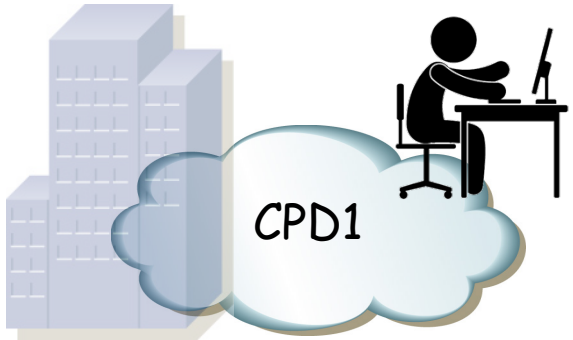
Interconexión de DCs: Introducción

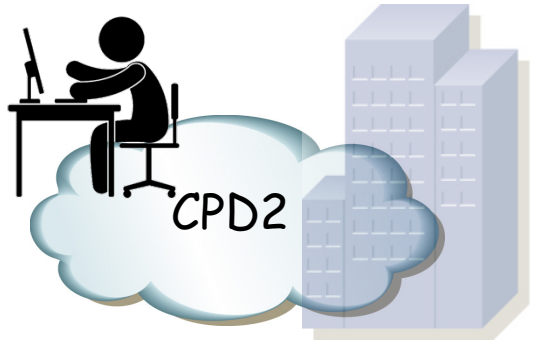
Múltiples DCs

- La palabra clave es “disponibilidad” (*availability*)
- Buscamos protección ante desastres:
  - Tsunamis, huracanes, inundaciones, terremotos, incendios
  - Fallos de larga duración de la red eléctrica (*black-outs*)
  - Violaciones de seguridad
- No es solo una cuestión de disponibilidad física sino que la lógica para coordinarlos debe funcionar correctamente también

Múltiples DCs

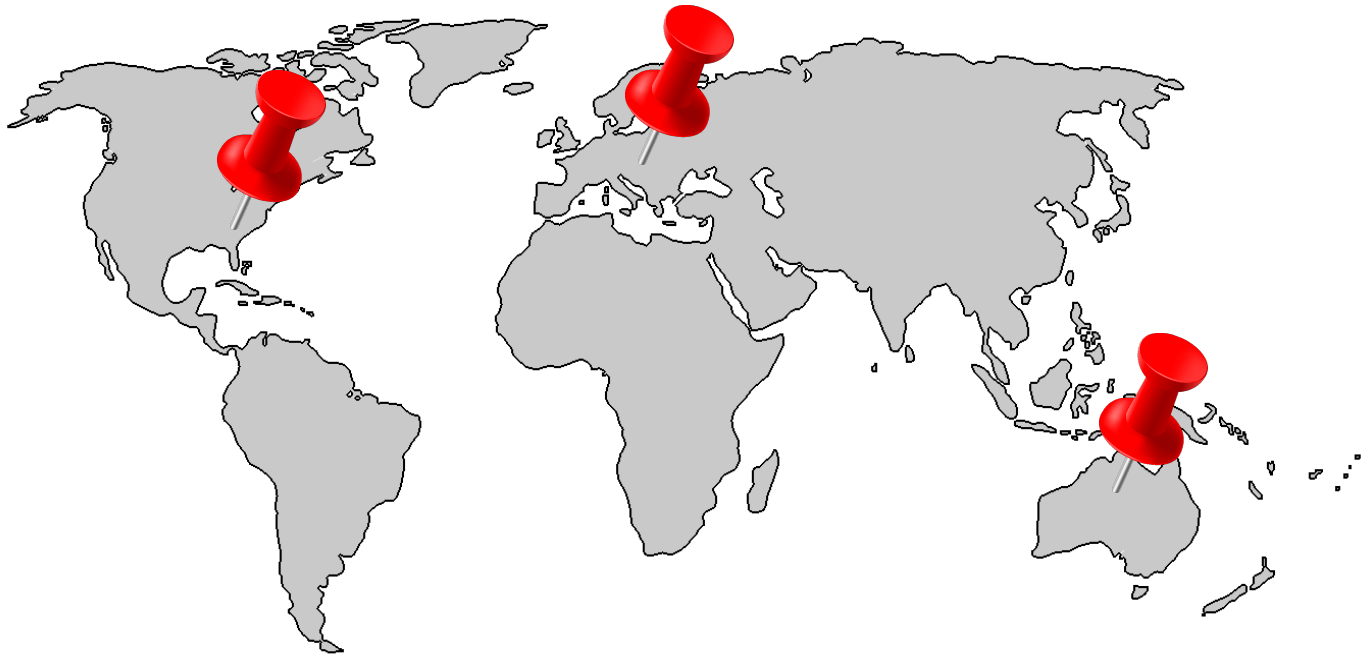
- Pueden trabajar por parejas en modo activo-standby
  - Uno de ellos cursa toda la carga de trabajo
  - El segundo monitoriza el estado del activo
  - Operaciones que modifiquen datos almacenados se sincronizan con el almacenamiento en el de respaldo
- Pueden trabajar en modo activo-activo
  - Necesitamos técnicas de reparto de carga entre los DCs
  - Así como (de nuevo) técnicas para sincronizar los datos entre ellos





Ubicación de DCs

- Alejados para que un problema “geográfico” no afecte a ambos
- Sin embargo podemos toparnos con limitaciones de retardo máximo para las aplicaciones distribuidas
- Por ejemplo la *replicación síncrona* se basa en devolver confirmación de haber almacenado el dato cuando se ha escrito en dos cabinas
- Si están en DCs alejados esto afectará al retardo de transacción
- Eso limita la distancia para reducir el tiempo de respuesta
- También protocolos como FC deben ajustarse para altos retardos (mayor RTT requiere mayor número de créditos para sacar provecho al BW)

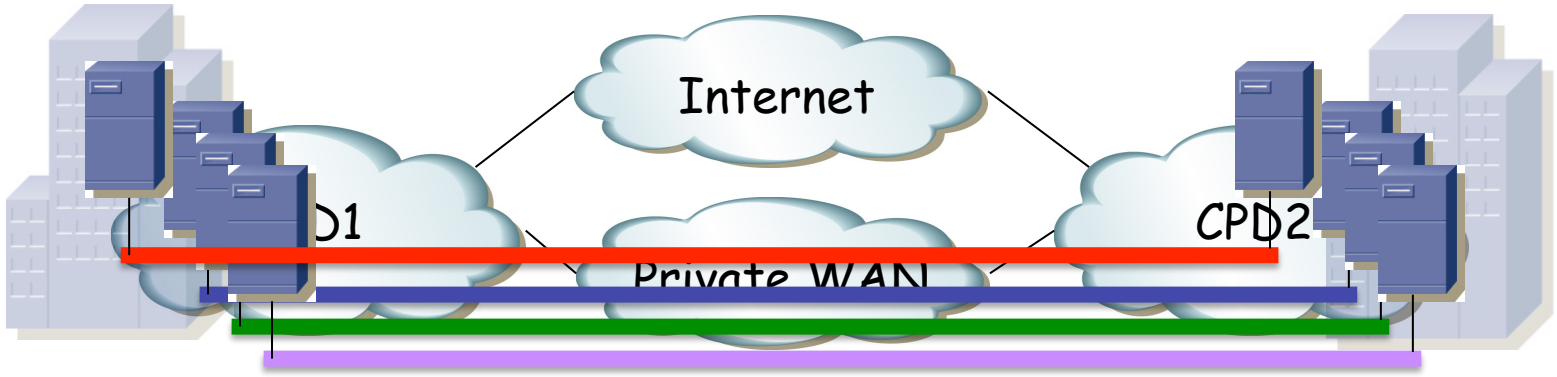


Múltiples DCs o sedes

- Habitualmente la interconexión se recomienda en capa 3
- Eso limita los problemas de capa 2 a cada DC
- Sin embargo muchas aplicaciones con funcionalidades de clustering requieren adyacencia en capa 2
  - Heartbeats o información de estado que envían multicast/broadcast
  - Nodos que comparten dirección IP y dirección MAC
- La movilidad de servidores (físicos o virtuales) requiere mantener la pertenencia a la misma VLAN
- O el crecimiento nos puede llevar a otro edificio
- Es decir, podemos necesitar extender las VLANs entre DCs
- Todo esto aplica tanto a interconexión de CPDs como de sedes remotas

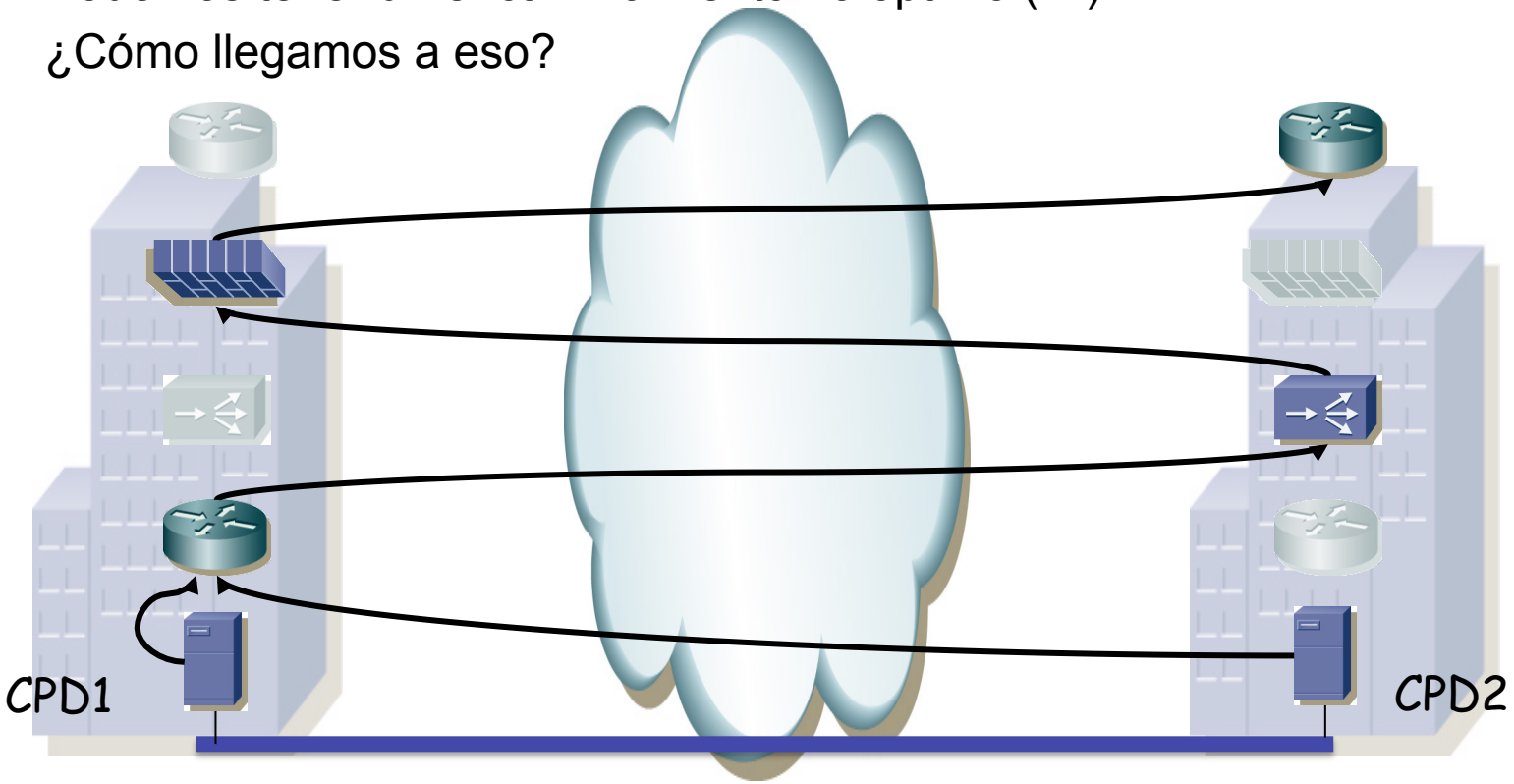
Multi-tenancy

- Múltiples clientes (miles) en un data-center
- Cada cliente debe tener una visión de la infraestructura como si estuviera solo y tuviera control total
- Virtualización en la red permite separar la red de esos usuarios



