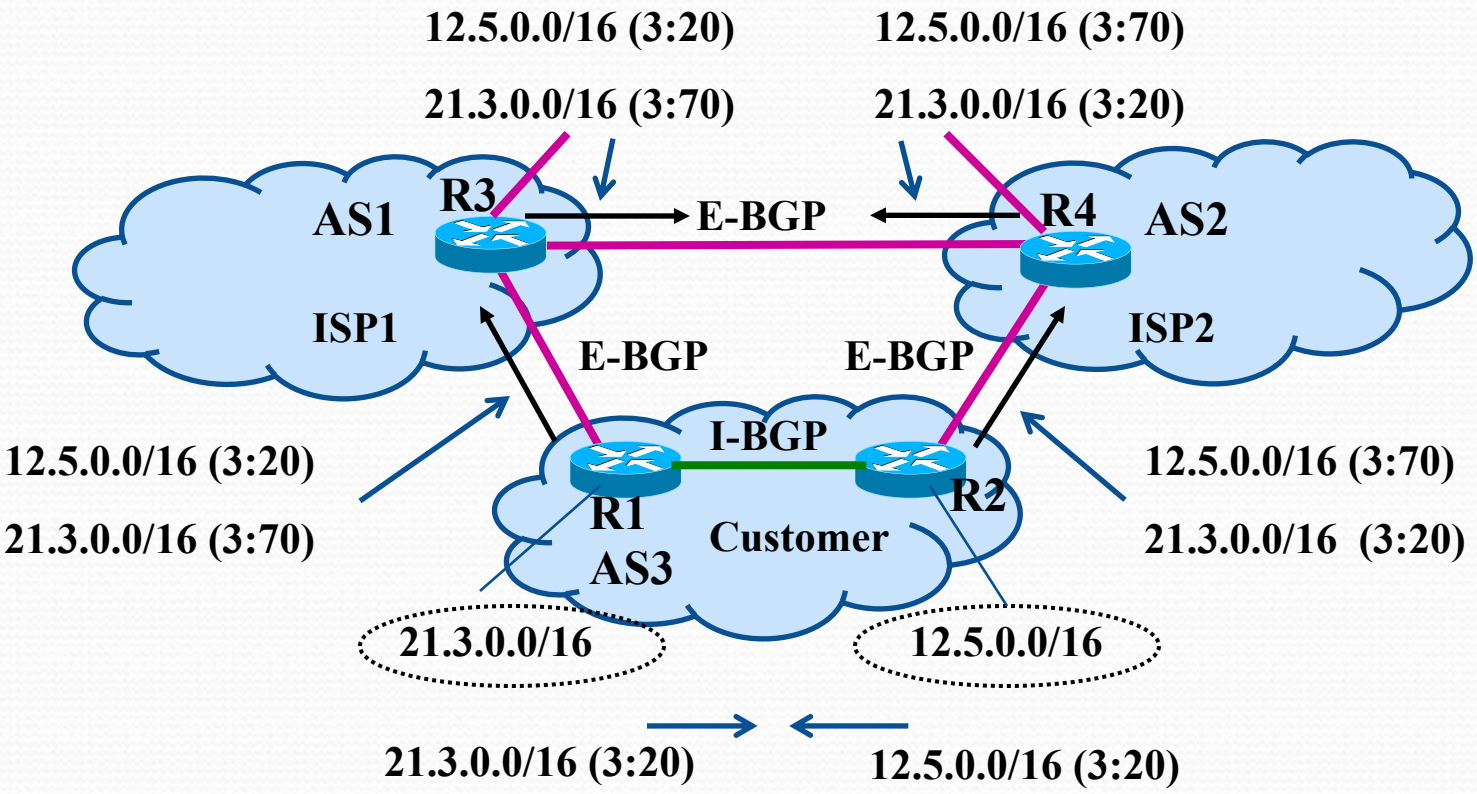


- **Other uses of Communities:**
  - Signal using a coded information. E.g.,
    - Community 325:**5**6**7**8 (customer information)
      - Field #1, **Value 5** (Type of Relationship)
      - Field #2, **Value 6** (Continent Code)
      - Field #3, **Value 7** (Region Code)
      - Field #4, **Value 8** (POP Code) → POP (Point of Presence) is an access point from one place to the rest of the Internet

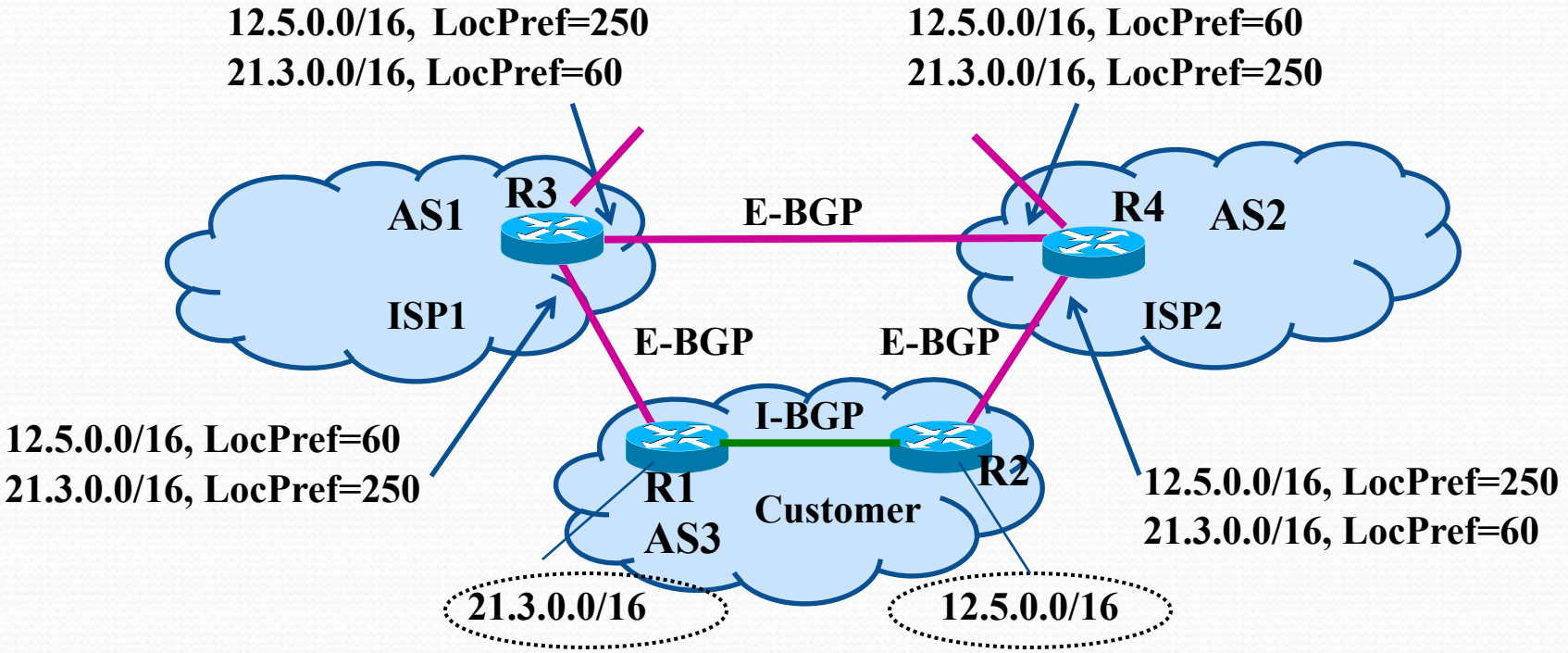
