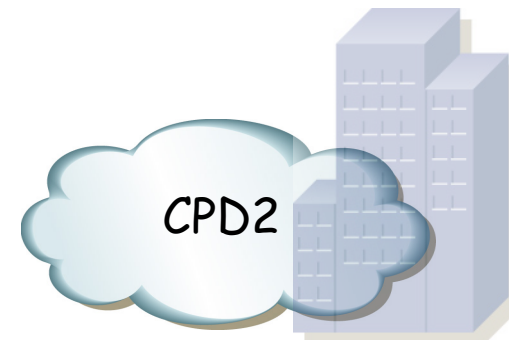
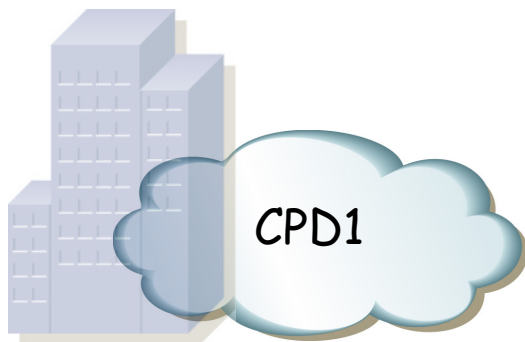


Interconexión de DCs: Introducción

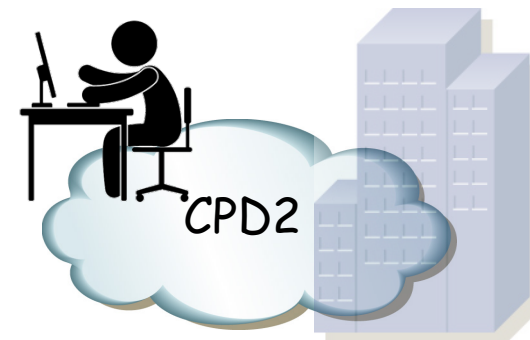
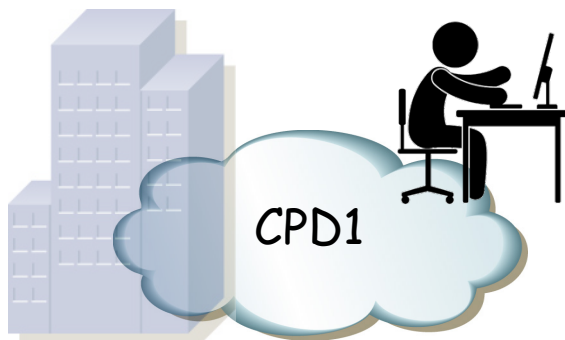
Múltiples DCs

- 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



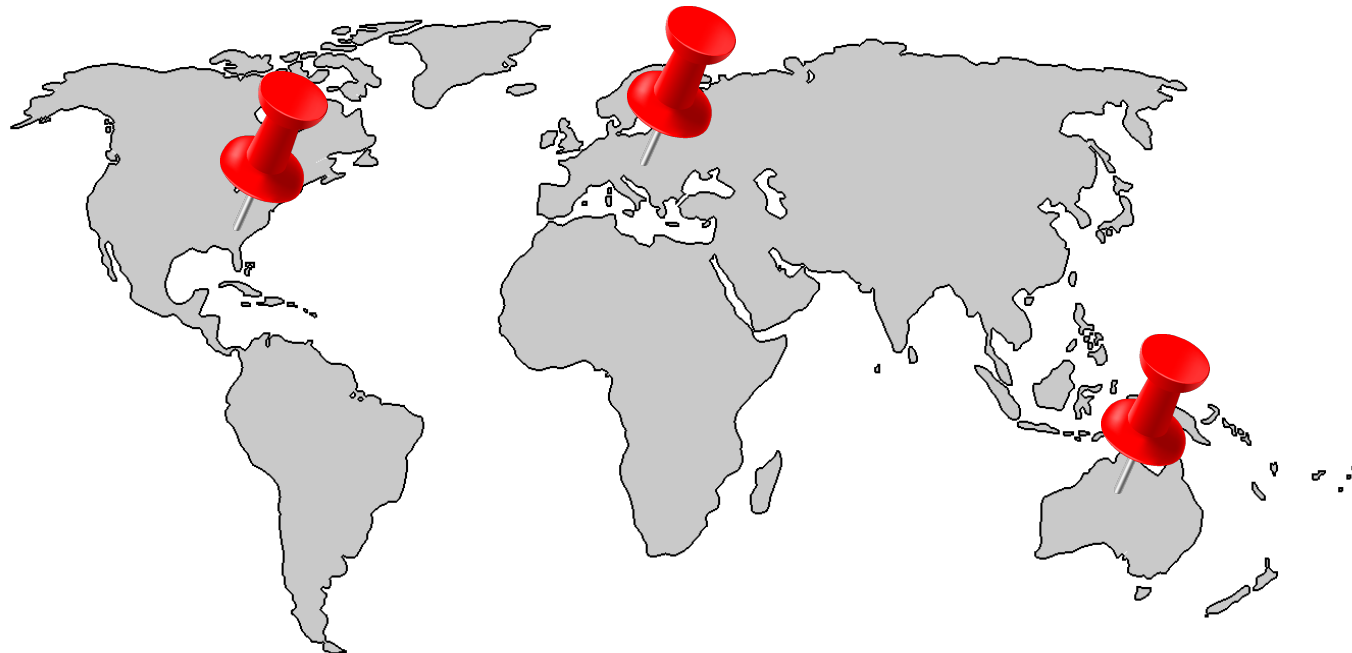
Múltiples DCs

- 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)



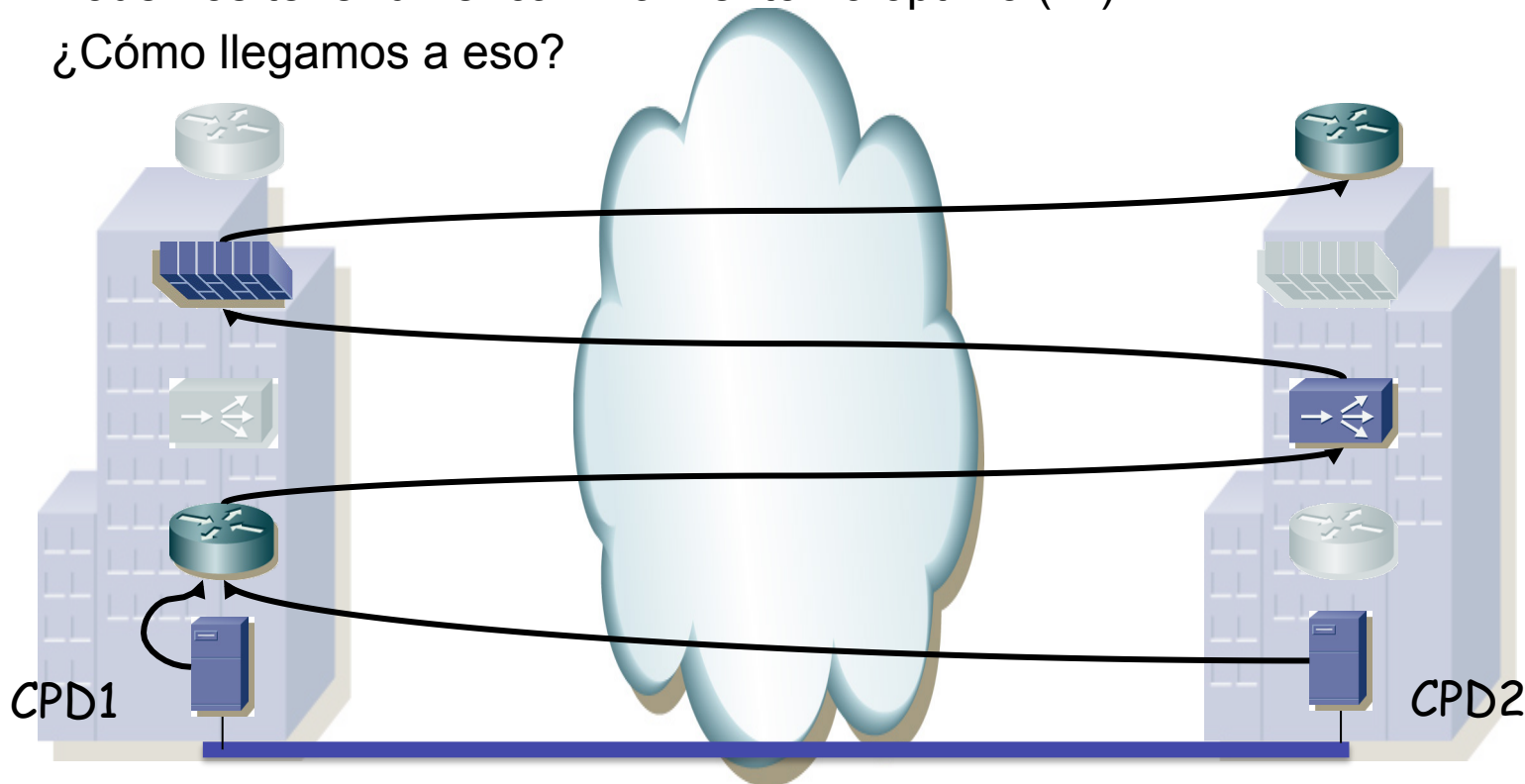
Múltiples DCs o sedes

- 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



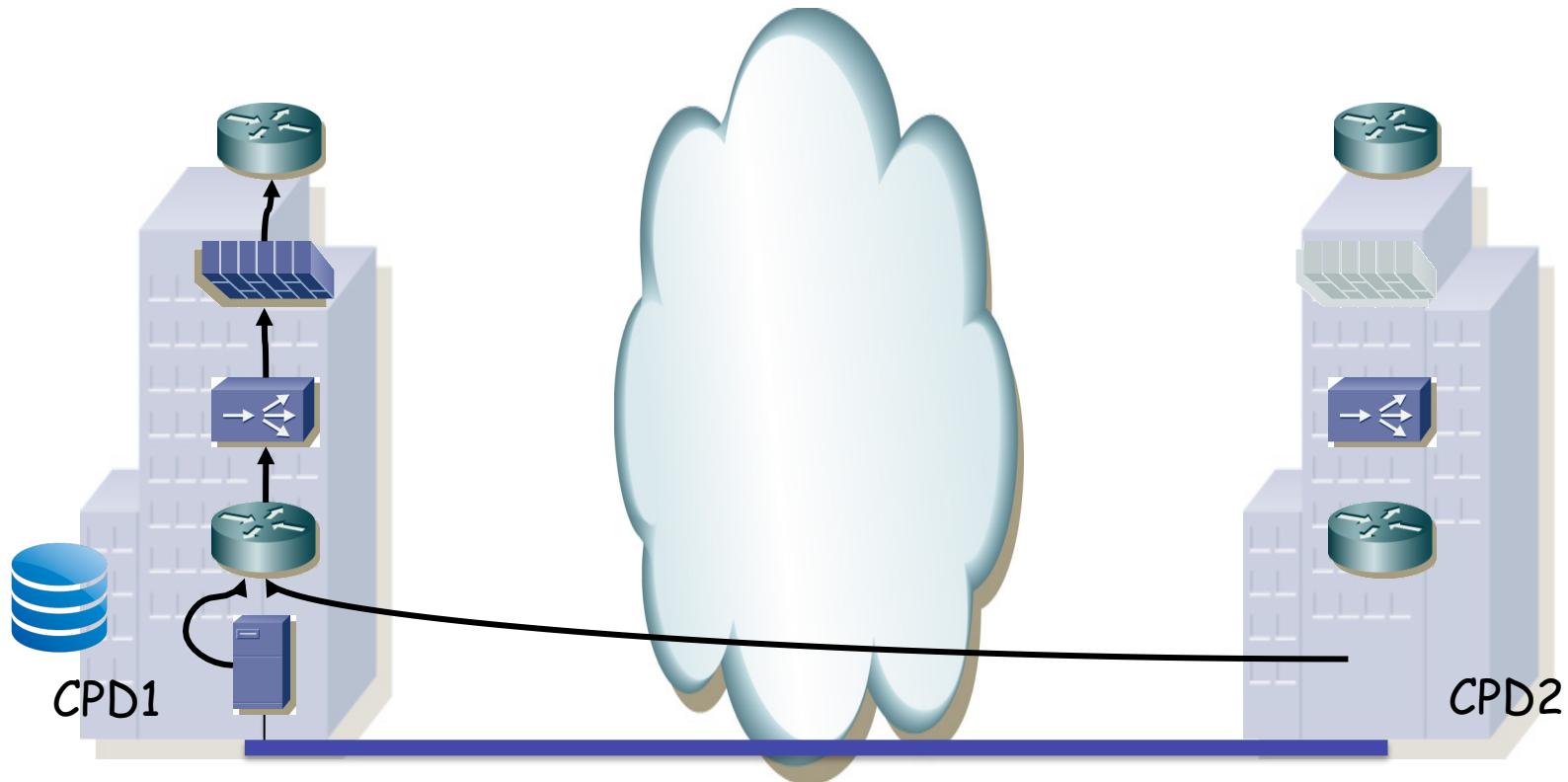
Problemas con extensión L2

- 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 (...)
- Puede ser el servidor accediendo a almacenamiento local a su DC, ahora va al otro DC



