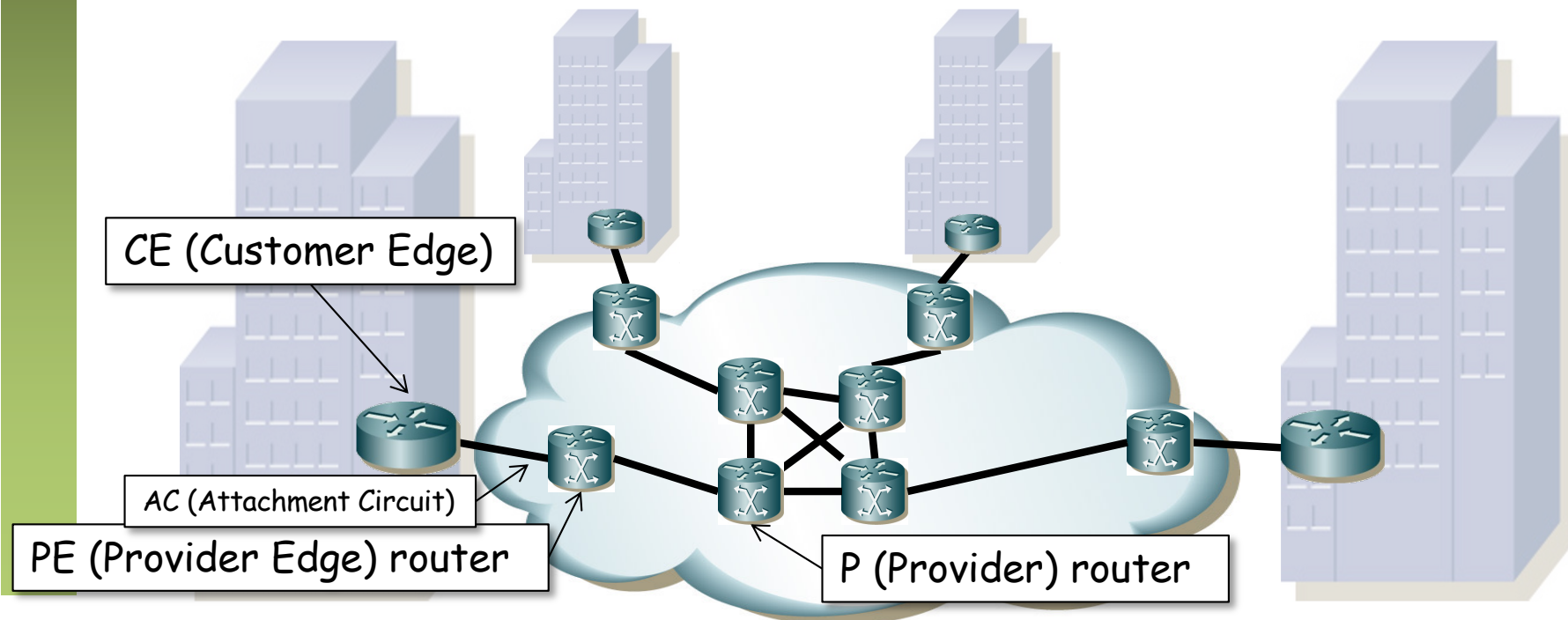


Layer 3 MPLS VPNs

Layer 3 VPNs

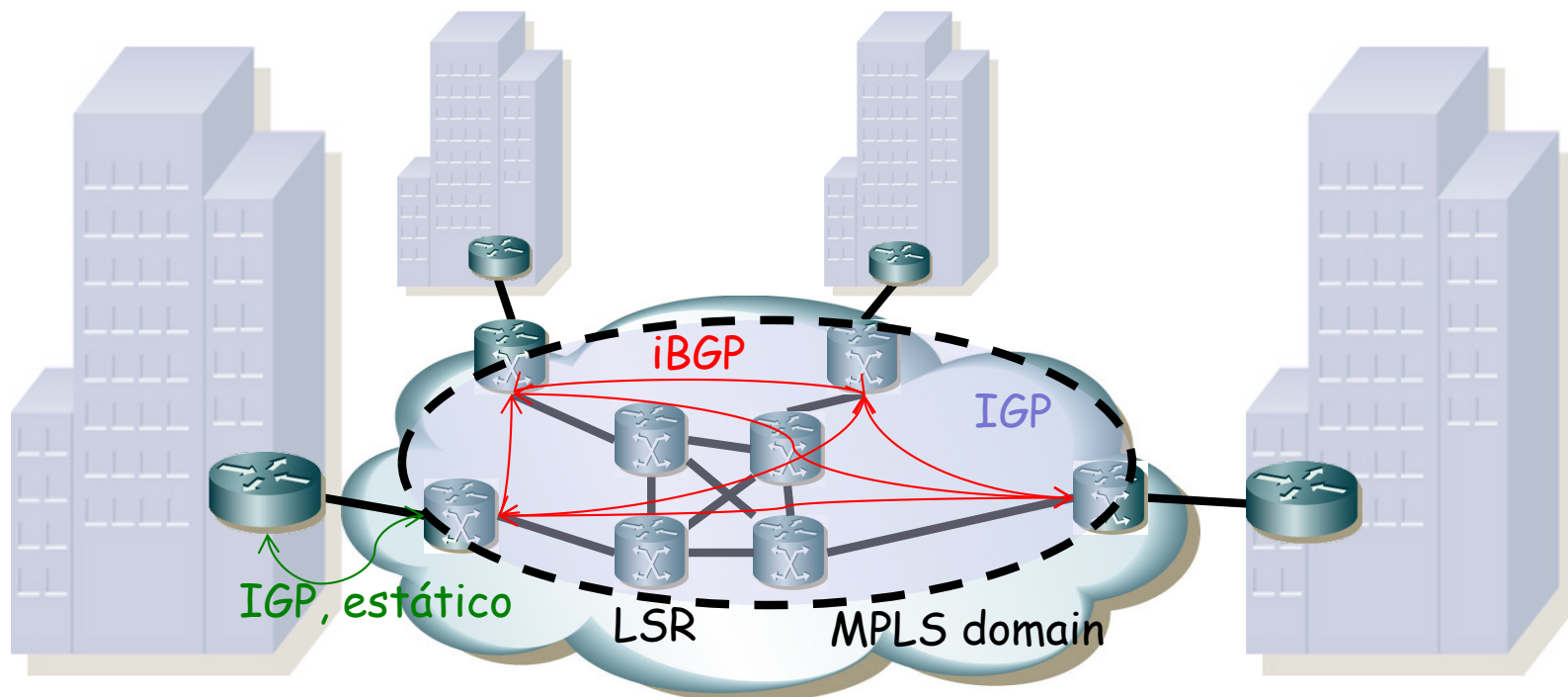
- RFC 4364 “BGP/MPLS IP Virtual Private Networks (VPNs)” (Cisco Systems y Juniper Networks, 2006)
 - *“This document describes a method by which a Service Provider may use an IP backbone to provide IP Virtual Private Networks (VPNs) for its customers.”*
- VPN para el transporte de paquetes IP entre sedes (*sites*)
- El backbone del proveedor de servicio es una red IP MPLS
- RFC 4760 “Multiprotocol Extensions for BGP-4” (Cisco, Sanoa, Juniper, 2007)
- Extensiones a BGP-4 para poder transportar información de otros protocolos de nivel de red: IPv6, IPX, L3VPN, etc
- En este caso, en lugar de transportar rutas IPv4 transportará rutas “VPN-IPv4”



L3VPN - Routing

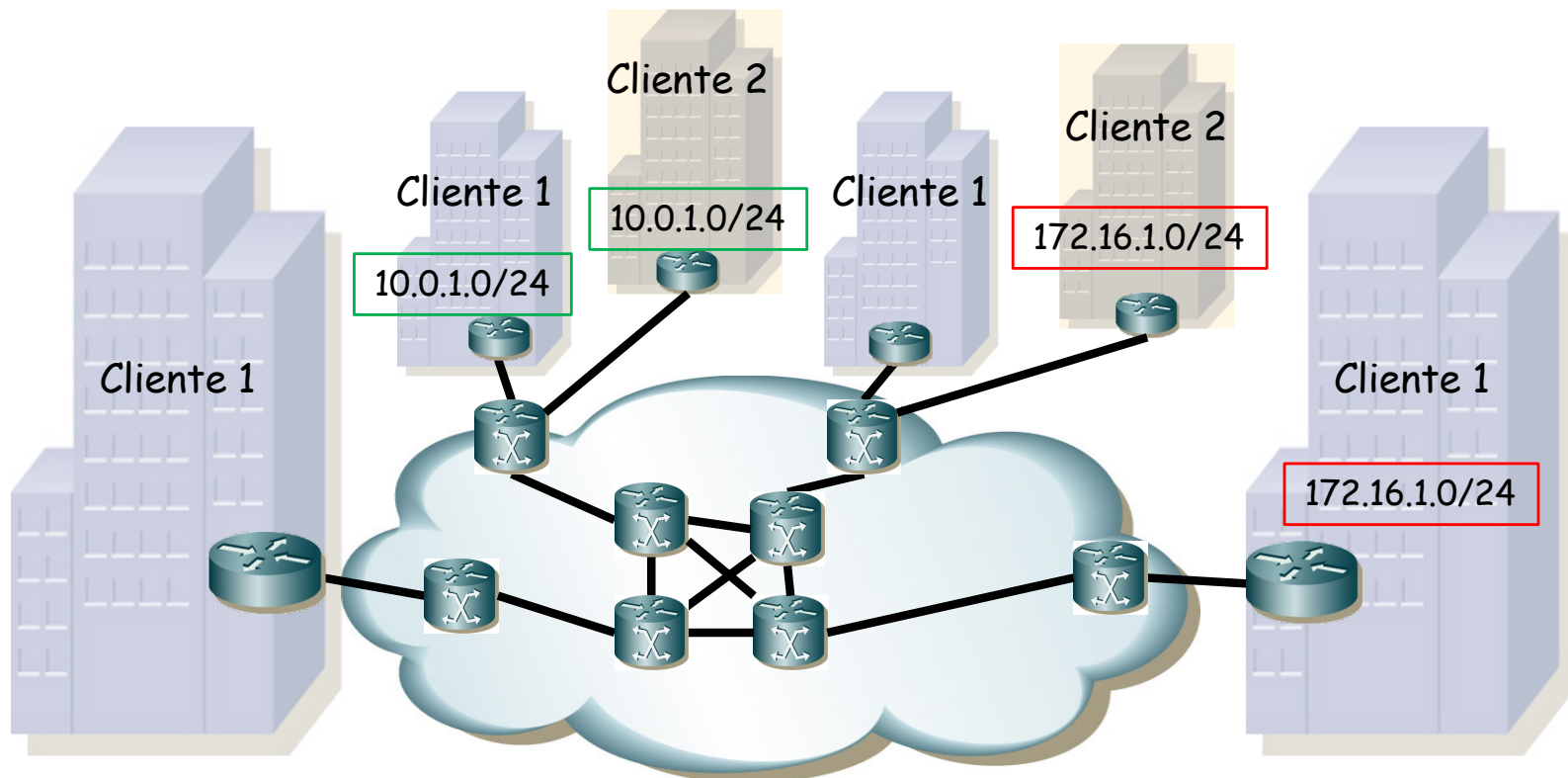
L3VPN: Routing

- Los CEs anuncian sus rutas a los PEs (con un IGP o rutas estáticas)
- Los PEs emplean MP-BGP para intercambiarse esas rutas (iBGP)
- El PE la distribuye al CE del mismo cliente (de la misma VPN)
- Los P y PE corren un IGP para tener alcanzabilidad interna
- Los CE son routers convencionales, no necesitan ninguna configuración de VPN ni emplean MPLS
- Los CEs no intercambian información de routing entre ellos, no son adyacentes
- La VPN no actúa como un overlay sino una red IP con otro gestor



L3VPN: Routing

- Dos VPNs pueden emplear espacios de direcciones IP que se solapan
- Los anuncios VPN-IPv4 (MP-BGP) incluyen un identificador (*Route Distinguisher* = *RD*, 64 bits) que las diferencia
- Anuncio de la AF (Address Family) VPN-IPv4 es un prefijo de 12 bytes:
`<Route Distinguisher>:<IP Prefix>`
- Permite también anunciar más de una vez el mismo prefijo IPv4 para el mismo cliente empleando distinto RD (múltiples rutas para el mismo destino)

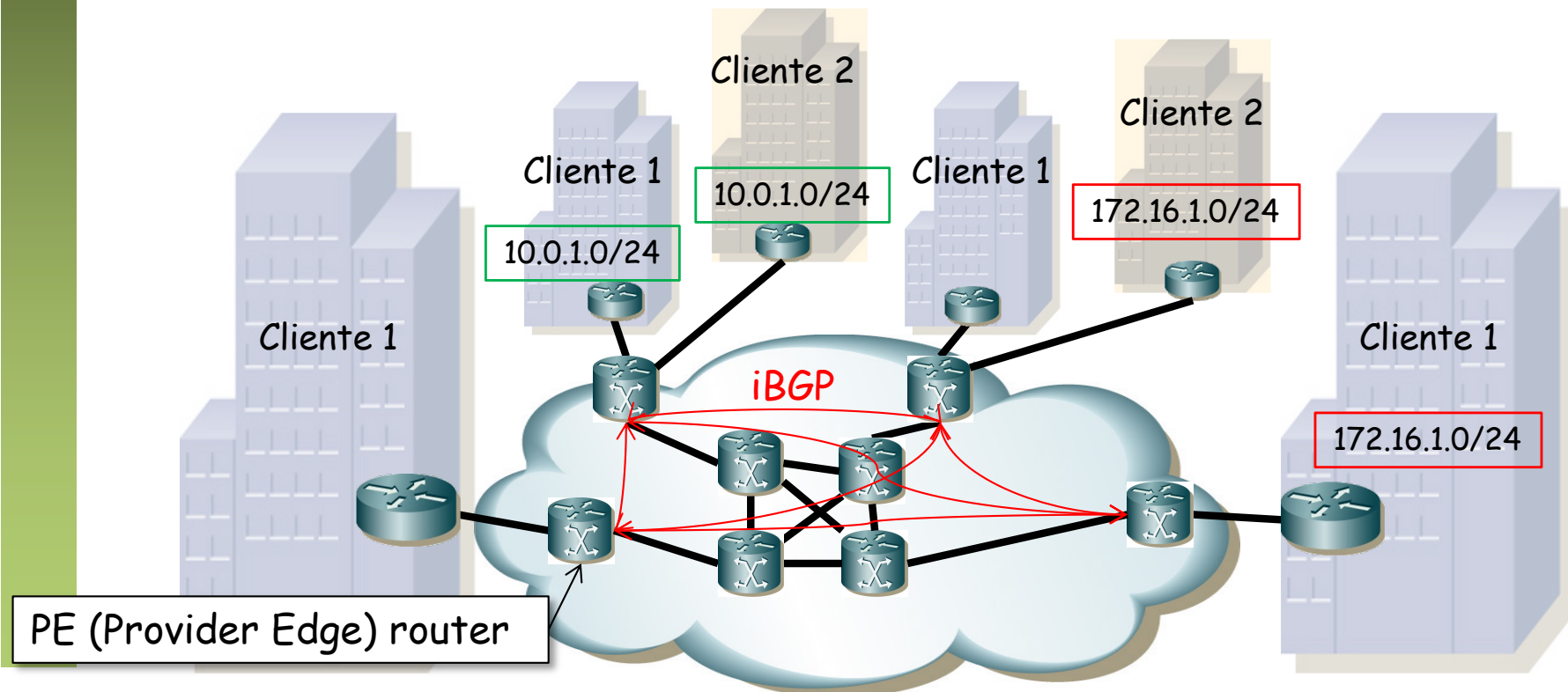


Ejemplo en mensaje BGP

- ▶ Frame 84: 203 bytes on wire (1624 bits), 203 bytes captured (1624 bits)
- ▶ Ethernet II, Src: Cisco_9d:ff:02 (a0:3d:6f:9d:ff:02), Dst: Cisco_cb:84:02 (a0:e0:af:cb:84:02)
- ▶ MultiProtocol Label Switching Header, Label: 23, Exp: 0, S: 1, TTL: 255
- ▶ Internet Protocol Version 4, Src: 10.100.100.6, Dst: 10.100.100.17
- ▶ Transmission Control Protocol, Src Port: 179, Dst Port: 52361, Seq: 104, Ack: 85, Len: 141
- ▼ Border Gateway Protocol – UPDATE Message
 - Marker: ffffffffffffffffffffffffffffffffff
 - Length: 112
 - Type: UPDATE Message (2)
 - Withdrawn Routes Length: 0
 - Total Path Attribute Length: 89
 - ▼ Path attributes
 - ▼ Path Attribute – MP_REACH_NLRI
 - ▶ Flags: 0x80, Optional, Non-transitive, Complete
 - Type Code: MP_REACH_NLRI (14)
 - Length: 48
 - Address family identifier (AFI): IPv4 (1)
 - Subsequent address family identifier (SAFI): Labeled VPN Unicast (128)
 - ▶ Next hop network address (12 bytes)
 - Number of Subnetwork points of attachment (SNPA): 0
 - ▼ Network layer reachability information (31 bytes)
 - ▼ BGP Prefix
 - Prefix Length: 112
 - Label Stack: 28 (bottom)
 - Route Distinguisher: 14:14
 - MP Reach NLRI IPv4 prefix: 10.1.6.0
 - ▼ BGP Prefix
 - Prefix Length: 120
 - Label Stack: 27 (bottom)
 - Route Distinguisher: 14:14
 - MP Reach NLRI IPv4 prefix: 10.100.100.1

L3VPN: Routing

- RD no tiene significado pero un formato habitual es <ASN>:<id>
- Así cada service provider tiene su espacio de valores RD
- Las sesiones iBGP son entre los PE, así que ...
- Los P no ven las rutas de las VPNs (evita problemas de escalabilidad)
- ¿Cómo se enruta si hay direcciones duplicadas y los routers centrales no ven esas rutas?

