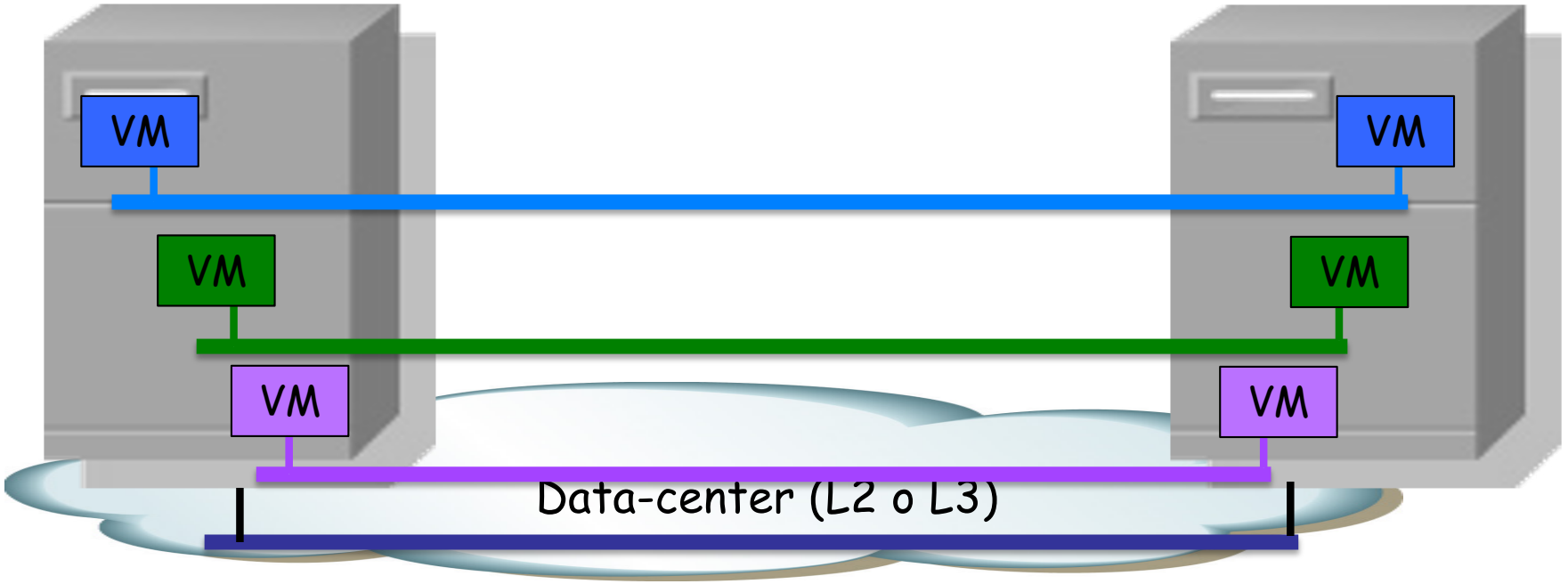
upna
Universidad Pública de Navarra
Nafarroako Unibertsitate Publikoa
Redes de Nueva Generación
Área de Ingeniería Telemática

