Xarxes de Computadors II

Tema 5: Encaminamiento inter-dominio: BGP

Davide Careglio

Temario

- ~~Tema 0. Repaso~~
- ~~Tema 1. Arquitectura y direccionamiento en Internet~~
- ~~Tema 2. Direccionamiento IPv6~~
- ~~Tema 3. Encaminamiento intra-dominio~~
- ~~Tema 5. Multiprotocol Label Switching~~
- Tema 5. Encaminamiento inter-dominio
- Tema 6. Conceptos avanzados

5. Encaminamiento inter-dominio

1. Introducción
2. Encaminamiento path-vector BGP
3. Funcionamiento de BGP
4. Establecimiento de una sesión BGP
5. Bases de datos BGP
6. Mensajes BGP
7. Atributos estándares
8. Algoritmo de selección de rutas
9. Escenarios comunes
10. Mejoras del BGP

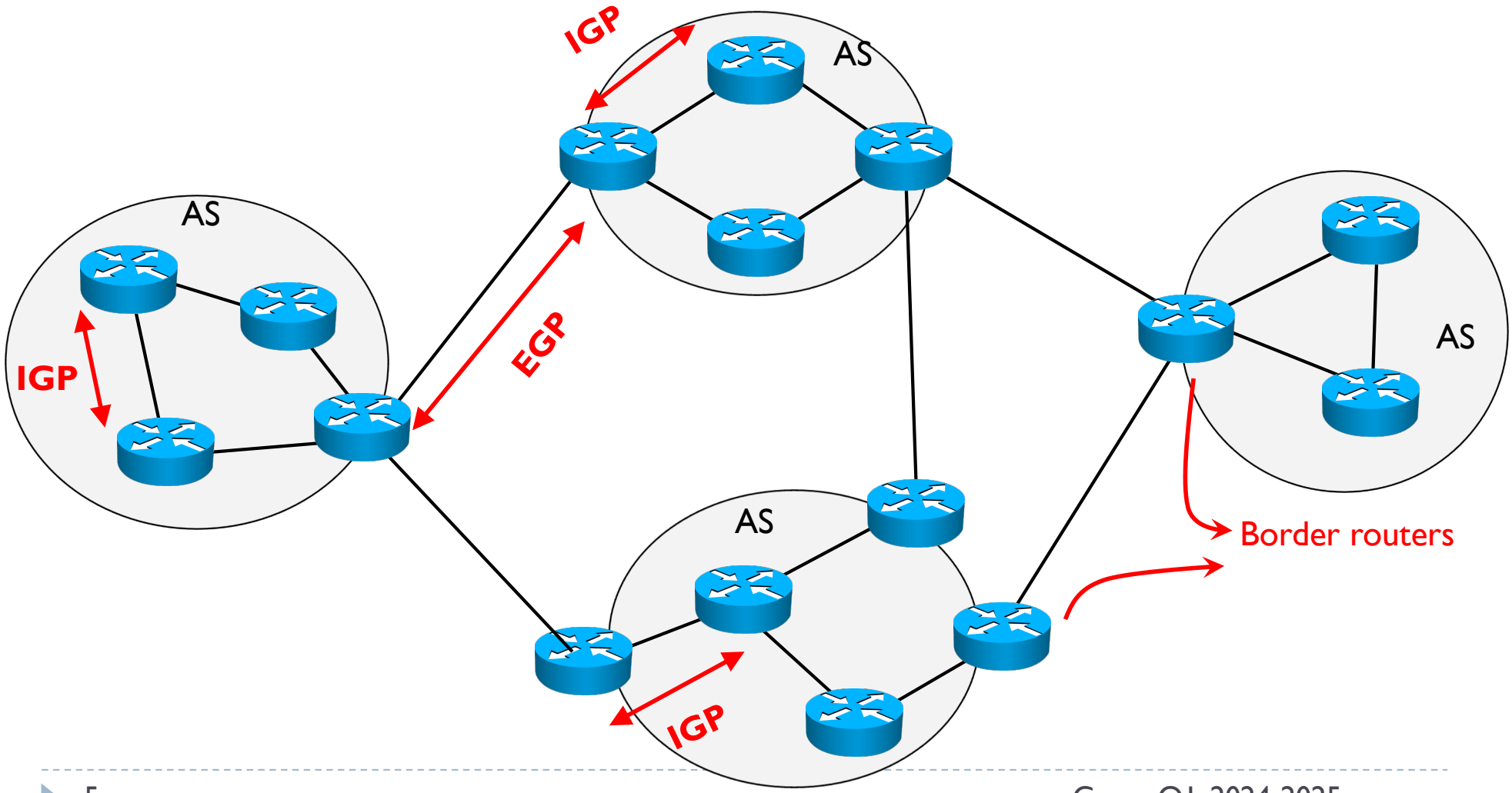
5. Encaminamiento inter-dominio

1. Introducción
2. Encaminamiento path-vector BGP
3. Funcionamiento de BGP
4. Establecimiento de una sesión BGP
5. Bases de datos BGP
6. Mensajes BGP
7. Atributos estándares
8. Algoritmo de selección de rutas
9. Escenarios comunes
10. Mejoras del BGP

5.1 – Introducción

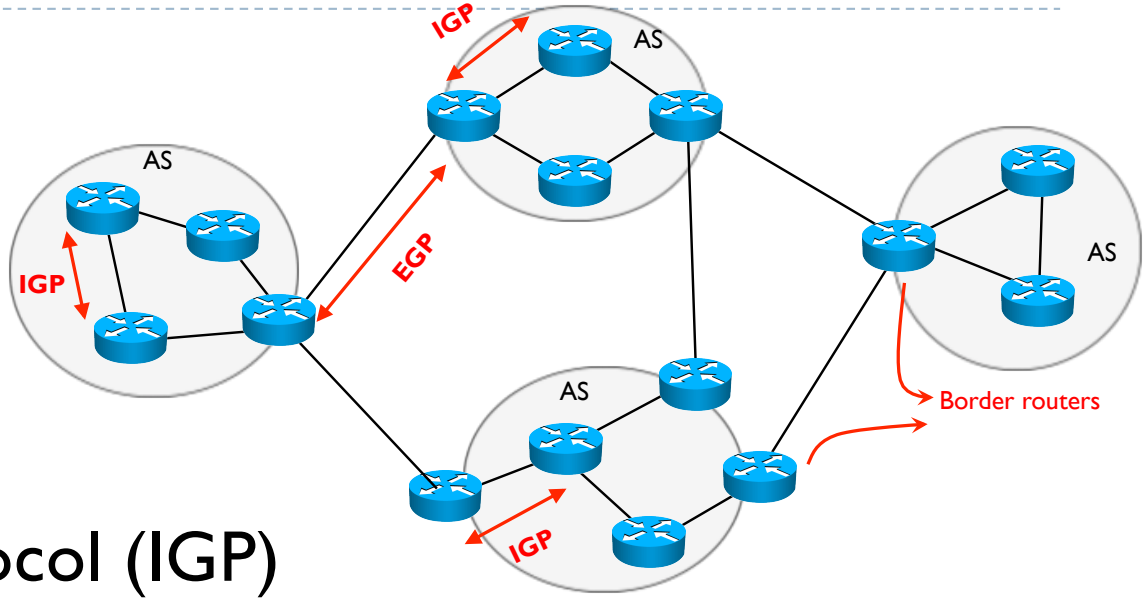
Intra-dominio vs. Inter-dominio

- Internet está formada por diferentes AS interconectados



5.1 – Introducción

Intra-dominio vs. Inter-dominio



- Interior Gateway Protocol (IGP)
  - RIP - RFC 2453 (versión 2) RFC 2080 para IPv6
  - OSPF - RFC 2328 (versión 2) RFC 5340 para IPv6
  - IS-IS - RFC 1142 RFC 5308 para IPv6
- Exterior Gateway Protocol (EGP)
  - EGP - RFC 904
  - BGP - RFC 1771 (versión 4) RFC 2545 para IPv6

5. Encaminamiento inter-dominio

1. Introducción
2. Encaminamiento path-vector BGP
3. Funcionamiento de BGP
4. Establecimiento de una sesión BGP
5. Bases de datos BGP
6. Mensajes BGP
7. Atributos estándares
8. Algoritmo de selección de rutas
9. Escenarios comunes
10. Mejoras del BGP

5.2 – Border Gateway Protocol (BGP)

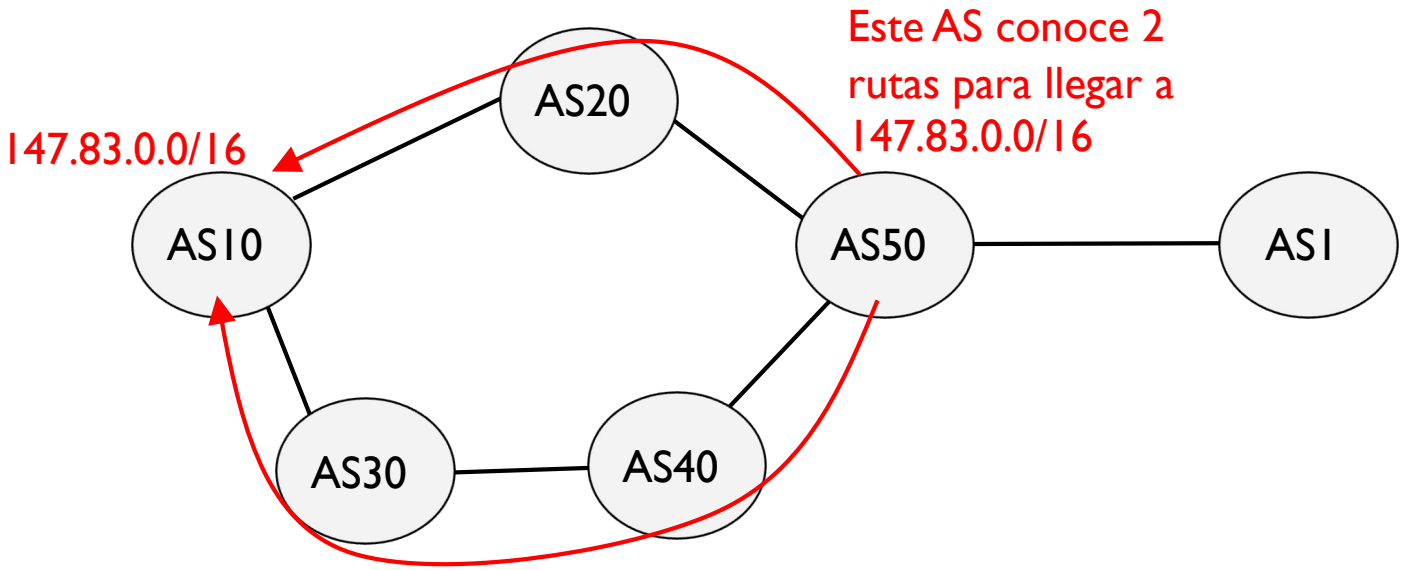
Encaminamiento path-vector

- RFC 1771 → (sustitución) 4271 → (actualización) 6286
- Protocolo de encaminamiento dinámico entre AS usado en Internet
- Basado en vector-distancia (realmente se dice que es path-vector)
  - El camino entre origen y destino se construye como composición de próximo salto (como RIP)
  - Eso implica que el conocimiento de un router depende de la decisión de otros
- Encaminamiento basado en políticas según unos atributos
  - Estos atributos se configuran en los routers y/o se distribuyen juntamente con los prefijos e influyen en la selección de la ruta
- Permite no enviar información considerado confidencial
  - p.e., topología del AS, número de routers, velocidad de transmisión
  - Solo se intercambian aquellos prefijos que se quiere que los AS vecinos sepan
  - En BGP se usa el termino genérico prefijos: ya que un prefijo puede no ser una red real

5.2 – Border Gateway Protocol (BGP)

Encaminamiento path-vector

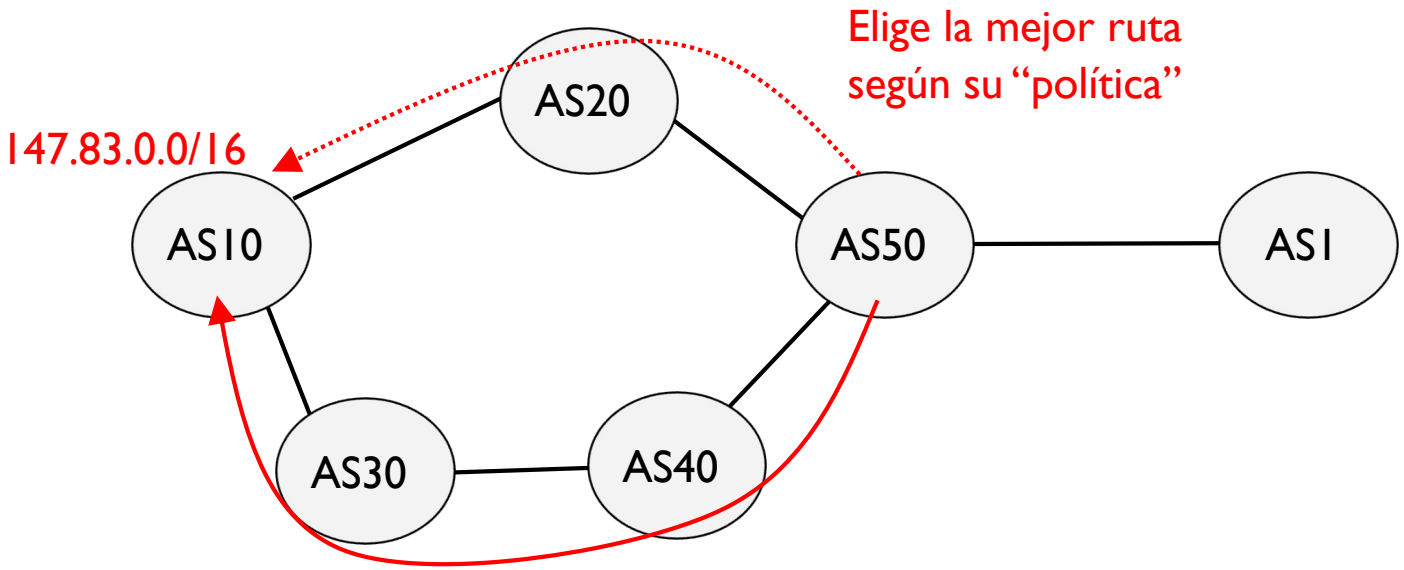
- Basado en vector-distancia (realmente se dice que es path-vector)
  - El camino entre origen y destino se construye como composición de próximo salto (como RIP)
  - Eso implica que el conocimiento de un router depende de la decisión de otros



5.2 – Border Gateway Protocol (BGP)

Encaminamiento path-vector

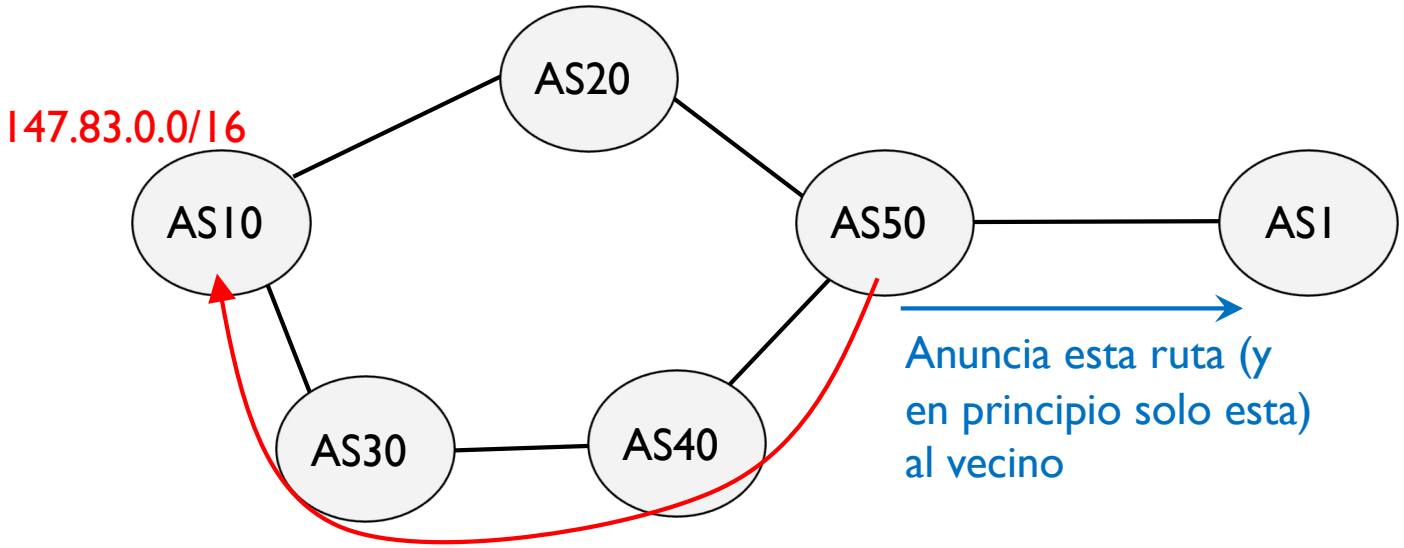
- Basado en vector-distancia (realmente se dice que es path-vector)
  - El camino entre origen y destino se construye como composición de próximo salto (como RIP)
  - Eso implica que el conocimiento de un router depende de la decisión de otros



5.2 – Border Gateway Protocol (BGP)

Encaminamiento path-vector

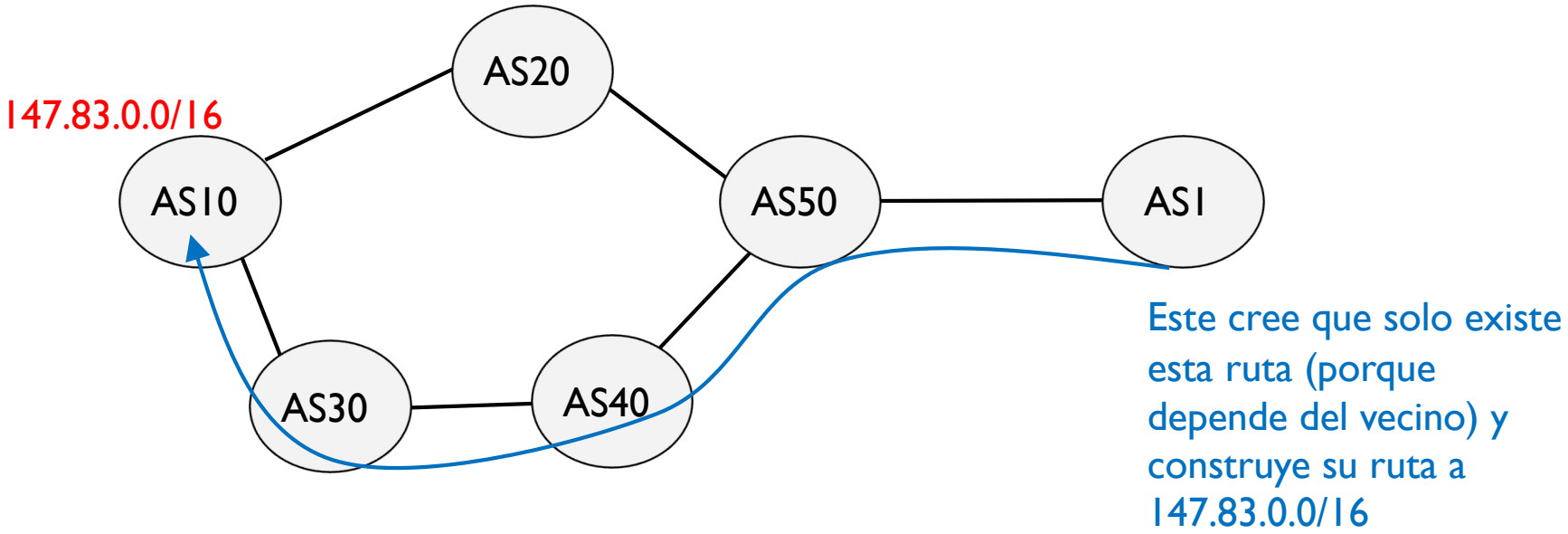
- Basado en vector-distancia (realmente se dice que es path-vector)
  - El camino entre origen y destino se construye como composición de próximo salto (como RIP)
  - Eso implica que el conocimiento de un router depende de la decisión de otros



5.2 – Border Gateway Protocol (BGP)

Encaminamiento path-vector

- Basado en vector-distancia (realmente se dice que es path-vector)
  - El camino entre origen y destino se construye como composición de próximo salto (como RIP)
  - Eso implica que el conocimiento de un router depende de la decisión de otros

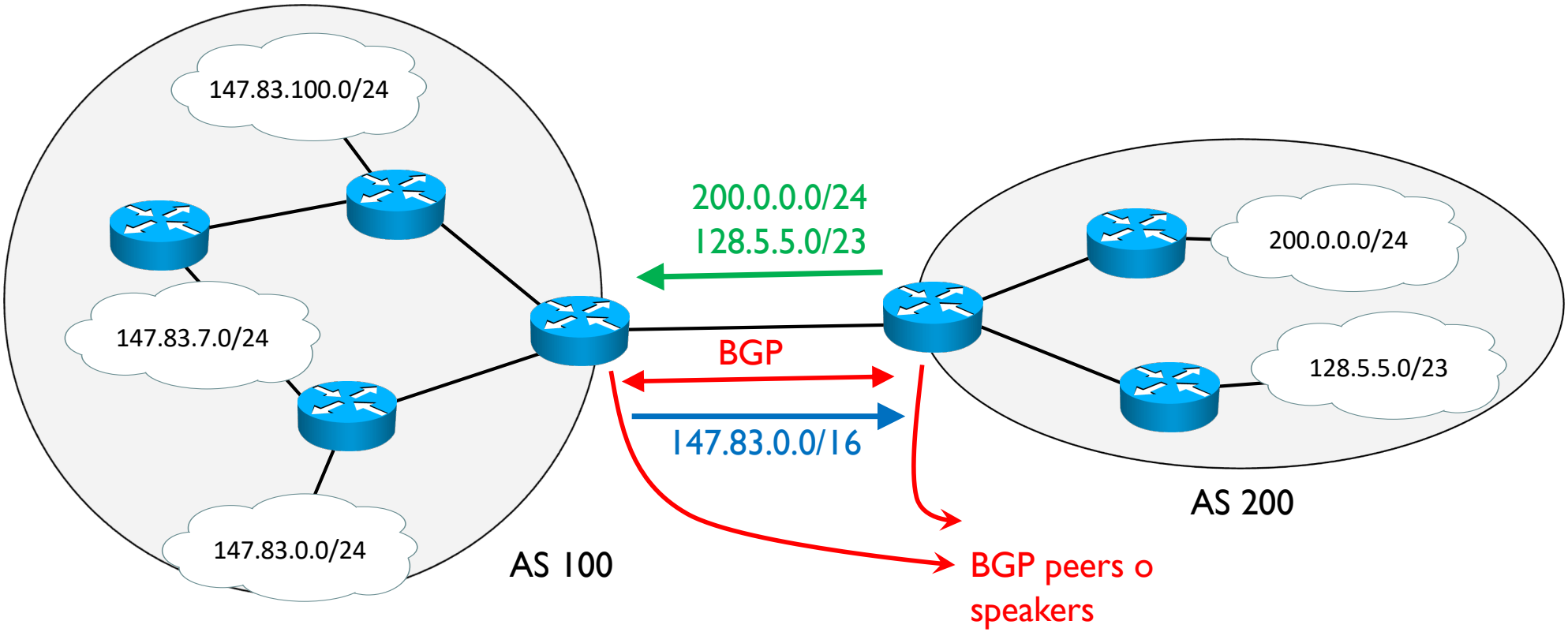


5. Encaminamiento inter-dominio

1. Introducción
2. Encaminamiento path-vector
3. Funcionamiento de BGP
4. Establecimiento de una sesión BGP
5. Bases de datos BGP
6. Mensajes BGP
7. Atributos estándares
8. Algoritmo de selección de rutas
9. Escenarios comunes
10. Mejoras del BGP

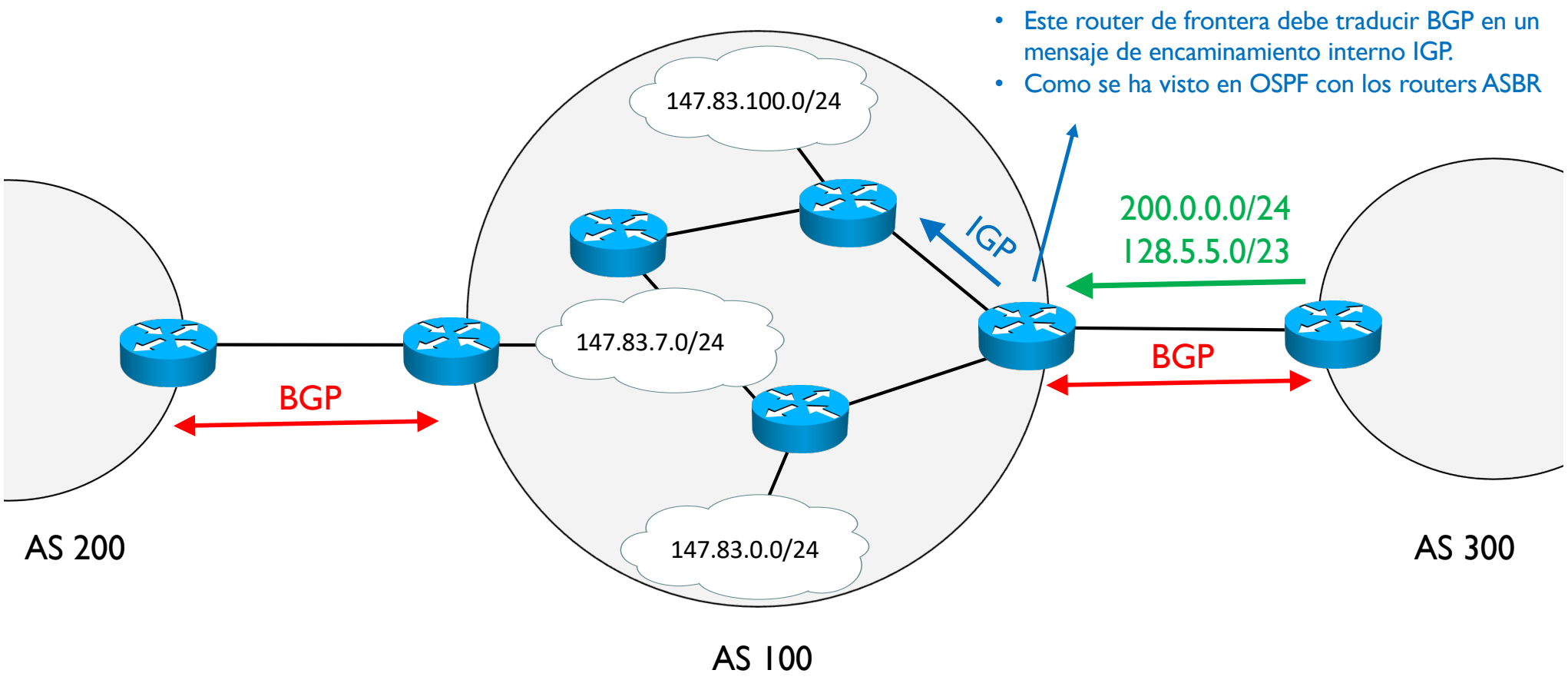
5.3 – Funcionamiento del BGP

- No se envía aquella información que se considera confidencial
  - En el ejemplo, el AS100 envía el prefijo que tiene asignado, sin especificar como se usa realmente en el AS
  - No se envían parámetros o métricas internas (i.e., aquella información relacionada con prestaciones, numero de routers, redes reales, tecnologías y protocolos empleados, etc.)



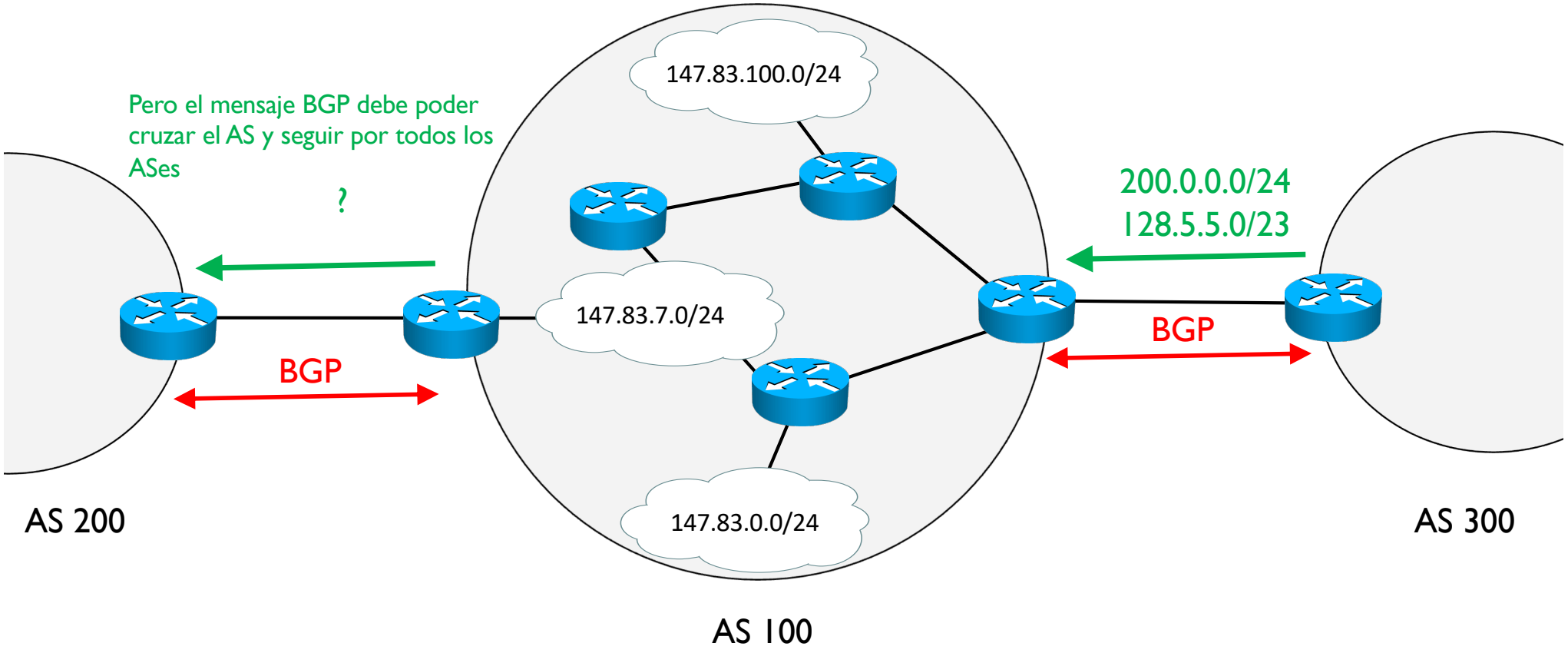
5.3 – Funcionamiento del BGP

external BGP vs internal BGP



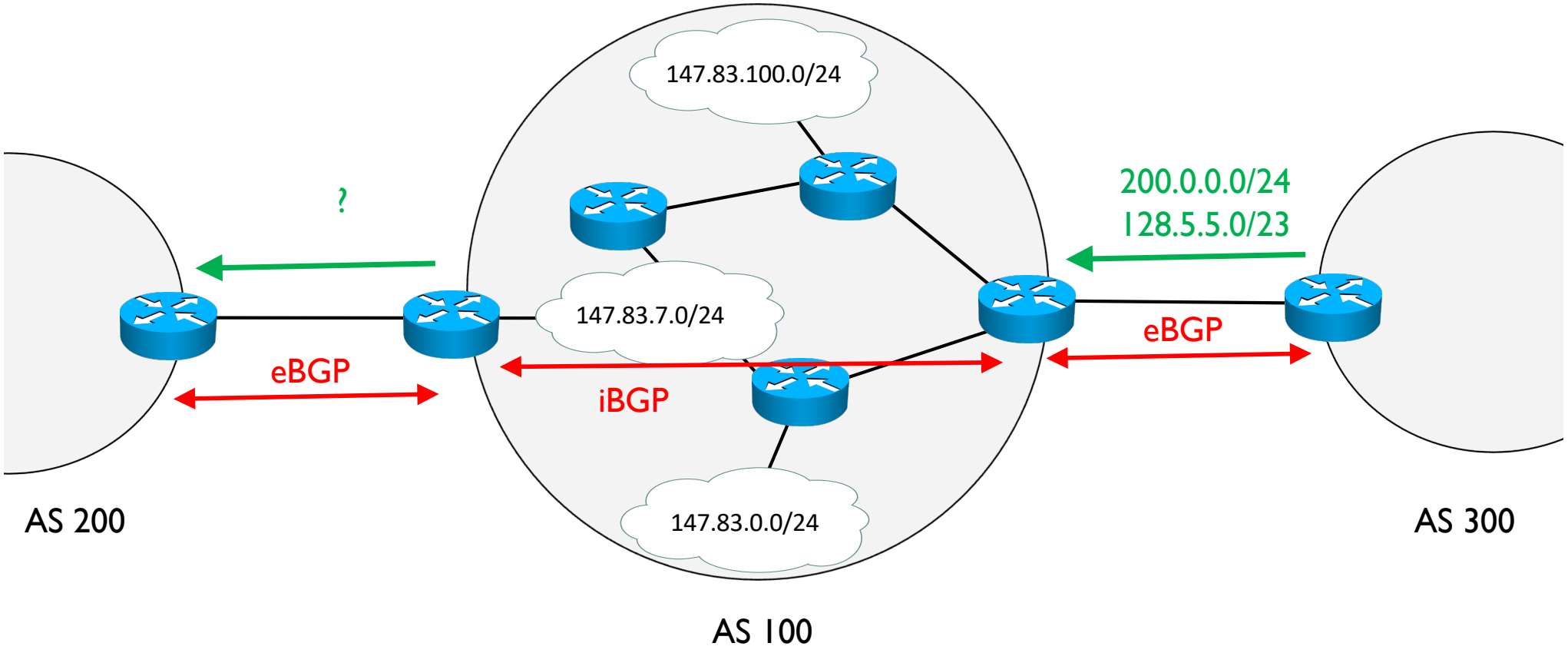
5.3 – Funcionamiento del BGP

external BGP vs internal BGP



5.3 – Funcionamiento del BGP

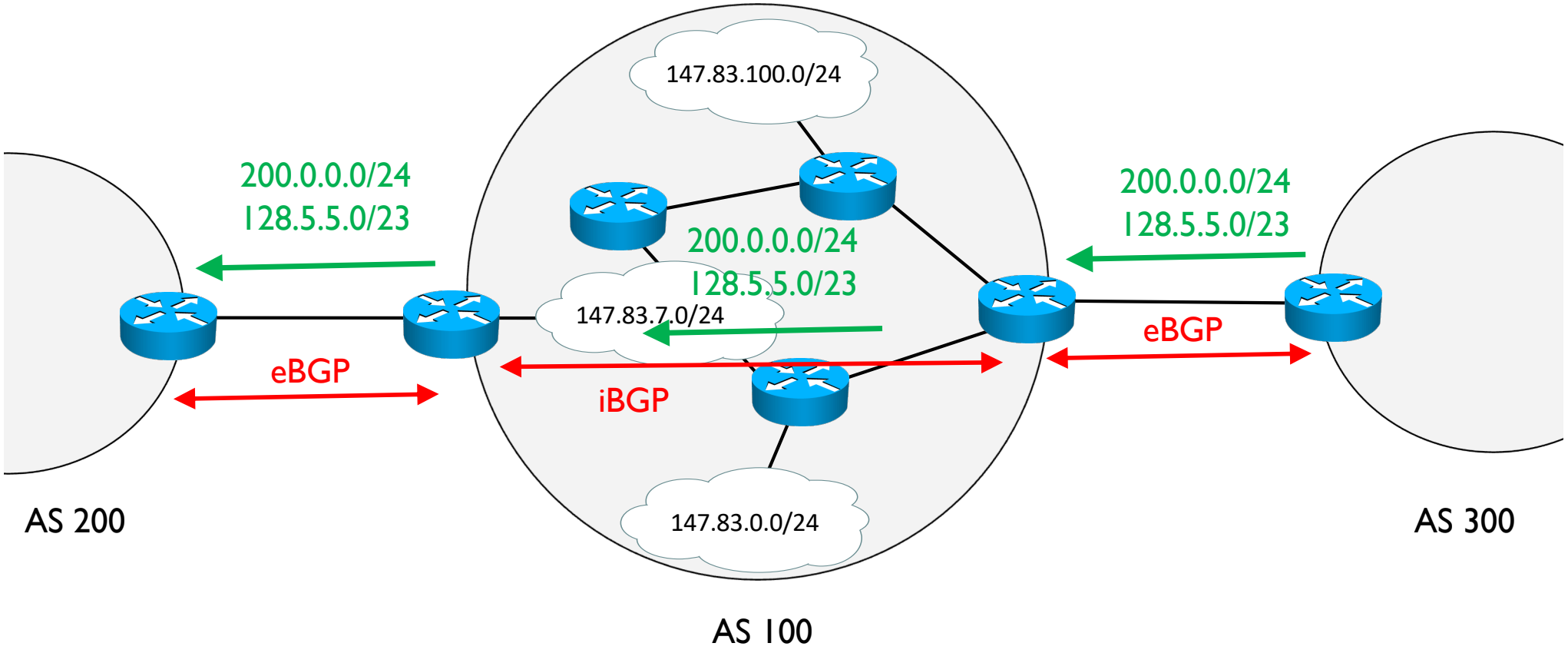
external BGP vs internal BGP



5.3 – Funcionamiento del BGP

external BGP vs internal BGP

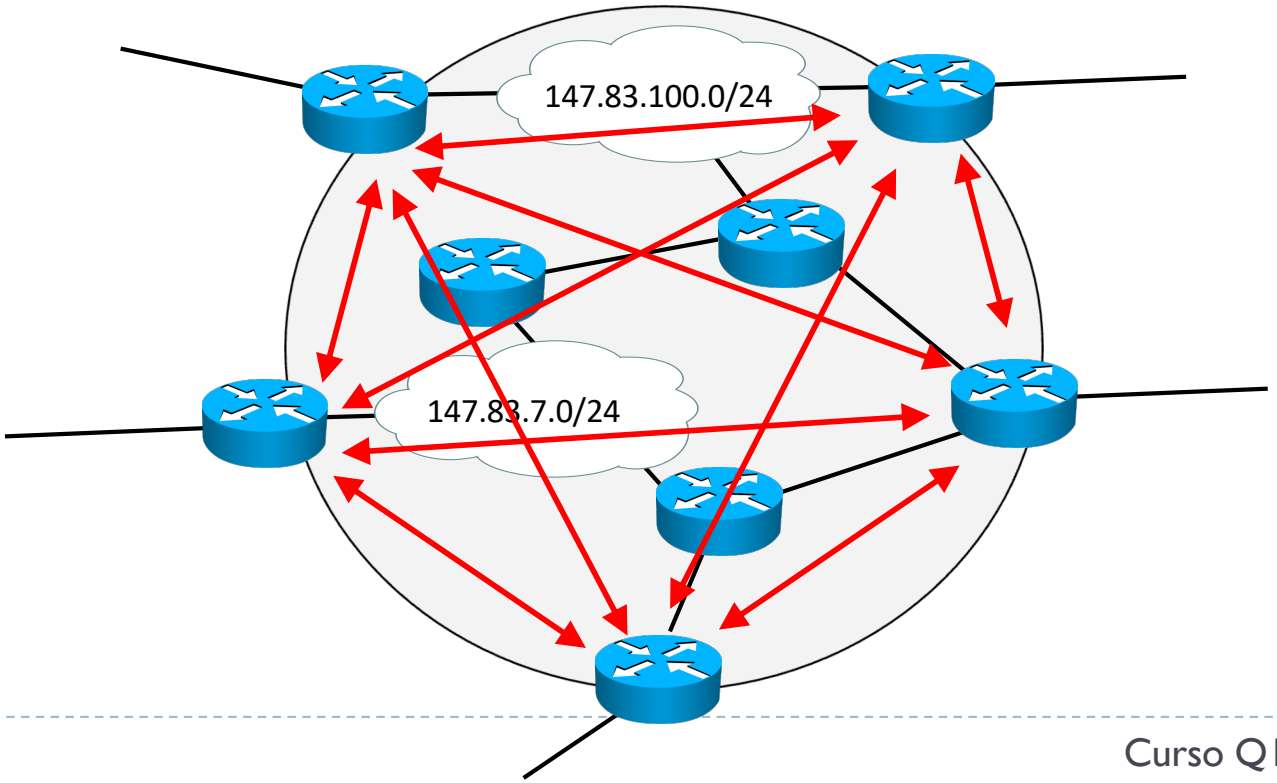
- De esta forma, los mensajes BGP van cruzando ASes y se van distribuyendo por todo Internet (además de quedarse internamente en cada AS)



5.3 – Funcionamiento del BGP

Sesiones iBGP

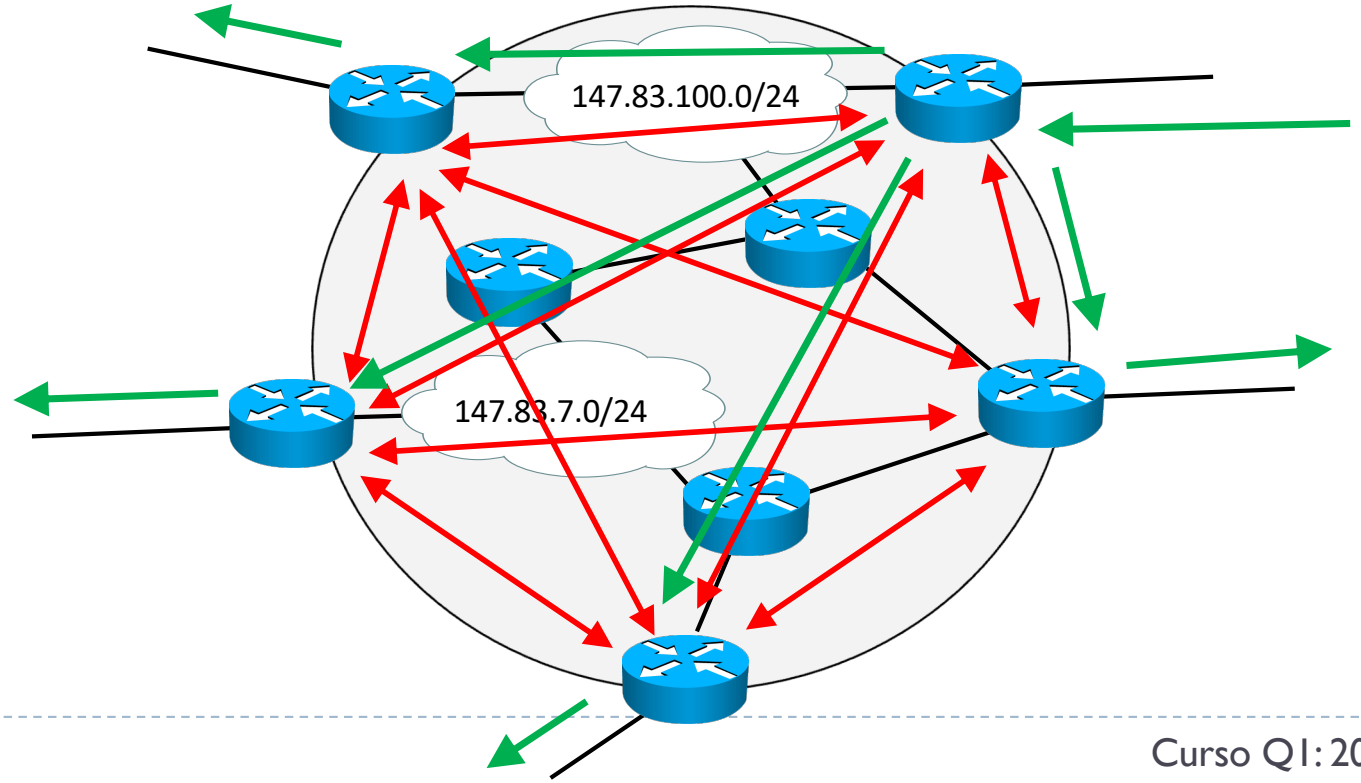
- Se necesita que todos los BGP speakers de un AS establezcan una sesión iBGP entre ellos
  - Se crea una full-mesh (malla completa) de iBGP (entre routers que tienen por lo menos una sesión BGP con otro AS)
  - De esta manera se evitan bucles de mensajes BGP



5.3 – Funcionamiento del BGP

Sesiones iBGP

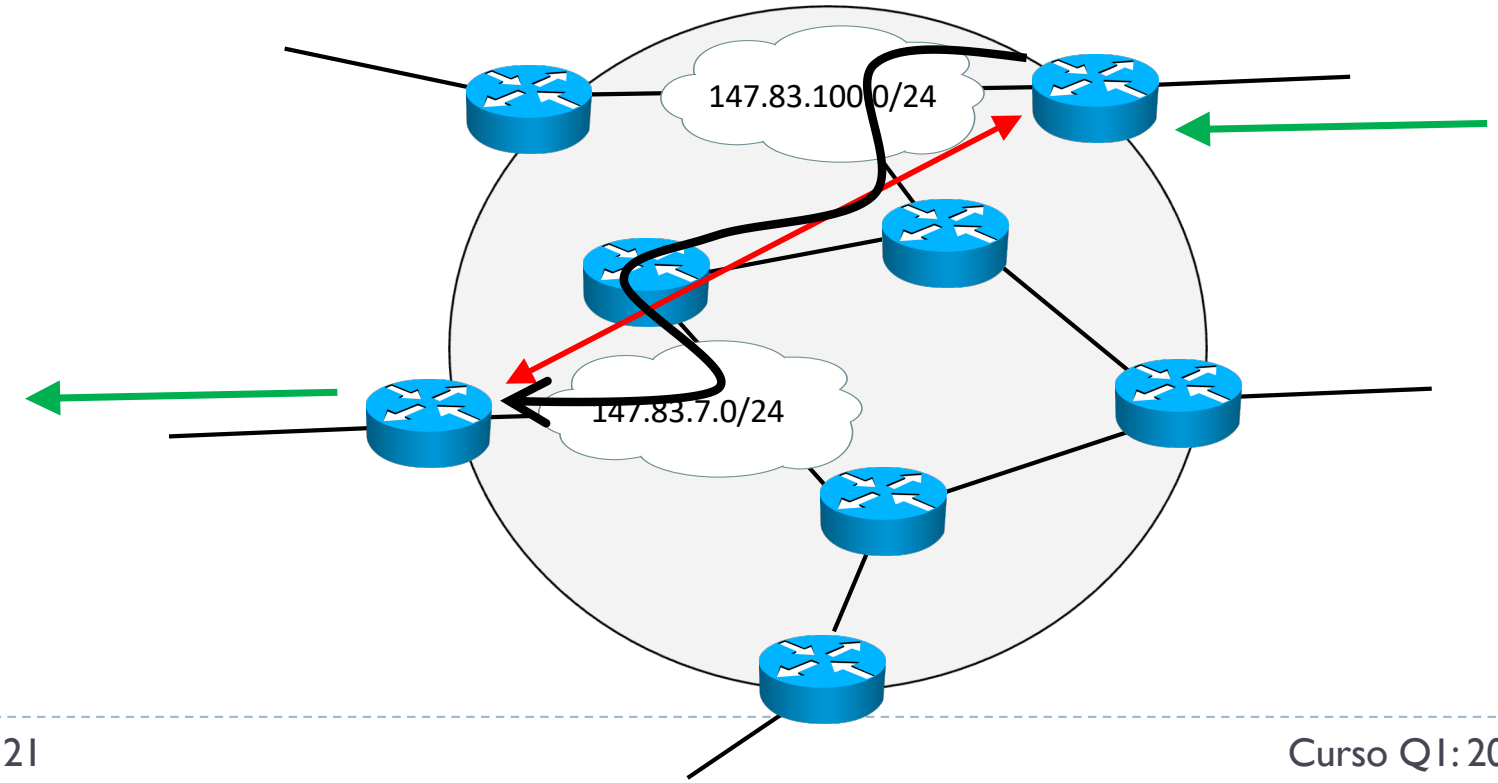
- Los iBGP speakers usan los mismos mensajes que los eBGP
- Los iBGP speakers anuncian solo los prefijos que aprenden de los eBGP pero no pueden re-enviar los recibidos de otro iBGP



5.3 – Funcionamiento del BGP

Sesiones iBGP

- Los mensajes iBGP se envían como si fueran paquetes normales encaminados según las tablas de encaminamiento de los routers
- El origen y destino de estos mensajes son los iBGP speakers
- No confundir iBGP con el protocolo de encaminamiento interno IGP



