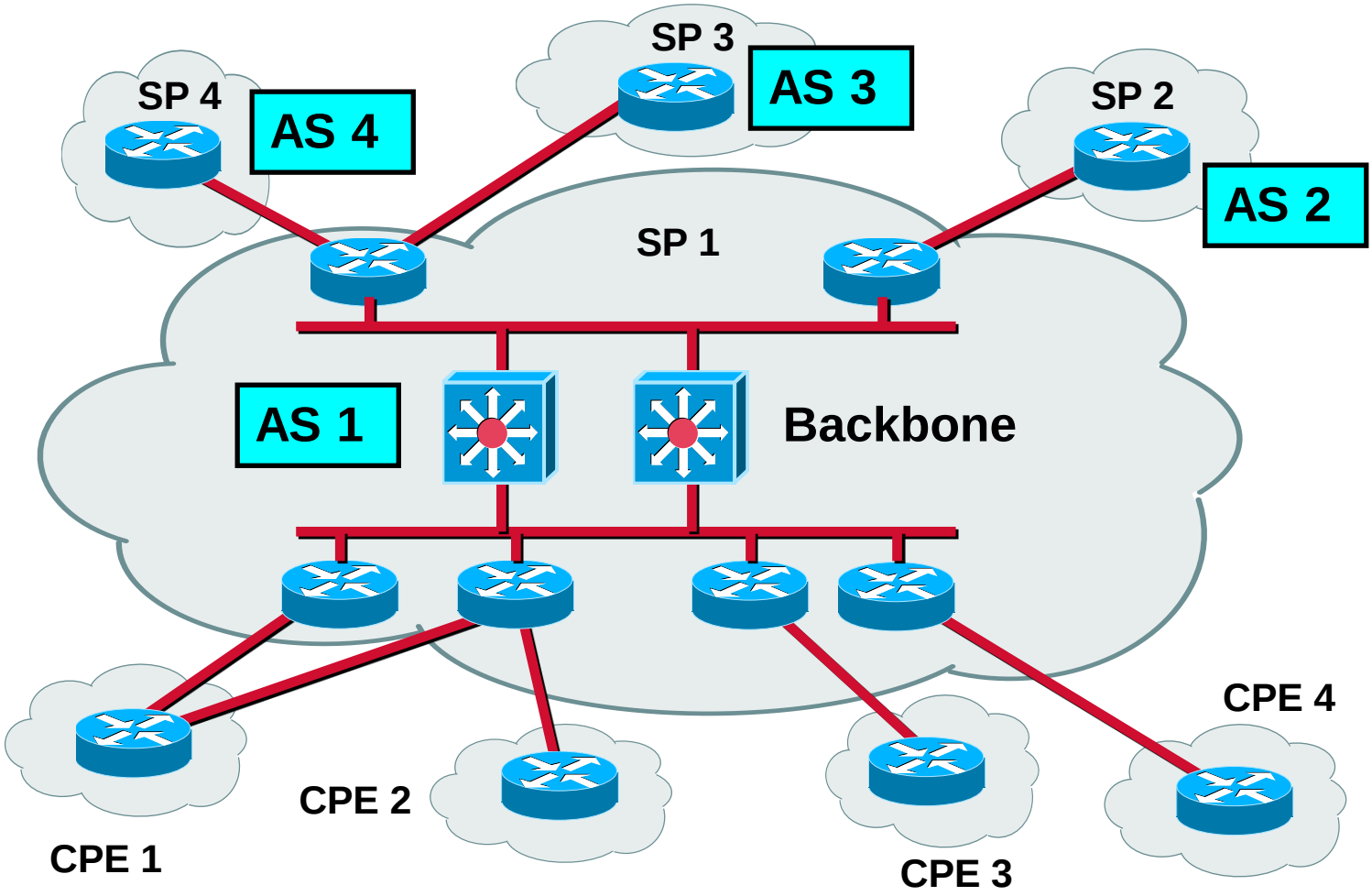
BGP

Border Gateway Protocol

Agenda (1)

- **Conceptos Fundamentales de BGP**
- **Análisis del protocolo (BGP-4)**
- **Atributos de BGP y políticas de control**
- **IBGP mesh. Soluciones**
- **Sumarización y anuncios (CIDR)**
- **Damping**
- **Algunos problemas**
- **Extensiones Multiprotocolo**
- **Salidas reales y datos de actualidad**

Paradigma de red de un proveedor



Sistemas Autónomos (AS) (1)

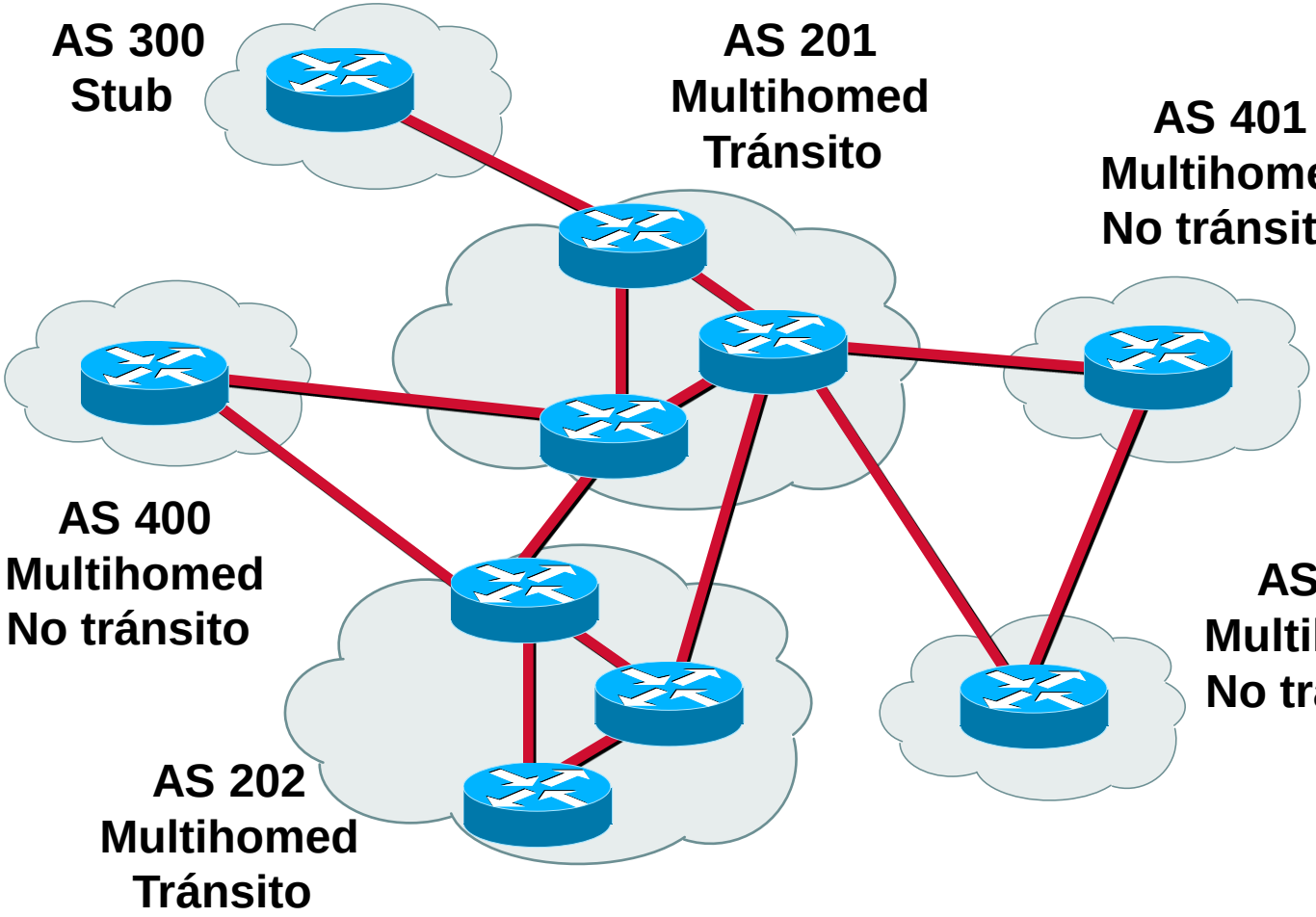
- Conjunto de Redes y enrutadores bajo una política común de enrutamiento, administrados por una única autoridad
- Para el “exterior” el AS se ve como una única entidad
- Cada AS tiene asignado un número en el rango de 1 a 65,535 (privados del 64512 en adelante)

  Ya se definió y comenzó a usarse extensión a 32 bits (RFC 4893)
- Pueden coexistir varios IGP dentro de un mismo AS

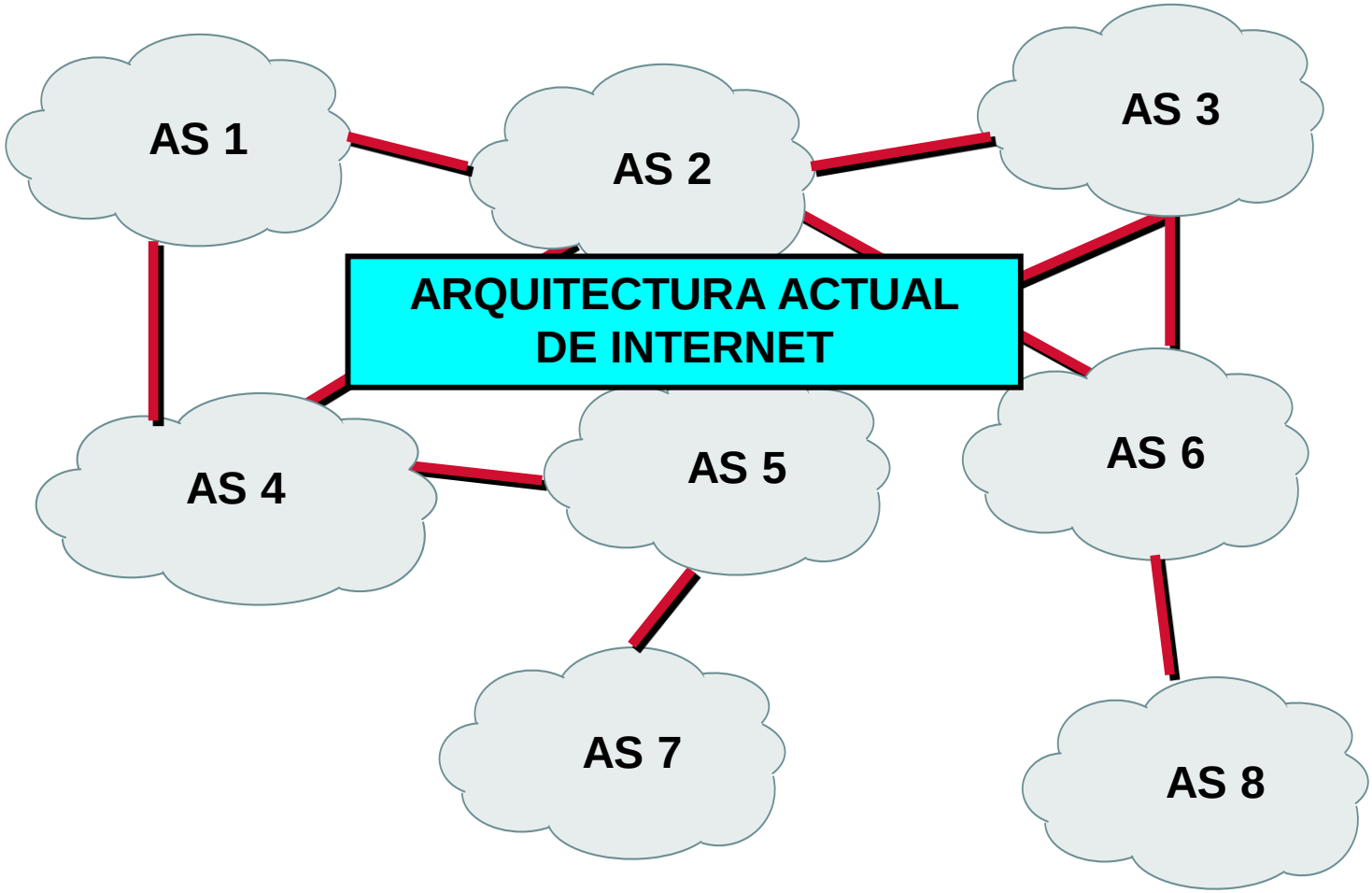
Sistemas Autónomos (AS) (2)

- Esta división en administraciones independientes permite trabajar con redes más pequeñas y manejables
- Tipos de AS:
  - Stub AS
  - AS Multihomed de no tránsito
  - AS Multihomed de tránsito

Sistemas Autónomos (AS) (3)



Sistemas Autónomos (AS) (4)



Internet (1)

- Interconexión de múltiples AS
- Es una red de redes
- No existe explícitamente un Backbone. Los mayores NSP (Network Service Provider) “hacen” las veces de Backbone

  Se les suele llamar proveedores “Tier 1”
- No hay una definición explícita de “Tier 1”, pero en general se asume que son aquellos proveedores que tienen presencia en todo el mundo y que “no le pagan a nadie” para llegar a todos los destinos

Internet (2)

- Pueden plantearse varias preguntas:
- ¿Cómo intercambiar información de ruteo entre los distintos AS?
- ¿Ruteo estático ó dinámico?
- ¿Influye el número de puntos de salida de un AS en la decisión?
- Para el caso dinámico: ¿Estado de enlace o Vector distancia?

Internet (3)

- ¿Por qué no usar un IGP: RIP, OSPF, otro?
  - Escalabilidad:
    - Sería como una única red (¿cómo expresar política de ruteo?)
    - Máx. número de hops acotado en algunos
    - Tamaño de las tablas de ruteo (y topología) inmanejable
  - Estabilidad:
    - Capacidad de adaptación a los cambios
    - Convergencia
    - Tráfico de Control inmanejable
  - Necesidad de políticas de enrutamiento

BGP

- BGP permite interconectar múltiples AS conformando la Internet que conocemos (Se dice que es la “goma” que mantiene unida internet)
- BGP es lo que se denomina un Interdomain routing protocol o InterAS routing protocol
- BGP es un EGP (Exterior Gateway Protocol)
- BGP-4 es hoy por hoy un estándar de facto en Internet

Ventajas de BGP (1)

- Escalabilidad: diseñado para manejar grandes tablas de rutas (full-routing implica del orden de 400.000 prefijos). No requiere excesivo tráfico de control
- Estabilidad: se adapta a los cambios fácil y rápidamente
- Sencillez: de la familia de los protocolos de vector distancia (no requiere estructura jerárquica ni conocimiento de la topología de red)

Ventajas de BGP (2)

- Soporta políticas de enrutamiento distintas y administración independiente por AS
- No impone restricciones al tamaño del bloque a anunciar
- Es un estándar
- Garantiza intercambio de información de ruteo libre de loops
- Es extensible

Agenda (2)

- Conceptos Fundamentales de BGP
- Análisis del protocolo (BGP-4)
- Atributos de BGP y políticas de control
- IBGP mesh. Soluciones
- Sumarización y anuncios (CIDR)
- Damping
- Algunos problemas
- Extensiones Multiprotocolo
- Salidas reales y datos de actualidad

BGP versión 4

- Definido originalmente en la RFC 1771 (1995)
  - Las versiones anteriores se manejaban con “clases”
- RFC 4271 (enero 2006, draft standard), pequeñas modificaciones a RFC 1771. Se adapta al uso real actual
- Se ha extendido mediante múltiples RFCs
- No se ve un sustituto a BGP en un futuro cercano

“Conexión BGP”

- Concepto de “peers” o “neighbors” (vecinos)

  No hay descubrimiento, deben configurarse explícitamente

- Utiliza TCP, puerto 179

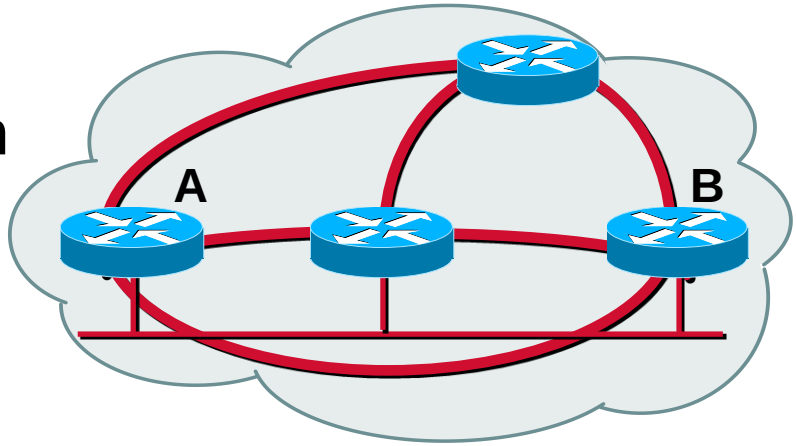
  La confiabilidad recae en la capa de transporte simplificando el protocolo

- La sesión TCP requiere de una capa de ruteo interno que me permita llegar al vecino (IGP, estático o directamente conectado)
- Se distingue entre BGP interno y externo

IBGP: BGP Interno (Internal BGP)

- Cuando los vecinos pertenecen al mismo AS
- Las conexiones representan sesiones BGP entre vecinos y no necesariamente links físicos

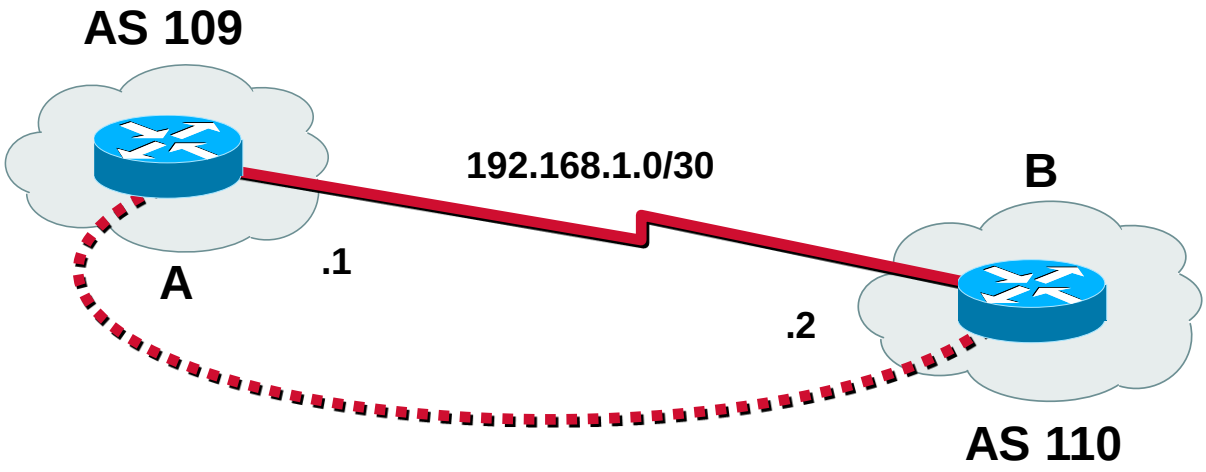
Full mesh en principio



AS 200

EBGP: BGP Externo (External BGP)

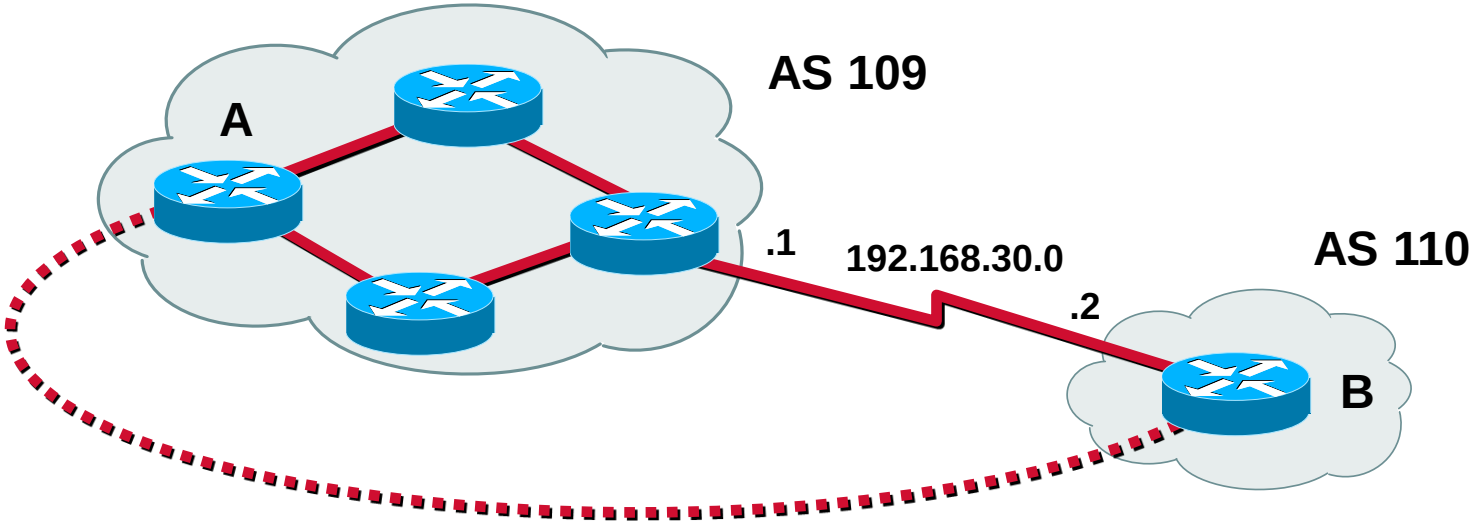
- Cuando los vecinos BGP pertenecen a distintos AS
- En general los vecinos se encuentran directamente conectados



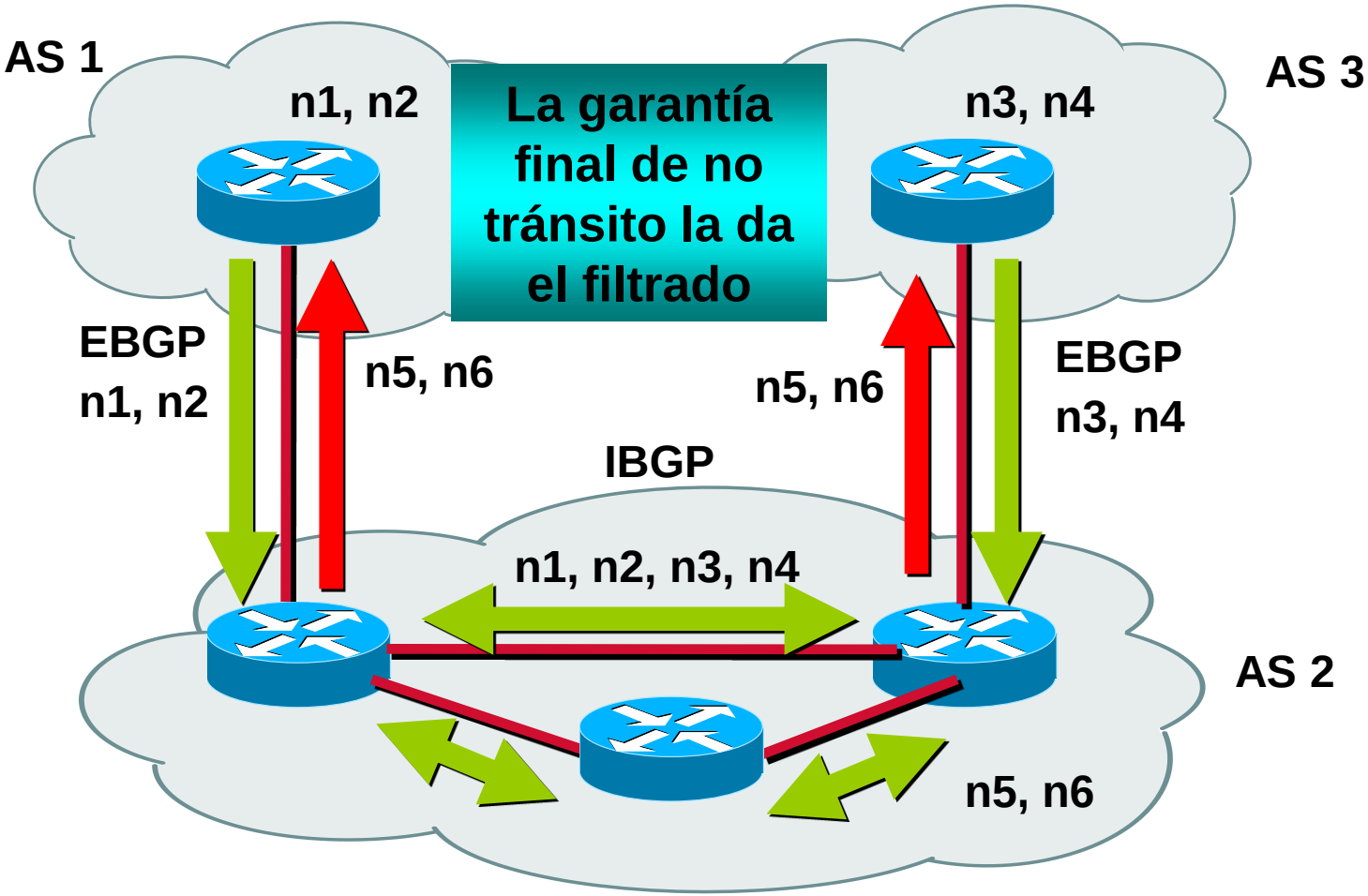
EBGP multihop

- Cuando los vecinos no se encuentran directamente conectados, se dice que la sesión entre ellos es EBGP-Multihop

