Interdomain routing

BGP-4 decision process

Enzo Mingozzi
Professor @ University of Pisa
enzo.mingozzi@unipi.it

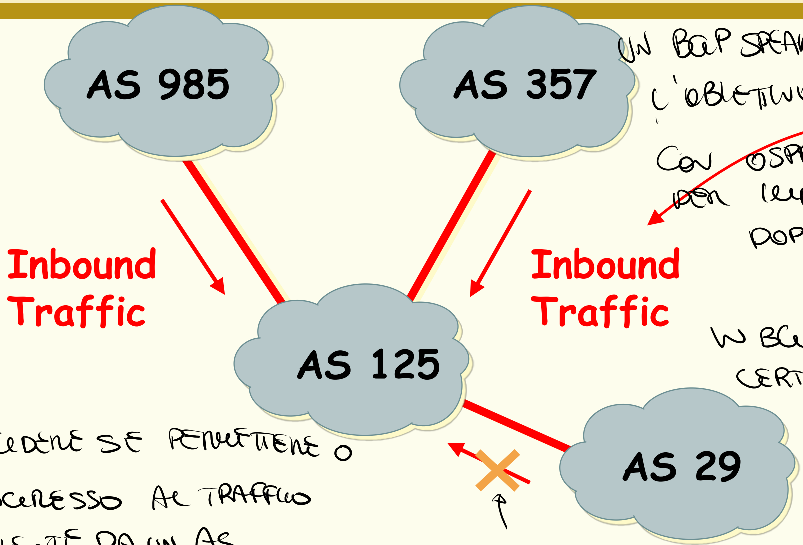
Route Filtering

UN BGP SPEAKER RICEVE BGP ROUTES, COME PROCESSA QUESTE INFORMAZIONI?
L'OBIETTIVO E' IMPARARE COME RAGGIUNGERE CERTE DESTINAZIONI_

CON OSPF O IS-IS CERCO DI USARE TUTTE LE INFORMAZIONI DISPONIBILI PER IMPARARE TUTTE LE ROTTE POSSIBILI VERSO TUTTE LE DESTINAZIONI_ DOPO FACCIO UNA VISITA COMPLETA DEL GRAFO CON DIJKSTRA_

IN BGP FACCIO UNA DECISIONE SU SE USARE O MENO UNA CERTA INFORMAZIONE_ FILTERING

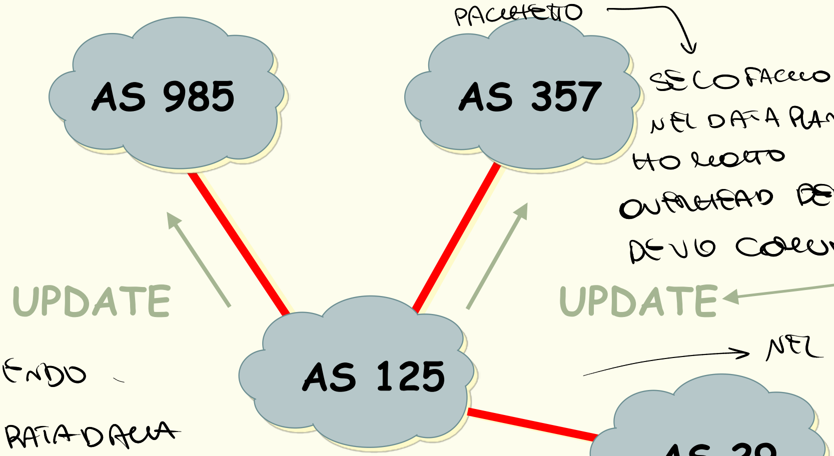
HO INFORMAZIONI MA NON LE PROPAGO ALLO SPEAKER



PUÒ DECIDERE SE PERMETTERE O NO L'INGRESSO AL TRAFFICO PROVENIENTE DA UN AS_

SCARTO IL PACCHETTO

SE LO FACCIO NEL DATA PLANE HO MOLTO OVERHEAD PERCHE' DEVO COMUNQUE PROCESSARE IL PACCHETTO



An autonomous system (AS 125) can identify the inbound traffic it is willing to accept from other neighbors by specifying the list of routes it advertises to its neighbors

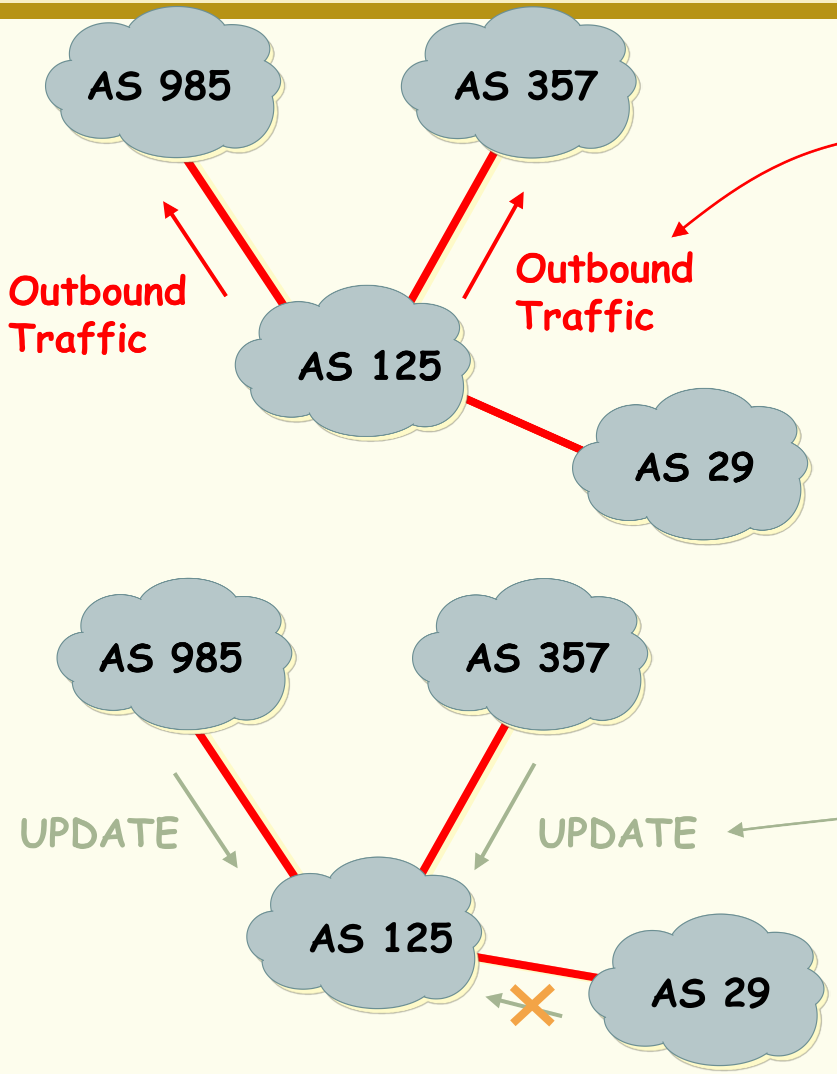
NEL CONTROL PLANE PIÙ SEMPLICEMENTE NON FACCIO ADVERTISING DELLA ROTTA_

RIMUOVENDO .

UN'ENTRATA DALLA TABELLA DI ROUTING ELIMINO UNA ROTTA

POLICY - BASED ROUTING

Route Filtering



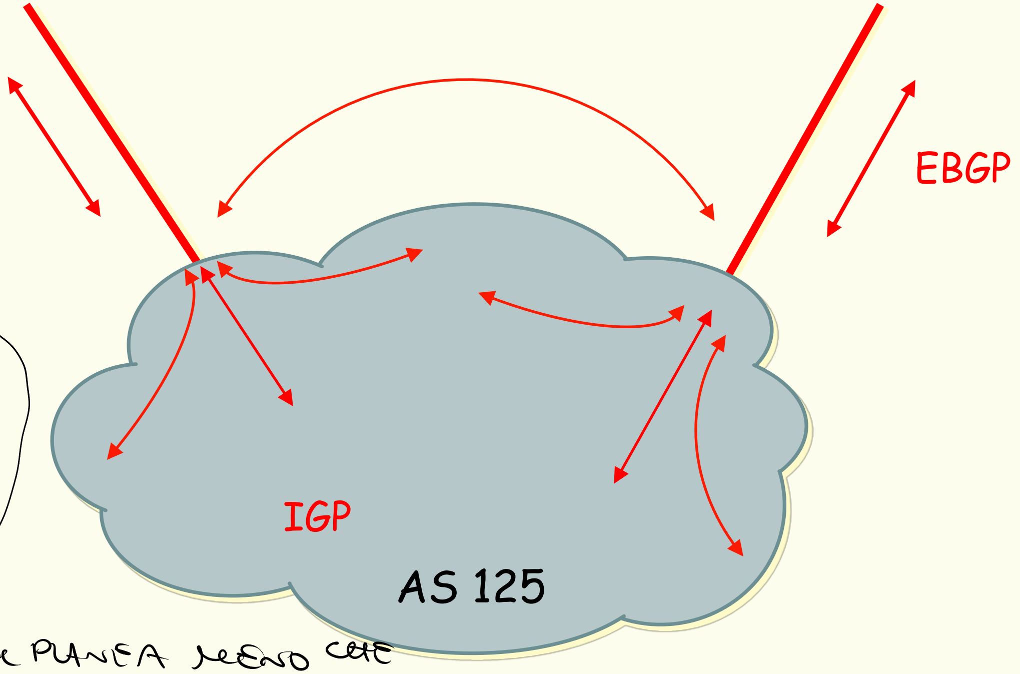
An AS can control what routes its outbound traffic uses by specifying the routes it accepts from its neighbors