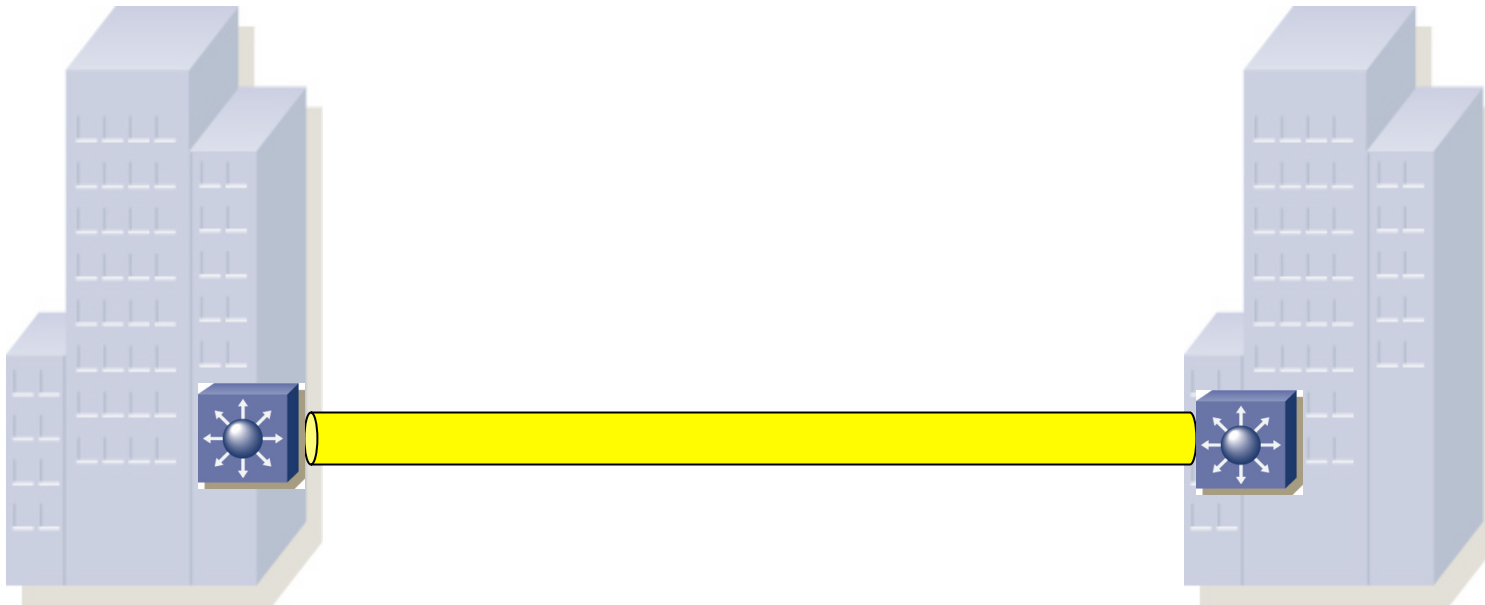
Interconexión del *storage*

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



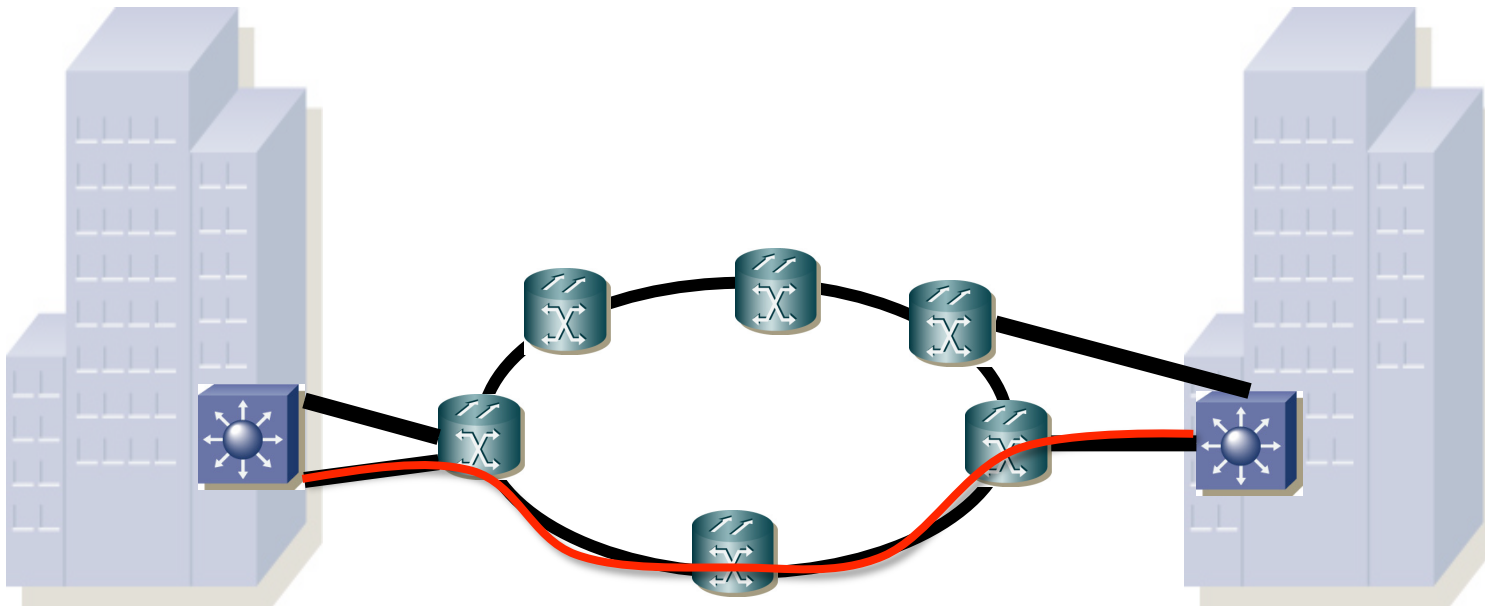
Interconexión por fibra

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



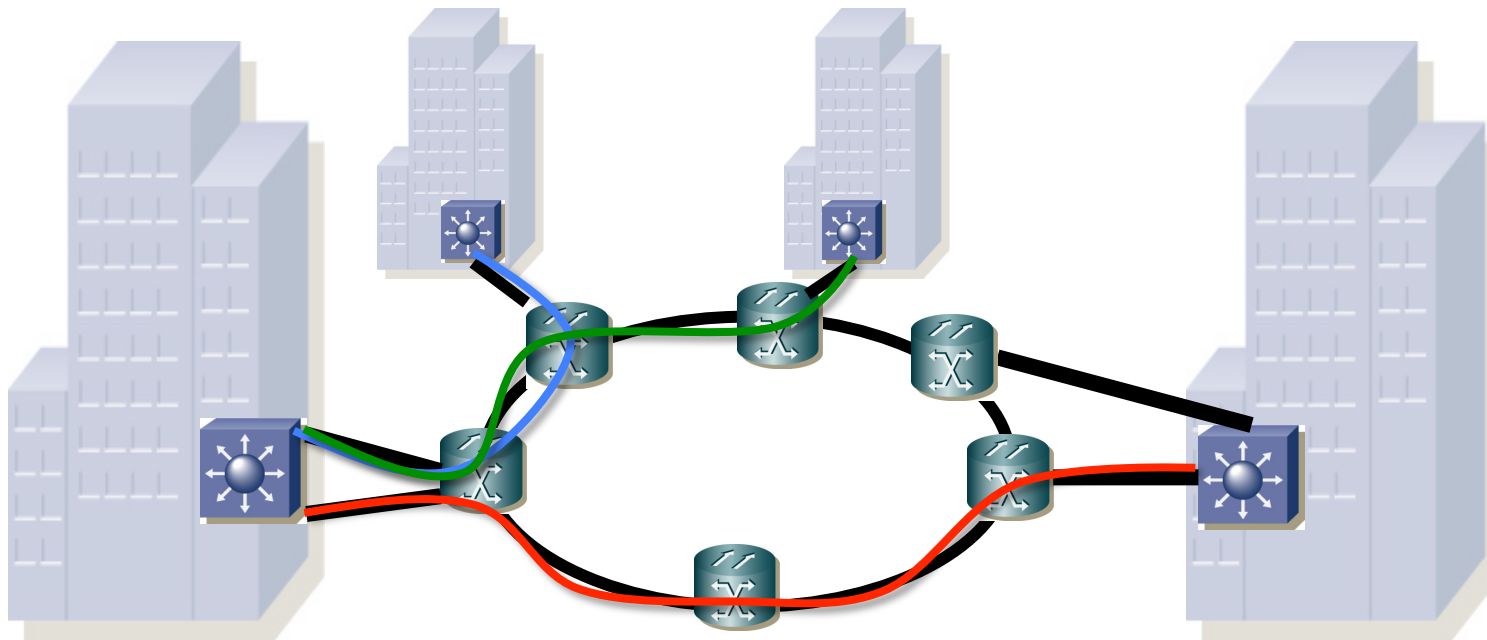
Interconexión por fibra

- 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



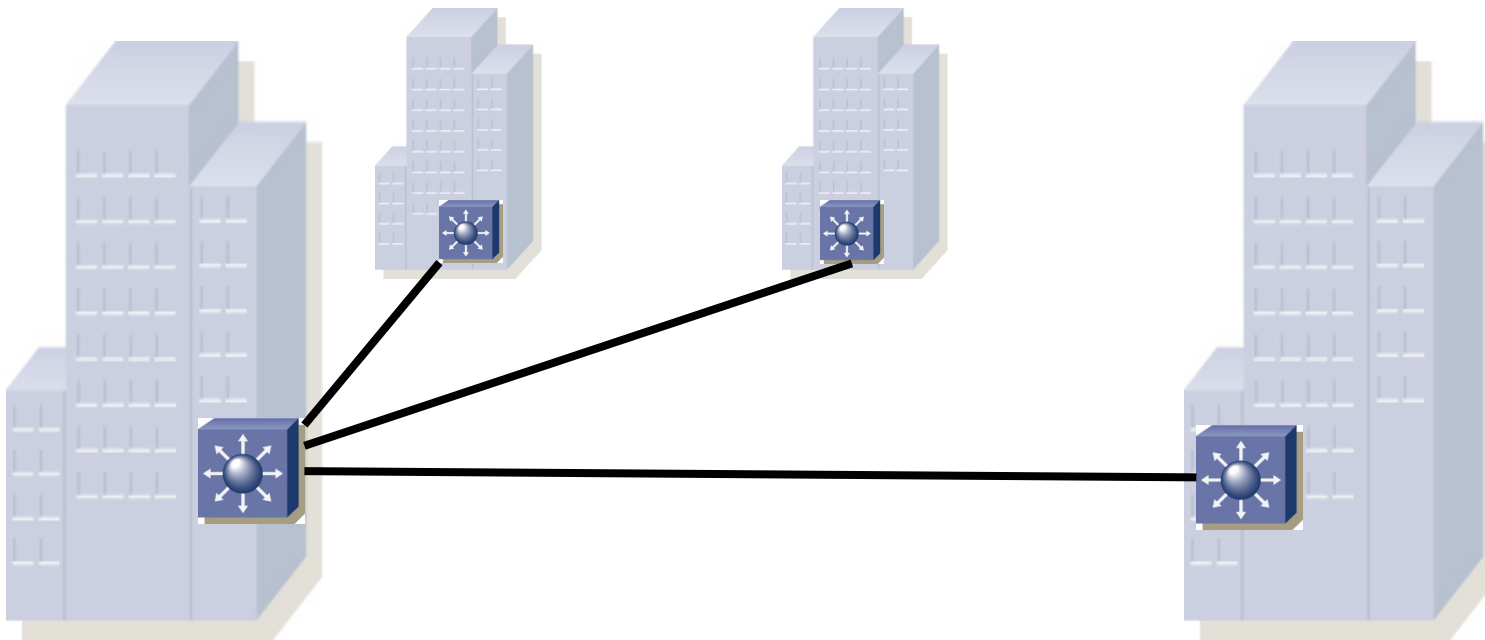
Múltiples *Sites*

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



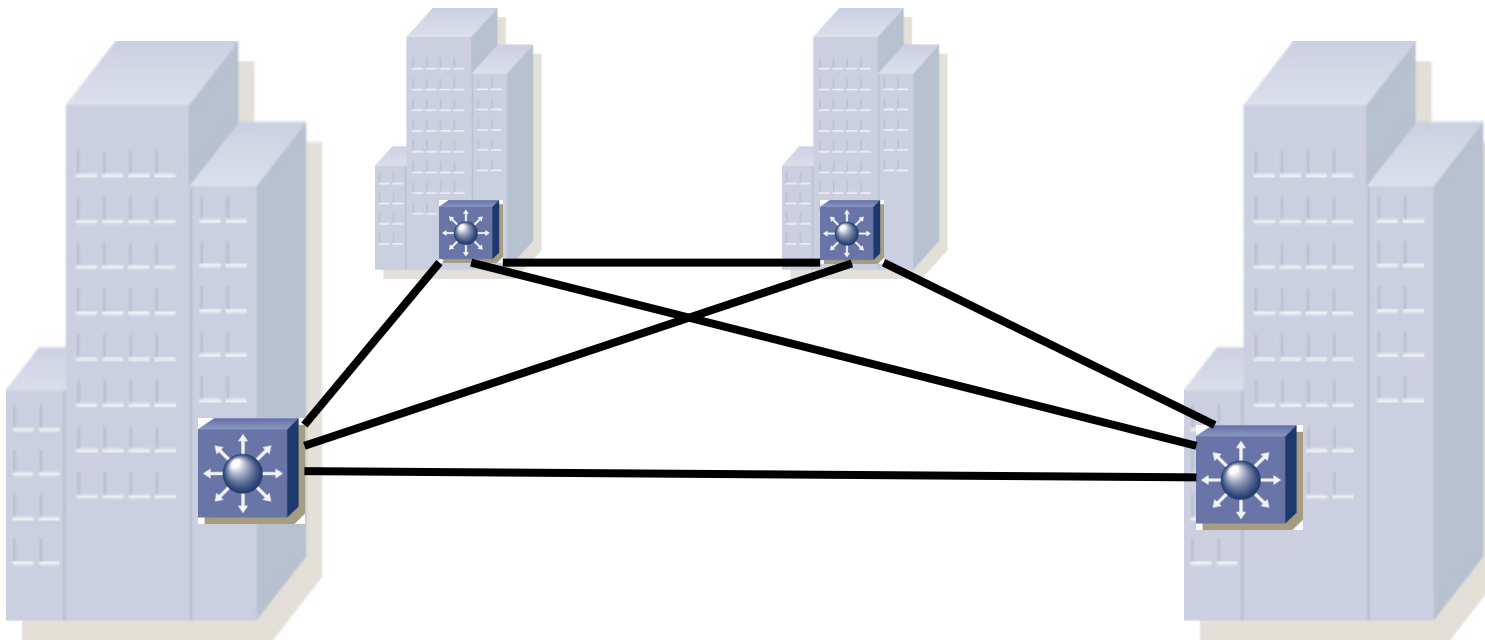
Múltiples *Sites*

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



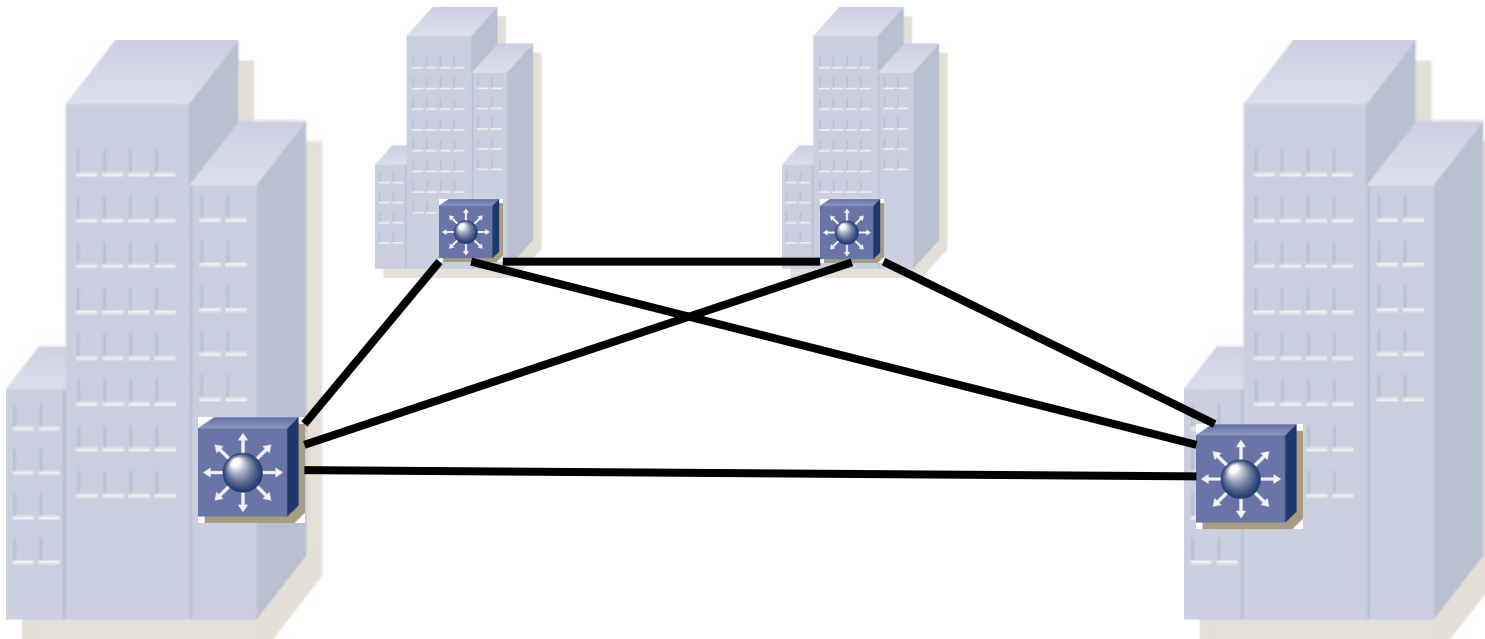
Múltiples *Sites*

- 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*