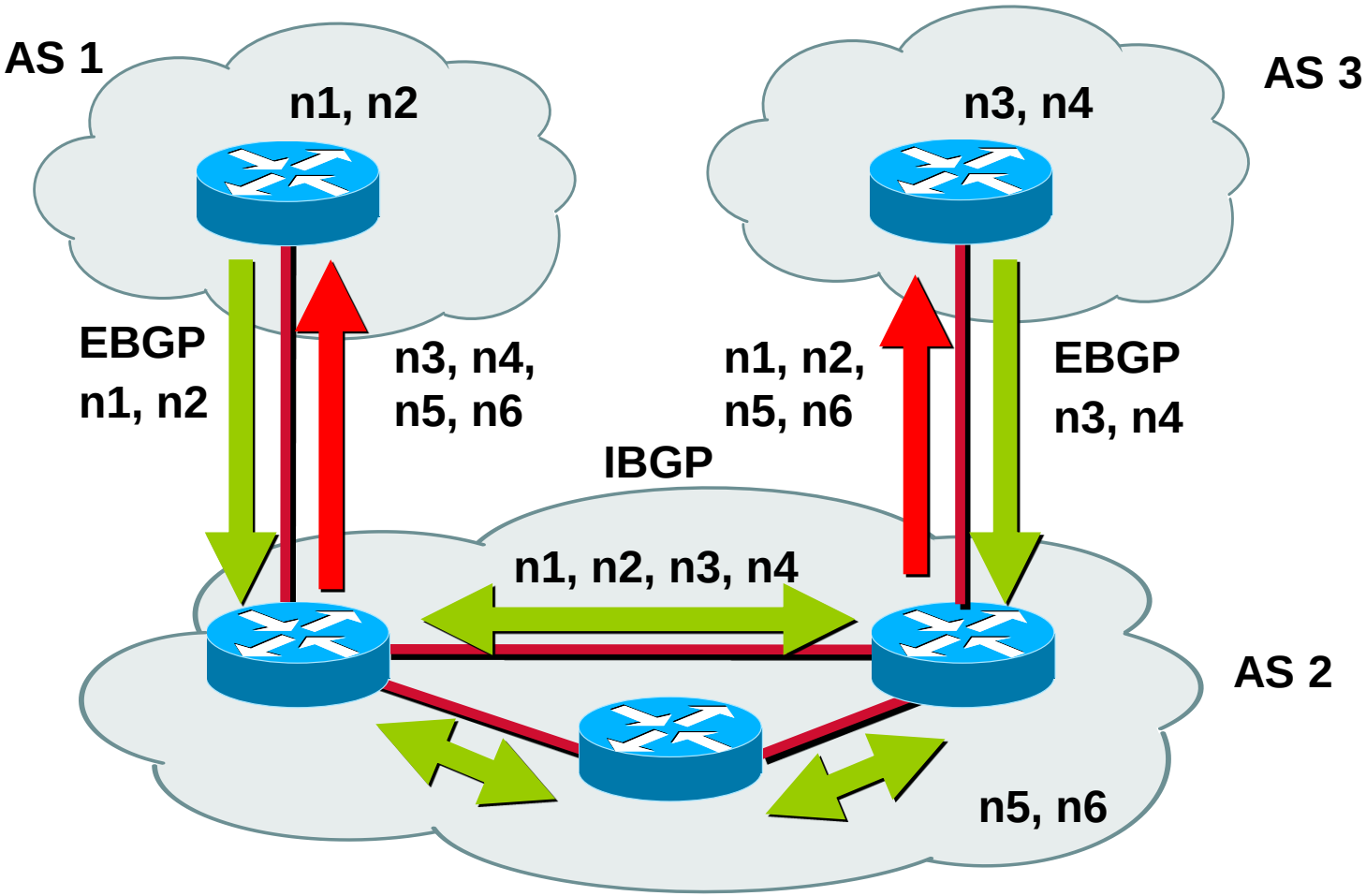
  Debe configurarse explícitamente



Ej. IBGP, EBGP no tránsito



Ej. IBGP, EBGP tránsito

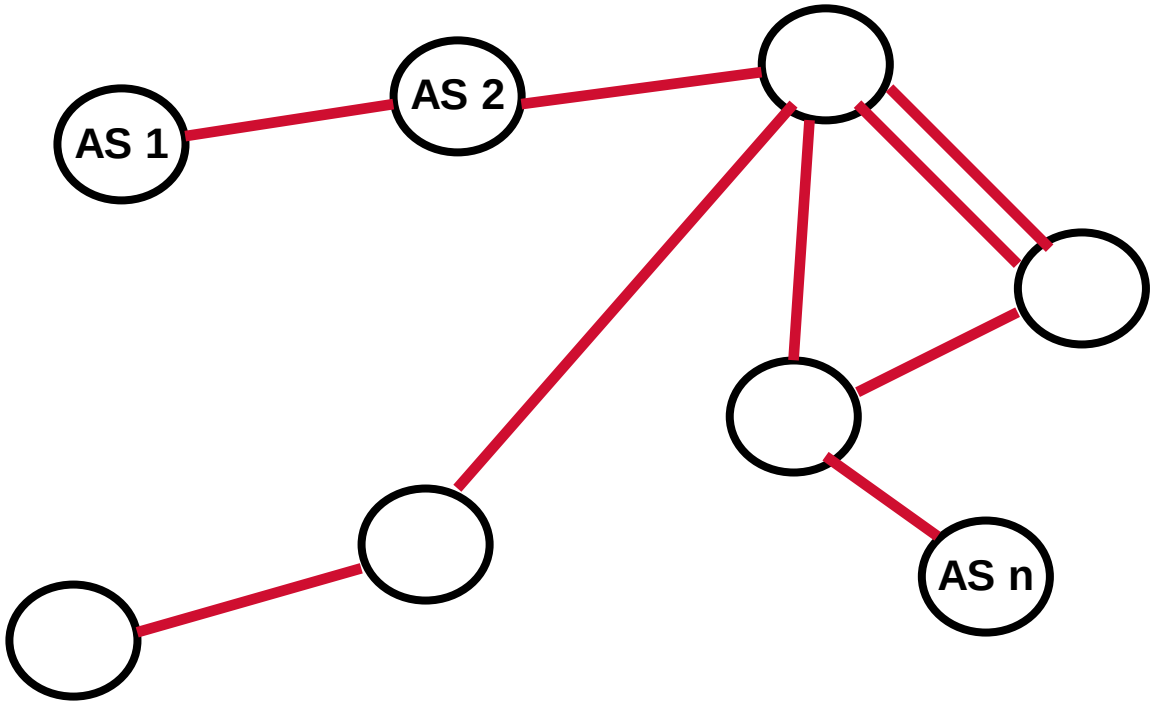


Cómo trabaja BGP (1)

- Se propaga cada prefijo, con un conjunto de atributos, incluyendo el camino de sistemas autónomos por el que pasó el anuncio
- Se dice que BGP es un “path vector protocol”

  Cada anuncio incluye una lista con la secuencia completa de AS que un paquete debe atravesar para llegar a la red de destino

Cómo trabaja BGP (2)



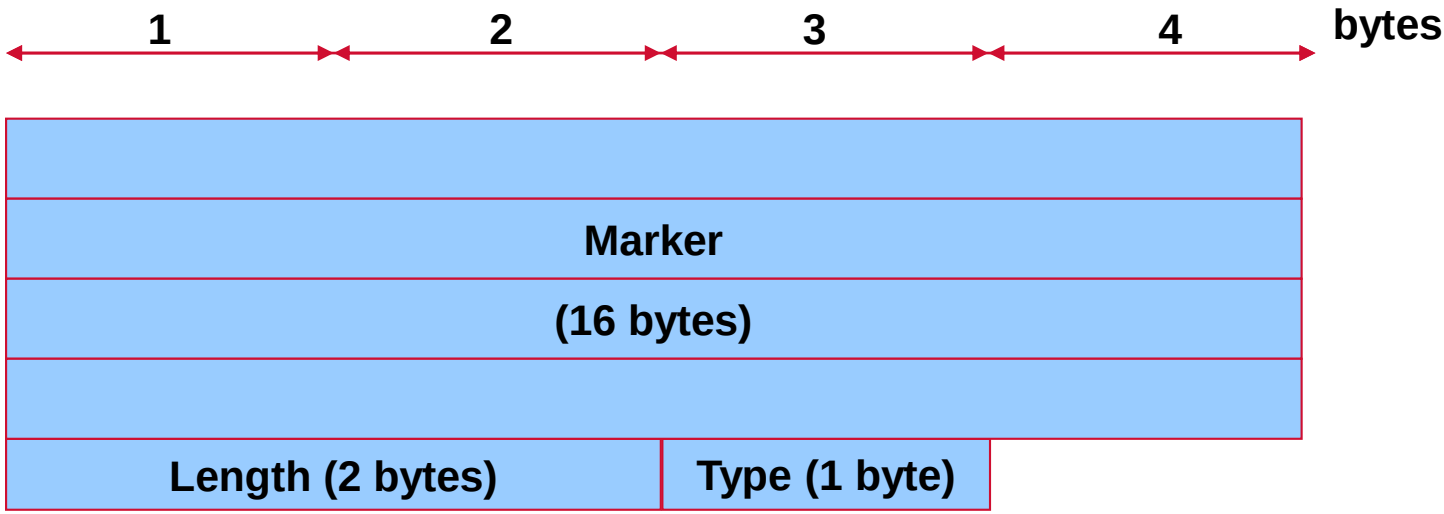
AS_Path Tree

Link BGP

Cómo trabaja BGP (3)

- ¿Cómo establece una sesión BGP sobre TCP entre dos peers?
- ¿Cómo se intercambia información de ruteo en BGP?
- ¿Qué tipo de mensajes intercambian dos peers de BGP?
- ¿Cómo se mantiene “viva” la sesión una vez establecida?

Encabezado (header) de BGP (1)



- **Total: 19 bytes**

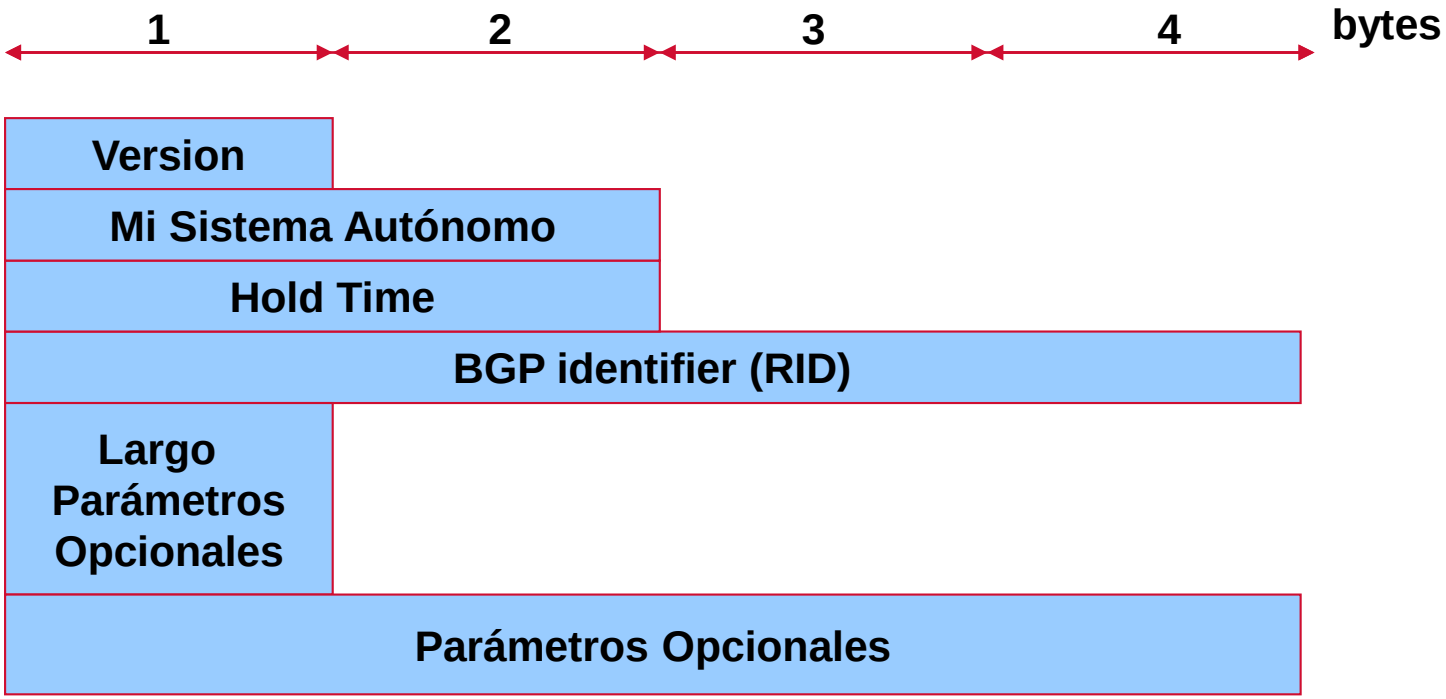
Encabezado de BGP (2)

- Marker: contiene una secuencia que puede ser predecida por el peer remoto
  - (en desuso) Autenticar los mensajes BGP recibidos en caso de usar autenticación
  - Secuencia de unos (binario). Detectar pérdida de sincronización en caso de no usar autenticación de mensajes
- Length: largo total del mensaje incluído el encabezado
- Type: Open, Update, Keepalive, Notification

Tipo de mensajes BGP

- OPEN: iniciar sesión BGP
- NOTIFICATION: condición de error
- UPDATE: alta o baja de rutas
- KEEPALIVE: confirmación periódica
- Tamaño de los mensajes:
  - Mínimo 19 bytes (sólo header)
  - Máximo 4096 bytes

Mensaje OPEN (1)



- 10 bytes mínimo

Mensaje OPEN (2)

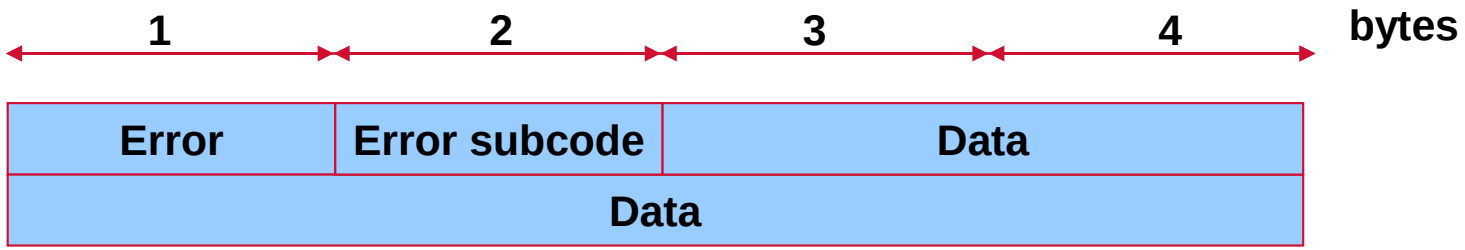
- Version: 4
- Mi AS: Sistema autónomo del emisor
  - Así se distingue IBGP de EBGP
  - Se verifica configuración
- BGP ID: ID del router que envía el mensaje
- Hold time:
  - Tiempo máximo en segundos que puede transcurrir sin recibir mensajes de UPDATE o KEEPALIVE
  - Este tiempo se negocia al iniciar la sesión (mínimo entre ambos extremos, no menos de 3 segundos)

Mensaje OPEN (3)

- Parámetros Opcionales: como su nombre lo indica, parámetros opcionales que se negocian al iniciar la relación de vecinos (Por ej. “capabilities”)
- Largo de parámetros opcionales:

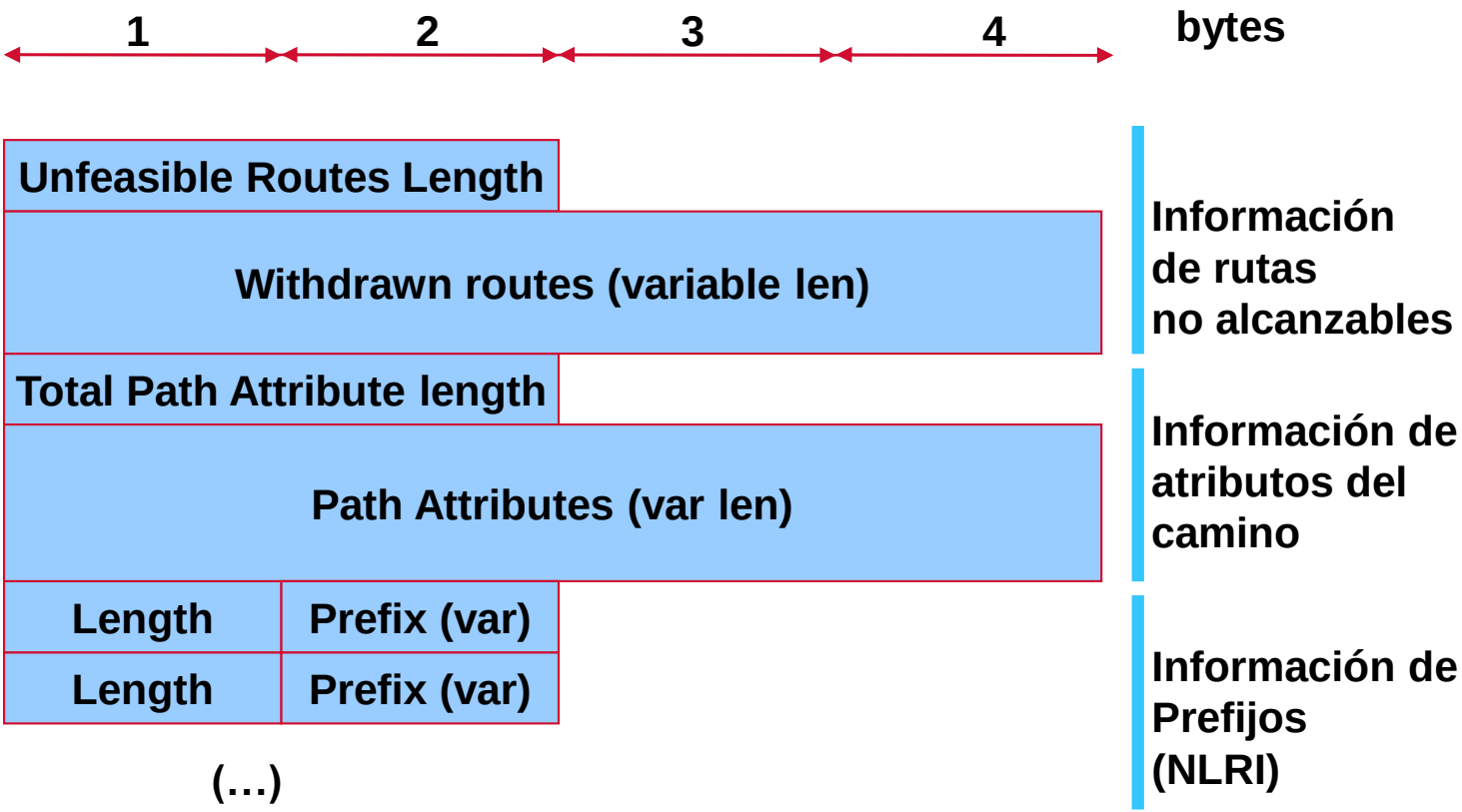
  “0” indica que no se negociarán parámetros opcionales

Mensaje NOTIFICATION



Error code	Error subcode
1- message Header Error	1: Connection Not sync 2: Bad message length 3: Bad message type
2-Open message error	1: Unsupported version numb 2: Bad Peer AS 3: Bad BGP identifier 4: Unsupported Optional Par. 5: Authent error 6: Unacceptable hold time
3-UPDATE message error	1: Malformed Attribute-list 2: Unrecognised well-know attr. 3: Missing well-know attribute (…)
4-Hold timer expired	NA
5-Finite state machine error	NA
6-Cease	NA

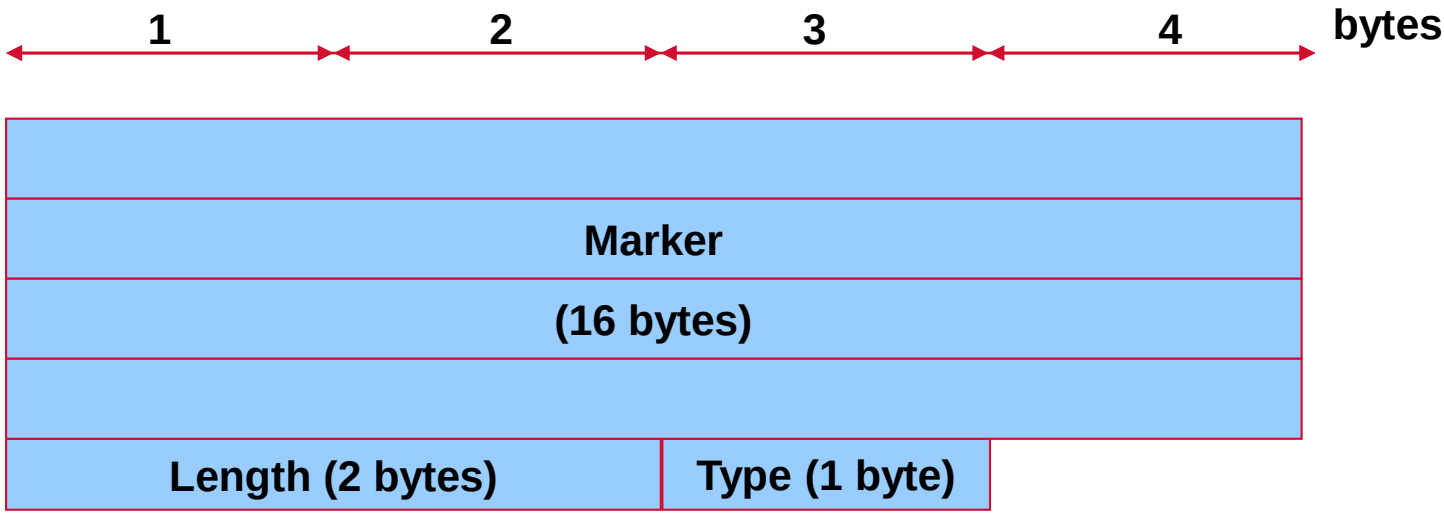
Mensaje UPDATE (1)



