BGP Técnicas de escalamiento

Talleres ISP

BGP Técnicas de escalamiento

- La especificación original de BGP y la implementación estaba bien para el Internet de la década de 1990
  - Pero no escalaba
- Cuestiones como el crecimiento de internet incluyen:
  - ¿Escalar la malla iBGP más allá de algunos *peers*?
  - Implementar una nueva política sin causar intermitencia y rutas mezcladas
  - ¿Mantener la red estable, escalable, tan bien como sea posible?

BGP Técnicas de escalamiento

- Actuales mejores prácticas en técnicas de escalamiento
  - Refrescamiento de ruta
  - Grupos-*peer* de Cisco
  - Reflectores de ruta (y confederaciones)
- Técnicas de escalamiento en desuso
  - Reconfiguración suave
  - Amortiguación de ruta intermitente

Reconfiguración dinámica

Cambios a la política no destructivos

Refrescamiento de ruta

- Restablecimiento de pares (*peers*) BGP se requiere después de cada cambio de política
  - Porque el enrutador no almacena prefijos que son rechazados por la política
- Reinicio duro de pares (*peers*) BGP:
  - Drena el *peering* BGP y consume CPU
  - Interrumpe severamente la conectividad para todas las redes
- Reinicio suave de pares (*peers*) BGP (o refrescamiento de ruta):
  - peering BGP permanece activo
  - Impacta solamente esos prefijos afectados por el cambio de política

Capacidad de refrescamiento de ruta

- Facilita cambios de política no abrupta
- No es necesaria configuración
  - Automáticamente negociada al establecimiento del *peer*
- No usa memoria adicional
- Requiere enrutadores *peering* que soporten "route refresh capability" – RFC2918
- Digale al peer que reenvíe el anuncio completo de BGP

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

- Reenviar el anuncio completo BGP al *peer*

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

Reconfiguración dinámica

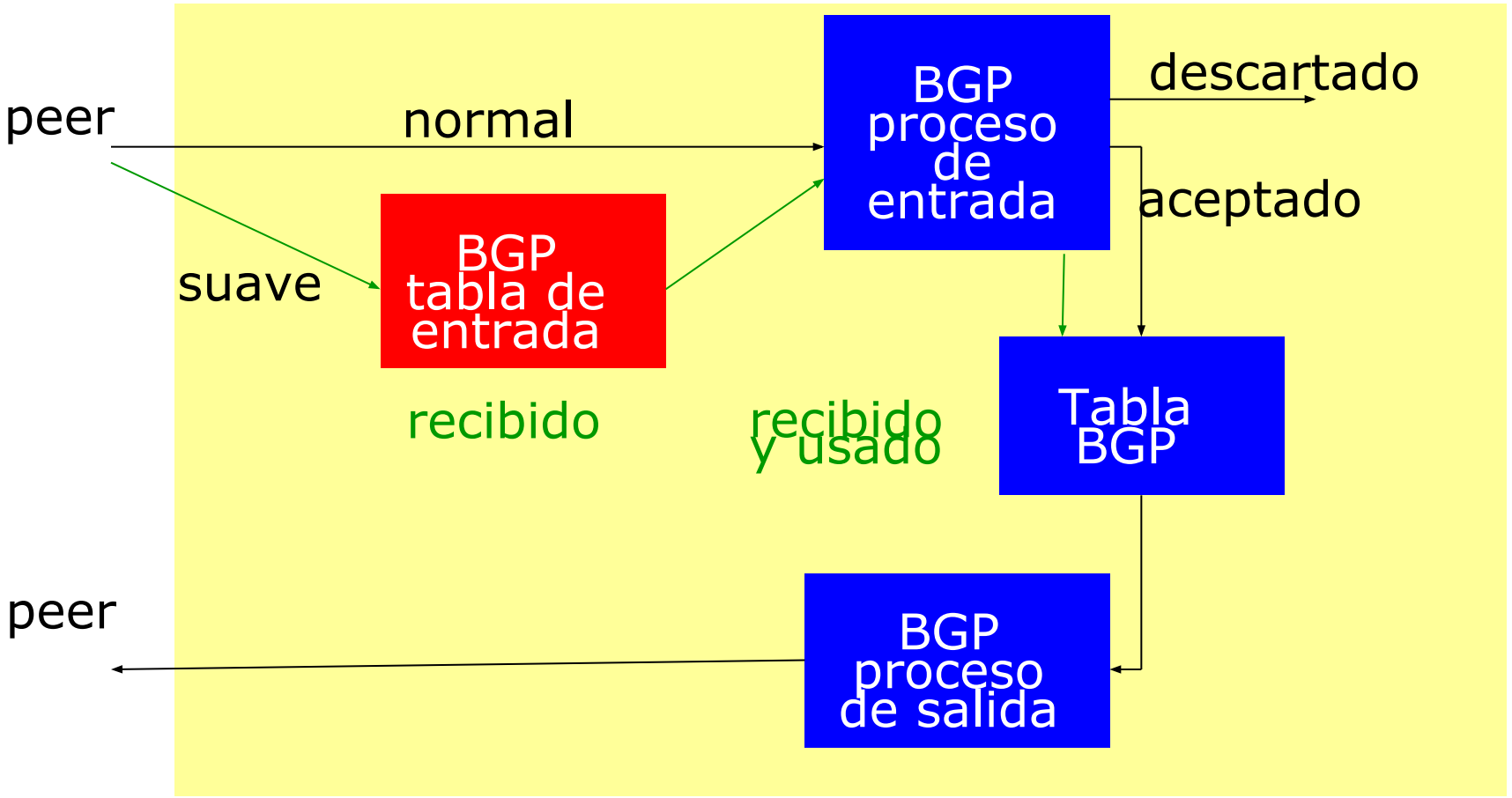
- Uso de la capacidad de refrescamiento de la ruta
  - Compatible con prácticamente todos los enrutadores
  - Averiguar a desde “show ip bgp neighbor”
  - No abrupto, “Bueno para el internet”
- Solamente como último recurso se hace el el reinicio duro (hard-reset) a un *peering* BGP

Considere el impacto que sea equivalente al reinicio del enrutador

Reconfiguración suave de Cisco

- Ahora en desuso — pero:
- Un enrutador normalmente almacena prefijos que han sido recibidos desde un *peer* después de la aplicación de la política
  - Habilitar reconfiguración suave significa que el enrutador también almacena prefijos/atributos recibidos anterior a cualquier aplicación de política
  - Usa más memoria para mantener prefijos cuyos atributos han sido cambiados o no han sido aceptados
  - Solamente útil en el momento en que el operator requiere saber cuales prefijos han sido enviados a un enrutador antes de la aplicación de cualquier política entrante

Reconfiguración suave de Cisco



Reconfiguración suave, configuración

```
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 does not need to be configured !
```