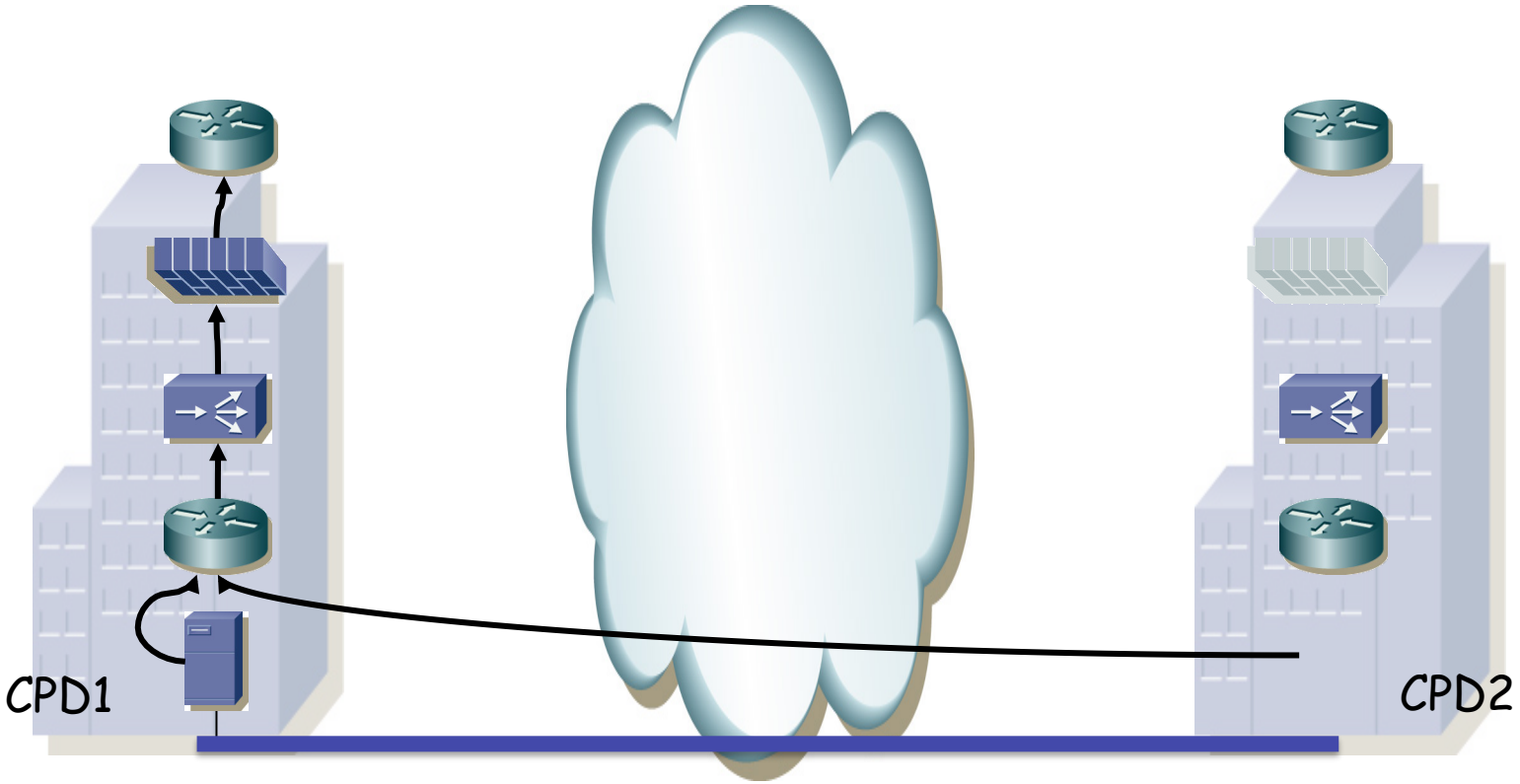
Problemas con extensión L2

- Entre los DCs hay que controlar el Broadcast, Unknown unicast y Multicast (BUM)
- ¿STP?
  - Problemas de escalabilidad
  - Fallo en la raíz afecta a los dos DCs
  - Si hay más de una interconexión seguramente desactive una
- Podemos tener un encaminamiento no óptimo (…)
- ¿Cómo llegamos a eso?



Tromboning

- ¿Cómo llegamos a eso?
- Por ejemplo porque hemos movido una máquina virtual (…)
- Una posible solución es emplear un FHRP entre los routers de ambos CPDs
- Y cortamos el tráfico del FHRP entre los CPDs para que ambos routers se elijan como maestros (activo/activo)



Interconexión del *storage*

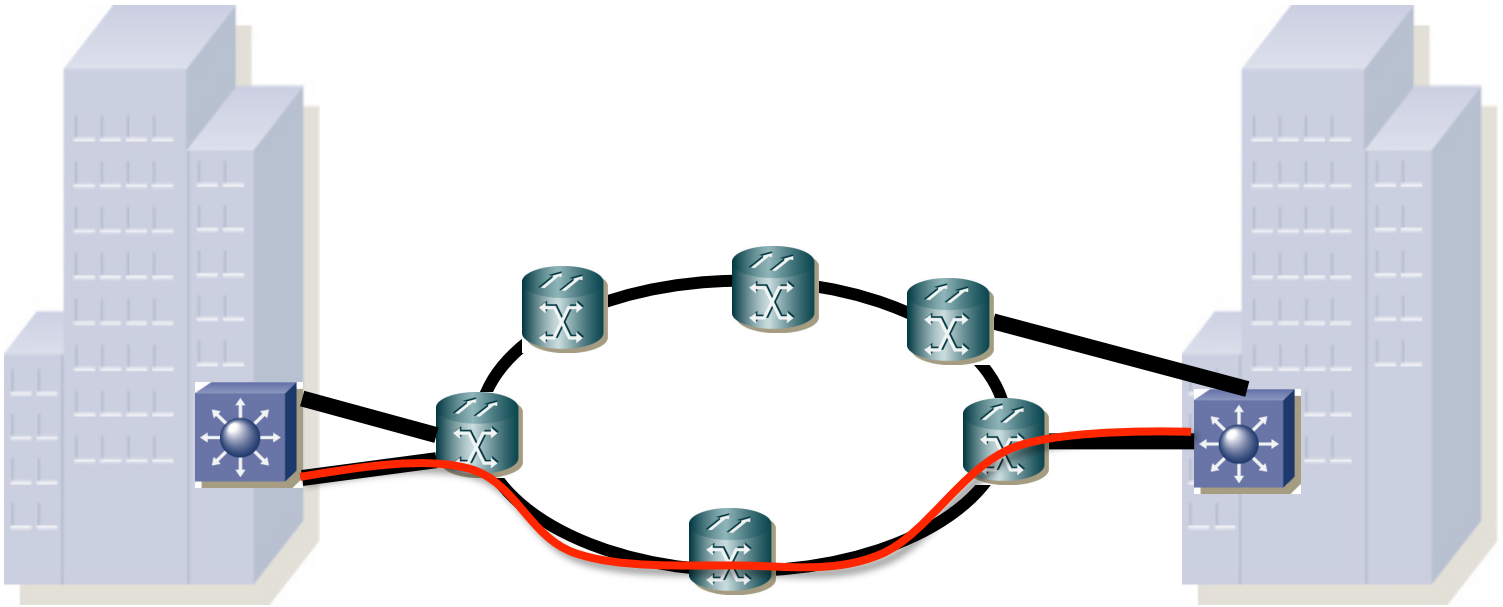
- La interconexión entre los DCs puede ser solo para sincronizar el almacenamiento
- "SAN extension"

Interconexión por fibra

- Se puede emplear *fibra oscura*
- Puede transportar múltiples wavelengths (CWDM, DWDM)
- (…)

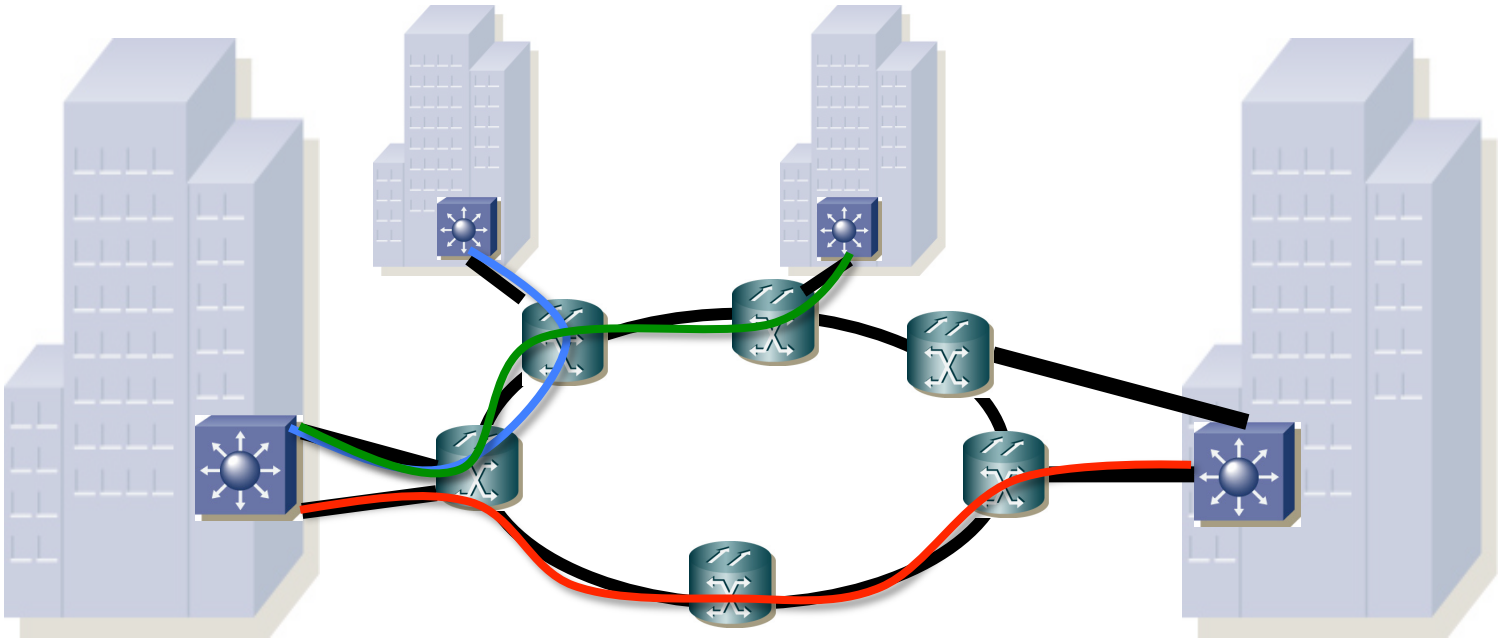
Interconexión por fibra

- Se puede emplear *fibra oscura*
- Puede transportar múltiples wavelengths (CWDM, DWDM)
- O se podría transportar una o varias wavelengths por una red de conmutación óptica
- Esta red puede dar protección
- La distancia sigue limitada pues da continuidad óptica (no hay OEO)
- Y es probable que queramos redundancia en el acceso a ella



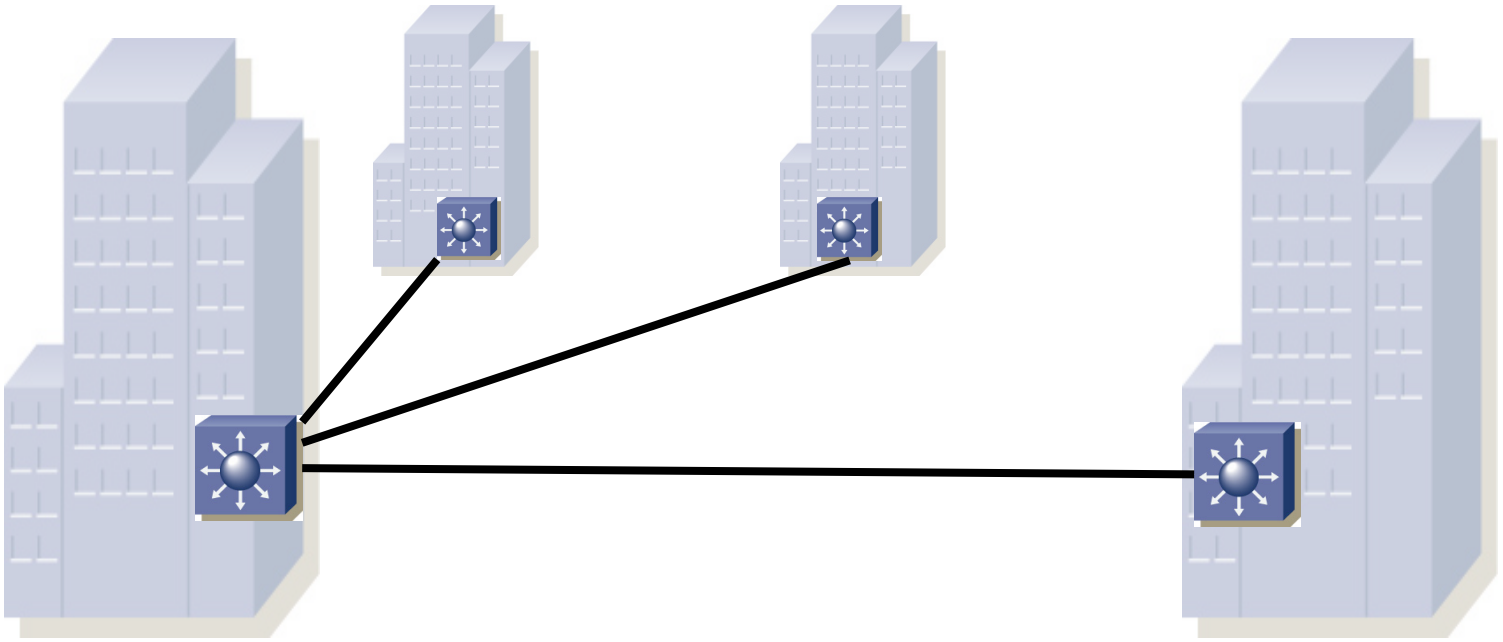
Múltiples *Sites*

- Independientemente de cómo se resuelva el transporte WAN
- ¿Cómo queda la interconexión “física”?
- (…)



