Algorithmes de routage IP

4 Le routage inter-domaine avec BGP

Filière F5 - Isima 2009-2010

Mickael Meulle
mickaelmeulle@gmail.com
Michael.meulle@orange-ftgroup.com

Le processus de décision BGP

- **Il est exécuté pour décider de la meilleure route vers chaque NLRI**
  - Spécifié dans RFC
  - Implémenté par les constructeurs de façon différente…
    - Cisco, Juniper, Alcatel
  - Le principe reste le même
- **Il permet de départager pas à pas des routes également préférées pour un même NLRI**
  - A chaque point de comparaison, on ne garde que la ou les routes qui sont les meilleures
  - Remarque: certaines routes peuvent ne pas être comparées dans certains cas particuliers
    - Exemple avec l'attribut MED et des routes reçues d'AS différents

Le principe du processus de décision BGP

- **1. Préférer les routes avec l'attribut Local_Pref maximal**
  - Correspond généralement au accords d'interconnexion
    - Local ou Originated > Client > Peer > fournisseur
- **2. Préférer les chemins d'AS les plus court**
  - Attribut AS_PATH
  - Les routeurs peuvent augmenter arbitrairement la longueur du chemin en répétant leur numéro d'AS
    - lors de l'import ou de l'export de la route
- **3. Préférer les routes avec le meilleur attribut Origin type**
  - IGP > EGP > incomplet
- **4. Préférer les MED les plus petit (si l'attribut est présent)**
- **5. Préférer les chemins externes des chemins internes (eBGP > iBGP)**
- **6. Préférer les routes avec métrique IGP minimale vers le NEXT_HOP**
- **7. Préférer la route avec la plus petite adresse IP du NEXT_HOP (router ID)**

NEXT_HOP

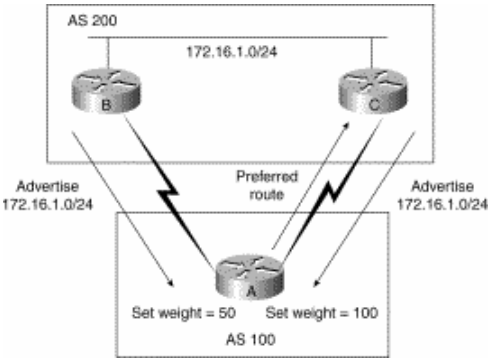
- **Attribut obligatoire : "well-known mandatory"**
- **L'adresse IP du routeur de bordure (dans le même AS) qui permettra la transit des paquets pour la route courante (le transit des paquets vers la ou les destinations spécifiées par la route)**
  - L'attribut NEXT HOP est placé par un routeur qui donne origine à une route
  - L'attribut NEXT HOP est modifié lorsqu'un routeur annonce une route à un routeur BGP d'un autre AS
  - Il permet d'indiquer une direction pour le trafic à destination d'un NRLI qui n'est pas forcément le routeur annonçant la route
- **Internal/external Next Hop**
  - Si un routeur appartient au même AS que son peer (voisin)
    - « Internal border router »
  - Sinon « External Border router »

NEXT_HOP

- **Lorsqu'un routeur reçoit une route BGP, l'adresse IP indiquée par l'attribut NEXT HOP doit être accessible ("reachable") du routeur en question.**
  - L'adresse IP de chaque next hop est généralement contenue dans un réseau annoncé via l'IGP.
- **Remarque:**
  - Un routeur ne se mets jamais en next hop pour une route
  - Un routeur qui donne origine à une annonce, ne mets pas le destinataire du message comme NEXT_HOP
  - Certains routeurs vérifient que le NEXT HOP indiqué par une route BGP est toujours accessible pour considérer la route comme utilisable (up)
    - "Next hop tracking"

L'attribut Weight

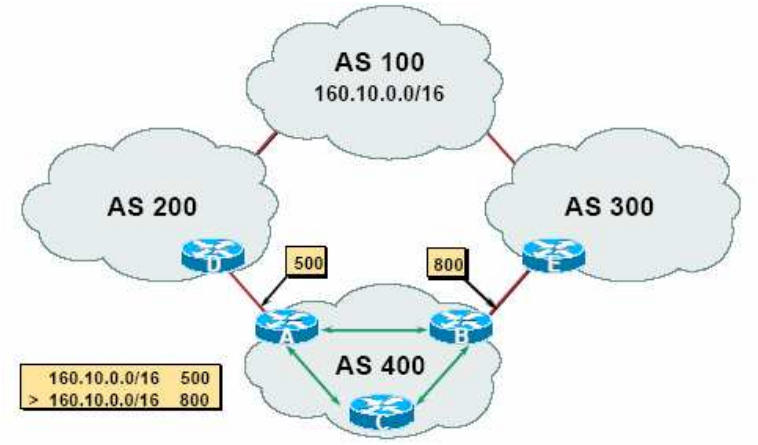
- **Essayer de ne pas l'utiliser!**
  - Spécifique aux routeurs Cisco
- **Cet attribut n'est pas propagé, il est local à chaque routeur**
  - Il ne fait pas partie du standard



Local Preference

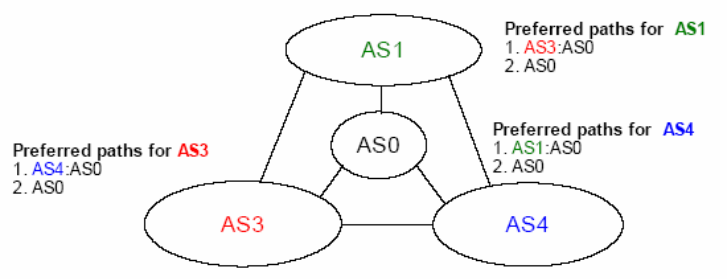
- **Attribut très utilisé pour préférer un AS de sortie**
  - Indique un degré de préférence
  - Plus l'attribut est grand pour une route, plus cette route sera préférée
- **Cet attribut est seulement propagé à l'intérieur de l'AS**
  - Il indique donc une préférence globale pour tout l'AS
  - Si l'attribut est propagé entre AS, il doit être supprimé
- **L'attribut est souvent utilisé pour rendre compte de considérations économiques**
  - Préférer les clients aux peers et les peers aux fournisseurs
- **La valeur par défaut est 100**