- Entonces cuando cambiamos la política, emitimos un comando exec

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

- Nota:
  - Cuando la "reconfiguración suave" está habilitada, no hay acceso a la capacidad de refrescamiento de la ruta
  - También lo hará una actualización suave

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

Grupos *peer* de Cisco

Grupos *peer* de Cisco

- Problema – como escalar iBGP
  - malla grande de iBGP es lenta en construir
  - Vecinos iBGP reciben la misma actualización
  - El CPU del enrutador se desperdicia en repetir los cálculos
- Solución – peer-groups
  - Grupos *peer* con la misma política de salida
  - Las actualizaciones son generadas una vez por grupo

Grupos peer - Ventajas

- Hace más fácil la configuración
- Hace la configuración menos propensa a errores
- hace la configuración más legible
- Menor carga de CPU del enrutador
- Malla iBGP se construye más rápido
- miembros pueden tener diferentes políticas de entrada
- ¡Puede ser usada para los vecinos de eBGP también!

Configurando un grupo *peer*

```
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
```

! note como 2.2.2.2 tiene un filtro de entrada diferente desde peer-group !

Configurando un grupo *peer*

```
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
```

Grupos *peer*

- Peer-groups son considerados obsoletos por Cisco:
  - Reemplazados por update-groups (código interno - no configurable)
- Pero son considerados todavía como buena práctica por muchos operadores de red
- Cisco introdujo peer-templates
  - Una versión mejorada de peer groups, lo que permite construcciones más complejas

Cisco's update-groups (1)

- Update-groups es un código interno del IOS, que se hizo cargo de las mejoras del rendimiento introduciendo peer-groups

```
Router1#sh ip bgp 10.0.0.0/26
BGP routing table entry for 10.0.0.0/26, version 2
Paths: (1 available, best #1, table default)
  Advertised to update-groups:
     1
  Refresh Epoch 1
  Local
    0.0.0.0 from 0.0.0.0 (10.0.15.241)
      Origin IGP, metric 0, localpref 100, weight 32768, valid...
```

- El comando "show" indica que el prefijo es manejado por update-group #1

Cisco's update-groups (2)

- La actualización del grupo en sí misma lista todos los peers los cuales tienen la misma actualización (idéntica):

```
Router1#sh ip bgp update-group 1
BGP version 4 update-group 1, internal, Address Family: IPv4 Unicast
  BGP Update version : 16/0, messages 0
  Topology: global, highest version: 16, tail marker: 16
  Format state: Current working (OK, last not in list)
                Refresh blocked (not in list, last not in list)
  Update messages formatted 11, replicated 13, current 0, refresh 0, limit 1000
  Number of NLRIs in the update sent: max 2, min 0
  Minimum time between advertisement runs is 0 seconds
  Has 13 members:
   10.0.15.242      10.0.15.243      10.0.15.244      10.0.15.245
   10.0.15.246      10.0.15.247      10.0.15.248      10.0.15.249
   10.0.15.250      10.0.15.251      10.0.15.252      10.0.15.253
   10.0.15.254
```

- Y este grupo tiene 13 miembros

Grupos peer

- Siempre configure peer-groups para iBGP
  - Incluso si hay solamente pocos pares iBGP
  - Más fácil de escalar la red en el futuro
  - Hace la configuración más legible
- Considere el uso de peer-groups para eBGP
  - Especialmente útil para múltiples clientes BGP usando el mismo AS (RFC2270)
  - También útil para Puntos de Intercambio:
    - Donde la política ISP es generalmente el misma en cada peer
    - Para el servidor de ruta donde todos los peer reciben las mismas actualizaciones de enrutamiento where all peers receive the same routing updates

