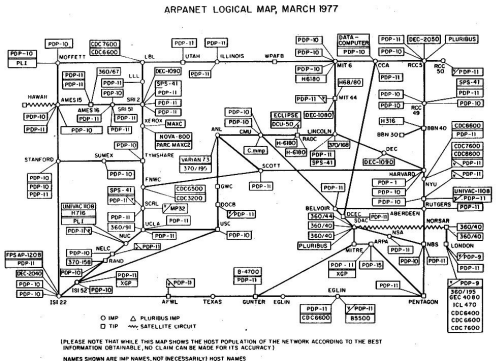
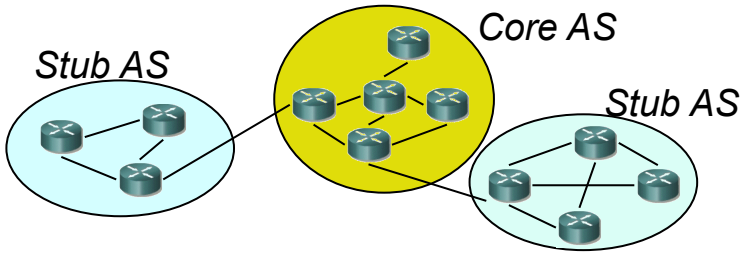
BGP: Introducción

Un poco más de historia

- En el comienzo de los 80 ARPANET empleaba GGP
  - *Gateway-to-Gateway Protocol*
  - Distance Vector
  - Todo router conocía ruta a todas las redes
  - Distancia en saltos
  - No escala



- RFC 827 (1982) propuso:
  - Migración a un sistema de redes autónomas interconectadas: *Autonomous Systems*
  - Añade un nuevo nivel en la jerarquía
  - Red de redes ⇒ Red de sistemas autónomos
  - Los routers *externos* compartirían información de enrutamiento mediante un protocolo llamado EGP
- GGP se convierte en el primer IGP, para el AS *Core*



EGP (RFC 904)

- Distance Vector
- Da alcanzabilidad de redes
- No tiene un algoritmo para buscar el mejor camino
- No envía suficiente información como para evitar ciclos
- Necesita una **topología sin ciclos**
- Intercambia mensajes entre *vecinos* o *peers*
  - *vecinos interiores (interior neighbors)*: en el mismo AS
  - Si no, son *exterior neighbors*
  - Configuración manual de quiénes son los vecinos

Core AS

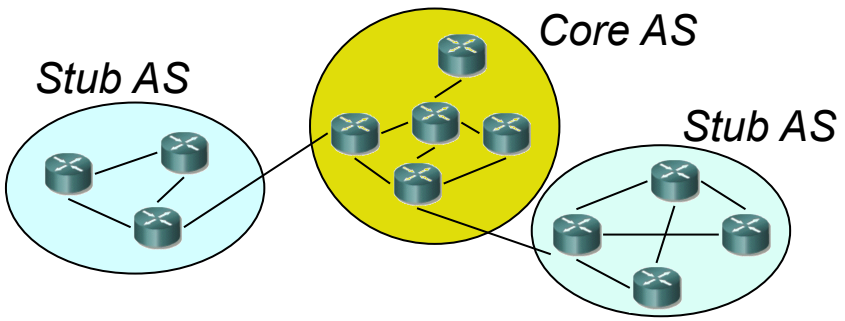
- Solo uno, a él se conectan los demás
- Sus routers pueden enviar información sobre redes diferentes de su AS

Stub AS

- Solo envían información sobre este AS

Problemas:

- Necesidad de topología más conectada
- Detectar bucles
- Reducir tiempos de convergencia
- Especificar *routing policies*



El nuevo protocolo

¿Distance vector?

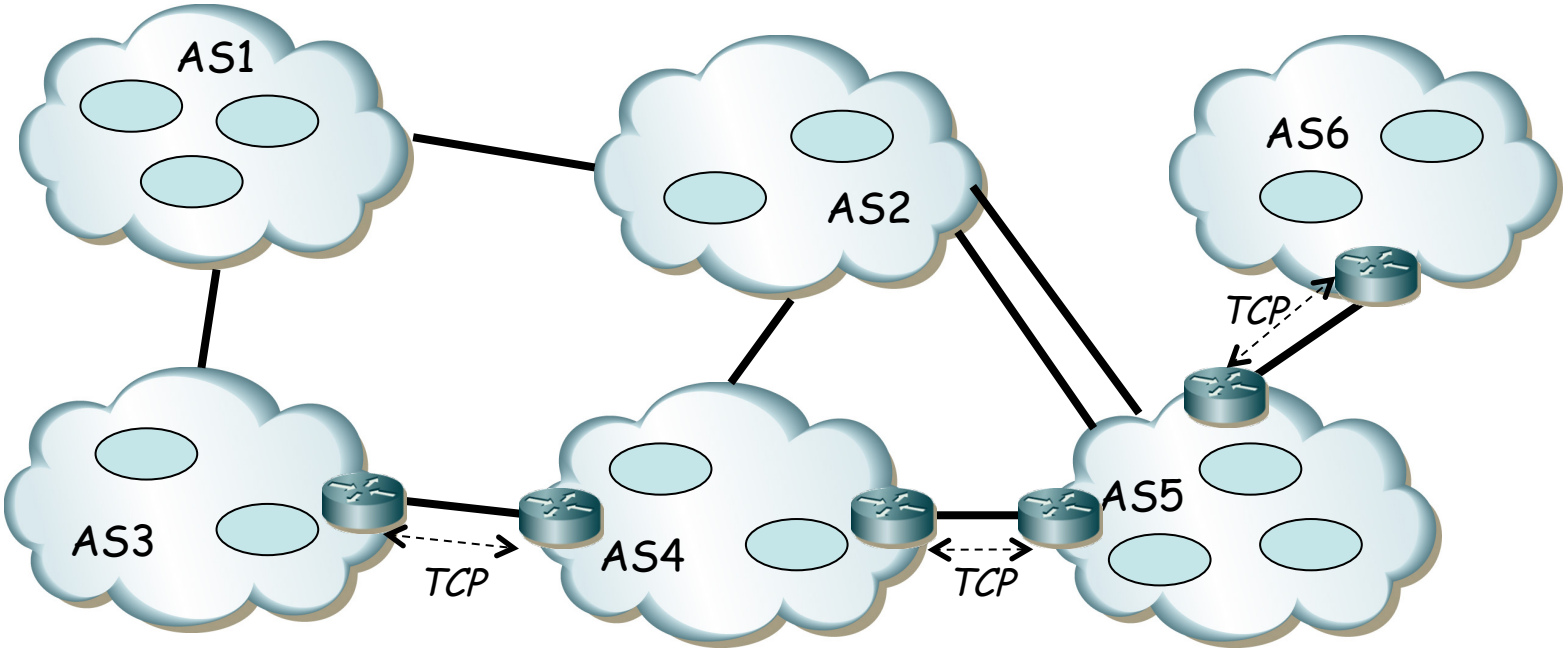
- En ocasiones no se quiere ir por la ruta más corta (políticas)
- Inestable

¿Link State?