Local Preference



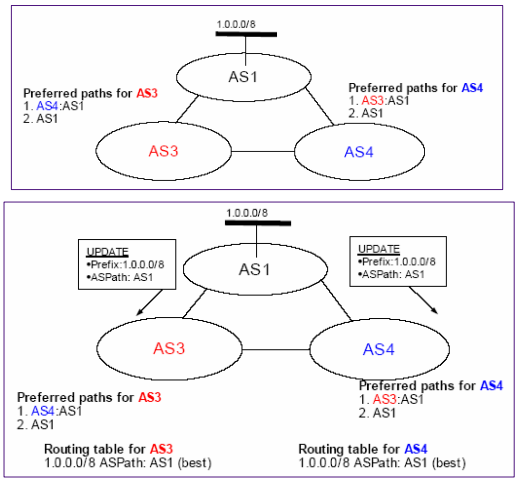
Limitation de Local Pref (1)

- **Quelquefois, la convergence n'est pas assurée**
  - On voit même apparaître des boucles



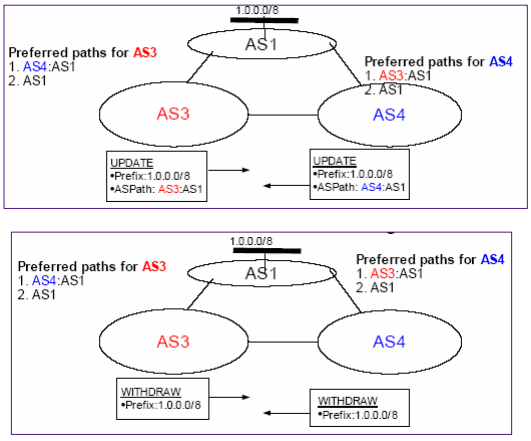
Limitation de Local Pref (2)

- **Quelquefois, la convergence n'est pas assurée**
  - Les configurations doivent être cohérentes



Limitation de Local Pref (3)

- **Quelquefois, la convergence n'est pas assurée**
  - Les configurations doivent être cohérentes



AS_PATH

- **Attribut obligatoire : "well-known mandatory"**
- **"Chemin d'AS"**
  - De droite à gauche
    - les AS successivement parcourus par l'annonce de la route
  - De gauche à droite
    - Les AS entre le routeur local et le routeur qui a donnée origine à la route
  - Ne contient pas l'AS du routeur local
- **Utilisation/intérêt :**
  - Pas de boucles au niveau AS. Tout routeur recevant une route avec son AS dans l'attribut AS_PATH supprime cette route!
  - Cela permet d'effectuer des décisions politiques en fonction de certains AS présents ou non dans l'AS_PATH
  - Une route avec un AS_PATH de plus petite longueur sera préférée à une route avec un AS_PATH de taille plus grande

AS_PATH

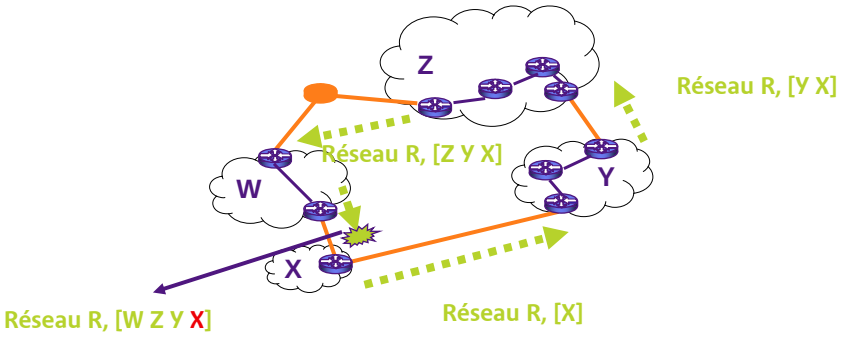
- **AS_PATH = C'est une séquence de segments**
- **Segment = <as_set> | <as_sequence>**
- **<as_set> = { ASX, ASY,ASZ}**
  - Type 1
  - Les AS X, Y et Z ont été traversés dans un ordre inconnu
- **<as_sequence> = ASX ASY ASZ**
  - Type 2
  - Les AS X, Y et Z ont été traversés dans l'ordre spécifié

AS_PATH

- **Modification de l'attribut lorsque les messages BGP sont propagés par eBGP**
  - Considérons le cas d'un routeur A qui propage un message au router B
    - Si A et B sont dans le même AS, alors l'attribut n'est pas modifié
    - Si A et B sont dans deux AS différents: mise à jour de l'attribut
      - *Si le premier segment est une séquence*
        Le routeur ajoute son AS au segment au début su segment (leftmost bit)
      - *Si le premier segment est un set*
        Le routeur ajoute un nouveau segment : son AS (seul) dans une séquence
  - Si un routeur A donne origine à un message et l'envoi au routeur B
    - Si le routeur B est dans le même AS, l'attribut AS_PATH est vide
    - Sinon le routeur B appartient à un autre AS et alors le routeur crée un segment: son AS (seul) dans une séquence

