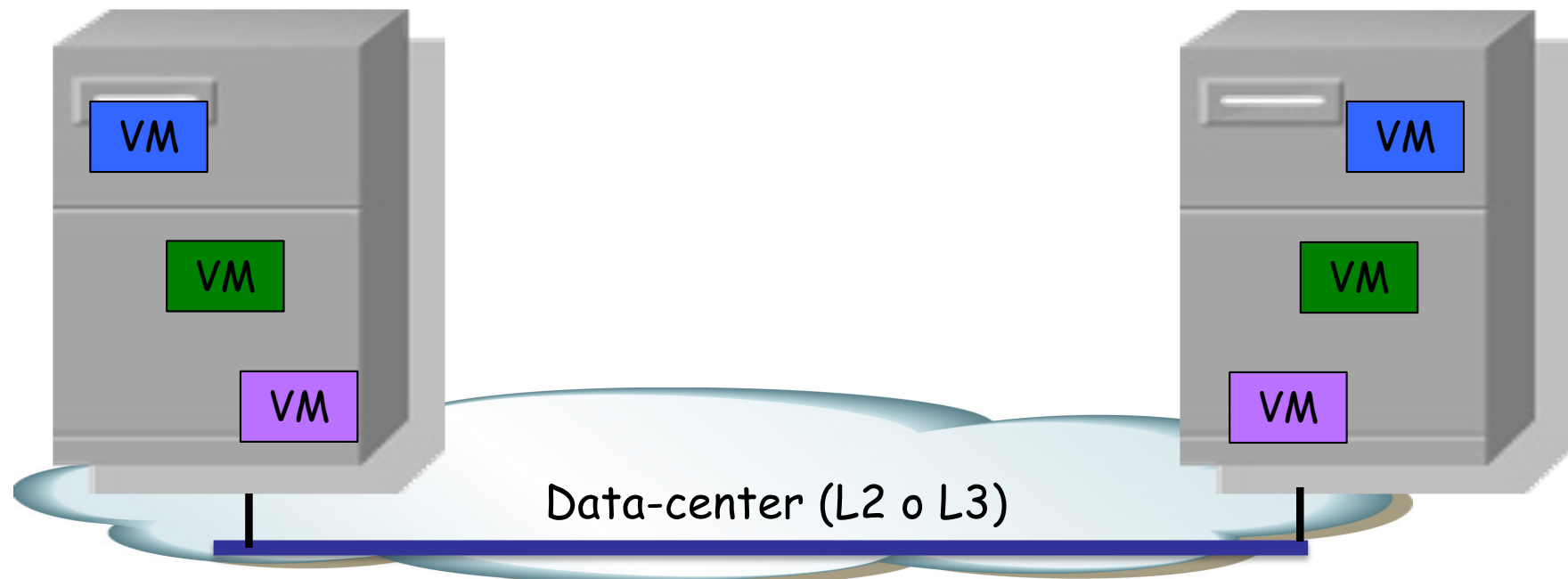


Overlays en el data center

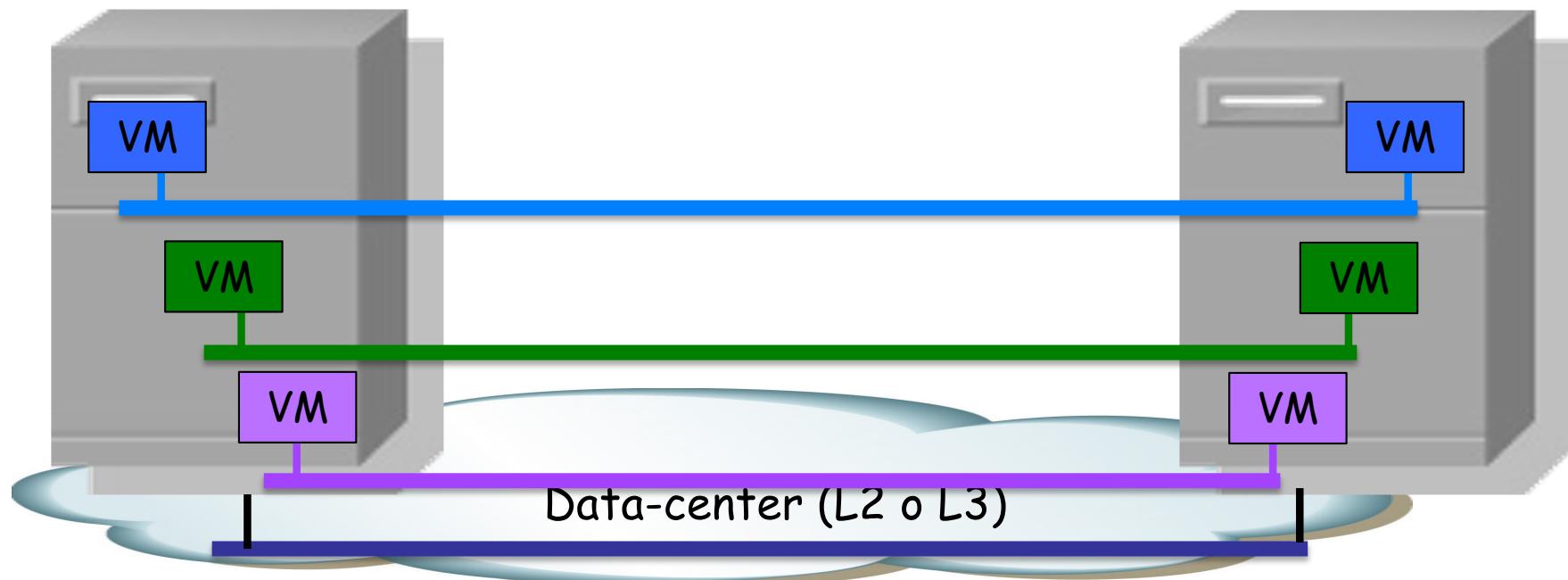
Multi-tenancy

- VMs (o contenedores) de diferentes clientes del DC o de diferentes departamentos de la empresa
- Las del mismo cliente deben poder estar en su propia red, aislada del resto
- (...)



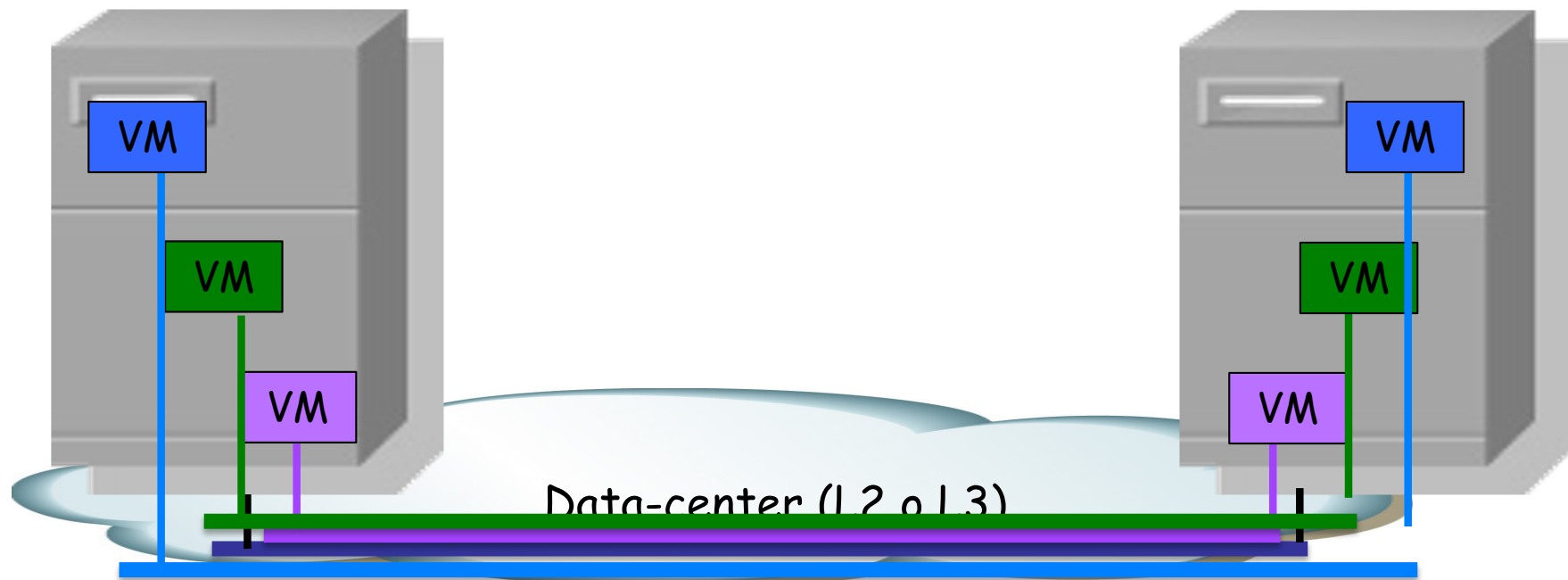
Multi-tenancy

- VMs (o contenedores) de diferentes clientes del DC o de diferentes departamentos de la empresa
- Las del mismo cliente deben poder estar en su propia red, aislada del resto
- Deben poder utilizar el direccionamiento que quieran sin colisión con otros clientes o con la *underlay network*
- Deben poder migrarse, sin hacer cambios en las VMs, por todo el DC