- No escala bien (aunque se puede lograr, como PNNI)
- Base de datos de LSPs grande
- Coste de calcular las rutas mediante Dijkstra

BGP

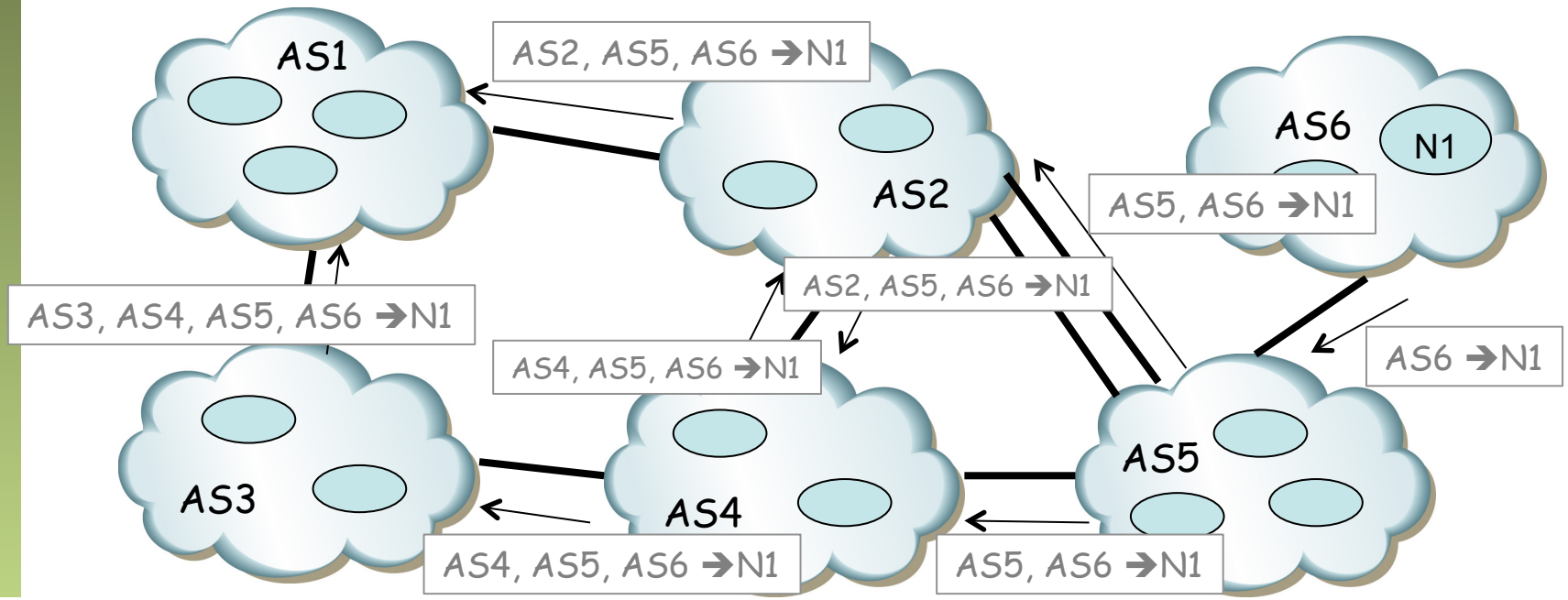
- *Border Gateway Protocol*
- BGP-4, RFC 4271
- BGP-4 primera versión classless
- Protocolo Interdomain estándar *de facto*
- Comunicación fiable mediante conexión TCP entre routers adyacentes
- Puerto 179



BGP

Path Vector

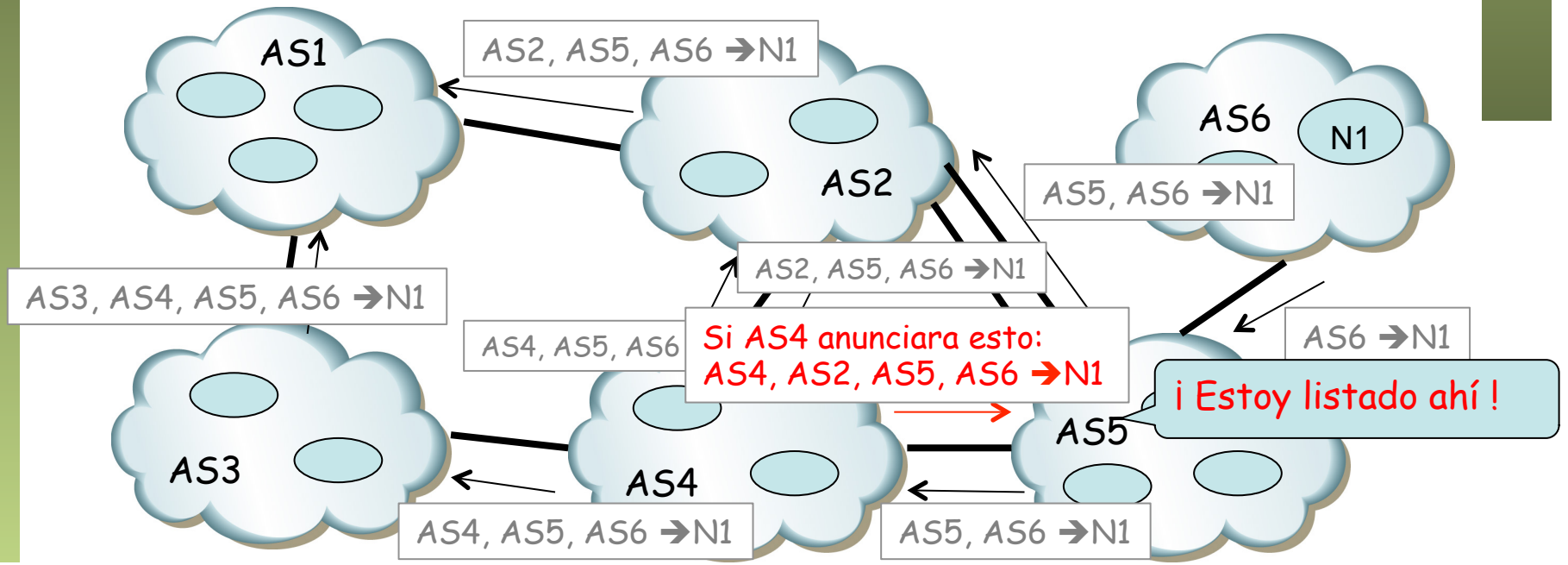
- Calcula caminos a prefijos
- Como DV recibe de vecinos, calcula sus rutas y envía a vecinos
- En vez de métrica anuncia la lista de AS en cada camino (. . .)
- Por defecto elige el camino que pasa por menor número de ASs



BGP

Path Vector

- Anunciar el camino permite evitar los ciclos
- El menor número de ASs no quiere decir que sea el menor número de saltos por routers



BGP: Mensajes

Mensajes

- Primero se establece la conexión TCP entre los dos *BGP speakers*
- Cuatro mensajes obligatorios

OPEN

- Tras establecerse la conexión
- Router especifica parámetros de operación: versión, identificador, AS number, *hold time*, *capabilities*, etc.
- Suele ir seguido de un intercambio de todas las rutas

KEEPALIVE

- Para comprobar periódicamente el *peering*
- Se da por rota la sesión si pasa el *hold time* sin recibirlo

NOTIFICATION

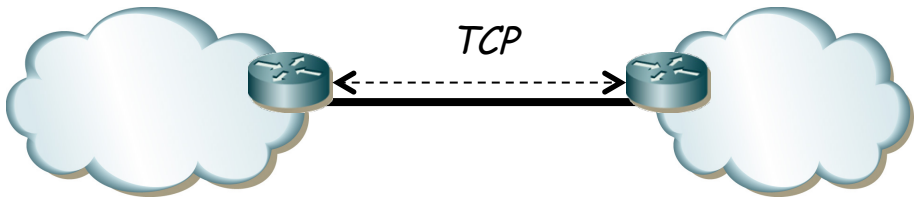
- Cuando se detecta un error
- Termina la conexión

UPDATE

- Anuncia información de enrutamiento (nuevas rutas o eliminar otras – *withdraw* –)
- Anuncia un solo camino por mensaje
- Anuncia cuando ha calculado una nueva mejor ruta al destino
- Si deja de poder alcanzarlo anuncia eso también
- Prefijo / Longitud
- ***Atributos del camino***: permiten a BGP elegir el mejor

ROUTE-REFRESH (opcional)

