GRAND DUCHY OF LUXEMBOURG
Ministry of Foreign Affairs

Directorate for Development Cooperation

European Union Africa
Infrastructure Trust Fund

Techniques de mise à l'échelle BGP (Scaling)

Techniques de mise à l'échelle BGP (BGP Scaling)

- La spécification et mise en œuvre originale de BGP étaient suffisantes pour l'Internet du début des années 1990
  - Mais pas d’évolution
- Les questions qui ont accompagné le développement de l'Internet inclus:
  - Intensification du maillage iBGP au-delà de quelques pairs?
  - Mettre en œuvre une nouvelle politique sans causer de "flaps” et “route churning"?
  - Maintenir le réseau stable, évolutif, et simple?

Techniques de mise à l'échelle BGP (BGP Scaling)

- Les meilleures techniques actuelles de mise à l'échelle
  - Route Refresh
  - Peer-groups
  - Route reflector (et confédérations)
- Techniques de mise à l'échelle obsolètes
  - Soft Reconfiguration
  - Route Flap Damping

Route Refresh

Changements de politique non destructifs

Actualisation d'itinéraire (Route Refresh)

- Changements de politique:
  - Nécessité d’une réinitialisation dure des pairs BGP après chaque changement de politique parce que le routeur ne stocke pas les préfixes qui sont rejetés par la politique
- Réinitialisation dure de pairs BGP:
  - Redémarre les sessions BGP
  - consomme du CPU
  - Perturbe fortement la connectivité de tous les réseaux
- Solution:
  - Actualisation d'itinéraire (Route Refresh)

Capacité d’actualisation des itinéraires

- Facilite les changements de politique sans interruption/perturbation
- Aucune configuration n'est nécessaire
  - Automatiquement négocié à l'établissement des pairs
- Aucune mémoire supplémentaire n’est utilisée
- Les routeurs doivent supporter la " Capacité d’actualisation des routes» - RFC2918
- Informe les pairs d’envoyer à nouveau des annonces BGP complètes

  `clear ip bgp x.x.x.x [soft] in`
- Envoie à nouveau des annonces BGP complètes aux pairs

  `clear ip bgp x.x.x.x [soft] out`

Reconfiguration dynamique

- Utiliser la capacité d'actualisation des routes
  - Pris en charge sur pratiquement tous les routeurs
  - visible avec "show ip bgp neighbor"
  - Non perturbateur, “Good For the Internet”

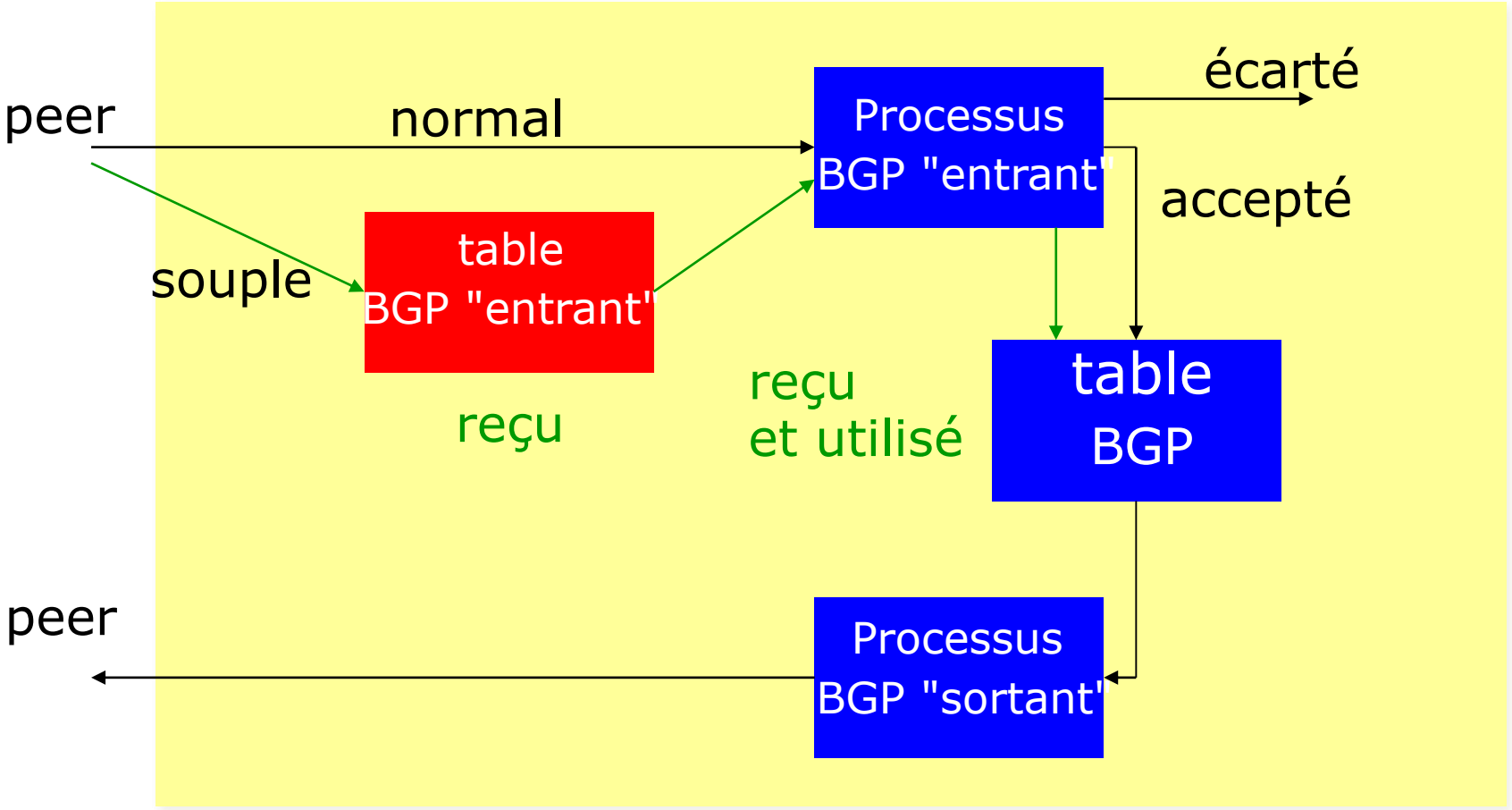
- Réaliser une réinitialisation hard d’un peering BGP seulement comme dernier recours

Considérez l'impact comme équivalent à un redémarrage du routeur

Soft reconfiguration de Cisco

- maintenant obsolète — mais:
- Un routeur stocke les préfixes qui ont été reçus des pairs, après application des politiques
  - L'activation de la soft reconfiguration signifie que le routeur stocke également les préfixes / attributs reçus avant toute application de politiques
  - Utilise plus de mémoire pour garder les préfixes dont les attributs ont été modifiés ou n'ont pas été acceptés
- Seulement utile quand l'opérateur nécessite de connaître quels préfixes ont été envoyés à un routeur avant l'application de toute politique entrante

Soft Reconfiguration de Cisco



Configurer Soft Reconfiguration

```
router bgp 100
 neighbor 1.1.1.1 remote-as 101
 neighbor 1.1.1.1 route-map infilter in
 neighbor 1.1.1.1 soft-reconfiguration inbound
! " Outbound " n'a pas besoin d'être configuré!
```