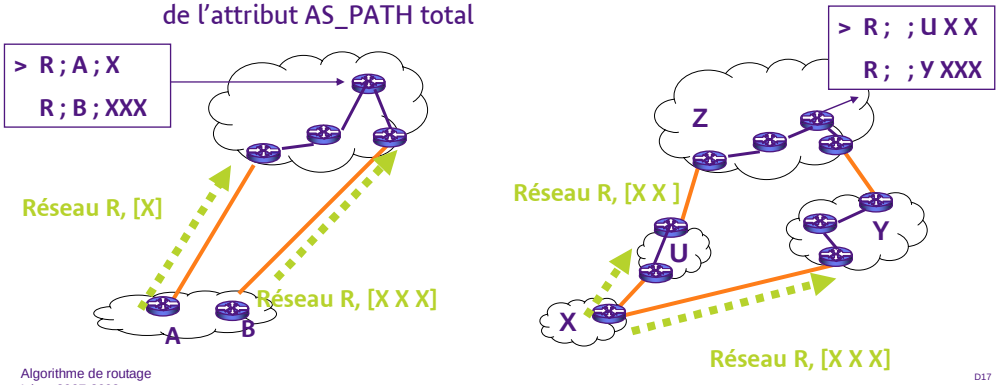
No Loops

- **Un routeur qui reçoit une route avec son AS à l'intérieur d'un segment de l'attribut AS_PATH supprime la route**
  - Détection de boucles



Les répétitions d'AS
"AS PATH prepending", "shifting"

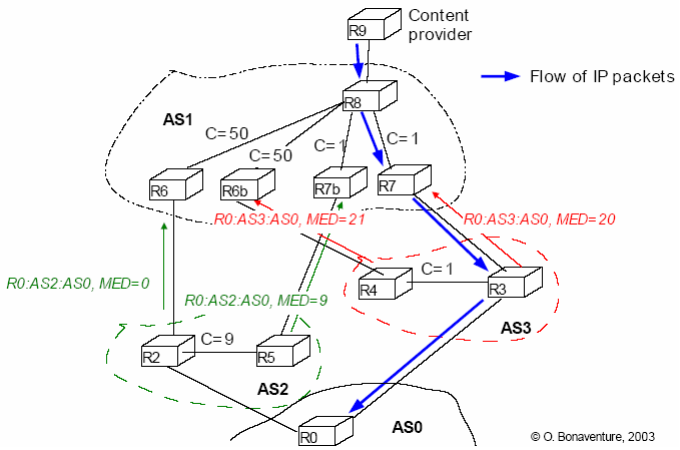
- **Ce sont lors des sessions eBGP que les routeurs insèrent leur AS dans une séquence**
  - L'AS peut être ajouté plusieurs fois pour augmenter la taille de l'attribut AS_PATH total



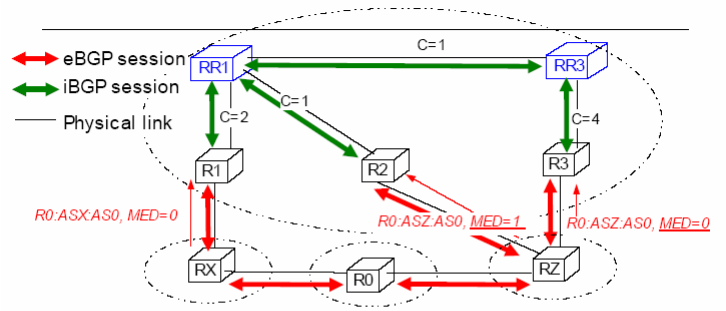
MED

- **Multi Exit Discriminator : MULTI_EXIT_DISC**
  - Attribut non obligatoire : "optional nontransitive"
  - Utilisé pour influencer un routeur dans son processus de décision BGP pour choisir un point de sortie d'un AS plutôt qu'un autre
    - Était appelé "Inter AS Metric" dans de précédentes versions de BGP
  - En pratique : utilisé pour discriminer les différents points d'entrée dans un même AS voisin.
- **Le MED est un attribut associé aux liens inter-AS**
  - Il est soit configuré par l'AS local soit envoyé par l'AS voisin
  - Si reçus par lien externes, l'attribut MED doit être propagé vers les routeurs internes de l'AS
  - L'attribut MED n'est jamais propagé vers des AS voisins
- **Les MED de deux routes ne sont quelquefois comparés que s'ils concernent le même AS voisin**
  - Sa valeur est aussi appelée Métrique
  - Les plus petits MED sont préférés

MED



MED Oscillations (1)

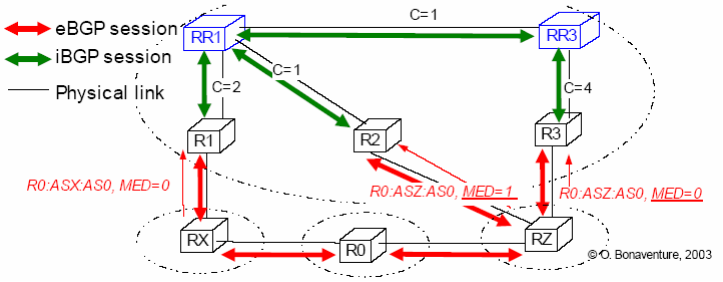


- Consider a single prefix advertised by R0 in AS0
  - R1, R2 and R3 always prefer their direct eBGP path
  - Due to the utilization of route reflectors, RR1 and RR3 only know a subset of the three possible paths
    - This limited knowledge is the cause of the oscillations

© O. Bonaventure, 2003

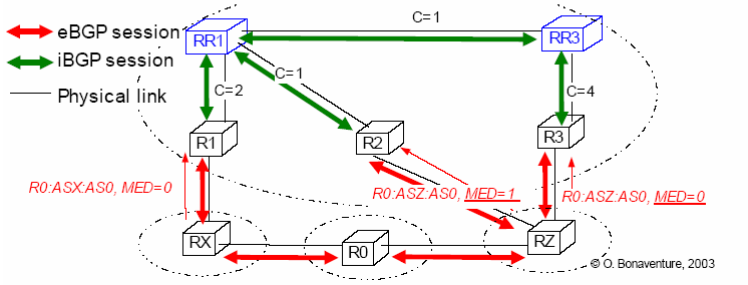
MED Oscillations (2)

- RR3's best path selection
  - If RR3 only knows the R3-RZ path, this path is preferred and advertised to RR1
  - RR3 knows the R1-RX and R3-RZ paths, R1-RX is best (IGP cost) and RR3 doesn't advertise a path to RR1
  - If RR3 knows the R2-RZ and R3-RZ paths, RR3 prefers the R3-RZ path (MED) and R3-RZ is advertised to RR1