Overlays en el data center

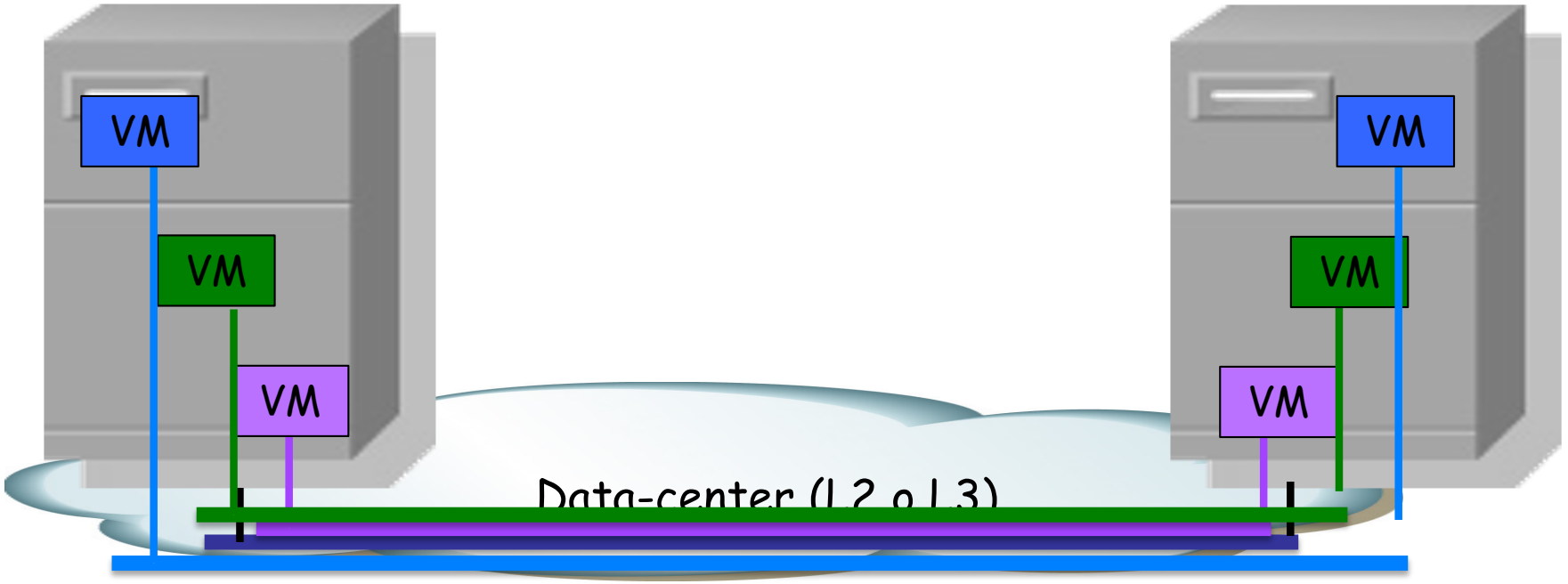
Multi-tenancy

- 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
- 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