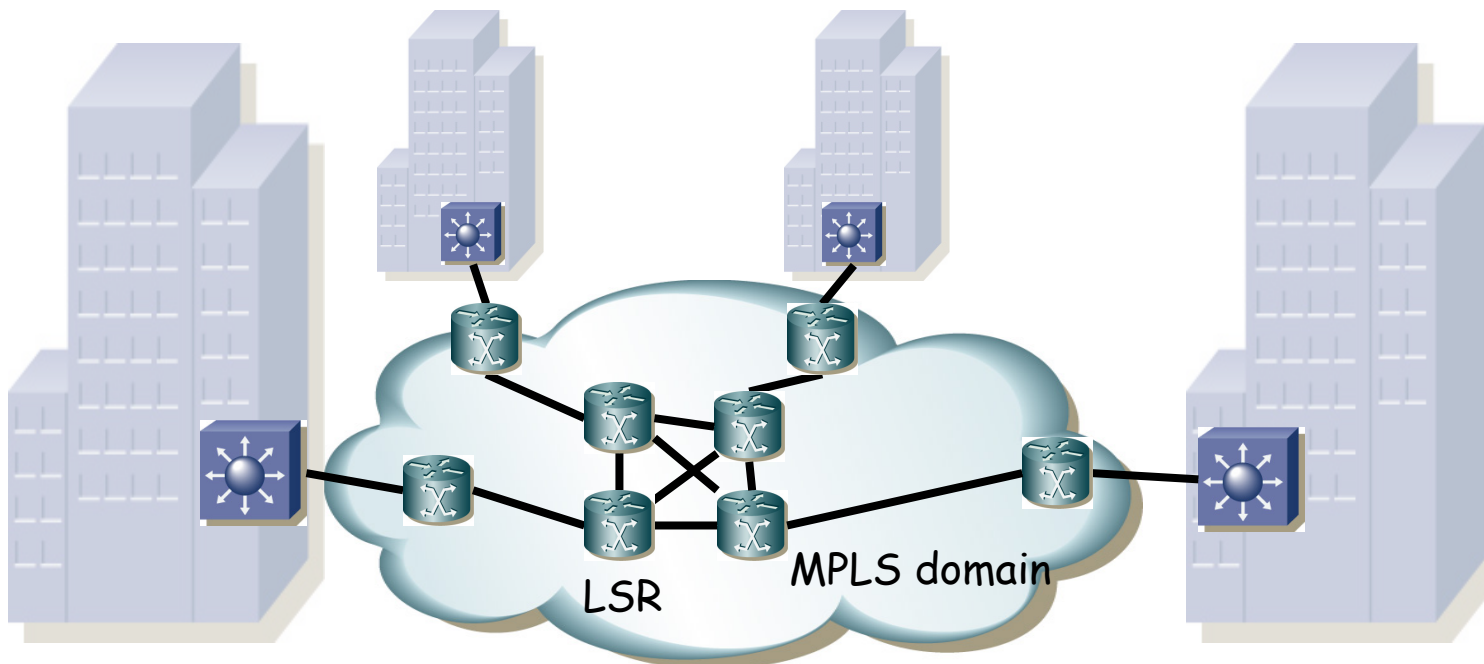
Circuitos

- 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

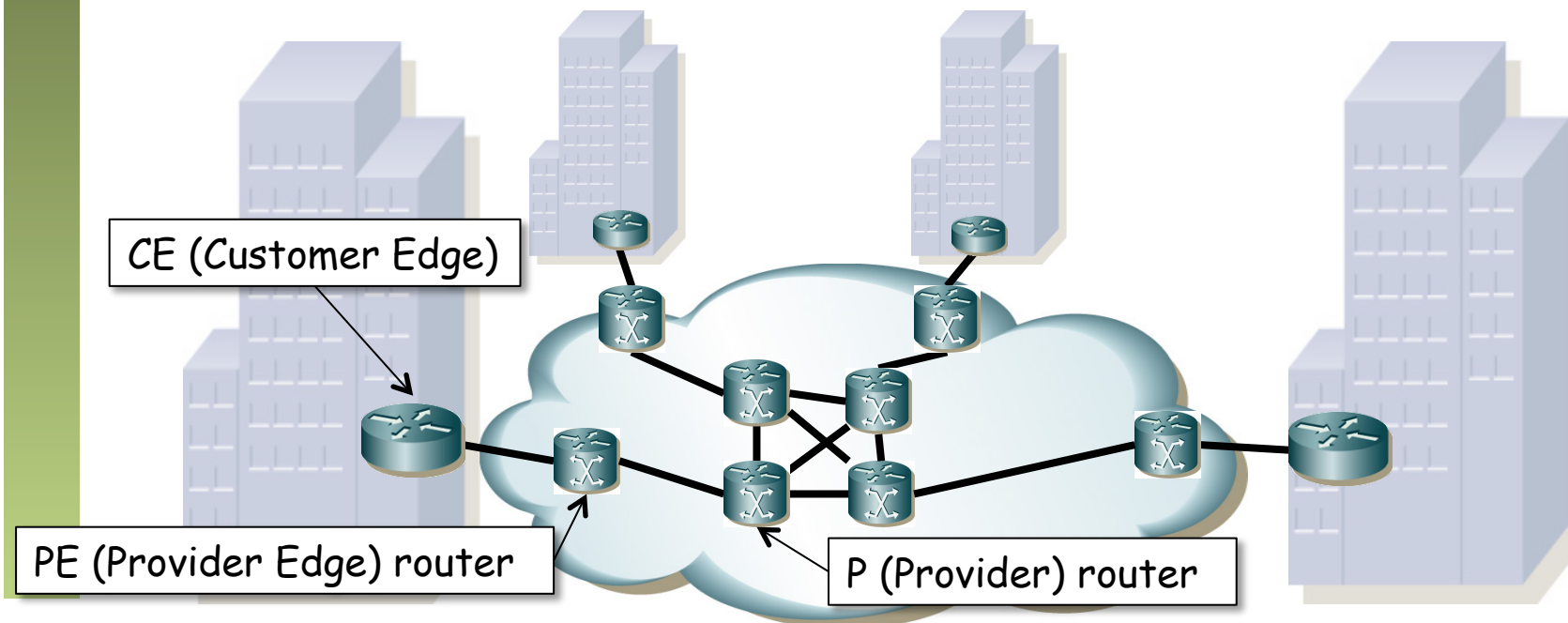
- En lugar de wavelengths o PVCs tenemos LSPs
- Recordemos que podemos encapsular Ethernet sobre MPLS (EoMPLS)
 - RFC 4448 “Encapsulation Methods for Transport of Ethernet over MPLS Networks”
- De hecho se suele decir que tenemos “AToM” o “Any Transport over MPLS”
- Los equipos de usuario van a poder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



Layer 3 MPLS VPNs

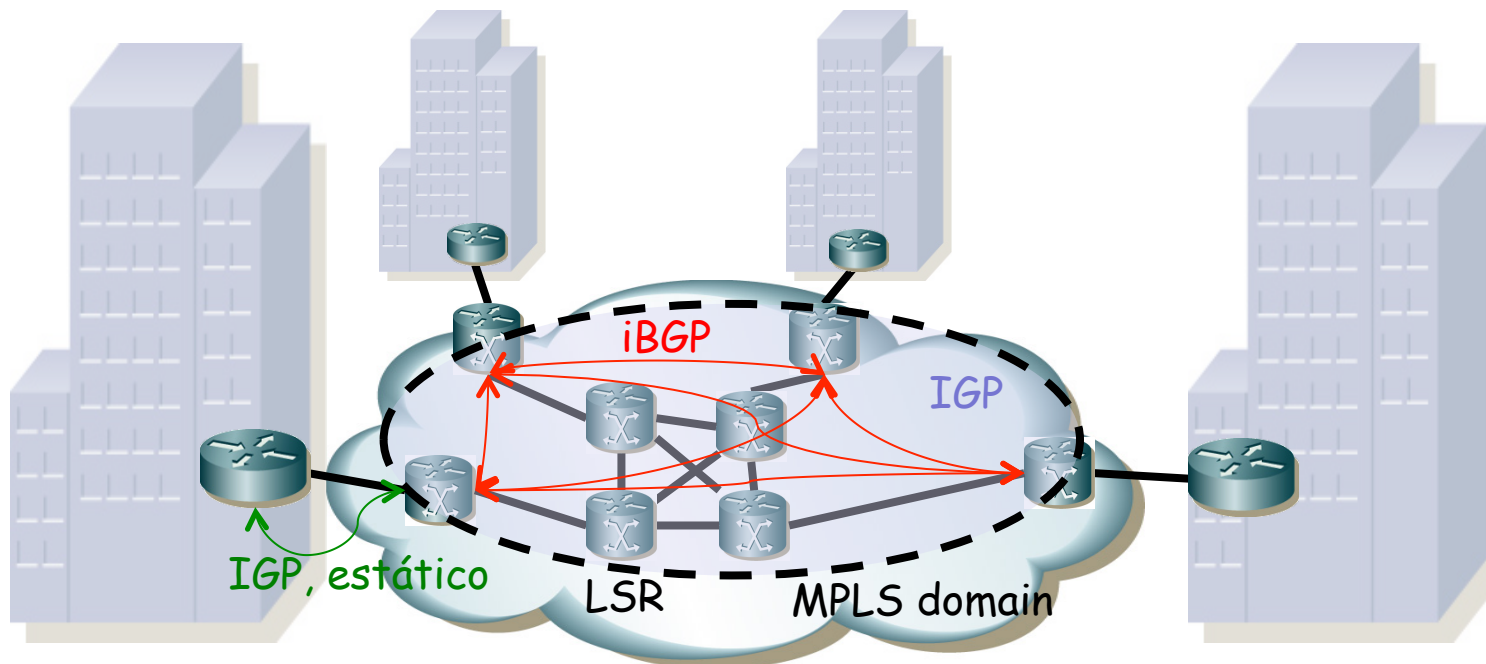
Layer 3 VPNs

- RFC 4364 “BGP/MPLS IP Virtual Private Networks (VPNs)” (Cisco Systems y Juniper Networks, 2006)
- 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”