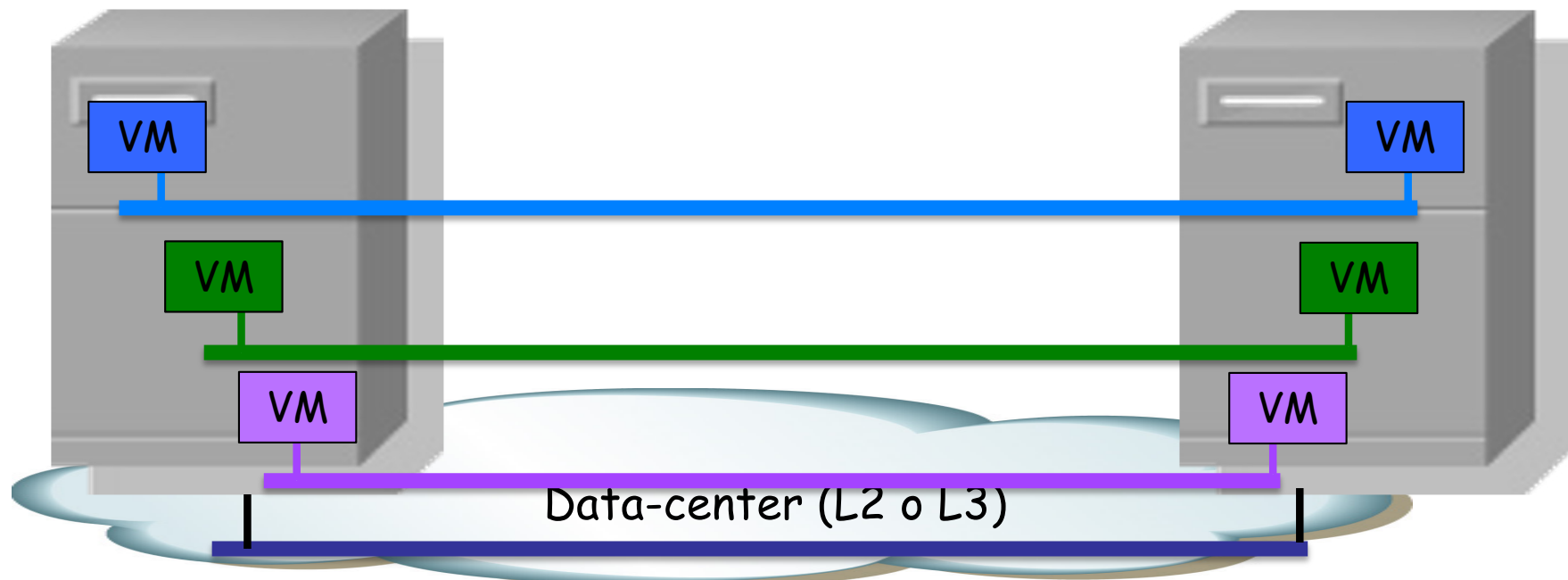
VLANs

- ¿Podríamos emplear VLANs?
- Mapear cada red de tenant a una VLAN del DC
- Limitado a 4094 tenants (menos las VLANs que requiera el DC)
- Se ven las MACs de todas las VMs en los conmutadores del DC
- Todo el DC capa 2
- Implica extender los dominios de broadcast por todo el DC



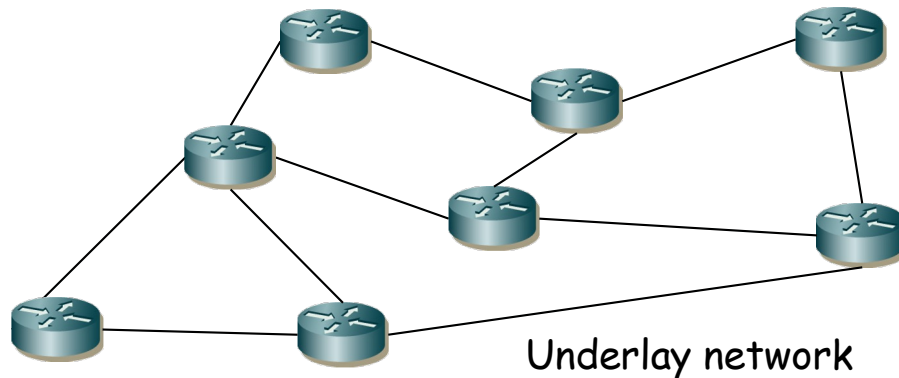
Multi-tenancy

- IETF WG nvo3 (Network Virtualization over Layer 3)
- RFC 7364: “Overlays for Network Virtualization”, (IBM, EMC, Cisco, AT&T, 2014)
- Overlay Network: una red virtual con **separación entre *tenants*** (inquilinos) **sin conocimiento por parte de la *underlay network*** de dichos tenants para el forwarding
- Desacople entre underlay y overlays



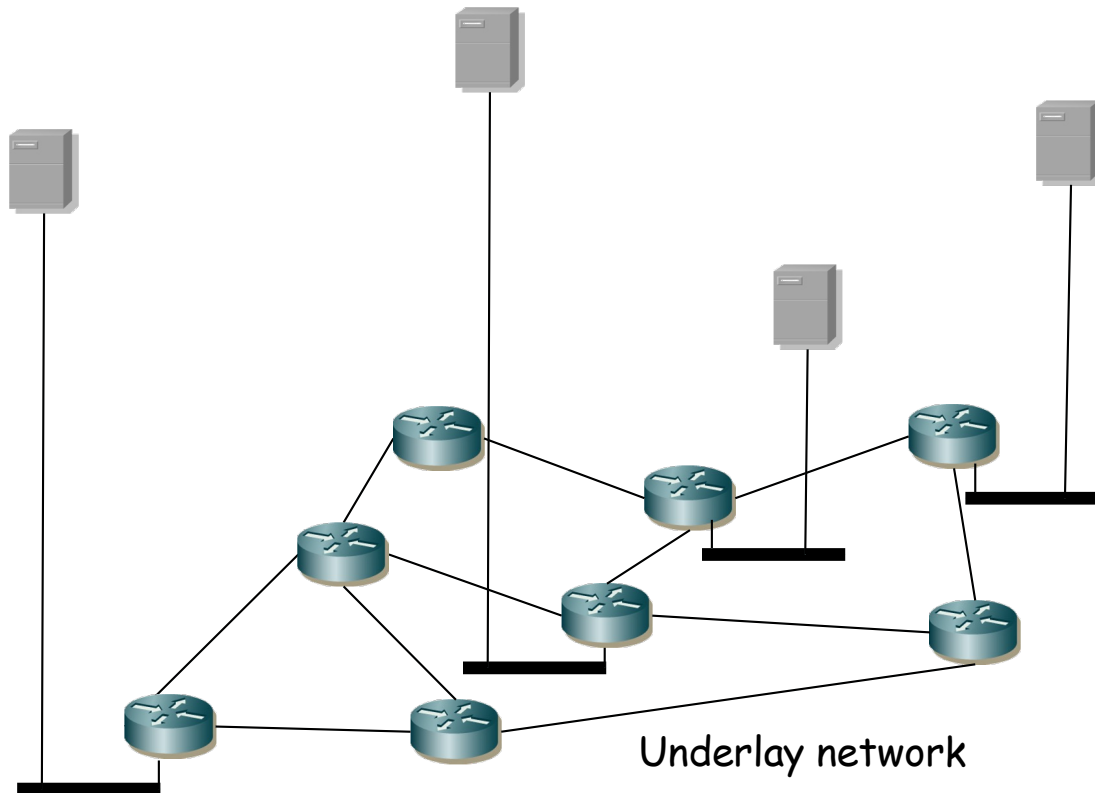
Overlay network

- El DC dispone una red (underlay network)
- Combinación de Ethernet, IP, MPLS, etc



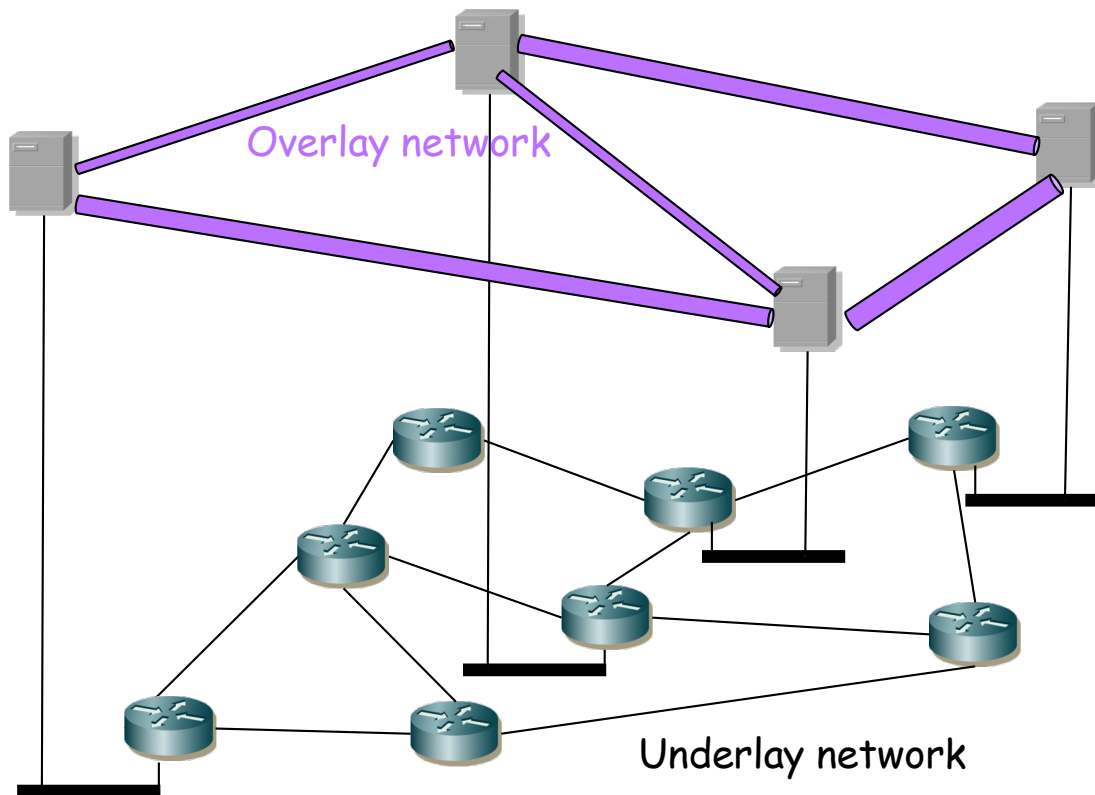
Overlay network

- El DC dispone una red (underlay network)
- Combinación de Ethernet, IP, MPLS, etc
- Los hosts donde corren las VMs del cliente están distribuidos por el DC