Mensaje UPDATE (2)

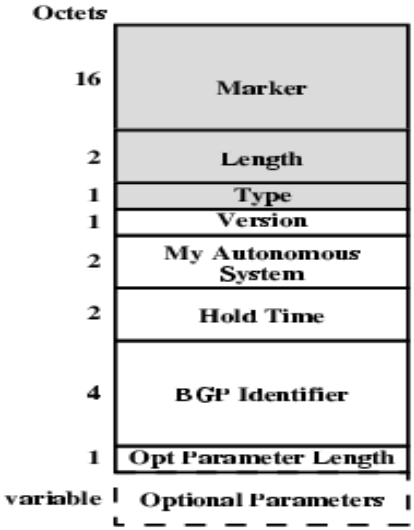
- Withdrawn routes: Prefijos anunciados previamente que ya no son alcanzables
- Path Attributes: Atributos de un determinado “camino”
- Información de NLRI (Network Layer Reachability Information): prefijos que comparten un camino y los atributos

Keepalive message

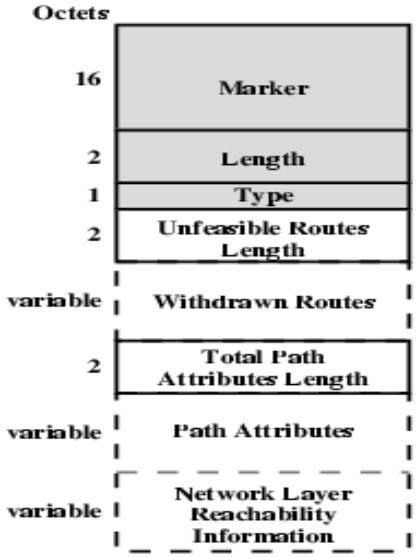


- Consiste simplemente en el Header
- 19 bytes intercambiados periódicamente

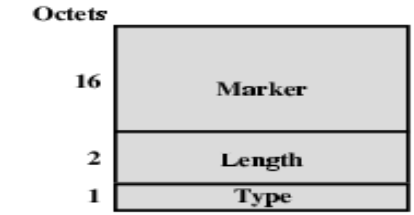
Resumen de mensajes



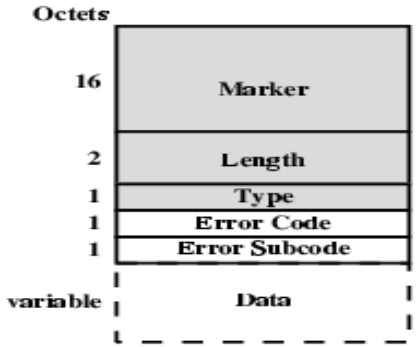
(a) Open Message



(b) Update Message

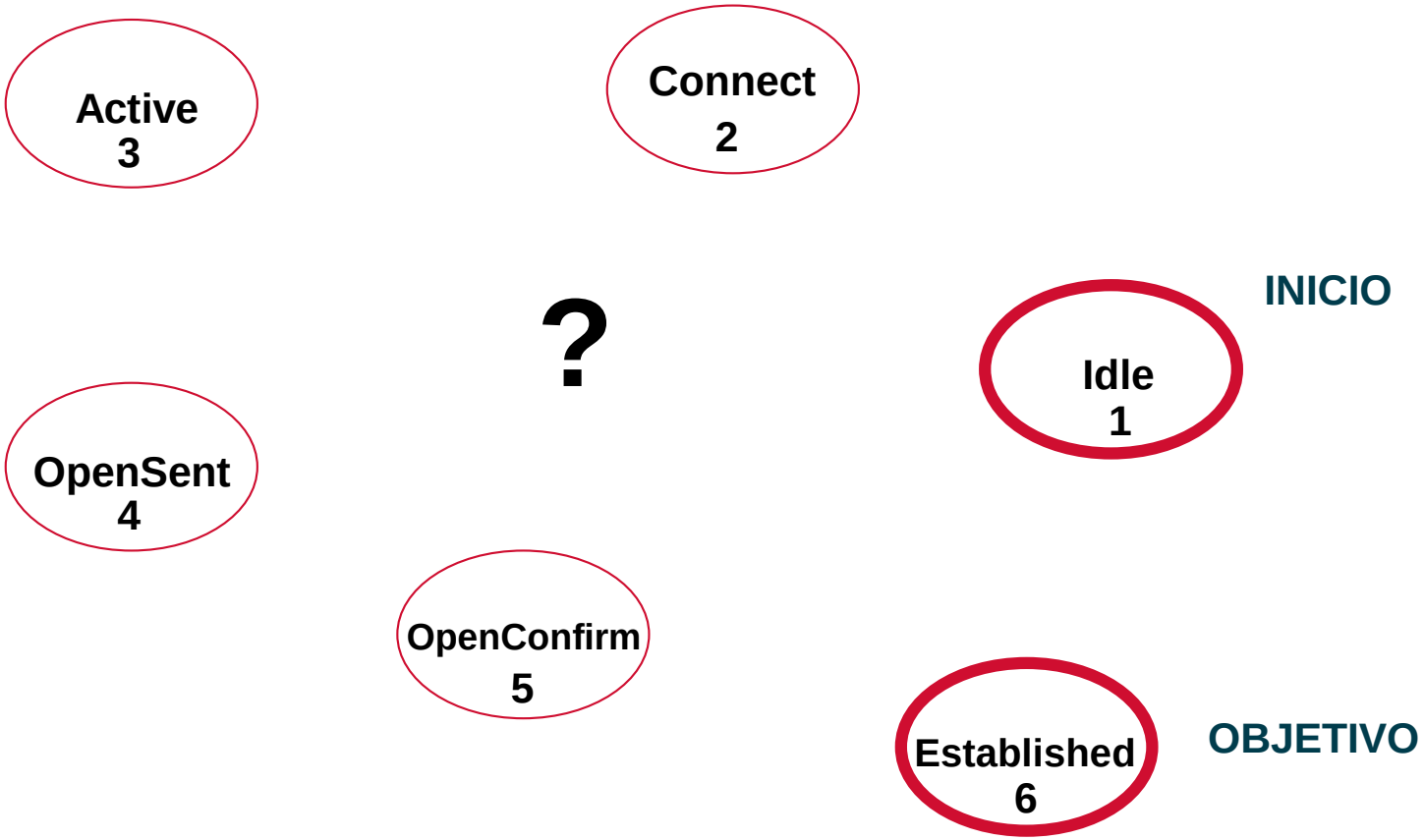


(c) Keepalive Message

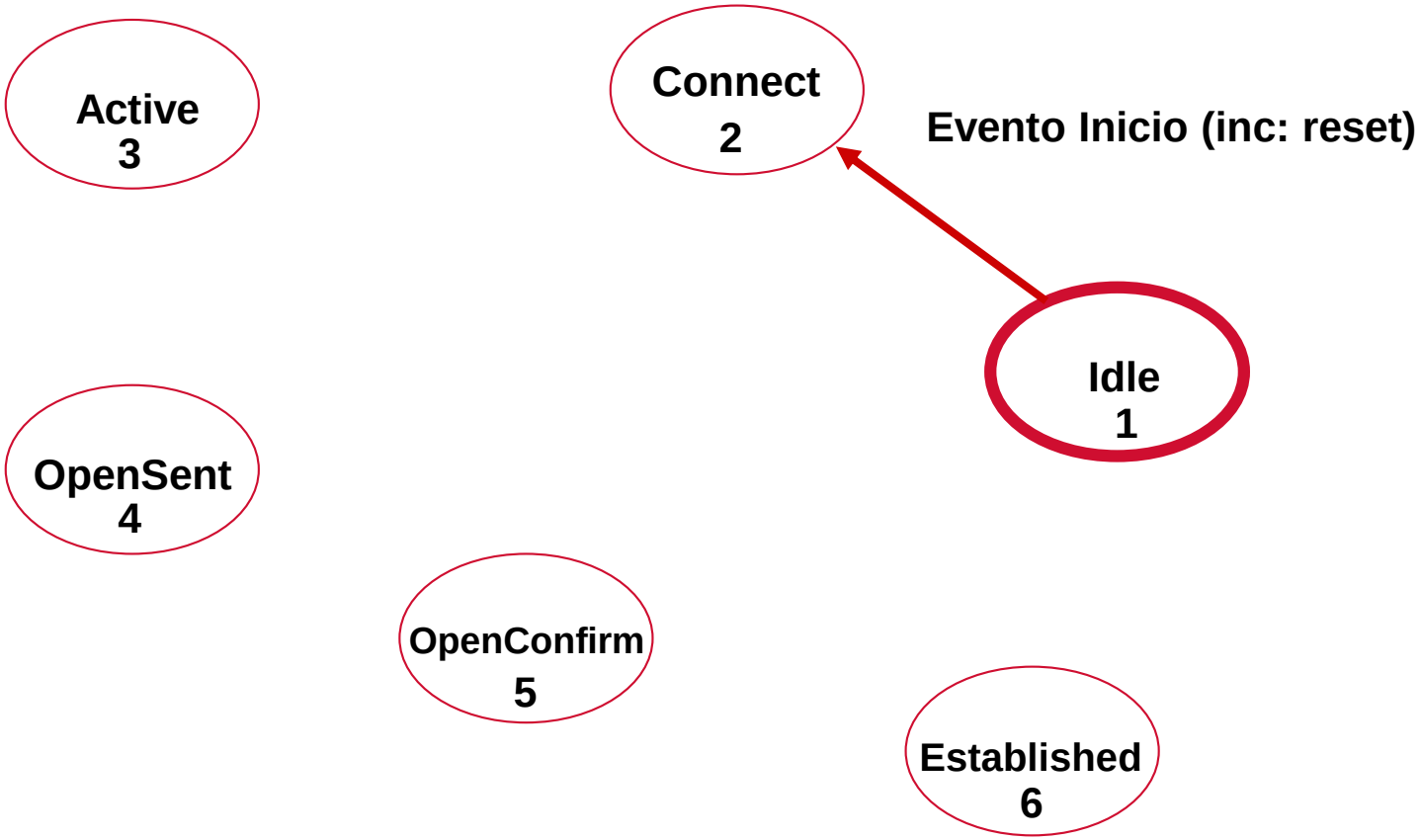


(d) Notification Message

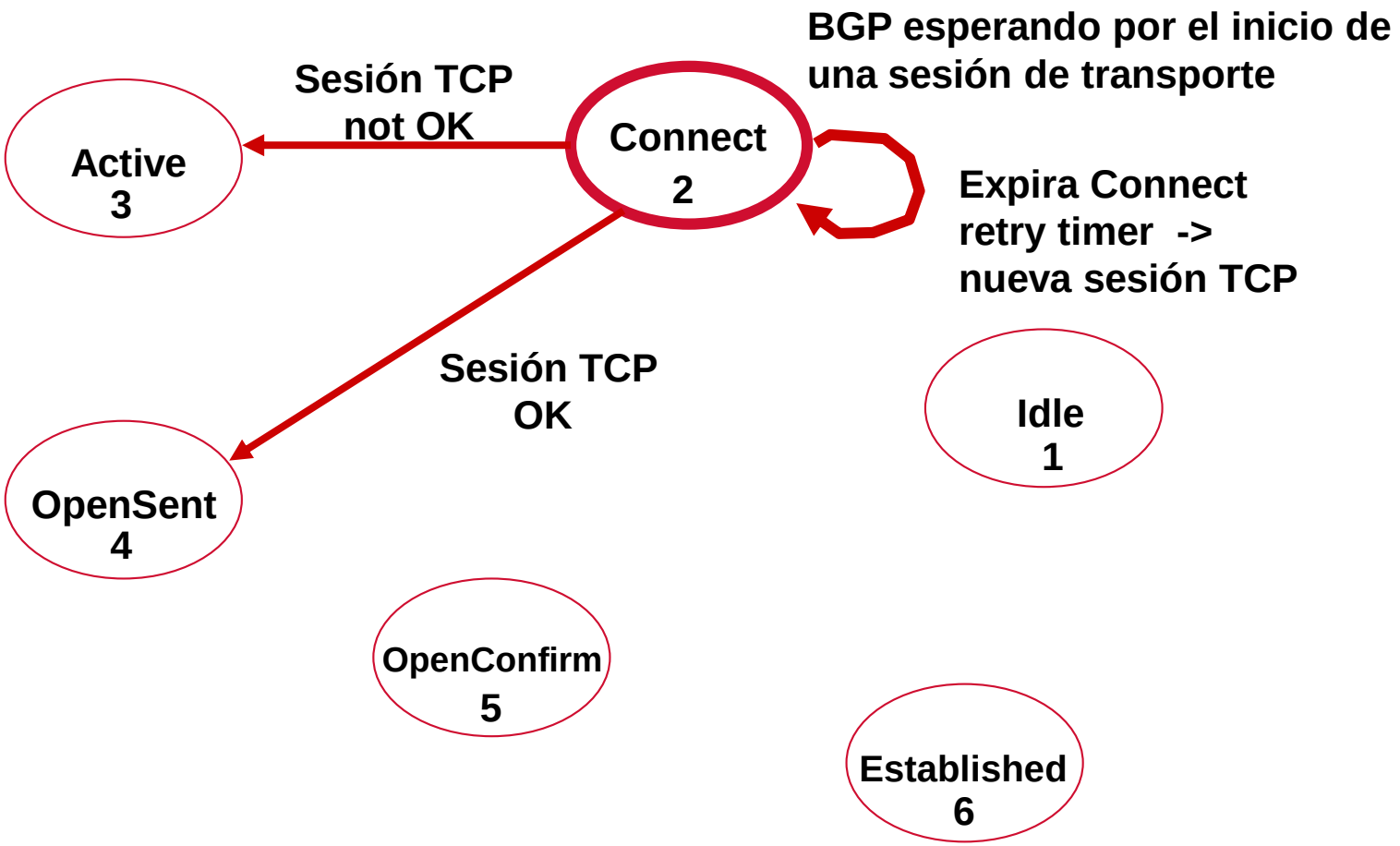
Inicio de una sesión BGP
Diagrama de Estados (RFC 1771)



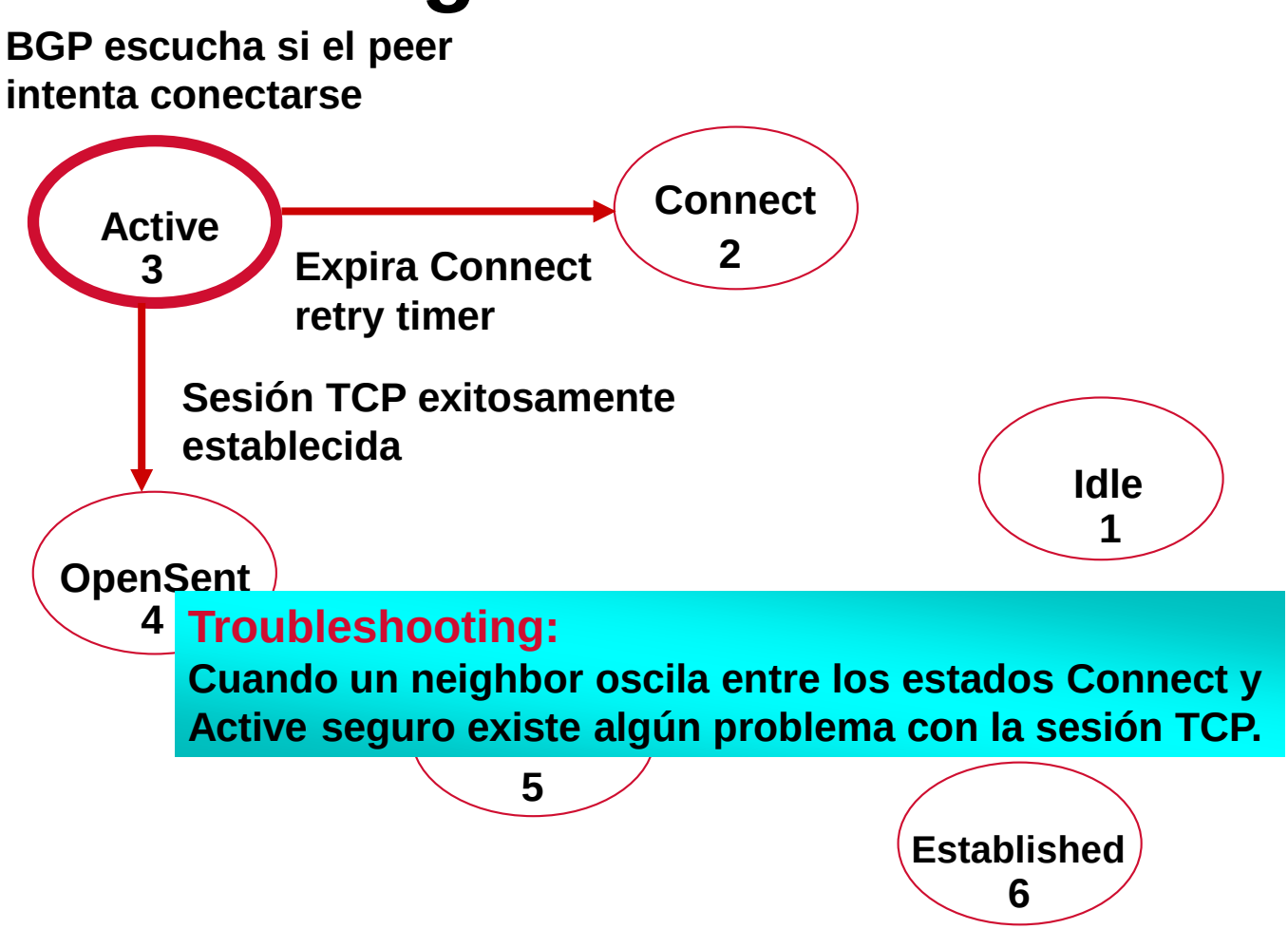
Inicio de una sesión BGP
Diagrama de Estados



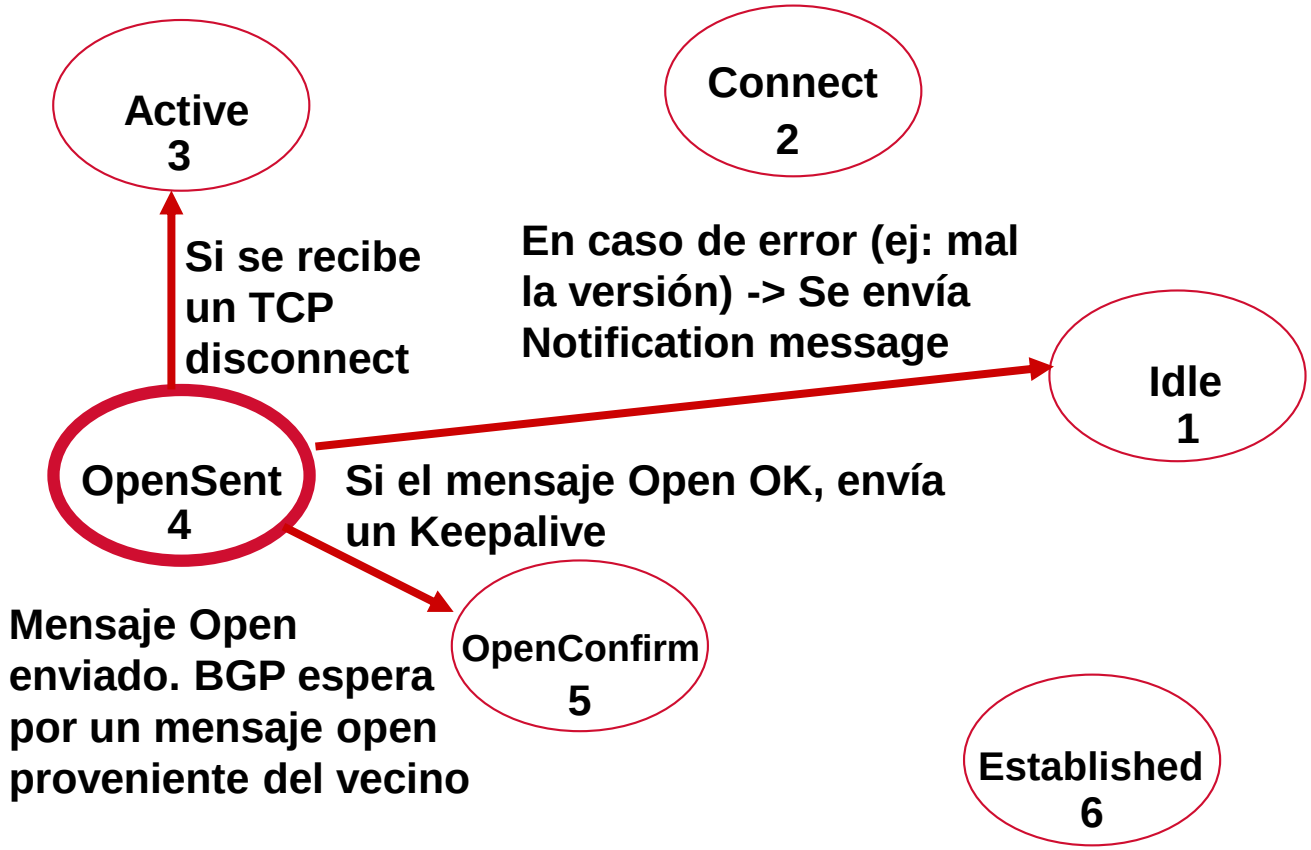
Inicio de una sesión BGP
Diagrama de Estados