- Puis, quand on change la politique, nous émettons une commande exec

  `clear ip bgp 1.1.1.1 soft [in | out]`

- Note:
  - Lorsque la soft reconfiguration est activée, il n'y a pas d'accès à la capacité Route Refresh
  - **`clear ip bgp 1.1.1.1 [in | out]`** va aussi faire une actualisation soft

Peer Groups

Peer Groups

- Problème - comment mettre iBGP à l'échelle
  - Maillage d'iBGP large - lent à construire
  - Les voisins iBGP reçoivent la même mise à jour
  - Le CPU des Routeurs gaspillé à faire des calculs répétitifs
- Solution - Les groupes (peer-groups)
  - Grouper les pairs ayant la même politique sortante
  - Les mises à jour sont générées une fois par groupe

Peer Groups - Avantages

- Rend la configuration plus facile
- Rend la configuration moins sujette à erreur
- Rend la configuration plus lisible
- Réduit la charge qui pèse sur le CPU du routeur
- Le maillage iBGP se construit plus rapidement
- Les membres peuvent avoir différentes politiques entrantes
- Peut aussi être utilisé pour les voisins eBGP!

Configurer un Peer Group

```
router bgp 100
 neighbor ibgp-peer peer-group
 neighbor ibgp-peer remote-as 100
 neighbor ibgp-peer update-source loopback 0
 neighbor ibgp-peer send-community
 neighbor ibgp-peer route-map outfilter out
 neighbor 1.1.1.1 peer-group ibgp-peer
 neighbor 2.2.2.2 peer-group ibgp-peer
 neighbor 2.2.2.2 route-map  infilter in
 neighbor 3.3.3.3 peer-group ibgp-peer
```

! noter comment 2.2.2.2 a un filtre entrant différent des groupes de pairs!

Configurer un Peer Group

```
router bgp 100
 neighbor external-peer peer-group
 neighbor external-peer send-community
 neighbor external-peer route-map set-metric out
 neighbor 160.89.1.2 remote-as 200
 neighbor 160.89.1.2 peer-group external-peer
 neighbor 160.89.1.4 remote-as 300
 neighbor 160.89.1.4 peer-group external-peer
 neighbor 160.89.1.6 remote-as 400
 neighbor 160.89.1.6 peer-group external-peer
 neighbor 160.89.1.6 filter-list infilter in
```

Des groupes de pairs (Peer Groups)

- Toujours configurer peer groups pour iBGP
  - Même s'il y n'a seulement que quelques pairs iBGP
  - Facile à mettre à l'échelle le réseau dans le futur
- Pensez à utiliser les peer groups pour eBGP
  - Particulièrement utile pour de multiples clients BGP utilisant le même AS (RFC2270)
  - Également utile aux points d'échange où les politiques ISP sont généralement les mêmes pour chaque pair
- Les peer groups sont jugés obsolètes
  - Mais sont encore largement considérés comme bonne pratique
  - Remplacés par update-groups (codage interne - non configurable)
  - Renforcés par peer-templates - (permettant des constructions plus complexes)

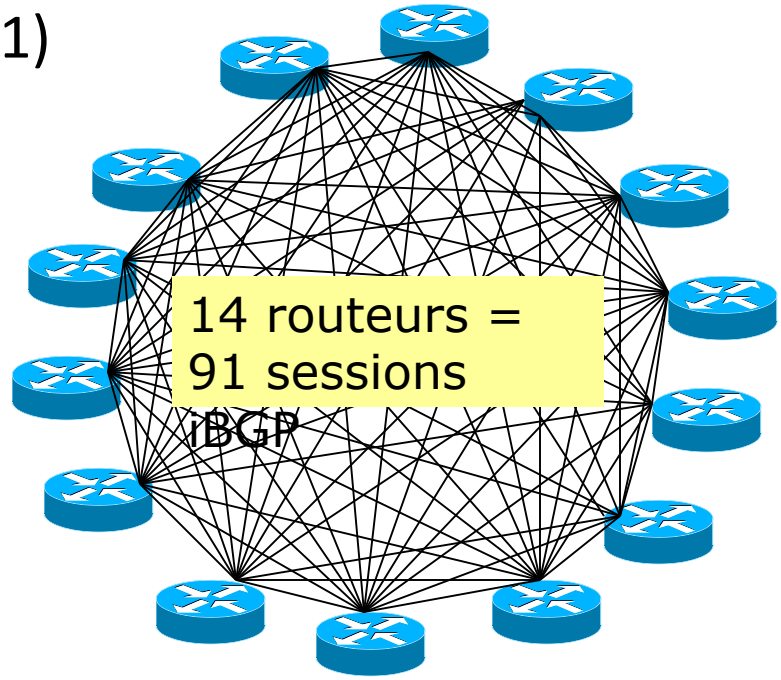
Route Reflector

Mise à l'échelle du maillage iBGP

Mise à l'échelle du maillage iBGP

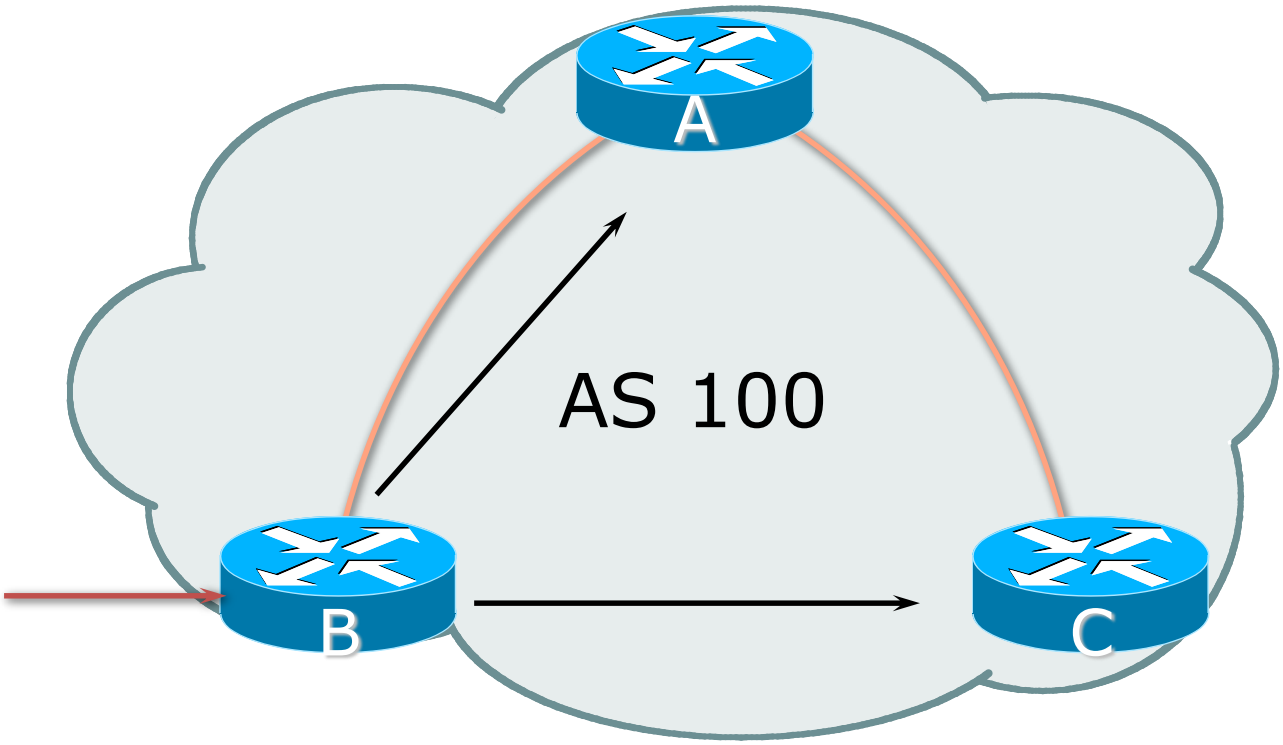
- Éviter un maillage iBGP ½ n (n-1)

n=1000 ⇒ presque un demi million de sessions ibgp !