Route Filtering

Filtering can also limit routing updates flowing from one protocol to another

There is the possibility of injecting BGP routes in the IGP as well as injecting the IGP or static routes into BGP

SONO USATI PER CREARE UNA FORWARDING TABLE NEL ↑ DATA PLANE



OSPF HA IL SUO DATABASE

BGP HA I SUOI DATABASE

ROTTE STATICHE IN UN FILE DI CONFIGURAZIONE

↓

RISORSE SEPARATE NEL CONTROL PLANE A MENO CHE NON SI CHIEDA AI PROTOCOLLI DI CONDIVIDERLE

BGP decision process

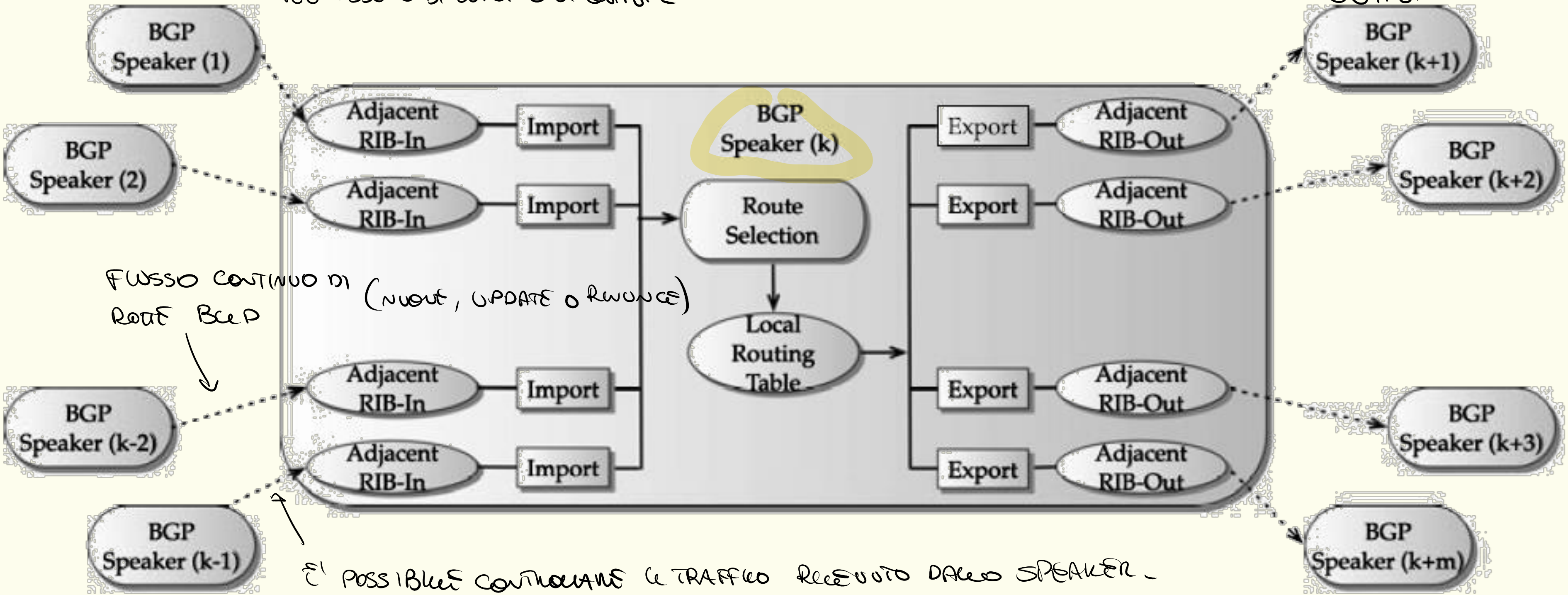
The BGP decision process consists of

1) **path selection**, and
2) **(aggregation and) dissemination** OUTPUT

IN VERITA' NON ESISTE UNA DIVISIONE DI QUESTO GENERE TRA OGNI SPEAKER PUO' ESSERE DI INPUT O DI OUTPUT.

INPUT

OUTPUT



FLUSSO CONTINUO DI ROUTE BGP (NUOVE, UPDATE O RINUNCE)

E' POSSIBILE CONTROLLARE IL TRAFFICO RICEVUTO DALLO SPEAKER.

ROUTE DINAMICHE

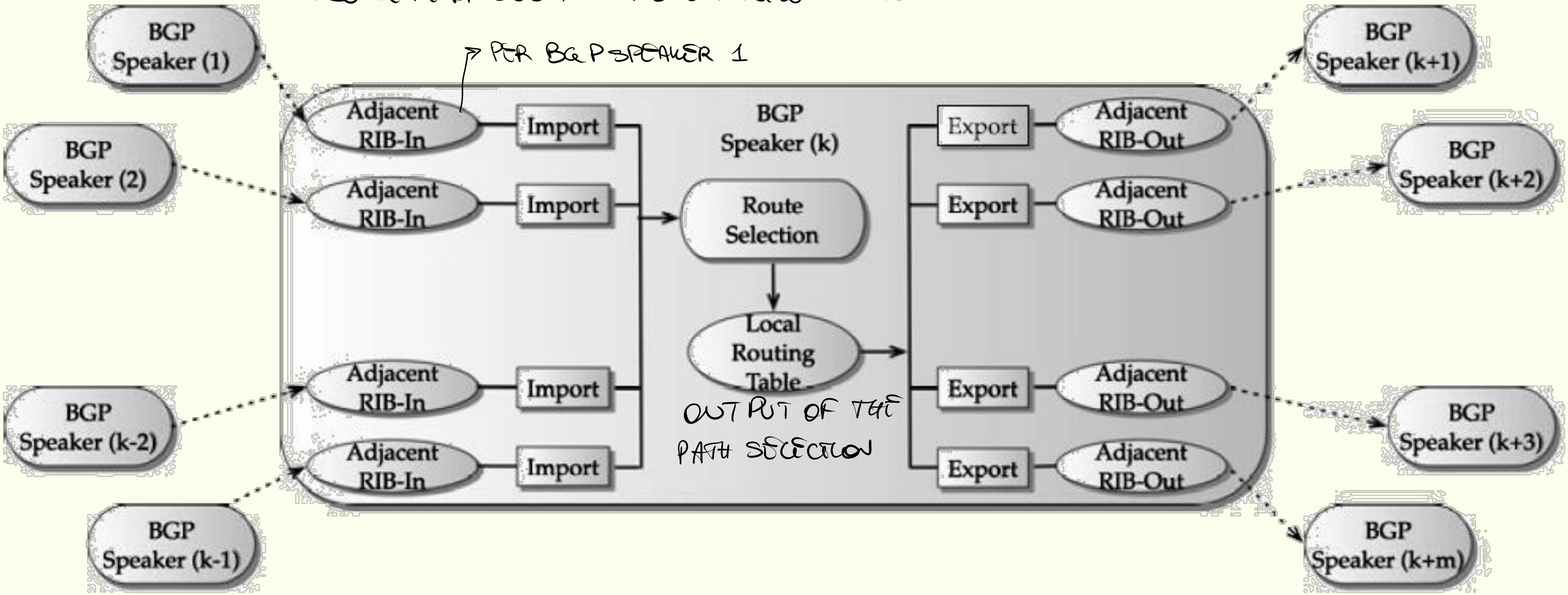
BGP decision process

Each BGP speaker maintains several **Routing Information Bases**

Adjacent RIBs-In (Adj-RIBs-In) stores AS level routing information for each IP prefix it has learned about from its neighbors through inbound UPDATE messages

QUANDO RICEVO UNA ROTTA BGP, PRENDO L'INFORMAZIONE E LA CONSERVO LOCALMENTE.

SE SELEZIONI UN PATH MA QUESTO NON E' DISPONIBILE, HO LOCALMENTE L'INFORMAZIONE CHE MI CONSENTE DI SELEZIONARE UN ALTRO PATH.