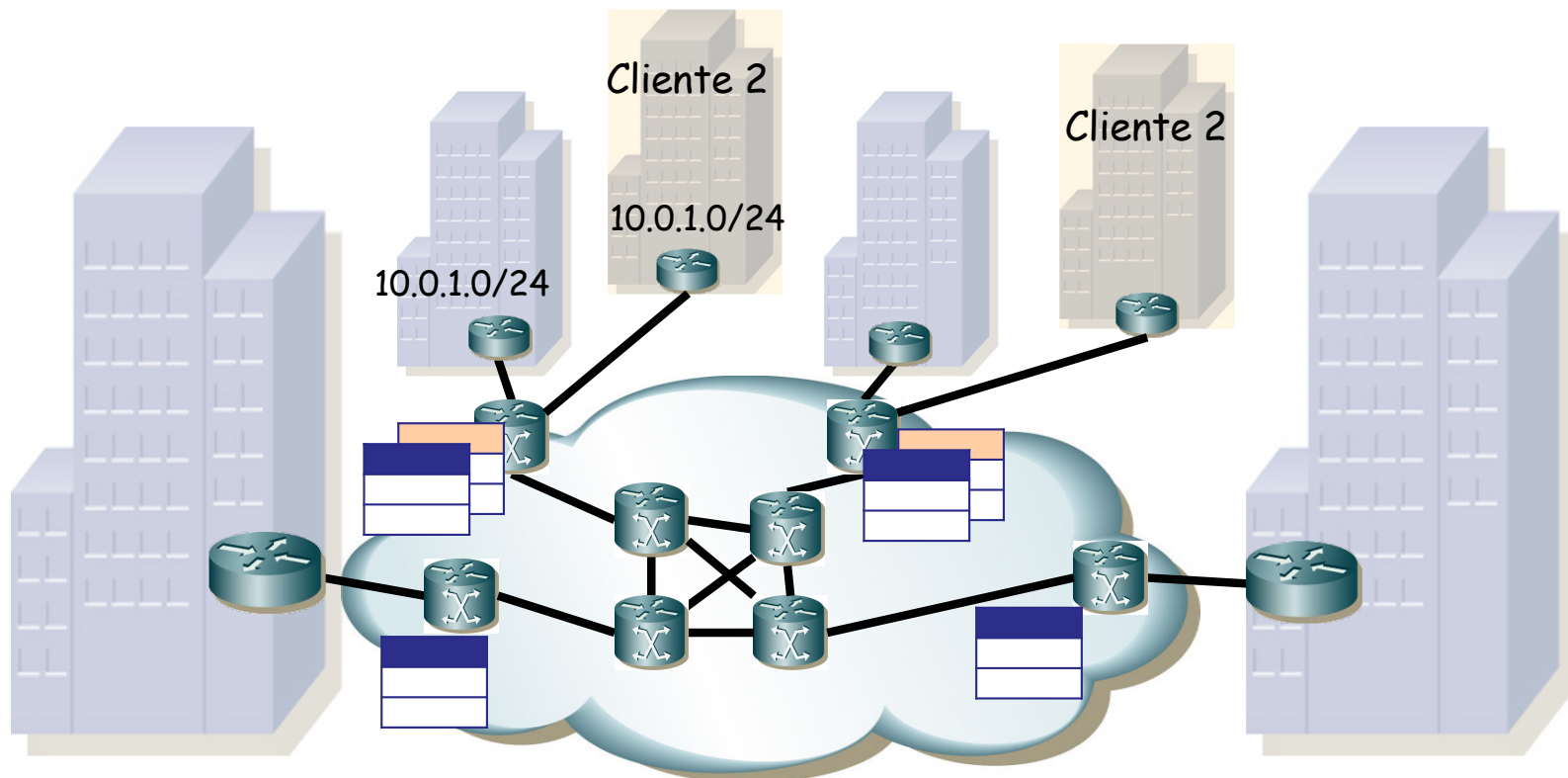


L3VPN - Routing

L3VPN - Forwarding

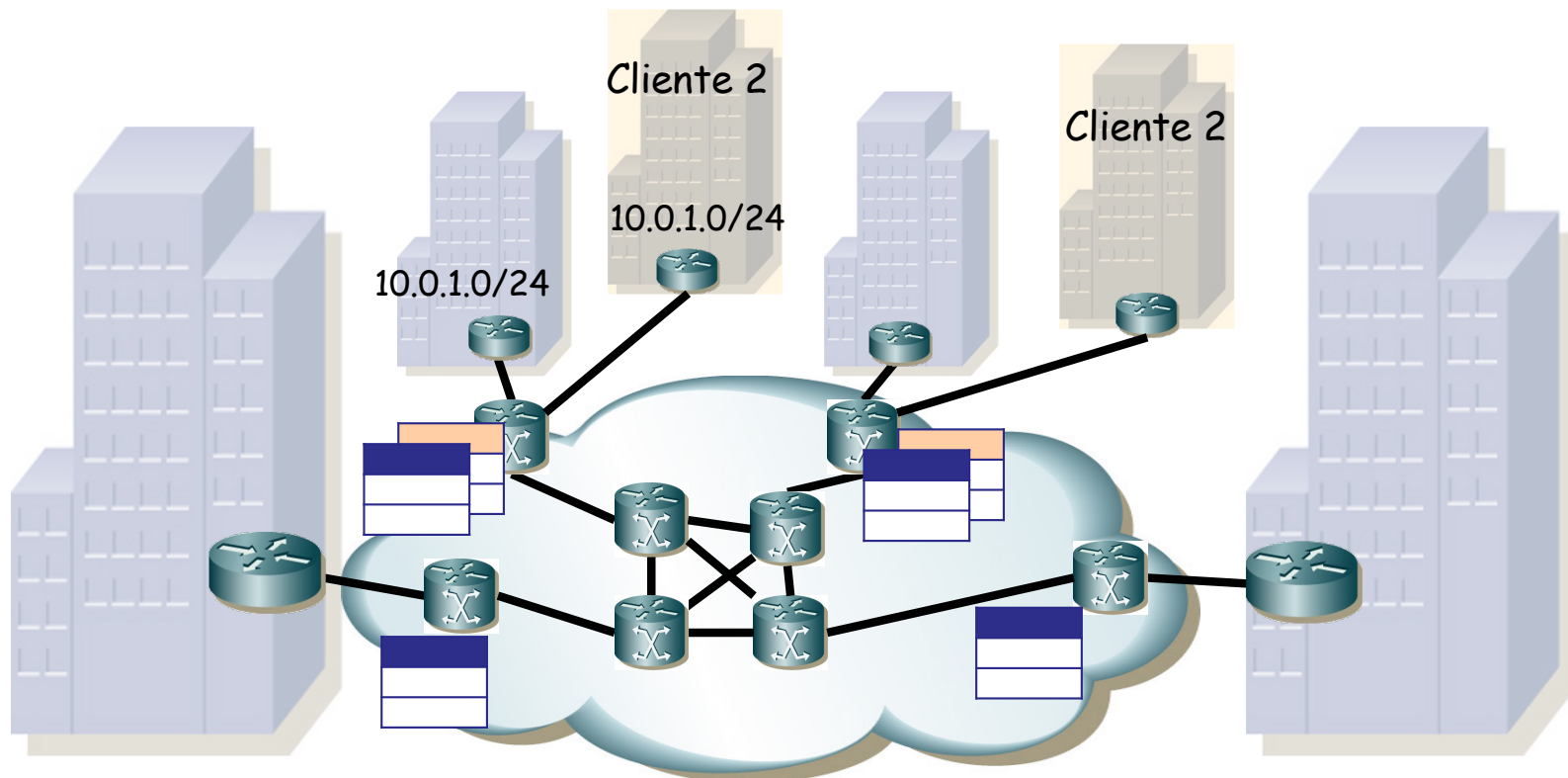
L3VPN: Forwarding

- Cada PE mantiene una tabla de rutas para cada cliente y además una tabla por defecto
- VPN o *VPN/Virtual Routing and Forwarding tables (VRFs)*
- Al recibir un paquete IP de un cliente consulta la VRF correspondiente



L3VPN: Routing

- Los anuncios MP-BGP de la AF VPN-IPv4 incluyen uno o más “*Route Targets*” (RT)
- Una VRF importa las rutas que traen los RT que desee
- El RT se transporta como una *extended community* (RFC 4360)
- Las comunidades son atributos opcionales transitivos (32 ó 64 bits)
- Son una forma de incluir una etiqueta numérica que quien la vea debe saber interpretar

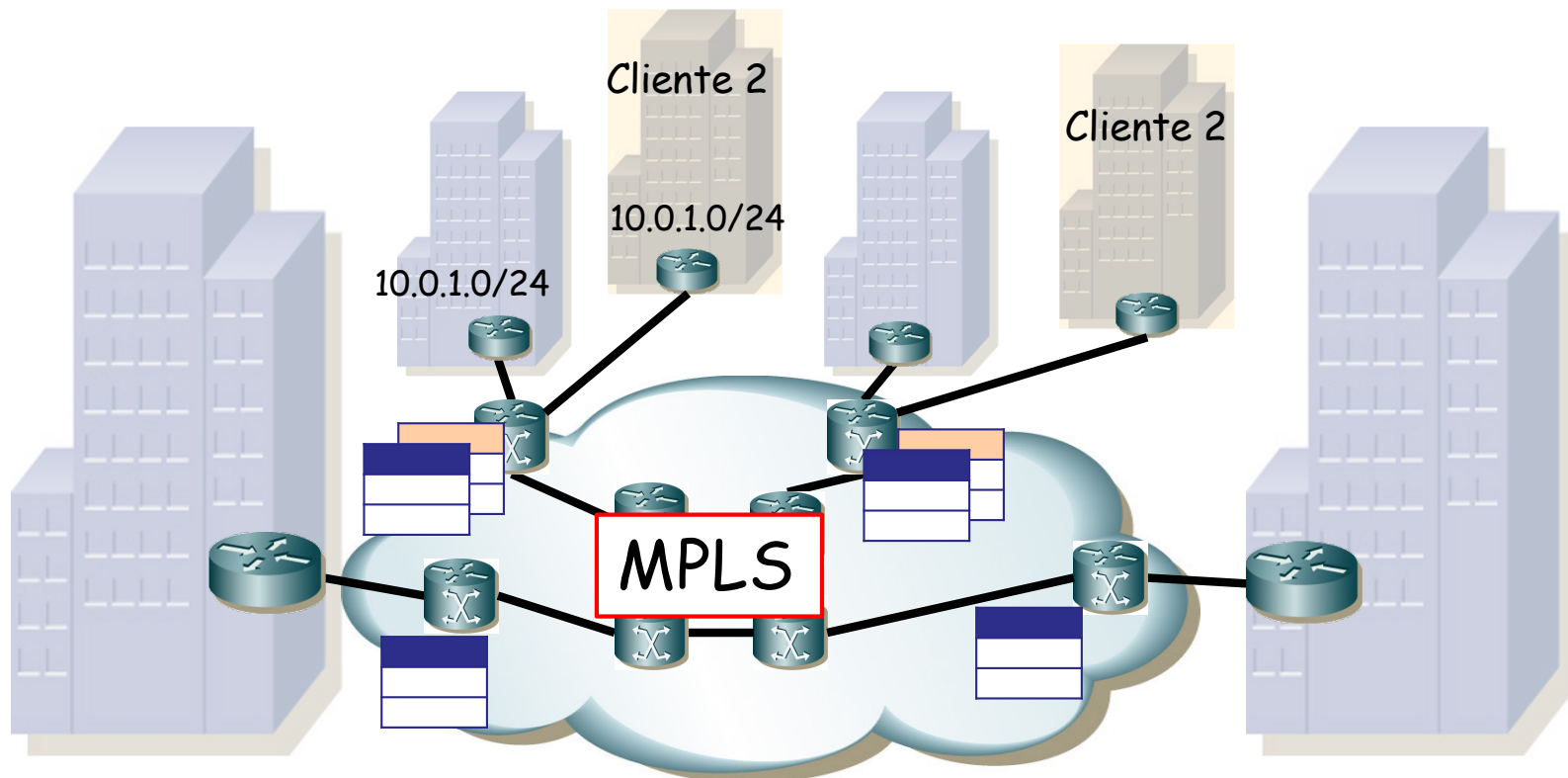


Ejemplo en mensaje BGP

- ▶ Frame 84: 203 bytes on wire (1624 bits), 203 bytes captured (1624 bits)
- ▶ Ethernet II, Src: Cisco_9d:ff:02 (a0:3d:6f:9d:ff:02), Dst: Cisco_cb:84:02 (a0:e0:af:cb:84:02)
- ▶ MultiProtocol Label Switching Header, Label: 23, Exp: 0, S: 1, TTL: 255
- ▶ Internet Protocol Version 4, Src: 10.100.100.6, Dst: 10.100.100.17
- ▶ Transmission Control Protocol, Src Port: 179, Dst Port: 52361, Seq: 104, Ack: 85, Len: 141
- ▼ Border Gateway Protocol – UPDATE Message
 - Marker: ffffffffffffffffffffffffffffffffff
 - Length: 112
 - Type: UPDATE Message (2)
 - Withdrawn Routes Length: 0
 - Total Path Attribute Length: 89
 - ▼ Path attributes
 - ▶ Path Attribute – MP_REACH_NLRI
 - ▶ Path Attribute – ORIGIN: INCOMPLETE
 - ▶ Path Attribute – AS_PATH: 65000
 - ▶ Path Attribute – MULTI_EXIT_DISC: 0
 - ▶ Path Attribute – LOCAL_PREF: 100
 - ▼ Path Attribute – EXTENDED_COMMUNITIES
 - ▶ Flags: 0xc0, Optional, Transitive, Complete
 - Type Code: EXTENDED_COMMUNITIES (16)
 - Length: 8
 - ▼ Carried extended communities: (1 community)
 - ▼ Route Target: 14:14 [Transitive 2-Octet AS-Specific]
 - ▶ Type: Transitive 2-Octet AS-Specific (0x00)
 - Subtype (AS2): Route Target (0x02)
 - 2-Octet AS: 14
 - 4-Octet AN: 14

L3VPN: Routing + Forwarding

- RD:prefijo y RT son plano de control
- El anuncio MP-BGP incluye una etiqueta MPLS
- Esta etiqueta es para el plano de datos



Ejemplo en mensaje BGP

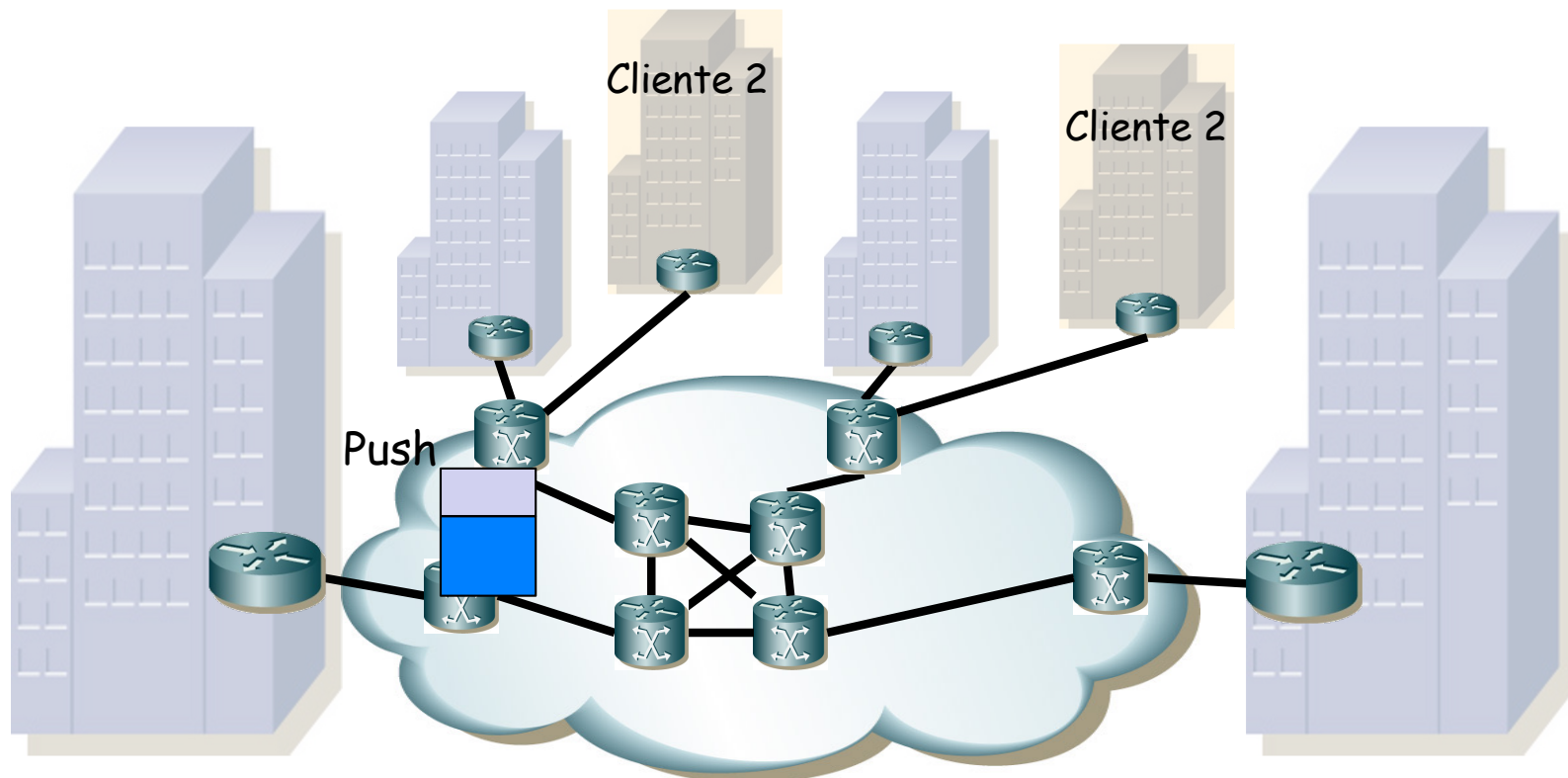
```
▶ Frame 84: 203 bytes on wire (1624 bits), 203 bytes captured (1624 bits)
▶ Ethernet II, Src: Cisco_9d:ff:02 (a0:3d:6f:9d:ff:02), Dst: Cisco_cb:84:02 (a0:e0:af:cb:84:02)
▶ MultiProtocol Label Switching Header, Label: 23, Exp: 0, S: 1, TTL: 255
▶ Internet Protocol Version 4, Src: 10.100.100.6, Dst: 10.100.100.17
▶ Transmission Control Protocol, Src Port: 179, Dst Port: 52361, Seq: 104, Ack: 85, Len: 141
▼ Border Gateway Protocol – UPDATE Message
  Marker: ffffffffffffffffffffffffffffffffff
  Length: 112
  Type: UPDATE Message (2)
  Withdrawn Routes Length: 0
  Total Path Attribute Length: 89
  ▼ Path attributes
    ▼ Path Attribute – MP_REACH_NLRI
      ▶ Flags: 0x80, Optional, Non-transitive, Complete
      Type Code: MP_REACH_NLRI (14)
      Length: 48
      Address family identifier (AFI): IPv4 (1)
      Subsequent address family identifier (SAFI): Labeled VPN Unicast (128)
      ▶ Next hop network address (12 bytes)
      Number of Subnetwork points of attachment (SNPA): 0
      ▼ Network layer reachability information (31 bytes)
        ▼ BGP Prefix
          Prefix Length: 112
          Label Stack: 28 (bottom)
          Route Distinguisher: 14:14
          MP Reach NLRI IPv4 prefix: 10.1.6.0
        ▼ BGP Prefix
          Prefix Length: 120
          Label Stack: 27 (bottom)
          Route Distinguisher: 14:14
          MP Reach NLRI IPv4 prefix: 10.100.100.1
```

24 bits Label
64 bits RD
32 bits dirección de red

Total: 120 bits
Prefix Length: 112
120-112 = 8 bits (es decir /24)

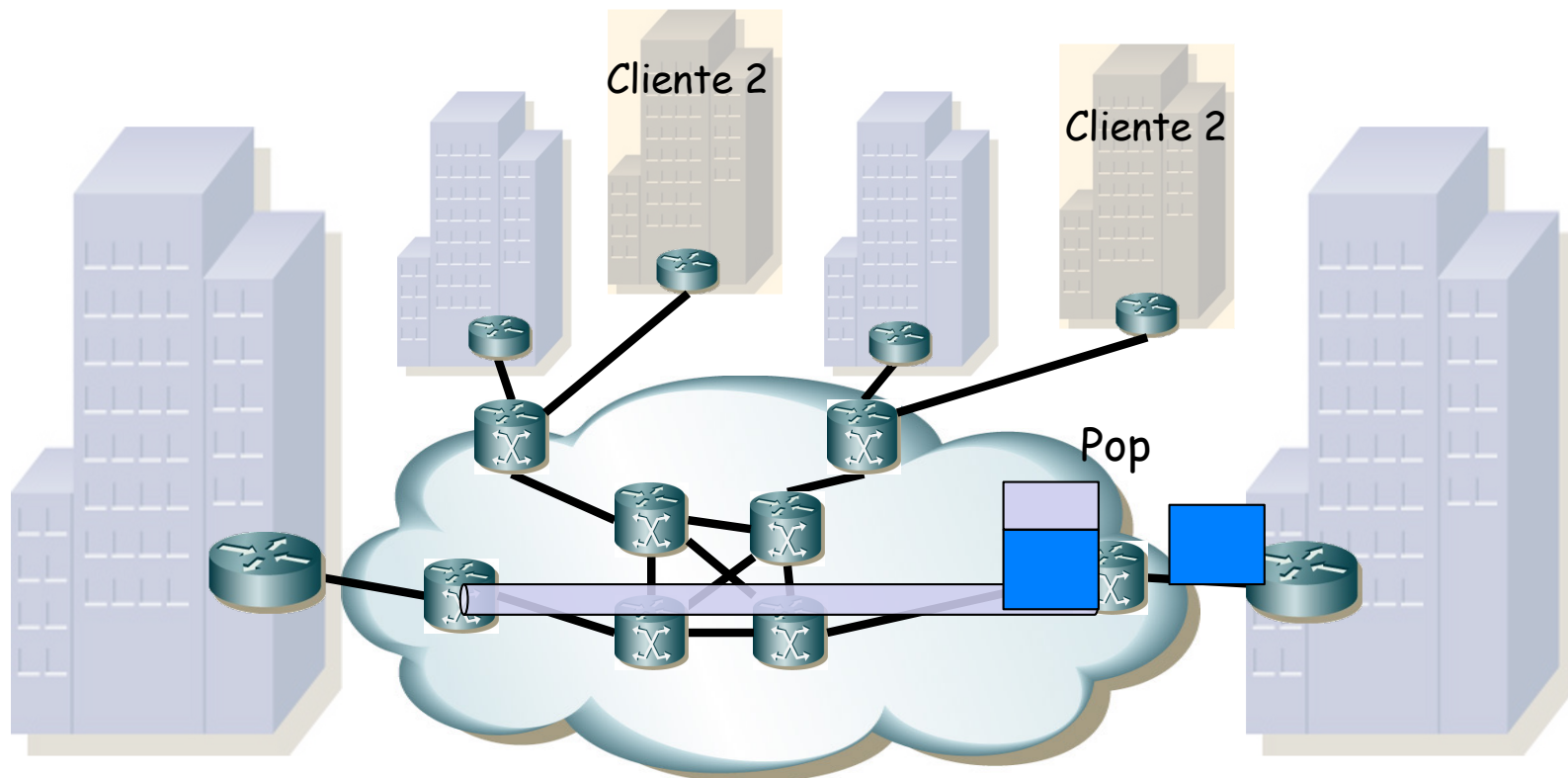
L3VPN: MPLS

- ¿Para qué esa etiqueta?
- Asociada a la ruta
- Paquetes de la VRF que siguen esa ruta se introducen en LSP con esa etiqueta (*VPN route label*)
- (...)



