Agenda (5)

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

¿Cómo se generan anuncios? Prefijos de AS por eBGP

- **Redistribución de un IGP (**origin incomplete**):**

  NO recomendado!!! (impacto en anuncios externos por eBGP)

(sí se suele utilizar redistribuir estáticas, hay otros escenarios)

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

- **Generación local (**origin IGP**):**

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 (desacomplar publicación a estado IGP)

Sumarizaciones - Agregado de prefijos de otros AS

- 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 agrega AS-SET o se elimina
- **Cisco:**

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

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

Agenda (6)

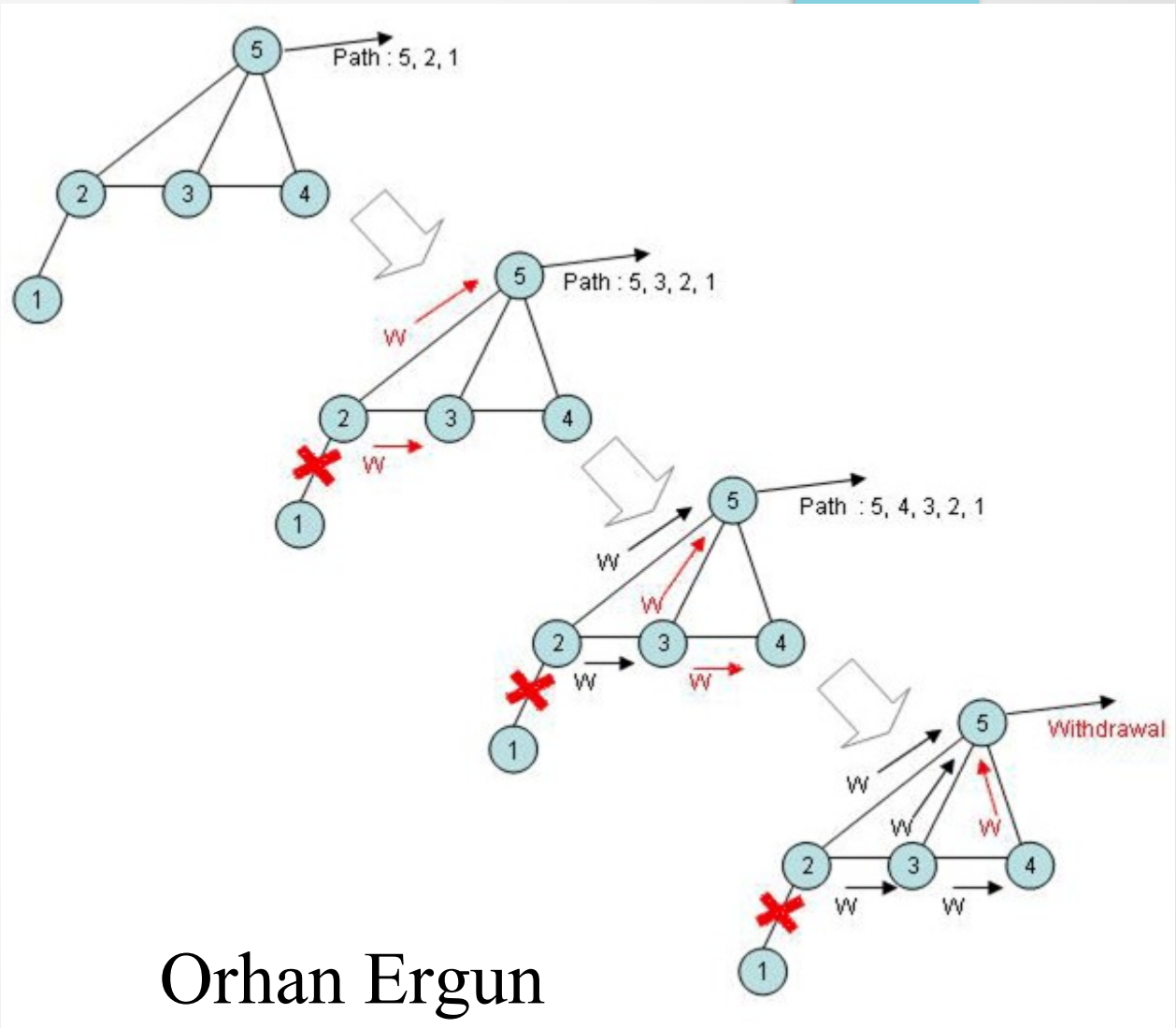
- Conceptos Fundamentales de BGP
- Análisis del protocolo (BGP-4)
- Atributos de BGP y políticas de control
- IBGP mesh y Alternativas
- Sumarización y anuncios (CIDR)
- Damping y problemas de convergencia
- Extensiones Multiprotocolo
- Seguridad de BGP
- Salidas reales y datos de actualidad
- Ejemplo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

Inestabilidad de rutas (1)

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

Inestabilidad de rutas (2)

- Las rutas inestables se traducen en la generación de gran cantidad de mensajes UPDATE de BGP
- La inestabilidad se propagará a todos los enrutadores que reciban esos anuncios
- BGP "**Path Hunting**": <u>en la Local-RIB están todos los caminos, solo se propaga el mejor.</u>



Inestabilidad de rutas (3)

- Formas de minimizar inestabilidades (inestabilidades hacia internet):
  - **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)

    Publicar redes generadas estáticamente, pero internamente se aprenden por IGP o iBGP rutas más específicas (por ejemplo publico al exterior /20 pero internamente tengo /24).

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

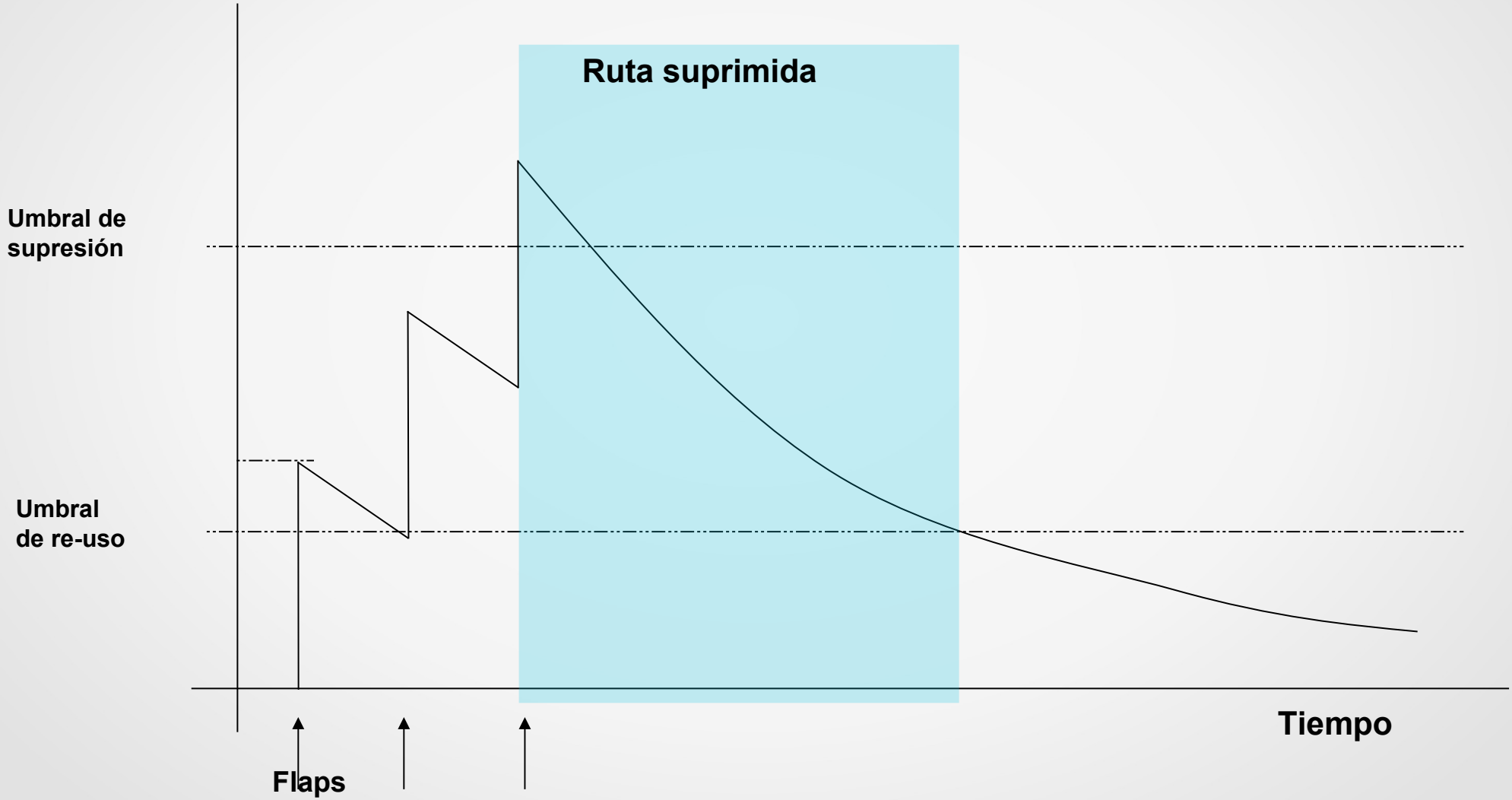
Route Flap Damping (2)

- 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 **AS**es clientes u otros **AS**es 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…
- **Penalización máxima** por encima de está no se acumula más penalización => penalización máxima de tiempo

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

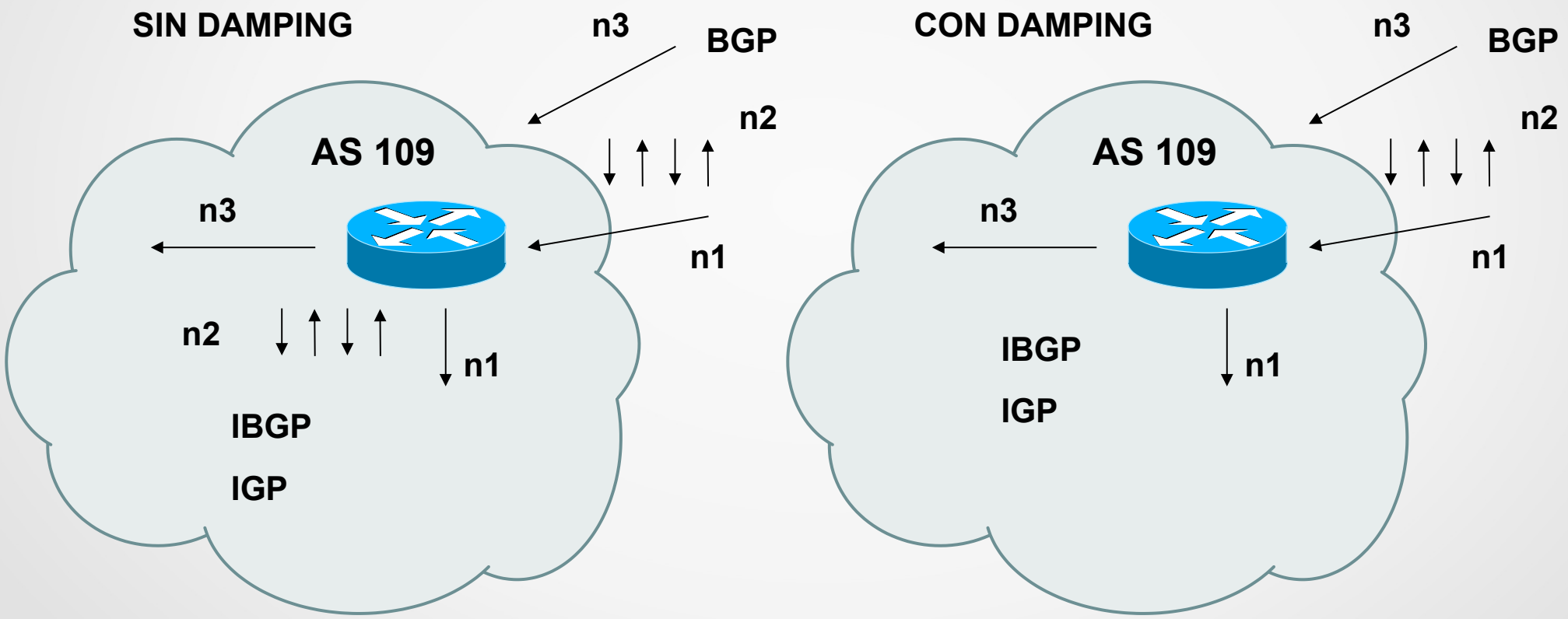
  umbral para re-usar: 750

  Tiempo medio: 15 min (el valor cae a la mitad)

  Tiempo máximo de supresión: 4 x tiempo medio (máximo de penalización), define un valor máximo de penalidad

Route Flap Damping (7)

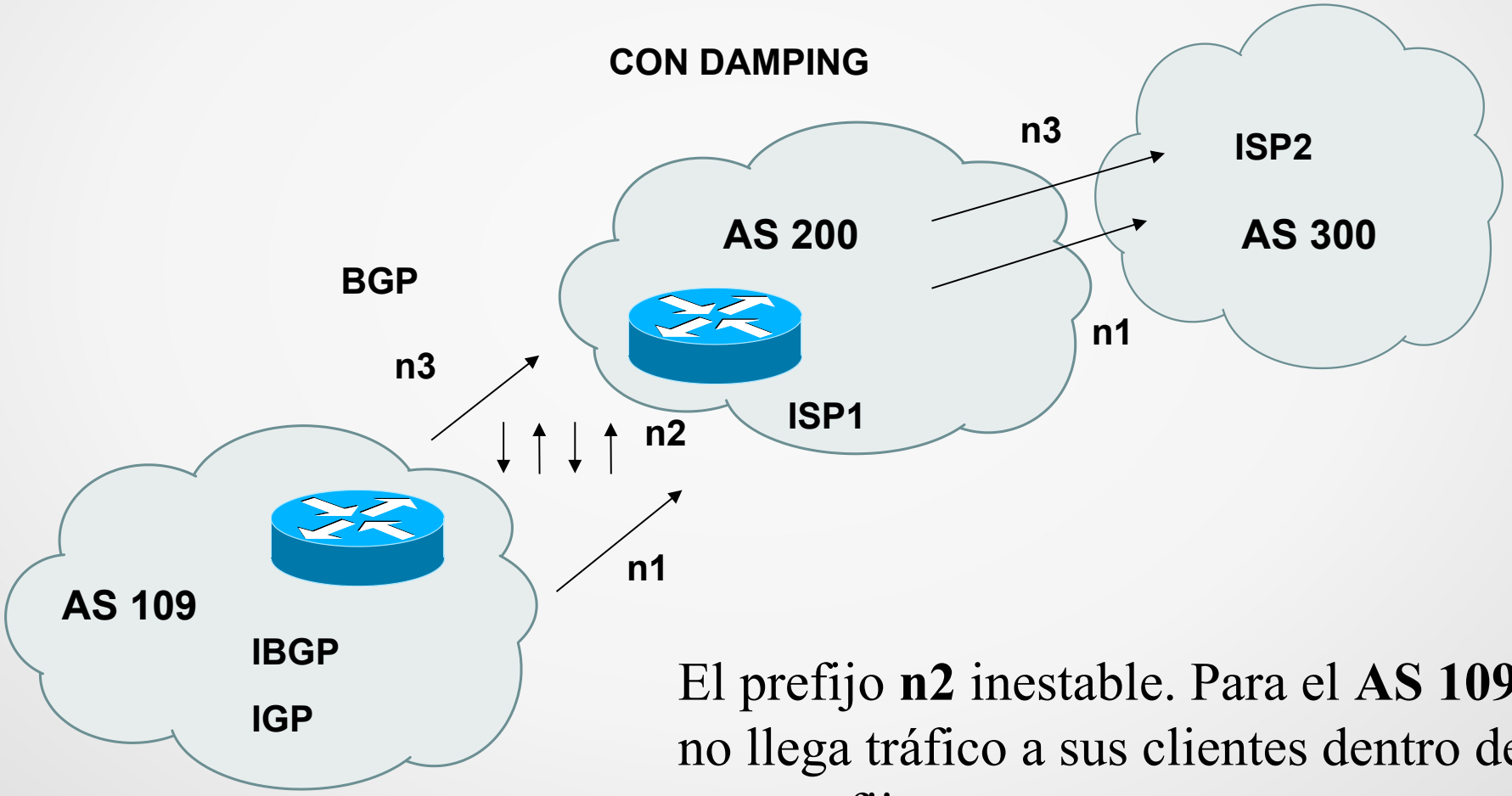
- Estabilidad dentro del AS:



El prefijo **n2** inestable.