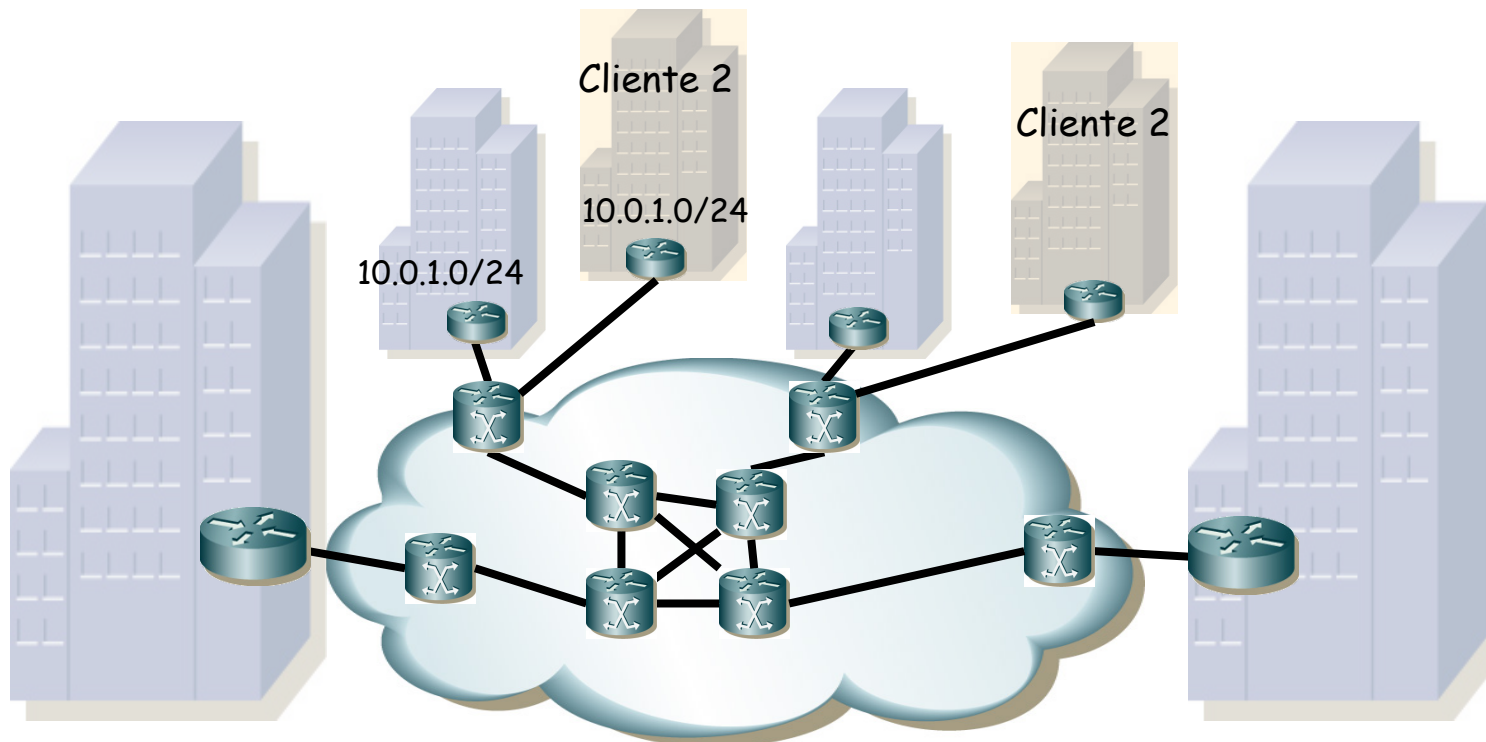
L3VPN: Routing

- Los CE anuncian sus rutas a los PE (con un IGP o rutas estáticas)
- Los PE 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 CE no intercambian información de routing entre ellos, no son adyacentes
- La VPN no actúa como un overlay sino una red IP con otro gestor



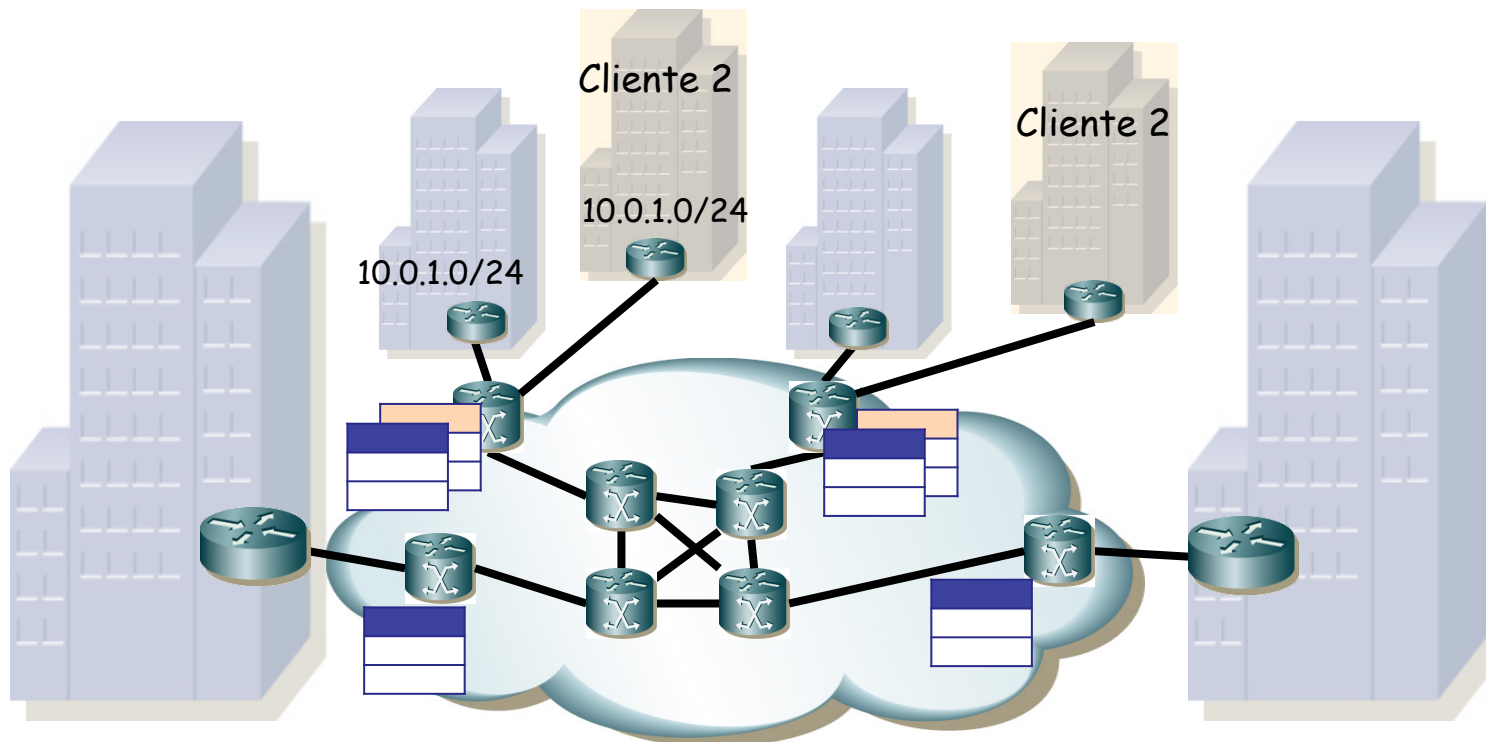
L3VPN: Routing

- Dos VPNs pueden emplear espacios de direcciones IP que se solapan
- Los anuncios VPN-IPv4 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?



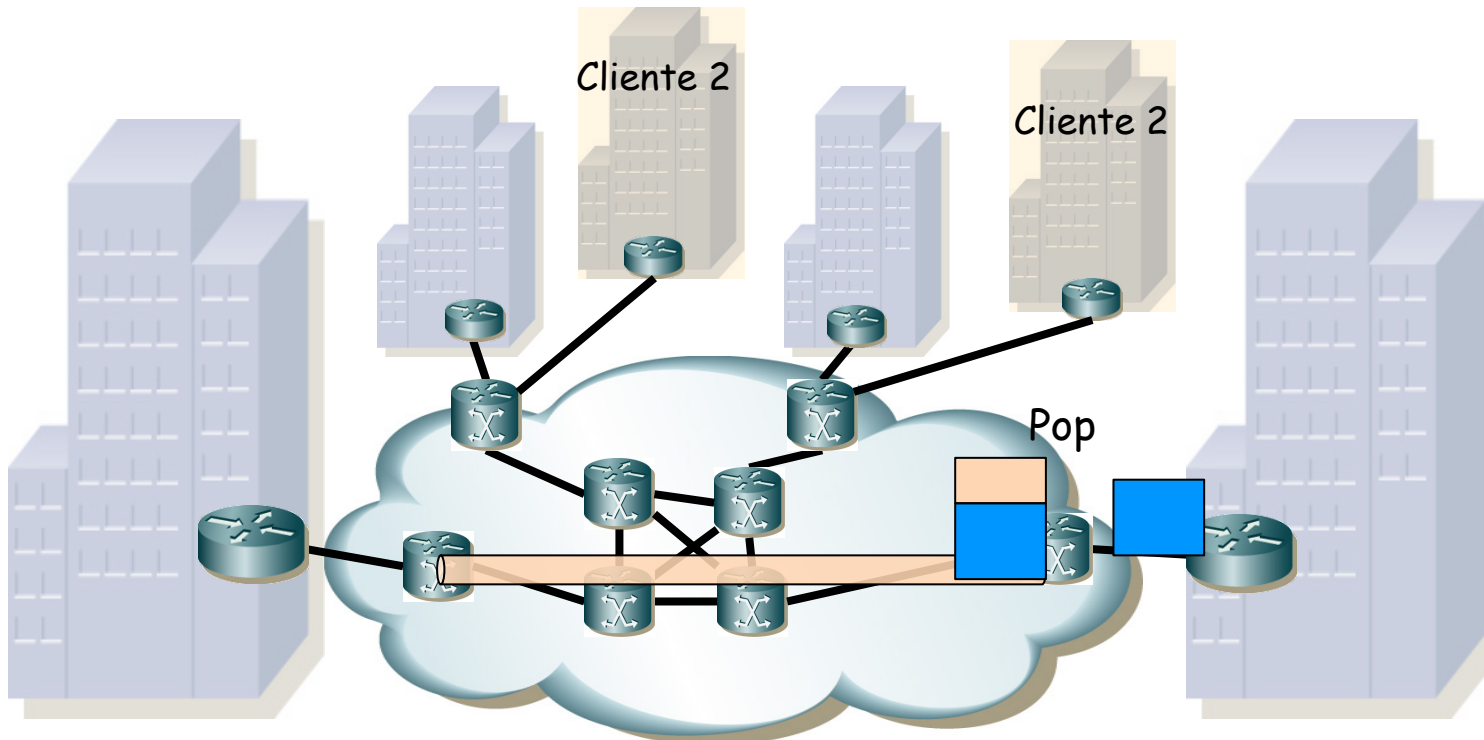
L3VPN: Forwarding

- Cada PE mantiene una tabla de rutas para cada VPN o *VPN/Virtual Routing and Forwarding tables (VRFs)* y además una tabla por defecto
- Cada 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 (o más) valor de RT
- Incluye una etiqueta MPLS (para el plano de datos)
- Una VRF importa las rutas con unos RT que desee (plano de control)