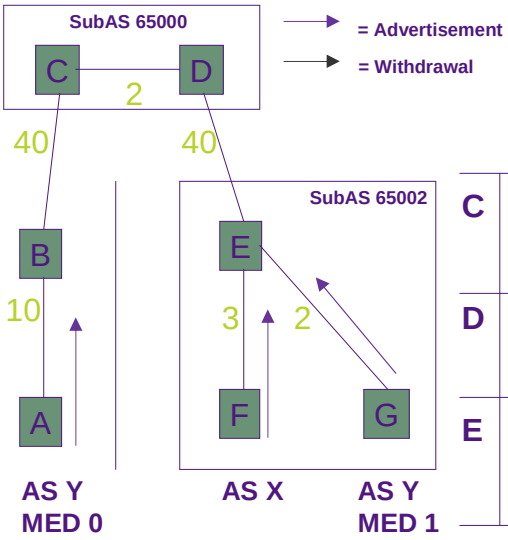


MED Oscillations (3)

- RR1's best path selection
  - If RR1 knows the R1-RX, R2-RZ and R3-RZ paths, R1-RX is preferred and RR1 advertises this path to RR3
    - But if RR1 advertises R1-RX, RR3 does not advertise any path !
  - If RR1 knows the R1-RX and R2-RZ paths, RR1 prefers the R2-RZ path and advertises this path to RR3
    - But if RR1 advertises R2-RZ, RR3 prefers and advertises R3-RZ !



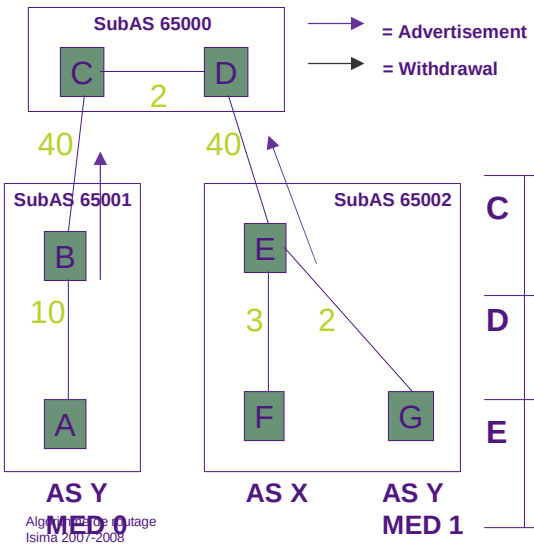
Oscillations avec MED (4)



Step 1
- **E selects Y1**

	AS_PATH	MED	IGP
C			
D			
E	X * Y	 1	3 2

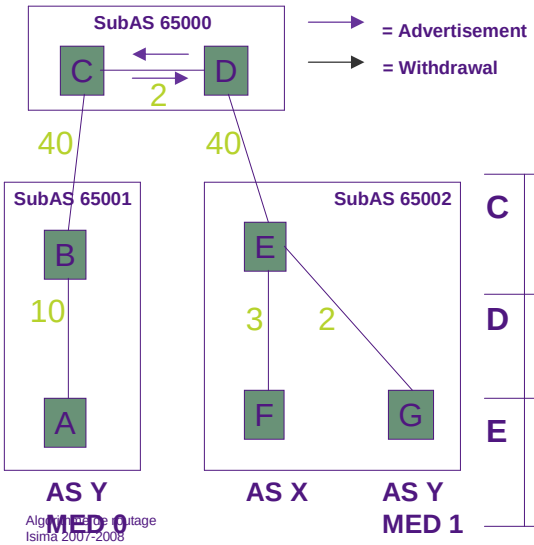
Oscillations avec MED (5)



Step 2
- **C selects Y0**
- **D selects Y1**

	AS_PATH	MED	IGP
C	* Y	0	50
D	* Y	1	42
E	X * Y	 1	3 2

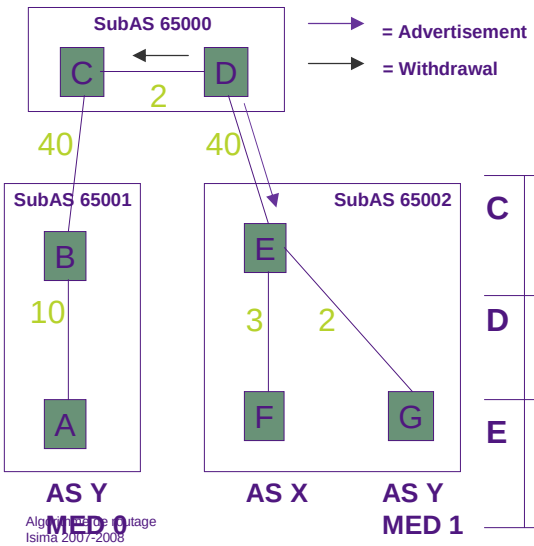
Oscillations avec MED (6)



Step 3
– D selects Y0

	AS_PATH	MED	IGP
C	* Y	0	50
	Y	1	44
D	* Y	0	52
	Y	1	42
E	X		3
	* Y	1	2

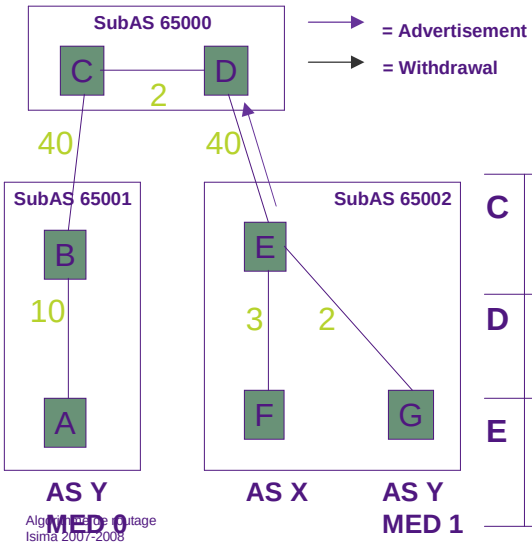
Oscillations avec MED (7)