5. Encaminamiento inter-dominio

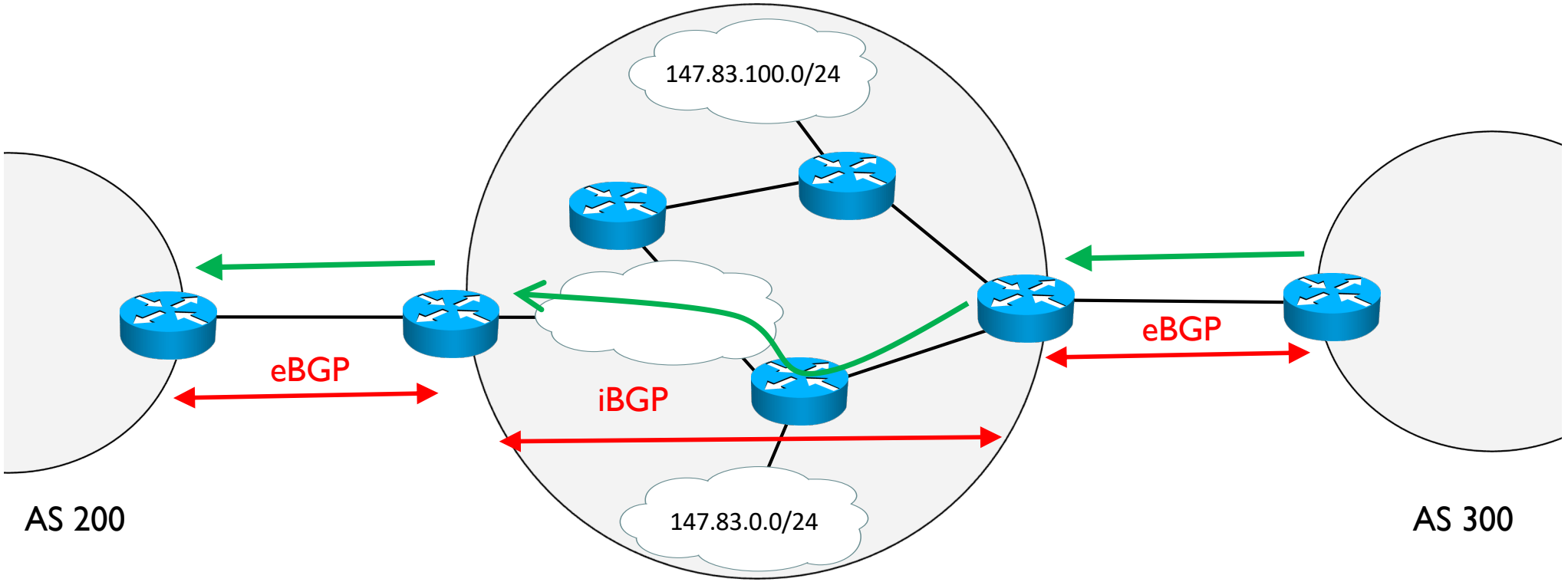
1. Introducción
2. Encaminamiento path-vector
3. Funcionamiento de BGP
4. Establecimiento de una sesión BGP
5. Bases de datos BGP
6. Mensajes BGP
7. Atributos estándares
8. Algoritmo de selección de rutas
9. Escenarios comunes
10. Mejoras del BGP

5.4 - Establecimiento sesión BGP

- A la hora de establecer una sesión BGP se puede elegir una interfaz real o una virtual. Se suele usar
  - una interfaz real para eBGP
  - una interfaz virtual para iBGP

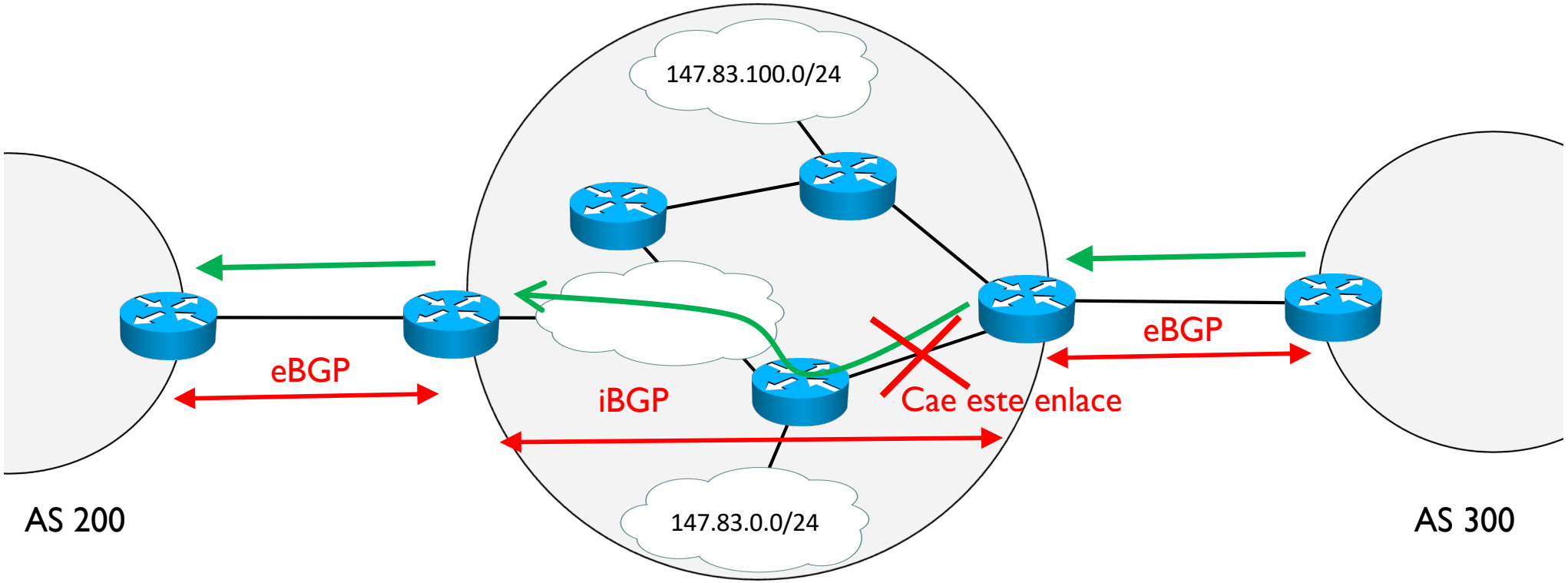
5.4 - Establecimiento sesión BGP

- A la hora de establecer una sesión BGP se puede elegir una interfaz real o una virtual. Se suele usar
  - una interfaz real para eBGP
  - una interfaz virtual para iBGP



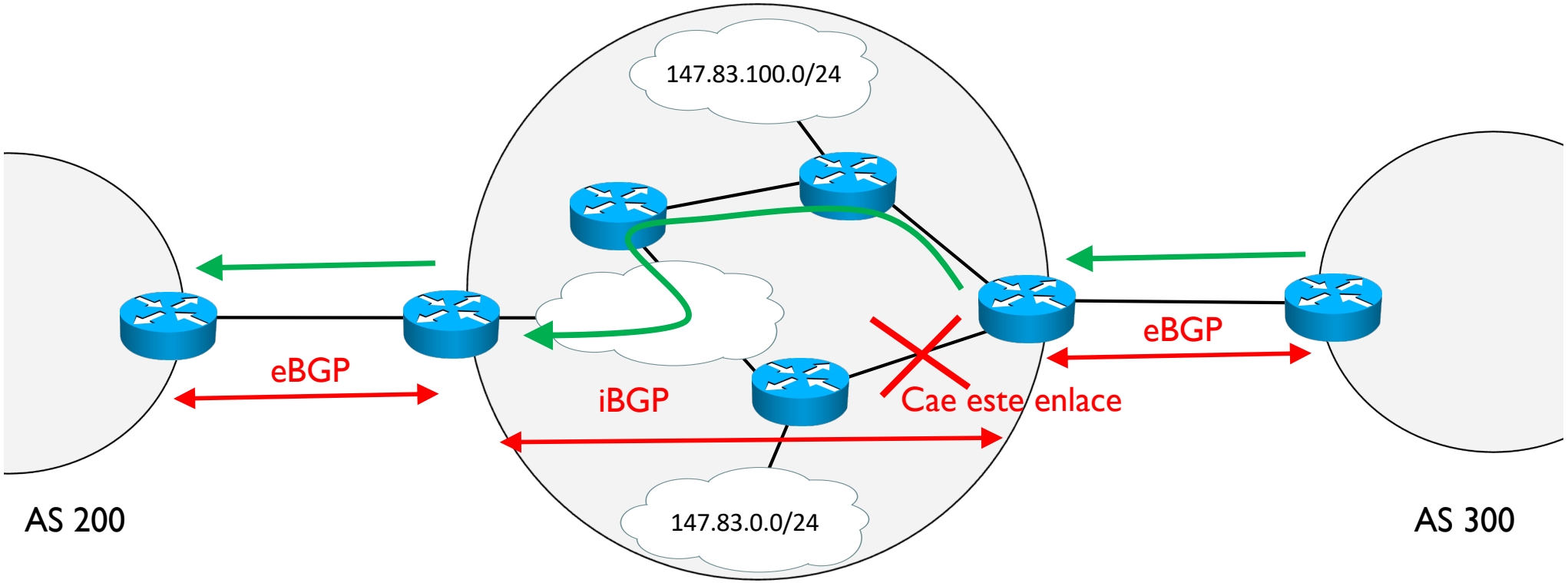
5.4 - Establecimiento sesión BGP

- Si se usara una interfaz real para iBGP y cae un enlace que se usa para encaminar los mensajes iBGP
- → Cae la sesión iBGP
- Hay que volver a configurarla manualmente usando otra interfaz



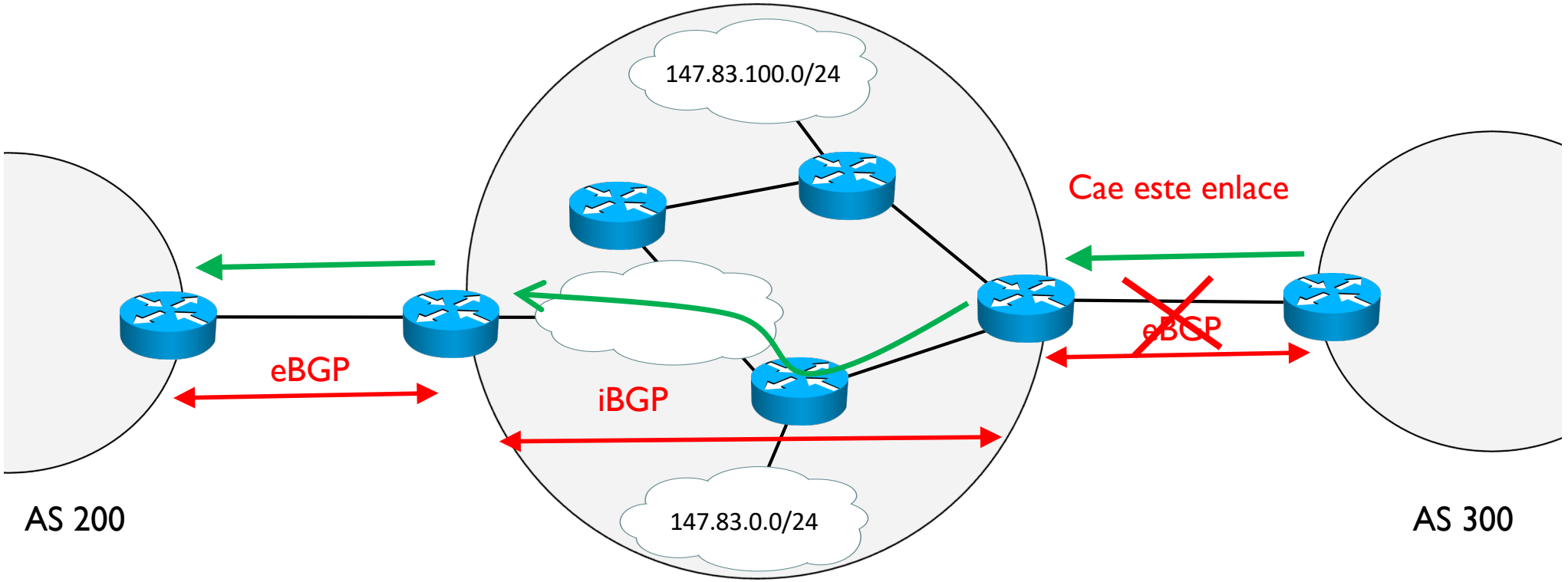
5.4 - Establecimiento sesión BGP

- Si en cambio se usara una interfaz virtual para iBGP y cae un enlace que se usa para encaminar los mensajes iBGP
- → el protocolo IGP (interno) encontraría otra ruta para estos mensajes
- La sesión iBGP sigue activa



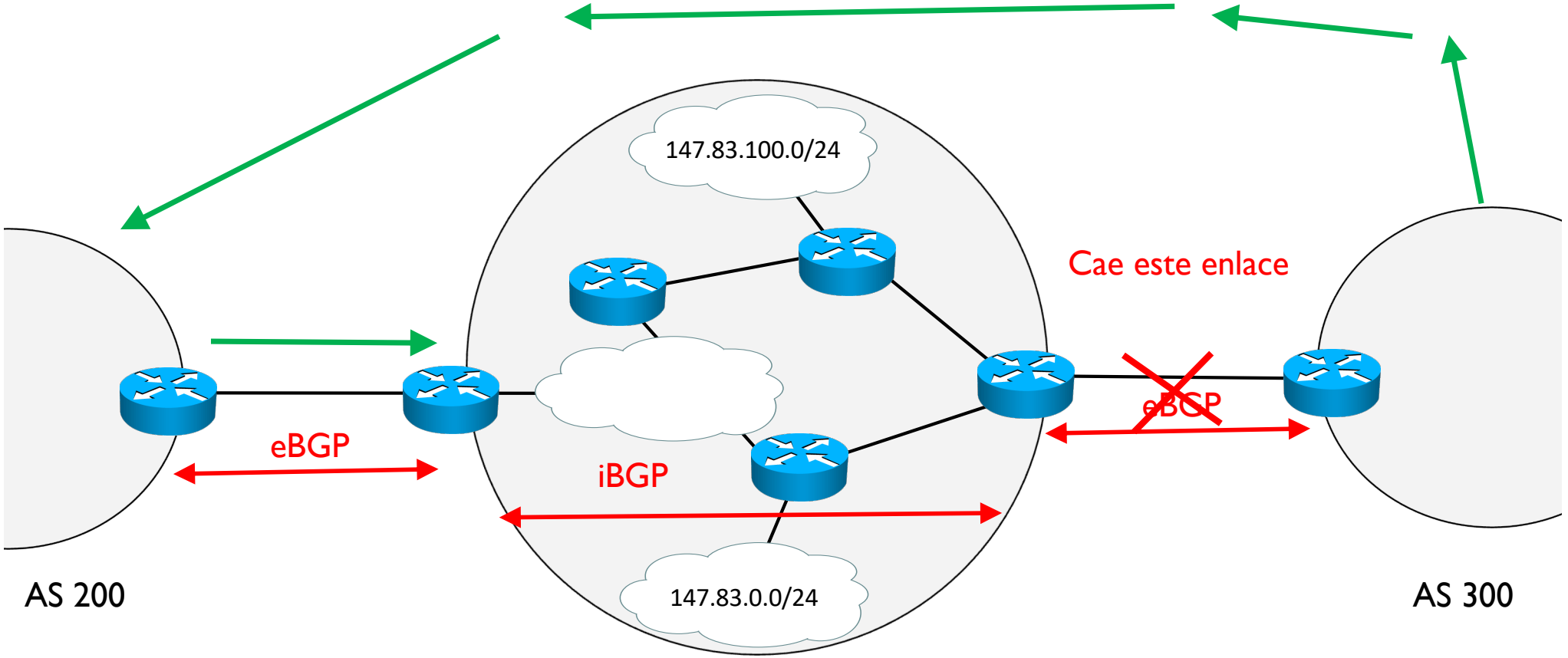
5.4 - Establecimiento sesión BGP

- Se usa una interfaz real en el caso eBGP porque se quiere que sea BGP que se ocupe de encontrar una ruta alternativa
- Y BGP busca otra ruta solo si se entera del fallo



5.4 - Establecimiento sesión BGP

- Se usa una interfaz real en el caso eBGP porque se quiere que sea BGP que se ocupe de encontrar una ruta alternativa
- Y BGP busca otra ruta solo si se entera del fallo

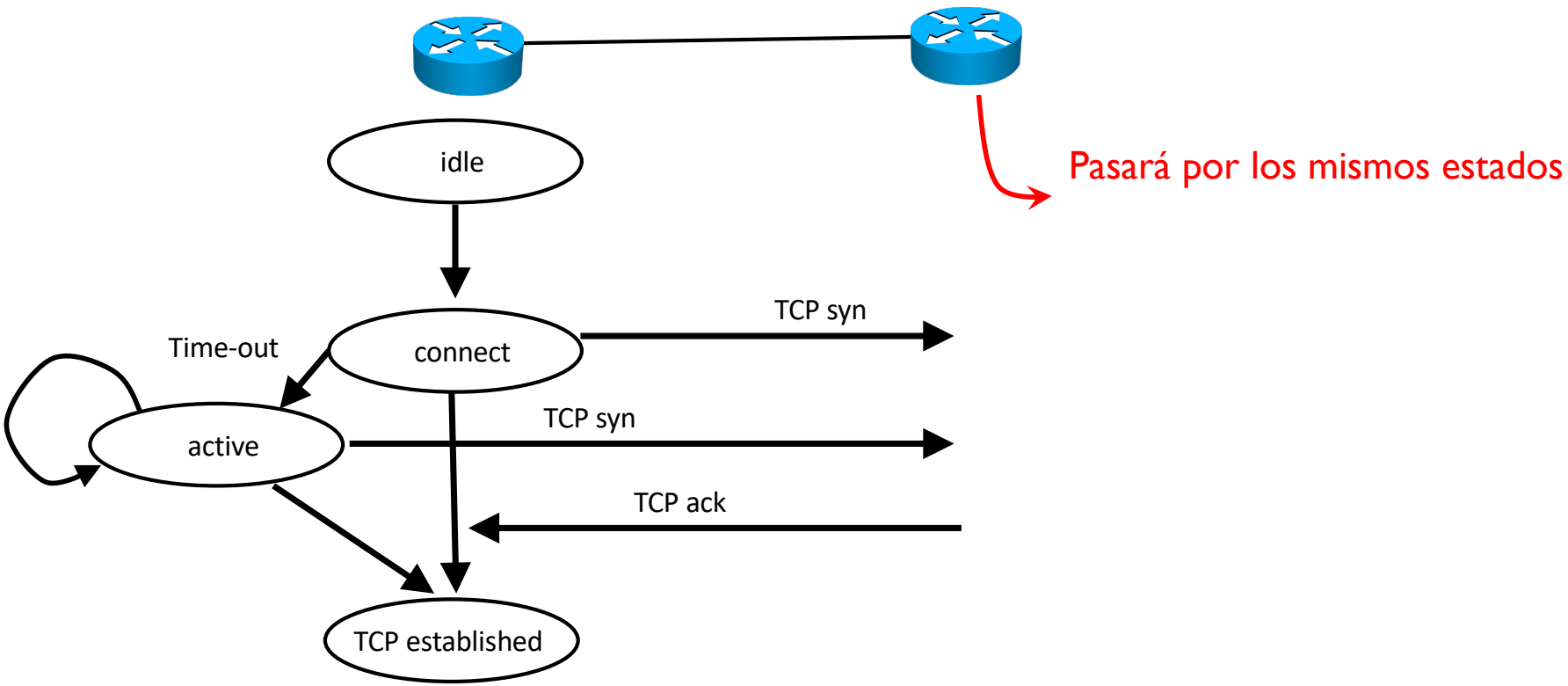


5.4 - Establecimiento sesión BGP

- La información que se intercambian los BGP speakers debe ser fiable
  - Se abren una conexión TCP entre ellos
  - Los dos extremos del TCP son los dos routers
  - Se usa el puerto 179 (BGP)

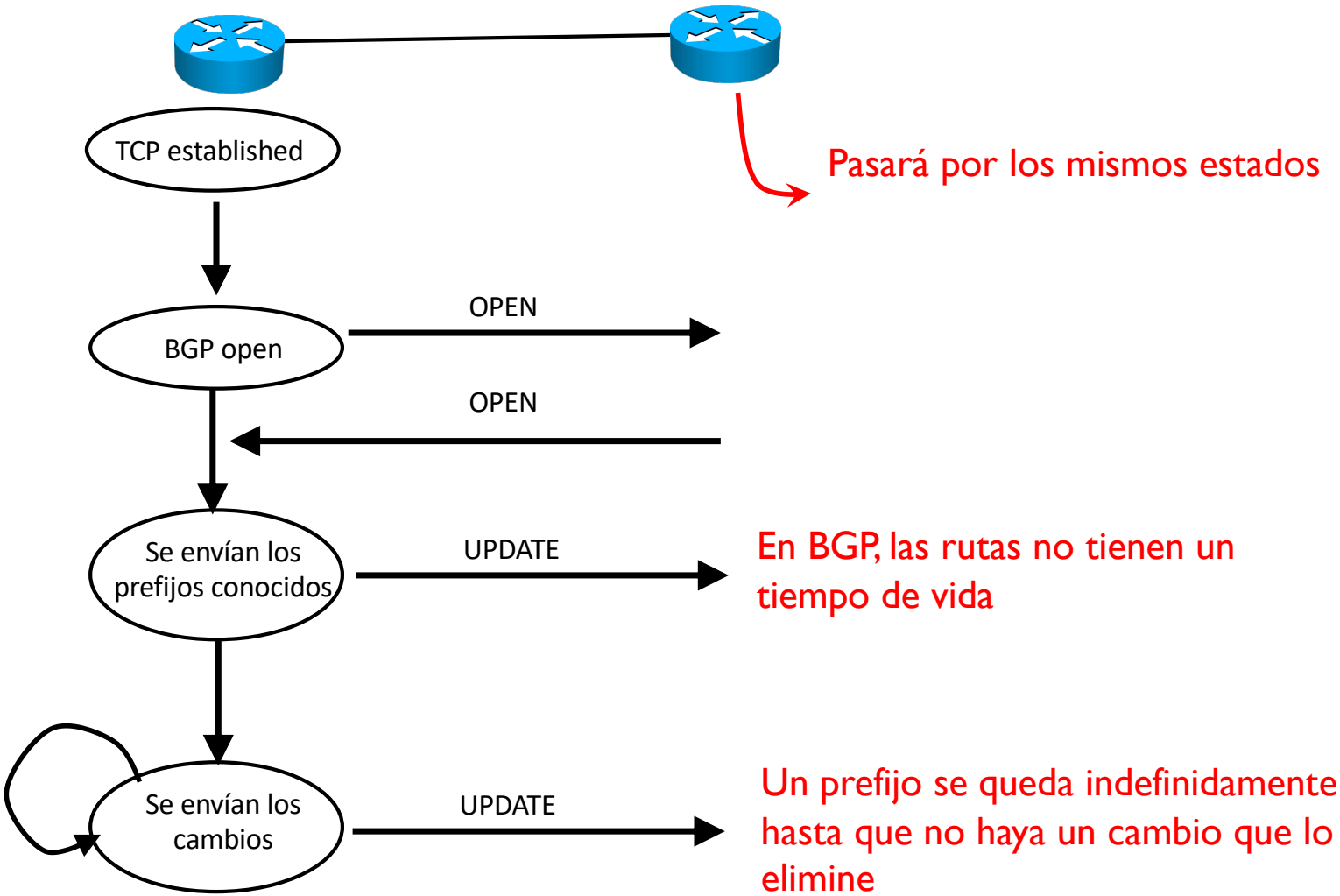
5.4 - Establecimiento sesión BGP

Estados



5.4 - Establecimiento sesión BGP

Estados

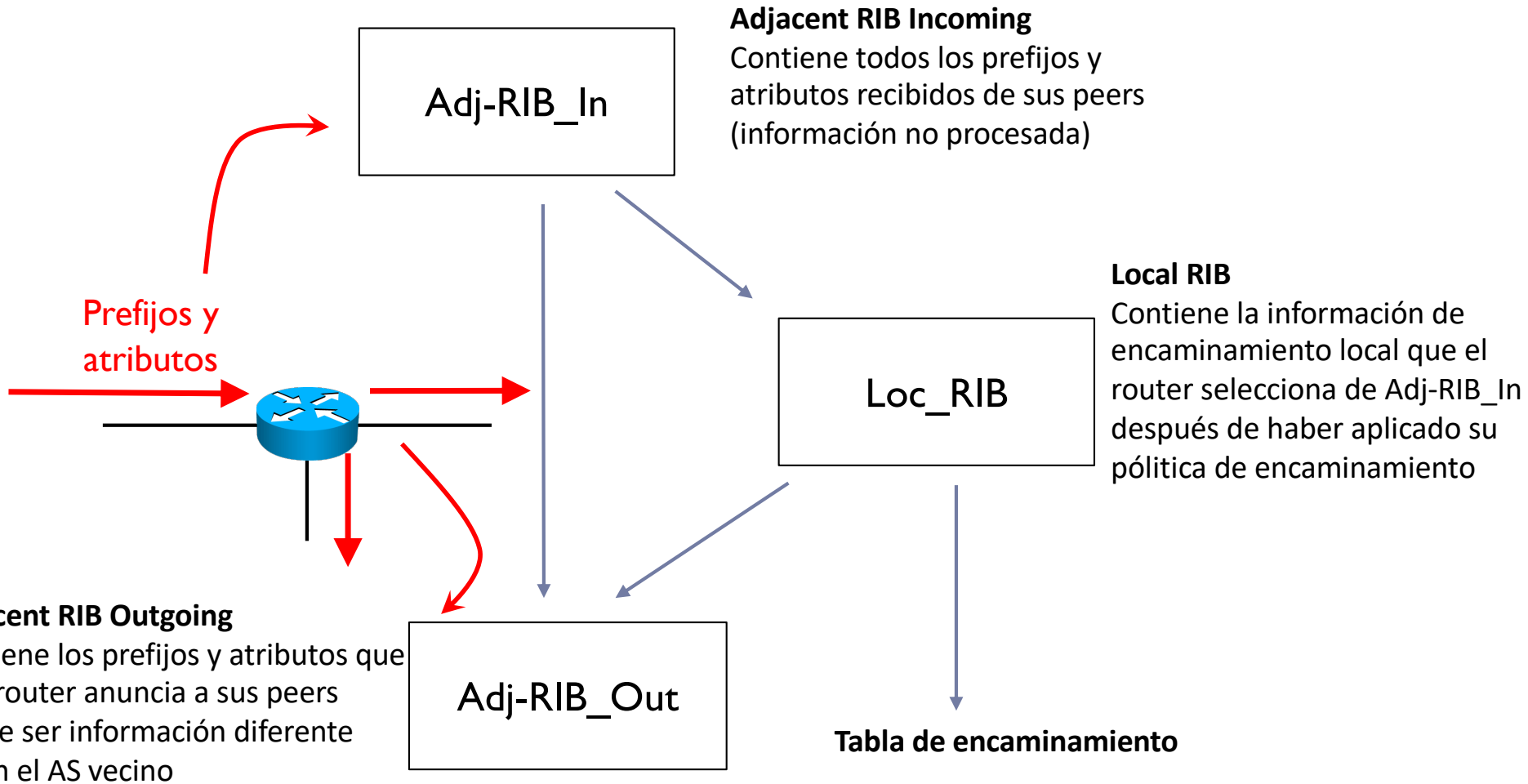


5. Encaminamiento inter-dominio

1. Introducción
2. Encaminamiento path-vector
3. Funcionamiento de BGP
4. Establecimiento de una sesión BGP
5. Bases de datos BGP
6. Mensajes BGP
7. Atributos estándares
8. Algoritmo de selección de rutas
9. Escenarios comunes
10. Mejoras del BGP

5.5 - Bases de datos en BGP

- Un router BGP mantiene 3 bases de datos

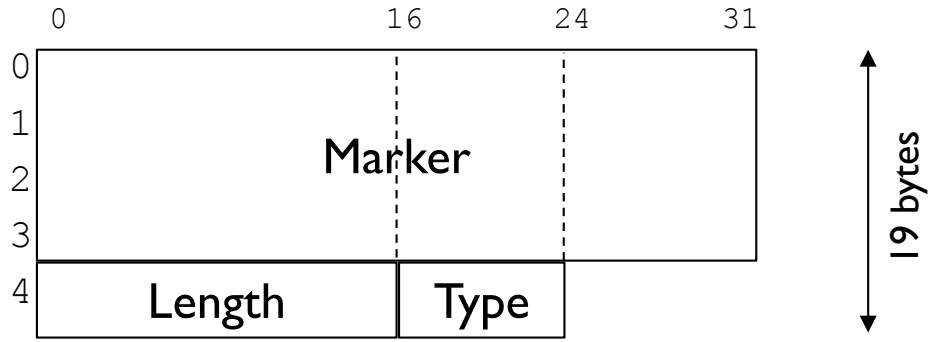


5. Encaminamiento inter-dominio

1. Introducción
2. Encaminamiento path-vector
3. Funcionamiento de BGP
4. Establecimiento de una sesión BGP
5. Bases de datos BGP
6. Mensajes BGP
7. Atributos estándares
8. Algoritmo de selección de rutas
9. Escenarios comunes
10. Mejoras del BGP

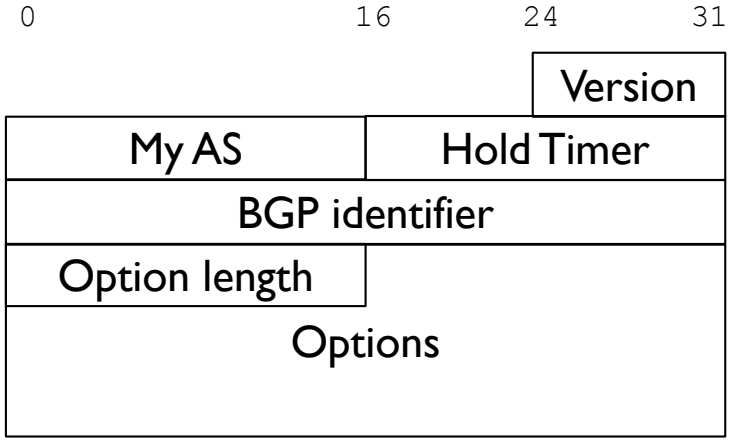
5.6 - Mensajes BGP

- Cabecera común



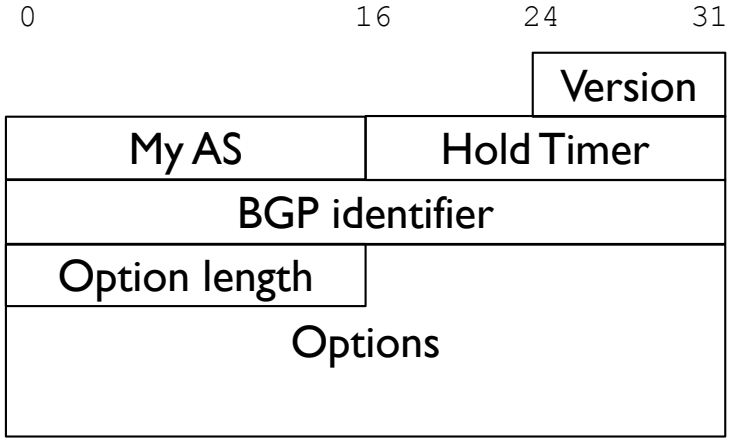
- Marker
  - Seguridad
  - Si primer mensaje o no hay seguridad, son todos 1
  - En otros casos, se aplica la seguridad negociada entre los dos peers
- Length
  - Longitud de todo el mensaje BGP (cabecera + payload)
- Type
  - 1. Open
  - 2. Update
  - 3. Notification
  - 5. Keepalive

5.6 - Mensajes BGP
OPEN

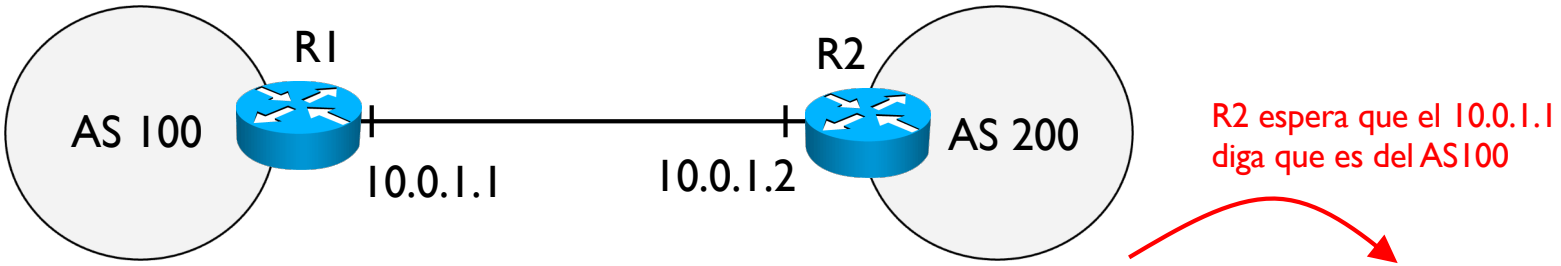


- Después de el establecimiento de la sesión TCP, los routers se envían un OPEN
- Los objetivos son
  - Identificarse
  - Negociar los parámetros del BGP

5.6 - Mensajes BGP
OPEN



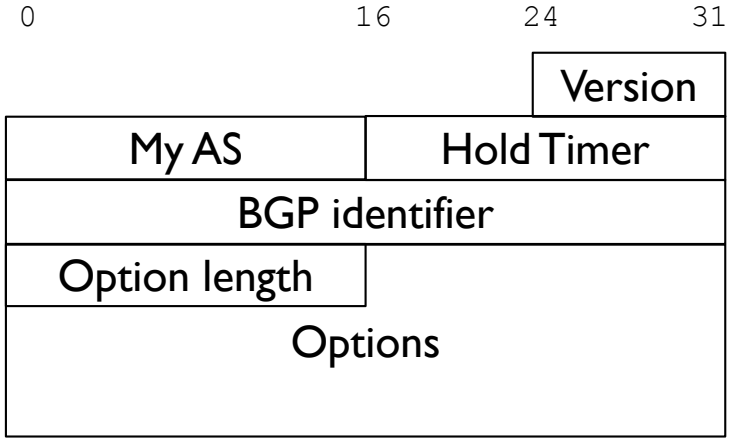
- Version: 4 (solo se usa esta versión actualmente)
- My AS:
  - El router envía el número de su AS
  - El receptor compara esta número con el que se espera recibir de este vecino




```
R1> neighbor 10.0.1.2 remote-as 200
```

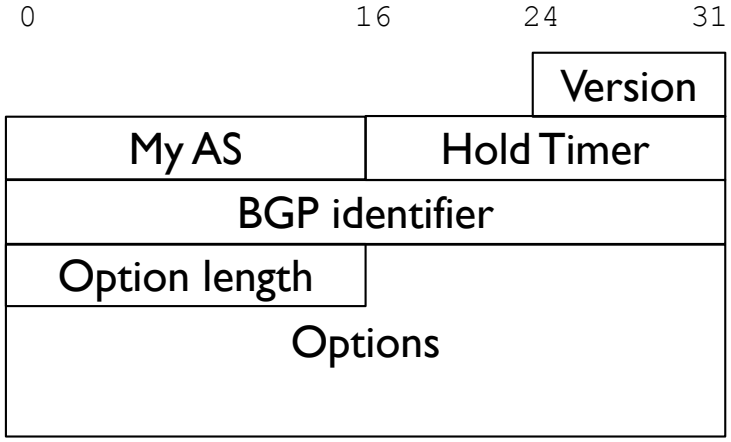
```
R2> neighbor 10.0.1.1 remote-as 100
```

5.6 - Mensajes BGP
OPEN



- Hold Timer
  - Especifica los segundos que el router quiere usar como Hold Timer
  - Es el tiempo máximo que un router puede esperar sin recibir mensajes BGP del otro peer (por defecto se usan 60 segundos)
    - Este Hold Timer se resetea cada vez que el router recibe un UPDATE o un KEEPALIVE
  - Generalmente se esperan n Hold Timer (por defecto 3) antes de considerar que ha pasado algo a la sesión BGP
    - Una vez pasados n Hold Timers, se cierra la sesión BGP que pasa a idle
  - Si los dos routers se envían 2 valores distintos, generalmente se usa el menor de los dos
  - Se puede configurar a 0, que significa que no se envían los mensajes KEEPALIVE

5.6 - Mensajes BGP
OPEN

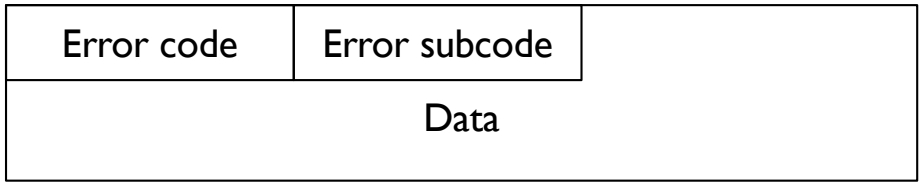


- BGP identifier
  - Es el RID del router
- Option length
  - Longitud del campo options
- Options
  - Contiene toda la información que un router quiere anunciar al otro sobre sus "capacidades" RFC 5492
  - Por ejemplo, métodos de autentificación y seguridad, mejoras como Comunidades, etc.

5.6 - Mensajes BGP
KEEPALIVE

- Se usa para verificar la conectividad entre dos peers
- Se envía cada vez que expira el Hold Timer establecido durante la etapa OPEN
  - Recordar que el Hold Timer se reinicia también si se envía un UPDATE
- El mensaje solo consiste de la cabecera común de 19 bytes

5.6 - Mensajes BGP
NOTIFICATION



- Si hay un error durante la sesión BGP, se usa este tipo de mensaje
- → Al enviar este mensaje, se cierra la sesión BGP
- Error code proporciona información general
- Ejemplo de errores
  - Error code: error en la cabecera común BGP — subcode: longitud incorrecta
  - Error code: error durante el proceso OPEN — subcode: versión 3 vs versión 4
    subcode: AS no aceptable
  - Error code: error al procesar un UPDATE — subcode: atributo no valido
  - Error code: Hold Timer expirado y no se han recibido mensajes
- Error code proporciona información general, error subcode proporciona detalles más específicos y el campo dato completa la información

5.6 - Mensajes BGP
UPDATE

Withdraw routes length	2 bytes
Withdraw routes	variable
Total path attribute length	2 bytes
Path attribute	variable
Network Layer Reachability Information (NLRI)	variable

- Se usa para enviar información de encaminamiento entre peers

- Withdraw routes length
  - Contiene la longitud del siguiente campo que es variable y puede ser 0 (no se eliminan rutas)
- Withdraw routes
  - Un peer puede notificar al otro que algunos prefijos previamente notificado ya no son validos y hay que borrarlos

5.6 - Mensajes BGP
UPDATE

Withdraw routes length	2 bytes
Withdraw routes	variable
Total path attribute length	2 bytes
Path attribute	variable
Network Layer Reachability Information (NLRI)	variable

- NLRI
  - Contiene la lista de prefijos anunciados por este peer
  - El número de prefijos está limitado por el tamaño máximo de un mensajes BGP (4096 bytes)
  - Formato:
    - longitud prefijo – net-id del prefijo
    - Por ejemplo para informar del prefijo 147.83.0.0/16 → se envía 16 - 147.83
  - La longitud de este campo no está especificada directamente pero se puede deducir por los otros campos

Longitud NLRI = Length – 19 bytes – 2 bytes – Longitud de withdraw routes – 2 bytes – Longitud de path attribute

- Length: Campo de la cabecera común
- 19 bytes: Longitud cabecera común
- 2 bytes: Longitud withdraw routes length
- Longitud de withdraw routes: Valor de Withdraw routes length
- 2 bytes: Longitud withdraw routes length
- Longitud de path attribute: Valor de Total path attribute Length

5.6 - Mensajes BGP
UPDATE

Withdraw routes length	2 bytes
Withdraw routes	variable
Total path attribute length	2 bytes
Path attribute	variable
Network Layer Reachability Information (NLRI)	variable

- Total path attribute length
  - Contiene la longitud del siguiente campo que es variable (no puede ser 0 ya que hay atributos obligatorios)
- Path attribute
  - La gran ventaja de BGP es que funciona a través de políticas que se determinan usando atributos
  - Una política indica las preferencias a la hora de seleccionar una ruta (veremos más adelante)
  - Estos atributos son comunes a todos los prefijos que se ponen en NLRI
  - Si hay prefijos que no tienen los mismos atributos, hay que separarlos en UPDATE diferentes
  - Se pueden definir atributos estandar o definir de nuevos

5.6 - Mensajes BGP
UPDATE

Withdraw routes length	2 bytes
Withdraw routes	variable
Total path attribute length	2 bytes
Path attribute	variable
Network Layer Reachability Information (NLRI)	variable

- Path attribute
- Estandar
  - ORIGEN
  - AS-PATH
  - NEXT HOP
  - MULTI-EXIT DISCRIMINATOR
  - LOCAL-PREFERENCE
  - AGGREGATOR
- Propietarios
  - Weight (CISCO)

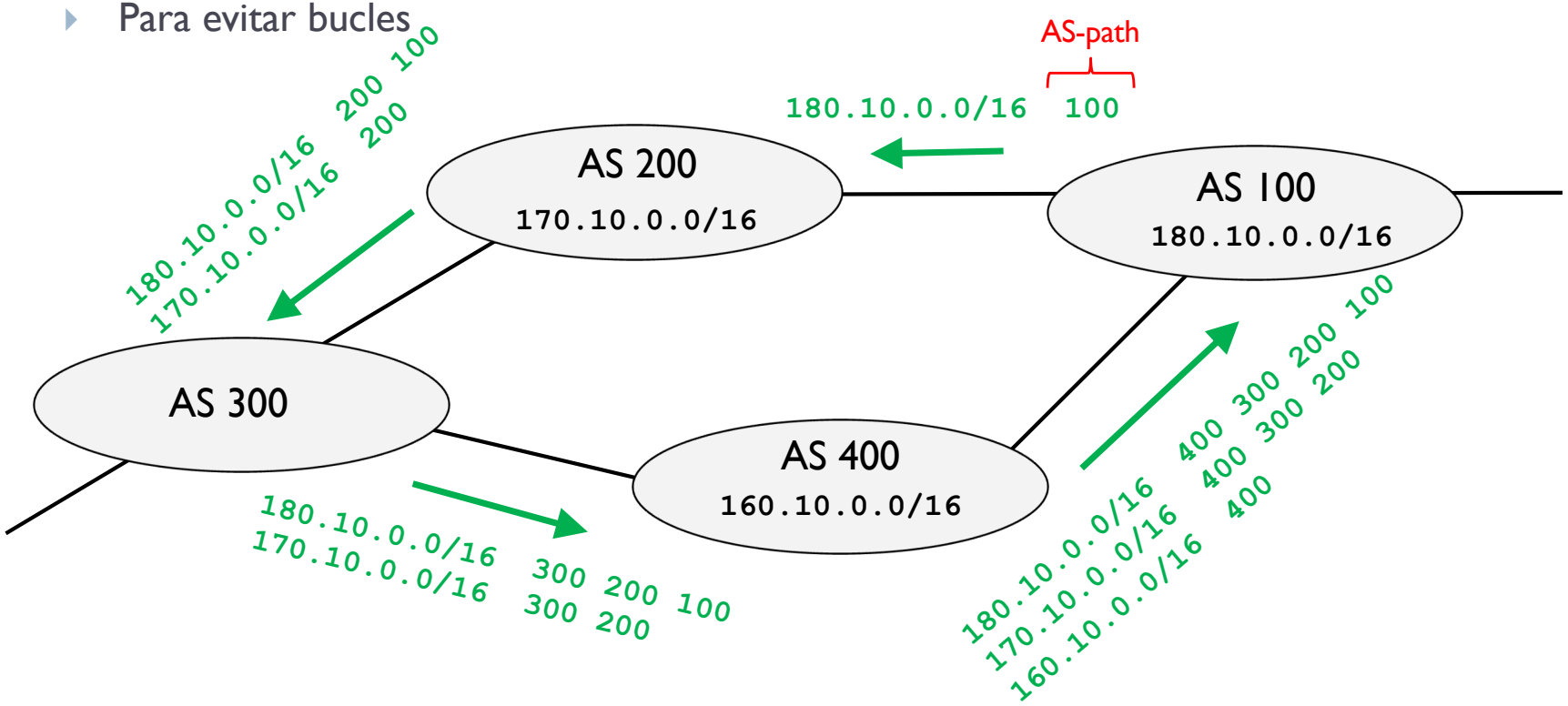
5. Encaminamiento inter-dominio

1. Introducción
2. Encaminamiento path-vector
3. Funcionamiento de BGP
4. Establecimiento de una sesión BGP
5. Bases de datos BGP
6. Mensajes BGP
7. Atributos estándares
8. Algoritmo de selección de rutas
9. Escenarios comunes
10. Mejoras del BGP

5.7 – Atributos estándares

AS-PATH

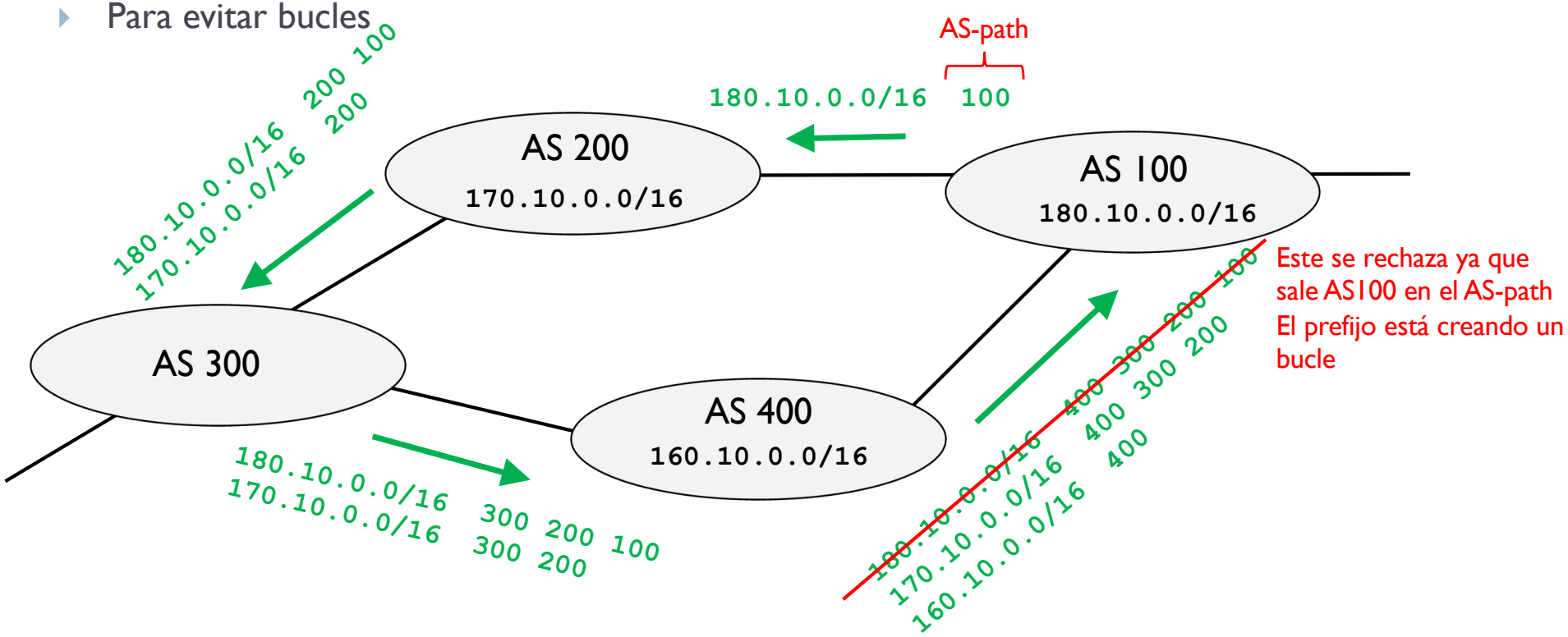
- Obligatorio
- Secuencia de ASN por donde ha pasado un prefijo
- Objetivos
  - Para evitar bucles



5.7 – Atributos estándares

AS-PATH

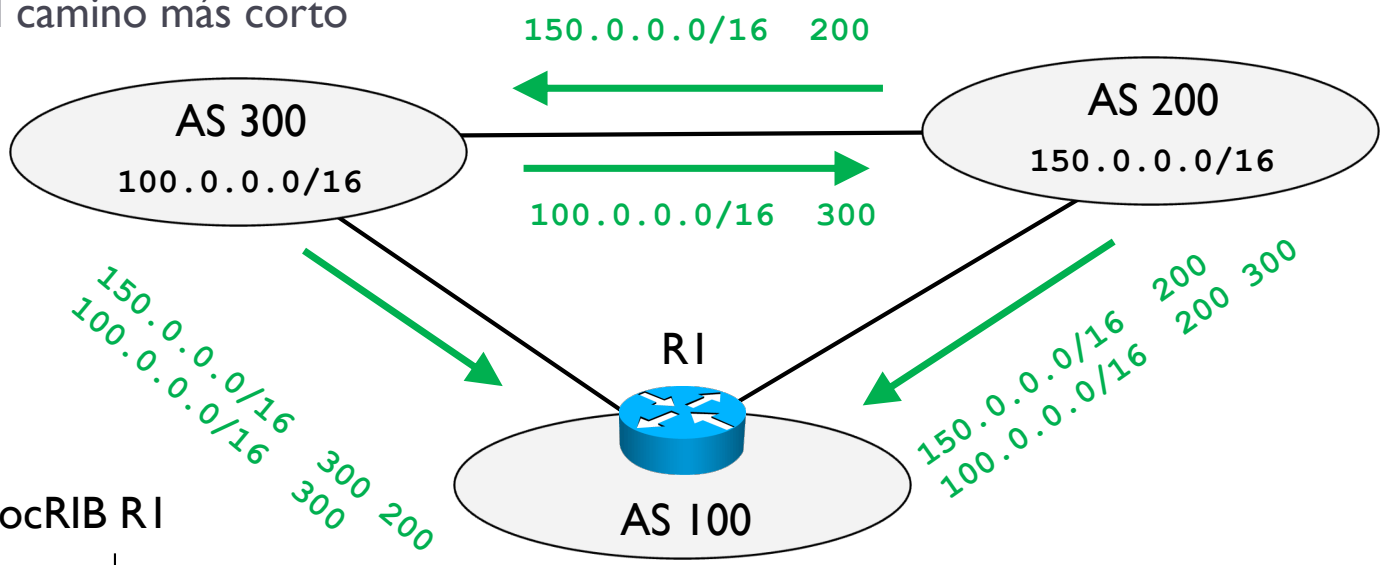
- Obligatorio
- Secuencia de ASN por donde ha pasado un prefijo
- Objetivos
  - Para evitar bucles