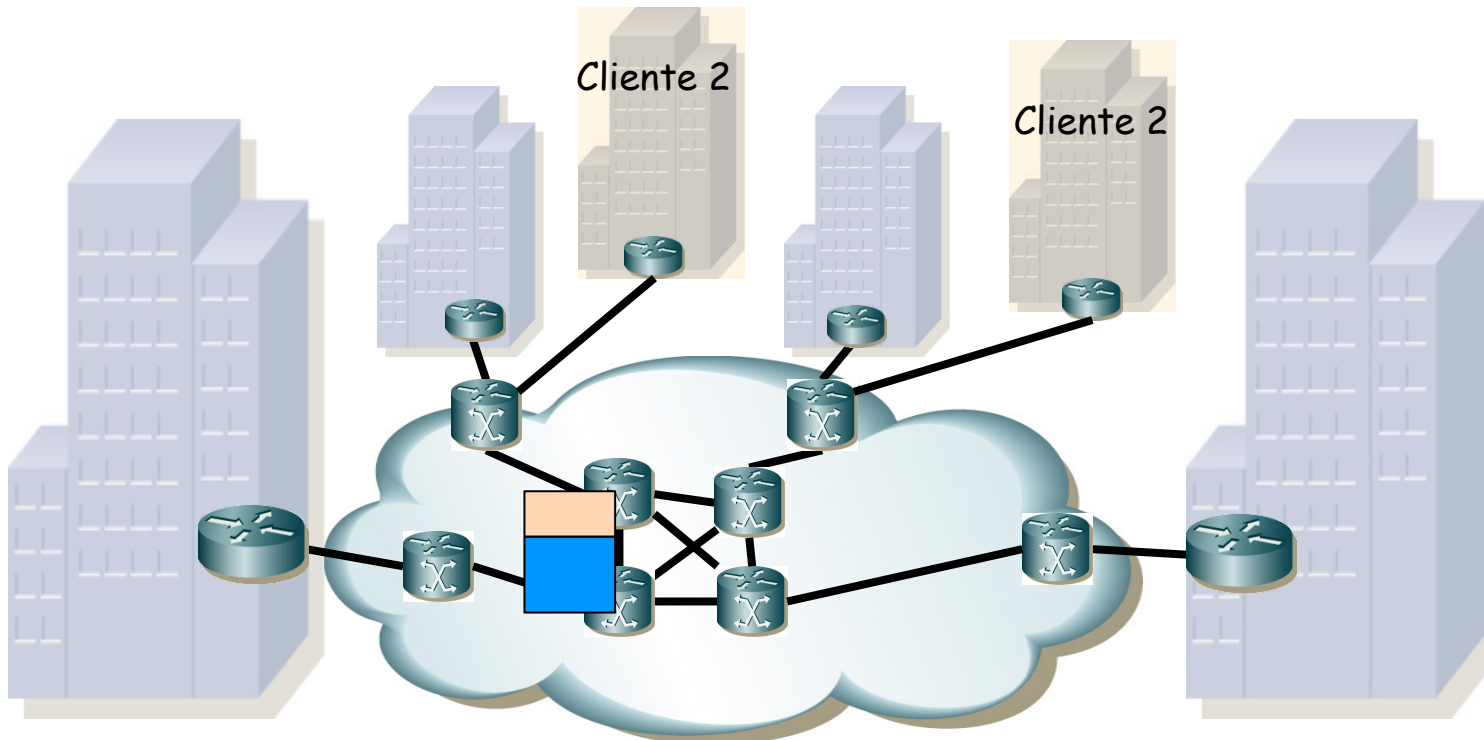
L3VPN: MPLS

- ¿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



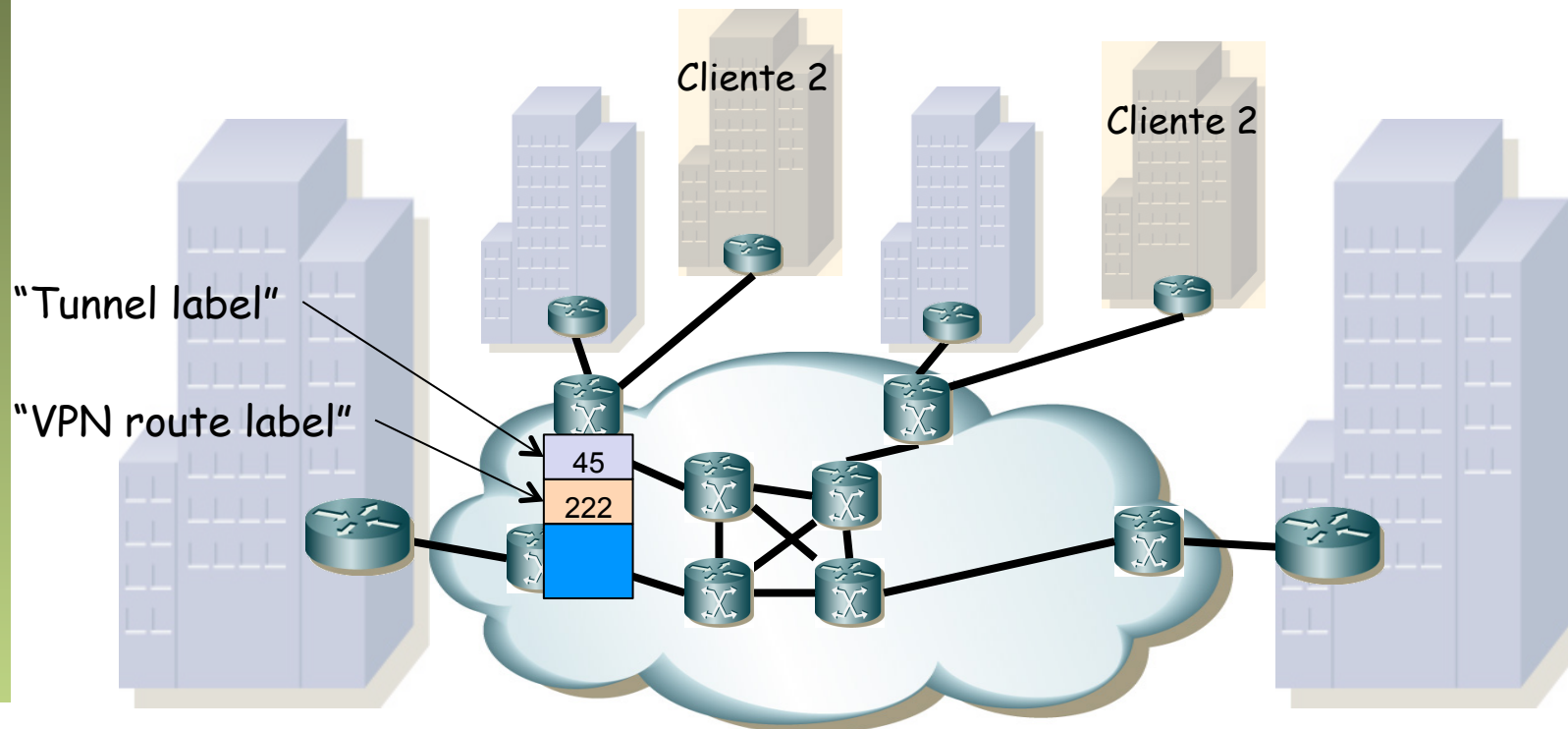
L3VPN: MPLS

- ¿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
- (...)



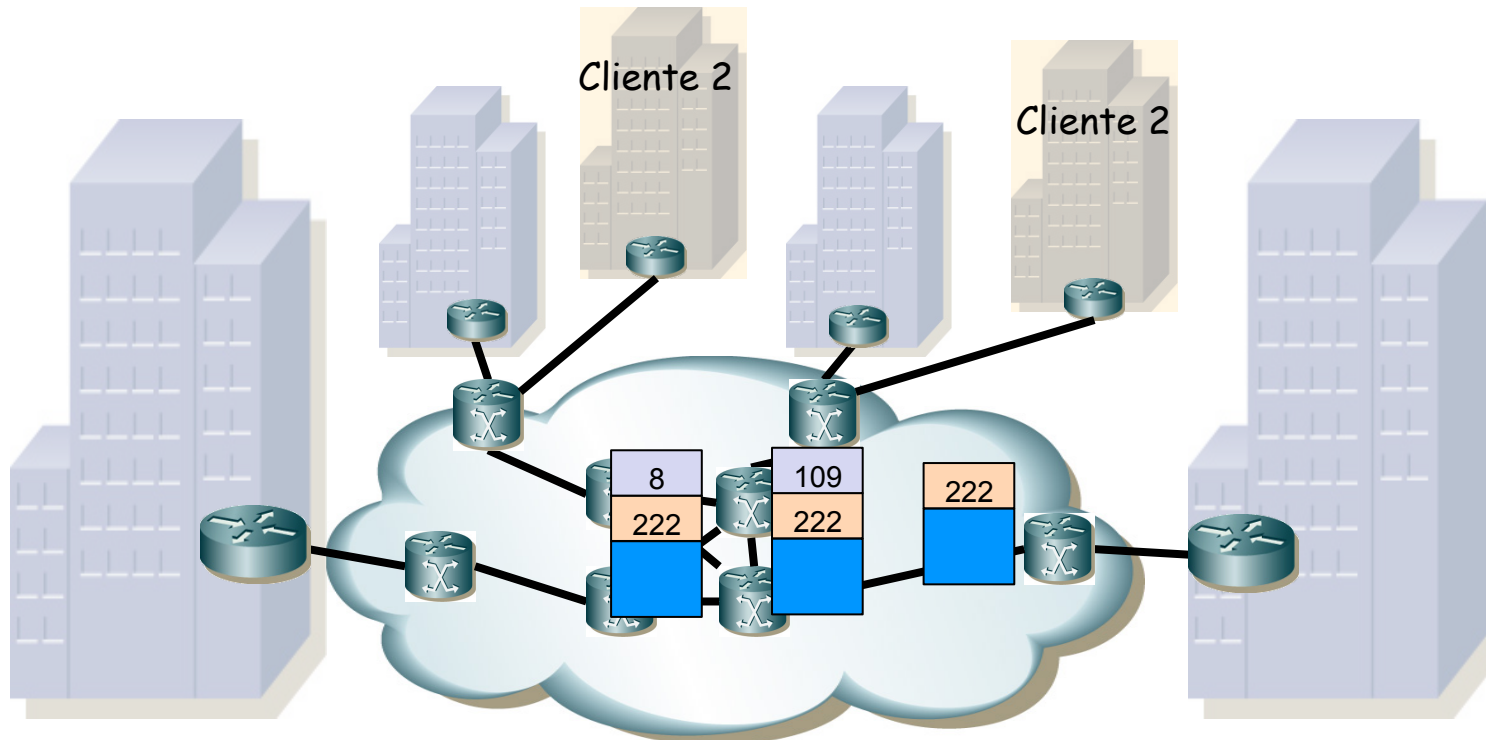
L3VPN: MPLS

- ¿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 (...)



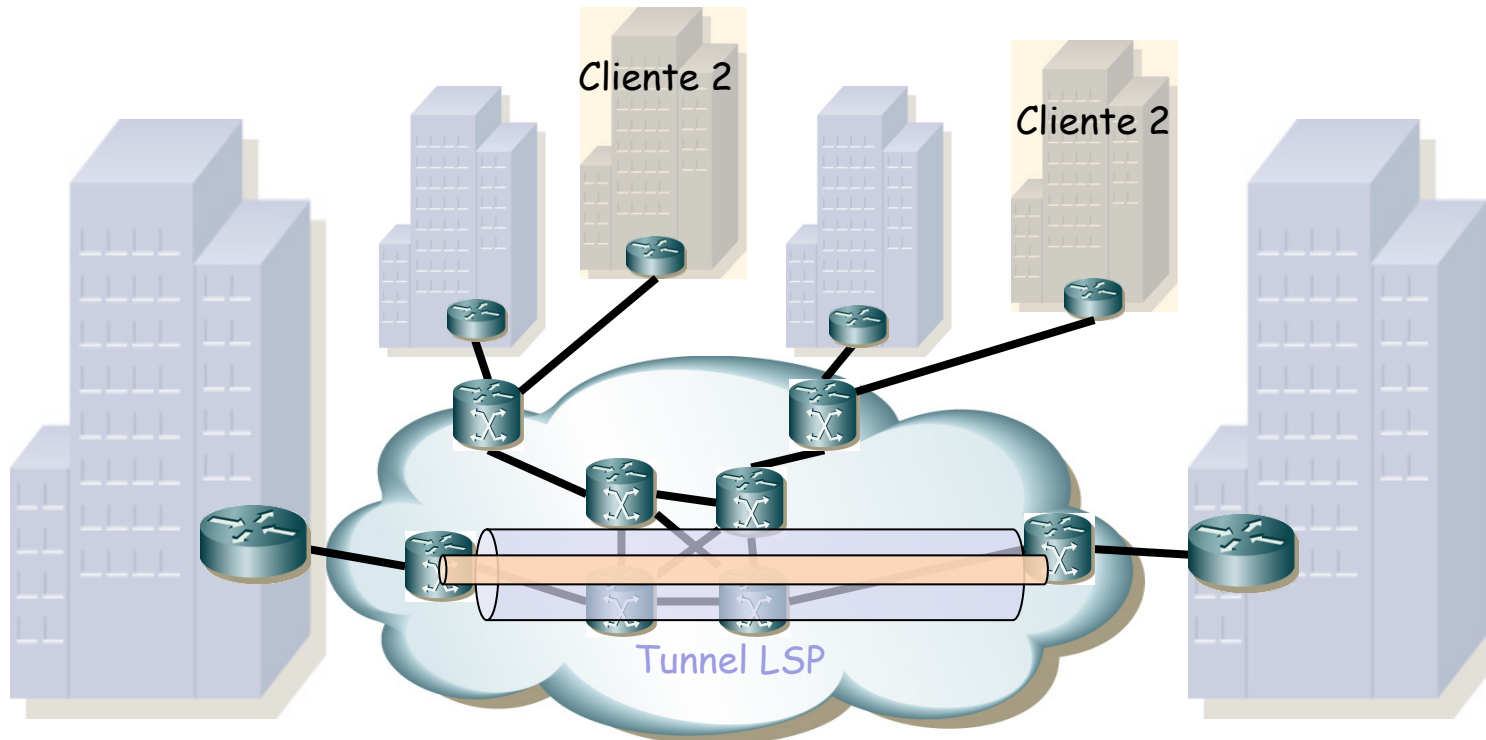
L3VPN: MPLS

- ¿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
- (...)



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

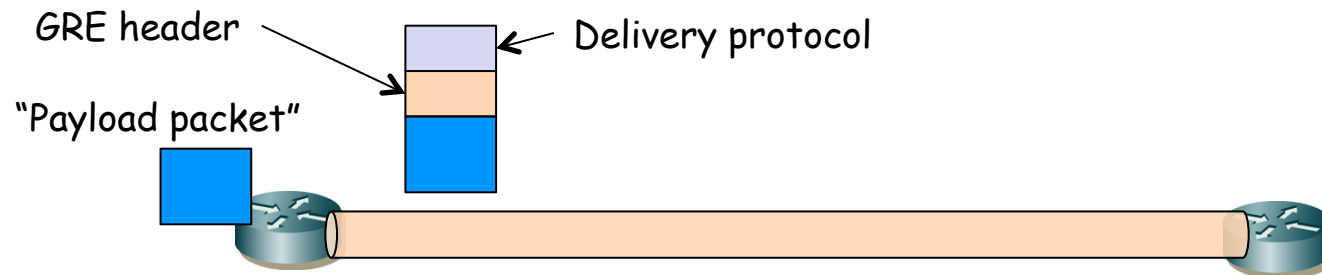


GRE

GRE

- RFC 2784 “Generic Routing Encapsulation (GRE)” (Procket Networks, Enron Communications, Cisco Systems, Juniper Networks, 2000)
- 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” (Cisco, 2000) (...)