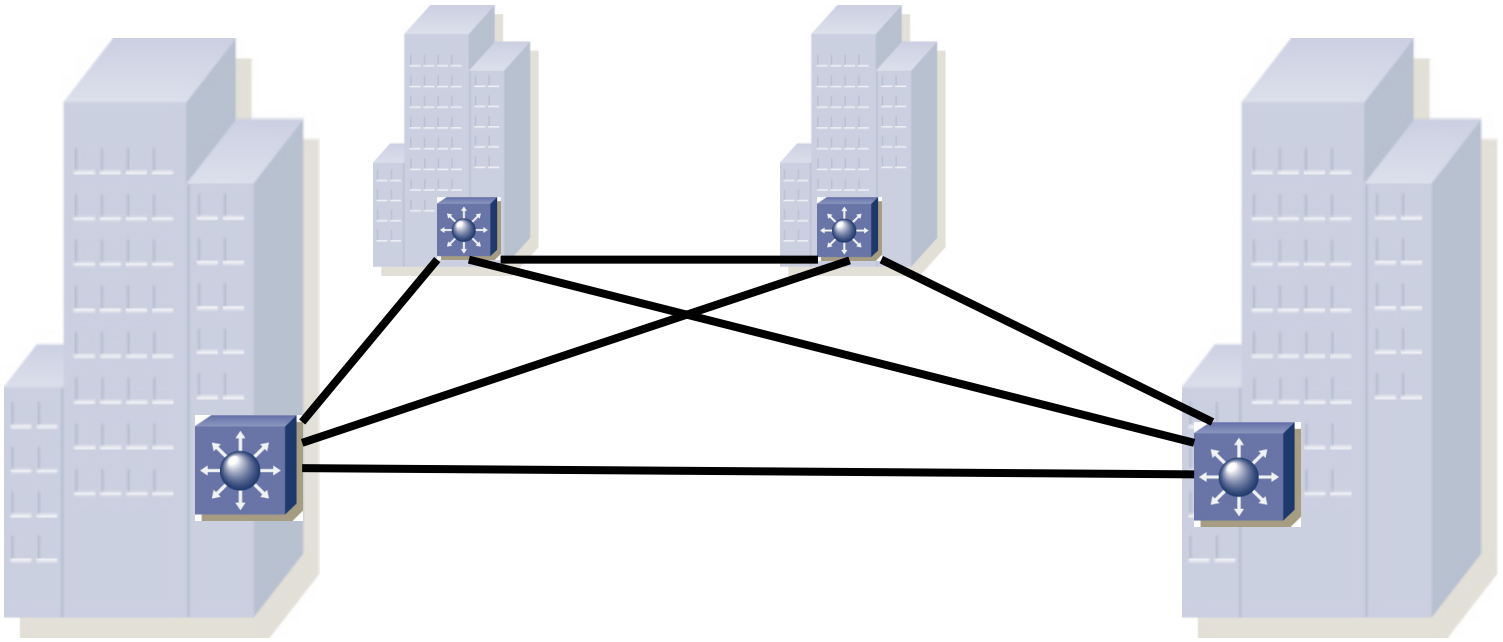
Múltiples *Sites*

- Independientemente de cómo se resuelva el transporte WAN
- ¿Cómo queda la interconexión "física"?
- Podemos tener un esquema *Hub&Spoke*
- (…)



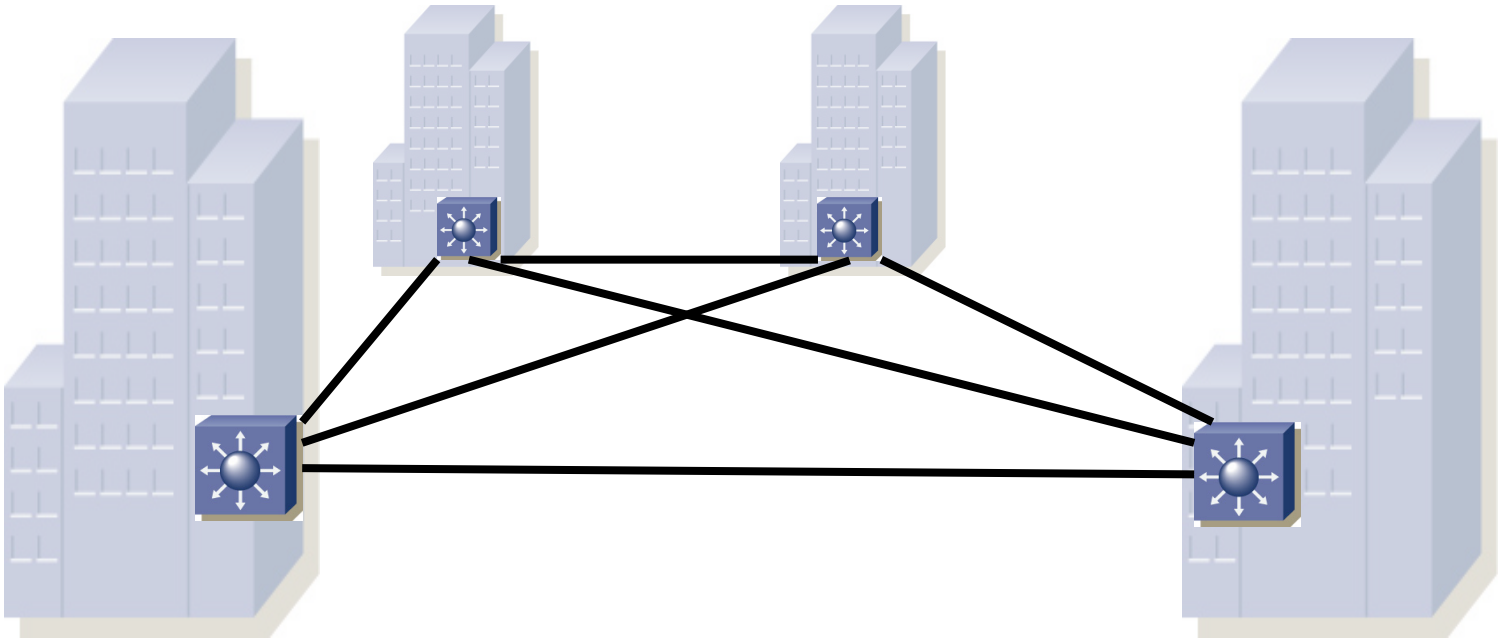
Múltiples *Sites*

- Independientemente de cómo se resuelva el transporte WAN
- ¿Cómo queda la interconexión "física"?
- Podemos tener un esquema *Hub&Spoke*
- También podemos tener un *mesh*
- En este caso hay que resolver esos bucles (STP, SPB, TRILL)
- Estamos hablando de *point-to-point VPNs*



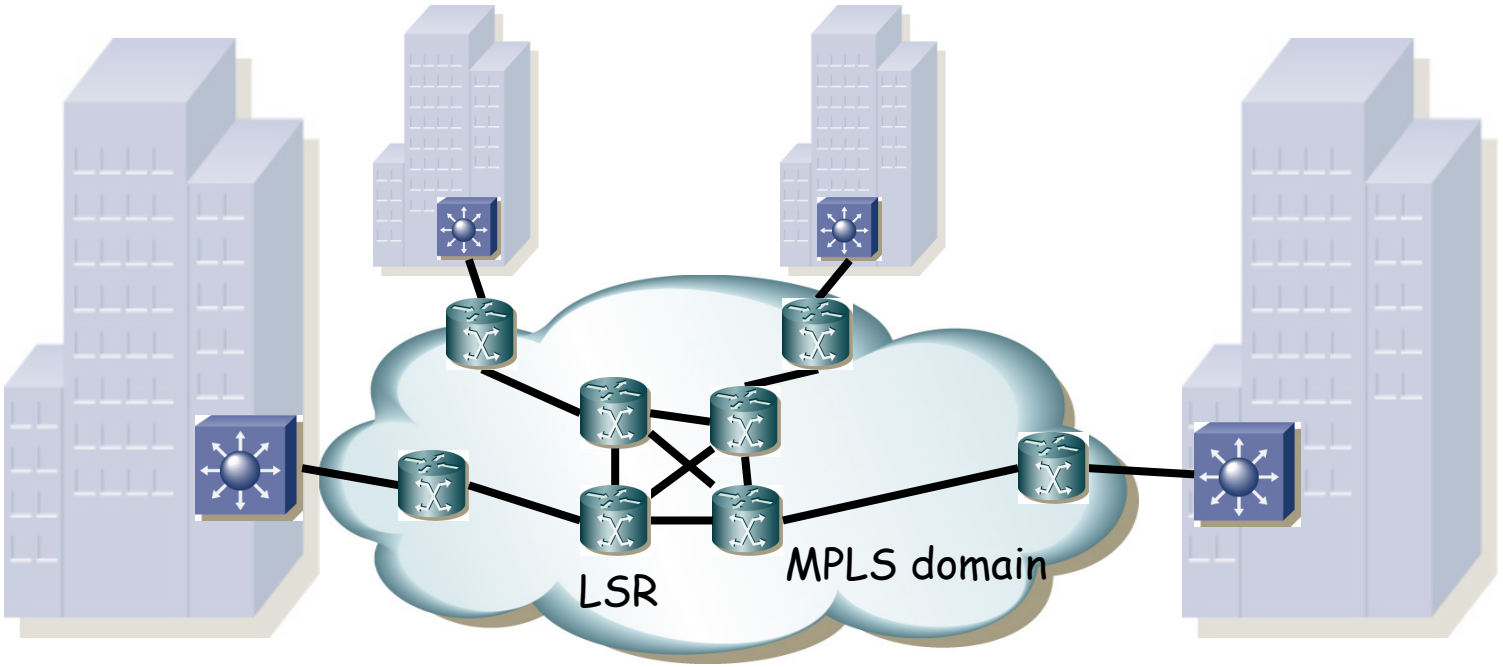
Circuitos

- Estos enlaces, en lugar de wavelengths, pueden ser algún tipo de "circuitos" o "circuitos virtuales"
  - SONET/SDH
  - ATM
  - Frame Relay
- Cualquiera de ellos permite transportar Ethernet o IP
- Serían L2 VPNs o L1 VPNs
- Hoy en día es habitual la solución MPLS



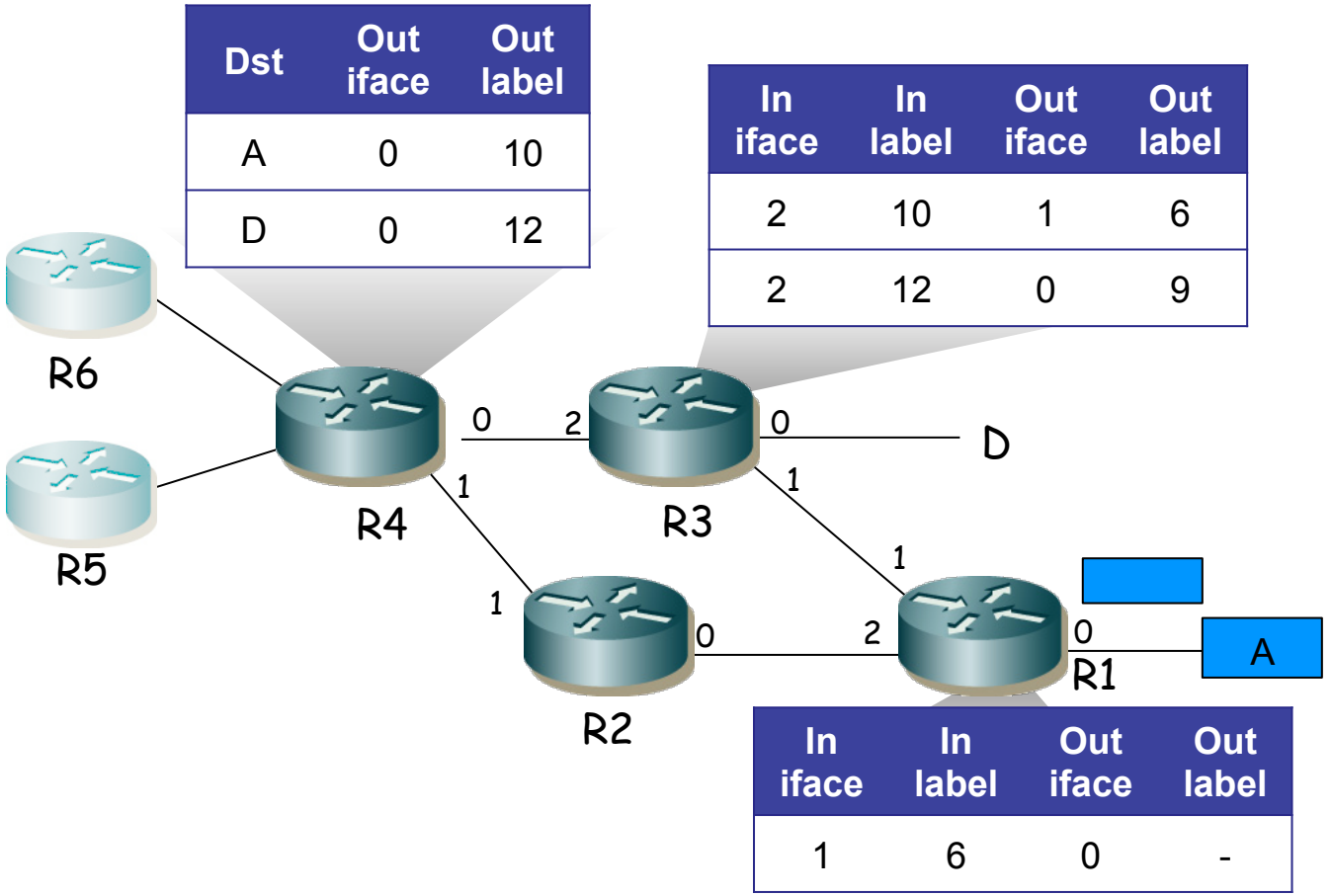
Interconexión MPLS

- En lugar de wavelengths o PVCs tenemos LSPs
- Recordemos que podemos encapsular Ethernet sobre MPLS (EoMPLS)
  - RFC 4448 “Encapsulation Methods for Transport of Ethernet over MPLS Networks”
- De hecho se suele decir que tenemos “AToM” o “Any Transport over MPLS”
- Los equipos de usuario van a poder ser capa 2 o capa 3



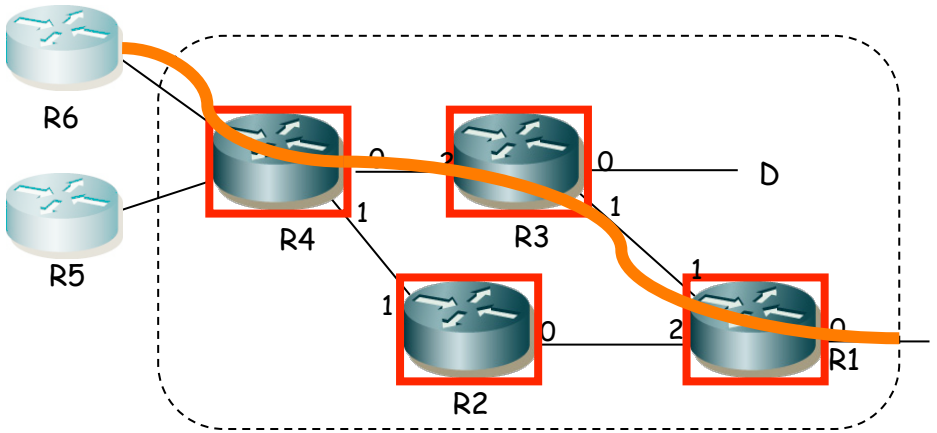
MPLS (recordatorio)