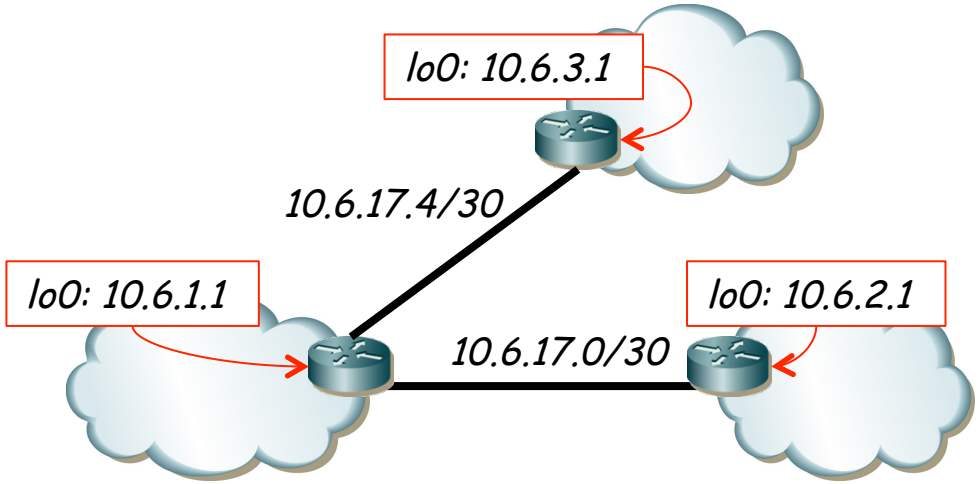
- Para pedir que vuelva a anunciar los prefijos que conoce



eBGP vs iBGP

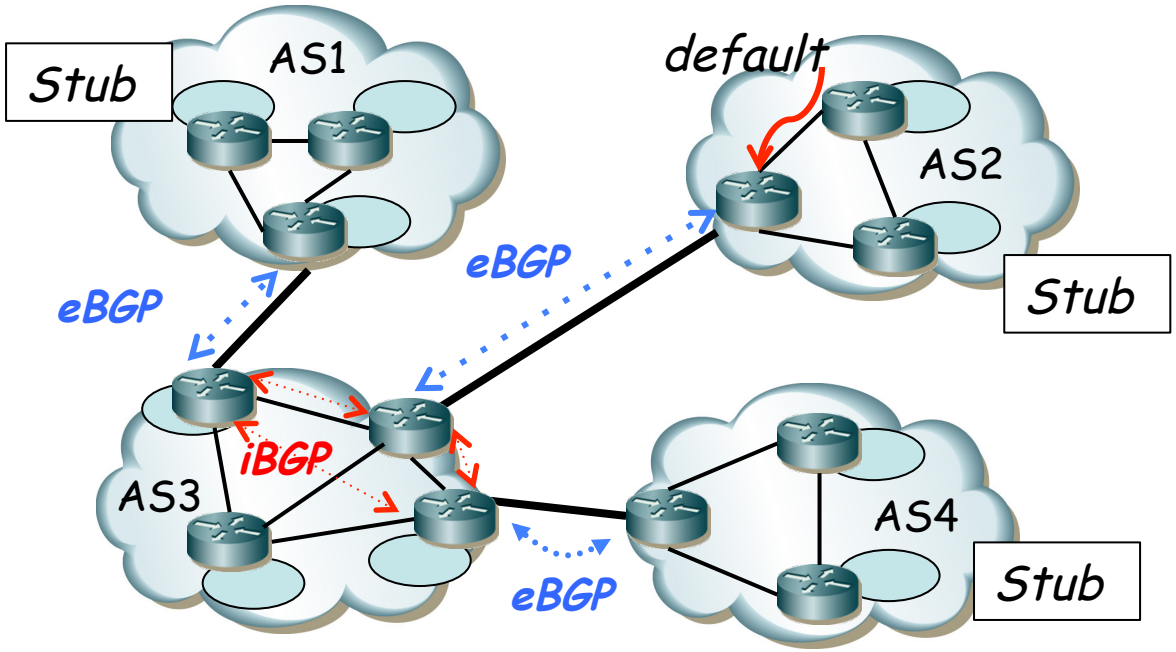
Direccionamiento

- El enlace entre dos ASBRs empleará un direccionamiento
- Frecuentemente es parte de la asignación de uno de los dos ASs
- Un router con más de un *peer* tendría una dirección diferente para cada uno, lo cual complica la gestión
- Necesita un identificador único del router, pero si es una de sus direcciones IP, ¿qué sucede si ese interfaz falla?
- Para tener una única identificación se emplea la de un interfaz de *loopback*
- Se envían los paquetes con TTL=2 si se usa *loopback* y con TTL=1 si no se usa



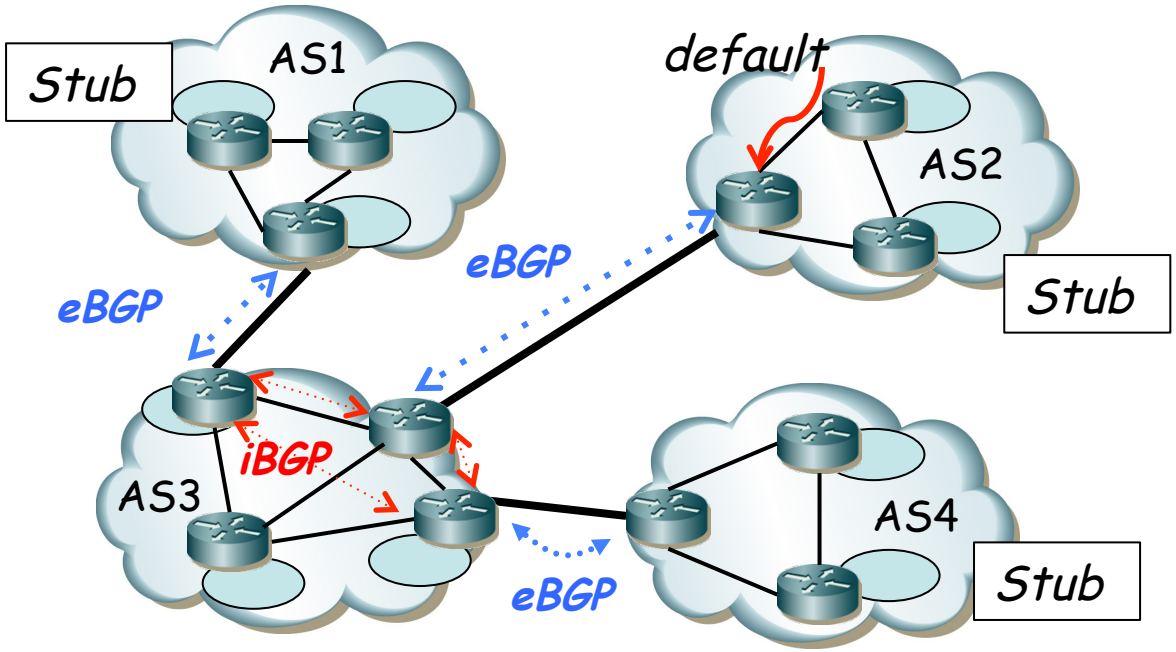
Peering en BGP

- Los *peers* de un proceso BGP pueden estar:
  - En otro AS: *external peer* ⇒ ***eBGP***
  - En el mismo AS: *internal peer* ⇒ ***iBGP***
- (…)



Peering en BGP

- En el mismo AS el *peering* iBGP forma una malla porque...
- No se pasan por iBGP prefijos aprendidos por iBGP
- Reconoce si es del mismo AS porque en el OPEN anuncia el ASN
- No interesa difundir todas las rutas al IGP (escalabilidad)
- iBGP permite que otros ASBRs aprendan los prefijos a anunciar
- El ASN se añade a la ruta al hacer anuncio a otro *eBGP*



ASBR = Autonomous System Border Router

Atributos en BGP

Path Attributes

- Son características de una ruta BGP

Tipos según se soporten:

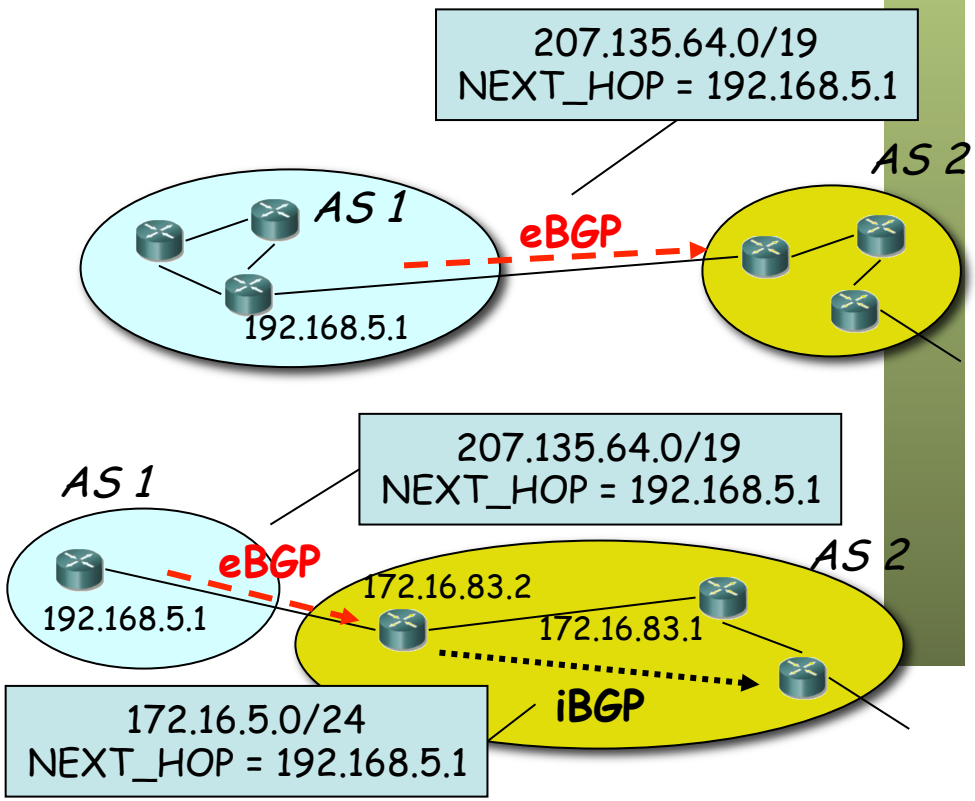
- *Well-known: mandatory* (en update) o *discretionary*
- *Optional: transitive* o *nontransitive*

ORIGIN (well-known mandatory)

- IGP, EGP o Incompleto (rutas estáticas)

NEXT_HOP (well-known mandatory)

- Si son *External Peers* es la IP del interfaz del router anunciante
- Si son *Internal Peers* y
  - Destino fuera del AS: IP del *peer* externo
  - Destino en el mismo AS: (…)



"*well-known*" : Debe soportarlo
"*Optional*" : No está obligado a soportarlo
"*mandatory*" : Debe aparecer en los mensajes
"*discretionary*" : Puede no aparecer en los mensajes
"*Transitive*" : Debe reenviarlo
"*Nontransitive*" : *No debe reenviarlo*

Path Attributes

- Son características de una ruta BGP

Tipos según se soporten:

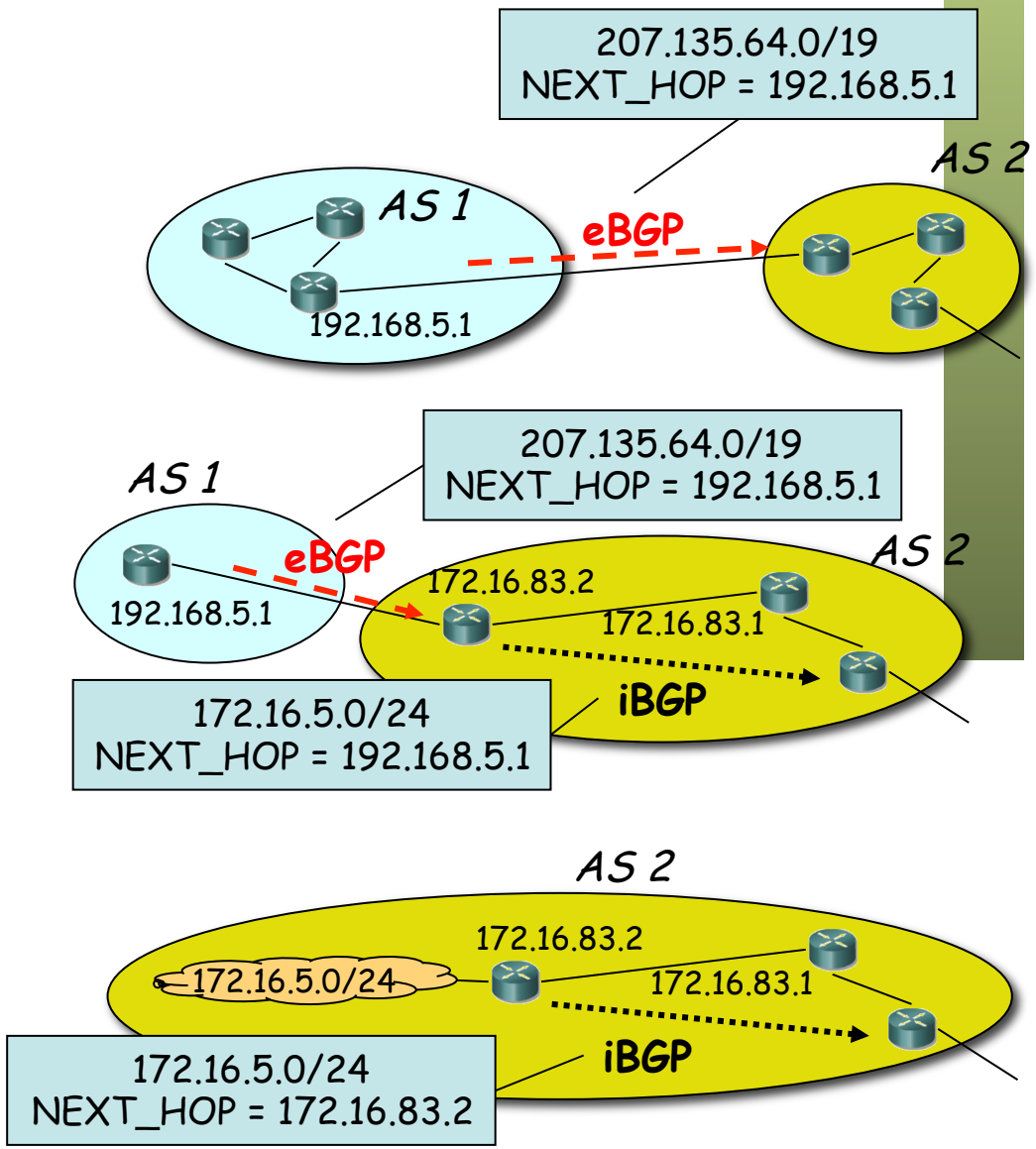
- *Well-known: mandatory* (en update) o *discretionary*
- *Optional: transitive* o *nontransitive*

ORIGIN (well-known mandatory)

- IGP, EGP o Incompleto (rutas estáticas)

NEXT_HOP (well-known mandatory)

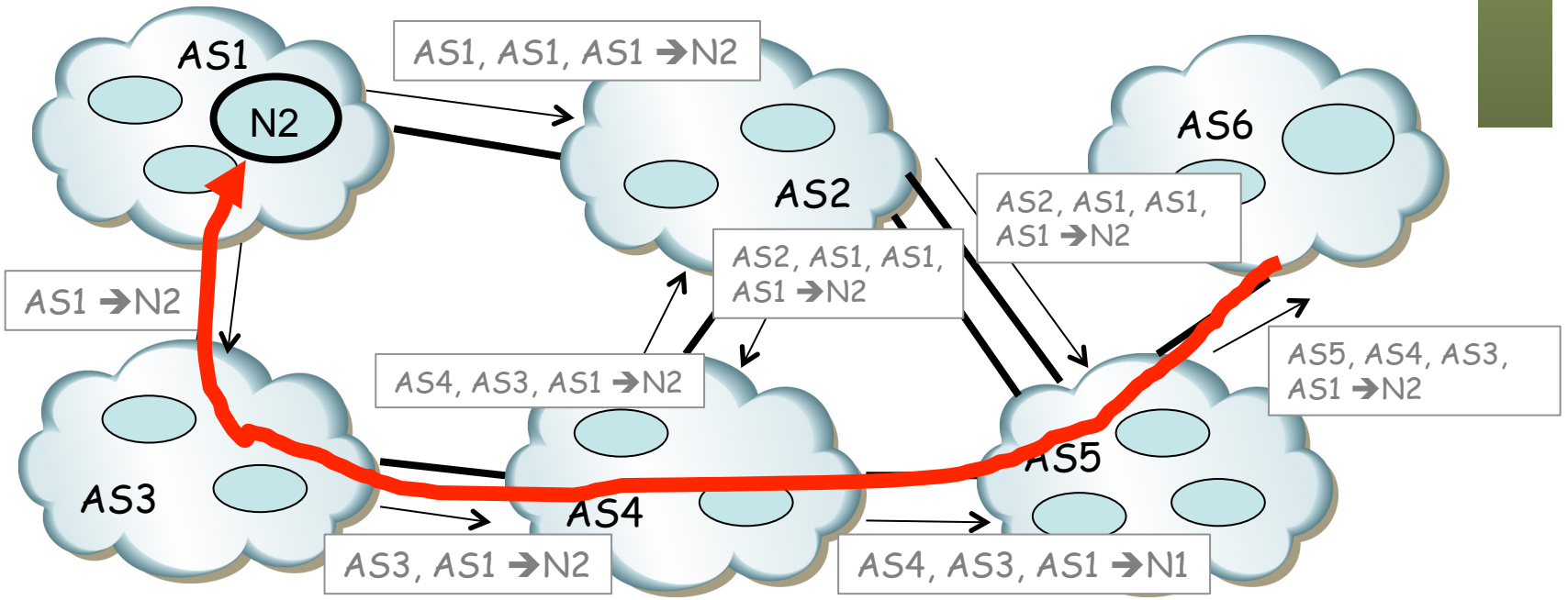
- Si son *External Peers* es la IP del interfaz del router anunciante
- Si son *Internal Peers* y
  - Destino fuera del AS: IP del *peer* externo
  - Destino en el mismo AS: IP del anunciante (*recursive lookups*)