Step 4
– E selects X

	AS_PATH	MED	IGP
C	* Y	0	50
D	* Y	0	52
	Y	1	42
E	* X		3
	Y	0	92
	Y	1	2

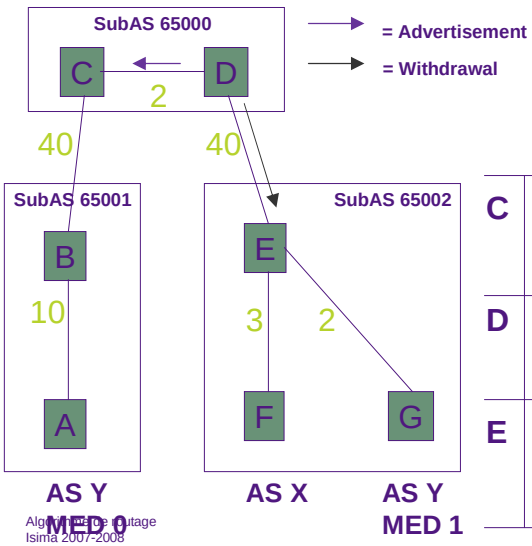
Oscillations avec MED (8)



Step 5
– D selects X

	AS_PATH	MED	IGP
C	* Y	0	50
D	Y	0	52
	* X		43
E	* X		3
	Y	0	92
	Y	1	2

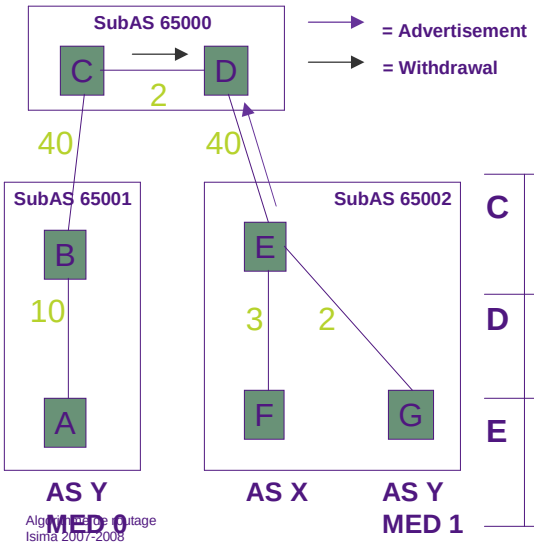
Oscillations avec MED (9)



Step 6
– C selects X
– E selects Y1

	AS_PATH	MED	IGP
C	Y	0	50
	* X		45
D	Y	0	52
	* X		43
E	X		3
	* Y	1	2

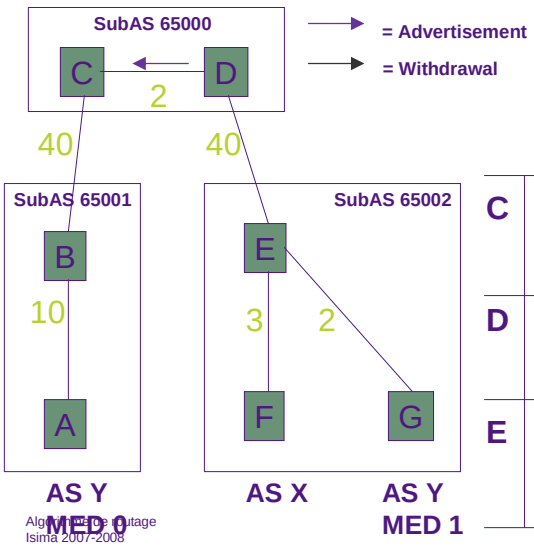
Oscillations avec MED (10)



Step 7
– D selects Y1

	AS_PATH	MED	IGP
C	Y	0	50
	* X		45
D			
	* Y	1	42
E	X		3
	* Y	1	2

Oscillations avec MED (11)



Step 8
– C selects Y0
– This is the same as Step 2
– BGP is in a loop

	AS_PATH	MED	IGP
C	* Y	0	50
	Y	1	44
D			
	* Y	1	42
E	X		3
	* Y	1	2

Oscillations avec MED: solution

- **Solution #1 – S'assurer que E a le chemin Y0**
  - Le routeurs BGP (BGP Peers) annonce plusieurs routes
    - BGP multipath…
  - Ajouter une session BGP
  - Un speaker BGP aurait besoin d'annoncer un chemin par groupe d' "AS voisin" [SI] le chemin vient d'un routeur interne. Cela forcerait dans l'exemple le routeur C a toujours annoncer la route vers Y0
- **Solution #2 – Éliminer le problème "Y0 < Y1 < X < Y0"**
  - Toujours comparer le MED!
    - Option "awlways compare med"

Origin Type: Code 1

- **Obligatoire : "well-known mandatory"**
  - propagé dans tous les messages
- **Correspond à la provenance de la route**
  - Permet de savoir comment la route a été injectée dans BGP
  - Cet attribut est fixé par les routeurs de l'AS qui annoncent originellement le préfixe
- **0 : IGP**
  - Le « Network Layer Reachability Information » est intérieur à l'AS
  - Le réseau a été injecté par configuration d'un routeur dans l'AS
- **1 : EGP**
  - Le NLRI a été connu par session EBGP
- **2 : imcomplete**
  - Autres cas : agrégation, redistribution, ou installation indirecte de la route dans BGP à l'intérieur de l'AS