Route Flap Damping (8)

- Inestabilidades fuera del AS:



El prefijo **n2** inestable. Para el **AS 109**, no llega tráfico a sus clientes dentro de este prefijo.

Route Flap Damping (9)

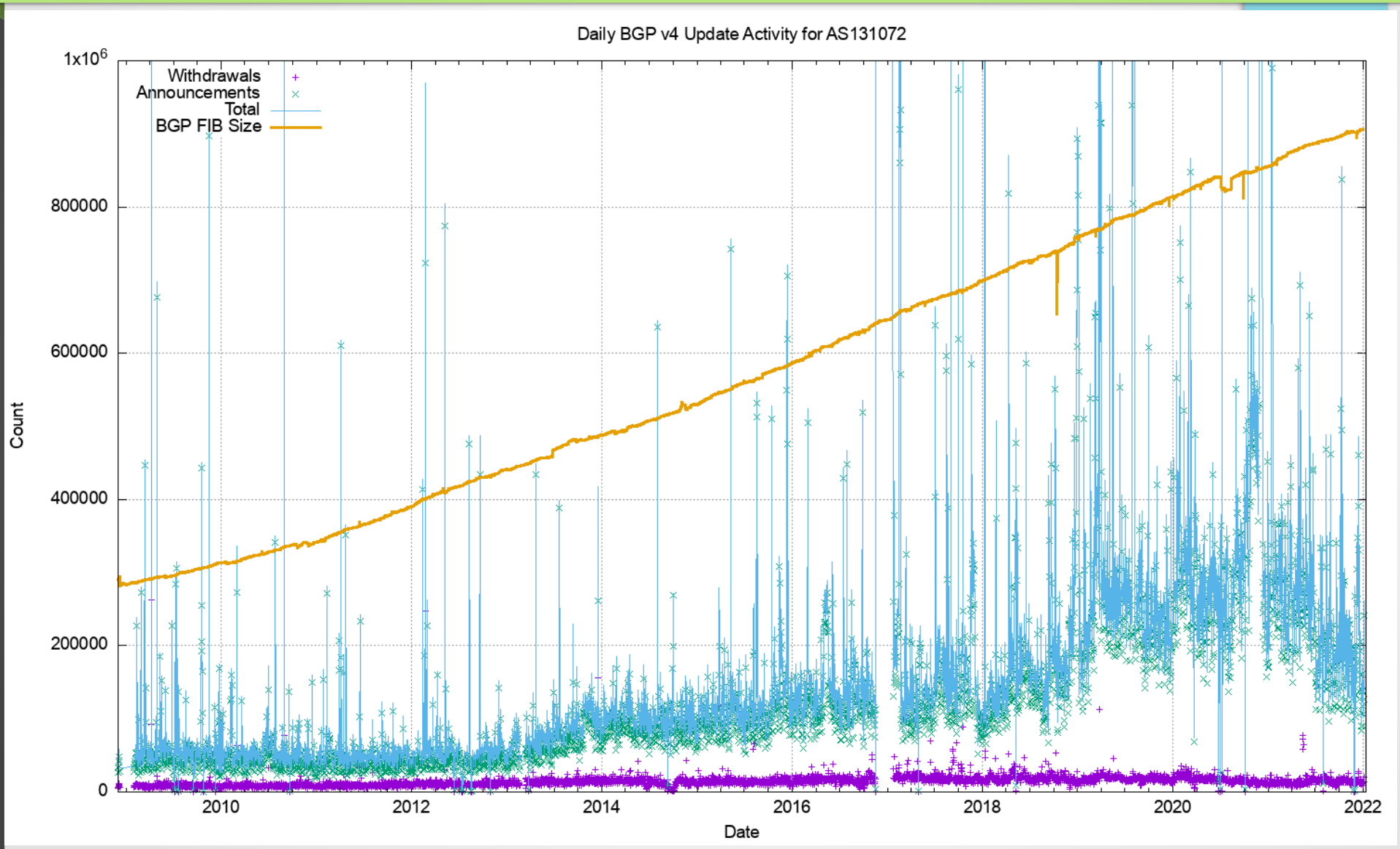
- Hay recomendaciones y estudios que indican que el Damping es más perjudicial que útil.

  En la actualidad (desde 2013) se considera que los routers tienen capacidad suficiente para hacer frente a la avalancha de mensajes.
- **Un cambio en una ruta, al propagarse, puede generar varios cambios en un AS remoto (**visto en un ejemplo previo**)**
- La desaparición de una ruta implica una búsqueda de un nueva mejor ruta/camino. El cambio de AS_PATH genera media penalidad, predisponiendo a que sea suprimido.
- Aún se sigue usando, por lo que debemos 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

Algunos problemas actuales – BGP Churn



Fuente: Labs APNIC

Agenda (7)

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

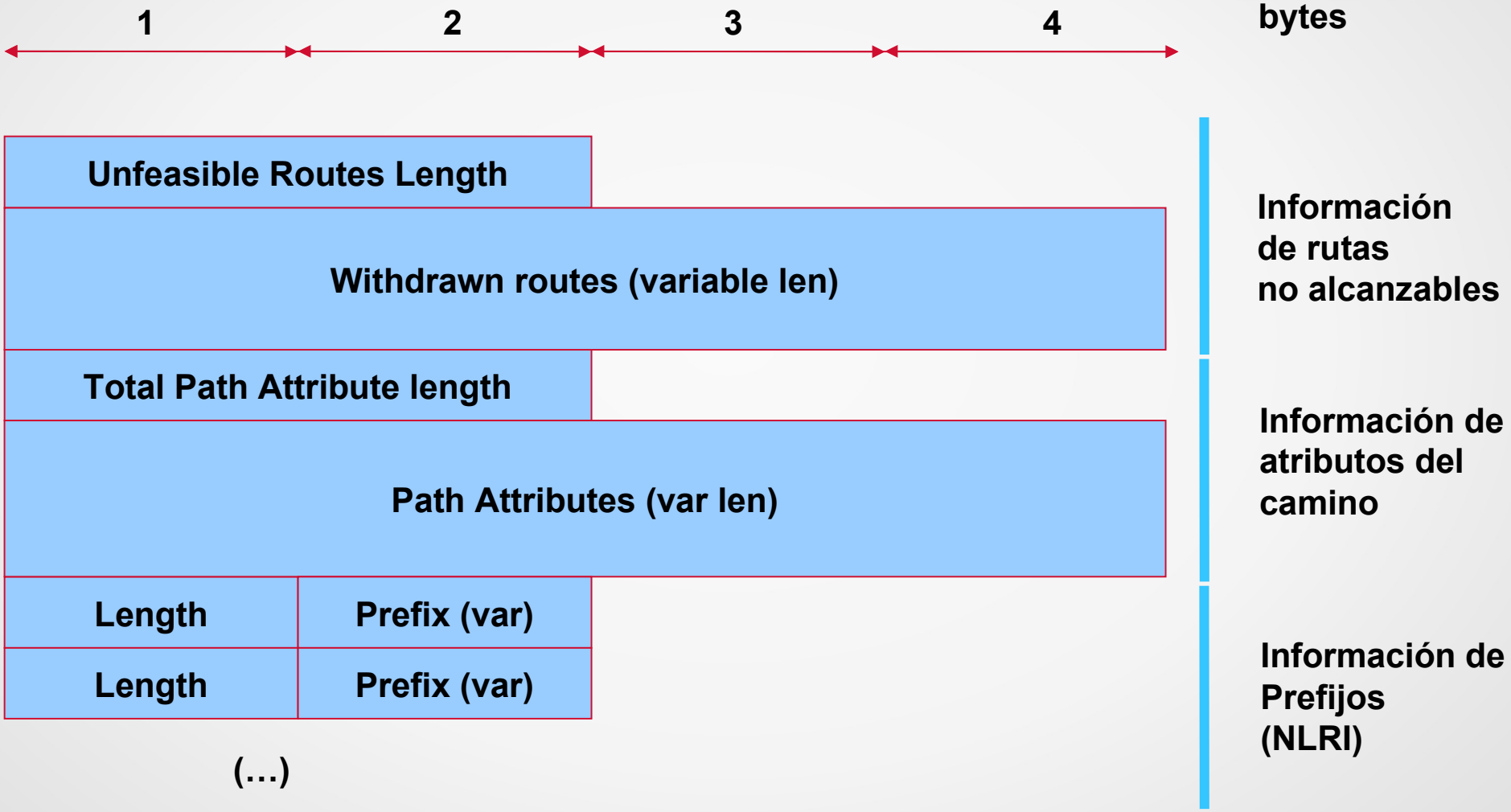
Extensiones Multiprotocolo

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

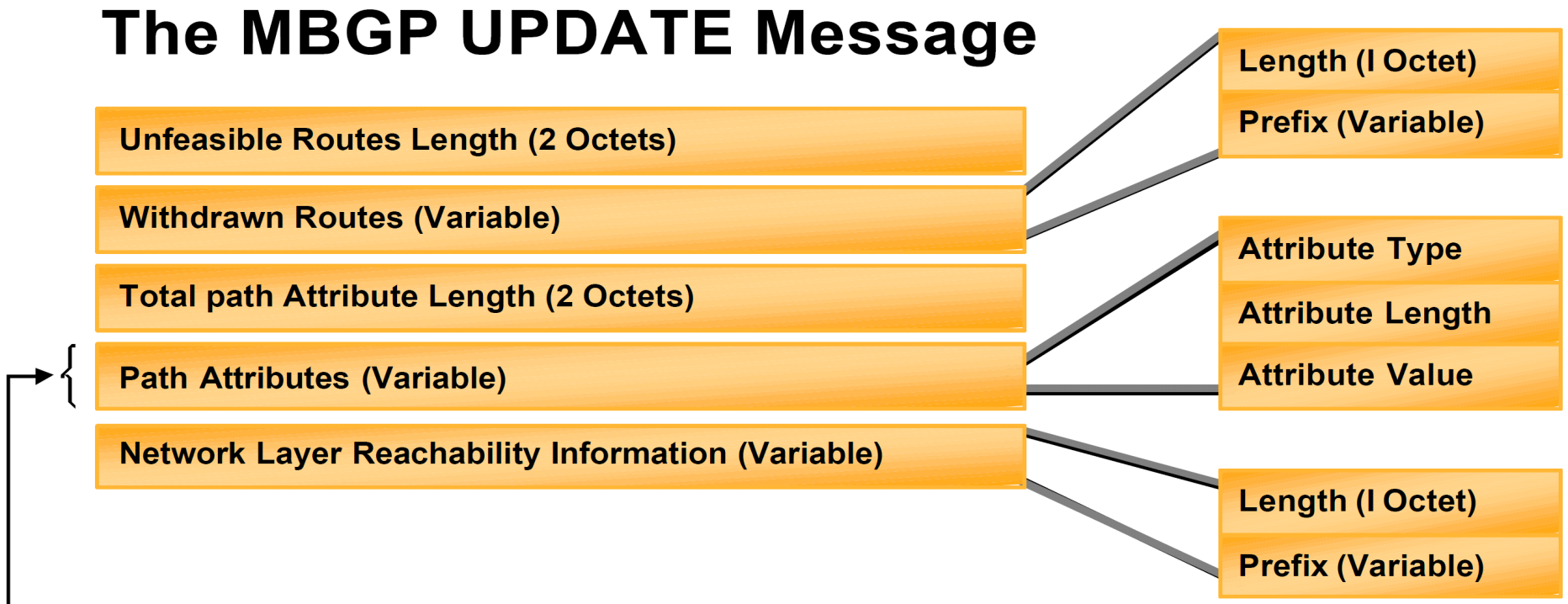
    Información de NLRI y Next Hop
  - **MP_UNREACH_NLRI**

    Reemplaza las rutas que se dejan de anunciar (withdrawn routes)
- Su utilización se negocia al inicio con el mensaje OPEN utilizando Capabilities (**MULTIPROTOCOL_EXTENSIONS 0x01**).

Mensaje UPDATE (RECORDANDO)

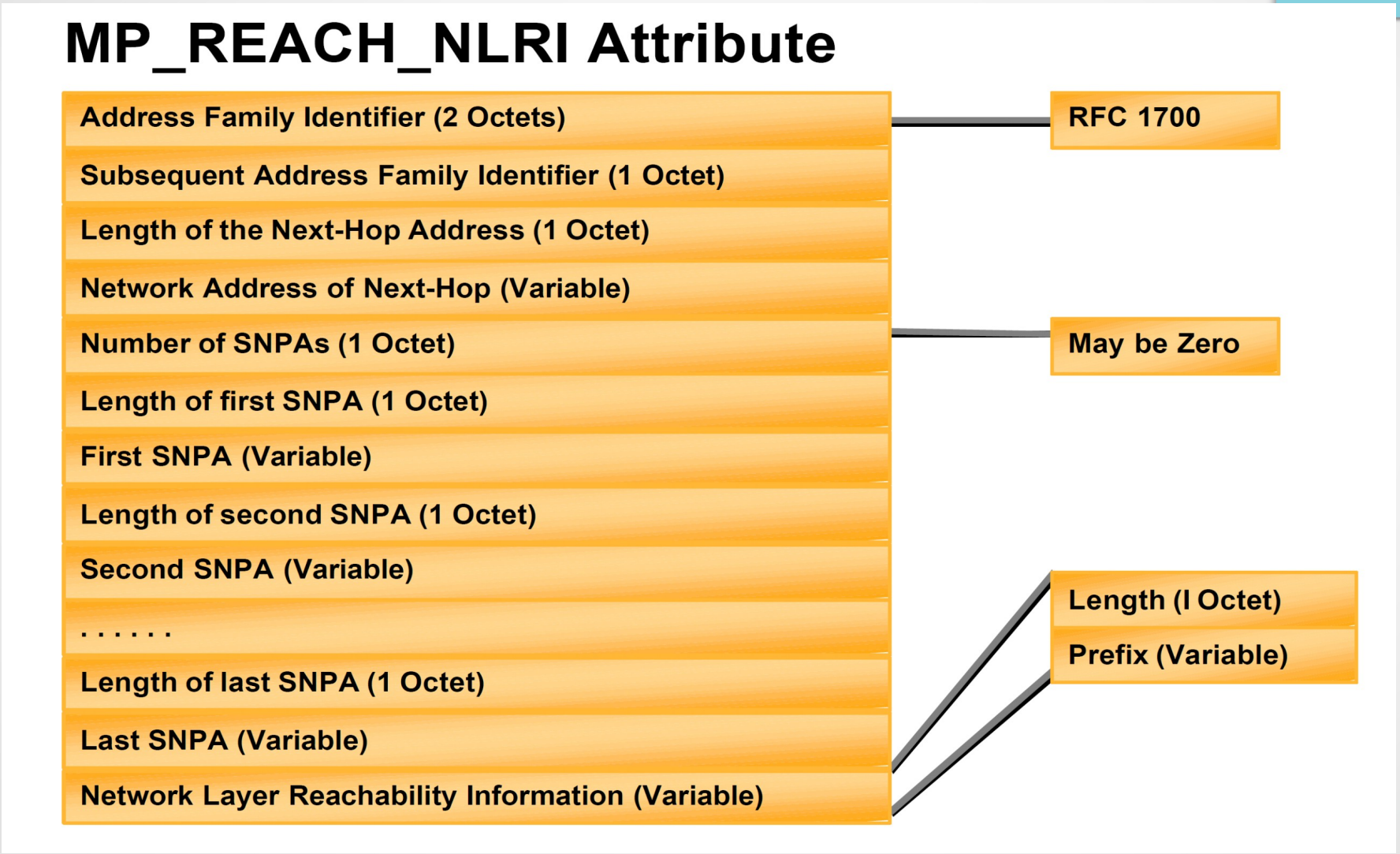


Extensiones Multiprotocolo - Update



- **New Multiprotocol Attributes added to Path Attributes:**
  - **MP_REACH_NLRI**
  - **MP_UNREACH_NLRI**

Atributo MP_REACH_NLRI



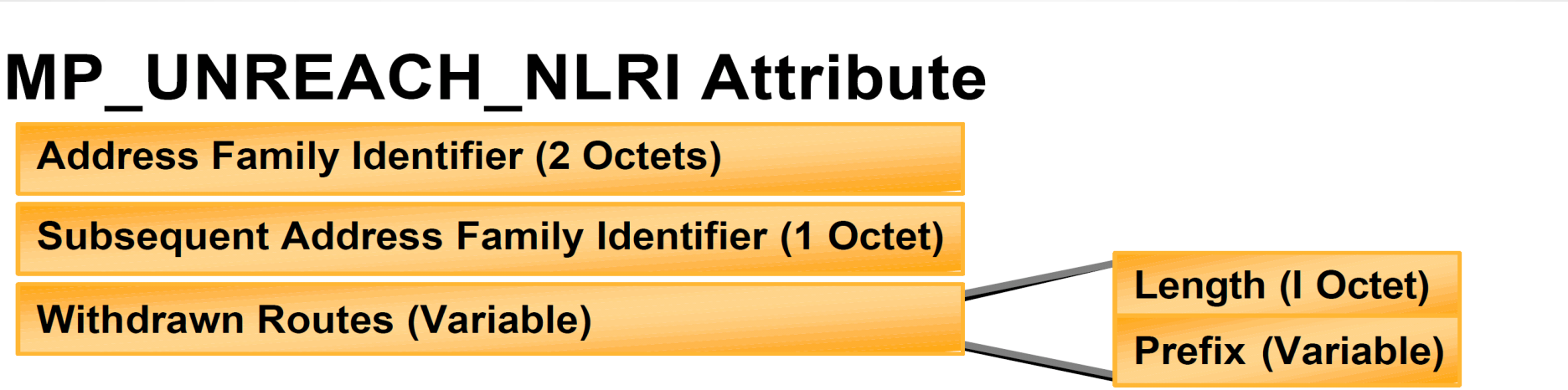
Extensiones Multiprotocolo – Address Family Id

- Address Family Information (AFI)

  Identifica el tipo de direcciones (2 bytes):

  - AFI = 1 (IPV4)
  - AFI = 2 (IPV6)
- Sub Address Family Information (SAFI)
  - Sub-AFI = 1 (unicast)
  - Sub-AFI = 2 (multicast)
  - Sub-AFI = 128 (MPLS-Label VPN Address)

Atributo MP_UNREACH_NLRI



IPv6 y BGP – Ejemplo de extensiones 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 y BGP (2)

- **MP_REACH_NLRI**

  Address family (IPv6) – AFI/SAFI = 2/1

  Next_Hop (IPv6)

  NLRI (prefijos)

- **MP_UNREACH_NLRI**

  – Prefijos que ya no son alcanzables

- Next_Hop (IPv6) se envía Link_Local y Unicast (como en IPv4). En iBGP es posible alterar el Next_Hop y no compartir un link con el peer.