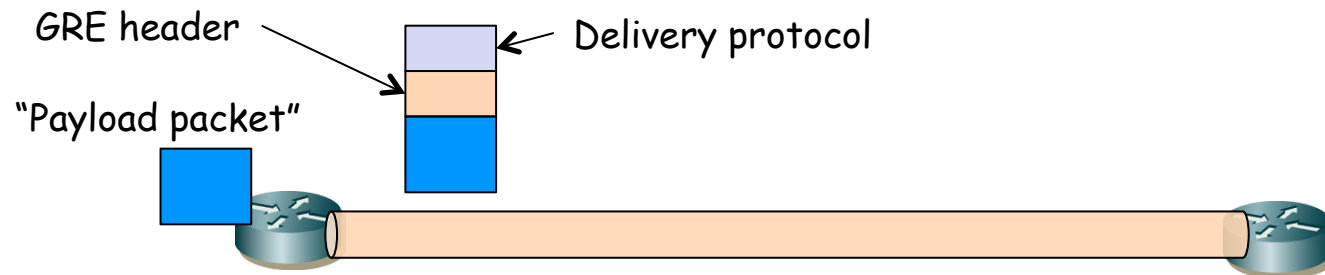
0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1
+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	C																														



GRE

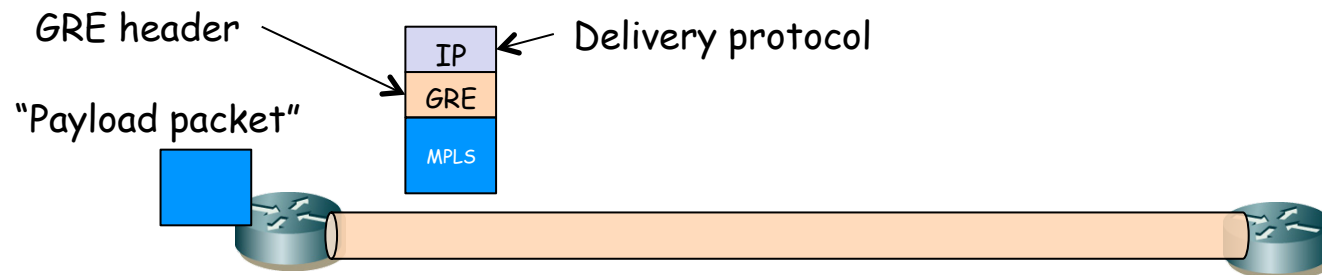
- RFC 2890 “Key and Sequence Number Extensions to GRE”
- “Key” sirve para distinguir flujos dentro del túnel
- “Sequence Number”
 - Si hay “key” entonces el número de secuencia es por “key”
 - Permite dar entrega en orden (aunque no fiable)
 - Si llega uno “anterior” lo descarta
 - Si llega uno que deja 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																		
+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+



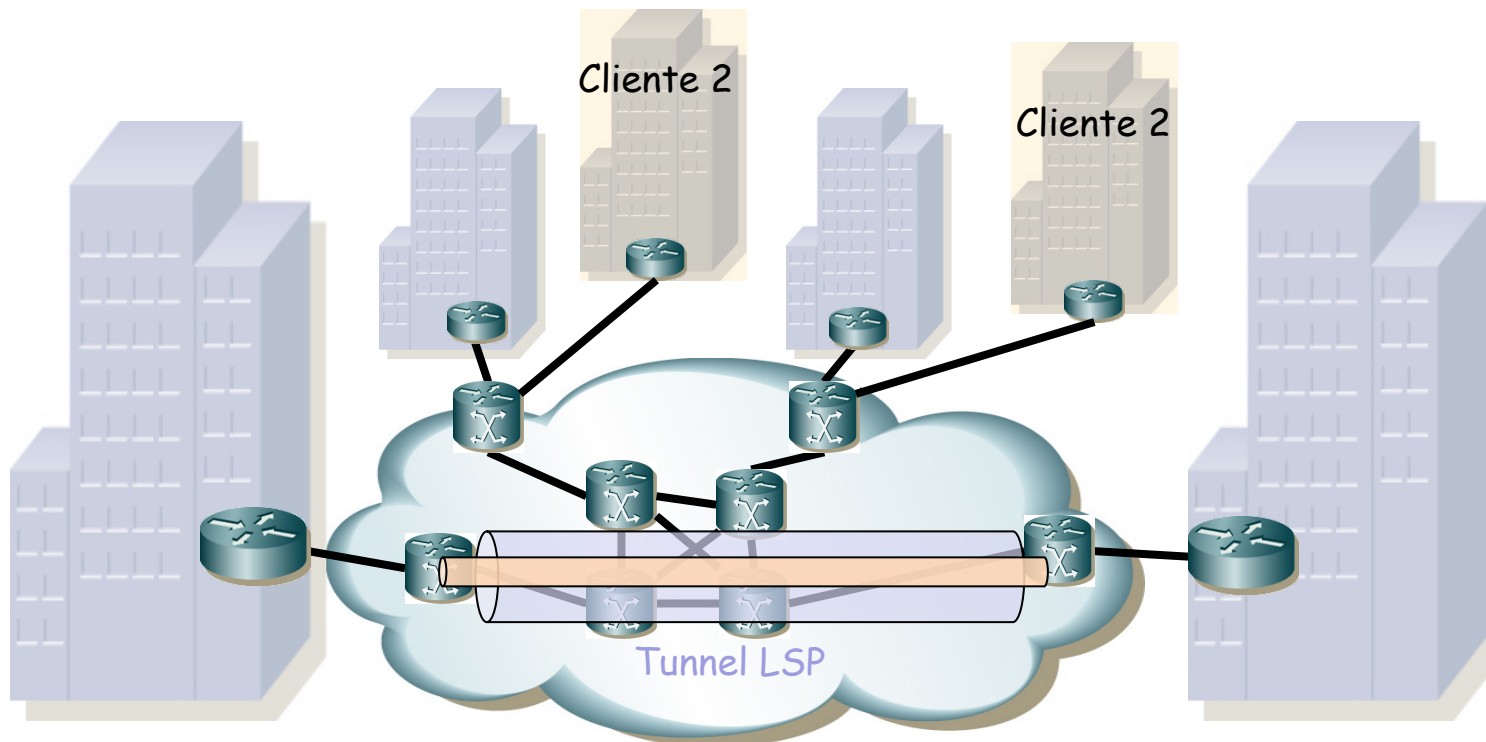
MPLS in GRE in IP

- 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)



L3VPN con GRE

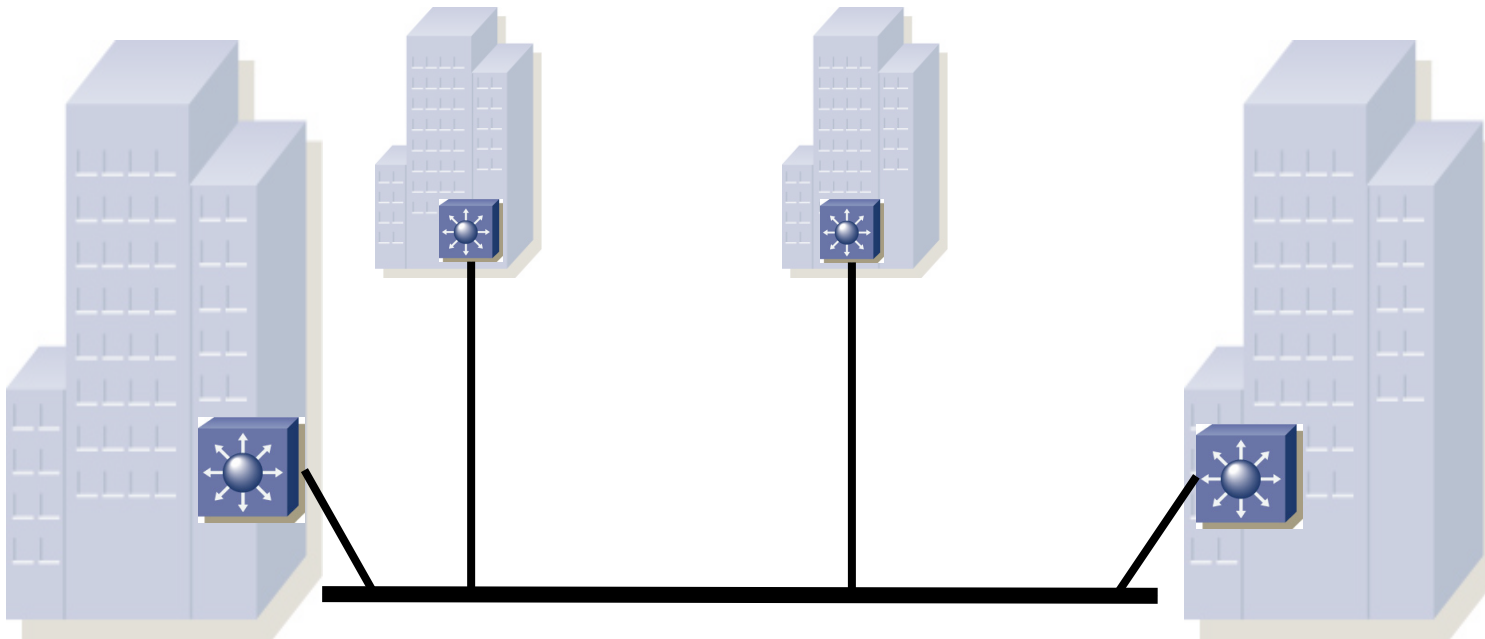
- Como decíamos, los túneles entre los PE pueden ser túneles GRE o simple IPoIP
- Entonces la red de transporte ya no necesita ser MPLS, es simple IP



VPLS

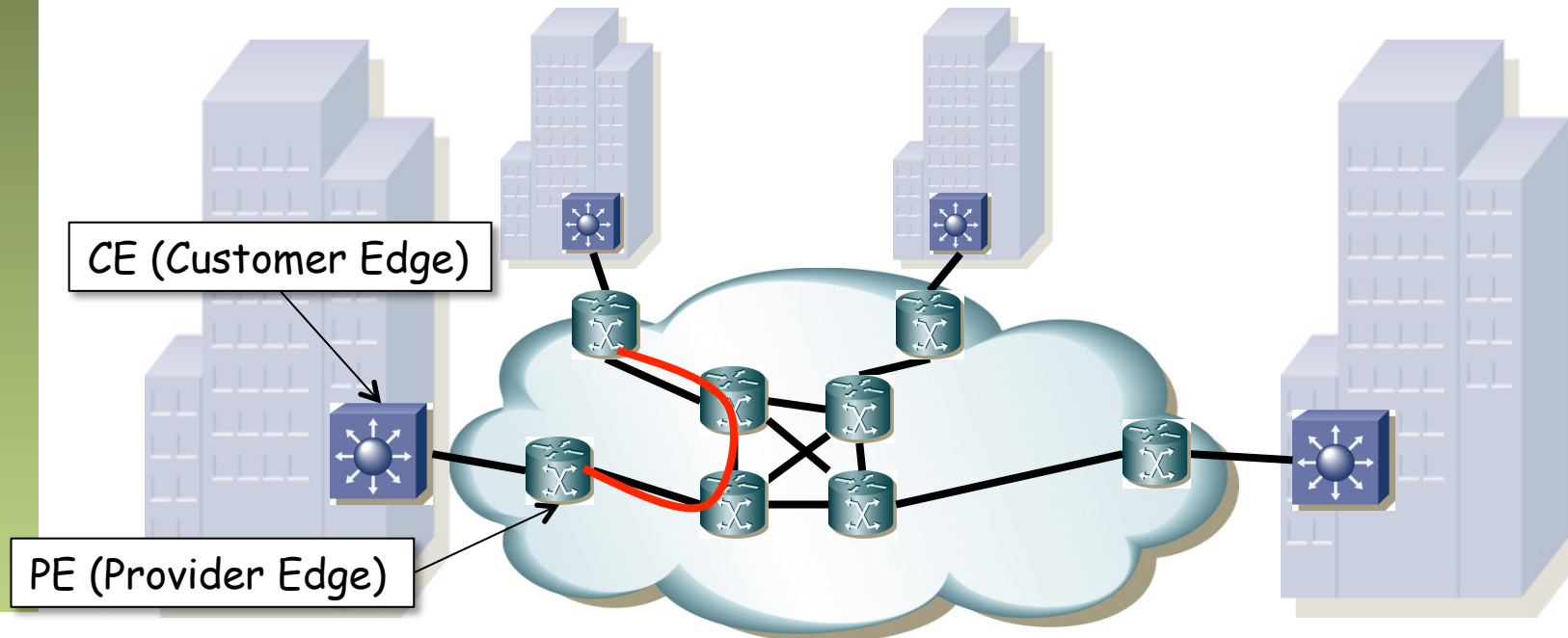
MPLS y VPLS

- “*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)