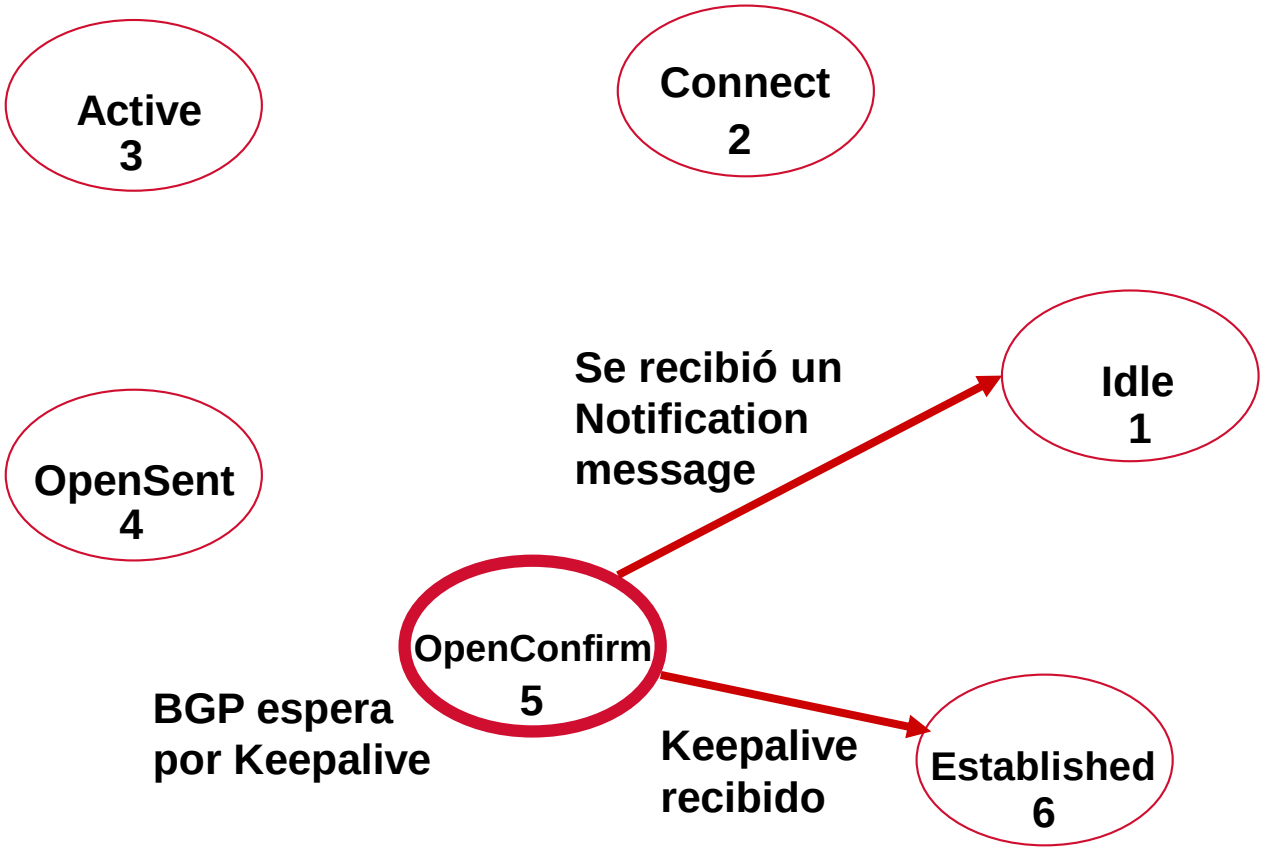
Inicio de una sesión BGP
Diagrama de Estados



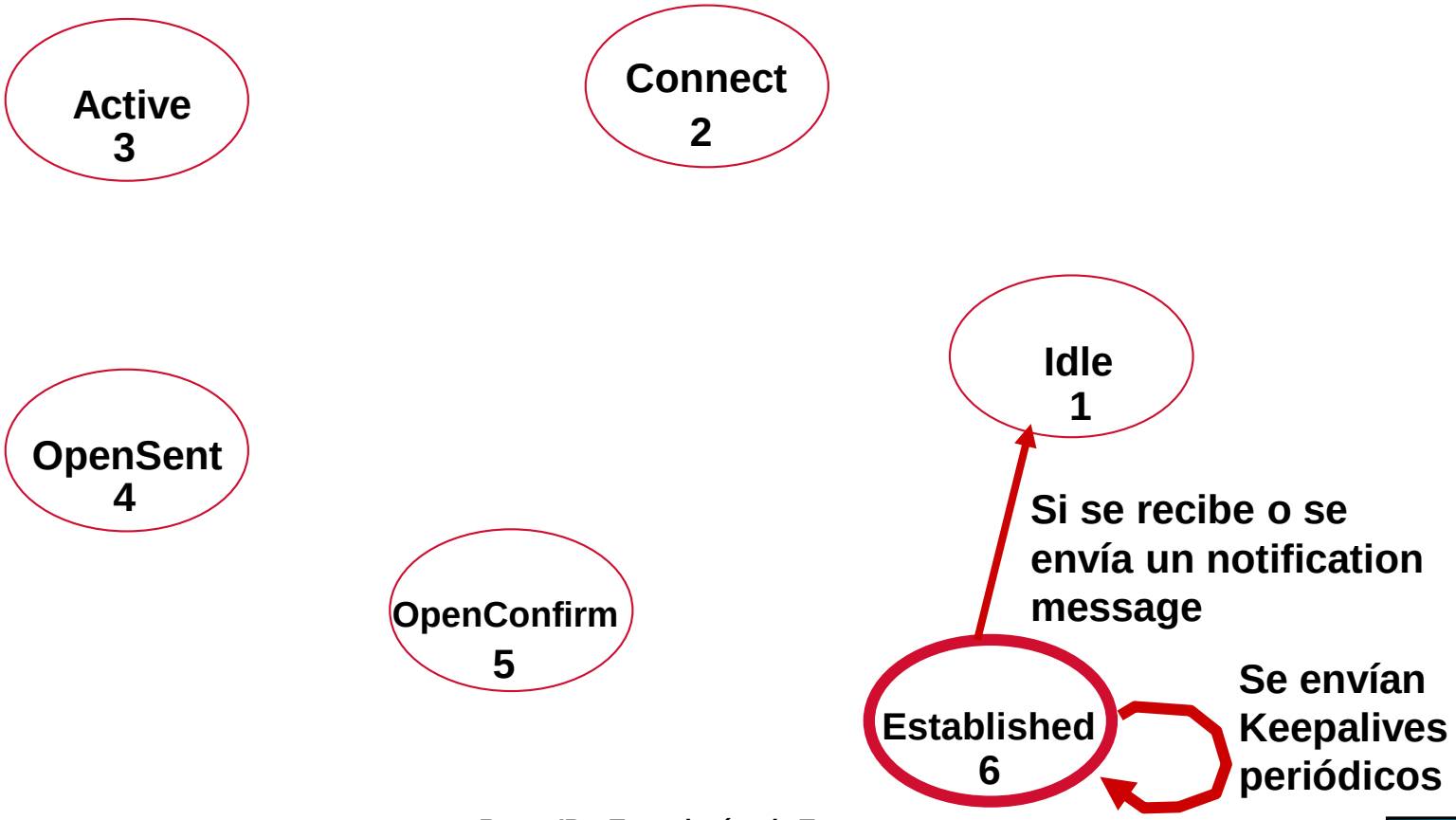
Inicio de una sesión BGP
Diagrama de Estados



Inicio de una sesión BGP
Diagrama de Estados

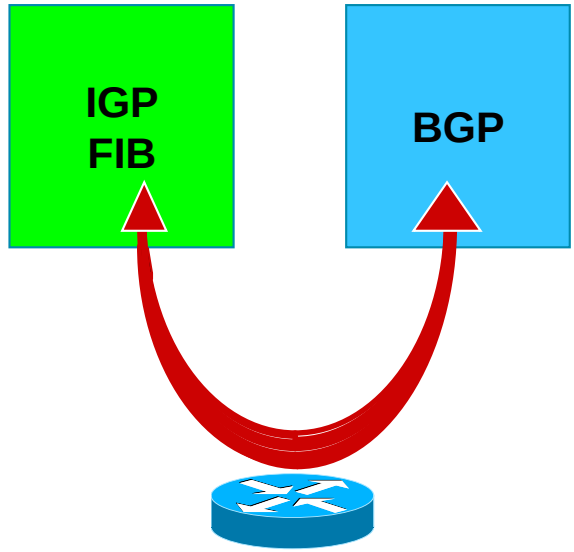


Inicio de una sesión BGP
Diagrama de Estados



Tablas de rutas en BGP

- BGP tiene sus propias tablas, independientes de la tabla de rutas
- Las tablas se intercambian entre peers al inicio de la sesión. Luego sólo actualizaciones incrementales



RIB (Routing Information Base)

- Tablas de rutas con sus atributos
- Conceptualmente, tres conjuntos de tablas:

  Adj-RIB-In: Rutas recibidas de un vecino. Tantas tablas como vecinos tenga

  Loc-RIB: Información local (lo que utilizo, luego de aplicarle políticas a las RIB-In)

  Adj-RIB-Out: Rutas para ser enviadas a los vecinos (una por vecino)

- La política se realiza a la entrada en Adj-RIB-In, y a la salida entre Loc-RIB y Adj-RIB-Out

Operación General (1)

- Un enrutador aprende múltiples caminos (paths) via BGP interno o externo
- Escoge “EL MEJOR” camino y lo instala en su tabla de forwarding
- El protocolo es susceptible a **Políticas** que se aplican para influenciar justamente la selección del “MEJOR” camino

Operación General (2)

- Se anuncia SOLO el MEJOR camino a cada destino
- Lo que un enrutador aprende por EBGP lo anuncia a TODOS sus peers
- Lo que un enrutador aprende por IBGP lo anuncia sólo a sus peers EBGP
  - Finalidad: Evitar Loops
  - Me obliga a tener una malla completa en iBGP
  - En redes grandes, hay soluciones para escalabilidad:
    - Reflectores de rutas
    - Confederaciones

Sincronización (1)

Regla: En los AS Multihomed de tránsito NO usar ni anunciar un prefijo hasta que una ruta que lo contenga haya sido aprendida por IGP

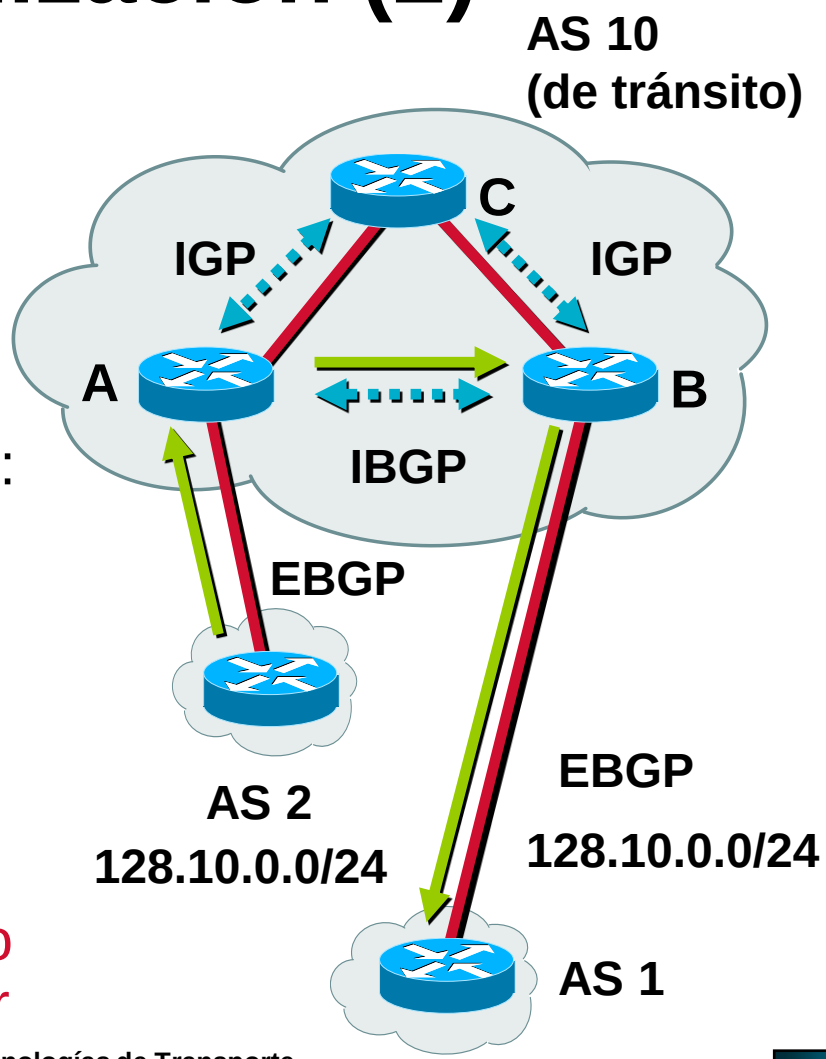
- Asegura la consistencia de la información en el interior del AS
- Evita “black holes” dentro del AS
- **Se trata de buscar topologías que permitan deshabilitar la sincronización**

Sincronización (2)

- A y B peers IBGP
- C no lo es

➢ Si la sincronización está apagada y el IGP no propagó la ruta a 128.10.0.0:

- B intenta alcanzar la red 128.10.0.0 via C
- C descarta los paquetes ya que no conoce una ruta a la 128.10.0.0
- El AS 1 recibe un anuncio al cual jamás podrá llegar

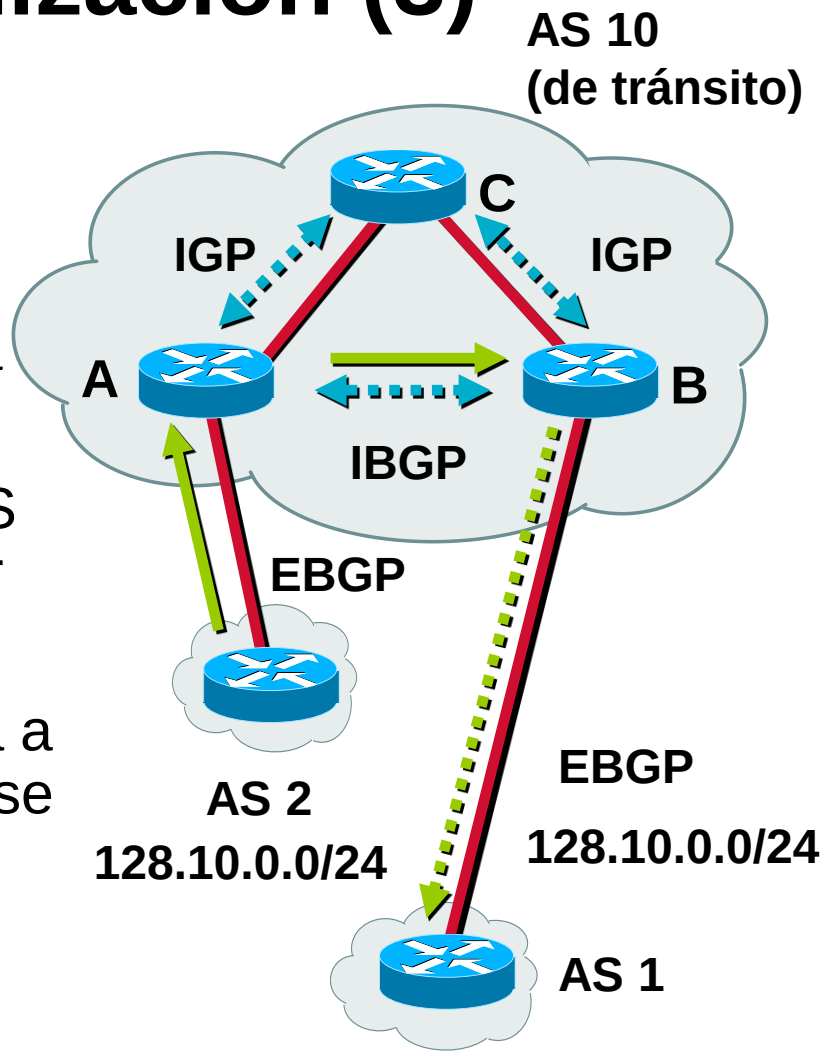


Sincronización (3)

- A y B peers IBGP
- C no lo es

➢ Si la sincronización está prendida:

- B no anuncia la red al AS 1 hasta no conocerla por IGP
- C debe conocer una ruta a la 128.10.0.0 via IGP -> se debe redistribuir la red aprendida por BGP en IGP en el enrutador A



Sincronización (4)

- Alternativas para evitar redistribuir en el enrutador A:
  - ➢ Deshabilitar la sincronización y correr BGP en todos los enrutadores del AS (al menos todos en el camino entre otros AS)
  - ➢MPLS (se verá luego)
- Política en caso de necesitar hacer la redistribución:
  - ➢ Hacerlo sólo para las redes de interés!!
- Es posible deshabilitarla si nuestro AS no oficiará de tránsito hacia otros AS

Agenda (3)

- Conceptos Fundamentales de BGP
- Análisis del protocolo (BGP-4)
- Atributos de BGP y políticas de control
- IBGP mesh
- Sumarización y anuncios (CIDR)
- Damping
- Algunos problemas
- Extensiones Multiprotocolo
- Salidas reales y datos de actualidad

Atributos de BGP (1)

- Los atributos de bgp son los que permiten tomar decisiones “complejas” sobre los caminos
- 4 Categorías:
  - **Well-Known Mandatory** (Obligatorios, bien conocidos). Deben ser reconocidos por todas las implementaciones de BGP y deben estar presentes en todo mensaje de UPDATE
  - **Well-Known Discretionary** (bien conocidos, opcionales). Deben ser reconocidos por todas las implementaciones de BGP, pero pueden o no aparecer en un mensaje de UPDATE

Atributos de BGP (2)

- **Optional Transitive (opcional, transitivo)**: no se requiere que sean soportados por todas las implementaciones de BGP. Deben ser reenviados aún en el caso de no ser soportados
- **Optional Nontransitive (opcional, no transitivo)**: no se requiere que sean soportados por todas las implementaciones de BGP. En caso de no ser reconocido, se ignora y no se pasa a otros vecinos BGP

Atributos de BGP (3)

Attribute flags	Tipo	Largo	Valor

Flags:

- Bit 0: Opcional/bien-conocido
- Bit 1: Transitivo
- Bit 2: Parcial
- Bit 3: Largo Extendido
- Bit 4-7: deben ser 0

Algunos Atributos de BGP

WKM
- Next_hop
- AS_path
- Origin

WKD
- Local preference
- Atomic aggregate

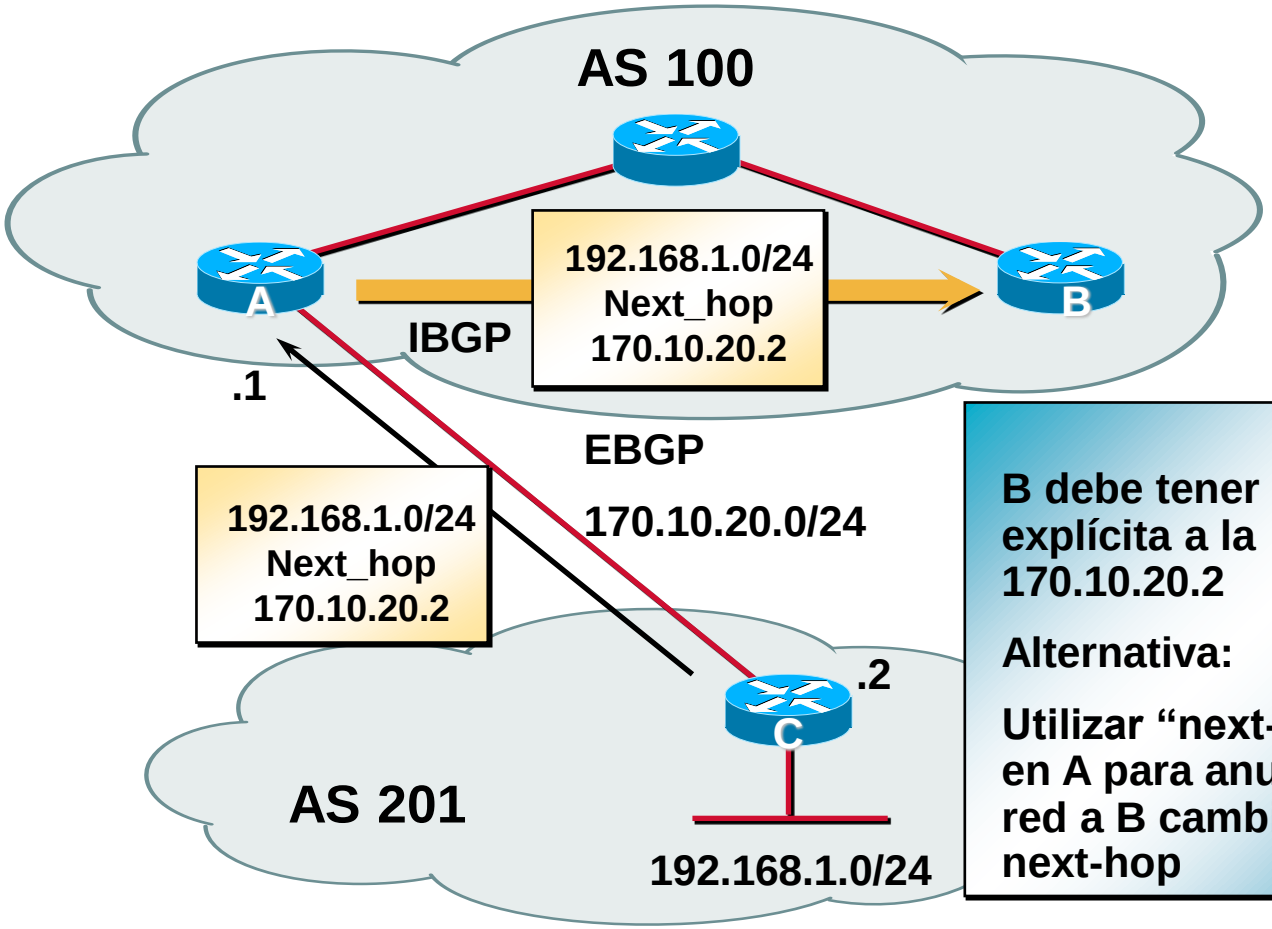
OT
- Aggregator
- Community

ONT
- Multi Exit Discriminator (MED)

NEXT_HOP (1)

- **NEXT_HOP indica la IP del vecino al cual enviarle los paquetes para alcanzar una red**
- **Varía según IBGP o EBGP:**
  - Para EBGP: dirección IP del peer que anunció la ruta (excepción posible en medios Multiacceso, p. ej. Ethernet)
  - Para iBGP: Redes que fueron inyectadas al AS via EBGP tienen como NEXT_HOP el anunciado por EBGP y se acarrea inalterado en IBGP
- **Si no se tiene una ruta específica a la IP del Next_hop no se debería usar la ruta. No alcanza con una ruta por defecto!!!**

NEXT_HOP (2)



B debe tener una ruta explícita a la 170.10.20.2

Alternativa:

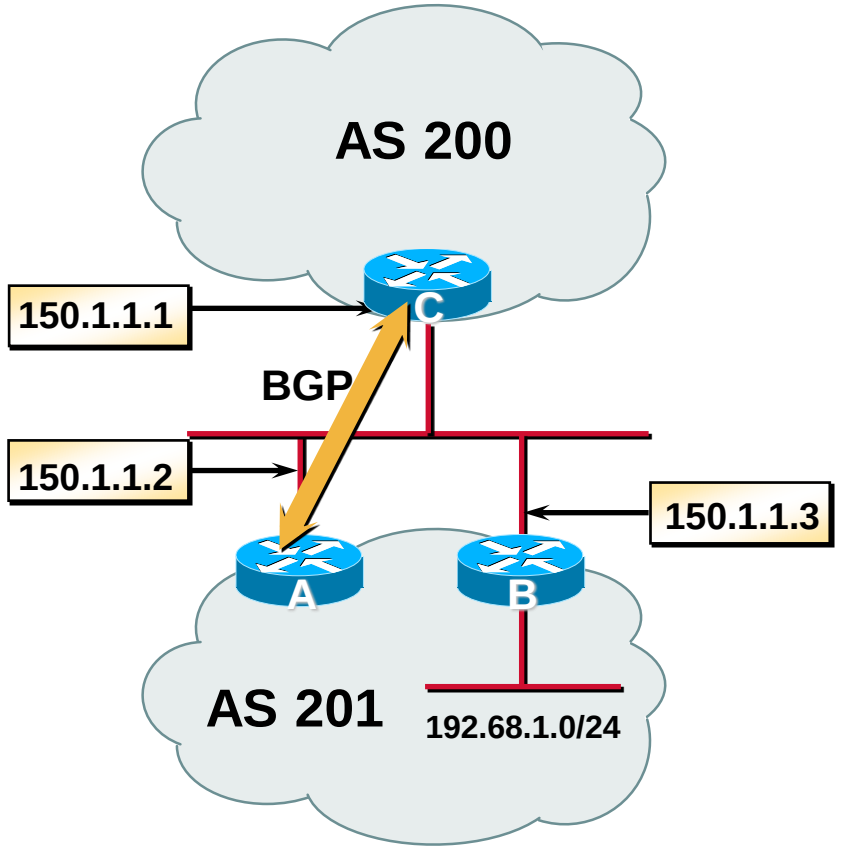
Utilizar "next-hop self" en A para anunciarle la red a B cambiando el next-hop