Path Attributes

AS_PATH (well-known mandatory)

- Secuencia de ASs hasta el destino
- Al mandar un *update* por eBGP se añade el ASN a la secuencia
- Si se manda por iBGP no se añade el ASN
- *AS path prepending:* añadir el ASN *más veces* para desalentar usar este camino (. . .)



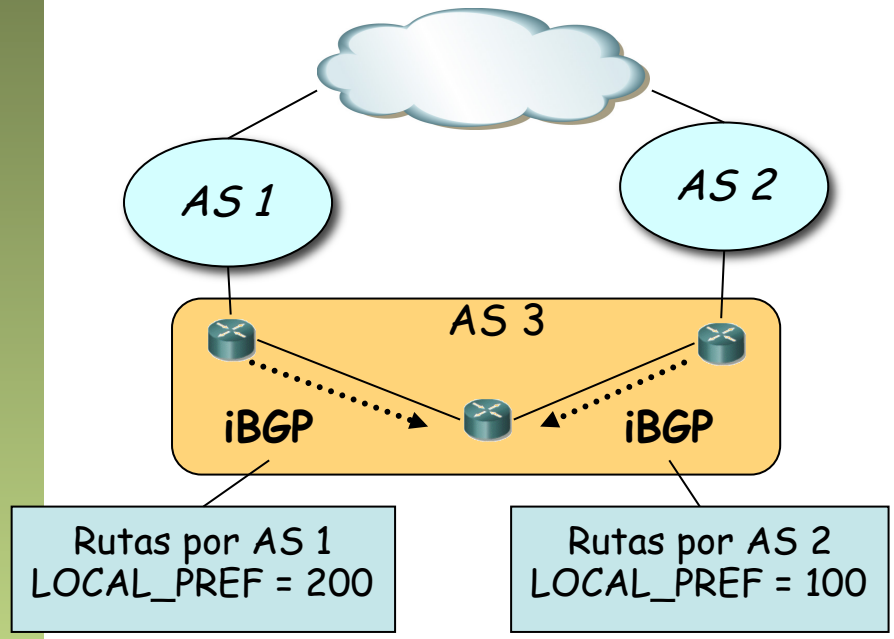
Path Attributes

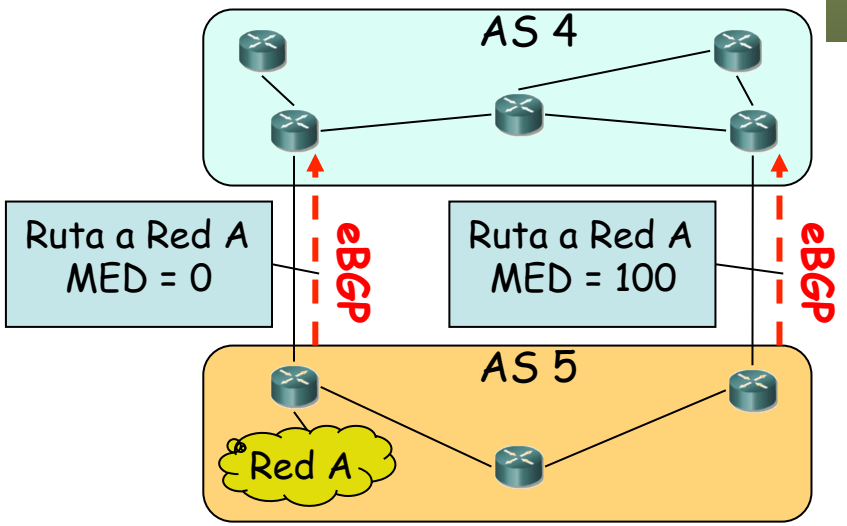
LOCAL_PREF (well-known discretionary, nontransitive)

- Solo en iBGP
- Comunica el grado de preferencia por una ruta
- La ruta de mayor valor es seleccionada

MED (optional, nontransitive)

- Multi-Exit-Discriminator
- Cuando hay múltiples links a un AS
- Anuncia el *ingress point* preferido
- Es una métrica y se selecciona el de menor MED
- No se propaga a más ASs (debe borrarlo al pasar la ruta a otro AS)





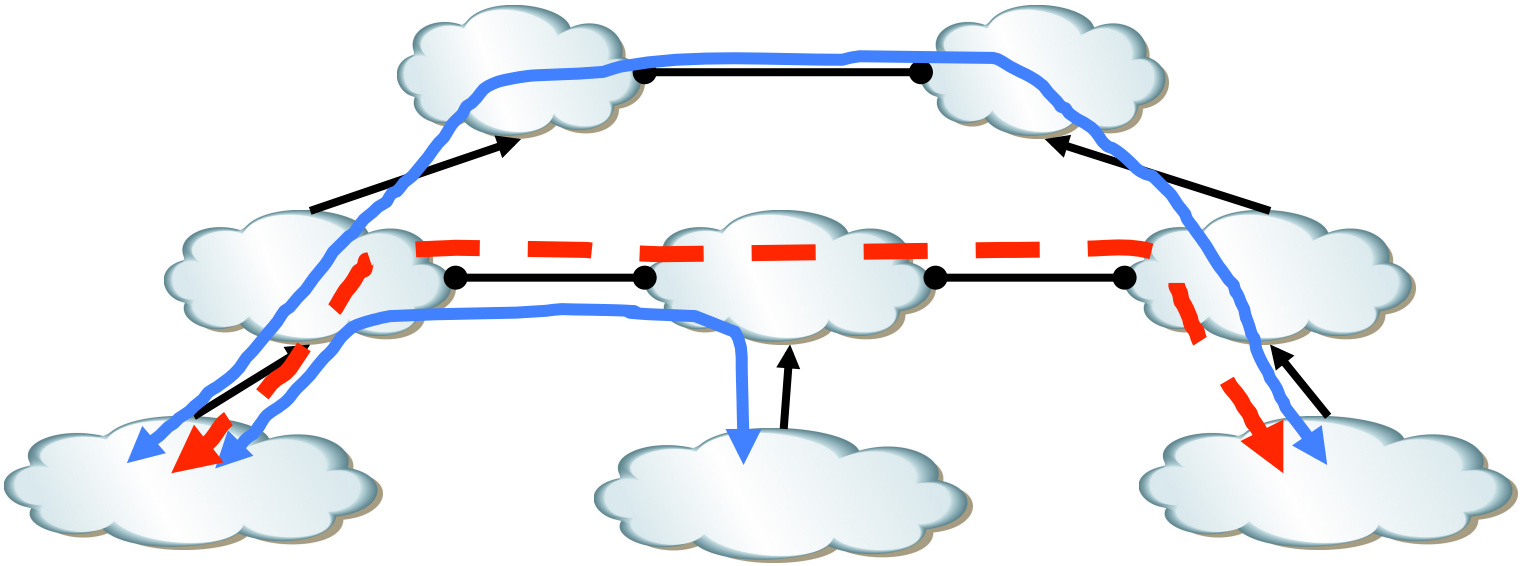
Un criterio de selección

1. Ruta con el mayor **LOCAL_PREF**
2. Si iguales, la ruta de **AS_PATH** más corto
3. Si iguales, la ruta de origen menor (**ORIGIN** IGP < EGP < Incomplete)
4. Si iguales y van al mismo AS, la de menor **MED**
5. Si igual, la de menor **métrica** del IGP hasta el NEXT_HOP
6. Si iguales y van al mismo AS, se puede instalar todas las rutas o escoger la de menor identificador de router