MPLS "forwarding"



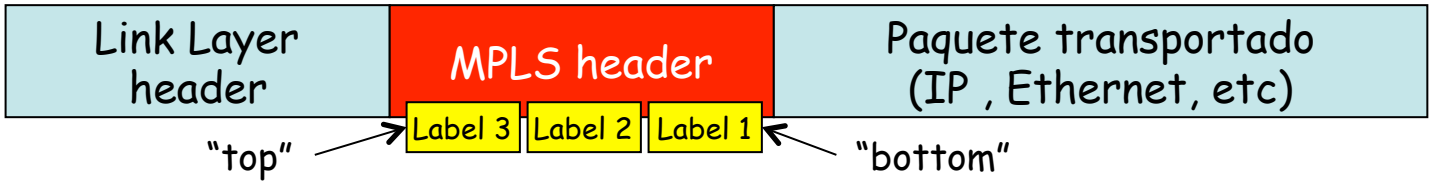
Terminología

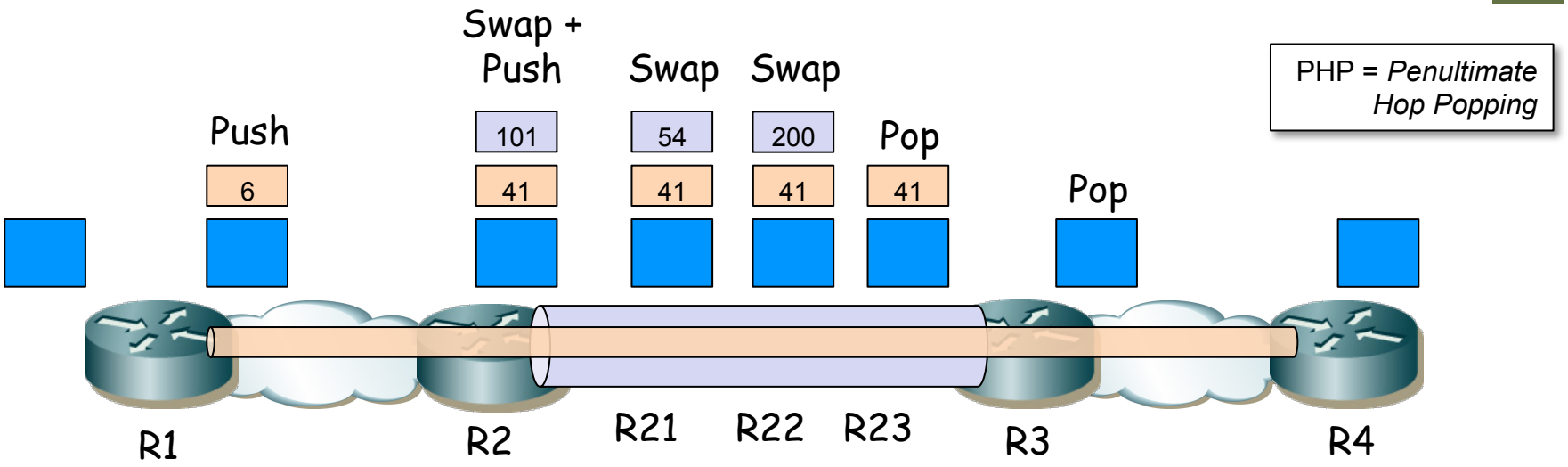
- "MPLS domain": conjunto contiguo de nodos MPLS bajo una misma administración
- "MPLS ingress node": nodo frontera de un dominio en su tarea como entrada de tráfico al mismo
- "MPLS egress node": nodo frontera de un dominio en su tarea como salida de tráfico del mismo
- "Label": etiqueta numérica, corta, longitud fija, identifica a un FEC localmente a un enlace
- "Label Switching Router (LSR)": nodo MPLS capaz de reenviar en base a etiquetas
- "Label Switched Path (LSP)": camino a través de LSRs



Label Stack

- La parte “superior” (“top”) de la pila comienza a continuación de la cabecera de nivel de enlace
- La parte “inferior” (“bottom”) de la pila está junto a la cabecera de nivel de red
- El procesado se basa siempre en la etiqueta exterior (“top”)



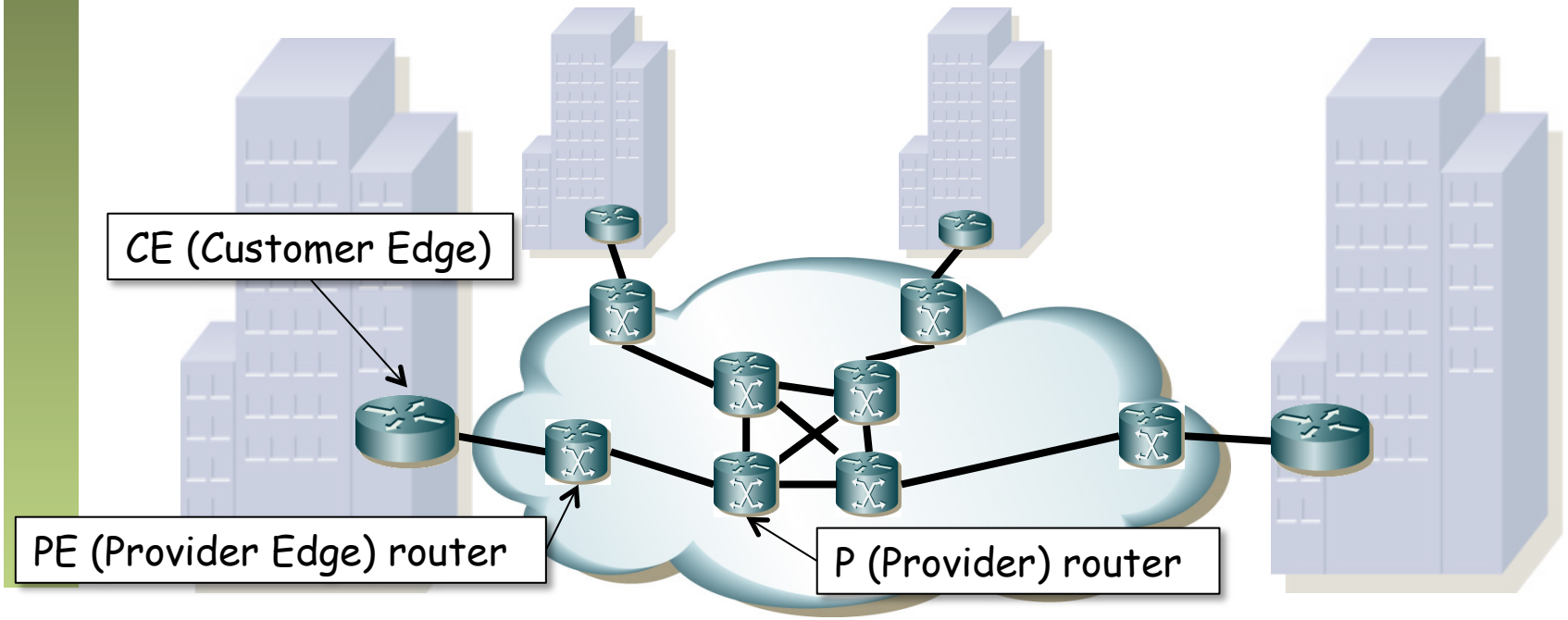


upna Universidad Pública de Navarra Nafarroako Unibertsitate Publikoa
Redes de Nueva Generación
Área de Ingeniería Telemática

Layer 3 MPLS VPNs

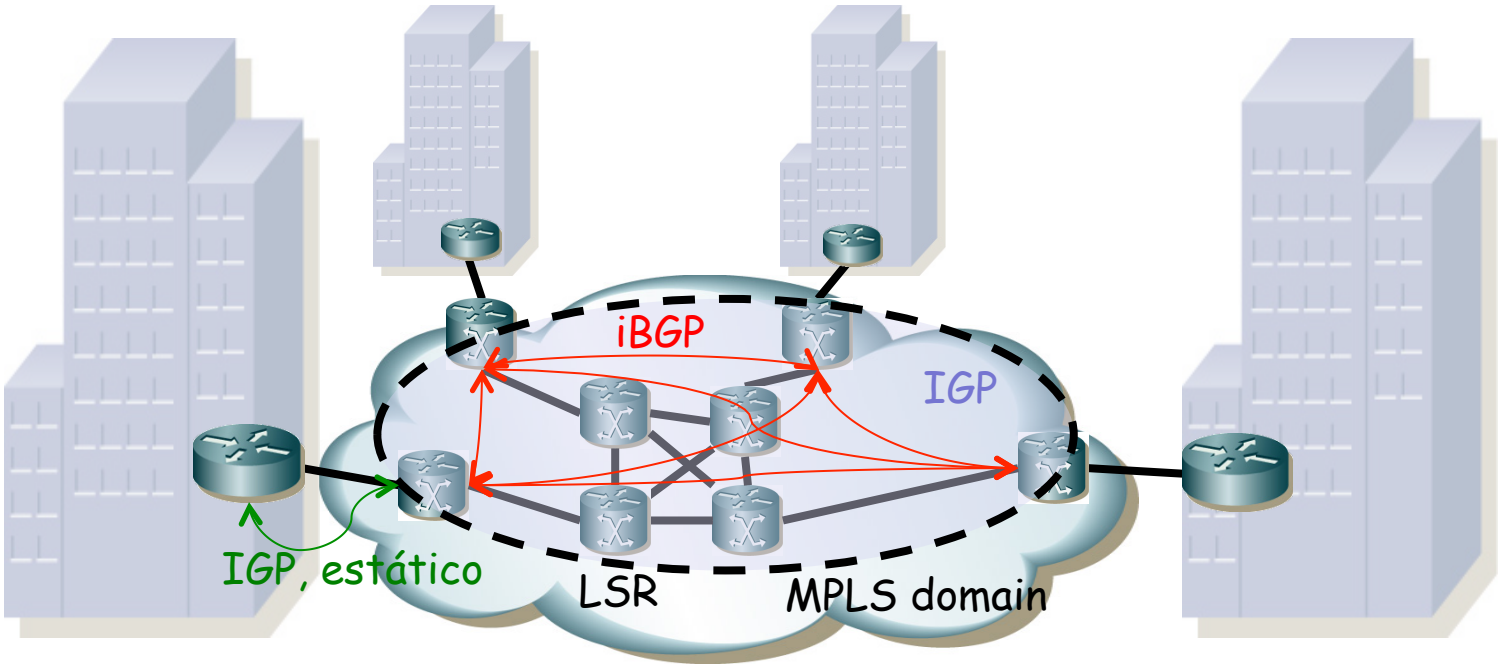
Layer 3 VPNs

- RFC 4364 “BGP/MPLS IP Virtual Private Networks (VPNs)”
- VPN para el transporte de paquetes IP entre sedes
- El backbone del proveedor de servicio es una red IP MPLS
- RFC 4760 “Multiprotocol Extensions for BGP-4”
- 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”



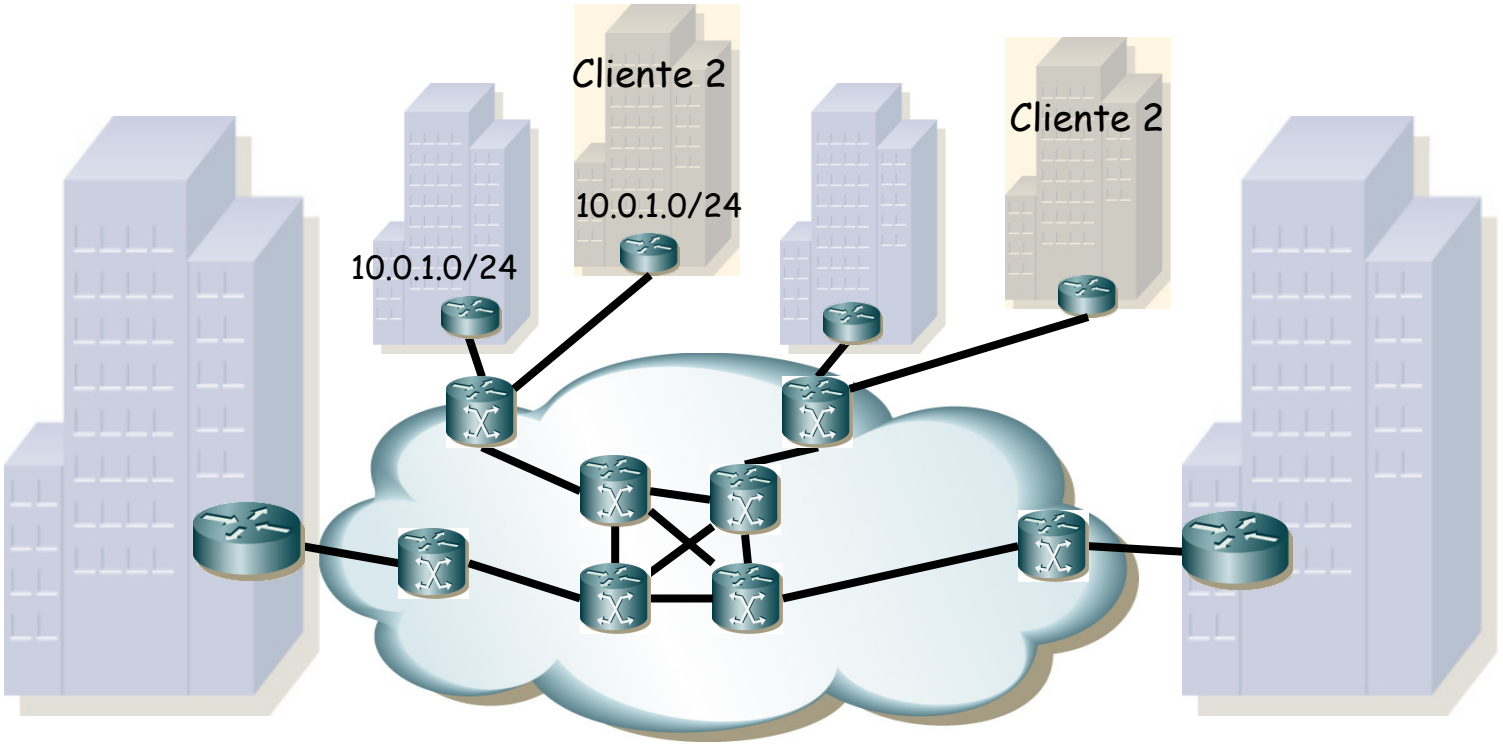
L3VPN: Routing

- 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 CEs 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



