

Routage BGP

Philippe ISORCE

D'après les cours de Luc SACCAVINI © INRIA, Eric FLEURY © INSA LYON

Taghrid ASFOUR © CPE LYON, Benoit Lourdelet © Cisco Systems



Remerciements

- Luc Saccavini, INRIA
- Laurent Toutain, ENST Bretagne
- Isabelle Chrisment, LORIA
- Nick McKeown, Stanford University
- Eric Fleury, INSA Lyon, CITI
- Taghrid Asfour, CPE Lyon
- Benoit Lourdelet, Cisco Systems

Système autonome : AS

- AS : Ensemble de routeurs et de réseaux sous une administration unique :
« *administrative domain* »
 - Entreprise, campus, opérateur, organisme,...
 - Les numéros d'AS « *AS number* » sont attribués :
 - Ressource rare : (2 bytes) 2^{16}
 - Délivrés par AFNIC-France et IANA
 - AS N° 64512 à 65535 numéros privés

Protocoles de routage extérieurs

Protocoles de routage	Intérieurs IGP	Extérieurs EGP
Objectifs	<ul style="list-style-type: none">■ Optimiser■ Fiabiliser	<ul style="list-style-type: none">■ Maintenir la connectivité■ Appliquer une stratégie de routage■ Sécuriser
Exemples	RIP, IGRP, OSPF	BGP (IBGP et EBGP)

Objectifs généraux de BGP

- Échanger des routes (du trafic) entre organismes indépendants
 - Opérateurs, ISP,...
 - Gros sites mono ou multi connectés
- Implémenter la politique de routage de chaque organisme
 - Respect des contrats passés entre organismes
 - Sûreté de fonctionnement
- Être indépendant des IGP utilisés en interne à l'organisme
- Supporter un passage à l'échelle (de l'Internet)
- Minimiser le trafic induit sur les liens
- Donner une bonne stabilité au routage

Principes généraux du protocole BGP

- Protocole de type ***PATH-vecteur***
- Chaque entité est identifiée par un numéro d'AS
- La granularité du routage est l'AS
- Le support de la session BGP est ***TCP:179***
 - garantie d'une bonne transmission des informations.
 - Envoi initial puis mise à jour
- Les sessions BGP sont établies entre les routeurs de bord d'AS
- Protocole point à point entre routeurs de bord d'AS
- Protocole symétrique
- Politique de routage → filtrage des routes apprises et annoncées
 - tout ou rien sur la route (annonce, prise en compte),
 - modification des attributs de la route pour modifier la préférence

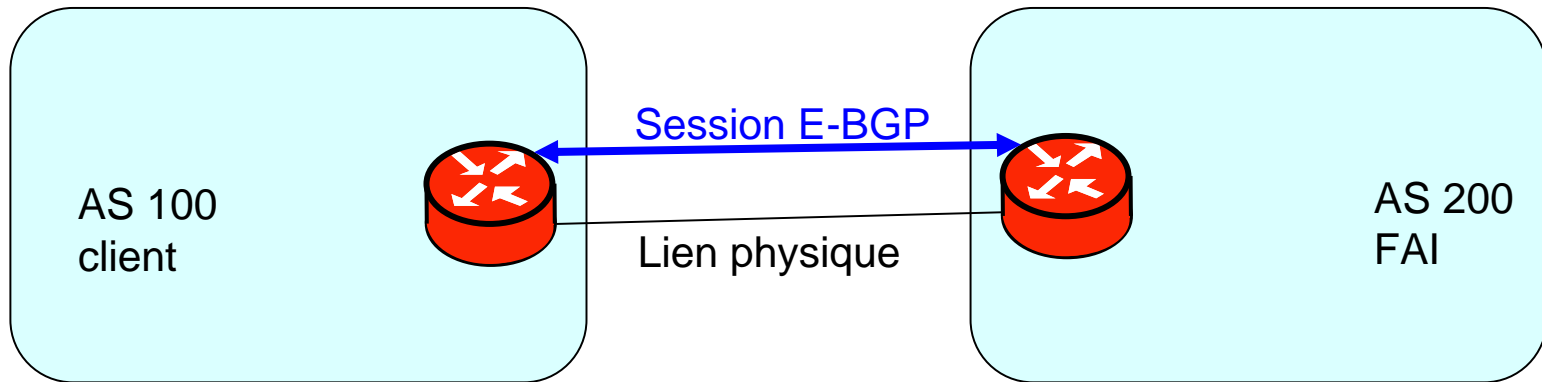
ne jamais oublier qu'annoncer une route vers un réseau c'est accepter du trafic à destination de ce réseau

BGP : Fonctionnement

- A l'initialisation :
 - Les deux routeurs BGP s'échangent leurs tables de routage.
 - Echange d'informations à chaque changement dans les tables de routage
 - Pas de mise à jour régulière comme dans RIP
- Chaque routeur envoie des messages « **heartbeats** » pour informer l'autre de sa viabilité.
- Chaque AS choisit lui même la route qu'il utilise sur la base des vecteurs de chemin reçus
 - La route choisie n'est pas forcément la plus courte
- Chaque AS choisit, en fonction de sa politique, les routes à déclarer aux voisins

Exemple de connexion BGP (1)

- Client connecté à un seul Fournisseur d'Accès Internet (FAI). Seuls les routeurs de bord de l'AS sont configurés.
- Les routeurs qui échangent leurs informations en BGP doivent être directement connectés (liaison point à point ou LAN partagé).
- L'utilisation de numéros d'AS privés est à éviter pour des AS terminaux (clients) car une connexion à un deuxième AS de transit (FAI) peut conduire à une configuration illégale.

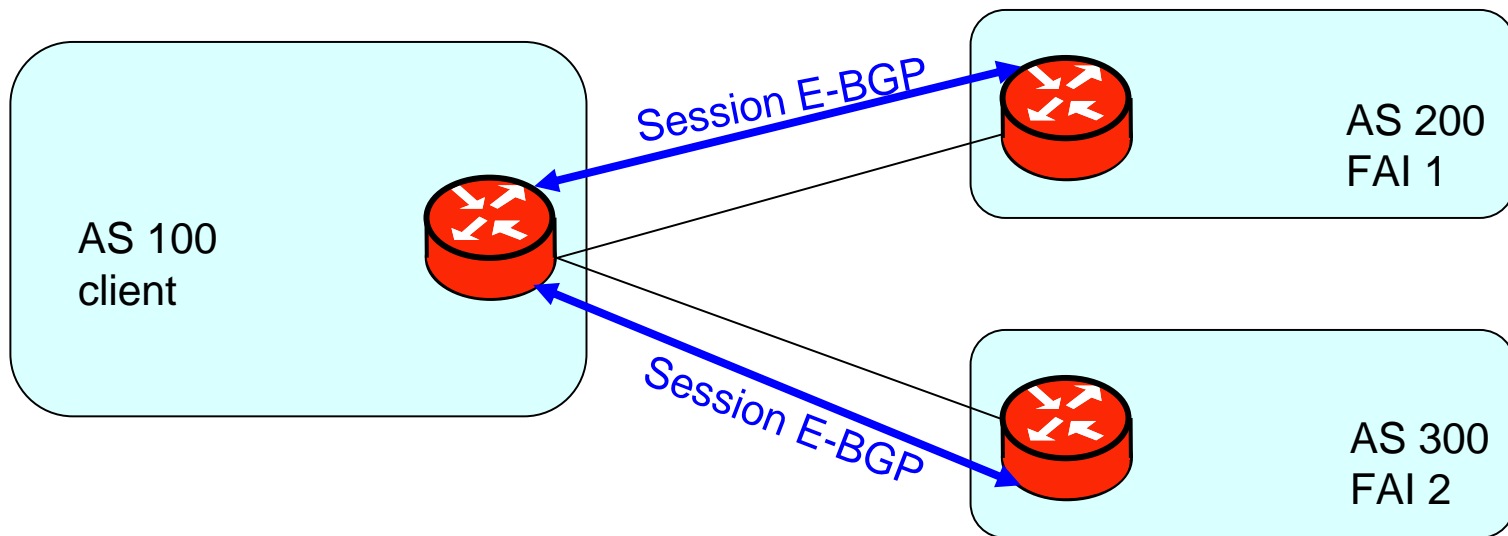


AS officiels (enregistrés) : de 1 à 64511

AS privés (non-enregistrés) : de 64512 à 65535

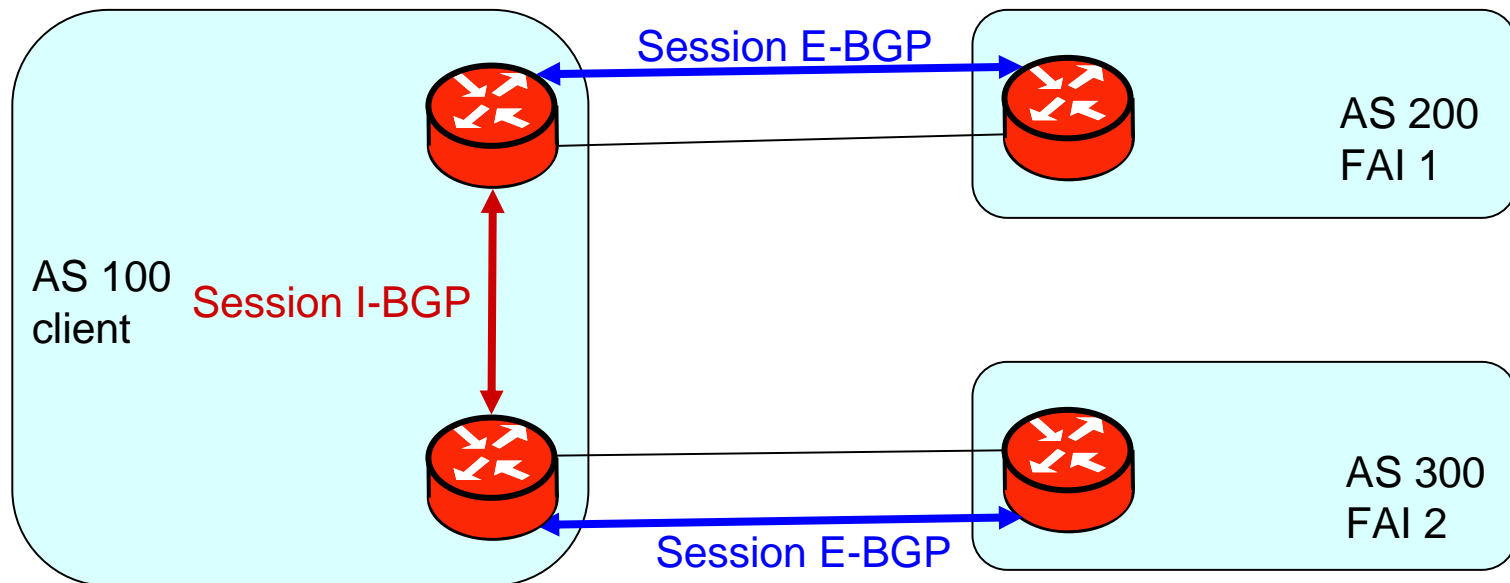
Exemple de connexion BGP (2)

- Client connecté à deux FAI
 - faire passer tout son trafic par FAI1 (AS 200) et garder sa liaison vers FAI2 (AS 300) en secours
 - Équilibrer son trafic entre FAI1 et FAI2.
- C'est le cas typique qui amène à utiliser le protocole de routage BGP pour réagir dynamiquement en cas de défaillance d'un lien.



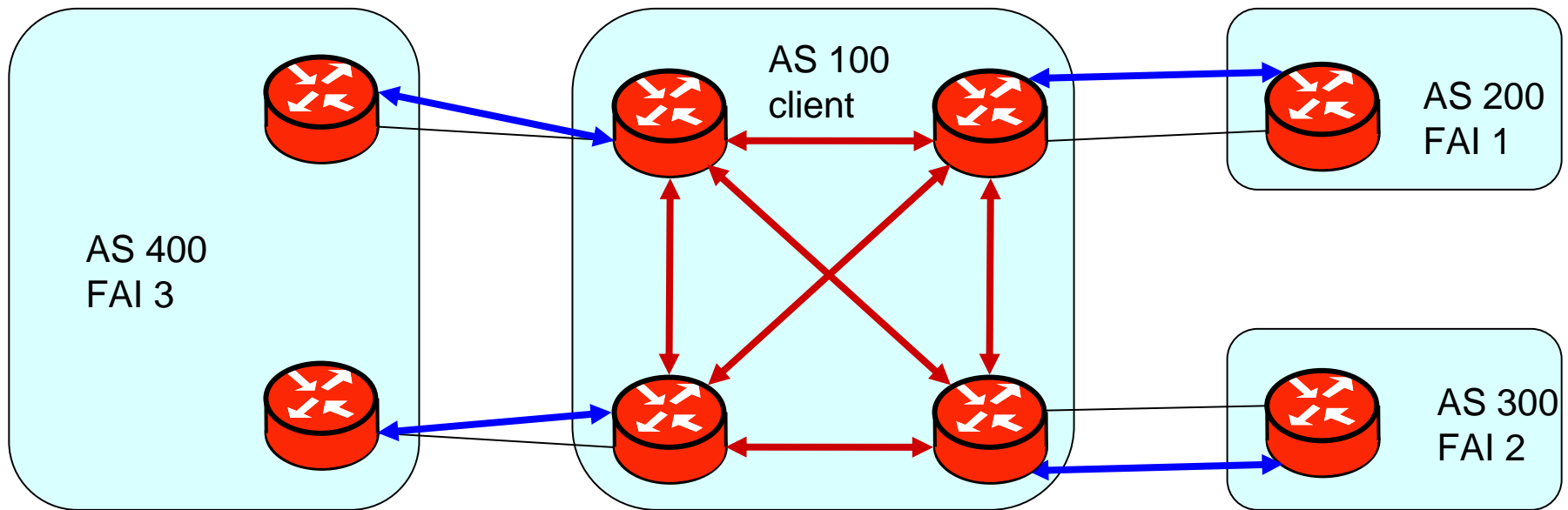
Exemple de connexion BGP (3)

- Client connecté à deux FAI par 2 routeurs
 - protection contre la défaillance de l'un d'entre eux ou de l'un de ses routeurs
- Connexion BGP entre les routeurs de bord de l'AS 100.
 - maintenir la cohérence entre les 2 routeurs qui doivent posséder les mêmes informations de routage
 - En BGP la granularité du routage est l'AS !



Exemple de connexion BGP (4)

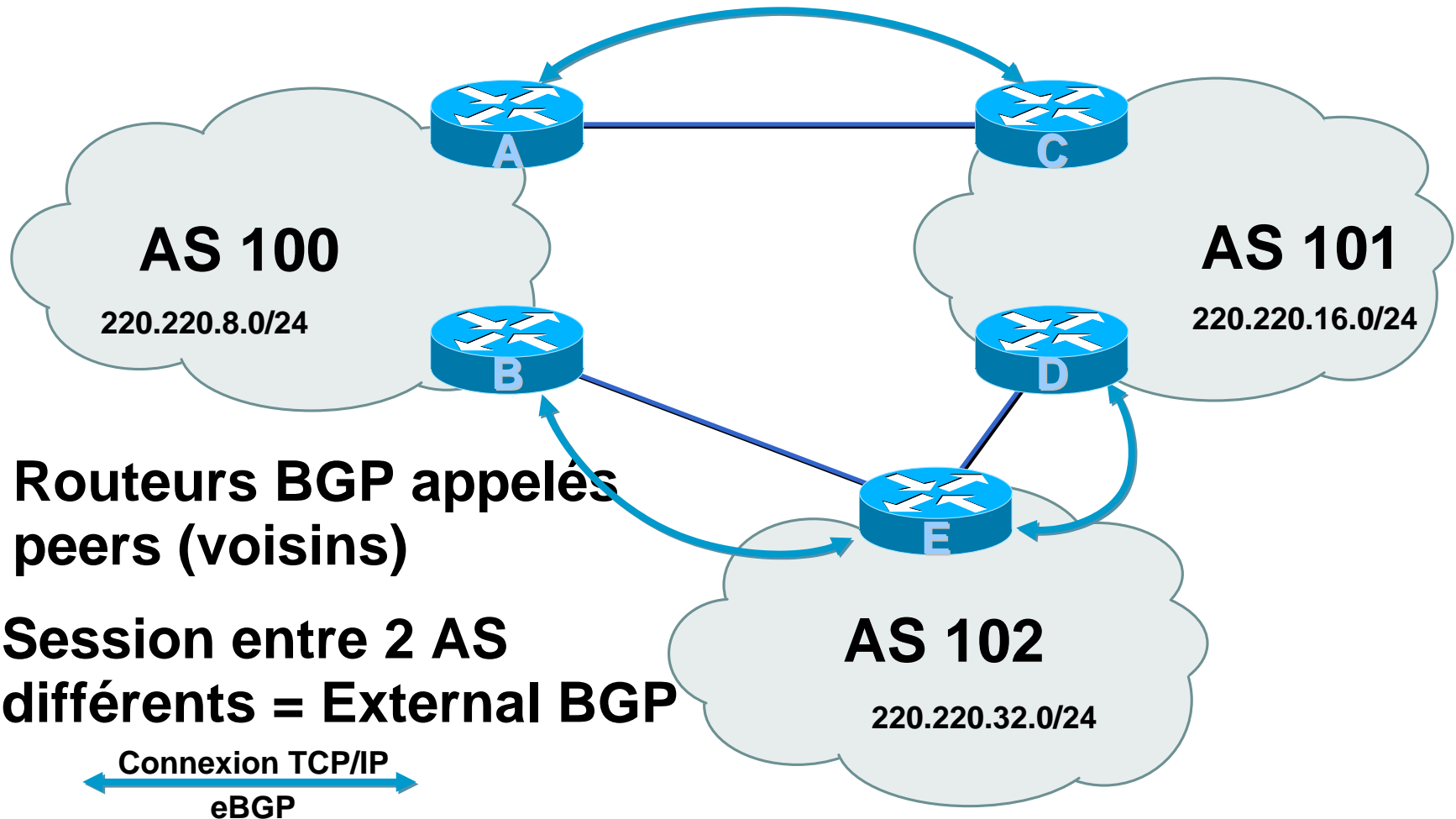
- Client connecté à 3 FAI avec redondance sur l'un
- Maillage complet de sessions I-BGP
- Pour les autres AS, les 4 routeurs de bord de l'AS 100 sont vus, du point de vue fonctionnel comme un seul routeur (avec 4 interfaces).