5.7 – Atributos estándares

AS-PATH

- Obligatorio
- Secuencia de ASN por donde ha pasado un prefijo
- Objetivos
  - Para evitar bucles
  - Para determinar el camino más corto



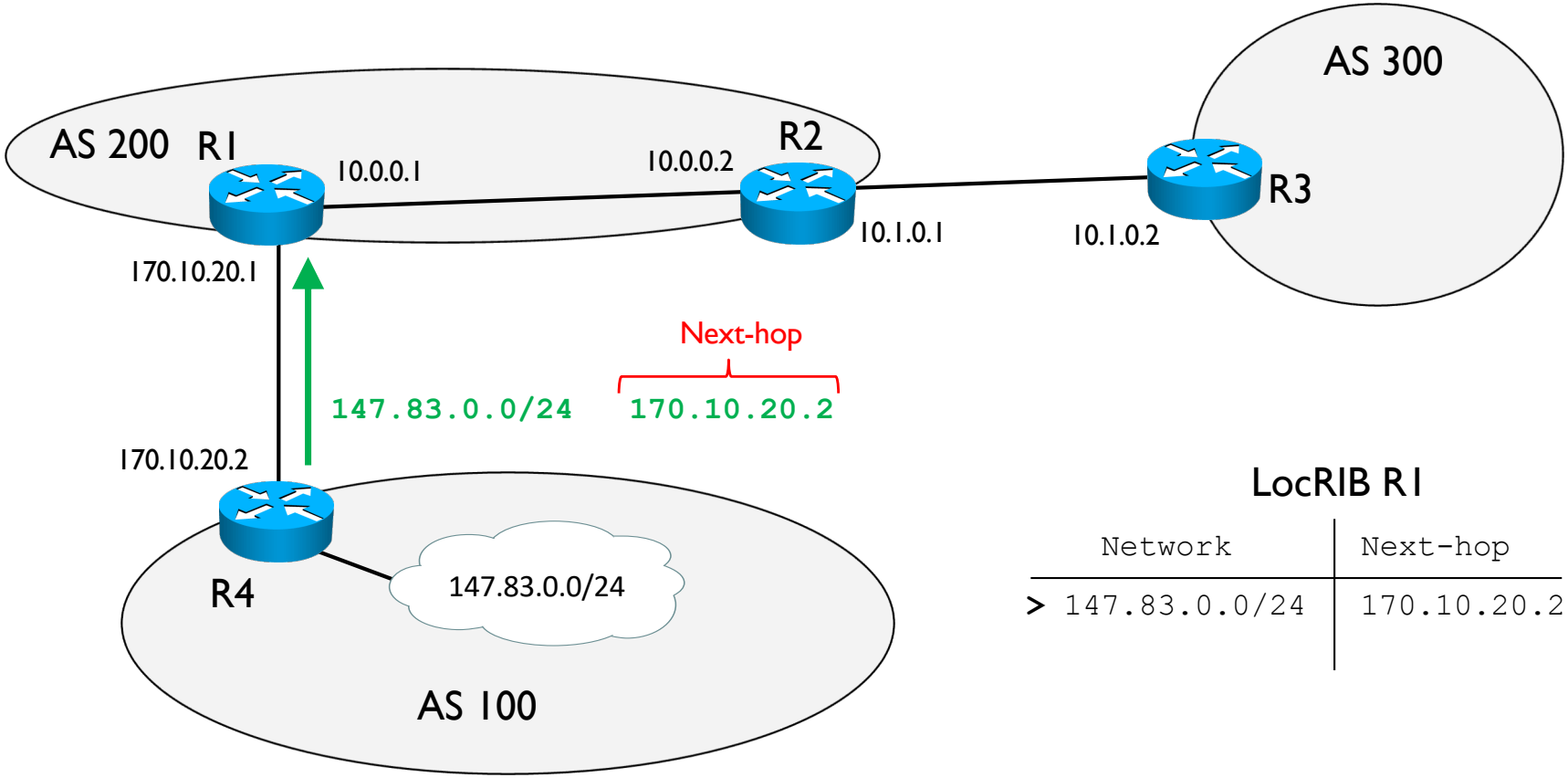
LocRIB R1

	Network	AS-path
>	100.0.0.0/16	300
		200 300
	150.0.0.0/16	300 200
>		200

BGP mantiene múltiples entradas
Para la tabla de encaminamiento, elige la "mejor" ruta y la marca con >
Si no hay otros atributos, elige la ruta con el AS-path más corto

5.7 – Atributos estándares
NEXT-HOP

- Obligatorio
- Indica la @IP del router que hace de gateway entre AS

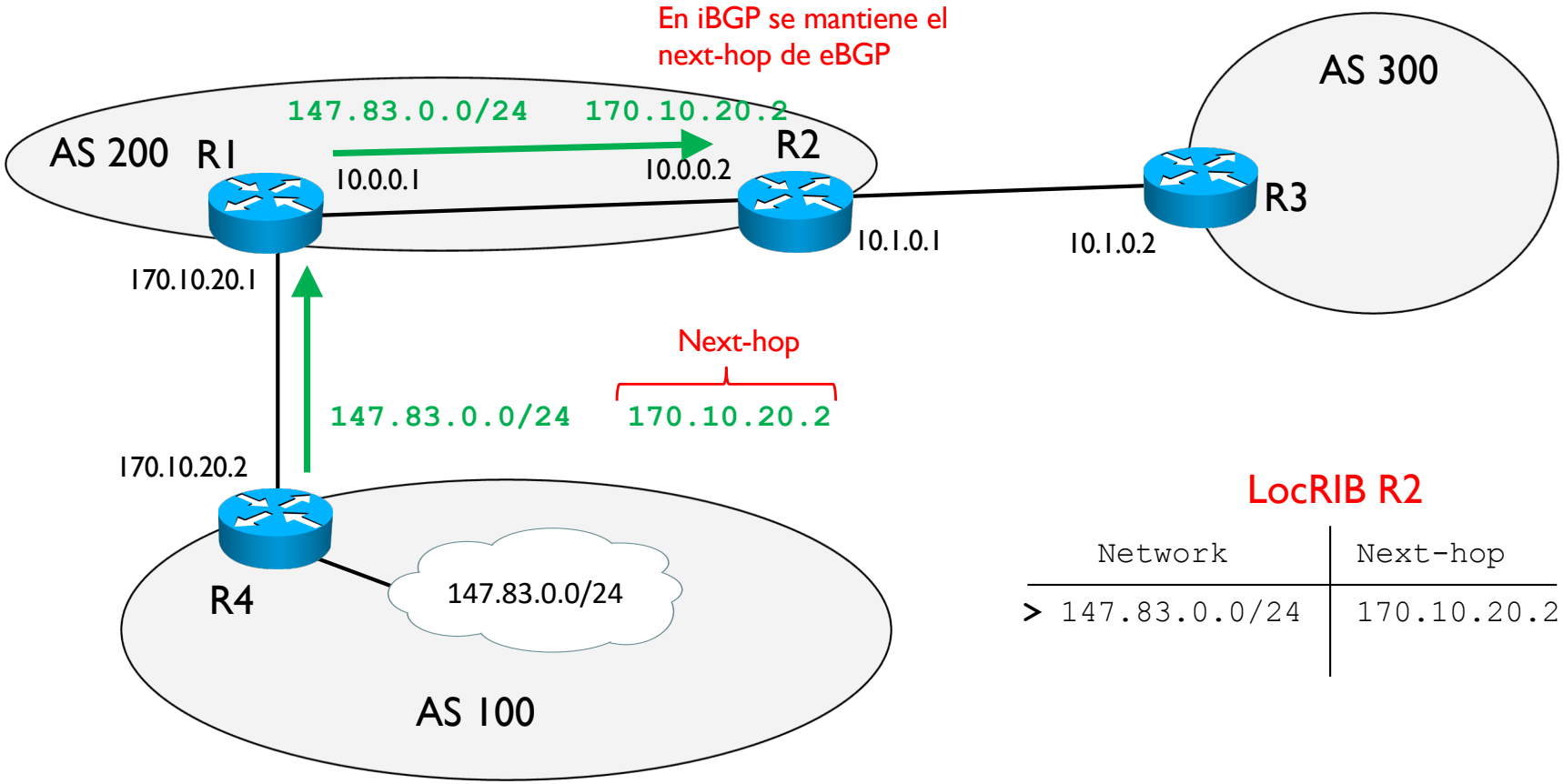


LocRIB R1

	Network	Next-hop
>	147.83.0.0/24	170.10.20.2

5.7 – Atributos estándares
NEXT-HOP

- Obligatorio
- Indica la @IP del router que hace de gateway entre AS

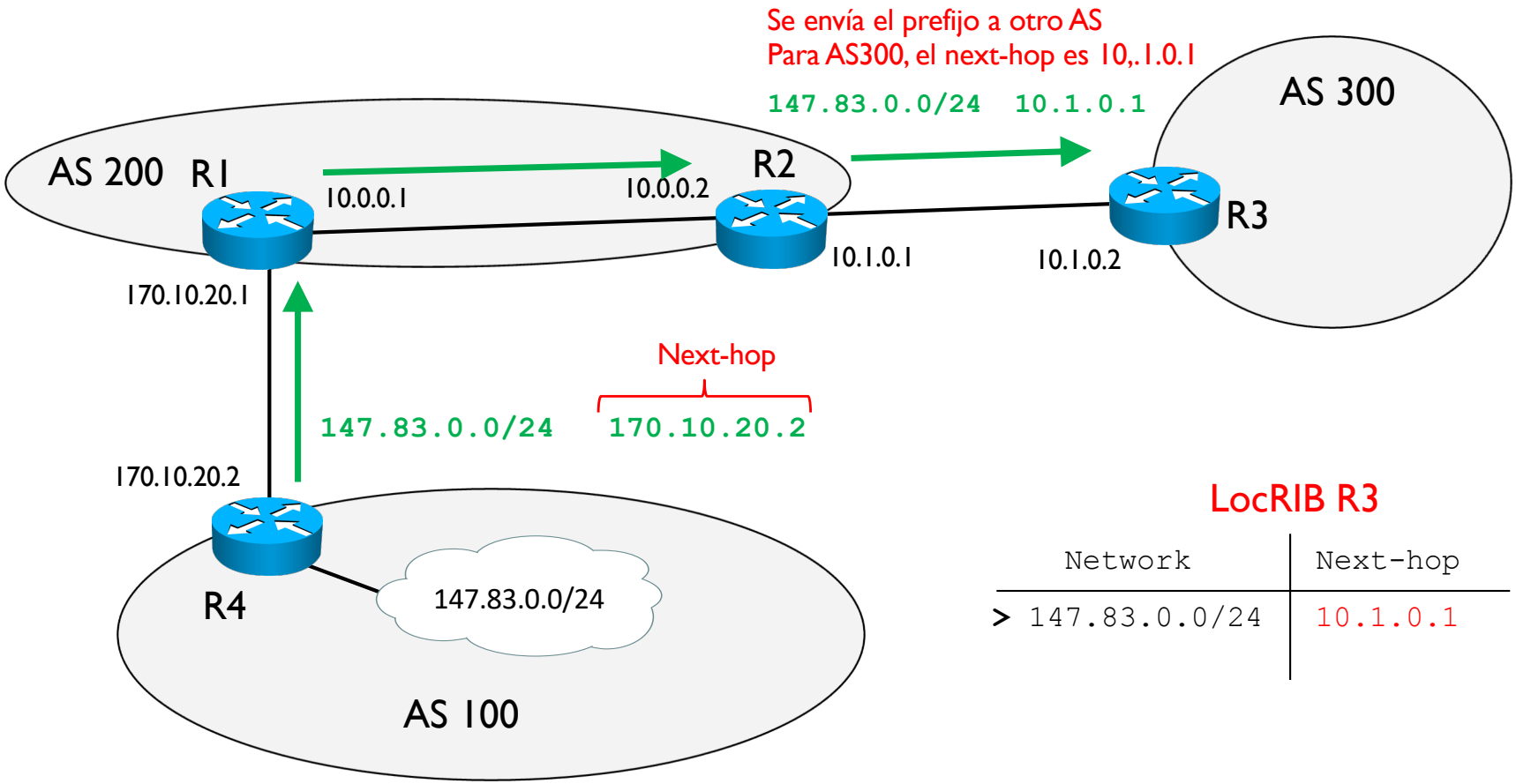


LocRIB R2

	Network	Next-hop
>	147.83.0.0/24	170.10.20.2

5.7 – Atributos estándares
NEXT-HOP

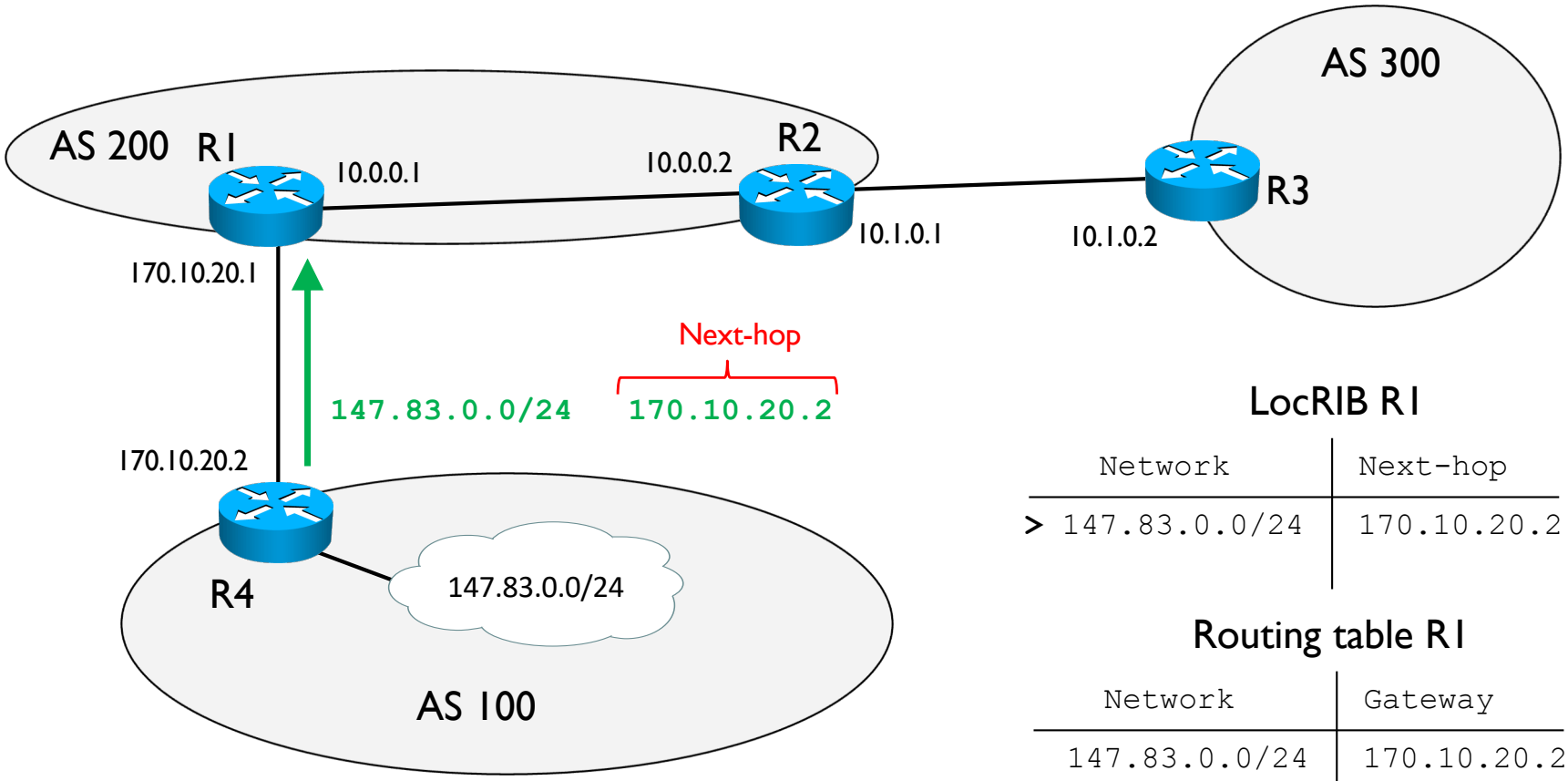
- Obligatorio
- Indica la @IP del router que hace de gateway entre AS



5.7 – Atributos estándares

NEXT-HOP

- No confundir gateway de una tabla de encaminamiento con next-hop
- Gateway es a nivel de routers, next-hop a nivel de AS



LocRIB R1

Network	Next-hop
> 147.83.0.0/24	170.10.20.2

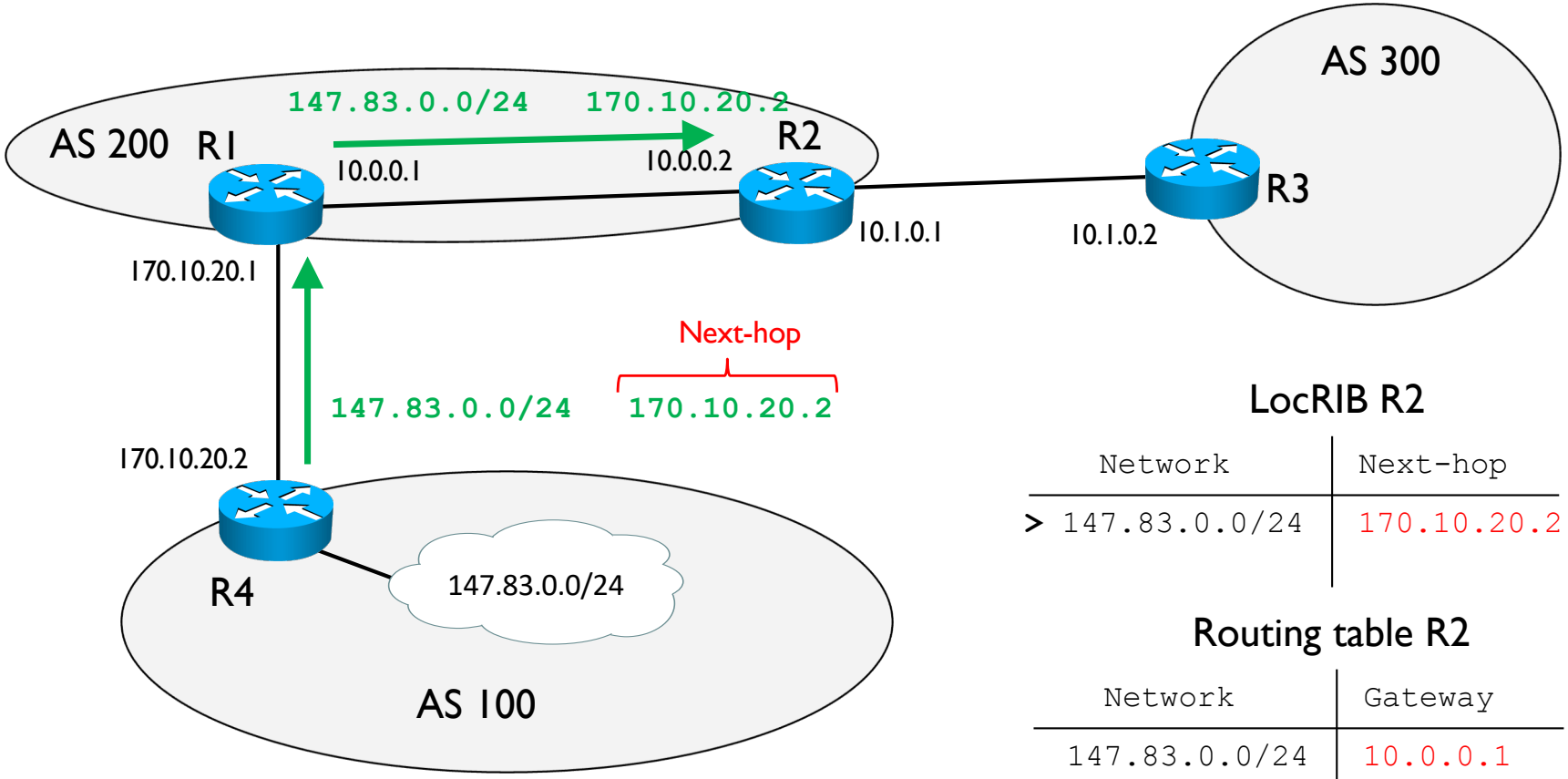
Routing table R1

Network	Gateway
147.83.0.0/24	170.10.20.2

5.7 – Atributos estándares

NEXT-HOP

- No confundir gateway de una tabla de encaminamiento con next-hop
- Gateway es a nivel de routers, next-hop a nivel de AS



LocRIB R2

Network	Next-hop
> 147.83.0.0/24	170.10.20.2

Routing table R2

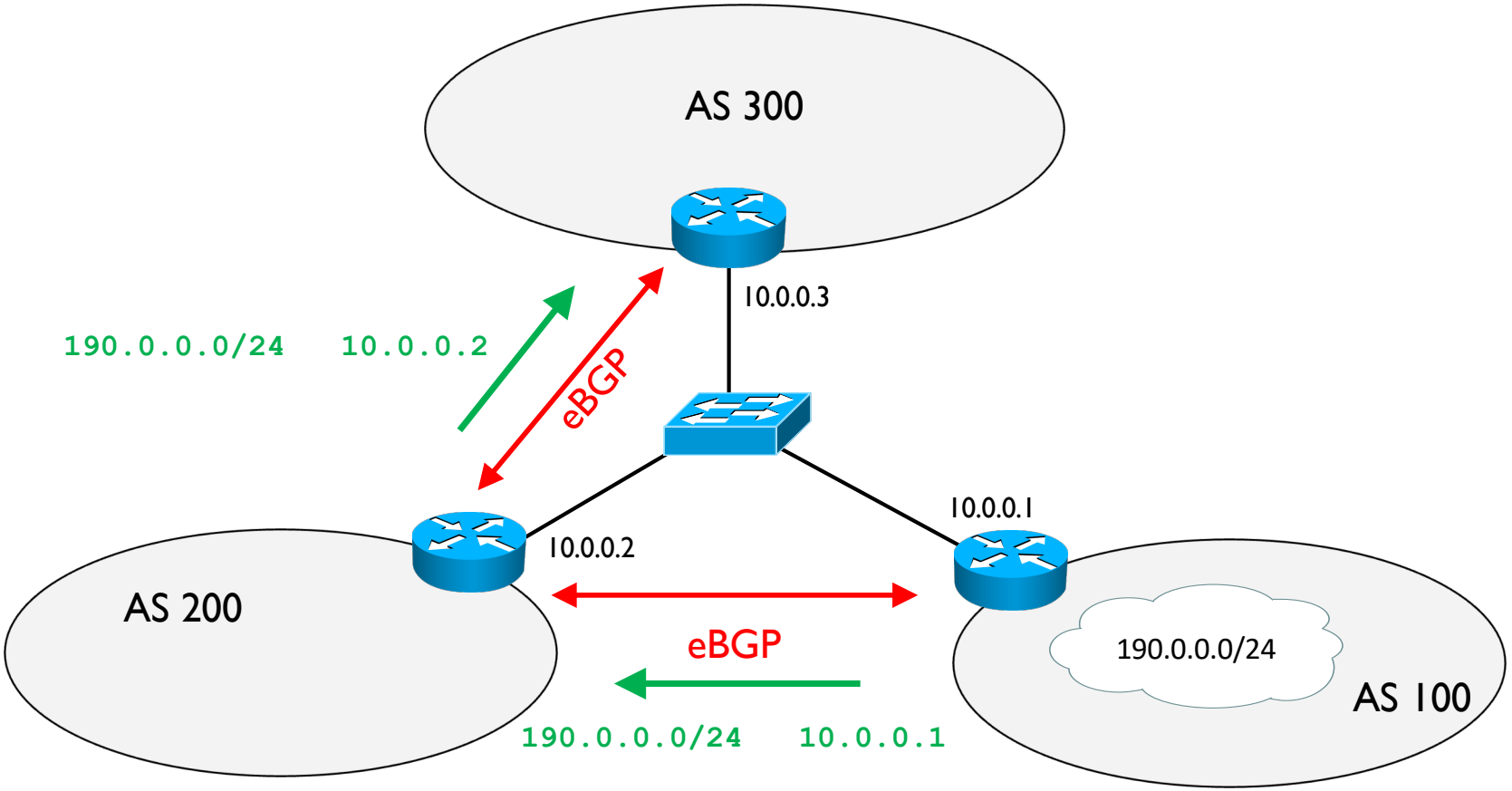
Network	Gateway
147.83.0.0/24	10.0.0.1

Para R2, gateway y next-hop coinciden

5.7 – Atributos estándares

NEXT-HOP third-party

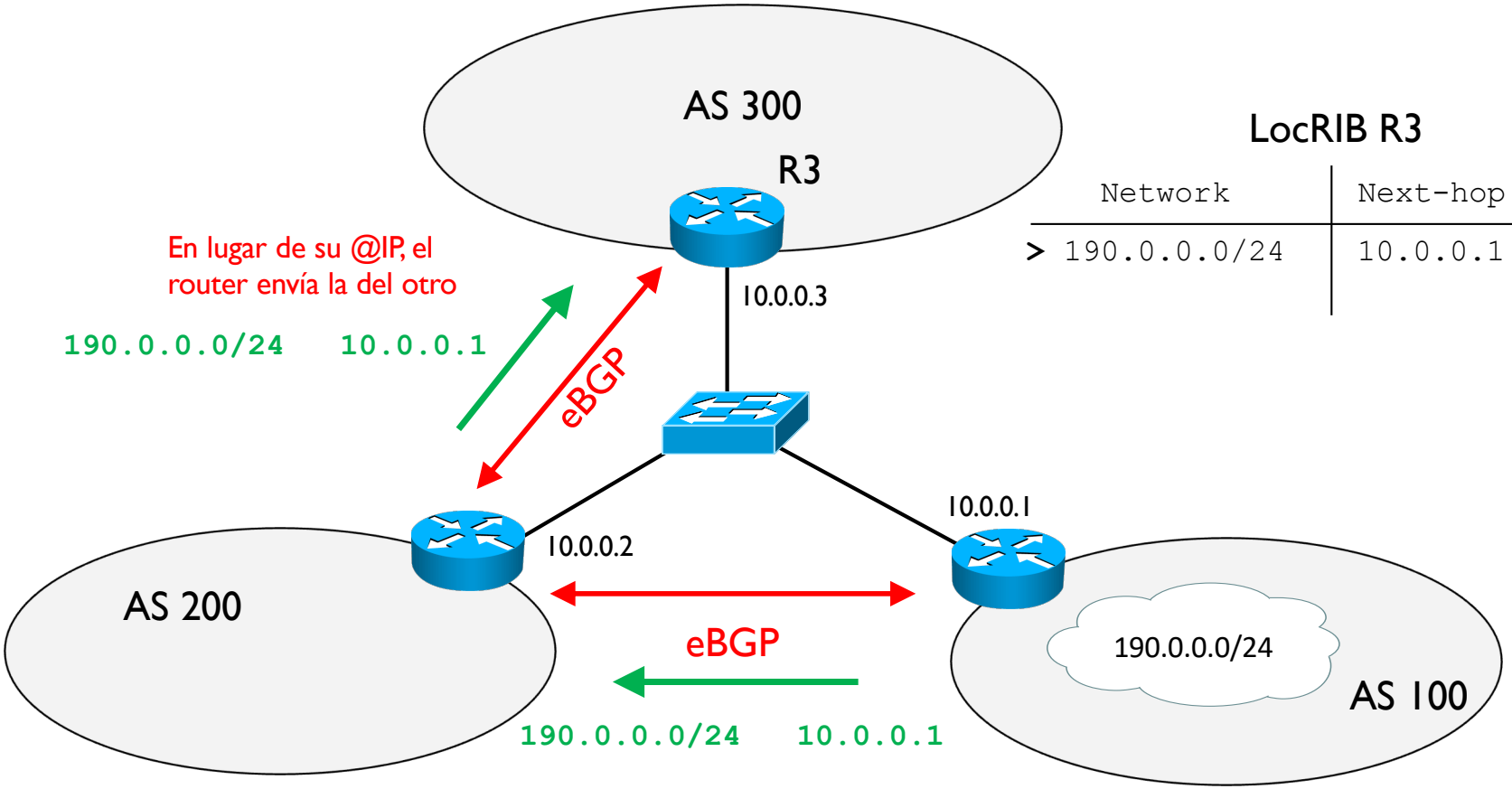
- Se usa para evitar procesar información inútilmente



5.7 – Atributos estándares

NEXT-HOP third-party

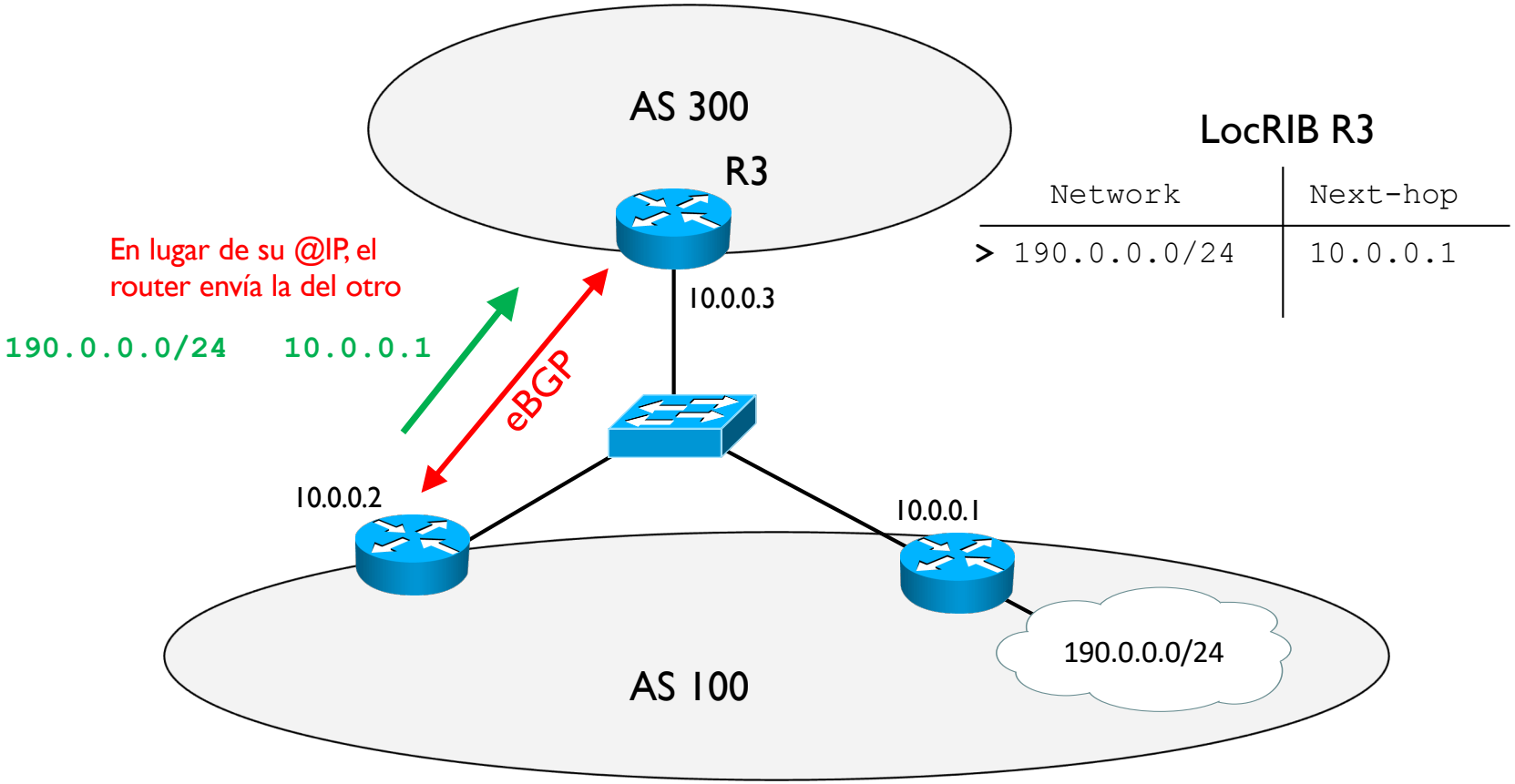
- Se usa para evitar procesar información inútilmente



5.7 – Atributos estándares

NEXT-HOP backdoor

- Se usa para desacoplar el router BGP del que procesa datagramas



5.7 – Atributos estándares
ORIGIN

- Obligatorio e histórico
- Determina como se ha aprendido un prefijo (el origen del prefijo)
  - IGP: aprendido a partir de un encaminamiento interno dinámico como RIP o OSPF
  - EGP: aprendido del protocolo EGP (protocolo externo para inter-AS que se usaba al principio conjuntamente con BGP, hoy en día ya no se usa)
  - Incompleto: aprendido de otro protocolo o el origen no se quiere anunciar (generalmente usado cuando se ha aprendido un prefijo de una ruta estática)

- En CISCO, suele aparecer en la tabla Loc_RIB al final del AS-path
  - IGP: se indica con i
  - EGP: se indica con e
  - Incompleto: se indica con ?

5.7 – Atributos estándares
AGGREGATOR

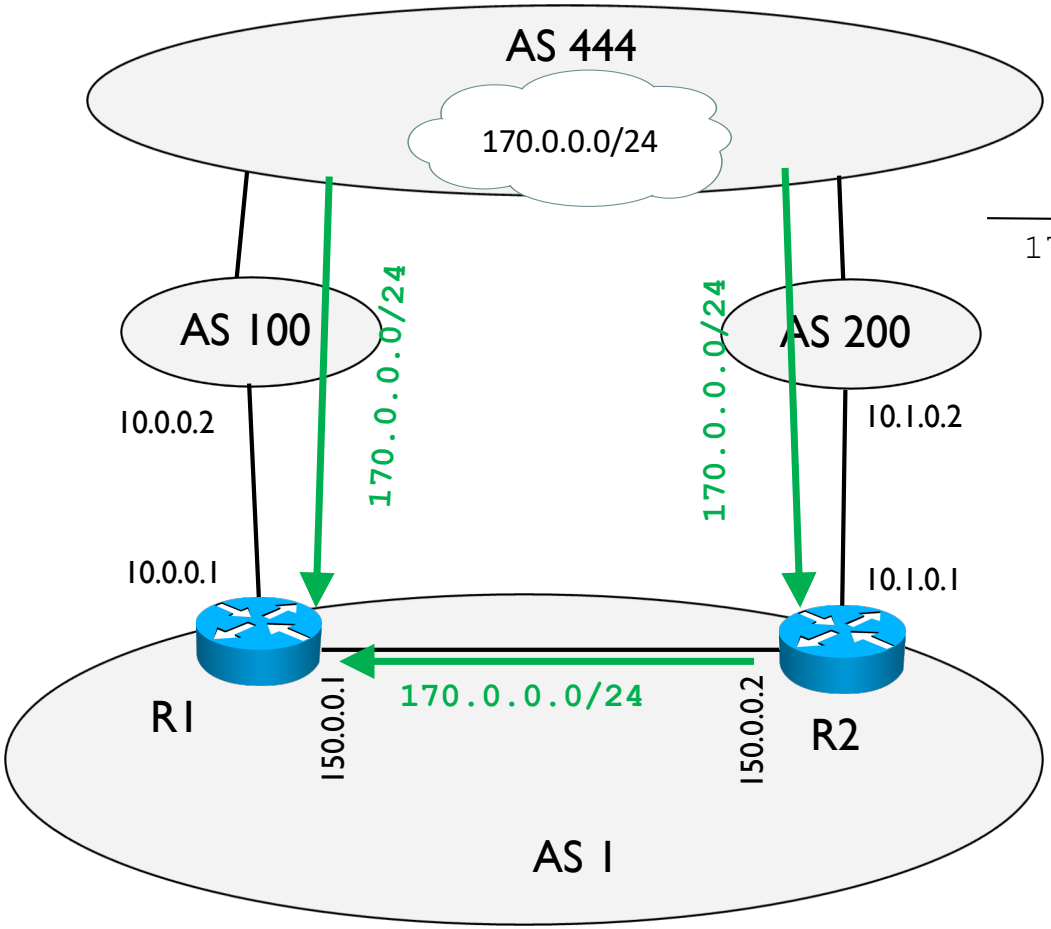
- Opcional
- Cuando un router BGP agrega prefijos, se puede usar este atributo para indicar en que AS se ha hecho esta agregación y que router RID lo ha hecho
- Es optativo notificar esta atributo, es decir se puede hacer la agregación y no usar el atributo para notificarlo a los demás routers

5.7 – Atributos estándares
LOCAL PREFERENCE

- Opcional
- Si no se usa, tiene 100 como valor por defecto
- Se usa para manipular la selección del mejor camino
- Se elige la ruta con el local preference más alto
- Los local preference tienen significado local, es decir interno al AS
  - No se anuncian por eBGP
  - Se anuncian por iBGP

5.7 – Atributos estándares
LOCAL PREFERENCE

- Ejemplo



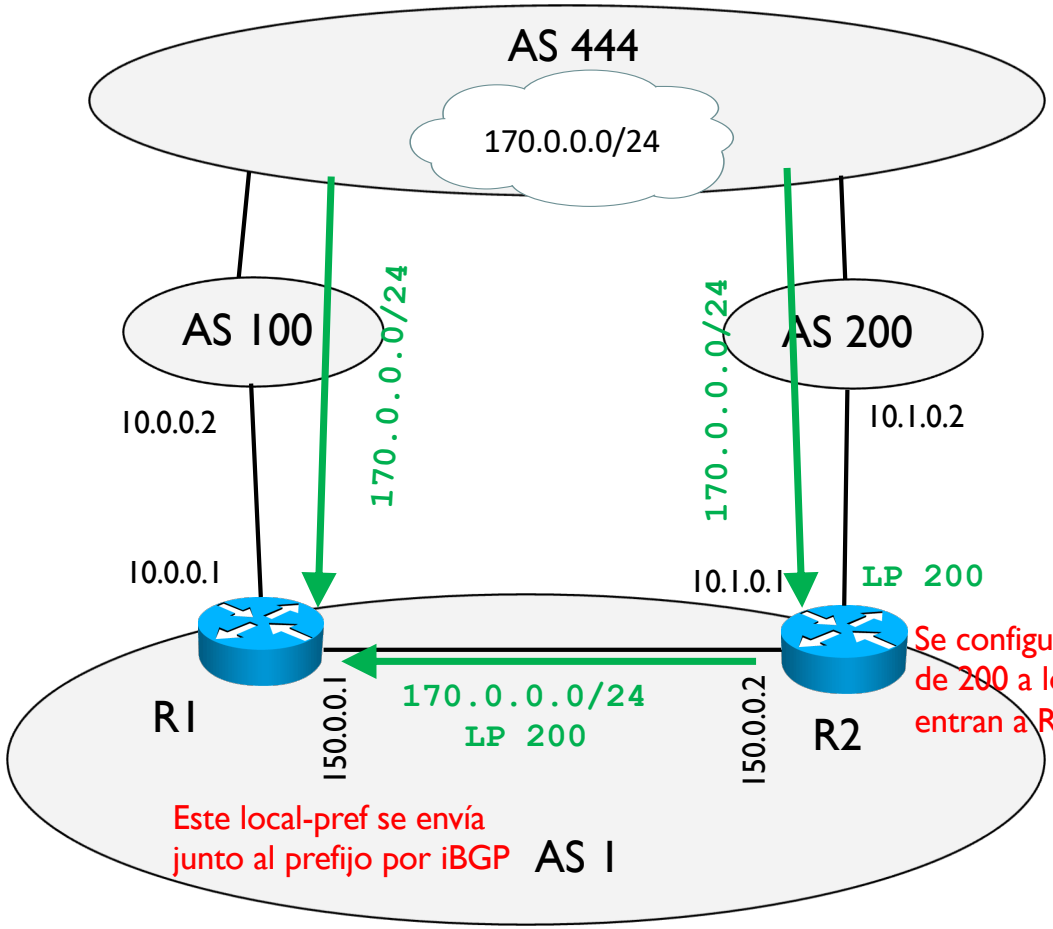
LocRIB R1

Network	Next-hop	Local-pref	AS-path
170.0.0.0/24	10.0.0.2	100	100 444
	10.1.0.2	100	200 444

¿Cuál es mejor?

5.7 – Atributos estándares
LOCAL PREFERENCE

- Ejemplo



Se configura un local-pref de 200 a los prefijos que entran a R2 por AS200

LocRIB R1

	Network	Next-hop	Local-pref
	170.0.0.0/24	10.0.0.2	100
>		10.1.0.2	200

R1 elige la ruta por R2 que va al AS200

Routing table R1

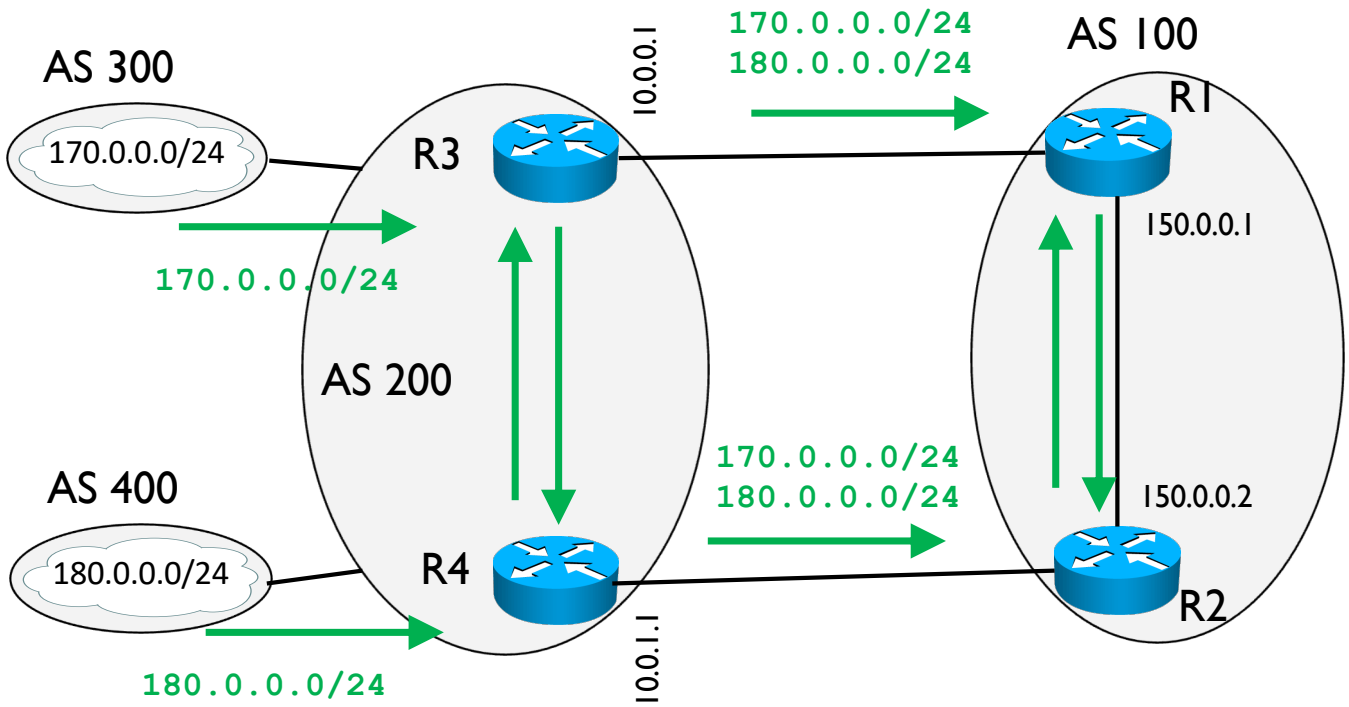
Network	Gateway
170.0.0.0/24	150.0.0.2

5.7 – Atributos estándares
MULTI EXIT DISCRIMINATOR (MED)

- Opcional
- Si no se usa, tiene 0 como valor por defecto
- Su valor se llama metric
- Se elige la ruta con el metric más bajo
- Un AS puede indicar al AS vecino cual enlace sería mejor usar entre varios disponibles
  - Libertad por parte del vecino de usar o rechazar esta sugerencia
- Los metric solo se transmiten entre dos AS vecinos

5.7 – Atributos estándares
MULTI EXIT DISCRIMINATOR (MED)

- Ejemplo

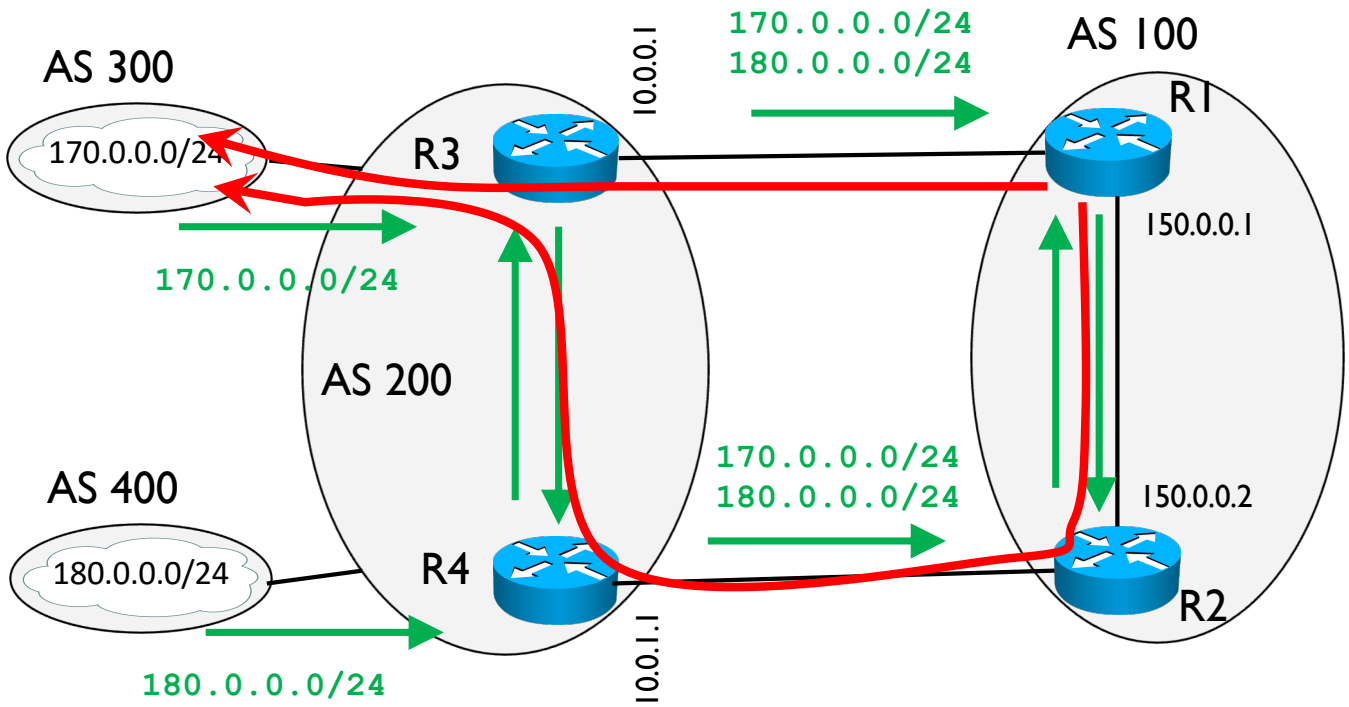


LocRIB R1

Network	Next-hop	metric
170.0.0.0/24	10.0.0.1	0
	10.0.1.1	0
180.0.0.0/24	10.0.0.1	0
	10.0.1.1	0

5.7 – Atributos estándares
MULTI EXIT DISCRIMINATOR (MED)

- Ejemplo



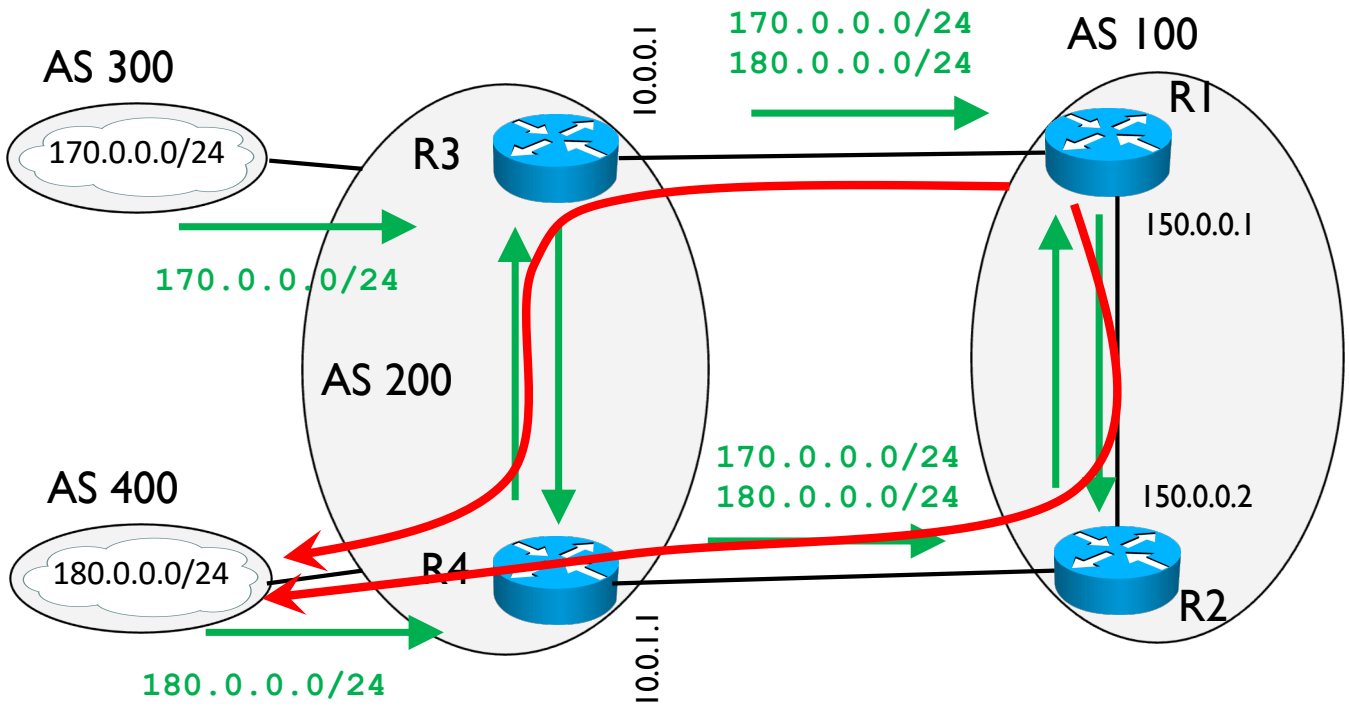
LocRIB R1

Network	Next-hop	metric
170.0.0.0/24	10.0.0.1	0
	10.0.1.1	0
180.0.0.0/24	10.0.0.1	0
	10.0.1.1	0

¿Cuál es mejor para llegar a 170.0.0.0/24?