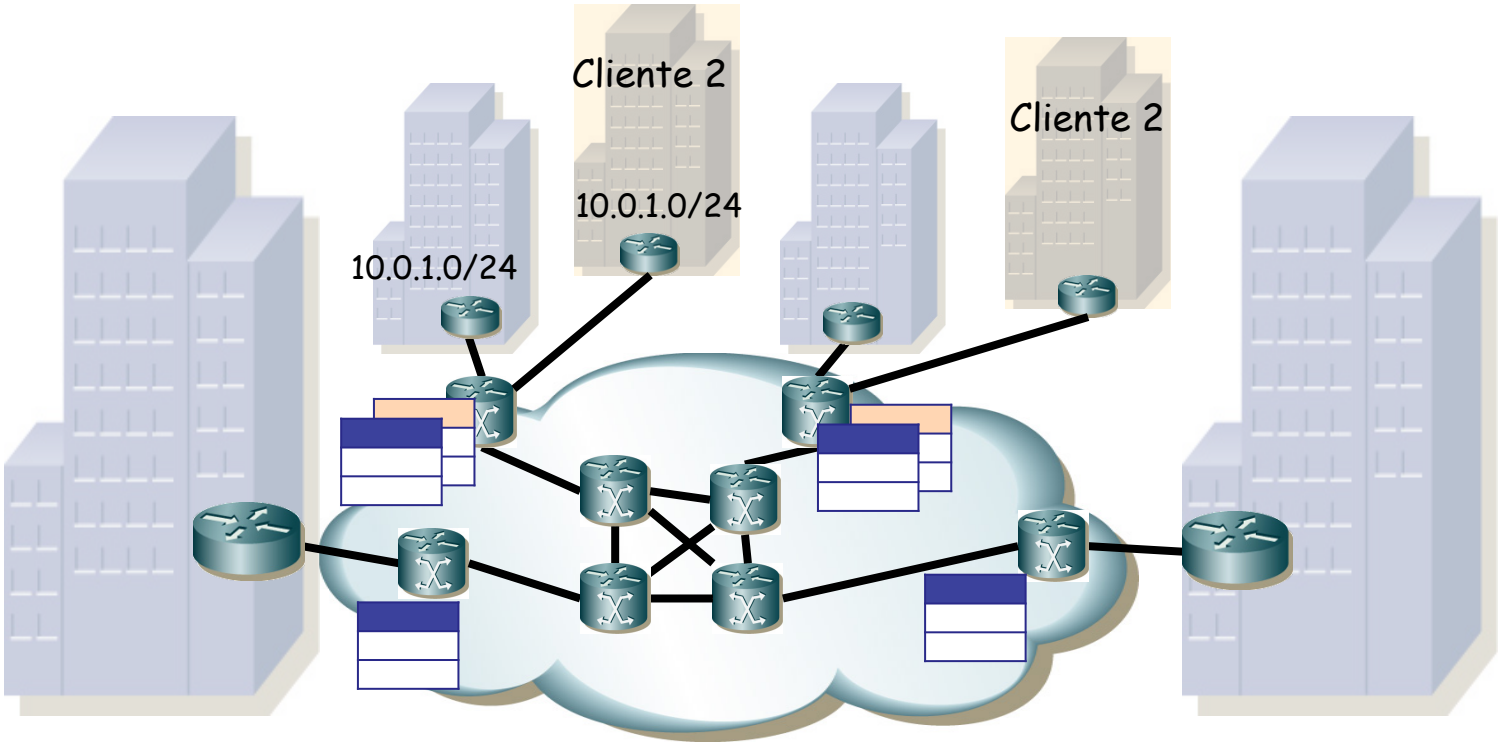
L3VPN: Routing

- Dos VPNs pueden emplear espacios de direcciones IP que se solapen
- Los anuncios VPN-IPv4 de esas subredes mediante BGP incluyen un identificador (*Route Distinguisher = RD, 8 bytes)* que las diferencia
- Cada service provider tiene su espacio de valores RD
- 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?



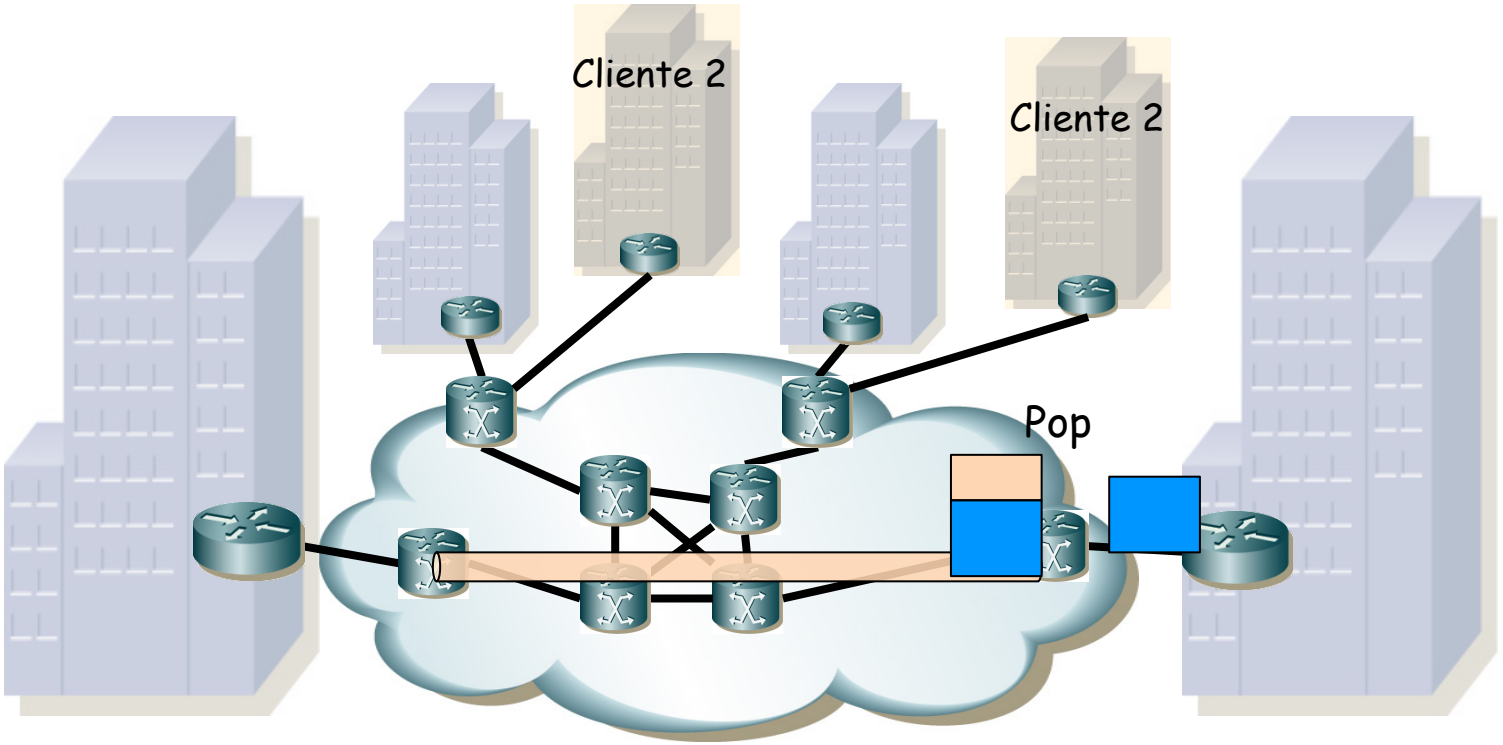
L3VPN: Forwarding

- Cada PE mantiene una tabla de rutas para cada VPN o *VPN/Virtual Routing and Forwarding tables (VRFs) y a*demás una tabla por defecto
- Cara VRF está asociada a un valor o más de *"Route Target" (RT)*
- Al recibir un paquete IP de un cliente consulta la VRF correspondiente
- Las rutas VPN-IPv4 se anuncian con un valor de RT a todas las VRF con ese valor (plano de control)
- Incluye una etiqueta MPLS (para el plano de datos)



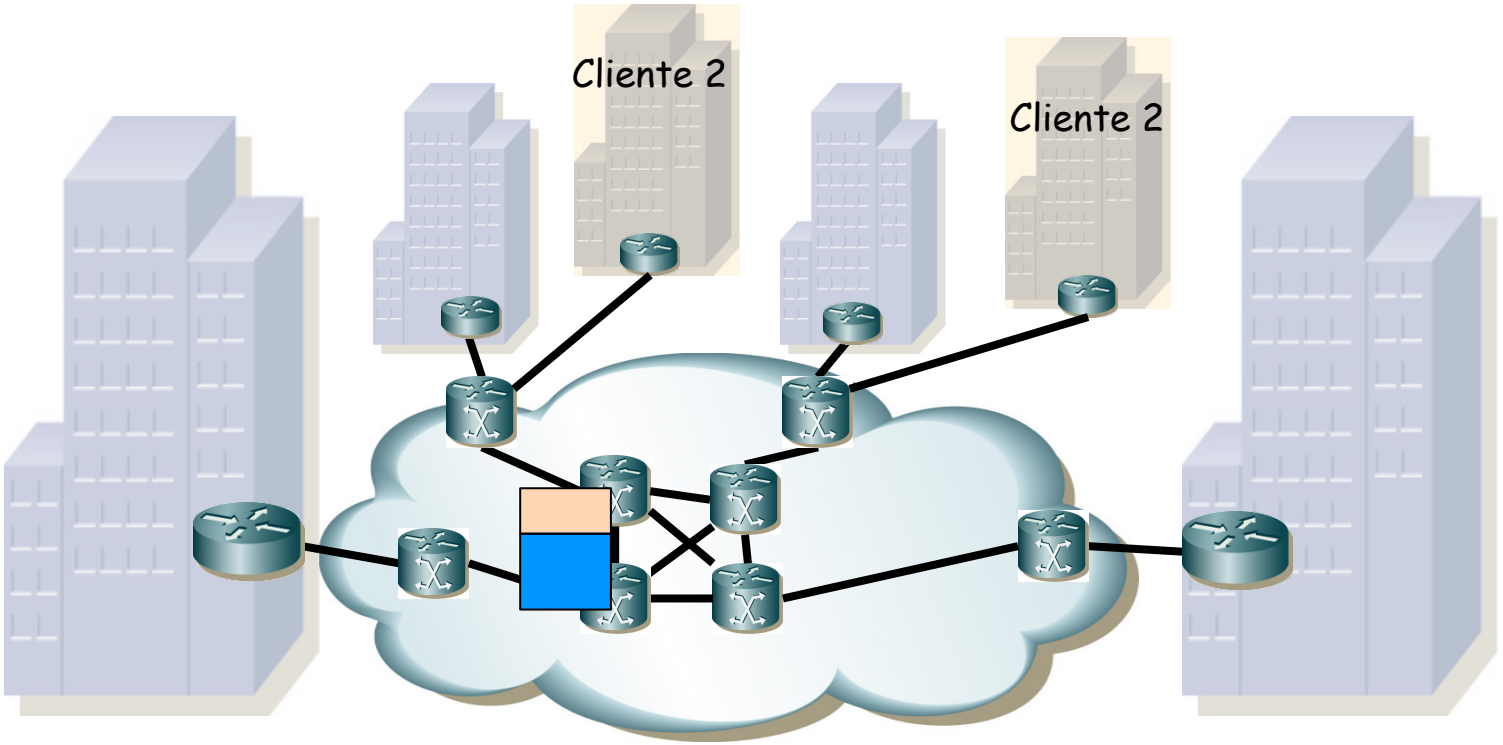
L3VPN: MPLS

- ¿Para qué esa etiqueta?
- Para que el PE de salida sepa a qué VRF pertenece el paquete
- No puede basarse en la dirección IP destino pues pueden estar duplicadas