L3VPN: MPLS

- Al salir el paquete en el PE de egreso se sabe a qué VRF pertenece en base a ese LSP en el que venía
- No puede usar la dirección IP destino ya que pueden estar duplicadas
- El paquete MPLS viene con la etiqueta que puso el de ingreso
- Es decir, PE de egreso hizo el anuncio de la ruta con la etiqueta, PE de ingreso la añadió y el paquete llegó con ella
- ¿No cambiaron la etiqueta los LSR del camino?

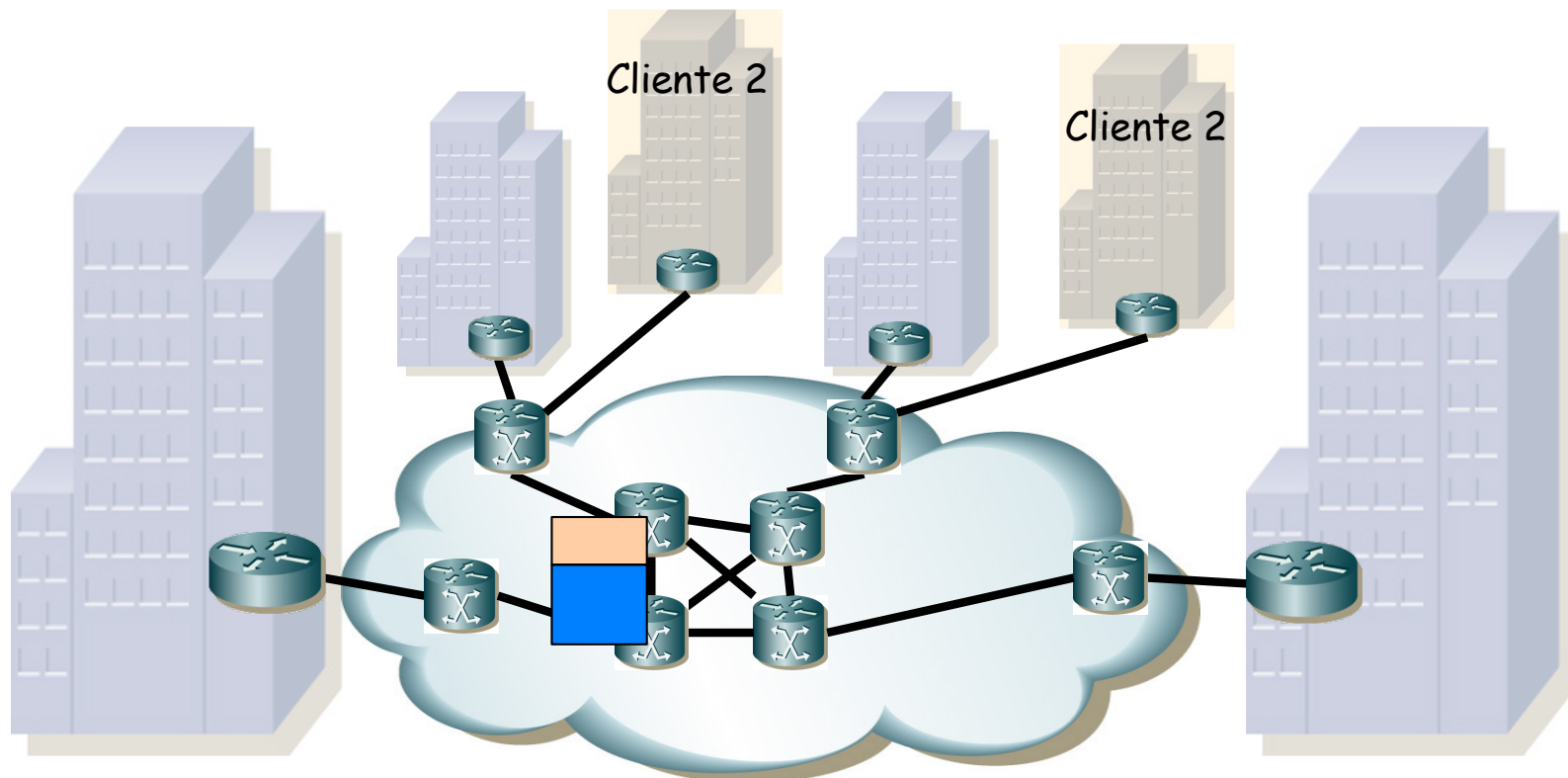


L3VPN - Forwarding

L3VPN – Forwarding en el Core

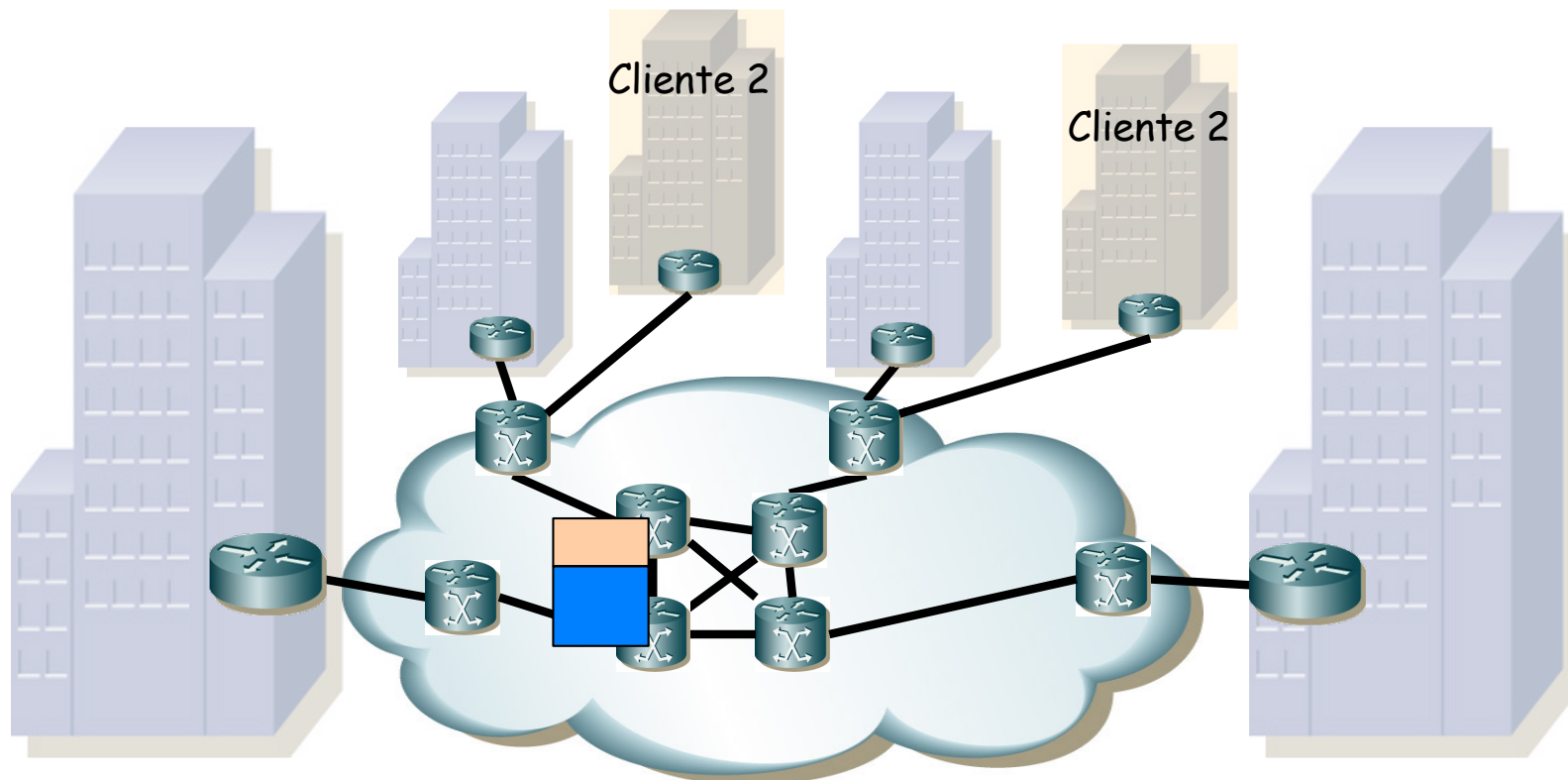
L3VPN: MPLS

- ¿Y en los P routers? ¿Reenvían en función de esa etiqueta?
- (...)



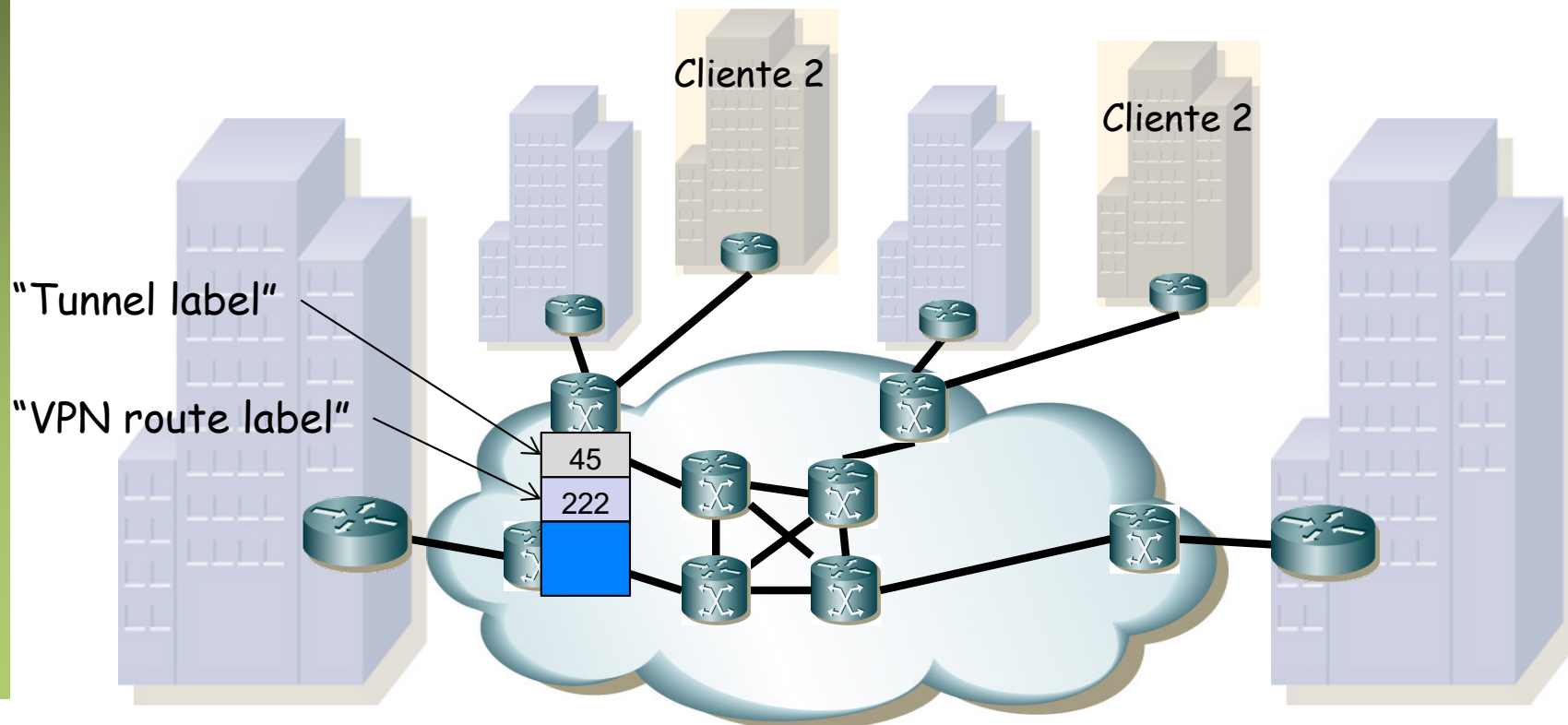
L3VPN: MPLS

- ¿Y en los P routers? ¿Reenvían en función de esa etiqueta?
- Tendríamos en ellos una gran cantidad de LSPs, para todas las VPNs
- Mala escalabilidad
- (...)



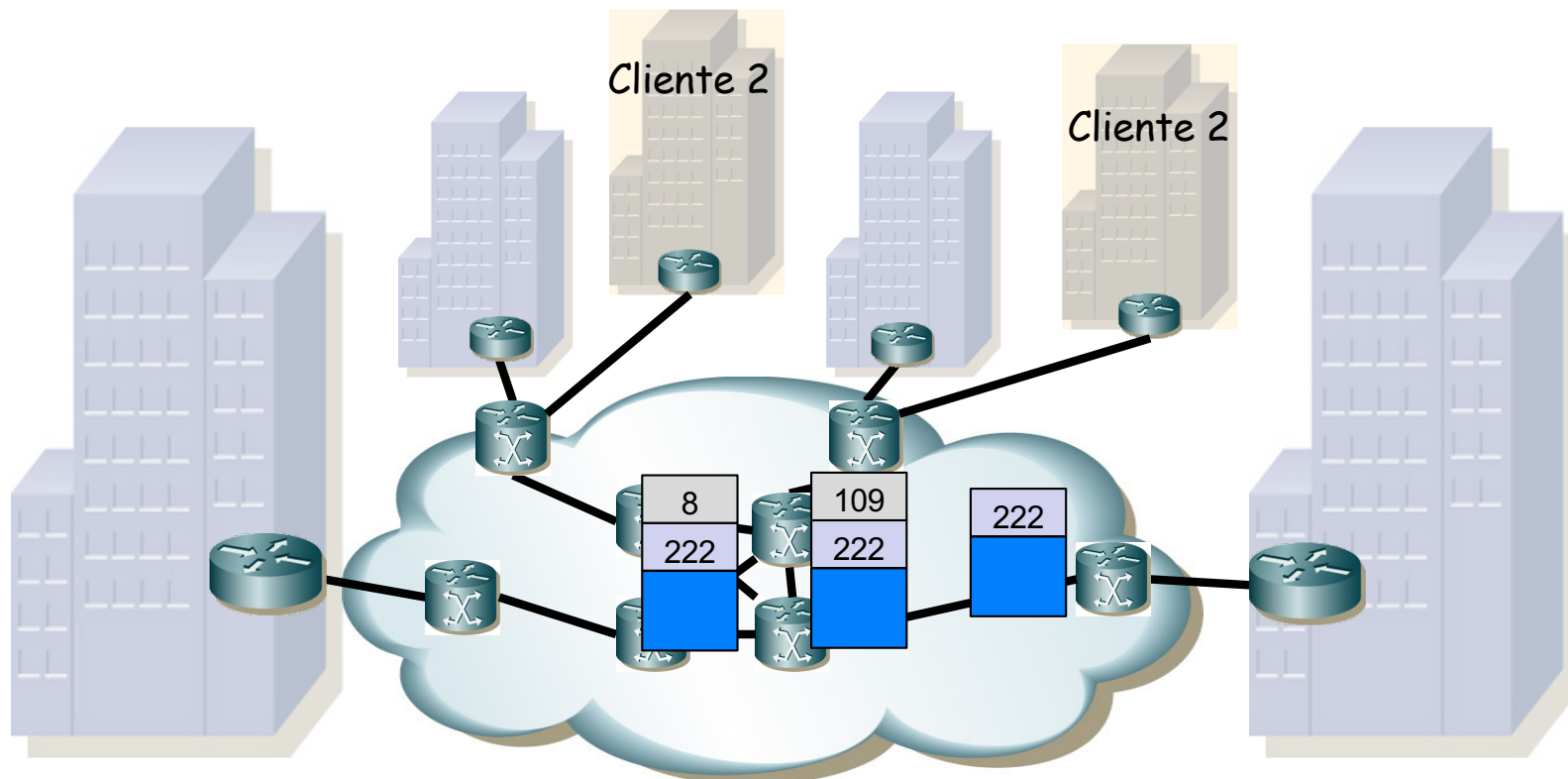
L3VPN: MPLS

- ¿Y en los P routers? ¿Reenvían en función de esa etiqueta?
- Se crean LSPs entre los PEs (“Tunnel LSPs”)
- Los P routers reenvían en función de esa etiqueta externa (...)



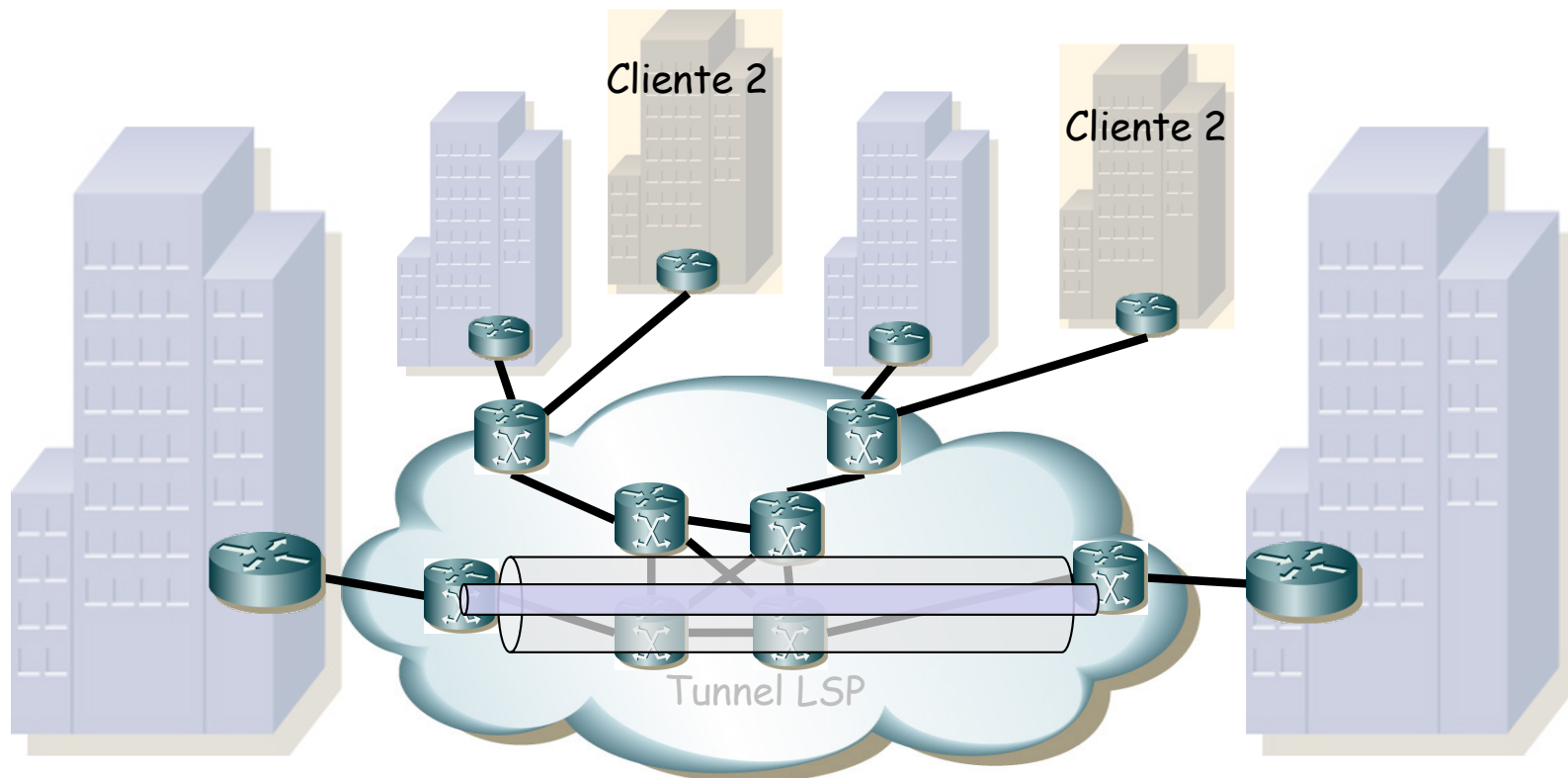
L3VPN: MPLS

- ¿Y en los P routers? ¿Reenvían en función de esa etiqueta?
- Se crean LSPs entre los PEs (“Tunnel LSPs”)
- Los P routers reenvían en función de esa etiqueta externa
- (...)