Third-Party NEXT_HOP en un medio multiacceso

- Ejemplo:
  - ➢A y B están en el mismo AS
  - ➢A le anuncia a C

    la red 192.68.1.0/24

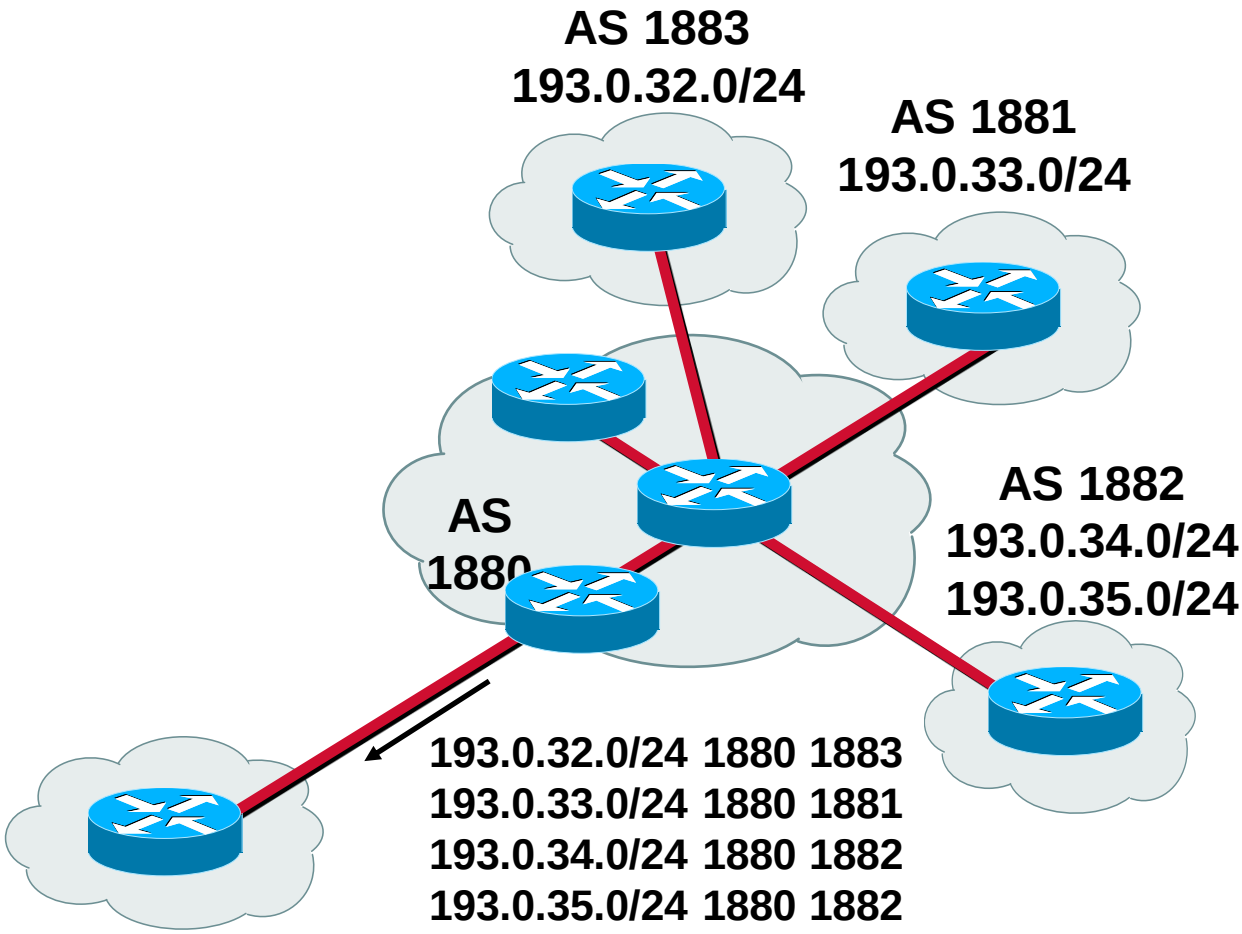
    con NEXT_HOP 150.1.1.3.
- Es más eficiente!



Atributo AS_Path (1)

- Lista de Sistemas Autónomos (AS) que un anuncio ha atravesado
- evita loops!!!
- 2 posibles componentes: AS_SEQUENCE y AS_SET
- AS-SET: {1881 1882 1883}
- Se usa como uno de los criterios para la elección del mejor camino (se prefiere un AS_Path más corto)

Uso del AS-Set (sumarización)

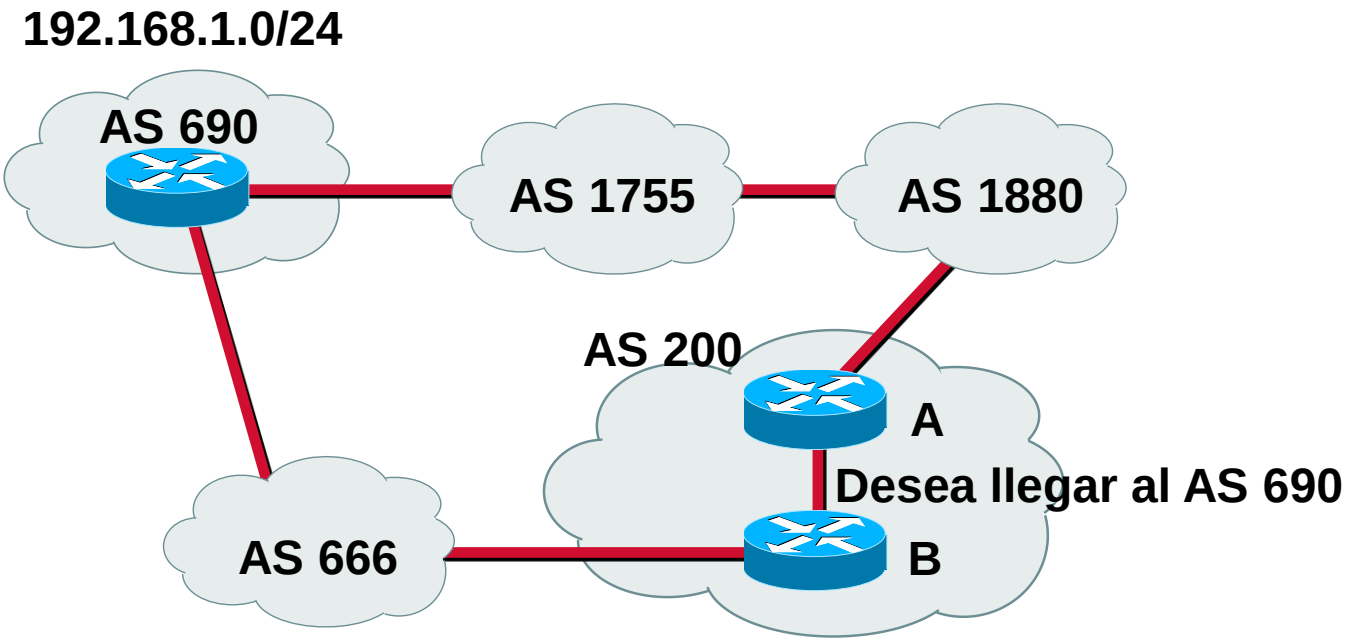


Uso del AS-Set (sumarización)

- El AS-SET se utiliza para indicar los sistemas autónomos que participaron en la formación del agregado
- La realidad es que esta forma de sumarización se utiliza muy poco
- RFC 6472 recomienda NO utilizar AS_SET

  (actualmente (2014) del orden de 700 entradas con AS-SET en la tabla global, 82 orígenes)

AS_Path (2)



- **El AS 200 conocerá la red 192.168.1.0/24:**

  192.168.1.0/24 1880 1755 690

  192.168.1.0/24 666 690 <= Preferible!!!

ORIGIN (origen de la ruta)

- Provee información acerca de cómo se generó la ruta:
- IGP

  La ruta se generó en el proceso BGP (configuración)
- EGP

  Viene de EGP (obsoleto)
- Incomplete

  La ruta se aprendió de otra manera. Por ejemplo surge de redistribuir rutas IGP en BGP

  Ej. más común: redistribución de estáticas

Local Preference

- Cuando existen múltiples caminos para el mismo destino, el atributo de Local Preference indica el camino preferido **administrativamente**. Define el punto de salida de mi red
- El camino con la mayor preferencia local es el elegido
- El atributo Local Preference sólo tiene sentido “local”, se propaga en el interior del AS por IBGP, no por EBGP

Multi Exit Discriminator (MED)

- Para influenciar el camino de vuelta
- Se puede usar para discriminar entre múltiples caminos al mismo AS
- No se propaga a otros vecinos
- Limitado en principio a múltiples enlaces con un mismo AS
- Muchas veces llamado “métrica”

Proceso de selección del mejor camino en BGP (1)

1. No considerar un prefijo IBGP hasta no estar sincronizado (si sincronización habilitada)
2. No considerar un prefijo si no existe una ruta al next_hop (o si al agregar el prefijo se genera un loop de resolución)
3. Preferir la ruta con mayor Local Preference (global dentro del AS)

Proceso de selección del mejor camino en BGP (2)

4. Si la ruta no fue localmente originada, elegir el AS_path de menor largo
5. Si los paths son de igual largo, elegir el prefijo con el menor ORIGIN type: IGP sobre EGP, y EGP sobre Incomplete

Proceso de selección del best path en BGP (3)

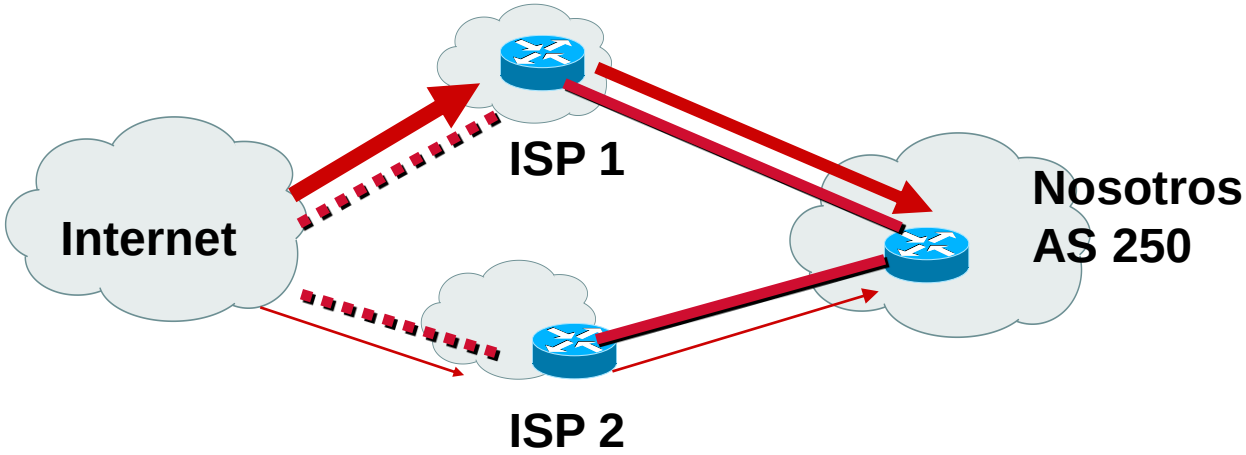
6. Si los orígenes son los mismos, y el AS vecino es el mismo, elegir el que tenga el menor atributo MED (ojo MED es Opcional!!!)
7. Preferir un anuncio externo antes que uno interno
8. Preferir el path a través del neighbor más próximo según el IGP
9. Preferir el path con el menor BGP router id

Influenciando el tráfico saliente

- Observar que la preferencia local es el primer atributo que se verifica
- Si recibo el mismo prefijo de más de un vecino, puedo elegir el camino de salida fijando un valor mayor de LOCAL-PREFERENCE

Influenciando el tráfico entrante utilizando atributos de BGP

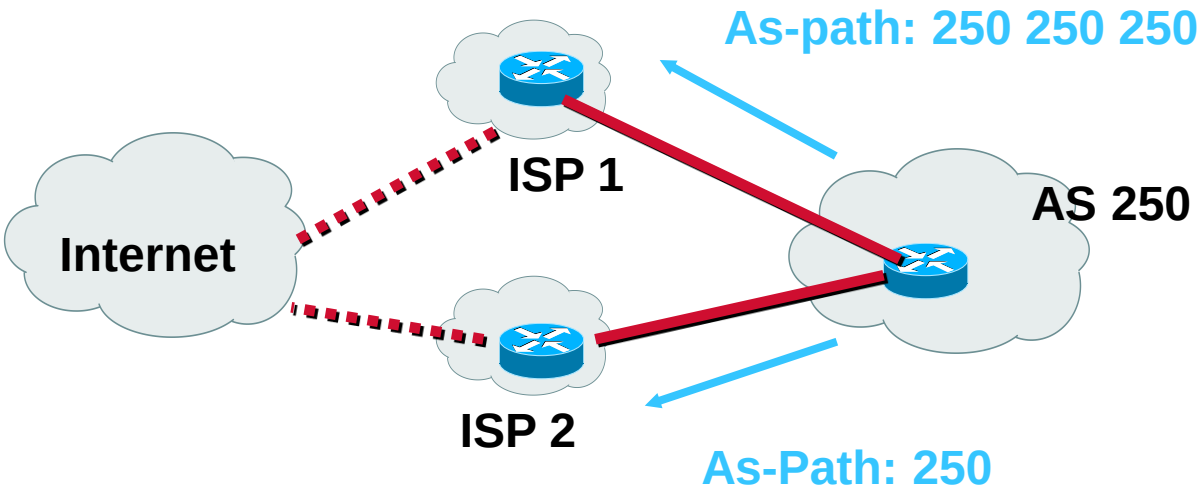
- MED. Muy limitado
- Práctica habitual: hacer prepends al AS-PATH (agregar copias de mi número de AS)
- Problema ejemplo: 80% del tráfico viene por el ISP 1:



“Solución”: prepends

route-map bajarprio permit 10

set as-path prepend 250 250



Los sistemas autónomos remotos verán un camino más largo por ISP1

Otras soluciones

- No publicar algunos prefijos por determinados enlaces
  - Problema: pierdo respaldo
- Publicar anuncios más específicos por el enlace descargado
- En general todas las soluciones son “ad-hoc”, estimando una corrección y luego verificando el efecto

Comunidades

- Atributo “Community”
- Opcional, transitivo
- Una comunidad es un grupo de destinos que comparten una propiedad común
- Usado para agrupar destinos y aplicar una política común
- Un prefijo puede pertenecer a varias comunidades
- En muchos equipos no se propaga por defecto

  neighbor *ip-address* send-community (en Cisco)

Comunidades (2)

- 32 bits
- Recomendación: número de AS en los primeros 16 bits
- set community *AS:community*
- Bien conocidas:

  internet

  no-export

  no-advertise

  local-AS

Políticas de Control (1)

- Filtrado de rutas

  Entrantes o salientes

  Al filtrar los anuncios entrantes, estoy definiendo el camino del tráfico saliente

  Al filtrar los anuncios salientes, estoy definiendo por donde vendrá el tráfico hacia mi AS, y si permito o no tránsito

- Manipulación de atributos

  Puedo cambiar los valores de los atributos para afectar el proceso de decisión (en mi AS o en los vecinos)

Políticas de Control (2)

- Tres pasos:
  1. Identificar las rutas o prefijos
  2. Permitir o negar las rutas
  3. Manipular los atributos
- Veremos algunos detalles en el laboratorio

Políticas de Control (3)

Listas de Distribución

- Por peer BGP
- Entrantes o Salientes
- Basadas en prefijos
- Ej: no anunciar al peer 200.108.192.1 el prefijo 172.16.10.128/25

Políticas de Control (4)

Listas de filtrado

- Permite filtrar rutas basándose en el AS_PATH
- Tanto entrantes como salientes
- Ej: no permitir anuncios cuyo AS_PATH comience con el AS 100

Políticas de control (5)

- Para políticas más complejas y manipulación de atributos, en Cisco se utilizan route-maps

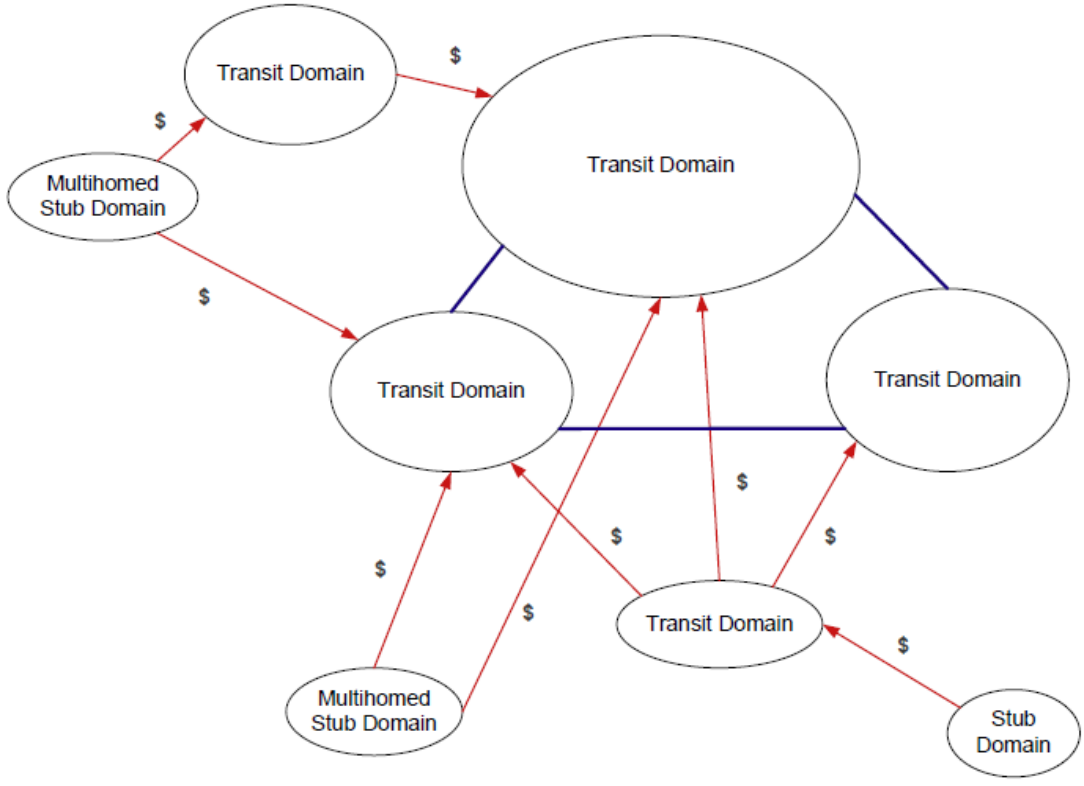
```
route-map pref permit 10
   match as-path 100
   set local-preference 250
route-map pref permit 20
   match ip address 1
   set local-preference 300
route-map pref permit 30
```

Valley-free routing

- En general las relaciones entre sistemas autónomos, se pueden clasificar de acuerdo a las relaciones comerciales entre las entidades
- En su forma más básica, se pueden clasificar en:
  - –Relaciones Cliente-Proveedor
    - •Cliente paga al proveedor
  - –Relaciones entre pares
    - •No hay intercambio de dinero

Valley-free routing

Types and Hierarchy of Autonomous Systems

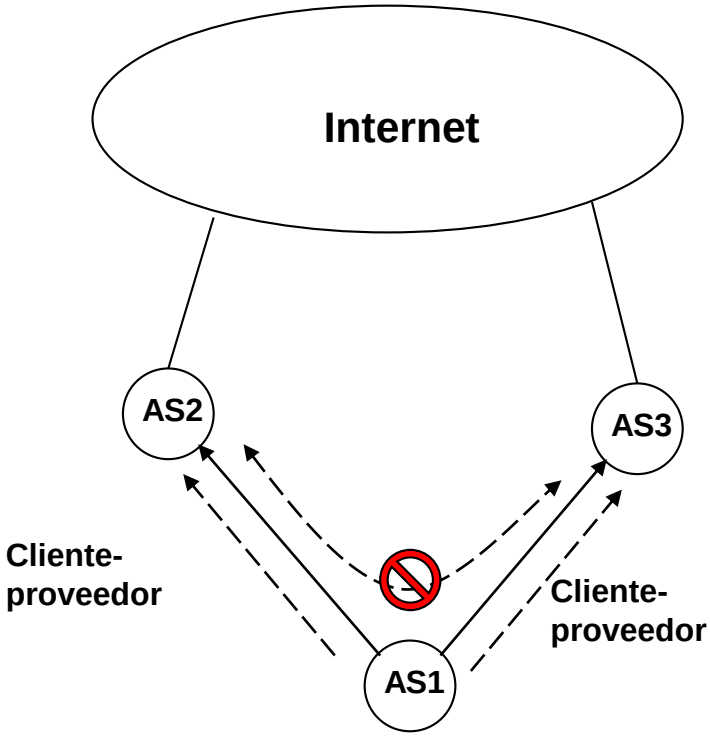


Marcelo Yannuzzi, Curso "Graphs on path vectors"

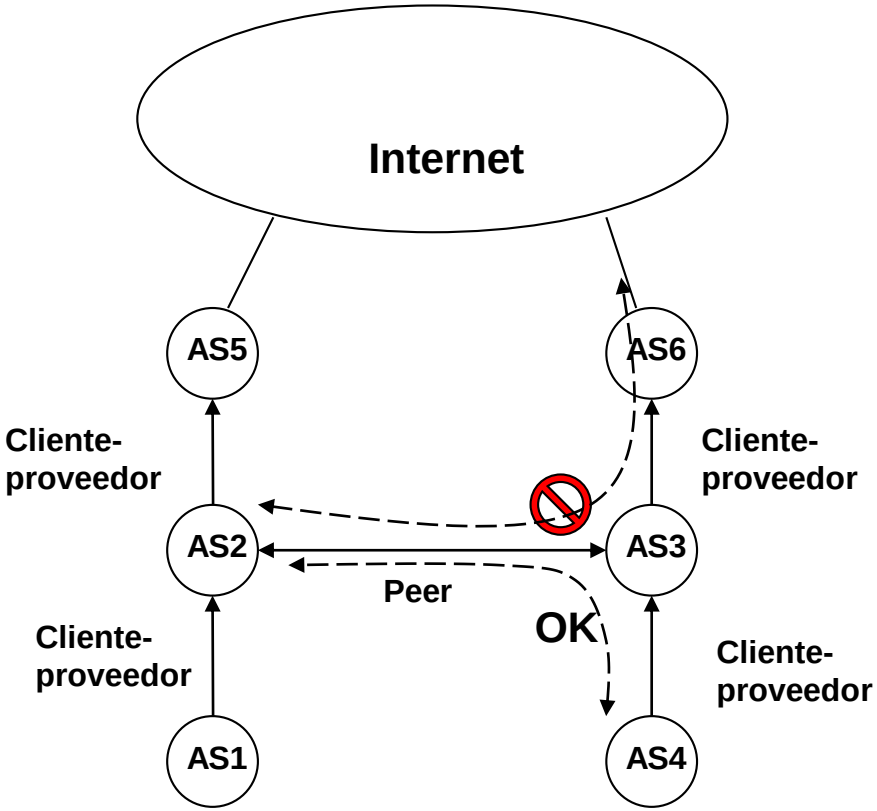
Valley-free routing

- En general, las políticas de enrutamiento siguen las relaciones comerciales
- No permito tránsito si no me pagan por ello
- La política mas básica que suele encontrarse sigue esta idea. Se le suele llamar “política libre de valles”
- En general también los valores de local preference siguen las relaciones comerciales

