LE ROUTAGE BGP-4

Août 2003

Luc.Saccavini@inria.fr

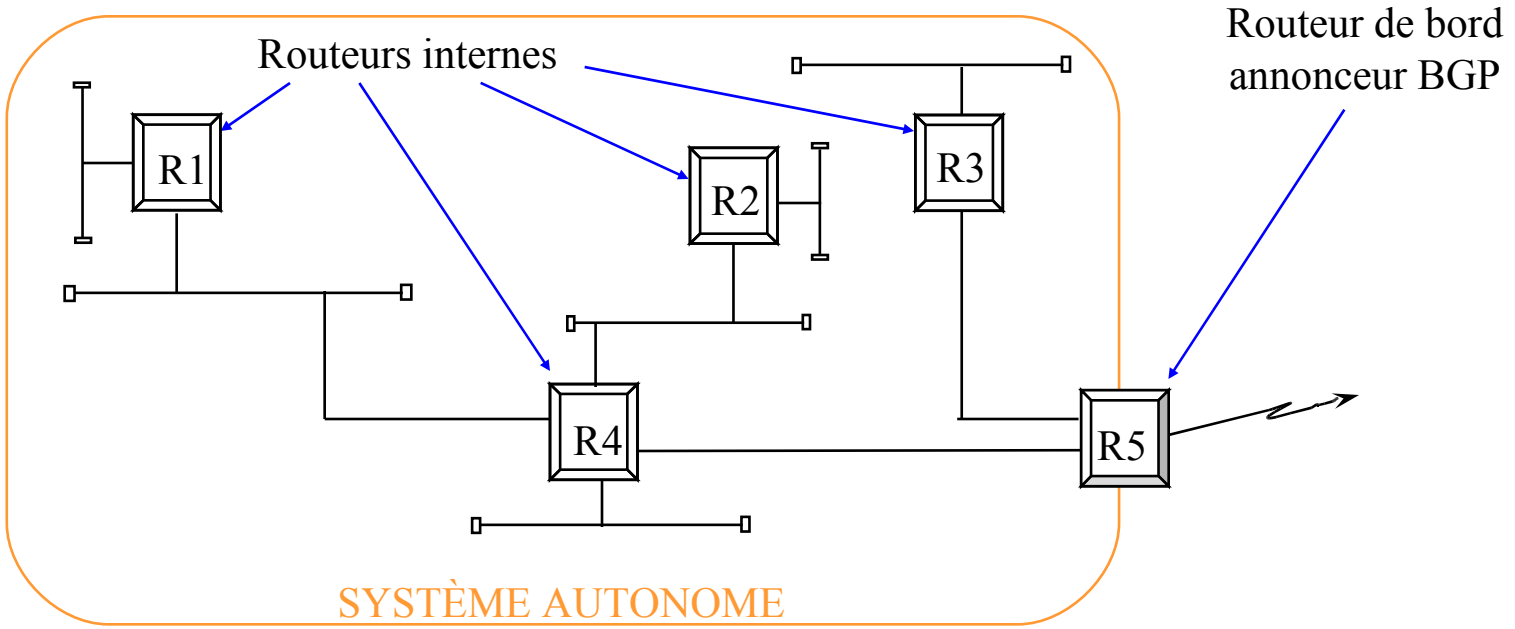
V1.6 (août 2003)

Classification des protocoles de routage

- Il existe 2 grandes familles de protocoles de routage
  - Les protocoles intérieurs (IGP)
    - Distance-vecteur : RIP, IGRP
    - État des liens : OSPF, IS-IS
    - Taille <100 routeurs, 1 autorité d'administration
    - Échange de routes, granularité = routeur
  - Les protocoles extérieurs (EGP)
    - EGP, BGP, IDRP
    - Taille = Internet, coopération d'entités indépendantes
    - Échange d'informations de routage, granularité = AS

Notion de système autonome (AS)

- Ensemble de routeurs sous une même entité administrative



Objectifs généraux du protocole BGP

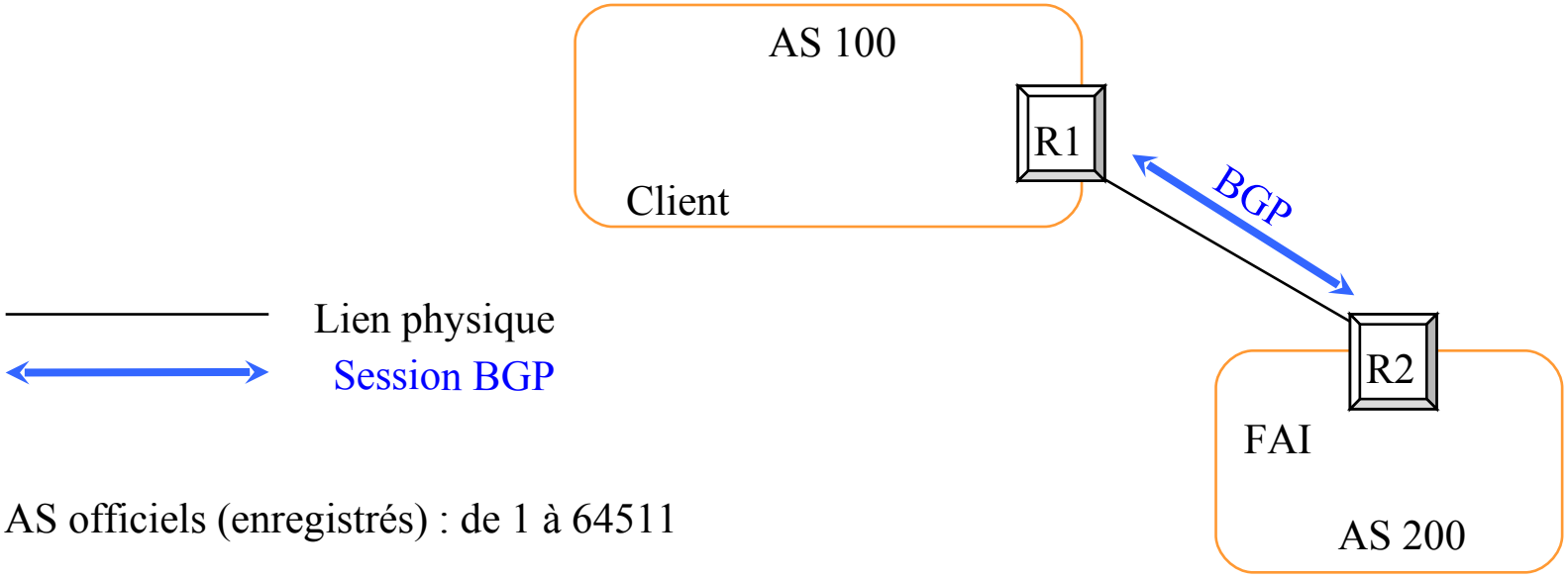
- Échanger des routes (du trafic) entre organismes indépendants
  - Opérateurs
  - Gros sites mono ou multi connectés
- Implémenter la politique de routage de chaque organisme
  - Respect des contrats passés entre organismes
  - Sûreté de fonctionnement
- Être indépendant des IGP utilisés en interne à un 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 le Système Autonome (AS)
- Le support de la session BGP est TCP (port 179)
- Les sessions BGP sont établies entre les routeurs de bord d'AS
- Protocole point à point entre routeurs de bord d'AS
- Protocole symétrique
- (un annonceur BGP n'est pas forcément un routeur)

Exemple de connexion BGP (1)

- Client connecté à un seul Fournisseur d'Accès Internet (FAI). Seuls les routeurs de bord de l'AS sont figurés.

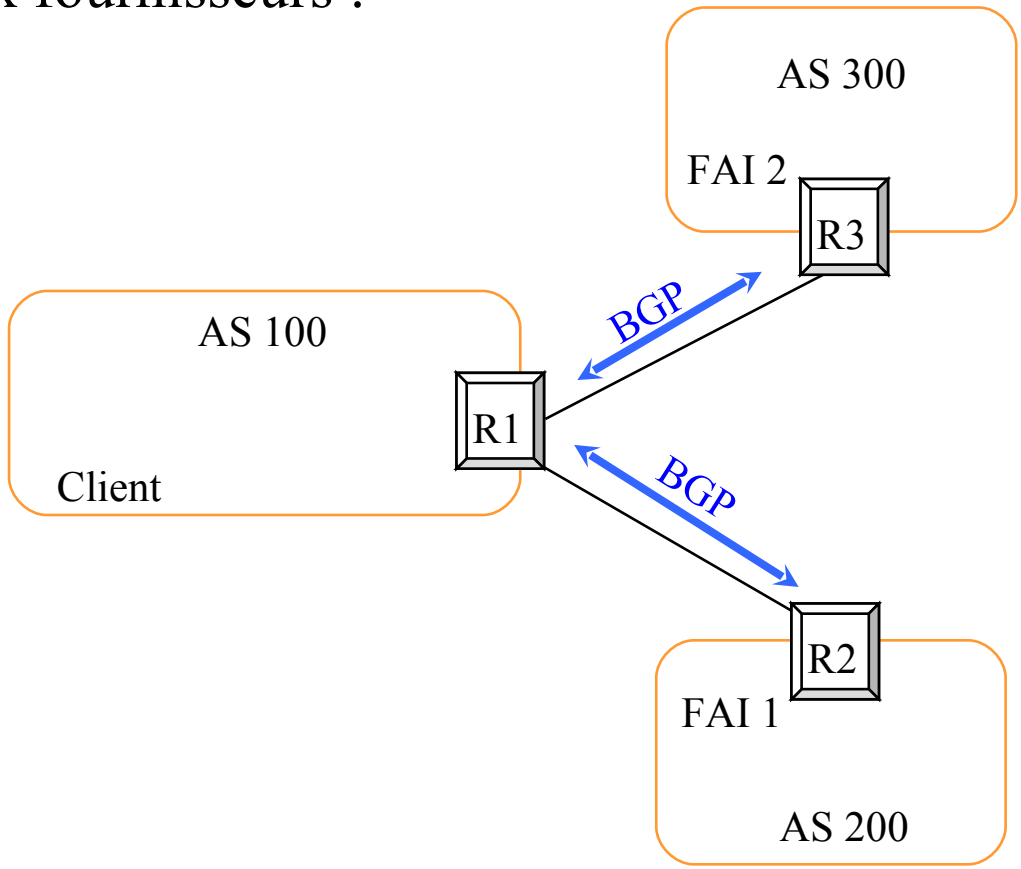


AS officiels (enregistrés) : de 1 à 64511

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

Exemple de connexion BGP (2)