Reflectores de ruta

Escalando la malla iBGP

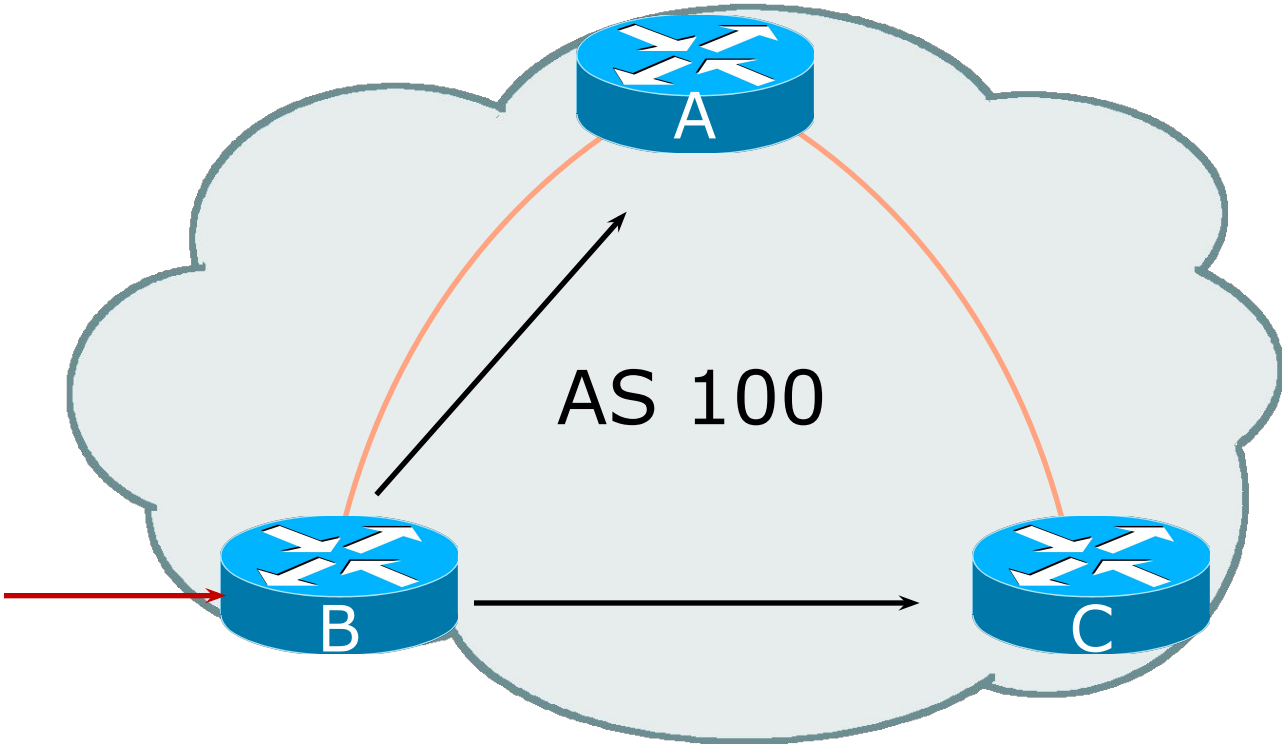
Escalando la malla iBGP

- Evita ½n(n-1) malla iBGP

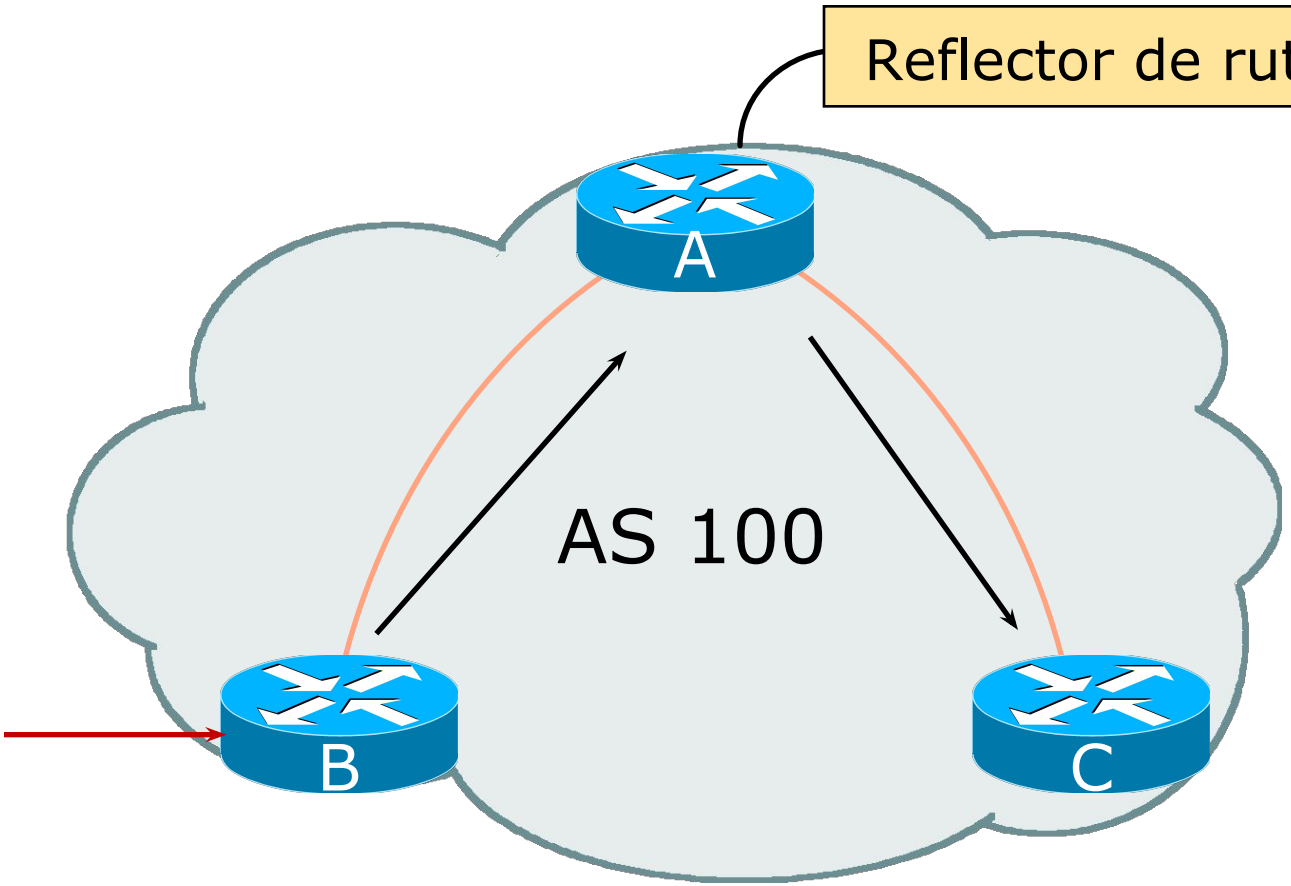
n=1000 ⇒ cerca de medio millón de sesiones ibgp

- Dos soluciones
  - Reflector de ruta – simple de implementar y ejecutar
  - Confederación - más complejo, tiene ventajas de casos de frontera

Reflector de ruta: principio

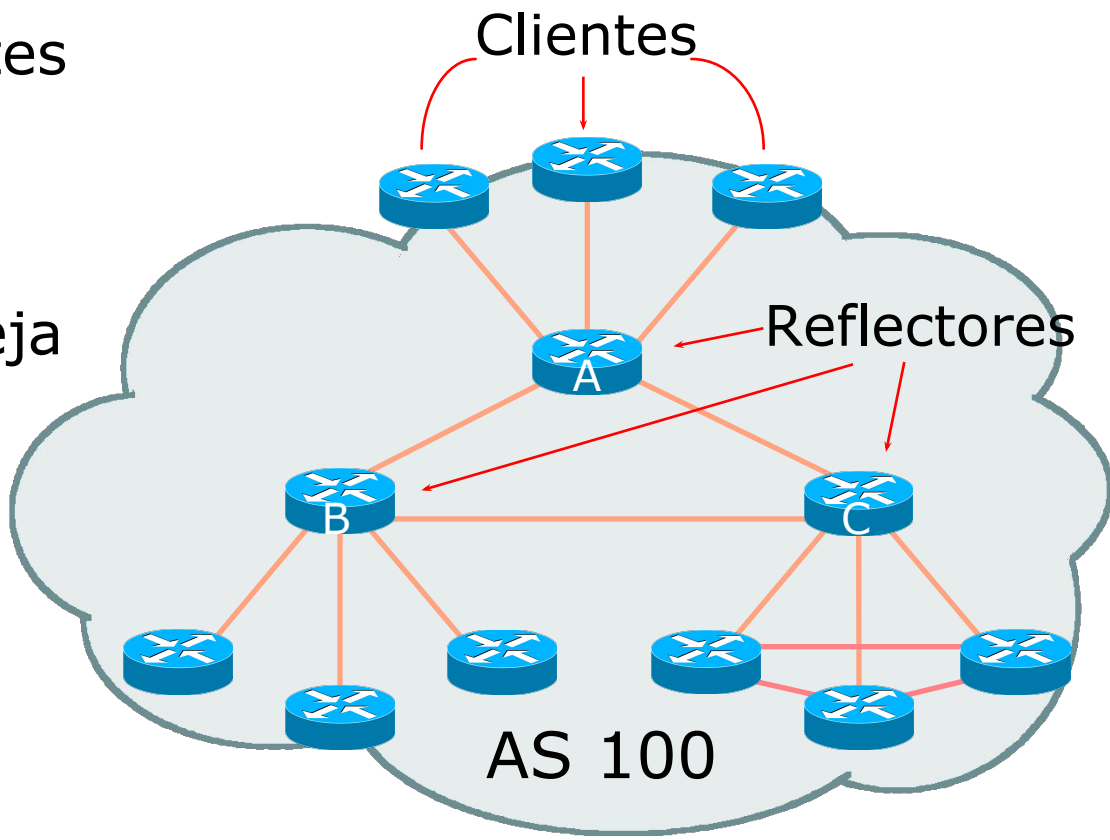


Reflector de ruta: principio



Enrutador reflector

- Reflector recibe las rutas desde los clientes y los no-clientes
- Elije el mejor camino
- Si la mejor ruta es desde el cliente, refleja a otro cliente y no cliente
- Si la mejor ruta es desde un no-cliente, refleja a clientes solamente
- Clientes no mallados
- Descrito en RFC4456