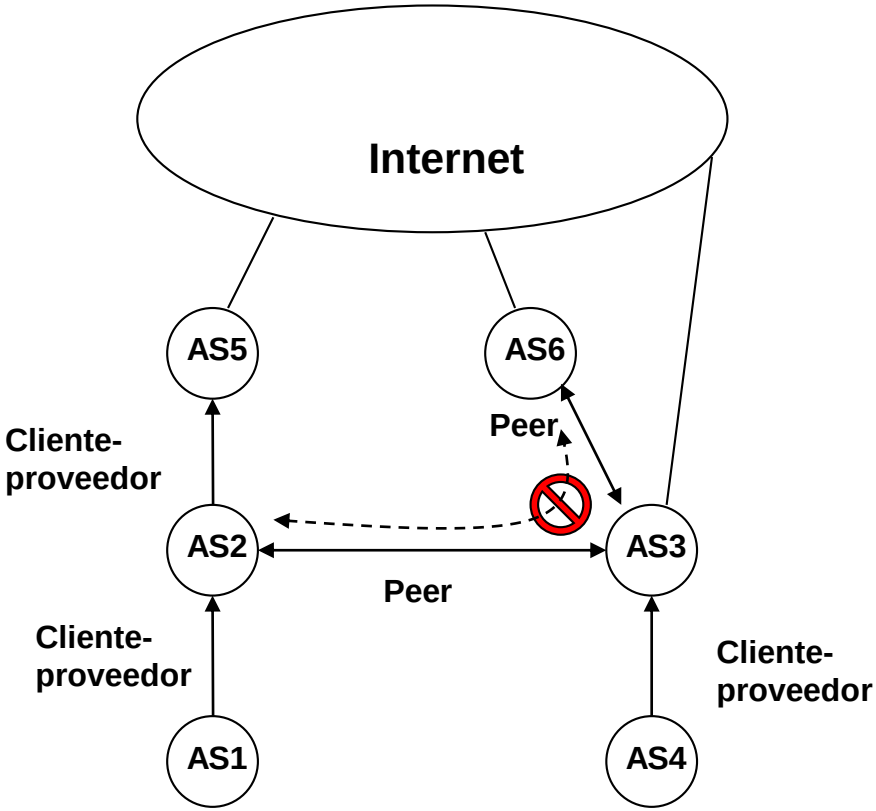
Tráfico no permitido: Proveedor-cliente-proveedor



Tráfico no permitido: peer-peer-proveedor



Tráfico no permitido
Peer-peer-peer



Agenda (4)

- Conceptos Fundamentales de BGP
- Análisis del protocolo (BGP-4)
- Atributos de BGP y políticas de control
- IBGP mesh
- Sumarización y anuncios (CIDR)
- Damping
- Algunos problemas
- Extensiones multiprotocolo
- Salidas reales y datos de actualidad

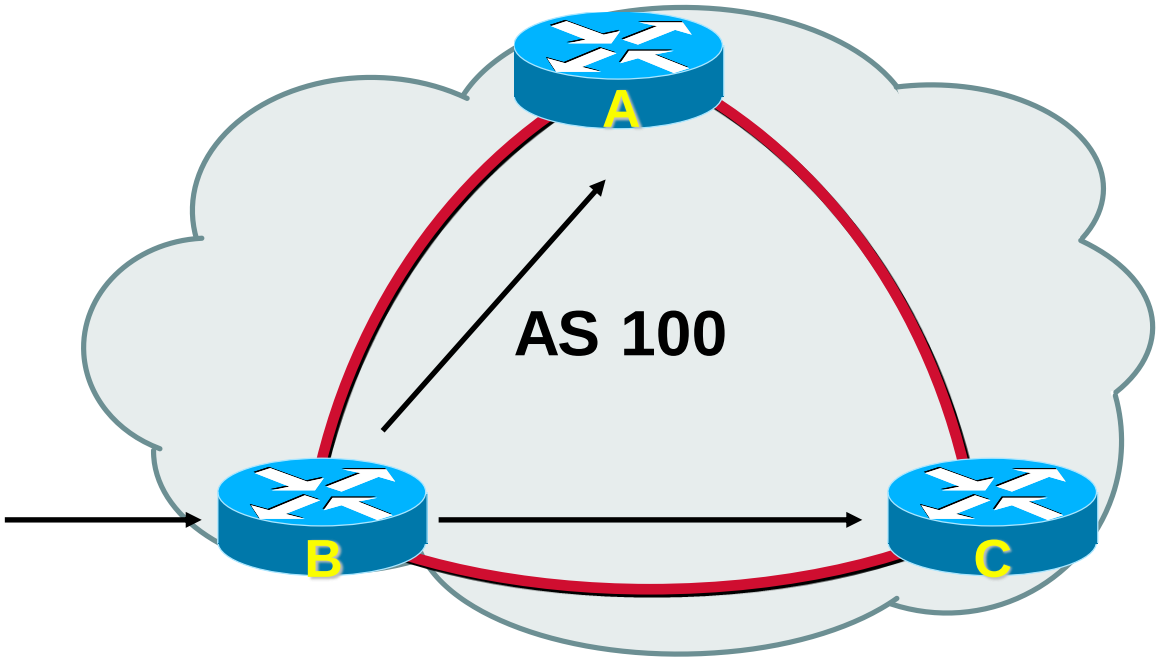
IBGP Mesh y soluciones

- En principio se precisa una sesión entre cada par de enrutadores hablando BGP

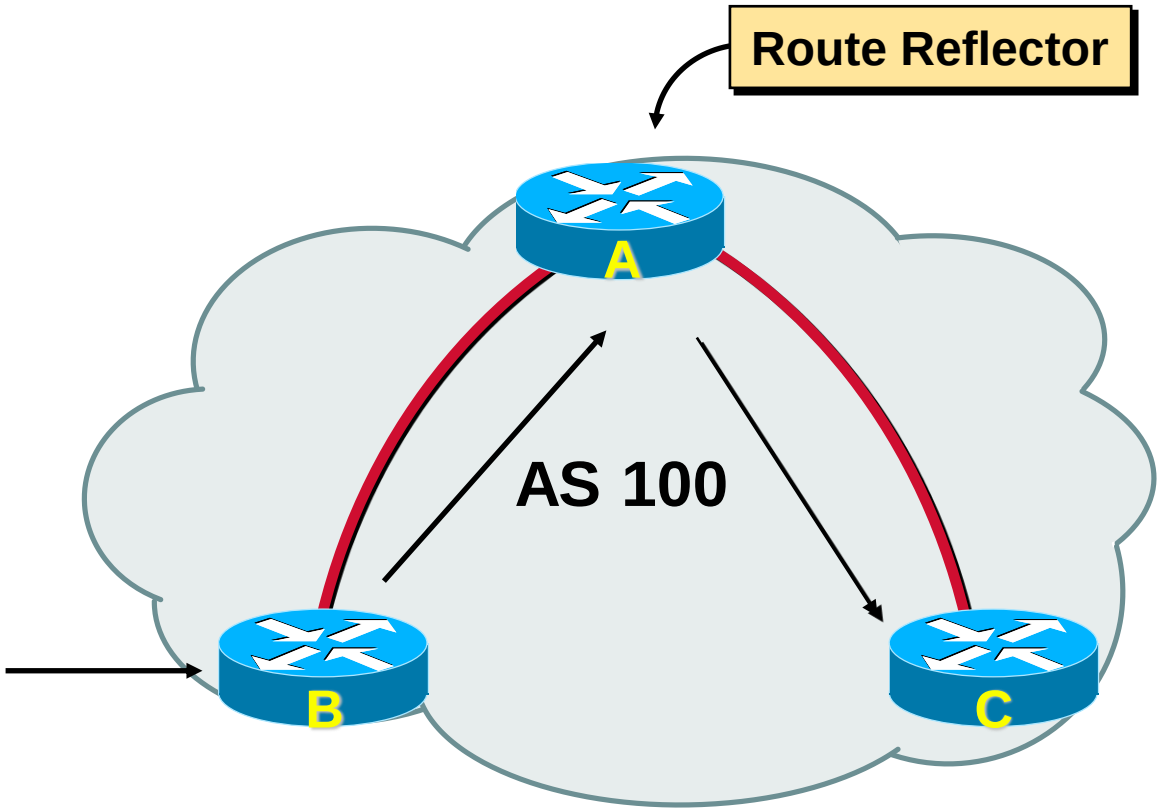
  Por la regla que impide propagar por IBGP lo aprendido por IBGP (para evitar loops)

- No escala
- Alternativas:
  - Reflectores de rutas (Route reflectors, RR)
  - Confederaciones

Funcionamiento normal con IBGP mesh



Funcionamiento con un Route Reflector



Route Reflector: Beneficios

- Evita el mesh IBGP
- Normalmente no altera el forwarding de los paquetes
- Pueden coexistir BGP peers normales
- Pueden configurarse múltiples RR por redundancia
- Puede haber una jerarquía de RR (varios niveles)
- Es fácil migrar de mesh a RR

Route Reflector: Definiciones

- Route reflector (RR)
- Cliente de reflector (RRC): peer BGP que recibe rutas internas repetidas por un RR
- Cluster: uno o más RR y sus clientes
- Cluster ID: identificación del cluster, importante cuando tengo más de un RR
- No-cliente: peer iBGP que no es RRC
- Normal BGP peer: no cliente, o externo

Route Reflector: Operación

- RR recibe anuncios de clientes y de no-clientes
- RR elige el mejor camino
- Si el mejor camino viene de un cliente => lo refleja tanto a sus no-clientes, como a sus clientes (excepto quien originó el mensaje)
- Si el mejor camino viene de un no-cliente => lo refleja sólo a los clientes
- Si el best path viene de EBGP, lo envía tanto a sus no-clientes, como a sus clientes

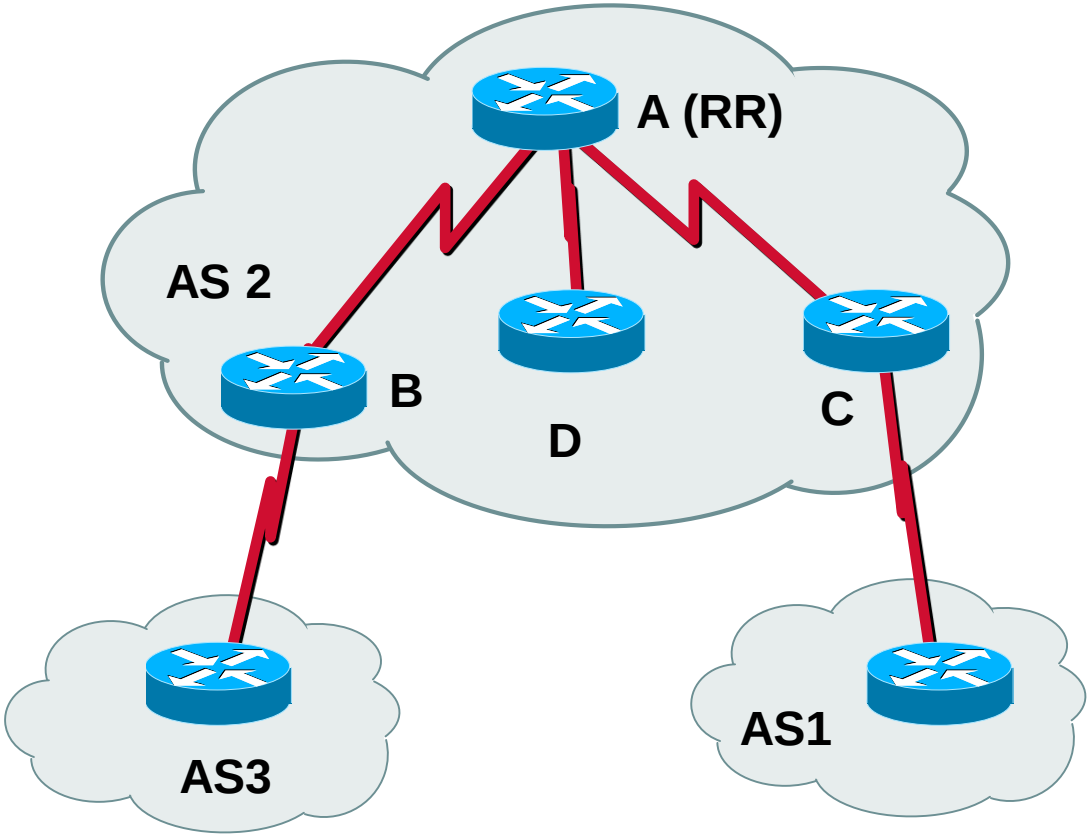
Route Reflector

- Permite dividir el sistema autónomo en múltiples clusters
- Al menos un RR y algunos clientes por cluster
- Route reflectors estarán fully-meshed
- No se necesita mesh entre clientes!!!

Recomendaciones

- Seguir la topología física en la medida de lo posible
- Evitar modificar los atributos de las rutas reflejadas
  - De ser necesario, tener cuidado para evitar loops
- En caso de múltiples reflectores en un cluster, configurar el mismo CLUSTER_ID
  - Esta recomendación está en discusión

Route Reflector: Ejemplo

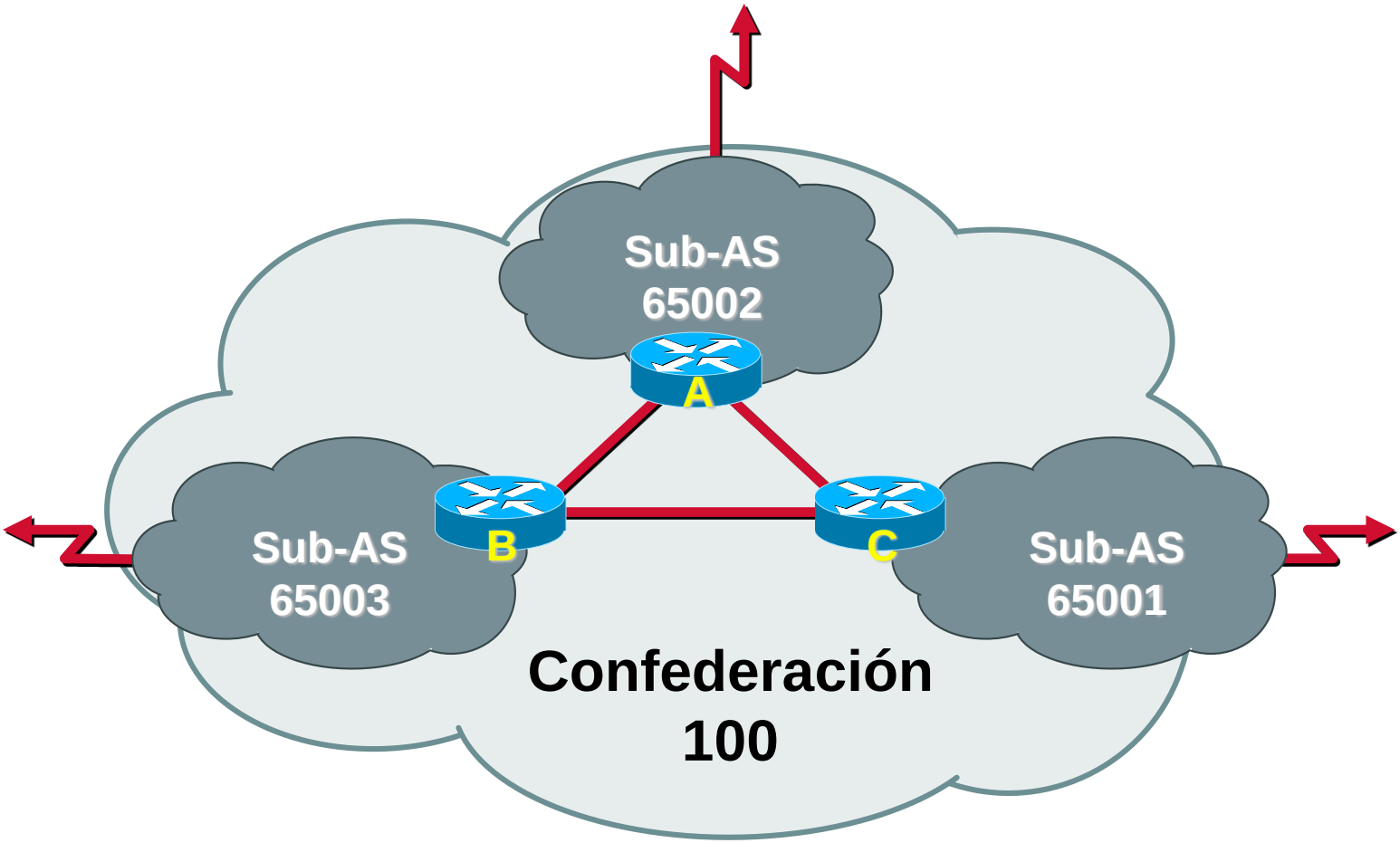


Confederaciones (1)

- Colección de Sistemas Autónomos - sub-AS
- Desde el mundo exterior se ve como un único AS
- Se usan los números de AS reservados para los sub-AS internos (AS privados : 64512 - 65535)
- Cada sub-AS en arquitectura fully-meshed
- EBGP entre los sub-AS

  Manteniendo MED, local-pref, next-hop

Confederaciones (2)



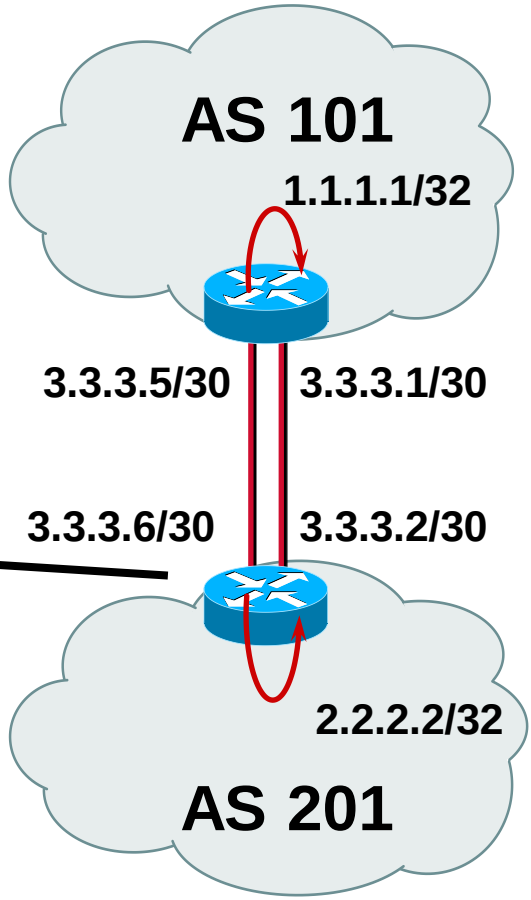
Confederaciones: Beneficios

- Soluciona y escala el problema de IBGP full-mesh
- Puede ser usado conjuntamente con Route Reflectors
- Admite la aplicación de políticas para enrutar tráfico entre los distintos sub-ASs

Redundancia y Balance de Carga

- Se usa ebgp-multihop
- Se usa una interfase loopback en cada router
- Se deben tener dos rutas explícitas en cada router hacia la loopback del peer
- Ejemplo de configuración:

```
router bgp 201
neighbor 1.1.1.1 remote-as 101
neighbor 1.1.1.1 update-source loopback0
neighbor 1.1.1.1 ebgp-multihop
!
ip route 1.1.1.1 255.255.255.255 3.3.3.1
ip route 1.1.1.1 255.255.255.255 3.3.3.5
```



Agenda (5)

- Conceptos Fundamentales de BGP
- Análisis del protocolo (BGP-4)
- Atributos de BGP y políticas de control
- IBGP mesh
- Sumarización y anuncios (CIDR)
- Damping
- Algunos problemas
- Extensiones multiprotocolo
- Salidas reales y datos de actualidad

¿Cómo se generan anuncios?

- Redistribución de un IGP: NO recomendado!

(sí se suele utilizar redistribuir estáticas)

```
router bgp 14234
   redistribute ospf
   redistribute static
```

- generación local

En cisco, comando “network”

```
router bgp 109
   network 198.10.0.0 mask 255.255.0.0
 !
ip route 198.10.0.0 255.255.0.0 null 0
```

-> Tiene que existir la ruta en la tabla de ruteo local. Se suele usar estática a interfaz NULL

Sumarizaciones

- Combinar diferentes rutas en un único anuncio
- Se anuncia como proveniente del propio AS
- Una componente del bloque debe existir en la tabla de rutas
- Pueden utilizarse los atributos “Aggregator” y “Atomic Agregate”
- AS-path: se convierte en AS-SET o se elimina
- Cisco:

  aggregate-address <red> <mascara> [as-set]

  “summary-only”: solo se propaga la ruta sumarizada

MD5 signature

- RFC 2385
- Protección contra segmentos TCP “insertados” en la conexión existente (especialmente TCP Resets)
- Popular últimamente
- Hash MD5 de encabezado IP/TCP + datos + key (clave). Enviado en opción TCP
- (cisco)

  neighbor <direccion ip> password <string>

Agenda (6)

- Conceptos Fundamentales de BGP
- Análisis del protocolo (BGP-4)
- Atributos de BGP y políticas de control
- IBGP mesh. Soluciones
- Sumarización y anuncios (CIDR)
- Damping
- Algunos problemas
- Extensiones Multiprotocolo
- Salidas reales y datos de actualidad.