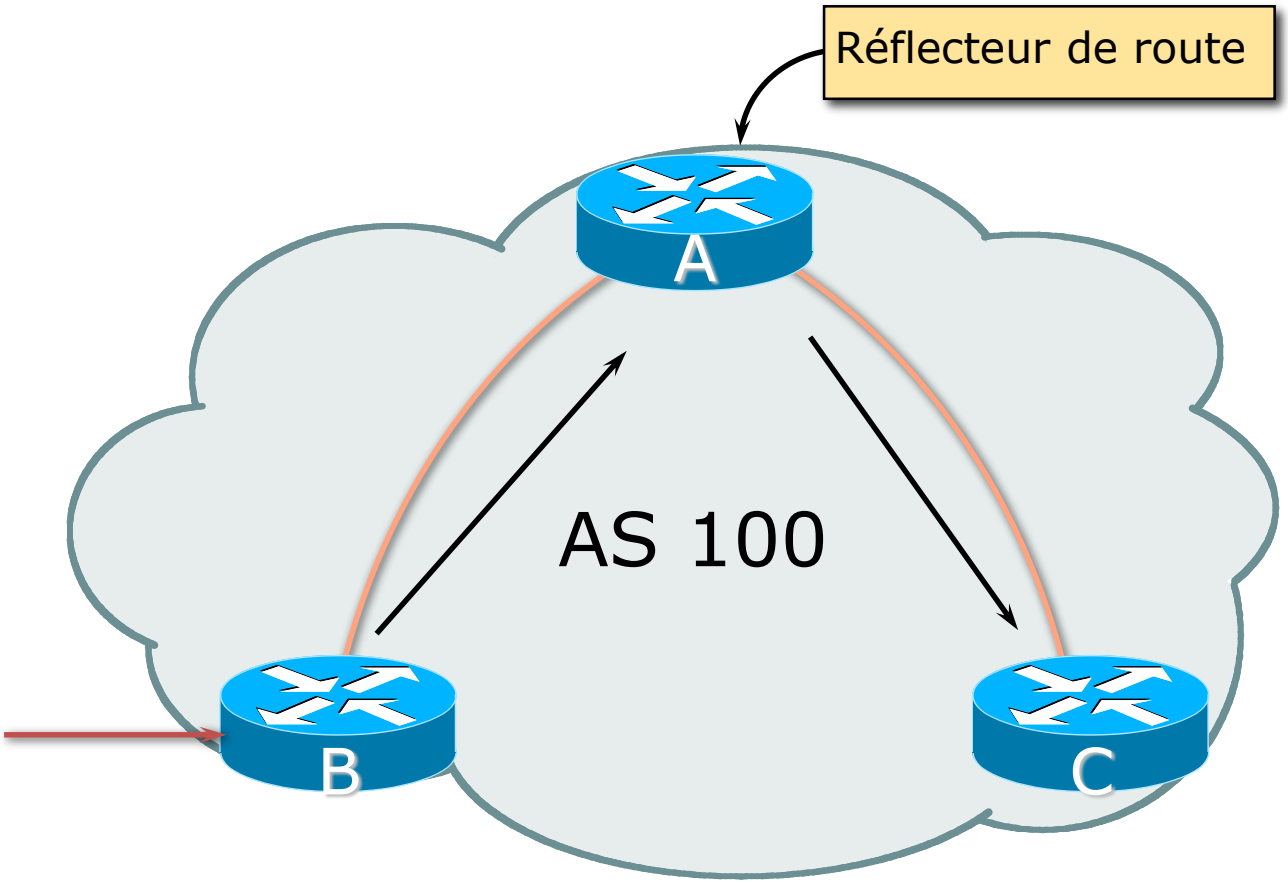


- Deux solutions
  - Réflecteur de route - plus simple à déployer et exécuter
  - Confédération - plus complexe, présente des avantages " corner case "

Réflecteur de route: Principe

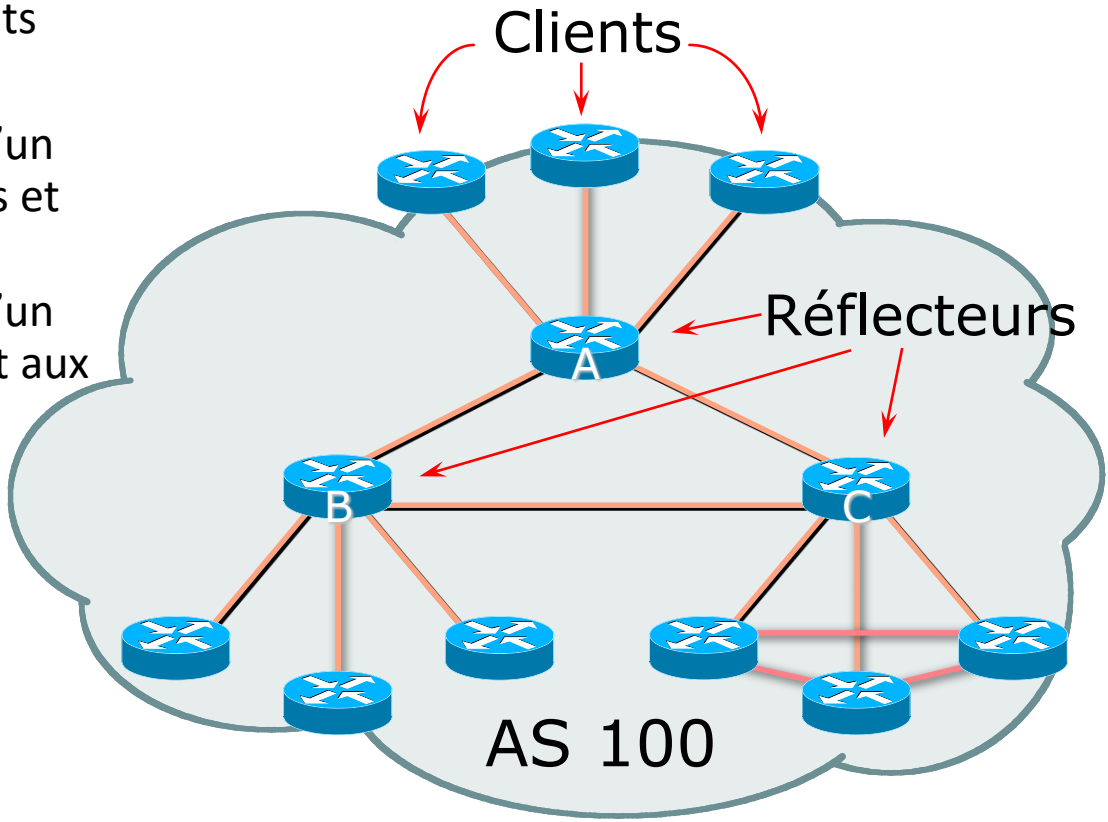


Réflecteur de route: Principe



Réflecteur de route

- Le réflecteur reçoit des itinéraires de la part de clients et de non-clients
- Sélectionne le meilleur chemin
- Si le meilleur chemin provient d'un client, réfléchir à d'autres clients et non-clients
- Si le meilleur chemin provient d'un non-client, réfléchir uniquement aux clients
- Clients non-maillés
- Décrit dans RFC4456