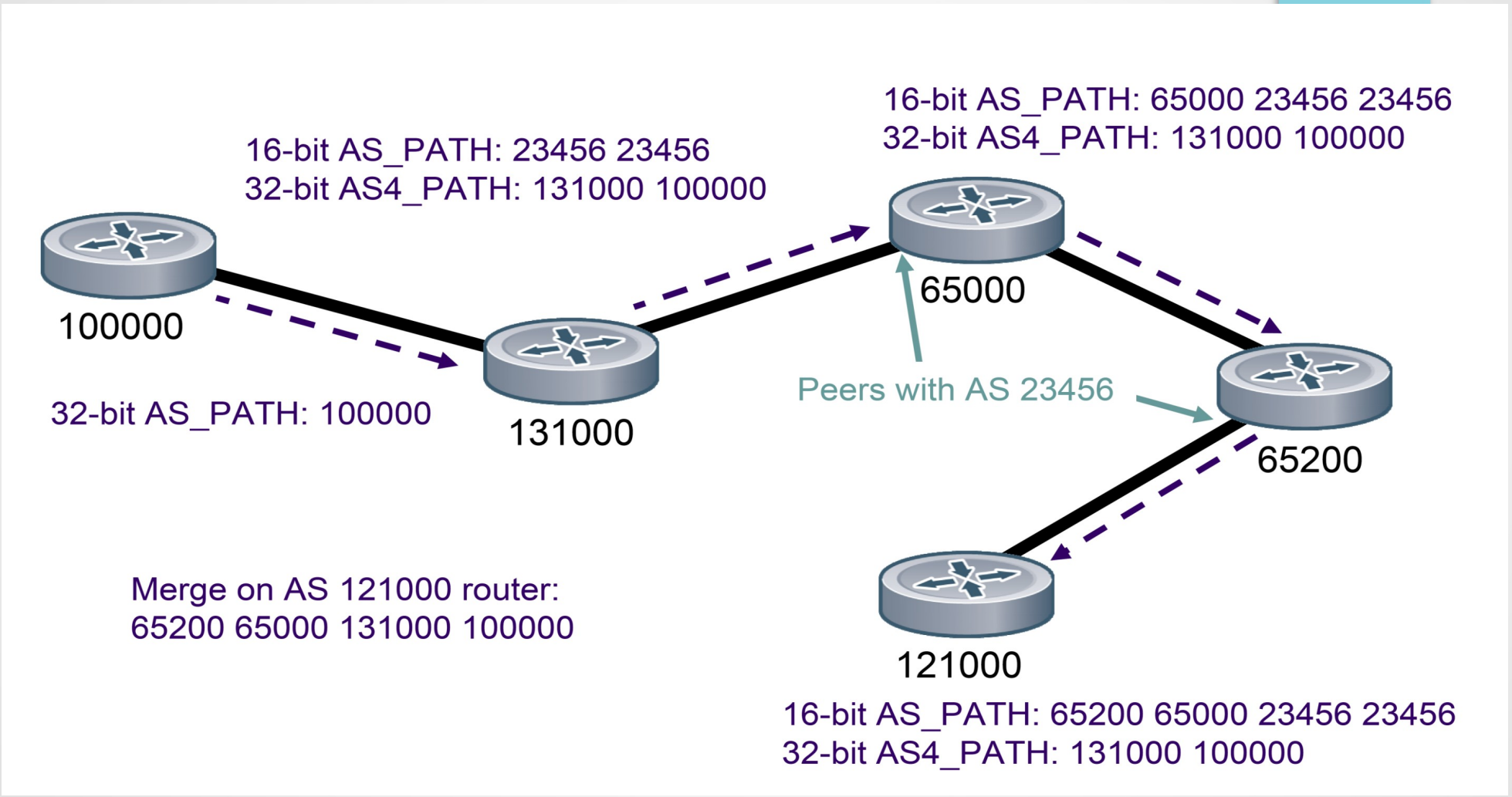
AS 32 bits – Ejemplo de capacidad de adaptación

- Se negocia al comienzo de la sesión BGP en el mensaje OPEN
- Nueva Capability (**FOUR_BYTES_ASN 0x41**)
- **¿Cuáles son los cambios?**:
- ¿My Autonomous System en el mensaje OPEN?
  - AS_TRANS = 23456
  - El "**verdadero**" número de AS se envía dentro de la **nueva Capability**.
- Si ambos soportan AS32 bits
  - Luego de establecida la sesión, se utilizan los atributos actuales **AS_PATH** y **AGGREGATOR**, pero se corrige el largo del AS a 32 bits.

AS 32 bits (2)

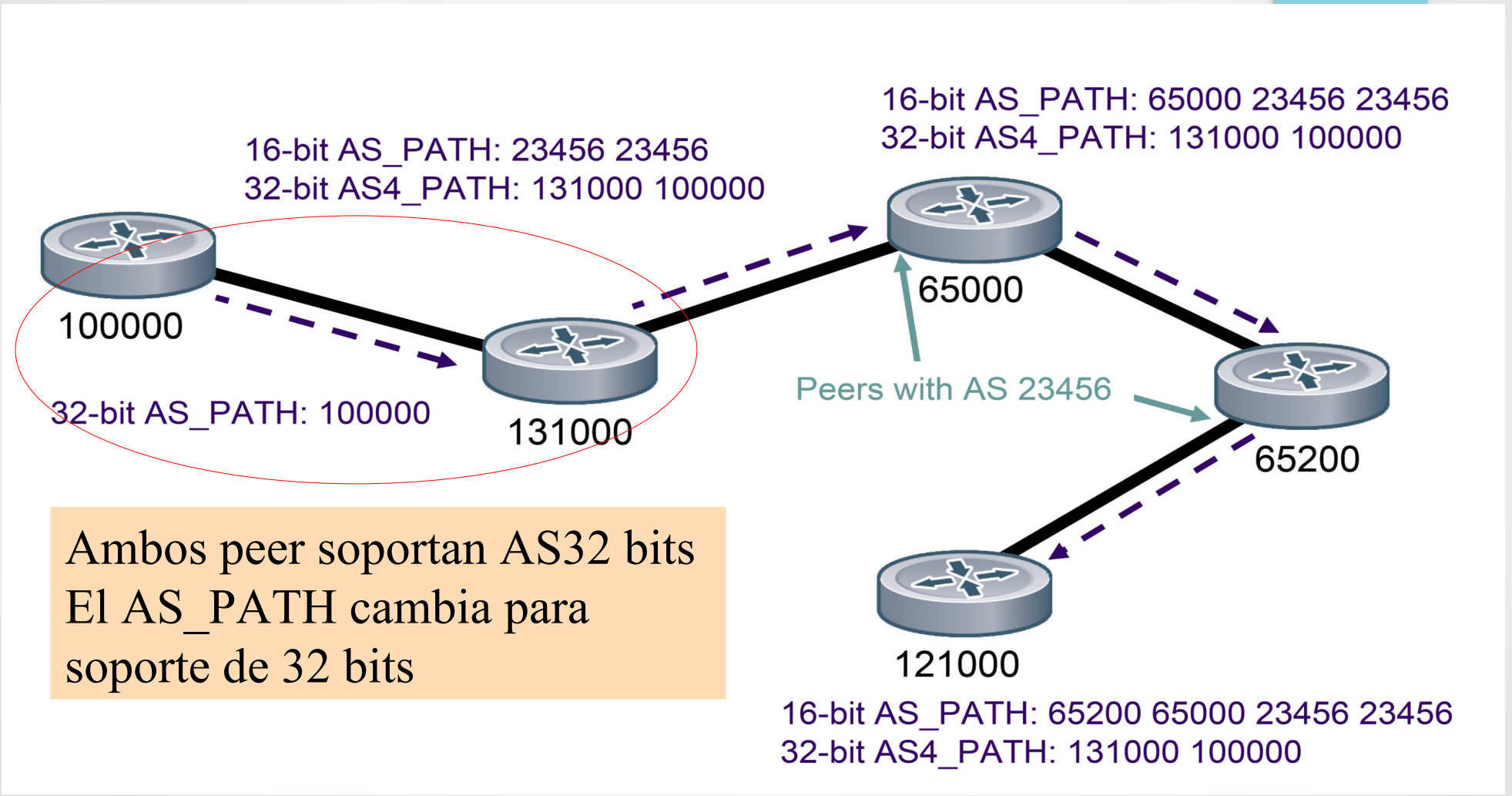
- Si el AS remoto no soporta AS 32 bits.
- Nuevos atributos (OT) **AS4_PATH** y **AS4_AGGREGATOR**
- En el AS_PATH y AGGREGATOR debe aparecer el **AS_TRANS**. Se mantiene el AS_PATH length
- En los nuevos atributos se guarda los valores con AS 32 bits.
- **¿Qué pasa con las Comunidades?**
- Se define un atributo llamado comunidad extendida (RFC 4360) de largo 8 bytes.
- Se utilizan comunidades extendidas (RFC 5668) para preservar **ASN:comunidad**

AS32 bits – Transición (3)



Hankins and Malyster, NANOG45

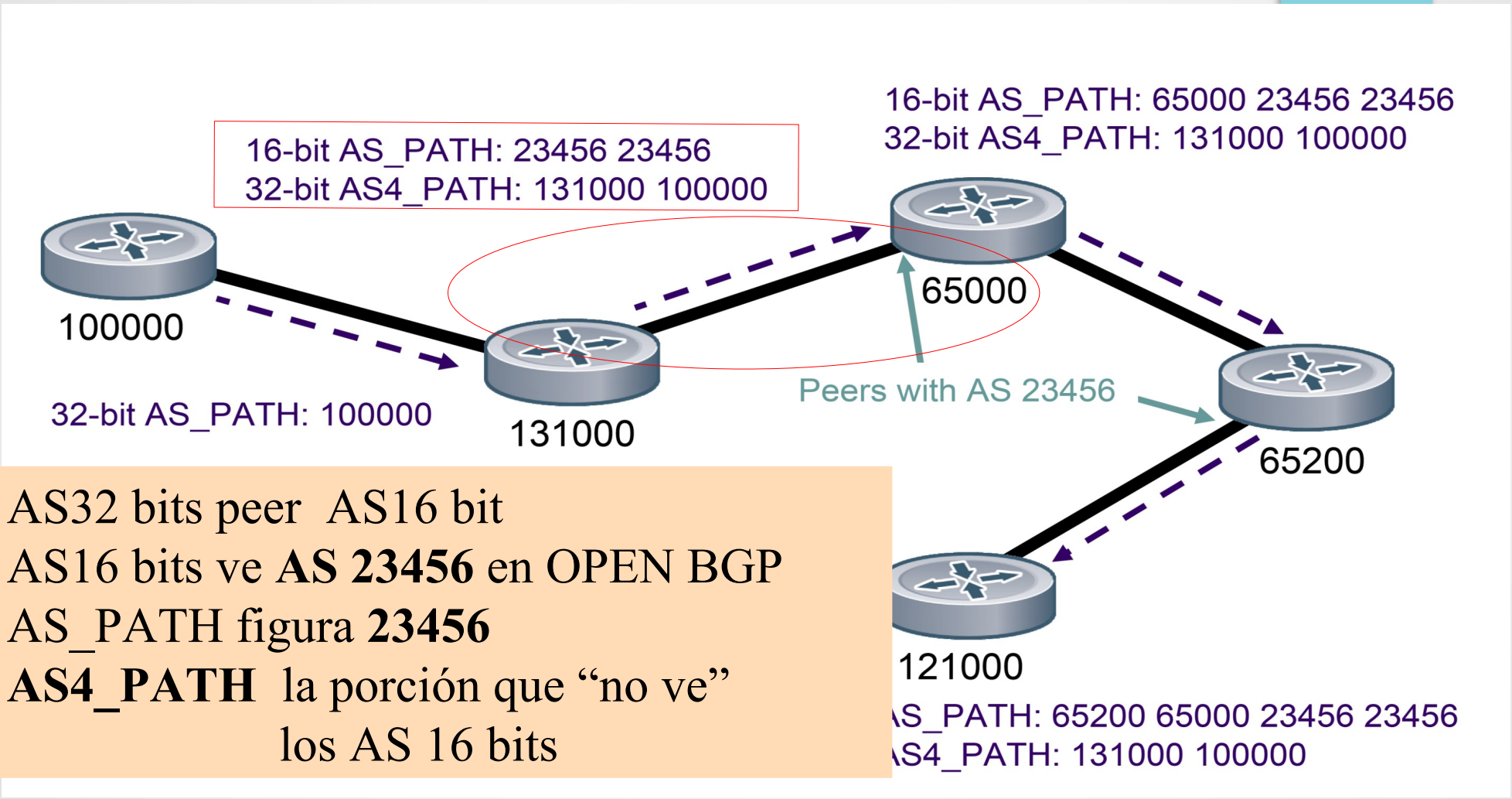
AS32 bits – Transición (3)