Inestabilidad de rutas (1)

- Flapeo de una ruta:
  - Ruta o camino “apareciendo y desapareciendo”
  - Modificaciones en el camino
- La inyección de rutas IGP en BGP dinámicamente genera flapeos
- Enlaces inestables generan flapeos
- Cambio de atributos o política => antiguamente se debía reinicializar la sesión TCP. Alternativas modernas: reconfiguración “on-the-fly” (por ejemplo RFC 2918, route-refresh)

Inestabilidad de rutas (2)

- Las rutas inestables se traducen en la generación de gran cantidad de mensajes BGP de UPDATE
- La inestabilidad se propagará a todos los enrutadores que reciban esos anuncios

Inestabilidad de rutas (3)

- Formas de minimizar inestabilidades:
  - Agregación (Supernets). ¿Dónde sumarizar?
    - Agregación en el borde del cliente
    - Agregación en el borde del SP
  - Desligar los anuncios de una ruta hacia el exterior de la propia existencia de la ruta en el AS (inyección estática de rutas hacia el exterior)

Route Flap Dampening (o Damping) RFC 2439

- PROBLEMA: el flapeo de rutas genera inestabilidades, consume ancho de banda y CPU de los enrutadores
- "SOLUCIÓN": reducir el alcance y la propagación de esas inestabilidades
- DAMPING: categoriza las rutas en dos grupos:
  - well-behaved
  - ill-behaved

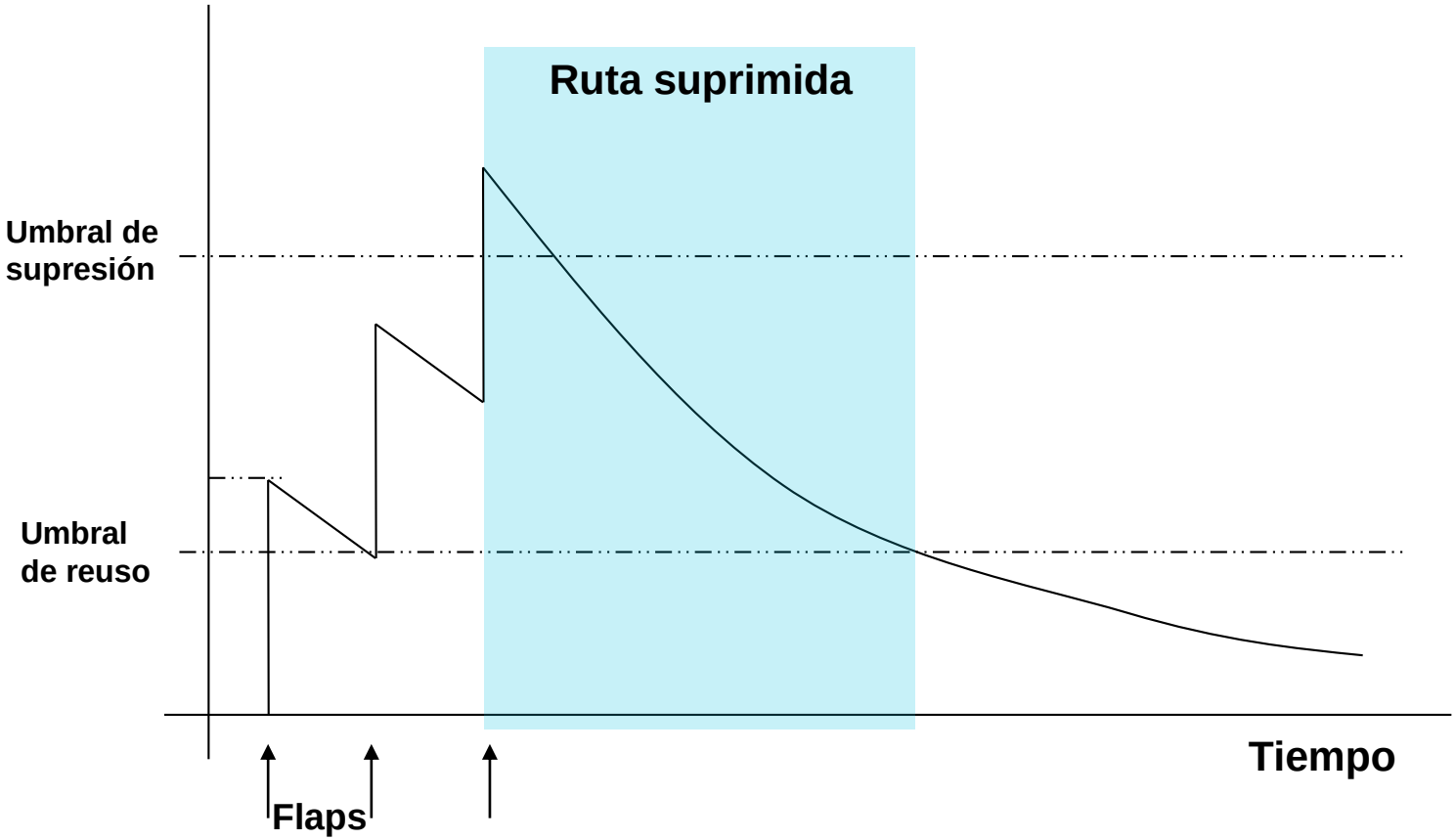
Route Flap Damping (2)

- Las rutas con mal comportamiento deben ser penalizadas de una manera que refleje las inestabilidades esperadas en las mismas a futuro
- Cada vez que una ruta flapea se la penaliza
- Se debe contar el número de veces que una ruta flapeo en un cierto período de tiempo

Route Flap Damping (3)

- Superado cierto umbral, la ruta se suprime y no se anuncia a otros peers de BGP (ya sean ASes clientes u otros SP)
- La ruta puede seguir siendo penalizada aún cuando ya haya sido suprimida
- Además la ruta puede permanecer penalizada aún cuando ya esté en estado estable (histéresis)

Route Flap Damping (4)



Route Flap Damping (5)

- Adicionar un entero (penalización) por cada flapeo
- Decaimiento exponencial de la penalización aplicada (lo fija quien penaliza, hay recomendaciones)
- Penalización por encima del umbral de supresión => no se anuncia la ruta
- Penalización por debajo del umbral de reutilización => se vuelve a anunciar la ruta

  Se asume que la ruta continuará con su comportamiento histórico...

Route Flap Damping (6)

- Los parámetros los elige quien penaliza
- Ej. Valores por defecto cisco:

  Se incrementa en 1000 cada flapeo (en 500 si cambian los atributos del anuncio)

  umbral de supresión: 2000

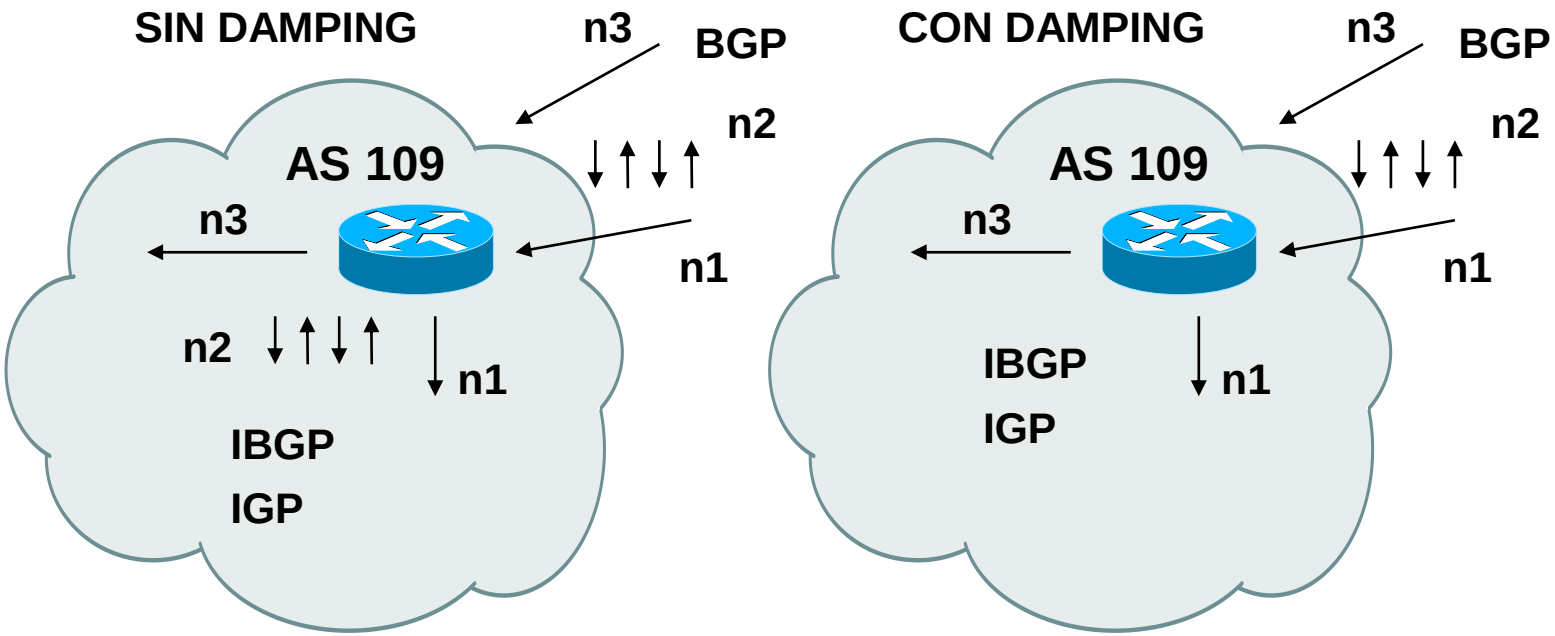
  umbral para reusar: 750

  Tiempo medio: 15 min

  Tiempo máximo de supresión: 4 x tiempo medio

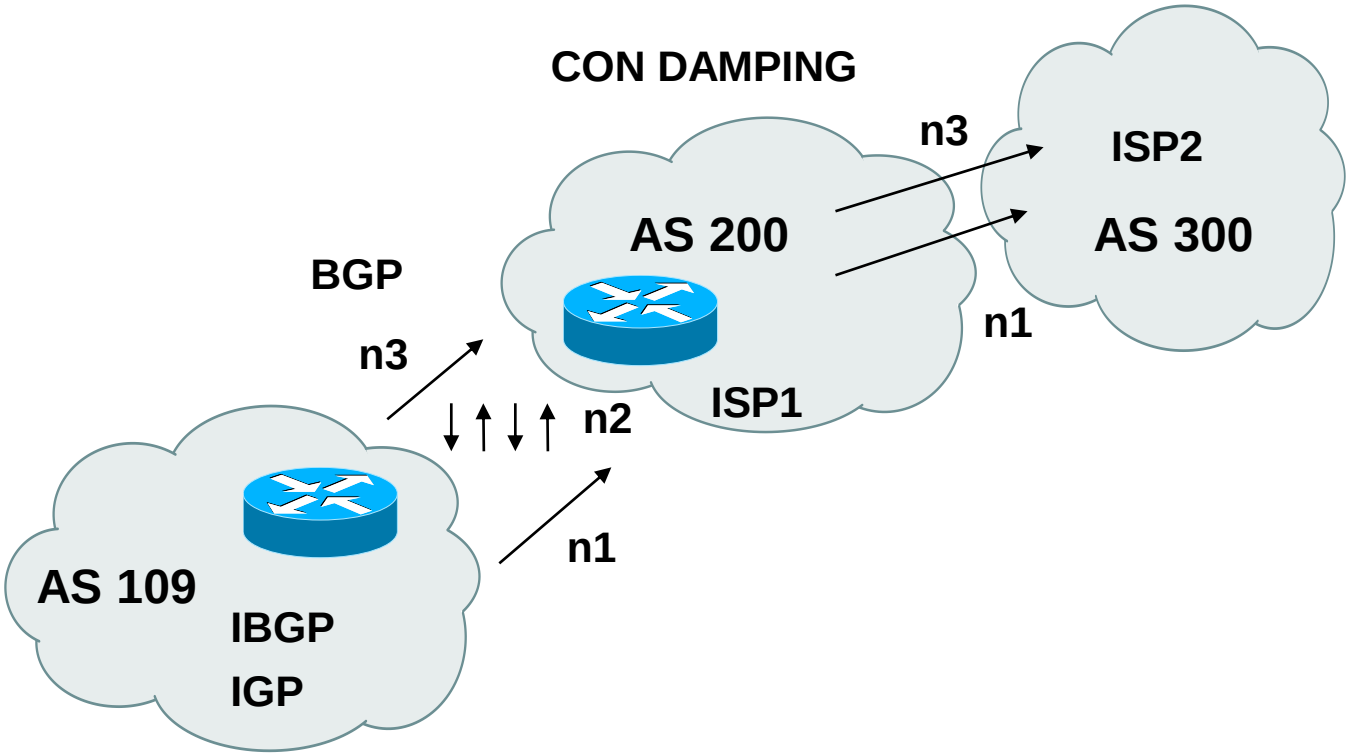
Route Flap Damping (7)

- **Estabilidad dentro del AS:**



Route Flap Damping (8)

- **Inestabilidades fuera del AS:**



Route Flap Damping (9)

- Están apareciendo recomendaciones y estudios que indican que el Damping es más perjudicial que útil
- Un cambio en una ruta, al propagarse, puede generar varios cambios en un AS remoto
- Aún se sigue usando, por lo que tenemos que tenerlo en cuenta

Algunos problemas actuales

- Velocidad de convergencia
  - –Estudios han demostrado que, pese a lo que se creía, BGP puede demorar tiempos muy altos (minutos/decenas de minutos) en converger globalmente
- MinRouteAdvertiseInterval (MRAI)
  - –Cada cuanto puedo propagar cambios a un prefijo
  - –Genera sus propios problemas
- “BGP churn”: tasa de anuncio de cambios
  - –Muy alta en algunos puntos de Internet
    - •Cientos de miles por día, con picos de miles por minuto

Extensiones Multiprotocolo

- RFC 2858
- Permite a BGP llevar información de otros protocolos (IPv4 Multicast, IPv6)
- 2 atributos (ONT):
- MP_REACH_NLRI

  Información de NLRI y Next Hop
- MP_UNREACH_NLRI

  Reemplaza las rutas que se dejan de anunciar (withdrawn routes)

IPv6 y BGP

- RFC 2545
- No hay mayores cambios en el funcionamiento. 3 páginas, mayormente indicando cómo usar las direcciones globales y link-local
- Se codifica en Extensiones Multiprotocolo
- Intentos de sustituir BGP han fracasado hasta ahora

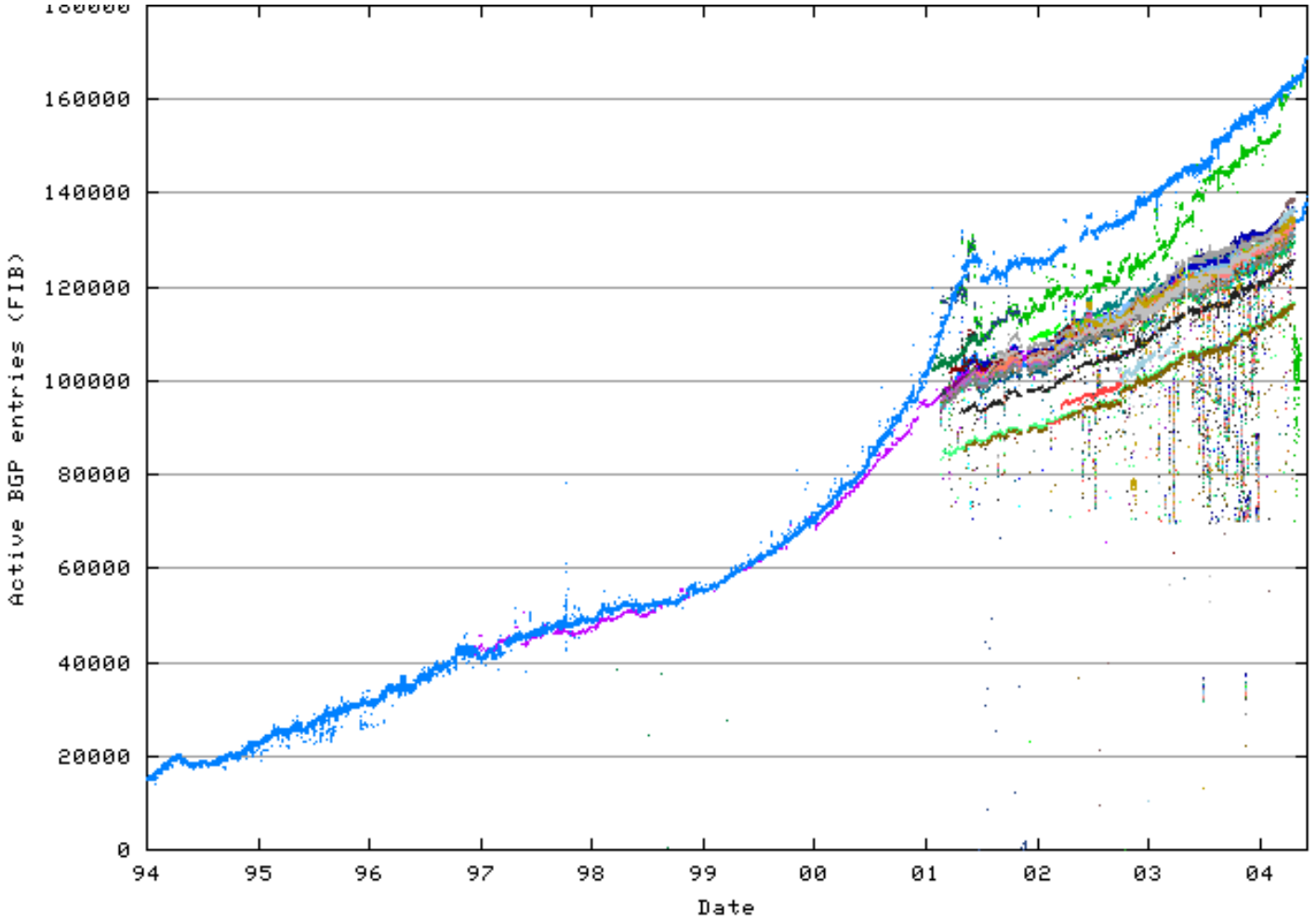
IPv6 (2)

- MP_REACH_NLRI
  - Address family (IPv6)
  - Next_Hop (IPv6)
  - NLRI (prefijos)
- MP_UNREACH_NLRI
  - – Prefijos que ya no son alcanzables

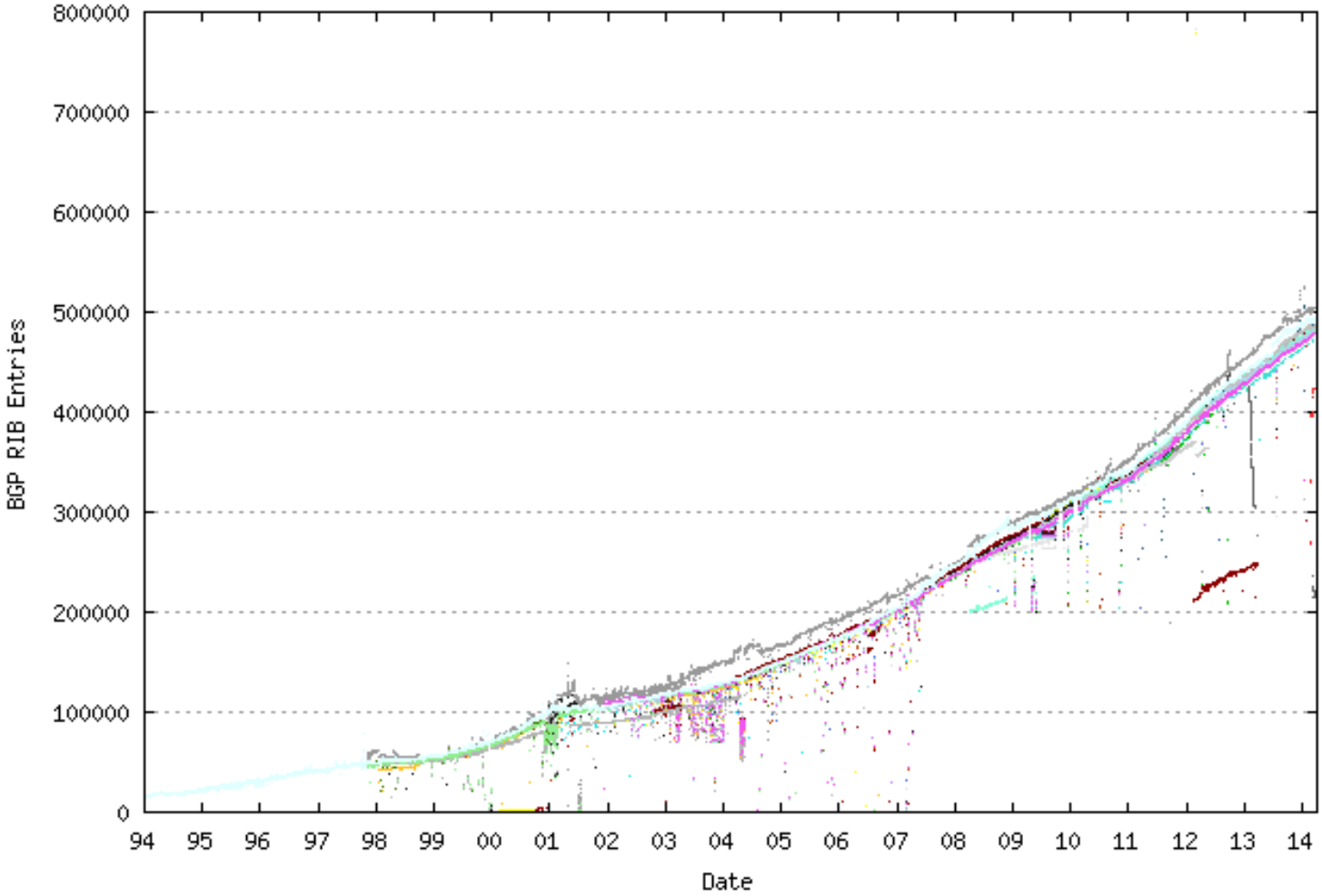
Agenda (8)

- Conceptos Fundamentales de BGP
- Análisis del protocolo (BGP-4)
- Atributos de BGP y políticas de control
- IBGP mesh. Soluciones
- Sumarización y anuncios (CIDR)
- Damping
- Algunos problemas
- Extensiones Multiprotocolo
- Salidas reales y datos de actualidad

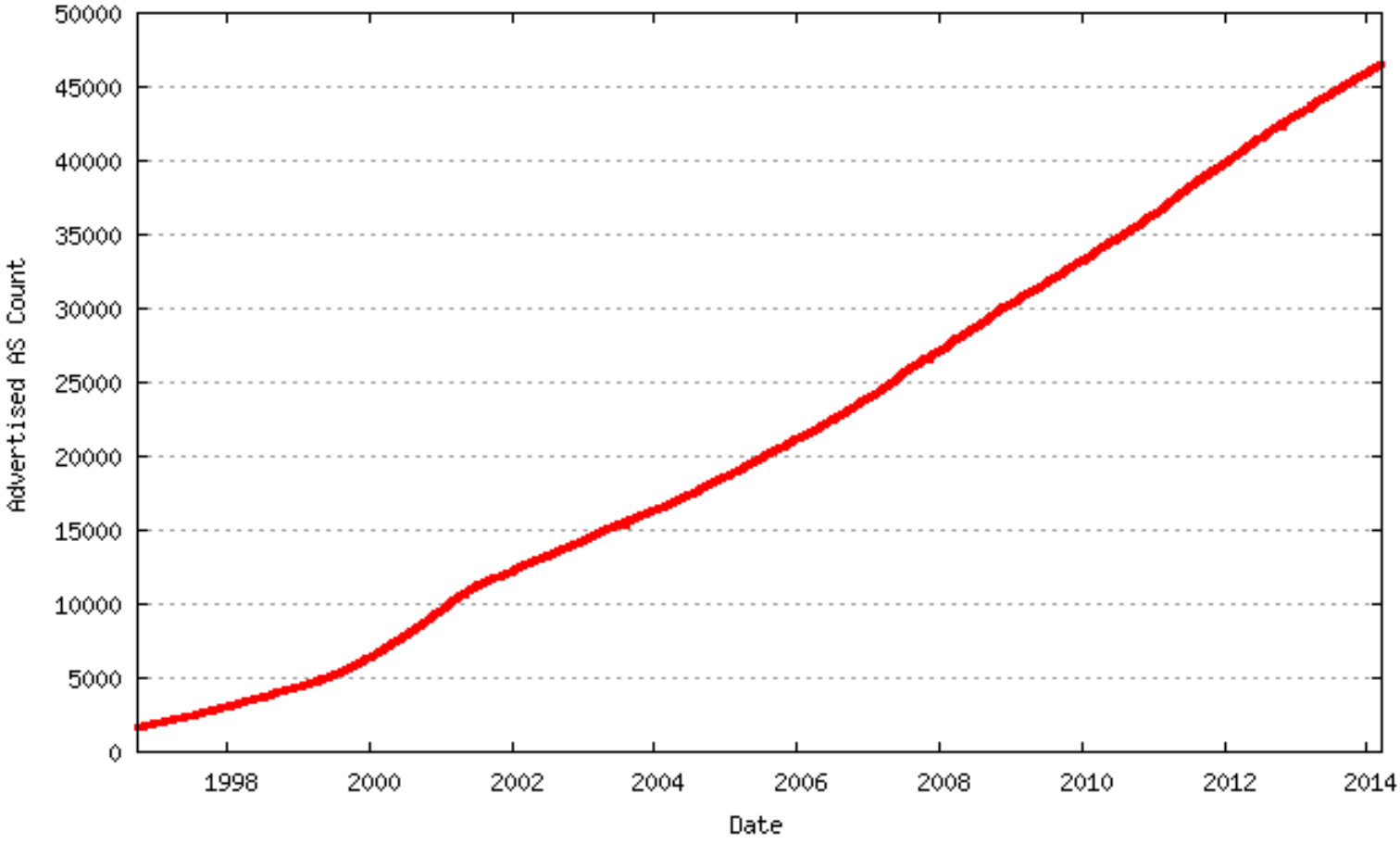
Prefijos en la DFZ (http://bgp.potaroo.net/)



Y 2014.....