Topologie de réflecteurs de route

- Diviser le backbone en plusieurs clusters
- Au moins un réflecteur de route et quelques clients par cluster
- Les réflecteurs de route sont entièrement maillés
- Les clients dans un cluster pourraient être entièrement maillés
- Un IGP unique pour transporter " next hop " et routes locales

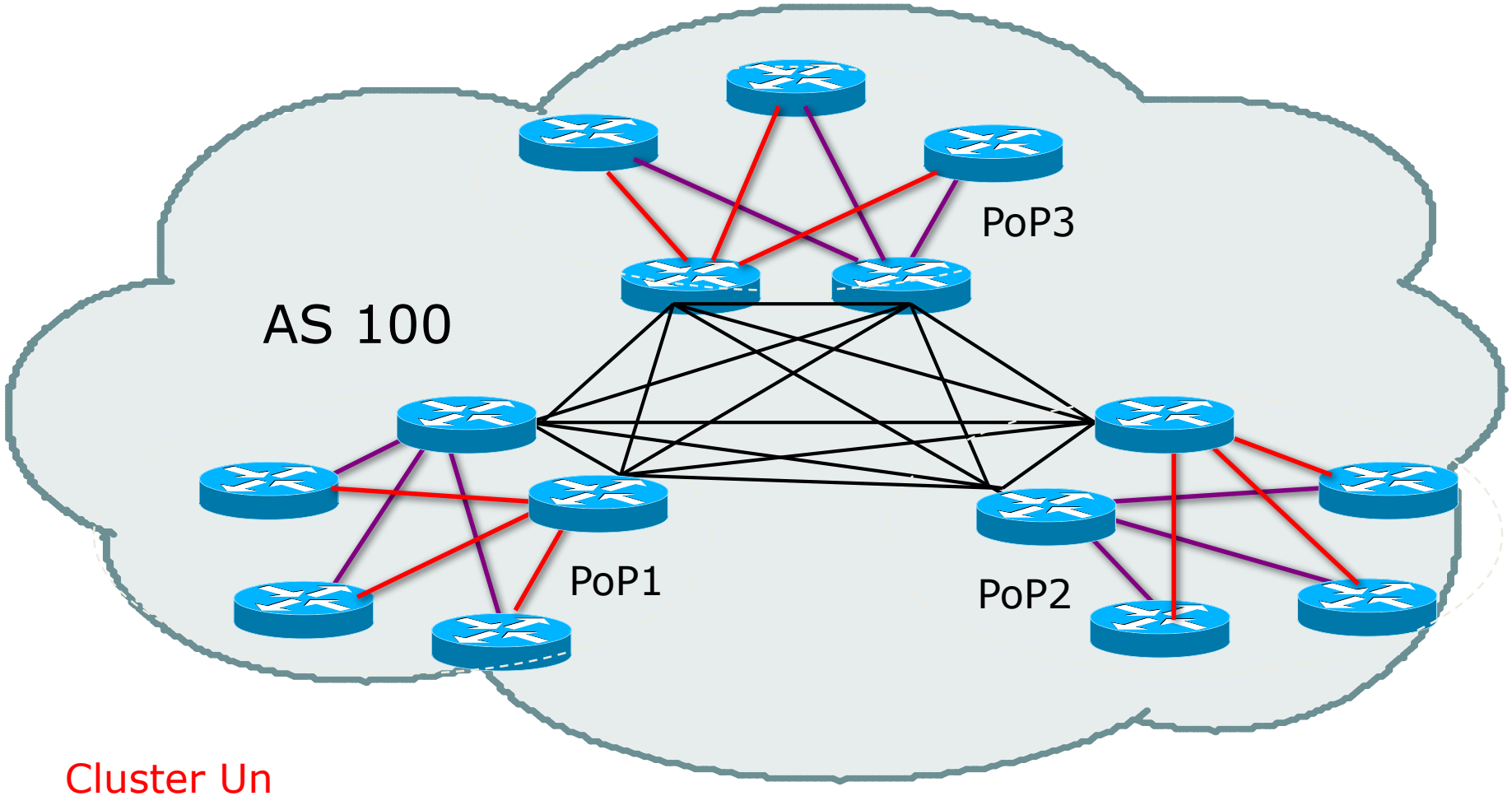
Réflecteurs de route : Éviter les boucles (Loop Avoidance)

- Originator_ID attribute
  - Porte le RID de l'auteur de la route dans l'AS local (créé par la RR)
- Cluster_list attribute
  - Le cluster-id local est ajouté lorsque la mise à jour est envoyée par le RR
  - Cluster-id est routeur-id (adresse de loopback)
  - NE PAS utiliser **`bgp cluster-id x.x.x.x`**

Réflecteurs de route: Redondance

- De multiples RR peuvent être configurés dans le même cluster - pas conseillé!
  - Tous les RR du cluster doivent avoir le même cluster-id (sinon c'est un autre cluster)
- Un routeur peut être un client de RR dans différents clusters
  - Courant aujourd'hui dans les réseaux ISP de superposer deux clusters - redondance obtenue de cette façon
  - $\rightarrow$ Chaque client dispose de deux RR = redondance