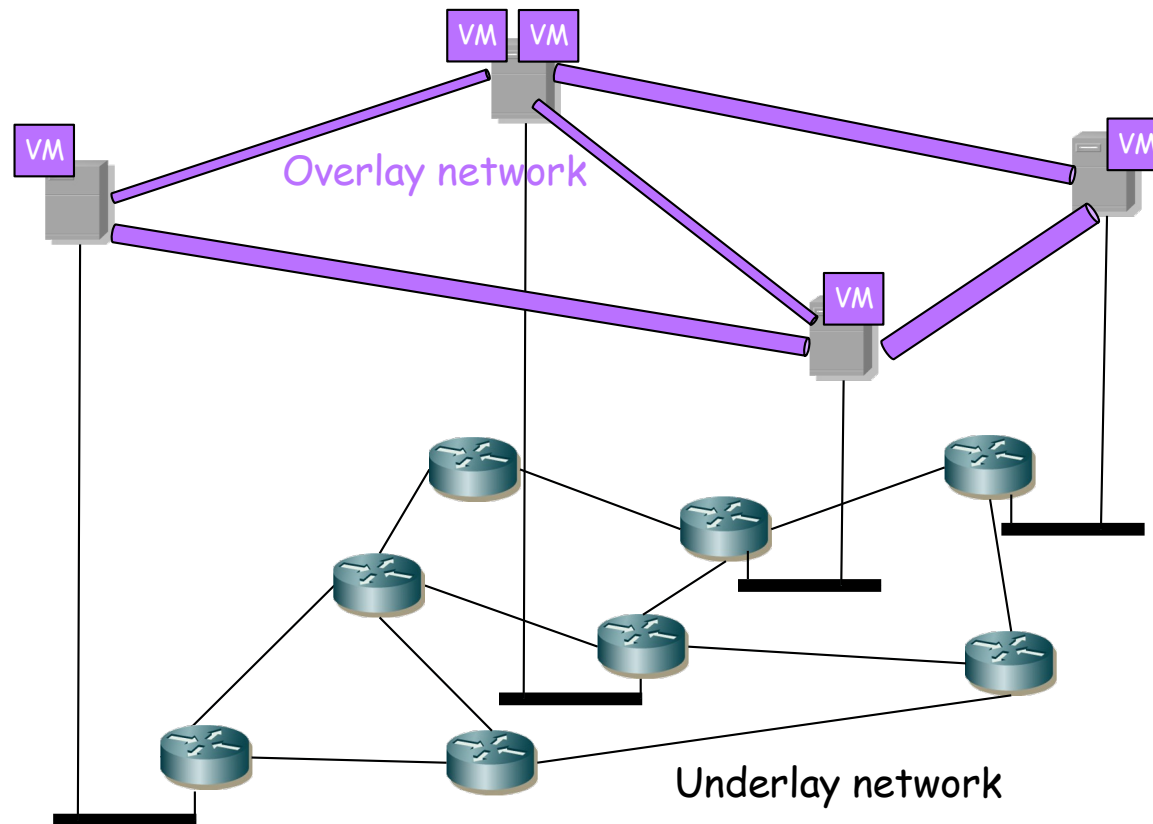
Overlay network

- El DC dispone una red (underlay network)
- Combinación de Ethernet, IP, MPLS, etc
- Los hosts donde corren las VMs del cliente están distribuidos por el DC
- Esos hosts crean la overlay mediante túneles



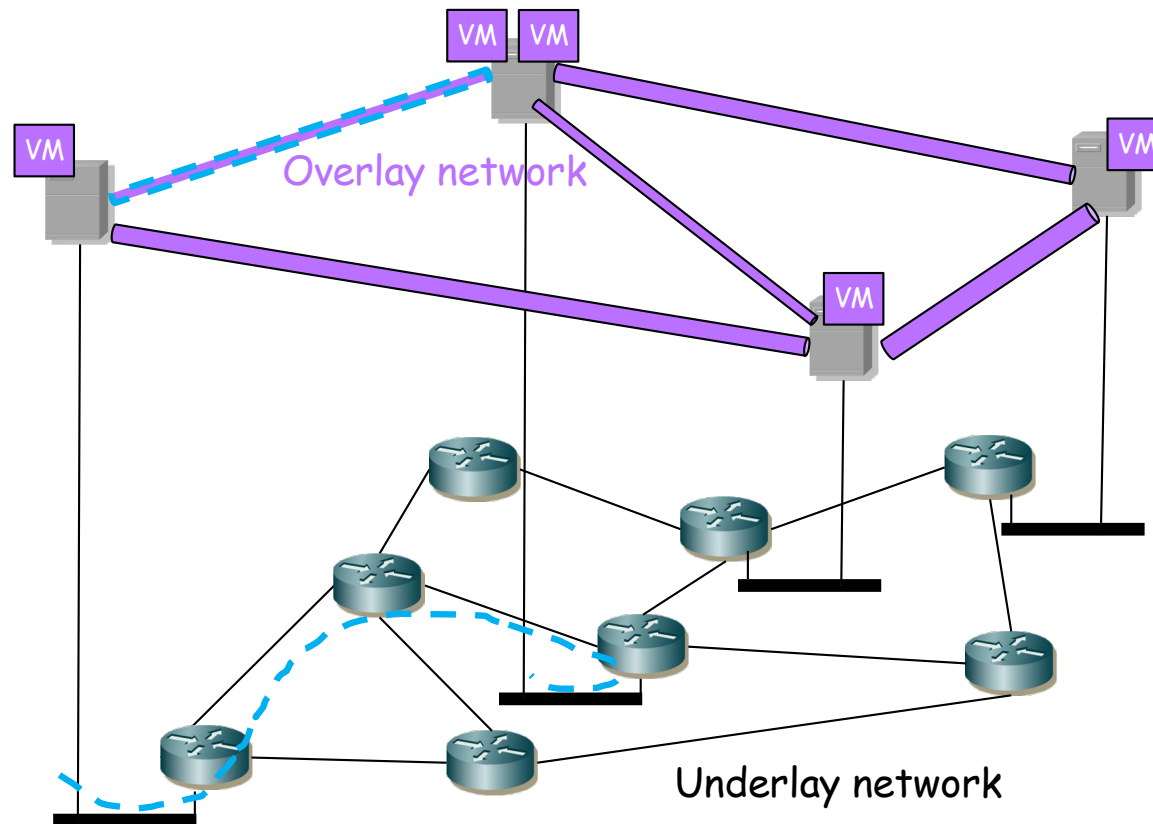
Overlay network

- Comunicación entre VMs:
 - Paquete de una VM de overlay se encapsula en el primer salto o NVE (*Network Virtualization Edge*) (Switch, router, vSwitch)
 - Túnel hasta el NVE remoto
 - La red reenvía en base a esta encapsulación, ignorando el contenido
 - El NVE de egreso desencapsula y entrega a la VM (o host físico) destino
- El paquete transportado puede ser IP o Ethernet