Ambos peer soportan AS32 bits
El AS_PATH cambia para soporte de 32 bits

Hankins and Malyster, NANOG45

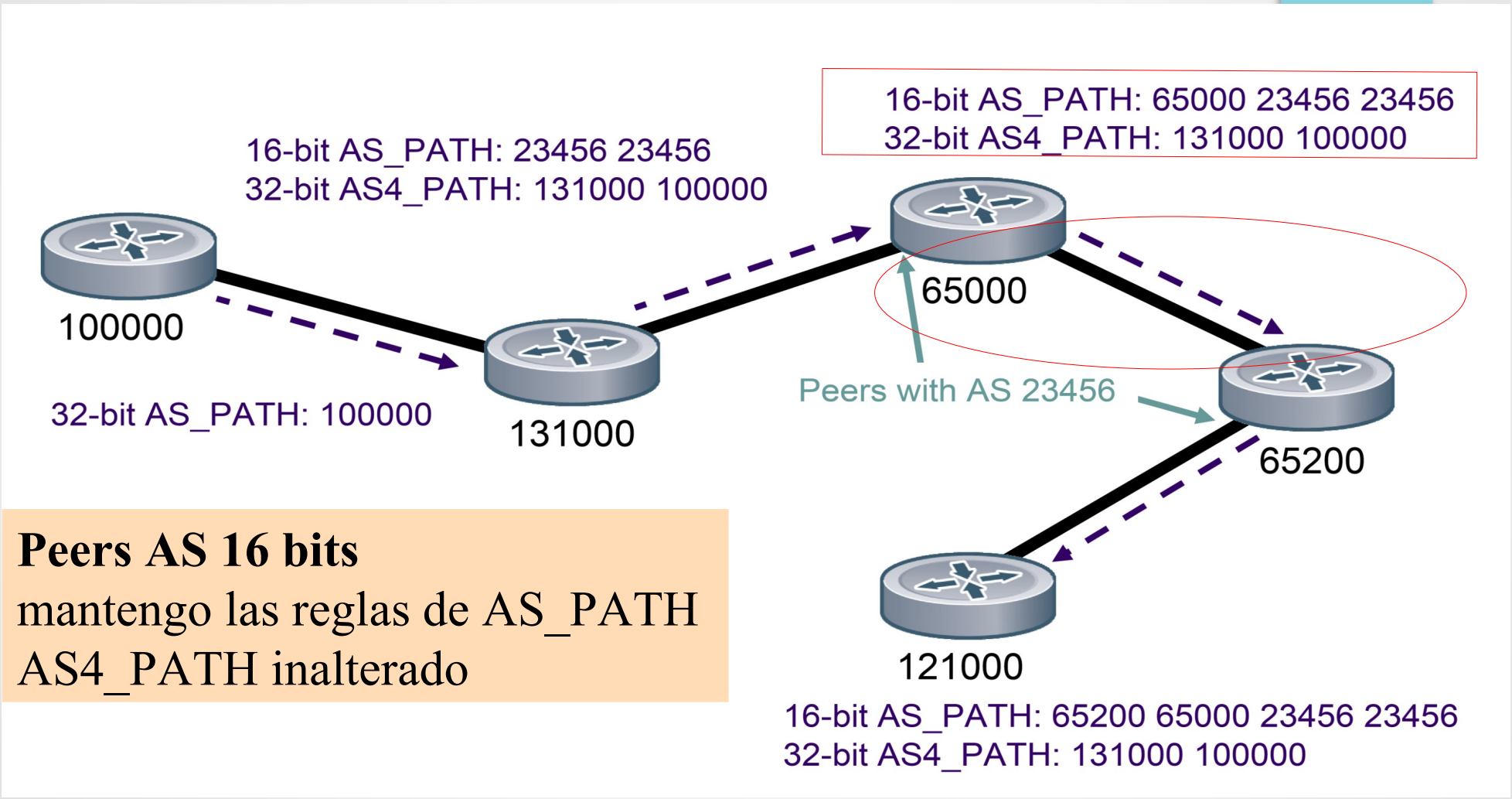
AS32 bits – Transición (3)



AS32 bits peer AS16 bit
AS16 bits ve **AS 23456** en OPEN BGP
AS_PATH figura **23456**
AS4_PATH la porción que "no ve"
los AS 16 bits

Hankins and Malyster, NANOG45

AS32 bits – Transición (3)

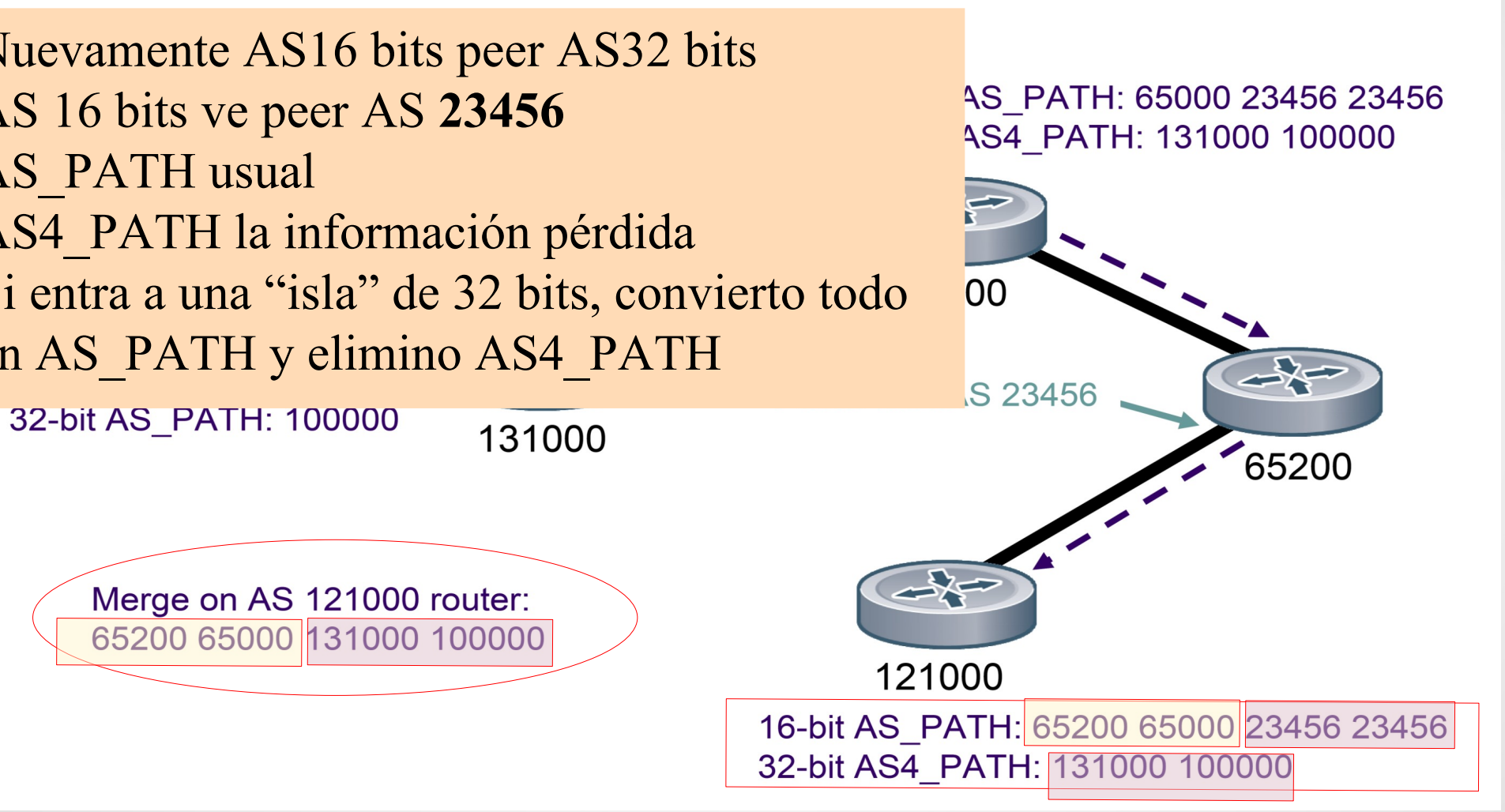


Peers AS 16 bits
mantengo las reglas de AS_PATH
AS4_PATH inalterado

Hankins and Malyster, NANOG45

AS32 bits – Transición (3)

Nuevamente AS16 bits peer AS32 bits
AS 16 bits ve peer AS **23456**
AS_PATH usual
AS4_PATH la información pérdida
Si entra a una "isla" de 32 bits, convierto todo en AS_PATH y elimino AS4_PATH



Hankins and Malyster, NANOG45

Agenda (8)

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

MD5 signature

- RFC 2385
- Protección contra segmentos TCP “insertados” en la conexión existente (especialmente TCP Resets)
- Actualmente costumbre en eBGP
- Hash MD5 de encabezado IP/TCP + datos + key (clave). **Enviado en opción TCP (option 19)**
- **(cisco)**

neighbor <direccion ip> password <string>

BGP secuestro de rutas

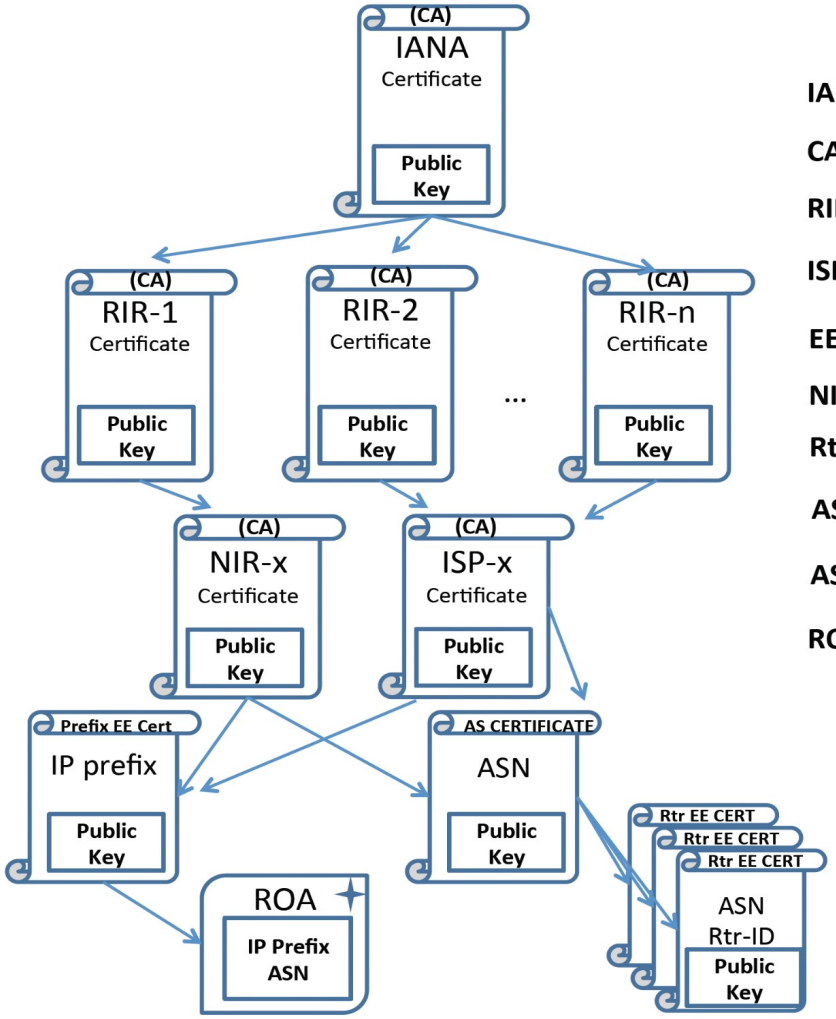
- **BGP prefix Hijacking:** publicación de prefijos de forma ilegítima por otro AS.
- ¿Por qué Ocurre? “Falla” en los filtros sobre que prefijos permitir desde un AS.
- En general debe filtrar el proveedor a sus clientes, luego es difícil de gestionar y se suelen usar reglas genéricas.
- **No hay forma de verificar los prefijos a que AS pertenecen, salvo que el proveedor se tome el trabajo.**

BGP prefix Hijacking: Incidentes públicos

- May 7, 2005: Google's May 2005 Outage
- February 24, 2008: Pakistan's attempt to block YouTube access within their country takes down YouTube entirely.
- April 8, 2010: Chinese ISP hijacks the Internet - China Telecom originated 37,000 prefixes not belonging to them in 15 minutes, causing massive outage of services globally.
- **February, 2014:** Canadian ISP used to redirect data from ISPs - 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.**
- **April 2017:** Russian telecommunication company Rostelecom (AS12389) originated 50 prefixes for numerous other Autonomous Systems. The hijacked prefixes belonged to financial institutions (**most notably MasterCard and Visa**), …...
- **January 2017: Iranian pornography censorship**
- 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).
- May 2019: Traffic to a public DNS run by Taiwan Network Information Center (TWNIC) was rerouted to an entity in Brazil (AS268869)
- **June 2019: Large European mobile traffic was rerouted through China Telecom** (AS4134)
- April 2020: A massive BGP hijack involving over 8800 prefixes affected companies such as **Akamai, Amazon and Alibaba** on April 1, 2020. Initiated by a Rostelecom user, the attack caused service disruptions throughout the world.
- **September 2020:** 500 prefixes wrongfully advertised as belonging to Telstra caused lengthy data detours via the Australian telecommunications company in September 2020. Telstra later apologised for the unintentional hijacking, stating the incident was caused by post verification testing to address an unrelated software bug

Propuesta RPKI and ROA

Preventing Prefix Hijacking: RPKI and ROA



IANA = Internet Assigned Numbers Authority

CA = Certification Authority

RIR = Regional Internet Registry

ISP = Internet Service Provider

EE Cert = End-Entity Certificate

NIR = National Internet Registry

Rtr = Router

AS = Autonomous System

ASN = Autonomous System Number

ROA = Route Origin Authorization (RFC 6482)

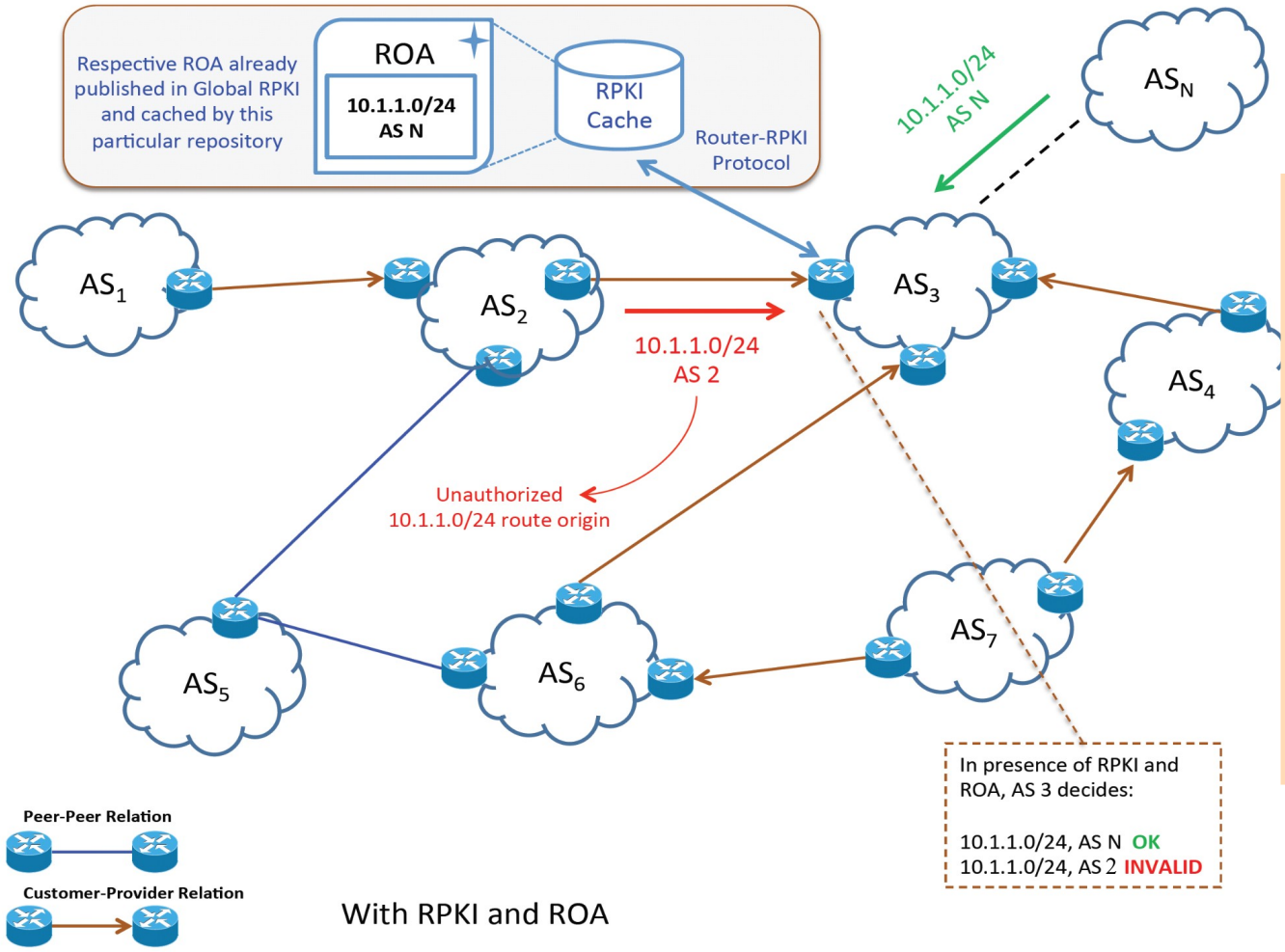
Administrative Resource Allocation Hierarchy

ROA: una firma digital del prefijo y AS habilitado a publicar. La delegación del árbol de autoridad es de las entidades que asignan los números de AS y rangos de direcciones IP

Marcelo Yannuzzi, curso "Graphs on path vectors"

RPKI and ROA

Preventing Prefix Hijacking: RPKI and ROA



La **verificación** se realiza "fuera de banda" (por fuera de BGP), aparece la figura de un caché para tener que realizar las verificación todas las veces.