- **Communities : automatic back-up routes in multi-homing**
  - AS3 wants multi-homing with ISP1 and ISP2. Traffic towards network 12.5.0.0/16 enters via ISP2 and traffic towards network 21.3.0.0/16 enters via ISP1, but both connections act as backup with respect the other network (they can not use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



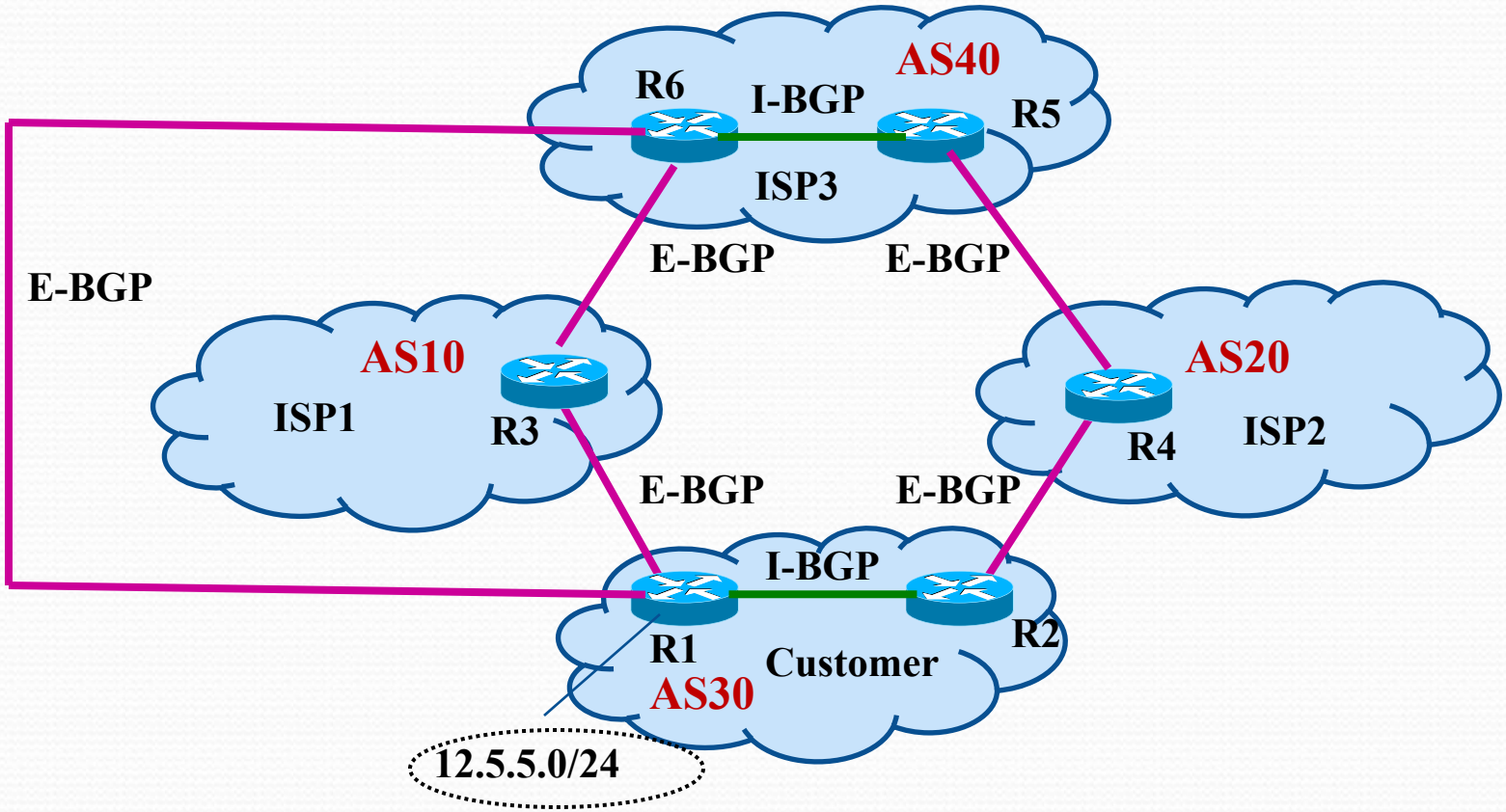
- **Communities: automatic back-up routes in multi-homing**