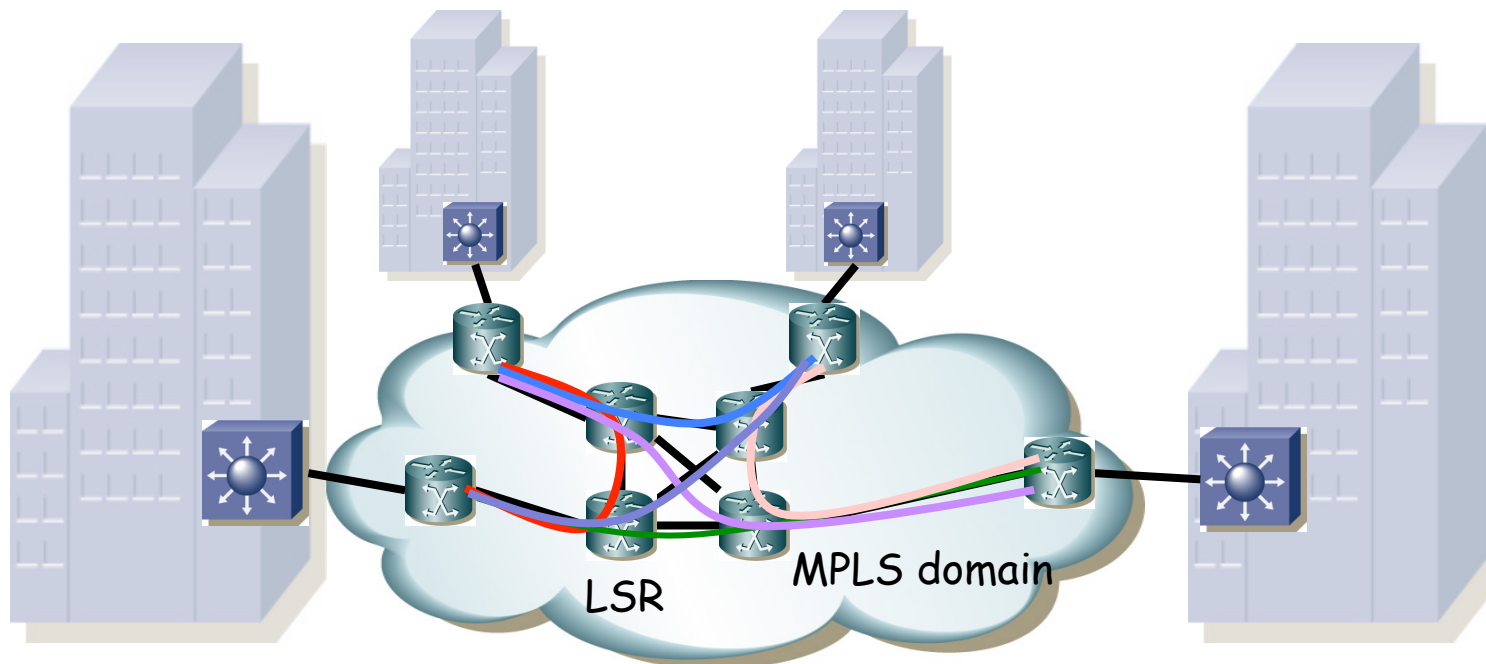
MPLS y VPLS

- 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



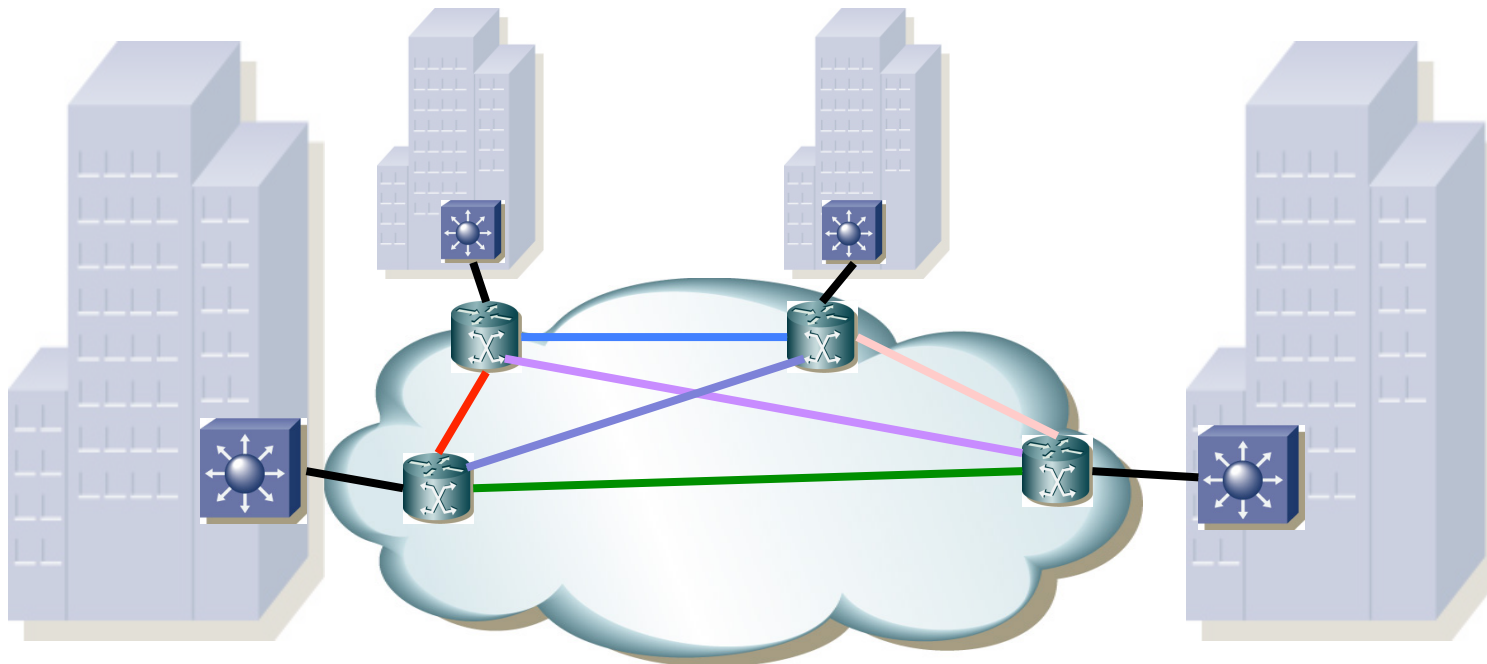
MPLS y VPLS

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



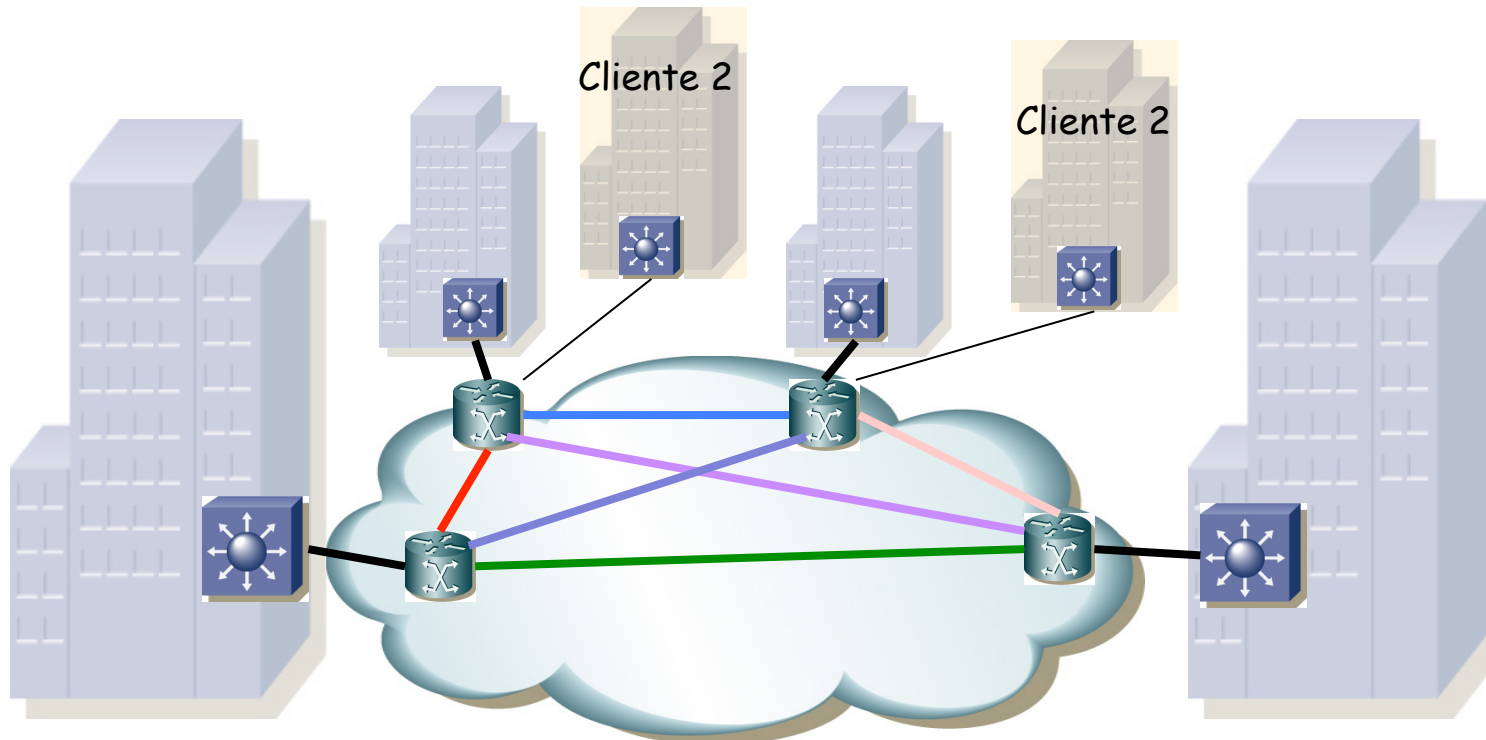
MPLS y VPLS

- 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 (...)



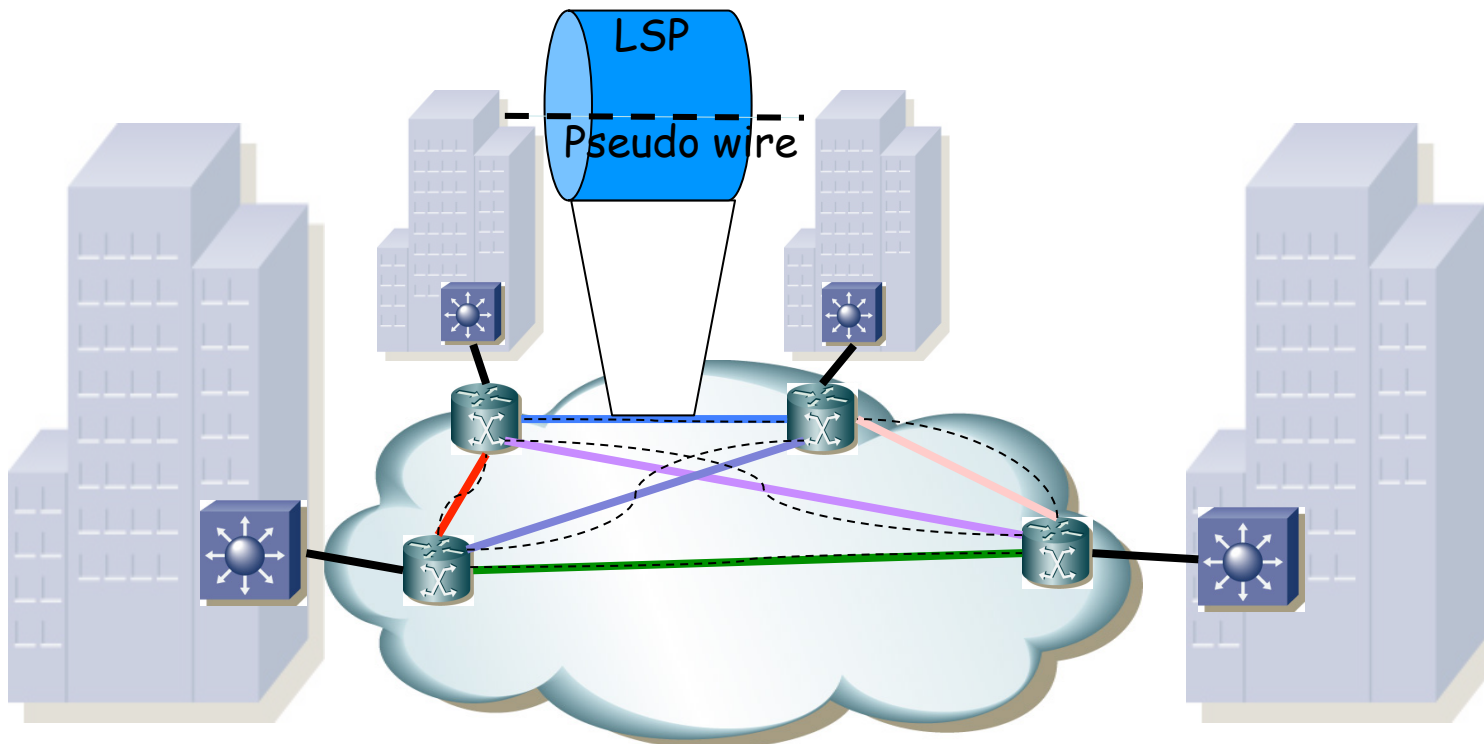
MPLS y VPLS

- 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?



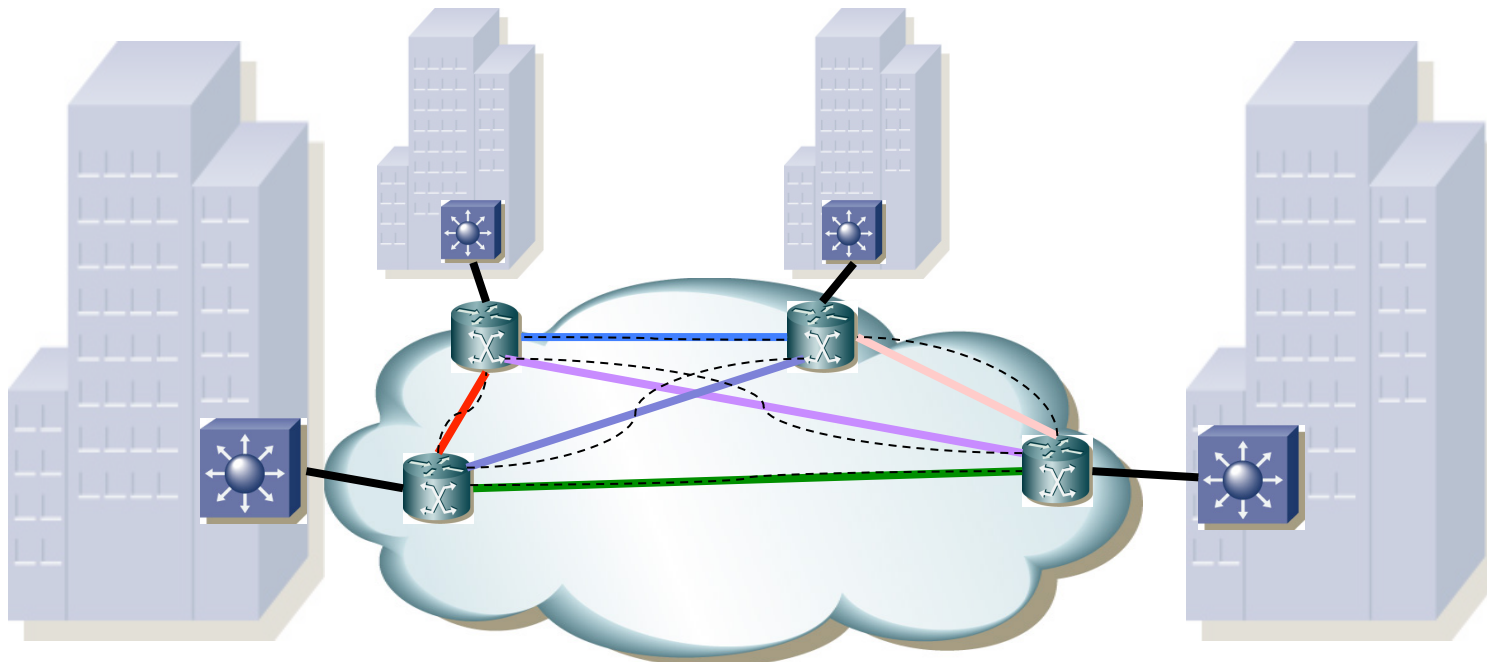
VPLS y PWE

- Por cada instancia VPLS (cada cliente) se establece un full mesh de *pseudo-wires* (PWs) entre los PEs
- 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