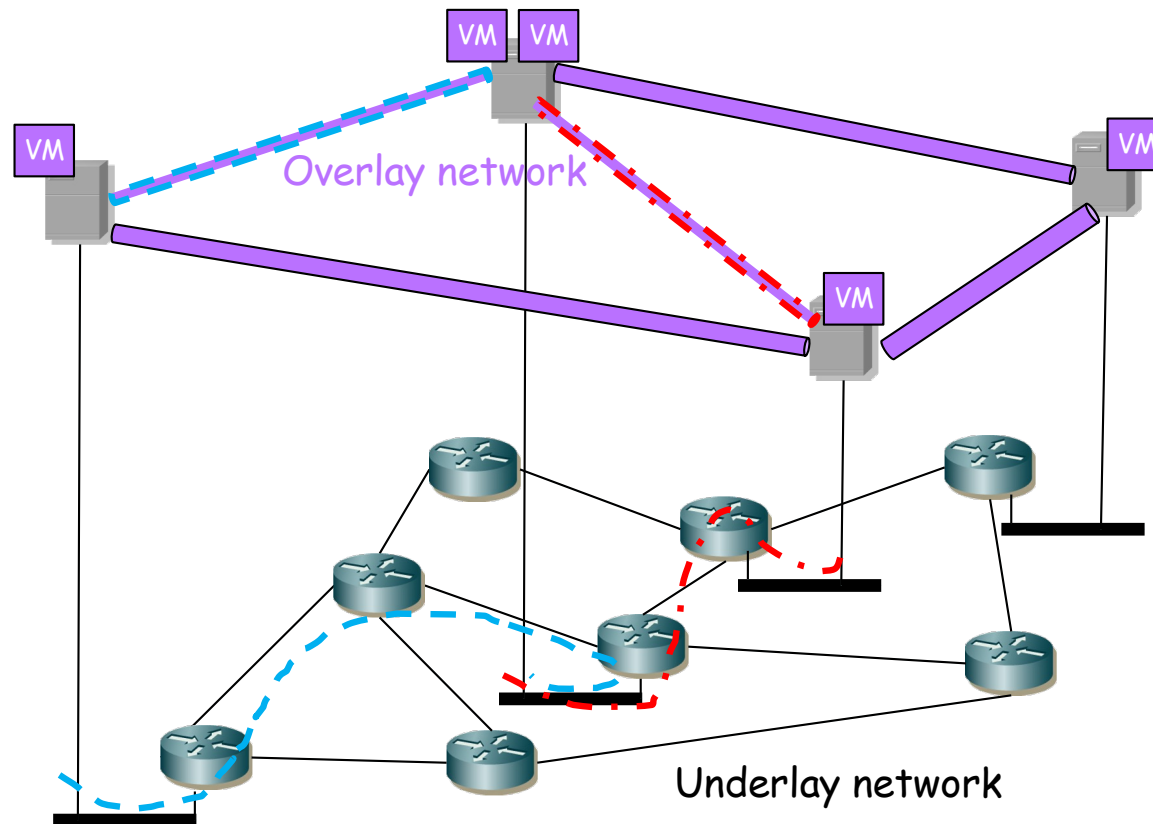
Overlay network

- El tráfico de cada túnel sigue el camino que elija la underlay



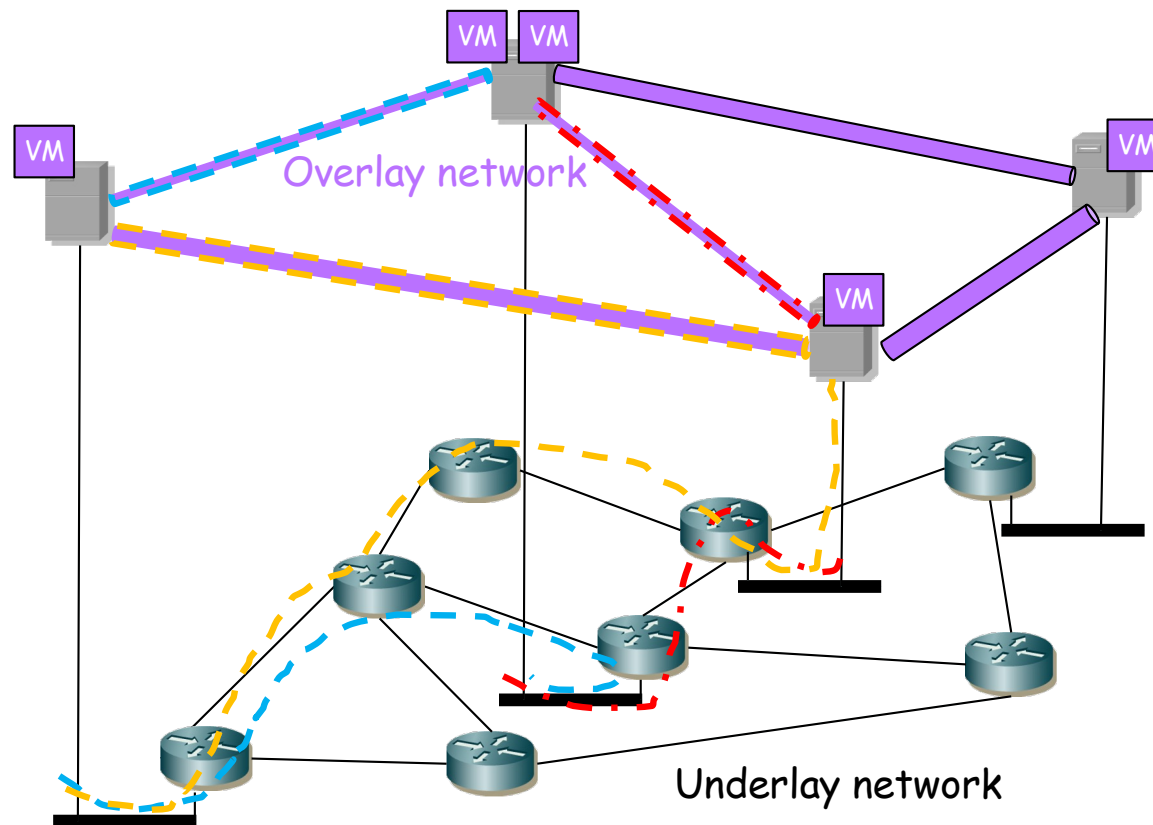
Overlay network

- El tráfico de cada túnel sigue el camino que elija la underlay



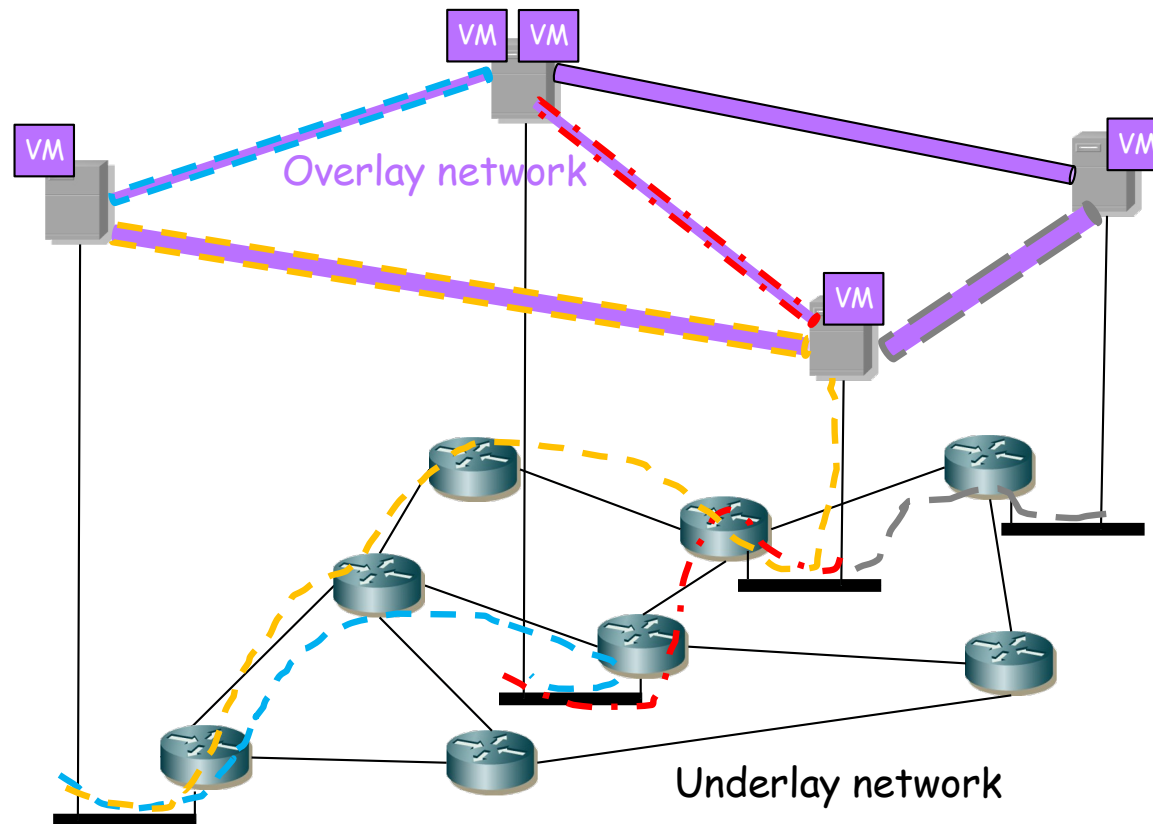
Overlay network

- El tráfico de cada túnel sigue el camino que elija la underlay



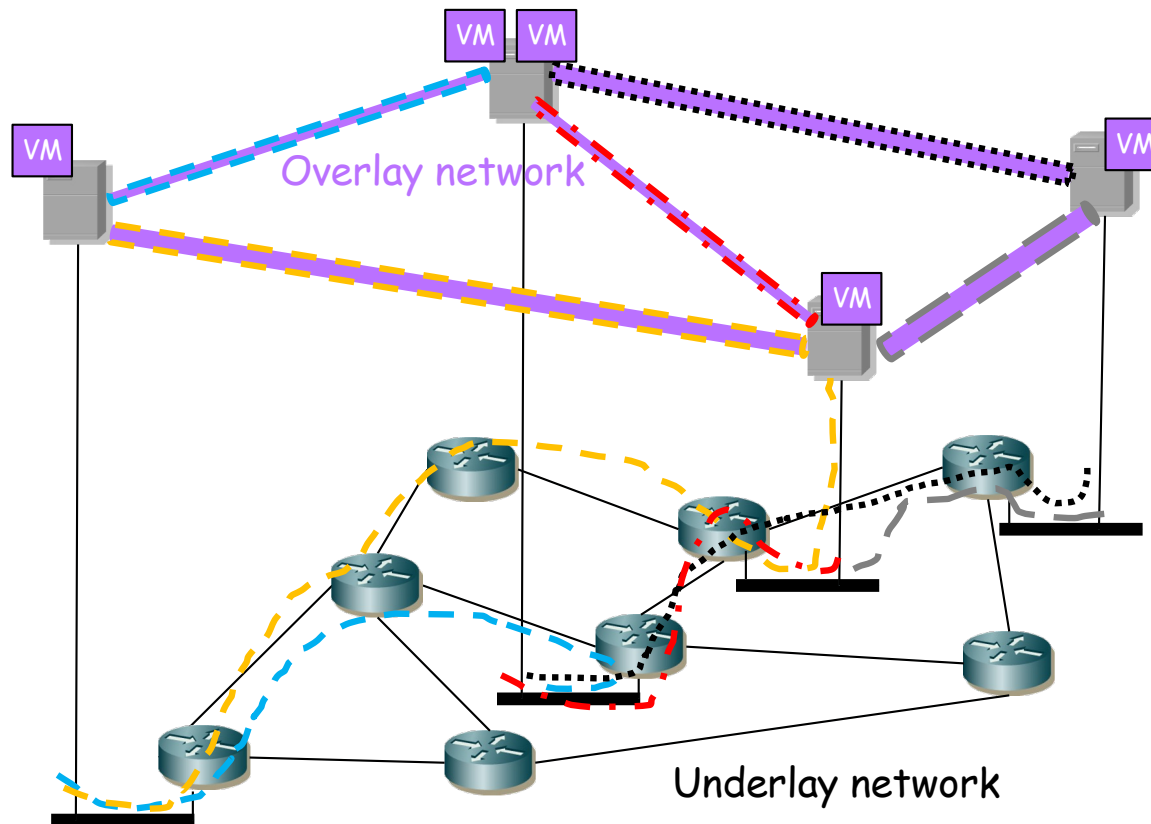
Overlay network

- El tráfico de cada túnel sigue el camino que elija la underlay



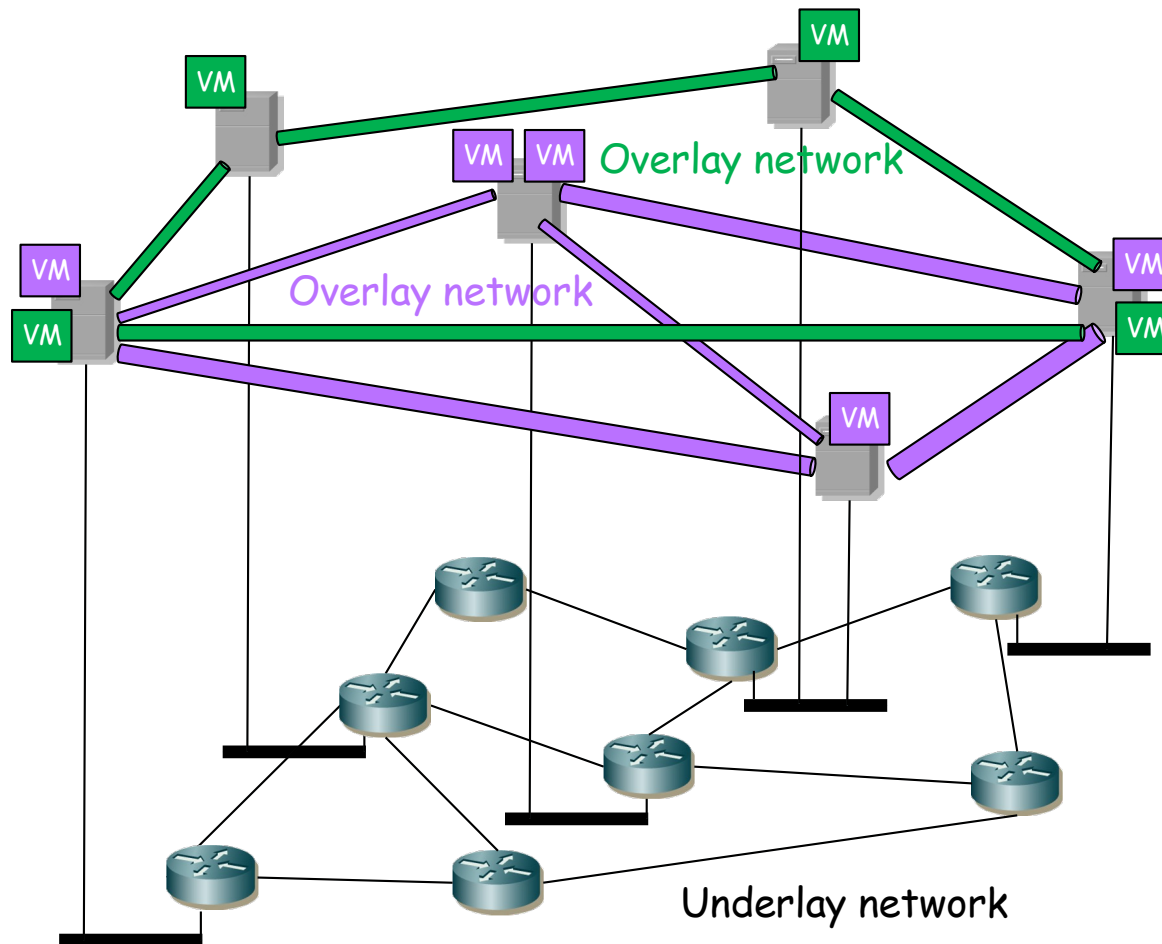
Overlay network

- El tráfico de cada túnel sigue el camino que elija la underlay
- Esos caminos podrían ser simétricos o asimétricos
- Transparente para la overlay