  - AS1 and AS2 react to community 3:20 activating LocalPref=60
  - AS1 and AS2 react to community 3:70 activating LocalPref=250



- **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# .

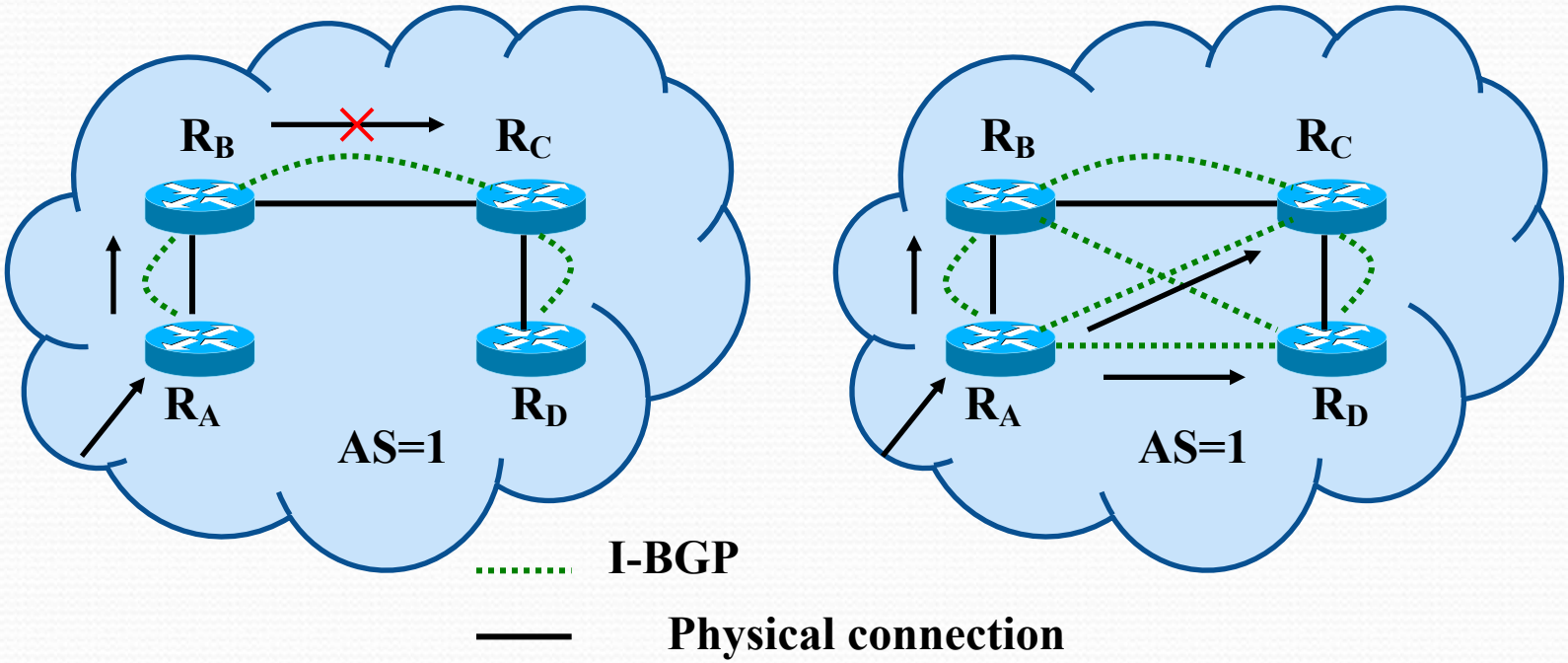
```
router bgp 30
neighbor 1.1.1.1 remote-as 40
neighbor 2.2.2.2  route-as 10
neighbor 3.3.3.3  route-as 30
network  12.5.5.0/24
neighbor 1.1.1.1 send-community
neighbor 2.2.2.2 send-community
neighbor 1.1.1.1  route-map Peer-R6 out
neighbor 2.2.2.2  route-map Peer-R3 out
!