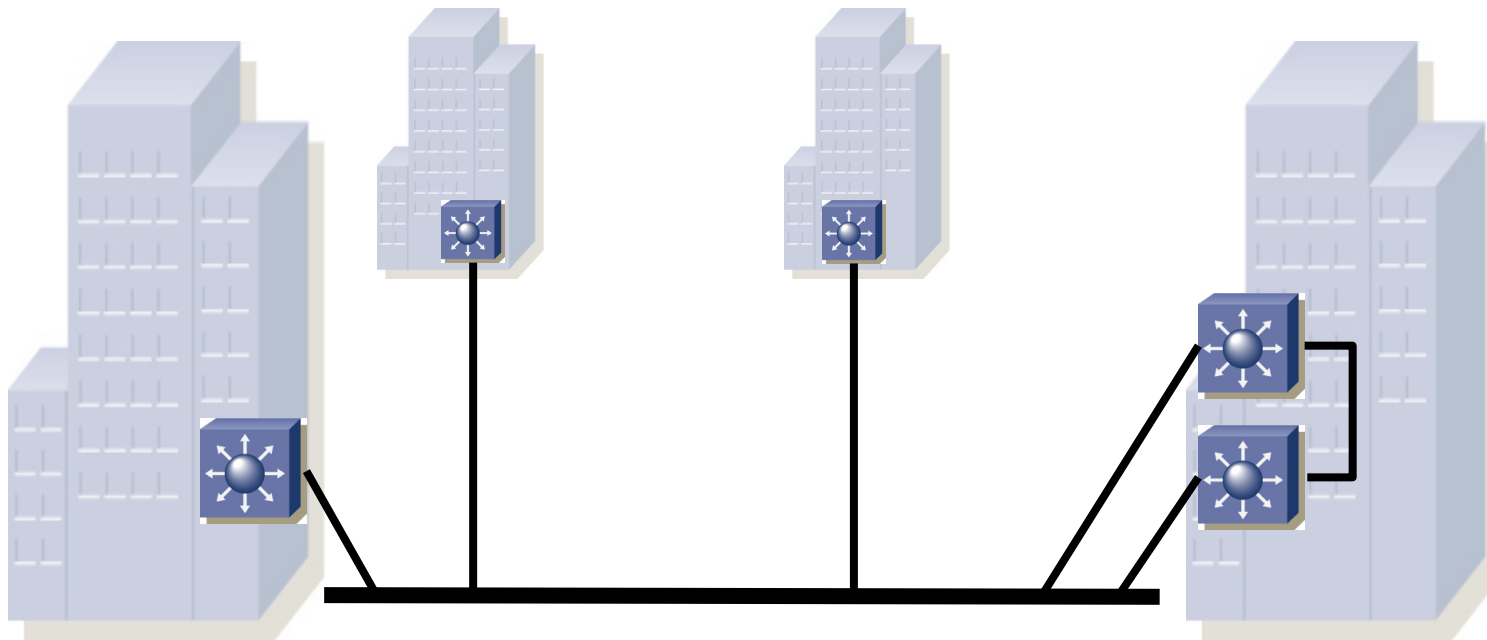
VPLS y PWE

- 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 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

- 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”
- Repito: el aprendizaje de direcciones MAC se hace en el plano de datos, es decir, con la dirección MAC origen de la trama recibida

Problemas en VPLS

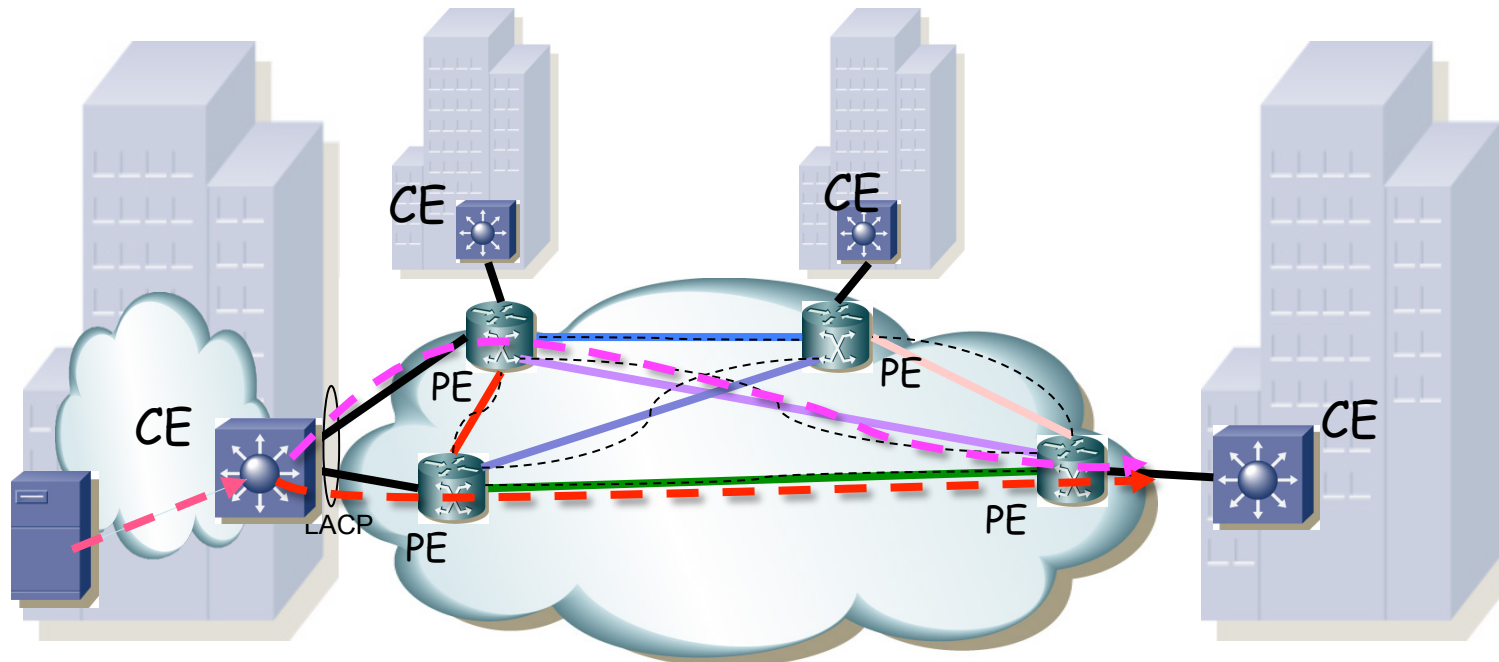
- Se deben establecer $N \times (N-1)/2$ pseudo-wires
- Problema de escalabilidad (cantidad de tráfico de control)
- Replicación de paquetes que sufren inundación:
 - Se lleva a cabo en el PE de entrada
 - Se dirigen punto-a-punto a cada otro PE del servicio
 - Mayor trabajo en el PE
 - Más uso de capacidad
 - 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



EVPN

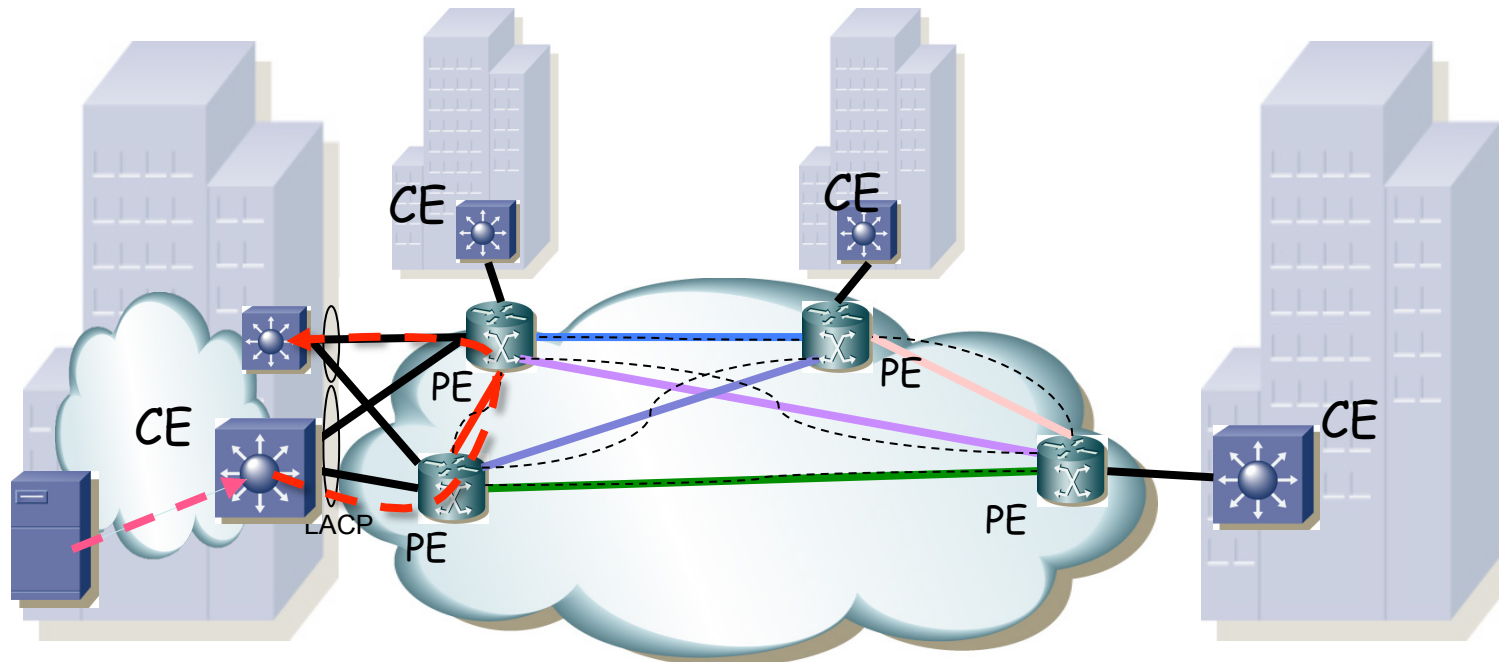
Limitaciones de VPLS

- 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 (...)
- Active-Active solo mediante soluciones propietarias (vPC, VSS, etc)



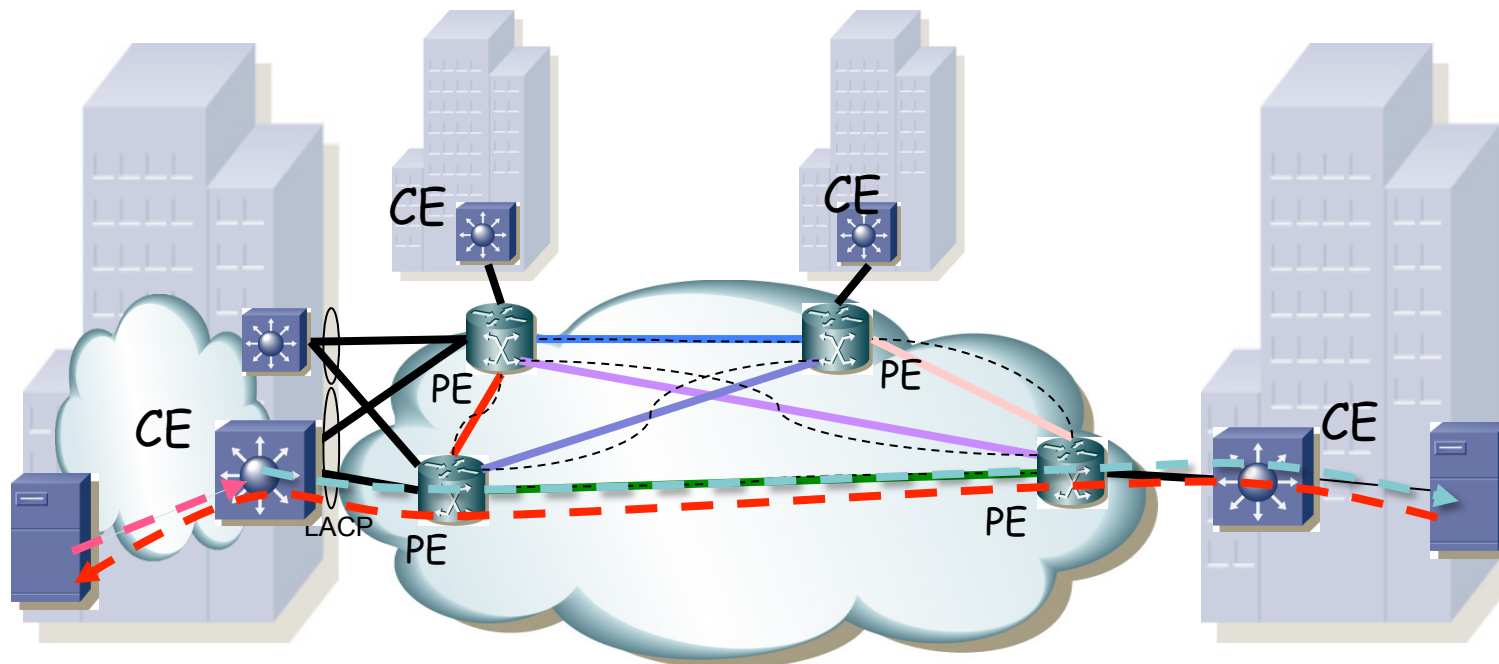
Limitaciones de VPLS

- 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 (...)



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)
- Ofrece una VPN capa 2, como VPLS
- Los PEs puede estar conectados mediante LSPs o túneles (IP/GRE)
- 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
- Es decir, 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