Marcelo Yannuzzi, curso "Graphs on path vectors"

RPKI and ROA (Root of Authority)

- **Formato ROA:** [AS, {prefix/mask, maxLen}+]
- Implementar las consultas al cache RPKI
- Requiere cambios en **algoritmo de decisión** de rutas de BGP y estados de prefijos.
- Validación de caminos (preferencia):

  – 0 = BGP_PFX_STATE_VALID (Lookup Successful)

  – 1 = BGP_PFX_STATE_NOT_FOUND (Not in the table)

  – 2 = BGP_PFX_STATE_INVALID (Lookup invalid - different origin AS or masklength not in the range)
- Mediante **políticas** puedo decidir solo aceptar los prefijos **válidos**

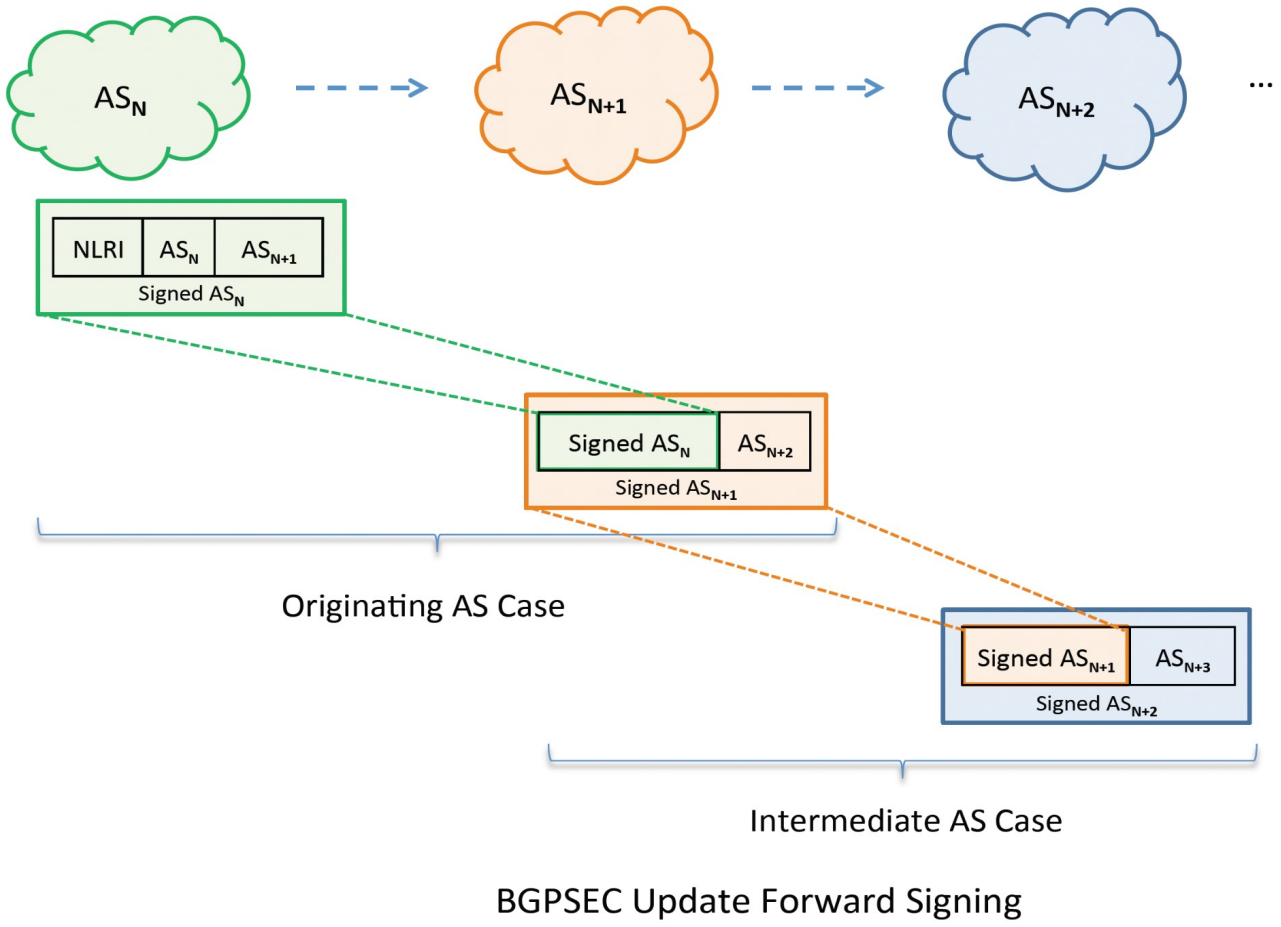
Path Hijacking: Secuestro de camino

- **Rute Hijacking:** Alteración del AS_PATH (ataques **men in the middle**, DoS, beneficios por cobro de tráfico)
- **Por más que validemos el origen, no estamos validado el camino.**
- Se conoce como **BGPSEC**. No solo hay que validar los peer sino también la cadena.

  Problema de adopción.
- Hay algunas propuestas, pero ninguna ha prosperado.
- Dificultades de implementación y computo.

BGPSEC (propuesta)

Preventing Route Hijacking: BGPSEC

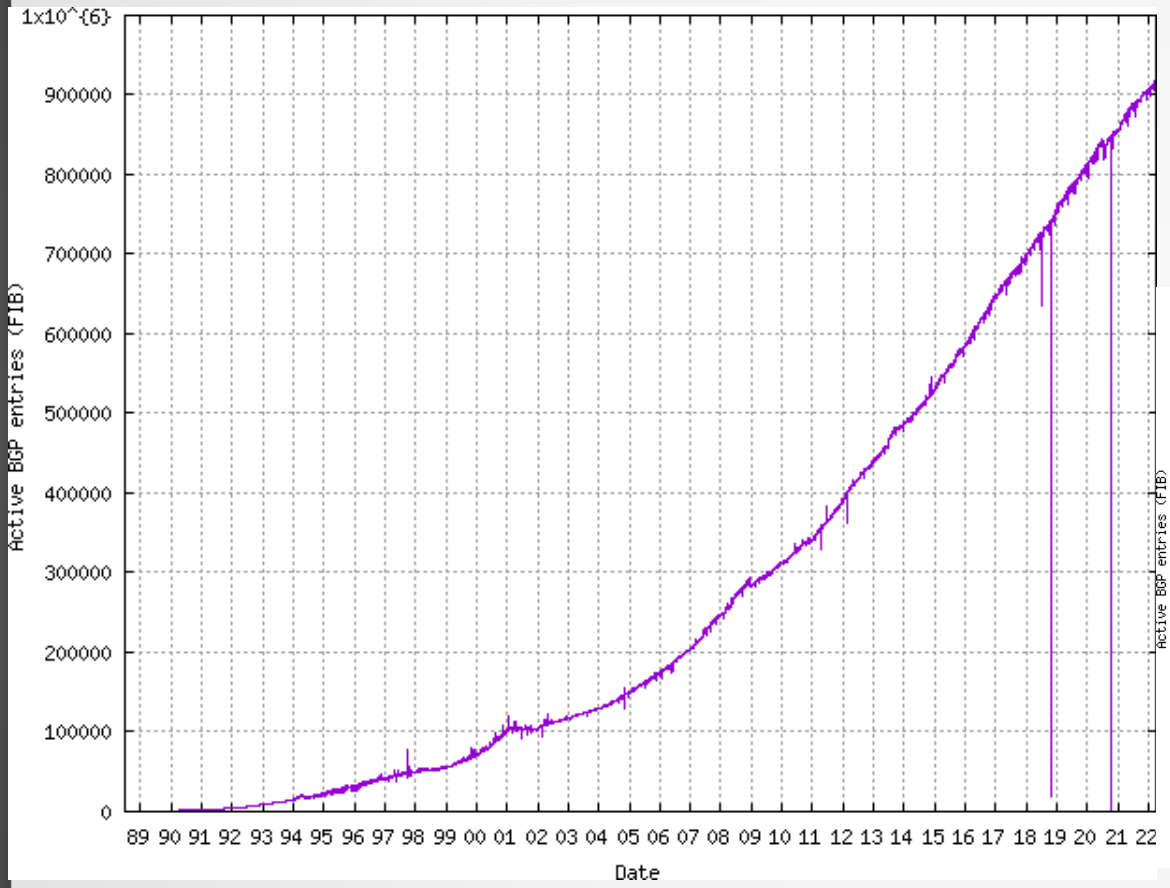


Marcelo Yannuzzi, curso "Graphs on path vectors"

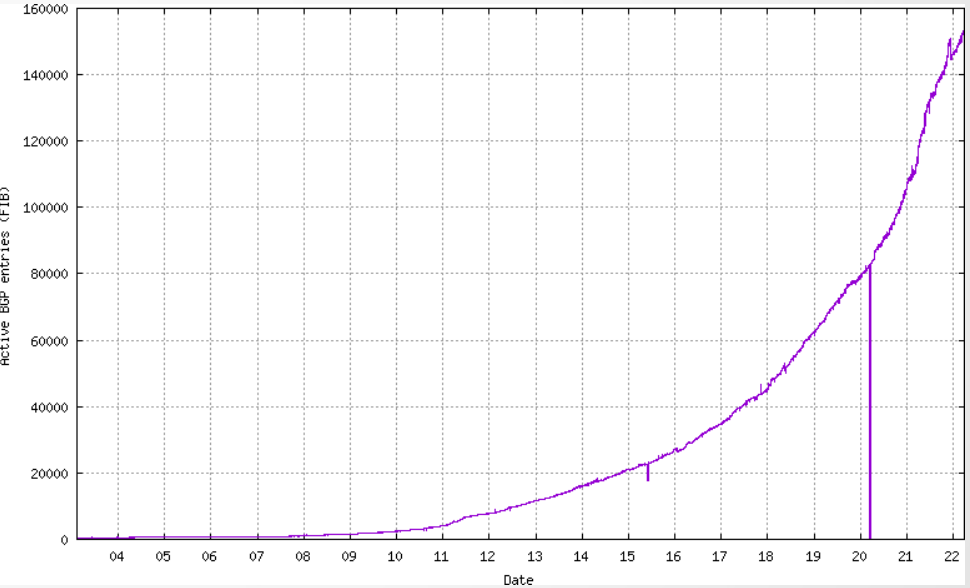
Agenda (9)

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

Prefijos en la DFZ (FIB http://www.cidr-report.org)

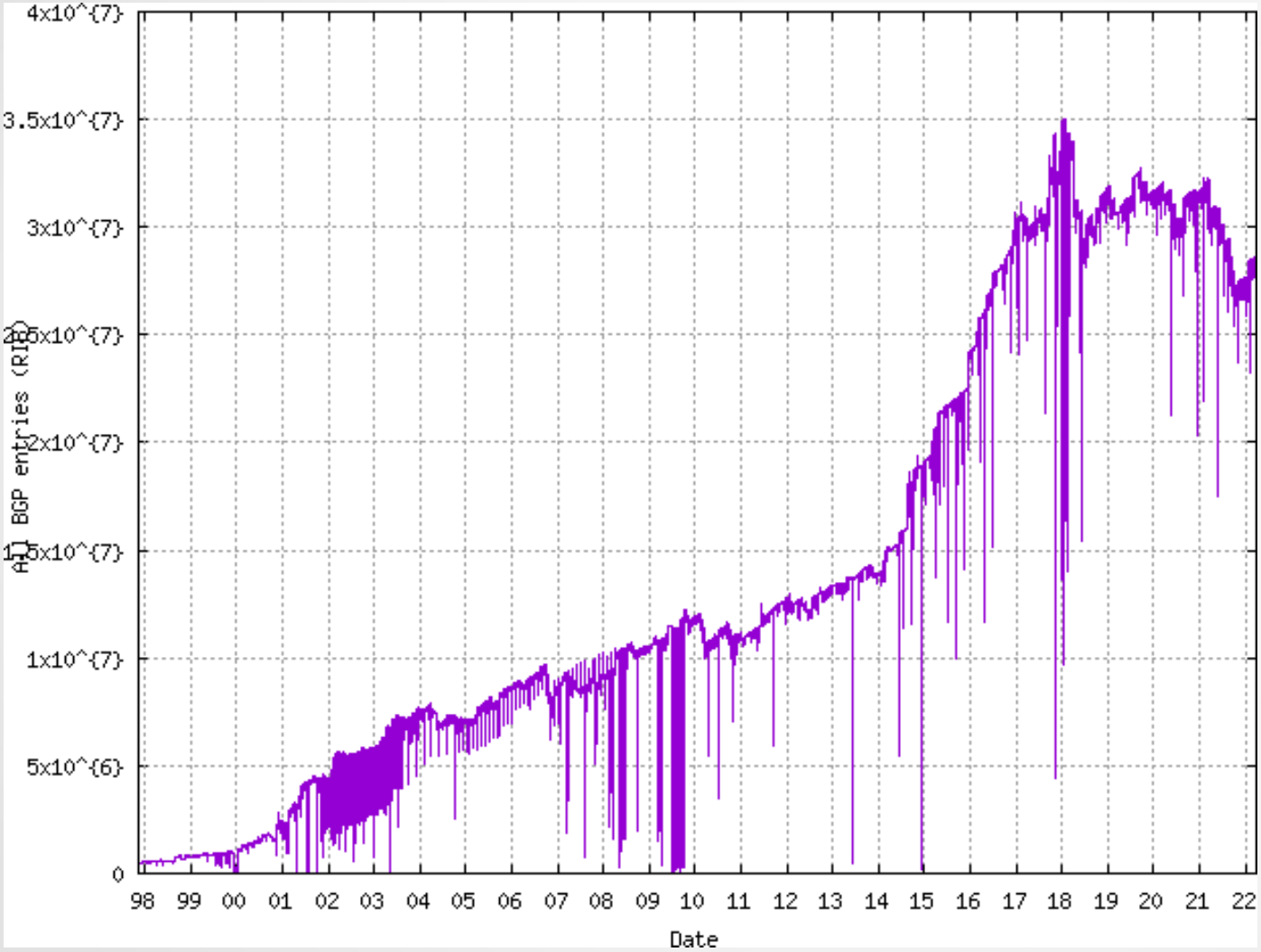


IPv4



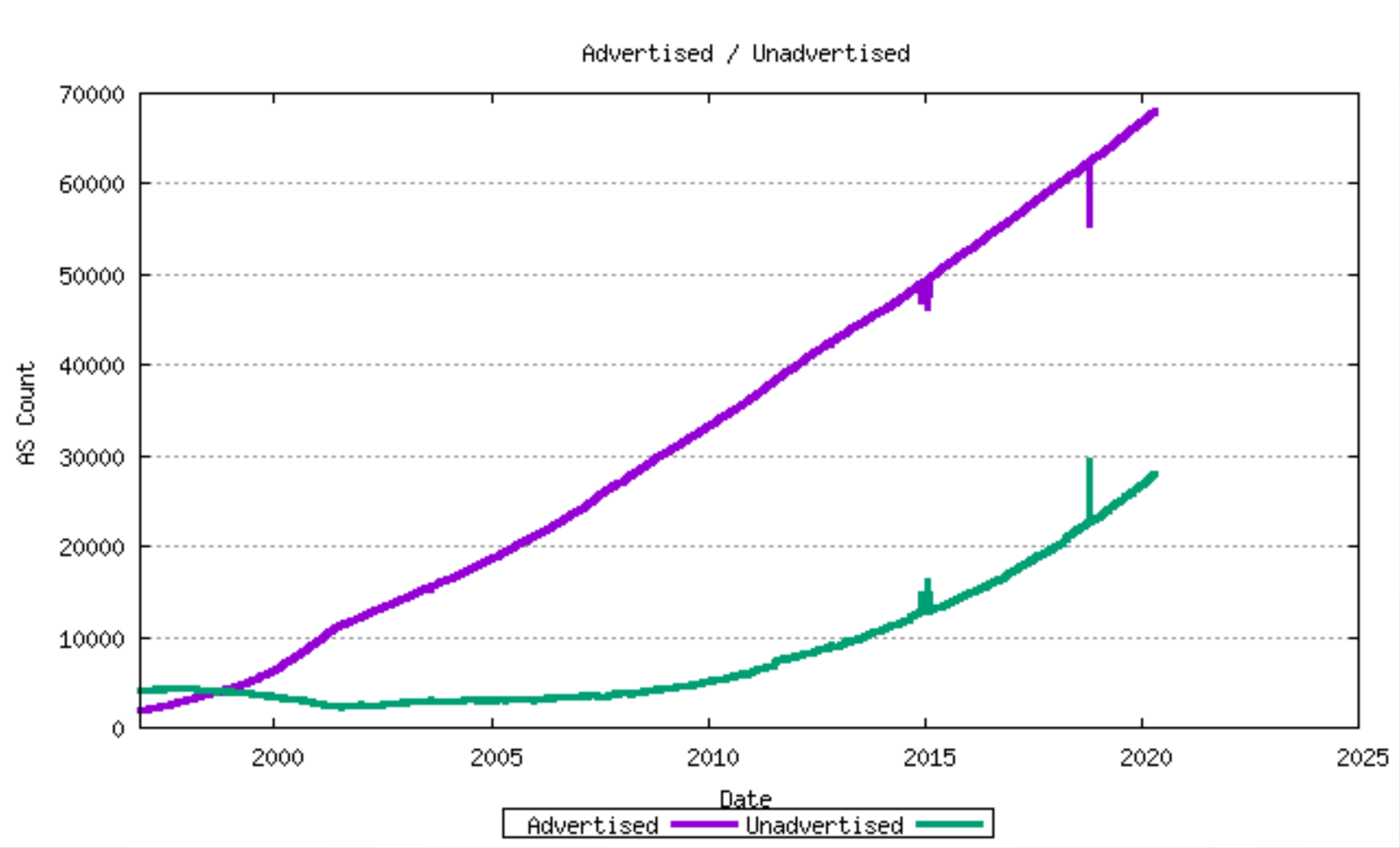
IPv6

Todas las BGP – RIB de IPv4 (http://bgp.potaroo.net/AS6447)