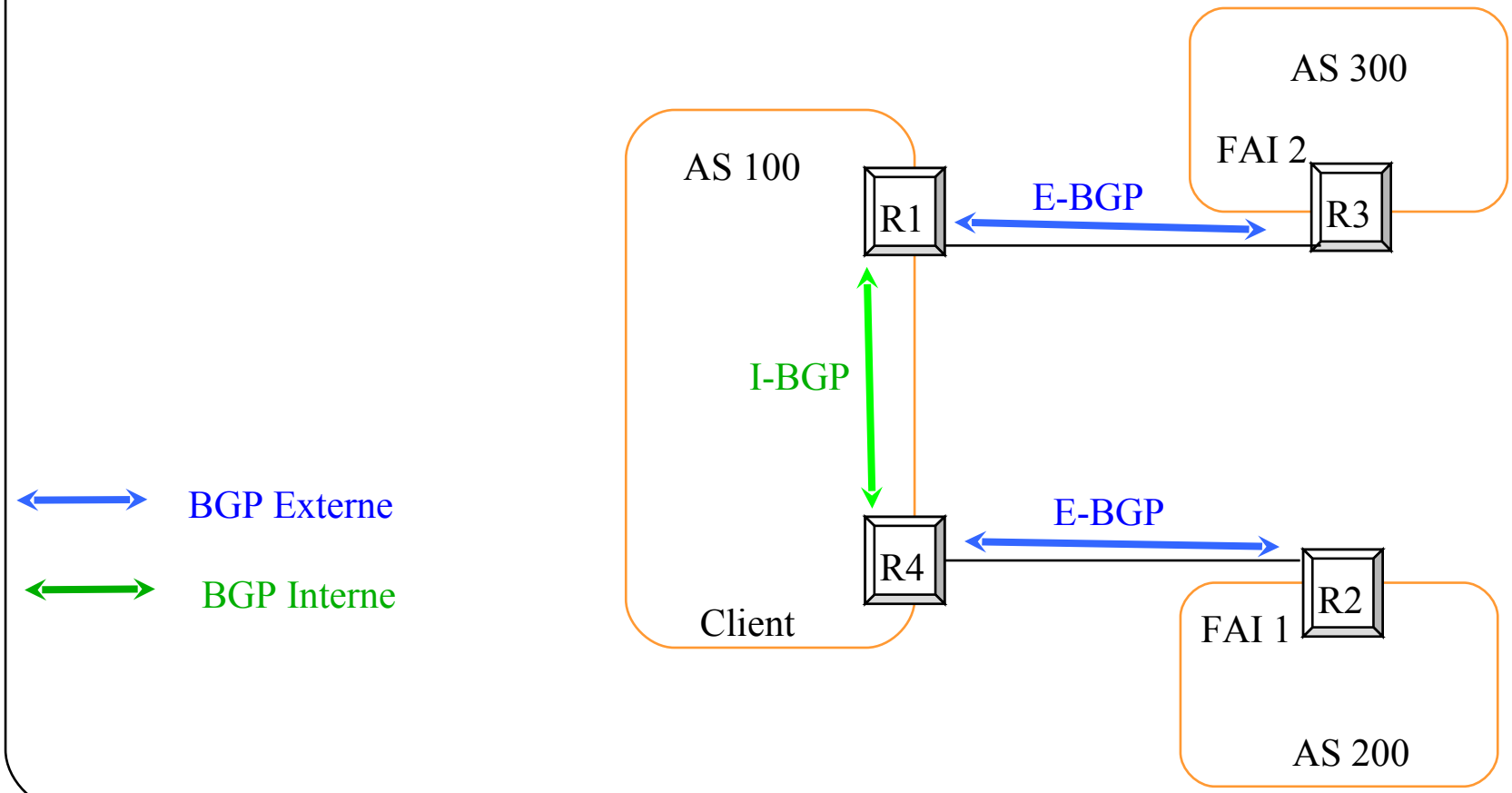
- Client connecté à deux fournisseurs :



R1 à deux voisins : R2 et R3

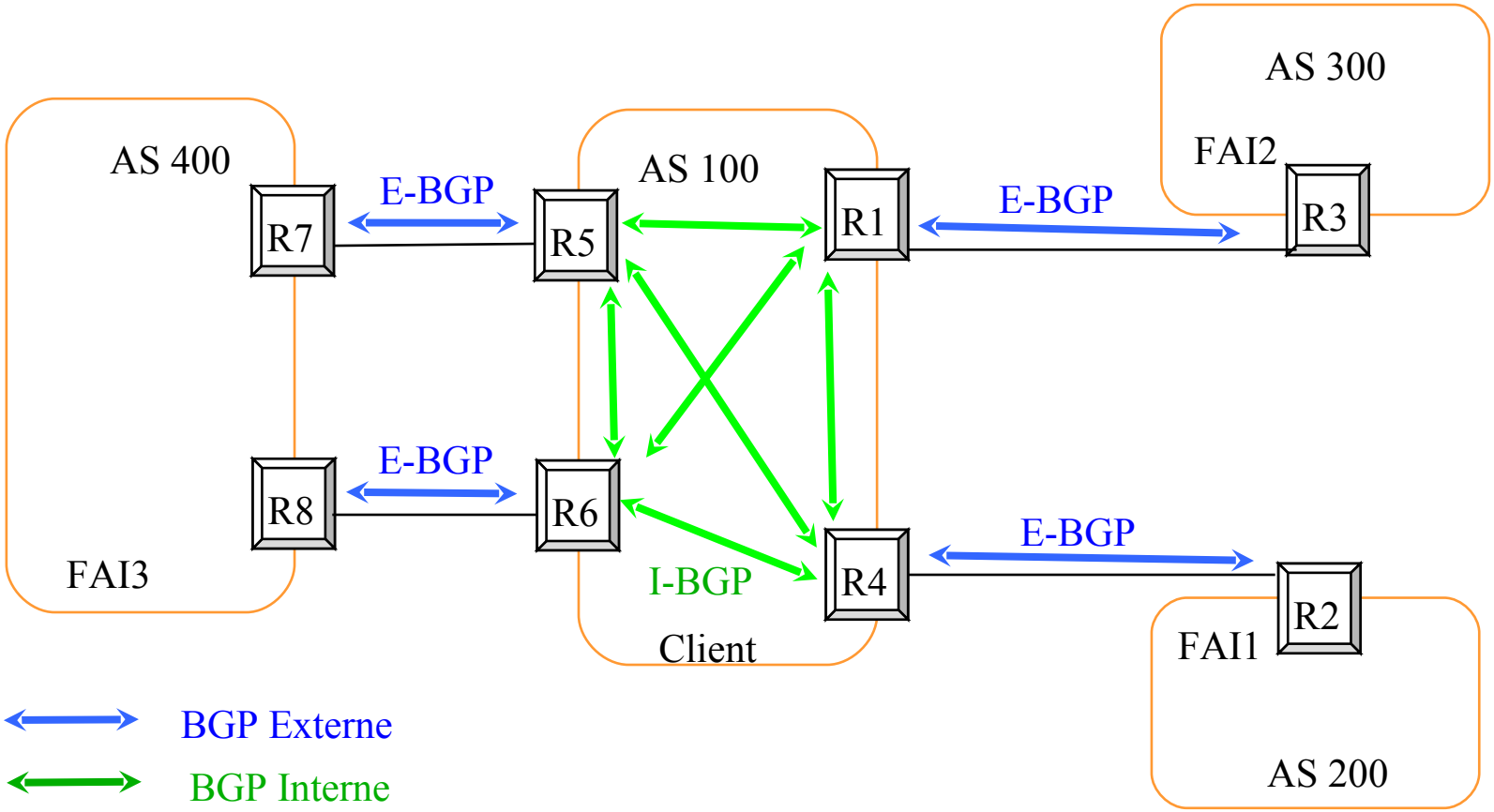
Exemple de connexion BGP (3)

- Client connecté à 2 fournisseurs par 2 routeurs différents :



Exemple de connexion BGP (4)

- Client connecté à 3 fournisseurs avec redondance sur l'un :



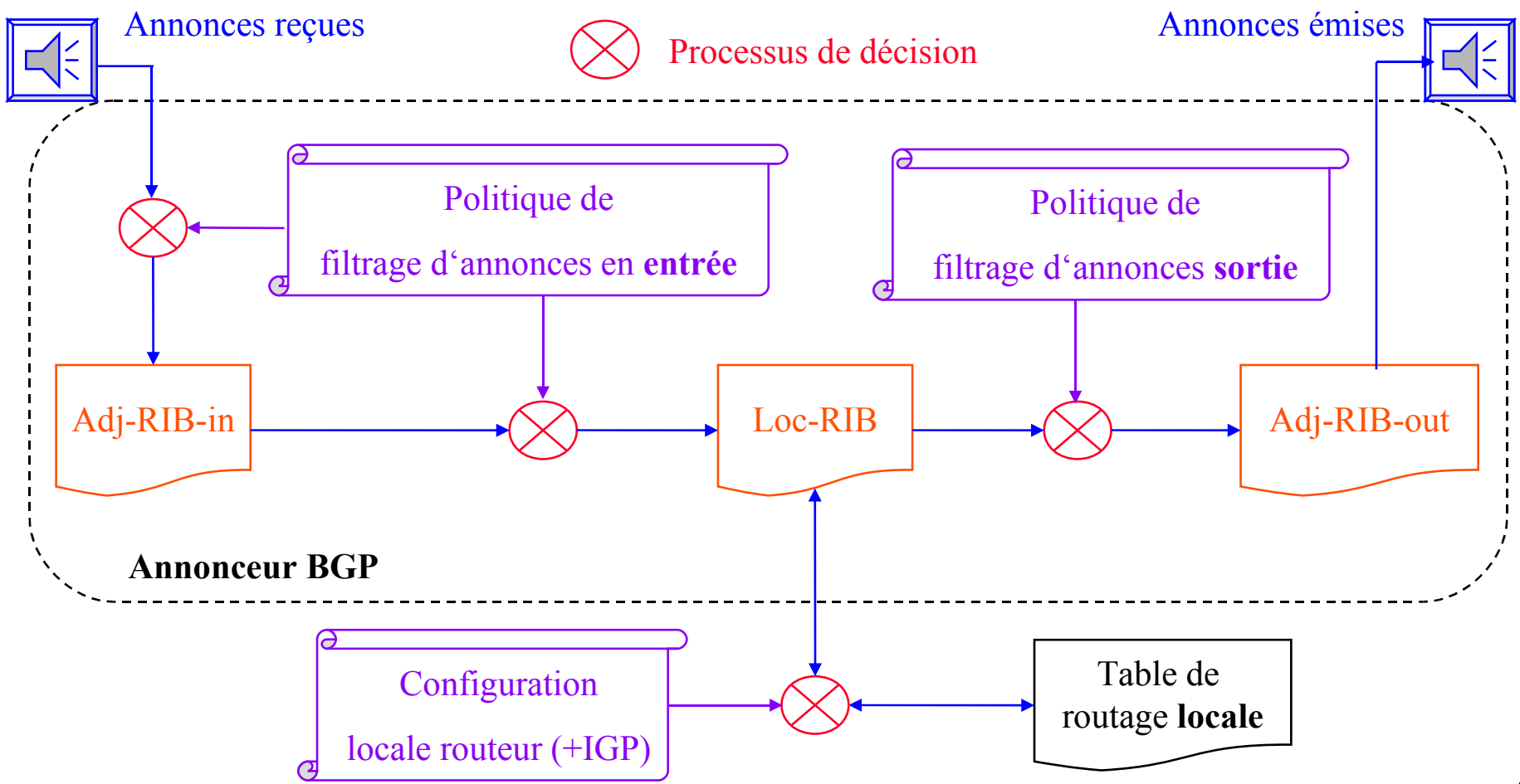
Règles pour les AS multi-connectés

- 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)

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
- Éventuellement sécurisés par MD5