Règles pour les AS multi-connectés

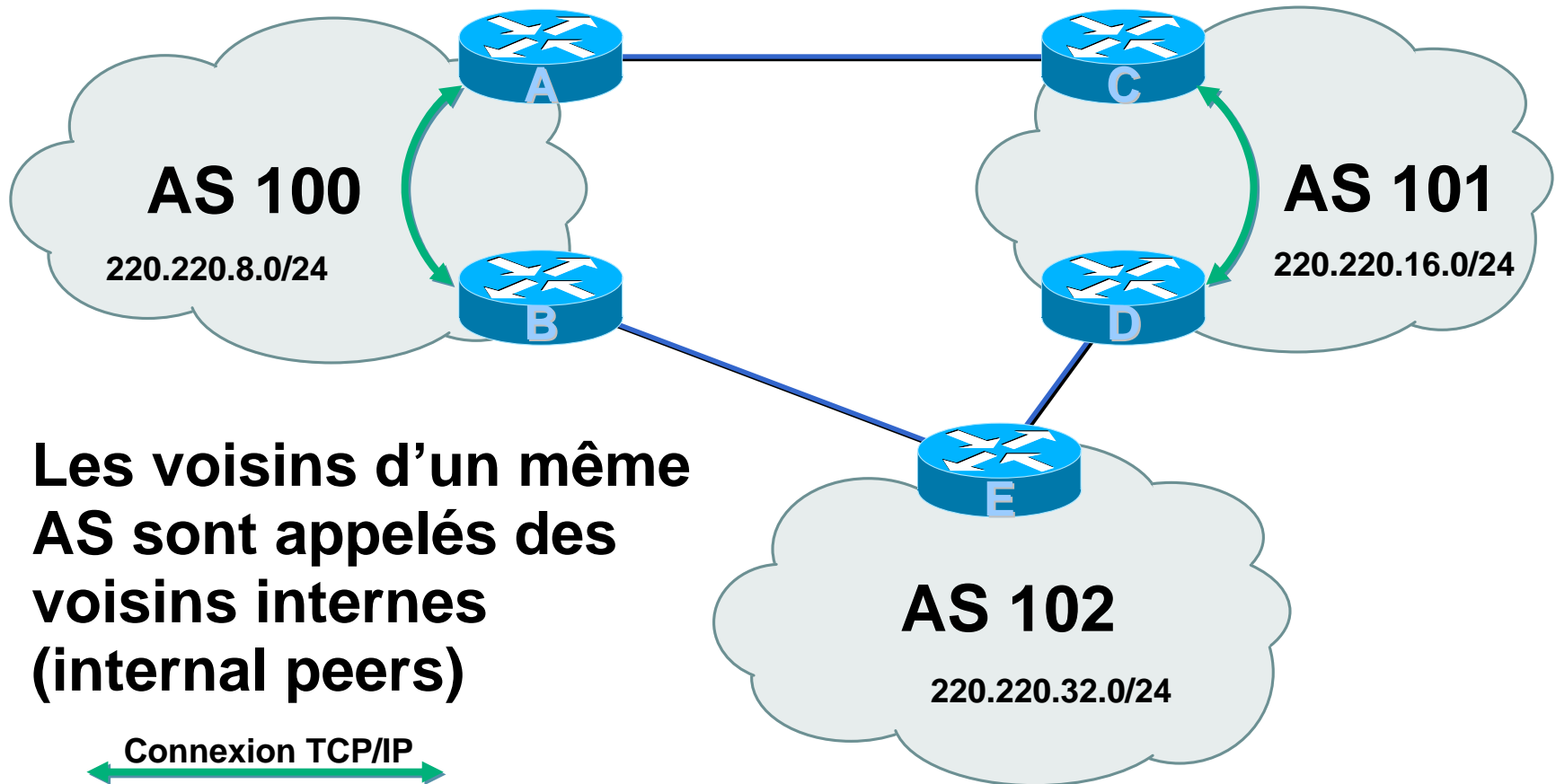
- Les routeurs de bord d'un même AS échangent leurs informations de routage en I-BGP
- Les connexions en I-BGP forment un maillage complet sur les routeurs de bord d'un AS
- Ce sont les IGP internes à l'AS qui assurent et maintiennent la connectivité entre les routeurs de bord qui échangent des informations de routage en I-BGP
- Le numéro d'AS est un numéro officiel (si connexions vers 2 AS différents)
- Dans un même AS, c'est bien l'IGP (ou le routage statique) qui est responsable de la connectivité interne de l'AS.
 - Si un routeur de bord ne peut pas atteindre une route de son AS (qui lui a été annoncée par un voisin interne par exemple), il ne la propagera pas à ses voisins BGP (externes ou internes).

Sessions eBGP



Note: les voisins eBGP doivent être directement raccordés.

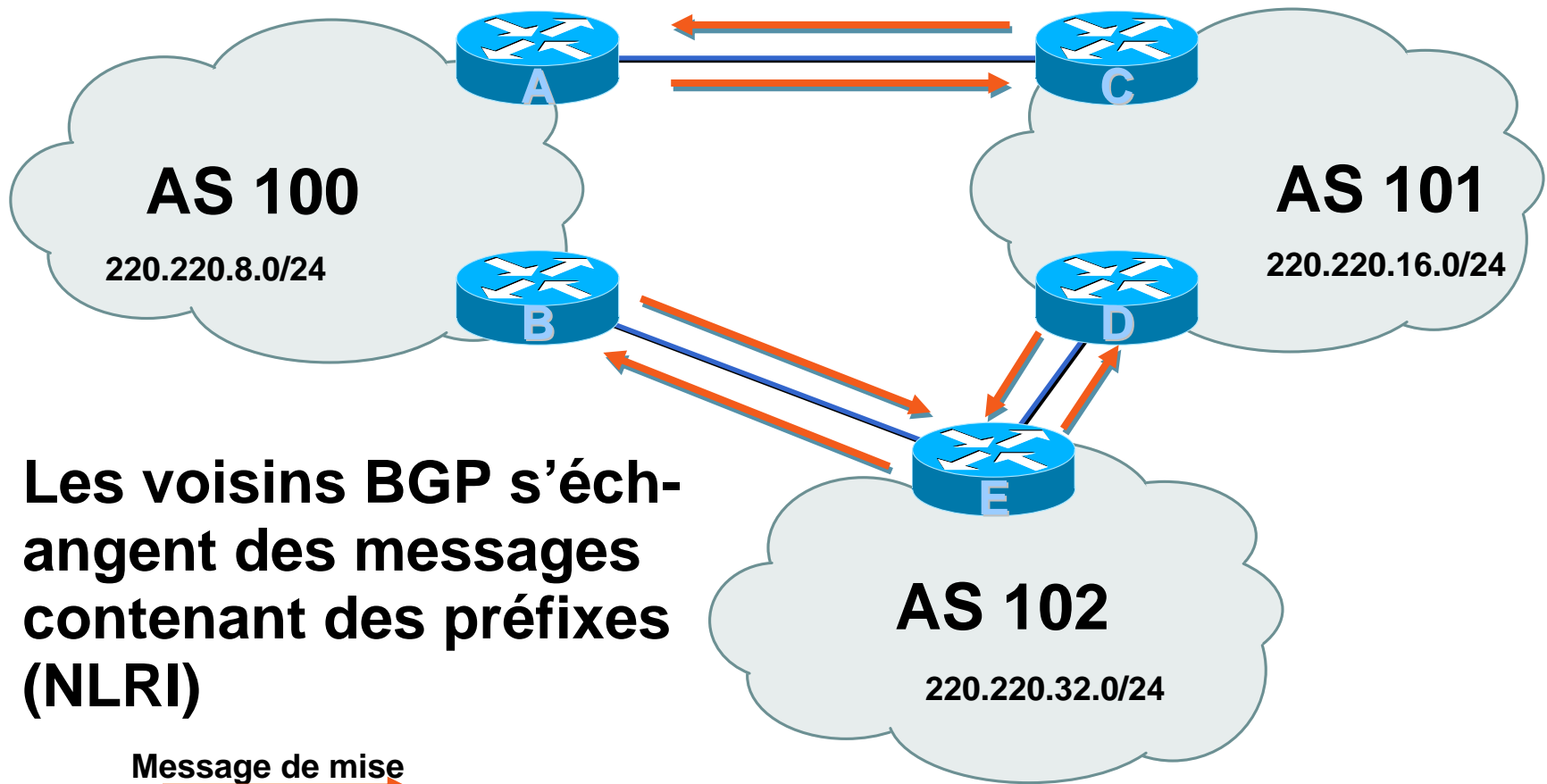
Sessions iBGP



Les voisins d'un même AS sont appelés des voisins internes (internal peers)

Note: les voisins iBGP peuvent ne pas être directement connectés.

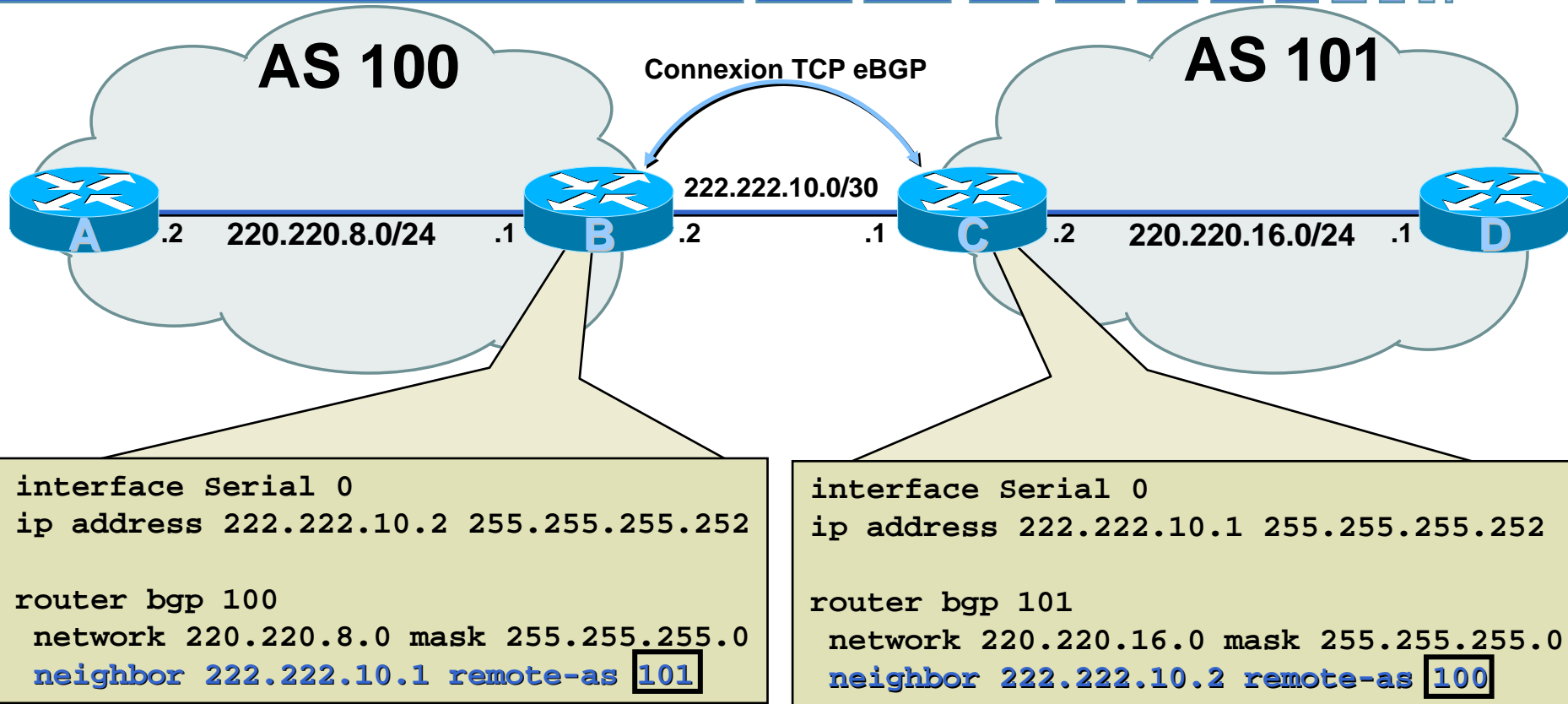
Sessions eBGP



Les voisins BGP s'échangent des messages contenant des préfixes (NLRI)

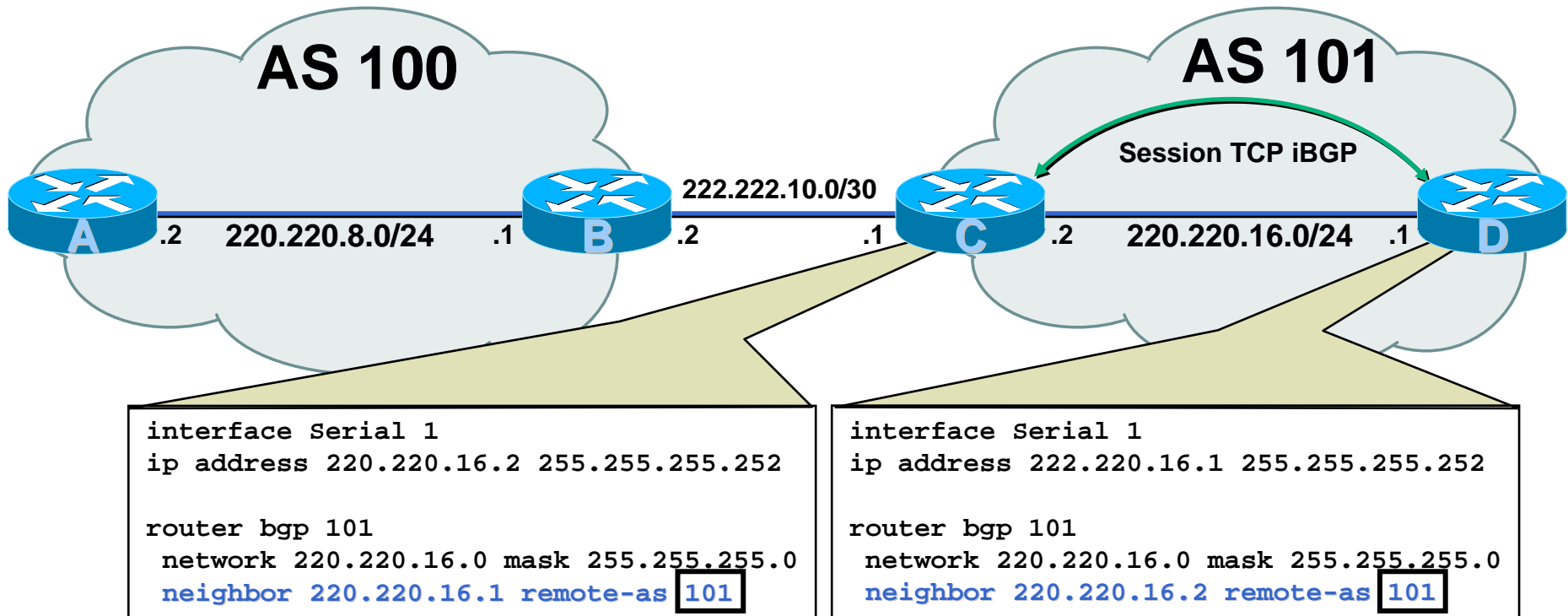
Message de mise
à jour BGP

Configuration des sessions eBGP



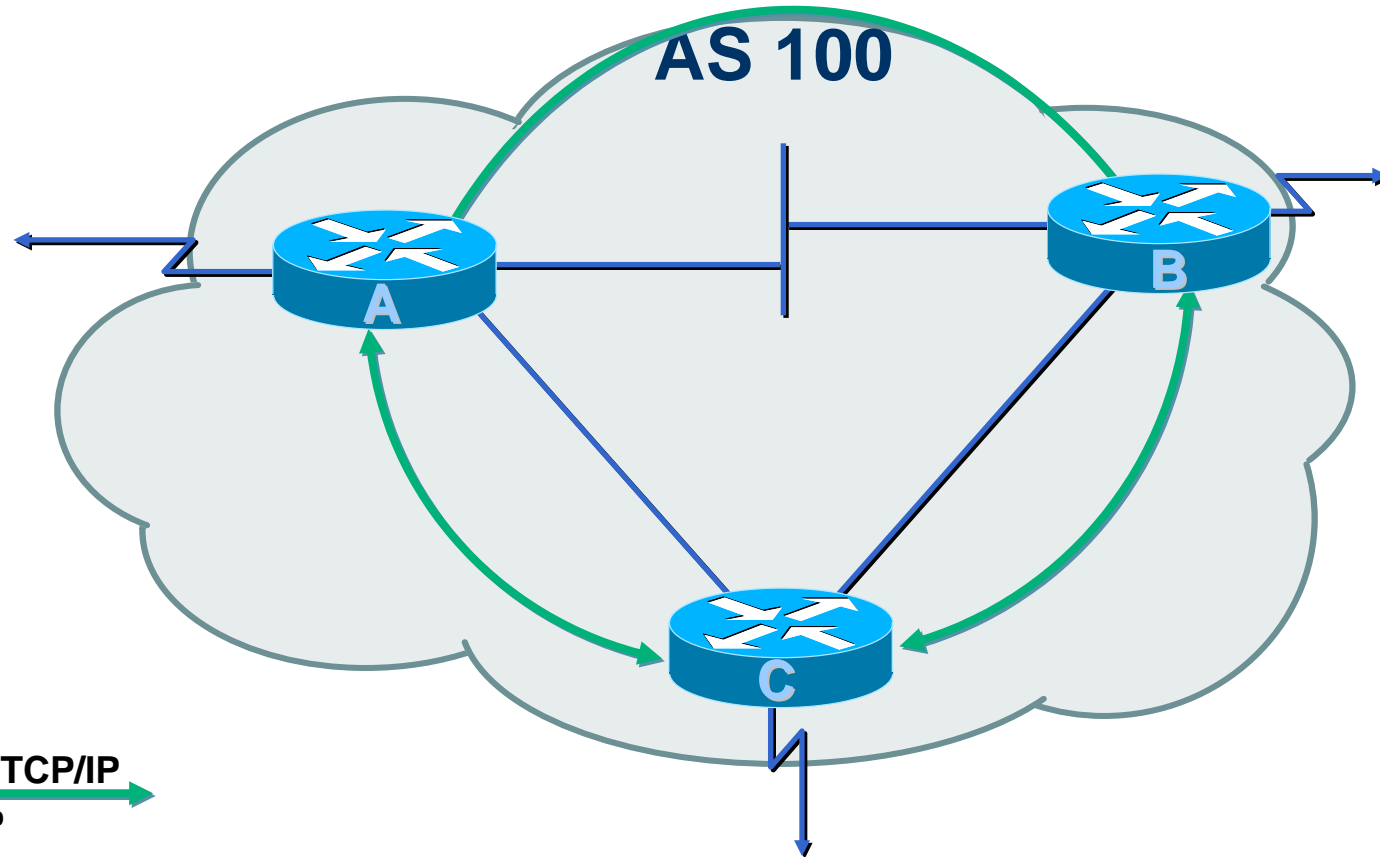
- Les sessions BGP sont établies en utilisant la commande BGP “neighbor” du routeur
 - Lorsque les numéros d’AS sont différents il s’agit d’une session BGP Externe (eBGP)