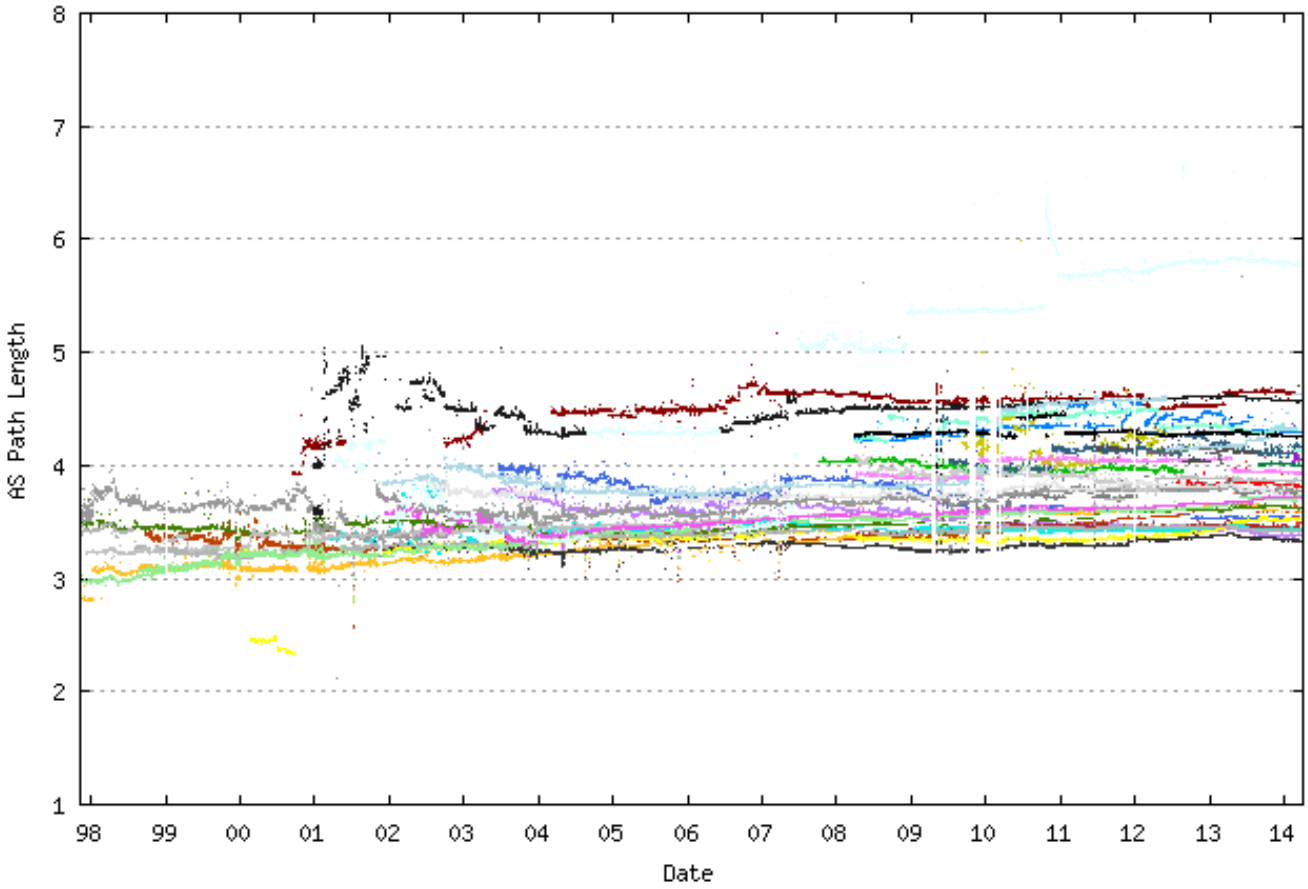

BGP RIB Entries
800000
700000
600000
500000
400000
300000
200000
100000
0
94
95
96
97
98
99
00
01
02
03
04
05
06
07
08
09
10
11
12
13
14
Date

Cantidad de AS activos (http://www.potaroo.net/tools/asn32/)

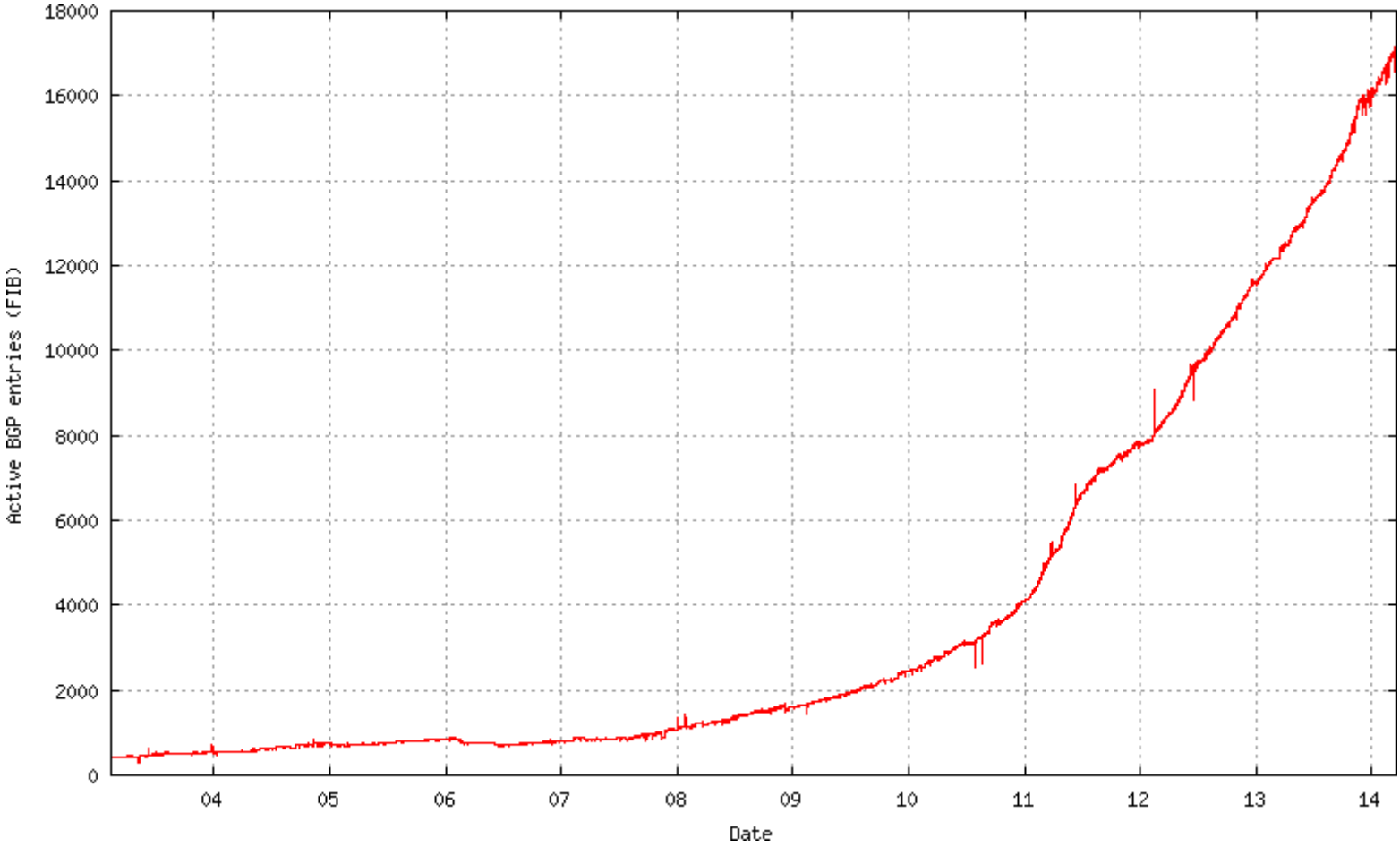


Largo promedio del AS-PATH

(http://bgp.potaroo.net/bgprpts/rva-index.html)



Publicaciones IPv6



Looking Glass y Route Servers

- ¿Cómo saber cómo se ven mis redes fuera de mi AS?
- Looking glass: consultas BGP, traceroute, ping.....
- Ejemplos:
- Ver en http://www.traceroute.org
- http://netmon.acad.bg/lg/
- https://www.sprint.net/lg/lg_start.php

BGP routing table entry for 164.73.0.0/16, version 47166667
Paths: (3 available, best #2, table Default-IP-Routing-Table)
Advertised to peer-groups:
edge
2828 701 6762 7303 6057 1797
206.111.0.17 (metric 92) from 165.117.162.201 (165.117.162.201)
Origin IGP, metric 100, localpref 100, valid, internal
Community: 2548:666
Originator: 165.117.162.231, Cluster list: 165.117.162.201, 165.117.162.200
2828 701 6762 7303 6057 1797
206.111.0.17 (metric 92) from 165.117.162.200 (165.117.162.200)
Origin IGP, metric 100, localpref 100, valid, internal, best
Community: 2548:666
Originator: 165.117.162.231, Cluster list: 165.117.162.200
2828 701 6762 7303 6057 1797
206.111.0.17 (metric 92) from 165.117.162.202 (165.117.162.202)
Origin IGP, metric 100, localpref 100, valid, internal
Community: 2548:666
Originator: 165.117.162.231, Cluster list: 165.117.162.202, 165.117.162.200

Route servers

- Servidores (enrutadores) que hablan BGP, accesibles públicamente
- Ejemplos:
- telnet://route-views.oregon-ix.net/
- telnet://route-server.ip.att.net/

route-server-eu>sh ip bgp

BGP table version is 63848938, local router ID is 212.62.0.13

Status codes: s suppressed, d damped, h history, * valid, > best, i - internal

Origin codes: i - IGP, e - EGP, ? - incomplete

Network	Next Hop	Metric LocPrf Weight Path
*>i 3.124.236.0/22	166.63.210.40	100 0 i
* i	166.63.210.41	100 0 i
*>i 4.0.0.0	166.63.210.40	100 0 3356 i
* i	166.63.210.41	100 0 3356 i
*>i 6.1.0.0/16	166.63.210.40	100 0 701 668 1455 i
* i	166.63.210.41	100 0 701 668 1455 i
*>i 6.2.0.0/22	166.63.210.40	100 0 701 668 1455 I

……

……

```
route-server>sh ip bgp | include 6057 1797
* 164.73.0.0   12.129.192.1       0 17233 7018 701 6762 7303 6057 1797 i
*              12.129.192.2       0 17233 7018 701 6762 7303 6057 1797 i
*>             199.106.200.1      0 17233 701 6762 7303 6057 1797 i

route-server>sh ip route 164.73.0.0
Routing entry for 164.73.0.0/16
  Known via "bgp 1838", distance 20, metric 0
  Tag 17233, type external
  Last update from 199.106.200.1 6d00h ago
  Routing Descriptor Blocks:
  * 199.106.200.1, from 199.106.200.1, 6d00h ago
      Route metric is 0, traffic share count is 1
      AS Hops 11, BGP network version 0
      Route tag 17233
```

```
route-server>show ip bgp 164.73.0.0
BGP routing table entry for 164.73.0.0/16, version 5147606
Paths: (3 available, best #3, table Default-IP-Routing-Table)
  Advertised to non peer-group peers:
  134.24.13.2 134.24.13.3
  17233 7018 701 6762 7303 6057 1797, (received & used)
    12.129.192.1 from 12.129.192.1 (12.129.192.1)
      Origin IGP, localpref 100, valid, external
      Community: 7018:5000 17233:666 17233:1001 17233:7018
  17233 7018 701 6762 7303 6057 1797, (received & used)
    12.129.192.2 from 12.129.192.2 (12.129.192.2)
      Origin IGP, localpref 100, valid, external
      Community: 7018:5000 17233:666 17233:1002 17233:7018
  17233 701 6762 7303 6057 1797, (received & used)
    199.106.200.1 from 199.106.200.1 (199.106.200.1)
      Origin IGP, localpref 100, valid, external, best
```

```
route-server>show ip route 12.129.192.1
Routing entry for 12.129.192.1/32
  Known via "static", distance 1, metric 0
  Routing Descriptor Blocks:
  * 12.129.193.233
      Route metric is 0, traffic share count is 1
```

```
route-server>traceroute www.fing.edu.uy
Translating "www.fing.edu.uy"...domain server (12.129.192.148) [OK]
Tracing the route to margarita.fing.edu.uy (164.73.32.3)

  1 mdf1-bi8k-1-ve-93.lax1.attens.net (12.129.193.233) [AS 17233] 0 msec 0 msec
  2 mdf1-gsr12-1-gig-1-0.lax1.attens.net (12.129.192.237) [AS 17233] 0 msec 0 ms
  3 gar3-p320.la2ca.ip.att.net (12.122.255.249) [AS 7018] 0 msec 0 msec 4 msec
.....................
  7 ggr1-p3100.sffca.ip.att.net (12.122.11.230) [AS 7018] 12 msec 8 msec 8 msec
  8 POS4-2.BR5.SAC1.ALTER.NET (204.255.174.177) [AS 701] 12 msec 12 msec 12
.....................
14 telecomitalia-gw1.customer.alter.net (65.208.85.114) [AS 701] 84 msec 84 mse
 15 bai1-san1-racc1.bai.seabone.net (195.22.220.216) [AS 6762] 208 msec 204 ms
 16 195.22.220.34 [AS 6762] 208 msec 208 msec 208 msec
 17 mun01rt-pos16-2-0.tasf.telecom.net.ar (200.3.32.133) [MPLS: Label 1292 Exp 0
.....................
20 icorecen2-backb.antel.net.uy (200.40.0.13) [AS 6057] 236 msec 236 msec 232
 21 ibgpcen1-f1-0.antel.net.uy (200.40.0.207) [AS 6057] 240 msec 240 msec 236
 22 seciu-ibgp.adinet.com.uy (200.40.160.9) [AS 6057] 240 msec 240 msec 236 m
 23 164.73.253.82 [AS 1797] !A  !A  *
```

Referencias

- RFC 4271 – RFC 1771 – BGP
- RFC 1772 – BGP – Aplicación
- Documentación varia de Cisco
- RFC 5065 – Confederaciones
- RFC 4456 - Route Reflectors
- http://www1.cs.columbia.edu/~ji/F02/
- Internet Routing Architectures. Sam Halabi. Cisco Press

Configuración En Cisco

- Definición de sistema autónomo y de vecinos

```
router bgp 65525
  neighbor 200.108.19.2 description cliente prueba
  neighbor 200.108.19.2 remote-as 20255
  neighbor 200.108.19.2 ebgp-multihop 10
  neighbor 200.108.19.2 update-source loopback 5
  neighbor 200.108.19.2 password 7 <password>
```

Configuraciones de ipv4

- En versiones actuales, la configuración de parámetros de IPv4 se encuentra en una sección address-family dentro de “router bgp”
- En versiones anteriores, se encuentra directamente en la configuración principal de BGP (bajo router bgp)
- En todas, la configuración de otras familias de direcciones está separada

Configuración IPv4

```
router bgp 65525
.....
 address-family ipv4
   neighbor 200.108.19.2 next-hop-self
   neighbor 200.108.19.2 prefix-list filtro in
   neighbor 200.108.19.2 filter-list 41 out
   neighbor 200.108.19.2 route-map mirmap in
   no auto-summary
   no synchronization
   exit-address-family
```

Generando anuncios

```
router bgp 65525
  network 192.168.20.0 mask 255.255.255.0
  redistribute rip
  redistribute connected
```

- Ojo, 192.168.20.0/24 debe estar en la tabla de ruteo para que se anuncie

Filtrado de prefijos. prefix-list

- Entrante o saliente
- Tradicionalmente, con access-lists. Hoy en día, se prefieren las prefix-lists
- Sintaxis:

ip prefix-list <nombre> <seq> <prefijo> ge <x> le <y>

ip prefix-list filtro 10 permit 192.168.0.0/16 ge 18 le 24

ip prefix-list filtro 30 deny 172.17.0.0/12 le 32

ip prefix-list filtro 40 permit 0.0.0.0/0 le 32

Filtrado por el as-path

- as-path access-lists
- El as-path se ve como una línea de caracteres, y se le puede aplicar una expresión regular
- Ejemplo:

  ip as-path access-list 41 deny _12345_

  ip as-path access-list 41 permit ^20255$

  ip as-path access-list 41 permit ^20255(_12345)*

.....

Expresiones regulares

- Sucesión de caracteres a machear
- Algunos caracteres especiales (ver próxima transparencia)
- Ejemplos:

  ^19422$ - machea con el as-path que contenga solamente el AS de Movistar

  ^19422 - machea con cualquier as-path que comience con 19422

  ^20255(_20255)*(_19422)*

Caracteres especiales

- . (punto) – Cualquier carácter
- * - cero o más secuencias del patrón
- + - una o más secuencias del patrón
- ? - cero o una secuencia del patrón
- ^ - Comienzo del string
- $ - fin del string
- _ - matchea espacio, “,”, {, }, (,), comienzo, fin
- [ ] - Indica un conjunto de caracteres a matchear

route-map

- Sucesión de bloques ordenados
- Cada bloque: match y set
- route-map nombre <permit|deny> seq

  Match <condiciones>

  Set < atributos>

Ejemplo

```
route-map pref permit 10
  match as-path 100
  set local-preference 250
route-map pref permit 20
  match ip address prefix-list filtro
  set local-preference 300
route-map pref permit 30
```

Route map

- Condiciones para el match:
  - as-path access-list
  - prefix-list o access-list (prefijos)
  - community-list
  - local-preference

Route-map (cont)

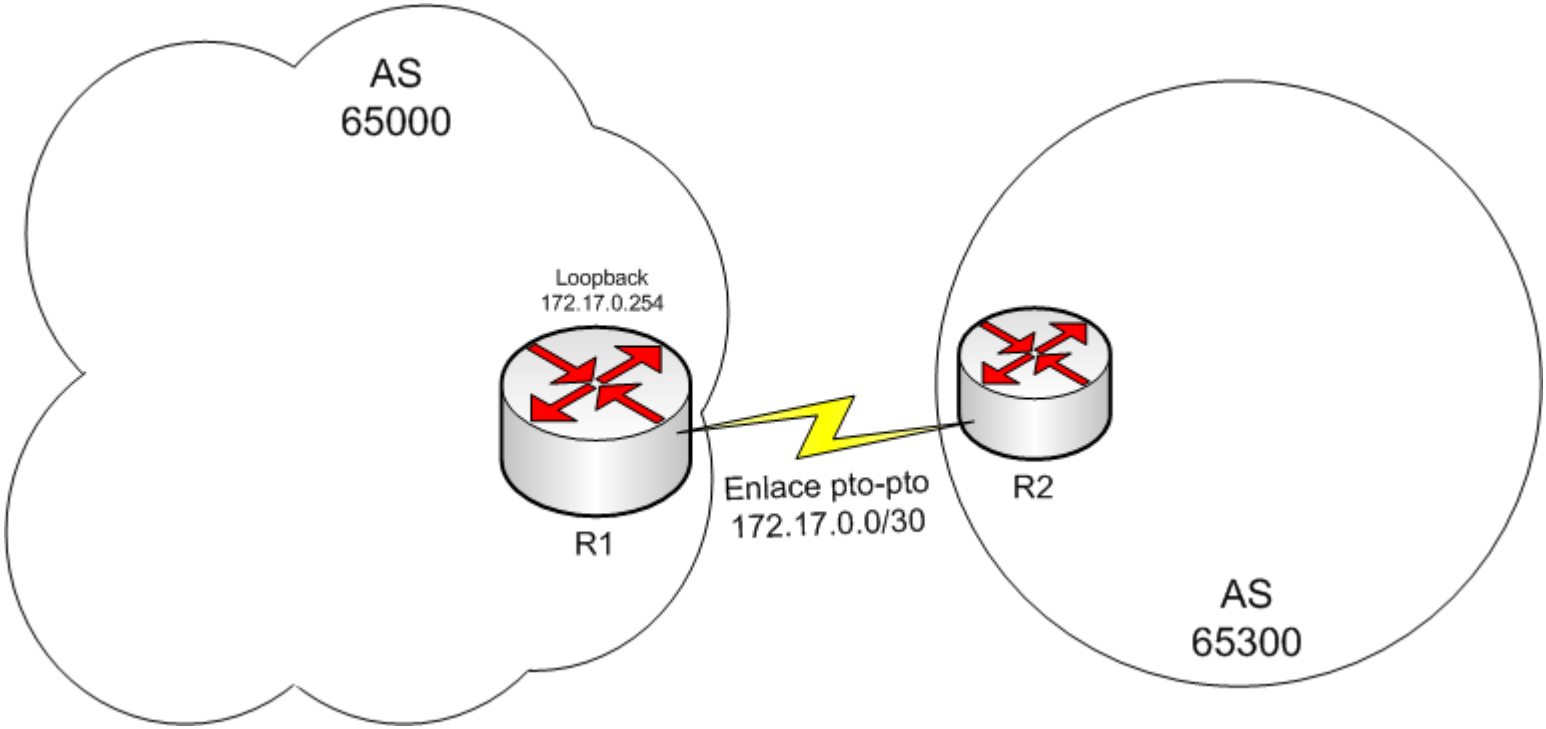
- Set:
  - set as-path prepend
  - set community
  - set comm-list
  - set dampening
  - set ip next-hop
  - set local-preference
  - set origin

Ejemplo: proveedor y su cliente

- Proveedor: AS 65000, cliente: AS 65300
- Prefijos del cliente:

  172.17.128.0/19

  172.17.160.0/20
- Prefijo más específico aceptado: /24
- Al cliente solo le interesa recibir la ruta por defecto


AS
65000
Loopback
172.17.0.254
R1
Enlace pto-pto
172.17.0.0/30
R2
AS
65300

Router del cliente

```
router bgp 65300
  neighbor 172.17.0.1 remote-as 65000
  neighbor 172.17.0.1 description mi proveedor
  address-family ipv4
    network 172.17.128.0 mask 255.255.224.0
    network 172.17.160.0 mask 255.255.240.0
    neighbor 172.17.0.1 prefix-list solodefault in
..
ip prefix-list solodefault seq 10 permit 0.0.0.0/0
ip prefix-list solodefault seq 20 deny 0.0.0.0/0 le 32
```

Router del proveedor

```
router bgp 65000
  router-id loopback 1
  neighbor 172.17.0.2 remote-as 65300
  neighbor 172.17.0.2 description cliente 65300
  address-family ipv4
    neighbor 172.17.0.2 prefix-list solodefault out
    neighbor 172.17.0.2 prefix-list cliente653 in
    neighbor 172.17.0.2 filter-list 50 in
```

```
ip prefix-list cliente653 seq 10 permit 172.17.128.0/19 le 24
ip prefix-list cliente653 seq 20 permit 172.17.160.0/20 le 24
ip prefix-list cliente653 seq 1000 deny 0.0.0.0/0 le 32
ip as-path access-list 50 permit ^65300$

! Si permito prepends
ip as-path access-list 50 permit ^65300(_65300)*$
```