route-map Peer-R6 permit 10
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 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, and avoiding loops,

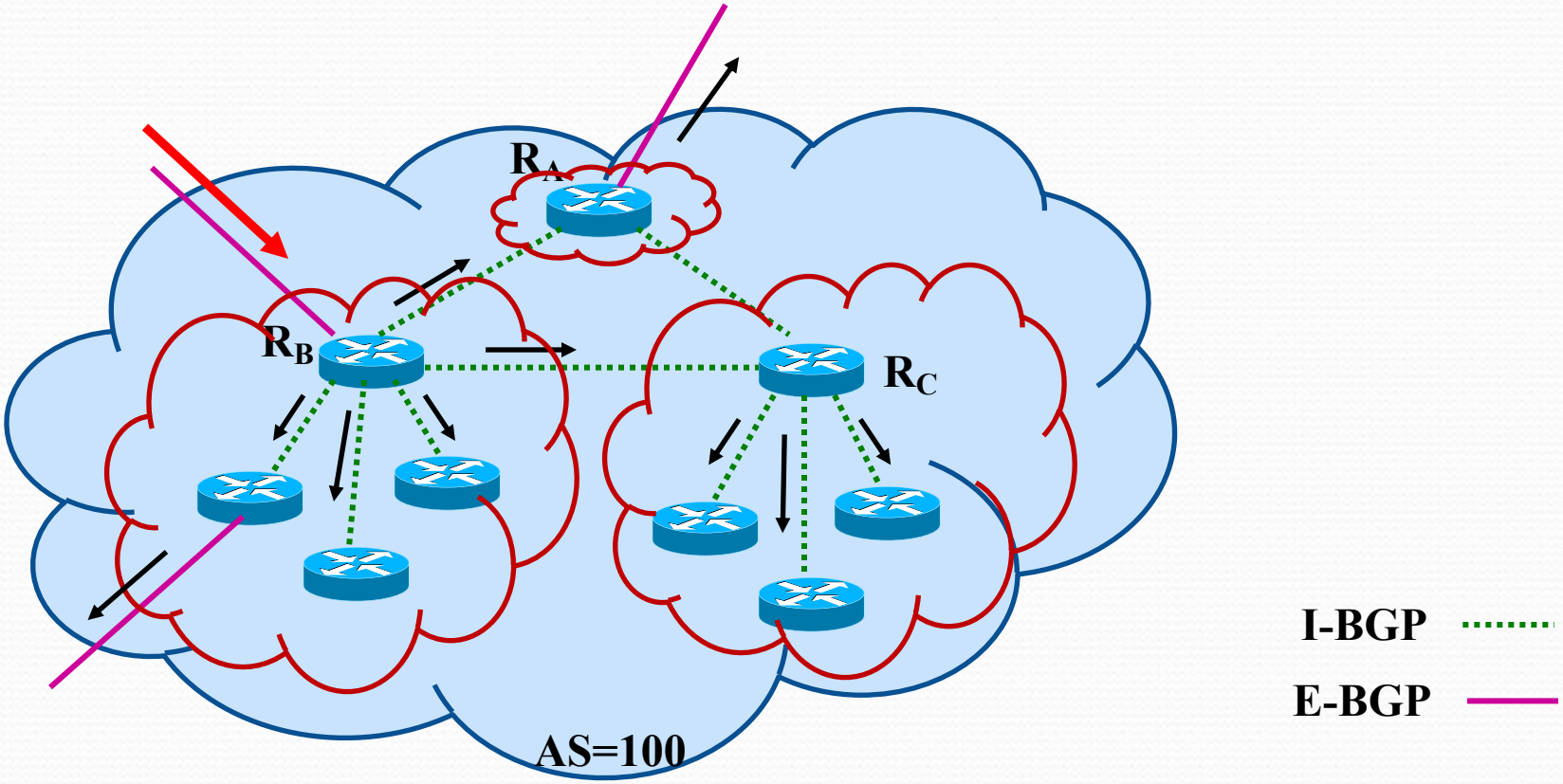
  - **Clusters** are used to define the network
    - A **route reflector** acts as **cluster-head**
    - Each route reflector maintain I-BGP sessions with its **customers** (routers that belong to the cluster)
    - The route reflectors should form a mesh network between them, but customers do not need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