Configuration de sessions iBGP



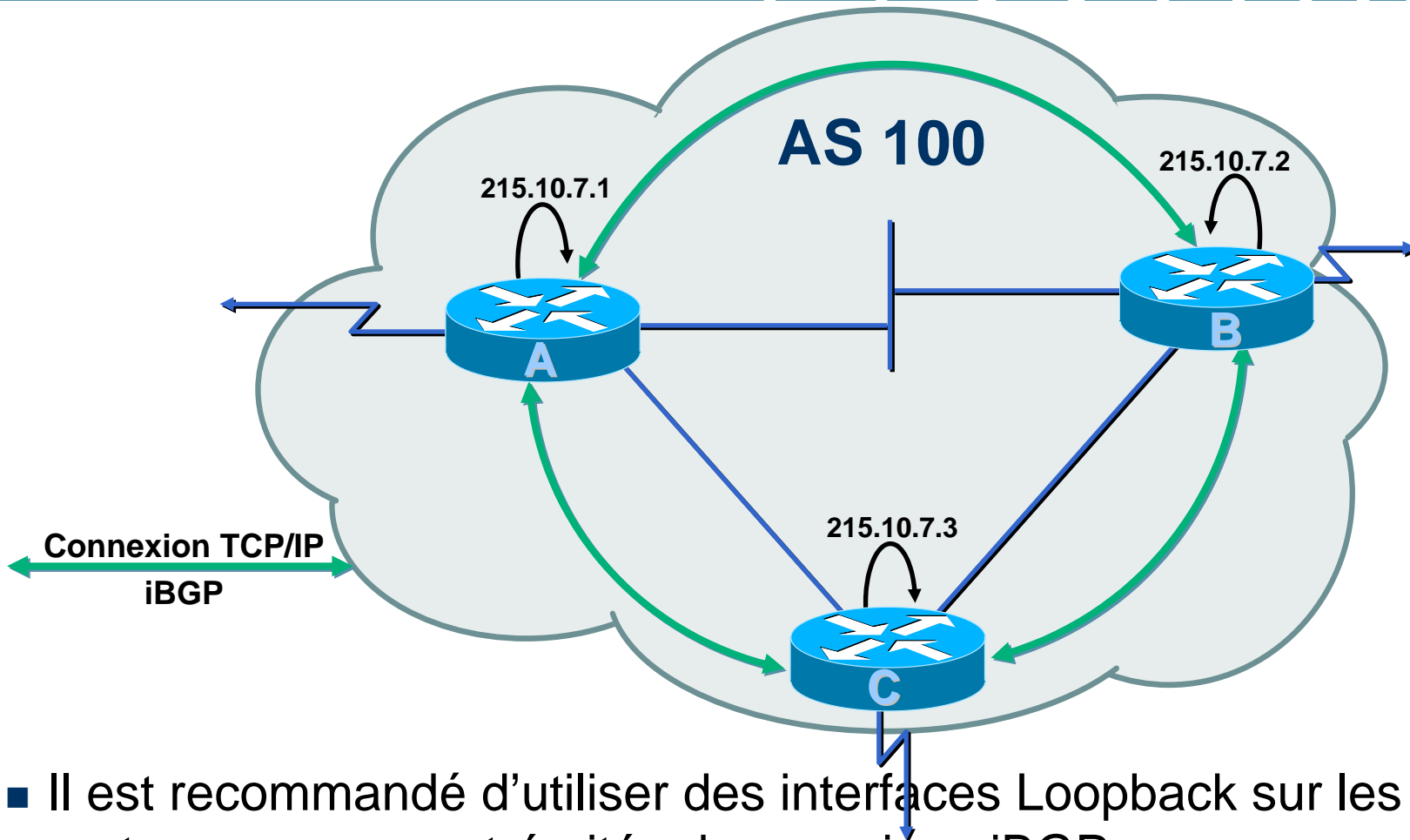
- Les sessions BGP sont établies en utilisant la commande BGP “neighbor” du routeur
 - Numéros d’AS différents -> BGP Externe (eBGP)
 - Numéros d’AS identiques -> BGP Interne (iBGP)

Configuration de sessions BGP



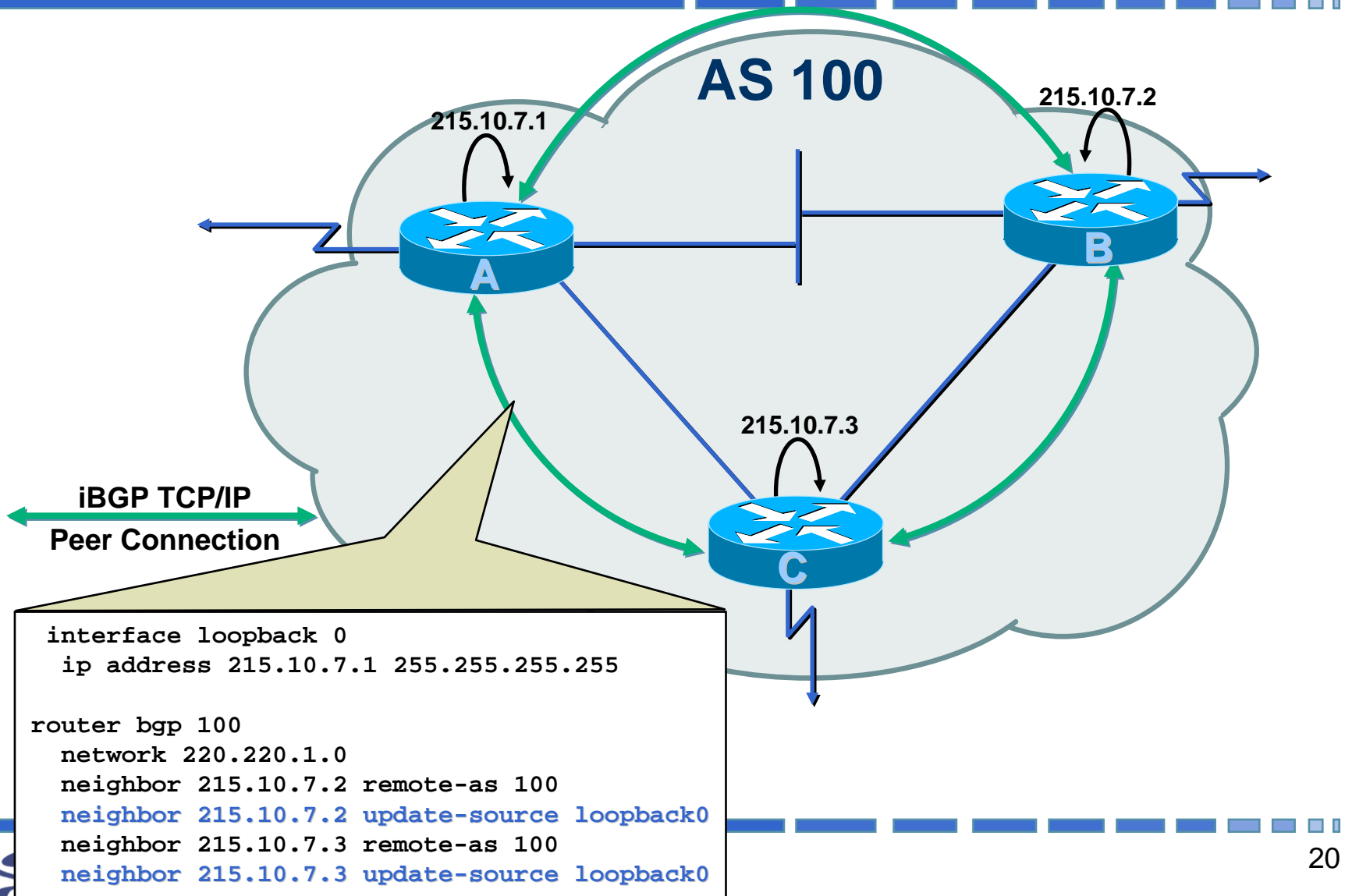
- Chaque routeur iBGP doit établir une session avec tous les autres routeurs iBGP du même AS

Configuration de sessions BGP

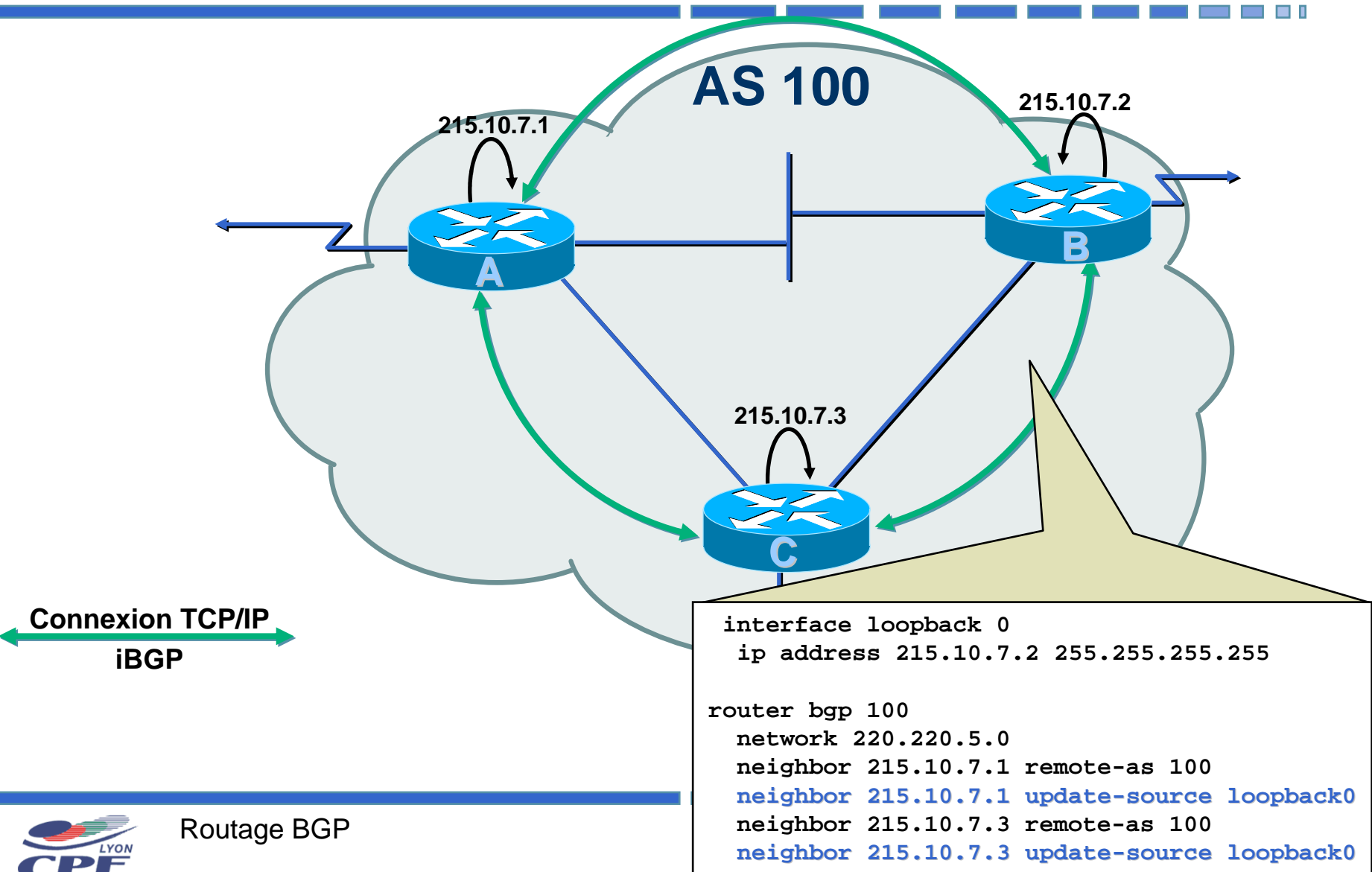


- Il est recommandé d'utiliser des interfaces Loopback sur les routeurs comme extrémités des sessions iBGP