BGP e Internet

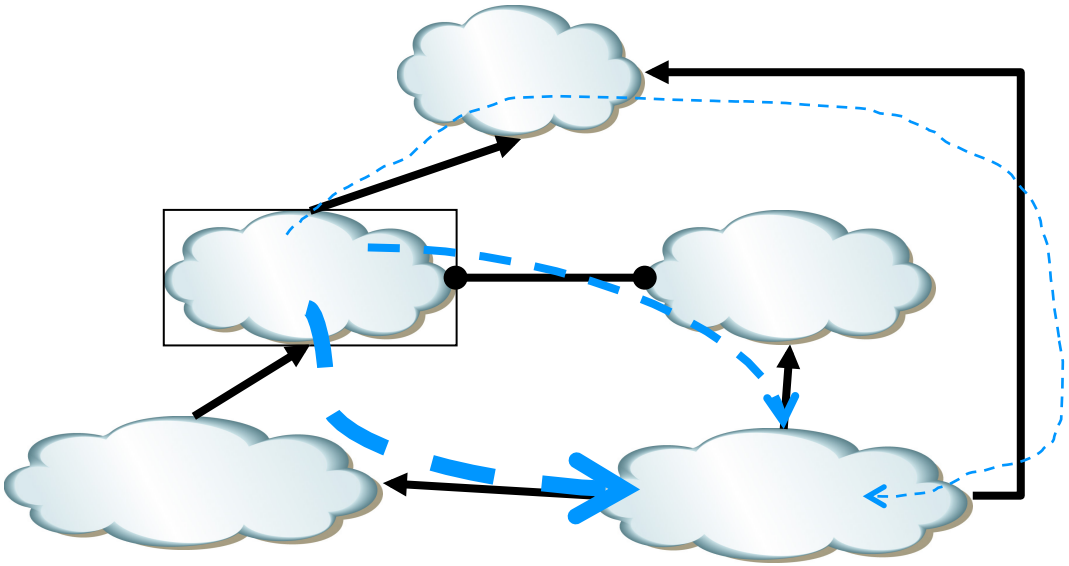
Jerarquía y economía

- En la Internet tenemos enlaces
  - Cliente-Proveedor (de pago)
  - Entre iguales (normalmente no se pagan)
- Por un enlace entre pares no se hace tránsito (...)
- Preferencia habitual:
  - (...)



Jerarquía y economía

- En la Internet tenemos enlaces
  - Cliente-Proveedor (de pago)
  - Entre iguales (normalmente no se pagan)
- Por un enlace entre pares no se hace tránsito (...)
- Preferencia habitual:
  1. Por cliente
  2. Por *peer*
  3. Por proveedor



Políticas

- Anunciar una ruta implica que se está dispuesto a encaminar el tráfico a ese destino

Los administradores pueden implementar diferentes políticas:

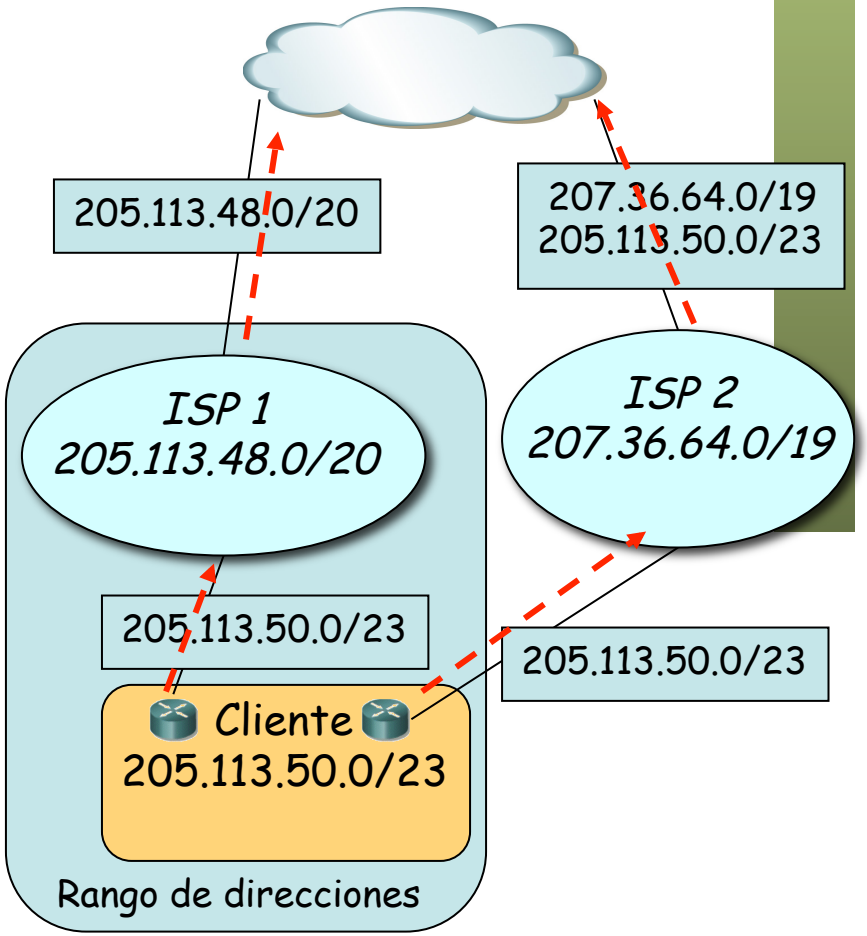
- No anunciar un destino a un vecino
- No usar caminos que pasen por cierto AS
- Ignorar el MED y usar *shortest-paths* (*hot potato routing*)
- Añadir varias veces su ASN
- Etc.

Problemas

- Hay políticas que no convergen
- Hay políticas que pueden converger dependiendo del orden de los mensajes
- Hay políticas que convergen pero dejan de hacerlo si un enlace se cae
- Dadas las políticas y la topología, decidir si convergerá es NP-completo

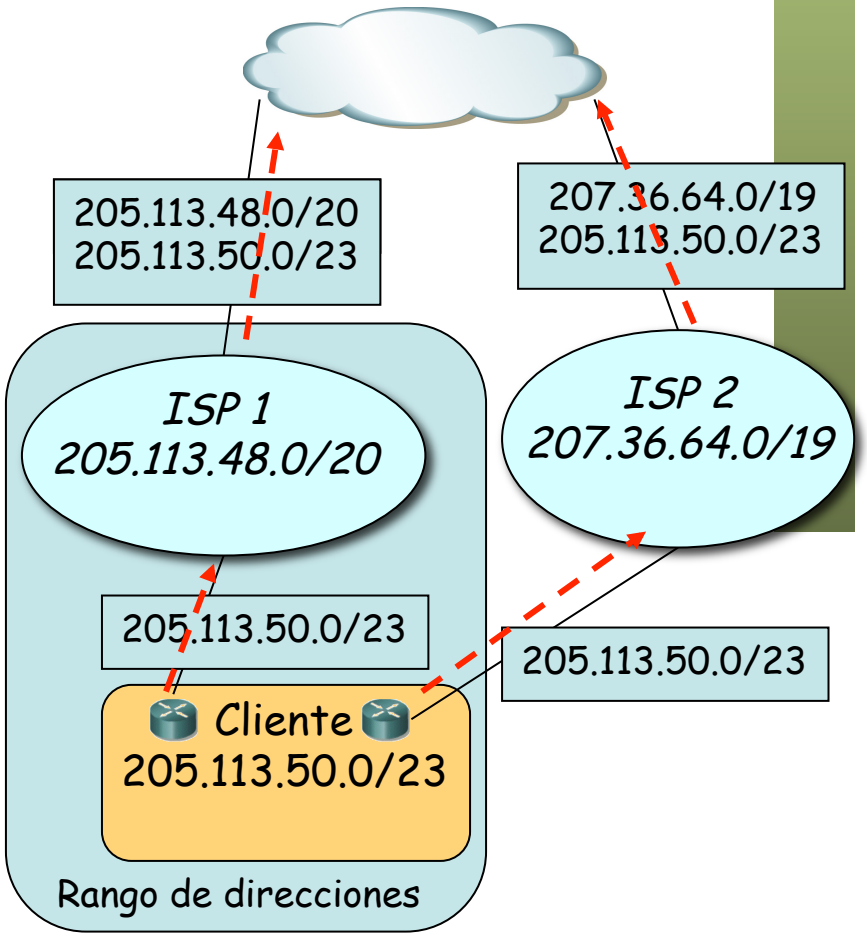
Multihoming

- Para ofrecer redundancia
- El rango de direcciones pertenece al ISP 1
- Habrá que anunciarlo también al ISP 2
- Ahora la ruta por ISP 2 es más específica
- *Address leaking:* ISP 1 debería anunciar también la ruta específica (…)