Réflecteurs de route: Redondance



Cluster Un

Cluster Deux

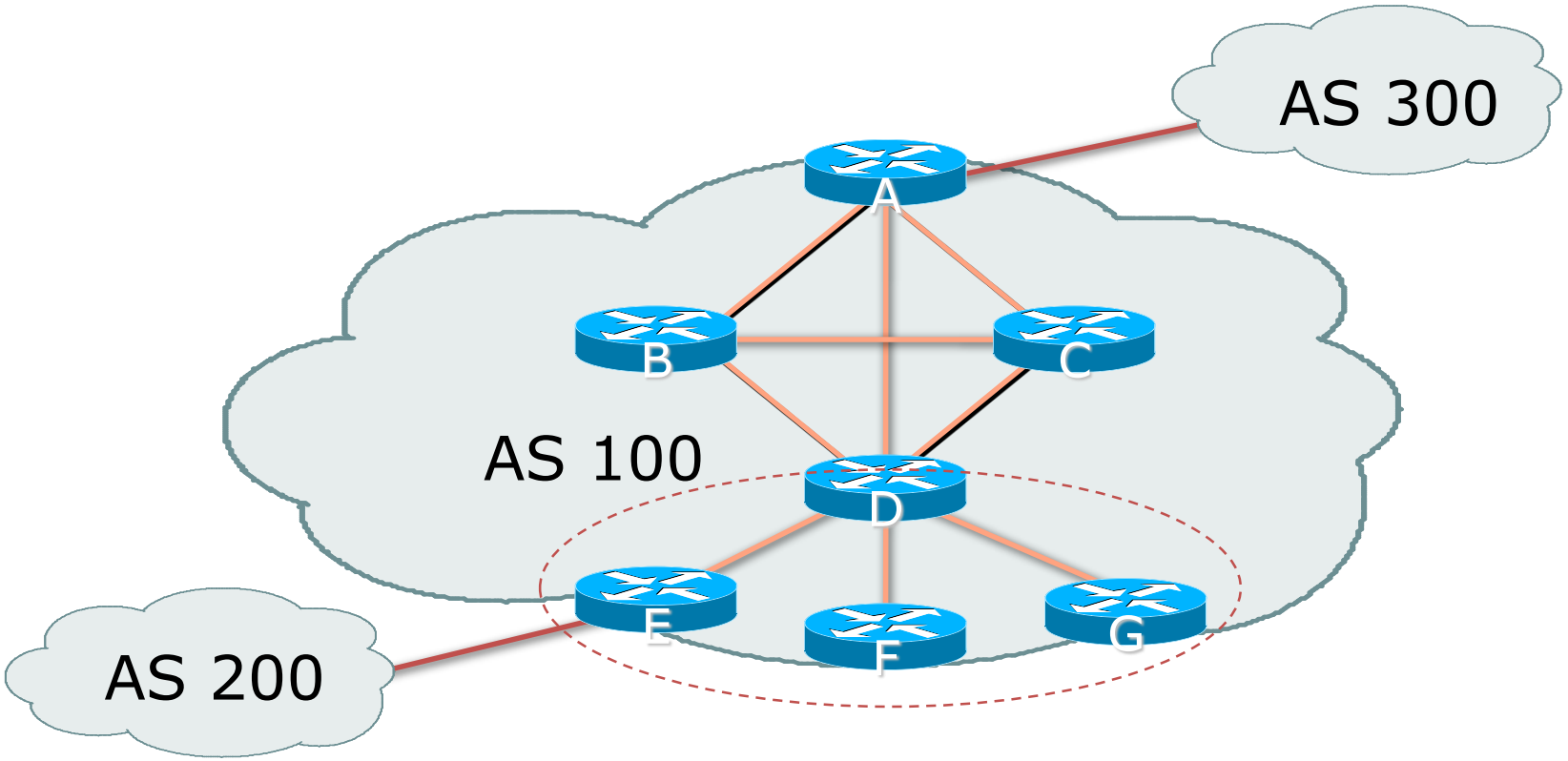
Réflecteur de route: Avantages

- Résout les problèmes de maillage iBGP
- Le transfert de paquets n'est pas affecté
- Des speakers BGP normaux co-existent
- Réflecteurs multiples pour la redondance
- Migration facile
- Plusieurs niveaux de réflecteurs de route

Réflecteurs de route: Migration

- Où placer les réflecteurs de route?
  - Suivez la topologie physique!
  - Cela permettra de garantir que la transmission de paquets ne sera pas affectée
- Configurer un RR à la fois
  - Éliminer les sessions iBGP redondantes
  - Placer un RR par cluster

Réflecteurs de route: Migration



- Migrer de petites parties du réseau, une partie à la fois.

Configurer un réflecteur de route

- Configuration du routeur D:

```
router bgp 100
 ...
 neighbor 1.2.3.4 remote-as 100
 neighbor 1.2.3.4 route-reflector-client
 neighbor 1.2.3.5 remote-as 100
 neighbor 1.2.3.5 route-reflector-client
 neighbor 1.2.3.6 remote-as 100
 neighbor 1.2.3.6 route-reflector-client
 ...
```

Techniques de mise à l'échelle BGP (BGP Scaling Techniques)

- Ces 3 techniques devraient être des exigences de base sur tous les réseaux ISP
  - Route Redresh (ou soft reconfiguration)
  - Peer Groups
  - Route Reflector

Confédérations BGP

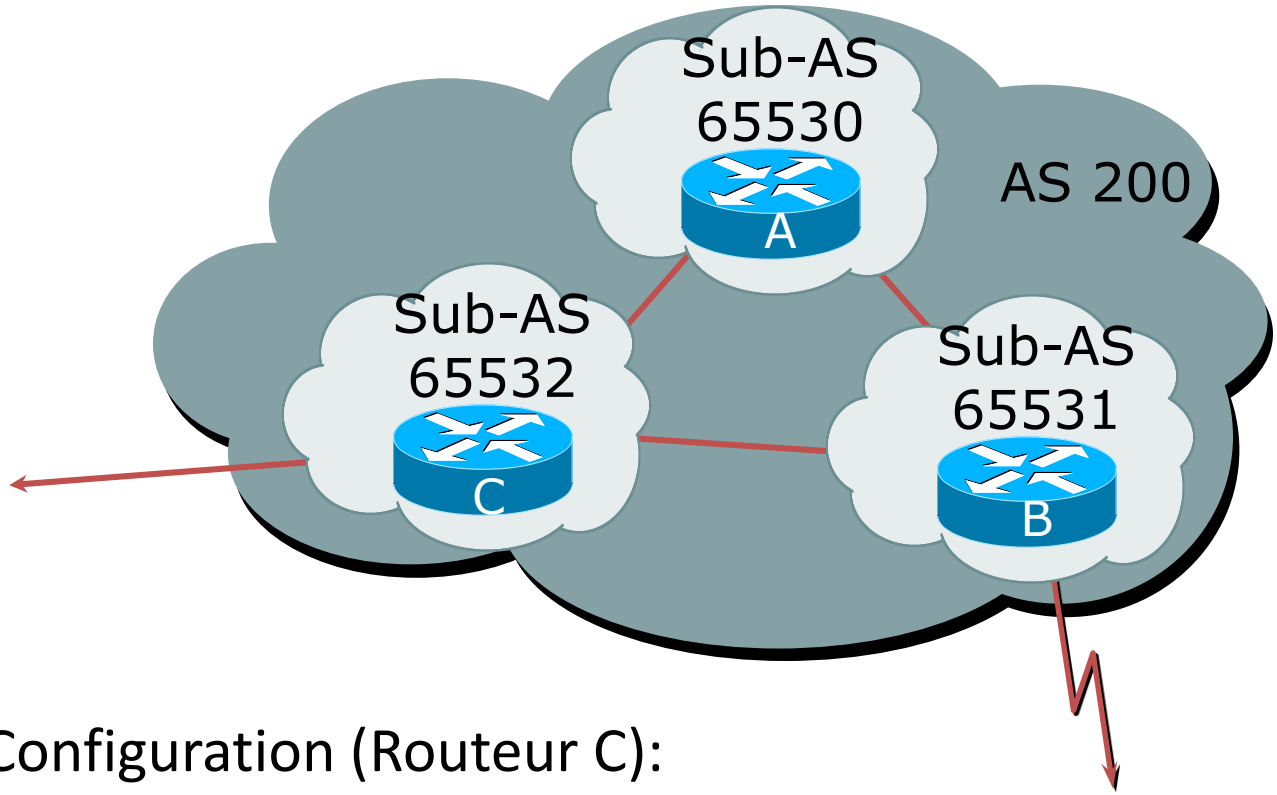
Confédérations

- Diviser l'AS en sous-AS
  - eBGP entre sous-AS, mais certaines informations iBGP sont maintenues
    - Préserver NEXT_HOP à travers les sous-AS (IGP porte cette information)
    - Préserver LOCAL_PREF et MED
- Habituellement, un seul IGP
- Décrit dans RFC5065