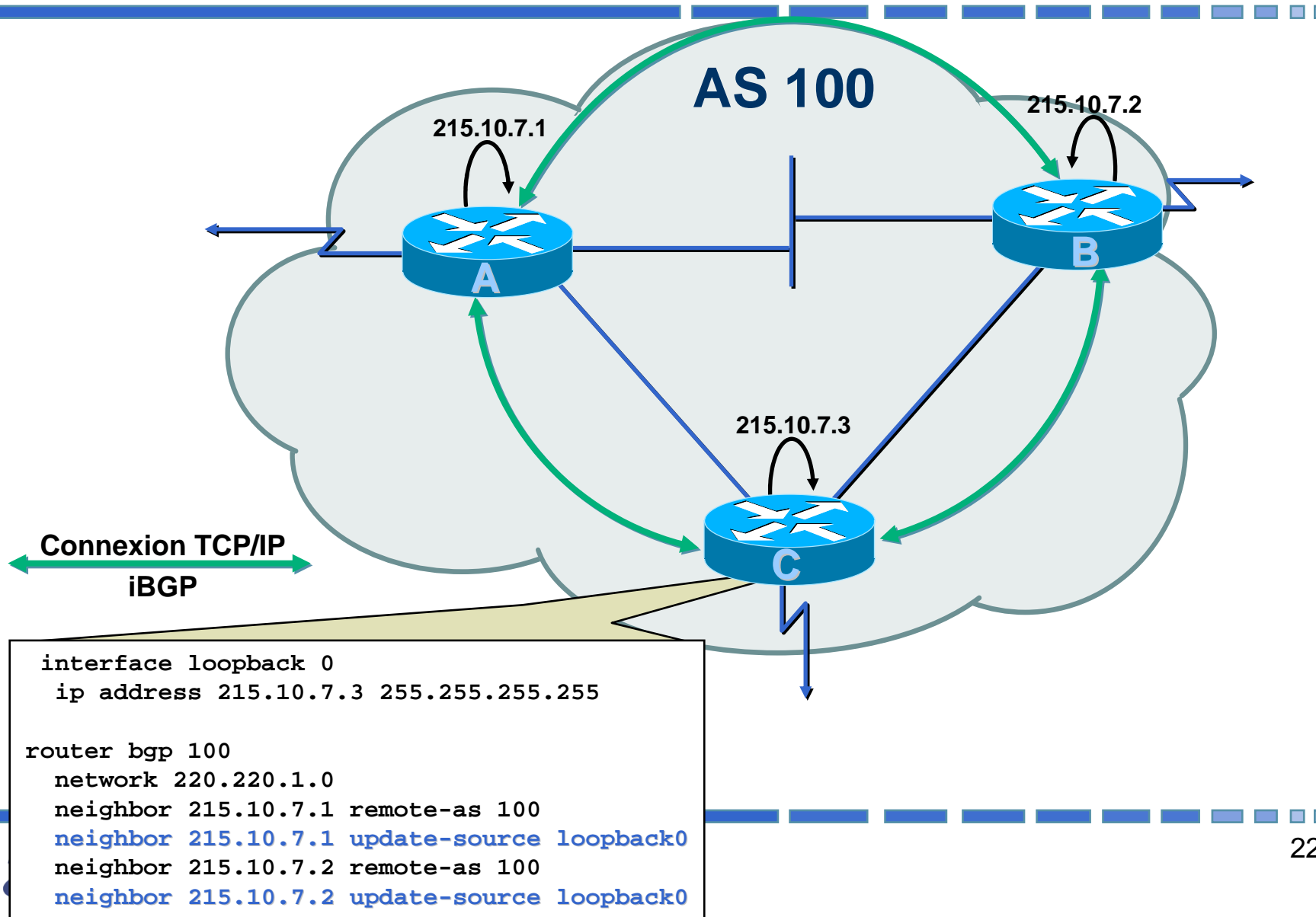
Configuration des sessions BGP



Configuration des sessions BGP



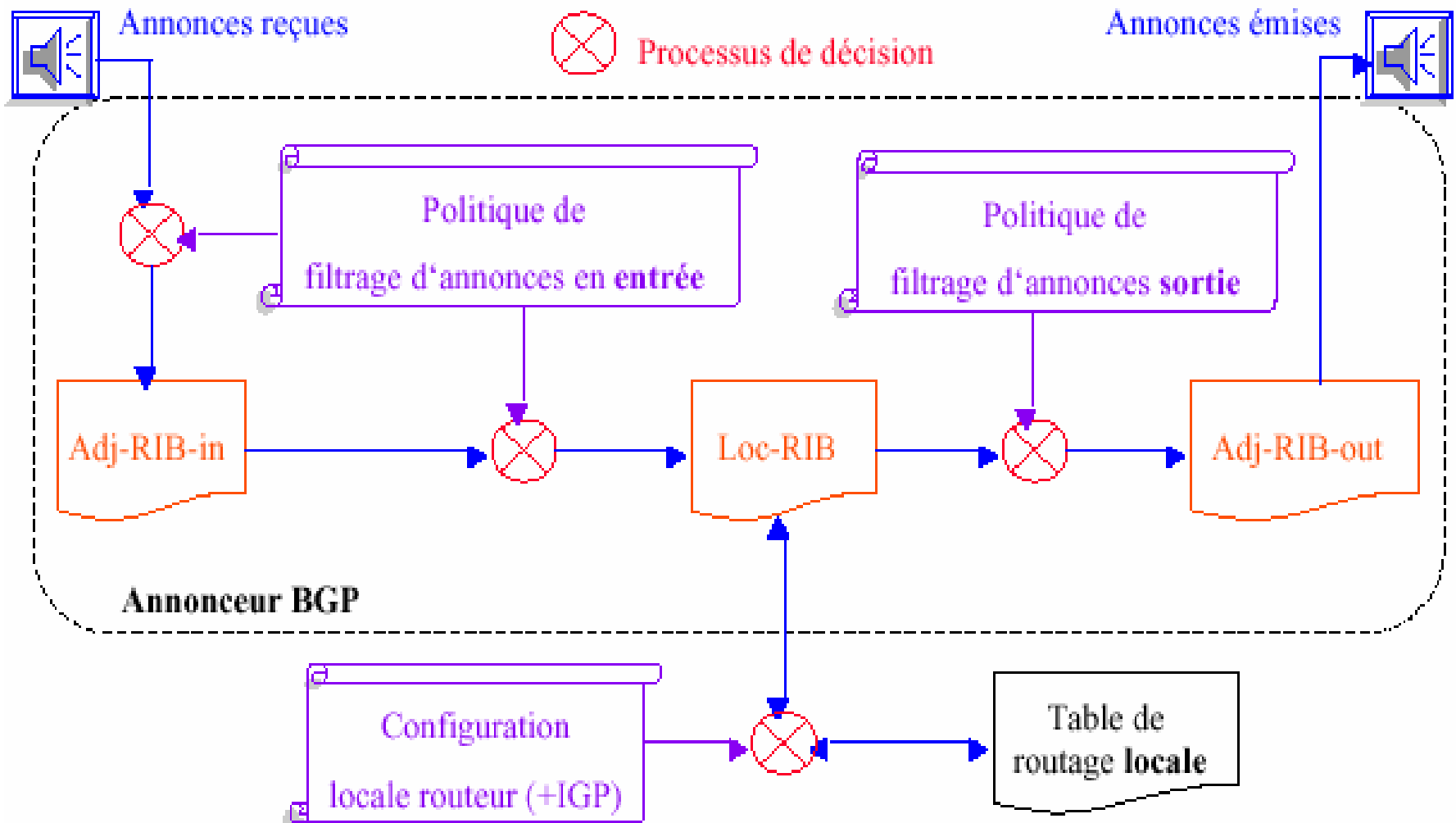
Configuration des sessions BGP



Les composants d'un annonceur BGP

- Une description des politiques de routage
 - entrée et sortie
- Des tables où sont stockées les informations de routage
 - En entrée : table **Adj-RIB-in**
 - En sortie : table **Adj-RIB-out**
 - En interne : table **Loc-RIB**
- Un automate implémentant le processus de décision
- Des sessions avec ses voisins pour échanger les informations de routage

Schéma fonctionnel du processus BGP



La vie du processus BGP

- Automate à 6 états, qui réagit sur 13 événements
- Il interagit avec les autres processus BGP par échange de 4 types de messages :
 - OPEN
 - KEEPALIVE
 - NOTIFICATION
 - UPDATE
- Taille des messages de 19 à 4096 octets
 - Marker (16 octets), Length (2 octets), Type (1 octet), Message contents (0-4077 octets)
- Éventuellement sécurisés par MD5

Le message OPEN

- 1^{er} message envoyé après l'ouverture de la session TCP.
Informe son voisin de :
 - Sa version de BGP (1 octet)
 - Son numéro d'AS (2 octets)
 - D'un numéro identifiant le processus BGP (4 octets)
- Propose une valeur de temps de maintien de la session,
 - **Hold time** : (2 octets)
 - Valeur suggérée : 90 secondes
 - Si 0 : maintien sans limite de durée
- Met le processus en attente d'un KEEPALIVE

Le message KEEPALIVE

- Confirme un OPEN
- Réarme le minuteur contrôlant le temps de maintien de la session
- Si temps de maintien non égal à 0
 - Est réémis toutes les 30 secondes (suggéré)
- Message de taille minimum (19 octets)
- En cas d'absence de modification de leur table de routage, les routeurs ne s'échangent plus que des messages KEEPALIVE toutes les 30 secondes, ce qui génère un trafic limité à environ 5bits/s au niveau BGP.

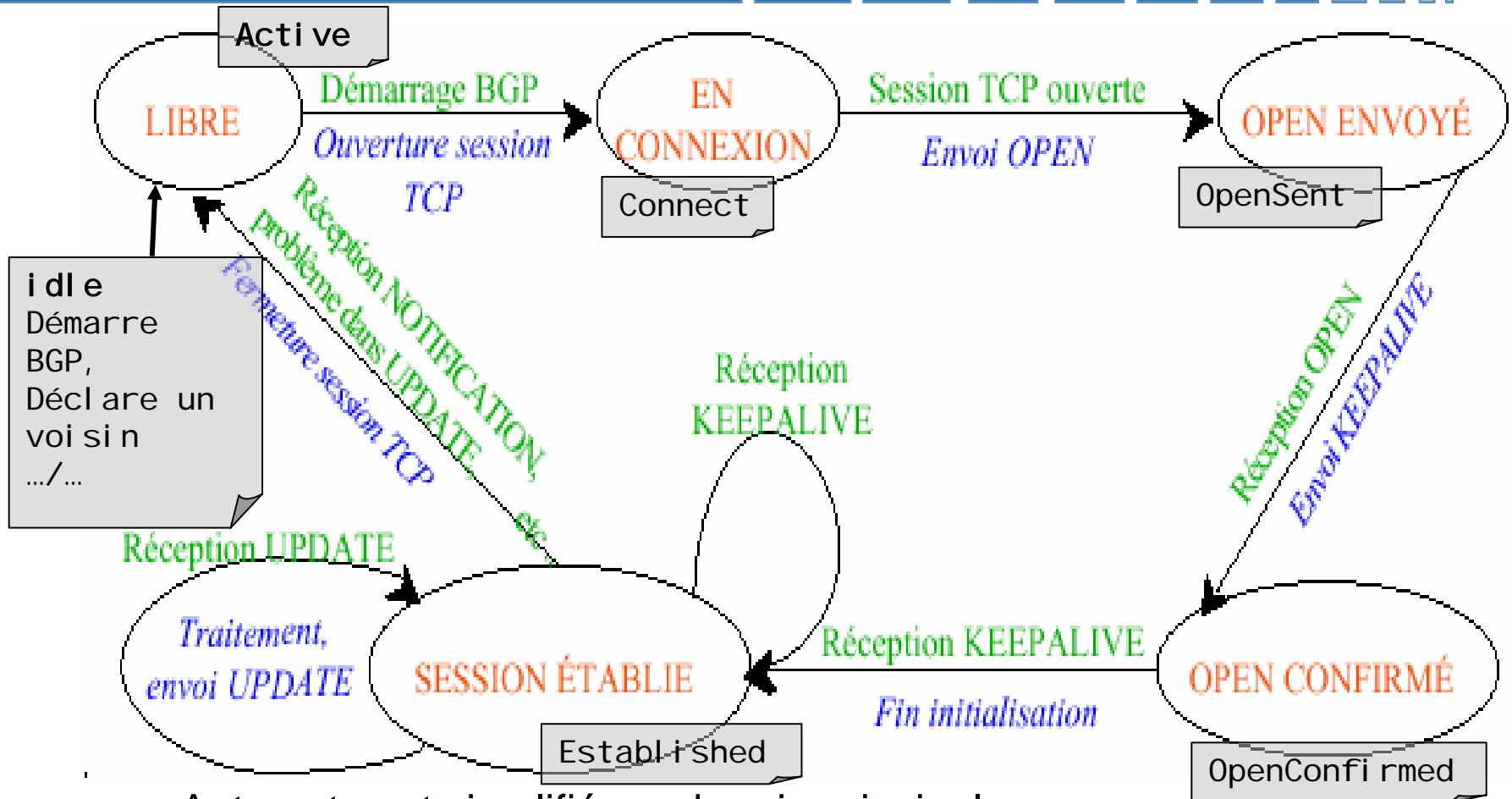
Le message NOTIFICATION

- Ferme la session BGP
- Fournit un code et un sous code renseignant sur l'erreur
- Ferme aussi la session TCP
- **Annule toutes les routes apprises par BGP**
 - peut provoquer des instabilités de routage injustifiées
 - un incident ne veut pas forcément dire que toutes les routes apprises précédemment sont devenues fausses
- Émis sur incidents :
 - Pas de KEEPALIVE pendant 90s (<*hold time*>)
 - Message incorrect
 - Problème dans le processus BGP

Le message UPDATE

- Sert à échanger les informations de routage
 - Routes à éliminer (éventuellement)
 - Ensemble des attributs de la route
 - Ensemble des réseaux accessibles (NLRI) *Network Layer Reachability Information*
 - Chaque réseau est défini par (préfixe, longueur)
- Envoyé uniquement si changement
- Active le processus BGP
 - Modification des RIB (Update, politique de routage, conf.)
 - Émission d'un message UPDATE vers les autres voisins

Le processus BGP : automate 6 à états



- Automate est simplifié au chemin principal

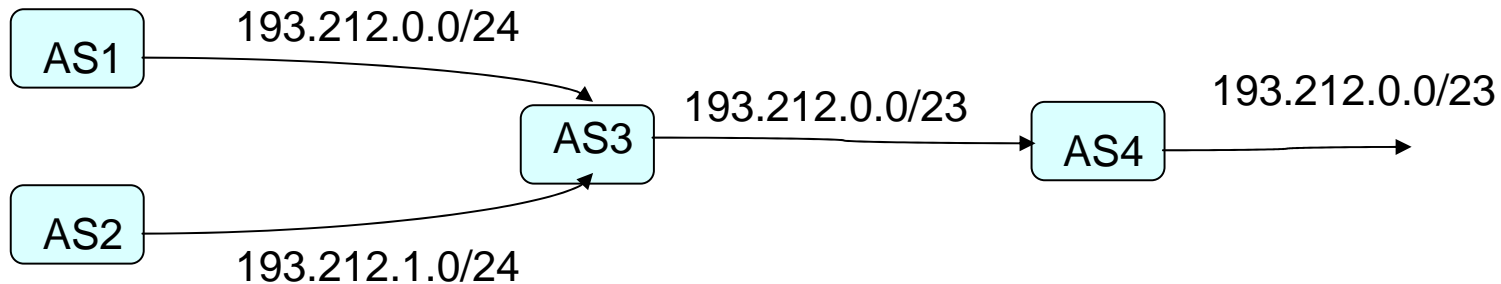
Le message UPDATE : attributs de la route

- Reconnus, obligatoires
□ **ORIGIN, AS_PATH, NEXT_HOP**
 - Reconnus, non obligatoires
□ **LOCAL_PREF, ATOMIC_AGGREGATE**
 - Optionnels, annonçables (transitifs ou non)
□ **MULTI_EXIT_DISC (MED), AGGREGATOR**
 - Optionnels, non-annonçables
□ **WEIGHT (spécifique à Cisco)**
- BGP doit savoir le traiter
- Porté illimitée

Agrégation

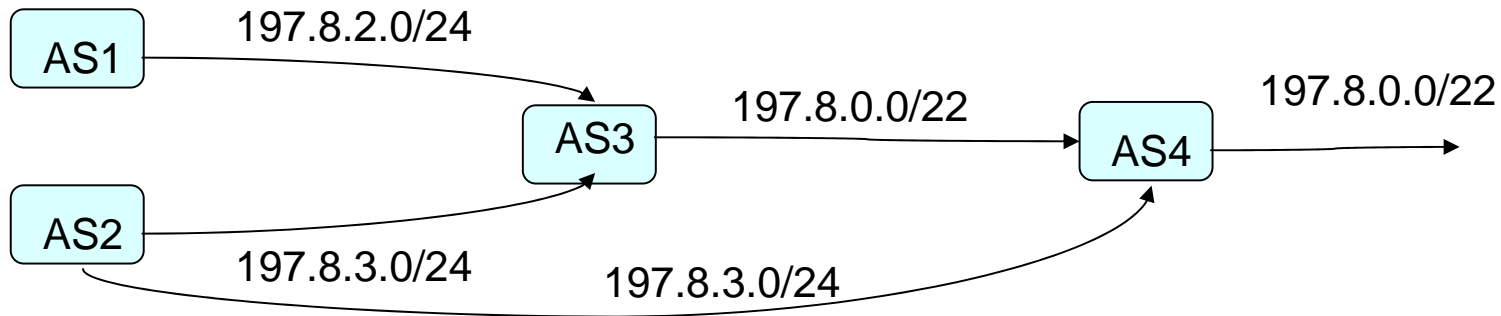
- Tout domaine sans route par défaut
 - Connaître toutes les routes
 - Dans les tables de routage IP
 - Dans les annonces BGP
- Agrégation permet de réduire le nombre de routes

Agrégation I



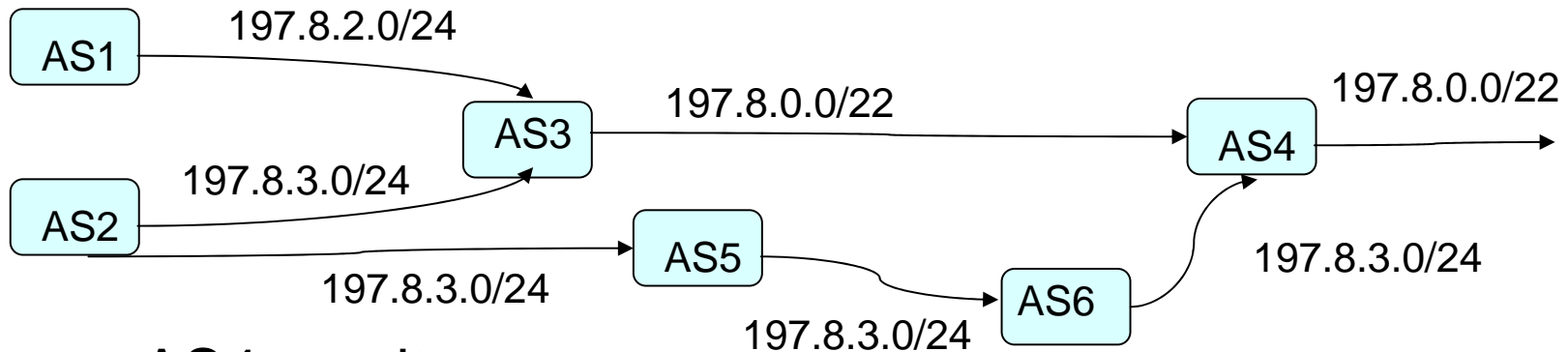
- AS1 : 193.212.0.0/24
- AS2 : 193.212.1.0/24
- AS3 : 193.212.0.0/23
- AS4 : 193.212.0.0/23
- AS_PATH 1
- AS_PATH 2
- AS_PATH 3 {1 2}
- AS_PATH 4 3 {1 2}

Agrégation II



- AS4 reçoit
 - 197.8.0.0/22
 - 197.8.3.0/24
 - AS_PATH 3 {1 2}
 - AS_PATH 2
- Les 2 routes sont injectées dans les tables de AS4
- Comment sont routés les paquets de n4 vers n2 ?

Agrégation III



- AS4 reçoit
 - 197.8.0.0/22
 - 197.8.3.0/24
 - AS_PATH 3 {1 2}
 - AS_PATH 6 5 2
- Les 2 routes sont reçues
- Seul les plus courtes sont injectées dans les tables de AS4
- Comment sont routés les paquets de n4 vers n2 ?