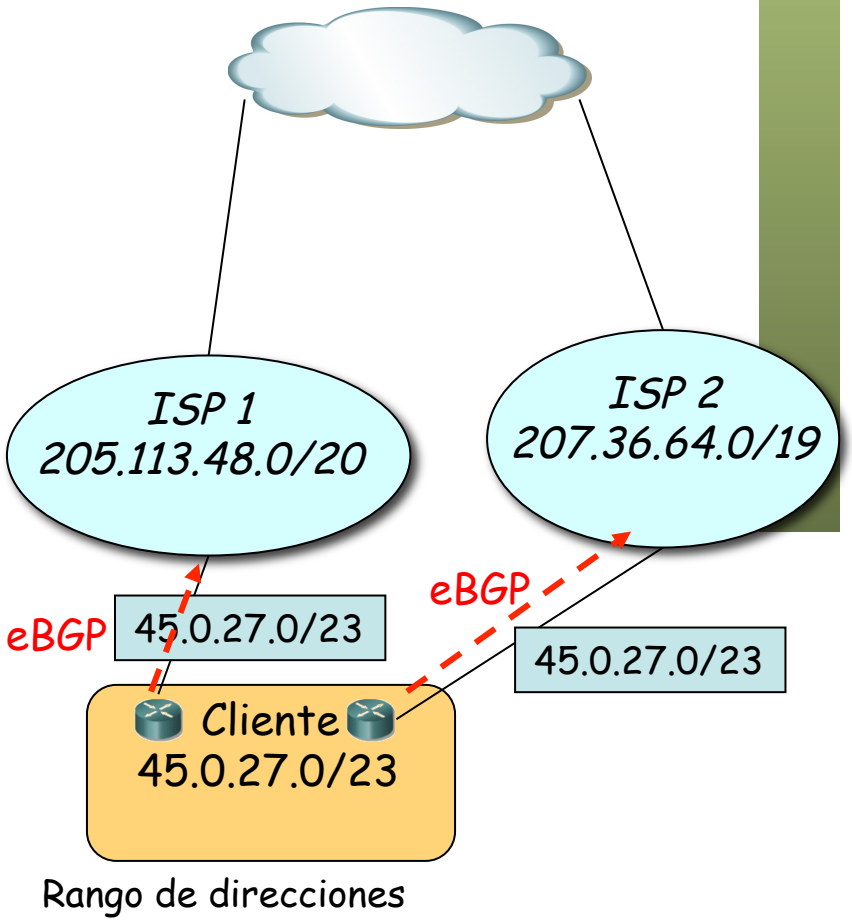
Multihoming

- Para ofrecer redundancia
- El rango de direcciones pertenece al ISP 1
- Habrá que anunciarlo también al ISP 2
- Ahora la ruta por ISP 2 es más específica
- *Address leaking:* ISP 1 debería anunciar también la ruta específica
- (…)



Multihoming

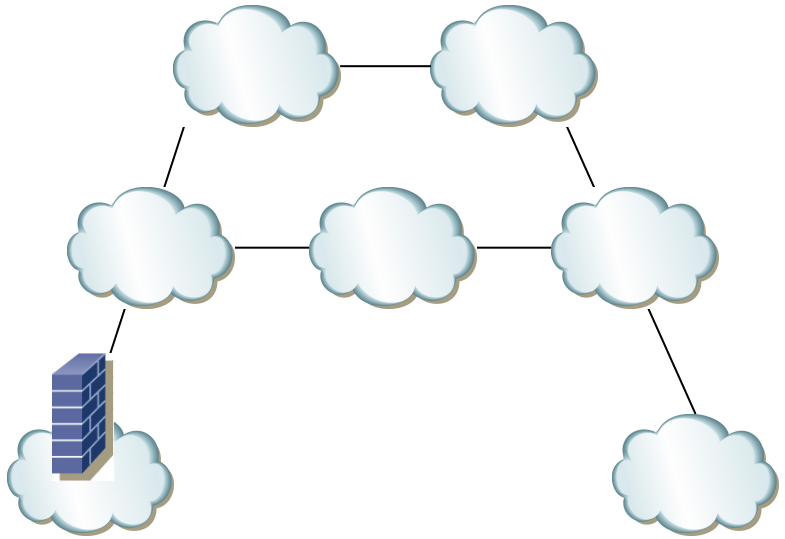
- Para ofrecer redundancia
- El rango de direcciones pertenece al ISP 1
- Habrá que anunciarlo también al ISP 2
- Ahora la ruta por ISP 2 es más específica
- *Address leaking:* ISP 1 debería anunciar también la ruta específica
- Más habitual tener un espacio de direcciones propio
- Ser un AS y correr BGP



Precauciones

Martians

- Algunos prefijos no se deben anunciar ni enrutar paquetes de ellos
- Ruta por defecto (0.0.0.0/0)
- Direccionamiento privado
  - 10.0.0.0/8
  - 172.16.0.0/12
  - 192.168.0.0/16
- *Link-local* (169.254.0.0/16)
- TEST_NET (192.0.2.0/24, etc.)
- Clases D y E (224.0.0.0/3)
- Reservados para IANA



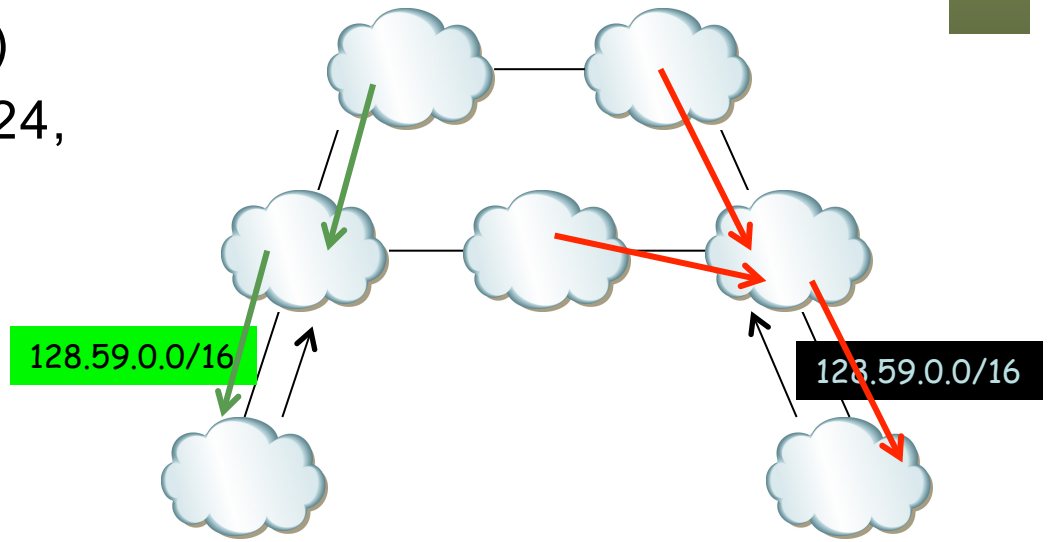
Precauciones

Martians

- Algunos prefijos no se deben anunciar ni enrutar paquetes de ellos
- Ruta por defecto (0.0.0.0/0)
- Direccionamiento privado
  - 10.0.0.0/8
  - 172.16.0.0/12
  - 192.168.0.0/16
- *Link-local* (169.254.0.0/16)
- TEST_NET (192.0.2.0/24, etc.)
- Clases D y E (224.0.0.0/3)
- Reservados para IANA