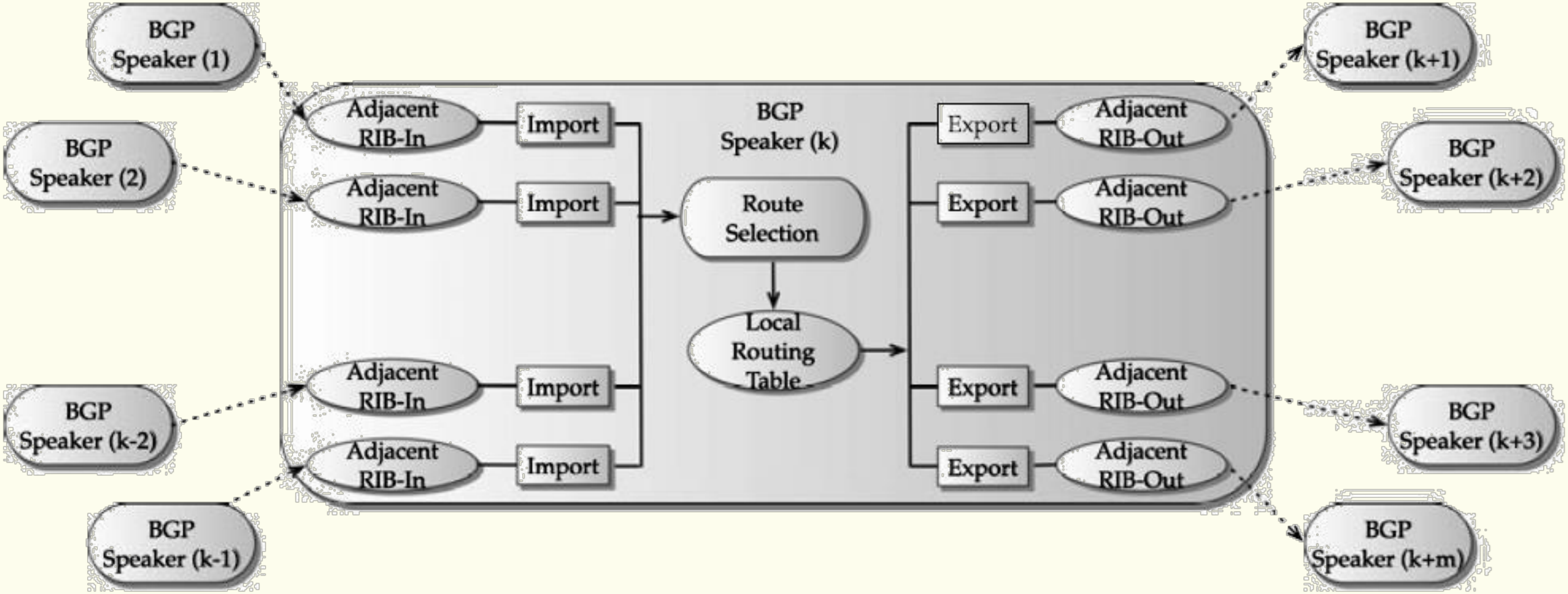
PER BGP SPEAKER 1

OUTPUT OF THE PATH SELECTION

BGP decision process

Each BGP speaker maintains several **Routing Information Bases**

Loc-RIB stores the routes that have been determined locally by the BGP speaker decision process, used for updating the forwarding table

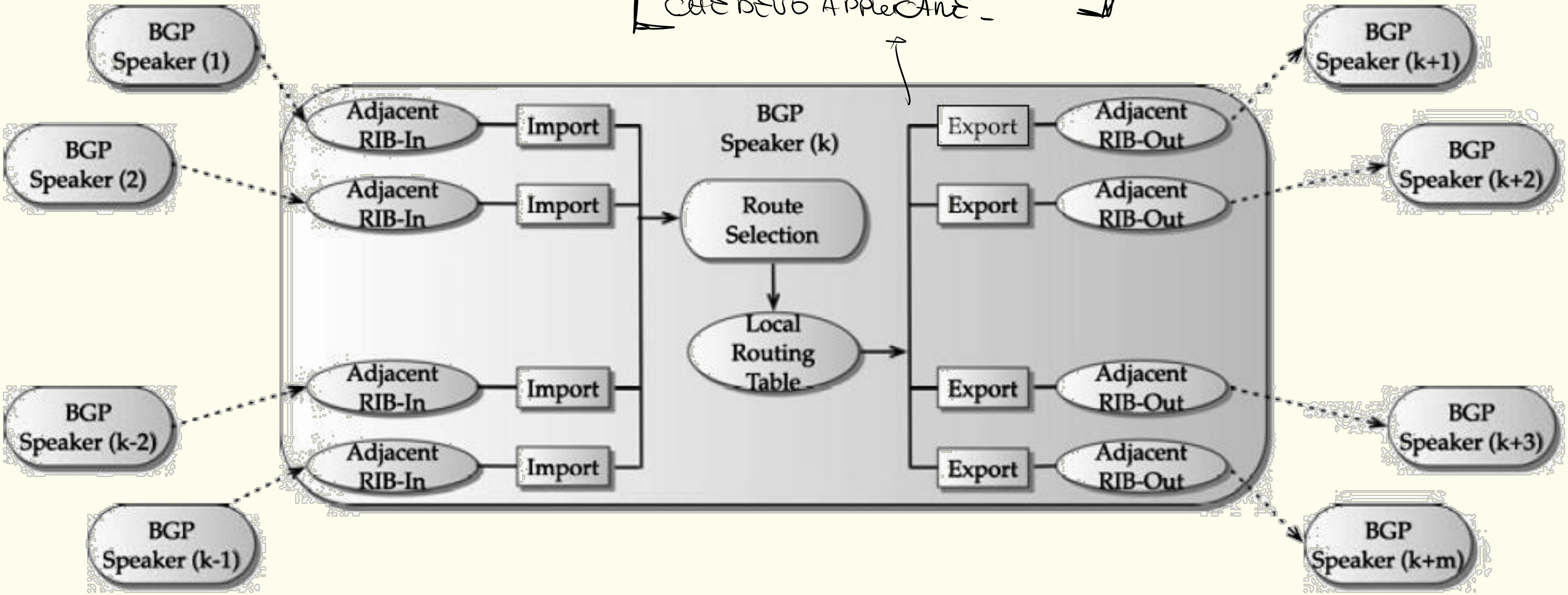


BGP decision process

Each BGP speaker maintains several **Routing Information Bases**

Adjacent RIBs-Out (Adj-RIBs-Out) stores the routes for advertisement to its neighboring BGP speakers through outbound UPDATE messages

[IN ACCORDO CON LE POLITICHE
CHE DEVO APPLICARE.]



BGP path selection

Two phases

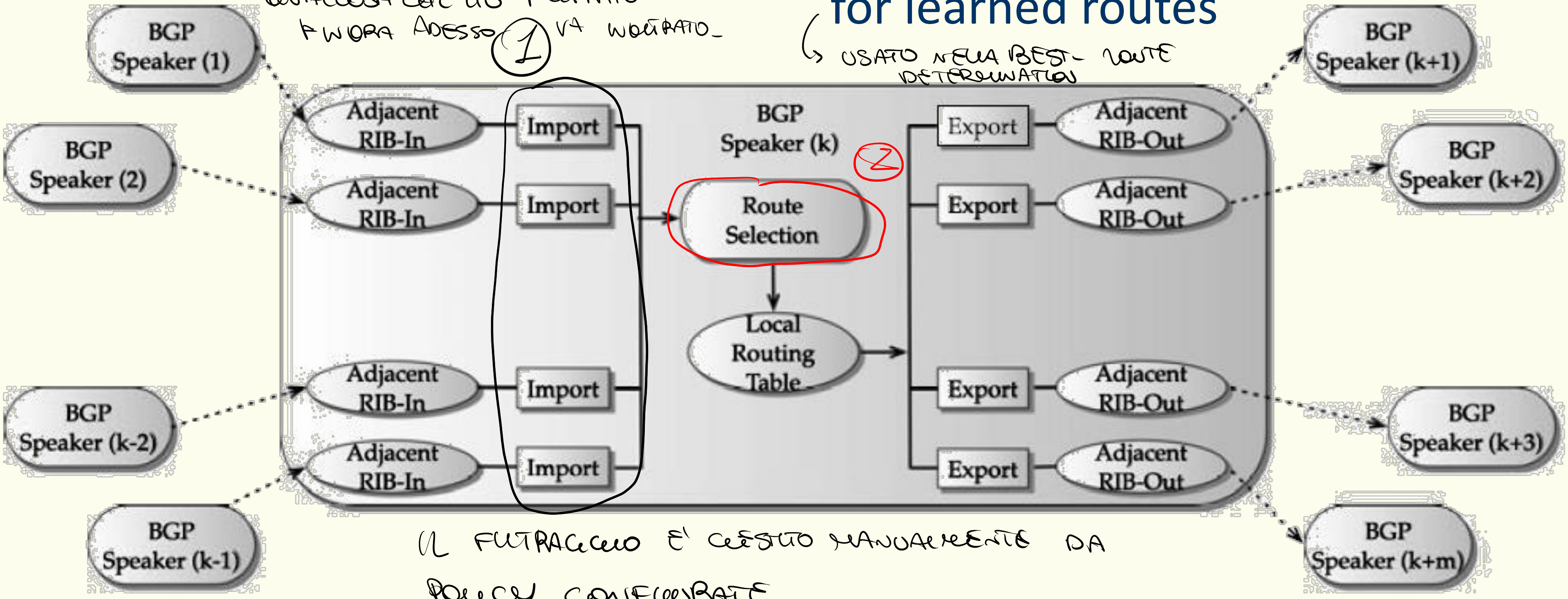
1) **Import policy and filtering**
2) Best route determination

- Filter out IP prefixes that are not allowed or that should not be reached via that peer
- Assess the degree of preference for learned routes

→ NUMERO

USATO NELLA BEST-ROUTE DETERMINATION

SE CAMBIO LE POLICY MAGARI QUALCOSA CHE HO FILTRATO FINORA ADESSO VA INOLTRATO.



IL FILTRAGGIO È GESTITO MANUALMENTE DA POLICY CONFIGURATE.