Les attributs de route obligatoires (1)

- **ORIGIN** : Donne l'origine de la route :
 - IGP (i) : la route est intérieure à l'AS d'origine
 - EGP (e) : la route a été apprise par **le protocole** EGP (historique car EGP non employé)
 - Incomplète (?) : l'origine de la route est inconnue ou apprise par un autre moyen (redistribution des routes statiques ou connectées dans BGP par exemple)

- **show ip bgp**

Les attributs de route obligatoires (2)

■ AS_PATH

- Donne la route sous forme d'une liste de segments d'AS
- Les segments sont ordonnés ou non (AS_SET)

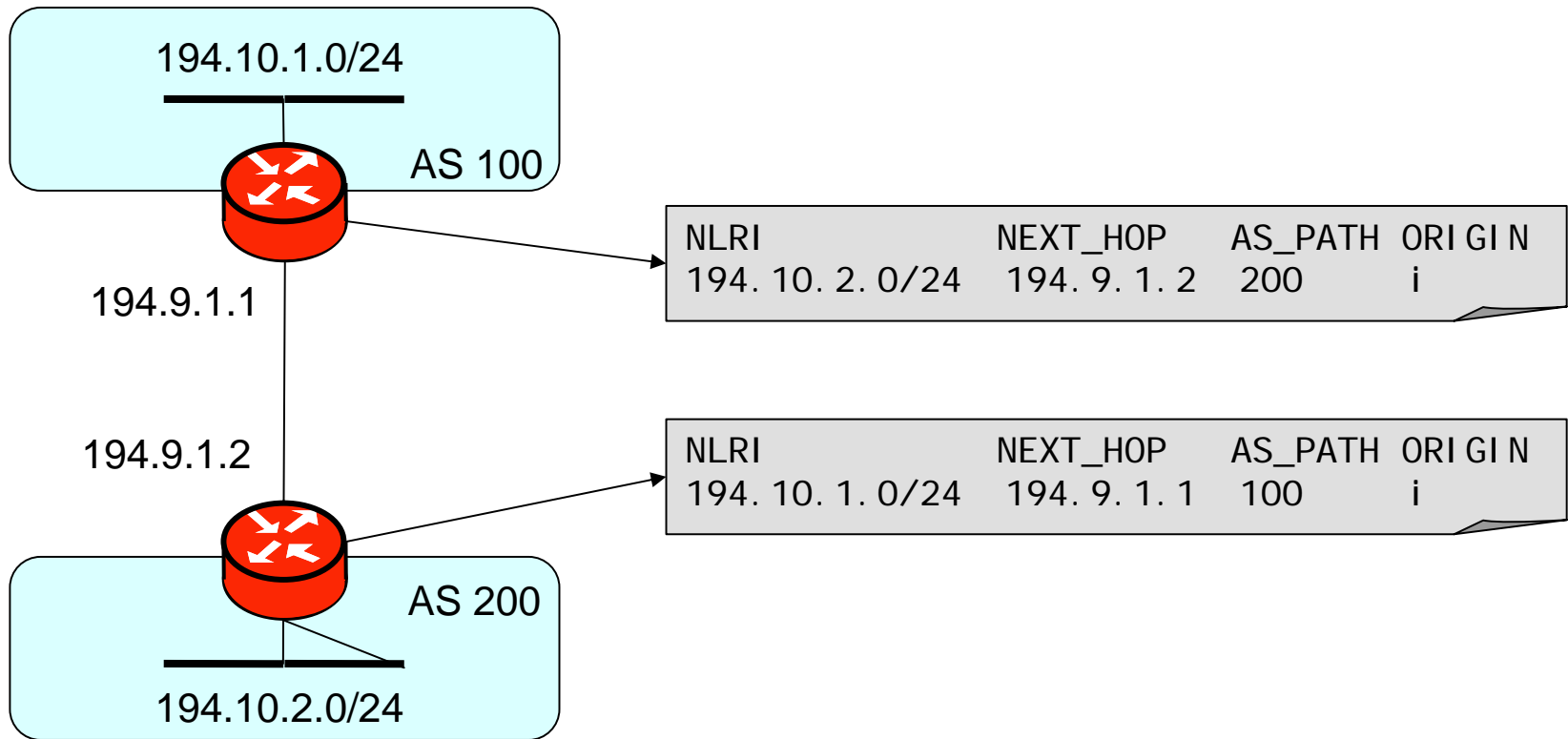
Les segments d'AS non ordonnés sont formés par un routeur qui a fait une opération d'agrégation. Ce dernier regroupe dans cet ensemble non ordonné tous les AS associés aux routes qu'il a agrégées. Cela permet aux autres routeurs de continuer à détecter d'éventuelles boucles concernant ces routes.

- Chaque routeur rajoute son numéro d'AS aux AS_PATH des routes qu'il a apprises avant de les ré-annoncer

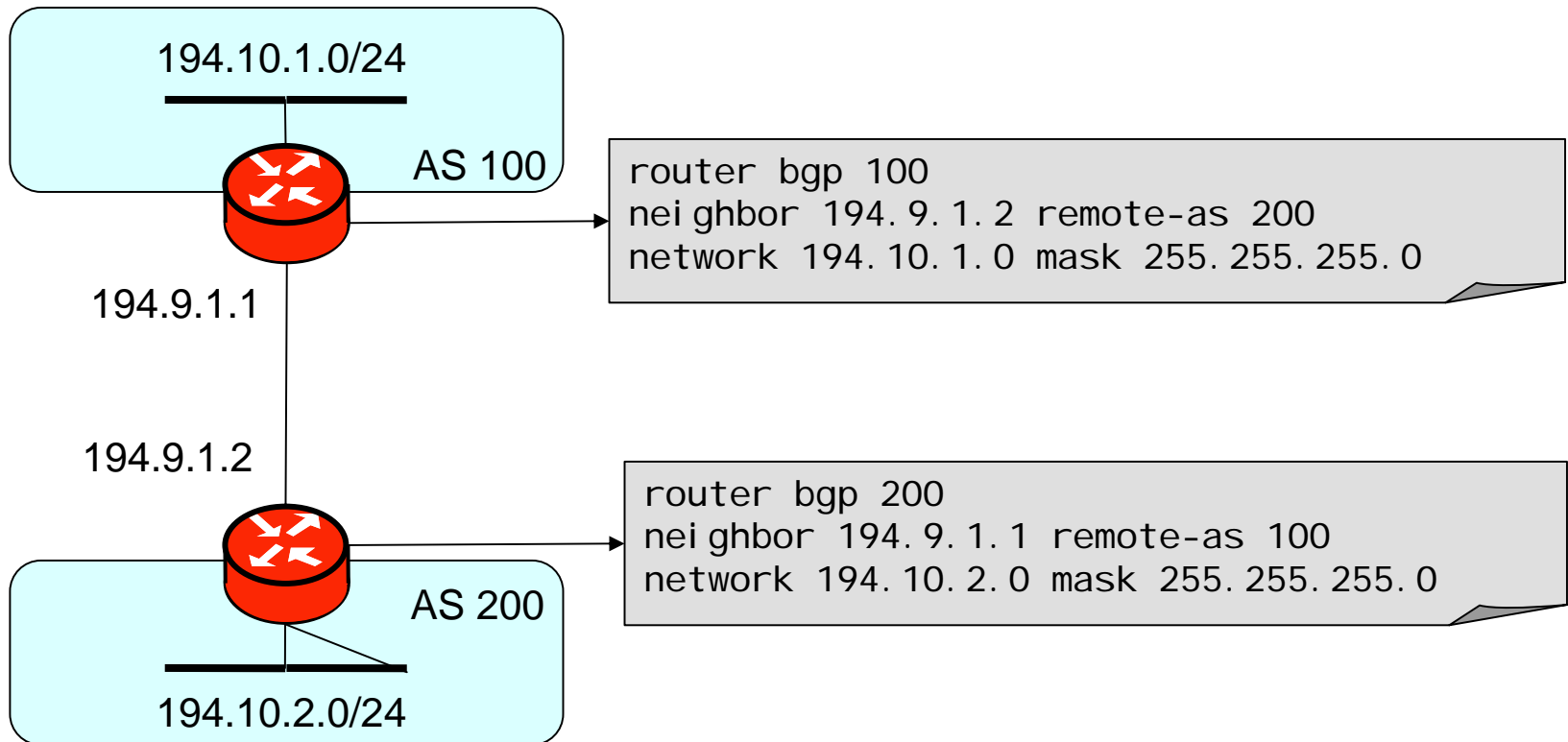
■ NEXT_HOP

- Donne l'adresse IP du prochain routeur qui devrait être utilisé (peut éviter un rebond si plusieurs routeurs BGP sont sur un même réseau local)