Le message OPEN

- 1er message envoyé après l'ouverture de la session TCP
- Informe son voisin de :
  - Sa version de BGP
  - Son numéro d'AS
  - D'un numéro identifiant le processus BGP
- Propose une valeur de temps de maintien de la session
  - 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 ré-émis toutes les 30 secondes (suggéré)
- Message de taille minimum (19 octets)

Le message NOTIFICATION

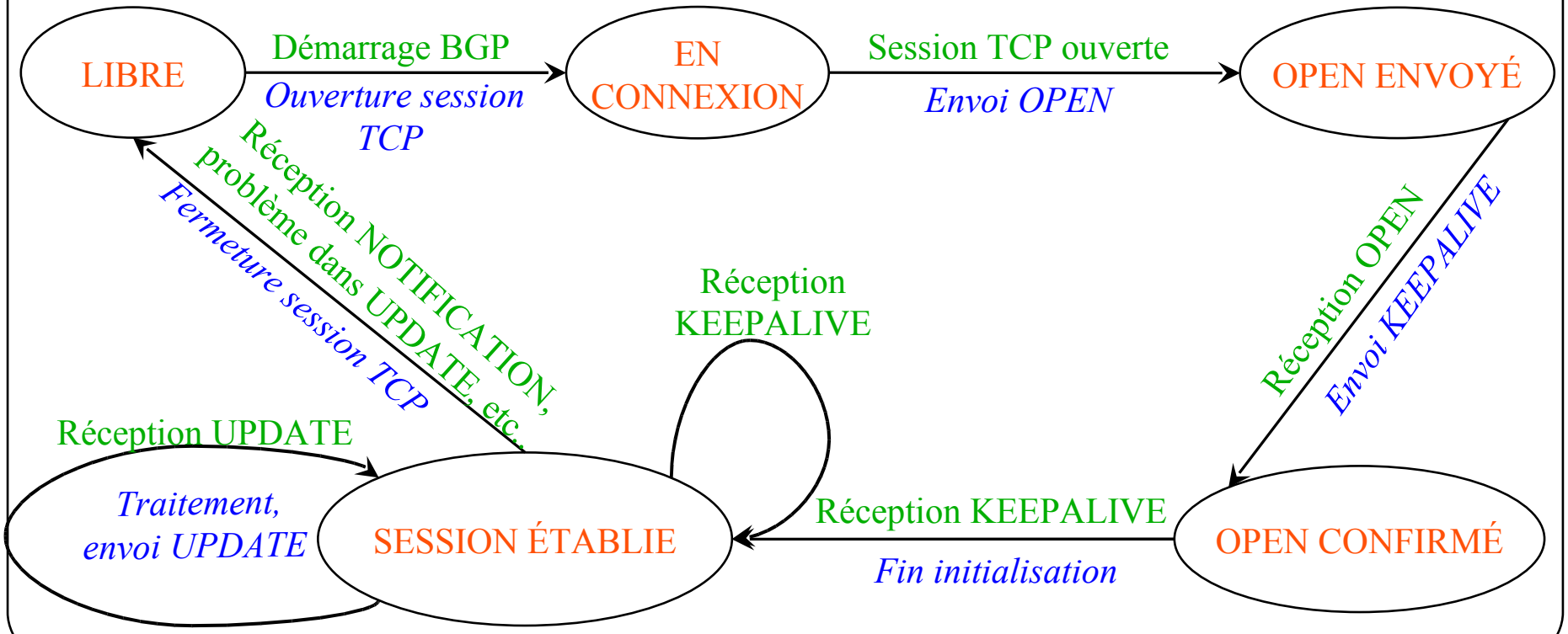
- Ferme la session BGP
- Fournit un code et un sous code renseignants sur l'erreur
- Ferme aussi la session TCP
- **Annule toutes les routes apprises par BGP**
- Émis sur incidents :
  - Pas de KEEPALIVE pendant 90s (<*hold time*>)
  - Message incorrect
  - Problème dans le processus BGP
  - ….

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)
    - Chaque réseau est défini par (préfixe, longueur)
- Envoyé uniquement si changement
- Active le processus BGP
  - Modification des RIB f(Update, politique de routage, conf.)
  - Émission d'un message UPDATE vers les autres voisins

Le processus BGP

- L'automate à états finis du processus BGP (simplifié au chemin principal, sans la gestion des incidents)



Le message UPDATE : attributs de la route

- Classés en 4 catégories :
  - 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)

Les attributs de route obligatoires (1)

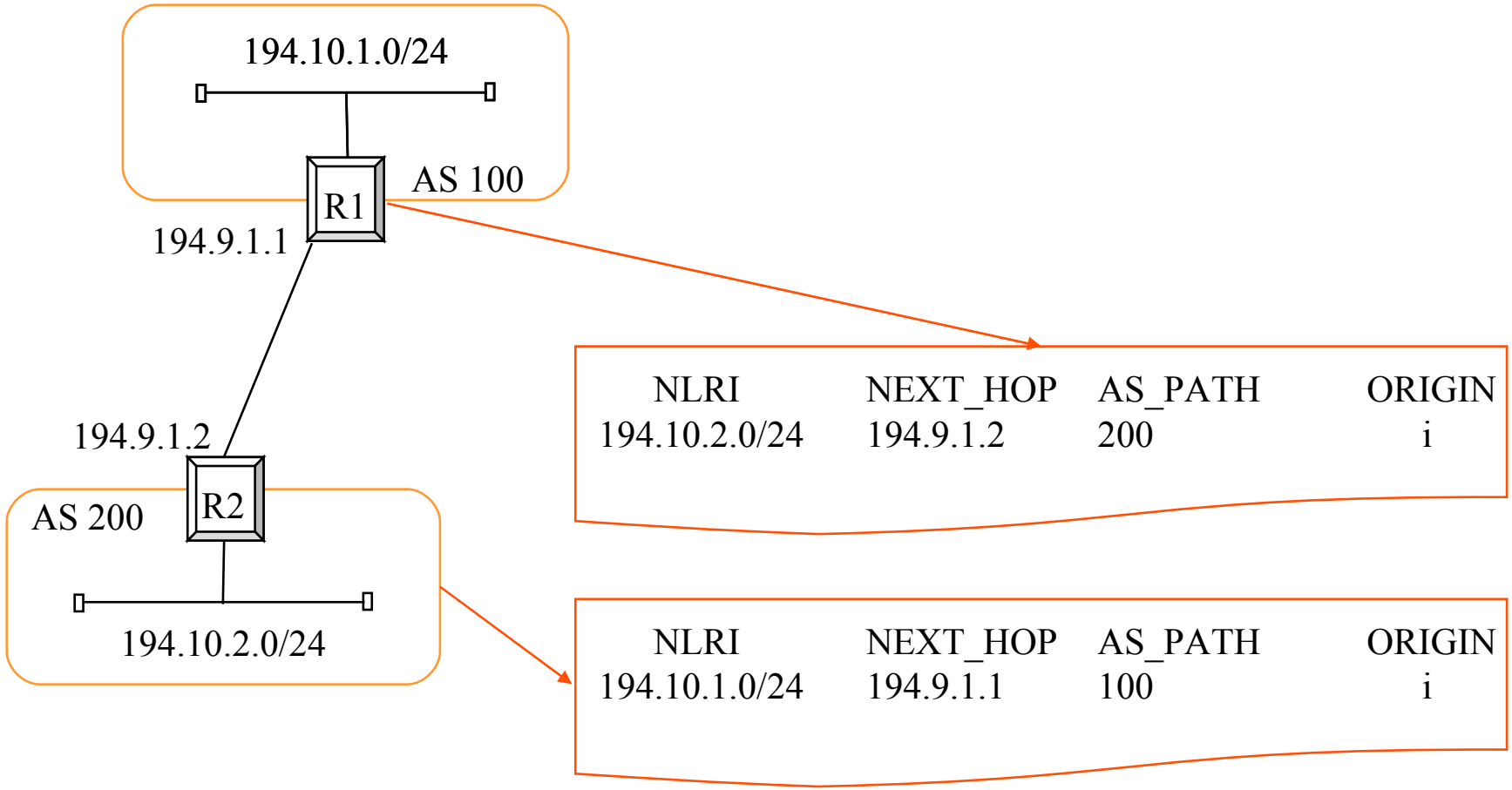
- ORIGIN
  - Donne l'origine de la route, peut prendre 3 valeurs :
    - IGP : la route est intérieure à l'AS d'origine
    - EGP : la route a été apprise par **le protocole** EGP
    - Incomplète : l'origine de la route est inconnue ou apprise par un autre moyen (redistribution des routes statiques ou connectées dans BGP par exemple)

Les attributs de route obligatoires (2)

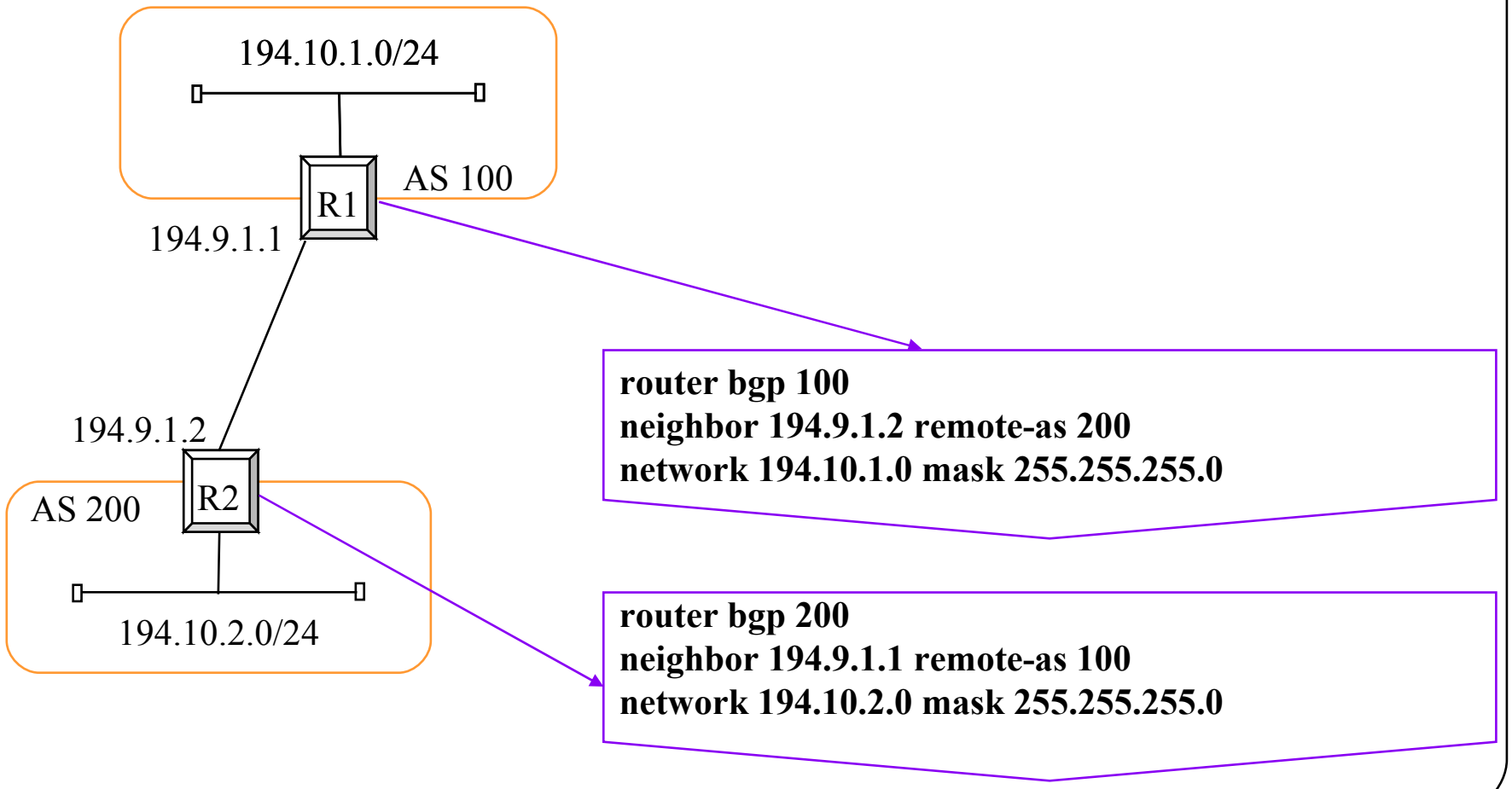
- AS_PATH
  - Donne la route sous forme d'une liste de segments d'AS
  - Les segments sont ordonnés ou non (AS_SET)
  - Chaque routeur rajoute son numéro d'AS aux AS_PATH des routes qu'il a apprises avant de les ré-annoncer