Topología de enrutador reflector

- Divide el troncal entre mútiples racimos
- Al menos un reflector de ruta y unos clientes por racimo
- Reflectores de ruta son mallados completamente
- Clientes en un racimo (cluster) podrían ser completamente mallados
- IGP individual para llevar al siguiente salto y rutas locales

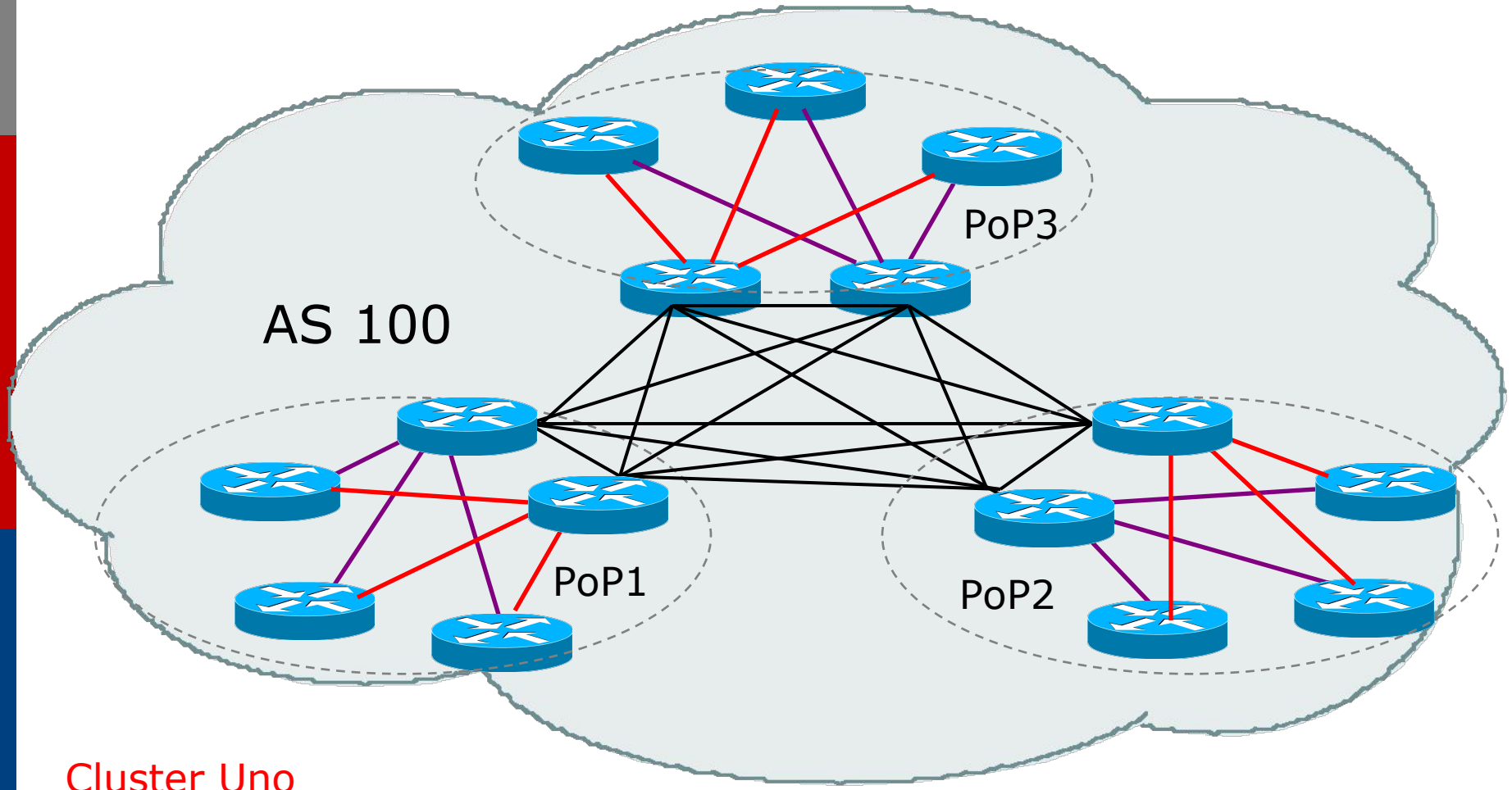
Reflectores de ruta: evitando el bucle

- Atributo originator_ID
  - Lleva el RID del creador de la ruta en el AS local (creado por el RR)
- Atributo cluster_list
  - El cluster_id local es agregado cuando la actualización es enviada por el RR
  - Cluster-id es router-id por defecto (usualmente la dirección de la interfaz loopback)
  - NO use **`bgp cluster-id x.x.x.x`** a menos que dos reflectores de ruta estén fisicamente/directamente

Reflectores de ruta: redundancia

- Múltiples RRs pueden ser configurados en el mismo cluster - no aconsejado
  - Todos los RRs en un cluster deben tener el mismo cluster_id (de otra manera es un cluster diferente)
- Un enrutador puede ser un cliente de RRs en diferentes clusters
  - Común hoy en día en redes ISP que superponen dos clusters - se logra redundancia de esa manera
  - → Cada cliente tiene dos RRs = redundancia