BGP path selection

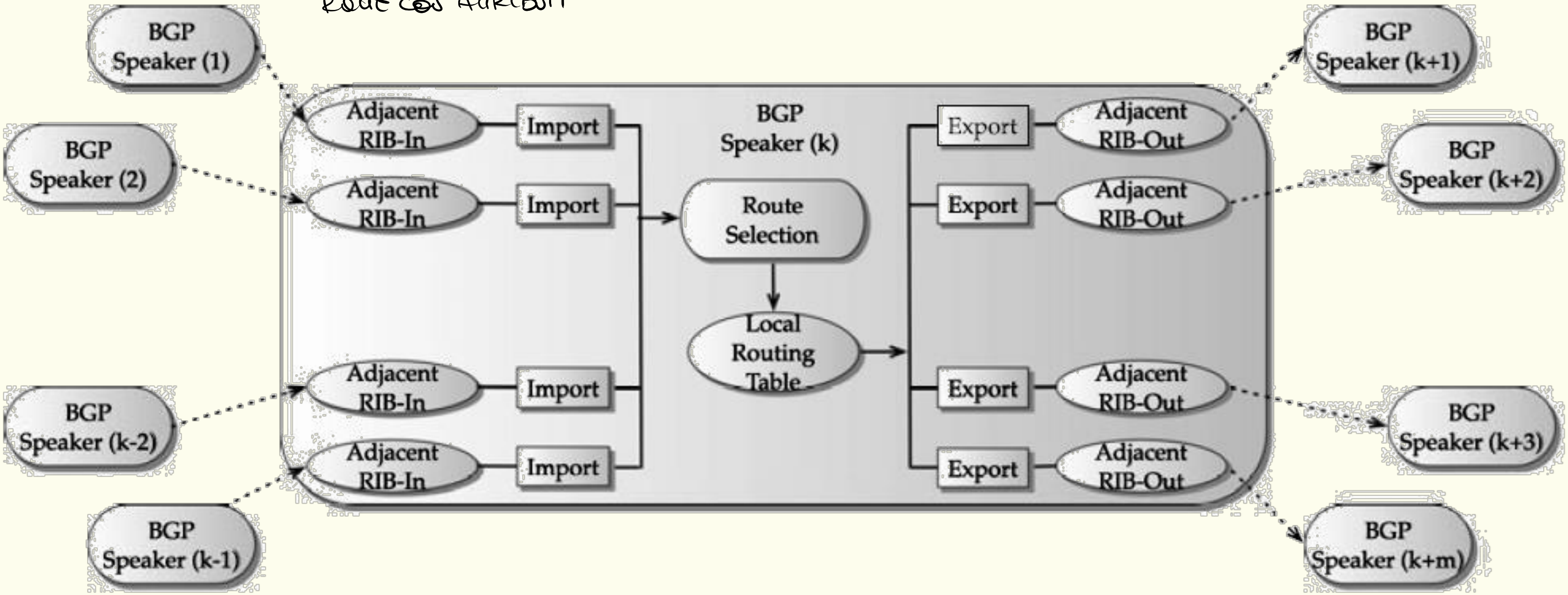
Two phases

1) Import policy and filtering
2) **Best route determination**

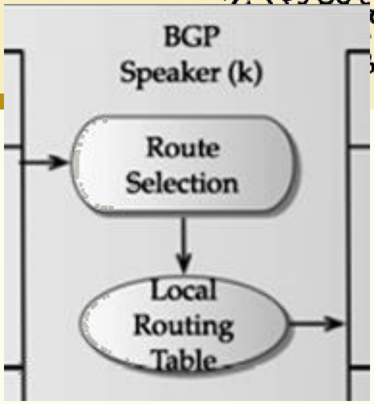
- Select the best route for each separate **imported** IP prefix

HO IMPORTATO INFORMAZIONI DA UN CERTO NUMERO DI SPEAKER ↓

ROUTE CON ATTRIBUTI



BGP path selection



- **Tie-breaking rules** when multiple routes are available to the same <u>**imported** IP prefix</u>

SE RICEVO SOLO UNA ROTTA OK, MA SE NE HO MOLTE?

ORDINE

1. Ignore routes for which the **NEXT–HOP attribute is not resolvable**
2. Apply the **degree of preference** assessed during the *import policy and filtering* phase (either on LOCAL–PREF if the announcement is received from an iBGP speaker, or any locally pre-configured decision) → PREFERISCO INTRA-AS ROUTING
3. Select the route that **originated locally** at the BGP speaker → SE E' STATA IMPARATA ATTRAVERSO REDISTRIBUZIONE ALLORA LA PREFERISCO
4. Select the route with the **shortest AS path**
5. Select the one with the **lowest ORIGIN** attribute (IGP, then EGP, then Incomplete)
6. Select the route with the **lowest MED** for eBGP routes (learned from the same AS)
7. Select the route received from **eBGP** over iBGP
8. Select the route with **shortest (internal) path to the NEXT–HOP router** (as determined by IGP)
9. Select the route learned from the eBGP neighbor with the **lowest BGP identifier**
10. Select the route from the iBGP neighbor with the **lowest BGP identifier**

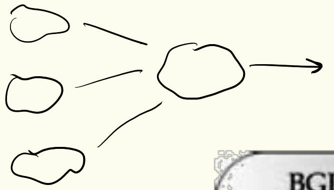
→ L'ULTIMO AS NEL PATH E' LA HOME NETWORK CON EGP HO SOLO UN SEGMENTO DEL PATH.

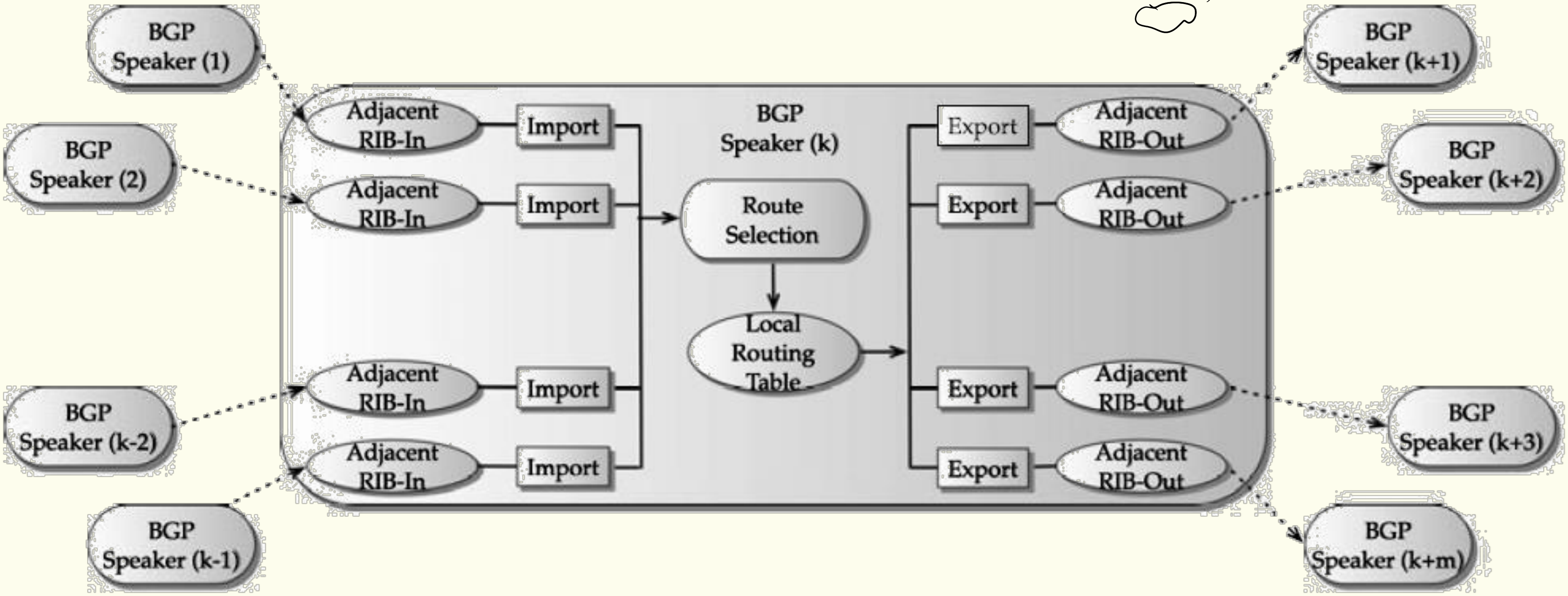
BGP route aggregation and dissemination

- *Optional* **route aggregation** based on CIDR: combine IP prefixes (*supernetting*) to reduce the number of networks announced to a downstream AS