Route Reflectors

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**
  - 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 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,…, 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:

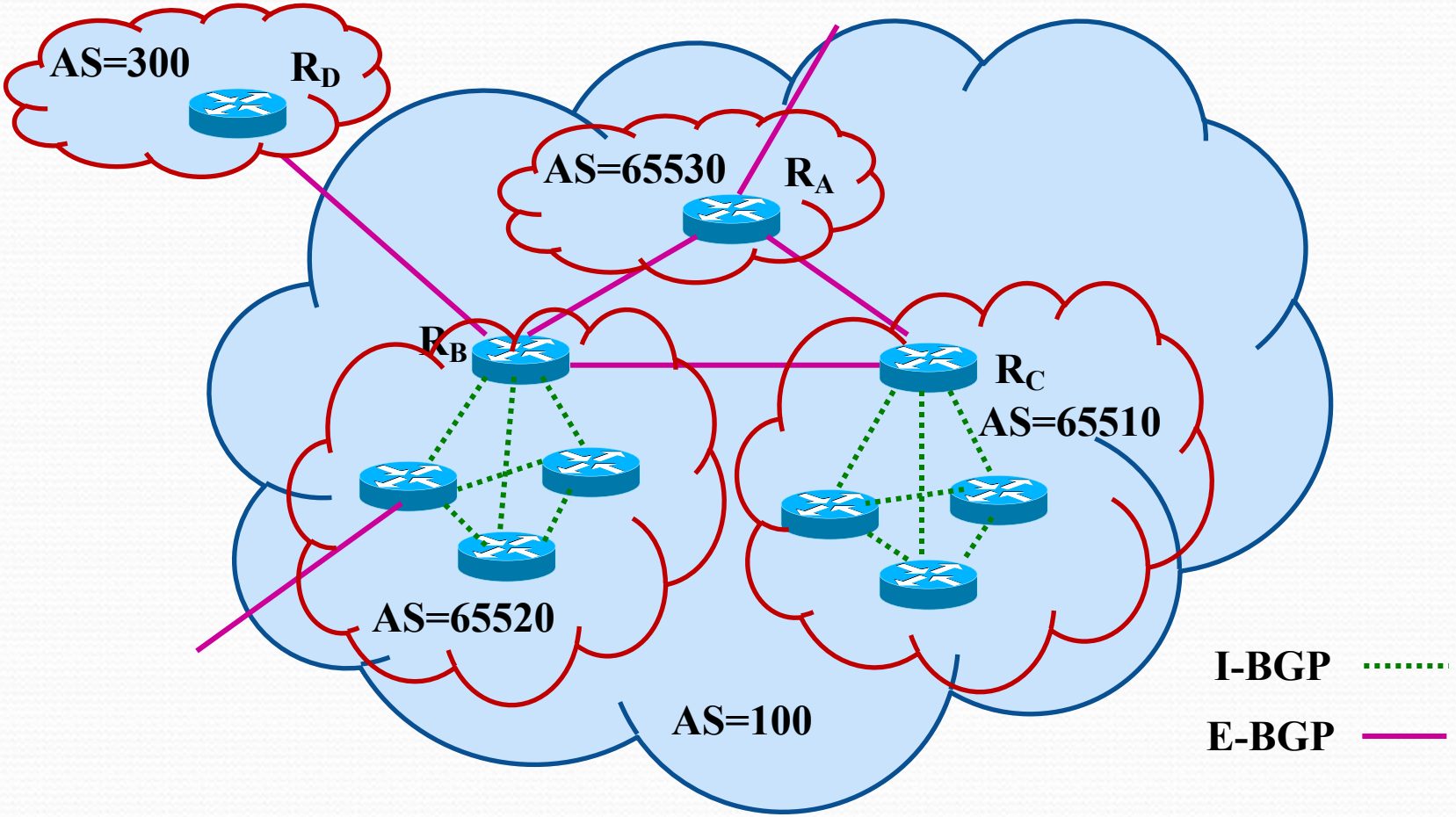
$$I-BGP = \sum_{i=1}^{N_R} NR_i + \frac{N_R(N_R-1)}{2}$$

    - 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:

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

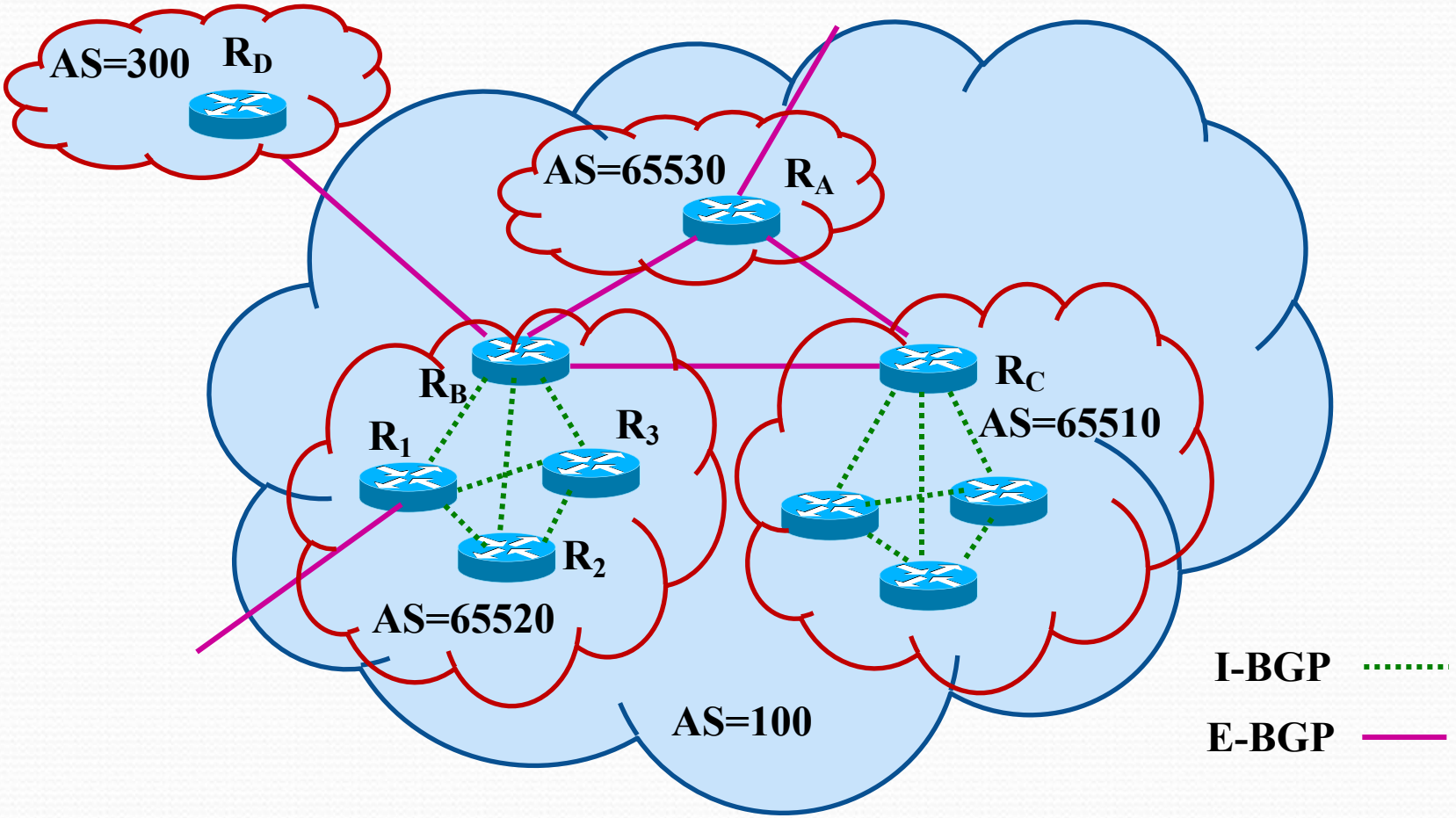
    - Then, the number of BGP sessions that have to be configured is:

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

    - 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

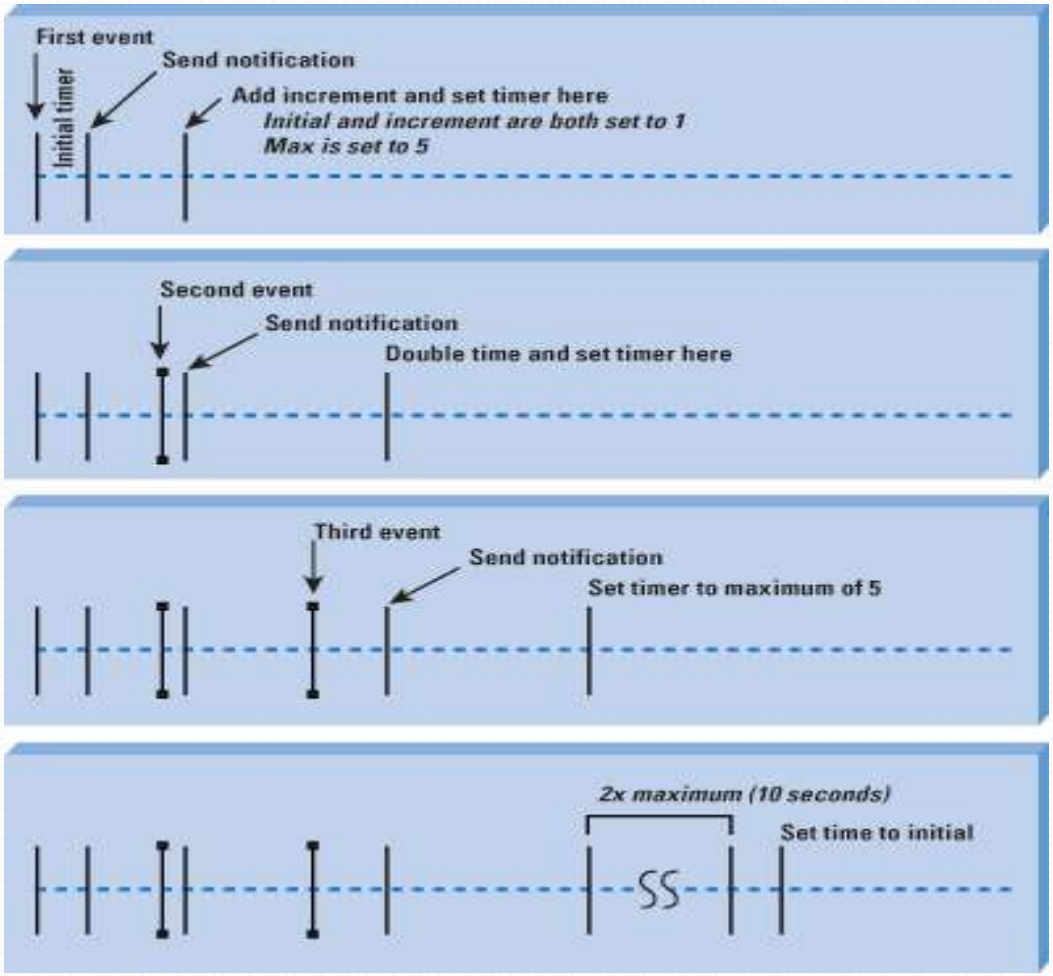
- **BGP convergence:**
  - **Flapping:** a link changes constantly from one state to other (up and down), provoking updating of messages and thus low network convergence, loops and network failures, "**meltdown**"