Reflectores de ruta: redundancia



Cluster Uno

Cluster Dos

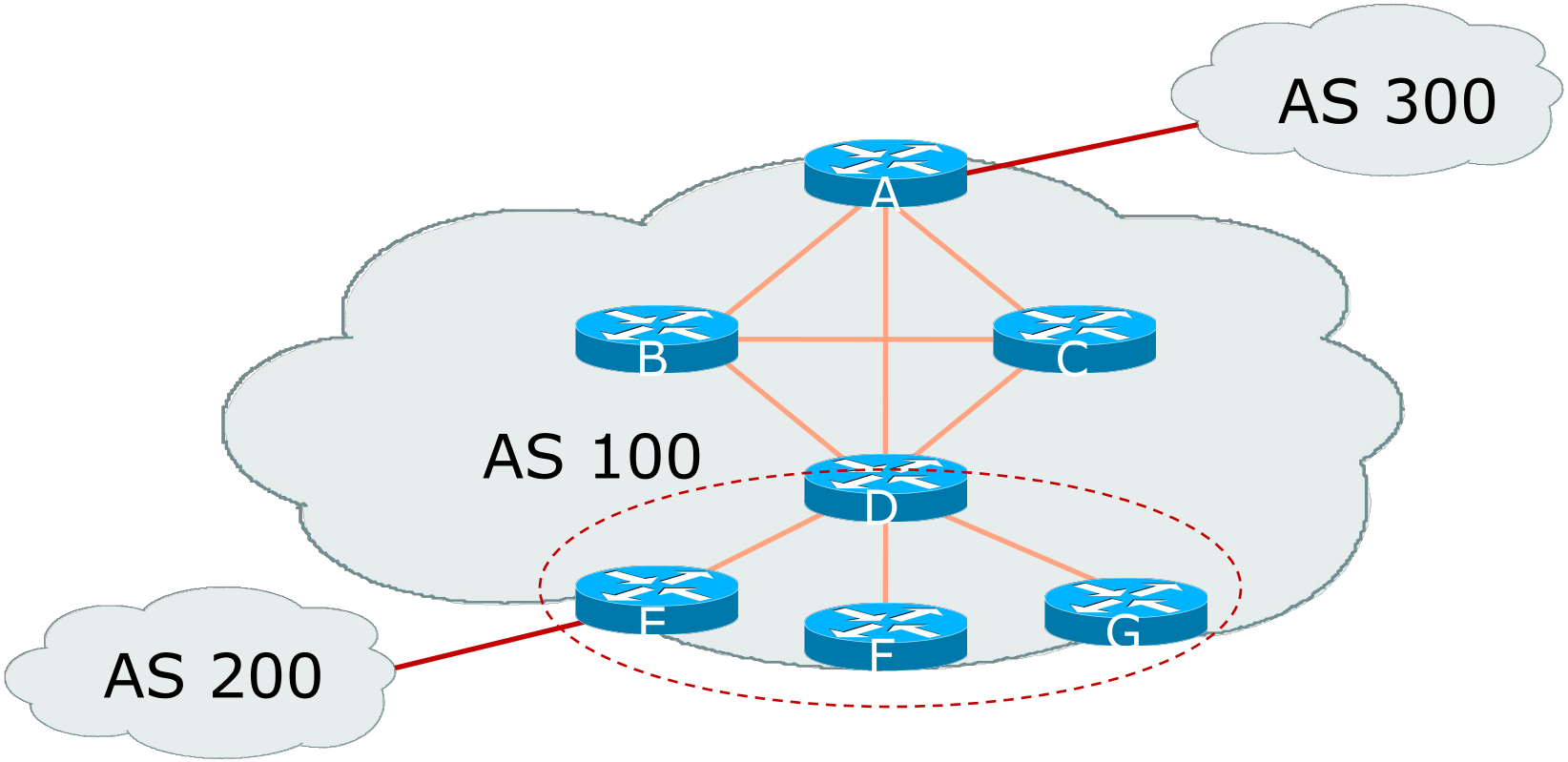
Reflectores de ruta: beneficios

- Resuelve el problema de malla iBGP
- El reenvío de paquetes no es afectado
- Hablantes BGP normales pueden coexistir
- Reflectores múltiples por redundancia
- Fácil migración
- Niveles múltiples de reflectores de ruta

Reflectores de ruta: migración

- ¿Dónde poner los reflectores de ruta?
  - ¡Siga la topología física!
  - Esto garantizará que el reenvío de paquetes no se verá afectado
- Configure un RR a la vez
  - Elimine sesiones iBGP redundantes
  - Coloque un RR por cluster