5.7 – Atributos estándares
MULTI EXIT DISCRIMINATOR (MED)

- Ejemplo



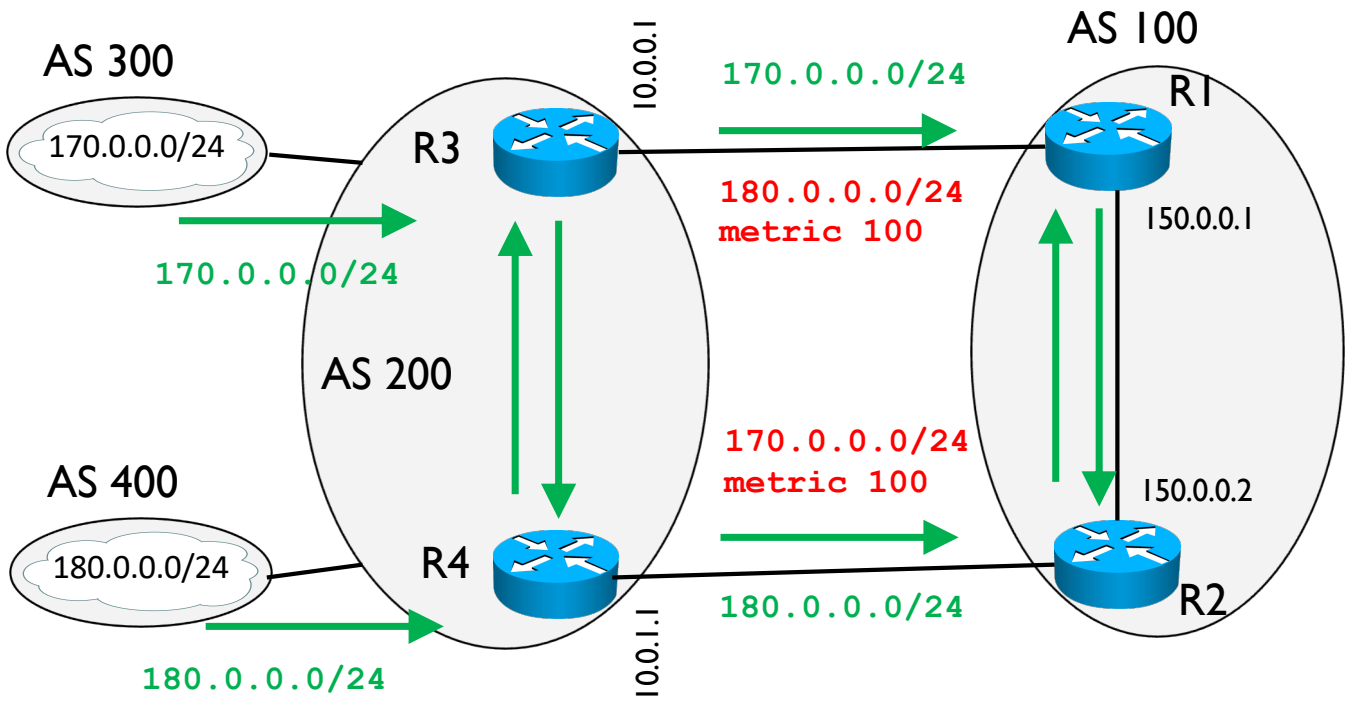
LocRIB R1

Network	Next-hop	metric
170.0.0.0/24	10.0.0.1	0
	10.0.1.1	0
180.0.0.0/24	10.0.0.1	0
	10.0.1.1	0

¿Cuál es mejor para llegar a 180.0.0.0/24?

5.7 – Atributos estándares
MULTI EXIT DISCRIMINATOR (MED)

- Ejemplo



El AS200 puede sugerir al AS100 las mejores rutas

En R3, se configura un metric de 100 al prefijo 180.0.0.0/24 de salida hacia R1

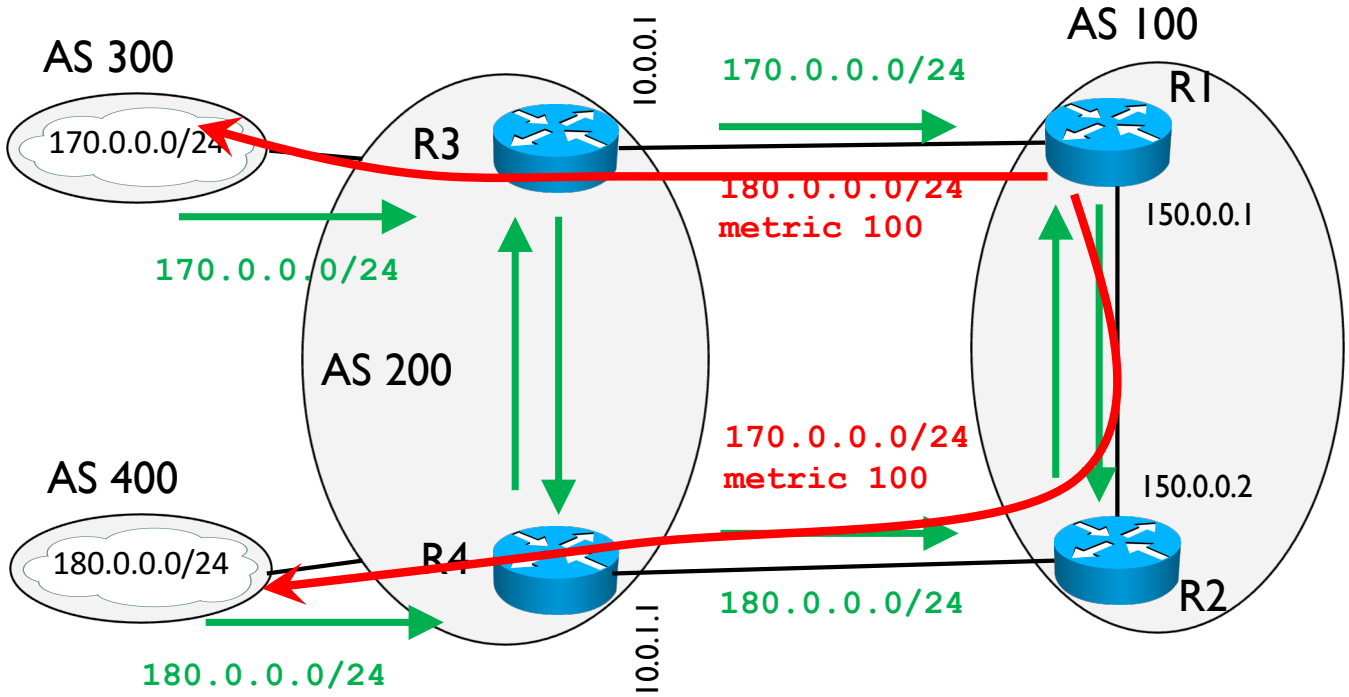
En R4, se configura un metric de 100 al prefijo 170.0.0.0/24 de salida hacia R1

LocRIB R1

	Network	Next-hop	metric
>	170.0.0.0/24	10.0.0.1	0
		10.0.1.1	100
	180.0.0.0/24	10.0.0.1	100
>		10.0.1.1	0

5.7 – Atributos estándares
MULTI EXIT DISCRIMINATOR (MED)

- Ejemplo



R1 elige las rutas con menor metric

LocRIB R1

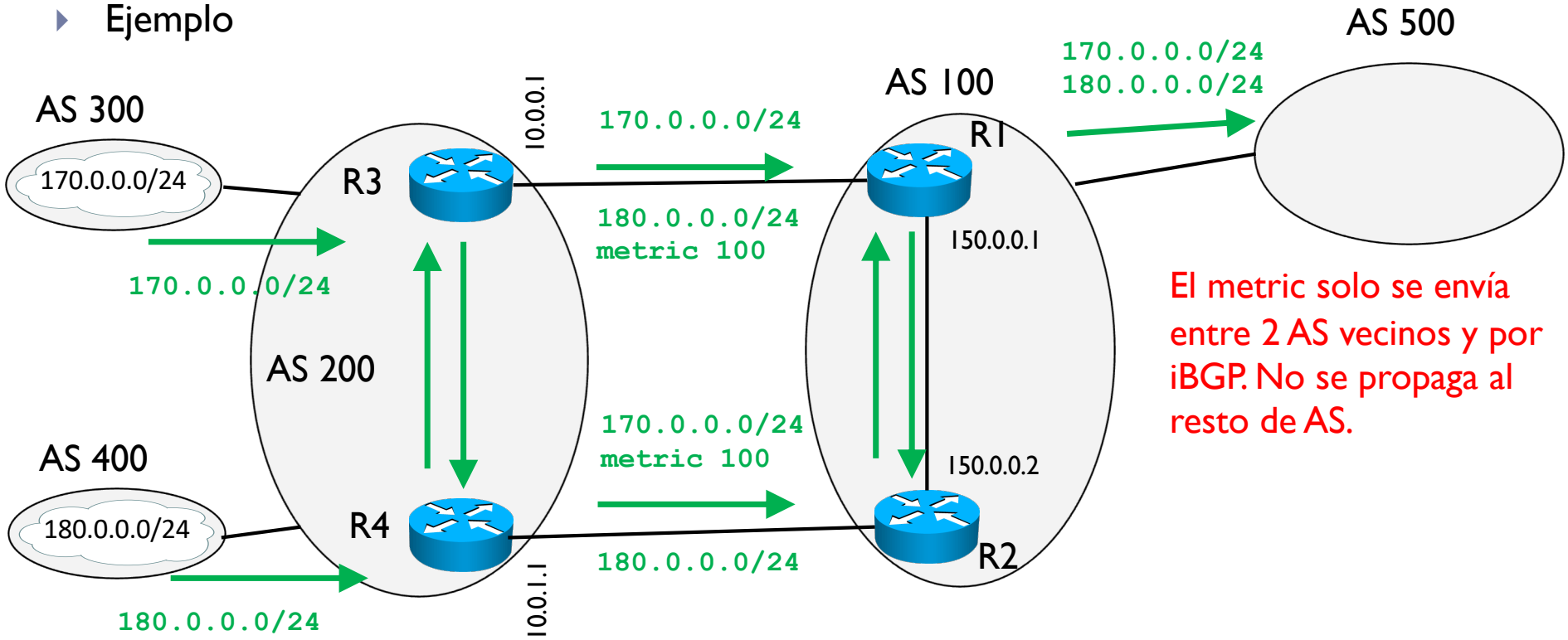
	Network	Next-hop	metric
>	170.0.0.0/24	10.0.0.1	0
		10.0.1.1	100
	180.0.0.0/24	10.0.0.1	100
>		10.0.1.1	0

Routing table R1

Network	Gateway
170.0.0.0/24	10.0.0.1
180.0.0.0/24	150.0.0.2

5.7 – Atributos estándares
MULTI EXIT DISCRIMINATOR (MED)

- Ejemplo



El metric solo se envía entre 2 AS vecinos y por iBGP. No se propaga al resto de AS.

LocRIB R1

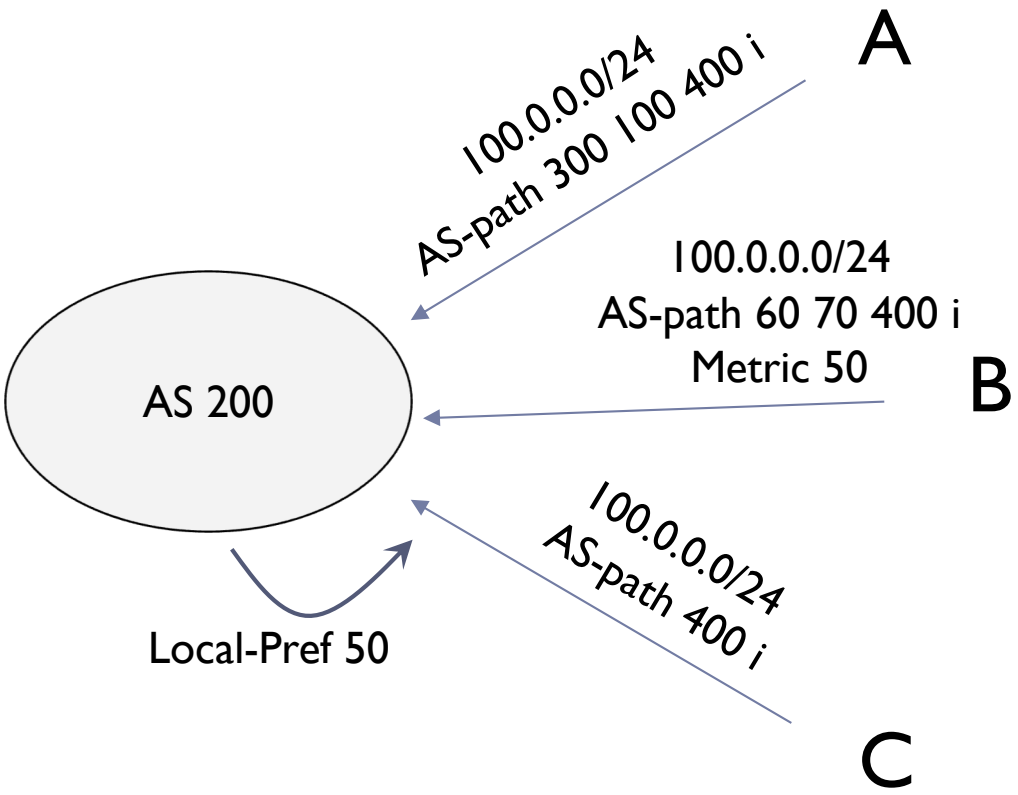
	Network	Next-hop	metric
>	170.0.0.0/24	10.0.0.1	0
		10.0.1.1	100
	180.0.0.0/24	10.0.0.1	100
>		10.0.1.1	0

5. Encaminamiento inter-dominio

1. Introducción
2. Encaminamiento path-vector
3. Funcionamiento de BGP
4. Establecimiento de una sesión BGP
5. Bases de datos BGP
6. Mensajes BGP
7. Atributos estándares
8. Algoritmo de selección de rutas
9. Escenarios comunes
10. Mejoras del BGP

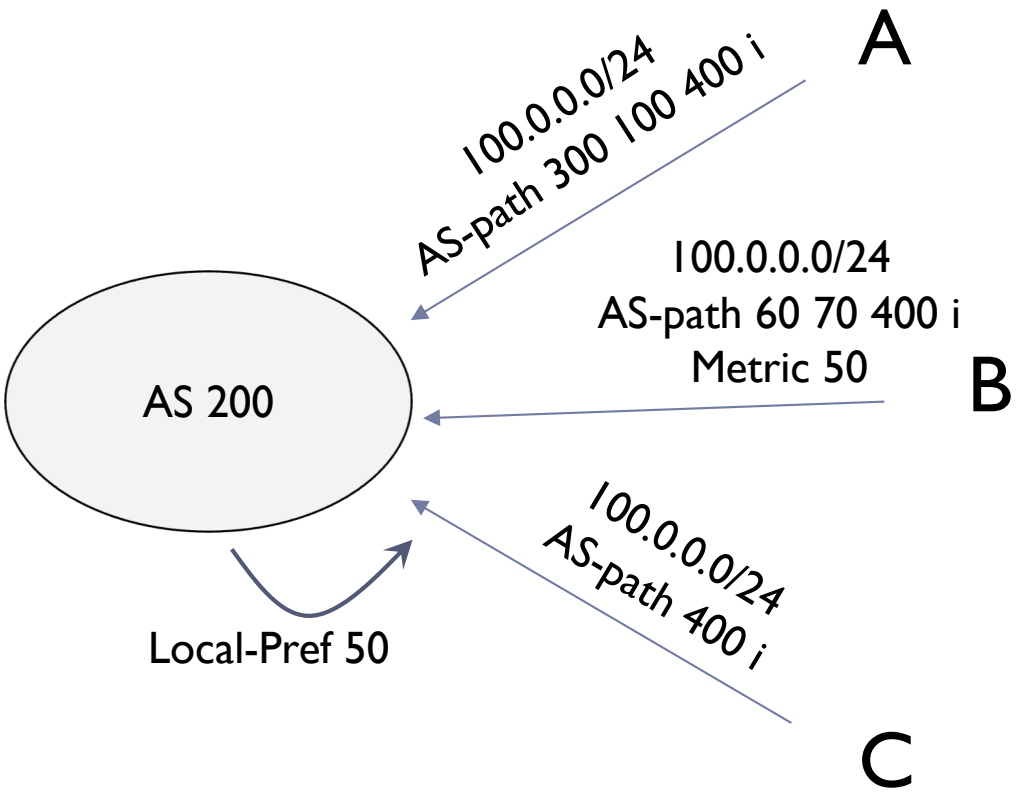
5.8 - Algoritmo de selección de rutas

- ¿Que ruta elige AS200?



5.8 - Algoritmo de selección de rutas

- ¿Que ruta elige AS200?



Si es por AS-path menor
→ Gana C

Si es por Metric menor
→ Gana A o C

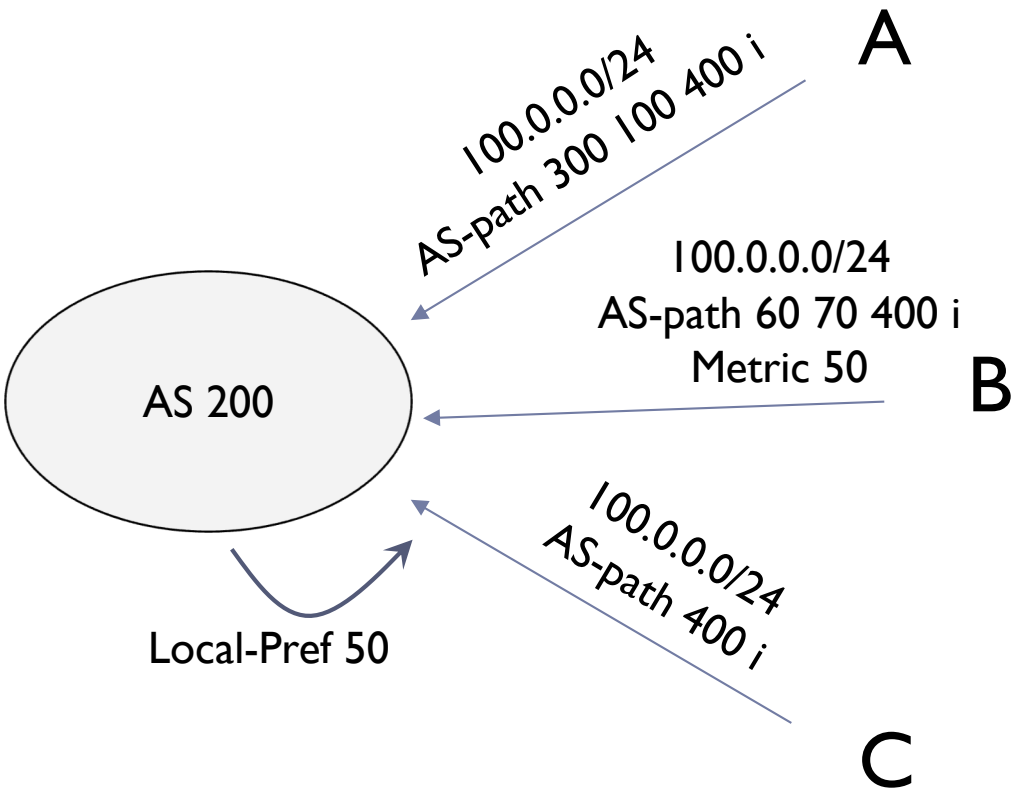
Si es por Local-Pref mayor
→ Gana A o B

5.8 - Algoritmo de selección de rutas

1) Ruta con mayor local preference
2) Ruta con menor AS-path
3) Ruta con este orden de ORIGEN
   IGP > EGP > incompleto
4) Ruta con menor Metric
5) …

5.8 - Algoritmo de sele

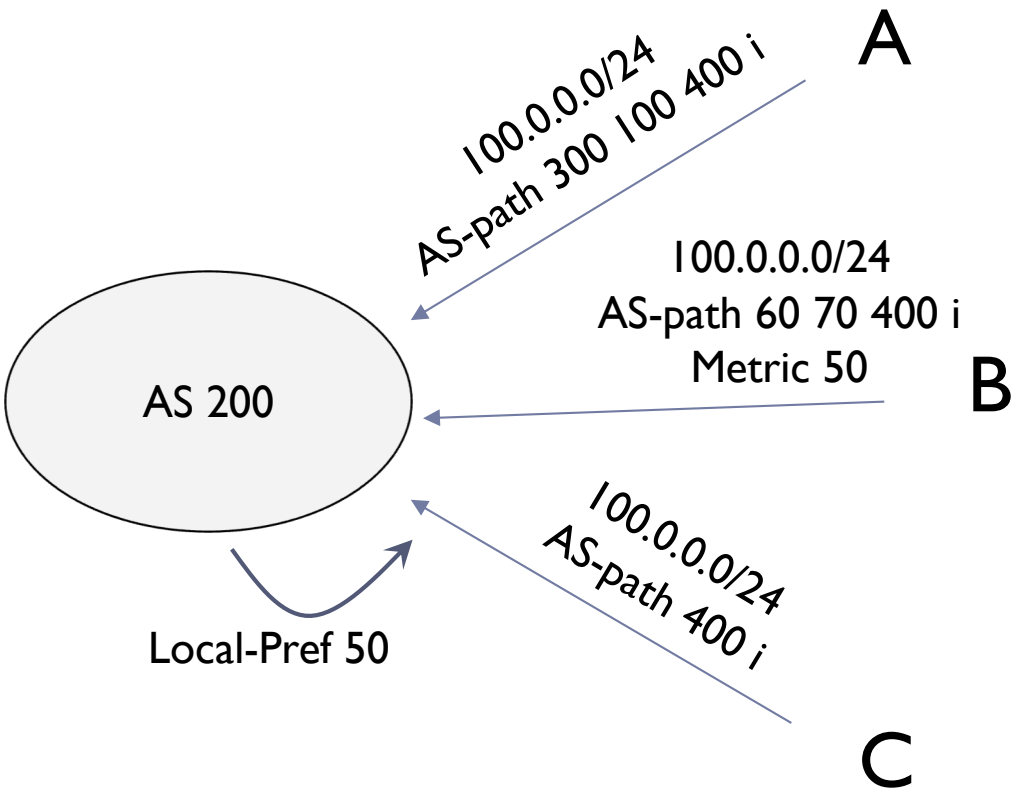
- ¿Que ruta elige AS200?



1) Ruta con mayor local preference
2) Ruta con menor AS-path
3) Ruta con este orden de ORIGEN
   IGP > EGP > incompleto
4) Ruta con menor Metric

5.8 - Algoritmo de sele

- ¿Que ruta elige AS200?

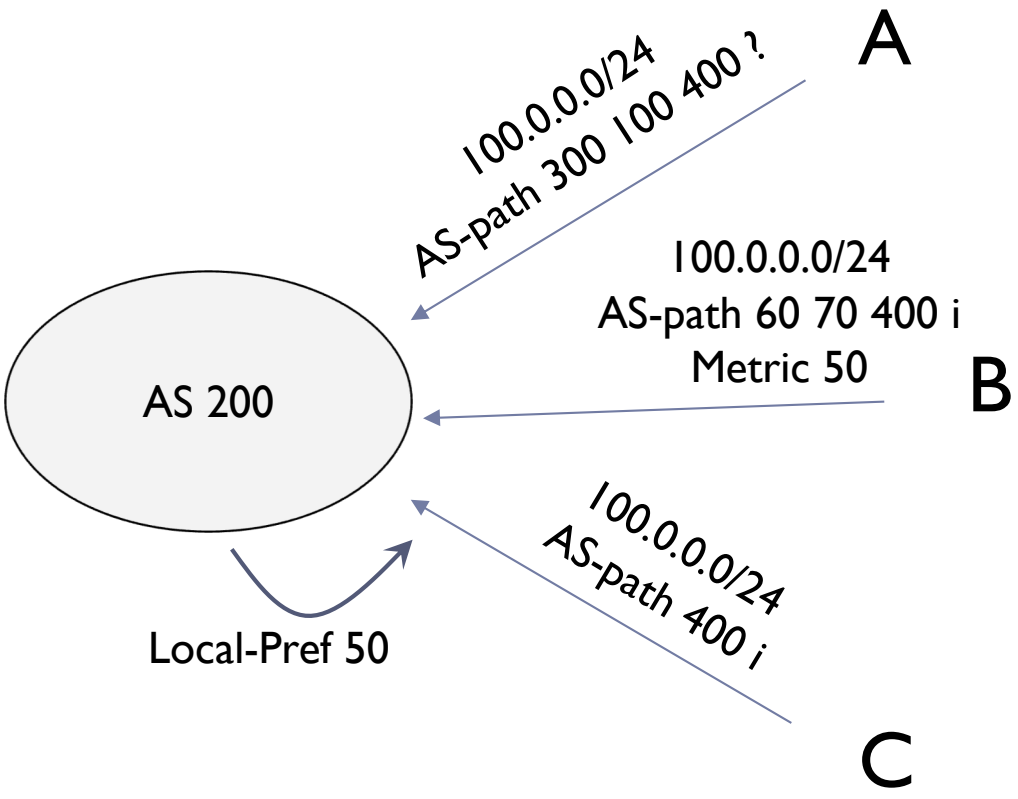


1) Ruta con mayor local preference
2) Ruta con menor AS-path
3) Ruta con este orden de ORIGEN
   IGP > EGP > incompleto
4) Ruta con menor Metric

Por el criterio 1)
Se excluye C
Por el criterio 2)
Empate entre A y B
Por el criterio 4)
Gana A

5.8 - Algoritmo de sele

- ¿Y si fuera así?

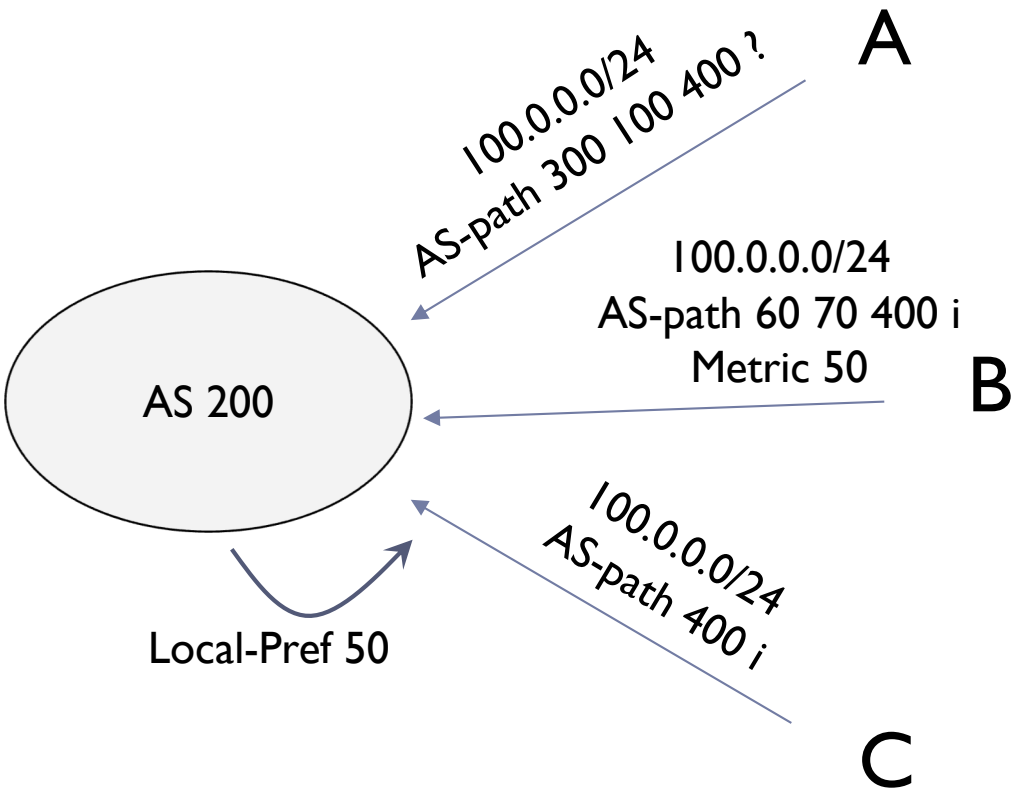


1) Ruta con mayor local preference
2) Ruta con menor AS-path
3) Ruta con este orden de ORIGEN
   IGP > EGP > incompleto
4) Ruta con menor Metric

5.8 - Algoritmo de sele

1) Ruta con mayor local preference
2) Ruta con menor AS-path
3) Ruta con este orden de ORIGEN
   IGP > EGP > incompleto
4) Ruta con menor Metric

- ¿Y si fuera así?



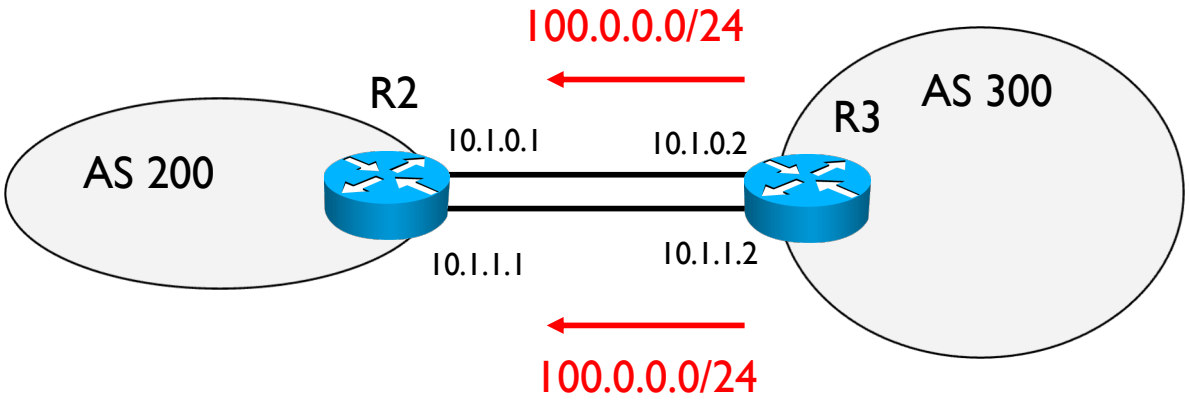
Por el criterio 1)
Se excluye C
Por el criterio 2)
Empate entre A y B
Por el criterio 3)
Gana B

5.8 - Algoritmo de selección de rutas

1) Ruta con mayor local preference
2) Ruta con menor AS-path
3) Ruta con este orden de ORIGEN
   IGP > EGP > incompleto
4) Ruta con menor Metric
5) Ruta aprendida por eBGP antes que por iBGP
6) Si dos o más rutas por iBGP, la que tiene menor coste IGP
7) Ruta aprendida antes (más antigua)
8) Ruta hacia el router con menor RID
9) … falta una … ¿qué más hay?

5.8 - Algoritmo de selección de rutas

- ¿Que ruta elige AS200?



5.8 - Algoritmo de selección de rutas

1) Ruta con mayor local preference
2) Ruta con menor AS-path
3) Ruta con este orden de ORIGEN
   IGP > EGP > incompleto
4) Ruta con menor Metric
5) Ruta aprendida por eBGP antes que por iBGP
6) Si dos o más rutas por iBGP, la que tiene menor coste IGP
7) Ruta aprendida antes (más antigua)
8) Ruta hacia el router con menor RID
9) Ruta hacia la interfaz de un mismo router con menor @IP

5.8 - Algoritmo de selección de rutas
Políticas de encaminamiento

- Entre un AS y los demás AS vecinos hay relaciones bien definidas
- Estas relaciones definen
  - Si hay un vinculo de tipo peer-to-peer o customer-provider
  - A quien se le proporciona transito
  - Quien le proporciona transito
- Herramientas más comunes
  - Filtrado por prefijo
  - Filtrado por AS
  - Script de programación

Se aplican a los BGP updates

5.8 - Algoritmo de selección de rutas

Políticas usando script

- Permiten controlar/filtrar pero también modificar información de encaminamiento
- Construcción
  - Definir que es lo que se quiere filtrar para aplicarle una determinada política
  - Definir que se quieres hacer con lo que se ha seleccionado. Para eso se usa un if/else en programación pero usando comandos específicos del OS del router

```
if CONDITION then ACTION
else if CONDITION2 then ACTION2
else if CONDITION3 then ACTION3
…
else DEFAULT_ACTION
```

  - Aplicar a la interfaz del router y en la dirección deseada

5.8 - Algoritmo de selección de rutas

Políticas usando script

- Comandos CISCO para if/else

```
route-map SCRIPT_NAME permit LINE1
match CONDICTION1
set ACTION 1.1
set ACTION 1.2
...
set ACTION 1.n

route-map SCRIPT_NAME permit LINE2
match CONDICTION2
set ACTION 2.1
set ACTION 2.2
...
```

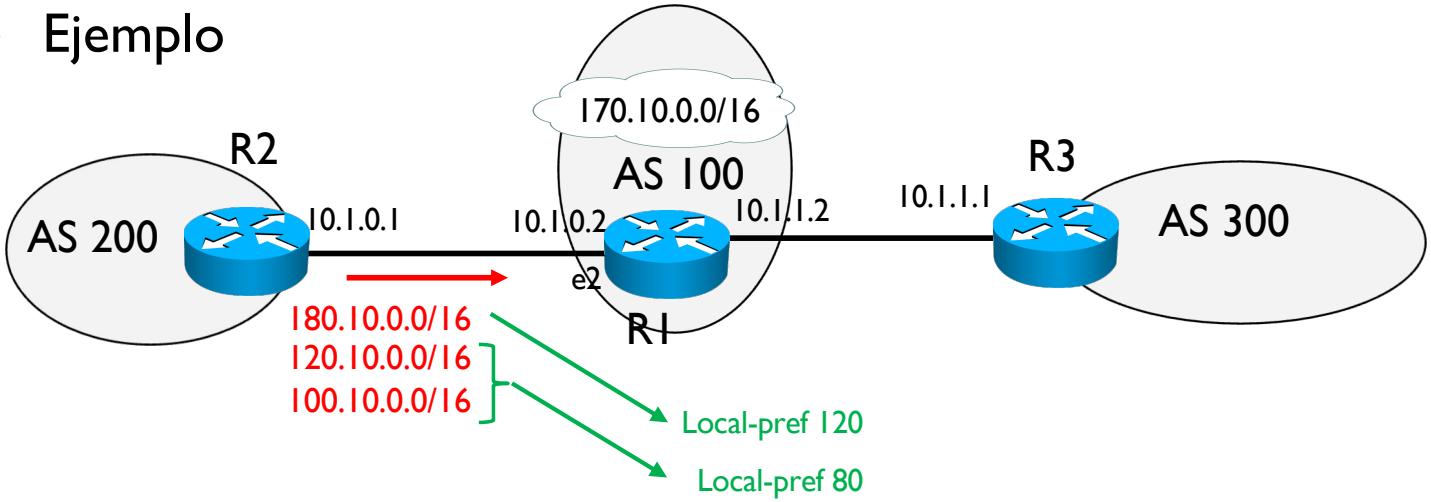
Números para indicar en que orden ejecutar estos if/else

Si no hay nada que diga el contrario, la regla por defecto es filtrar todo lo que no cumple con las condiciones anteriores

5.8 - Algoritmo de selección de rutas

Políticas usando script

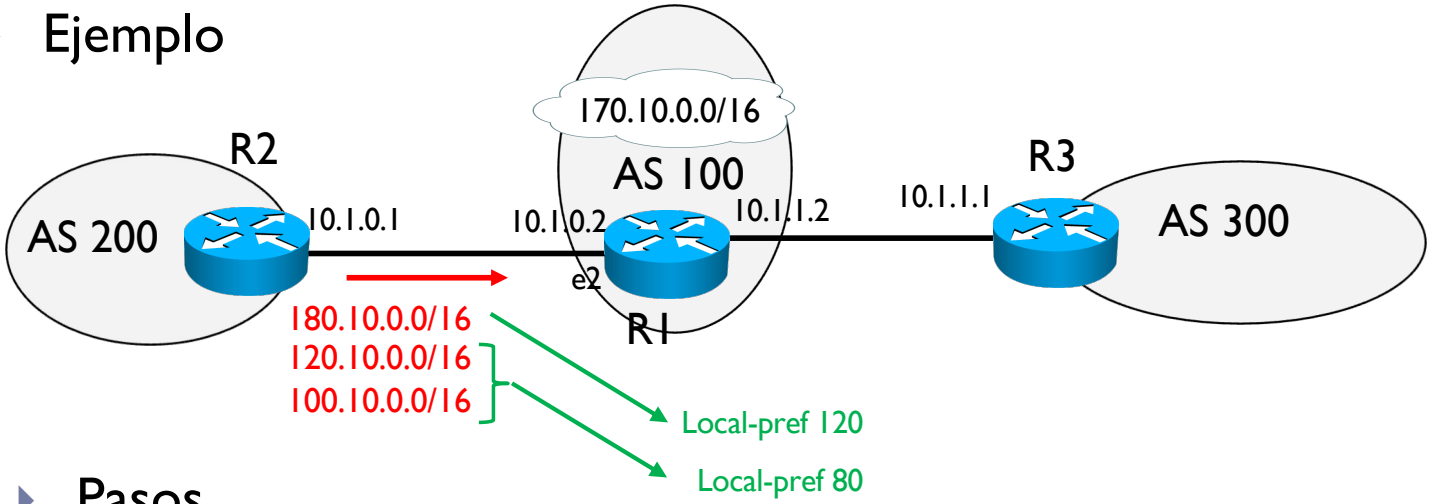
- Ejemplo



5.8 - Algoritmo de selección de rutas

Políticas usando script

- Ejemplo

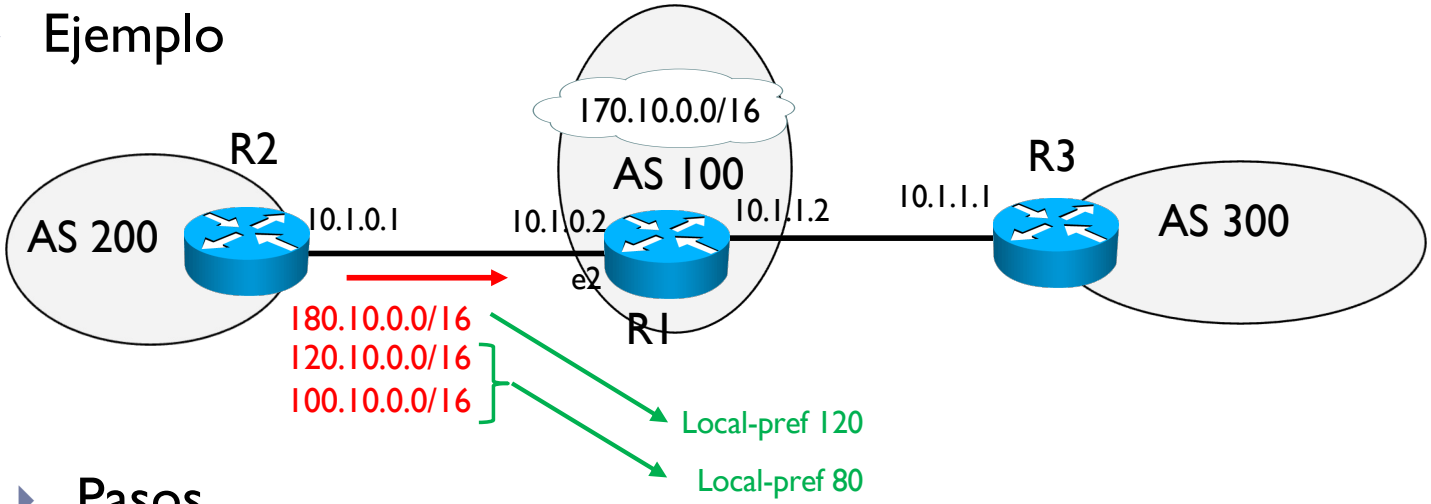


- Pasos
  - Definir que es lo que se quiere filtrar para aplicarle una determinada política
  - Definir que se quieres hacer con lo que se ha seleccionado. Para eso se usa un if/else en programación pero usando comandos específicos del OS del router
  - Aplicar a la interfaz del router y en la dirección deseada

5.8 - Algoritmo de selección de rutas

Políticas usando script

- Ejemplo



- Pasos
  - Definir que es lo que se quiere filtrar para aplicarle una determinada política

```
R1# access-list 1 permit 180.10.0.0/16
R1# access-list 2 permit 120.10.0.0/16
R1# access-list 2 permit 100.10.0.0/16
```

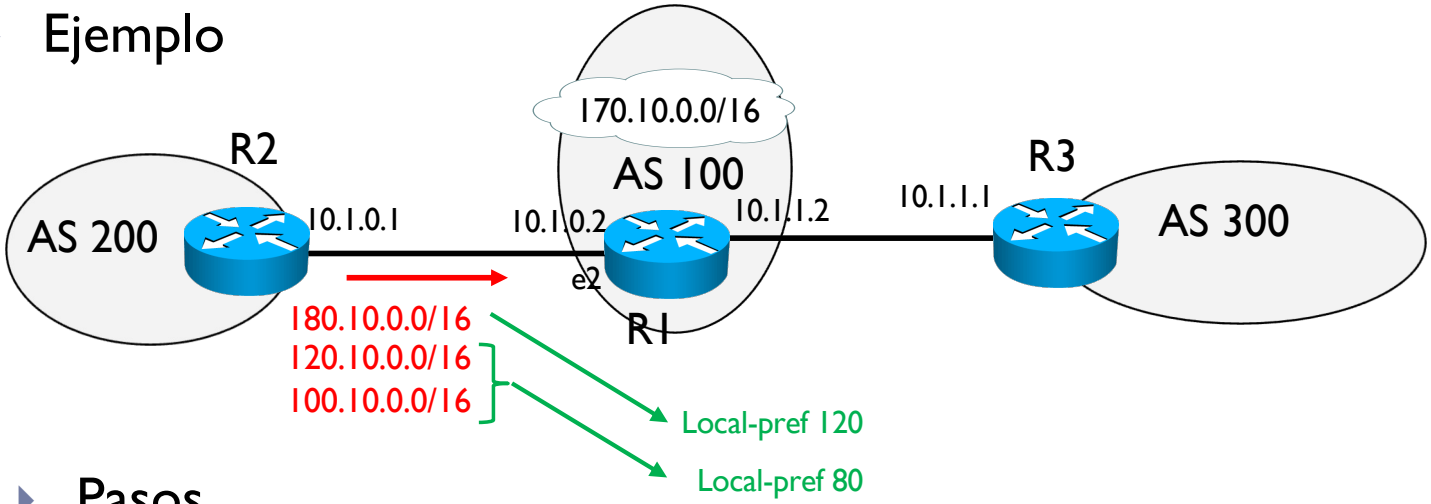
De todos los prefijos que llegan, la lista 1 selecciona este prefijo

De todos los prefijos que llegan, la lista 1 selecciona estos dos prefijos

5.8 - Algoritmo de selección de rutas

Políticas usando script

- Ejemplo



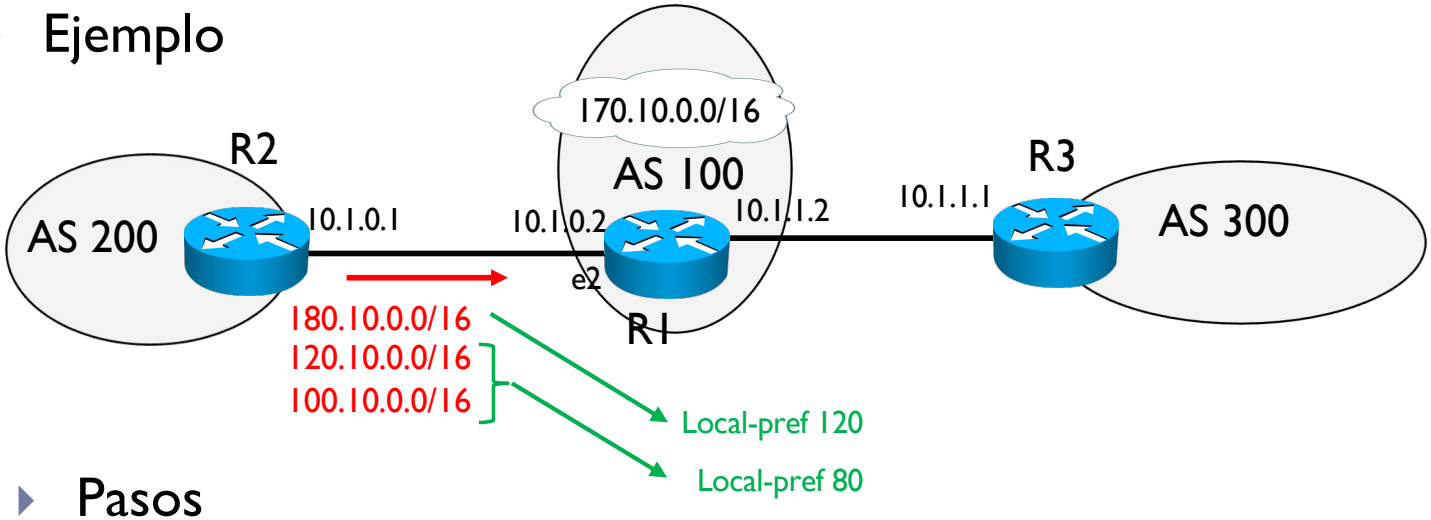
- Pasos
  - Definir que se quieres hacer con lo que se ha seleccionado

```
R1# route-map POL1 permit 10
R1# match ip address 1          → Lo que se selecciona de la lista 1
R1# set local-preference 120

R1# route-map POL1 permit 20
R1# match ip address 2          → Lo que se selecciona de la lista 2
R1# set local-preference 80
```

5.8 - Algoritmo de selección de rutas

Políticas usando script

- Ejemplo



- Pasos
  - Aplicar a la interfaz del router y en la dirección deseada

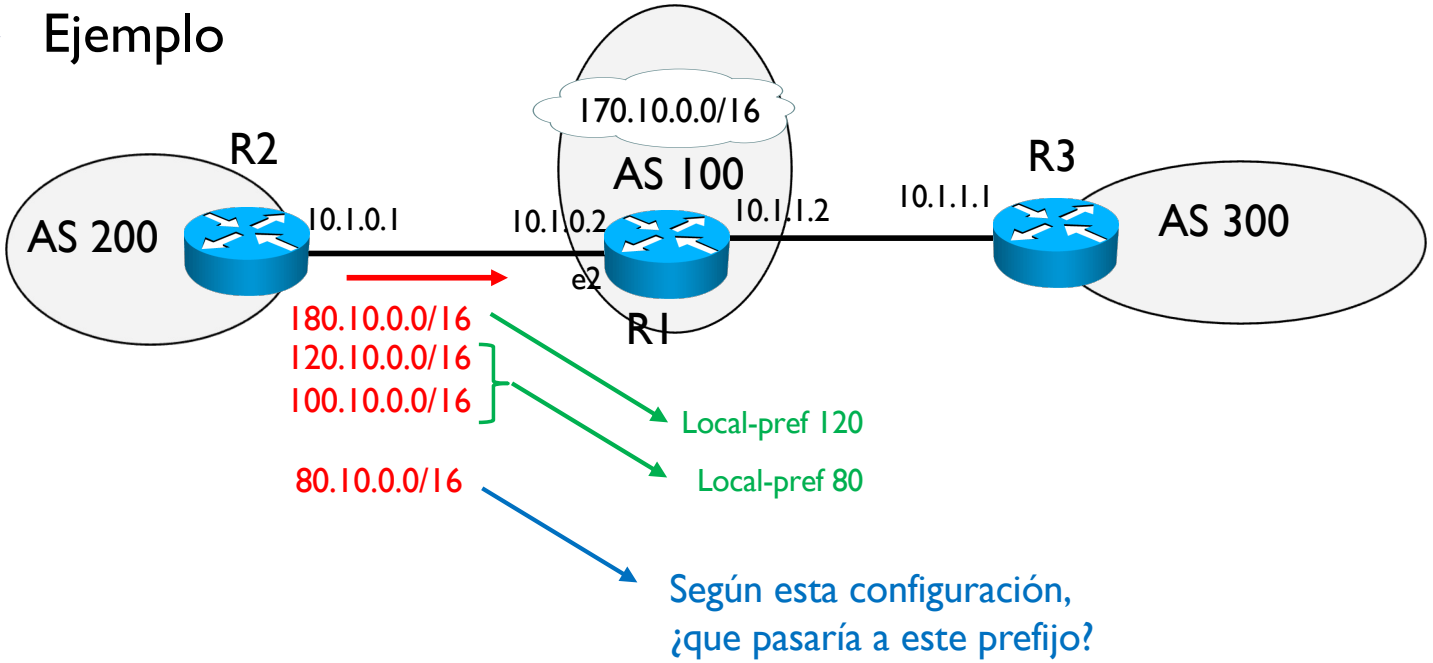
```
R1# router bgp 100
R1# neighbor 10.1.0.1 remote-as 200
R1# neighbor 10.1.0.1 route-map POL1 in
```