- 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)

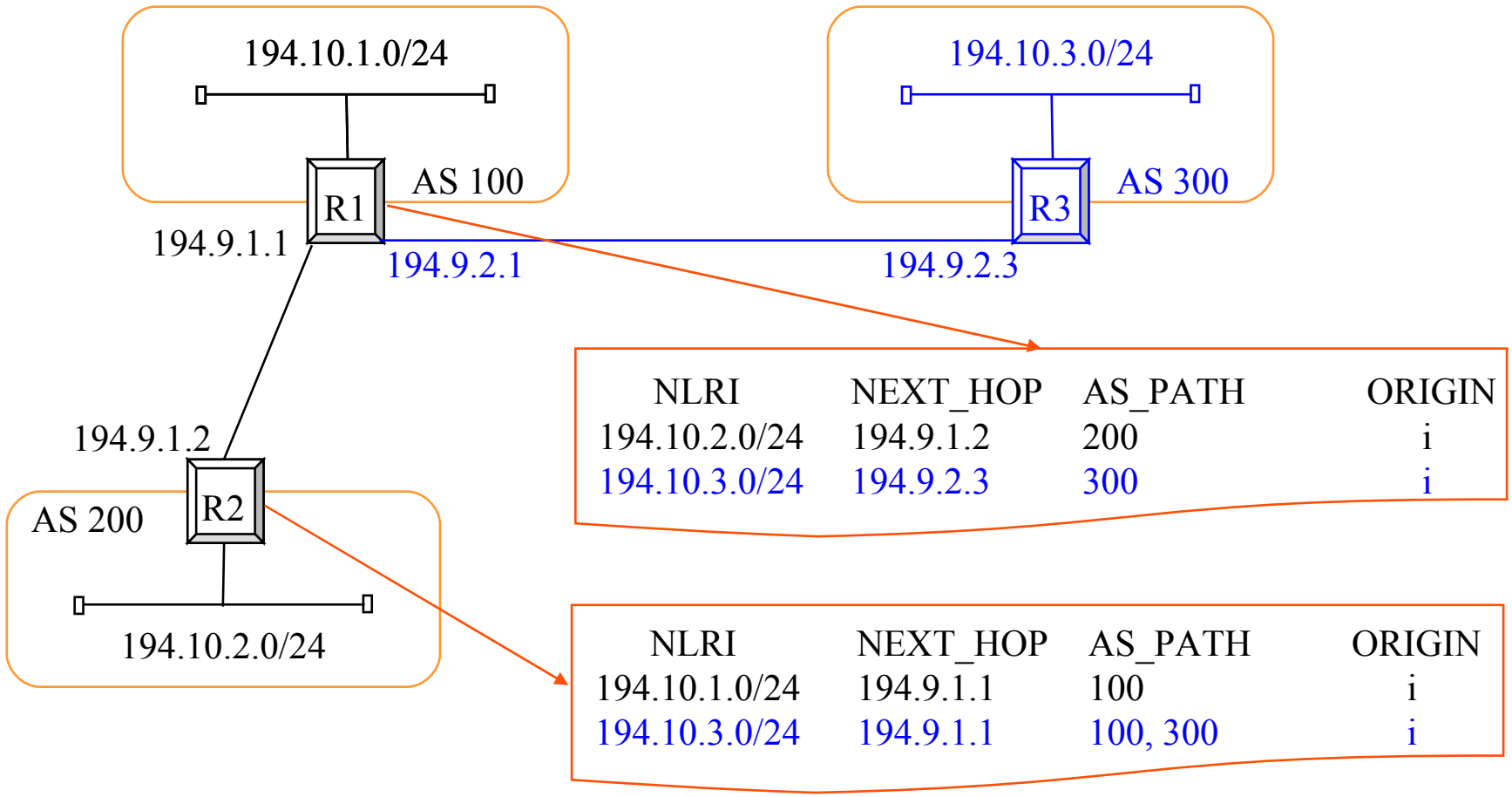
Exemple 1 : tables Adj-RIB-in



Exemple 1 : configuration sur IOS



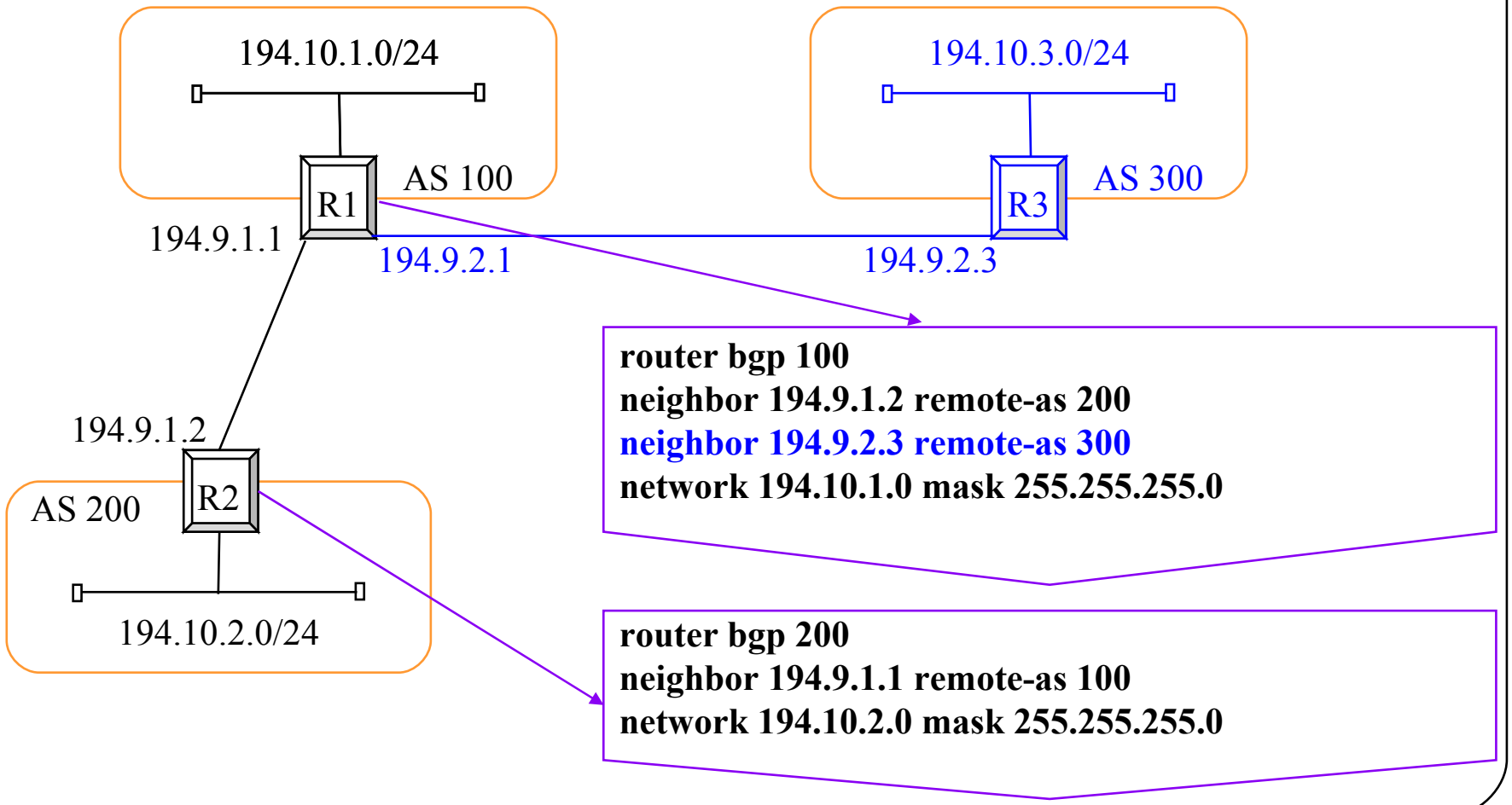
Exemple 2 : tables Adj-RIB-in



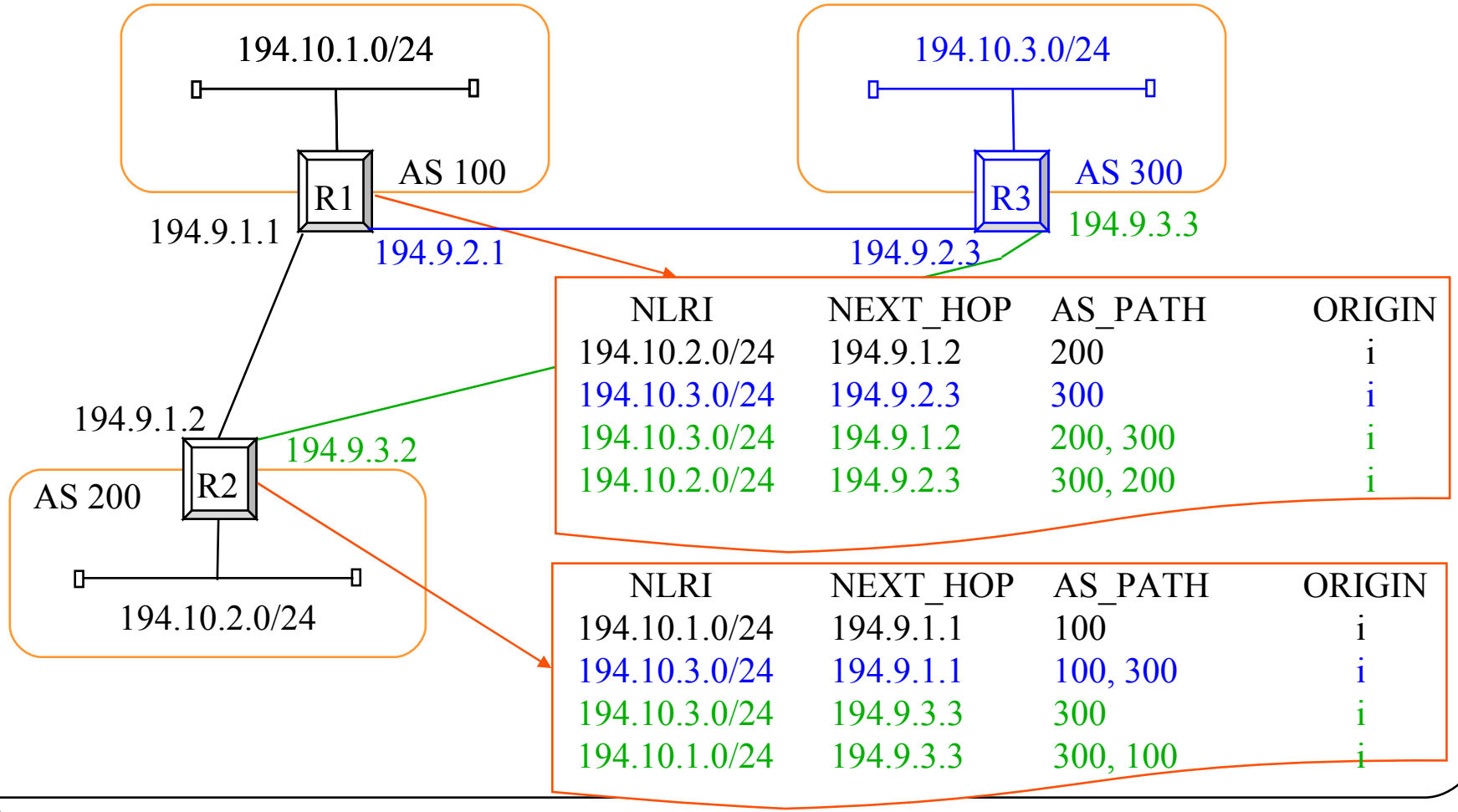
NLRI	NEXT_HOP	AS_PATH	ORIGIN
194.10.2.0/24	194.9.1.2	200	i
194.10.3.0/24	194.9.2.3	300	i