Reflectores de ruta: migración



- Migre pequeñas partes de la red, una parte a la vez

Configurando un reflector de ruta

- Configuración Router 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
 ...
```

Técnicas de escalamiento BGP

- Estas 3 técnicas deberían ser requerimientos esenciales en todas las redes ISP
  - Refrescamiento de ruta (o reconfiguración suave)
  - Grupos Peer
  - Reflectores de Ruta

Confederaciones BGP

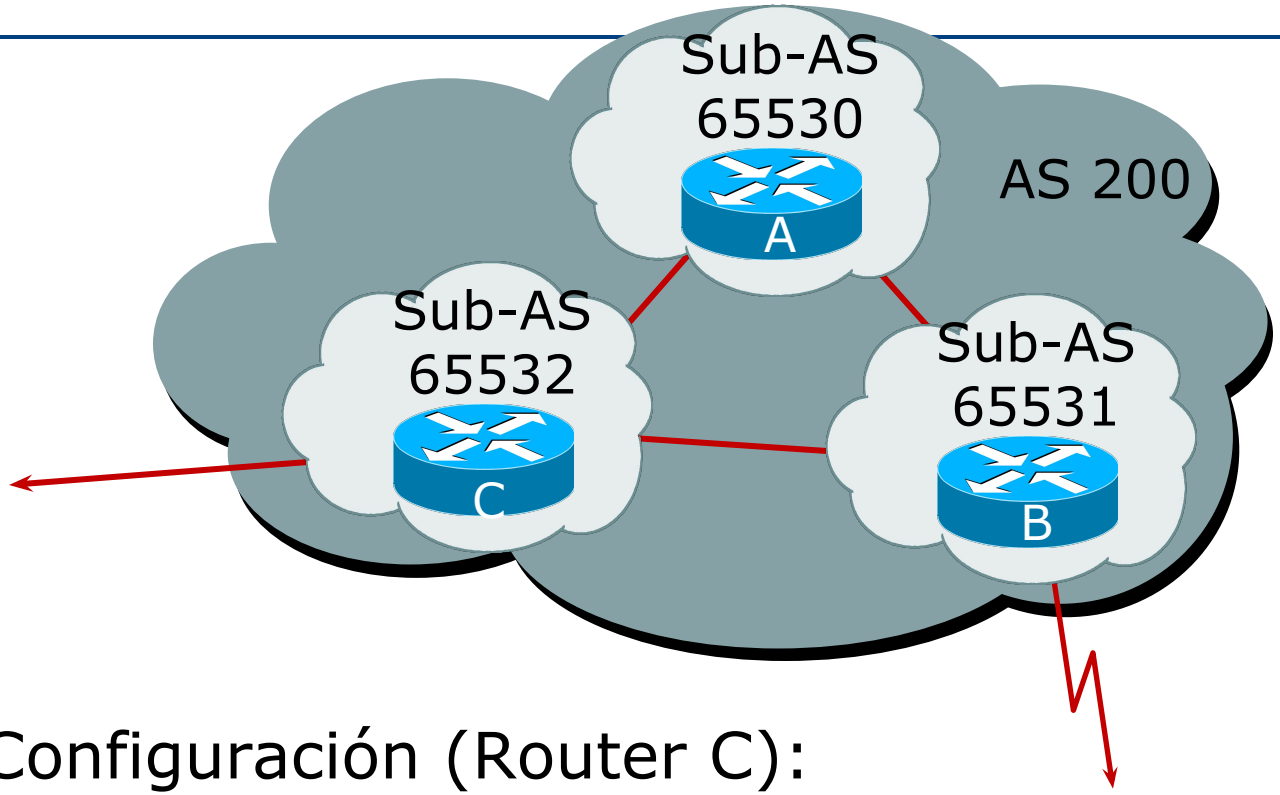
Confederaciones

- Divide el AS en un sub-AS
  - eBGP entre sub-AS, pero alguna información de iBGP es conservada
  - Se preserva el NEXT_HOP a través del sub-AS (IGP lleva esta información)
    - Se preserva el LOCAL_PREF y MED
- Usualmente un individuo IGP
- Descrito en RFC5065

Confederaciones

- Visible hacia el mundo exterior como un único AS -- "Identificador de confederación"
  - Cada sub-AS usa el número desde su espacio privado (64512-65534)
- Hablantes iBGP en el sub-AS son completamente mallados
  - El número total de vecinos es reducido limitando el requerimiento de malla completa solo a los peers en el sub-AS
  - También se puede utilizar Reflector de Ruta dentro de un sub-AS

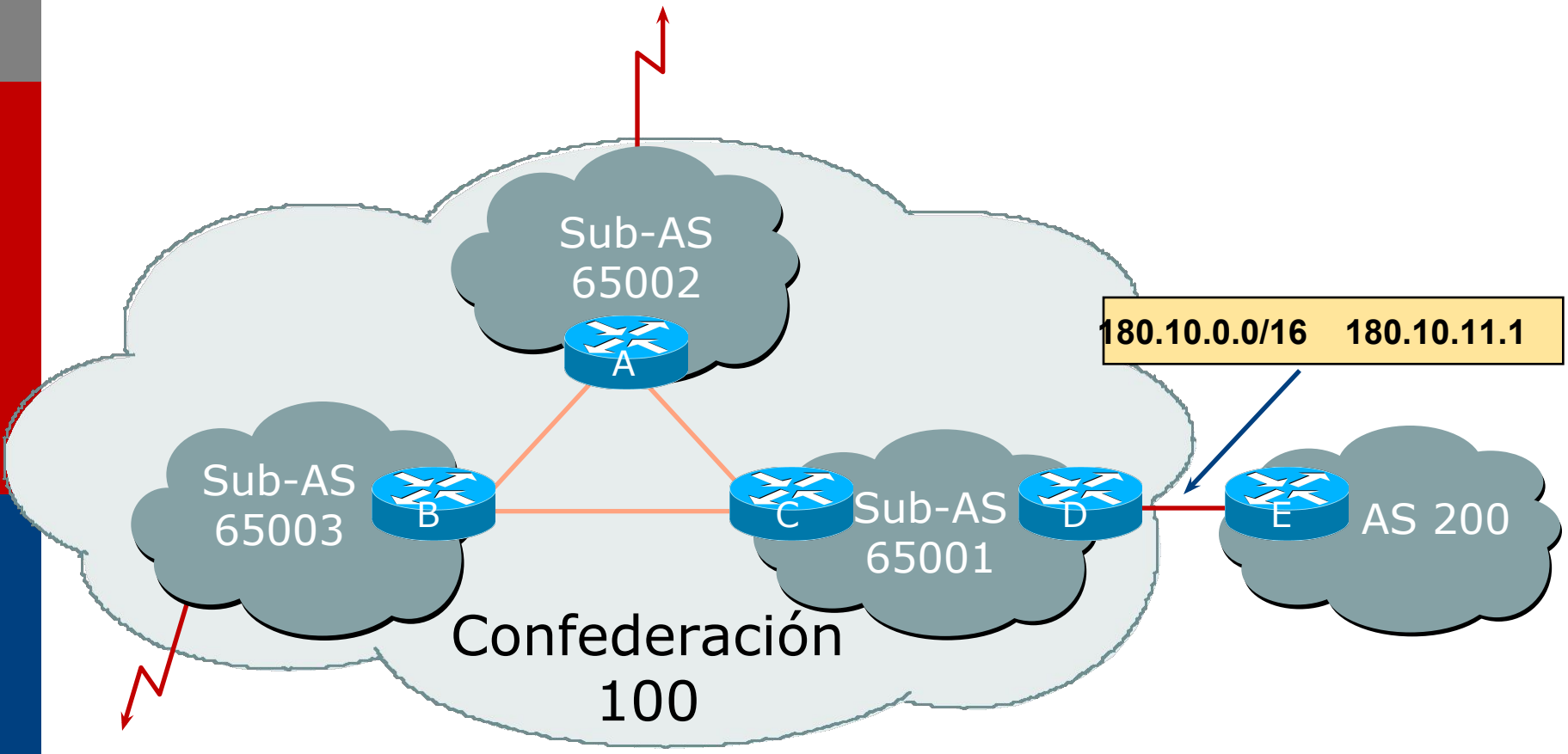
Confederaciones



- Configuración (Router 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
```