POSSO PUBBLICIZZARE UN INSIEME DI PREFISSI IP CON UNA SOLA SUPERNET

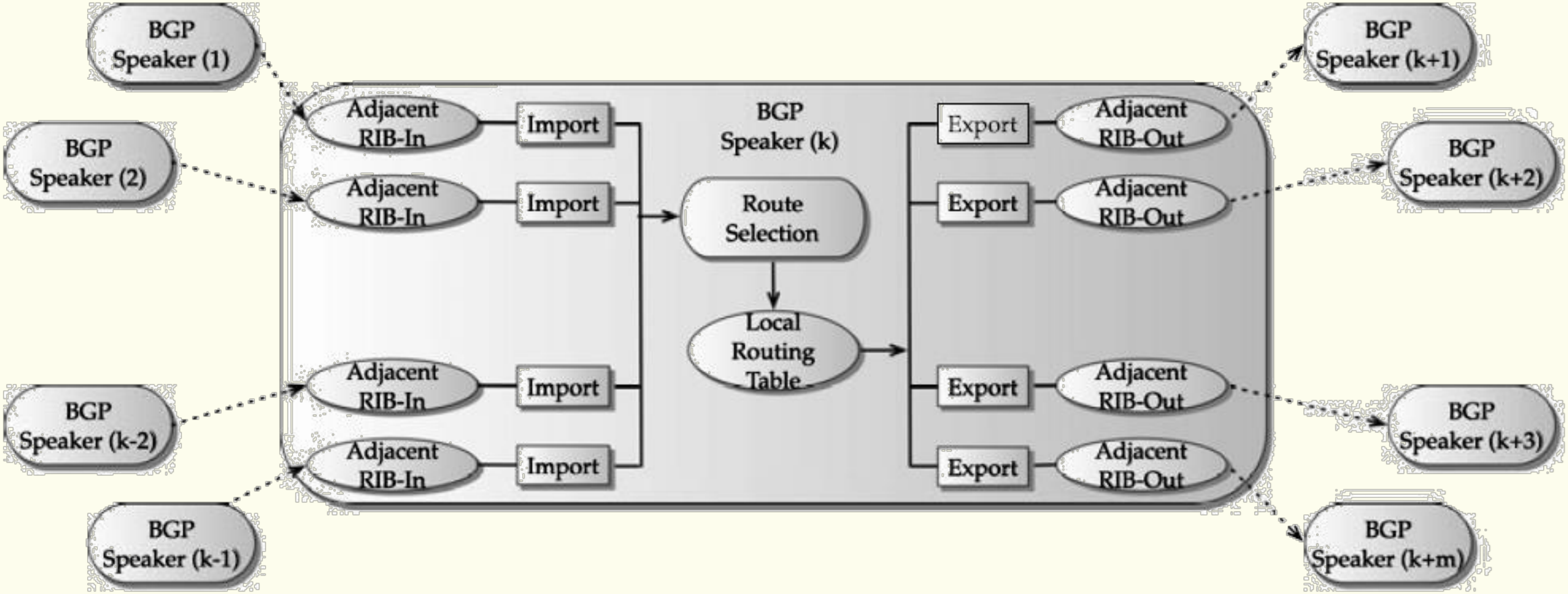
NON SOLO PER QUELLI COLLOCATI NELLO STESSO AS MA ANCHE IN AS DIVERSI





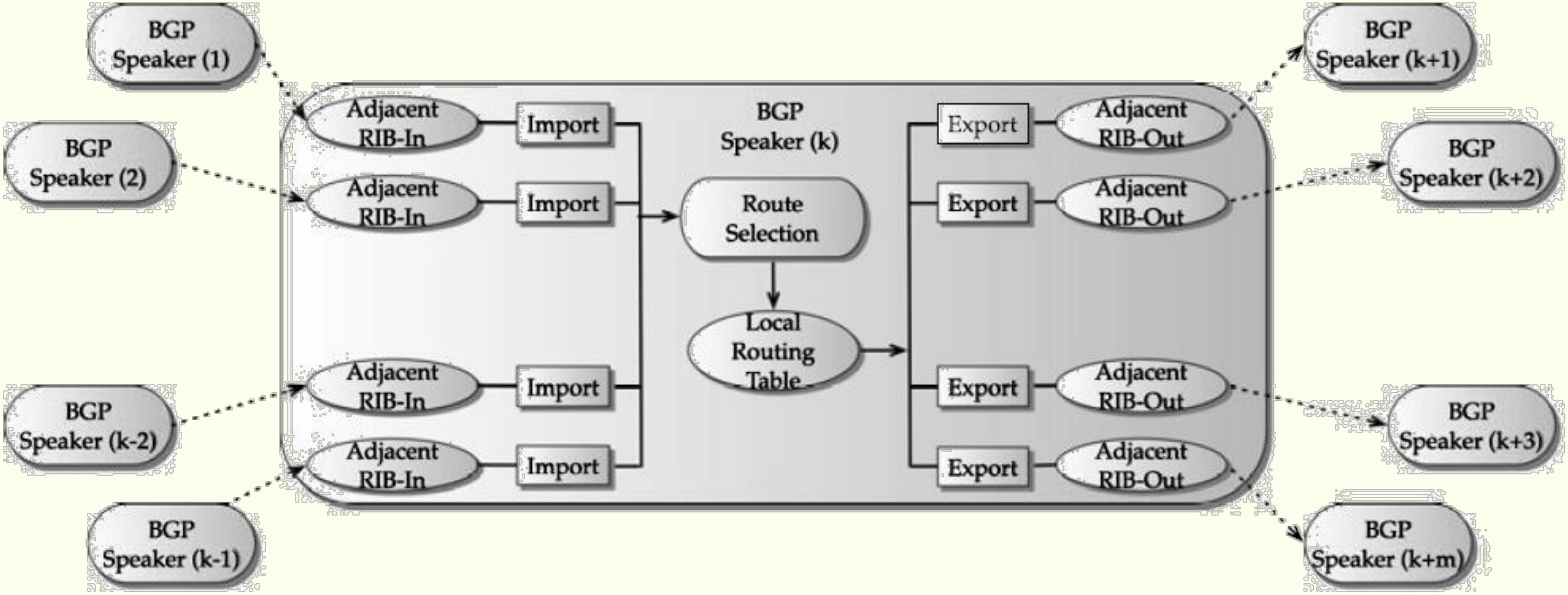
BGP route aggregation and dissemination

- A BGP speaker applies an **export policy** before propagating routes to other BGP speakers
- Export policies are separate per neighboring BGP speaker

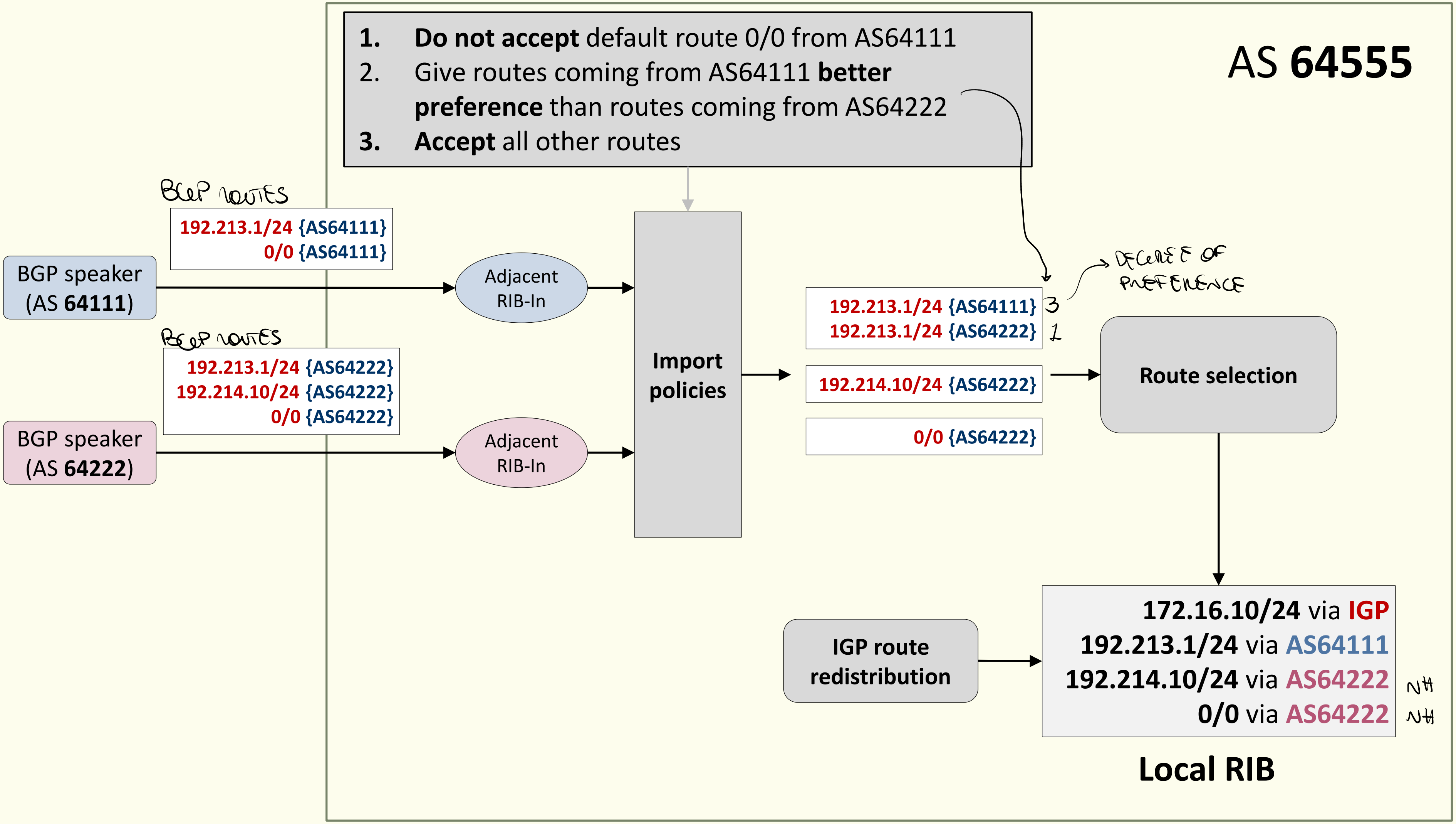


BGP decision process

- **Policy-based routing**: import and export policies are placed at a BGP speaker by a network administrator due to business relations or peering arrangement, i.e., external factors



Example



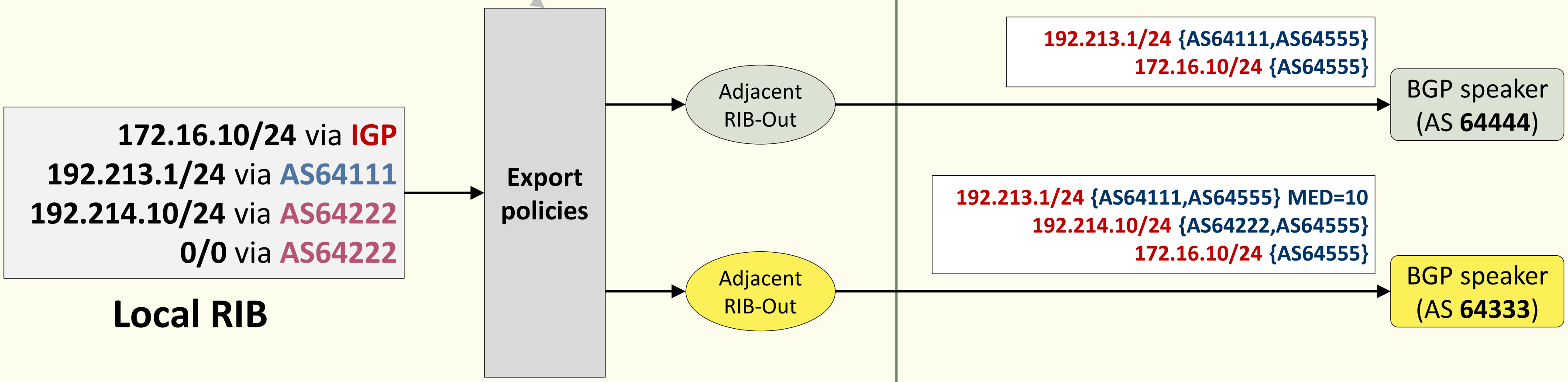
1. **Do not accept** default route 0/0 from AS64111
2. Give routes coming from AS64111 **better preference** than routes coming from AS64222
3. **Accept** all other routes