NLRI	NEXT_HOP	AS_PATH	ORIGIN
194.10.1.0/24	194.9.1.1	100	i
194.10.3.0/24	194.9.1.1	100, 300	i

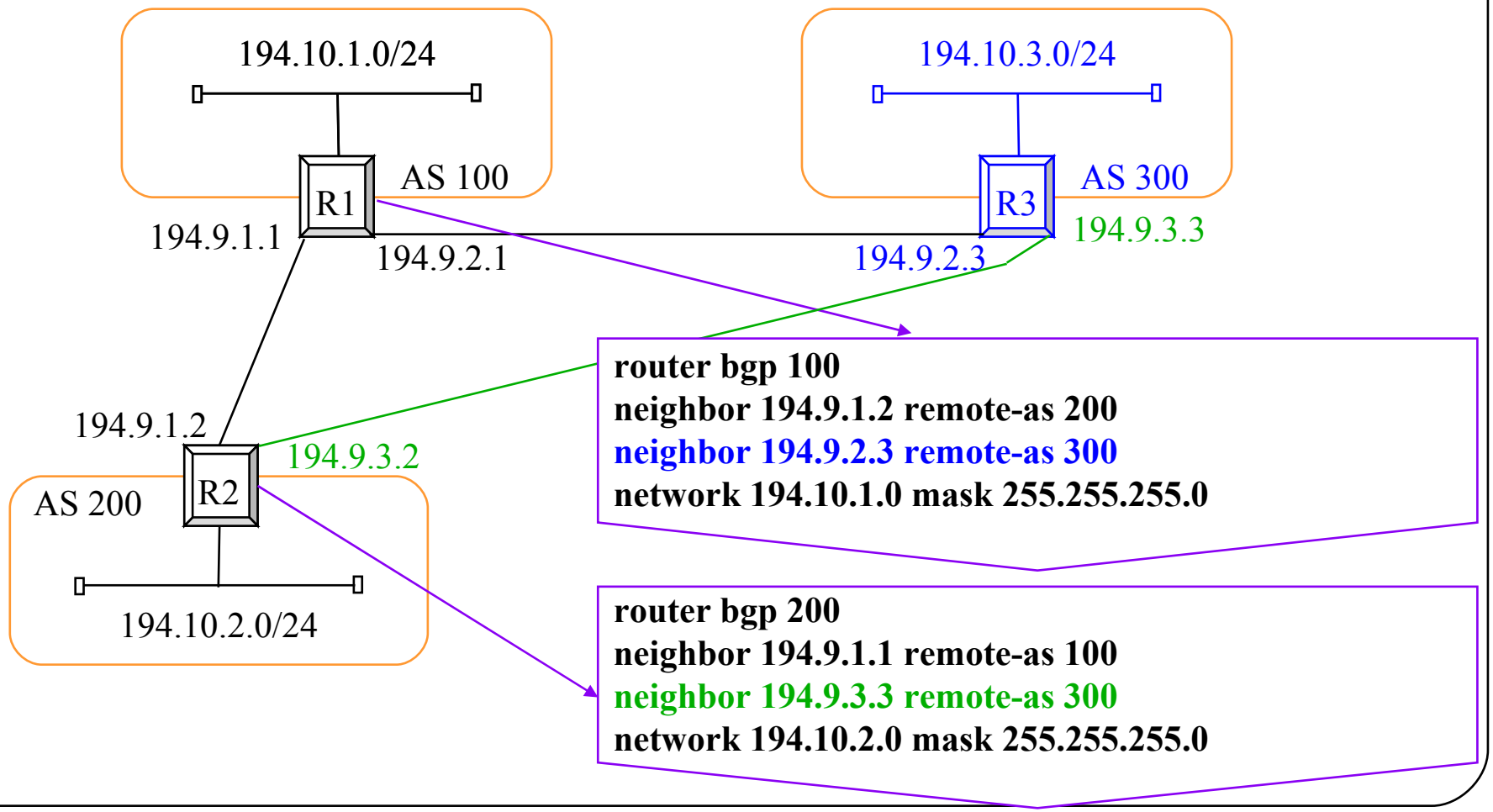
Exemple 2 : configuration sur IOS



Exemple 3 : tables Adj-RIB-in



Exemple 3 : configuration sur IOS



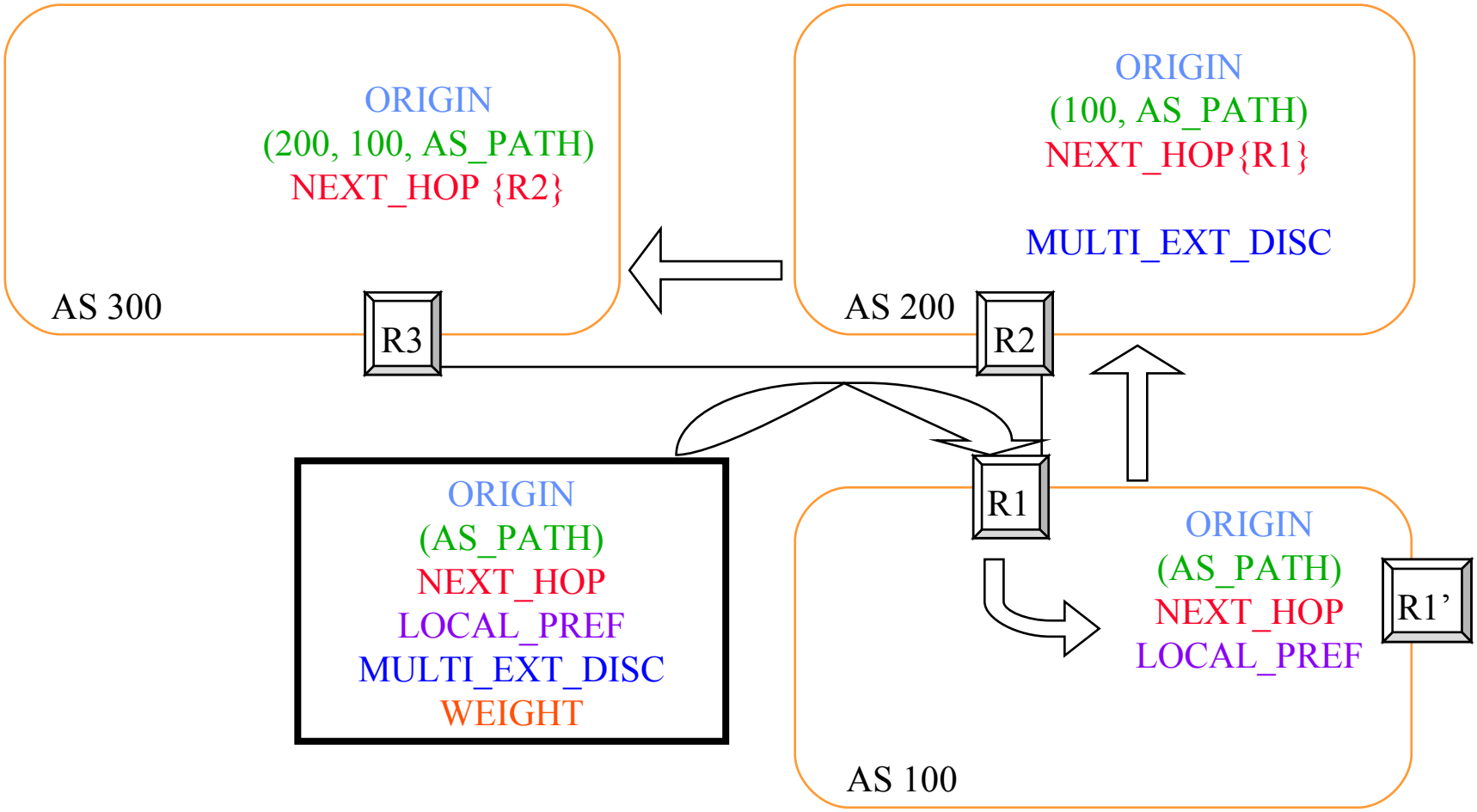
Les attributs de route optionnels (1)

- LOCAL_PREF (non transitif, discretionary)
  - Pondère la priorité donnée aux routes en interne à l'AS
  - Jamais annoncé en E-BGP
- ATOMIC_AGGREGATE (transitif, discretionary)
  - Indicateur d'agrégation
  - Quand des routes plus précises ne sont pas annoncées
- AGGREGATOR (transitif)
  - Donne l'AS qui a formé la route agrégée
  - L'adresse IP du routeur qui a fait l'agrégation

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
  - Trois valeurs reconnues
    - no-export : pas annoncé aux voisins de la confédération
    - no-advertise : pas annoncé aux voisins BGP
    - no-export-subconfed : pas annoncé en E-BGP

La portée de quelques attributs de route



Le processus de décision (1)

- 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 : application de la politique de routage de l'AS.