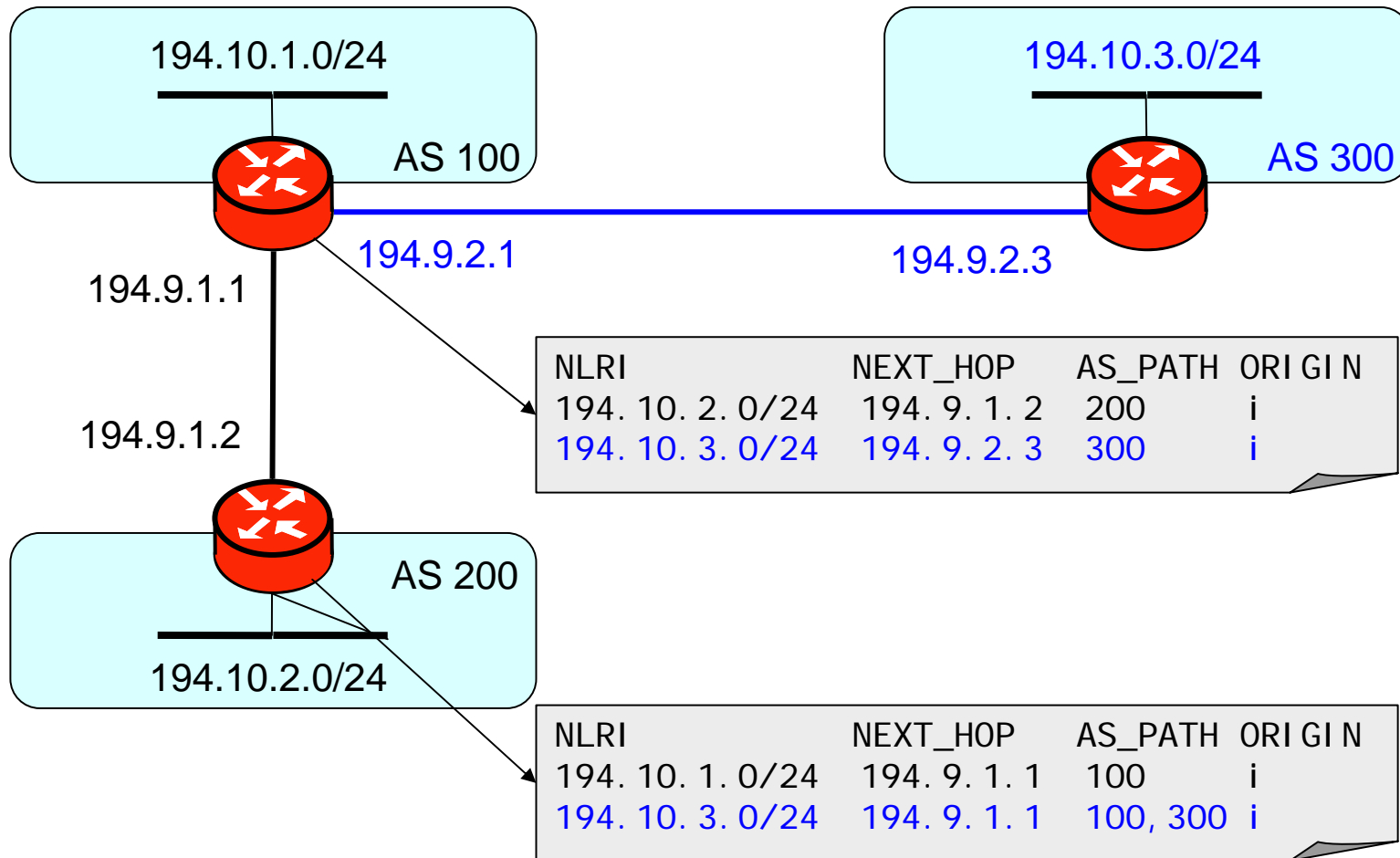
Exemple 1 : tables Adj-RIB-in



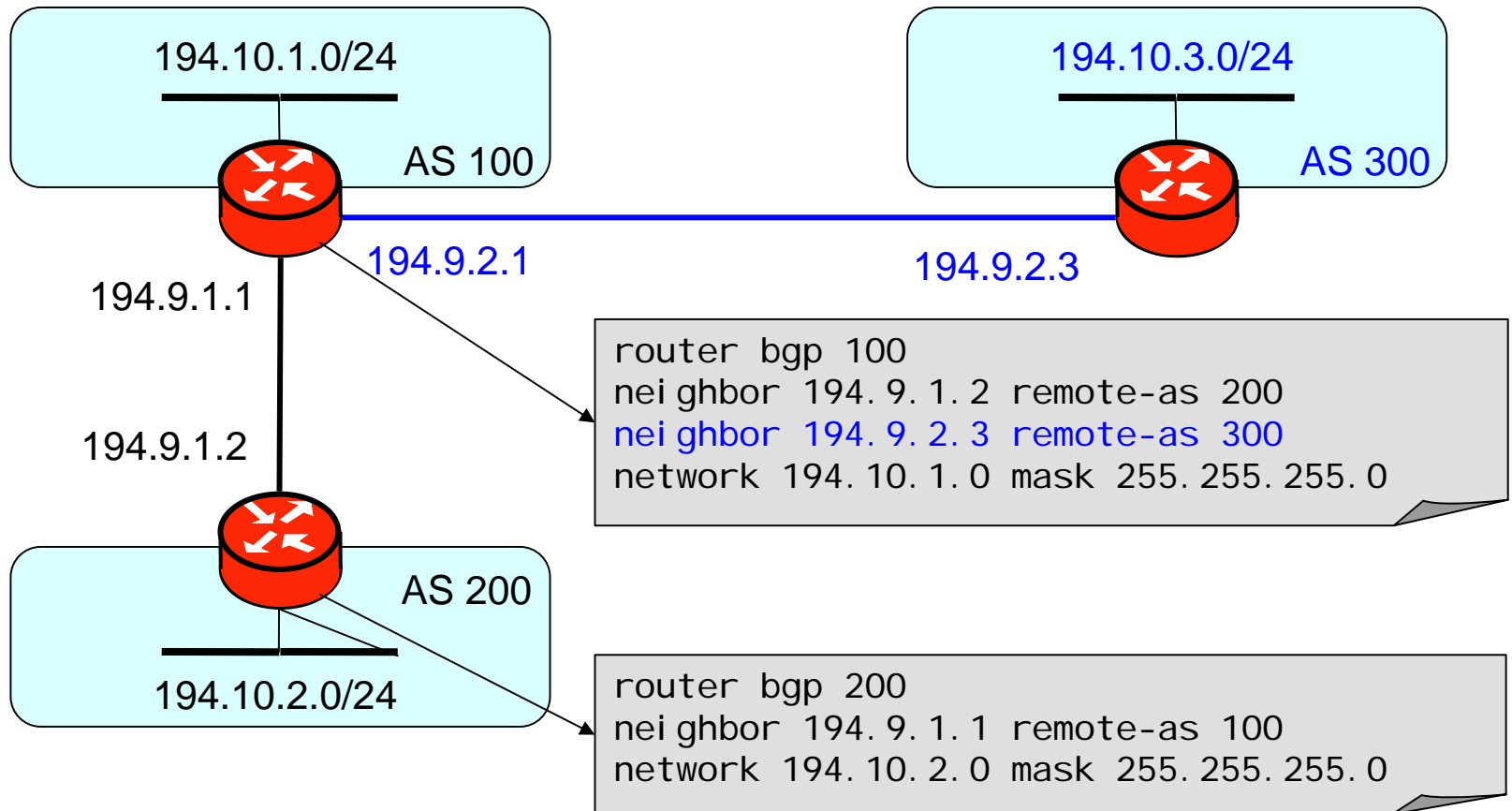
Exemple 1 : configuration sur IOS Cisco



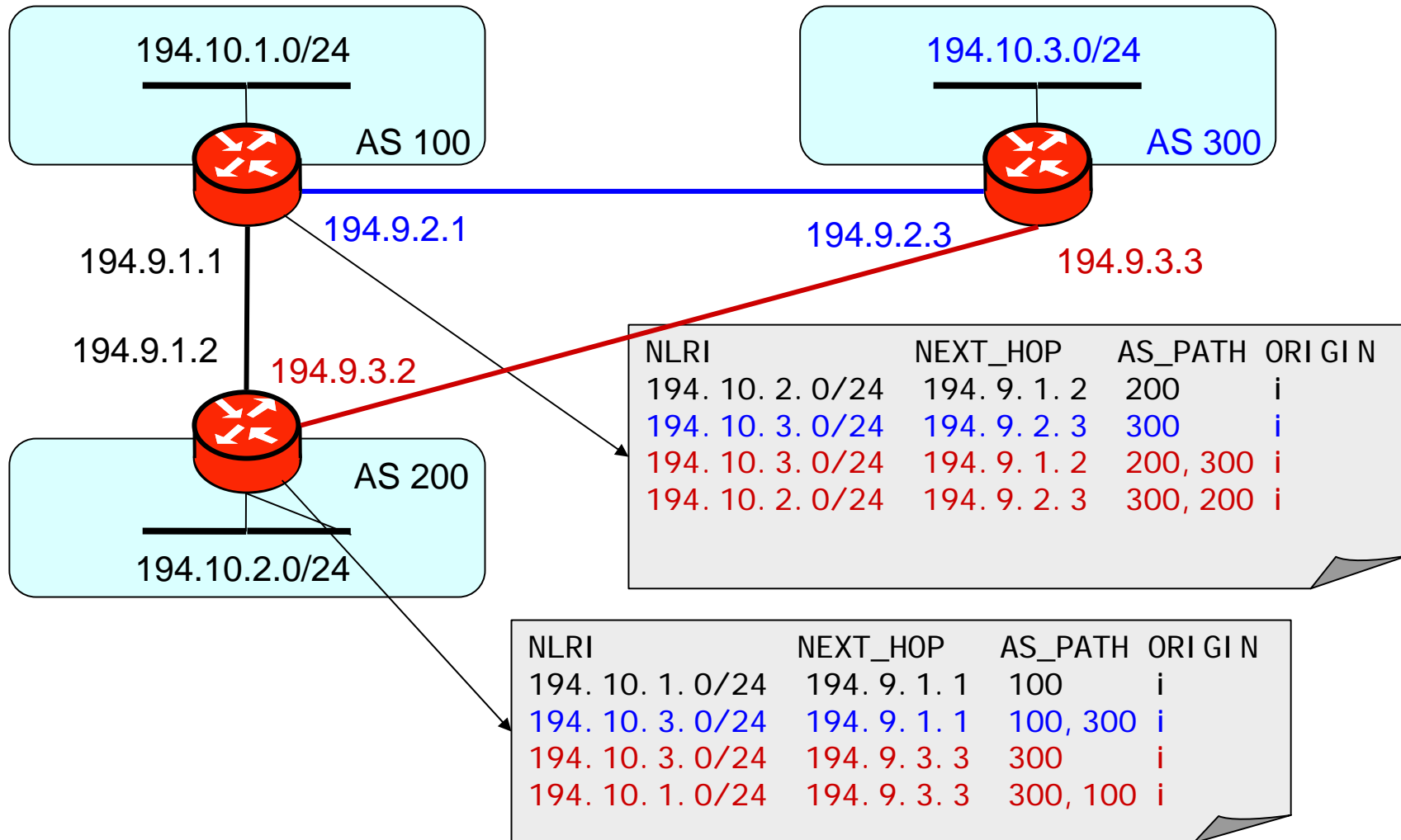
Exemple 2 : tables Adj-RIB-in



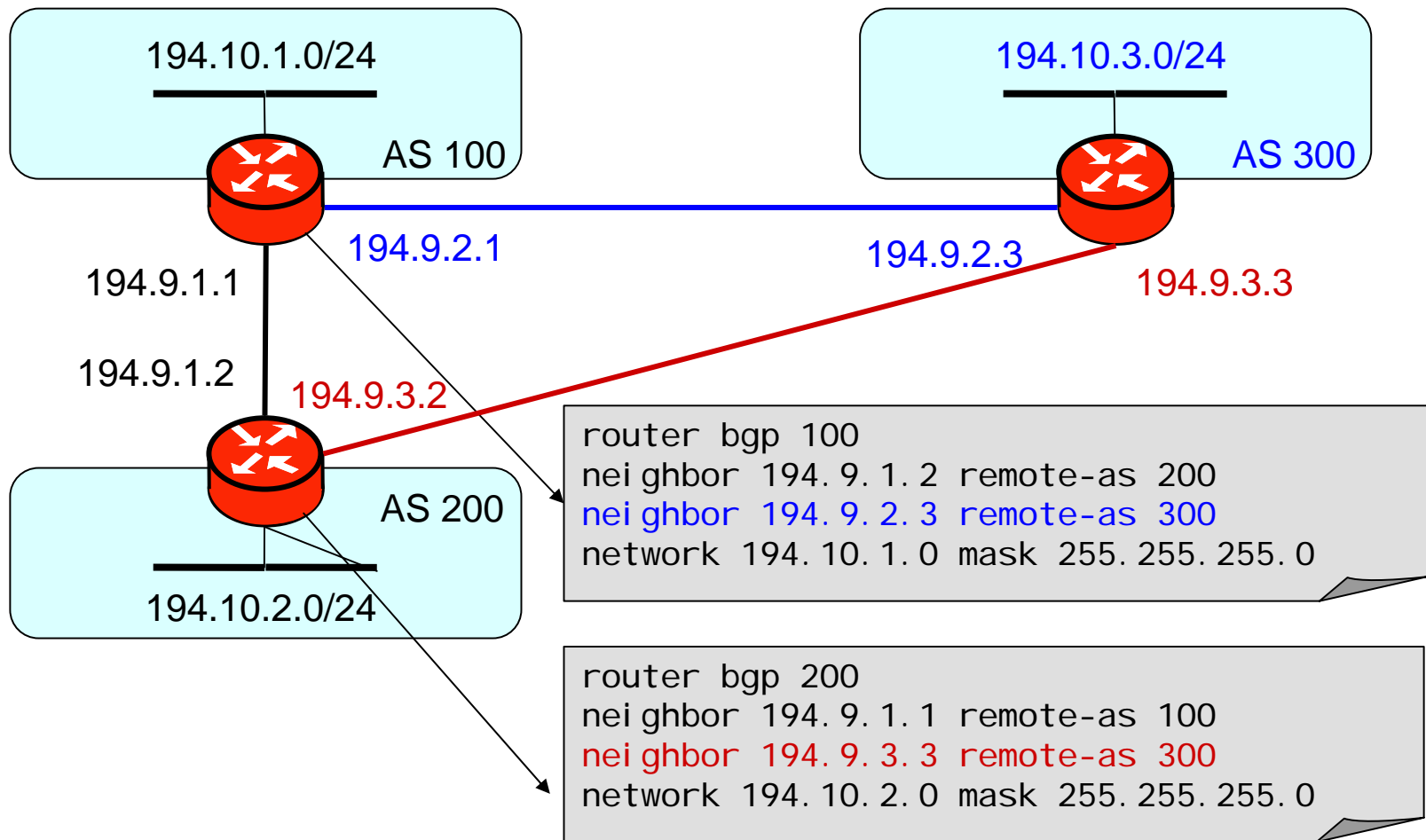
Exemple 2 : configuration sur IOS



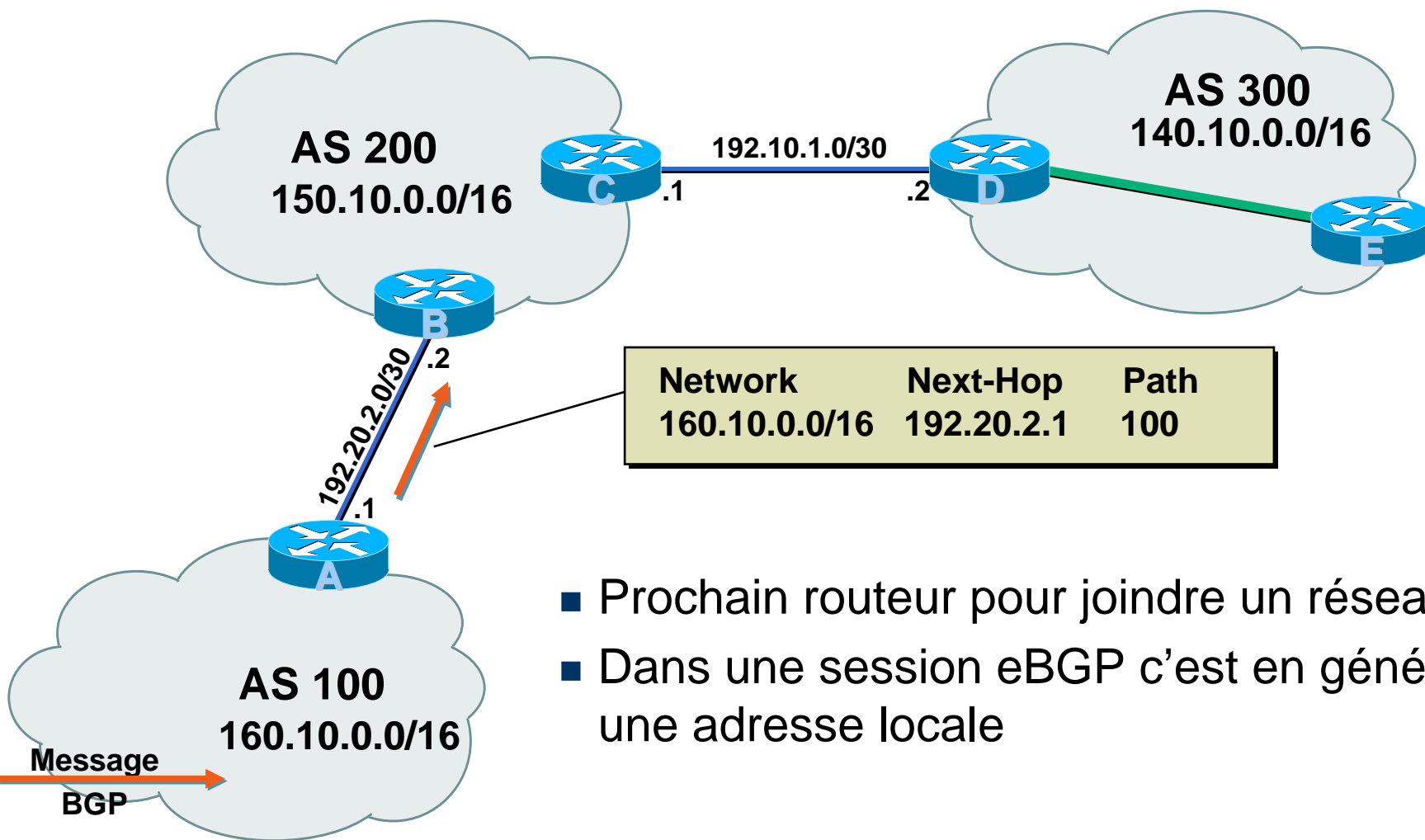
Exemple 3 : tables Adj-RIB-in



Exemple 3 : tables Adj-RIB-in

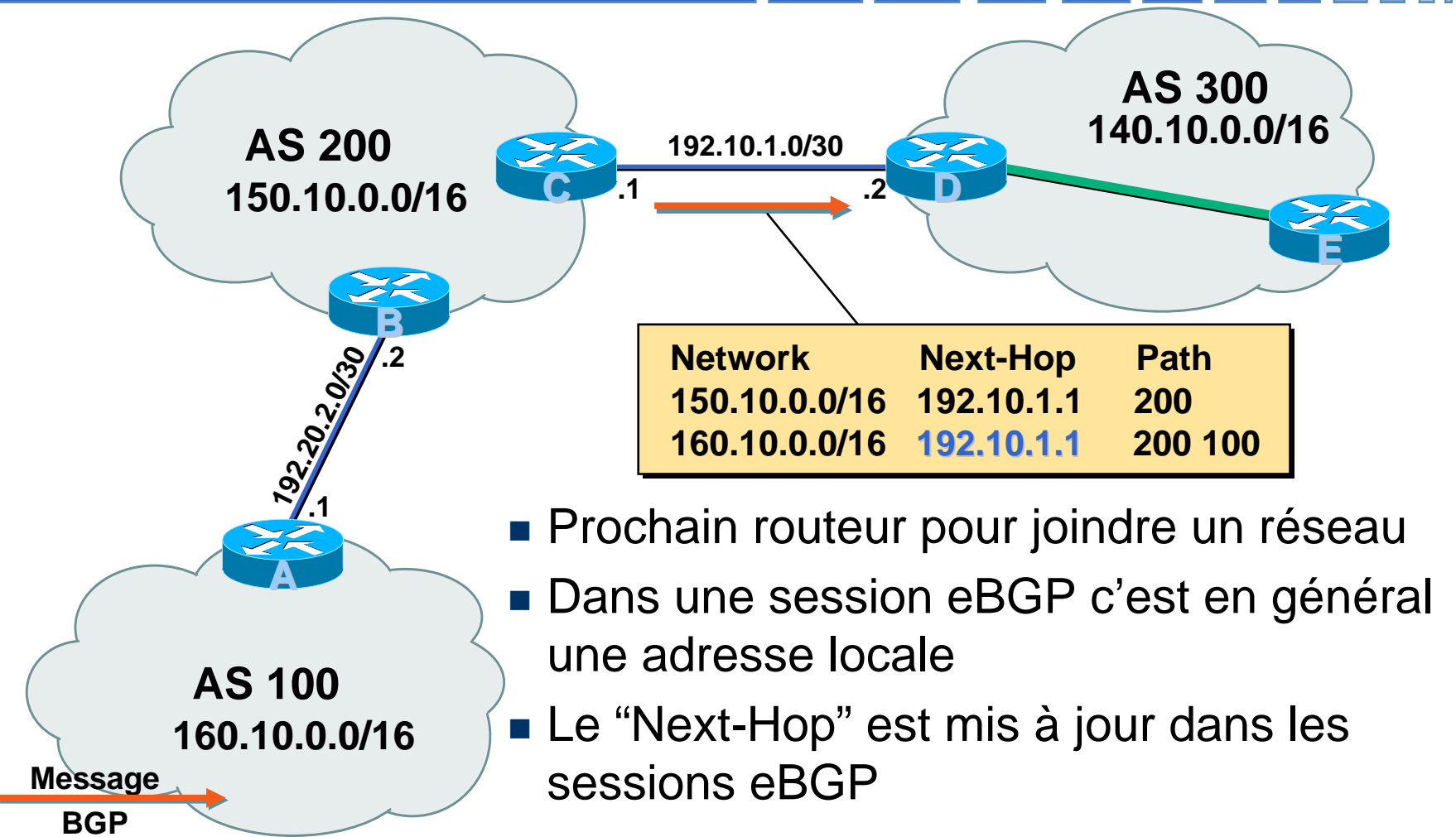


Attribut "Next-Hop"



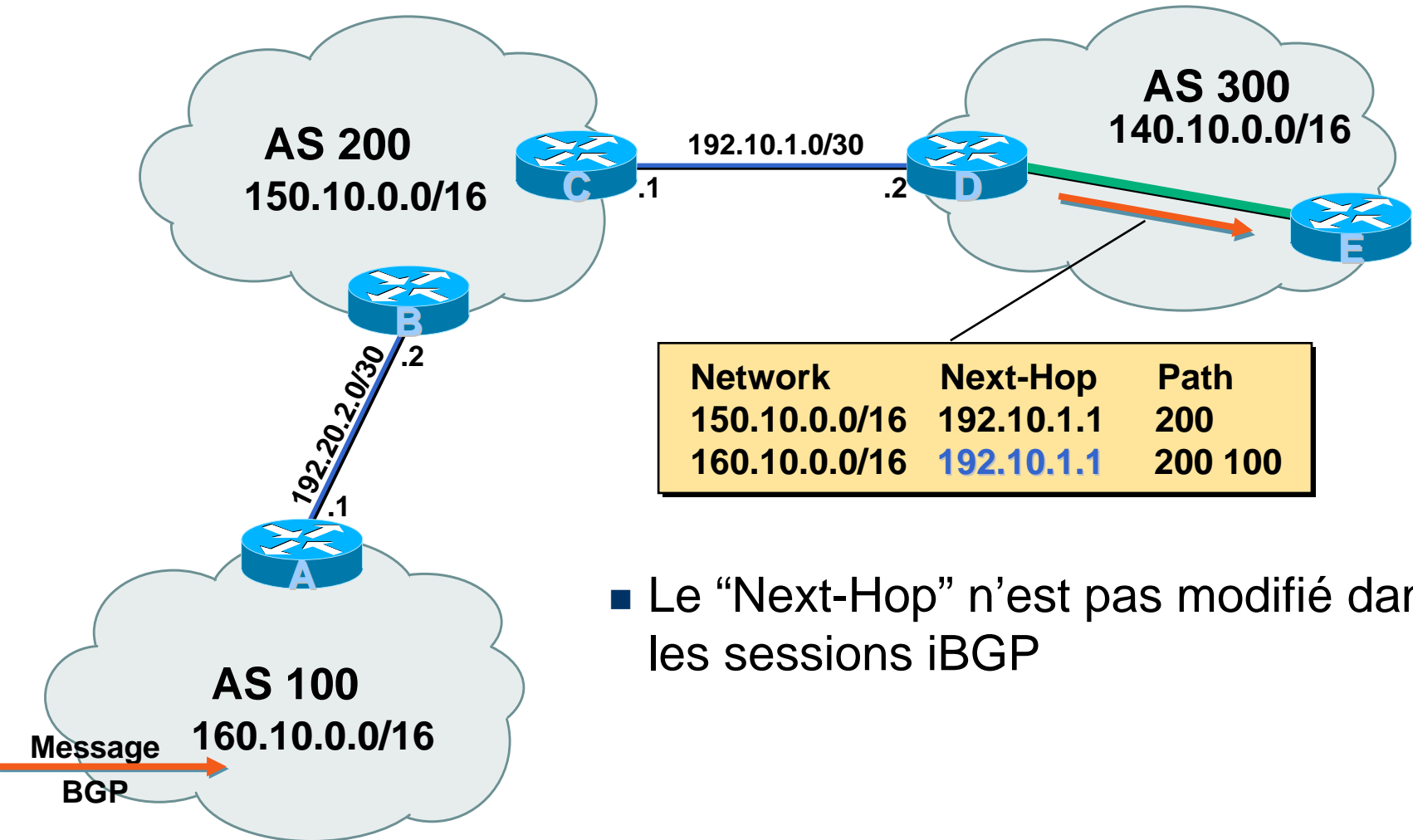
- Prochain routeur pour joindre un réseau
- Dans une session eBGP c'est en général une adresse locale

Attribut "Next-Hop"



- Prochain routeur pour joindre un réseau
- Dans une session eBGP c'est en général une adresse locale
- Le "Next-Hop" est mis à jour dans les sessions eBGP

Attribut "Next-Hop"



- Le "Next-Hop" n'est pas modifié dans les sessions iBGP

Table du routeur BGP

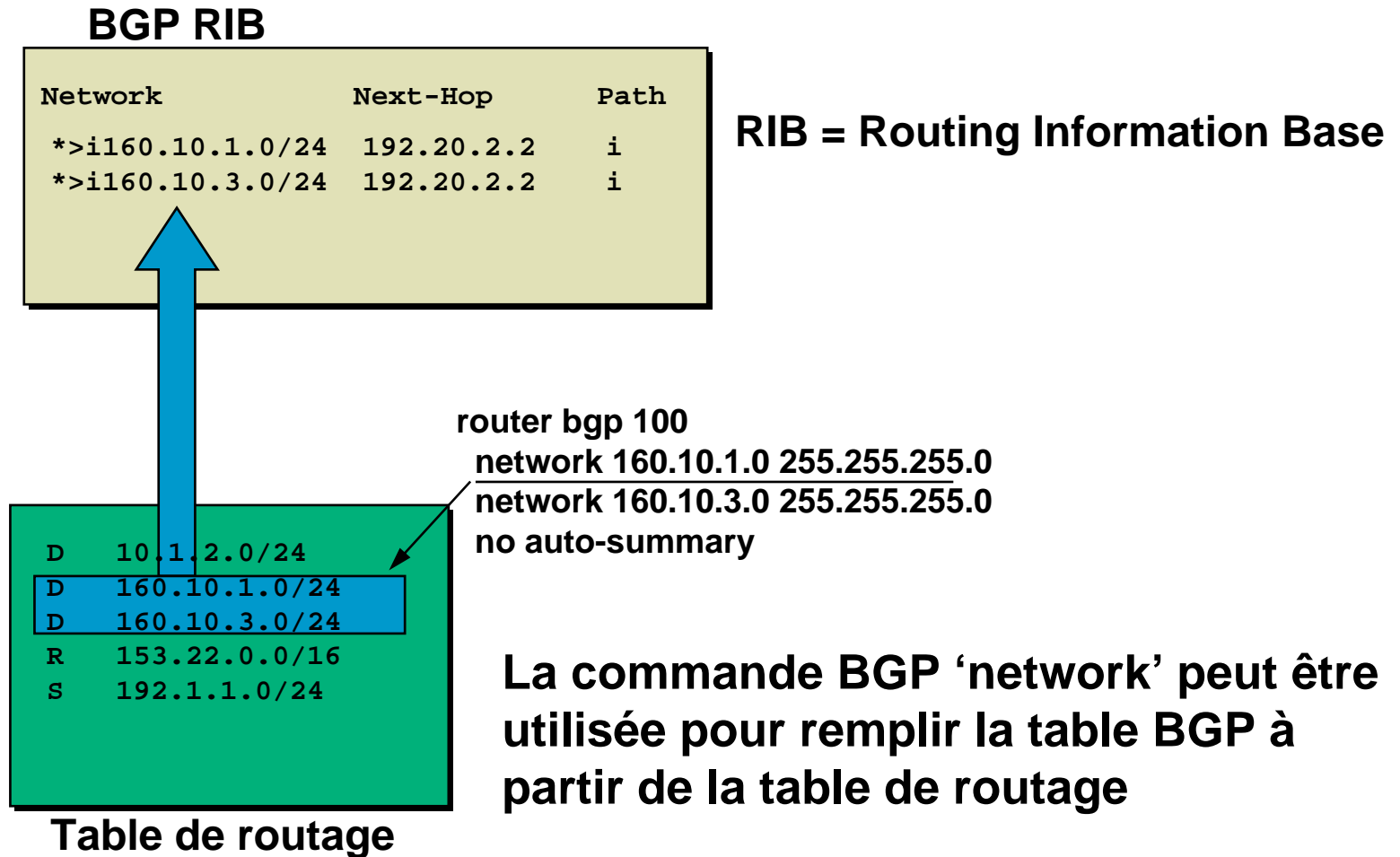


Table du routeur BGP

BGP RIB

Network	Next-Hop	Path
*> 160.10.0.0/16	0.0.0.0	i
* i	192.20.2.2	i
s> 160.10.1.0/24	192.20.2.2	i
s> 160.10.3.0/24	192.20.2.2	i

router bgp 100

network 160.10.0.0 255.255.0.0

aggregate-address 160.10.0.0 255.255.0.0 summary-only
no auto-summary

D	10.1.2.0/24
D	160.10.1.0/24
D	160.10.3.0/24
R	153.22.0.0/16
S	192.1.1.0/24

Route Table

La commande BGP 'aggregate-address' permet d'installer dans la table BGP une route agrégée dès que au-moins un sous-réseau est présent