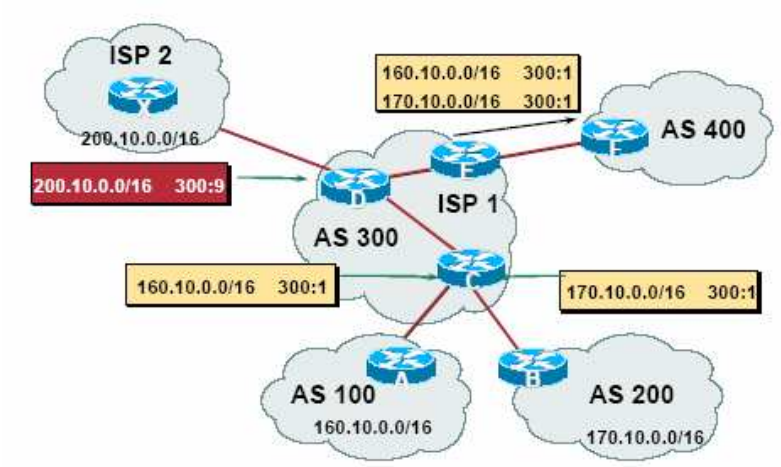
Communities

- **Optionnel et transitif (RFC 1997)**
  - C'est un marquage ("tag") qui permet de transporter une information quelconque sur une route dans un AS ou entres AS
- **Fournit un moyen de regrouper des routes (oudes destinations) qui partagent des "attributs" communs**
  - pour le traffic engineering. Elles ne sont pas utilisées dans le processus de décision BGP mais plutôt pour déclencher l'application de règles définies dans la politique de routage en fonction de la présence d'une valeur ou non
  - Les règles définies dans la politique de routage sont beaucoup plus simples à écrire avec les communautés
- **Les communautés ont la signification qu'on leur attribue, un opérateur va par exemple :**
  - Associer une communauté pour chaque route en fonction de son point d'entrée dans l'AS et de son point de sortie de l'AS.
  - Associer une communauté pour indiquer qu'une route est à préferer ou non par un AS voisin (indiquer une route de backup)

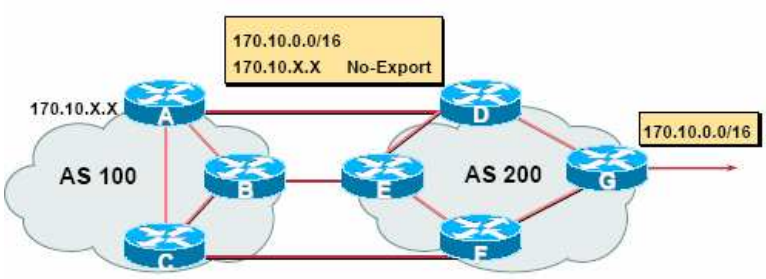
Communities

- **Fournit un moyen « scalable » de définir une bonne politique de routage**
- **Communautés: numéro sur 32 bits**
  - Numéro16bit : Numéro16bit
    - Exemple: [numéro de l'AS]:[numéro de la communauté]
      - *5511:2000 -> la communauté 2000 de l'AS 5511*
- **Il existe des communautés pré-définies qui sont reconnues par les implémentations**
  - NO_EXPORT : pas d'annonces aux routeurs des AS voisins
  - NO_ADVERSTISE: une route marquée avec cette communauté ne sera annoncée à aucun routeur voisin
  - …
- **L'attribut EXTENED_COMMUNITIES**
  - Une extension aux communautés avec quelques différences
    - De taille 64 bits
    - L'espace de valeur est plus structuré

Communities

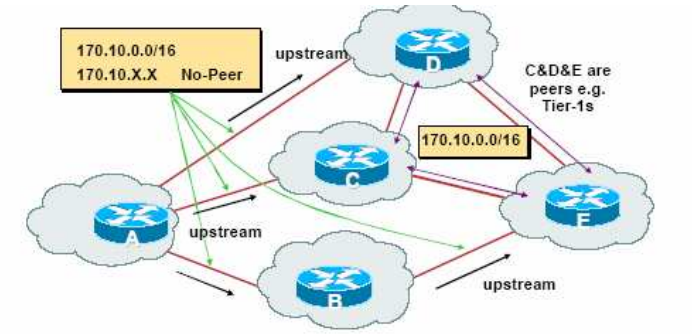


No Export



- AS100 announces aggregate and subprefixes
  aim is to improve loadsharing by leaking subprefixes
- Subprefixes marked with no-export community
- Router G in AS200 does not announce prefixes with no-export community set

No-Peer

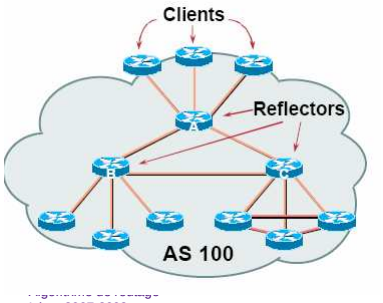


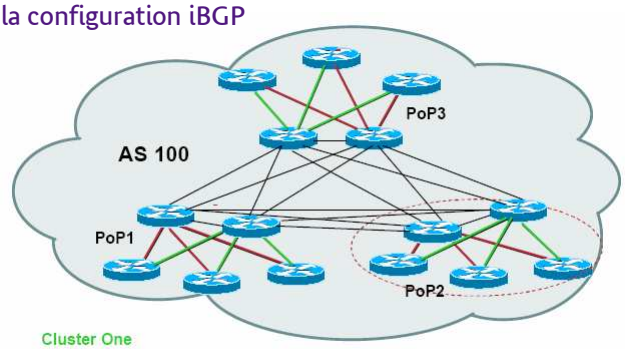
- Sub-prefixes marked with no-peer community are not sent to bi-lateral peers
  They are only sent to upstream providers

Plan

Route Relfectors

- **Pas de redistribution de route iBGP vers un voisin iBGP**
  - Limite le nombre de messages échangés et termine la propagation…
  - Oblige un full mesh des routeurs dans un AS en théorie!
- **Permet une alternative au Full Mesh iBGP**
  - Certains routeurs vont devenir esclaves des Route Reflector (RR)
  - Les RR font partie du Full Mesh
  - Permet de hiérarchiser la configuration iBGP