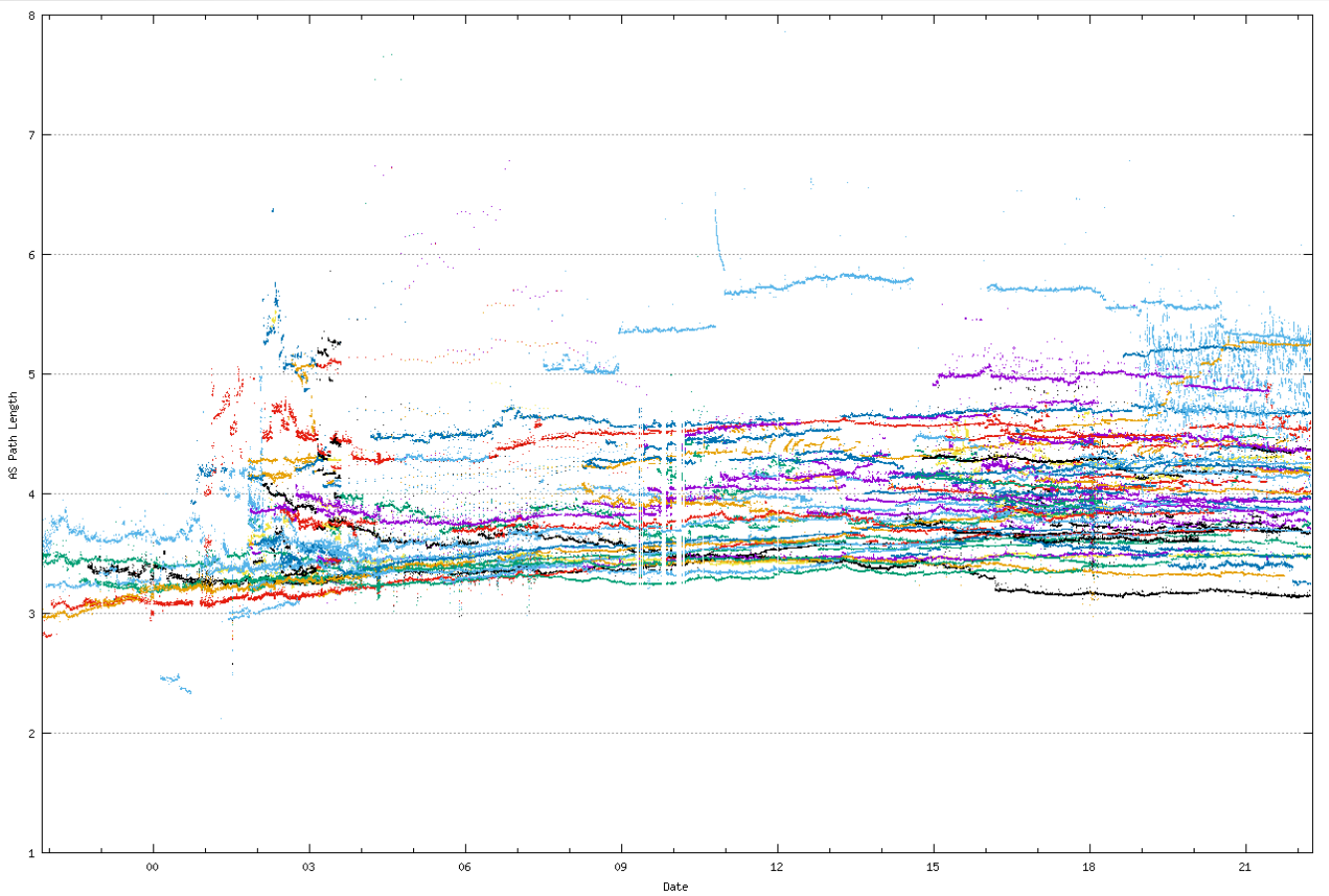
- En 2022 (AS6447):
  - FIB DFZ~937k
  - RIB ~ 28,6 M
  - RIB/FIB ~ 30
  - Recuerden que la RIB son los prefijos y los atributos de caminos,

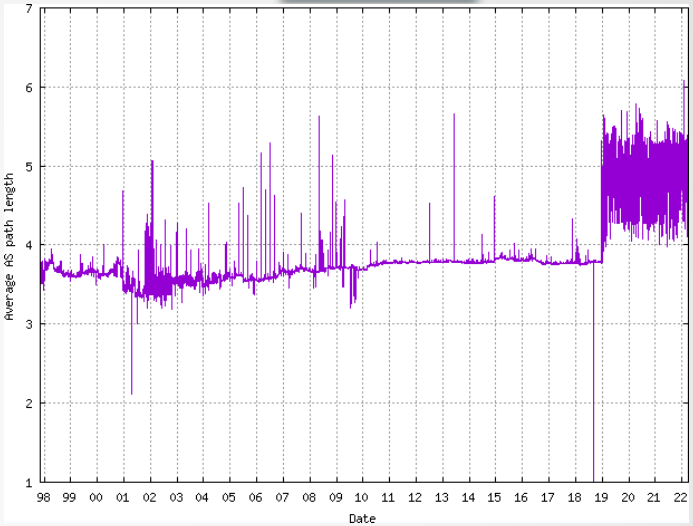
Cantidad de AS (http://www.potaroo.net/tools/asn32/)



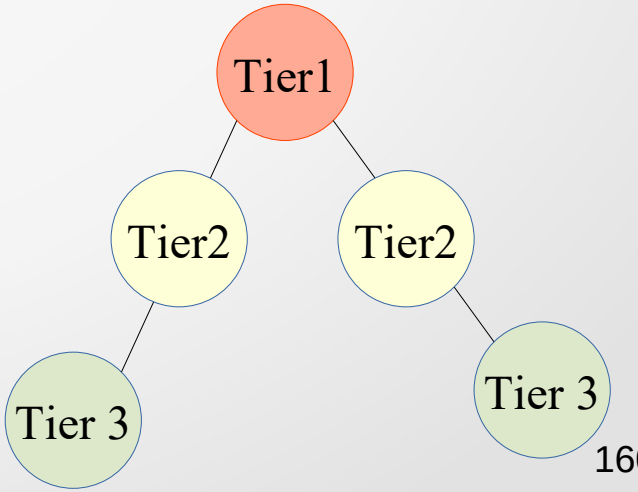
Largo promedio del AS-PATH

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





Esquema Conceptual

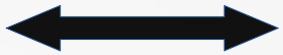


Looking Glass y Route Servers

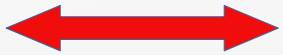
- ¿Cómo saber cómo se ven mis redes fuera de mi AS?
- **Looking glass:** consultas BGP, traceroute, ping.....
- **Ejemplos:**
- http://www.traceroute.org/ (camino)
- http://netmon.acad.bg/lg/ (looking glass)
- https://www.sprint.net/lg/lg_start.php (looking glass)
- https://www.us.ntt.net/support/looking-glass/ (looking glass)
- https://www.rediris.es/red/lg/ (looking glass)

Ejemplo: salida sprint “show bgp 164.73.0.0/16”

BGP routing table entry for **164.73.0.0/16**
Versions:
  Process bRIB/RIB SendTblVer
  Speaker 282797841 282797841
Last Modified: Mar 31 07:44:46.447 for 1d14h
Paths: (2 available, best #1)
  Advertised IPv4 Unicast paths to update-groups (with more than one peer):
   0.3 0.5 0.6
  Advertised IPv4 Unicast paths to peers (in unique update groups):
   144.228.242.75 144.228.243.242 160.81.104.38 208.76.14.223
  Path #1: Received by speaker 0
  Advertised IPv4 Unicast paths to update-groups (with more than one peer):
   0.3 0.5 0.6
  Advertised IPv4 Unicast paths to peers (in unique update groups):
   144.228.242.75 144.228.243.242 160.81.104.38 208.76.14.223
  6057 1797
   144.228.241.140 (metric 703) from 144.228.241.7 (144.228.241.140)
    Origin IGP, localpref 100, valid, internal, best, group-best
    Received Path ID 0, Local Path ID 1, version 282797841
    Community: 1239:500 1239:999 1239:1000 1239:1026
    Extended community: RT:6057:6057
    Originator: 144.228.241.140, Cluster list: 144.228.241.7

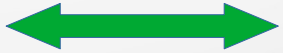


Dos caminos en RIB

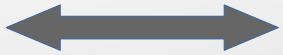


AS-PATH

Localpref presente



Sprint AS 1239



Reflector

Route servers

- Servidores (enrutadores) que hablan BGP, accesibles públicamente. Solo puedo consultar BGP (sin traceroute o ping)
- **Ejemplos:**
- telnet://route-views.oregon-ix.net/ (no está disponible)
- telnet://route-views6.routeviews.org
- telnet://route-server.ip.att.net/
- **Lista:** https://www.routeservers.org/

Route server – show ip bgp

route-views>**sh ip bgp**

BGP table version is 81485376, local router ID is 128.223.51.103

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

r RIB-failure, S Stale, m multipath, b backup-path, f RT-Filter,

x best-external, a additional-path, c RIB-compressed,

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

RPKI validation codes: V valid, I invalid, N Not found

	Network	Next Hop	Metric	LocPrf	Weight	Path
Nr>	0.0.0.0	162.251.163.2			0	53767 3257 i
V*	1.0.0.0/24	162.251.163.2			0	53767 3257 13335 i
V*		154.11.12.212	0		0	852 13335 i
V*		195.208.112.161			0	3277 3267 13335 i
V*>		**4.68.4.46**	**0**		**0**	**3356 13335 i**
V*		134.222.87.1	0		0	286 3257 13335 i

.....

Todos los caminos que pasaron las políticas (**RIB**), luego solo se utiliza el **mejor** en la **FIB**

Route server – show ip bgp (2)

route-views>sh ip bgp | inc 6057 1797

```
N*   164.73.0.0      162.251.163.2                    0 53767 3257 174 6057 1797 i
N*                   212.66.96.126                    0 20912 3257 1239 6057 1797 i
N*                   206.24.210.80                    0 3561 209 3356 174 6057 1797 i
N*>                  144.228.241.130       719        0 1239 6057 1797 i
```

route-views>sh ip bgp 164.73.0.0

BGP routing table entry for 164.73.0.0/16, version 75900492

Paths: (28 available, best #14, table default)

Not advertised to any peer

Refresh Epoch 1

1239 6057 1797

144.228.241.130 from 144.228.241.130 (144.228.241.130)

Origin IGP, metric 719, localpref 100, valid, external, best

path 7FE169CBDF08 RPKI State not found

rx pathid: 0, tx pathid: 0x0

Route server – show ip route 164.73.0.0

route-views>sh ip route 164.73.0.0
Routing entry for 164.73.0.0/16
  Known via "bgp 6447", distance 20, metric 719
  Tag 1239, type external
  Last update from 144.228.241.130 5d21h ago
  Routing Descriptor Blocks:
  * **144.228.241.130**, from 144.228.241.130, 5d21h ago
      Route metric is 719, traffic share count is 1
      AS Hops 3
      Route tag 1239
      MPLS label: none

Route Server – Ejemplos de ROA

route-views>show ip bgp 164.73.0.0/16
BGP routing table entry for 164.73.0.0/16,
version 75900492
Paths: (28 available, best #14, table default)
Not advertised to any peer
Refresh Epoch 1
53767 3257 174 6057 1797
162.251.163.2 from 162.251.163.2 (162.251.162.3)
Origin IGP, localpref 100, valid, external
Community: 3257:8922 3257:30669 3257:50002 3257:51200 3257:51205 53767:5000
path 7FE16030B908 **RPKI State not found**
rx pathid: 0, tx pathid: 0

route-views>show ip bgp 200.40.0.0/16
BGP routing table entry for 200.40.0.0/16,
version 50757039
Paths: (27 available, best #21, table default)
Not advertised to any peer
Refresh Epoch 1
53767 3257 6461 6057
162.251.163.2 from 162.251.163.2 (162.251.162.3)
Origin IGP, localpref 100, valid, external
Community: 3257:8936 3257:30512 3257:50002 3257:51200 3257:51205 53767:5000
path 7FE0DAD077B8 **RPKI State valid**
rx pathid: 0, tx pathid: 0

Looking Glass (rediris–CICA) - traceroute

inet.0: 798323 destinations, 1045166 routes (798322 active, 0 holddown, 1 hidden)

+ = Active Route, - = Last Active, * = Both

164.73.0.0/16 *[BGP/170] 20:37:26, MED 60, localpref 161, from 130.206.206.250

AS path: 20965 27750 1797 I, validation-state: unverified

> to **130.206.245.125** via ae3.0

BGP

traceroute to 164.73.32.5 (164.73.32.5), 30 hops max, 52 byte packets

1 130.206.245.125 (130.206.245.125) 10.980 ms 7.778 ms 7.672 ms **<AS 766>**

2 rediris.mx1.mar.fr.geant.net (62.40.124.192) 21.617 ms 21.531 ms 21.611 ms **<AS 20965>**

3 ae8.mx1.gen.ch.geant.net (62.40.98.73) 27.789 ms 27.869 ms 27.874 ms

4 ae6.mx1.par.fr.geant.net (62.40.98.183) 34.995 ms 35.072 ms 34.975 ms

5 redclara.par.fr.geant.net (62.40.125.169) 137.468 ms 139.401 ms 136.398 ms **<AS 27750>**

6 cl-us.redclara.net (200.0.204.59) 263.065 ms 263.369 ms 263.112 ms

MPLS Label=24003 CoS=0 TTL=1 S=1

7 ar-cl.redclara.net (200.0.204.89) 262.701 ms 262.683 ms 262.652 ms

8 rau-ar.redclara.net (200.0.204.178) 272.689 ms 272.937 ms 272.609 ms

9 164.73.128.150 (164.73.128.150) 272.698 ms 273.058 ms 272.661 ms **<AS 1797>**

10 * * *

11 164.73.250.146 (164.73.250.146) 271.424 ms 270.186 ms 270.285 ms

12 164.73.32.126 (164.73.32.126) 270.035 ms 269.871 ms 270.070 ms

13 - 30 * * *

IP

Agenda (10)