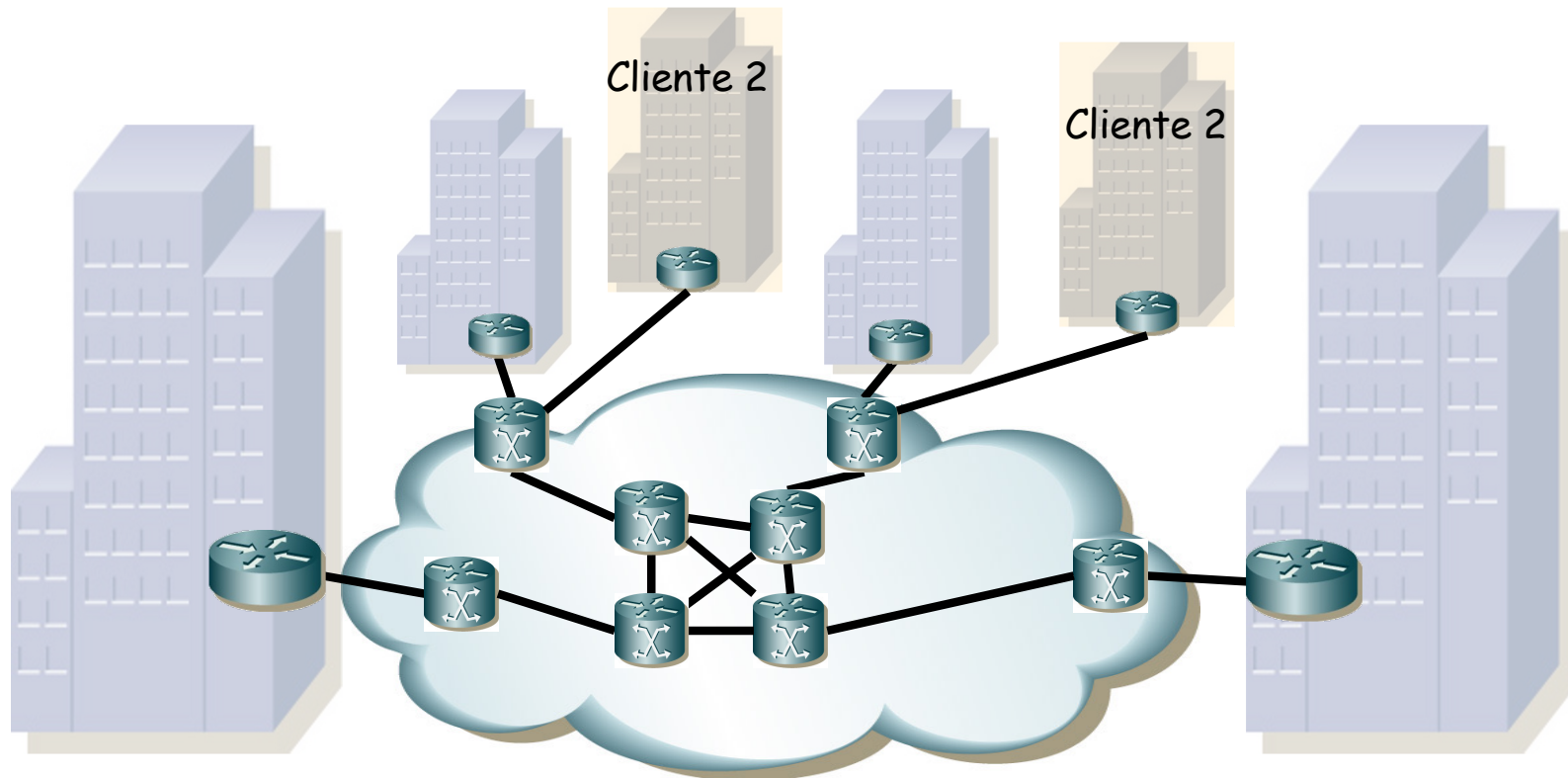


L3VPN: MPLS

- ¿Y en los P routers? ¿Reenvían en función de esa etiqueta?
- Se crean LSPs entre los PEs (“Tunnel LSPs”)
- Los P routers reenvían en función de esa etiqueta externa
- Un full-mesh entre los PEs que compartan VRF

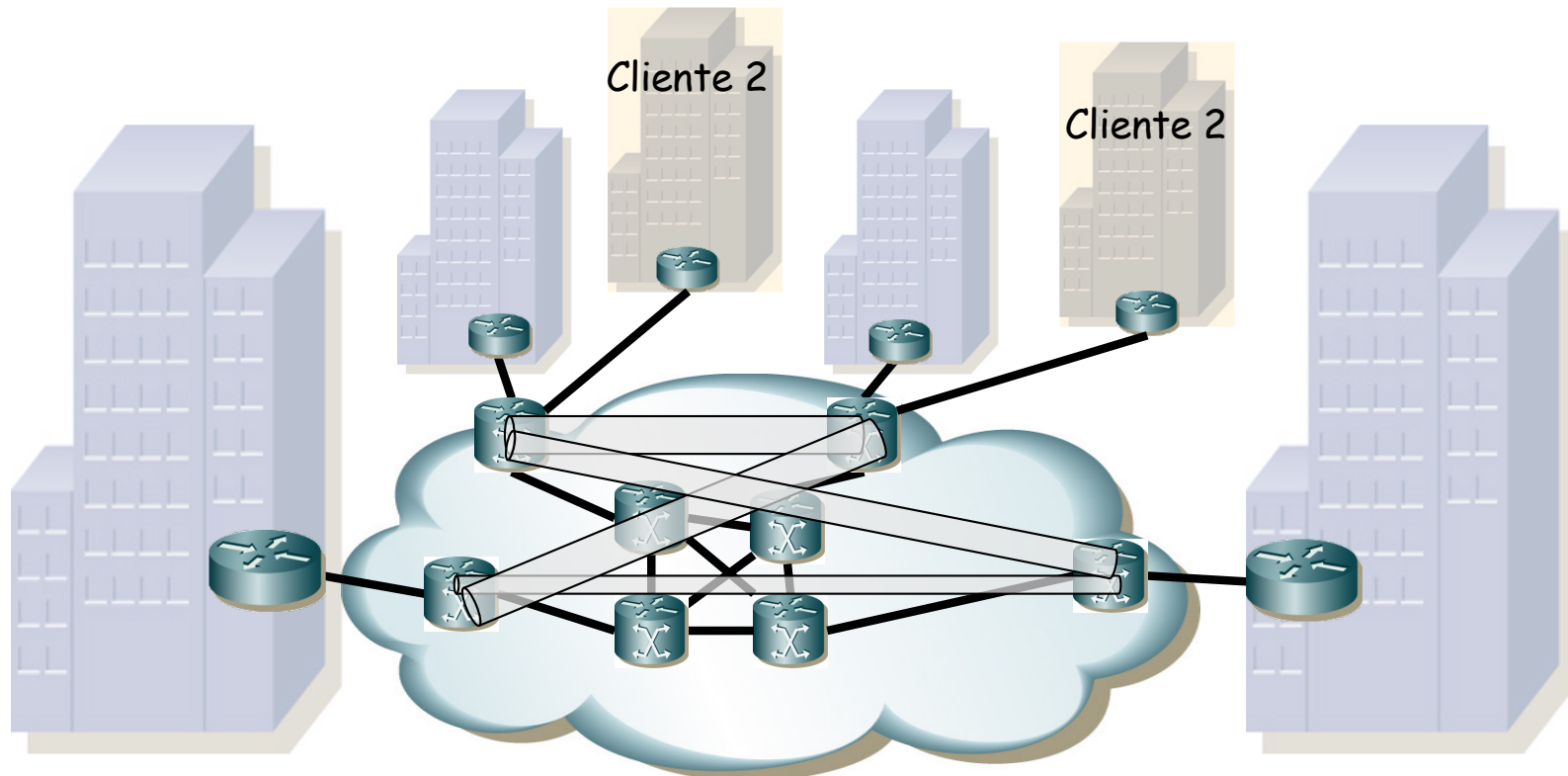


Recapitulando



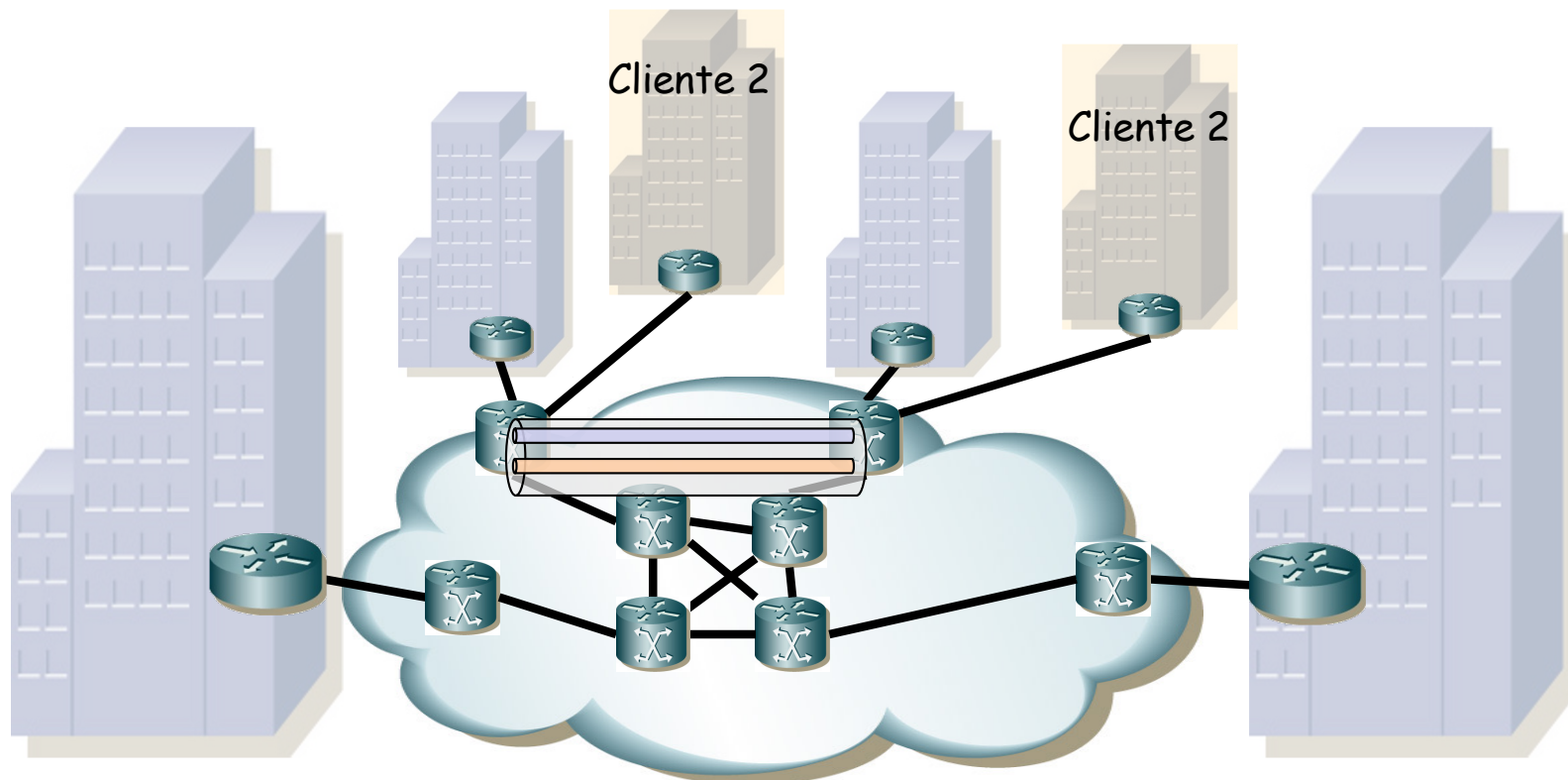
Recapitulando

- Un LSP entre cada pareja de PEs



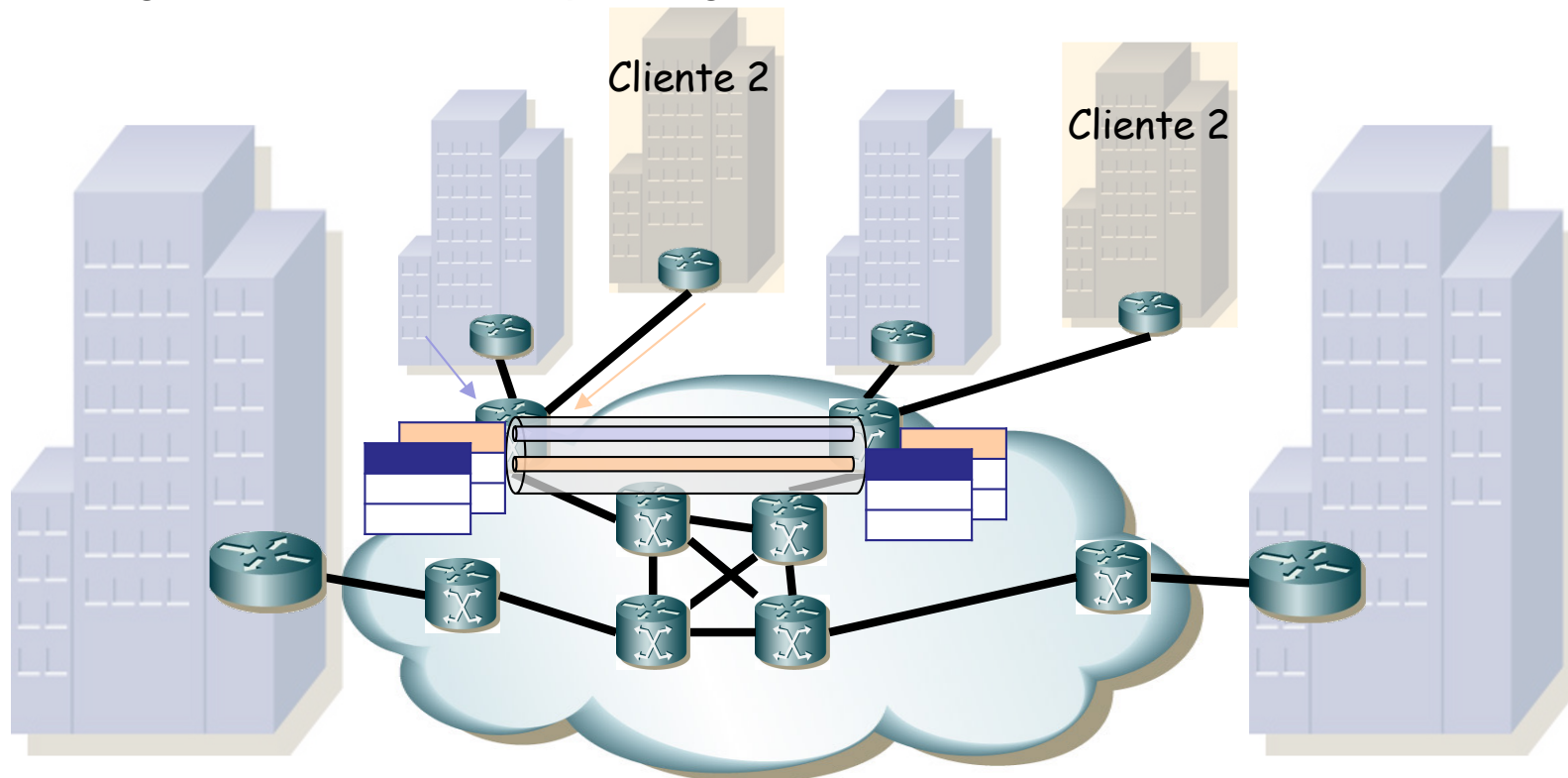
Recapitulando

- Un LSP entre cada pareja de PEs
- Hacen de túnel para los LSPs de cada cliente



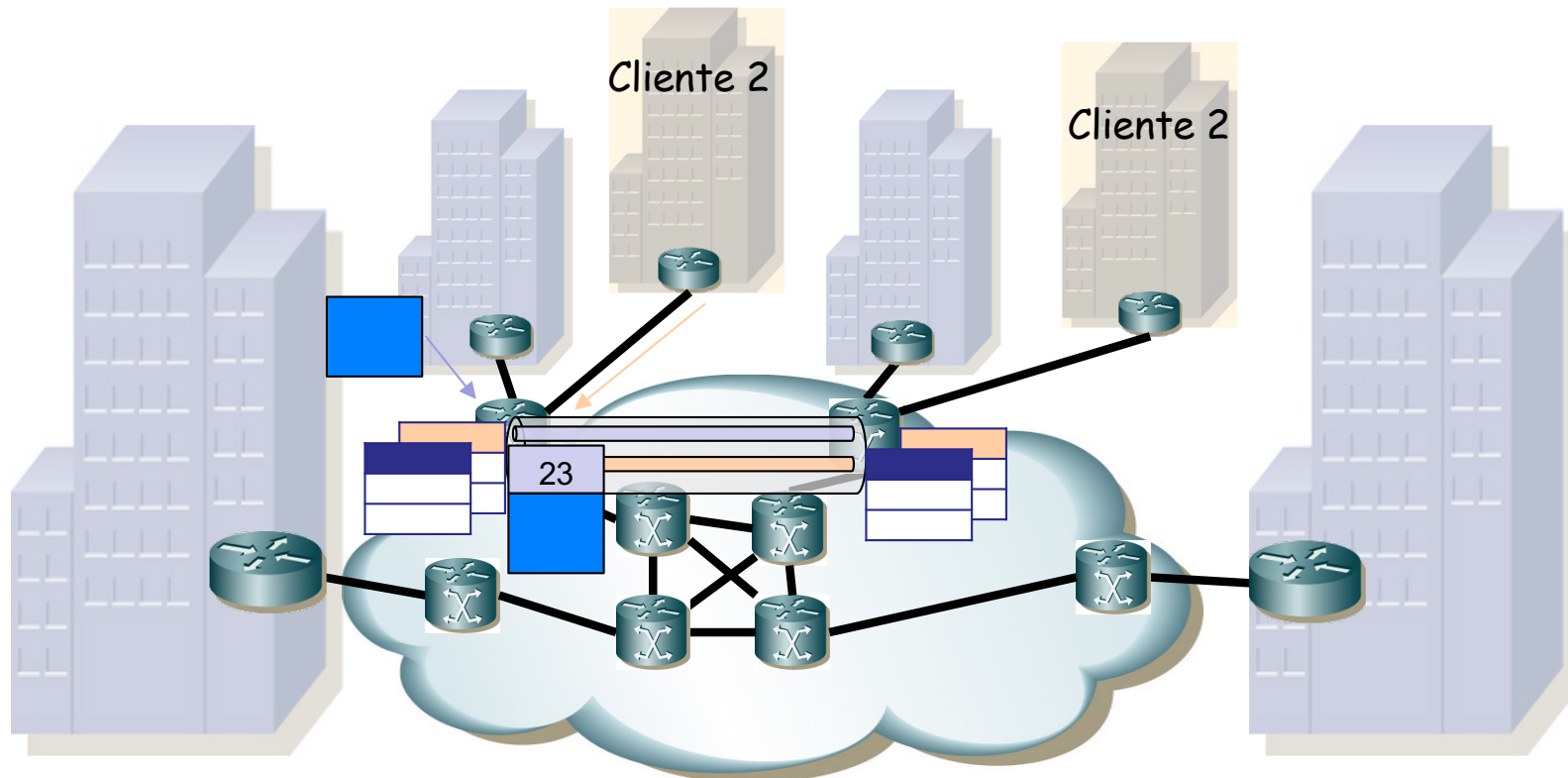
Recapitulando

- Un LSP entre cada pareja de PEs
- Hacen de túnel para los LSPs de cada cliente
- Cada PE una VRF para cada cliente
- Dado el interfaz de entrada (el AC) del paquete distingue al cliente y su VRF
- La ruta en la VRF indica la etiqueta MPLS a añadir y el puerto de salida
- Aprendió esa ruta y su etiqueta asociada mediante MP-BGP
- Asoció esa ruta a esa VRF en base al RT
- Distingue esa ruta de otra que tenga la misma dirección IP con el RD