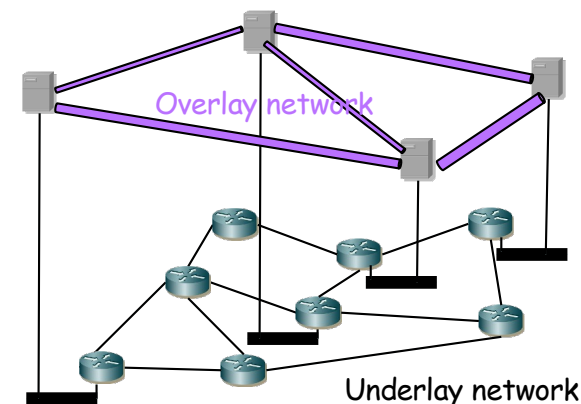
Overlay network

- Cada Virtual Network (VN) es una *overlay*



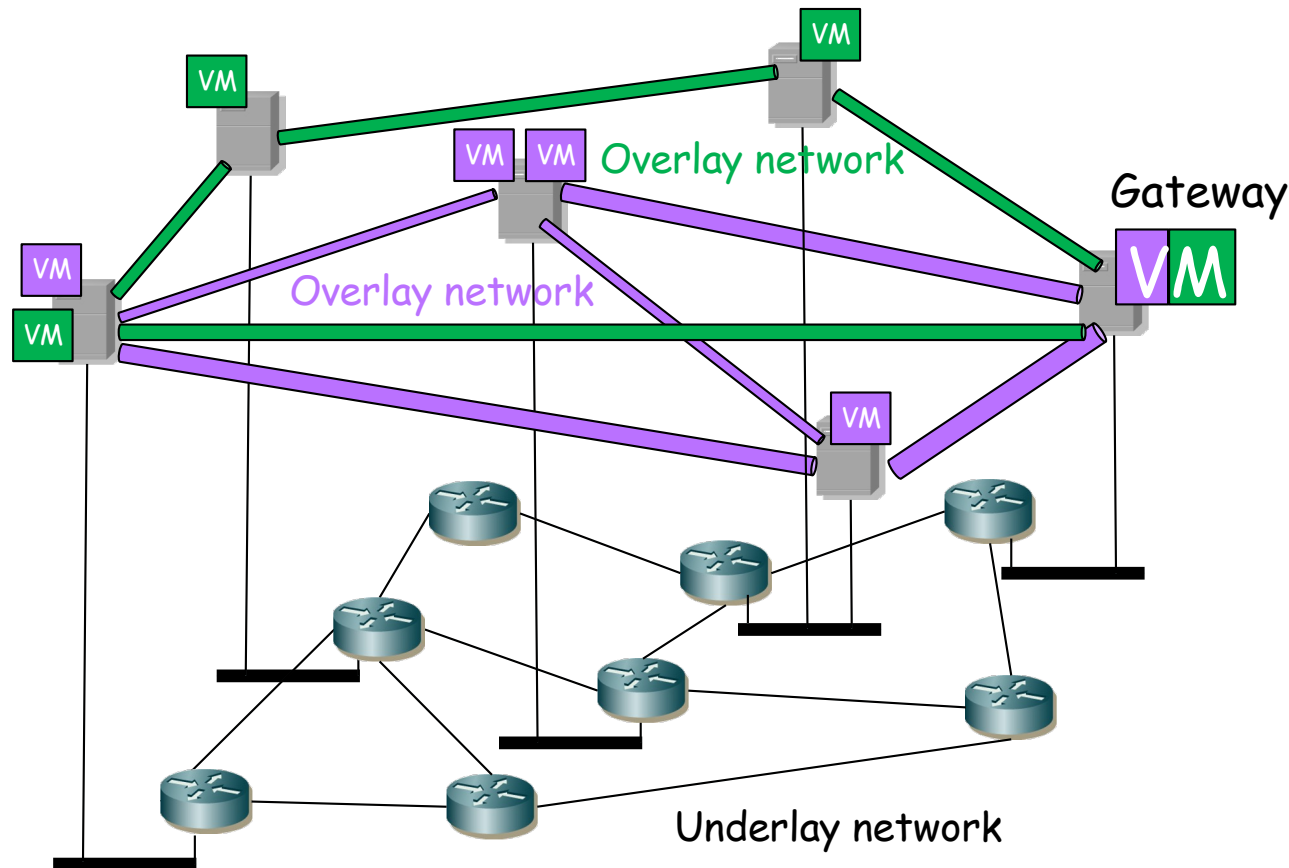
Overlays

- Debe permitir gran número de overlay networks
- Miembros de la overlay muy dispersos por el DC
- VMs de la overlay muy dinámicos (creación, destrucción, on, off, move)
- Sin requerir cambios en la underlay network
- Permiten que las tablas de direcciones MAC de los conmutadores de la underlay no crezcan con el número de VMs
- Para ello intentan evitar que los conmutadores del núcleo aprendan las direcciones MAC de las VMs (hosts de overlay)
- Esto lo van a hacer encapsulando las tramas Ethernet de los hosts extremo
- Para entornos con mucho tráfico este-oeste en vez de norte-sur



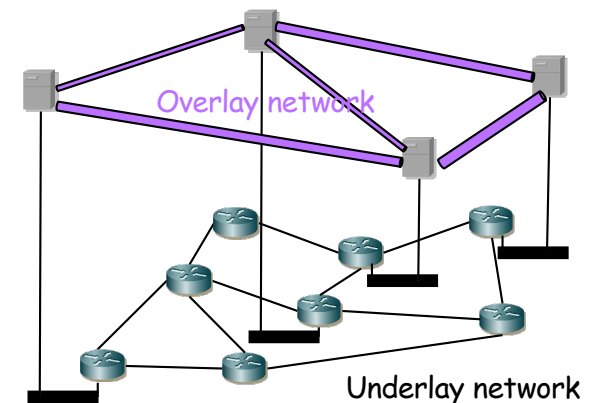
Comunicación al exterior

- Un equipo hará de *gateway*
- Puede ser por ejemplo un equipo con interfaces en dos overlays
- O con otro interfaz en una subred de la underlay
- Puede ser una VM, un vSwitch o un equipo físico
- Si enruta a otra overlay deben no colisionar sus espacios de direcciones



Overlays

- Alternativas existentes:
 - BGP/MPLS IP o Ethernet VPNs
 - TRILL (Transparent Interconnection of Lots of Links)
 - SPB (Shortest Path Bridging)
 - NVGRE (Network Virtualization using GRE)
 - OTV (Overlay Transport Virtualization)
 - VXLAN (Virtual Extensible LAN)
 - FabricPath (TRILL)
 - LISP (Locator/ID Separation Protocol)
 - Geneve (Generic Network Virtualization Encapsulation)