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

Ejemplos de Configuración en Cisco

- **En los laboratorios utilizaremos FRRouting de Linux Fundation, la CLI (Comand Line Interface) está basada en la CLI de CISCO. Su utilización es como referencia a una implementación**
- Definición de sistema autónomo y de vecinos

```
router bgp 65525 (My AS)
  neighbor 200.108.19.2 description cliente prueba
  neighbor 200.108.19.2 remote-as 20255          (IP neighbor, remote AS)
  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 – bajo address family

router bgp 65525

neighbor 200.108.19.2 remote-as 20255 **(se declara el vecino)**

.....

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

Configuración IPv6 – bajo address family

router bgp 65525

neighbor 2800:840:5::1 remote-as 65525 **(se declara el vecino)**

.....

address-family ipv6 unicast

neighbor 2800:840:5::1 **activate** **(intercambio BGP de address-family)**

neighbor 2800:840:5::1 prefix-list pfl-filtrov6 in

neighbor 2800:840:5::1 route-map rmap-65525v6 in

no auto-summary

no synchronization

exit-address-family

BGP - Generando anuncios

router bgp 65525

network 192.168.20.0 mask 255.255.255.0

redistribute rip

redistribute connected

- **Network:** La red 192.168.20.0/24 **debe estar en la tabla de ruteo** (**FIB**) para que se anuncie (**origin IGP**)

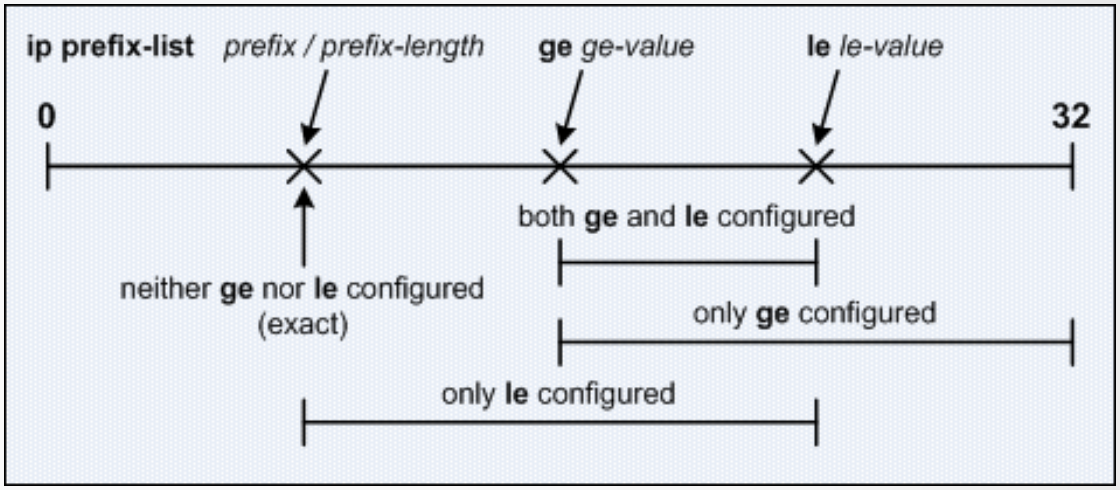
  Forzar la entrada en la FIB: **ip route 192.168.20.0/24 Null0**

- **Redistribute:** Redistribución desde otro protocolo (**origin incomplete**), puede utilizarse con filtros.

  No es habitual en peers eBGP, si se utiliza en otros escenarios.

CISCO 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 172.1.0.0/12 ge 20

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 (permito todo)

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
- ip as-path access-list **name/number** <permit|deny>
- **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)*

……

FRR: bgp as-path access-list <nombre>

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)* - matchea con cualquier as-path que comience con 20255, permite preepends del AS 20255 y luego puede contener el AS 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

BGP Políticas Complejas - **route-map**

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

  Match <condiciones>

  Set < atributos>

- Puedo tener mas de una condición de **match**, cero o más de un **set**

Ejemplo: route -map

route-map rmap-pref permit **10**

match as-path 100

set local-preference 250

route-map rmap-pref permit **20**

match ip address prefix-list filtro

set local-preference 300

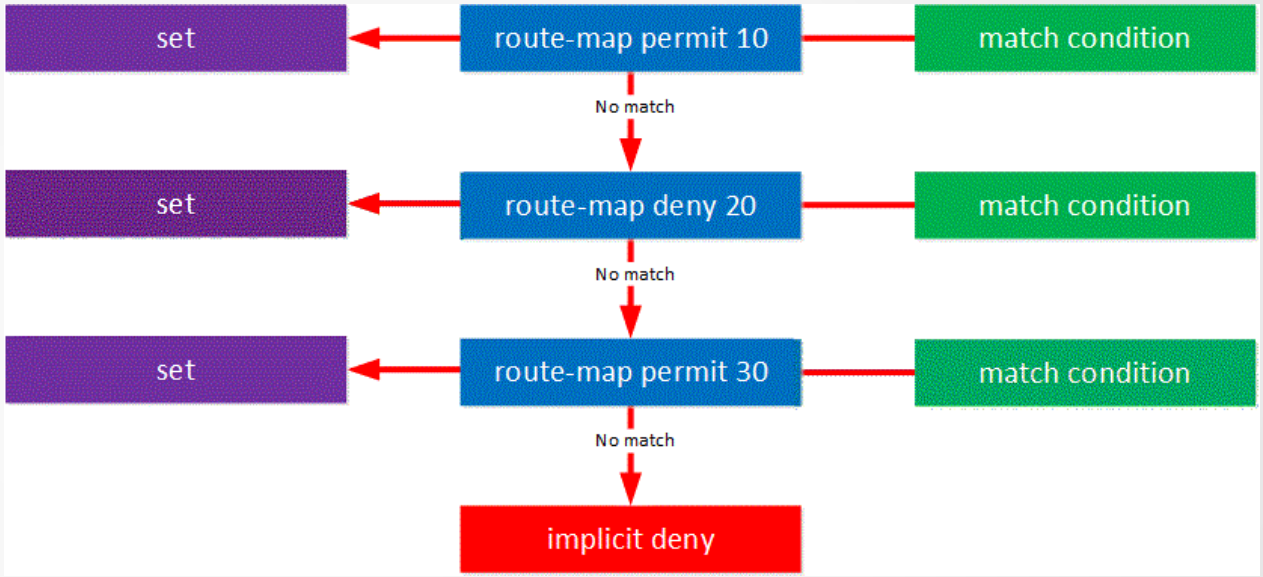
route-map rmap-pref permit **30**

match as-path 150

match ip address-prefix-list pfl-cliente150

set local-preference 500

set community 65005:10001



Route map – **Match conditions**

- **Condiciones para el match:**

  as-path access-list

  prefix-list o access-list (prefijos)

  interface

  peer

  rpki

  community-list

  local-preference

Route-map – **Set Attributes**

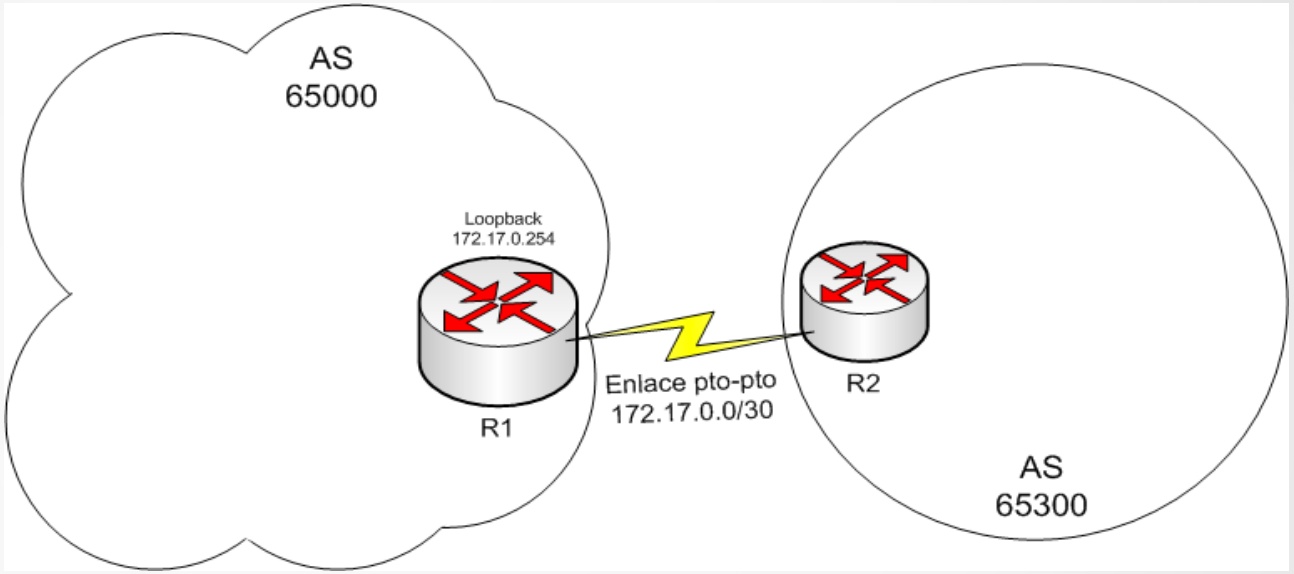
- **Set:**
  - set aggregator
  - set atomic-aggregate
  - set as-path prepend
  - set community
  - set comm-list
  - set dampening
  - set ip next-hop
  - set local-preference
  - set origin

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

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 pfl-solodefault in
..
ip prefix-list pfl-solodefault seq 10 permit 0.0.0.0/0
ip prefix-list pfl-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 pfl-solodefault out
    neighbor 172.17.0.2 prefix-list pfl-cliente653 in
    neighbor 172.17.0.2 filter-list 50 in
```

Router del proveedor (2)

```
ip prefix-list pfl-cliente653 seq 10 permit 172.17.128.0/19 le 24
ip prefix-list pfl-cliente653 seq 20 permit 172.17.160.0/20 le 24
ip prefix-list pfl-cliente653 seq 1000 deny 0.0.0.0/0 le 32
!
ip prefix-list pfl-solodefault seq 10 permit 0.0.0.0/0
ip prefix-list pfl-solodefault seq 20 deny 0.0.0.0/0 le 32
!
ip as-path access-list as-path-cliente653 permit ^65300$
! Si permito prepends
ip as-path access-list as-path-cliente653 permit ^65300(_65300)*$
!
route-map rmap-cliente653 permit 10
match as-path as-path-cliente653
```

Loopback Interface on routing

- Una interfaz de loopback **“pertenece” al router de forma independiente de las interfaces físicas** (interfaces lógicas).
- Siempre está “arriba” o disponible, salvo que el no lo este el router.
- En caso de utilizarse para formar adjacencias entre dos routers, tiene como ventaja de estar **disponible mientras al menos una interfaz física esta disponible y este participando del IGP**.
- Útiles para el acceso a gestión o reportes de alarmas u otros intercambios donde interese dialogar con el router (independientemente de la interfaz).
- Se suele utilizar como valor de **router-id** en varios protocolos (OSPF, LDP, BGP)

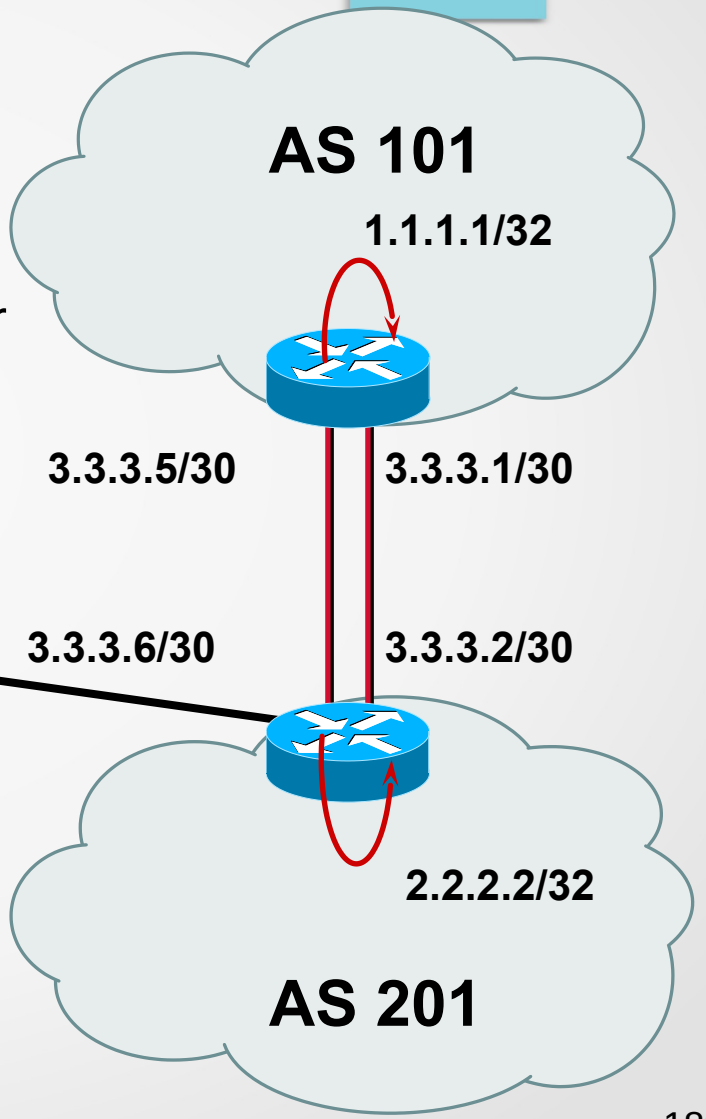
Loopback Interface on routing

```
logging source-interface Loopback0
snmp-server trap-source IPv4 Loopback0
!
ntp
 source Loopback0
!
interface Loopback0
 ipv4 address 200.58.155.129 255.255.255.255
!
router ospf core
 router-id 200.58.155.129
!
router bgp 19422
 bgp router-id 200.58.155.129
 neighbor 200.58.155.130 remote-as 19422
 neighbor 200.58.155.130 update-source Loopback0
```

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
```



BGP Multipath

- RFC 7911 (2016)
- Permite enseñar más de un camino.
- **ADD-PATH Capability** (ambos interlocutores deben soportarlo)
- Extended NLRI

extended NLRI

```
+-------- -------------------------+
| Path Identifier (4 octets) |
+----------------------------------+
| Length (1 octet)           |
+----------------------------------+
| Prefix (variable)          |
+----------------------------------+
```

NLRI (original)

- Sin esta opción, enseñar un nuevo camino se considera como re-escribir el actual (implícitamente doy de baja el camino vigente y de alta el nuevo camino)
- **Nota:** Observar que el NRLI tiene un **id** de path, la baja es por prefijo y **id**.

BGP Multipath - ECMP

- Por defecto asume **caminos paralelos** (coinciden weigth, local-preference, AS-path, Origin, MED, IGP metric).
- Se limita la cantidad de caminos diferentes.

```
router bgp 100
    maximum-paths 2
```

- Si queremos considerar caminos con el mismo AS-path lenght.