De todo lo que le envía R2, aplicamos el script POL1 en sentido de entrada respecto a R1

5.8 - Algoritmo de selección de rutas

Políticas usando script

- Ejemplo




```
R1# access-list 1 permit 180.10.0.0/16
R1# access-list 2 permit 120.10.0.0/16
R1# access-list 2 permit 100.10.0.0/16
```

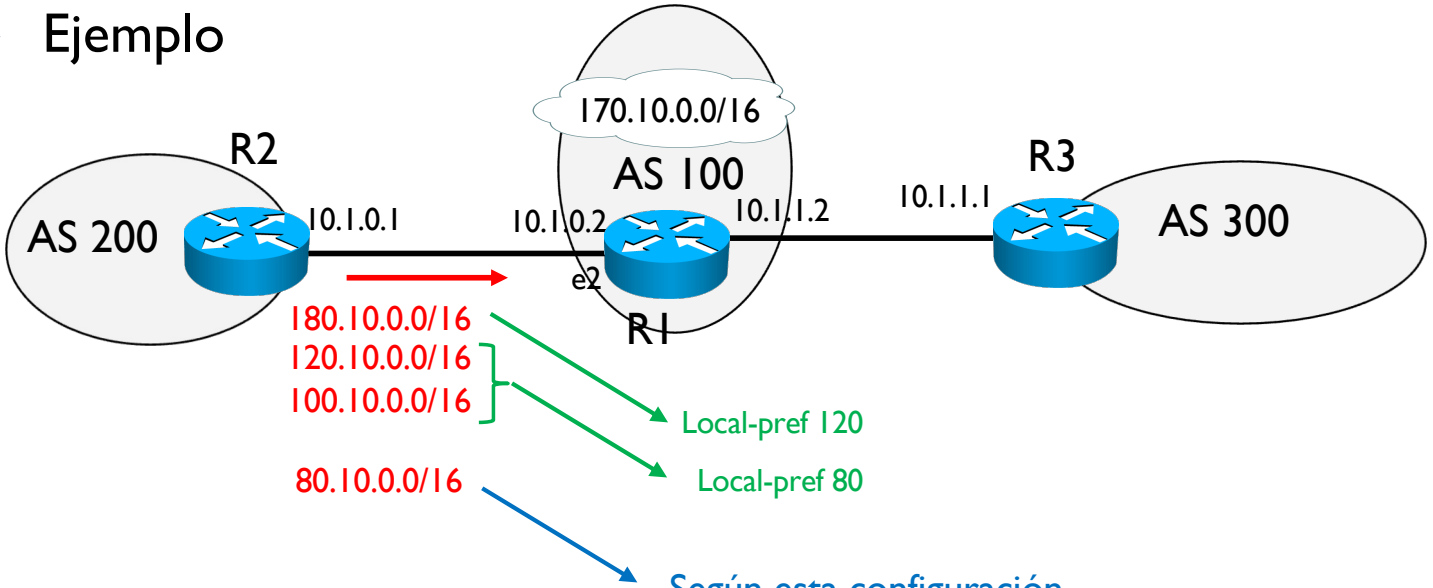
```
R1# router bgp 100
R1# neighbor 10.1.0.1 remote-as 200
R1# neighbor 10.1.0.1 route-map POL1 in
```

```
R1# route-map POL1 permit 10
R1# match ip address 1
R1# set local-preference 120

R1# route-map POL1 permit 20
R1# match ip address 2
R1# set local-preference 80
```

5.8 - Algoritmo de selección de rutas

Políticas usando script

- Ejemplo




```
R1# access-list 1 permit 180.10.0.0/16
R1# access-list 2 permit 120.10.0.0/16
R1# access-list 2 permit 100.10.0.0/16
```

```
R1# router bgp 100
R1# neighbor 10.1.0.1 remote-as 200
R1# neighbor 10.1.0.1 route-map POL1 in
```

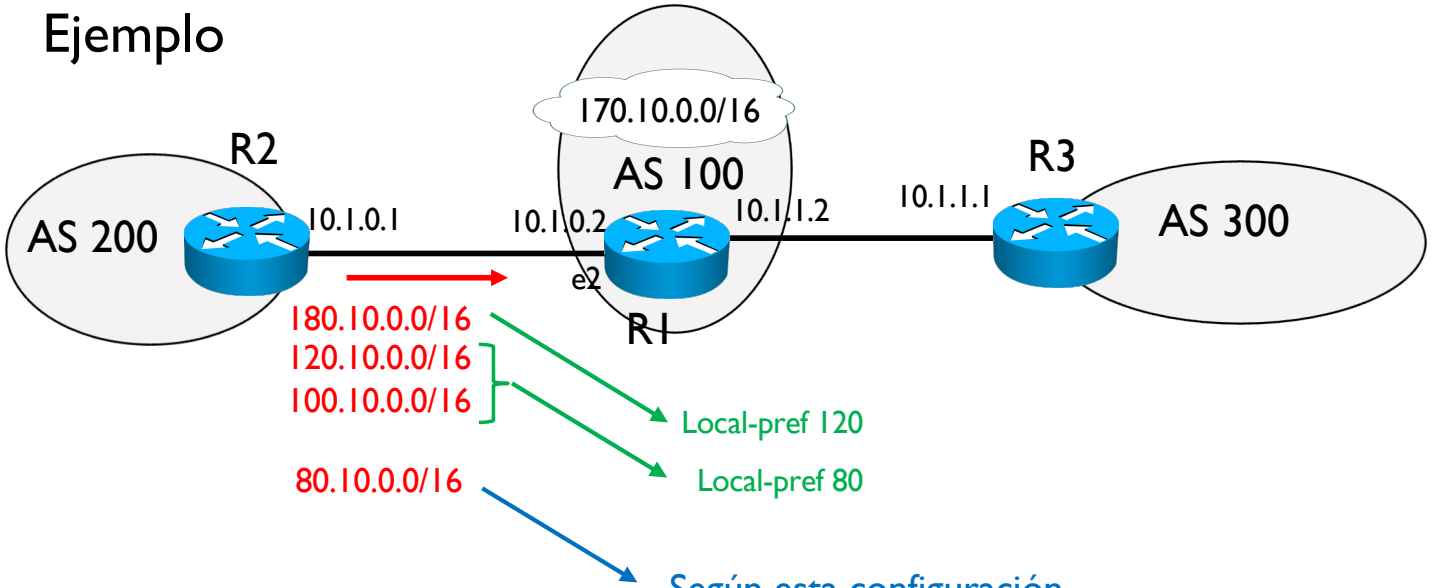
```
R1# route-map POL1 permit 10
R1# match ip address 1
R1# set local-preference 120

R1# route-map POL1 permit 20
R1# match ip address 2
R1# set local-preference 80
```

→ Hay un else final por defecto que filtra todo lo que queda

5.8 - Algoritmo de selección de rutas

Políticas usando script

- Ejemplo




```
R1# access-list 1 permit 180.10.0.0/16
R1# access-list 2 permit 120.10.0.0/16
R1# access-list 2 permit 100.10.0.0/16

R1# router bgp 100
R1# neighbor 10.1.0.1 remote-as 200
R1# neighbor 10.1.0.1 route-map POL1 in
```

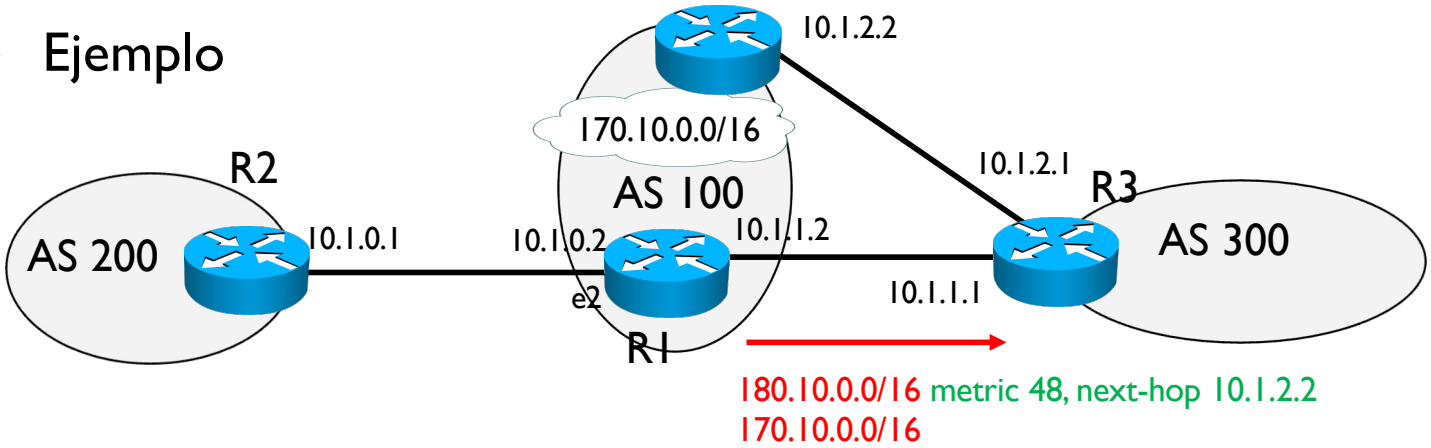
```
R1# route-map POL1 permit 10
R1# match ip address 1
R1# set local-preference 120

R1# route-map POL1 permit 20
R1# match ip address 2
R1# set local-preference 80

R1# route-map POL1 permit 1000
```

Esta línea para modificar esta regla por defecto y aceptar todo sin modificaciones

5.8 - Algoritmo de selección de rutas

Políticas usando script

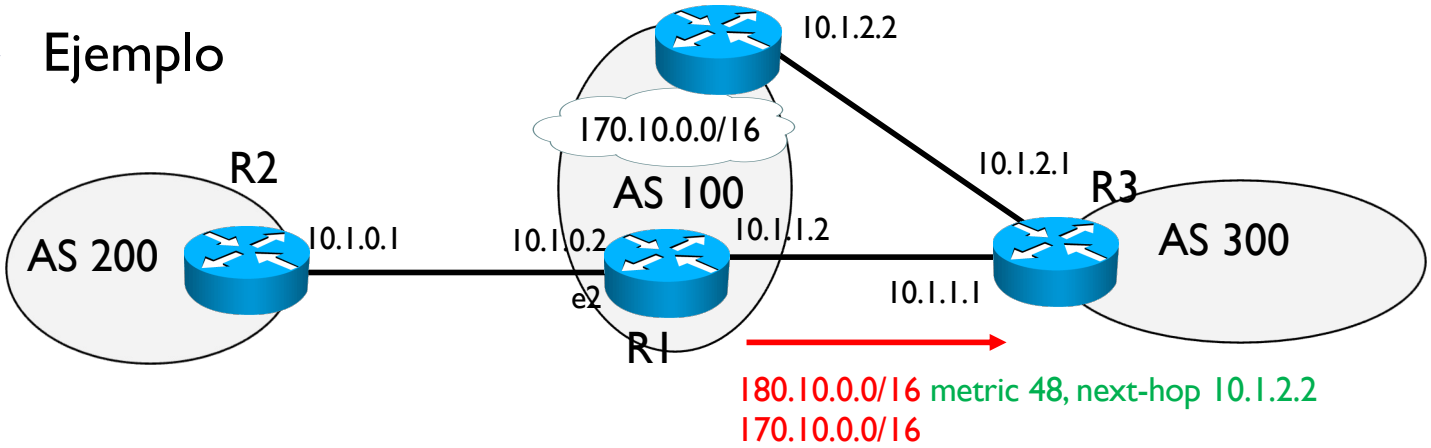
- Ejemplo



5.8 - Algoritmo de selección de rutas

Políticas usando script

- Ejemplo



- Pasos
  - Definir que es lo que se quiere filtrar para aplicarle una determinada política

```
R1# access-list 1 permit 180.10.0.0/16
R1# access-list 3 permit 170.10.0.0/16
```

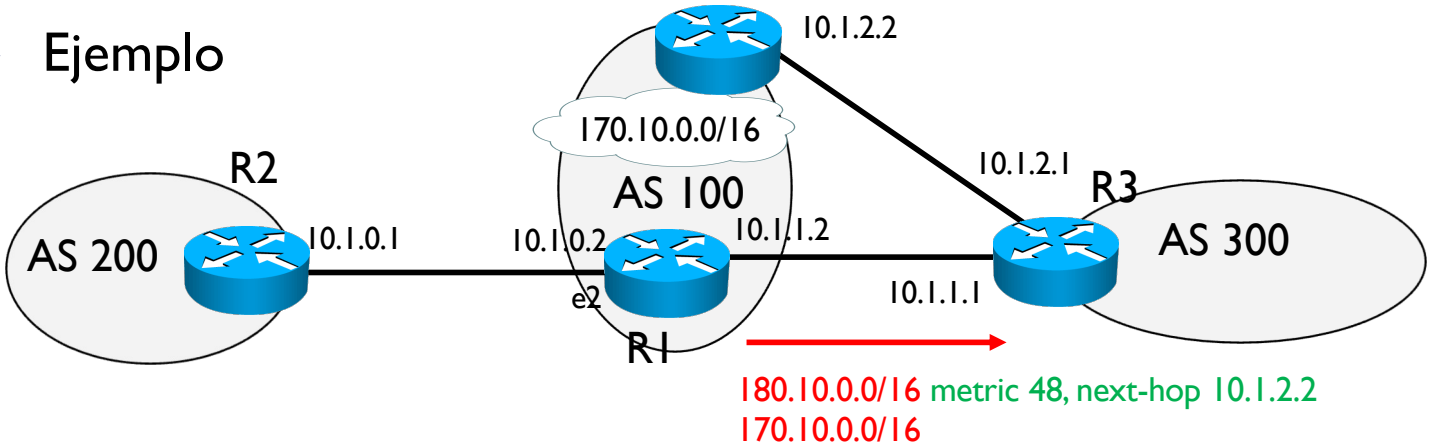
Ya existe de la política de entrada

Nueva lista

5.8 - Algoritmo de selección de rutas

Políticas usando script

- Ejemplo



- Pasos
  - Definir que se quieres hacer con lo que se ha seleccionado

Script diferente, otro nombre

```
R1# route-map POL2 permit 10
R1# match ip address 1
R1# set metric 48
R1# set next-hop 10.1.2.2

R1# route-map POL2 permit 20
R1# match ip address 3
```

Al prefijo de la lista 1 se aplica un metric de 48 y next hop 10.1.2.2

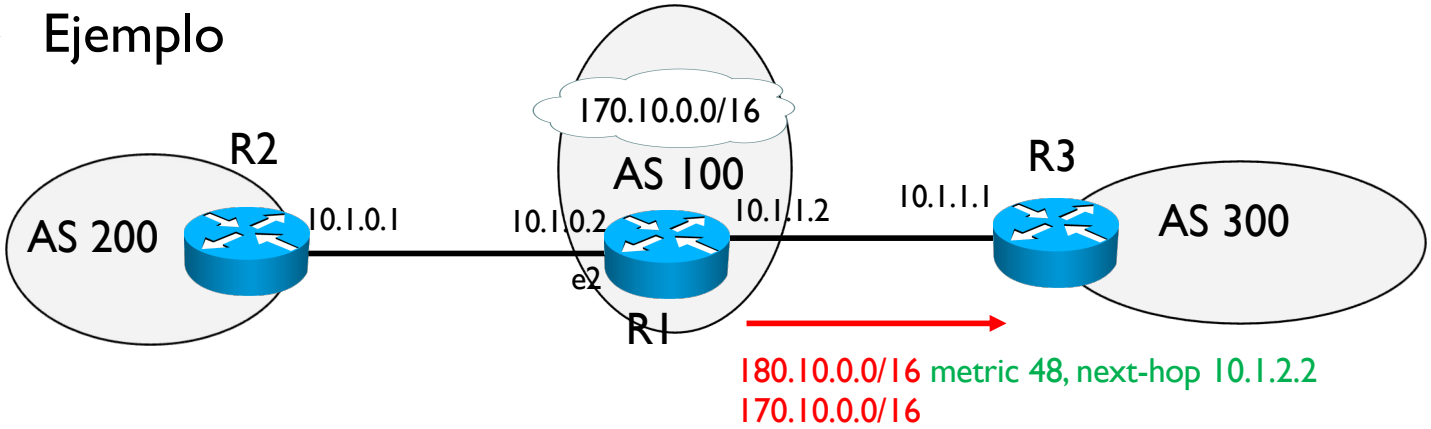
Simplemente se permite y no se hace ninguna acción

Todos los demás prefijos se filtran por la regla por defecto

5.8 - Algoritmo de selección de rutas

Políticas usando script

- Ejemplo



- Pasos
  - Aplicar a la interfaz del router y en la dirección deseada

```
R1# router bgp 100
R1# neighbor 10.1.1.1 remote-as 300
R1# neighbor 10.1.1.1 route-map POL2 out
```

5. Encaminamiento inter-dominio

1. Introducción
2. Encaminamiento path-vector
3. Funcionamiento de BGP
4. Establecimiento de una sesión BGP
5. Bases de datos BGP
6. Mensajes BGP
7. Atributos estándares
8. Algoritmo de selección de rutas
9. Escenarios comunes
10. Mejoras del BGP

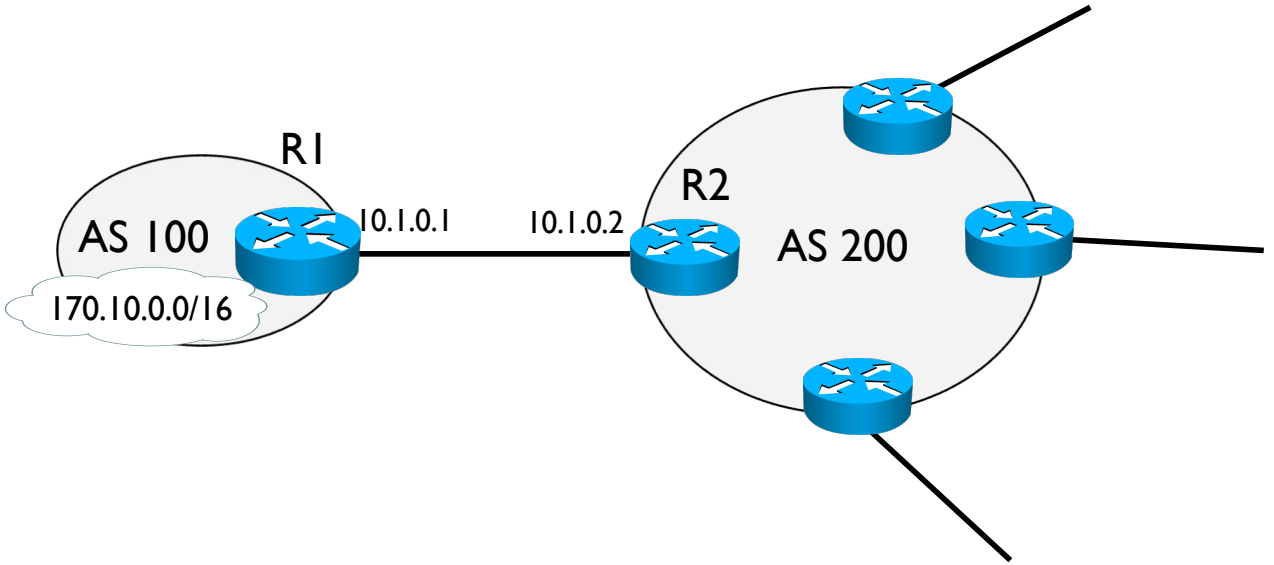
5.9 - Escenarios más comunes

- Stub
- Stub multi-homed
- Multi-homed
- Transito

5.9 - Escenarios más comunes

Stub

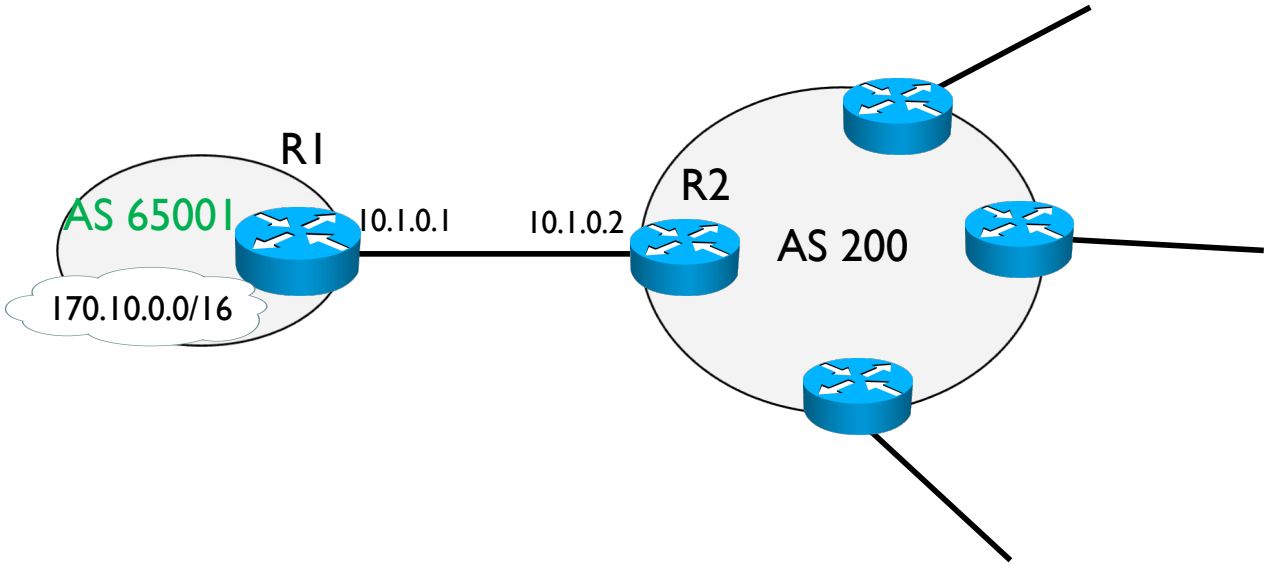
- Un AS que tiene una única conexión con otro AS
- Ejemplo AS 100
  - El AS 100 es customer (no proporciona transito a nadie)
  - Su vecino AS 200 es su provider (le proporciona transito)



5.9 - Escenarios más comunes

Stub

- Un AS que tiene una única conexión con otro AS
- Ejemplo AS 100
  - El AS 100 es customer (no proporciona transito a nadie)
  - Su vecino AS 200 es su provider (le proporciona transito)



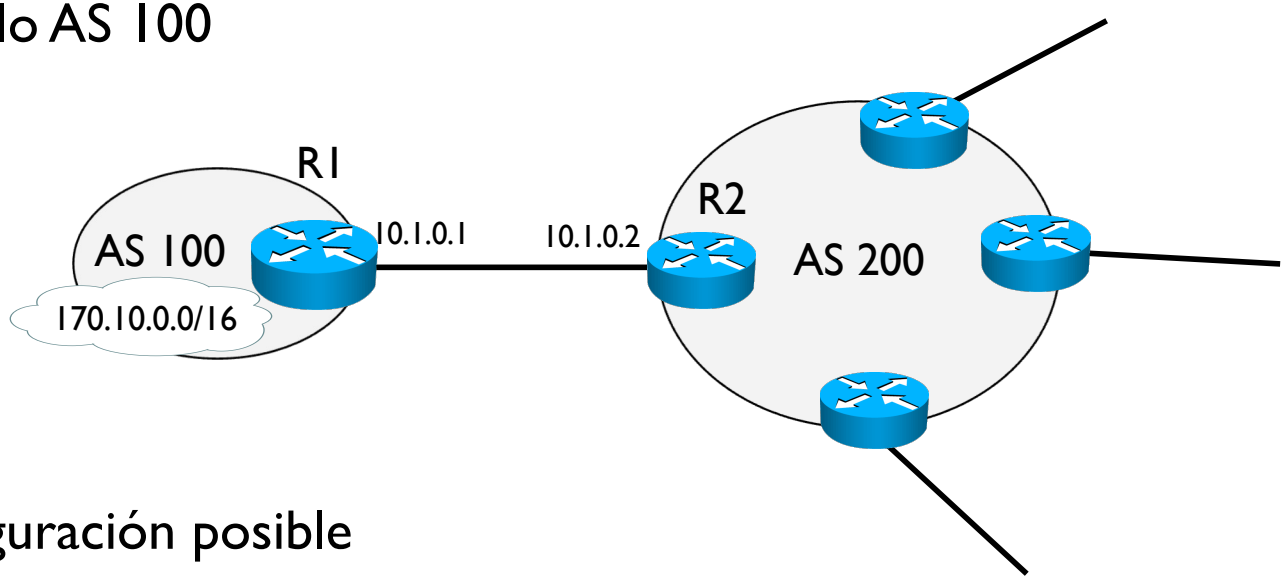
El AS puede usar un ASN privado (64,512 - 65,534)

En este caso, el provider distribuye el/los prefijos de este AS indicando que él es el origen

5.9 - Escenarios más comunes

Stub

- Ejemplo AS 100

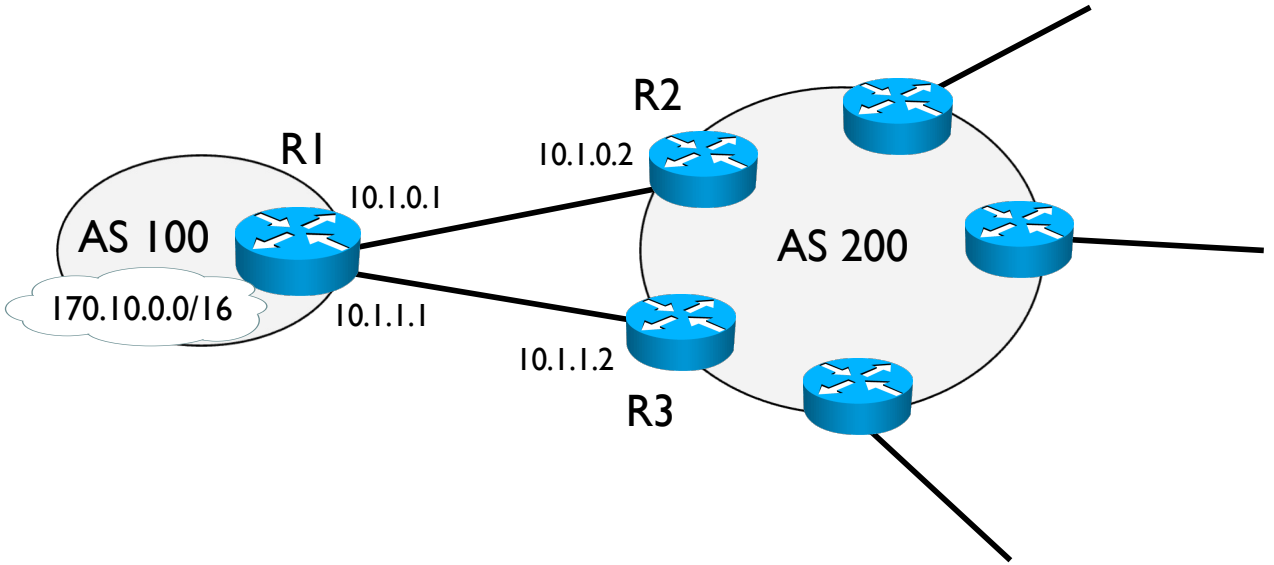


- Configuración posible
- R1
  - Debe anunciar el prefijo 170.10.0.0/16 a R2
  - Filtra todo lo que dice R2 y configura R2 como ruta por defecto
- R2
  - Debe filtrar todos los prefijos que no sean 170.10.0.0/16
  - No hace falta que anuncie prefijos a R2

5.9 - Escenarios más comunes

Stub multi-homed

- Un AS que tiene 2 o más conexiones a un mismo AS
- Ejemplo AS 100
  - El AS 100 es customer (no proporciona transito a nadie)
  - Su vecino AS 200 es su provider (le proporciona transito)



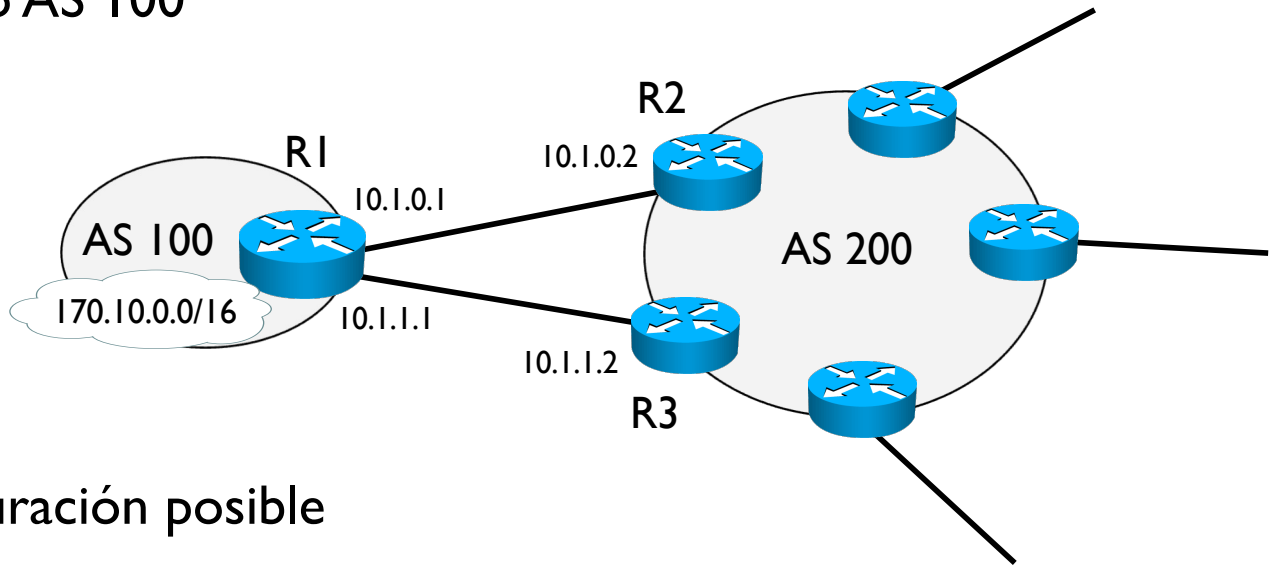
Como en el caso anterior, el AS 100 podría usar un ASN privado (64,512 - 65,534)

El provider distribuye el/los prefijos de este AS indicando que él es el origen

5.9 - Escenarios más comunes

Stub multi-homed

- Ejemplo AS 100

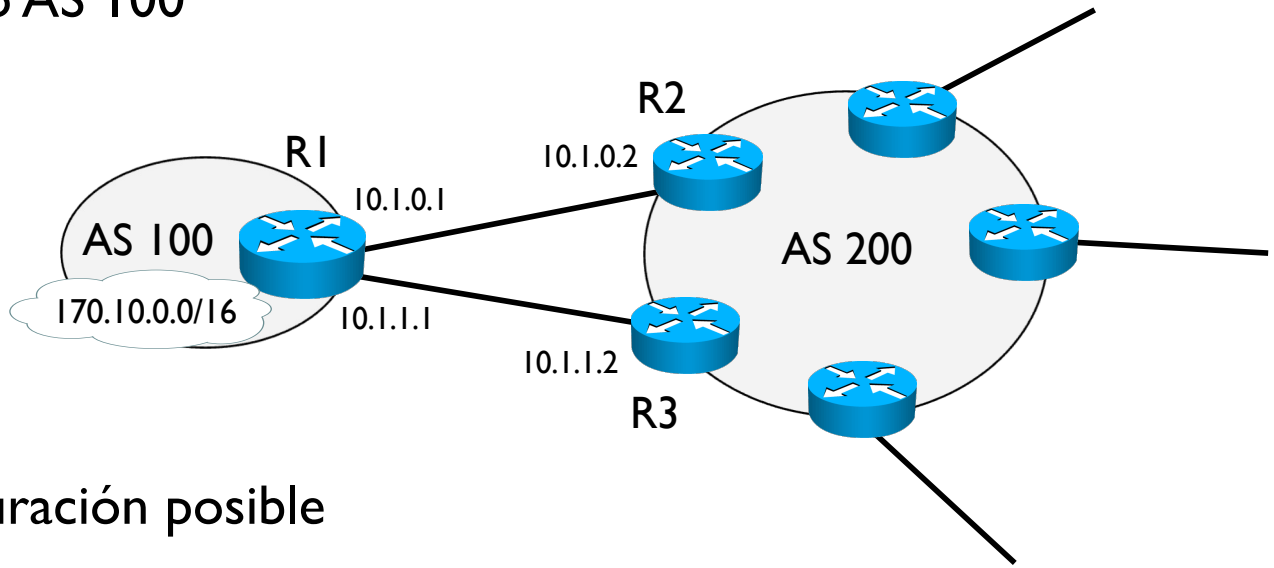


- Configuración posible
- R1 tiene tres posibilidades
  - Usar exclusivamente R2 y activar el eBGP con R3 si falla el enlace con R2
  - Usar R2 como ruta preferida y R3 como backup
  - Hacer balanceo de carga, usando los dos routers equitativamente

5.9 - Escenarios más comunes

Stub multi-homed

- Ejemplo AS 100

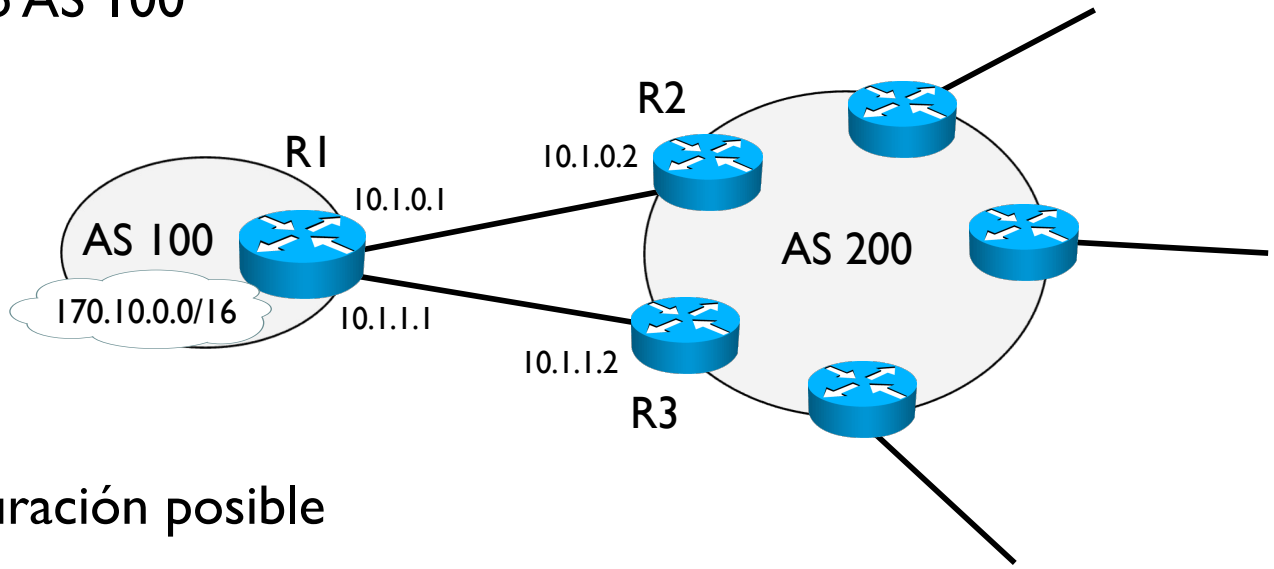


- Configuración posible
- Usar exclusivamente R2 y activar el eBGP con R3 si falla el enlace con R2
  - R1 anuncia el 170.10.0.0/16 a R2
  - R1 filtra todo lo que dice R2 y configura una ruta por defecto a R2
  - Tener esta misma configuración lista para R3 pero con la sesión eBGP cerrada

5.9 - Escenarios más comunes

Stub multi-homed

- Ejemplo AS 100

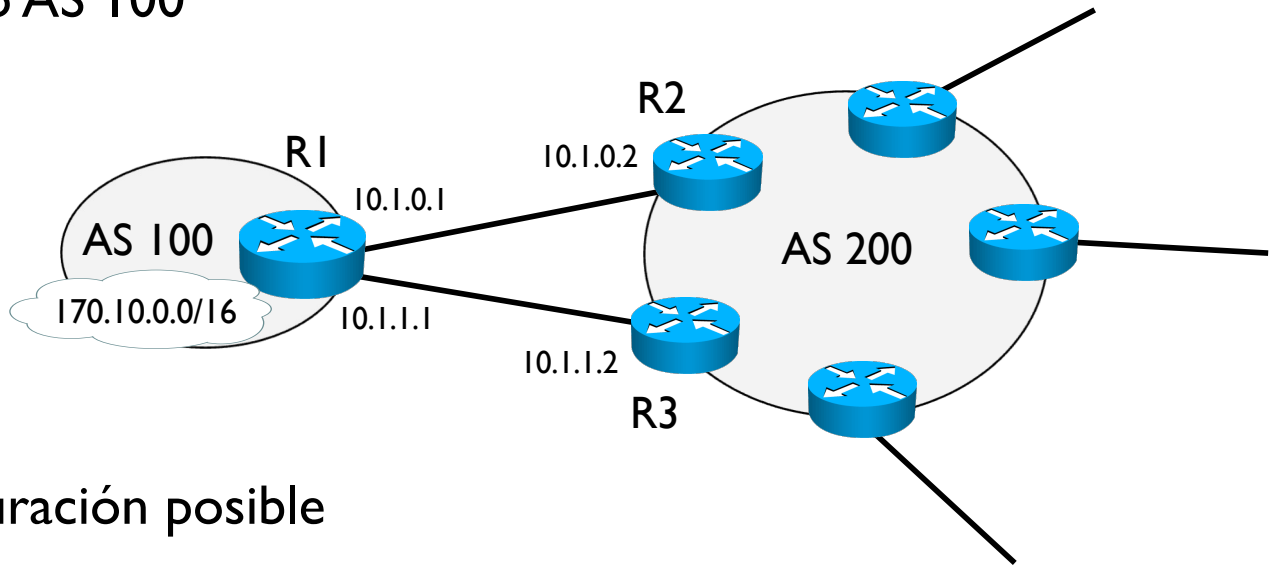


- Configuración posible
- Usar R2 como ruta preferida y R3 como backup
  - R1 anuncia el prefijo 170.10.0.0/16 a R2
  - R1 anuncia el prefijo 170.10.0.0/16 a R3 con metric 50
  - R1 acepta todos los prefijos que envía R2
  - R1 acepta todos los prefijos que envía R3 pero pone local-preference 50

5.9 - Escenarios más comunes

Stub multi-homed

- Ejemplo AS 100

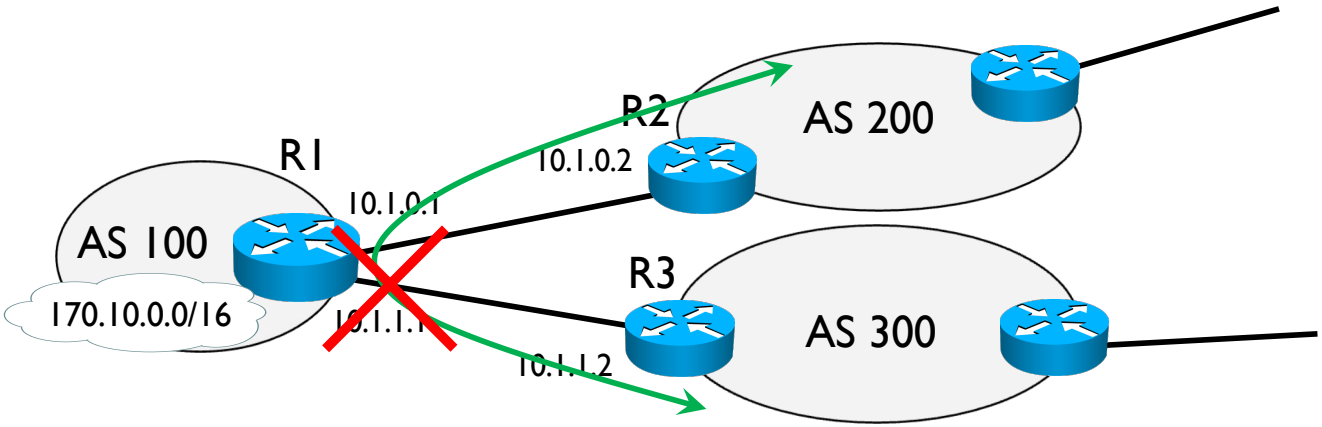


- Configuración posible
- Hacer balanceo de carga, usando los dos routers equitativamente
  - R1 anuncia la mitad de su prefijo 170.10.0.0/17 a R2
  - R1 anuncia la otra mitad 170.10.128.0/17 a R3
  - R1 acepta algunos prefijos de R2 (por ejemplo los <121.0.0.0/8) y filtra el resto
  - R1 acepta los otros de prefijos de R3 (por ejemplo los >=121.0.0.0/8) y filtra el resto

5.9 - Escenarios más comunes

Multi-homed

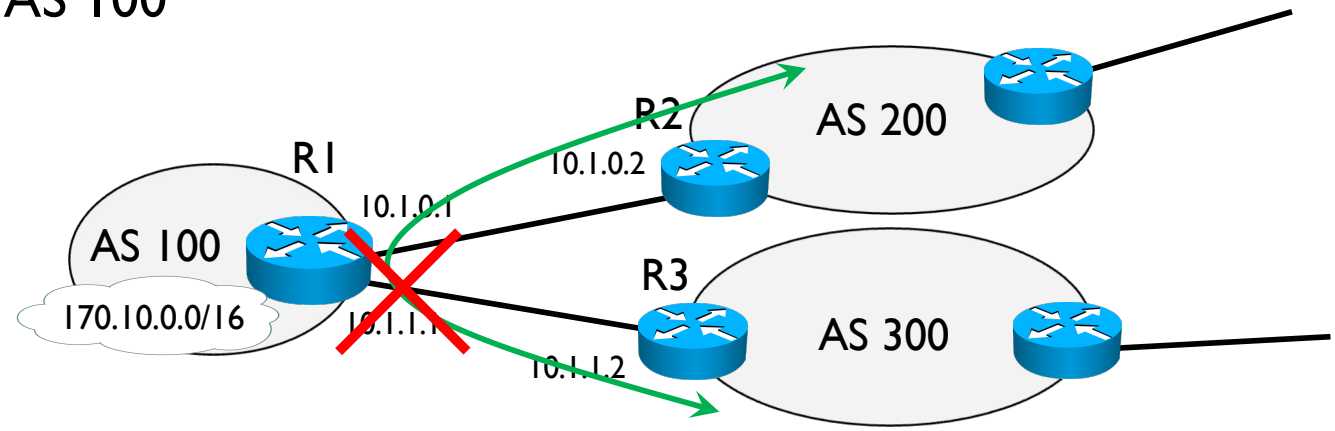
- Un AS que tiene 2 o más conexiones a diferentes AS pero no proporciona transito entre ellos
- Ejemplo AS 100
  - El AS 100 es customer (no proporciona transito a nadie)
  - Sus vecinos son sus providers (les proporcionan transito)



5.9 - Escenarios más comunes

Multi-homed

- Ejemplo AS 100

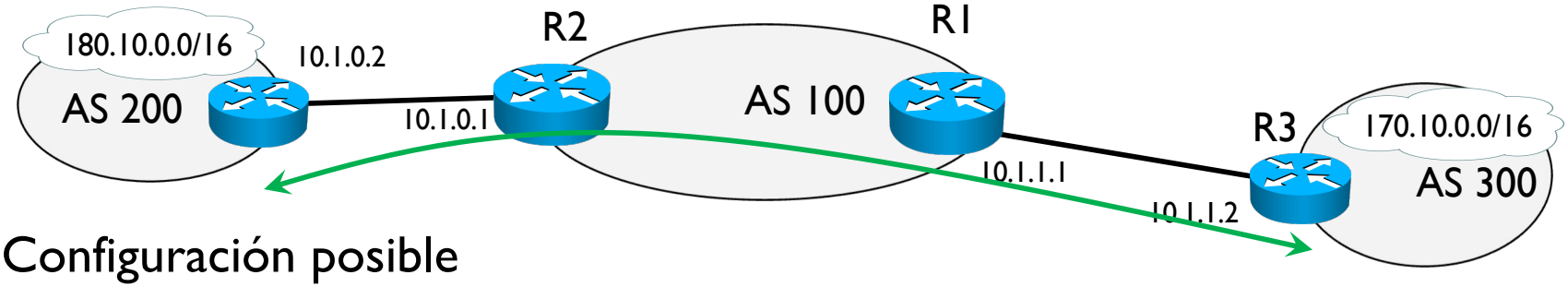


- Configuración posible
- Caso parecido al anterior, pero además R1 no debe explícitamente dar transito a los vecinos
- R1 tiene dos posibilidades
  - Usar R2 como ruta preferida y R3 como backup
  - Hacer balanceo de carga, usando los dos AS vecinos equitativamente

5.9 - Escenarios más comunes

Transito

- Un AS que tiene 2 o más conexiones a diferentes AS y proporciona transito entre ellos
- Ejemplo AS 100
  - AS 100 proporciona transito a AS300 y AS 200
  - Un mismo AS pero puede ser de transito para algunos vecinos y multi-homed para otros, depende del tipo de relación/contrato que hay entre AS



- Configuración posible
- Según el contrato con los vecinos, hay que configurar las políticas de encaminamiento correctas
  - Filtrar los prefijos no salen en el contrato
  - Modificar atributos y manipular la selección de ruta según la política aplicada

5.9 - Escenarios más comunes

Route leaks

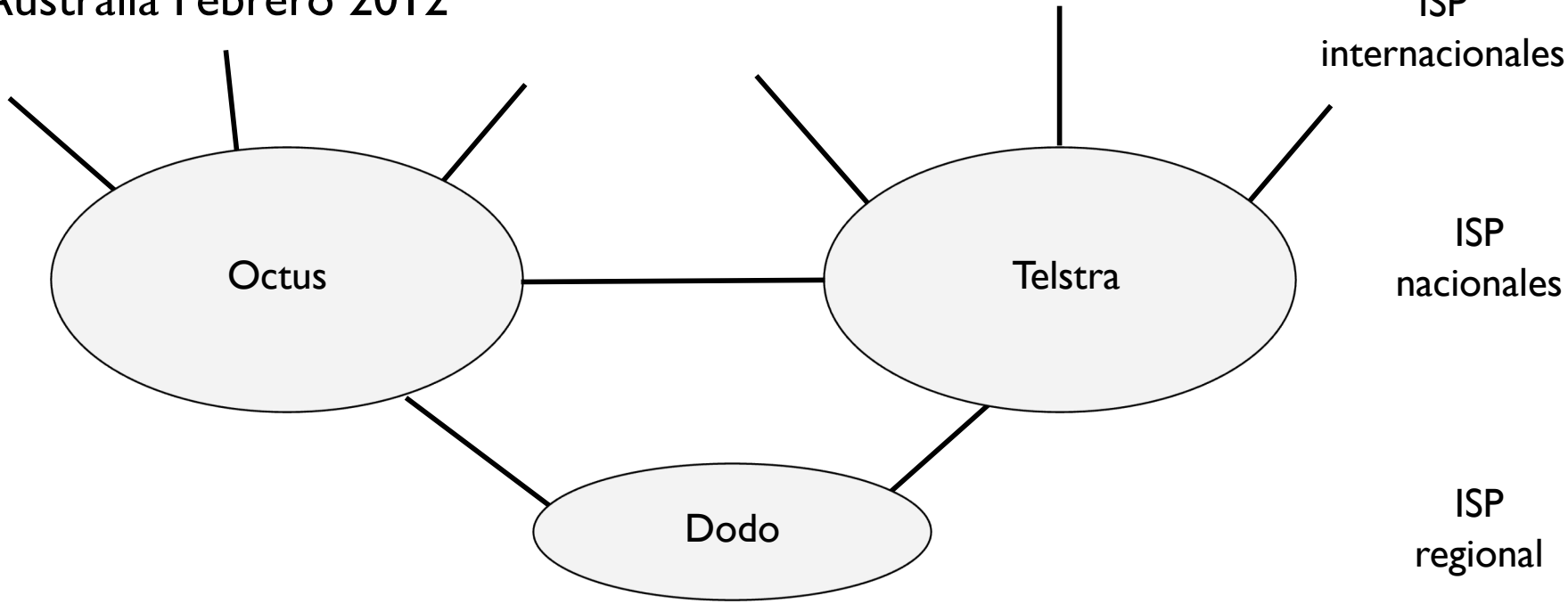
- Cuando se distribuye accidentalmente o intencionadamente uno o más prefijos de manera errónea (BGP hijacking)
- Primer caso documentado: "AS 7007 incident", abril 1997

- April 1997: The "AS 7007 incident"[7]
- December 24, 2004: TTNet in Turkey hijacks the Internet[8]
- May 7, 2005: Google's May 2005 Outage[9]
- January 22, 2006: Con Edison Communications hijacks big chunk of the Internet[10]
- February 24, 2008: Pakistan's attempt to block YouTube access within their country takes down YouTube entirely.[11]
- November 11, 2008: The Brazilian ISP CTBC - Companhia de Telecomunicações do Brasil Central leaked their internal table into the global BGP table.[12] It lasted over 5 minutes. Although, it was detected by a RIPE route server and then it was not propagated, affecting practically only their own ISP customers and few others.
- April 8, 2010: Chinese ISP hijacks the Internet[13]
- July 2013: The Hacking Team aided Raggruppamento Operativo Speciale (ROS - Special Operations Group of the Italian National Military police) in regaining access to Remote Access Tool (RAT) clients after they abruptly lost access to one of their control servers when the Santrex IPv4 prefix *46.166.163.0/24* became permanently unreachable. ROS and the Hacking Team worked with the Italian network operator Aruba S.p.A. (AS31034) to get the prefix announced in BGP in order to regain access to the control server.[14]
- February, 2014: Canadian ISP used to redirect data from ISPs.[15] - In 22 incidents between February and May a hacker redirected traffic for roughly 30 seconds each session. Bitcoin and other crypto-currency mining operations were targeted and currency was stolen.
- January 2017: Iranian pornography censorship.[16]
- April 2017: Russian telecommunication company Rostelecom (AS12389) originated 37 prefixes[17] for numerous other Autonomous Systems. The hijacked prefixes belonged to financial institutions (most notably MasterCard and Visa), other telecom companies, and a variety of other organizations.[18] Even though the possible hijacking lasted no more than 7 minutes it is still not clear if the traffic got intercepted or modified.
- December 2017: Eighty high-traffic prefixes normally announced by Google, Apple, Facebook, Microsoft, Twitch, NTT Communications, Riot Games, and others, were announced by a Russian AS, DV-LINK-AS (AS39523).[19][20]
- April 2018: Roughly 1300 IP addresses within Amazon Web Services space, dedicated to Amazon Route 53, were hijacked by eNet (or a customer thereof), an ISP in Columbus, Ohio. Several peering partners, such as Hurricane Electric, blindly propagated the announcements.[21]
- July 2018: Iran Telecommunication Company (AS58224) originated 10 prefixes of Telegram Messenger.[22]
- November 2018: US-based China Telecom site originated Google addresses.[23]
- May 2019: Traffic to a public DNS run by Taiwan Network Information Center (TWNIC) was rerouted to an entity in Brazil (AS268869).[24]
- June 2019: Large European mobile traffic was rerouted through China Telecom (AS4134)[25][26]

5.9 - Escenarios más comunes

Route leaks

- Ejemplo debido a mala configuración
- Australia Febrero 2012

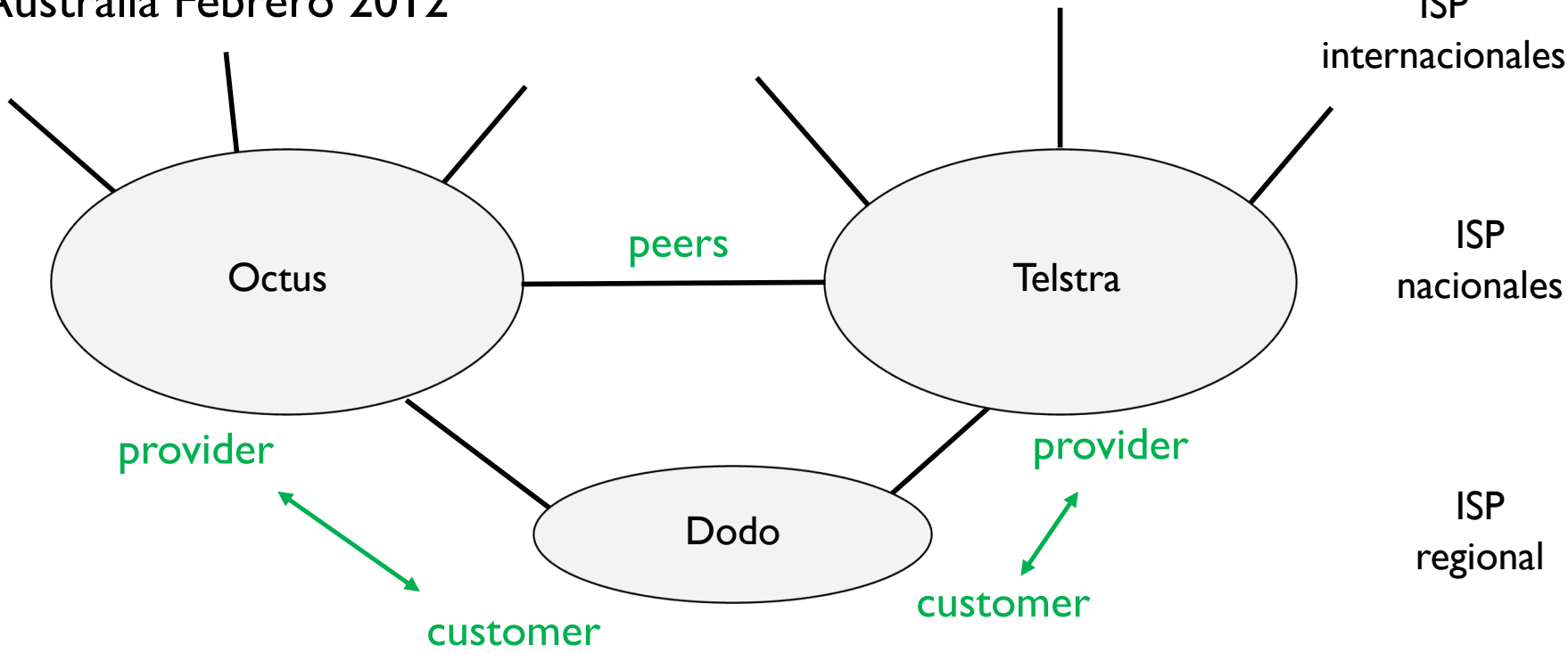


¿qué tipo de escenario se debería configurar en Dodo?