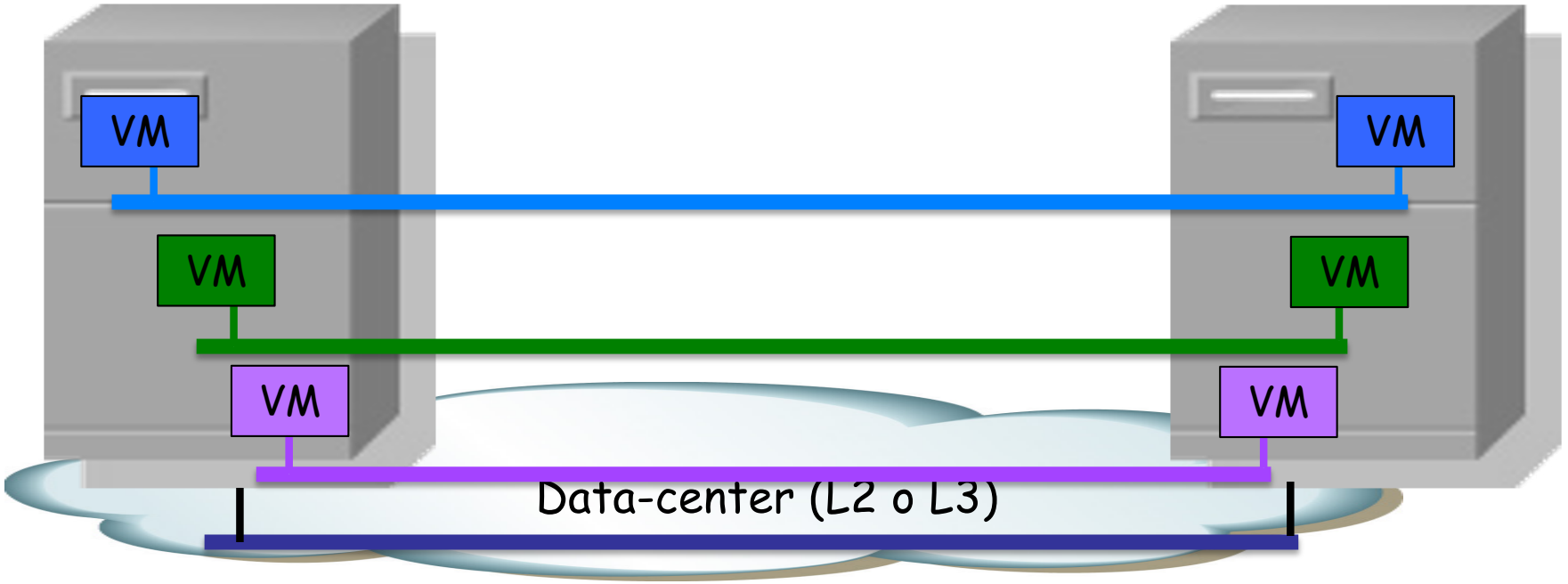
VLANs

- ¿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



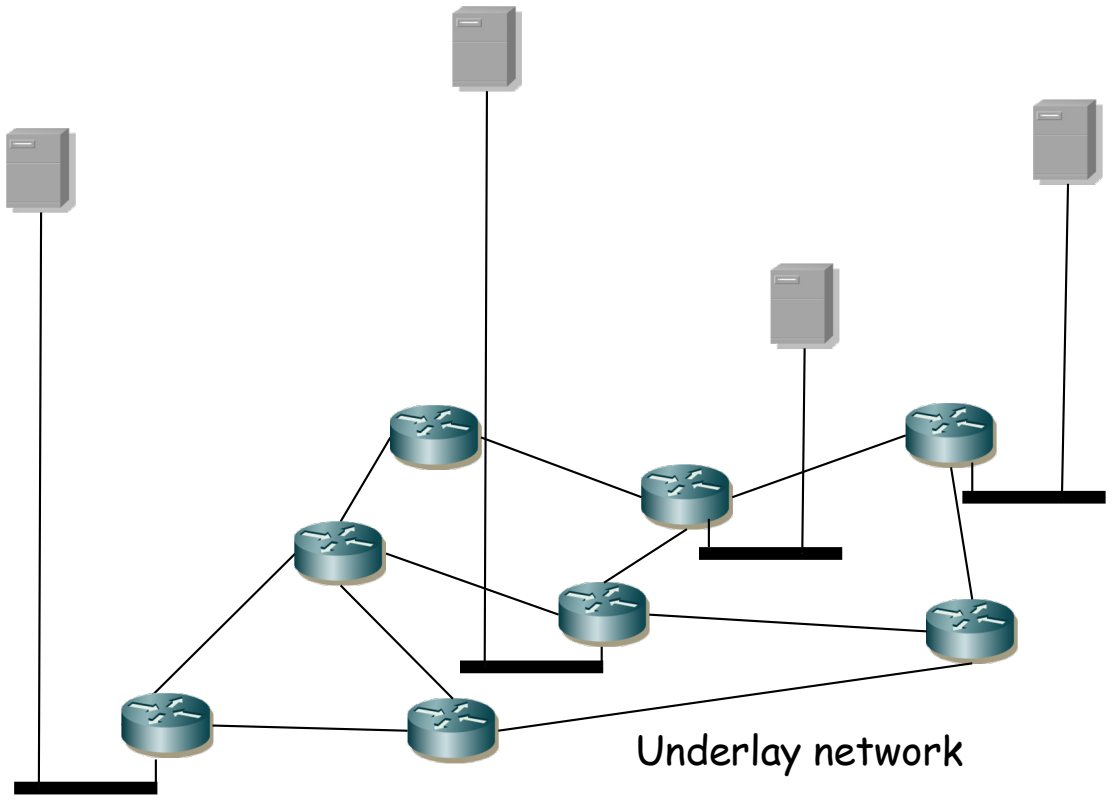
Multi-tenancy

- 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