Le processus de décision (2)

- Critères de choix entre 2 routes (priorités décroissantes) :
  - 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 ré-annoncée 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

L'annonce des routes internes d'un AS

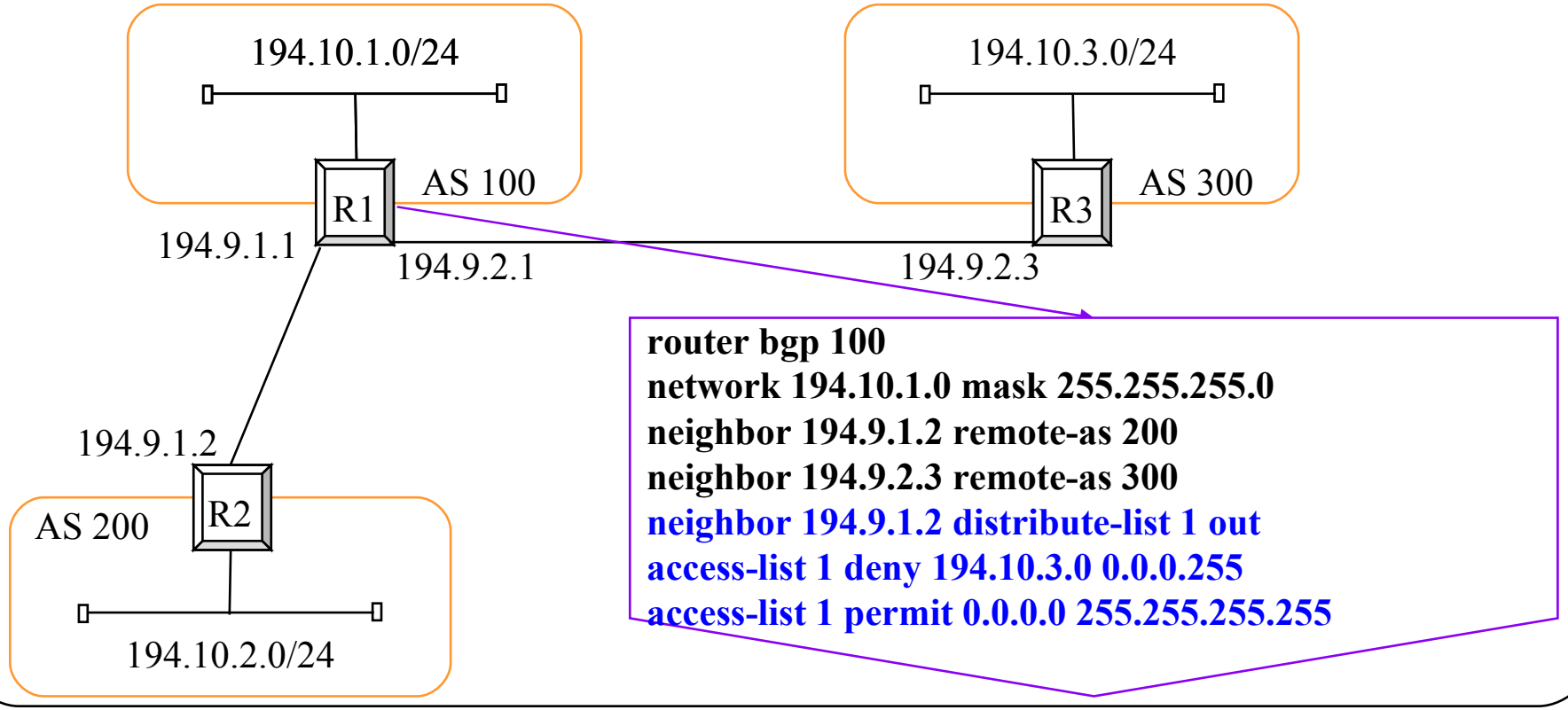
- Statique
  - Pas d'instabilité de routage, mais trous noirs possibles
  - Exemples en IOS
    - ***redistribute [static|connected]*** -> ORIGIN: Incomplete
    - ***network <adresse réseau>*** -> ORIGIN: IGP
- Dynamique
  - Suit au mieux l'état du réseau, nécessite du filtrage
  - Exemples en IOS
    - ***redistribute <paramètres de l'IGP>*** -> ORIGIN: IGP

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

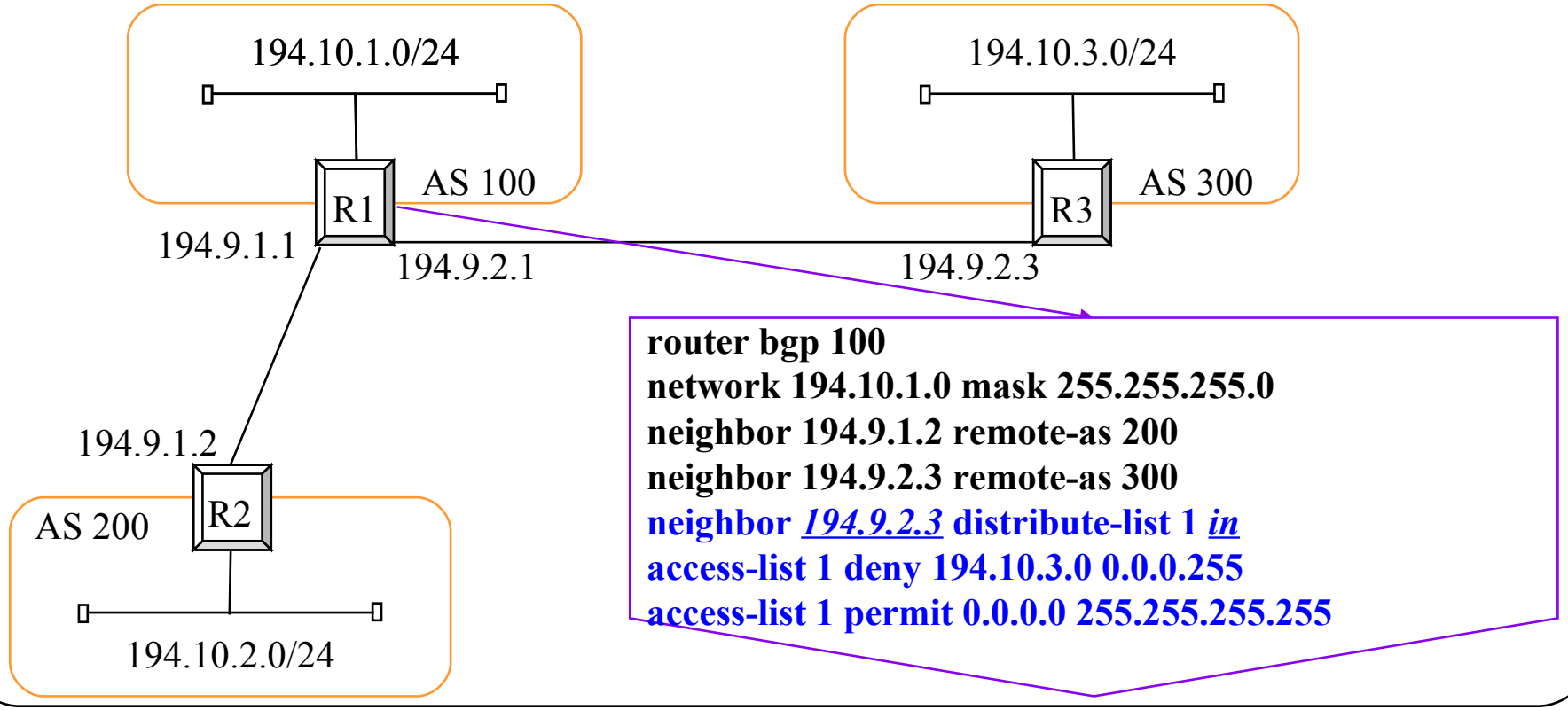
Politique de routage : exemple de filtrage de réseaux sur IOS

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



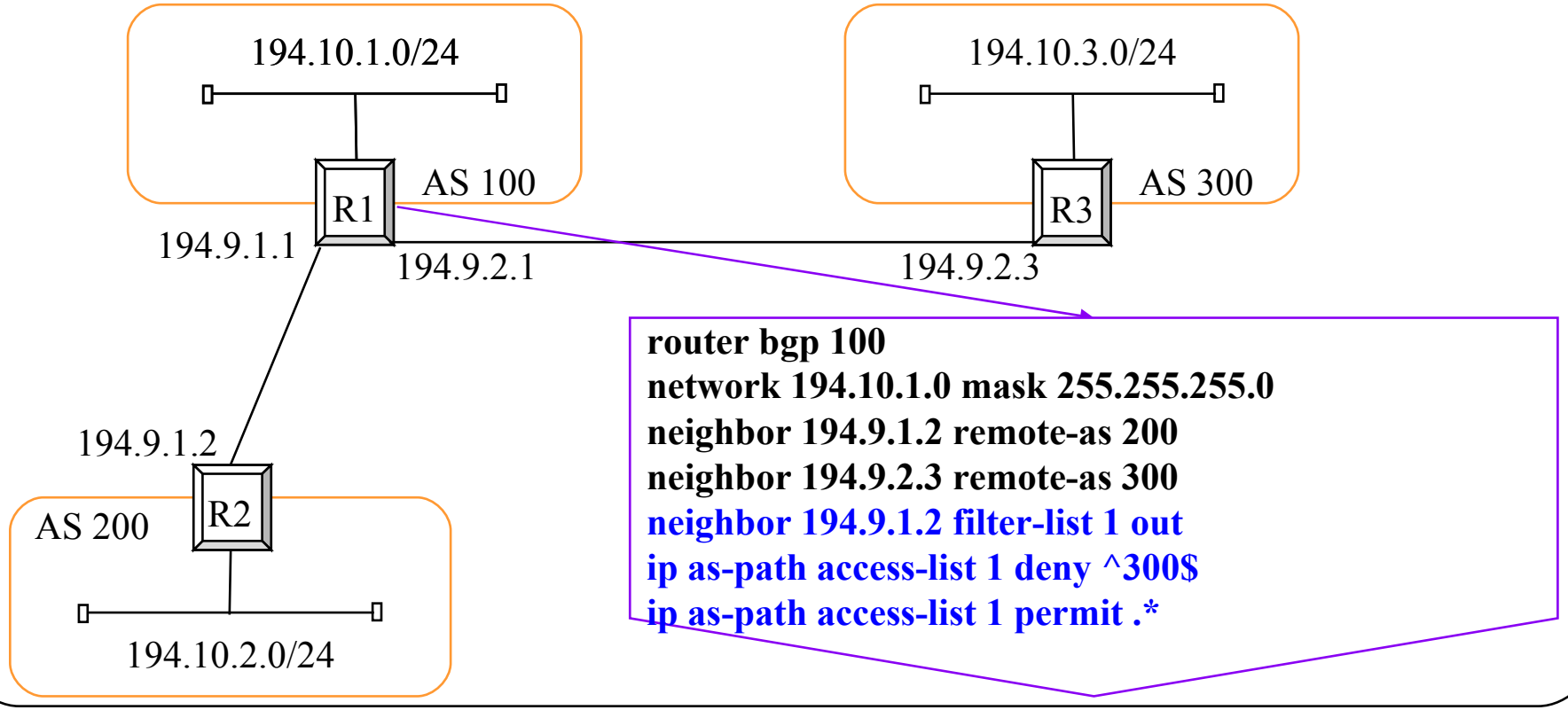
Politique de routage : exemple de filtrage de réseaux sur IOS