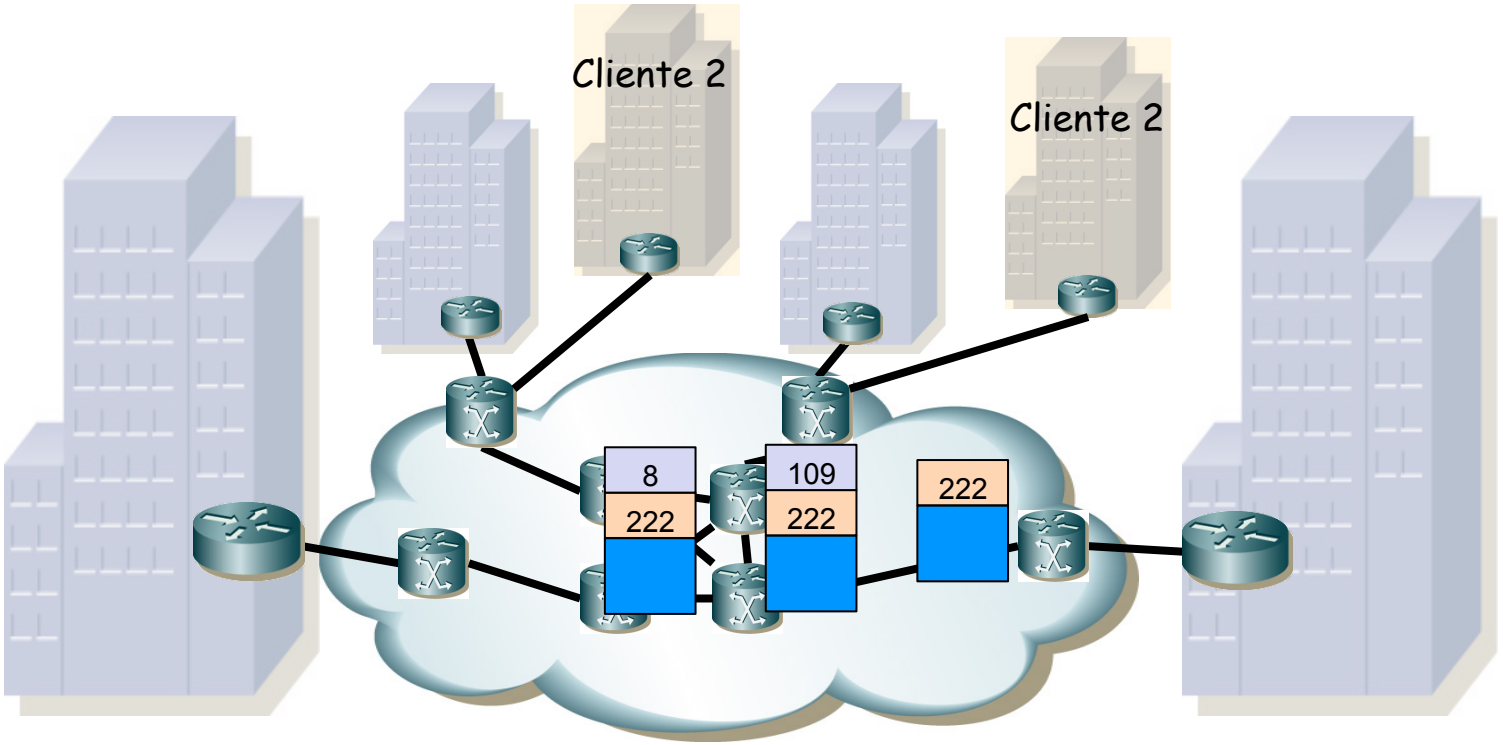
L3VPN: MPLS

- ¿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
- (…)



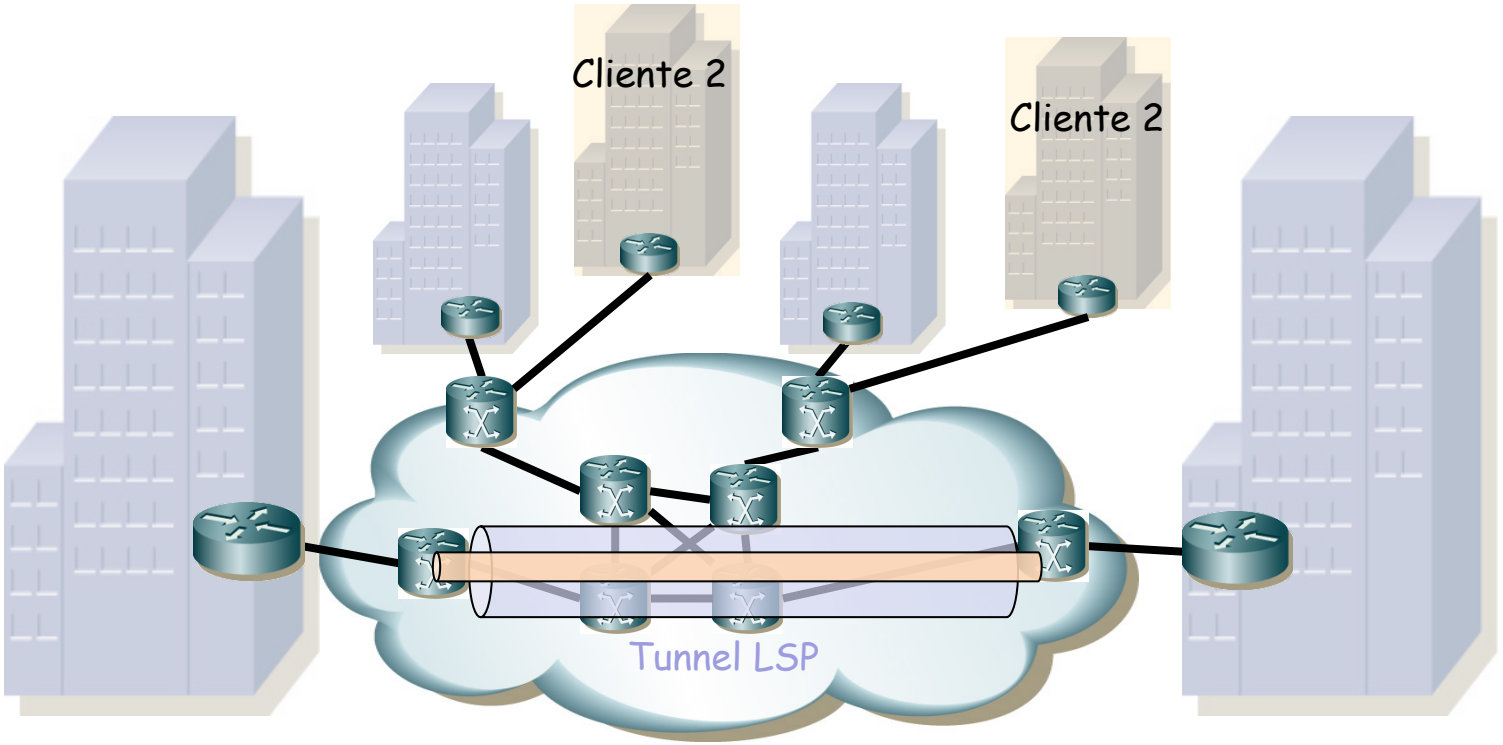
L3VPN: MPLS

- ¿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

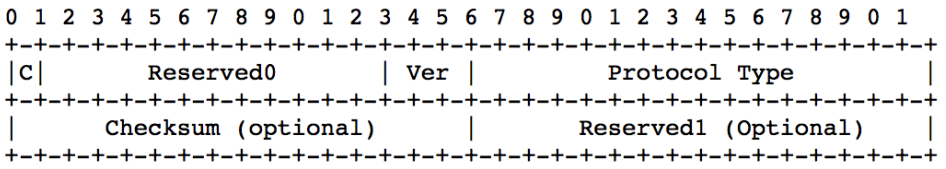
- ¿Y en los P routers? ¿Reenvían en función de esa etiqueta?
- Se crean LSPs entre los PEs (“Tunnel LSPs”)
- Los P routers reenvían en función de esa etiqueta externa
- Un full-mesh entre los PEs que compartan VRF
- Podrían ser otro tipo de túneles (GRE o IP en IP, RFC 4797), lo cual elimina el requerimiento de una red de transporte MPLS

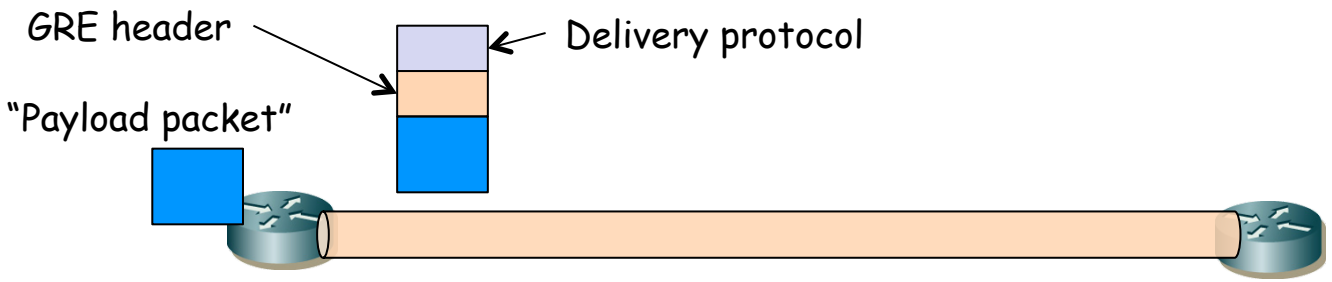


GRE

GRE

- RFC 2784 “Generic Routing Encapsulation (GRE)”
- Encapsular un nivel de red en otro nivel de red
- PPTP (Point-to-Point tunneling Protocol) usa algo similar a GRE
- La cabecera básica GRE ocupa 8 bytes
- Uno de los campos es un Ethertype (*Protocol Type*)
- La versión anterior (RFC 1701) tenía más campos que desaparecen en esta
- Aunque algunos se recuperan en la RFC 2890 “Key and Sequence Number Extensions to GRE” (…)