Table du routeur BGP

BGP RIB

Network	Next-Hop	Path
*> 160.10.0.0/16	0.0.0.0	i
* i	192.20.2.2	i
s> 160.10.1.0/24	192.20.2.2	i
s> 160.10.3.0/24	192.20.2.2	i
*> 192.1.1.0/24	192.20.2.2	?

router bgp 100
network 160.10.0.0 255.255.0.0
redistribute static route-map foo
no auto-summary

access-list 1 permit 192.1.0.0 0.0.255.255

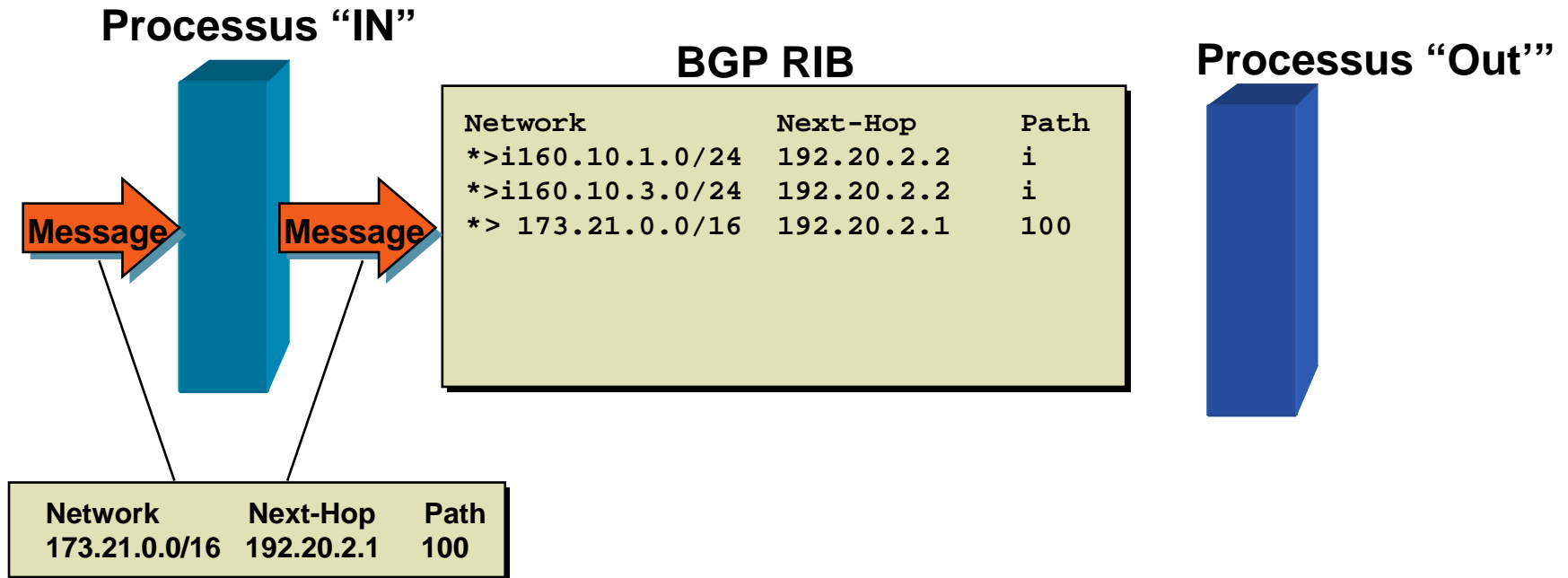
route-map foo permit 10
match ip address 1

La commande BGP 'redistribute' permet de remplir la table BGP à partir de la table de routage en appliquant des règles spécifiques

D	10.1.2.0/24
D	160.10.1.0/24
D	160.10.3.0/24
R	153.22.0.0/16
S	192.1.1.0/24

Route Table

Table du routeur BGP



- **Le processus BGP "in" (entrée)**
 - reçoit les messages des voisins
 - place le ou les chemins sélectionnés dans la table BGP
 - le meilleur chemin (best path) est indiqué avec le signe ">"

Table du routeur BGP

Processus "IN"



BGP RIB

Network	Next-Hop	Path
*>i160.10.1.0/24	192.20.2.2	i
*>i160.10.3.0/24	192.20.2.2	i
*> 173.21.0.0/16	192.20.2.1	100

Processus "OUT"



Network	Next-Hop	Path
160.10.1.0/24	192.20.2.2	200
160.10.3.0/24	192.20.2.2	200
173.21.0.0/16	192.20.2.1	200 100

Modification du "next-hop"

- Le processus BGP "out" (sortie)
 - message construit à partir des informations de la table BGP
 - modification du message selon configuration
 - envoi du message aux voisins

Table du routeur BGP

BGP RIB

Network	Next-Hop	Path
*>i160.10.1.0/24	192.20.2.2	i
*>i160.10.3.0/24	192.20.2.2	i
*> 173.21.0.0/16	192.20.2.1	100

D	10.1.2.0/24
D	160.10.1.0/24
D	160.10.3.0/24
R	153.22.0.0/16
S	192.1.1.0/24
B	173.21.0.0/16

Table de routage

- Le meilleur chemin est installé dans la table de routage du routeur si :
 - Le préfixe et sa taille sont uniques
 - la valeur “distance” du protocole est la plus faible

Les attributs de route optionnels (1)

■ LOCAL_PREF

- ☐ Pondère la priorité donnée aux routes en interne à l'AS
- ☐ Jamais annoncé en E-BGP (en interne donc !)
- ☐ Pris en compte avant la longueur de AS_PATH

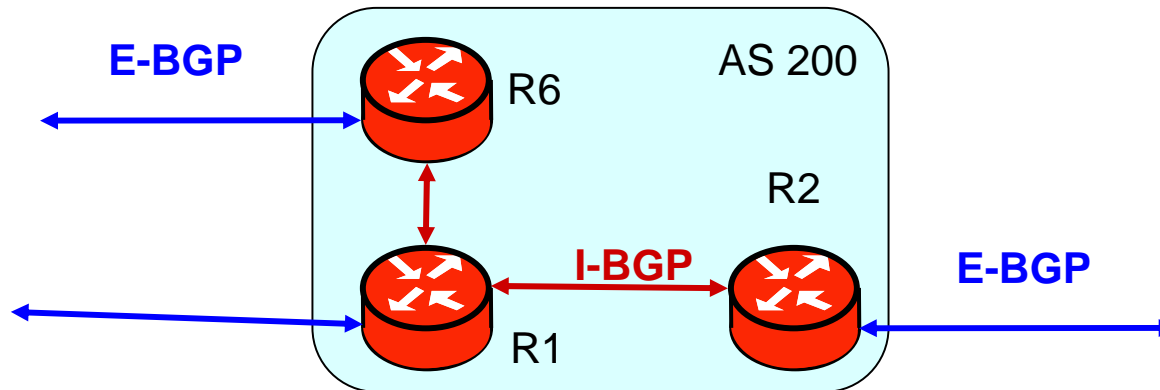
■ ATOMIC_AGGREGATE

- ☐ Indicateur d'agrégation
- ☐ Quand des routes plus précises ne sont pas annoncées

■ AGGREGATOR

- ☐ Donne l'AS qui a formé la route agrégée
- ☐ L'adresse IP du routeur qui a fait l'agrégation

LOCAL-PREF



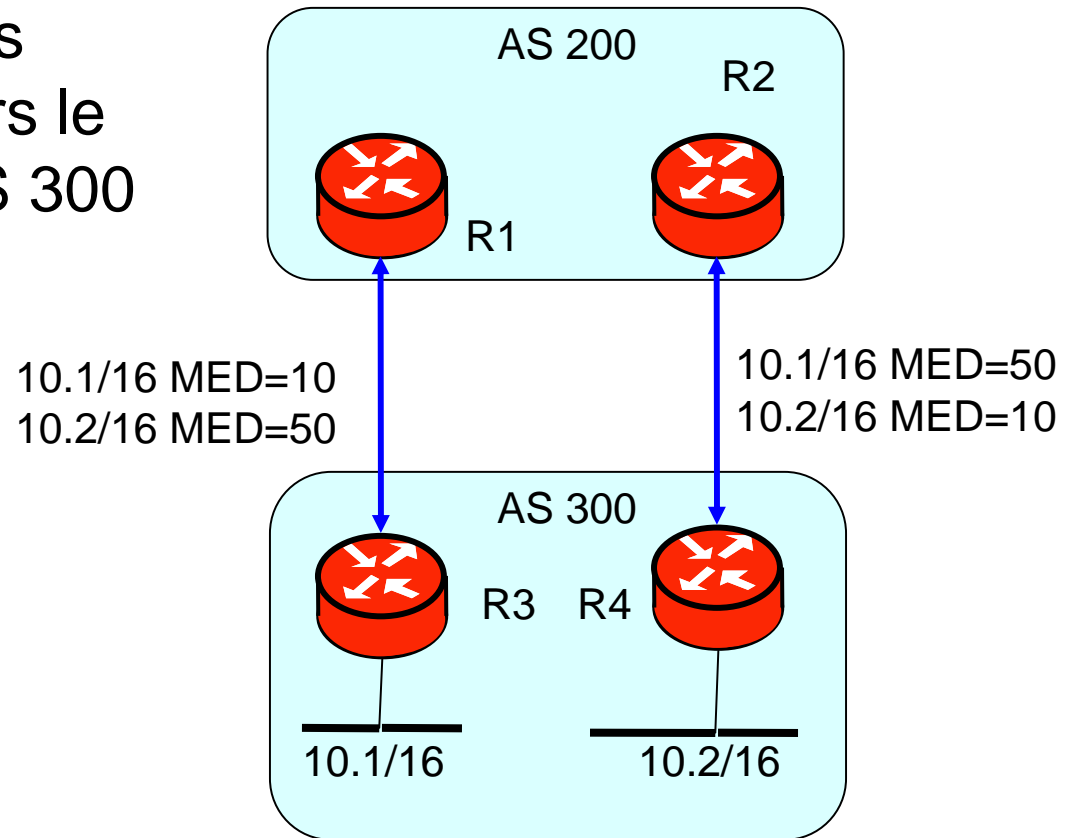
- Mis en œuvre par les routeurs à la réception de route sur E-BGP
 - Propagé sans changement par I-BGP
- R6 associe $\text{pref}=100$, R2 $\text{pref}=10$
- R1 choisit la plus grande préférence
- Bgp default local-preference *pref-value*

Les attributs de route optionnels (2)

- MULTI_EXT_DISC ou MED (non transitif)
 - Permet de discriminer les différents points de connexion d'un AS multi-connecté (plus faible valeur préférée)
- WEIGHT (non transitif, spécifique Cisco)
 - Pondère localement (au routeur) la priorité des routes BGP
- COMMUNITY (transitif)
 - Pour un ensemble de routeurs ayant une même propriété
 - no-export : pas annoncé aux voisins de la confédération
 - no-advertise : pas annoncé aux voisins BGP
 - no-export-subconfed : pas annoncé en E-BGP

MULTI-EXIT-DISC (MED) I

- Si AS 200 accepte les MEDs, le trafic va vers le lien privilégié par l'AS 300
- Plus petite MED



Le processus de décision

- Il est enclenché par une annonce de route
- Il se déroule en trois phases
 - Calcul du degré de préférence de chaque route apprise
 - Choix des meilleures routes à installer dans **RIB-Loc**
 - Choix des routes qui vont être annoncées
- Il applique aux informations de routage un traitement basé sur
 - Critères techniques : suppression boucles, optimisations...
 - Critères administratifs : politique de routage de l'AS
 - une annonce de route doit avoir son **NEXT_HOP** *routable*
 - Une route interne est annoncée par un routeur s'il sait la joindre.
 - Une route externe est annoncée par un routeur s'il sait joindre le NEXT_HOP.
 - Une route dont l'attribut NEXT_HOP est l'adresse IP du voisin n'est pas annoncée à ce voisin (qui la connaît déjà!).

Critères de choix entre 2 routes

- WEIGHT (propriétaire Cisco, plus grand préféré)
- LOCAL_PREF le plus grand
- Route initiée par le processus BGP local
- AS_PATH minimum
- ORIGIN minimum (IGP -> EGP -> Incomplete)
- MULTI_EXT_DISC minimum
- Route externe préférée à une route interne (à l'AS)
- Route vers le plus proche voisin local (au sens de l'IGP)
- Route vers le routeur BGP de plus petite adresse IP

Différences entre E-BGP et I-BGP

- Une annonce reçue en I-BGP n'est pas réannoncée en I-BGP
- L'attribut LOCAL_PREF n'est annoncé qu'en I-BGP
- Seuls les voisins E-BGP doivent être directement connectés
- Les annonces I-BGP ne modifient pas l'AS_PATH
- Les annonces I-BGP ne modifient pas le NEXT_HOP
- Le MED n'est pas annoncé en I-BGP

Interaction BGP – OSPF

■ Redistribution

☐ Routes apprises par BGP → IGP (OSPF)

- ☐ Redistribution de BGP dans OSPF
- ☐ OSPF propage les routes via des LSAs de type 4 à tous les routeurs du nuage OSPF (**aire**)

■ Injection

☐ Route apprise par BGP est écrite dans la table de routage du routeur.

- ☐ Pas de propagation

■ Synchronisation

L'annonce des routes internes d'un AS

■ Statique

- Pas d'instabilité de routage, mais trous noirs possibles
 - redistribute [static|connected]
 - → ORIGIN : Incomplete
 - network <adresse réseau>
 - → ORIGIN : IGP

■ Dynamique

- Suit au mieux l'état du réseau, nécessite du filtrage
 - **redistribute <paramètres de l'IGP>**
 - → ORIGIN : IGPz

La politique de routage

- Elle peut influencer :
 - Le traitement des routes reçues
 - Le traitement des routes annoncées
 - L'interaction avec les IGP de l'AS

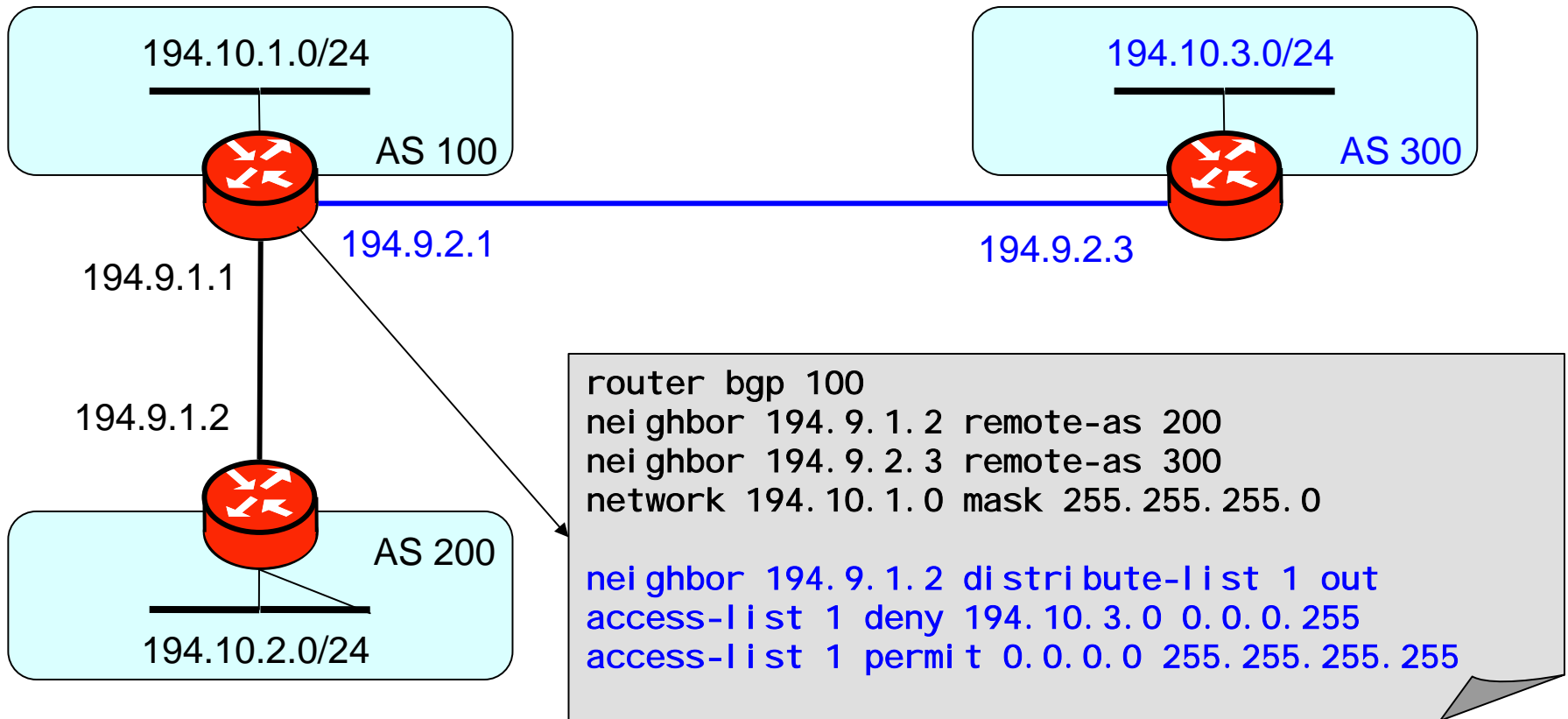
- En pratique elle s'exprime par :
 - Du filtrage de réseaux
 - Du filtrage de routes (AS_PATH)
 - De la manipulation d'attributs de routes