5.9 - Escenarios más comunes

Route leaks

- Ejemplo debido a mala configuración
- Australia Febrero 2012

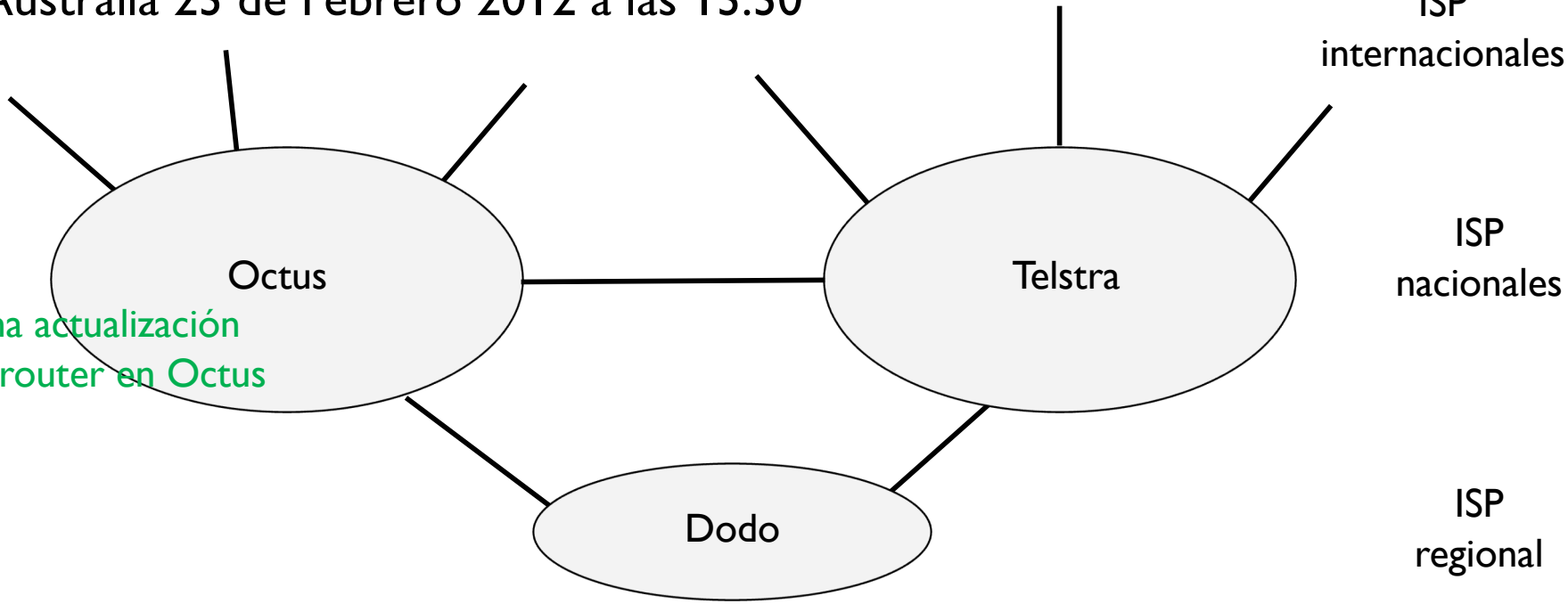


¿qué tipo de escenario se debería configurar en Dodo? Multi-homed

5.9 - Escenarios más comunes

Route leaks

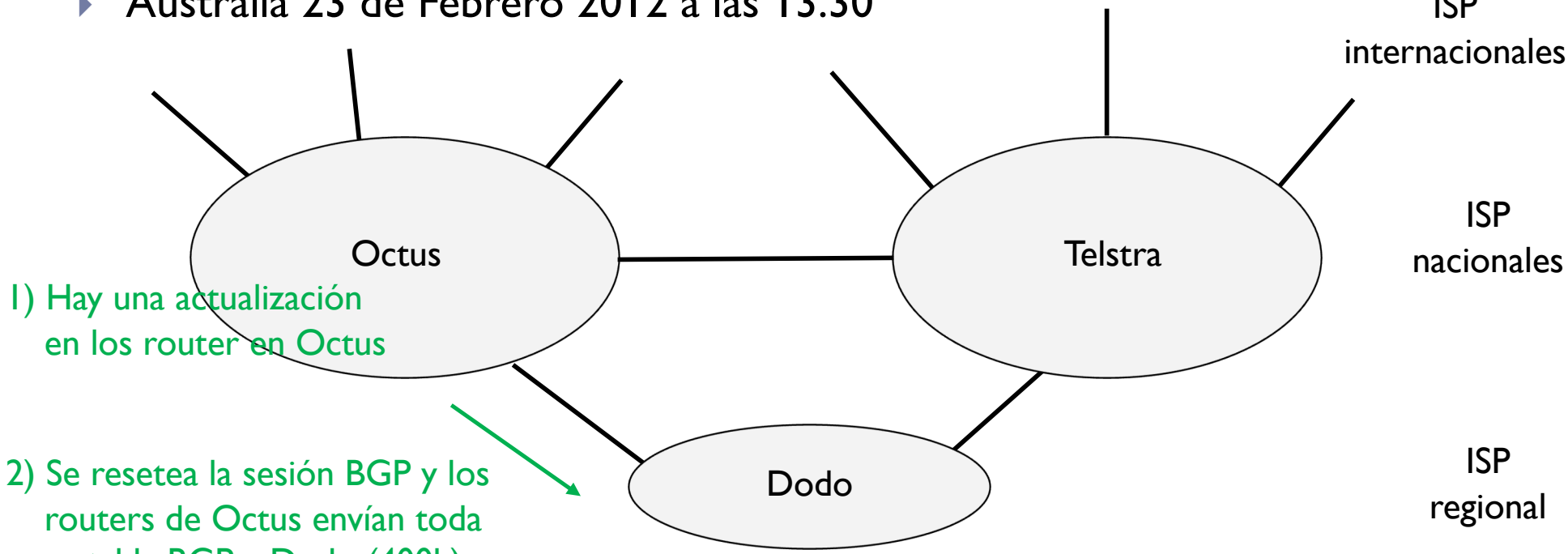
- Ejemplo debido a mala configuración
- Australia 23 de Febrero 2012 a las 13.30



5.9 - Escenarios más comunes

Route leaks

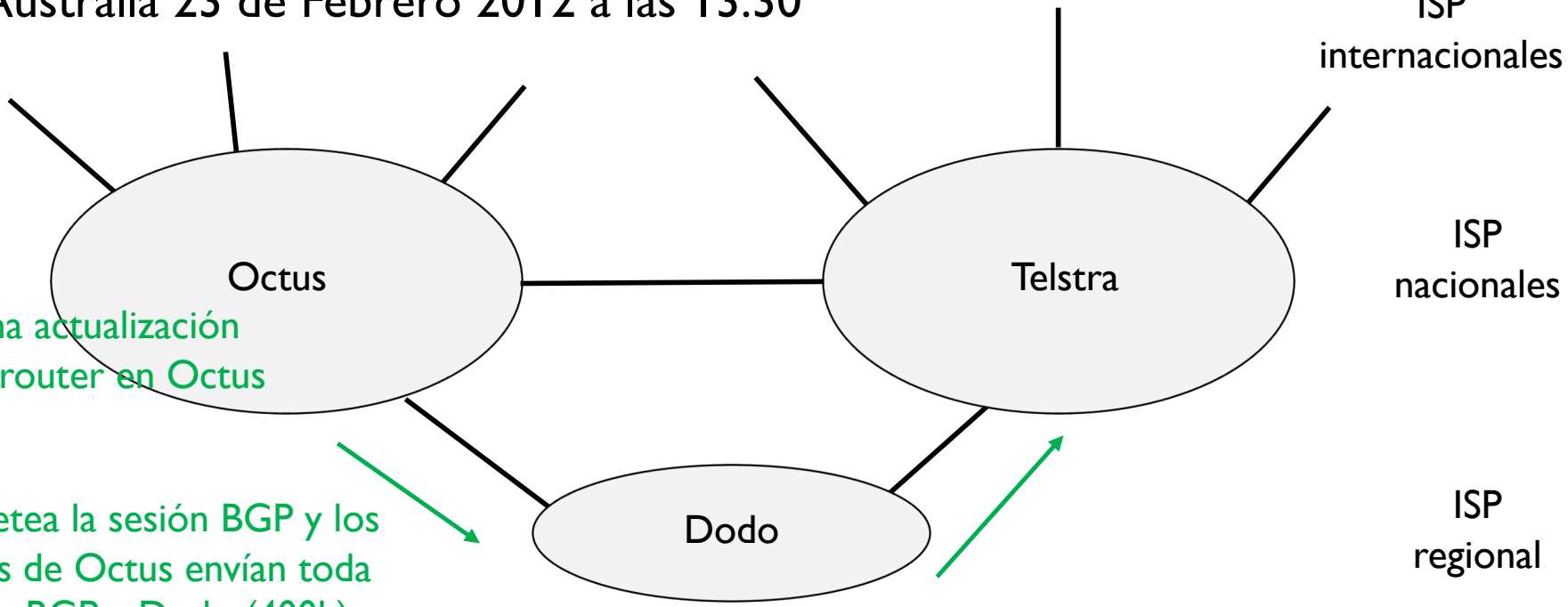
- Ejemplo debido a mala configuración
- Australia 23 de Febrero 2012 a las 13.30



5.9 - Escenarios más comunes

Route leaks

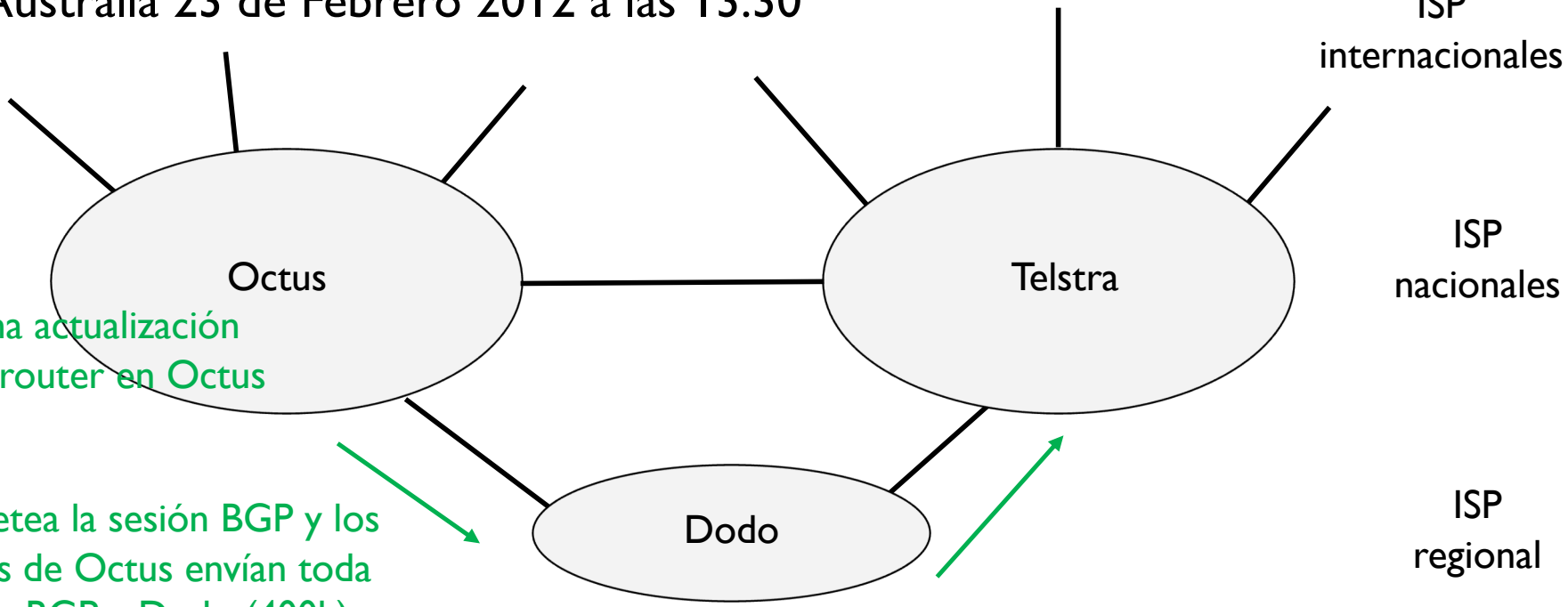
- Ejemplo debido a mala configuración
- Australia 23 de Febrero 2012 a las 13.30



5.9 - Escenarios más comunes

Route leaks

- Ejemplo debido a mala configuración
- Australia 23 de Febrero 2012 a las 13.30



1) Hay una actualización en los router en Octus

2) Se resetea la sesión BGP y los routers de Octus envían toda su tabla BGP a Dodo (400k)

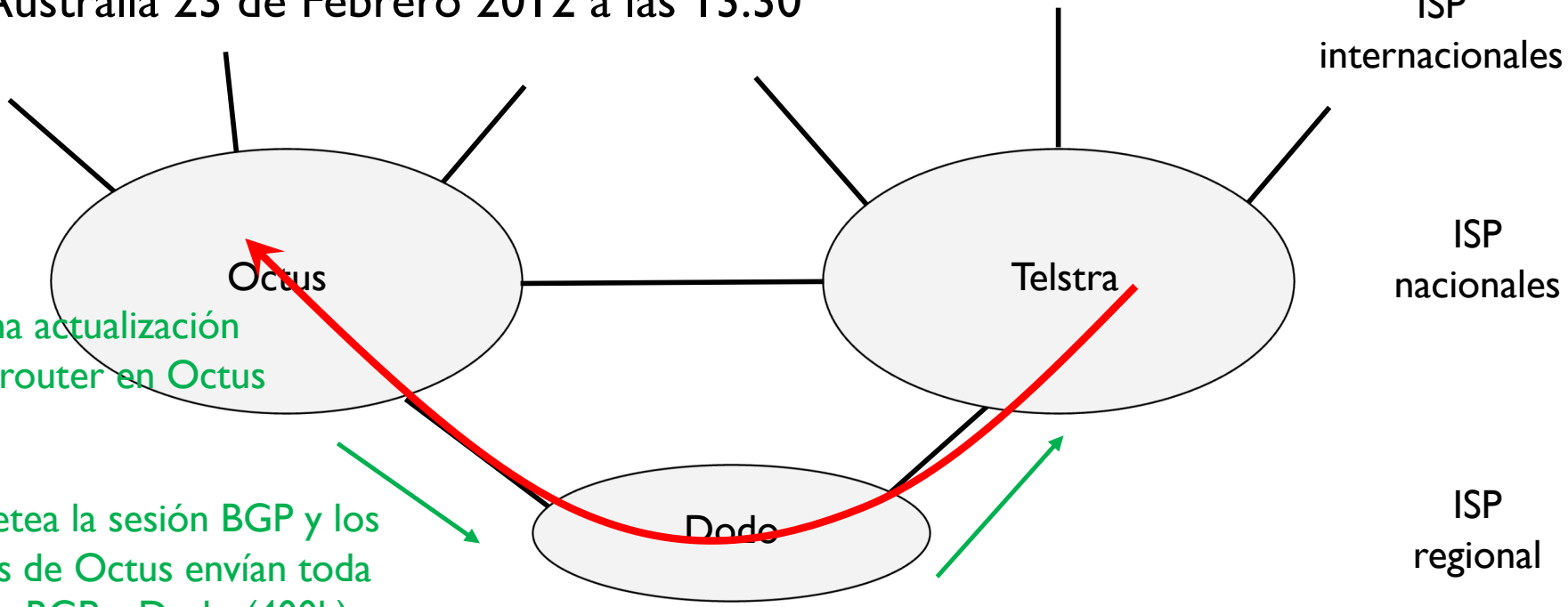
3) Dodo no debería proporcionar transito, pero lo hace y reenvía todo a Telstra junto a los prefijos propios de Dodo

4) Telstra aplica la política que, independientemente del AS-path, siempre es preferible pasar por un customer (no le paga) que por un peer o un provider. Configura sus rutas a estos 400k pasando por Dodo

5.9 - Escenarios más comunes

Route leaks

- Ejemplo debido a mala configuración
- Australia 23 de Febrero 2012 a las 13.30

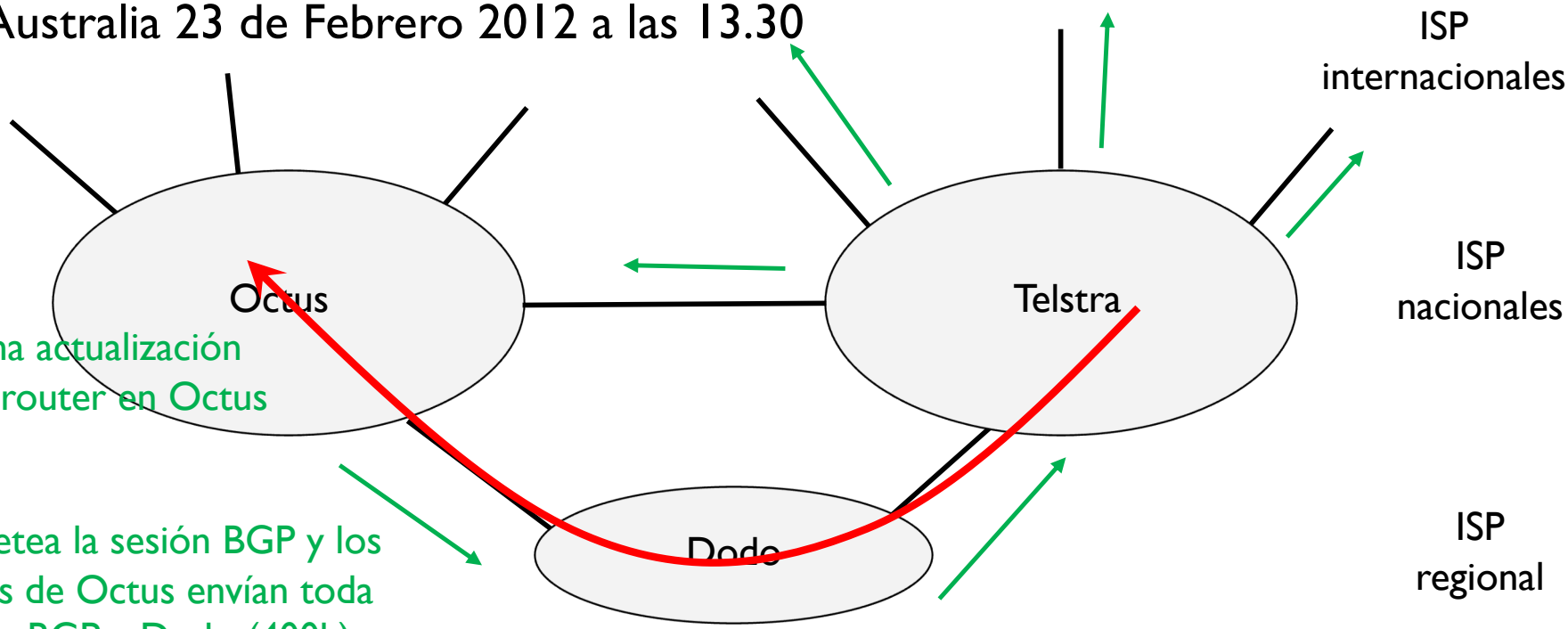


5.9 - Escenarios más

Route leaks

- Ejemplo debido a mala configuración
- Australia 23 de Febrero 2012 a las 13.30

5) Telstra envía estas nuevas rutas (400k) a los otros ISP.
Estos están bien configurado y usan un parámetro "max prefix limits". Si reciben un BGP update con un numero de prefijos que supera este limite, cierran la conexión



1) Hay una actualización en los router en Octus

2) Se resetea la sesión BGP y los routers de Octus envían toda su tabla BGP a Dodo (400k)

3) Dodo no debería proporcionar transito, pero lo hace y reenvía todo a Telstra junto a los prefijos propios de Dodo

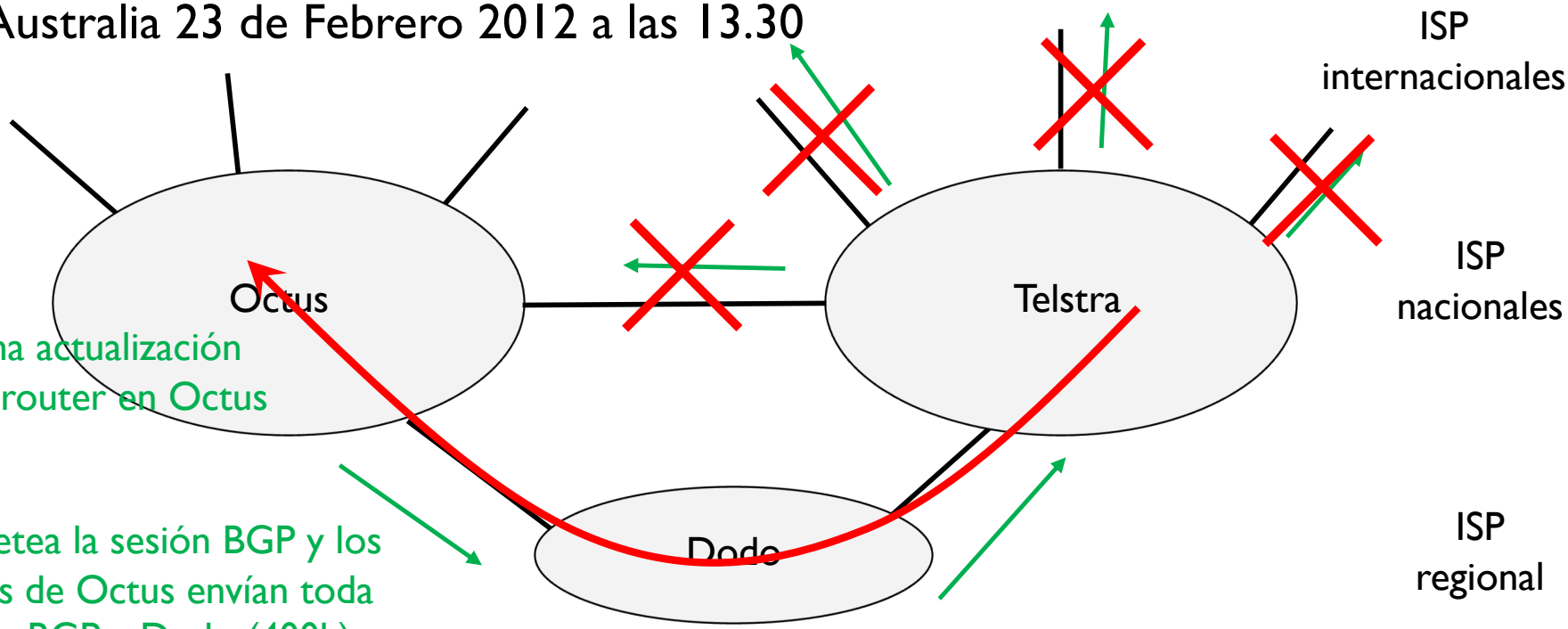
4) Telstra aplica la política que, independientemente del AS-path, siempre es preferible pasar por un customer (no le paga) que por un peer o un provider. Configura sus rutas a estos 400k pasando por Dodo

5.9 - Escenarios más

Route leaks

- Ejemplo debido a mala configuración
- Australia 23 de Febrero 2012 a las 13.30

5) Telstra envía estas nuevas rutas (400k) a los otros ISP.
Estos están bien configurado y usan un parámetro "max prefix limits". Si reciben un BGP update con un numero de prefijos que supera este limite, cierran la conexión



1) Hay una actualización en los router en Octus

2) Se resetea la sesión BGP y los routers de Octus envían toda su tabla BGP a Dodo (400k)

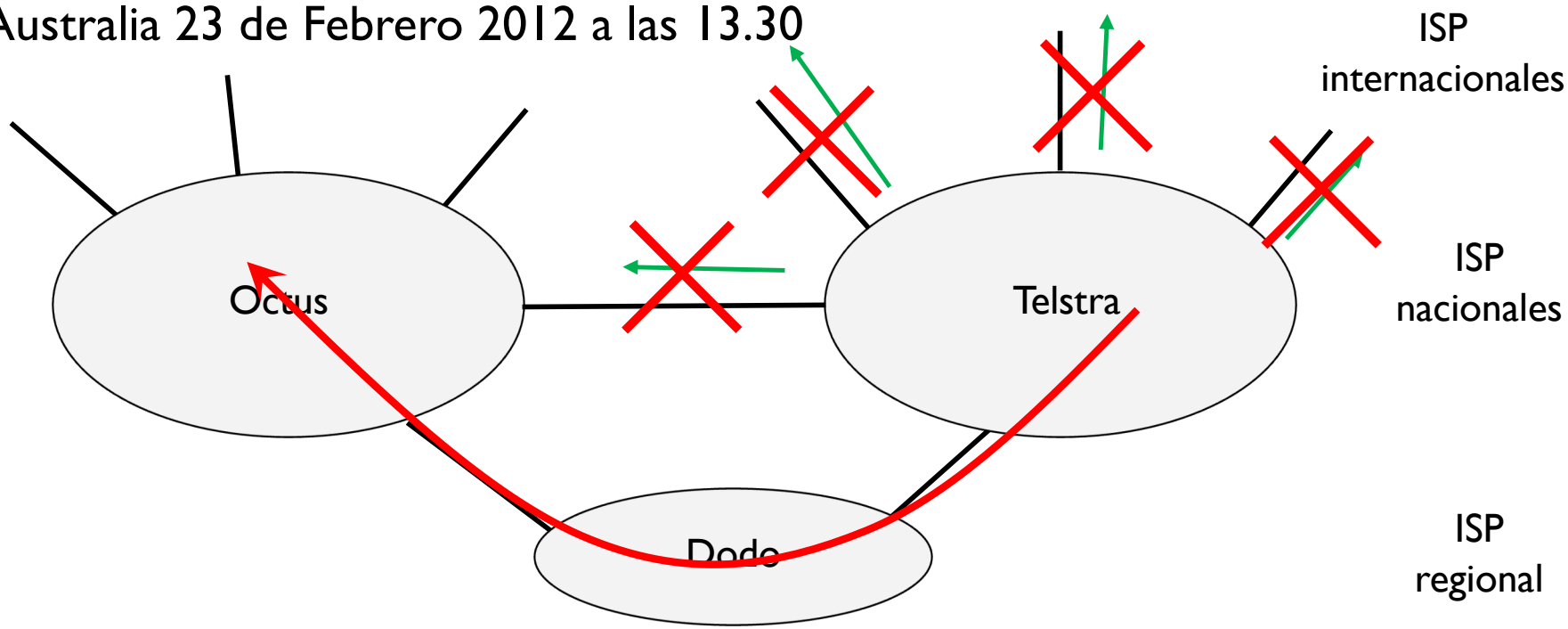
3) Dodo no debería proporcionar transito, pero lo hace y reenvía todo a Telstra junto a los prefijos propios de Dodo

4) Telstra aplica la política que, independientemente del AS-path, siempre es preferible pasar por un customer (no le paga) que por un peer o un provider. Configura sus rutas a estos 400k pasando por Dodo

5.9 - Escenarios más comunes

Route leaks

- Ejemplo debido a mala configuración
- Australia 23 de Febrero 2012 a las 13.30

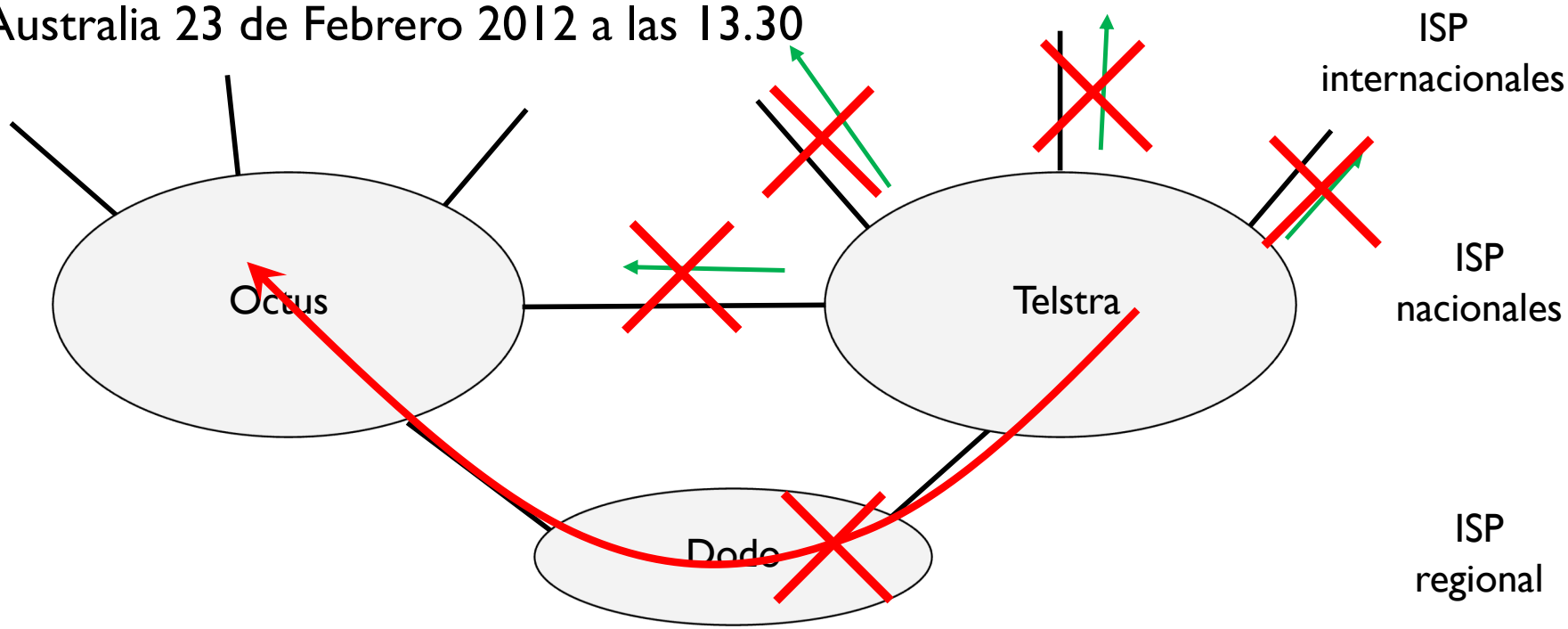


5) Dodo es un ISP regional que ahora debe soportar el tráfico de millones de australianos (Telstra es uno de los mayores ISP australianos y tiene varios otros AS customers)

5.9 - Escenarios más comunes

Route leaks

- Ejemplo debido a mala configuración
- Australia 23 de Febrero 2012 a las 13.30



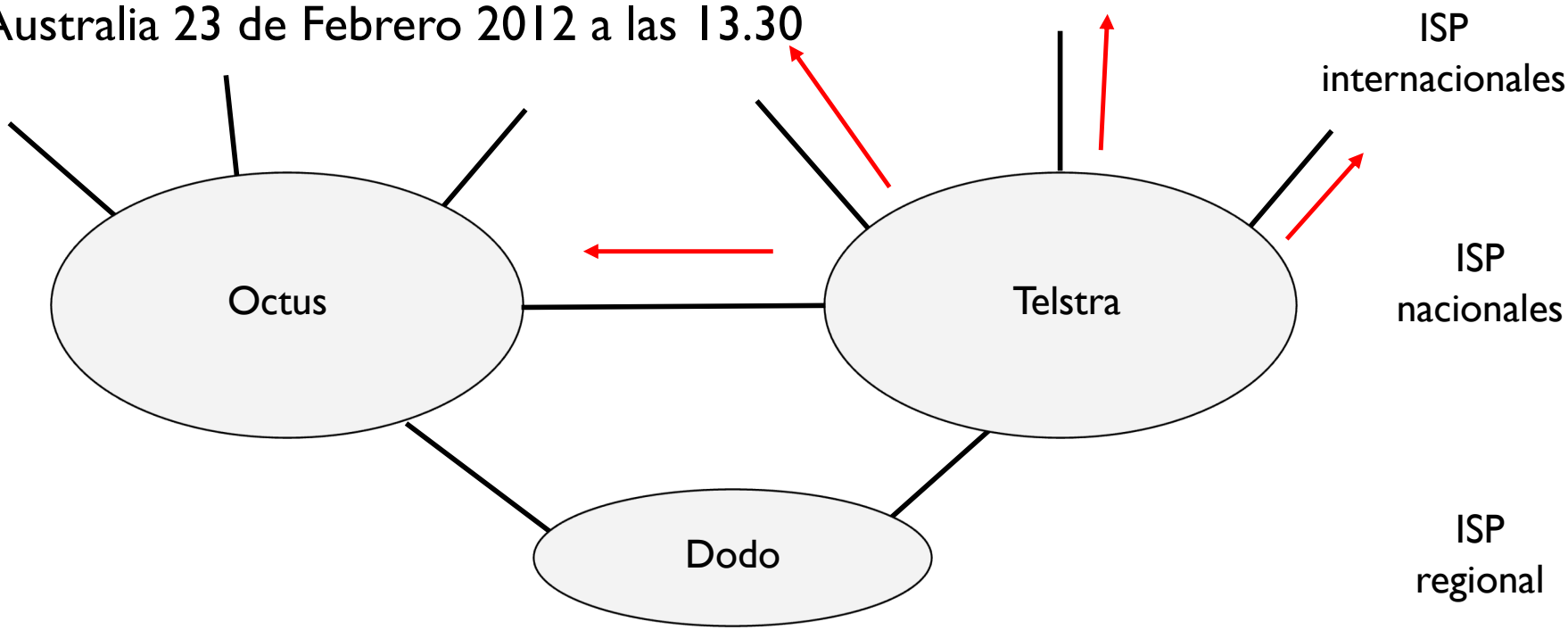
5) Dodo es un ISP regional que ahora debe soportar el tráfico de millones de australianos (Telstra es uno de los mayores ISP australianos y tiene varios otros AS customers)

→ **Los routers de Dodo caen → Telstra se queda totalmente desconectado**

5.9 - Escenarios más comunes

Route leaks

- Ejemplo debido a mala configuración
- Australia 23 de Febrero 2012 a las 13.30



5) Dodo es un ISP regional que ahora debe soportar el tráfico de millones de australianos (Telstra es uno de los mayores ISP australianos y tiene varios otros AS customers)

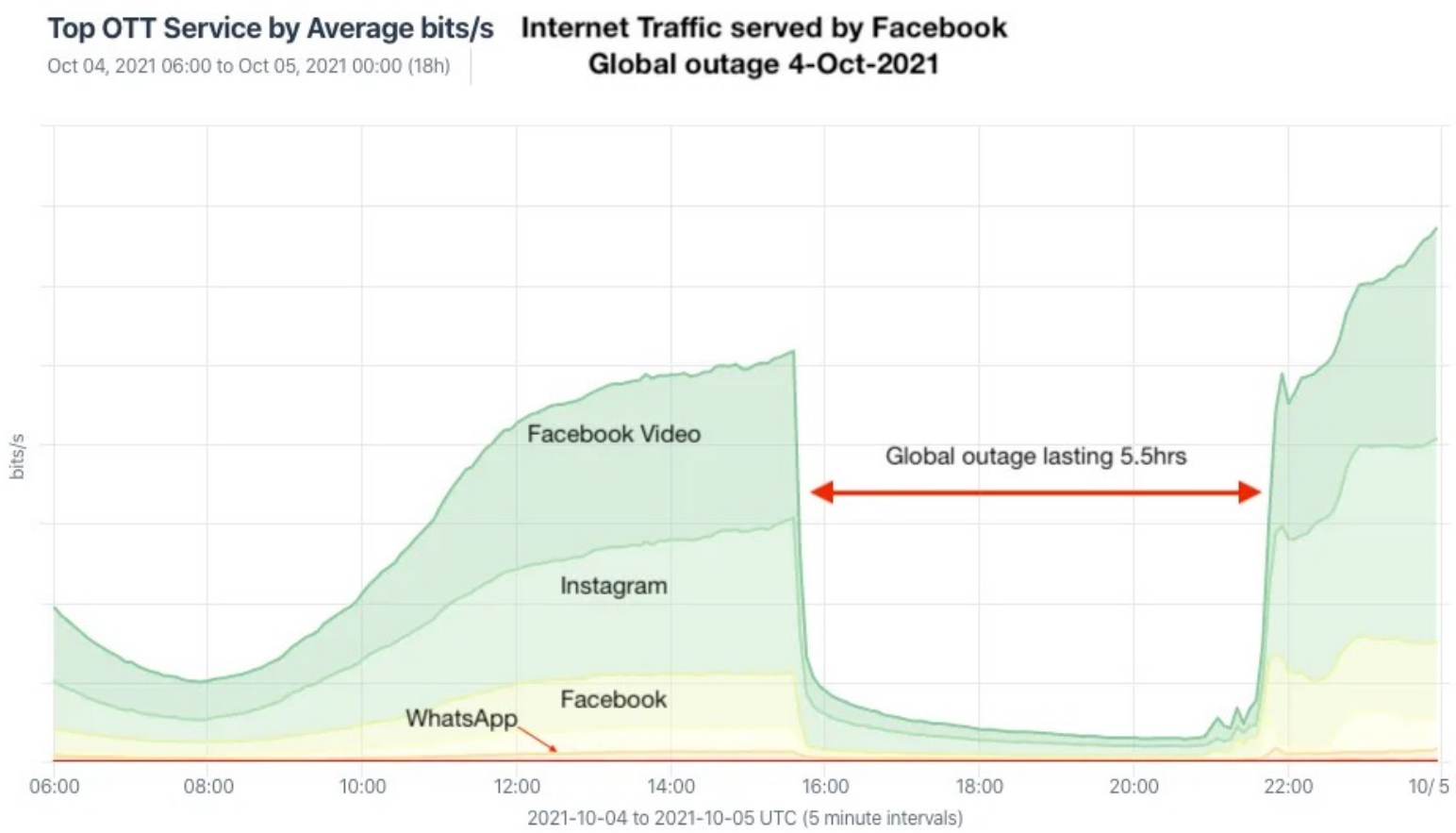
→ **Los routers de Dodo caen → Telstra se queda totalmente desconectado**

→ **46 minutos más tarde se reestablecen las sesiones con los ISP internacionales y vuelve la normalidad**

5.9 - Escenarios más comunes

Route leaks

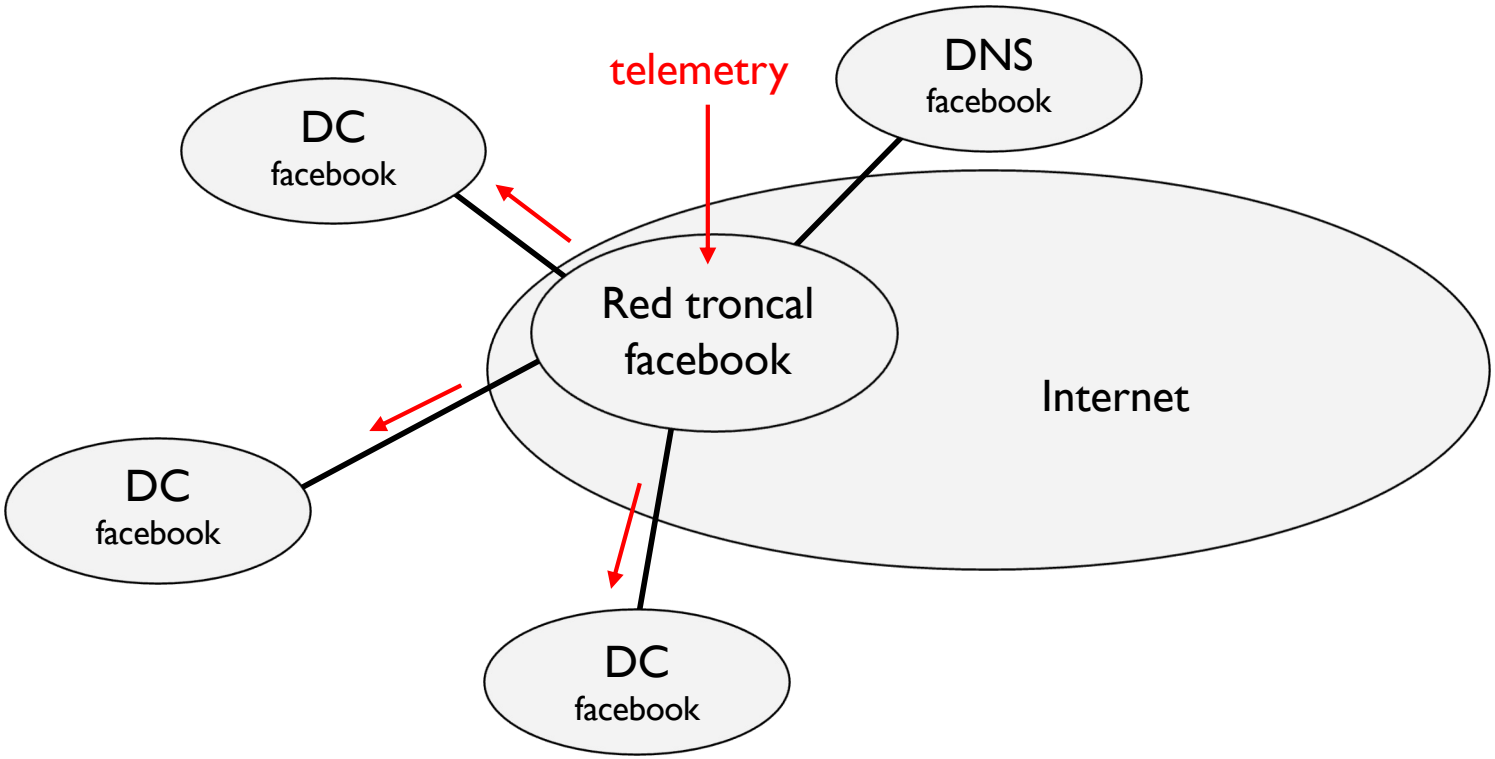
- Facebook y sus subsidiarias
  - 4 de Octubre de 2021



5.9 - Escenarios más comunes

Route leaks

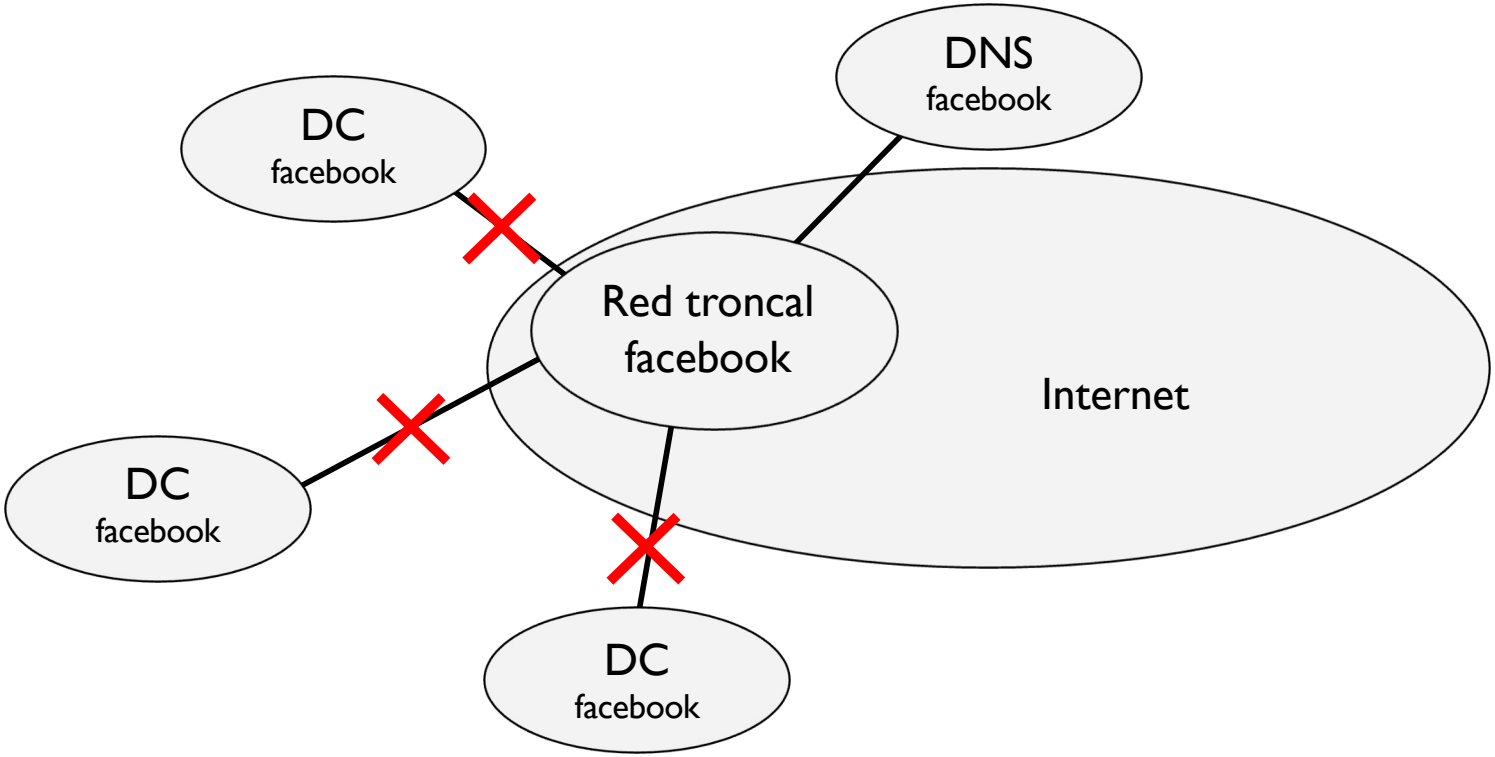
- Facebook y sus subsidiarias
  - A las 15:39 UTC, durante un mantenimiento, se envío un comando para verificar la capacidad de la red troncal