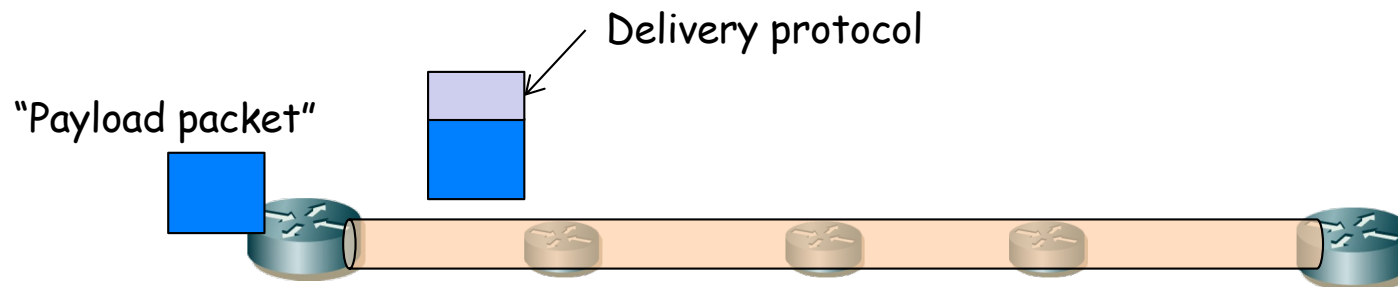


Overlays en el data center

Túneles básicos

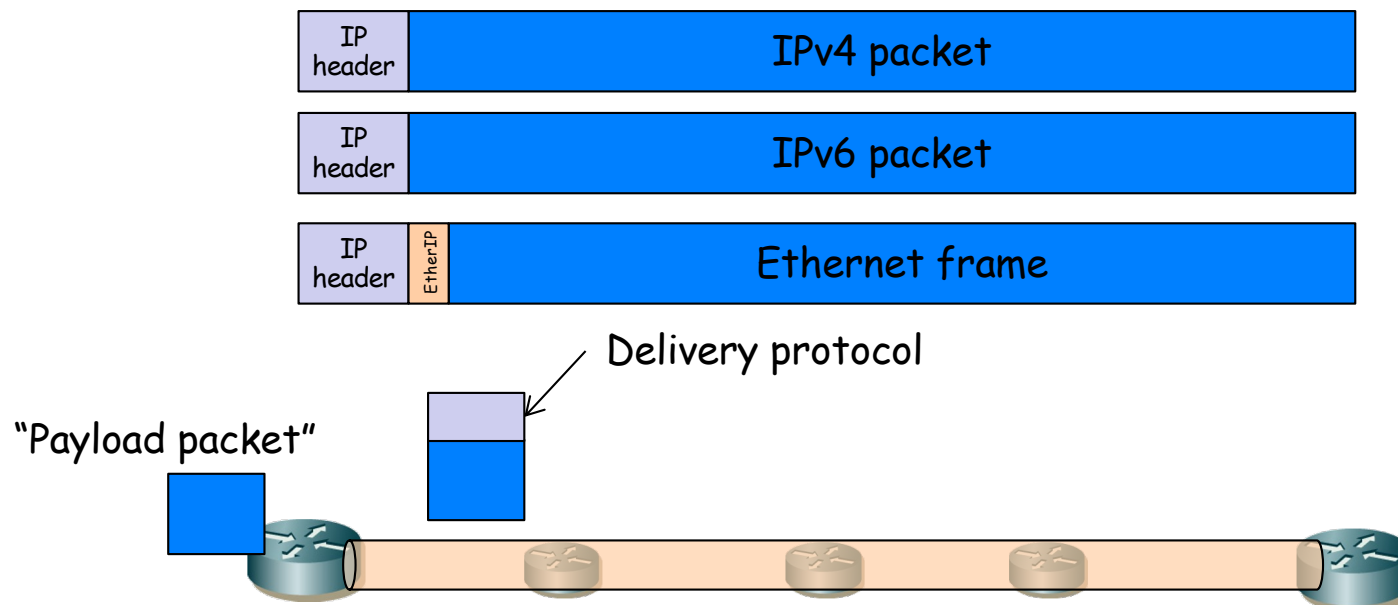
Túnel directo sobre IP

- Un extremo introduce el paquete a transportar en un paquete IP
- El paquete IP va dirigido a la dirección del otro extremo del túnel
- La red intermedia encamina en función de esa dirección destino, independiente del contenido
- ¿Qué podemos transportar dentro de IP?
 - (...)



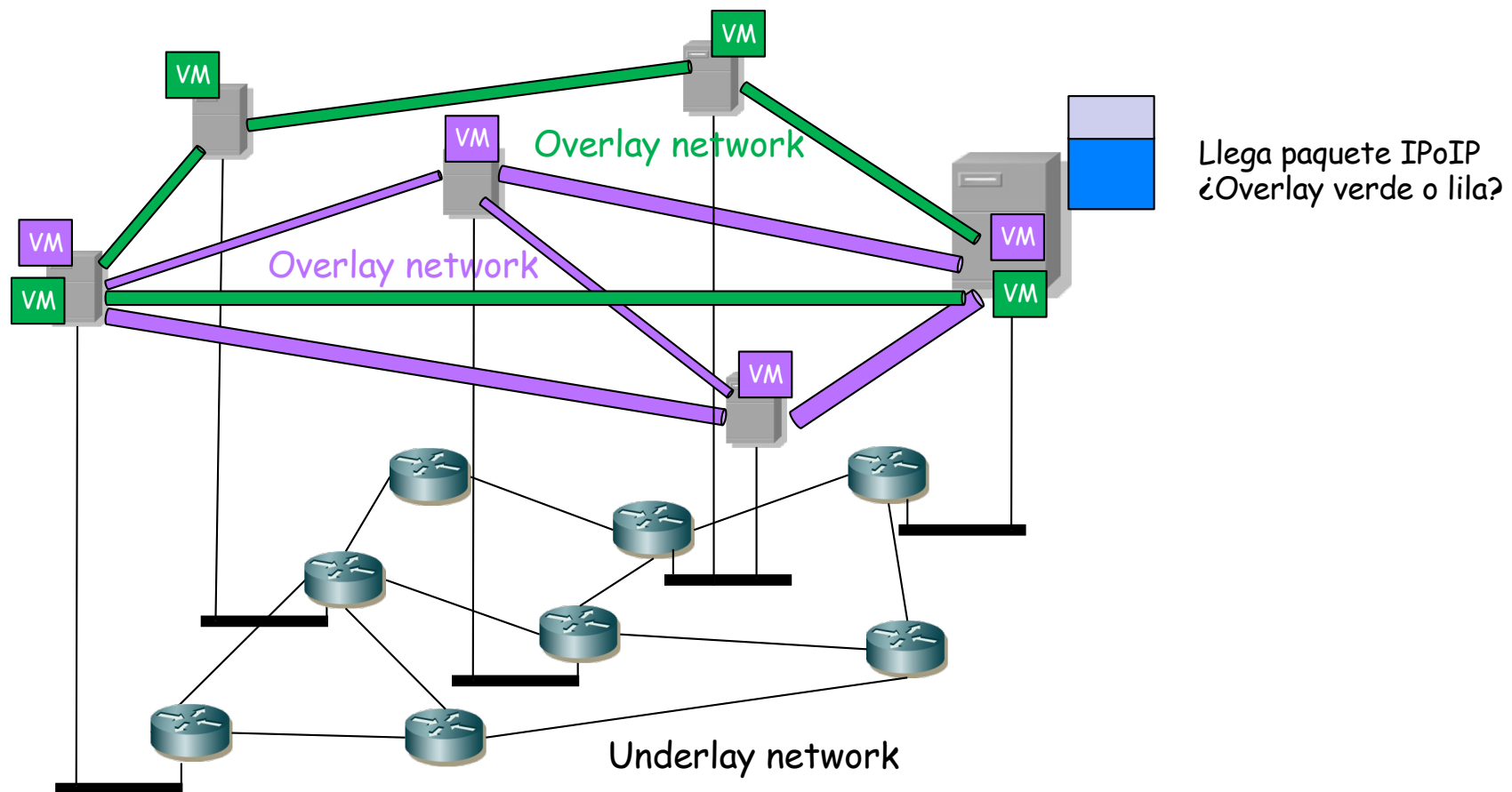
Túnel directo sobre IP

- Un extremo introduce el paquete a transportar en un paquete IP
- El paquete IP va dirigido a la dirección del otro extremo del túnel
- La red intermedia encamina en función de esa dirección destino, independiente del contenido
- ¿Qué podemos transportar dentro de IP?
 - Protocol = 4 : IPv4
 - Protocol = 41 : IPv6
 - Protocol = 97 : Ethernet-within-IP Encapsulation (RFC 3378), cabecera EtherIP de 2 bytes seguida de trama Ethernet



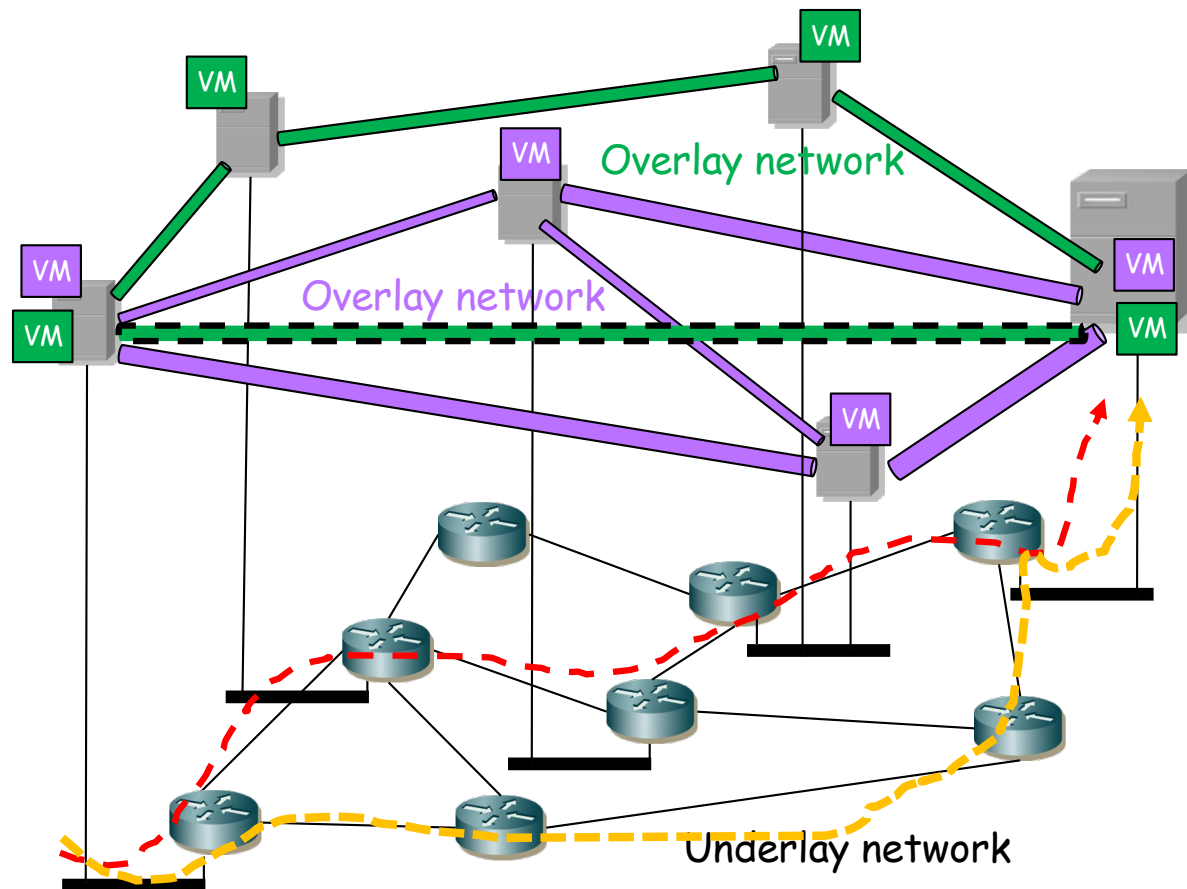
Uso en overlays

- No existe un identificador de la overlay en el paquete recibido
- Una vez extraído el paquete contenido no se puede saber para qué overlay presente en el destino va dirigido
- Se podría identificar por la dirección IP a la que va dirigido (una dirección IP en cada host para cada overlay)



Uso en overlays

- Todo un túnel es un mismo flujo IP-a-IP
- No podemos aprovechar ECMP en la underlay si queremos evitar desorden