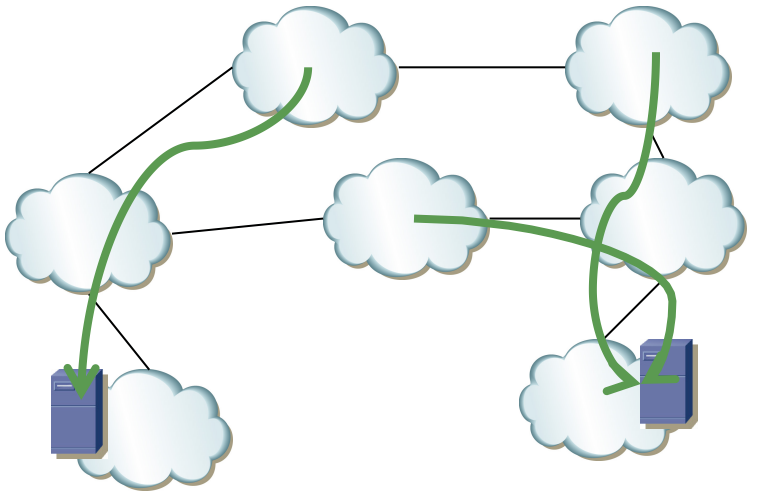
Black holes

- Si un AS anuncia un prefijo al que no está conectado
- El real puede dejar de ser accesible desde ciertas redes
- O puede hacer pasar tráfico por él



Anycast

- Servidores con misma dirección IP (contenido replicado o no)
- Todos en la misma red física o en diferentes
- Anuncios por ejemplo por diferentes proveedores
- Clientes acceden a servidor según proximidad
- Permite distribución de contenidos
- También se puede hacer en el IGP
- Ejemplo: F-root name server



Otras características

- Agregación de rutas
  - Gracias a CIDR
  - Combinar prefijos de dos o más ASs y anunciar el combinado
- *Route Reflectors*
  - Mejorar escalabilidad de iBGP (que crea un *full-mesh*)
  - Un router reflector actúa como un concentrador
- *Confederations*
  - Mejora escalabilidad de iBGP
  - Dividir AS en varios de forma que entre ellos sea eBGP
  - La confederación tiene un ASN y cada sub-AS puede tenerlo o usar uno privado
- *Route Flap Dampening*
  - Para evitar rápidas oscilaciones en una ruta
  - Aumenta el tiempo de convergencia

Ejemplos

Ejemplos

- Probad un *Looking Glass*, por ejemplo: http://www.rediris.es/red/lg
- Ejemplo: Desde CIEMAT, AS PATH a 169.229.216.200 :

```
Espere, por favor...
Please wait...
169.229.0.0/16     *[BGP/170] 3d 11:11:11, MED 320, localpref 161, from 130.206.206.250
                      AS path: 20965 11537 2153 25 I, validation-state: unverified
                    [BGP/170] 3d 11:11:11, MED 220, localpref 150
                      AS path: 20965 11537 2153 25 I, validation-state: unverified
                    [BGP/170] 3w6d 12:23:20, MED 23041, localpref 110
                      AS path: 174 3356 3356 3356 2152 2152 2152 25 I, validation-state: unverified
                    [BGP/170] 3w6d 12:23:20, MED 23041, localpref 110
                      AS path: 174 3356 3356 3356 2152 2152 2152 25 I, validation-state: unverified
{master}
```

Ejemplos

- Otro *Looking Glass*: http://lg.cern.ch
- Ejemplo: Desde LE1, show ip bgp summary:

```
BGP4 Summary
 Router ID: 194.12.136.65   Local AS Number: 513
 Confederation Identifier: not configured
 Confederation Peers:
 Maximum Number of IP ECMP Paths Supported for Load Sharing: 4
 Number of Neighbors Configured: 17, UP: 17
 Number of Routes Installed: 343, Uses 29498 bytes
 Number of Routes Advertising to All Neighbors: 954 (610 entries), Uses 29280 bytes
 Number of Attribute Entries Installed: 213, Uses 19170 bytes
 Neighbor Address  AS#         State   Time          Rt:Accepted Filtered Sent     ToSend
 172.24.46.2       513         ESTAB   44d23h 9m     33          0        152      0
 192.16.166.2      24167       ESTAB   4d 4h22m      5           0        33       0
 192.16.166.6      3152        ESTAB   22d11h12m     5           0        40       0
 192.16.166.34     34878       ESTAB   35d21h 9m     11          0        35       0
 192.16.166.50     39590       ESTAB   1d11h48m      8           4        37       0
 192.16.166.58     43115       ESTAB   17d22h53m     1           0        39       0
 192.16.166.66     43475       ESTAB   71d 3h19m     1           0        39       0
 192.16.166.82     36391       ESTAB   13h57m45s     2           0        38       0
 192.16.166.90     43          ESTAB   22d11h11m     3           0        37       0
 192.16.166.154    137         ESTAB   17d19h19m     1           0        39       0
 192.16.166.166    17579       ESTAB   16d20h58m     1           0        39       0
 192.16.166.170    59624       ESTAB   6d20h52m      2           0        40       0
 192.16.166.174    2875        ESTAB   6d20h55m      1           0        39       0
 192.16.166.182    59624       ESTAB   2d11h19m      2           0        40       0
 192.65.183.9      20641       ESTAB   71d 3h21m     150         0        3        0
 194.12.136.1      513         ESTAB   71d 3h28m     100         0        152      0
 194.12.136.29     513         ESTAB   71d 3h28m     14          0        152      0
```

Ejemplos

- Otro *Looking Glass*: http://lg.cern.ch
- Ejemplo: Desde EE2, show ip bgp 130.206.162.158:

```
Number of BGP Routes matching display condition : 6
Status codes: s suppressed, d damped, h history, * valid, > best, i internal
Origin codes: i - IGP, e - EGP, ? - incomplete
    Network            Next Hop        MED    LocPrf     Weight Path
*>  130.206.0.0/16     192.65.184.173  20     65000      100    559 20965 766 i
*   130.206.0.0/16     192.65.184.69   20     65000      100    559 20965 766 i
*i  130.206.0.0/16     192.65.184.1    20     65000      100    559 20965 766 i
*   130.206.0.0/16     195.141.200.17  10     64000      100    6730 6730 174 766 766 766 766 766 i
*   130.206.0.0/16     192.91.246.125  35     65000      100    11537 11537 20965 766 i
*   130.206.0.0/16     192.91.246.109  30     65000      100    10764 11537 20965 766 i
      Last update to IP routing table: 17d13h50m2s, 2 path(s) installed:
      Route is advertised to 4 peers:
                 192.65.184.158(1297)                                        192.65.184.3(513)
192.65.184.24(513)
       192.65.184.1(513)
```

Ejemplos

- http://www.routeviews.org
- Ejemplo: direcciones de equipos con sesión BGP y que aceptan telnet

```
$ telnet route-views.routeviews.org
Trying 128.223.51.103...
Connected to route-views.routeviews.org.
Escape character is '^]'.
******************************************************************
                    Oregon Exchange BGP Route Viewer
          route-views.oregon-ix.net / route-views.routeviews.org
BLA BLA BLA…

route-views>show ip bgp summary
BGP router identifier 128.223.51.103, local AS number 6447
BGP table version is 23102224, main routing table version 23102224
541514 network entries using 134295472 bytes of memory
17374508 path entries using 2084940960 bytes of memory
2820317/95311 BGP path/bestpath attribute entries using 699438616 bytes of memory
2546156 BGP AS-PATH entries using 124593136 bytes of memory
1 BGP ATTR_SET entries using 40 bytes of memory
76674 BGP community entries using 8194260 bytes of memory
356 BGP extended community entries using 12838 bytes of memory
…
BGP using 3051475282 total bytes of memory
```

Ejemplos

Resumen

- Es el EGP empleado en Internet
- Path Vector
- I-BGP y E-BGP
- La influencia de las políticas es crítica