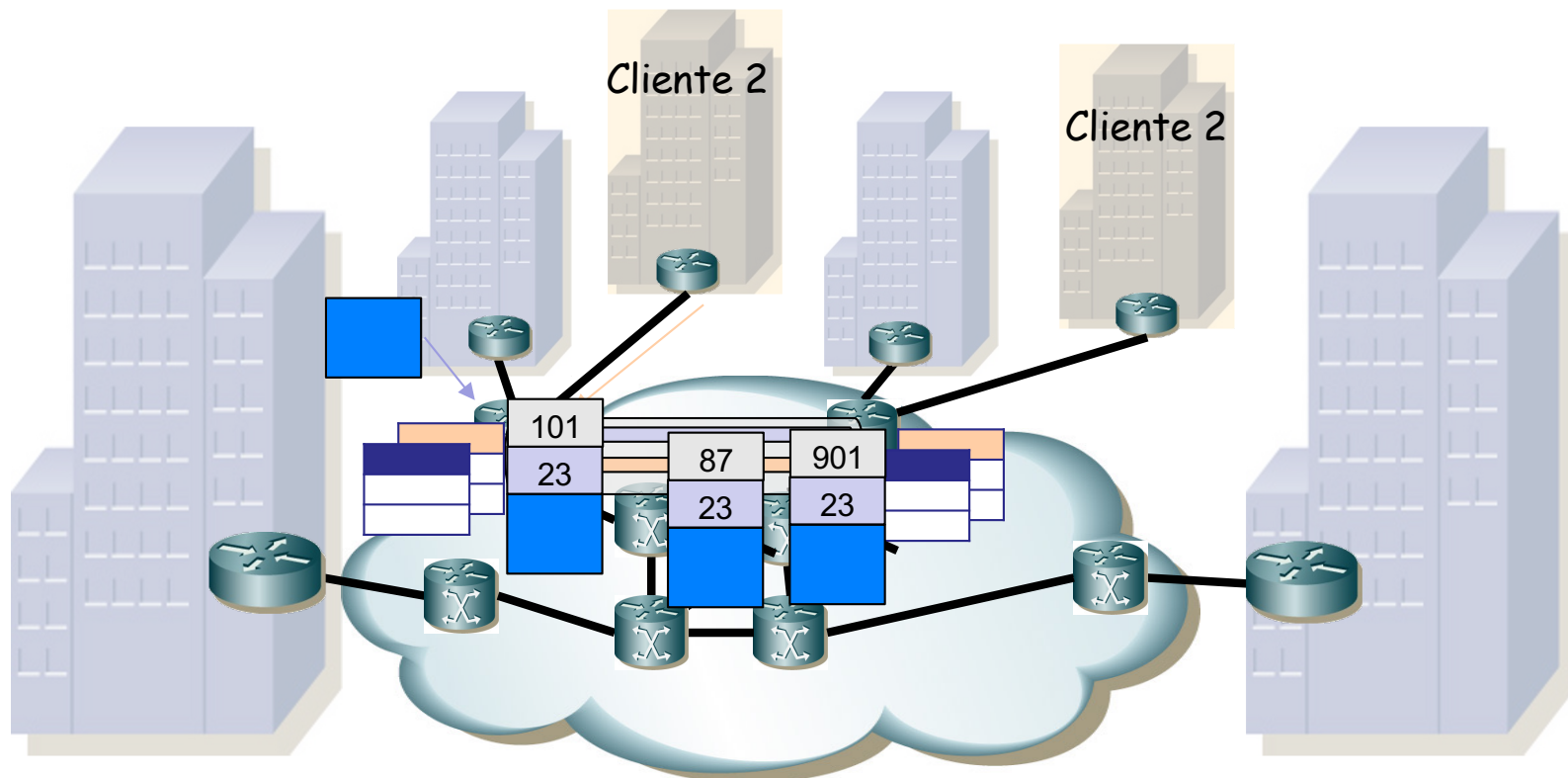
Recapitulando

- Esa etiqueta corresponde al LSP del usuario
- El puerto de salida es el LSP entre los dos PEs
- Se le añade otra etiqueta que identifica al Tunel LSP entre esos dos PEs



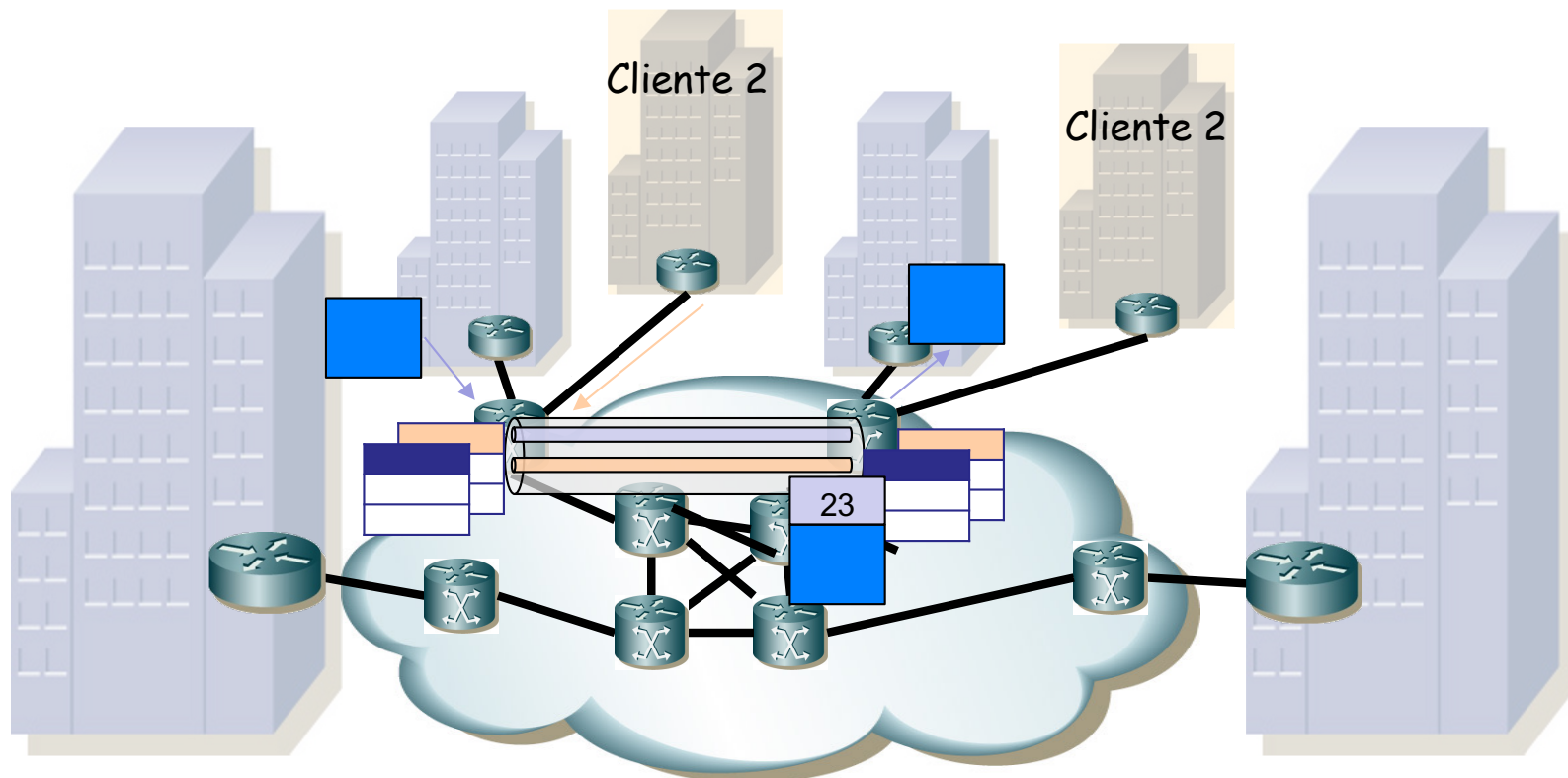
Recapitulando

- Esa etiqueta corresponde al LSP del usuario
- El puerto de salida es el LSP entre los dos PEs
- Se le añade otra etiqueta que identifica al Tunel LSP entre esos dos PEs
- Los P routers reenvían en función de esa etiqueta externa
- Los P routers no "ven" la etiqueta interna y por lo tanto no ven los LSPs de usuarios, solo la malla entre los PEs



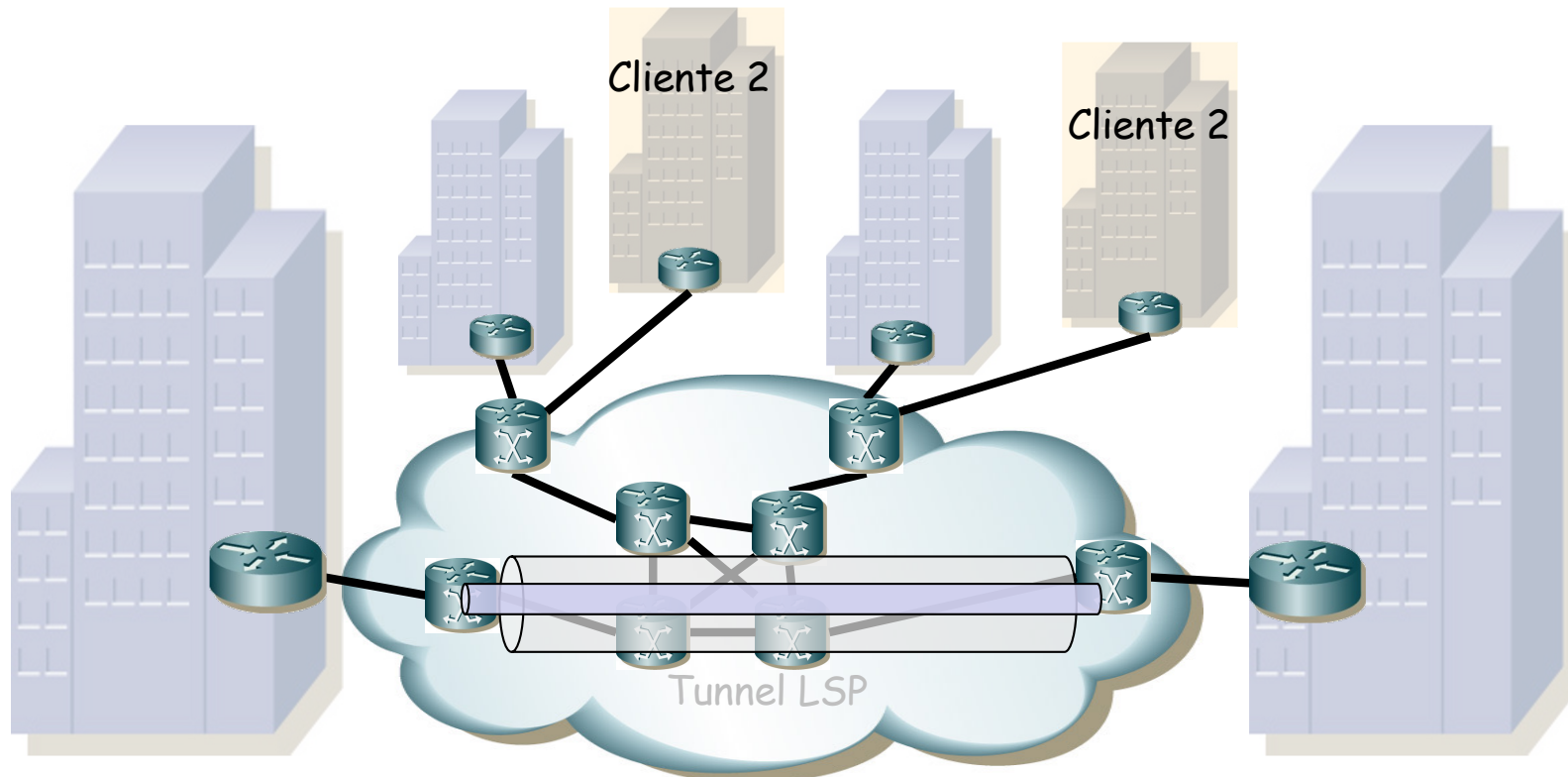
Recapitulando

- En el PE de egreso la etiqueta del LSP permite identificar la VRF y con ella la tabla de rutas



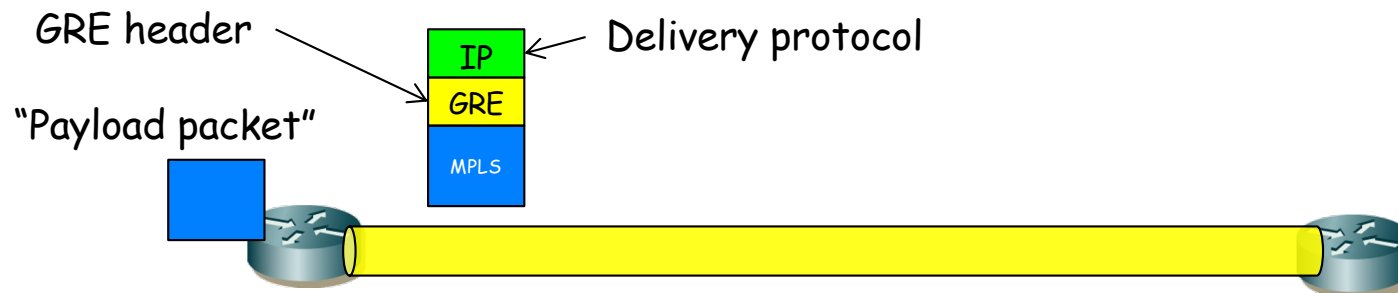
L3VPN con GRE

- Los túneles entre los PE pueden ser túneles GRE o simple IPoIP (RFC 4797)
- Entonces la red de transporte ya no necesita ser MPLS, es simple IP



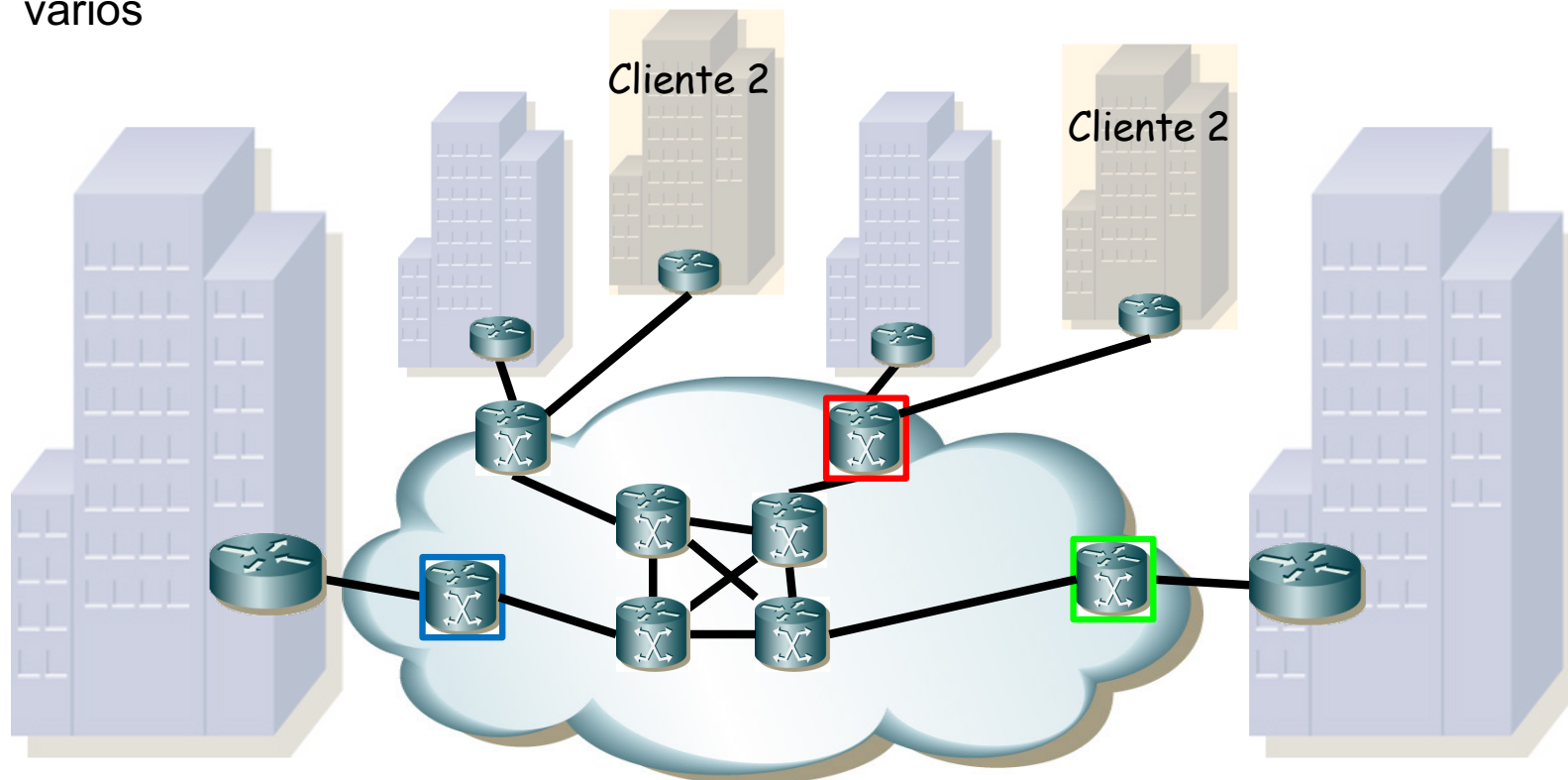
MPLS in GRE in IP

- RFC 4023 “Encapsulating MPLS in IP or Generic Routing Encapsulation (GRE)” (Motorola, Juniper, Cisco, 2005)
- El “*delivery protocol*” podría ser IP (protocol = 47 = GRE)
- El “*payload packet*” podría ser MPLS (Ethertype 0x8847 para unicast y ese mismo ó 0x8848 para multicast, RFC 5332)
- EoMPLSoGRE = Ethernet over MPLS over GRE
- Al transportarse sobre IP puede emplear IPSec
- RFC 4023 contempla también que MPLS se transporte directamente sobre IP, lo cual es más eficiente (sin GRE, protocolo 137 sobre IP)
- Puede haber motivos para tener GRE (exista el túnel con anterioridad, la implementación del equipo lo requiera en su fastpath, etc)