5.9 - Escenarios más comunes

Route leaks

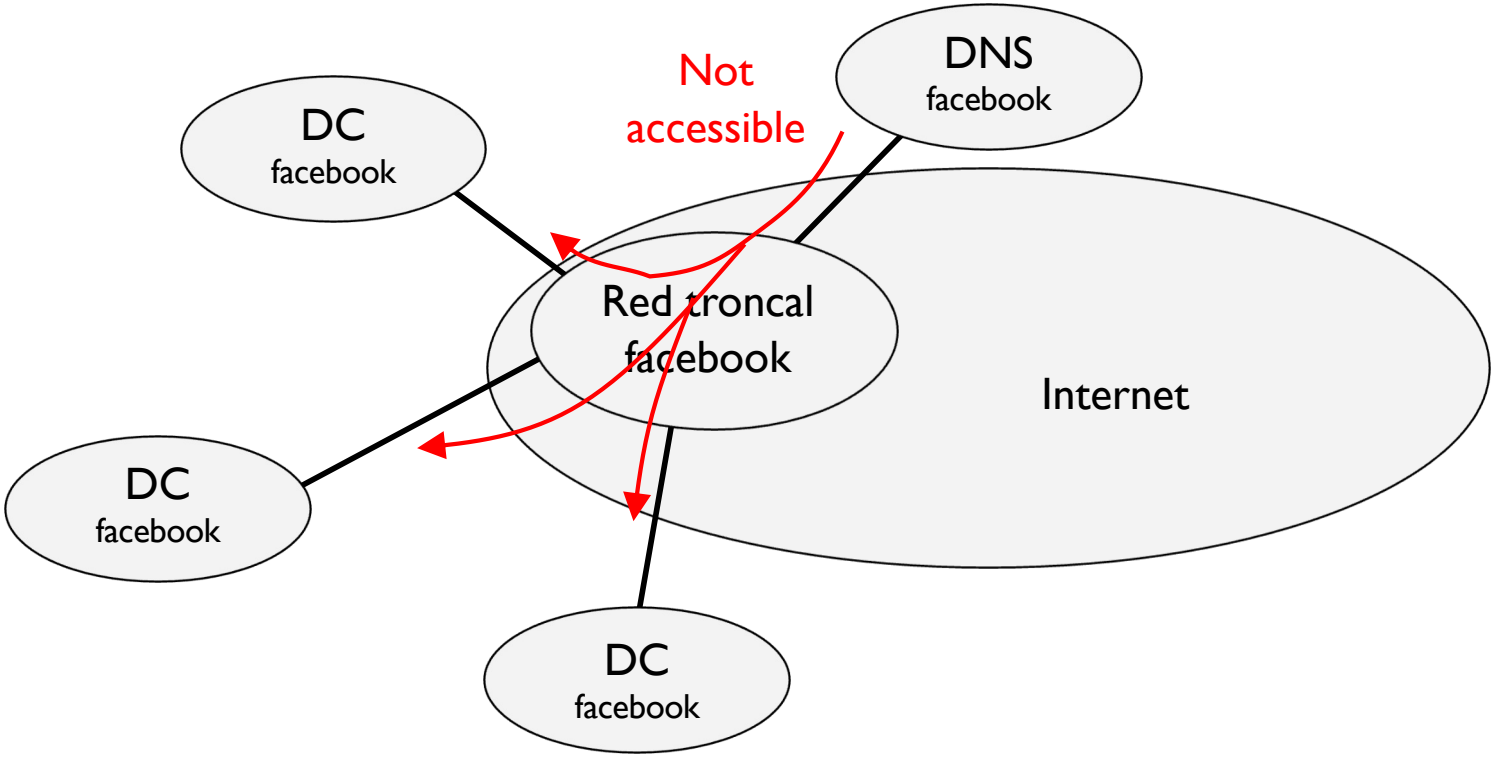
- Facebook y sus subsidiarias
  - Este comando causó una desconexión accidental de los DCs



5.9 - Escenarios más comunes

Route leaks

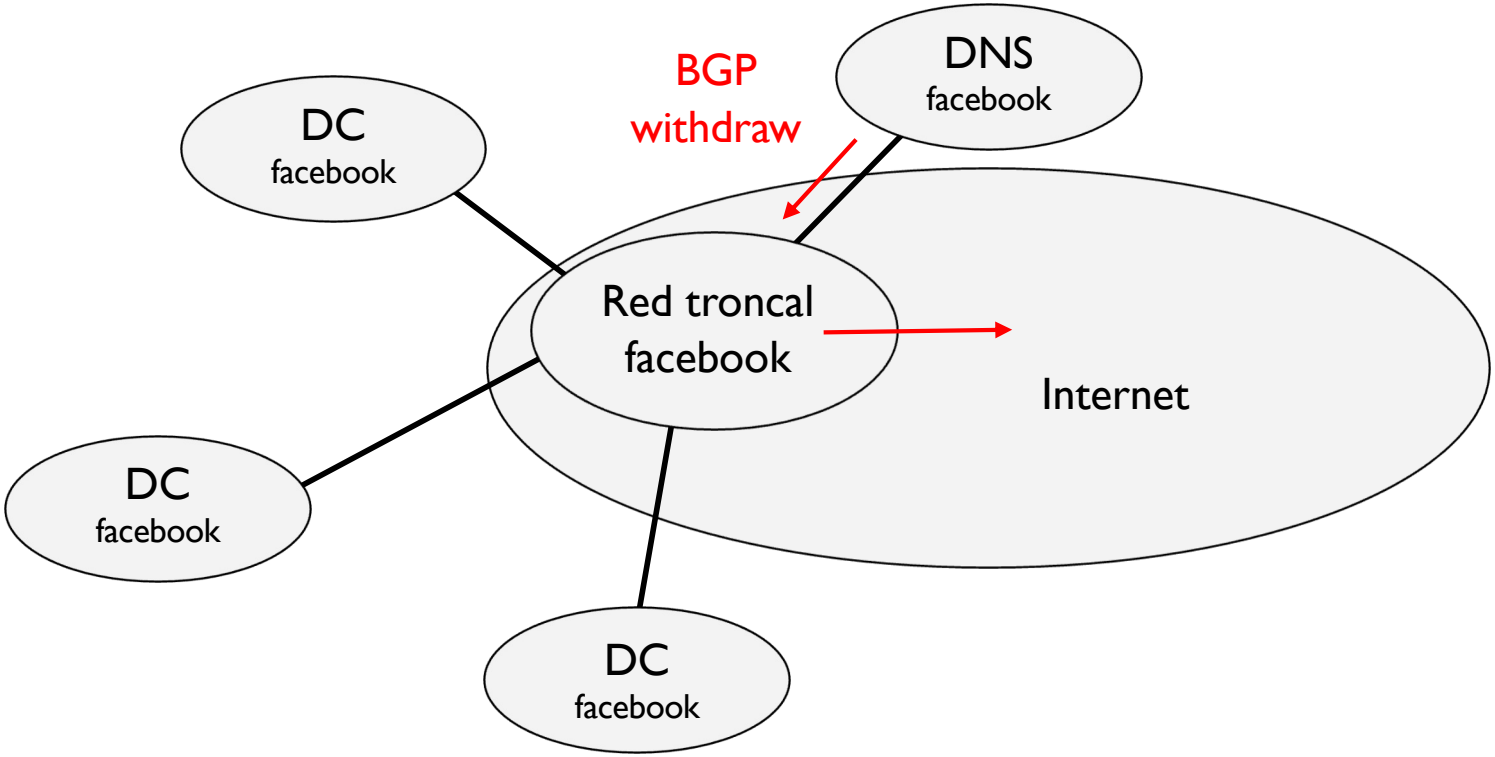
- Facebook y sus subsidiarias
  - Los servidores DNS de Facebook estaban diseñados para retirar sus rutas BGP (withdraw) si no podían conectarse a los centros de datos de Facebook



5.9 - Escenarios más comunes

Route leaks

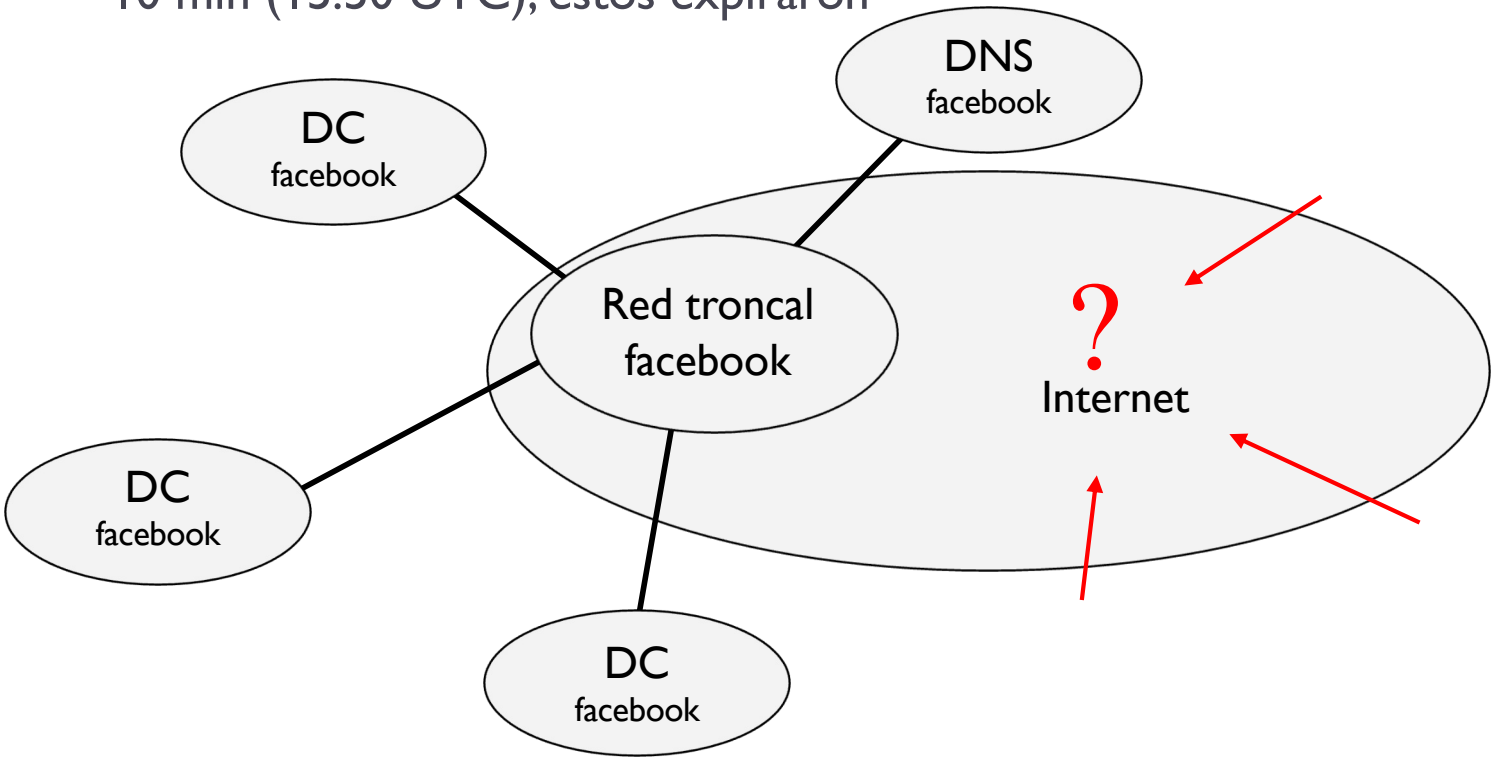
- Facebook y sus subsidiarias
  - Los servidores DNS de Facebook estaban diseñados para retirar sus rutas BGP (withdraw) si no podían conectarse a los centros de datos de Facebook



5.9 - Escenarios más comunes

Route leaks

- Facebook y sus subsidiarias
  - Los usuarios no pueden resolver los nombres relacionados con Facebook ya que no hay rutas para llegar a los servidores DNS de Facebook
  - Los otros servidores DNS de Internet tenía resolución en cache, una vez pasado 10 min (15:50 UTC), estos expiraron



5.9 - Escenarios más comunes

Route leaks

- Facebook y sus subsidiarias
  - A las 20:50 UTC, se re-anunciaron los prefijos BGP de Facebook
  - A las 21:05 UTC, los servidores DNS de Facebook ya podían resolver nombres
  - A las 22 UTC, las aplicaciones volvían a funcionar correctamente

- ¿Por qué se tardó tanto?
  - 5 horas y media para reactivar BGP
  - Casi 7 horas para las aplicaciones

5.9 - Escenarios más comunes

Route leaks

- Se necesitaban badges para entrar en las instalaciones de los DC de Facebook
- Al querer entrar, el lector verificaba los datos de los badges en los servidores de Facebook
- Los servidores de Facebook eran inalcanzable
- Nadie podía entrar (ni salir)

Was just on phone with someone who works for FB who described employees unable to enter buildings this morning to begin to evaluate extent of outage because their badges weren't working to access doors.

8:51 pm · 4 Oct 2021

```
    There are people now trying to gain access to the peering routers to implement
fixes, but the people with physical access is separate from the people with knowledge
of how to actually authenticate to the systems and people who know what to actually
do, so there is now a logistical challenge with getting all that knowledge unified.

    Part of this is also due to lower staffing in data centers due to pandemic
measures.
```

5.9 - Escenarios más comunes

Route leaks

- Tipos de ataques intencionados
  - Un AS anuncia un prefijo diciendo que es el origen cuando realmente no lo es
  - Un AS anuncia un prefijo con mayor mascara (más especifico) que el verdadero AS
  - Un AS anuncia una ruta más corta que la real y esta ruta no existe
- Resultado: los paquetes van por donde no toca y se capturan o se pierden

- December 2017: Eighty high-traffic prefixes normally announced by Google, Apple, Facebook, Microsoft, Twitch, NTT Communications, Riot Games, and others, were announced by a Russian AS, DV-LINK-AS (AS39523).[19][20]
- July 2018: Iran Telecommunication Company (AS58224) originated 10 prefixes of Telegram Messenger.[22]
- June 2019: Large European mobile traffic was rerouted through China Telecom (AS4134)[25][26]

5.9 - Escenarios más comunes

Route leaks

- Posibles soluciones
  - Filtrar por AS (descartar a priori los sospechosos o nuevos)
  - Análisis de eventos pasados
  - Usar un Internet Routing Registry centralizado
    - Cada AS debería registrar sus propios prefijos y las políticas aplicadas a los demás
  - Usar blockchain para verificar BGP
    - https://datatracker.ietf.org/meeting/105/materials/slides-105-dinrg-a-blockchainbased-testbed-for-bgp-verification

5. Encaminamiento inter-dominio

1. Introducción
2. Encaminamiento path-vector
3. Funcionamiento de BGP
4. Establecimiento de una sesión BGP
5. Bases de datos BGP
6. Mensajes BGP
7. Atributos estándares
8. Algoritmo de selección de rutas
9. Escenarios comunes
10. Mejoras del BGP

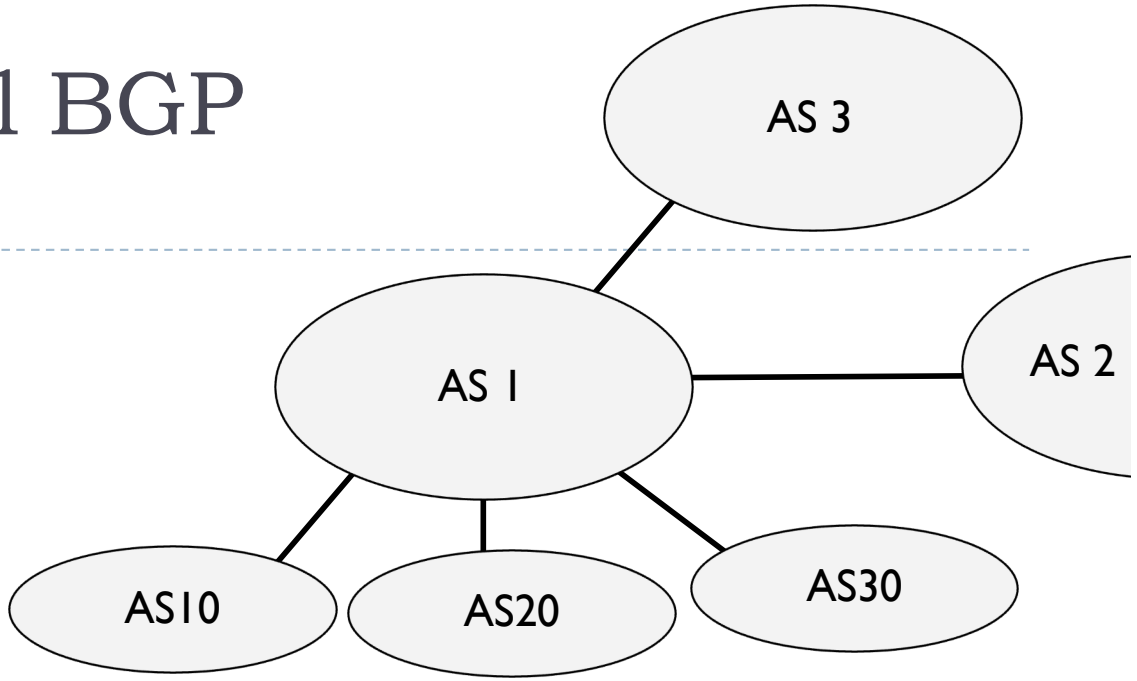
5.10 - Mejoras del BGP

- Community
- Route Flap Damping
- Escalabilidad en iBGP
  - Route reflection
  - Sub-AS confederation

5.10 - Mejoras del BGP
Community

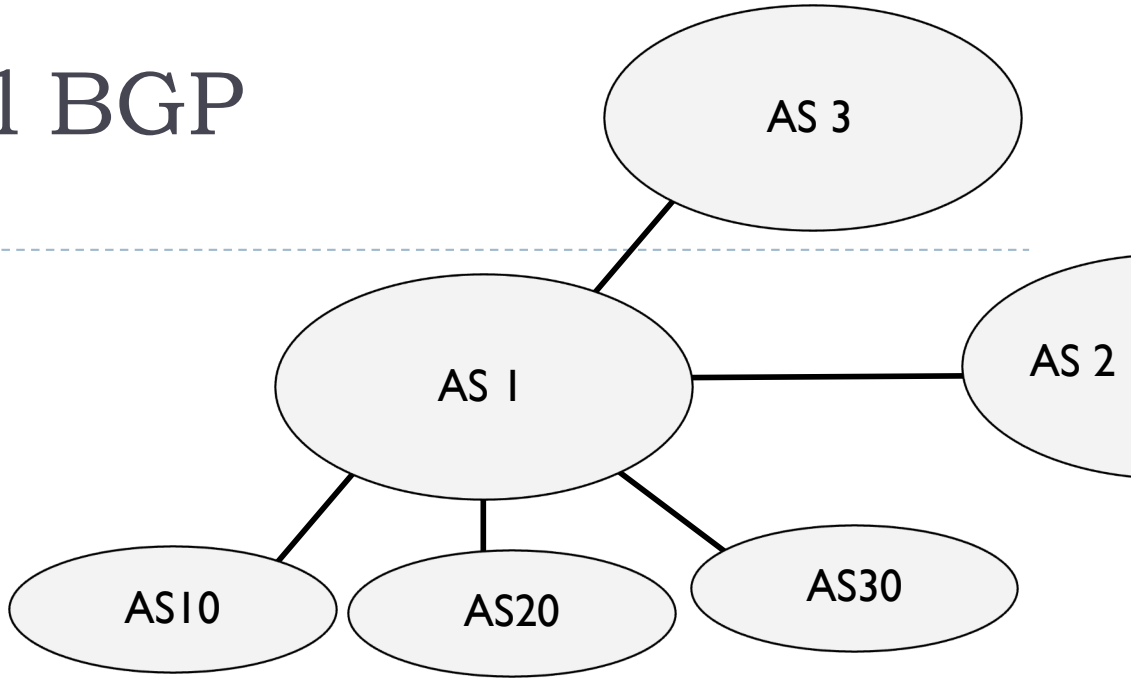
- Atributo opcional
- Facilita la configuración de políticas de encaminamiento
  - Permite agrupar destinos/prefijos en comunidades
  - A una misma comunidad se aplican las mismas políticas
- Se usan etiquetas de 32 bits para identificar una comunidad y se envían por BGP update
  - Generalmente se usa el formato <ASN:numero de 16 bits>

5.10 - Mejoras del BGP
Community



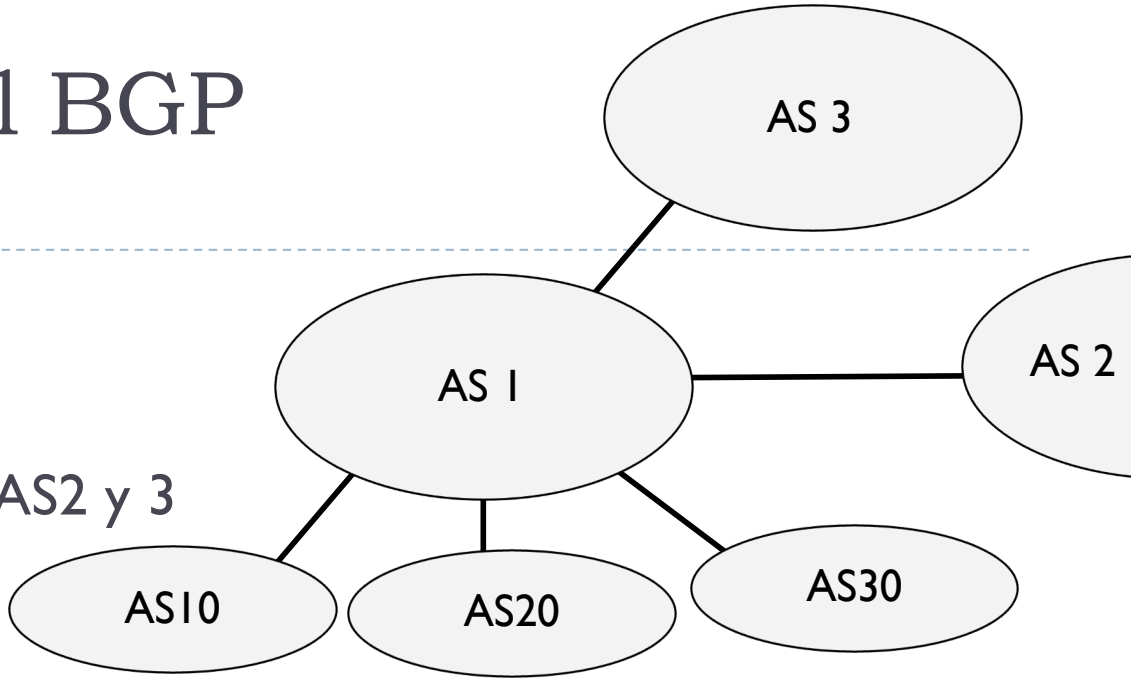
- Ejemplo
  - AS 10, 20 y 30 son stub
  - AS 1 proporciona transito
- Configuración AS 1
  - Debe filtrar los prefijos no previstos en el contrato de AS10, 20, 30
- Problemas
  - Puede ser una lista muy grande
  - Puede ser que los prefijos de AS10,20,30 varíen en el tiempo
    - Se hacen más grandes, tienen más clientes, cambian de rango de IPs, pasan a IPv6
  - Debe constantemente cambiar la configuración de sus routers para adaptarse a AS10,20,30 → tiempo, dinero

5.10 - Mejoras del BGP
Community



- Ejemplo
  - AS 10, 20 y 30 son stub
  - AS 1 proporciona transito
- Configuración con comunidades
  - AS 1 define la comunidad 1:123
  - En la sesión eBGP entre AS1 y AS10,20,30 se activa la opción comunidad
  - AS10,20,30 envían sus BGP updates con la comunidad 1:123
  - AS 1 proporciona transito solo a los prefijos con comunidad 1:123
- Si hay un cambio en los prefijos de AS10,20,30
  - AS10,20,30 simplemente envían los cambios con comunidad 1:123
  - AS 1 no debe tocar nada

5.10 - Mejoras del BGP
Community

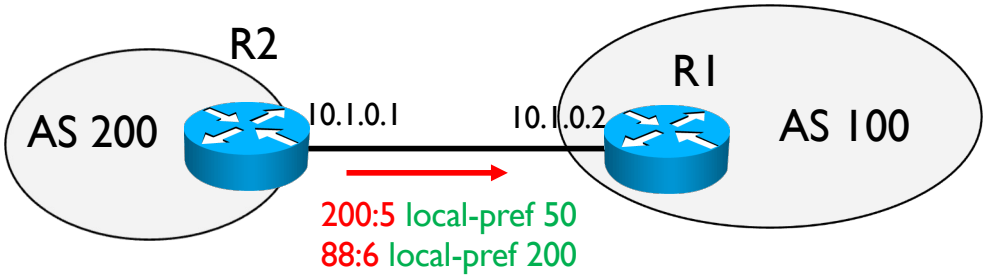


- Ejemplo
  - AS 1 es multi-homed
  - No proporciona transito a AS2 y 3
- Configuración con comunidades
  - AS 1 define la comunidad 1:23
  - En la sesión eBGP entre AS1 y AS2 y 3 se activa la opción comunidad
  - AS 2 y 3 envían sus prefijos con comunidad 1:23
  - AS 1 se queda con los prefijos pero no los re-envían

5.10 - Mejoras del BGP

Community

- Ejemplo CISCO

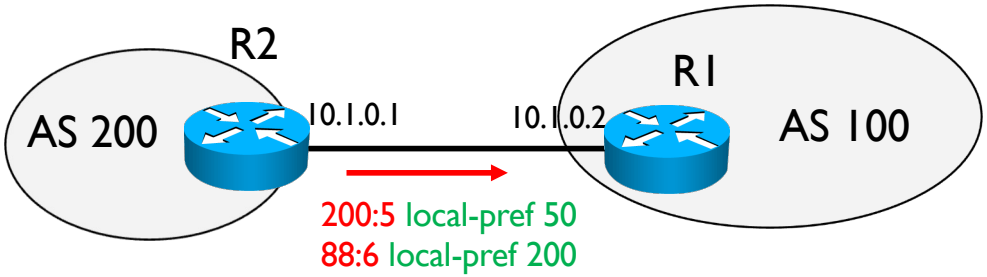



```
R1# ip community-list 1 permit 200:5
R1# ip community-list 2 permit 88:6

R1# route-map COM permit 10
R1# match community 1
R1# set local-preference 50
R1# route-map COM permit 20
R1# match community 2
R1# set local-preference 200

R1# router bgp 100
R1# neighbor 10.1.0.1 remote-as 200
R1# neighbor 10.1.0.1 route-map COM in
```

5.10 - Mejoras del BGP

Community

- Ejemplo CISCO




```
R1# ip community-list 1 permit 200:5
R1# ip community-list 2 permit 88:6

R1# route-map COM permit 10
R1# match community 1
R1# set local-preference 50
R1# route-map COM permit 20
R1# match community 2
R1# set local-preference 200

R1# router bgp 100
R1# neighbor 10.1.0.1 remote-as 200
R1# neighbor 10.1.0.1 route-map COM in
```

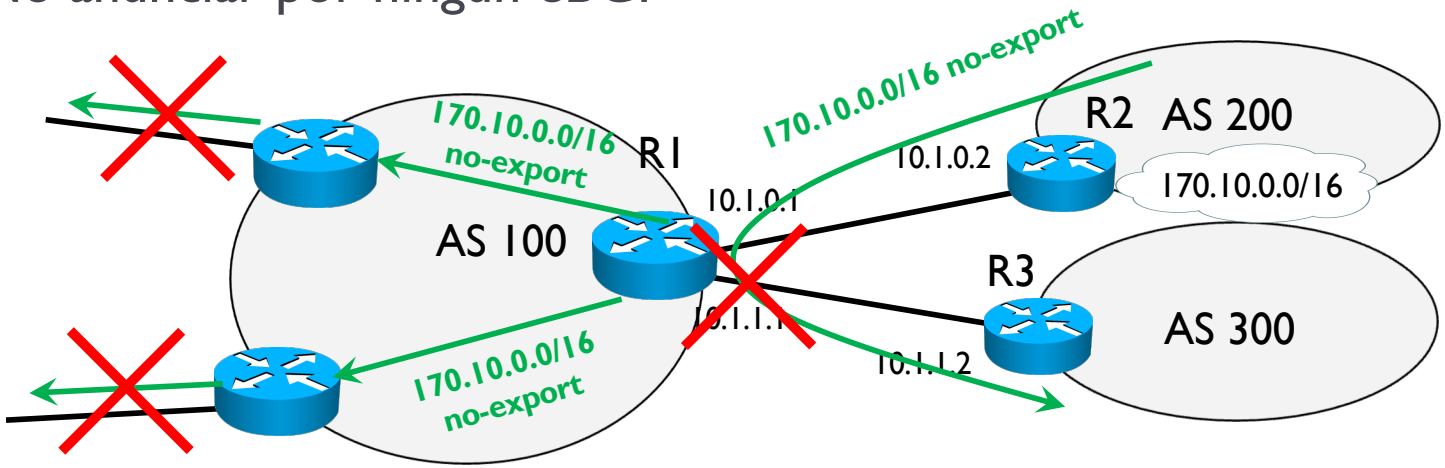
Todos los demás prefijos se filtran por la regla por defecto
Para dejar los demás prefijos sin cambios

```
ip community-list 3 permit Internet
route-map COM permit 100
match community 3
```

5.10 - Mejoras del BGP

Community

- Hay comunidades ya definidas por defecto
- no-export 65535:65281
  - No anunciar por ningún eBGP



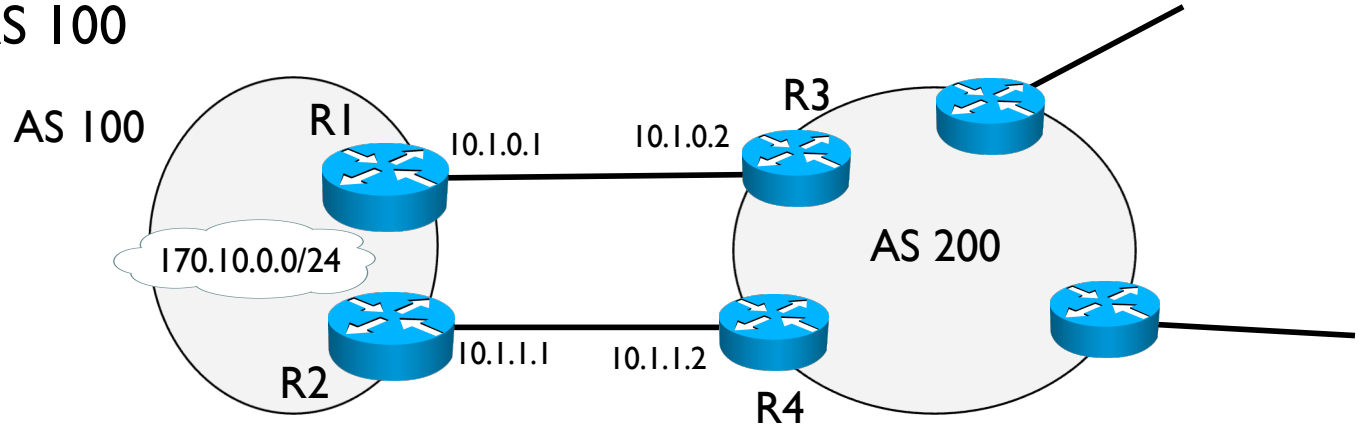

```
R2# access-list 1 permit 170.10.0.0/16

R2# route-map NONC permit 10
R2# match ip address 1
R2# set community no-export
```

```
R2# router bgp 200
R2# network 170.10.0.0/16
R2# neighbor 10.1.0.1 remote-as 100
R2# neighbor 10.1.0.1 send-community
R2# neighbor 10.1.0.1 route-map NONC out
```

5.10 - Mejoras del BGP

Community para stub multi-homed

- Ejemplo AS 100



- Hacer balanceo de carga y protección a la vez
  - R1 anuncia la mitad de su prefijo 170.10.0.0/25 a R3 con no-export
  - R1 anuncia el prefijo entero 170.10.0.0/24 a R3
  - R2 anuncia la otra mitad 170.10.0.128/25 a R4 con no-export
  - R2 anuncia el prefijo entero 170.10.0.0/24 a R4
  - R1 acepta algunos prefijos de R3 (por ejemplo los <121.0.0.0/8) y configura una ruta por defecto a R3
  - R2 acepta los otros de prefijos de R4 (por ejemplo los >=121.0.0.0/8) y configura una ruta por defecto a R4

5.10 - Mejoras del BGP
Community

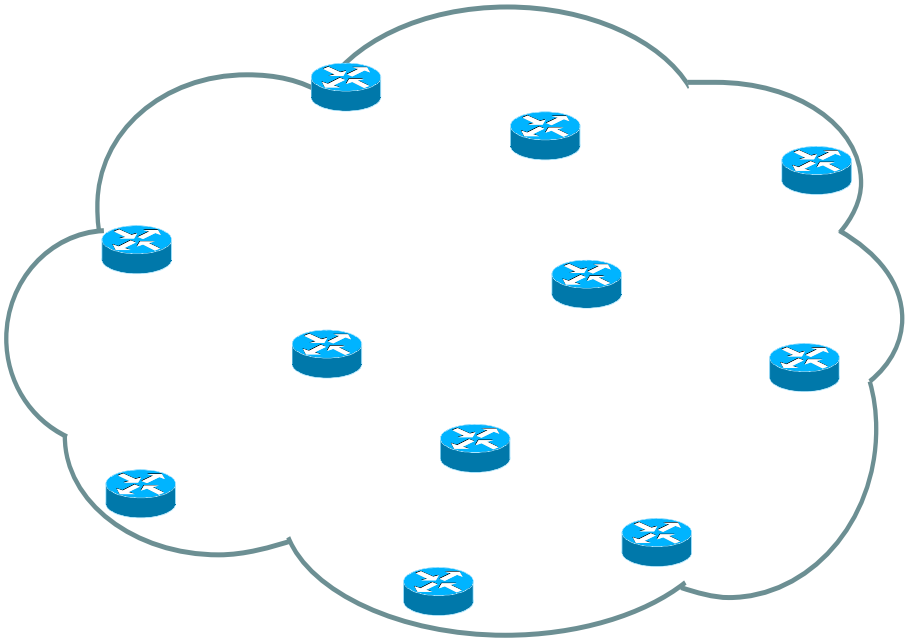
- Hay comunidades ya definidas por defecto
- no-advertise 65535:65282
  - No anunciar por ningún BGP (ni eBGP, ni iBGP)
  - Se lo queda solo el primer router
- no-export-subconfed 65535:65283
  - No anunciar a ningún otro router fuera del sub-AS
  - Usado en confederación de sub-AS

5.10 - Mejoras del BGP
Escalabilidad del iBGP

- Uno de los grandes problemas de BGP es la escalabilidad
  - Es un protocolo de los años 80 cuando Internet no tenía el tamaño que tiene hoy (ni se preveía)
- En el caso de iBGP
  - Se necesita un full-mesh de sesiones iBGP
  - Cada sesión necesita una conexión TCP
  - Demasiados mensajes "innecesarios"
  - Por ejemplo, un AS con 100 routers
    - Cada router mantiene 99 conexiones TCP
    - 9900/2 sesiones en total en el AS
  - Hoy en día hay ASes con grado 5,000
- Solución
  - Quitar la necesidad de tener una malla completa de sesiones iBGP

5.10 - Mejoras del BGP
Route Reflection

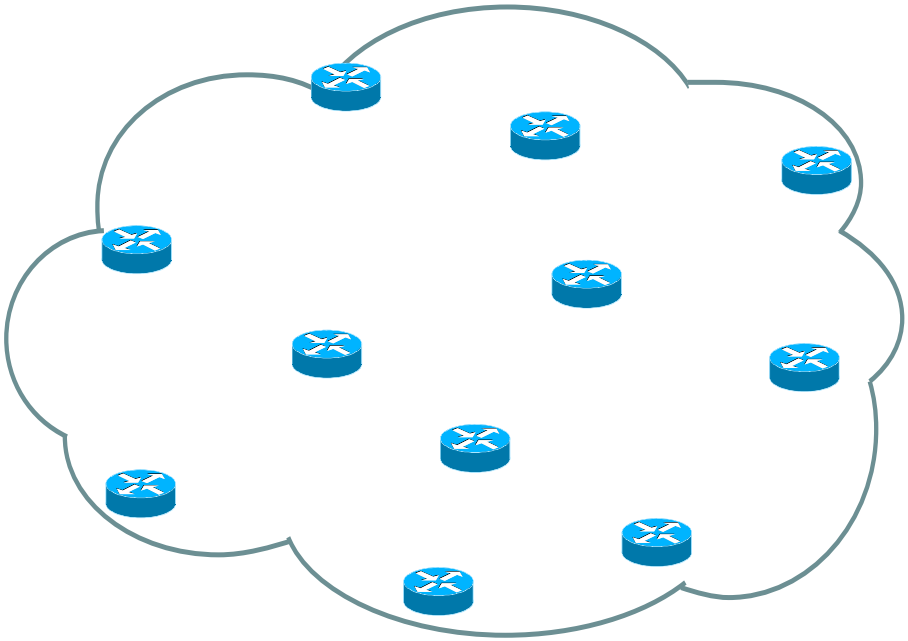
- Se organiza en AS en clusters
  - En cada cluster, se elige 1 o más routers llamados Route Reflector (RR)
  - El resto de routers de un cluster serán clientes de estos RR



- Se supone que todos son routers BGP
  - Tienen una sesión hacia otro AS (conexiones no dibujadas en la figuras)
  - Hay otros routers internos no-BGP que tampoco se han dibujado
- Sin esta mejora, se necesitarían
  - … sesiones BGP en total

5.10 - Mejoras del BGP
Route Reflection

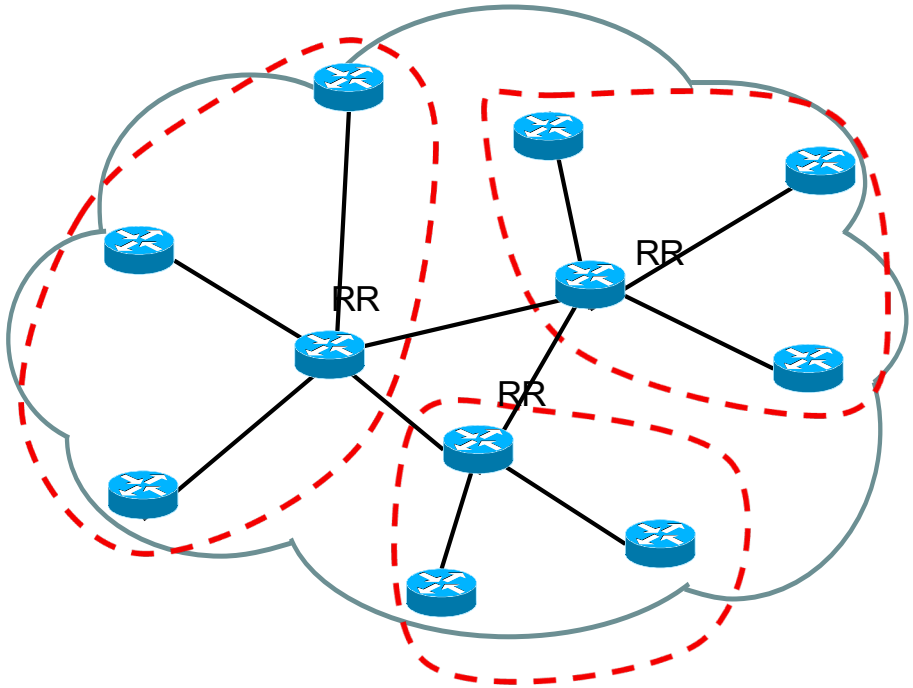
- Se organiza en AS en clusters
  - En cada cluster, se elige 1 o más routers llamados Route Reflector (RR)
  - El resto de routers de un cluster serán clientes de estos RR



- Se supone que todos son routers BGP
  - Tienen una sesión hacia otro AS (conexiones no dibujadas en la figuras)
  - Hay otros routers internos no-BGP que tampoco se han dibujado
- Sin esta mejora, se necesitarían
  - 11x10 / 2 = 55 sesiones BGP en total

5.10 - Mejoras del BGP
Route Reflection

- Se organiza en AS en clusters
  - En cada cluster, se elige 1 o más routers llamados Route Reflector (RR)
  - El resto de routers de un cluster serán clientes de estos RR

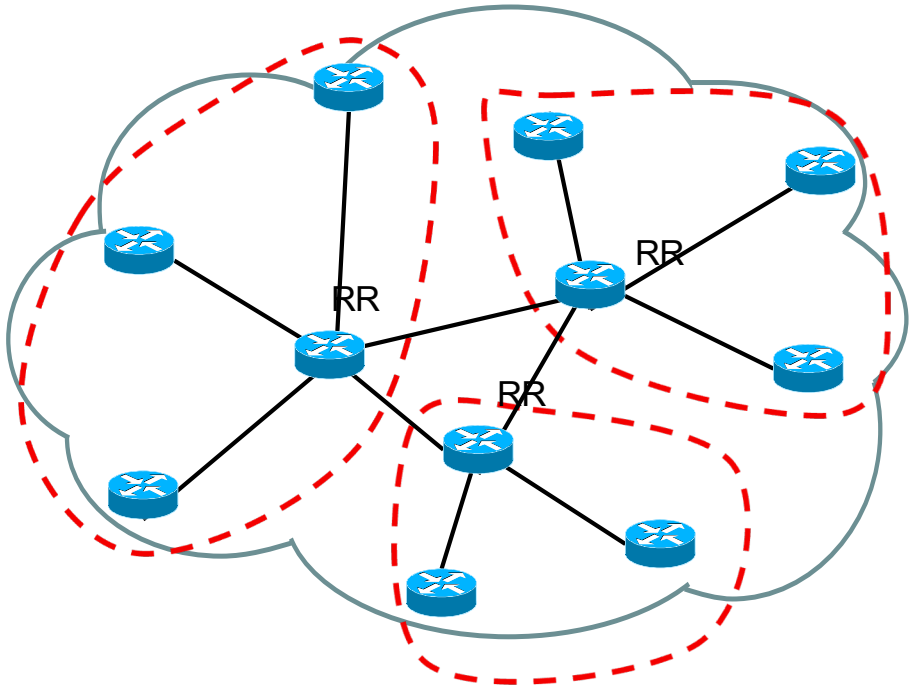


- Con RR, se crean los clusters
  - Se elige 1 RR por cluster
  - Los otros routers se conectan con 1 única sesión al RR de su cluster
  - Los RR se conectan entre sí en malla completa

5.10 - Mejoras del BGP

Route Reflection

- Se organiza en AS en clusters
  - En cada cluster, se elige 1 o más routers llamados Route Reflector (RR)
  - El resto de routers de un cluster serán clientes de estos RR

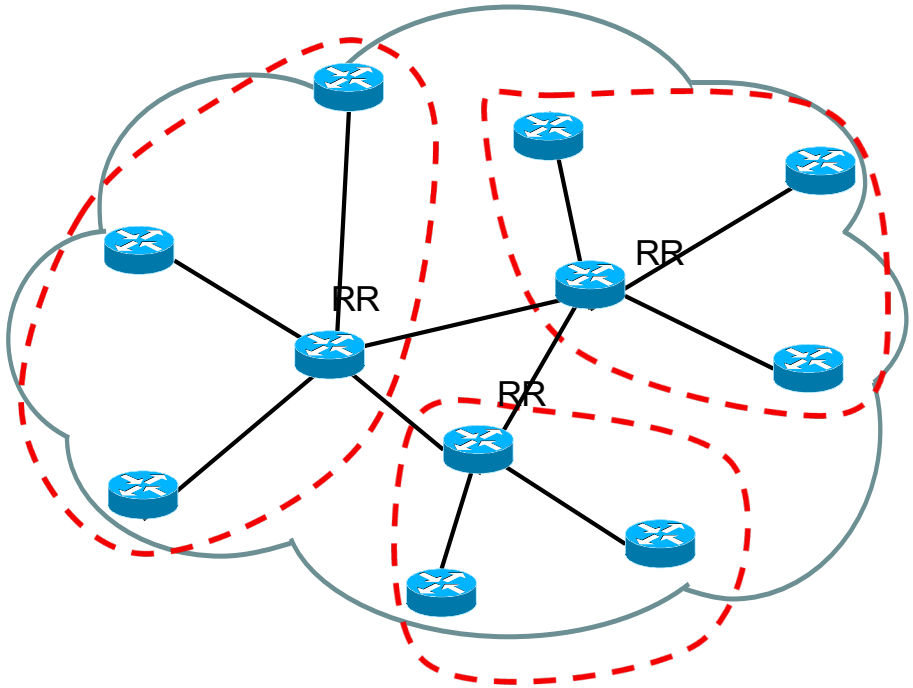


- Con RR, se crean los clusters
  - Se elige 1 RR por cluster
  - Los otros routers se conectan con 1 única sesión al RR de su cluster
  - Los RR se conectan entre sí en malla completa
- Reglas para re-enviar BGP updates
  - Router cliente
    - Si se recibe por eBGP, lo puede re-enviar por eBGP a otros AS y al RR de su cluster
    - Si se recibe por iBGP, solo puede re-enviarlo por eBGP
  - Router RR
    - Lo debe re-enviar a todos excepto por donde ha venido

5.10 - Mejoras del BGP

Route Reflection

- Se organiza en AS en clusters
  - En cada cluster, se elige 1 o más routers llamados Route Reflector (RR)
  - El resto de routers de un cluster serán clientes de estos RR

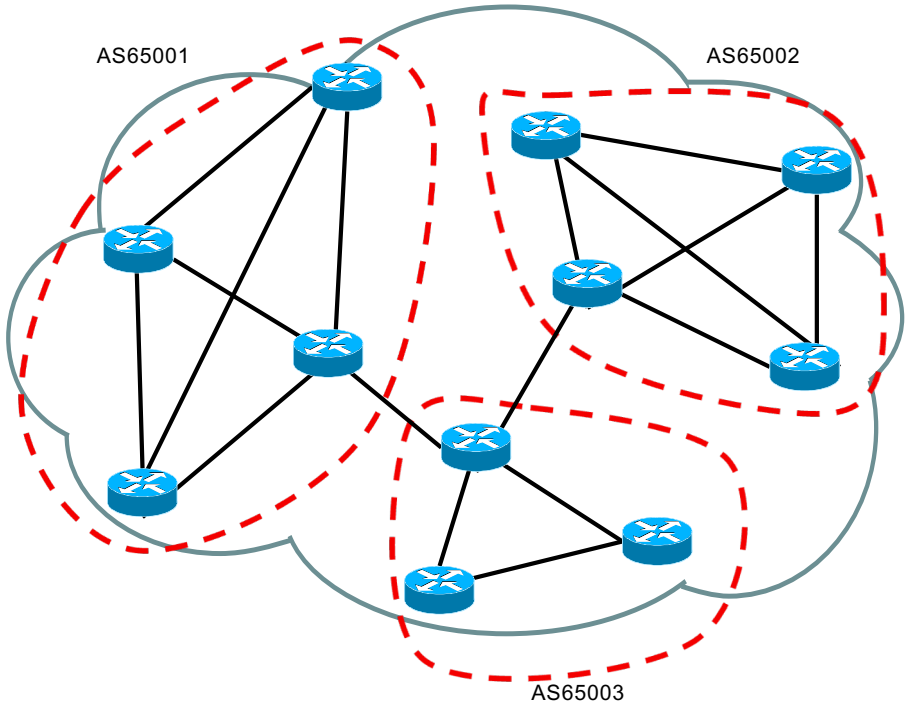


- Con RR, se crean los clusters
  - Se elige 1 RR por cluster
  - Los otros routers se conectan con 1 única sesión al RR de su cluster
  - Los RR se conectan entre sí en malla completa
- Reglas para re-enviar BGP updates
  - Router cliente
    - Si se recibe por eBGP, lo puede re-enviar por eBGP a otros AS y al RR de su cluster
    - Si se recibe por iBGP, solo puede re-enviarlo por eBGP
  - Router RR
    - Lo debe re-enviar a todos excepto por donde ha venido
- Se necesitan 2 nuevos atributos para evitar bucles
  - originator-id: RID del router que ha creado el mensaje BGP. Lo asigna el RR del cluster
  - cluster-list: parecido al AS-path entre AS, pero usando identificadores para los clusters