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



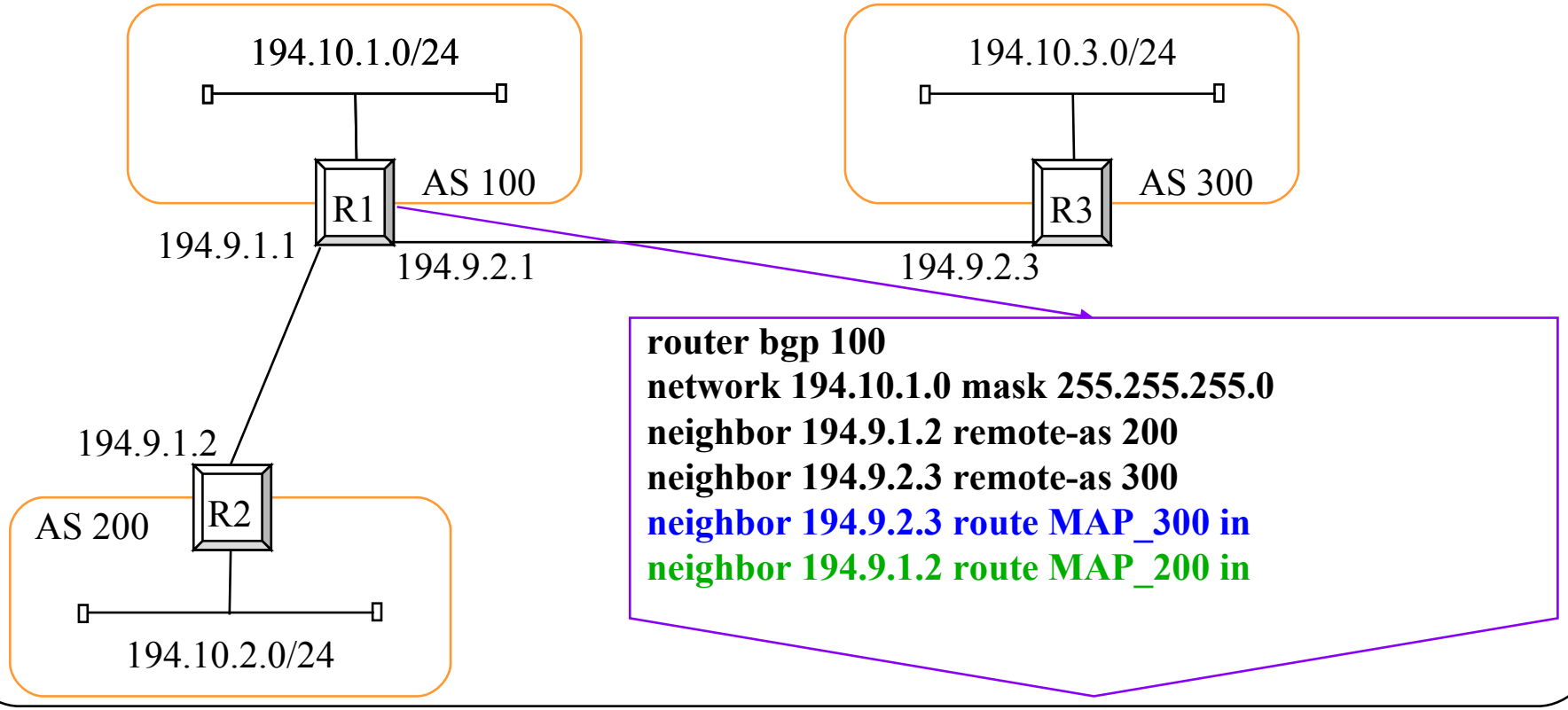
Politique de routage : exemple de filtrage de routes sur IOS

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



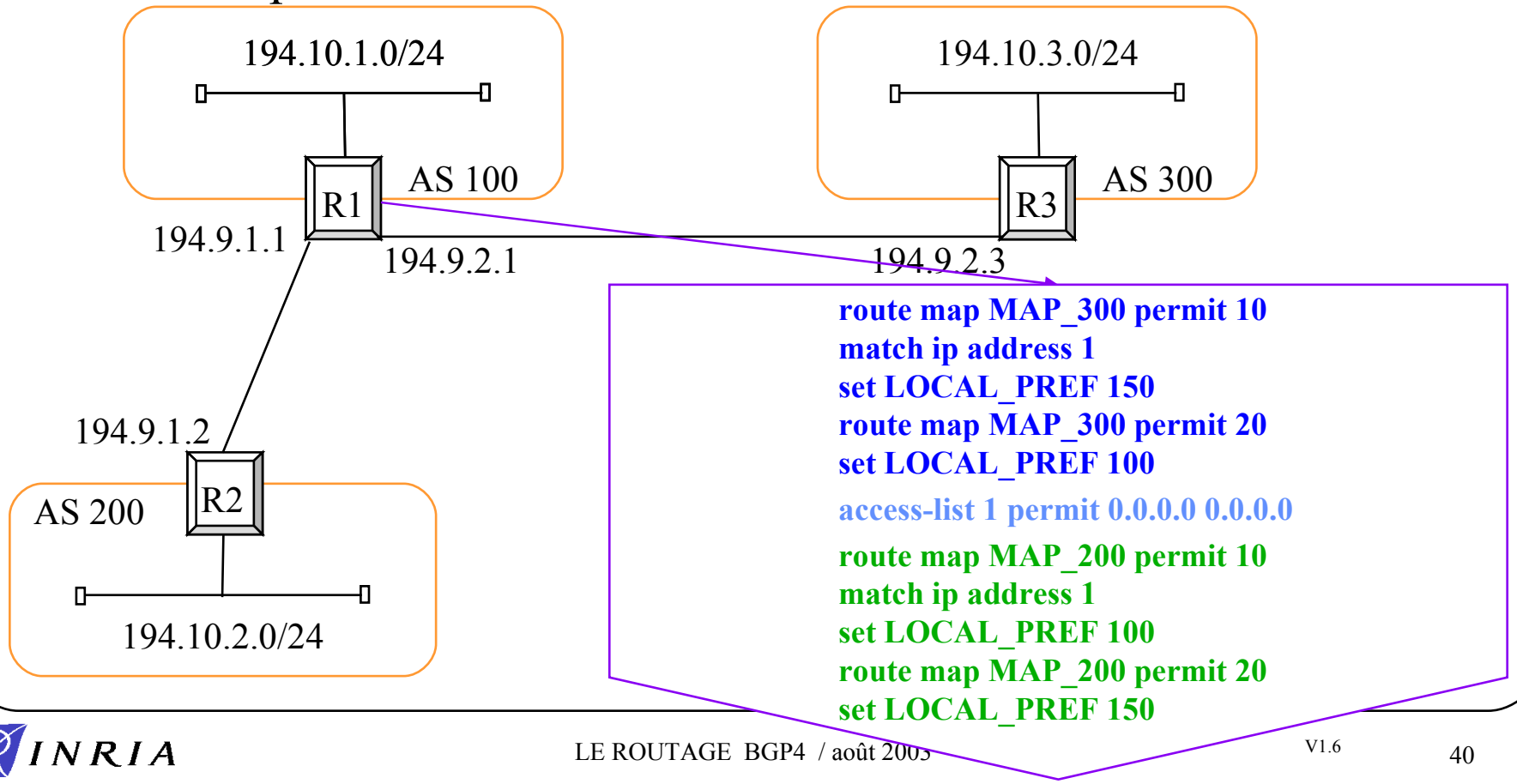
Politique de routage : exemple de manipulation d'attribut sur IOS

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



Politique de routage : exemple de manipulation d'attribut sur IOS (suite)

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

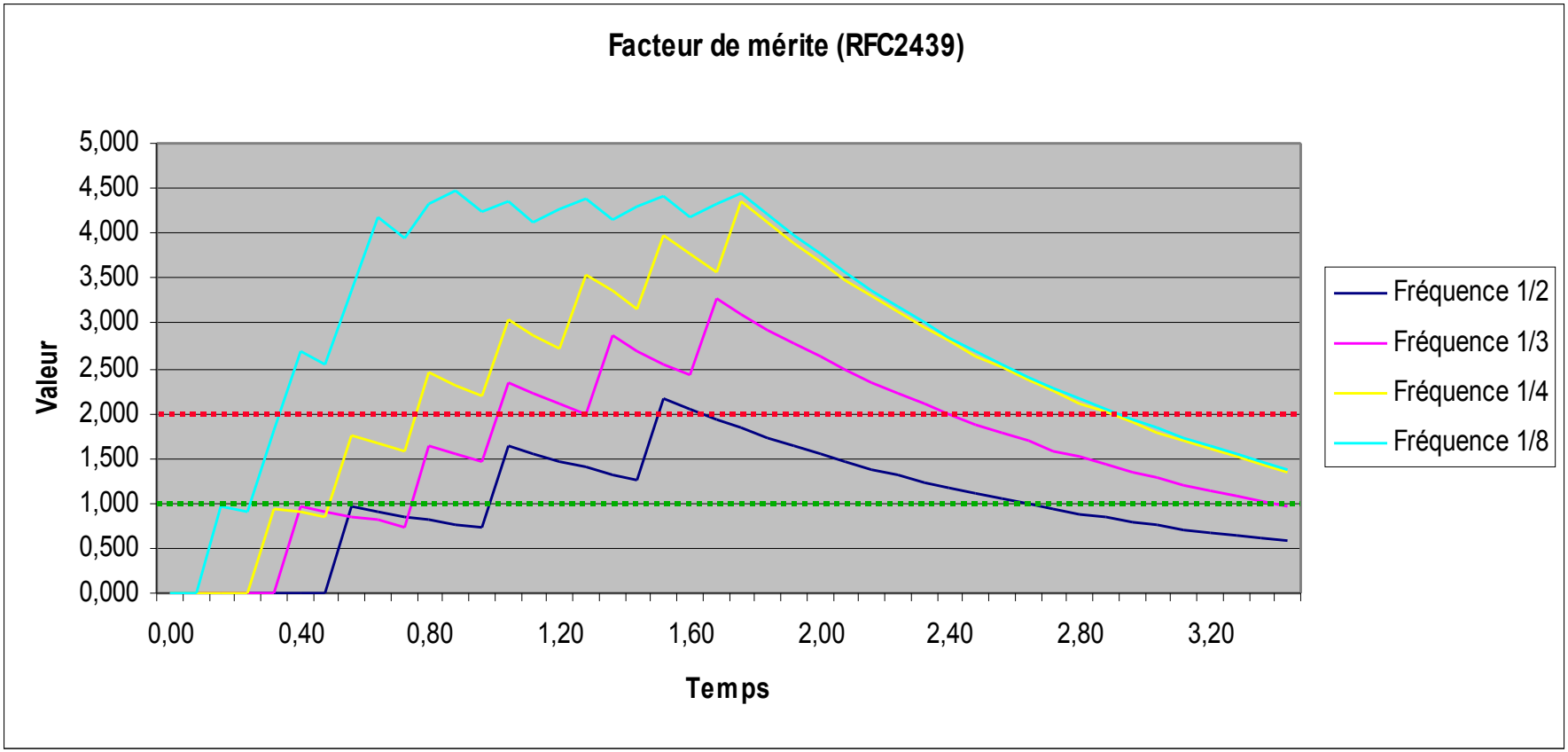


Optimisations : stabilité du routage (1)

- Les routes instables sont pénalisées
  - À chaque instabilité => X points de pénalité
  - Si Pénalité > limite L1 => route supprimée
  - Si Pénalité < limite L2 => route rétablie
  - Si : pas de nouvelle pénalité pendant T1 => Pénalité/2
  - Si Pénalité < limite L3 => on oublie tout
- Ne concerne que les annonces E-BGP