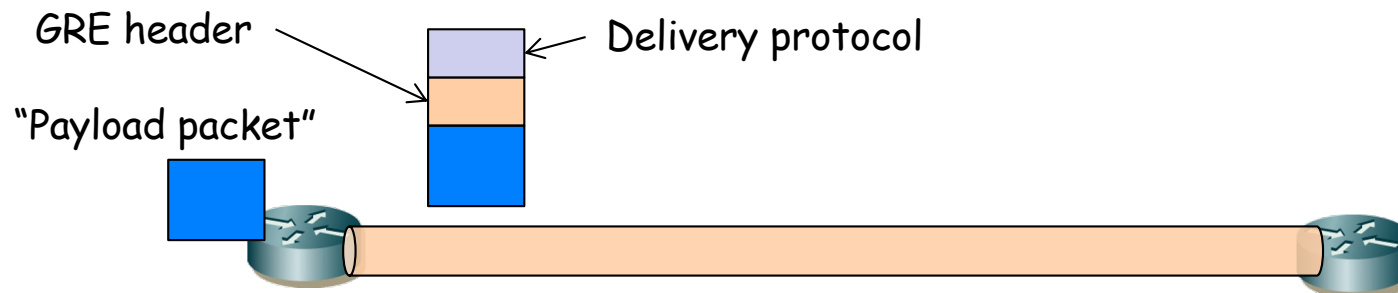


GRE

GRE

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

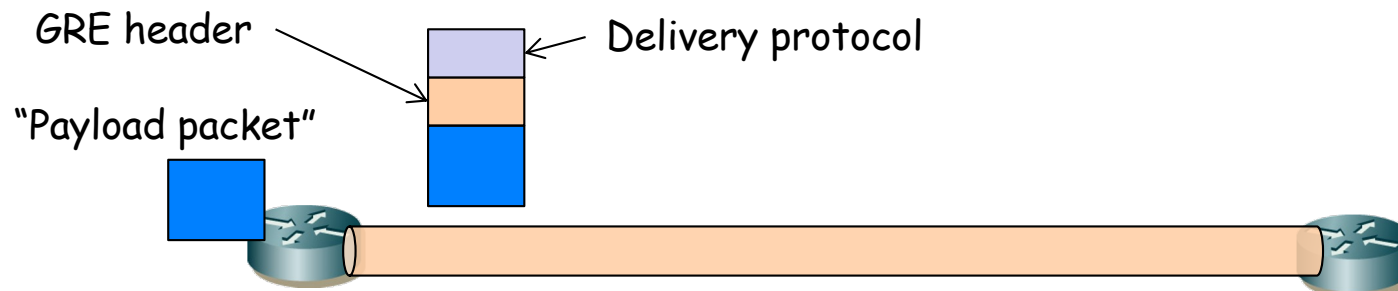
0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1
+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	C																														
+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+



GRE

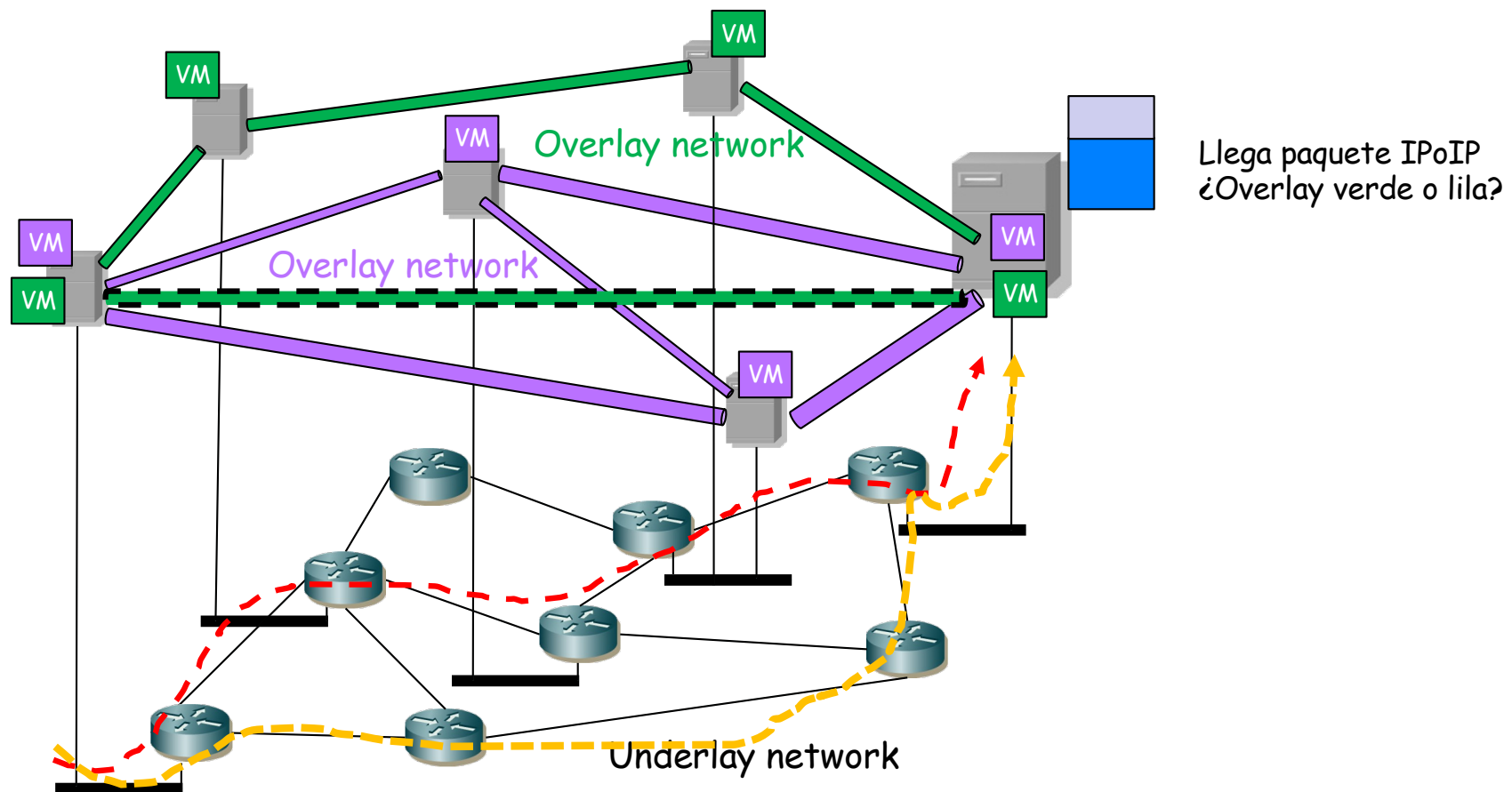
- RFC 2890 “Key and Sequence Number Extensions to GRE”
- “Key” sirve para distinguir flujos dentro del túnel
- “Sequence Number”
 - Si hay “key” entonces el número de secuencia es por “key”
 - Permite dar entrega en orden (aunque no fiable)
 - Si llega uno “anterior” lo descarta
 - Si llega uno que deja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



Uso en overlays

- Muchos chips de conmutador pueden calcular el hash para ECMP usando el campo key de GRE, permitiendo el reparto multipath
- No existe un identificador de la overlay en el paquete recibido
- Se podría identificar por la dirección IP a la que va dirigido (una dirección IP en cada host para cada overlay)