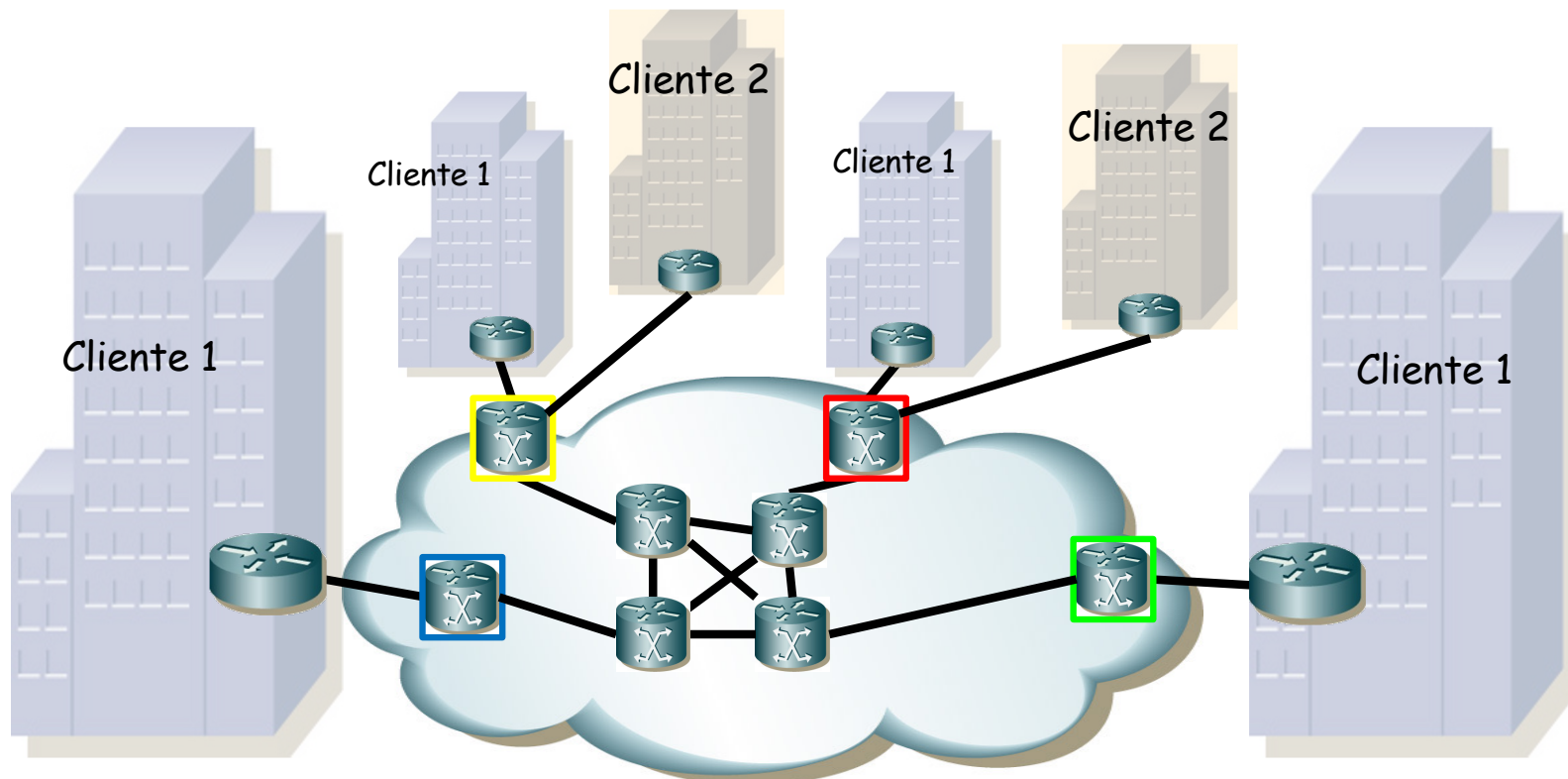
RD y RT

- RD permite anuncios del mismo prefijo que se puedan diferenciar
- Se suele asociar a una VRF
- RT permite importar en una VRF los anuncios con ciertos RTs
- En el anuncio de los prefijos se pone el RT que emplearemos para importar en otras VRFs
- Esto se puede usar para que unos PEs importen (para el mismo cliente) diferentes rutas
- Unos pueden importar por ejemplo los anuncios de un solo RT y otros los de varios



RD y RT

- Por ejemplo, para el cliente 1 cada PE podría hacer sus anuncios con un RT diferente
- Podríamos decirle a los PEs amarillo, azul y verde que importen del RT rojo
- Y decirle al PE rojo que importe de los RTs amarillo, azul y verde
- Con eso verde, azul y amarillo puedes comunicarse con rojo pero no entre ellos



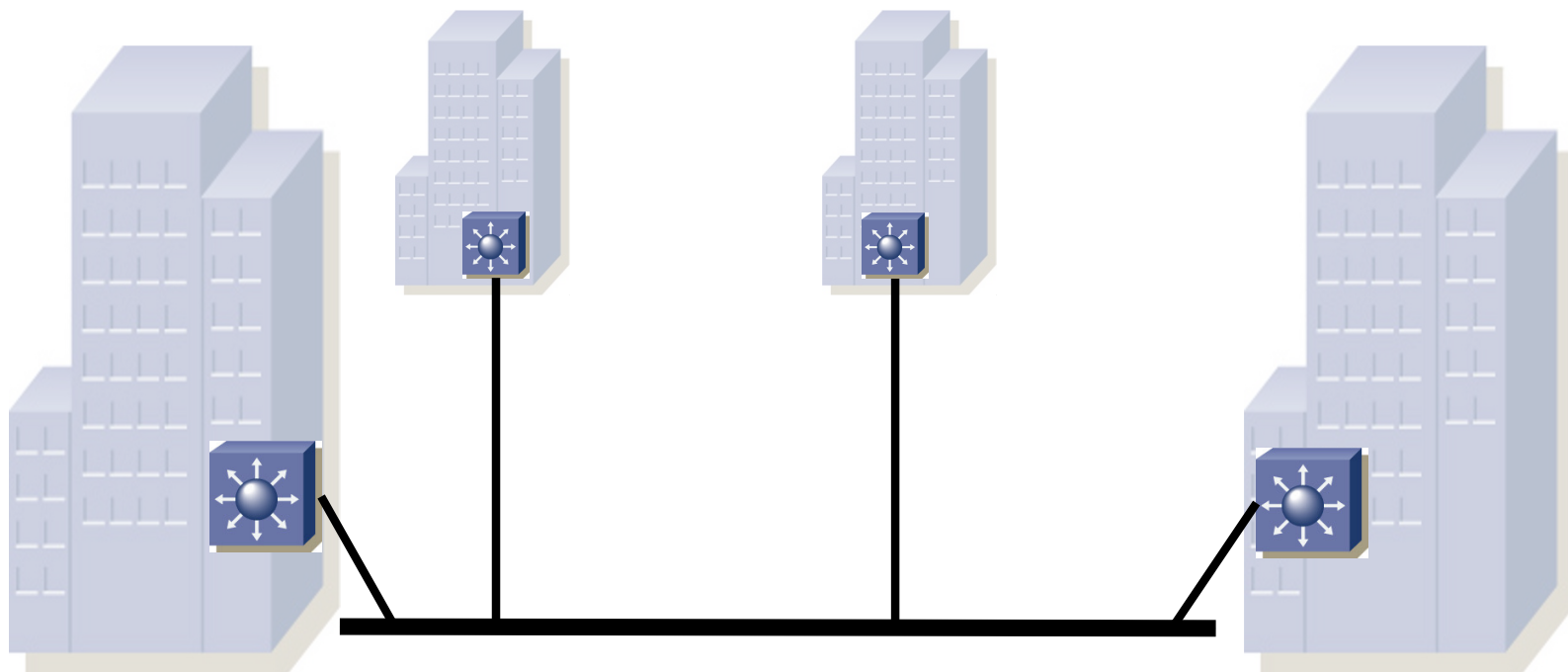
L3VPN – Forwarding en el Core

Layer 3 MPLS VPNs

VPLS

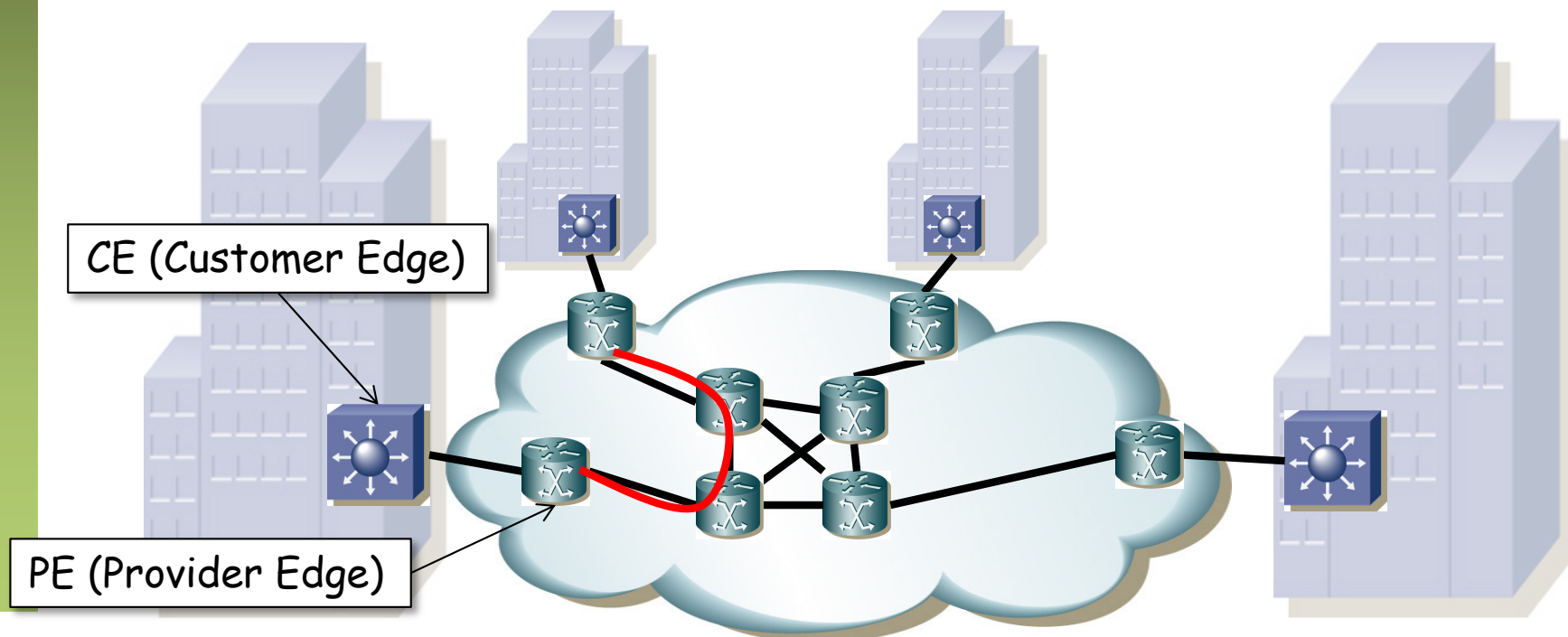
MPLS y VPLS

- “*Virtual Private LAN Service*”, una VPN layer 2 (RFC 4664, Acreo y Cisco, 2006)
- Interconecta múltiples *sites* en un solo dominio puenteado
- Todos los extremos se comportan como si estuvieran en una LAN
- *E-LAN Service*
- Transporta Ethernet así que sobre ella el cliente puede usar IP o cualquier otro protocolo
- Los equipos de usuario (Customer Edge) pueden ser switches o routers
- Transporte MPLS u otra solución de túneles (GRE, L2TP, IPsec)



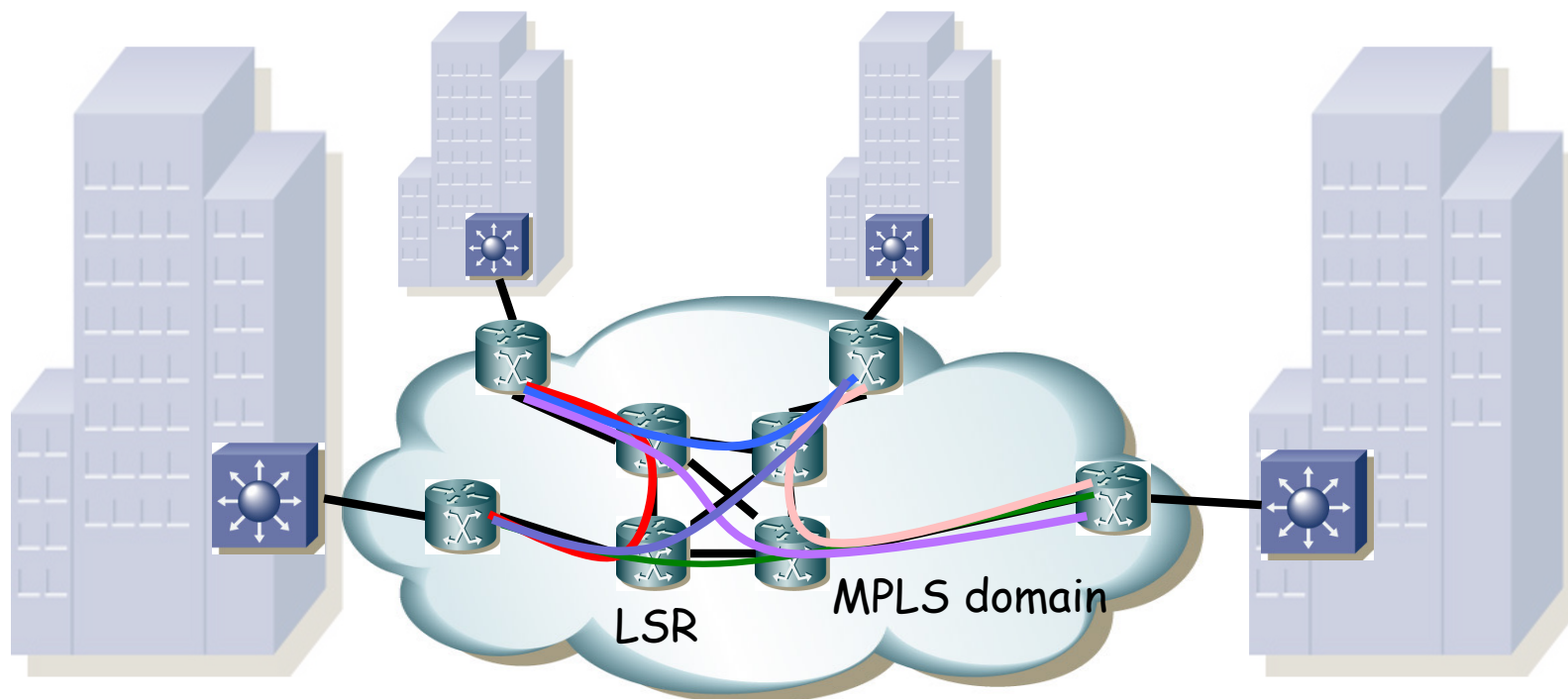
MPLS y VPLS

- El dominio MPLS puede transportar las tramas MPLS sobre IP o sobre otra tecnología
- La red puede dar servicio VPLS a más de un cliente
- El PE hace aprendizaje de direcciones MAC y replicación de tramas de forma independiente para cada cliente
- No interfiere el servicio de un usuario al otro (pueden por ejemplo emplear el mismo direccionamiento IP)
- Los equipos frontera establecen entre ellos los LSPs necesarios para el servicio multiacceso



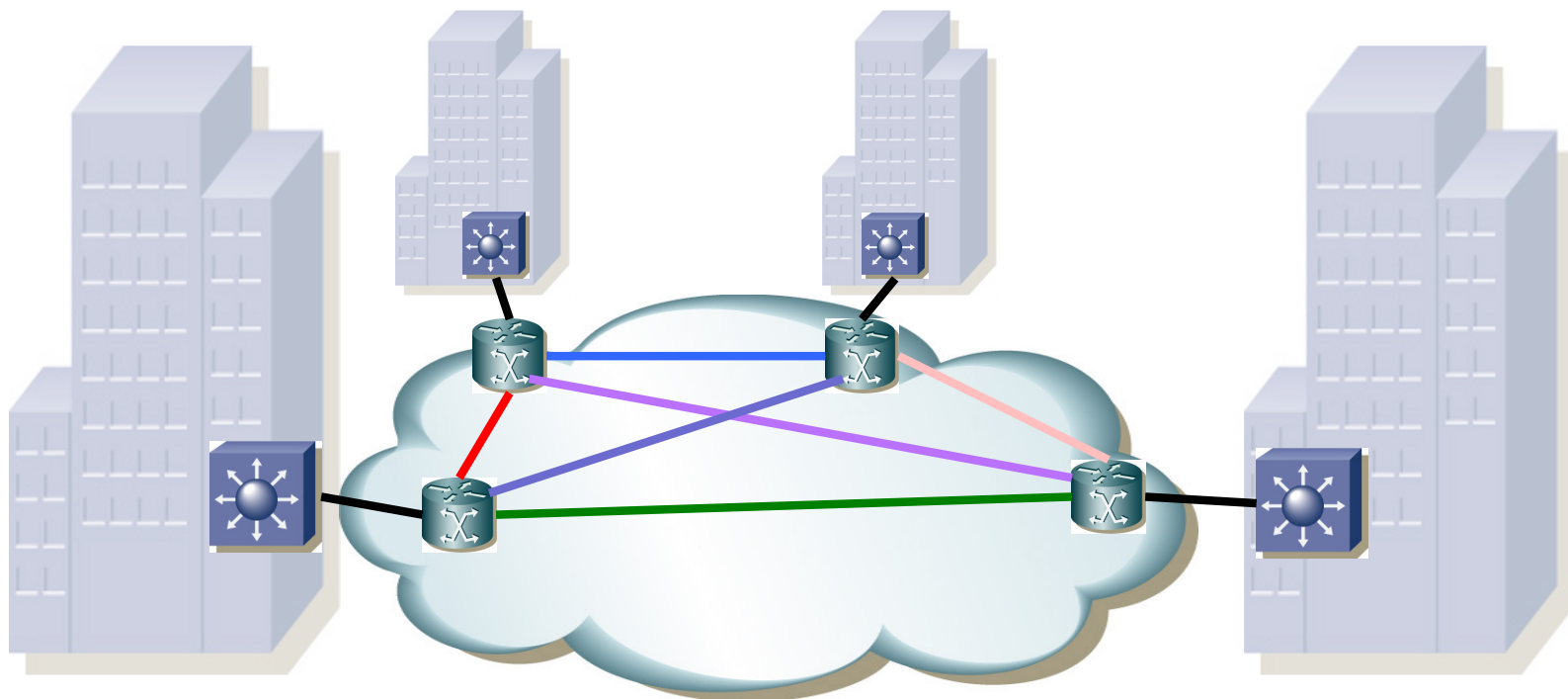
MPLS y VPLS

- Establecen un *full-mesh* entre los nodos frontera
- Para ello disponen de RSVP-TE (o LDP)
- (...)