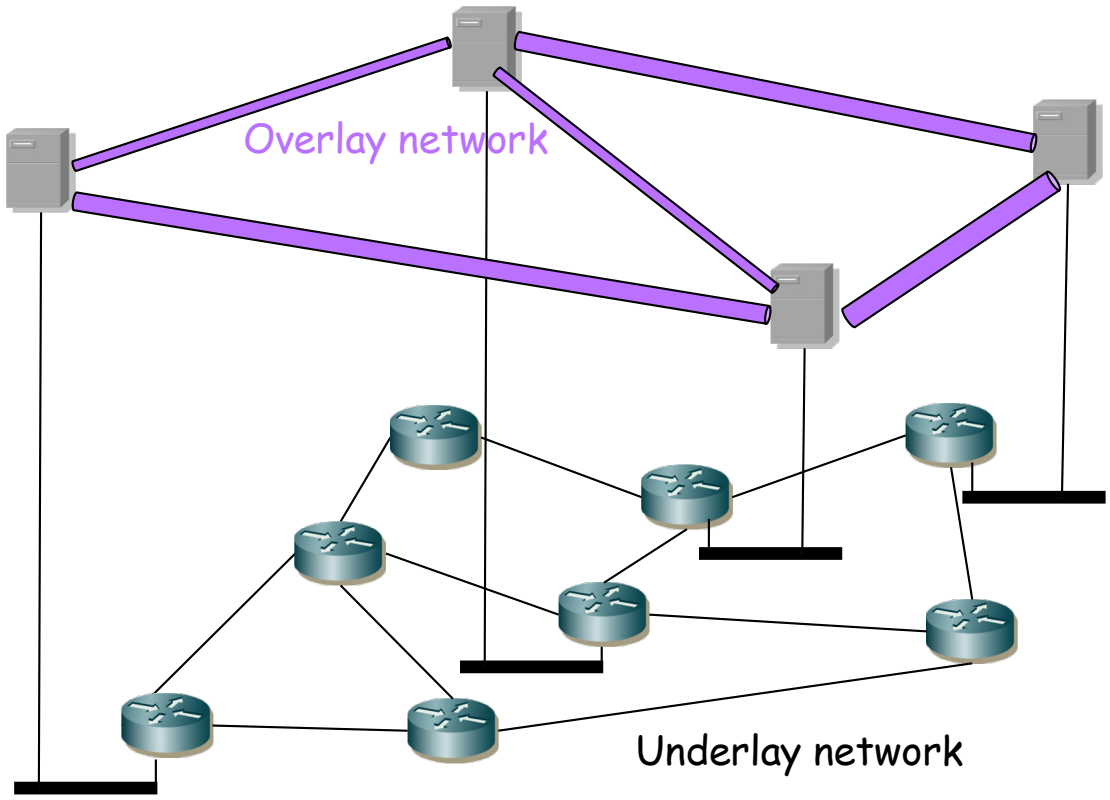
Overlay network

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



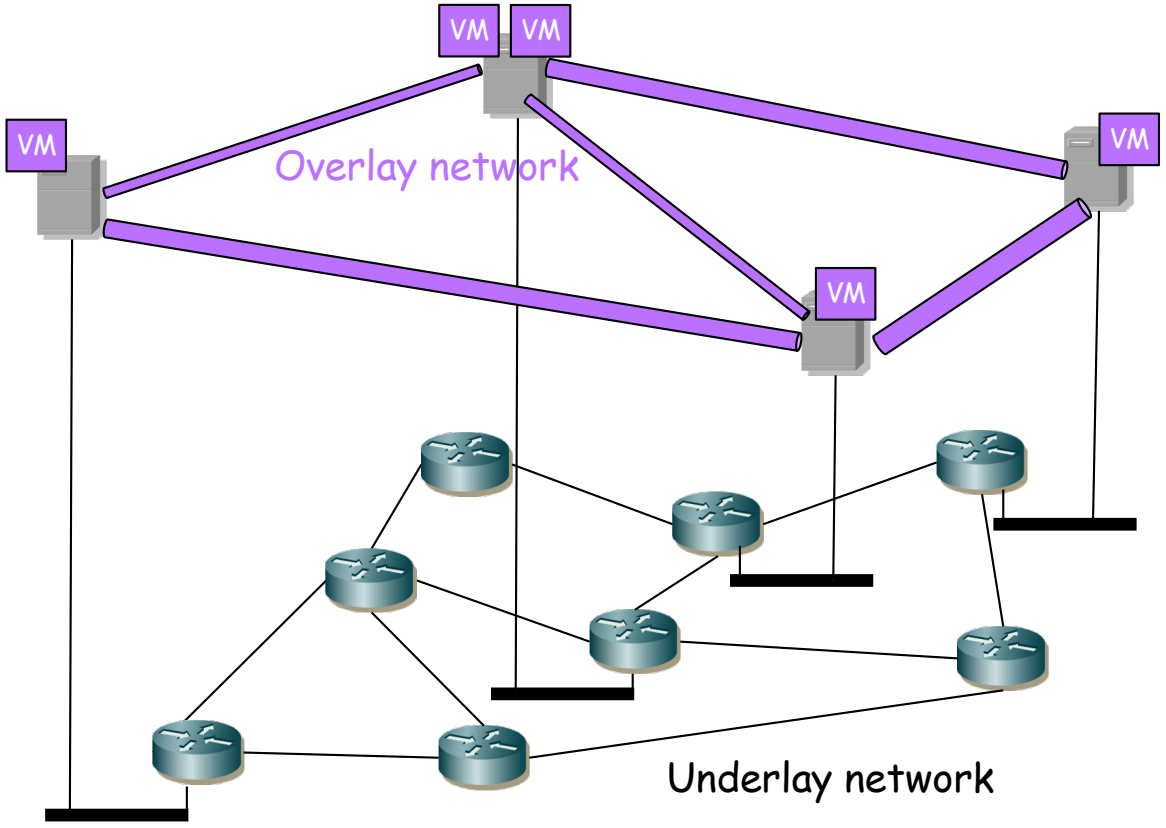
Overlay network

- 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