```
router bgp 100
 bgp log-neighbor-changes
 bgp bestpath as-path multipath-relax
 maximum-paths 2
```

- ¿Puedo usar dos entradas en la FIB con la misma métrica?

  SI. ECMP (Equal Cost Multi Path): balancea el tráfico entre los caminos.

  Típicamente por flujo (protocolo, IP origen, puerto origen, IP destino, puerto destino), aunque se puede realizar por paquete.

BGP Multipath - tradeoff

- Es posible elegir valores de máxima cantidad de caminos de forma diferente para iBGP y eBGP.
- Los comandos puede diferir el lugar donde se puede aplicar.
- Debemos entender bien el costo-beneficio:
- **Riesgo**:
  - Más caminos en la RIB, mas memoria.
  - Balanceo por caminos muy diferentes: diferentes retardos y variaciones de retardo.
- **Beneficios:**
  - Ante un problema, el witdrawn baja el camino, pero no tengo problema de path-hunting (más estable).
  - Distribución de carga entre las salidas (eBGP).
  - En iBGP potencia las redundancias internas.

IGP y BGP en un Service Provider

- **Dentro de un ISP tenemos tres tipos de redes:**
  - **Redes de clientes**
  - **Redes externas** al AS
  - **Redes de infraestructura** (WAN/LAN interconexión, Loopback y DNS)

    Redes de servicios (Distribución de contenido, cache, etc), dependiendo de la implementación podrían ser externas al AS o como un cliente.
- ¿Hay alguna mas importante que las otras?
- El camino “óptimo” al Next-Hop y la rápida adaptación a los cambios es crítico para el funcionamiento.
- La tabla completa de internet IPv4 ~ 900k prefijos, en un ISP “habitualmente” menos de 30k prefijos generados desde el IGP.

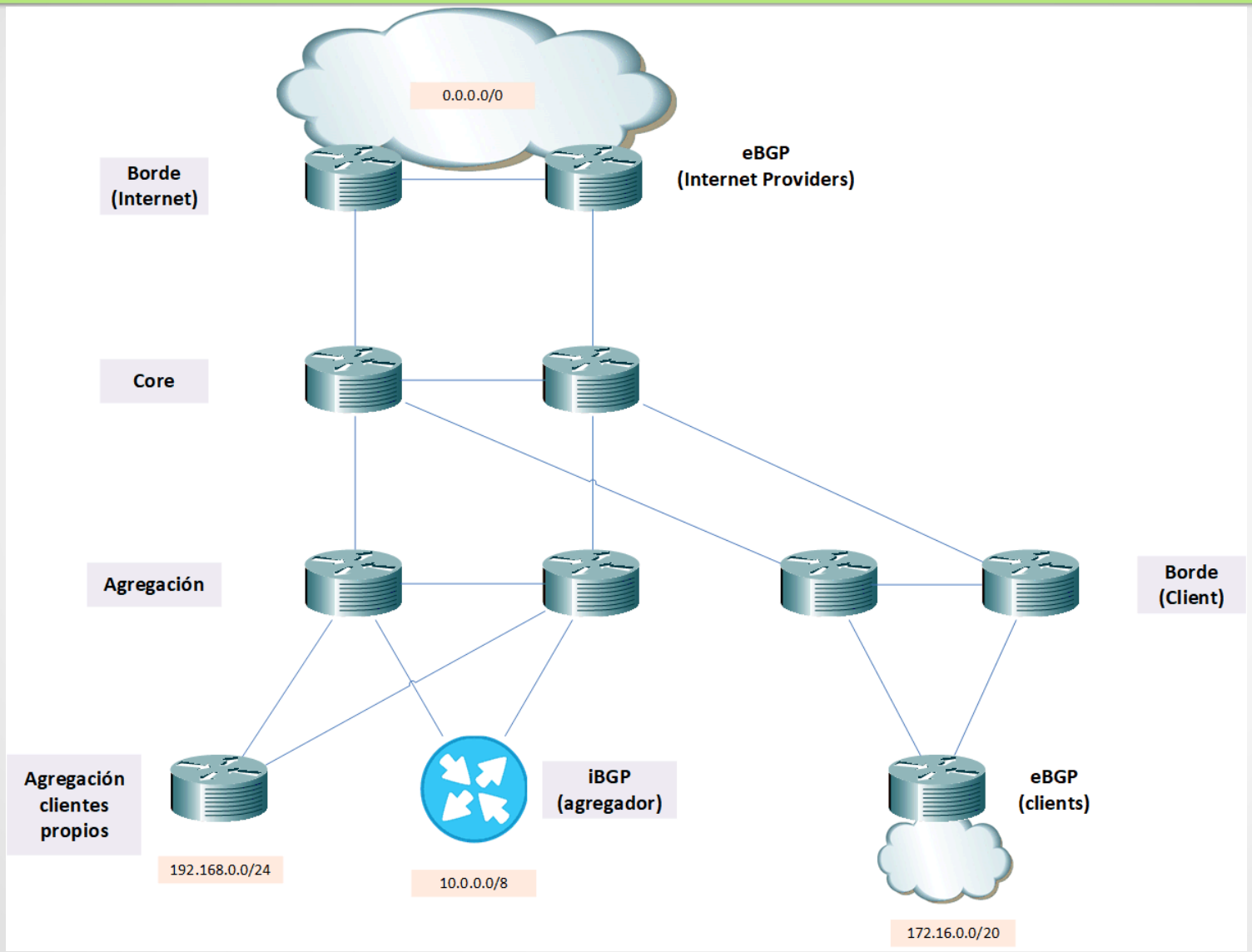
IGP y BGP – Best Practise in SP (II)

- **Utilizar IGP para crear la infraestructura**. Obtengo una topología simple, solo con los links y las loopback.
- Utilizar **eBGP** para el intercambio de prefijos **con otros AS** y para implementar políticas de intercambio de prefijos.
- Utilizar **iBGP** para propagar internamente, de forma total o parcial, los **prefijos de otros AS**; y **los prefijos de clientes**.
- Notar la diferencia entre un AS cliente de tránsito y un cliente que utiliza direccionamiento IP perteneciente al SP.
- Por el IGP aprendo como llegar al next-hop que propaga BGP.
- **Levantar las sesiones iBGP desde interfaces de loopback**, de forma de aprovechar la redundancias de caminos que descubre el IGP.

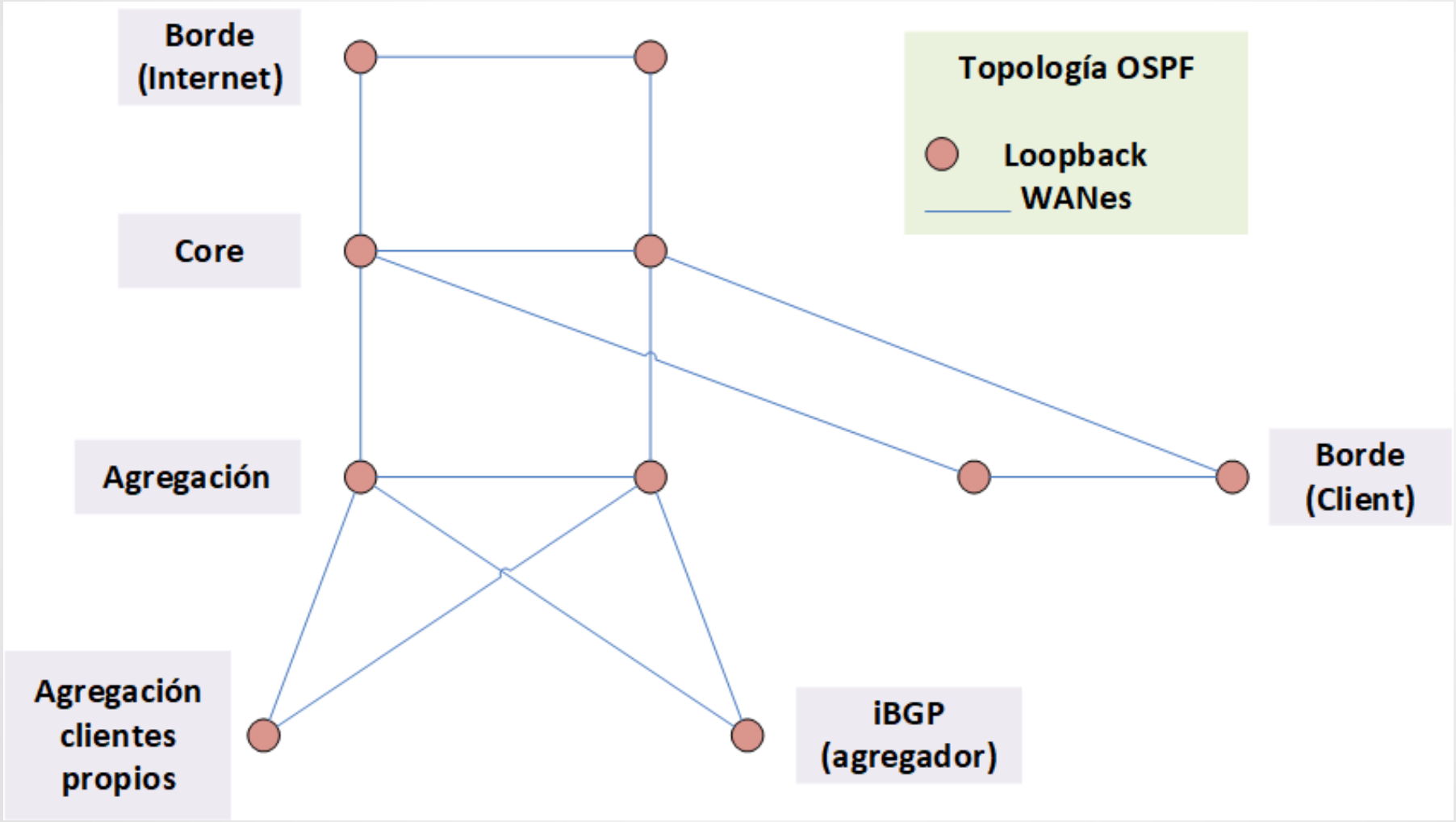
IGP y BGP – Best Practise in SP

- Las publicaciones **eBGP** generarlas de forma independiente del IGP. **Generarlas localmente** (network) de forma que se publiquen /M e internamente se conozcan como /N (más especificas, M>N).
- **Dividir en áreas y sumarizar al backbone (ABR)**. Normalmente requiere un plan de direccionamiento adecuado.
- No utilizar sincronismo en iBGP y utilizar r**eflectores** u otro mecanismo para escalar las sesiones iBGP.
- En caso de utilizar reflectores, utilizar **diferentes cluster-id**, **no modificar** el next-hop de iBGP.
- Revisar los timers de sesiones iBGP.

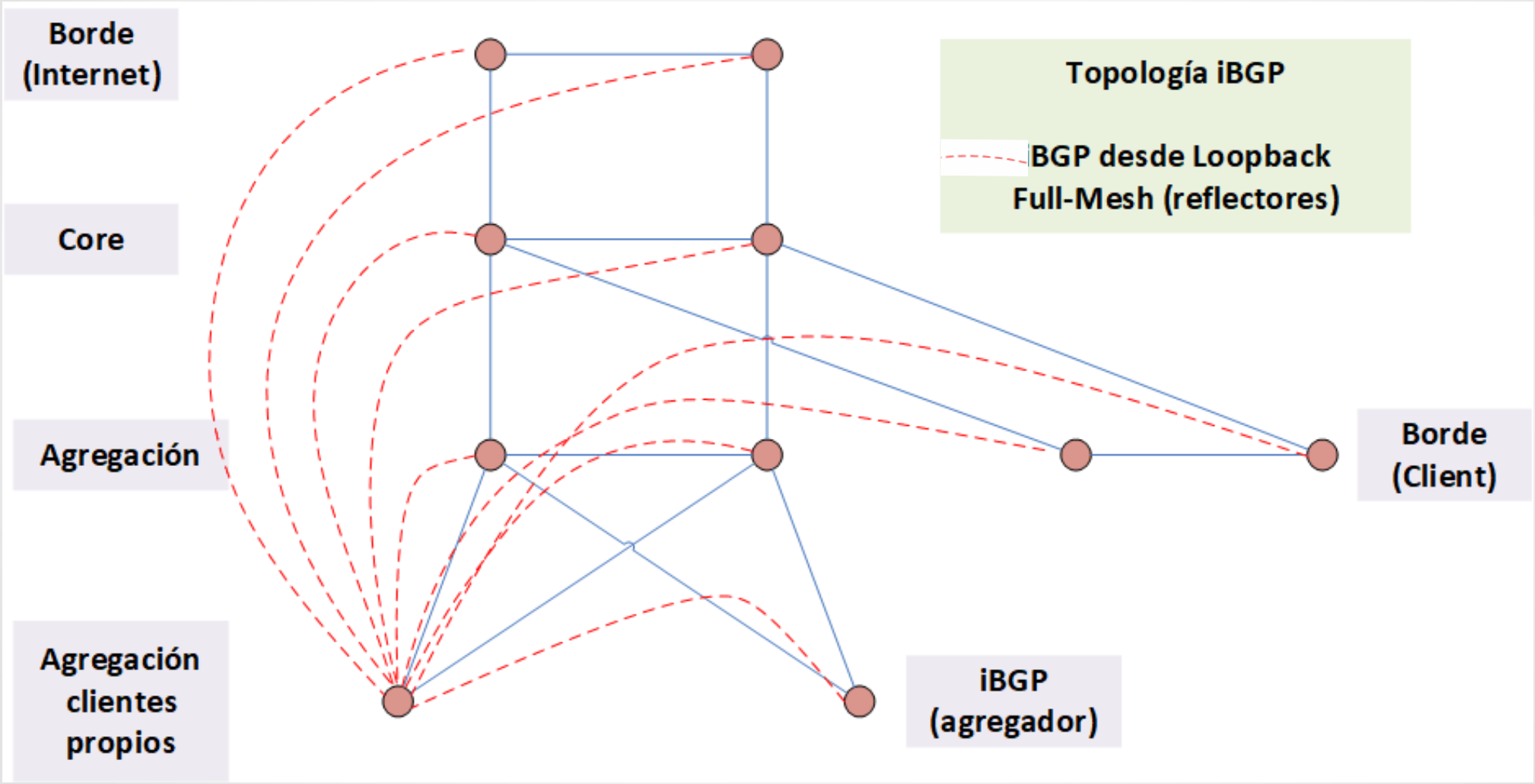
IGP y BGP - Best Practise SP (III)



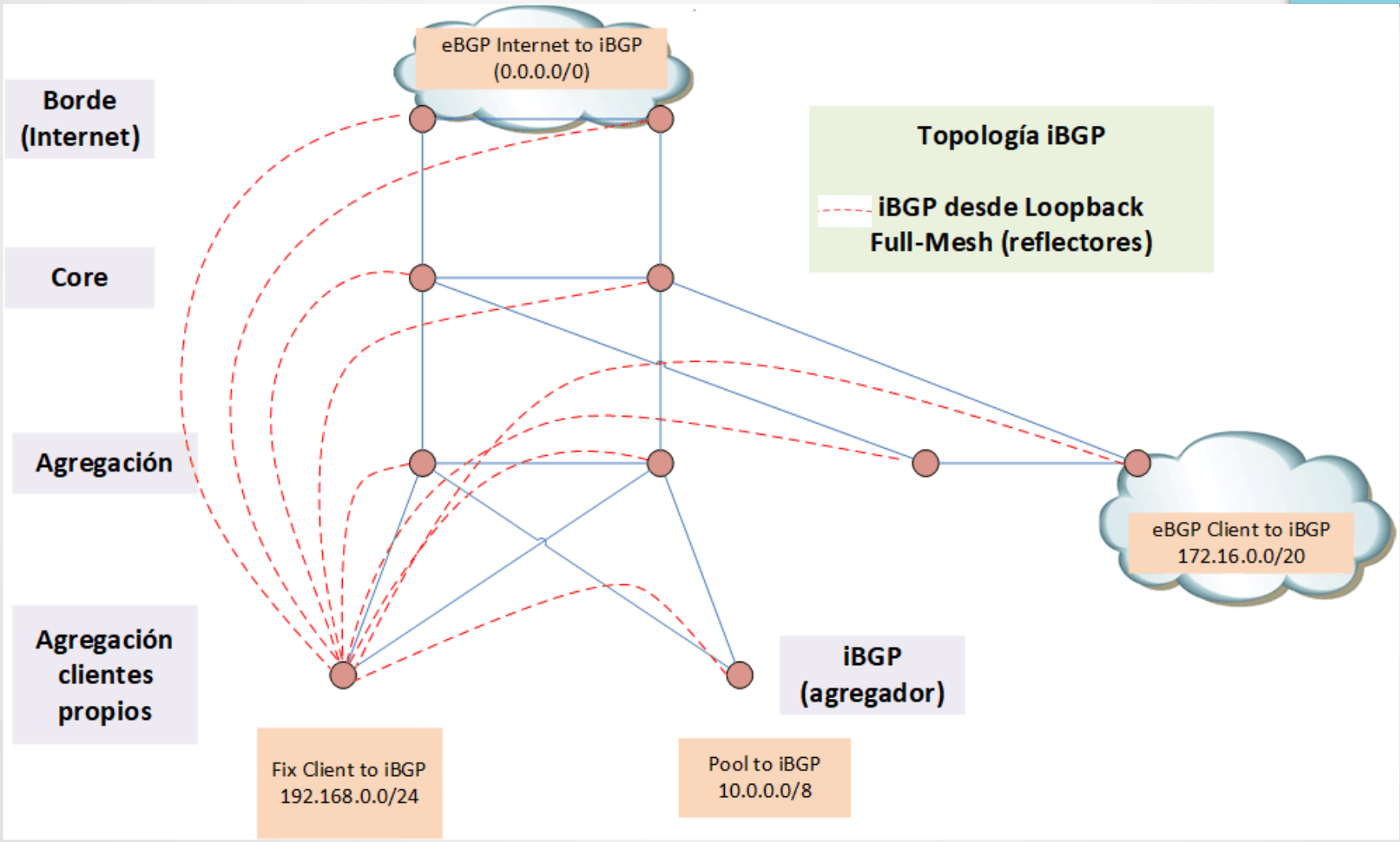
IGP y BGP - Best Practise (IV)



IGP y BGP - Best Practise (V)



IGP y BGP - Best Practise (VI)



Referencias (RFC)

- RFC 4271 – RFC 1771 – BGP
- RFC 1772 – BGP – Aplicación
- RFC 2385 – MD5 Signature
- RFC 2439 – Route Flat Damping
- RFC 2545 – BGP Multiprotocol Extension for IPv6
- RFC 2858 – Multiprotocol Extension for BGPv4
- RFC 2858 – Route Refresh Capability for BGPv4
- RFC 4360 – Extended Comunity Attribute
- RFC 4456 - Route Reflectors

Referencias (RFC)

- RFC 4893 – AS 32 bits
- RFC 5065 – Confederaciones
- RFC 5668 – 4-Octet AS Specific BGP Extended Community
- RFC 6472 – Recommendation for Not Using AS_SET and AS_CONFED_SET in BGP
- RFC 7454 – BGP Operations and Security
- RFC 7911 – Advertisement of Multiple Paths in BGP

Referencias

- http://www1.cs.columbia.edu/~ji/F02/
- Internet Routing Architectures. Sam Halabi. Cisco Press
- Designing for CISCO Networks Service Architectures, Al-shawi & Laurent, CISCO Press.
- Bulding Reliable Networks with Border Gateway Protocol, van Beijnum, O'Reilly
- Varias especificadas en las transparencias
- Documentación variada de Cisco