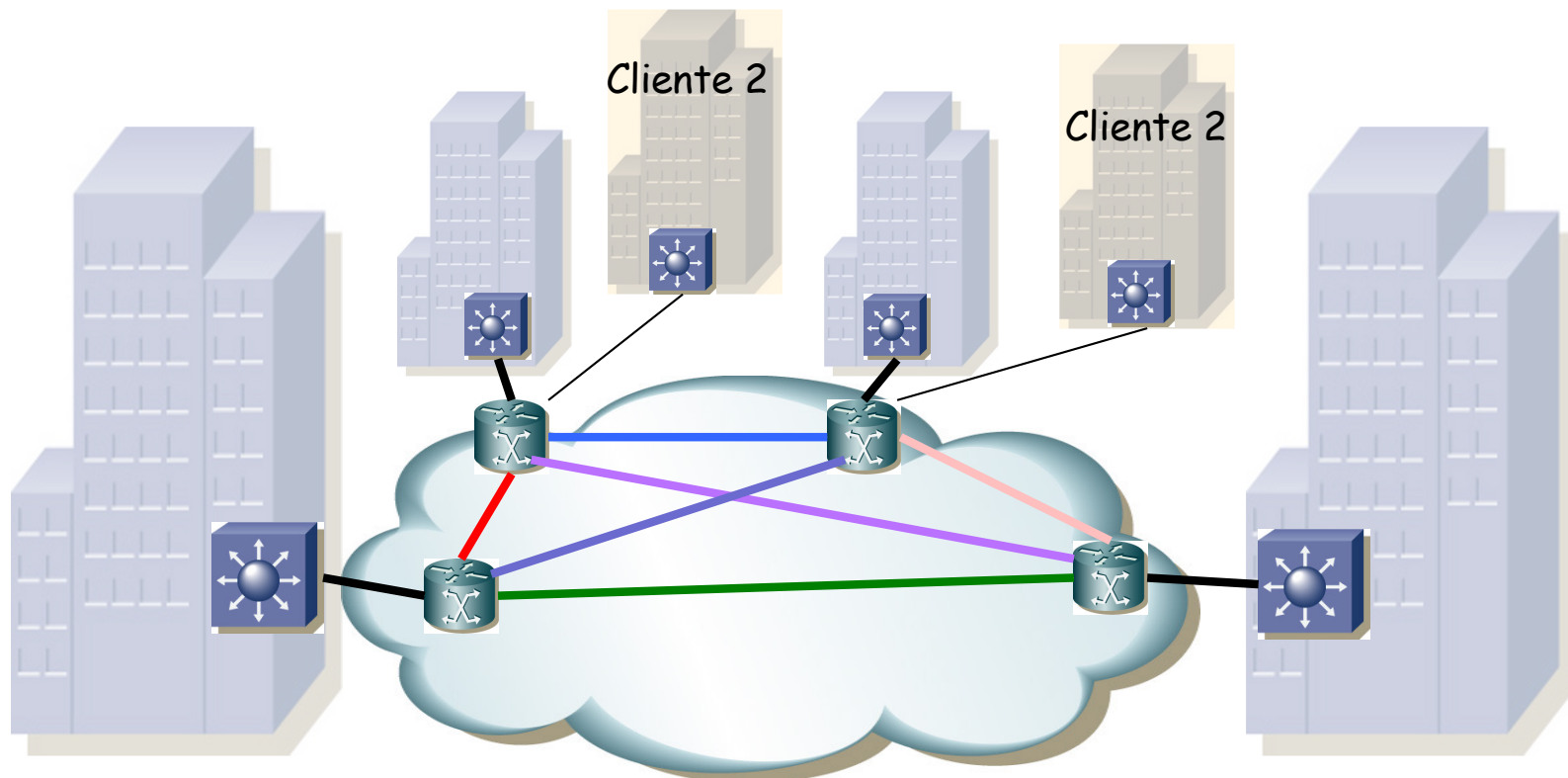
MPLS y VPLS

- Establecen un *full-mesh* entre los nodos frontera
- Para ello disponen de RSVP-TE (o LDP)
- Esos LSPs son globales al servicio VPLS, no particulares para cada cliente
- Es decir, puede haber otras LANs creadas con VPLS, para las sedes de otra empresa, y emplearán los mismos LSPs (...)



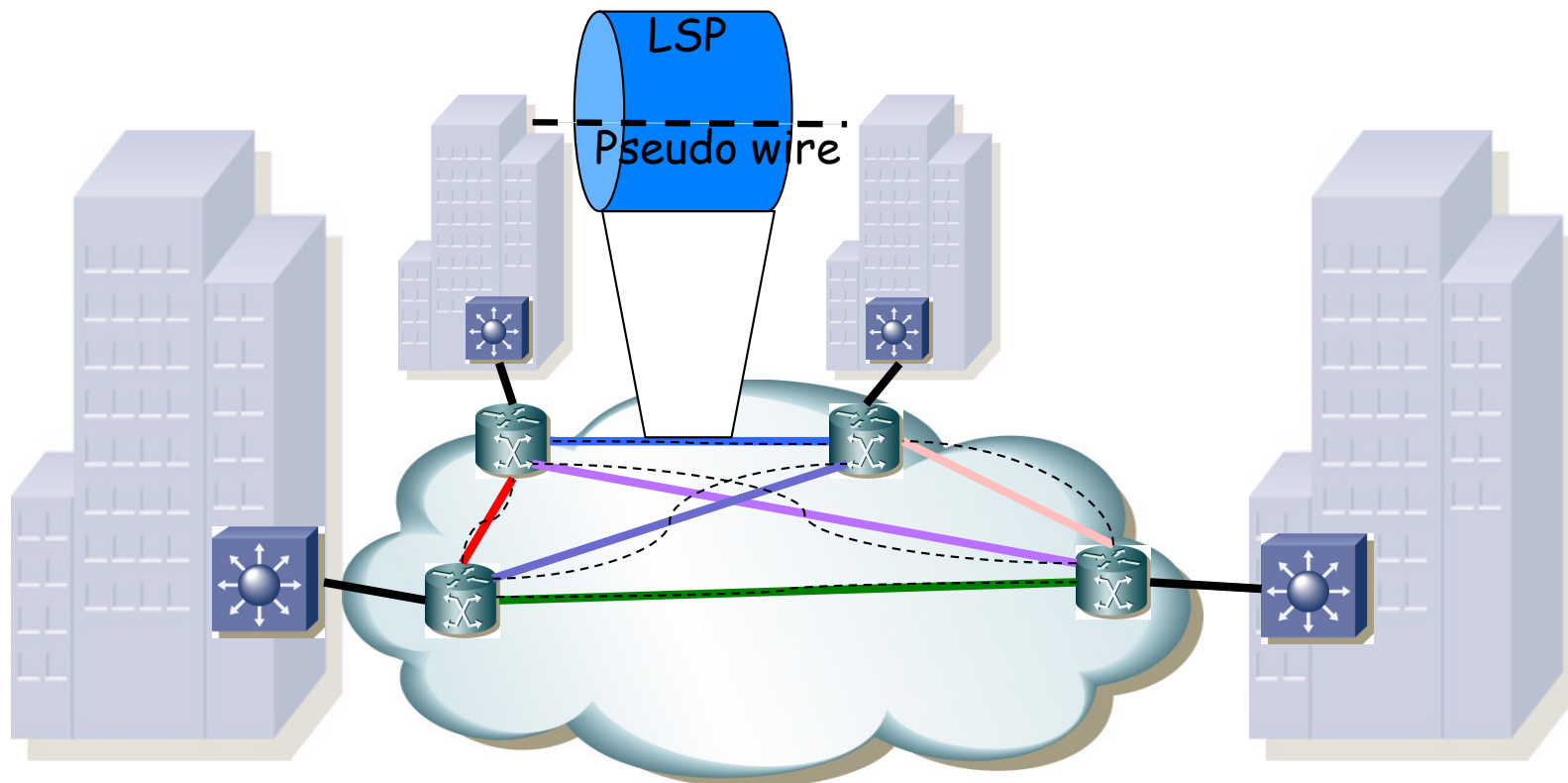
MPLS y VPLS

- Establecen un *full-mesh* entre los nodos frontera
- Para ello disponen de RSVP-TE (o LDP)
- Esos LSPs son globales al servicio VPLS, no particulares para cada cliente
- Es decir, puede haber otras LANs creadas con VPLS, para las sedes de otra empresa, y emplearán los mismos LSPs
- ¿Y para diferenciar a los clientes?



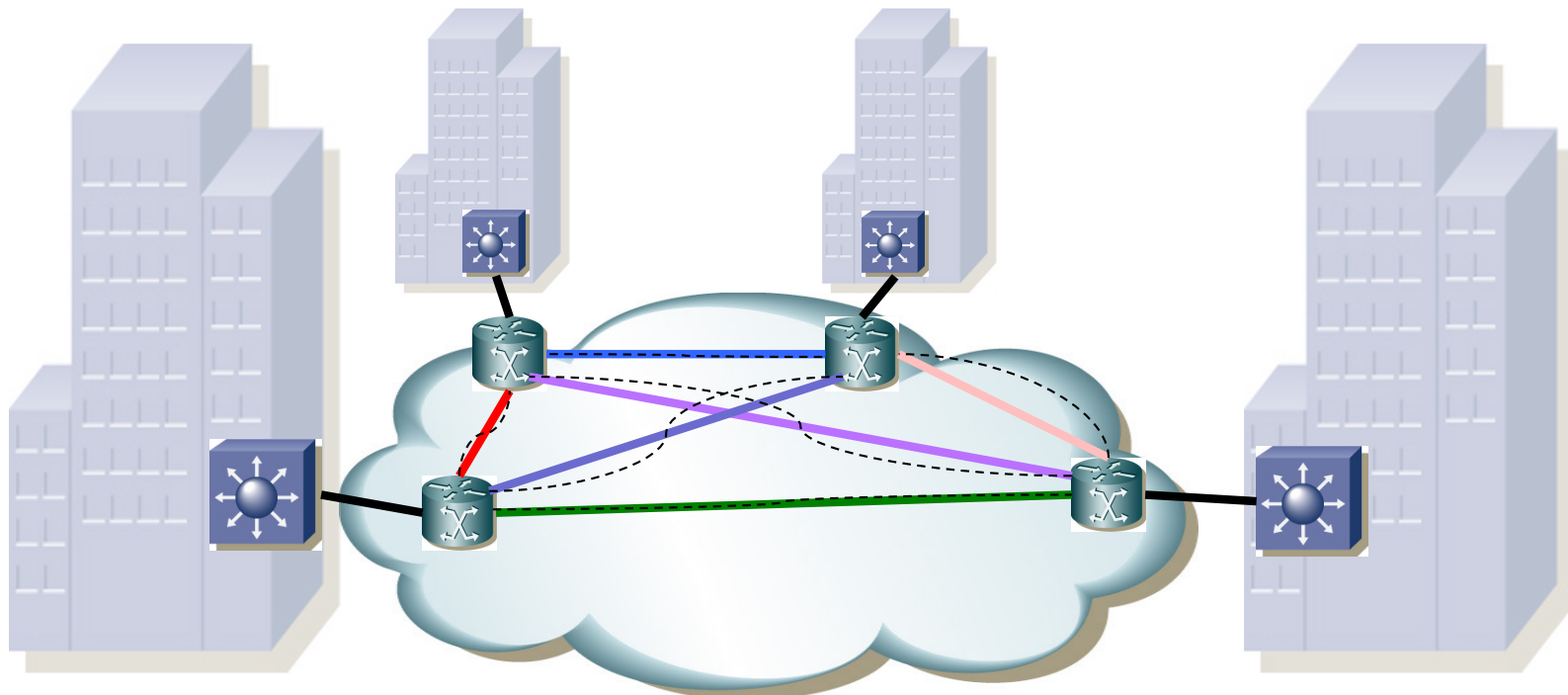
VPLS y PWE

- Por cada instancia VPLS (cada cliente) se establece un full mesh de *pseudo-wires* (PWs) entre los PEs (en el fondo la segunda etiqueta MPLS)
- RFC 3985 “Pseudo Wire Emulation Edge-to-Edge (PWE3) Architecture” (Cisco Systems, Overture Networks, 2005)
- Un PW emula un circuito, por ejemplo para transportar un E1 o un PVC ATM
- También puede transportar Ethernet, AAL5, SDH, etc



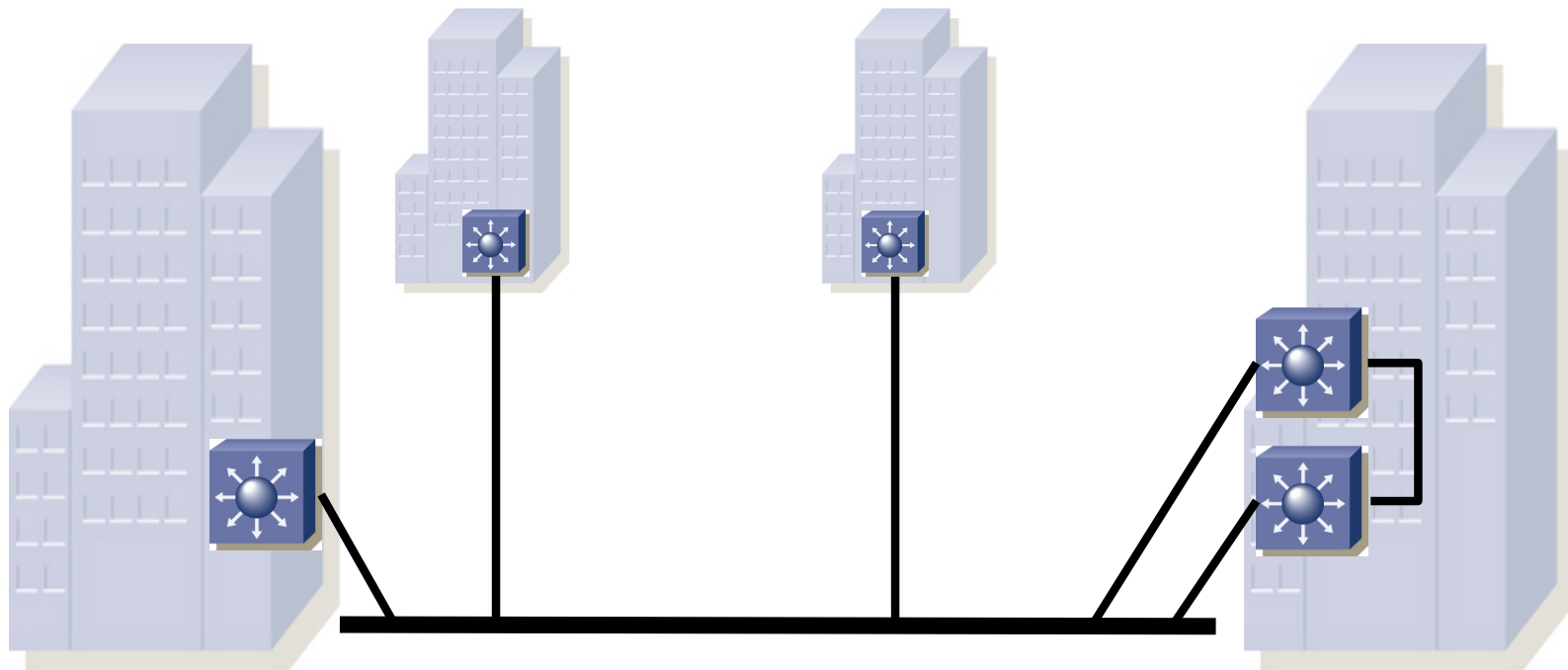
VPLS y PWE

- El full-mesh de PWs hace que los PE puedan enviarse directamente los unos a los otros
- No necesitan hacer reenvío y no hace falta resolver posibles bucles
- Simplemente se implementa una solución que se llama de “*split horizon*”:
 - Un PE no debe reenviar tráfico de un PW a otro en el mismo mesh VPLS
- El aprendizaje de direcciones MAC se hace en los PEs en el plano de datos (con la llegada de tramas Ethernet)



VPLS y PWE

- Sí puede haber ciclos, pero creados por el usuario para obtener redundancia
- En ese caso podrá emplear STP
- Las BPDUs se transportarían normalmente por el mesh VPLS



VPLS Control Plane

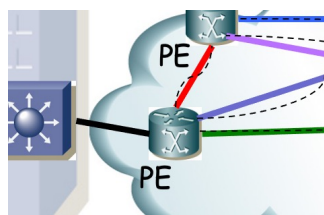
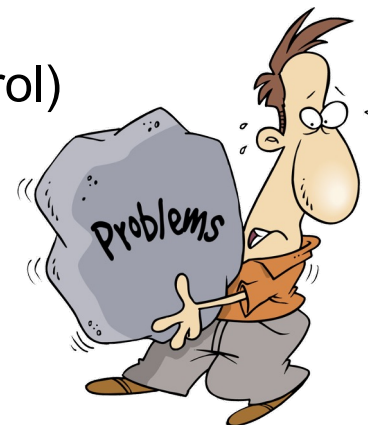
- Alternativas para el establecimiento de los pseudo-wires:
 - RFC 4761 “Virtual Private LAN Service (VPLS) Using BGP or Auto-Discovery and Signaling”
 - RFC 4762 “Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling”
- Repito: el aprendizaje de direcciones MAC se hace en el plano de datos, es decir, con la dirección MAC origen de la trama recibida (*flood and learn*)

VPLS

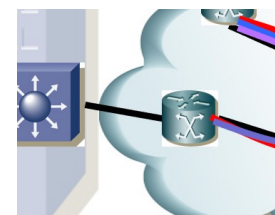
VPLS - Problemas

Problemas en VPLS

- Se deben establecer $N(N-1)/2$ pseudo-wires
- Problema de escalabilidad (cantidad de tráfico de control)
- Replicación de paquetes que sufren inundación:
 - Se lleva a cabo en el PE de entrada
 - Se dirigen punto-a-punto a cada otro PE del servicio
 - Mayor trabajo en el PE
 - Más uso de capacidad
 - Mayor retardo (si hay que enviar N veces la trama por N PWs que se implementan sobre el mismo LSP irán en serie)



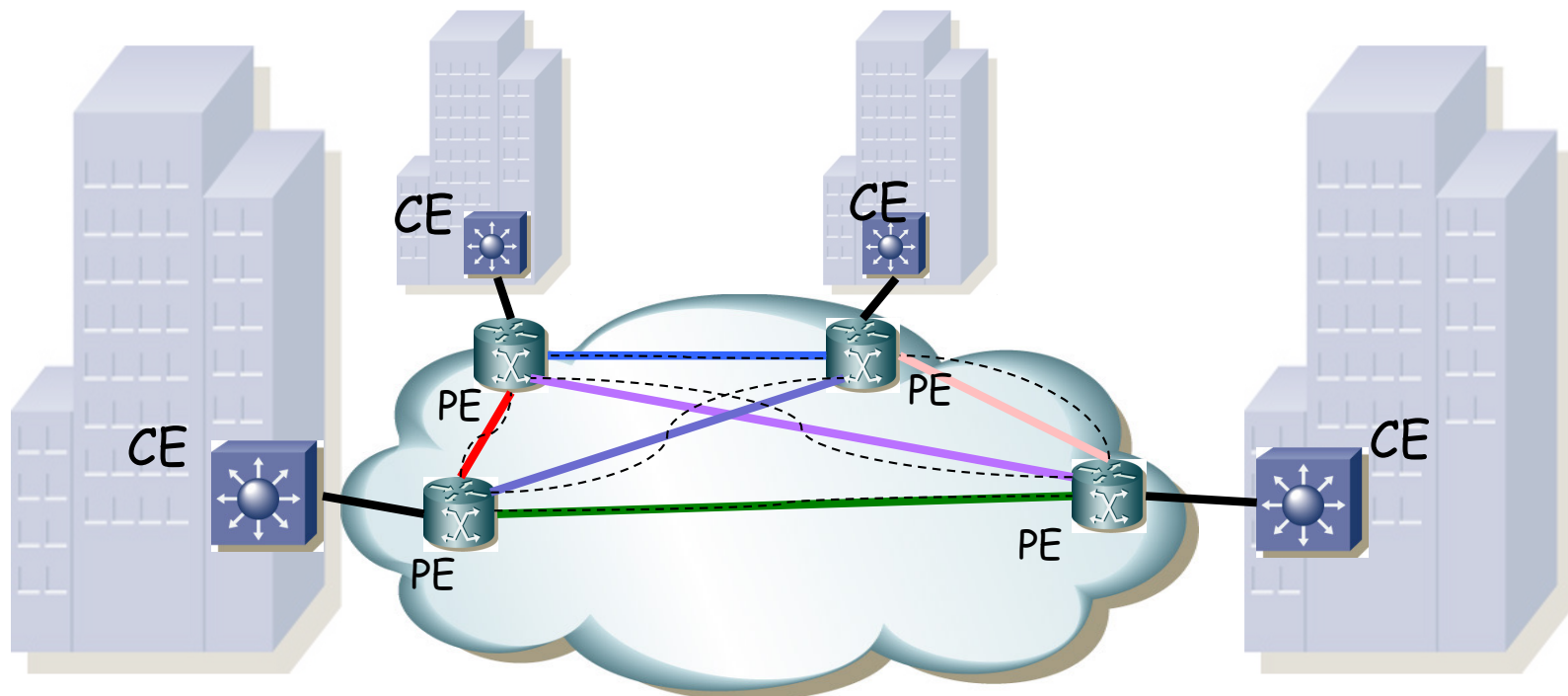
3 PWEs pero un solo interfaz físico



- Si se añade un acceso del cliente, a un PE diferente, se deben crear los PWs, lo cual implica reconfigurar los demás PEs
- Para despliegues pequeños
- Mejoras:
 - H-VPLS (Hierarchical VPLS)
 - Hierarchical BGP VPLS