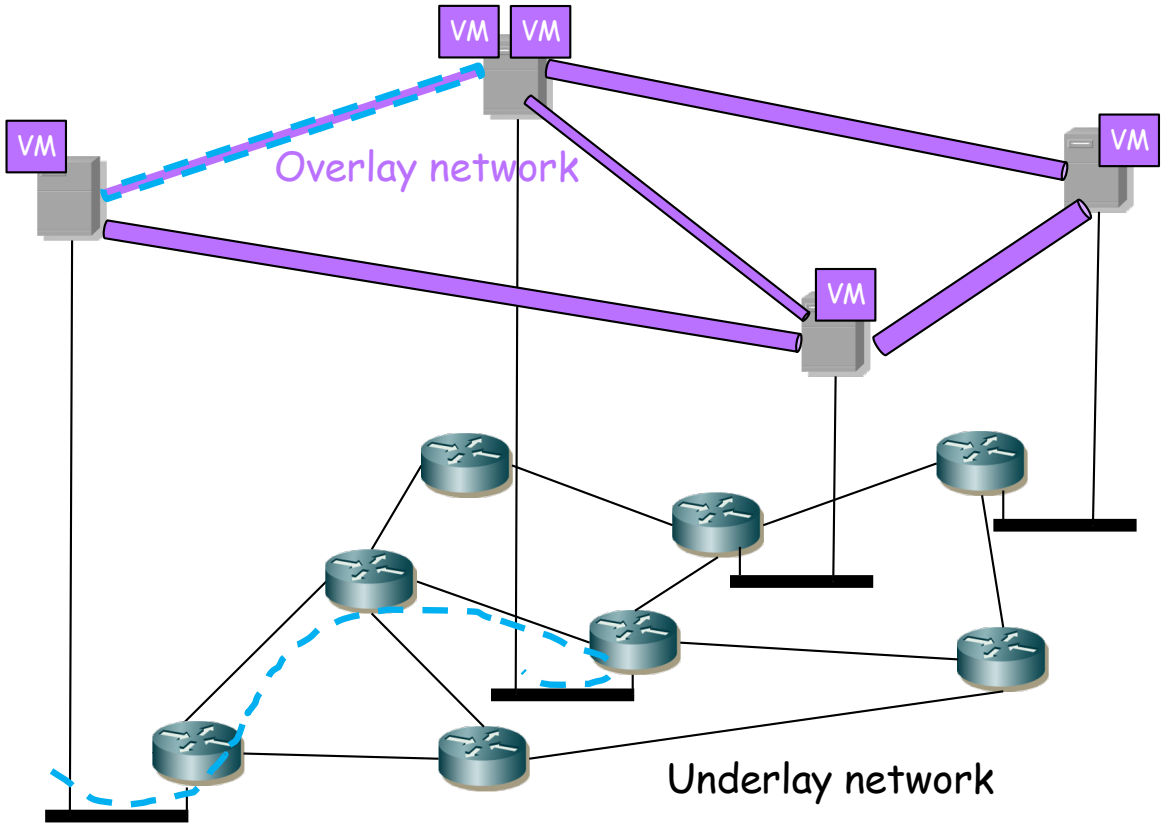
Overlay network

- 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



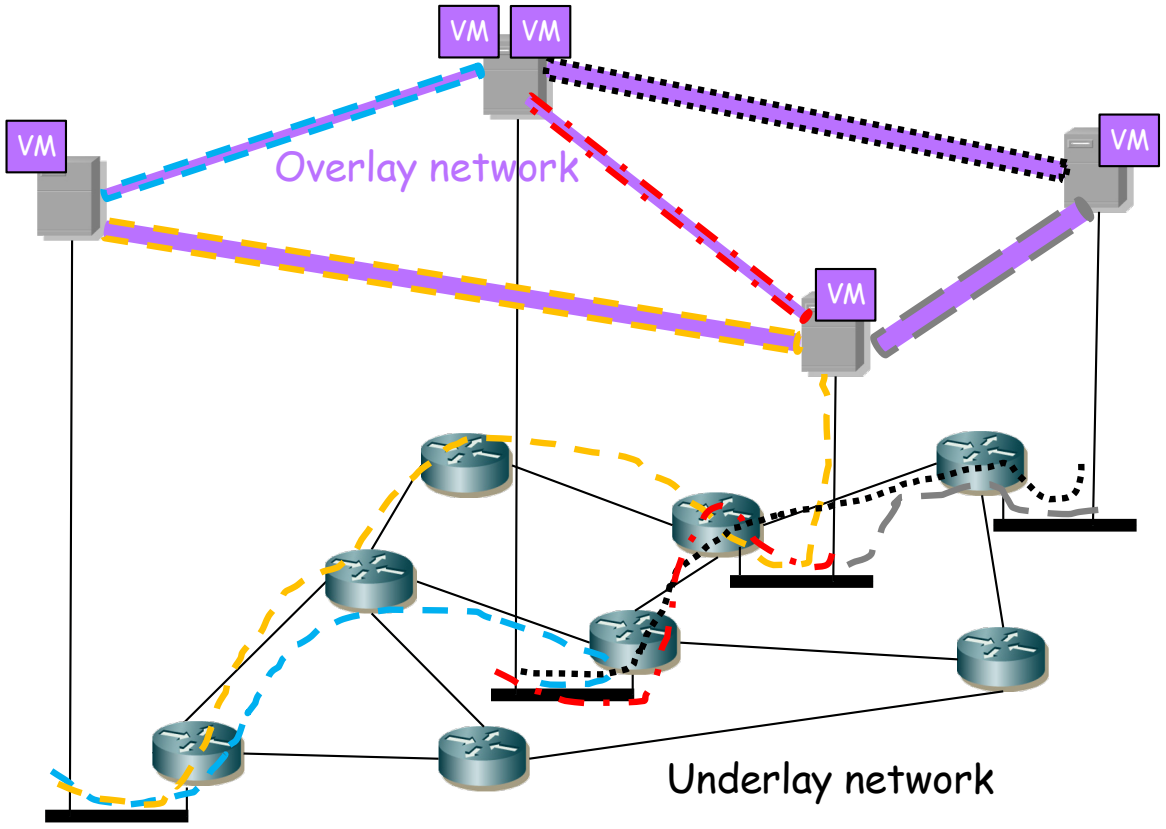
Overlay network

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