  - Solutions:
    - L2 has to wait before announcing that the link has failed L3 ("debouncing 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**

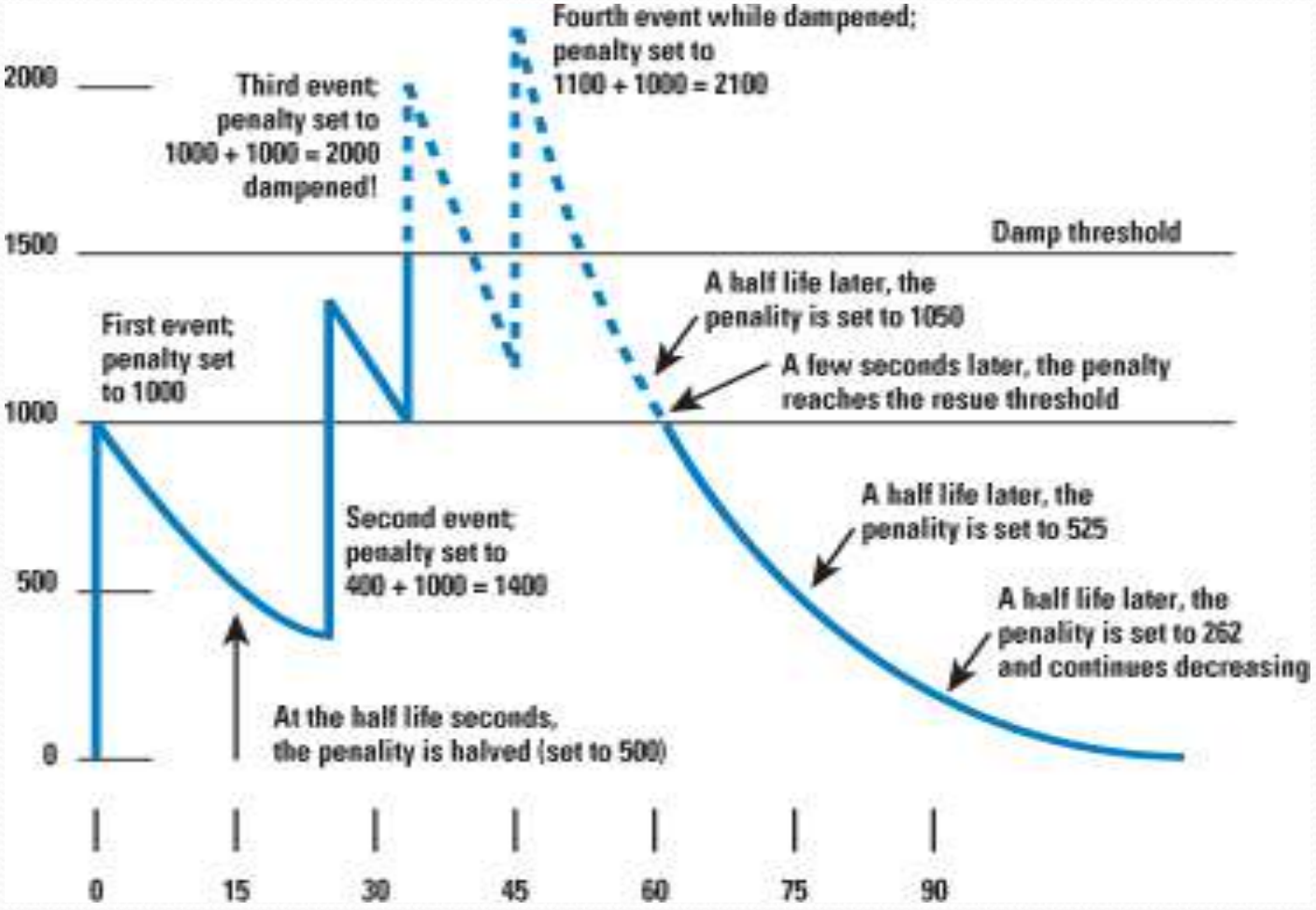
→ **Tout=5 s**

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 → dampening threshold
  - Reuse Limit: 750 → reuse threshold
  - 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

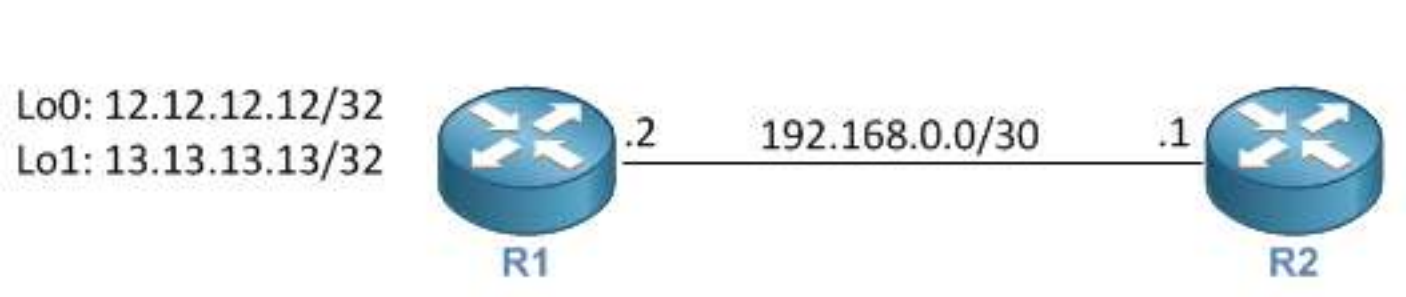
R1(conf)# **route-map** damp-R1 **permit** 20

!