Confederaciones: Next Hop



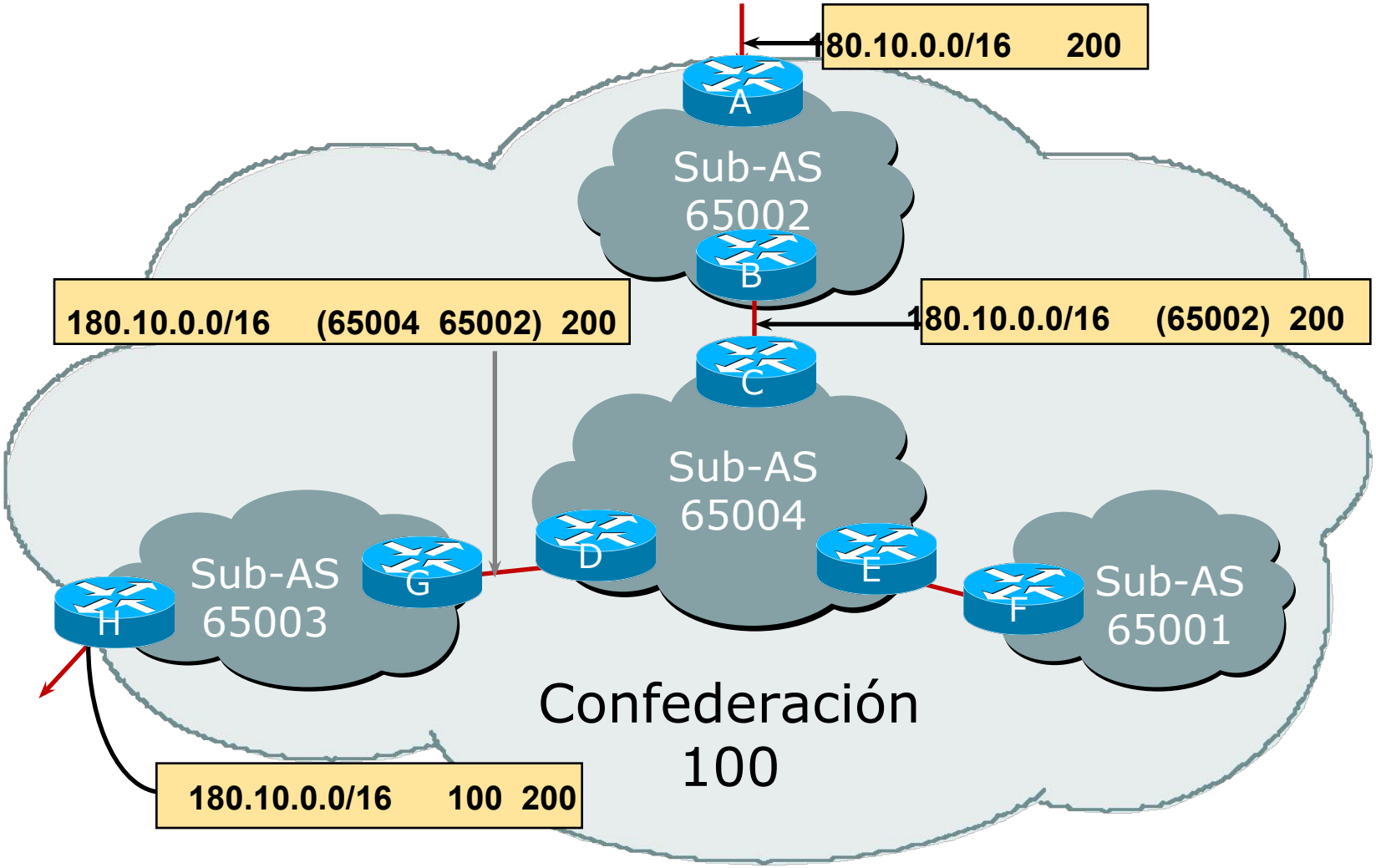
Confederación: Principio

- Preferencia local y MED tienen influencia en la selección del *path*
- Preserva preferencia local y MED a través de sub-AS de frontera
- Distancia administrativa Sub-AS eBGP path

Confederaciones: Evitar Loop

- El recorrido de Sub-AS se atraviesa como parte del
- Secuencia del AS y longitud de la trayectoria del AS
- Confederación de frontera
- Secuencia del AS debería ser saltada durante la comparación de MED

Confederaciones: AS-Sequence



Decisiones de propagación de ruta

- Igual que con BGP "normal":
  - Desde el peer en el mismo sub-AS → solamente para peers externos
  - Desde peers externos → hacia todos los vecinos
- "peers externos" referirse a
  - Peers afuera de la confederación
  - Peers in un diferente sub-AS
    - Preserva LOCAL_PREF, MED y NEXT_HOP

Confederaciones (cont.)

- Ejemlpo (cont.):

```
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
```

Más puntos acerca de las confederaciones

- Puede fácilmente "absorber" otros ISPs dentro de su ISP
  - ej., si un ISP compra otro ISP
  - (puede utilizar la característica local-as para hacer una cosa similar)
- Puede utilizar reflectores de ruta con confederaciones sub-AS para reducir la malla del sub-AS iBGP

Confederaciones: beneficios

- Resuelve el problema de malla iBGP
- El reenvío de paquetes no es afectado
- Puede ser utilizado con otros reflectores de ruta
- Las políticas podrían ser aplicadas para enrutar tráfico entre sub-ASs

Confederaciones: advertencias

- Número mínimo de sub-AS
- Sub-AS jerárquico
- Mínima interconectividad entre sub-ASs
- Diversidad de trayectoria
- Migración difícil
  - BGP reconfigurado en el sub-AS
  - Debe ser aplicado a través de la red

RRs o Confederaciones

	Conectividad a Internet	Jerarquía multi nivel	Control de política	Escalabilidad	Complejidad de la migración
Confederaciones	Cualquier lugar en la red	Sí	Sí	Media	Medio a Alta
Reflectores de Ruta	Cualquier lugar en la red	Sí	Sí	Muy alta	Muy baja

Muchos de los proveedores de servicios de redes ahora implementan Reflectores de Ruta en un día

Amortiguación de intermitencia de la ruta

Estabilidad de la red para los 1990s

¡Inestabilidad de la red para el siglo 21!

Amortiguación de la intermitencia de la ruta

- Por muchos años la amortiguación de la intermitencia (flap damping) de la ruta fue una práctica fuertemente recomendada
- Ahora se desalienta fuertemente ya que causa mucha mayor inestabilidad en la red de lo que cura
- Pero primero, la teoría…

Amortiguación de la intermitencia de la ruta

- Intermitencia de ruta
  - Subiendo y bajando la trayectoria o cambio en el atributo
    - BGP s retira seguido de UPDATE = 1 intermitencia
    - Vecino eBGP va arriba/abajo NO es una intermitencia
  - Ondulaciones a través de toda la Internet
  - Desperdicio de CPU
- El propósito de la amortiguación es reducir el alcance de la propagación de la intermitencia de ruta