5.10 - Mejoras del BGP

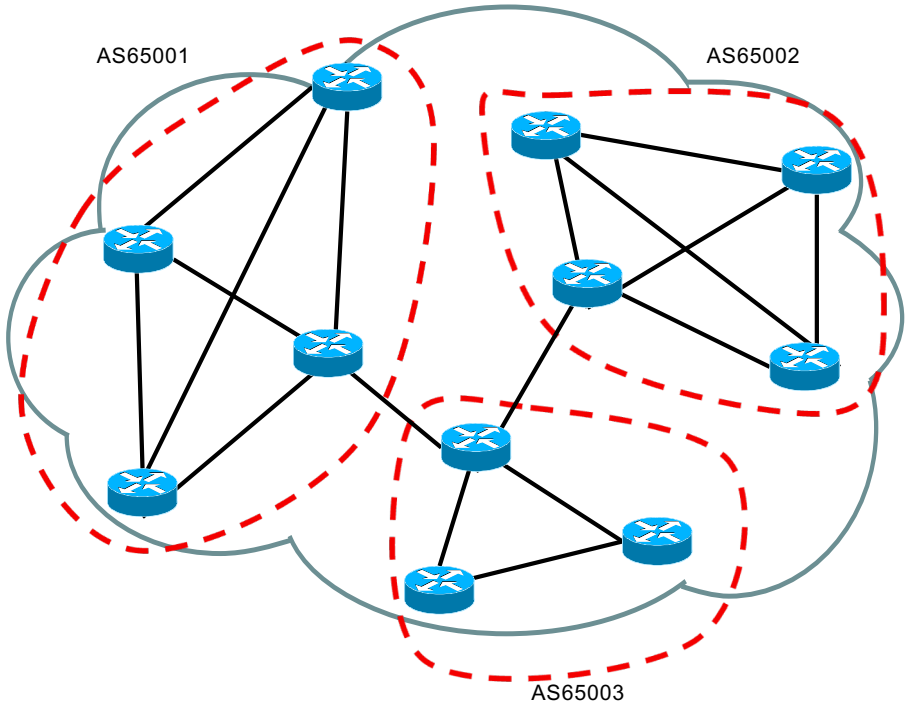
Sub-AS confederation

- Se divide el AS en sub-AS
  - Cada sub-AS es como si fuera un AS y se mantiene la malla completa
  - Se usan ASN privados para los sub-AS



5.10 - Mejoras del BGP
Sub-AS confederation

- Se divide el AS en sub-AS
  - Cada sub-AS es como si fuera un AS y se mantiene la malla completa
  - Se usan ASN privados para los sub-AS

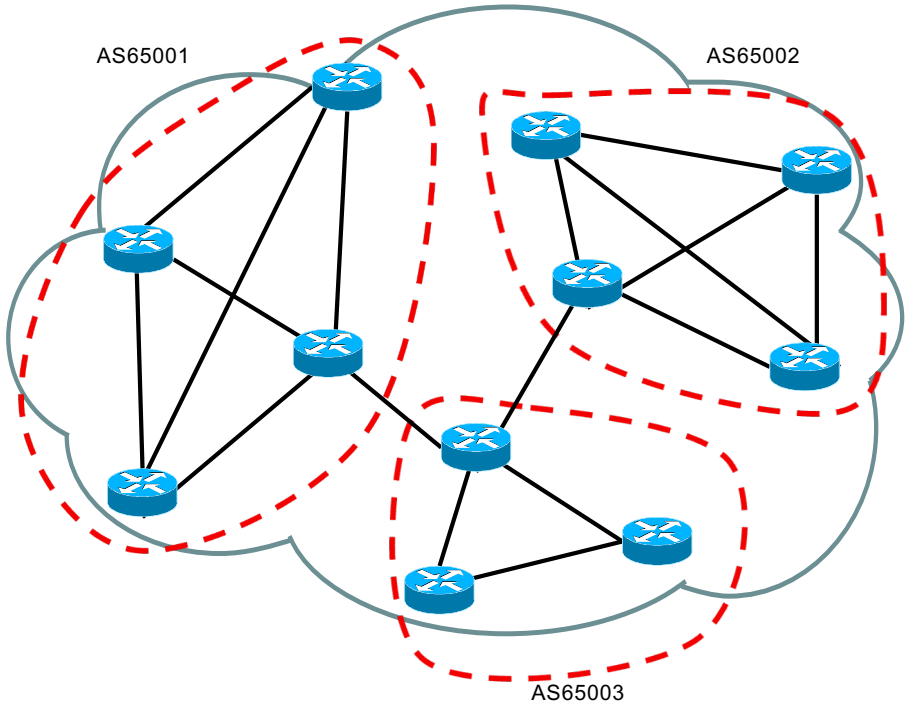


- En cada sub-AS
  - Hay una malla completa
  - Las sesiones BGP siguen llamándose iBGP
- Entre sub-AS
  - Se abren las sesiones que se consideran necesarias
  - Estas sesiones BGP se suelen llamar eiBGP

5.10 - Mejoras del BGP

Sub-AS confederation

- Se divide el AS en sub-AS
  - Cada sub-AS es como si fuera un AS y se mantiene la malla completa
  - Se usan ASN privados para los sub-AS



- En cada sub-AS
  - Hay una malla completa
  - Las sesiones BGP siguen llamándose iBGP
- Entre sub-AS
  - Se abren las sesiones que se consideran necesarias
  - Estas sesiones BGP se suelen llamar eiBGP
- Reglas para re-enviar BGP updates
  - Si se recibe por eBGP, se puede re-enviar por iBGP, eiBGP y por eBGP a otros AS
  - Si se recibe por eiBGP, se puede re-enviar por iBGP, eiBGP y por eBGP a otros AS
  - Si se recibe por iBGP, se puede re-enviar solo eiBGP y por eBGP a otros AS

5.10 - Mejoras del BGP
Route Flap Damping

- Problema
  - Los route flaps hacen que la convergencia de los routers BGP sea muy lenta y hacen el sistema más inestable

- Route flap: cuando un router
  - Anuncia y elimina constantemente uno o más prefijos
  - Cambia continuamente el valor de uno o más atributos que modifican la selección de rutas

5.10 - Mejoras del BGP
Route Flap Damping

- Por lo tanto, ya que un router BGP envía un BGP update cada vez que se añade o se elimina un prefijo o se cambia un atributo, el route flapping causa
  - Consumo de CPU
  - Cortes de servicio
  - Propagación del problema en todo Internet

- Causas más comunes que generan route flapping
  - Errores hardware, p.e. interfaces muy desgastadas
  - Errores software, p.e. un bug en el SO
  - Malas configuraciones del router
  - Cables en mal estado

5.10 - Mejoras del BGP
Route Flap Damping

- Solución propuesta
  - Route Flap Damping (RFD) RFC 2439

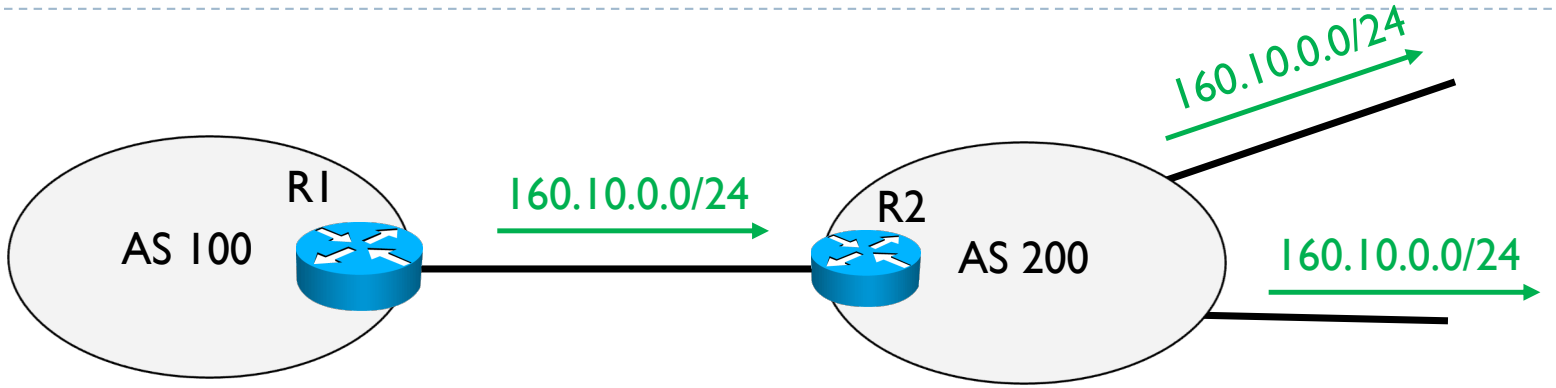
- Idea
  - Parar de anunciar los prefijos que causan "demasiados" flaps
  - ¿Como definir "demasiado"? Se usan penalidades

5.10 - Mejoras del BGP
Route Flap Damping

- Funcionamiento por prefijo
  - Se definen 4 parámetros
    - Reuse-limit (valor por defecto 750)
    - Suppress-limit (2000)
    - Half-life (15 minutos)
    - Maximum suppress-limit (60 minutos)
- Reglas
  - Penalidad de 1000 por cada flap y de 500 por cambio de atributo
  - Si penalidad > suppress-limit → el prefijo no se anuncia
  - Si penalidad < reuse-limit → el prefijo se vuelve a anunciar
  - La penalidad se reduce en el tiempo según en exponencial función de Half-life
  - Penalidad = 0 si no hay cambios durante un Maximum suppress-limit

5.10 - Mejoras del BGP

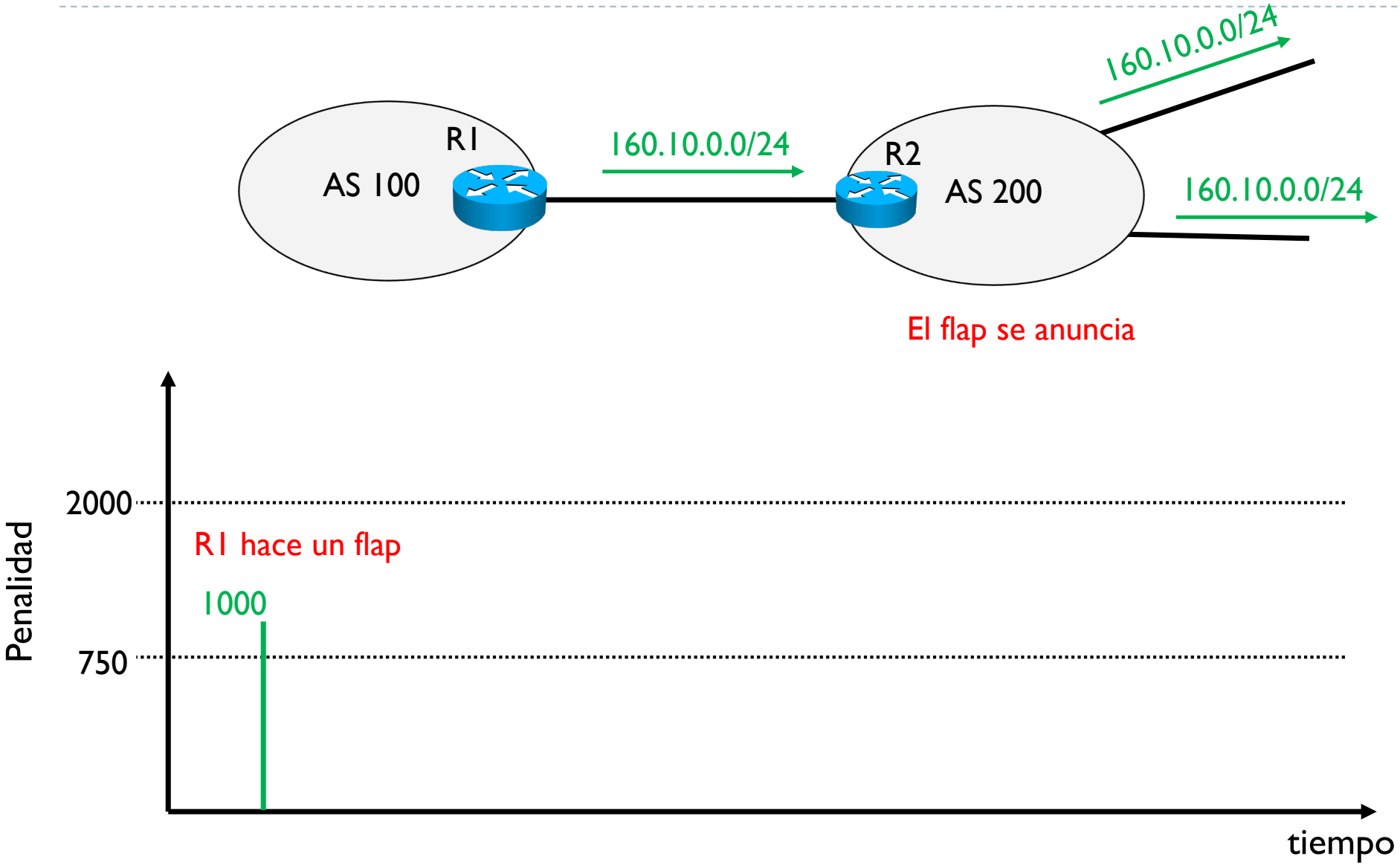
Ejemplo de funcionamiento del RFD



El prefijo se anuncia

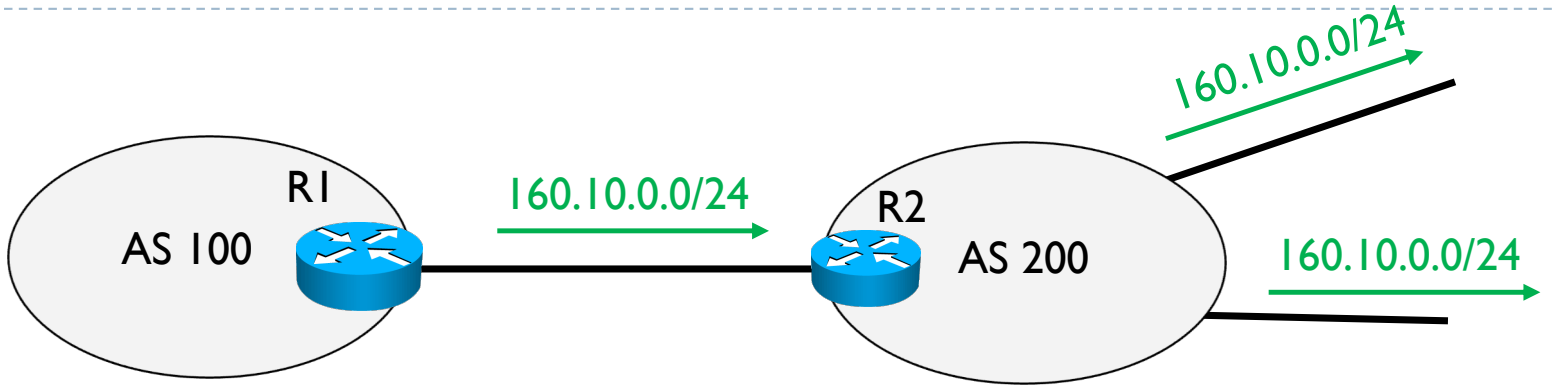
5.10 - Mejoras del BGP

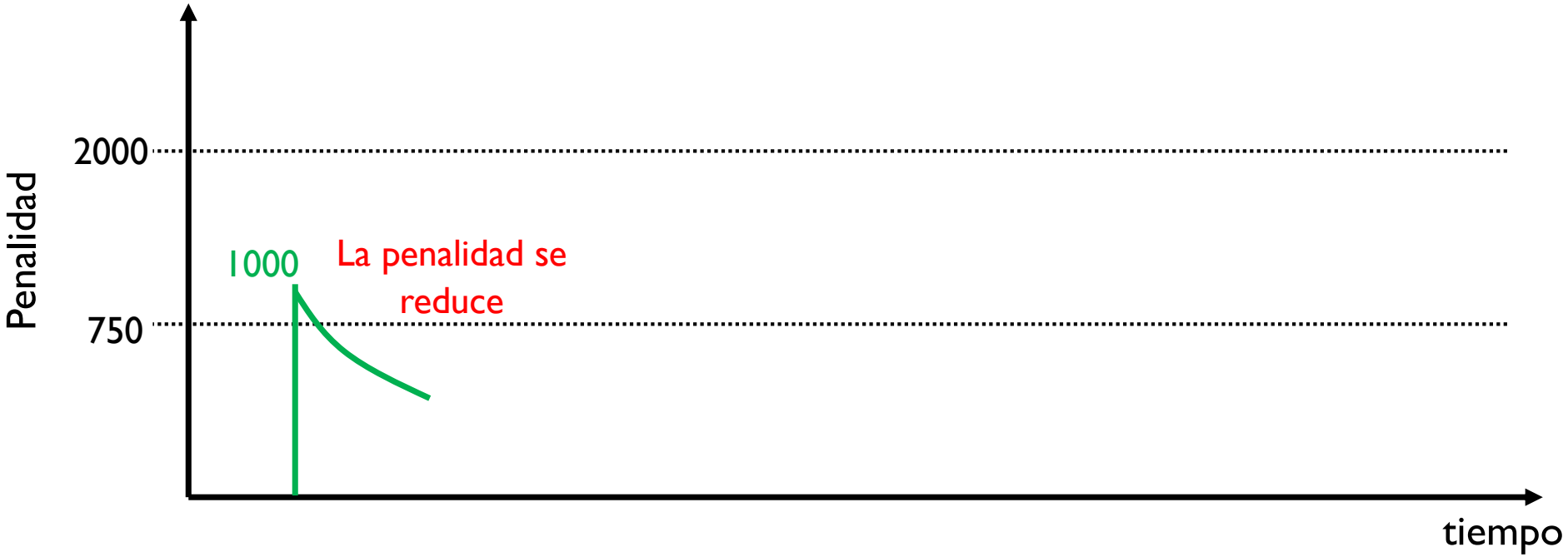
Ejemplo de funcionamiento del RFD



5.10 - Mejoras del BGP

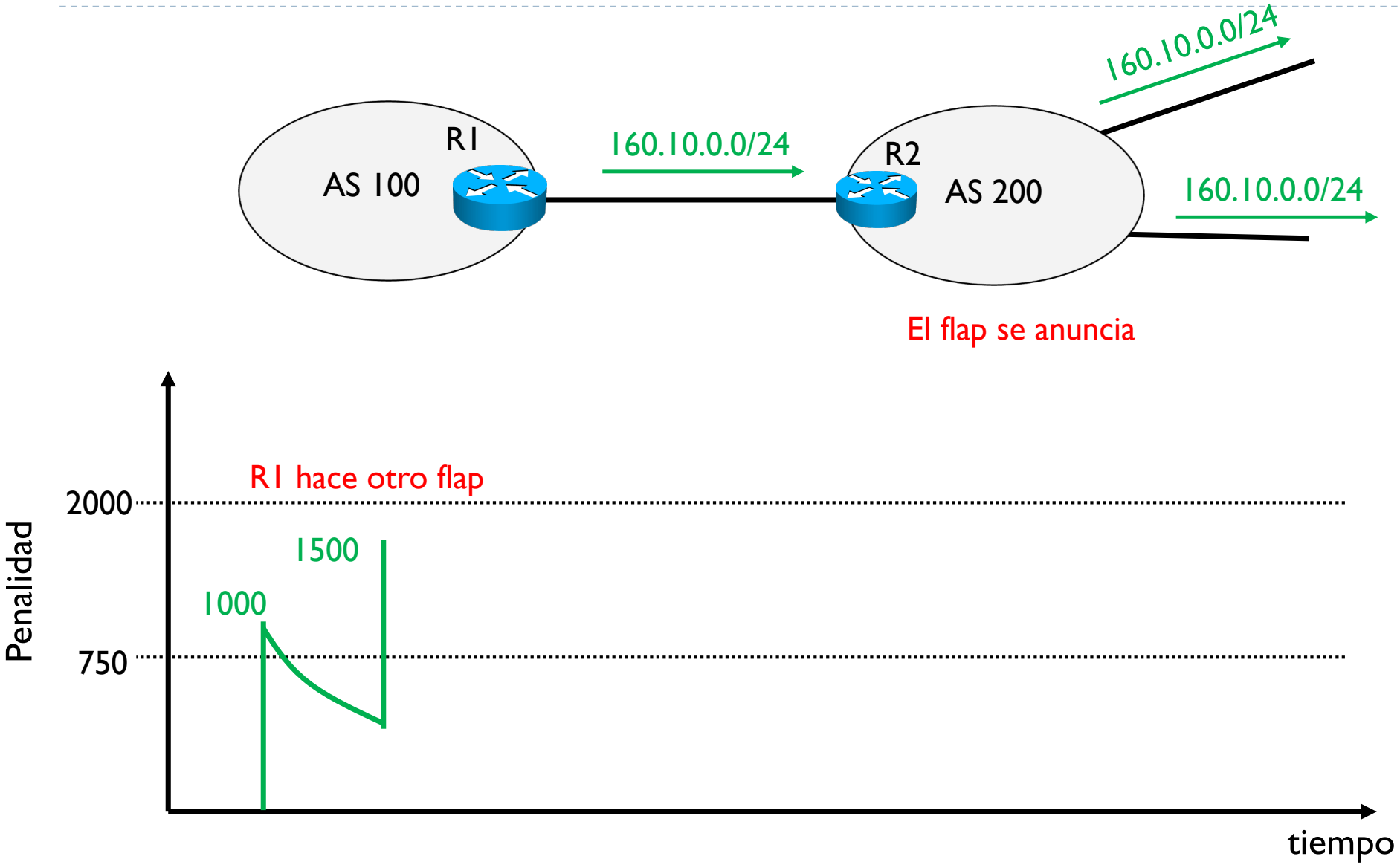
Ejemplo de funcionamiento del RFD





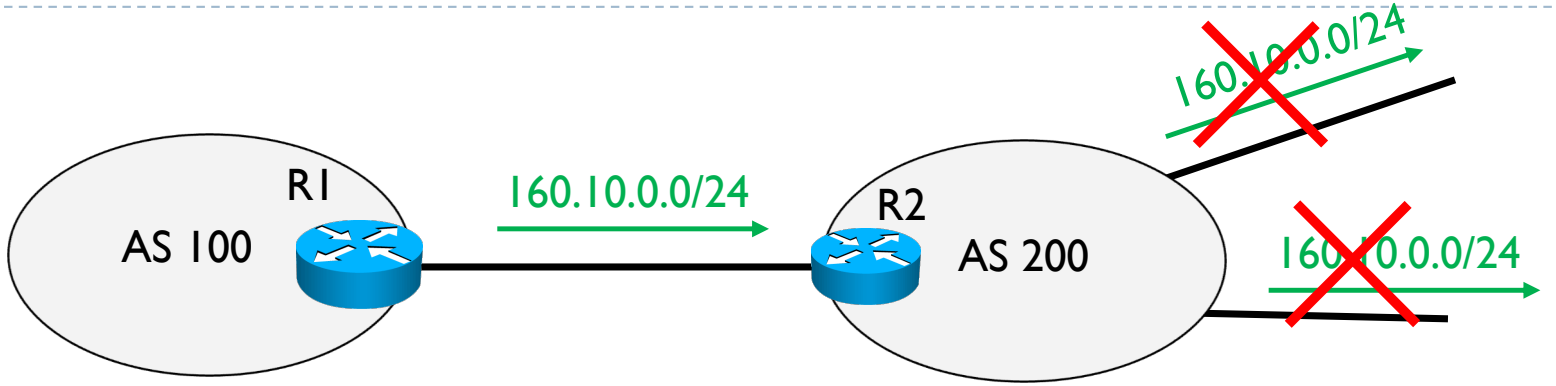
5.10 - Mejoras del BGP

Ejemplo de funcionamiento del RFD

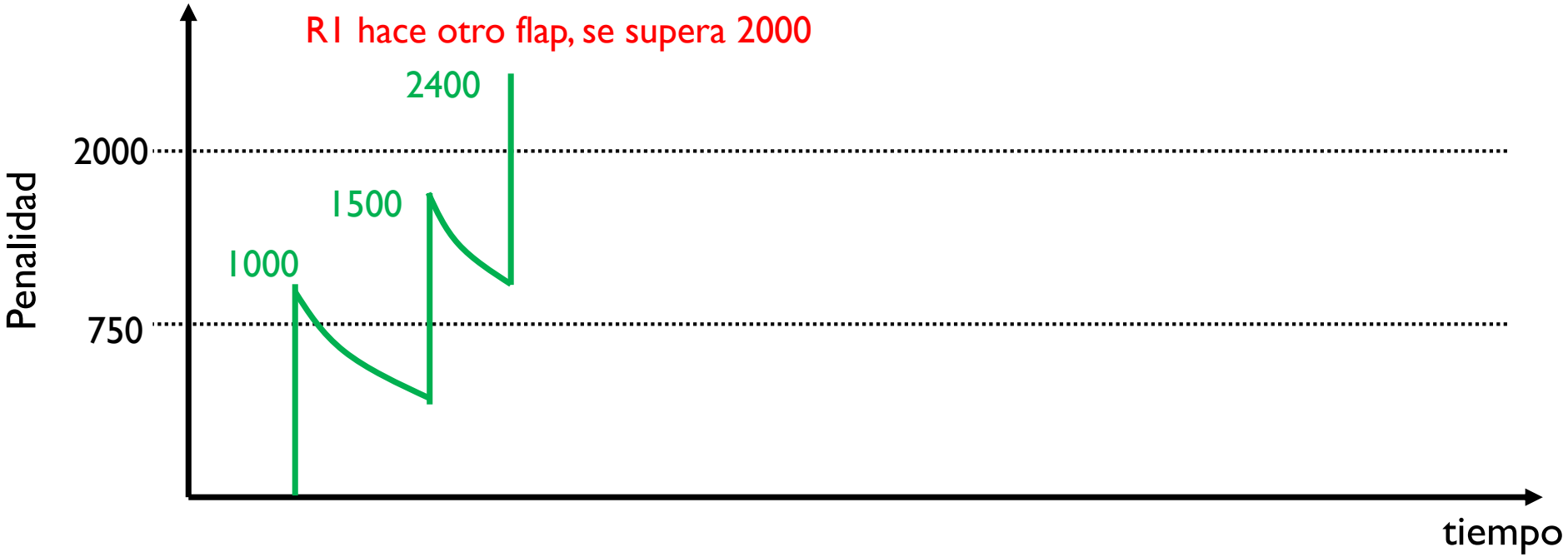


5.10 - Mejoras del BGP

Ejemplo de funcionamiento del RFD

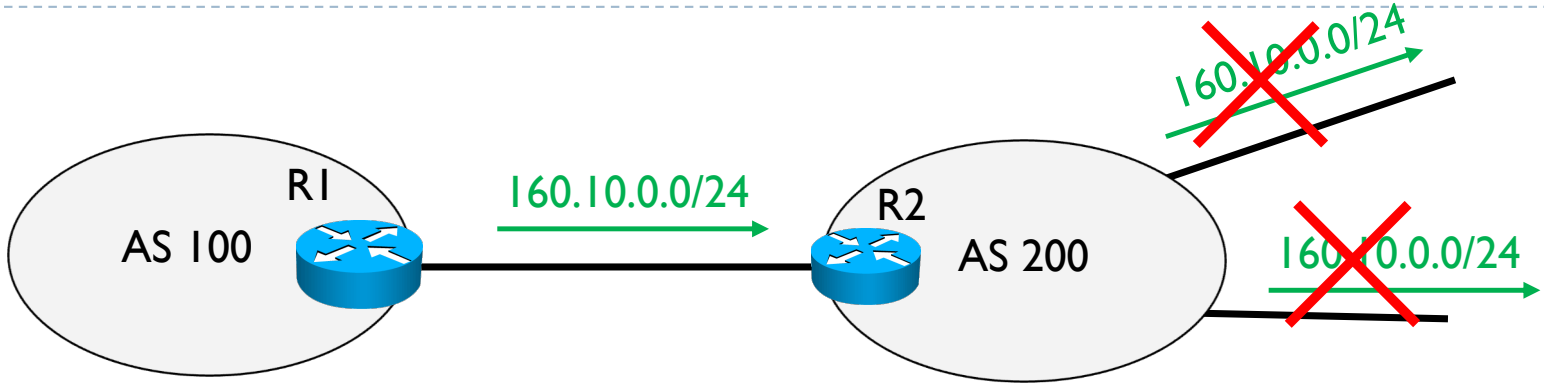


El flap ya no se anuncia

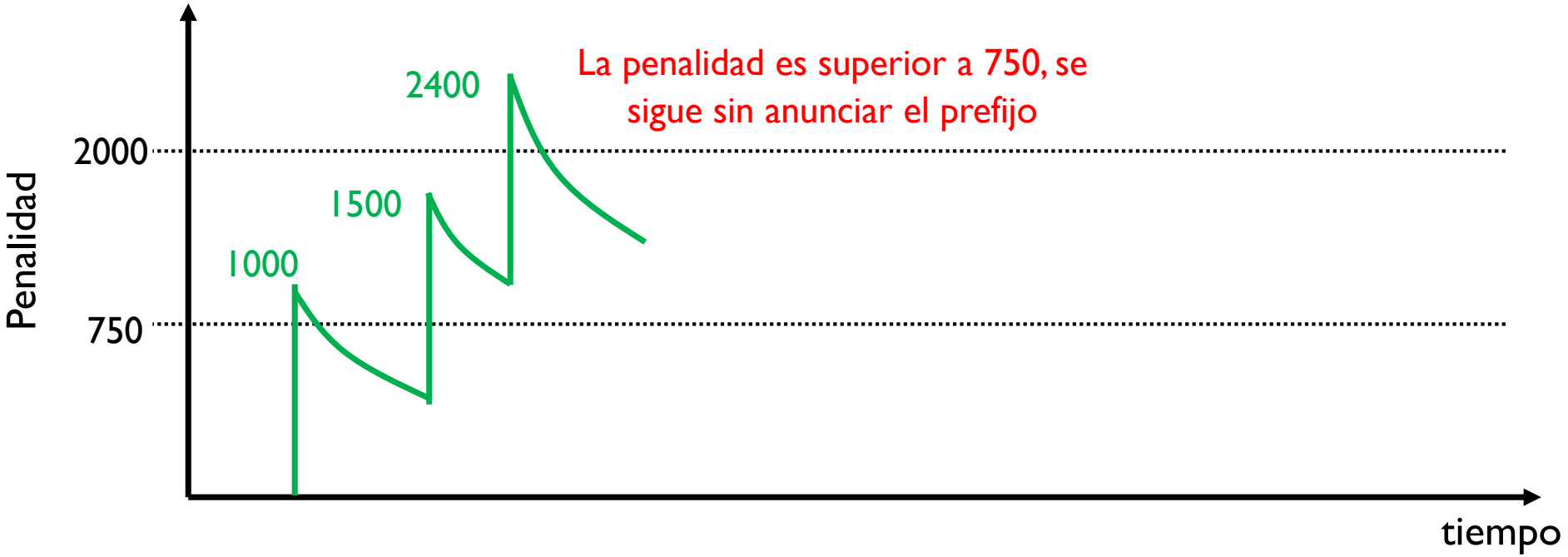


5.10 - Mejoras del BGP

Ejemplo de funcionamiento del RFD

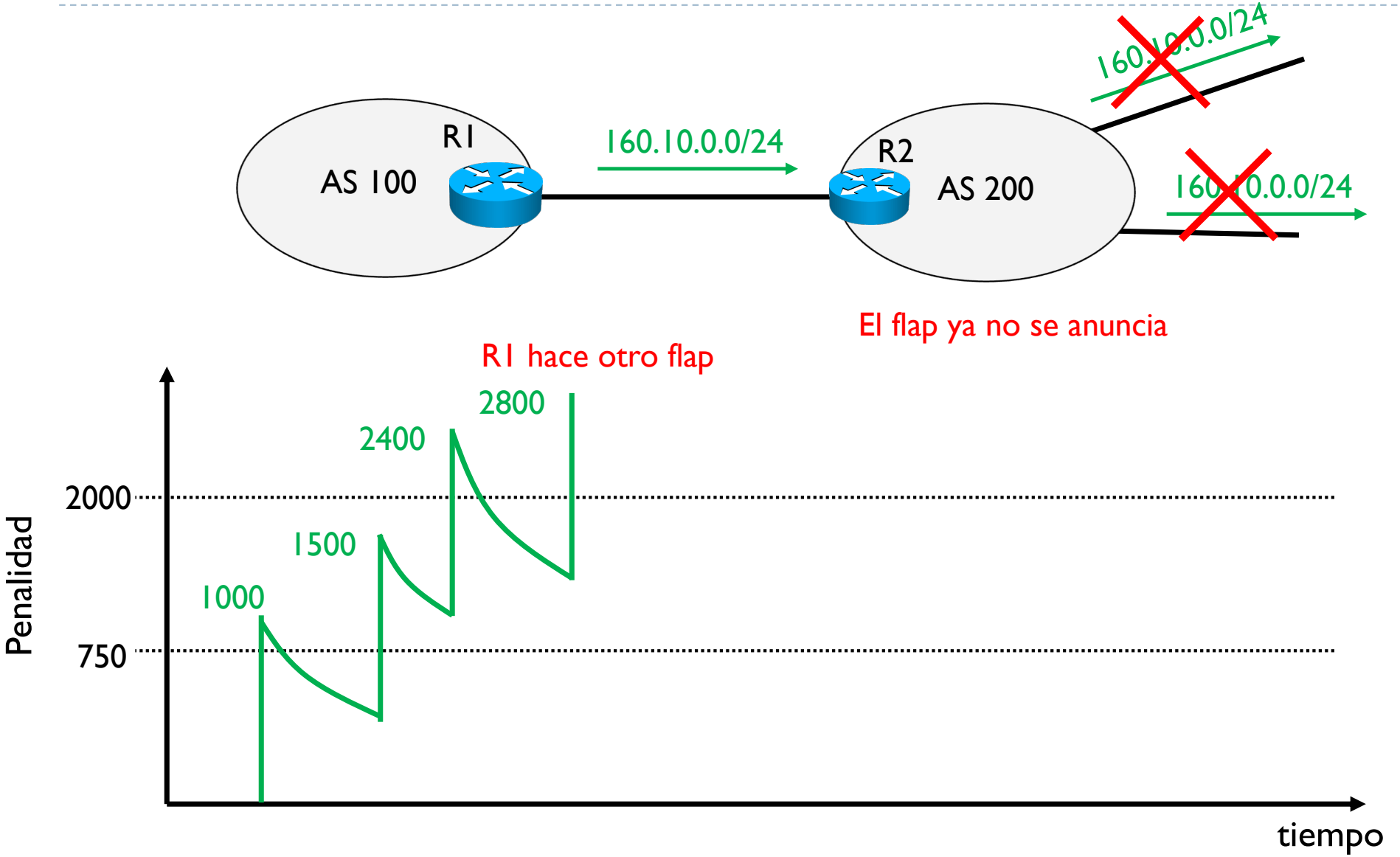


El flap ya no se anuncia



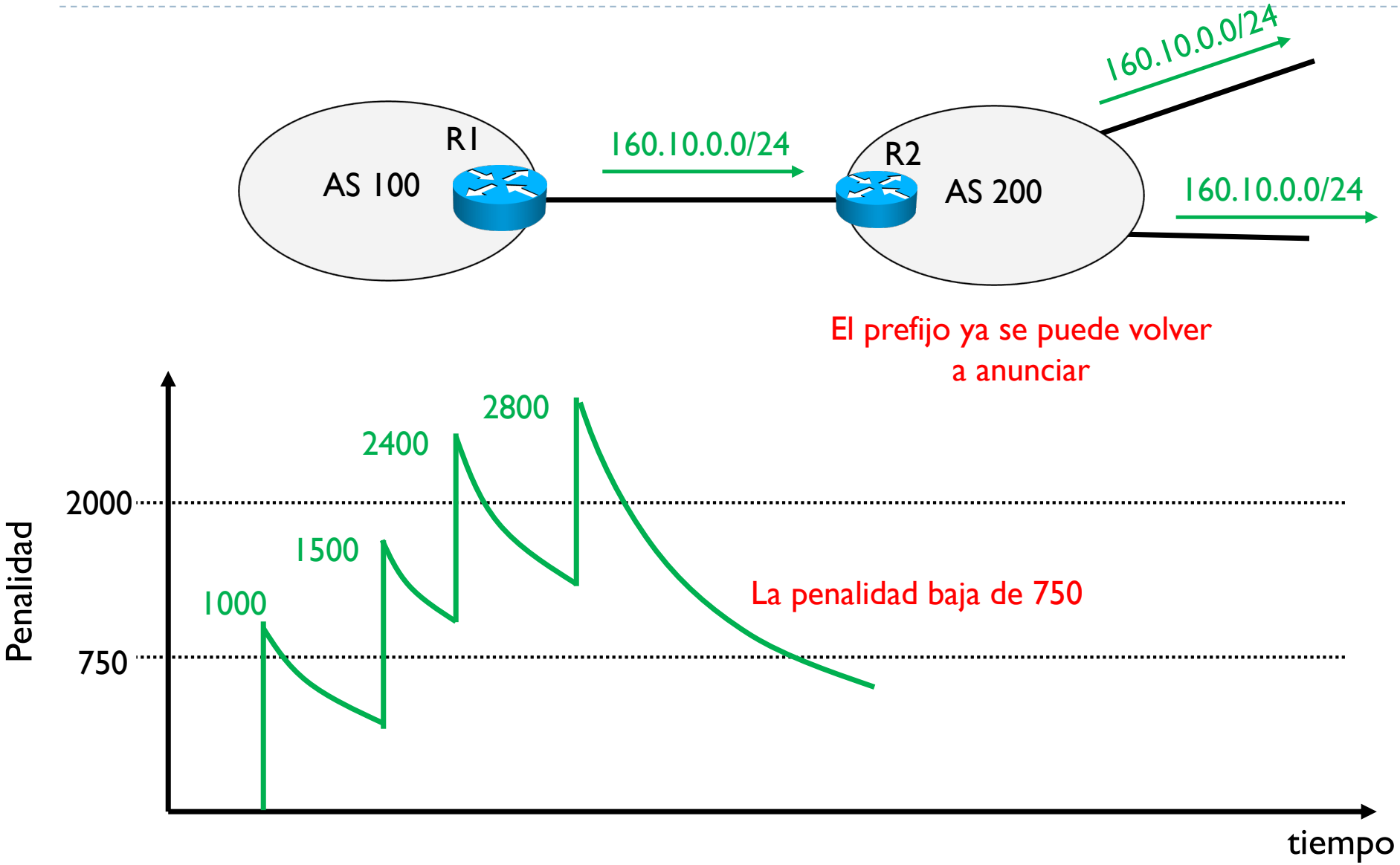
5.10 - Mejoras del BGP

Ejemplo de funcionamiento del RFD



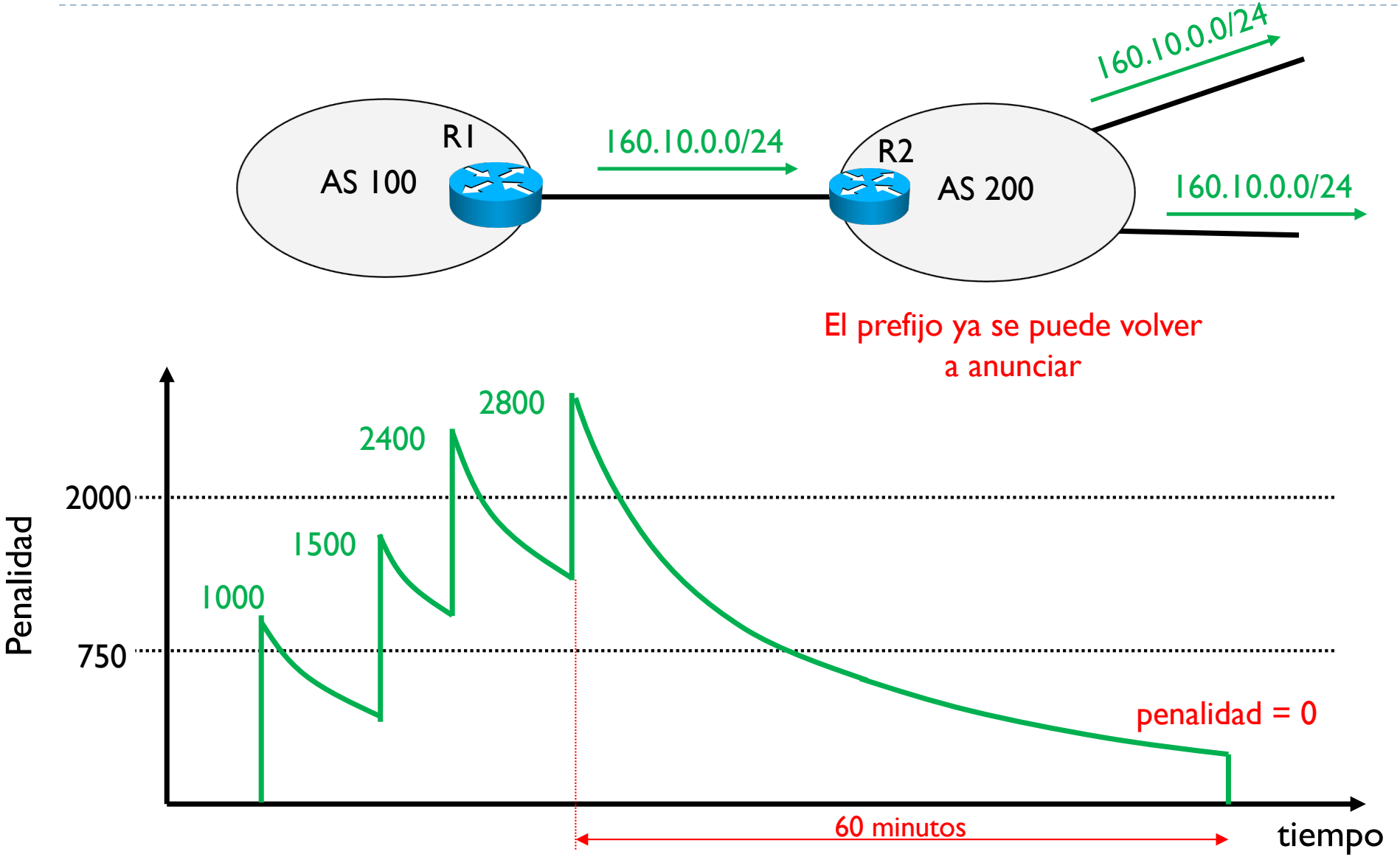
5.10 - Mejoras del BGP

Ejemplo de funcionamiento del RFD



5.10 - Mejoras del BGP

Ejemplo de funcionamiento del RFD



5.10 - Mejoras del BGP
Route Flap Damping

- Del 1998 (estándar) al 2006
  - Se empezó a usar en todos los AS
- Del 2006 al 2013
  - Se consideró que RFD causaba más problemas que soluciones
  - No anunciar un prefijo causaba otros flaps inecesarios
  - Además diferentes routers podían implementar valores diferentes, creando inestabilidad
  - Además los routers eran más potentes que en los 90 y podían aguantar tantos BGP updates
  - Se recomendó no usarlo

5.10 - Mejoras del BGP
Route Flap Damping

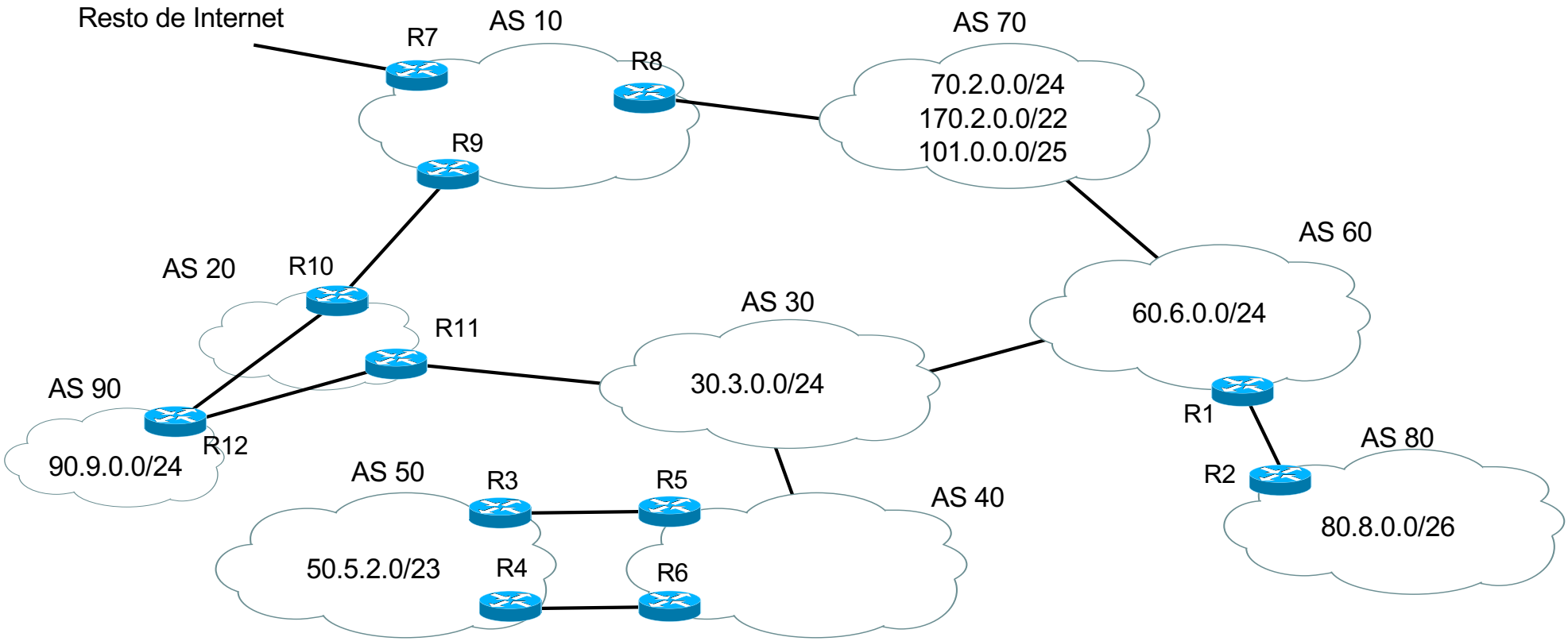
- Del 2013 a hoy
  - Ripe-580
  - Se recomienda volver a usarlo pero con suppress-limit a 6000
  - Una investigación descubrió que el 3% de los prefijos causan el 36% de BGP updates y el 0.01% el 10% de updates
  - Aumentando el umbral, se penalizan aquellos pocos prefijos que causan el mayor número de problemas

Xarxes de Computadors II

Tema 5: Encaminamiento inter-dominio: BGP

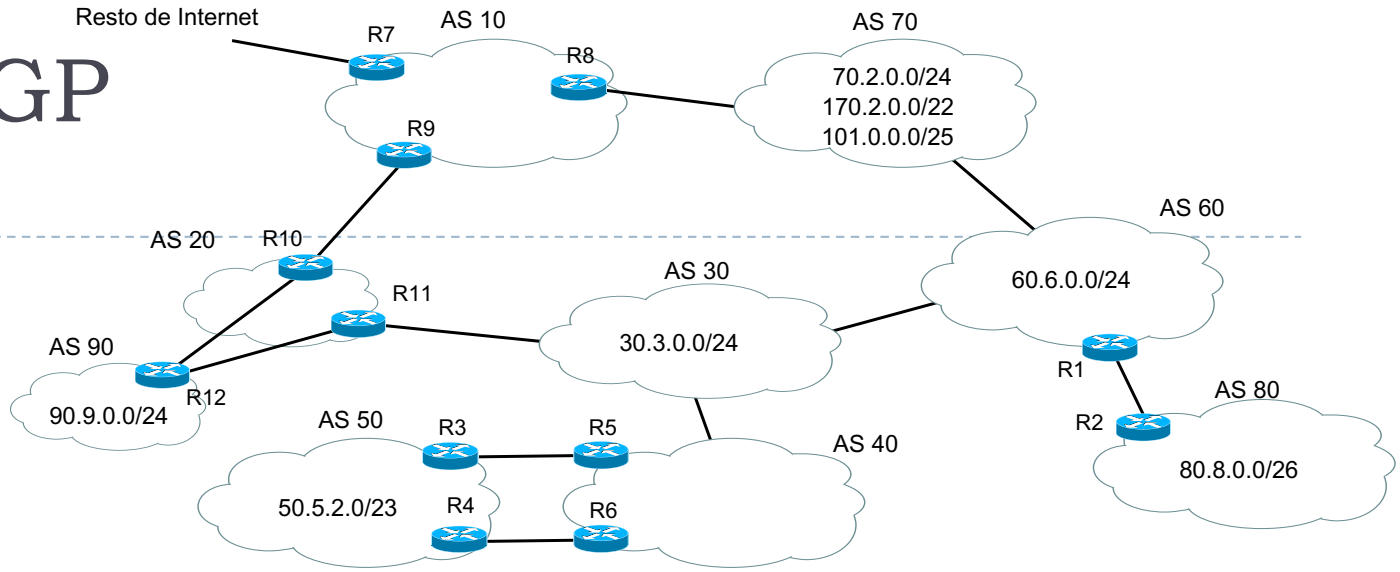
Davide Careglio

Tema 5 - BGP
Problemas



Tema 5 - BGP

Problemas



1. De que tipo son los AS 90, 80, 50 y 10
2. Indicar como se configuraría el AS 80
3. Indicar como se configuraría el AS 50 para tener balanceo de carga y protección
4. Identificar los prefijos que anuncia el AS 10 al resto de Internet, indicando el AS-path
5. Indicar como cambiaría el punto 4. si el AS60 fuese multihomed con el AS 70 y 30
6. Indicar como cambiaría el punto 4. si el AS70 fuese multihomed con el AS 10 y 60
7. Suponiendo que el AS 70 es de transito, indicar como habría que configurar el AS 10 para que usara la ruta 20-30-60-80 para llegar al 80.8.0.0/26
8. Indicar si se podría conseguir lo mismo que el punto 7. pero desde el AS 70
9. Indicar si se podría conseguir lo mismo que el punto 7. pero desde el AS 20

Tema 5 - BGP

Problemas