Confédérations

- Visible au monde extérieur comme un AS unique - " identificateur de Confédération "
  - Chaque sous-AS utilise un nombre de l'espace privé (64512-65534)
- Les routers iBGP dans les sous-AS sont entièrement maillés
  - Le nombre total de voisins est réduit en limitant l'exigence de maillage complet qu’aux seuls pairs dans les sous-AS

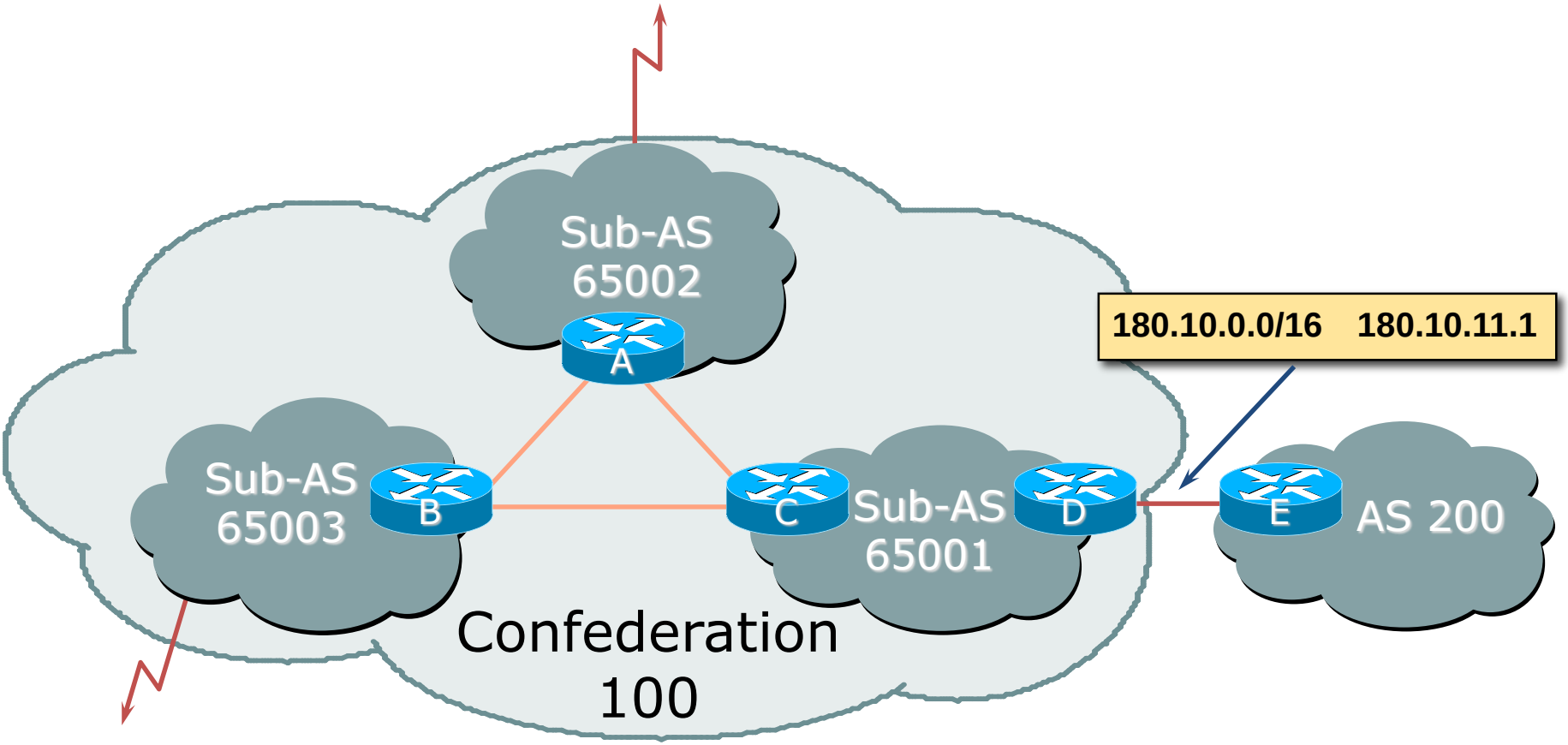
Confédérations



- Configuration (Routeur C):

```
router bgp 65532
 bgp confederation identifier 200
 bgp confederation peers 65530 65531
 neighbor 141.153.12.1 remote-as 65530
 neighbor 141.153.17.2 remote-as 65531
```