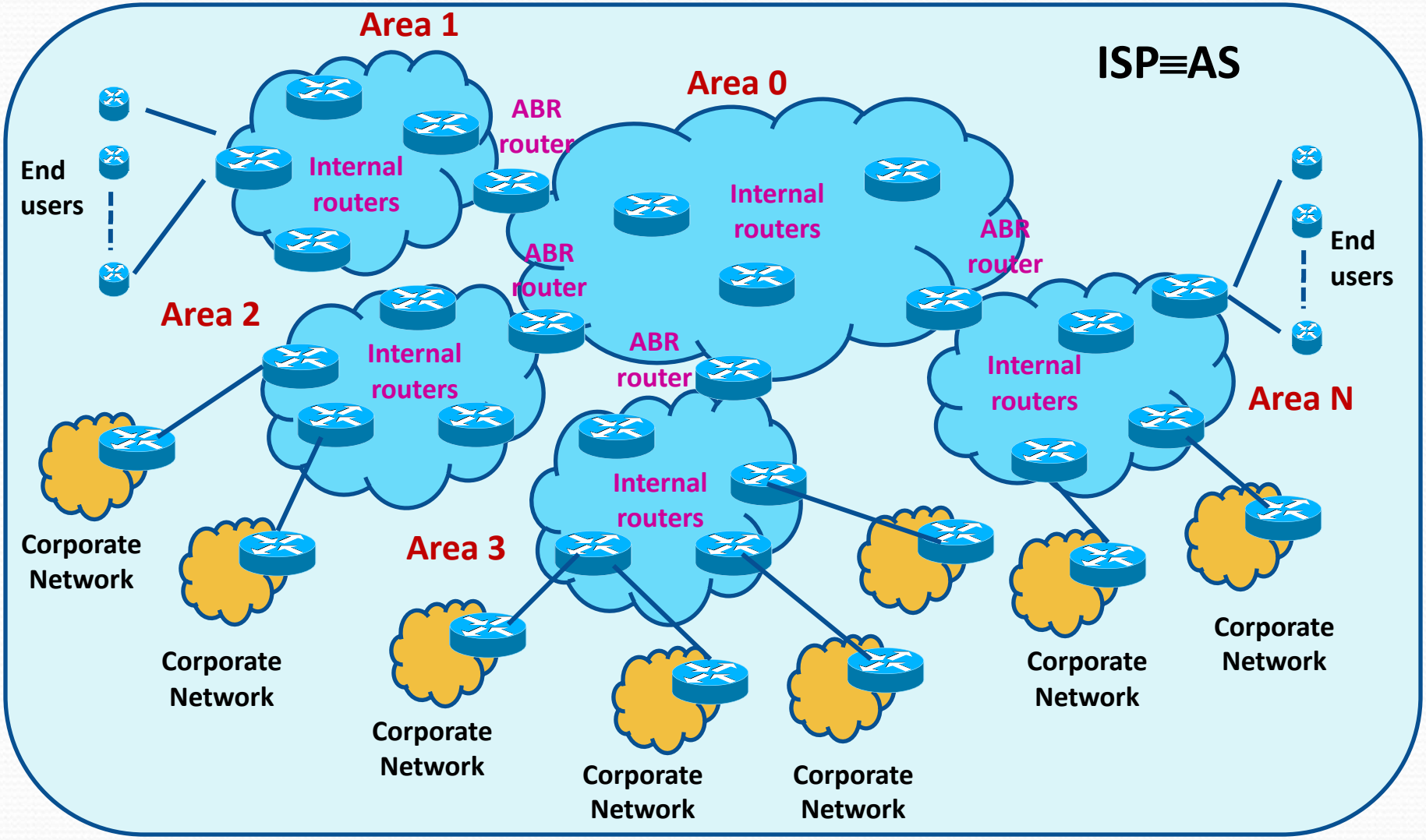
R1(conf)# **router** bgp 200

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