Example

AS **64555**

1. **Do not propagate** the default route 0/0
2. **Do not advertise** 193.214.10/24 to AS6444
3. Give 192.213.1/24 a **metric of 10** when sent to AS64333

MED ATTRIBUTE



172.16.10/24 via **IGP**
192.213.1/24 via **AS64111**
192.214.10/24 via **AS64222**
0/0 via **AS64222**

Local RIB

Export policies

Adjacent RIB-Out

192.213.1/24 {AS64111,AS64555}
172.16.10/24 {AS64555}

BGP speaker (AS **64444**)

Adjacent RIB-Out

192.213.1/24 {AS64111,AS64555} MED=10
192.214.10/24 {AS64222,AS64555}
172.16.10/24 {AS64555}

BGP speaker (AS **64333**)

Internal BGP scalability

ABBIAMO BISOGNO UNA FULL MESH DI SESSIONI BGP

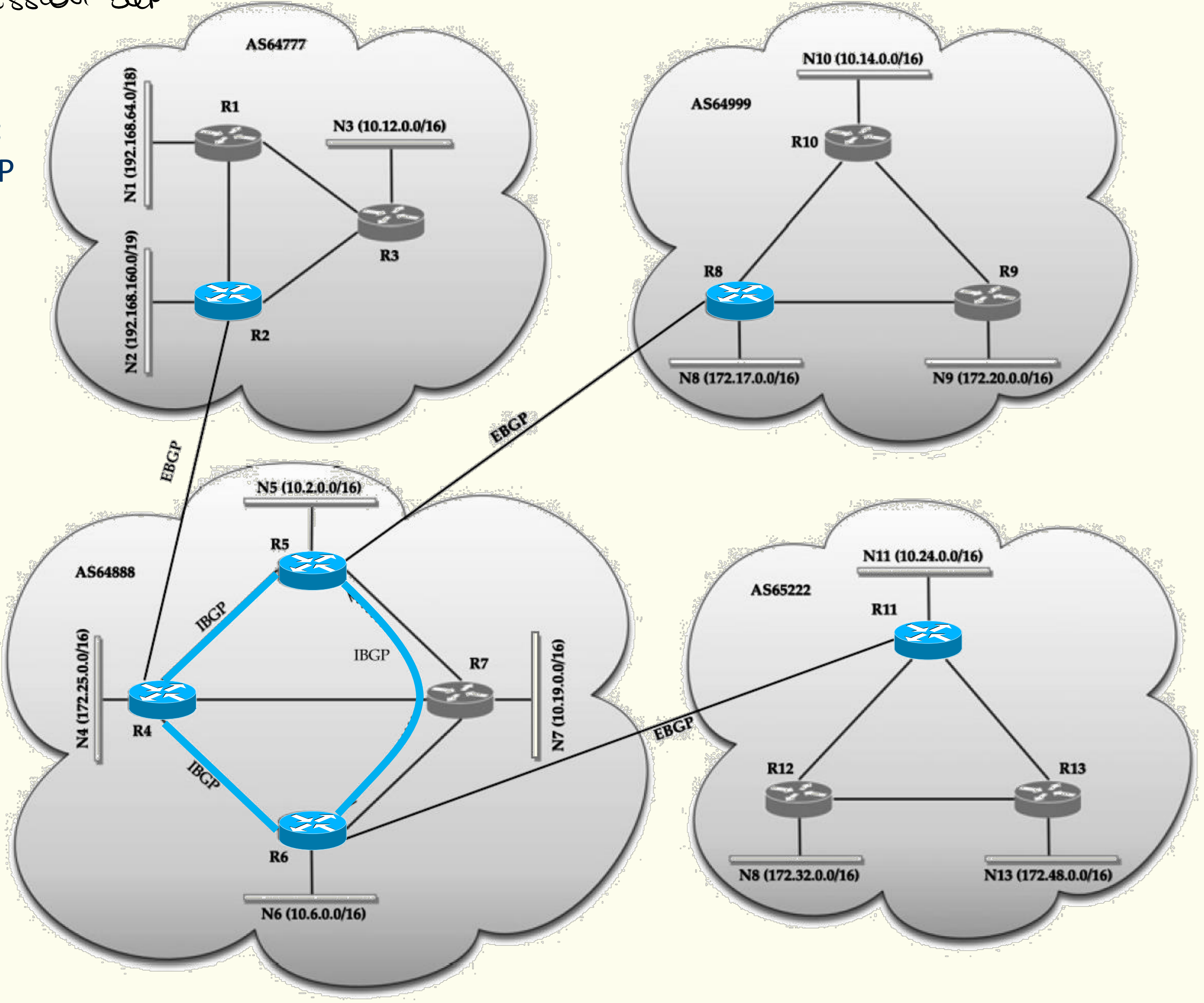
Rule 1 A BGP speaker can advertise IP prefixes it has learned from an eBGP speaker to a neighboring iBGP speaker; similarly, a BGP speaker can advertise IP prefixes it has learned from an iBGP speaker to an eBGP speaker

Rule 2 An iBGP speaker cannot advertise IP prefixes it has learned from an iBGP speaker to another peer iBGP speaker

Two reasons:

1. Avoid looping of BGP route updates within the AS
2. No need to advertise internal routes

A full mesh iBGP connectivity is needed



Internal BGP scalability

Rule 1 A BGP speaker can advertise IP prefixes it has learned from an eBGP speaker to a neighboring iBGP speaker; similarly, a BGP speaker can advertise IP prefixes it has learned from an iBGP speaker to an eBGP speaker

Rule 2 An iBGP speaker cannot advertise IP prefixes it has learned from an iBGP speaker to another peer iBGP speaker

Two reasons:

1. Avoid looping of BGP route updates within the AS
2. No need to advertise internal routes

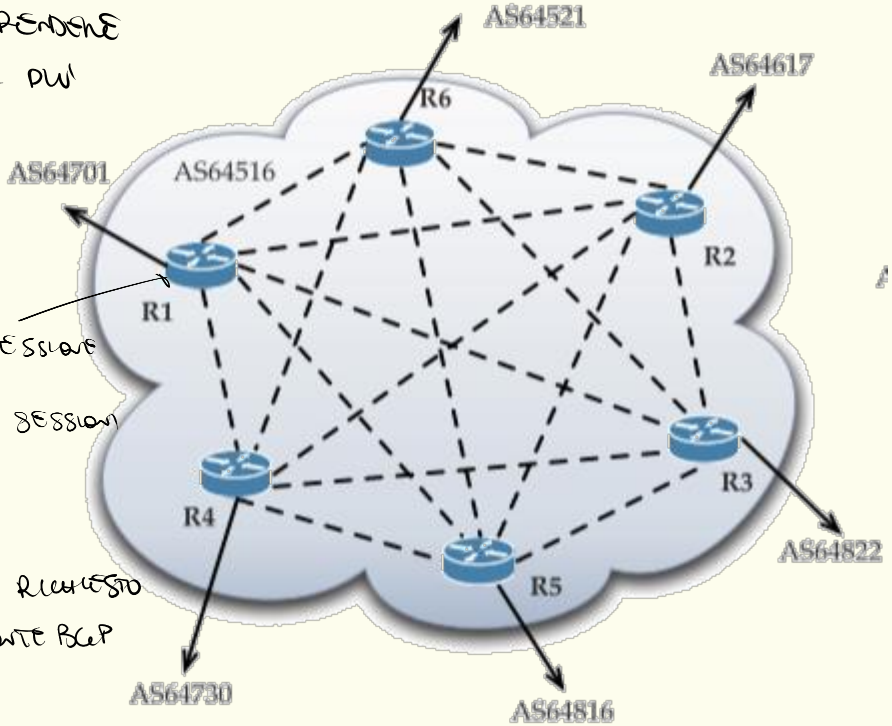
⇓

A full mesh iBGP connectivity is needed

QUADRATICAMENTE!

n iBGP speakers → $n(n-1)/2$ iBGP sessions each speaker handling $n-1$ sessions

COME POSSIAMO RENDERE QUESTO SISTEMA PIÙ SCALABILE?