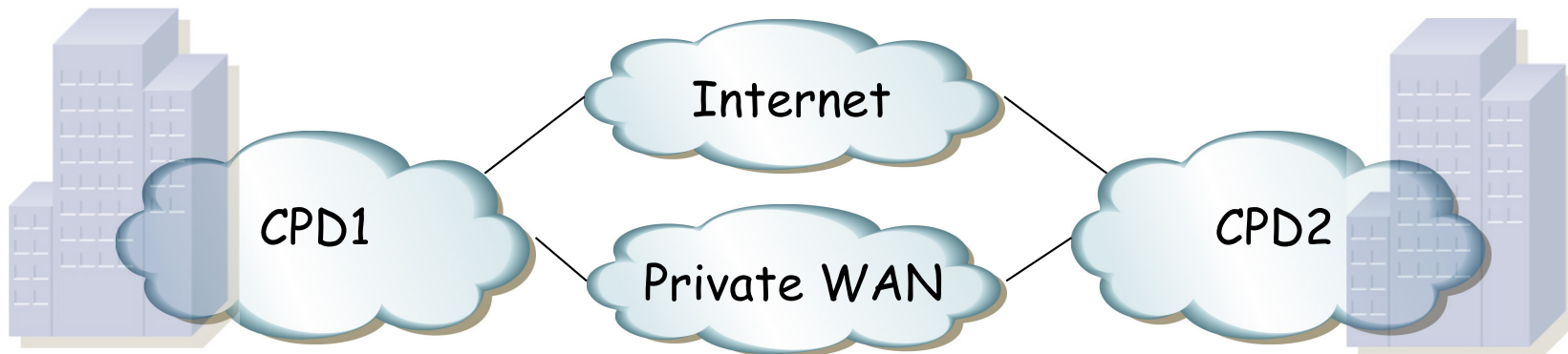
GRE



VXLAN

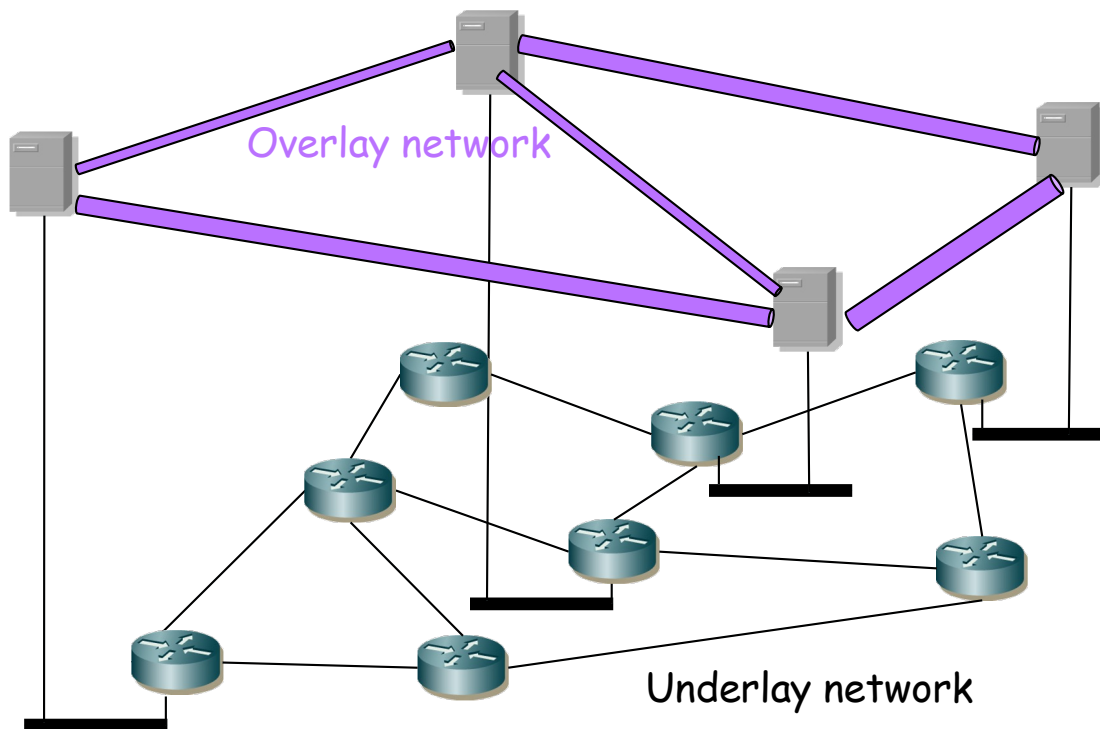
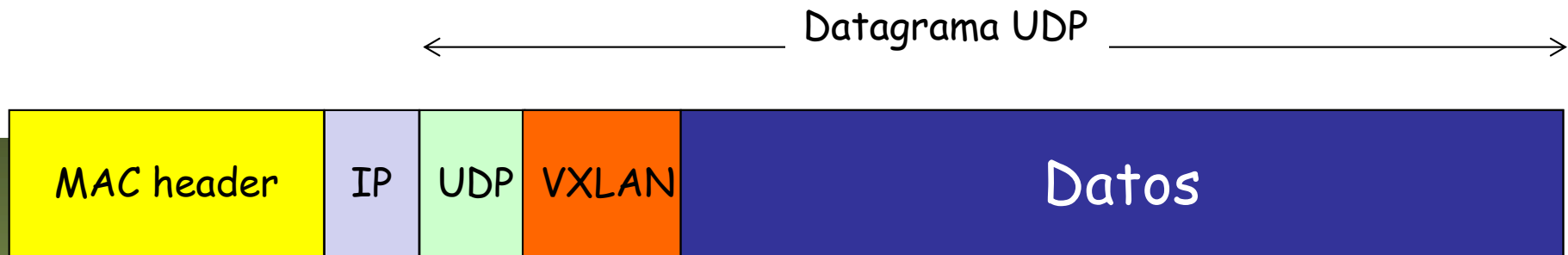
VXLAN

- RFC 7348 “Virtual eXtensible Local Area Network (VXLAN): A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks” (agosto 2014)
- RFC Informativa firmada por Cisco, VMware, Intel, Red Hat, Arista y Cumulus Networks
- Diseñado para un entorno de host virtualizado
- Emplea un esquema de overlay de capa 2 sobre capa 3 (o sea, un túnel), en el mismo data center o en otro



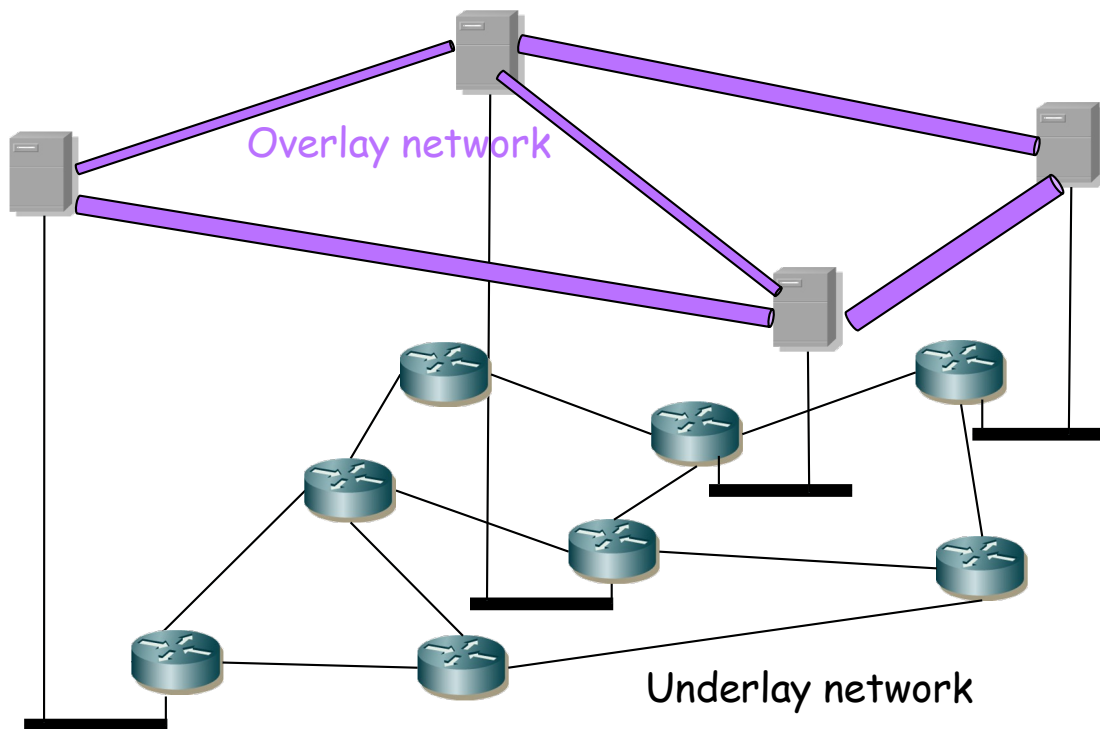
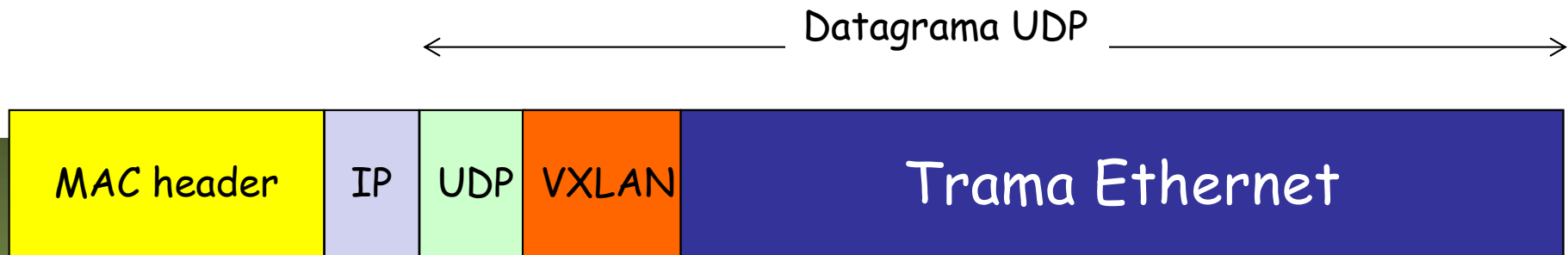
VXLAN

- Túnel sobre capa 4 pues hace el transporte sobre UDP
- Cabecera VXLAN con identificador de la Virtual Network (VNI)



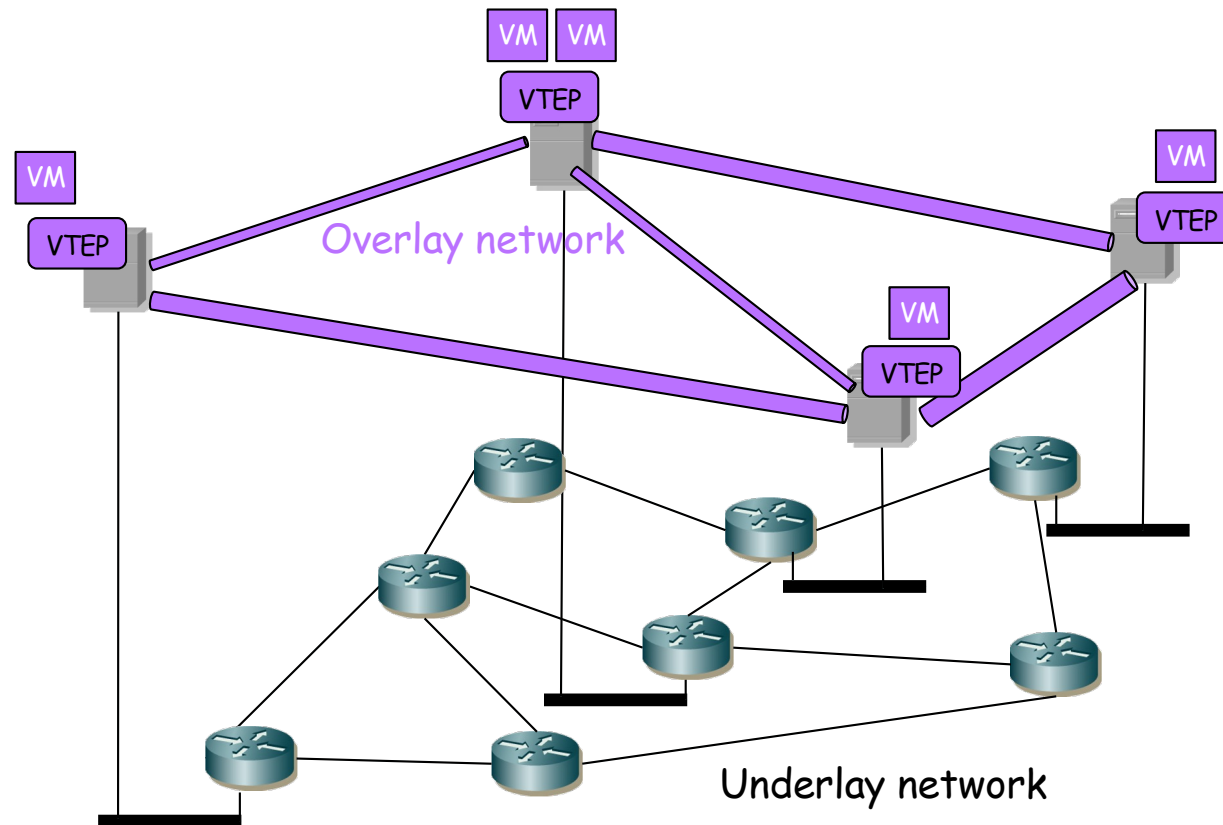
VXLAN

- Contenido: trama Ethernet entregada por la VM



VXLAN

- El extremo del túnel es el VTEP (VXLAN Tunnel EndPoint)
- Puede estar en un hypervisor o en un switch físico cercano



Uso en overlays

- La dirección IP origen no identifica a la overlay, sino el VNI
- Otra overlay, otro VNI