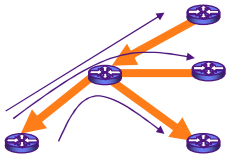


Cluster One
Cluster Two

Route Relfectors

- **Configuration d'un Route Reflector**
  - défini voisin par voisin
  - Un cluster-id est associé pour chaque "client"
    - CLUSTER_LIST est un attribut BGP qui stocke les clusters traversés au même titre que l'AS_PATH stocke les AS traversés. Cela permet d'éviter qu'un message ne se re-propage indéfiniment
  - Aucune configuration du client
- **Schéma de la retransmission effectuée par un route reflector**

Transmission d'une route en provenance d'un client

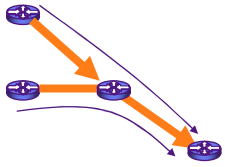


Propagation d'une route iBGP

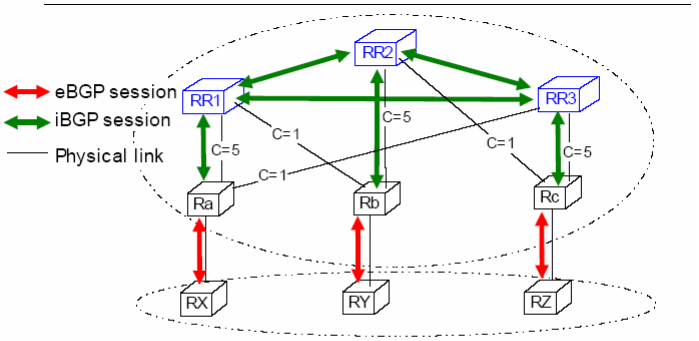
Session iBGP route reflector vers client

Session iBGP

Transmission d'une route

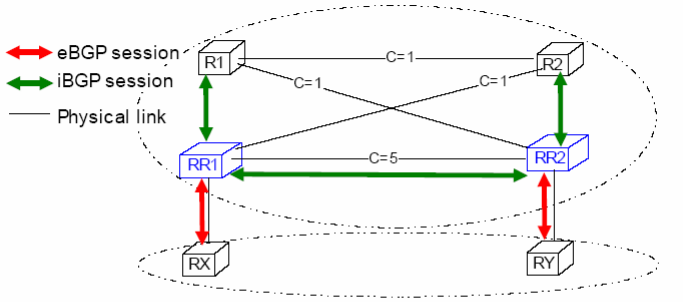


Problèmes avec Route Reflector



- Consider one prefix advertised by RX,RY,RZ
  - Ra, Rb, and Rc will all prefer their direct eBGP path
  - RR1, RR2 and RR3 will never reach an agreement

Problèmes avec Route Reflector

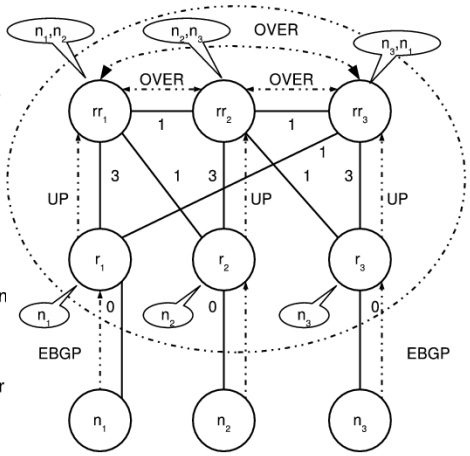


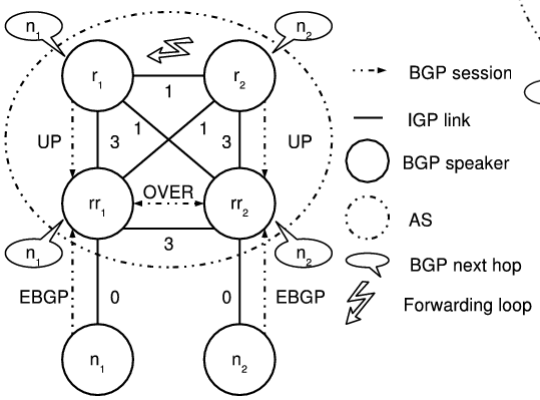
- Consider a prefix advertised by RX and RY
  - BGP routing will converge
    - RR1 (and R1) prefer path via RX, RR2 (and R2) prefer path via RY
  - But forwarding of IP packets will cause loop !
    - R1 sends packets towards prefix via R2 (to reach RX, its best path)
    - R2 sends packets towards prefix via R1 (to reach RY, its best path)