ALMENO UNA SESSIONE EBGP E n-1 SESSIONI iBGP

↓

TROPPO OVERHEAD RICHIESTO PER OGNI ROUTE BGP

↳ NON POSSIAMO SEMPRE USARE UNA FULL MESH

Internal BGP scalability

Rule 1 A BGP speaker can advertise IP prefixes it has learned from an eBGP speaker to a neighboring iBGP speaker; similarly, a BGP speaker can advertise IP prefixes it has learned from an iBGP speaker to an eBGP speaker

Rule 2 An iBGP speaker cannot advertise IP prefixes it has learned from an iBGP speaker to another peer iBGP speaker

Two reasons:

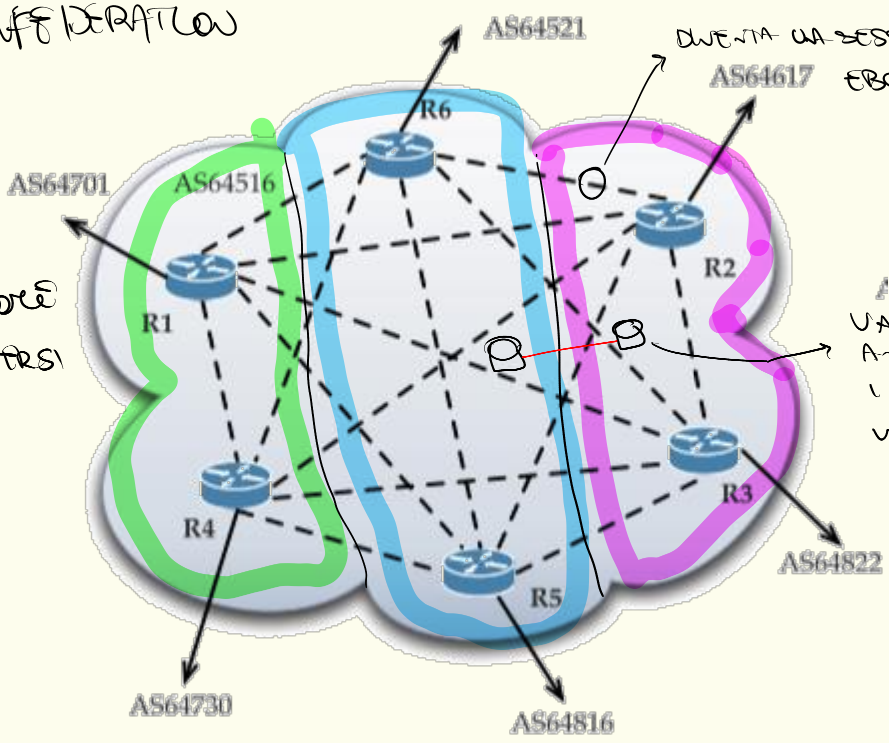
1. Avoid looping of BGP route updates within the AS
2. No need to advertise internal routes

A full mesh iBGP connectivity is needed

***n* iBGP speakers → *n*(*n* – 1)/2 iBGP sessions each speaker handling *n*–1 sessions**

AS CONFEDERATION

TRATTATI COME 3 AS DIVERSI



DIVENTA UNA SESSIONE EBGP

VALE ANCHE PER I ROUTER INTERNI,

POSSIAMO CONTROLLARE LA TOPOLOGIA DELLA RETE DECIDENDO CHI PUO' PARLARE CON CHI.

Route reflector

PRIMO APPROCCIO:

OSPF CON TUTTI I ROUTER COLLEGATI A LINK BROADCAST, UN ROUTER VIENE ELETTO A DESIGNATED ROUTER E SARA' L'UNICO NEIGHBOR

- One or more iBGP speakers act as **concentration routers** ~ DESIGNATED ROUTER
- The other iBGP speakers establish only one BGP session to a route reflector (route reflector **clients**)
- Each route reflector with its clients form a **cluster**, identified by a **CLUSTER-ID**

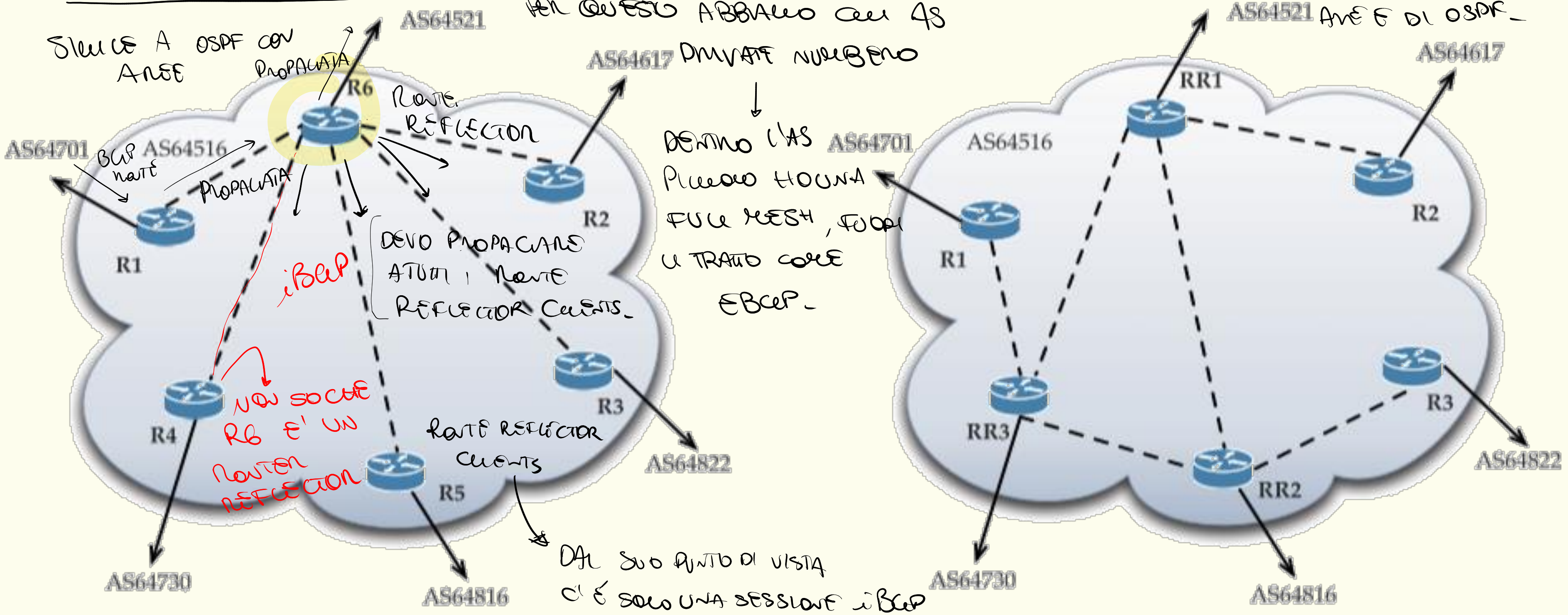
ALTRO APPROCCIO:

AS - CONFEDERATION → PARTIZIONIAMO UN AS IN AS PIU' PICCOLI E USIAMO BGP ~ EQUIVALENTE AL PARTIZIONAMENTO IN AREE DI OSPF.

PER QUESTO ABBIAMO GLI AS PRIVATE NUMBERED

↓

DENTRO L'AS PICCOLO HO UNA FULL MESH, FUORI IL TRATTO COME EBGP.



SIMILE A OSPF CON AREE

PROPAGATA

ROUTE REFLECTOR

BGP ROUTE

PROPAGATA

iBGP

DEVO PROPAGARE A TUTTI I ROUTE REFLECTOR CLIENTS.

NON SO CHE R6 E' UN ROUTER REFLECTOR

ROUTE REFLECTOR CLIENTS

DAL SUO PUNTO DI VISTA C'E' SOLO UNA SESSIONE iBGP INVECE CHE N-1

LE OPERAZIONI DESCRITTE POSSONO APPLICARSI A TUTTO L'AS (UN SOLO CLUSTER) OPPURE A VARI CLUSTER.

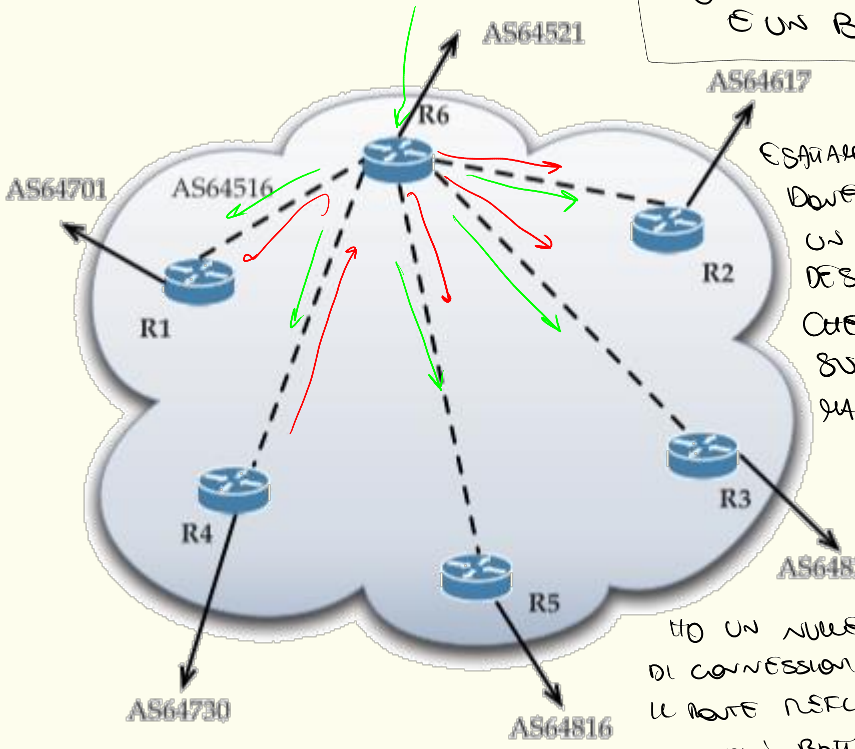
CLUSTER = SOTTO-INSIEME DI BGP SPEAKERS DOVE SELEZIONO UN ROUTE REFLECTOR.