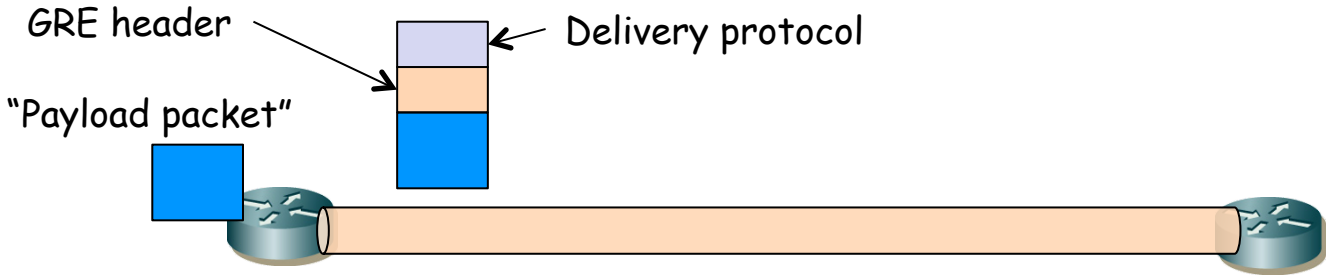




GRE

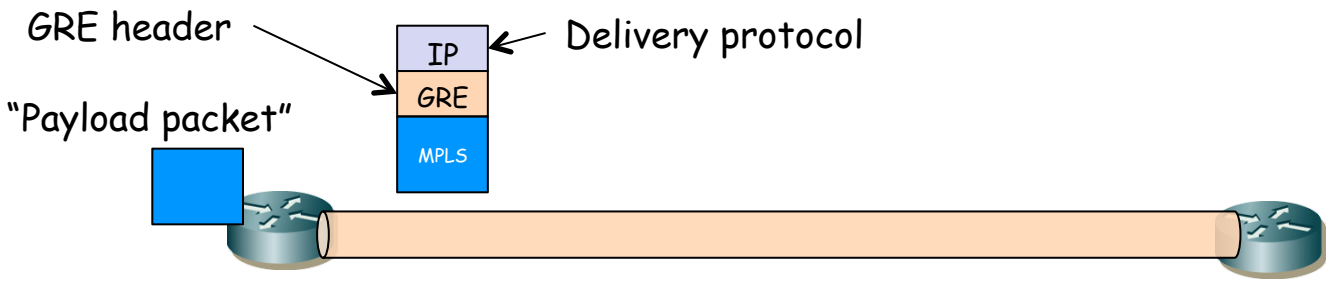
- RFC 2890 “Key and Sequence Number Extensions to GRE”
- “Key” sirve para distinguir flujos dentro del túnel
- “Sequence Number”
  - Si hay “key” entonces el número de secuencia es por “key”
  - Permite dar entrega en orden (aunque no fiable)
  - Si llega uno “anterior” lo descarta
  - Si llega uno que deja un hueco puede guardarlo intentando reconstruir la secuencia
  - Pasado cierto tiempo sin lograr reconstruir los reenvía

```
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|C| |K|S| Reserved0       | Ver |         Protocol Type         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|      Checksum (optional)      |       Reserved1 (Optional)    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                          Key (optional)                       |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                  Sequence Number (Optional)                   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```



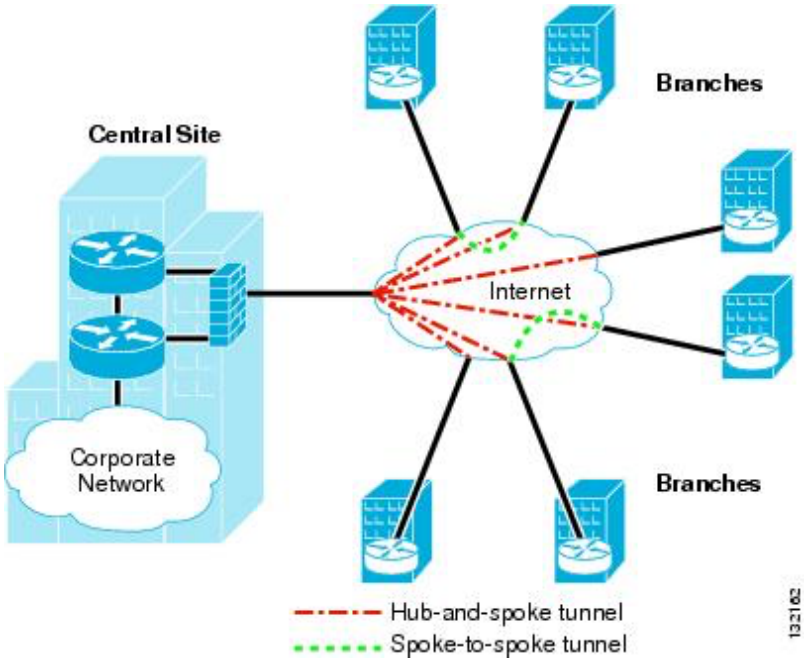
MPLS in GRE in IP

- RFC 4023 "Encapsulating MPLS in IP or Generic Ruting Encapsulation (GRE)"
- 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)



mGRE y DMVPN

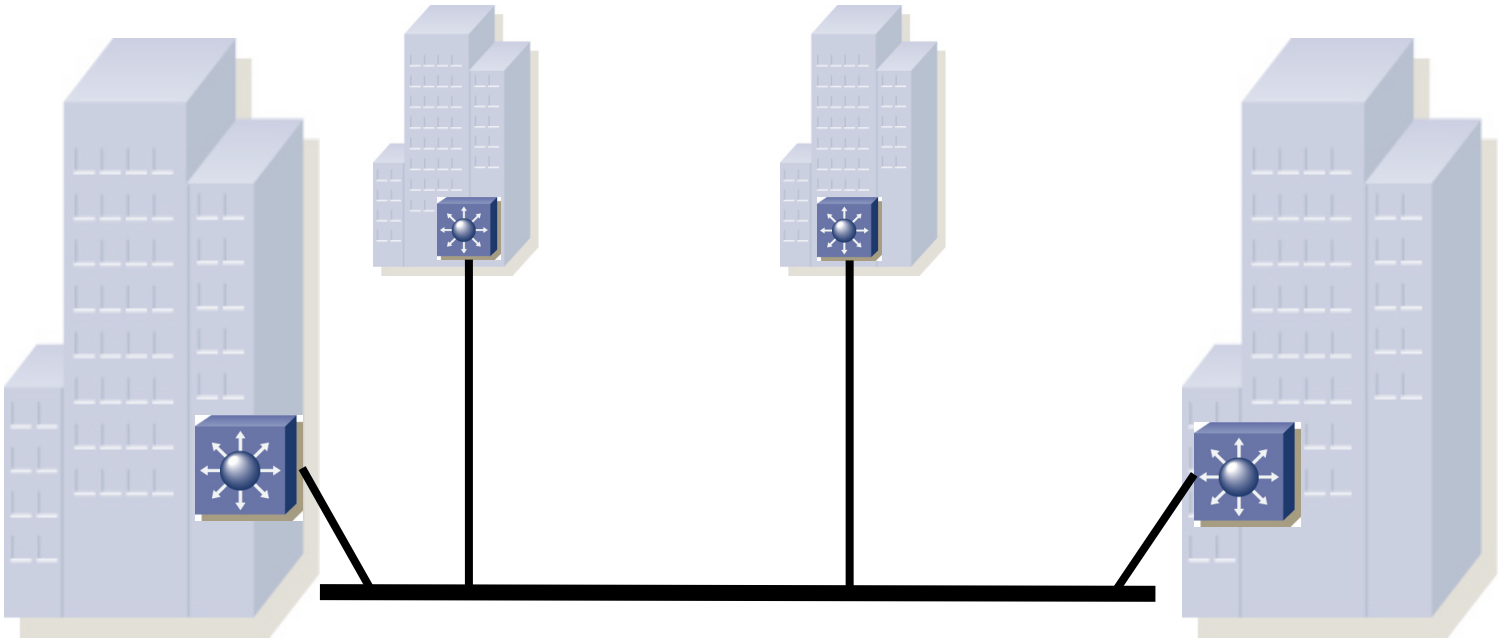
- mGRE = Multipoint GRE
- DMVPN = Dynamic Multipoint VPN
- Solución propietaria de Cisco



VPLS

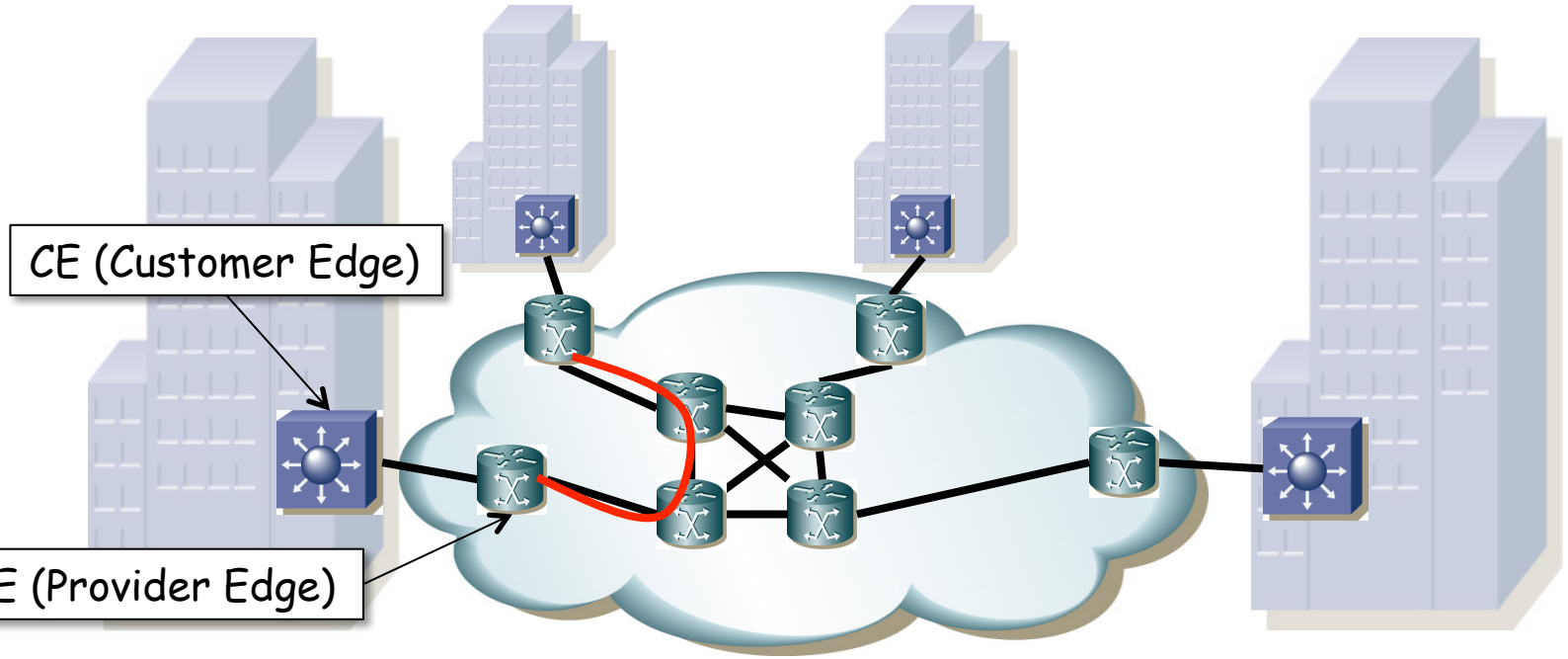
MPLS y VPLS

- *"Virtual Private LAN Service",* una VPN layer 2