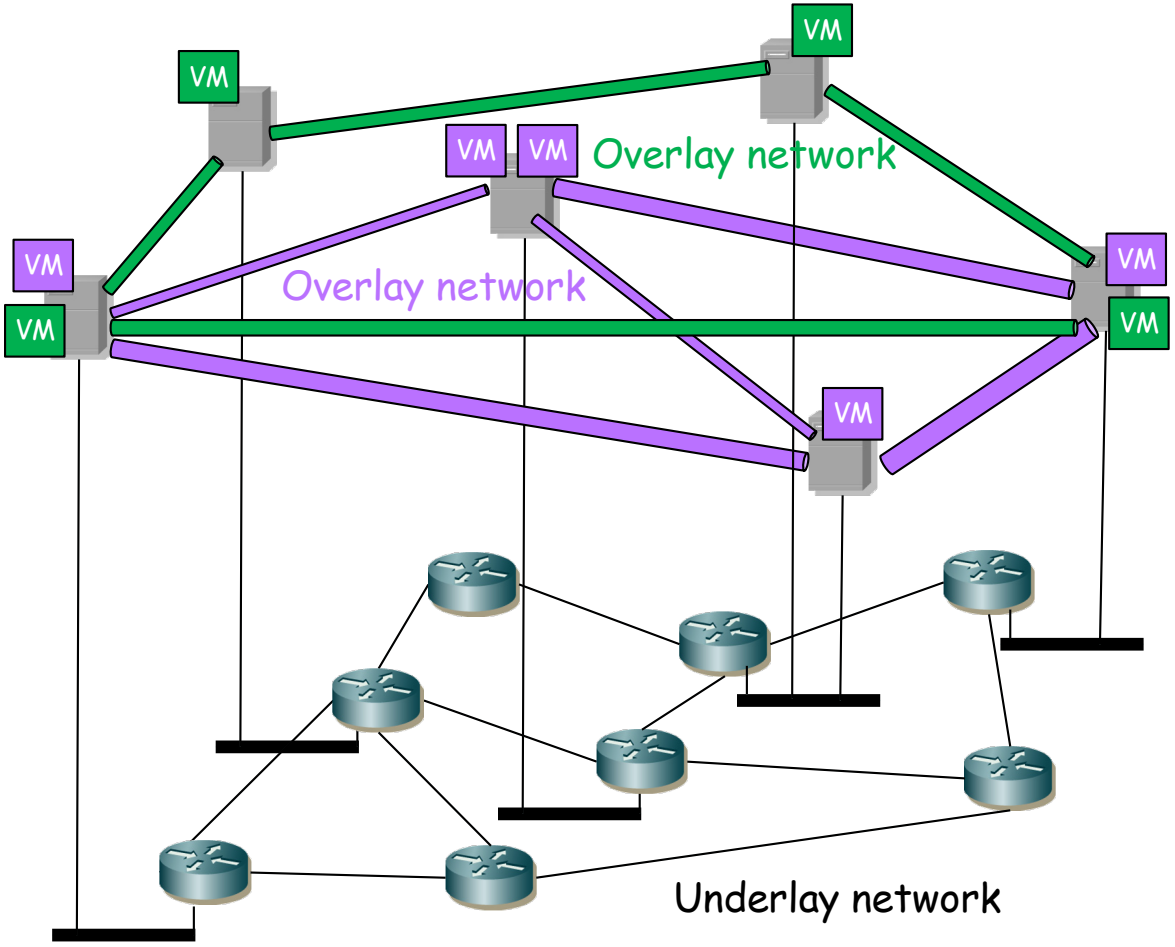
Overlay network

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



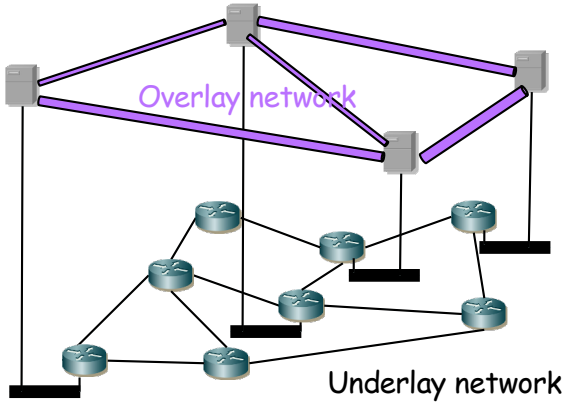
Overlay network

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