Confédérations: Next Hop



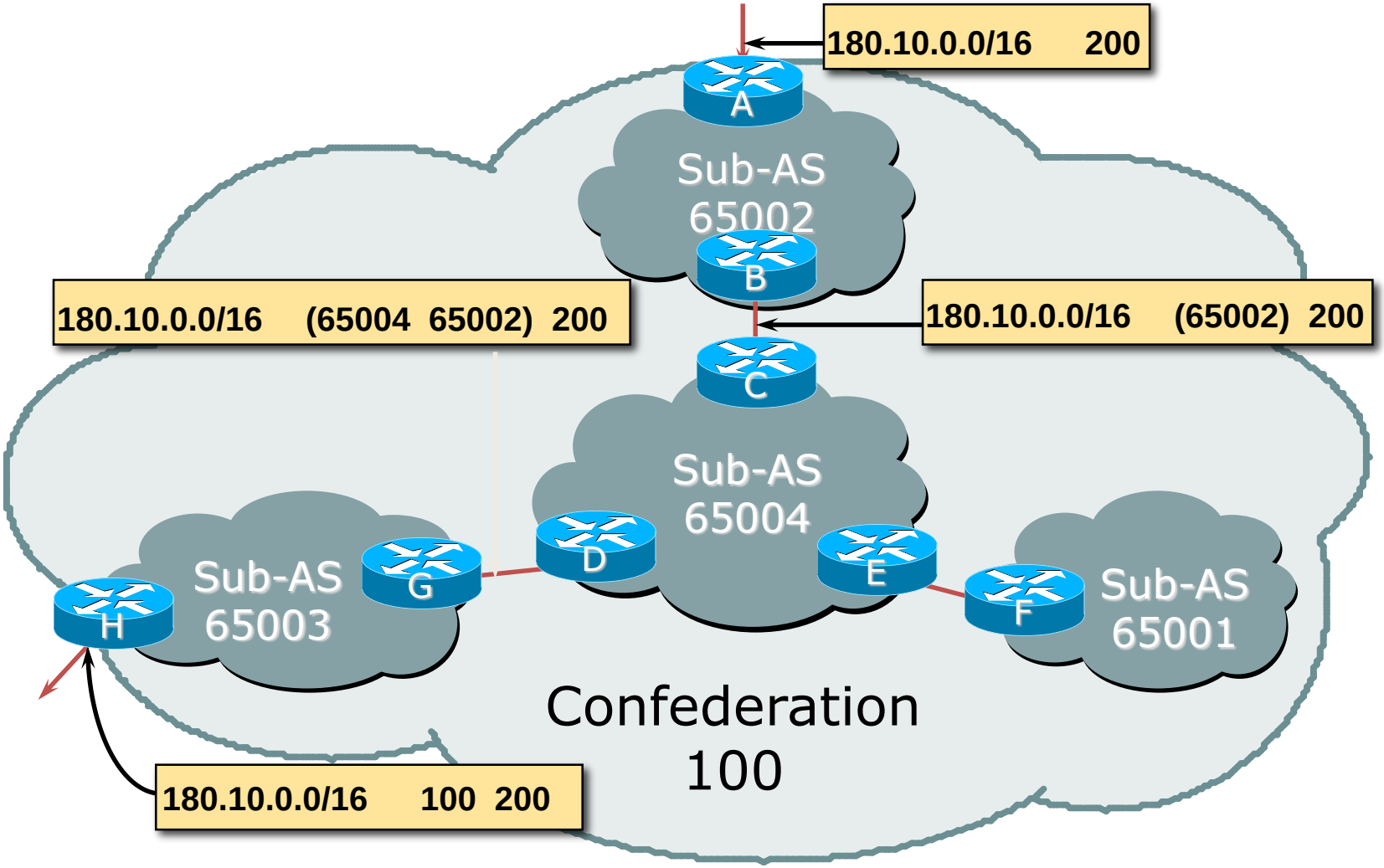
Confédération: Principe

- La préférence locale et MED influencent la sélection de chemin
- Préserver la préférence locale et MED à travers les frontières des sous-AS
- Sub-AS eBGP path administrative distance

Confédérations: éviter les boucles (Loop Avoidance)

- Les traversés sous-AS sont effectuées dans le cadre de l'AS-path
- Séquence AS et longueur de trajet AS
- Les limites des confédérations
- Séquence AS doit être omise lors de la comparaison MED

Confédérations: séquences -AS



Les décisions de propagation de route

- Même chose qu'avec BGP «normal»:
  - À partir de pairs dans le même sous-AS → uniquement vers des pairs externes
  - À partir de pairs externes → vers tous les voisins
- «Pairs externes» désigne
  - des pairs en dehors de la confédération
  - Pairs dans un autre sous-AS
    - préservent LOCAL_PREF, MED et NEXT_HOP

Confédérations (suite)

- Exemple (suite):

```
BGP table version is 78, local router ID is 141.153.17.1
Status codes: s suppressed, d damped, h history, * valid, >
  best, i - internal
Origin codes: i - IGP, e - EGP, ? - incomplete
Network          Next Hop      Metric LocPrf Weight Path
*> 10.0.0.0     141.153.14.3    0     100      0    (65531) 1 i
*> 141.153.0.0 141.153.30.2    0     100      0    (65530) i
*> 144.10.0.0  141.153.12.1    0     100      0    (65530) i
*> 199.10.10.0 141.153.29.2    0     100      0    (65530) 1 i
```

Plus de points sur les confédérations

- Peut soulager "l'absorption" d'autres ISP dans votre ISP
  - par exemple, si un ISP achète un autre ISP
  - (peut utiliser les caractéristiques d'un AS local pour faire la même chose)
- Vous pouvez utiliser des réflecteurs de routes avec la confédération sous-AS pour réduire le maillage iBGP du sous-AS

Confédérations : Avantages

- Résout les problèmes de maillage iBGP
- Le transfert de paquets n'est pas affecté
- Peut être utilisé avec des réflecteurs de route
- Des politiques pourraient être appliquées pour acheminer le trafic entre les sous-AS

Confédérations: Mises en garde

- Nombre minimal de sous-AS
- Hiérarchie des sous-AS
- Inter-connectivité minimale entre les sous-AS
- Diversité de trajet
- Migration difficile
  - BGP reconfiguré en sous-AS
  - doit être appliquée à travers le réseau

RR ou Confédérations

	Connectivité Internet	Hiérarchie à niveaux multiples	Politique Contrôle	Évolutivité	Complexité des migrations
Confédérations	N'importe où dans le réseau	Oui	Oui	Moyen	Moyen à élevé
Réflecteurs de route	N'importe où dans le réseau	Oui	Oui	Très élevé	Très faible

La plupart des nouveaux réseaux fournisseurs de services désormais déploie des réflecteurs de route dès la première journée

Route Flap Damping

Stabilité réseau pour les années 1990

Instabilité réseau pour le 21e siècle!

Route Flap Damping

- Pendant de nombreuses années, Route Flap Damping était une pratique fortement recommandée
- Maintenant, elle est fortement déconseillée car elle provoque beaucoup plus d’instabilité réseau qu’elle n’en guérit.
- Mais d'abord, la théorie ...

Route Flap Damping

- Route flap
  - Montée et descente de trajectoire (path) ou changement dans l'attribut
    - BGP WITHDRAW followed by UPDATE = 1 flap
    - eBGP neighbour going down/up is NOT a flap
  - Se répercute à travers tout l'Internet
  - Gaspille du CPU
- Damping vise à réduire la portée de propagation des " route flap "

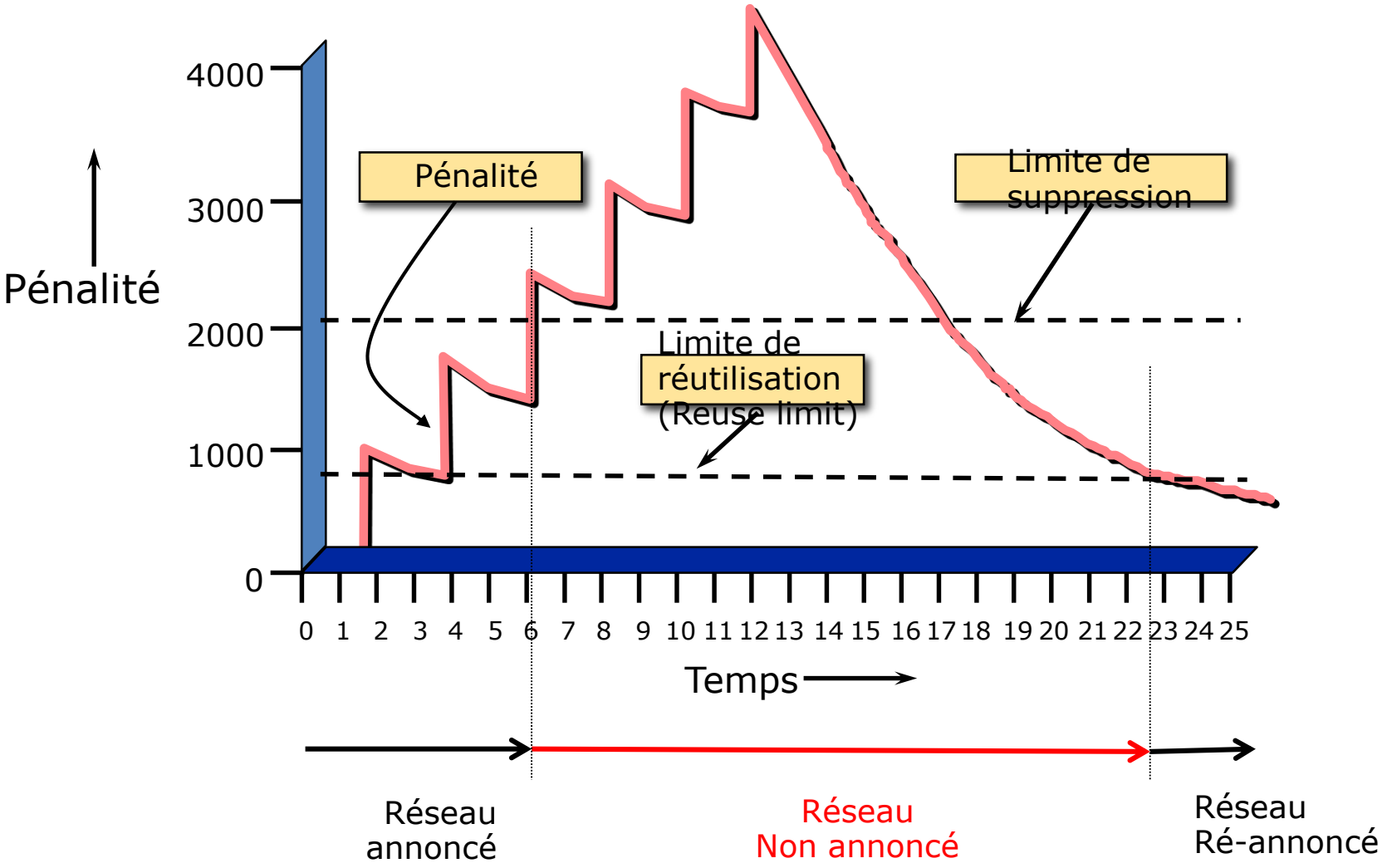
Route Flap Damping (suite)