Filtrage de routes

- Associer une *access list* à un voisin
- Neighbor *ID* distribute-list *no-of-the-list* [in/ou]
- Définir une access list
 - Bit non significatif (inverse du netmask)
 - Si pas d'action en fin de liste
 - Appliquer le 'deny everything else'
- Access-list *No-of-the-list* [deny/permit] IP-@ *non-sig-bit*

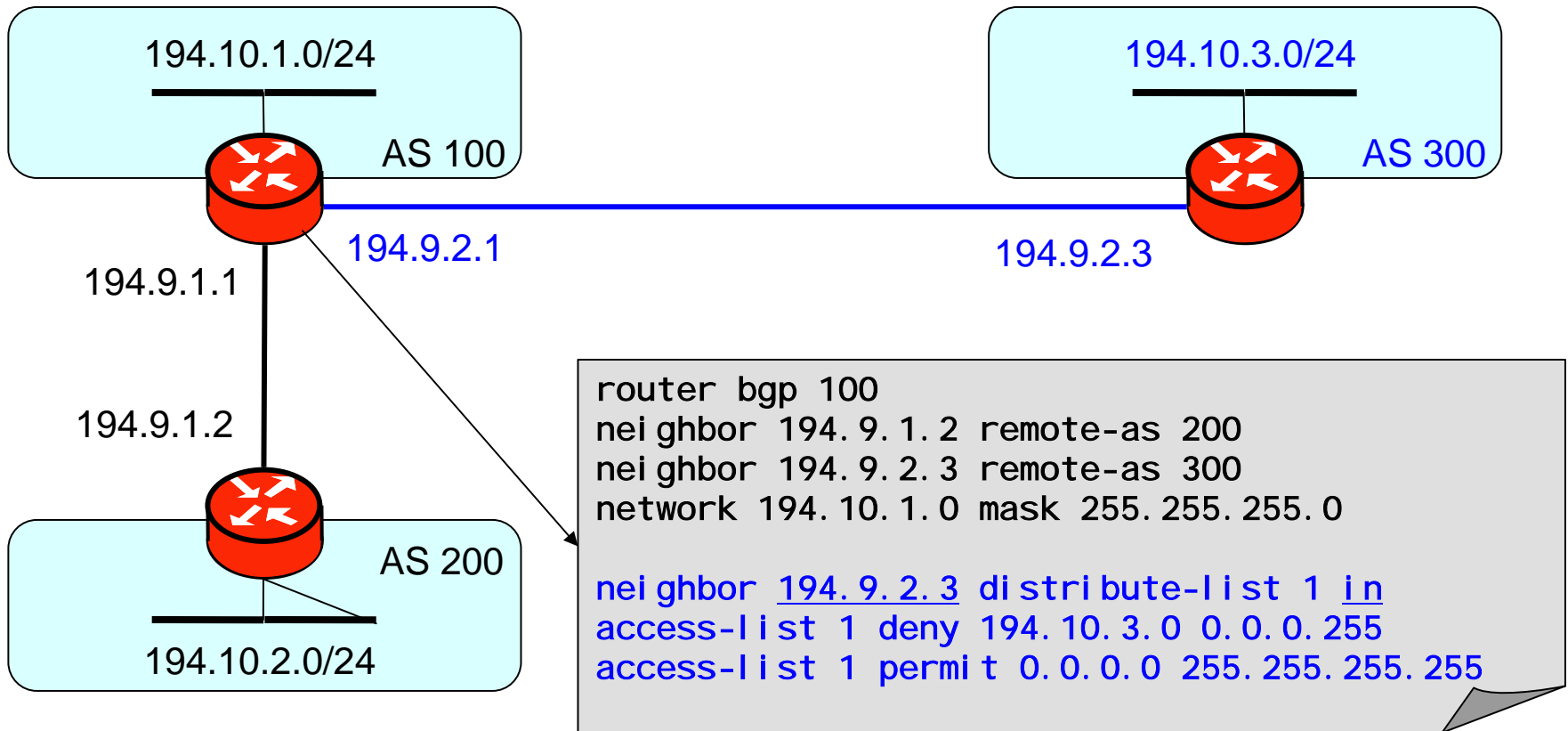
Politique de routage

- Filtrage des réseaux annoncés
 - AS100 ne veut pas servir d'AS de transit pour le réseau 194.10.3.0/24 de l'AS300



Politique de routage

- Filtrage des réseaux annoncés
 - Idem mais en plus l'AS100 ne connaît plus 194.10.3.0/24 de l'AS300



Filtrage de chemin

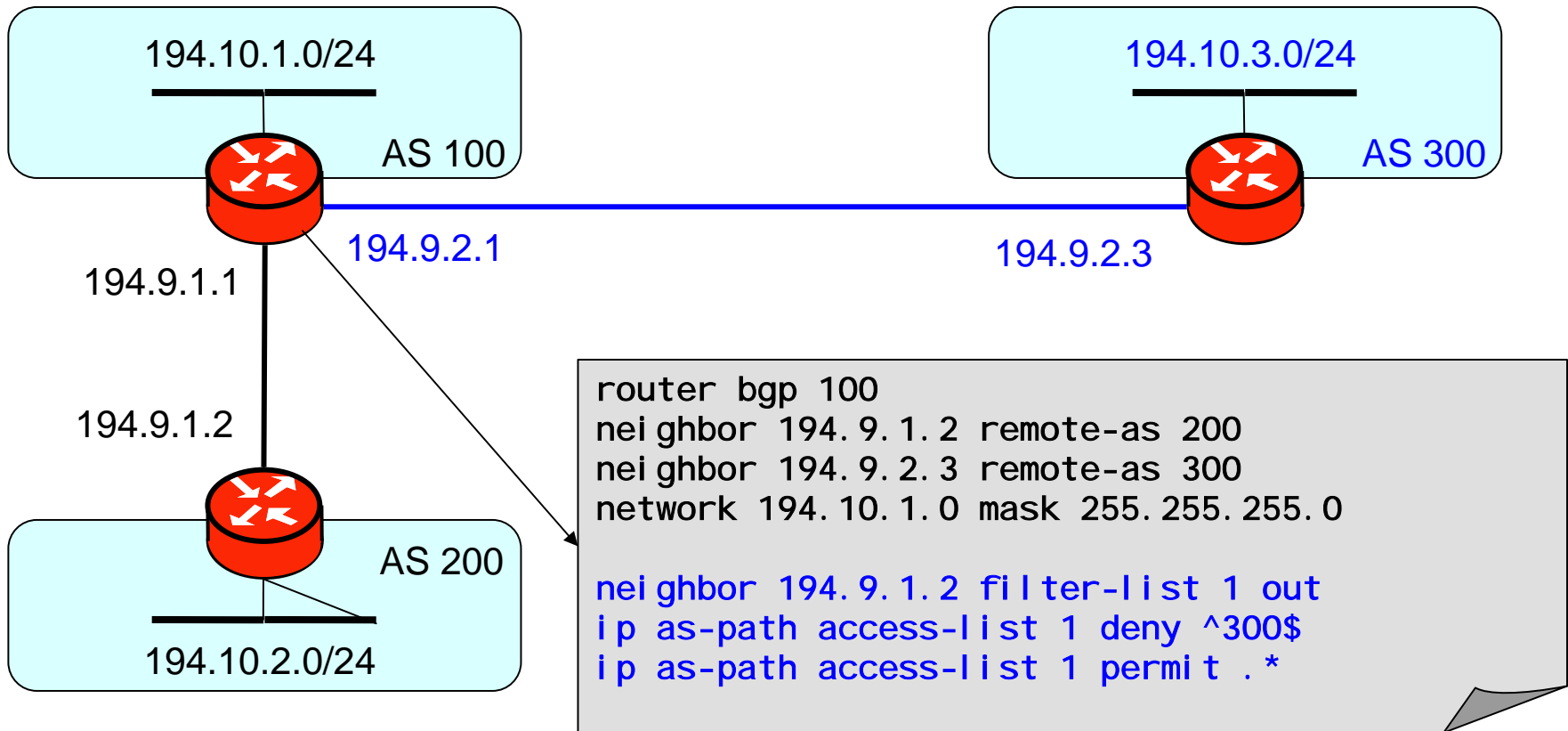
- Associer une *filter list* à un voisin
- Neighbor *ID* filter-list *no-of-the-list* [in/ou]
- Définir une *filter list*
- Ip as-path access-list *No-of-the-list* [deny/permit] regexpr
- Regular expression
 - ^ début du chemin
 - \$ fin du chemin
 - . Tout caractère
 - ? Un caractère
 - _ matches ^ \$ () 'space'
 - * toute répétition
 - + au moins une répétition

Filtrage de chemins

- ^\$
 - route locale seulement (AS_PATH vide)
- .*
 - toutes les routes
- ^300\$
 - AS_PATH = 300
- ^300_
 - toutes les routes en provenance de 300
- _300\$
 - Toutes les routes originaires de 300
- _300_
 - Toutes les routes passant par 300

Politique de routage

- Filtrage des AS_PATH annoncés
 - AS100 ne veut pas servir d'AS de transit pour tous les réseaux internes d'AS300

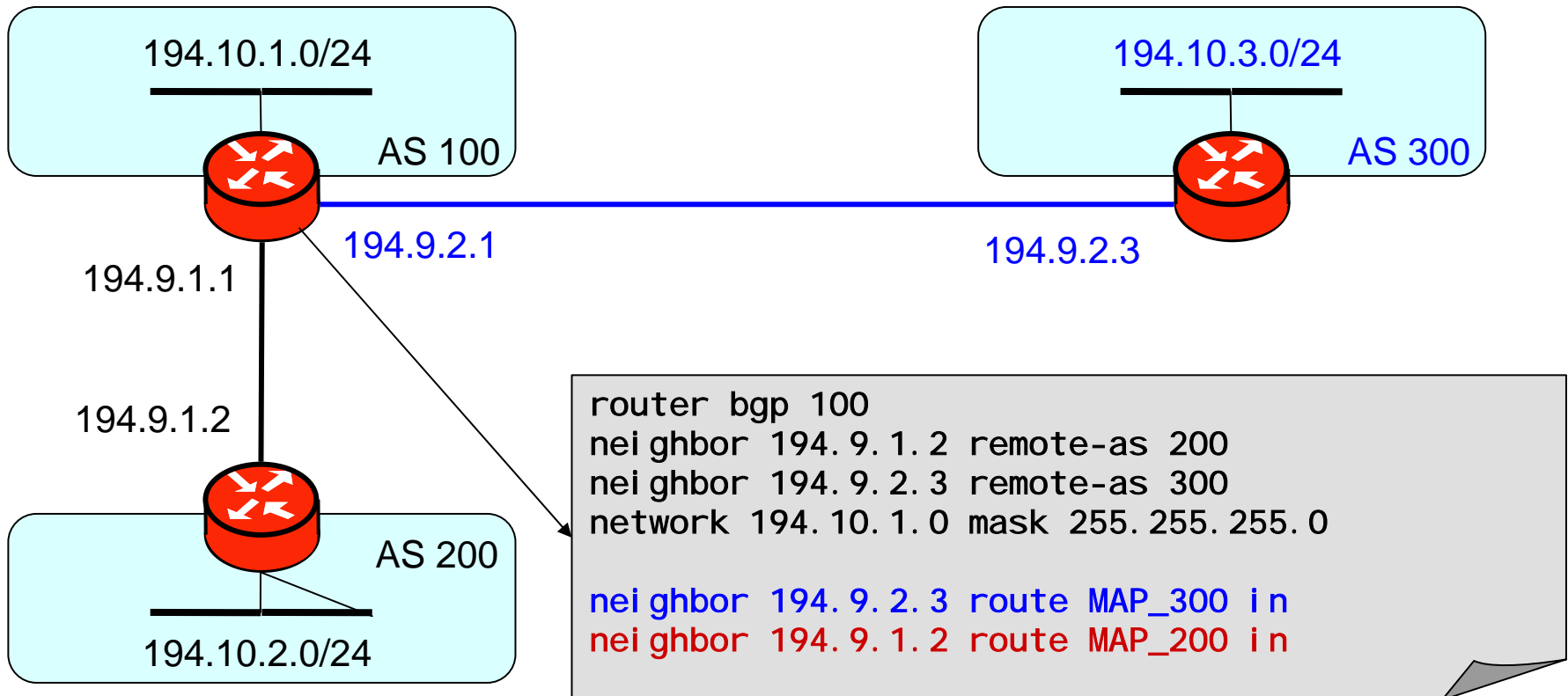


Route maps

- Route-map *map-tag* [permit|deny] *instance-no*
- *First instance-condition*: set match
- *Next-instance-condition*: set match
- ...
- Route-map SetMetric permit 10
- Match ip address 1
- Set metric 200
- Route-map SetMetric permit 20
- Set metric 300
- Access-list 1 permit 194.10.3.0 0.0.0.255

Politique de routage

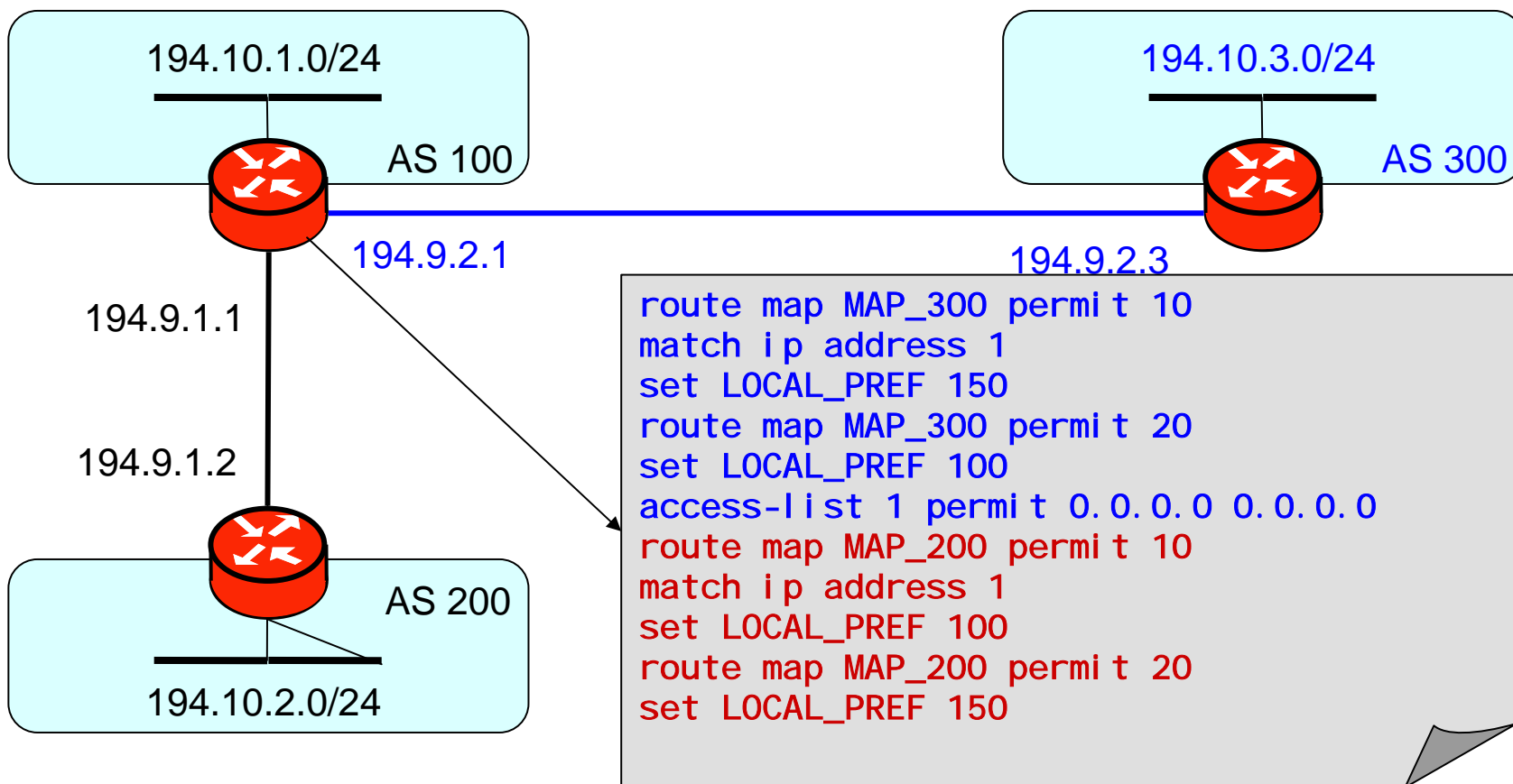
- Filtrage par route map :
- AS100 veut privilégier la route par défaut annoncée par AS300



Politique de routage (suite)

- Filtrage par route map :

- AS100 veut privilégier la route par défaut annoncée par AS300



Références :

- Building Reliable networks with the Border Gateway Protocol BGP, *Iljitsch van Beijnum*
 - Editions O'REILLY
- BGP-4 Command and Configuration Handbook, *William R. Parkhurst*
 - CiscoPress
- RFC 1771, 1105, 1163, 1267, 1918
- www.cisco.com
- <http://www.ietf.org/rfc/rfc1918.txt>