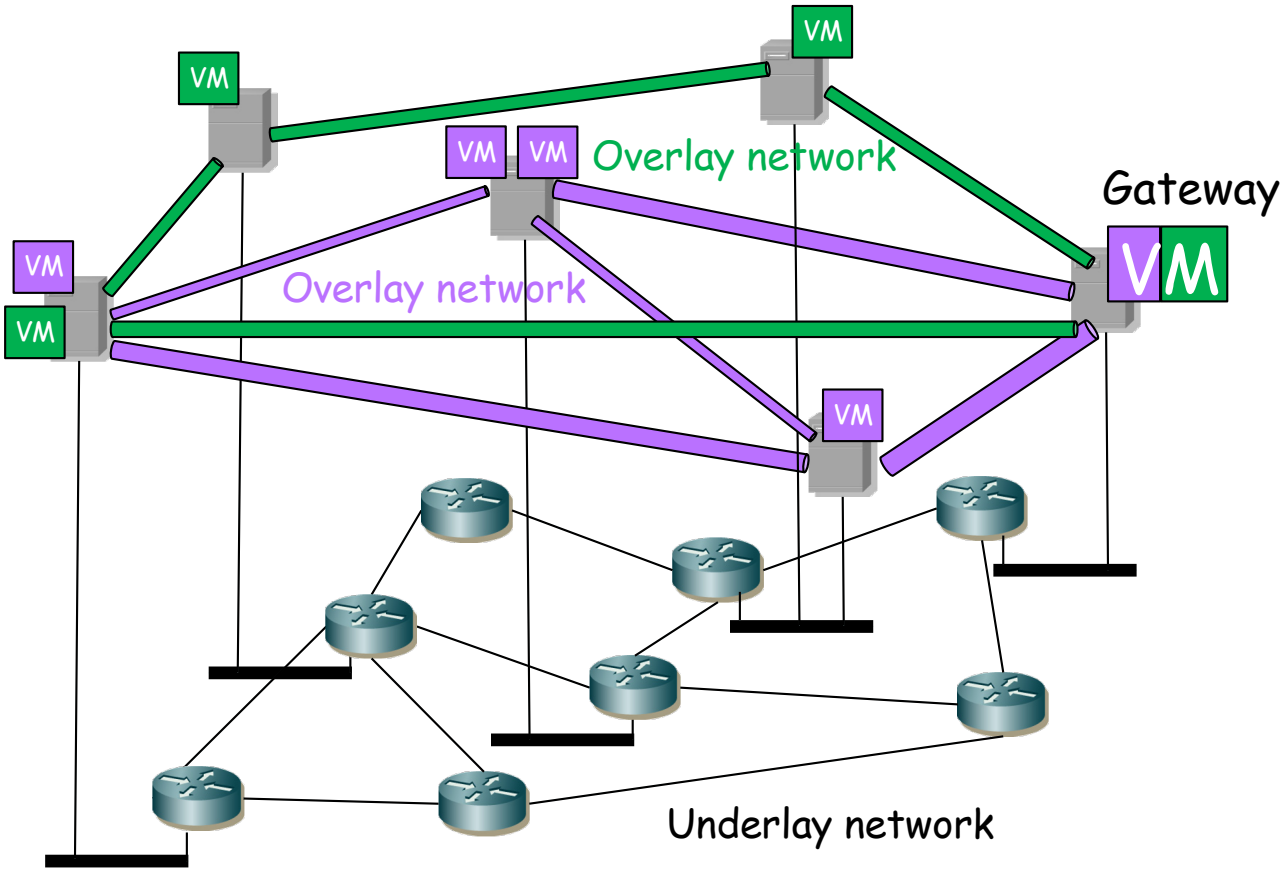
Overlays

- 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



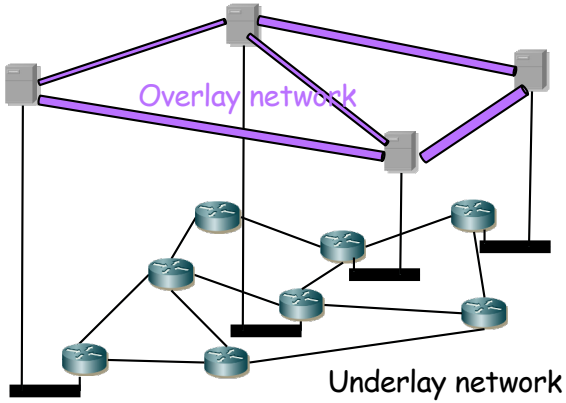
Comunicación al exterior