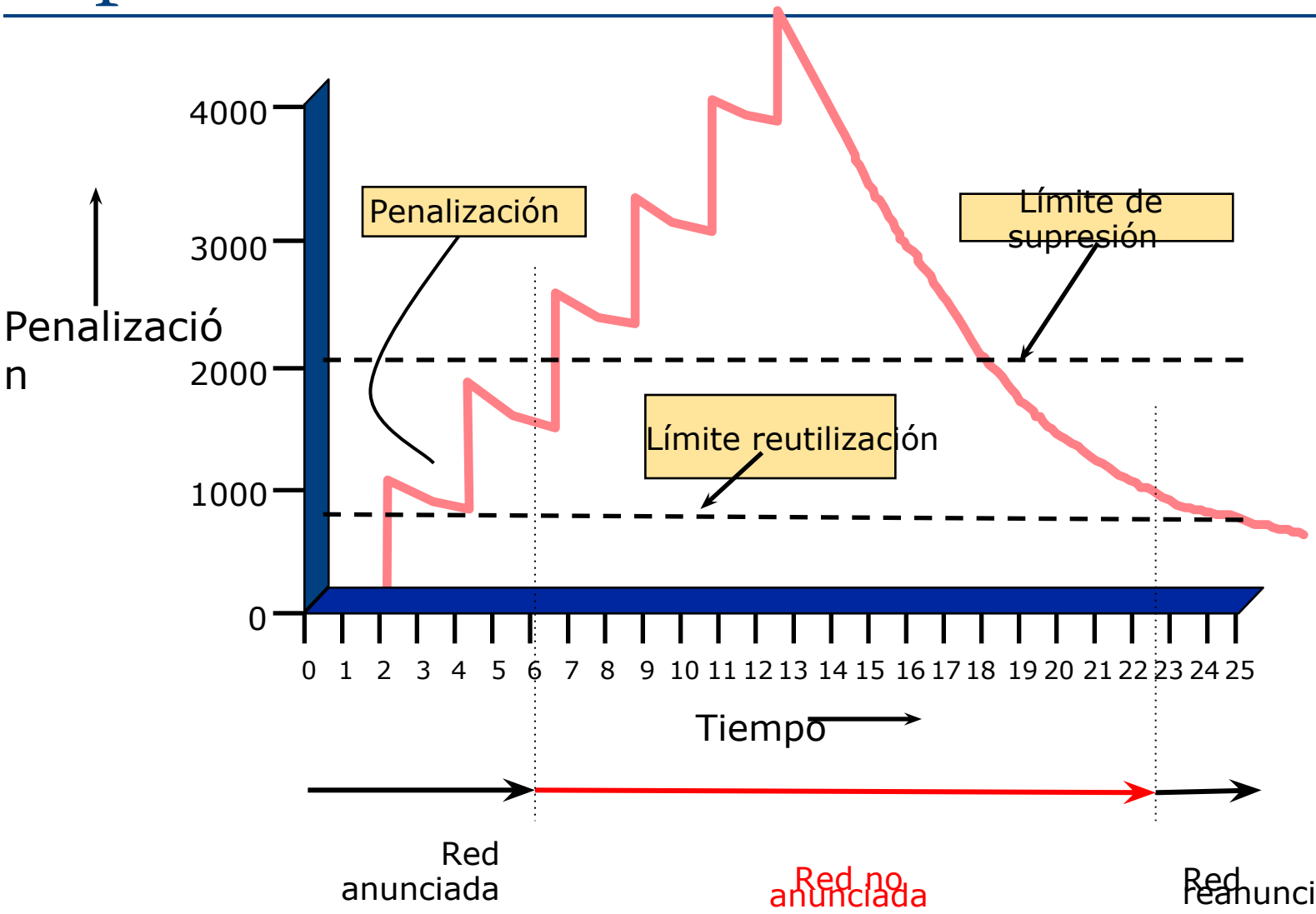
Amortiguación de la intermitencia de la ruta (cont.)

- Requisitos
  - Rápida convergencia para cambios de ruta normales
  - La historia predice el comportamiento futuro
  - Suprime rutas oscilatorias
  - Anuncia rutas estables
- Implementación descrita enRFC 2439

Operación

- Agregar penalización (1000) por cada intermitencia
  - Cambiar un atributo obtiene una penalización de 500
- Penalización de decaimiento exponencial
  - Vida media determina la tasa de decaimiento
- Penalización por encima del límite de supresión
  - No anuncia ruta a los peers BGP
- Penalización decayó por debajo de la límite de reutilización
  - Republicación de ruta a los peers BGP

Operación


Penalización
4000
3000
2000
1000
0
Penalización
Límite de supresión
Límite reutilización
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
Tiempo
Red anunciada
Red no anunciada
Red reanunciada

Operación

- Solo se aplica a anuncios entrantes de peers eBGP
- Trayectorias alternativas todavía son utilizables
- Controlado por:
  - Vida media (por defecto 15 minutos)
  - Límite de reutilización (por defecto 750)
  - Límite de supresión (por defecto 2000)
  - Tiempo máximo de supresión (por defecto 60 minutos)

Configuración

- Amortiguación fija

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

- Amortiguación selectiva y 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
```

Operación

- Se requiere precaución al establecer los parámetros
- Penalización debe ser inferior a la reutilización límite en el tiempo máximo de supresión
- Tiempo de supresión máxima y media vida deben permitir que la penalización sea mayor que el límite de supresión

Configuración

- Ejemplos – ✗
  - bgp dampening 15 500 2500 30
    - Límite de reutilización de 500 significa que la penalización máxima es 2000 - no hay prefijos suprimidos como penalización ya que no pudo exceder el límite de supresión
- Ejemplos – ✓
  - bgp dampening 15 750 3000 45
    - Límite de reutilización de 750 significa que la penalización máxima posible es 6000 - el límite de supresión es fácilmente alcanzado

Matemáticas

- El máximo valor de penalización es:

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

- Siempre asegurese que el límite de supresión es MENOR que la penalización máxima, de otra manera no habría amortiguación de la ruta

Historia de la amortiguación de intermitencia de la ruta

- Primera implementaciones sobre internet durante 1995
- Incumplimientos del proveedor demasiado severos
  - Recomendaciones de RIPE Routing Working Group en ripe-178, ripe-210, y ripe-229
  - http://www.ripe.net/ripe/docs
  - Pero muchos ISPs simplemente cambiaron por los valores por defect del proveedor sin pensar

Problemas serios:

- "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
- Varios trabajos sobre convergencia de enrutamiento por Craig Labovitz y Abha Ahuja hace unos años
- "Happy Packets"
  - En estrecha relación de trabajo por Randy Bush

Problema 1:

- Una trayectoria intermitente:
  - Hablantes BGP toman la siguiente mejor trayectoria, anuncia a todos los peers, contador de intermitencia incrementado
  - Esos peers ven el cambio en la mejor trayectoria, contador de intermitencia incrementado
  - Después de algunos saltos, los peers ven múltiples cambios simplemente causados por una sola intermitencia → el prefijo se suprime

Problema 2:

- Diferentes implementaciones BGP tienen diferentes tiempos de tránsito para los prefijos
  - Algunos mantienen el prefijo por algún tiempo antes de anunciar
  - Otros anuncian inmediatamente
- Correr hasta el final de la línea causa apariencia de intermitencia, causado por un simple anuncio o cambio de trayectoria → el prefijo es suprimido

Solución:

- Desconfiguración del amortiguamiento de intermitencia de la ruta afectará seriamente en el acceso a:
  - Su red y
  - El Internet
- Más antecedentes contenidos en documento de la RIPE Routing Working Group:
  - www.ripe.net/ripe/docs/ripe-378
- Recomendaciones ahora en:
  - www.rfc-editor.org/rfc/rfc7196.txt y www.ripe.net/ripe/docs/ripe-580

BGP Técnicas de escalamiento

Talleres ISP FIN