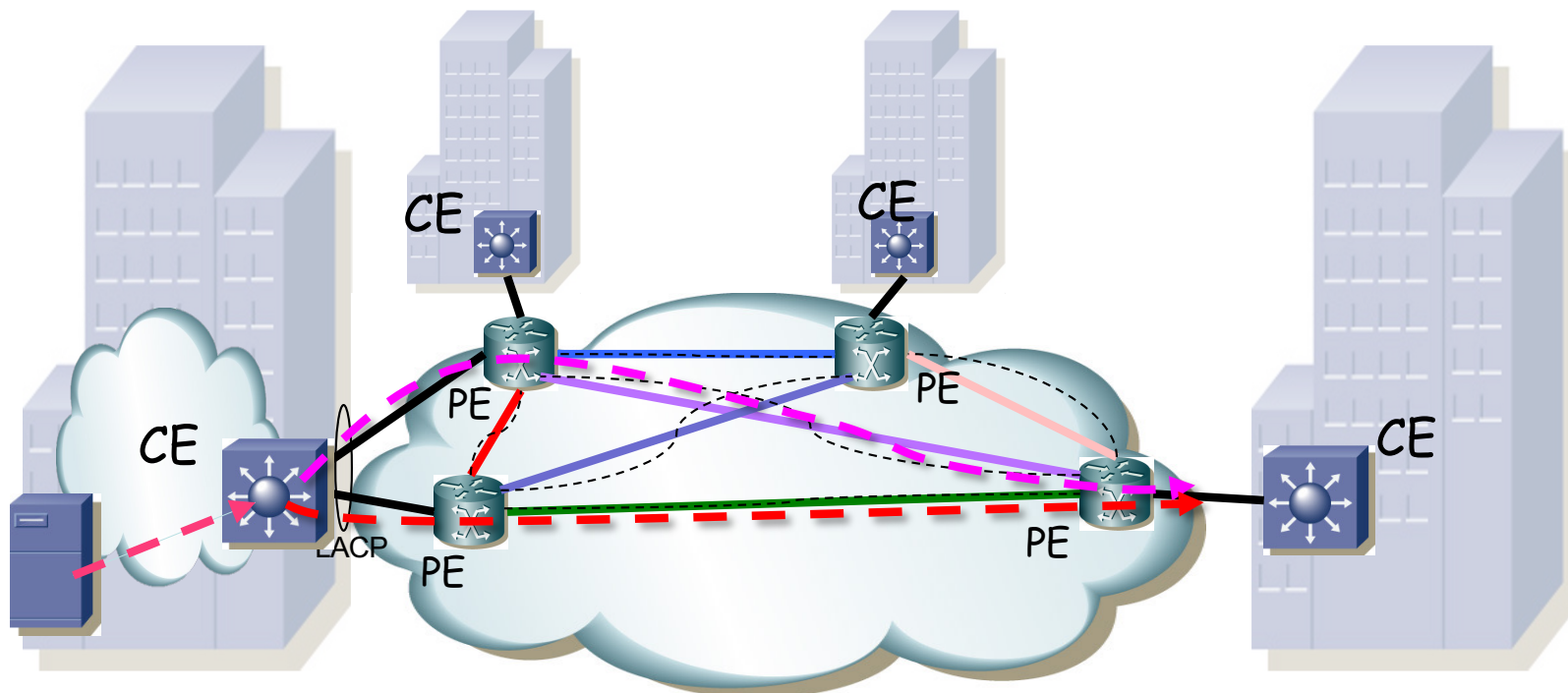
Limitaciones de VPLS

- Solo soporta *Single-Active Redundancy Mode* (no *all-active*)
- (...)



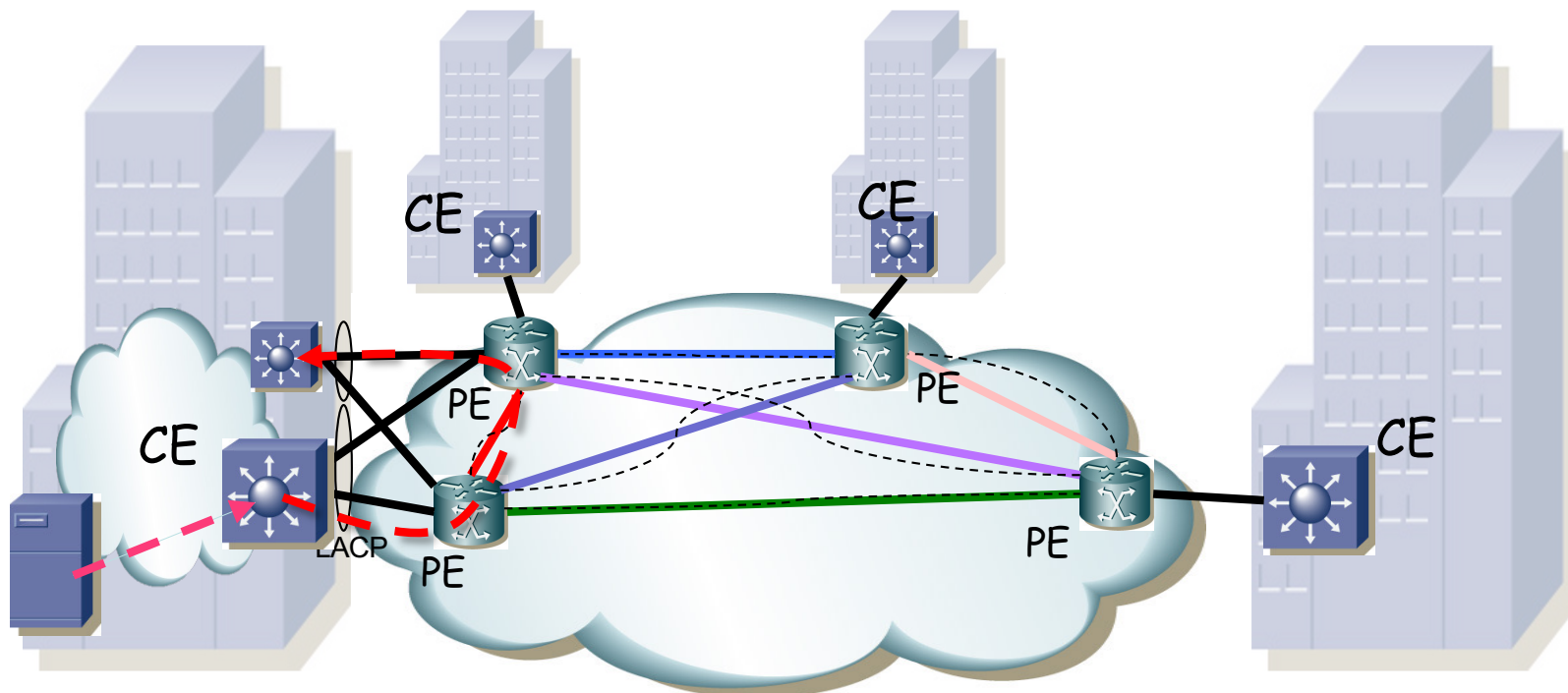
Limitaciones de VPLS

- Solo soporta *Single-Active Redundancy Mode* (no *all-active*)
- Ejemplo:
 - CE un LAG hacia dos PEs
 - Reparte tráfico entre ellos
 - Al llegar a otro PE llega la misma dirección MAC origen por dos PWs, saltando la MAC aprendida de uno a otro (...)



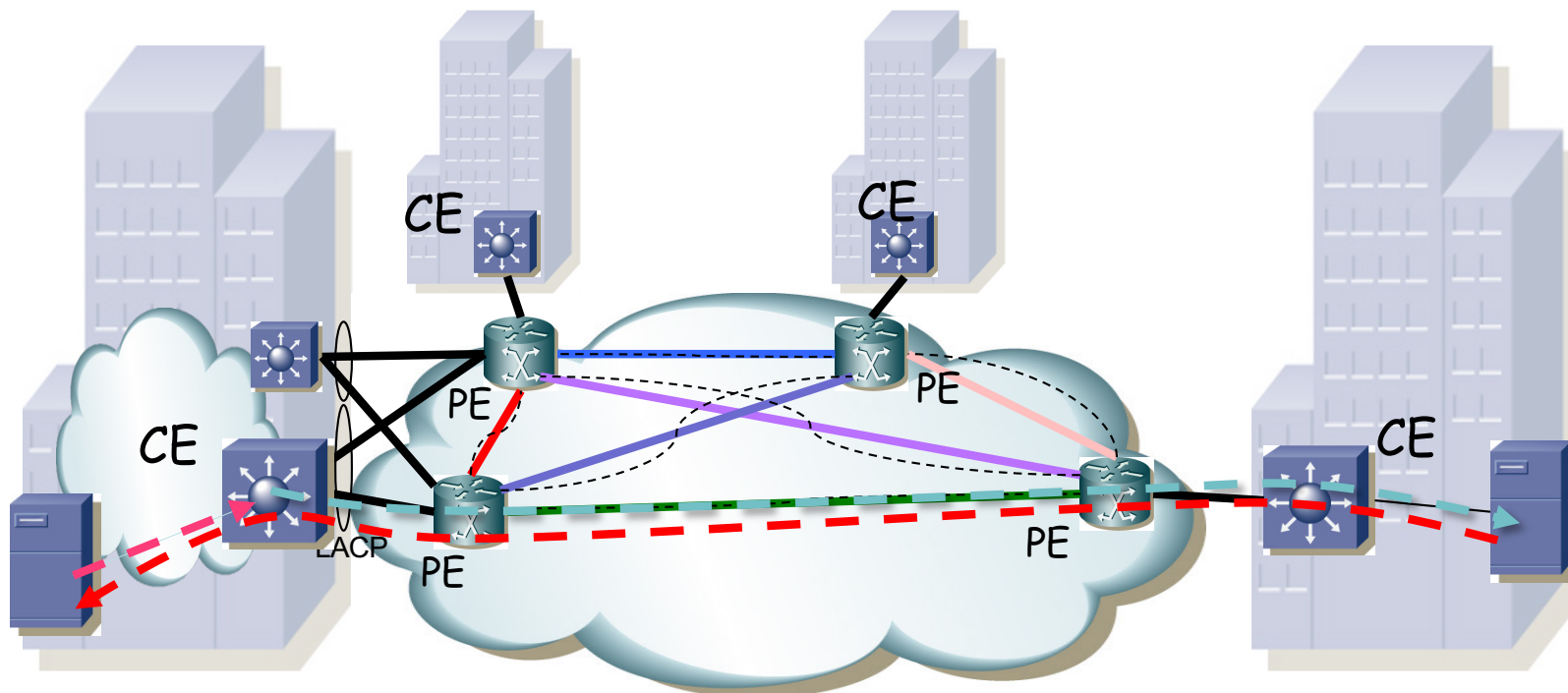
Limitaciones de VPLS

- Solo soporta *Single-Active Redundancy Mode* (no *all-active*)
- Ejemplo:
 - Dos CEs con LAGs (redundancia en el CE)
 - BUM es enviado por PE a todos los demás PEs y puede volver por el otro CE (...)



Limitaciones de VPLS

- Solo soporta *Single-Active Redundancy Mode* (no *all-active*)
- Ejemplo:
 - No hay load balancing en el sentido de respuesta si el hash lleva a emplear uno solo de los PEs (...)
 - El otro extremo solo aprende la dirección MAC por un PW
 - En cualquier caso no la puede aprender por dos PWs (...)
- Active-Active solo mediante soluciones propietarias (vPC, VSS, etc)

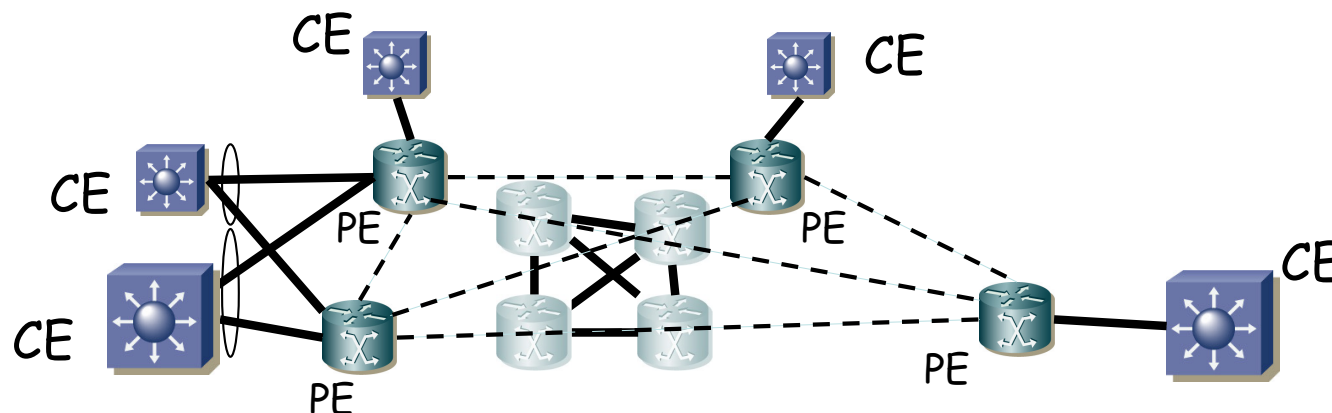


VPLS - Problemas

EVPN

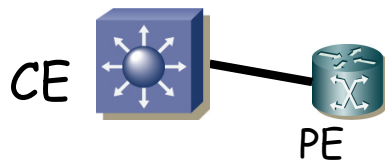
MPLS-EVPN

- RFC 7209 “Requirements for Ethernet VPN (EVPN)”, Cisco, Arktan, AT&T, Verizon, Alcatel-Lucent, Bloomberg (2014)
- RFC 7432 “BGP MPLS-Based Ethernet VPN”, Cisco, Arktan, Verizon, Bloomberg, AT&T, Juniper, Alcatel-Lucent (2015)
- Emplea un plano de control **BGP**, como una L3VPN
- Aprendizaje de direcciones MAC en el plano de control en lugar de en el plano de datos
- Ofrece una VPN capa 2, como VPLS
- Permite redundancia en el acceso (*all-active*) y balanceo de carga por flujo
- Los PEs pueden estar conectados mediante LSPs o túneles (IP/GRE)
- Abandona el aprendizaje en el plano de datos (*no flood and learn*)



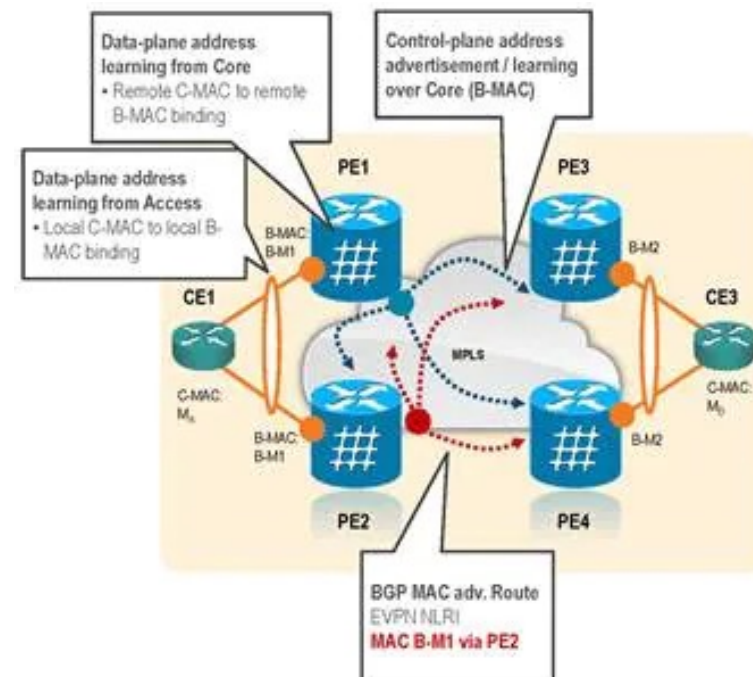
MPLS-EVPN

- BGP distribuye Ethernet MACs (opcional el par MAC-IP)
- PEs anuncian direcciones MAC aprendidas del CE, junto con una etiqueta MPLS, al resto de PEs mediante MP-BGP
- Queda abierto cómo aprende el PE del CE (puede ser plano de datos)
- Igual que las L3VPNs emplea *route distinguishers* y *route targets*
- Su uso principal es en DCI



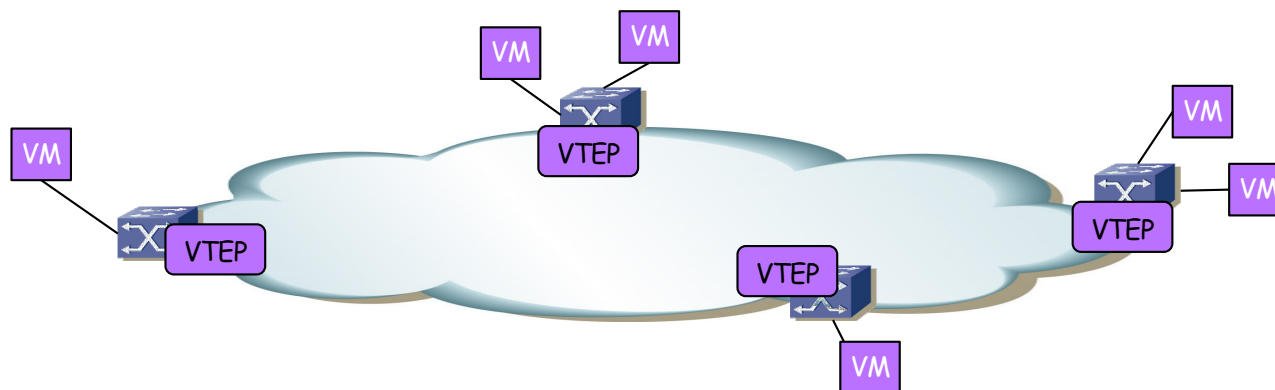
PBB-EVPN

- PBB-EVPN (RFC 7623) combina EVPN con plano de datos PBB (802.1ah)
- El core es MPLS
- La frontera son PEs de PBB que aprenden en el plano de datos las C-MACs asociadas a las B-MACs
- BGP anuncia las B-MACs
- Mejor escalabilidad en la RIB al no aprender las C-MACs



VXLAN-EVPN

- Plano de datos VXLAN (underlay IP)
- VTEPs corren proceso MP-BGP
- Anuncian las direcciones MAC de los hosts de la VXLAN
- Puede hacer *ARP suppression*
 - VTEP hace *ARP snooping* e inspecciona ARP request/response
 - VTEP aprende relación MAC/IP de cliente
 - VTEP anuncia esa relación mediante MP-BGP al resto de VTEPs
 - A partir de aquí, cualquier VTEP al hacer *ARP snooping* y ver una pregunta respecto a esa IP contesta él con la MAC aprendida



Otros

- Los anteriores son servicios E-LAN según define el MEF (antes el Metro Ethernet Forum)
- EVPN-VPWS (Virtual Private Wire Service) para servicio punto-a-punto (E-LINE)
- EVPN-ETREE para servicio en árbol (E-TREE)

EVPN