Problèmes avec Route Reflector

Oscillation de routage

- rr1,rr2,rr3 ne convergent jamais





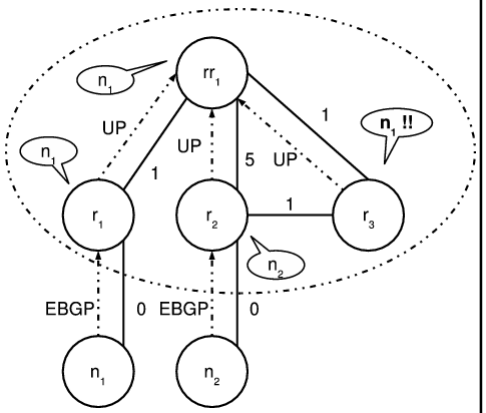
Boucle de routage

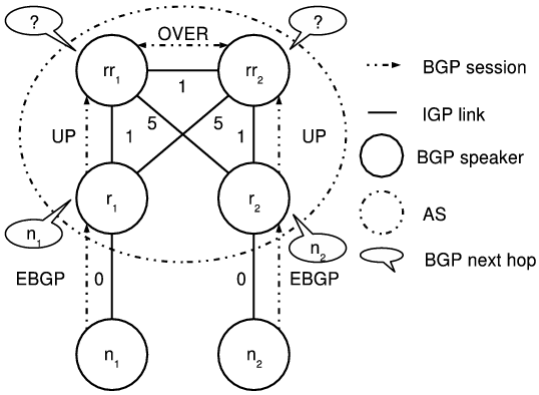
- Entre r1 et r2

D43

Problèmes avec Route Reflector

R3 n'apprend pas son point de sortie le plus proche





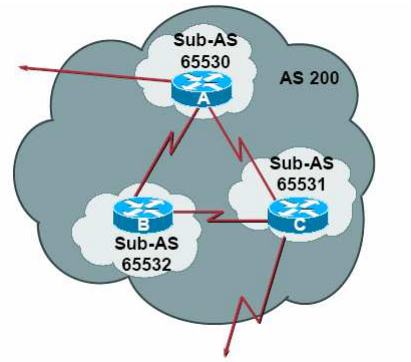
Non déterminisme

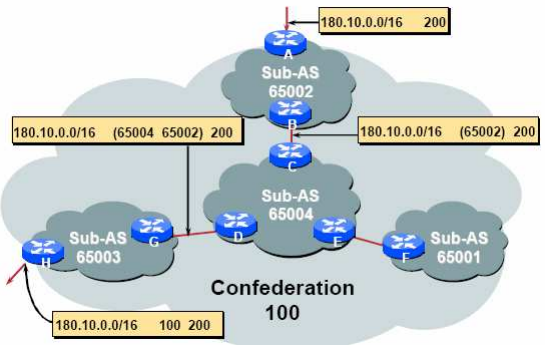
D44

Confédérations

- **Pour utiliser des sous-AS à l'intérieur d'un même AS (RFC 3065)**
  - Les sous-AS portent des numéros d'AS privés
  - eBGP entre sous-AS
  - Des informations iBGP sont quand même conservées
  - Préserve Local Pref, MED, Next HoP
- **Généralement un unique IGP est mis en place**
- **A l'extérieur de l'AS, les sous As sont invisibles**
  - Chaque sous-As utilise un numéro d'AS privé
- **Les routeurs sont en Full-Mesh dans chaque sous-AS**
- **Deux nouveaux sous-attributs dans les AS PATH sont ajoutés**
  - AS_CONFED_SEQUENCE : confédérations traversées dans un même AS
  - AS_CONFED_SET : équivalent de AS-SET amis pour les confédérations. Ils apparaissent lorsque l'on procède à des agrégations

Confédérations



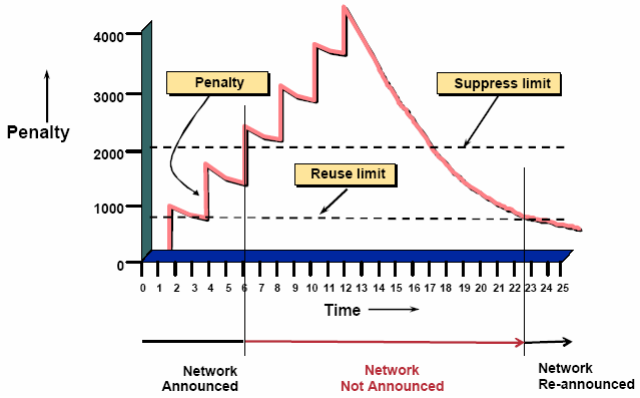


Temps de convergence inter-domaine

- **Il existe un mécanisme pour retarder la transmission d'updates**
  - Dépend des implémentations
    - Sur les routeurs Cisco: MRAI (Minimim Route Advertizement Interval)
    - Pas de mécanisme de la sorte sur les routeurs Juniper
  - Cela permet à un routeur de retarder l'envoi de mises à jour donc d'attendre d'avoir éventuellement reçu plusieurs routes avant de propager sa meilleure route
    - Par défaut configuré à 30 secondes sur les liaisons eBGP
  - Cela absorbe certaines instabilités mais peut retarder le temps de convergence suivant les cas

Convergence et stabilité

- **Il existe un mécanisme pour pénaliser les update trop fréquentes**
  - Route Flap Dampening



Agrégation

- **Le mécanisme d'agrégation de route permet de résumer plusieurs routes en une seule**
  - Permet de réduire le nombre de messages échangés
  - Permet de réduire la taille des tables de routage dans certains cas
- **L'agrégation permet de cacher**
  - Des NLRI : en regroupant plusieurs préfixes en un seul
  - De l'information topologique : en cachant les différentes AS PATH à partir du point d'agrégation
    - On utilise les AS-SET lorsqu'on agrège
- **Lorsque des préfixes sont inclus les uns dans les autres et que l'on agrège**
  - On fixe l'attribut ATOMIC AGGREGATE

Quelques pratiques pour de meilleures performances

- **Peer-Groups**
  - Dans la configuration d'un routeur BGP on peut définir des règles par groupes de voisins
    - Règles canoniques qui permettent dans un premier temps de simplifier les configurations de routeurs et l'implantation de la politique de routage
  - Peut permettre d'optimiser la réplication
    - Lorsqu'un routeur doit envoyer le même paquet à plusieurs voisins, il n'est quelquefois pas obligé de re-construire le message pour chaque voisin
- **"Update packing"**
  - Un routeur peut envoyer le même message BGP (avec les mêmes attributs) pour un ensemble de NLRI
- **Autre pratiques de "tuning" au niveau de la couche de transport**
  - Fast external fall-over
  - Optimisations TCP:
    - TCP Path Maximum transmission Unit (MTU) et Pcaket Buffer Overflow
- **La liste est loin d'être exhaustive…;)**

Quelques remarques et extensions

- **Sécurité des sessions BGP**
  - BGP-MD5 (rfc 2385)
- **BGP Multipath**
  - Permet à un routeur d'annoncer plusieurs chemins pour un même NLRI
    - Dans certains cas cela permet de résoudre des problèmes dus aux routes manquantes du fait de l'utilisation de route-reflector
- **Graceful restart**
  - Permet à un routeur d'annoncer qu'il va redémarrer aux autres routeurs
    - Cf. "overload-bit" dans lees protocoles IGP comme IS-IS
- **Modification du processus de décision BGP**
  - Disponible sur les routeurs cisco notamment…
- **En route vers des routeurs de plus en plus ouverts….**
  - Les routeurs Quagga progressent
  - Annonce du nouvel JunOS en décembre 2007 !