Route reflector

QUANDO CONFIGURO UN ROUTE REFLECTOR SCALO COMPLESSITA' DEI CLIENT DA n A 1, MA IL ROUTE REFLECTOR SCALA ANCORA IN $(n-1)$.

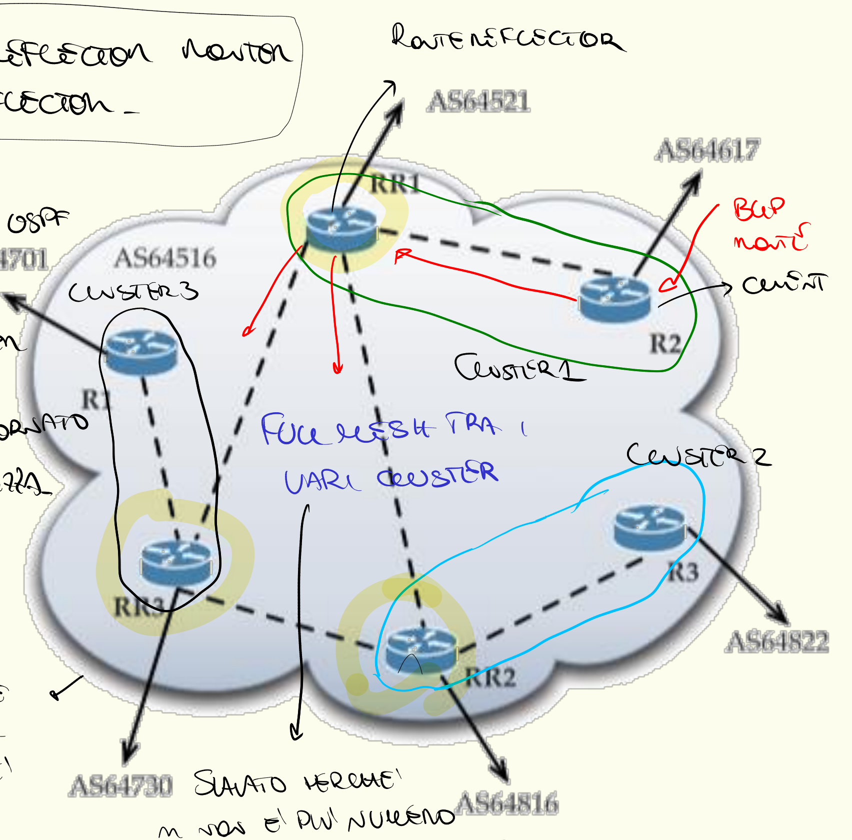
SE HO UN PROBLEMA DI SCALABILITA' CON UN ROUTE REFLECTOR DIVIDO AS IN CLUSTER E IN QUESTO MODO SCALO

- Announcement received **from another route reflector** → reflect/pass it to its clients
- Announcement received **from a route reflector client** → reflect to another route reflector
- Announcement received **from an eBGP speaker** → reflect to all other route reflectors and clients



OVVIAMENTE HO UN REFLECTOR ROUTER E UN BACK-UP REFLECTOR.

ESATTAMENTE COME IN OSPF DOVE AVEVAMO UN BACKUP DESIGNATED ROUTER CHE TIENE IL SUO DB AGGIORNATO MA NON PUBBLICIZZA.

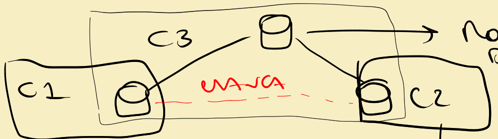


HO UN NUMERO MAGGIORE DI CONNESSIONI INTERNE MA IL ROUTE REFLECTOR NON E' PIU' BOTTLENECK.

SCALATO PERCHE' n NON E' PIU' NUMERO DI CLIENTS MA DI CLUSTER

NEL CASO DELLA AS CONFEDERATION (ALTRO APPROCCIO) STIAMO CONSIDERANDO ANCHE I ROUTER INTERNI, INVECE QUI SONO I BGP SPEAKER.

Route reflector



ROUTE REFLECTOR DEI ROUTE REFLECTOR

POSSO FARE ANCHE CLUSTER DI CLUSTER!

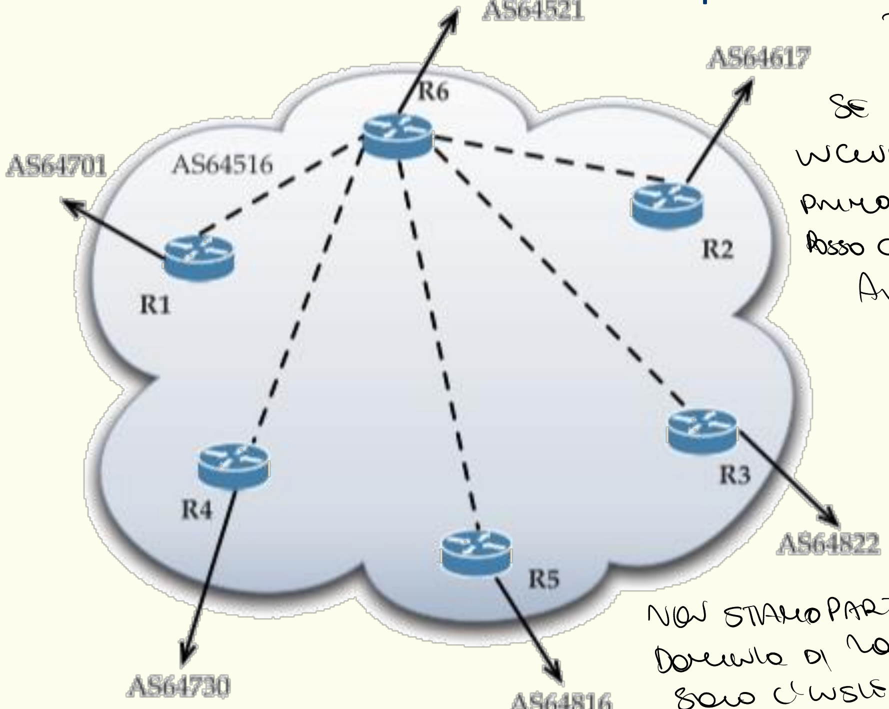
NON STIAMO TORNANDO AD UNA PARTIAL MESH, PERCHE' ABBIAMO UNA GERARCHIA.

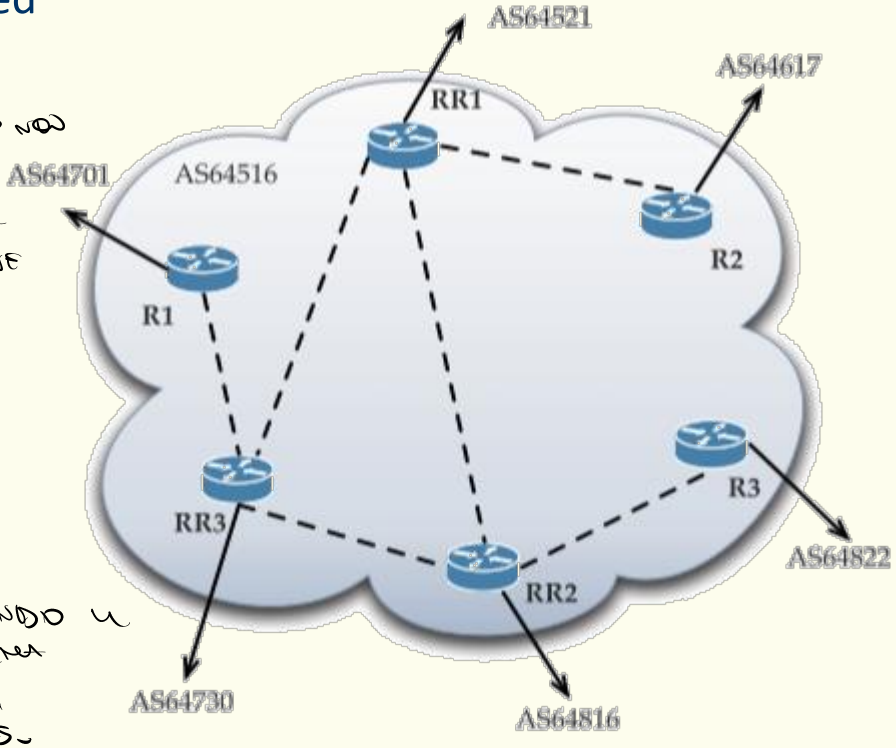
- Route reflectors must form a **full mesh connectivity among themselves**!
- How to avoid routing loops? Two additional attributes

ANCHE QUESTO PUO' PORTARE A PROBLEMI DI SCALABILITA'

1. **ORIGINATOR–ID**: identifies a route reflector through its 4-byte router ID, added only by the originating route reflector → SE LO INCONTRO VUOL DIRE CHE STO LOOPANDO E SCARTO
2. **CLUSTER–LIST**: stores a sequence of 4-byte CLUSTER–ID values to indicate the path of clusters that an advertised IP prefix has visited

SE IL LOOP NON INCLUDE IL PRIMO ROUTER POSSO COMUNQUE AVERLO.





NON STIAMO PARTIZIONANDO IL DOMINIO DI ROUTING MA SOLO L'INSIEME DEI BGP SPEAKERS.

References

- D. Medhi, K. Ramasamy, **Network Routing: Algorithms, Protocols, and Architectures**, 2nd/ed. Morgan Kaufmann, ©2018
- RFC
  - **RFC4271**, A Border Gateway Protocol 4 (BGP-4), Jan. 2006
  - **RFC4360**, BGP Extended Communities Attribute, Feb. 2006