- Interconecta múltiples *sites* en un solo dominio puenteado
- Todos los extremos se comportan como si estuvieran en una LAN
- 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)



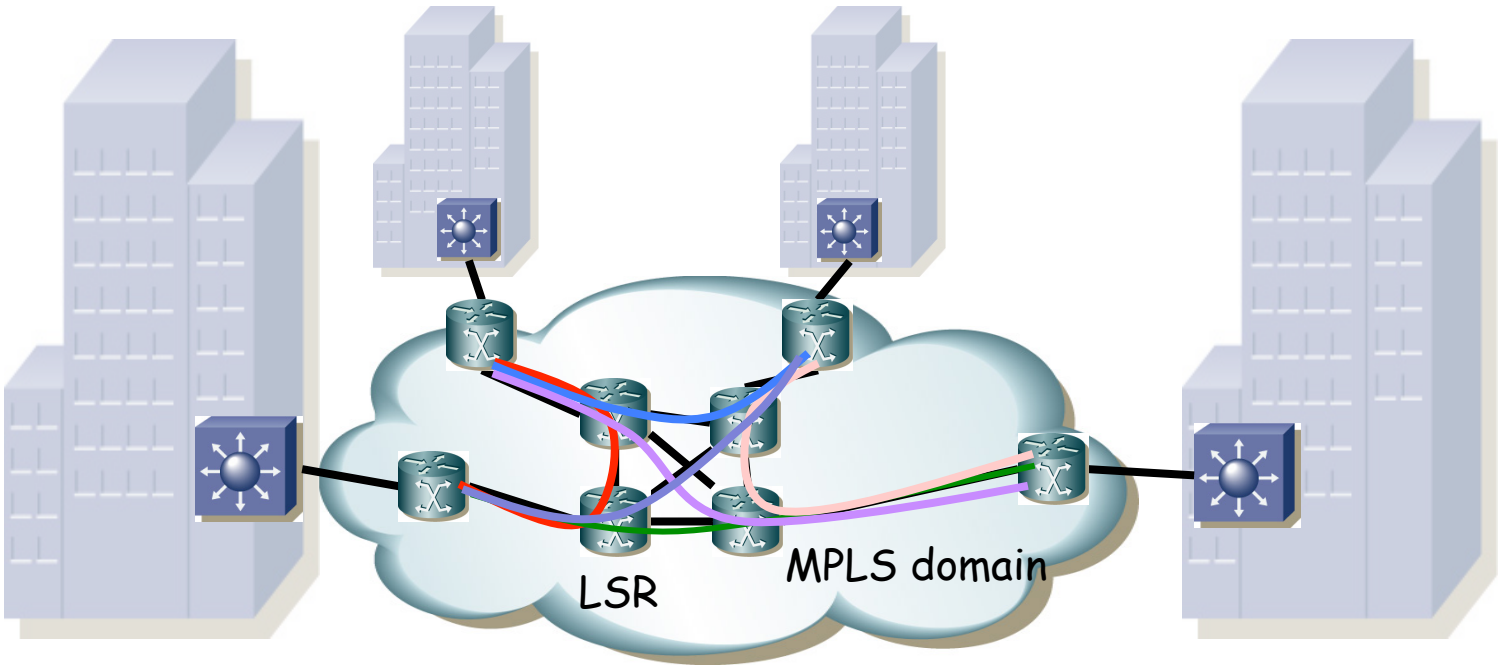
MPLS y VPLS

- 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



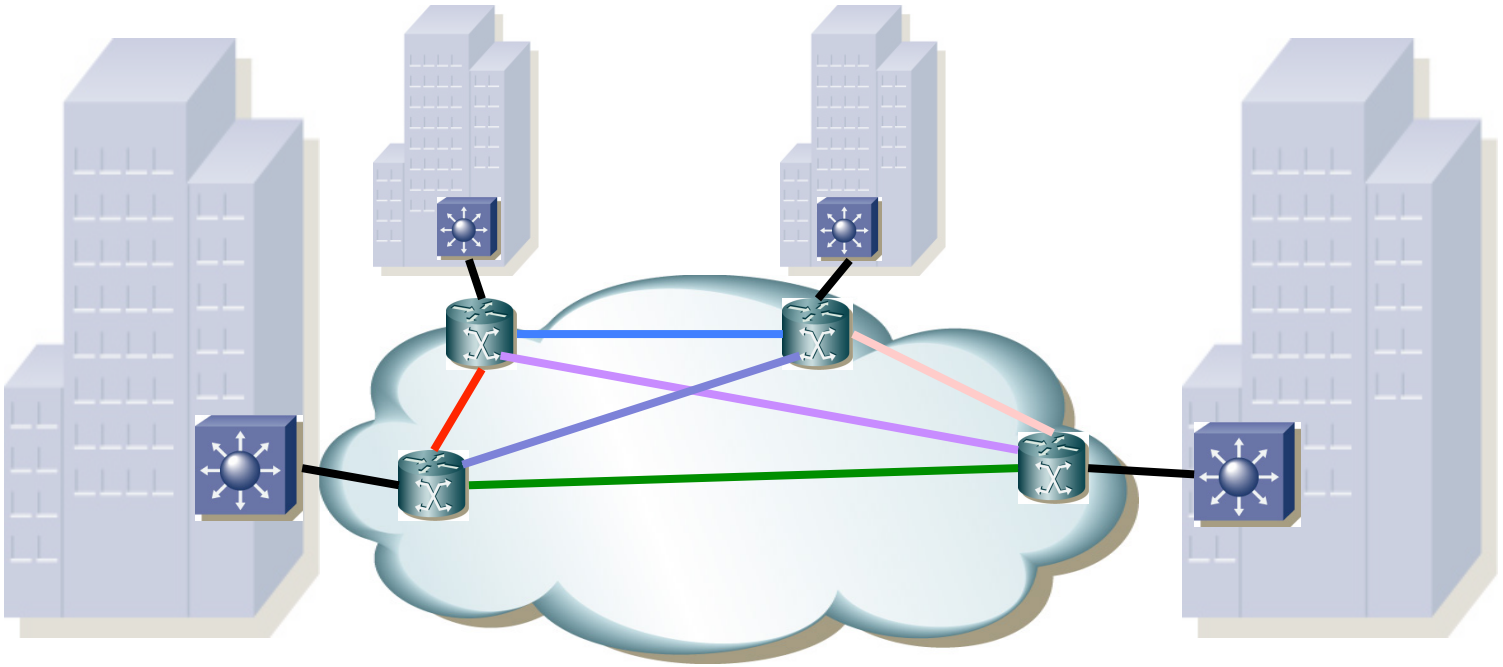
MPLS y VPLS

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



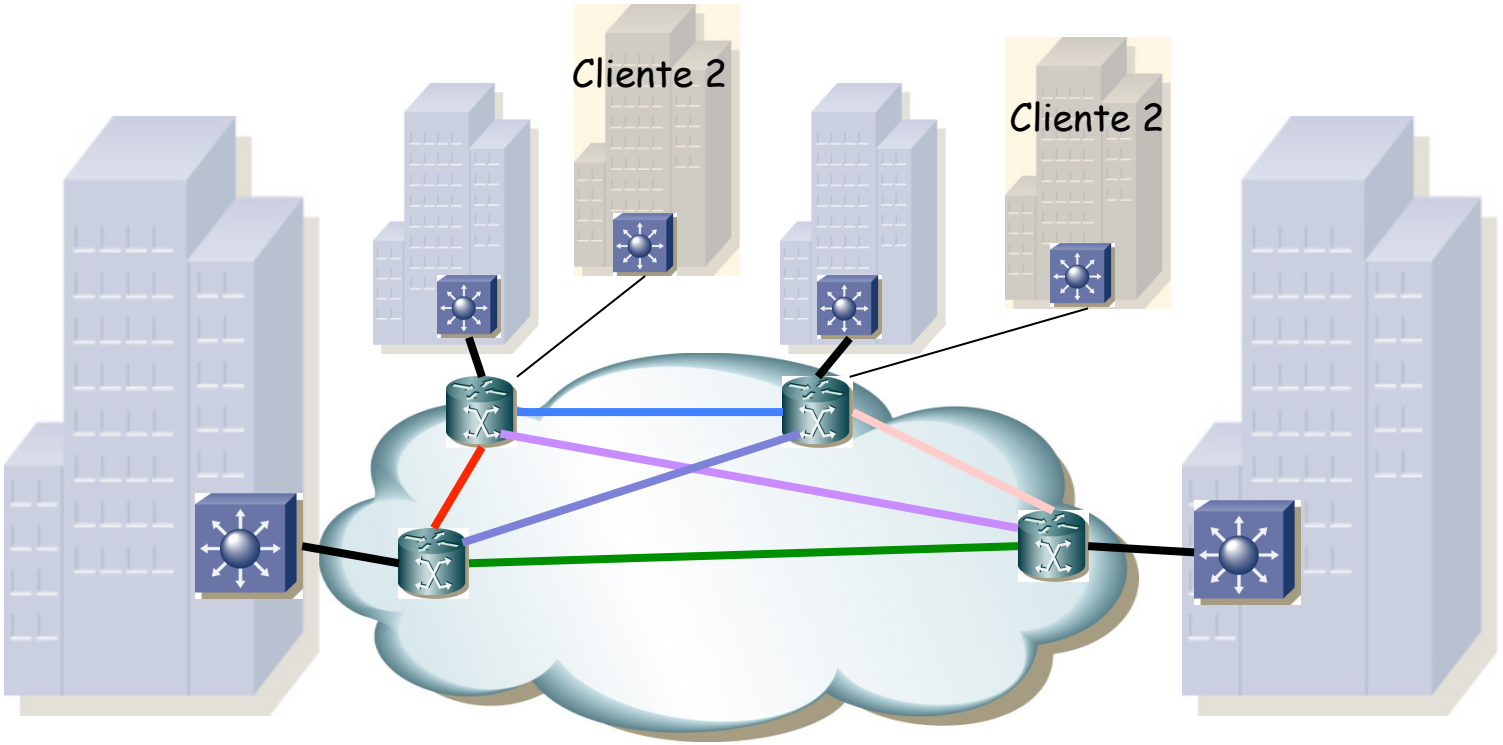
MPLS y VPLS

- 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 (…)



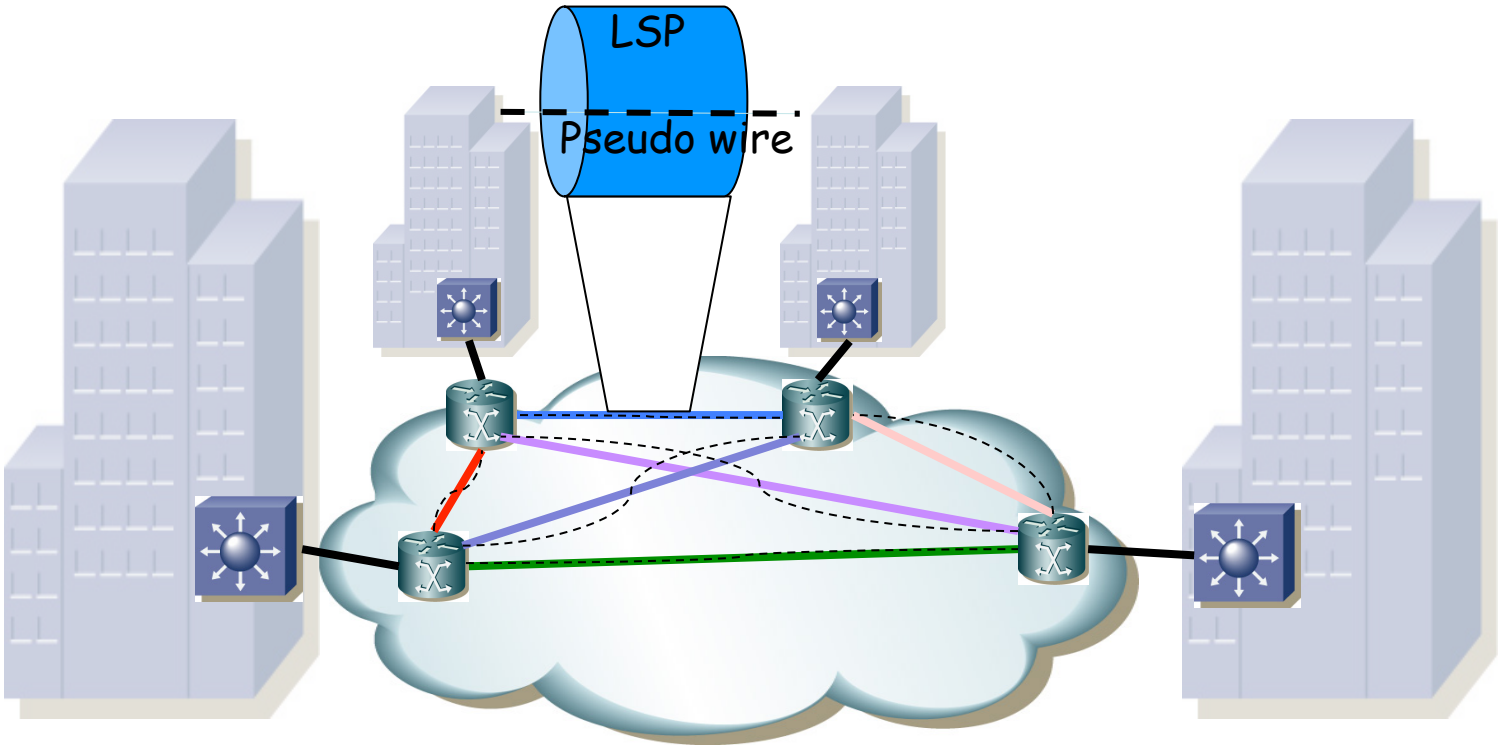
MPLS y VPLS

- 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?



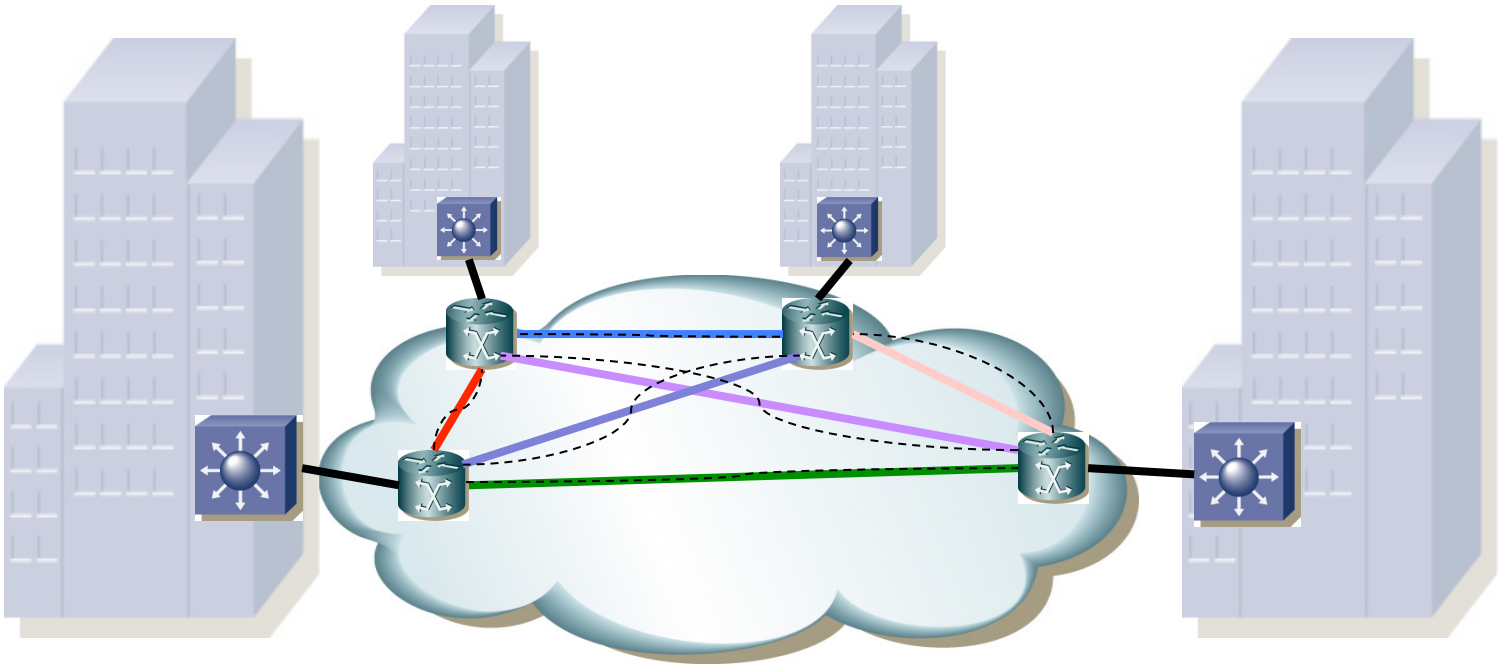
VPLS y PWE

- Por cada instancia VPLS (cada cliente) se establece un full mesh de *pseudo-wires* (PWs) entre los PEs
- RFC 3985 "Pseudo Wire Emulation Edge-to-Edge (PWE3) Architecture"
- Un PW emula un circuito, por ejemplo para transportar un E1 o un PVC ATM
- También puede transportar Ethernet, AAL5, SDH, etc



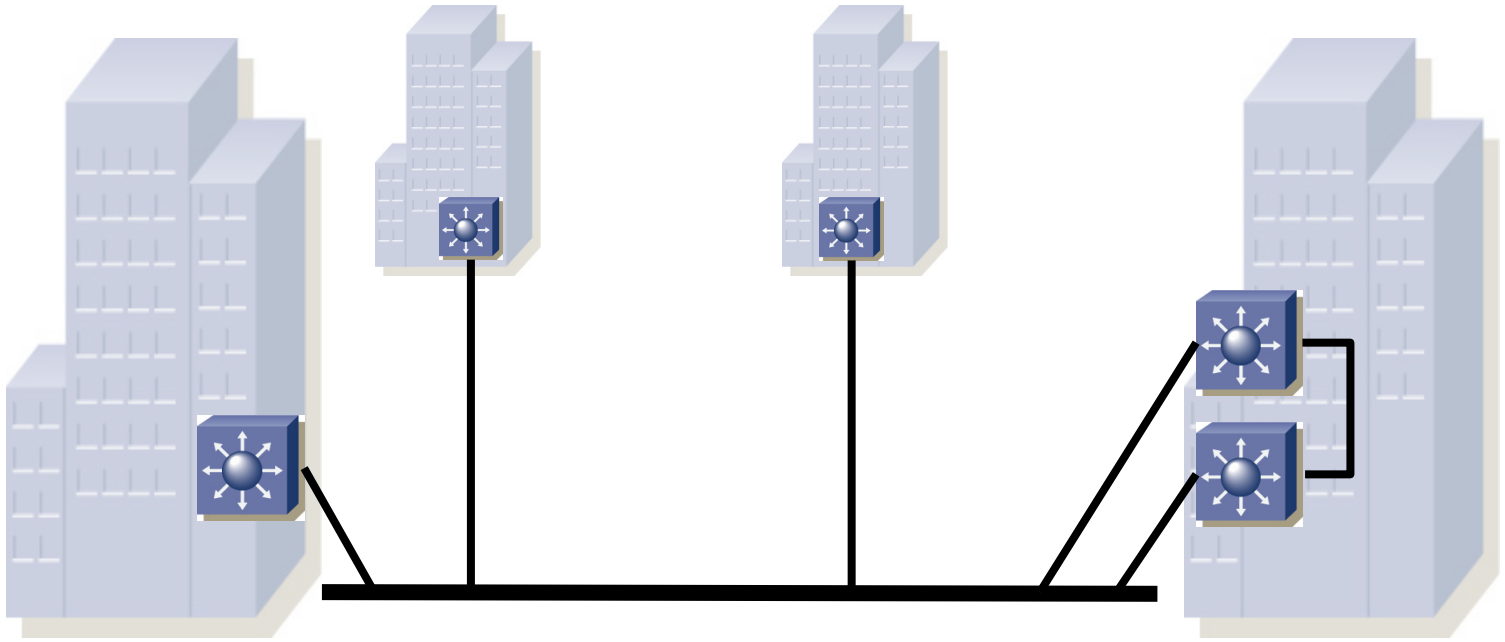
VPLS y PWE

- El full-mesh de PWs hace que los PE puedan enviarse directamente los unos a los otros
- Es decir, no necesiten hacer reenvío
- Así 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



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

- Dos 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”

Problemas en VPLS

- Se deben establecer Nx(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)
- 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