Optimisations : stabilité du routage (2)

- Allure du facteur de mérite associé à une route instable



Optimisations : contrôle du trafic BGP

- On peut agir sur différents minuteurs
  - MinRouteAdvertisementInterval
  - MinASOriginationInterval
  - La gigue dans la fréquence des annonces
- On peut réduire le volume des informations annoncées
  - NLRI agrégés
  - AS_PATH condensés

Optimisation : sécurisation des échanges BGP

- Mesures natives au protocole
  - Session BGP = {@IP1,numéro AS1},{@IP2,numéro AS2}
  - Signature MD5 de chaque message
- Compléments : mesures standard au niveau TCP ou IP
  - Filtrage du port 179
- MAIS : a toutes les vulnérabilités de TCP ou IP
  - Déni de service

Optimisations : les réflecteurs de routes

- Permet d'éviter une croissance en N^2 des sessions I-BGP
- Mais rajoute un point de panne singulier
- On met donc plusieurs réflecteurs de route par AS

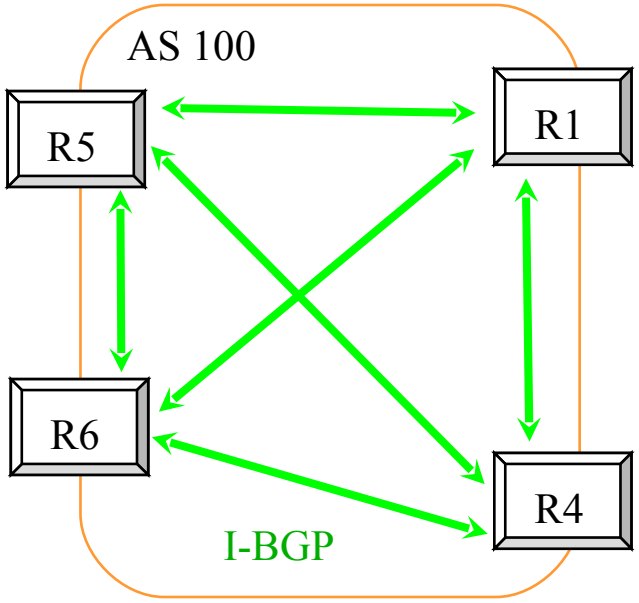


Schéma sans réflecteur de routes

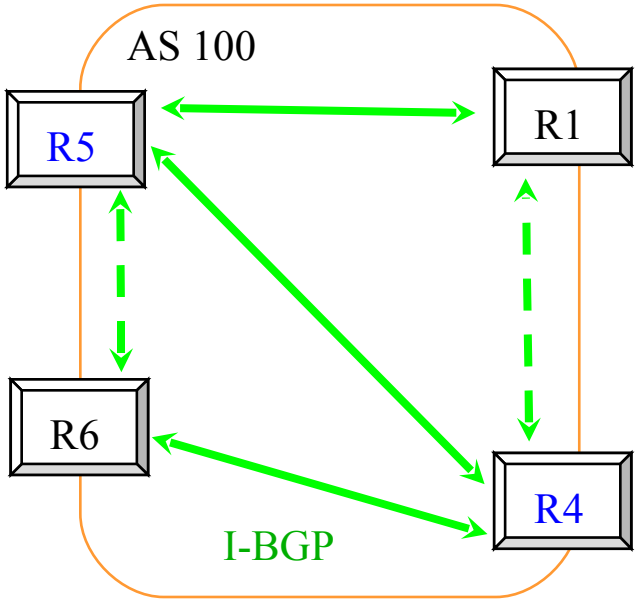


Schéma avec 2 réflecteurs de routes (R4 et R5)

Extensions : les confédérations d'AS

- Permet de réduire le nombre de sessions I-BGP
- En divisant l'AS en mini-AS (ou sous AS)
- Les routeurs de bord d'un mini-AS établissent des sessions
  - I-BGP entre eux (maillage complet)
  - E-BGP avec leurs voisins d'autres AS
  - Pseudo E-BGP avec leurs voisins d'autres minis-AS
- Vu de l'extérieur, la confédération d'AS apparaît comme un seul et unique AS

Extensions : les groupements de routeurs

- Les routeurs BGP d'un groupement partagent la même politique de routage (ex. routes maps, filtres d'annonces, …)
- Cette politique est définie sur l'un des routeurs du groupement
- Elle est propagée automatiquement sur les autres routeurs
- Un routeur du groupement peut modifier localement sa politique de routage (mais ne la propage pas aux autres)

Extensions : les serveurs de route

- Sur un grand point d'échange on peut avoir :
  - 100 fournisseurs d'accès Internet
  - Plus de 115 000 routes annoncées (août 2002)
- Ce qui peut impliquer :
  - Jusqu'à 10 000 sessions TCP !
- Solution : les serveurs de route
  - Réduit le nombre de sessions (quelques unes par fournisseur d'accès)

Extensions : le routage multi-protocole (IPv6)

- Dans BGP, seuls 3 attributs de route de dépendent d'IPv4
  - NLRI, NEXT_HOP, (AGGREGATOR)
- Pour rendre BGP multi-protocole, on introduit 2 attributs de route supplémentaires
  - MP_REACH_NLRI (optionnel, non-transitif)
  - MP_UNREACH_NLRI (optionnel, non-transitif)
- L'attribut de route MP_REACH_NLRI contient des triplets
  - *Adress_family* (ex. IPv4, IPv6, IPX), NEXT_HOP, NLRI
- Un message UPDATE contient MP_REACH_NLRI et les autres attributs de route déjà vus (ORIGIN, LOCAL_PREF...)

Exemple de configuration BGP en IPv6 (Zebra)

```
router bgp 65400
 bgp router-id 192.108.119.167
 ipv6 bgp neighbor 2001:660:281:8::1 remote-as 1938
 ipv6 bgp neighbor 2001:660:281:8::1 prefix-list filtre_nlri in
 ipv6 bgp neighbor 2001:660:281:8::1 filter-list filtre_as in
!
ipv6 prefix-list filtre_nlri description    Refus des annonces de son préfixe et du 2002::/16
ipv6 prefix-list filtre_nlri seq 5 deny 3ffe:305:1014::/48 le 128
ipv6 prefix-list filtre_nlri seq 10 deny 2002::/16 le 128
ipv6 prefix-list filtre_nlri seq 15 permit any
!
ip as-path access-list filtre_as deny 1938 2200 5511 *
ip as-path access-list filtre_as permit .*
```

Extensions : le routage multicast (MBGP)

- Vu comme un cas particulier du routage multi-protocole
- Utilisation de la notion de sous famille d'adresse
- Implémentations récentes (IOS, ….)

Extensions : annonce de capacité

- Standardisé initialement en mai 2000 par le RFC2842 (statut PS)
- Standardisé définitivement en novembre 2002 par le RFC3392 (DS)
- Introduit un paramètre optionnel : *capabilities*
- Annonce les capacités fonctionnelles d'un routeur lors de l'OPEN
- Permet une mise à niveau automatique des fonctionnalités utilisées dans cette session BGP
- Permettra des mises à niveau des implémentations de BGP non synchrones

Bibliographie : principaux RFC

- RFC1657 Definition of Managed Objects for the Fourth Version of Border Gateway Protocol (BGP-4) . S. Willis, J. Burruss, and J. Chu. 06/1995.(DS)
- RFC1771 A Border Gateway Protocol 4 (BGP-4). Y. Rekhter, T. Li. 03/1995.(DS)
- RFC1772 Application of the Border Gateway Protocol in the Internet. Y Rekhter, P. Gross. 03/1995.(DS)
- RFC1773 Experience with the BGP-4 protocol. P. Traina. 03/1995.(INFO)
- RFC1774 BGP-4 Protocol Analysis. P. Traina, Editor. 03/1995.(INFO)
- RFC1997 BGP Communities Attribute. R. Chandra, P. Traina & T. Li. 06/1996. (PS)
- RFC1998 An Application of the BGP Community Attribute in Multi-home Routing. E. Chen & T. Bates. 06/1996.(INFO)
- RFC2042 Registering New BGP Attribute Types. B. Manning. 01/1997.(INFO)
- RFC2385 Protection of BGP Sessions via the TCP MD5 Signature Option. A. Heffernan. 08/1998.(PS)
- RFC2439 BGP Route Flap Damping. C.Villamizar, R.Chandra, R.Govindan. 11/1998. (PS)

Bibliographie : principaux RFC

- RFC2457 Definitions of Managed Objects for Extended Border Node. B. Clouston, B. Moore. 11/1998. (PS)
- RFC2545 Use of BGP-4 Multiprotocol Extensions for IPv6 Inter-Domain Routing. P. Marques, F. Dupont. 03/1999. (PS)
- RFC2547 BGP/MPLS VPNs. E. Rosen, Y. Rekhter. 03/1999. (Status: INFO)
- RFC2796 BGP Route Reflection - An Alternative to Full Mesh IBGP. T. Bates, R. Chandra, E. Chen. 04/2000. (Updates RFC1966) (PS)
- RFC2858 Multiprotocol Extensions for BGP-4. T. Bates, Y. Rekhter, R. Chandra, D. Katz. 06/2000. (PS)
- RFC2918 Route Refresh Capability for BGP-4. E. Chen, 09/2000. (PS)
- RFC3065 Autonomous System Confederations for BGP. P. Traina, D. McPherson, J. Scudder. 02/2001.(PS)
- RFC3107 Carrying Label Information in BGP-4. Y.Rekhter, E.Rosen. 02/2001.(PS)
- RFC3345 Border Gateway Protocol (BGP) Persistent Route Oscillation Condition. D. McPherson, V. Gill, D. Walton, A. Retana, 08/2002. (Status: INFO)
- RFC3392 Capabilities Advertisement with BGP-4. R. Chandra, J. Scudder. 11/2002. (DS)

Bibliographie : livres

- Le routage dans l'Internet, C. Huitema, Eyrolles, 1994
- Interconnections with bridges and routers, R, Perlman, Addison-Wesley, 1996
- Internet Routing Architectures, B.Halabi, Cisco Press, 1997
- BGP4 Inter-Domain Routing in the Internet, J. W. Stewart III, Addison-Wesley, 1999

Bibliographie : Sites web

- www.rsng.net : Route Server Next generation Project
- www.merit.net : Nombreuses informations sur les points d'échange de trafic entre opérateurs des USA.
- www.gated.org : Site de distribution du logiciel gated (payant) qui implemente la plupart des logiciels de routage (dont BGP4)
- www.zebra.org : Site de distribution du logiciel zebra (licence GPL) qui implemente la plupart des logiciels de routage (dont BGP4)
- www.caida.org : Propose des outils de métrologie réseau, beaucoup de données sur le trafic.
- www.merit.edu/~ipma/ : outils de mesure de performances, beaucoup d'informations sur les tables BGP de certains routeurs des points d'échange
- www.ep.net : Liste des points déchange
- www.ra.net : Routing Arbiter Project
- telnet://route-server.cerf.net : Accès en ligne a un routeur BGP
- http://www.cisco.com/univercd/cc/td/doc/cisintwk/ics/icsbgp4.htm : Manuel de référence des commandes BGP sur IOS de Cisco.
- www.mcvax.org/~jhma/routing/ : nombreuses statistiques sur les tables de routage BGP