- Exigences
  - Convergence rapide pour les changements de route normaux
  - L'histoire prédit les comportements futurs
  - Supprime les routes oscillantes
  - Annonce les routes stables
- Mise en œuvre décrit dans la RFC 2439

Opération

- Ajouter une pénalité (1000) pour chaque " flap "
  - Un changement dans l'attribut reçoit une pénalité de 500
- Exponentially decay penalty
  - la demi-vie détermine le taux de décomposition
- Pénalité au-dessus de limite de suppression
  - ne pas annoncer (advertise) les routes aux pairs BGP
- Penalty decayed below reuse-limit
  - Ré-annoncer les routes aux pairs BGP
  - Pénalité remise à zéro quand elle est à la moitié de la limite de réutilisation (reuse limit)

Opération


Pénalité
4000
3000
2000
1000
0
Pénalité
Limite de suppression
Limite de réutilisation (Reuse limit)
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
Temps
Réseau annoncé
Réseau Non annoncé
Réseau Ré-annoncé

Opération

- Uniquement appliqué aux annonces entrantes depuis les pairs eBGP
- Chemins alternatifs encore utilisables
- Contrôlé par :
  - La demi-vie (par défaut 15 minutes)
  - limite de réutilisation (par défaut 750)
  - limite de suppression - (par défaut 2000)
  - temps de suppression maximum (par défaut 60 minutes)

Configuration

- Damping fixé (Fixed damping)

```
router bgp 100
bgp dampening [<half-life> <reuse-value> <suppress-
  penalty> <maximum suppress time>]
```

- Damping sélectif et variable

```
bgp dampening [route-map <name>]
route-map <name> permit 10
 match ip address prefix-list FLAP-LIST
 set dampening [<half-life> <reuse-value>
  <suppress-penalty> <maximum suppress time>]
ip prefix-list FLAP-LIST permit 192.0.2.0/24 le 32
```

Opération

- Soins requis lors de la configuration de paramètres
- La pénalité doit être inférieure à la limite de réutilisation au temps de suppression maximum
- Temps de suppression maximum et demi-vie doivent permettre à la pénalité d'être plus grande que la limite de suppression

Configuration

- Exemples – ✘
  - bgp dampening 15 500 2500 30
    - Limite de réutilisation de 500 signifie que la pénalité maximale possible est de 2000 - aucun préfixe supprimé à titre de pénalité ne peut excéder la limite de suppression
- Exemples – ✓
  - bgp dampening 15 750 3000 45
    - Limite de réutilisation de 750 signifie que la pénalité maximale possible est de 6000 – la limite de suppression est facilement atteinte

Maths!

- Valeur maximale de la pénalité est

$$\text{max-penalty} = \text{reuse-limit} \times 2^{\left\langle \frac{\text{max-suppress-time}}{\text{half-life}} \right\rangle}$$

- Veillez toujours à ce que la limite de suppression soit INFÉRIEURE à la pénalité maximale sinon il n'y aura pas de route damping

Histoire de Route Flap Damping

- Premières mises en œuvre sur l'Internet en 1995
- Paramètres par défaut des vendeurs trop rigoureux
  - Recommandations du groupe de travail sur RIPE Routing dans ripe-178, ripe-210, et ripe-229
  - http://www.ripe.net/ripe/docs
  - Mais de nombreux ISP se sont tout simplement positionnés sur les valeurs par défaut des vendeurs sans réfléchir

Problèmes graves:

- "Route Flap Damping Exacerbates Internet Routing Convergence"
  - Zhuoqing Morley Mao, Ramesh Govindan, George Varghese & Randy H. Katz, August 2002
- “What is the sound of one route flapping?”
  - Tim Griffin, June 2002
- Divers travaux sur la convergence de routage par Craig Labovitz et Abha Ahuja il y a quelques années
- “Happy Packets”
  - Des travaux étroitement liés par Randy Bush et al

Problème 1:

- One path flaps:
  - Les Speaker BGP choisissent la prochaine meilleure route, annoncent à tous les pairs, et le compteur flap est incrémenté
  - Ces pairs voient le changement du meilleur chemin, le compteur flap est incrémenté
  - Après quelques hops, les pairs voient de multiples changements simplement causés par un seul flap $\rightarrow$ le préfixe est supprimé

Problème 2:

- Différentes implémentations de BGP ont différents temps de transmission pour les préfixes
  - Certains restent sur des préfixes pour un certain temps avant d'annoncer (advertising)
  - D'autres annoncent immédiatement
- La course à la ligne d'arrivée provoque l'apparition de flapping, causé par une simple annonce ou changement de chemin $\rightarrow$ le préfixe est supprimé

Solution:

- NE PAS utiliser Route Flap Damping quoi que vous fassiez!
- RFD compromet inutilement l'accès à:
  - Votre réseau et
  - l'Internet
- Plus d'informations contenues dans les recommandations du groupe de travail sur le Routage RIPE:
  - www.ripe.net/ripe/docs/ripe-378.[pdf,html,txt]
- Des travaux sont en cours pour essayer de trouver des moyens de rendre RFD utilisable:
  - http://datatracker.ietf.org/doc/draft-ymbk-rfd-usable/

Reconnaissance et attribution

Cette présentation contient des informations initialement développées et maintenues par les organisations et individu suivants et prévues pour le projet AXIS de l'Union africaine

Cisco ISP/IXP Workshops

Philip Smith: - pfsinoz@gmail.com

www.apnic.net

GRAND DUCHY OF LUXEMBOURG
Ministry of Foreign Affairs

Directorate for Development Cooperation

European Union Africa
Infrastructure Trust Fund

Techniques de mise à l'échelle BGP (BGP Scaling Techniques)

Fin