- 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)

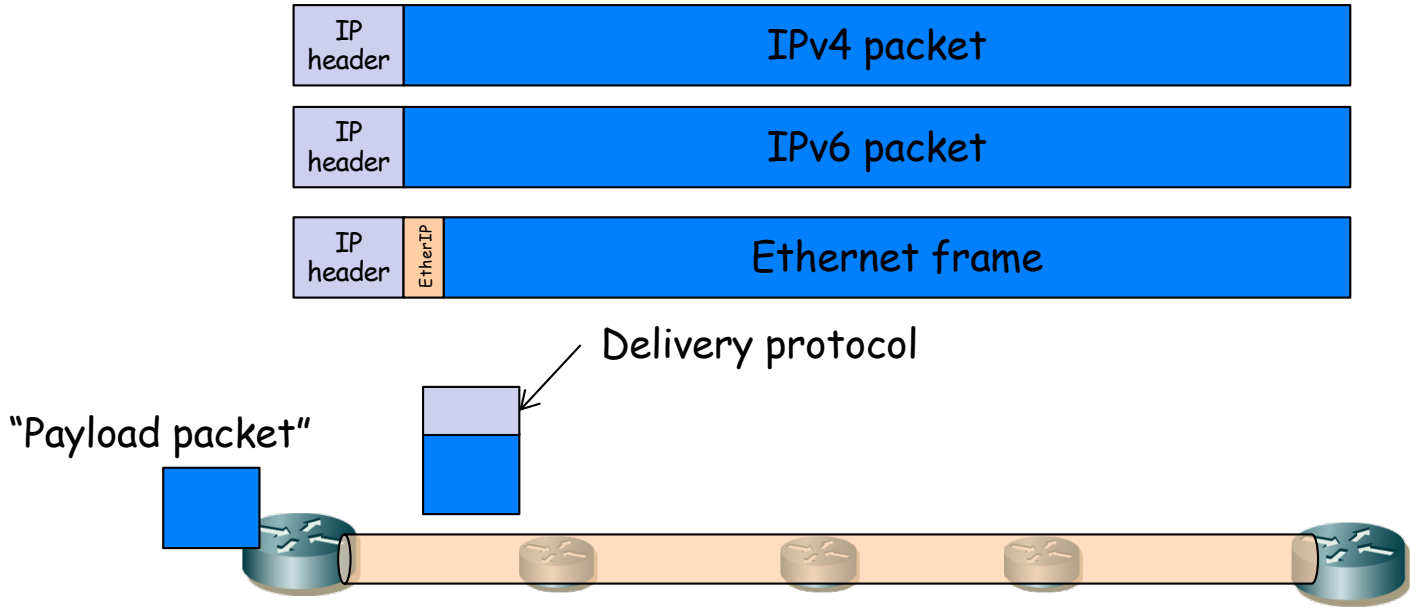


upna
Universidad Pública de Navarra
Nafarroako Unibertsitate Publikoa
Redes de Nueva Generación
Área de Ingeniería Telemática

Túneles básicos

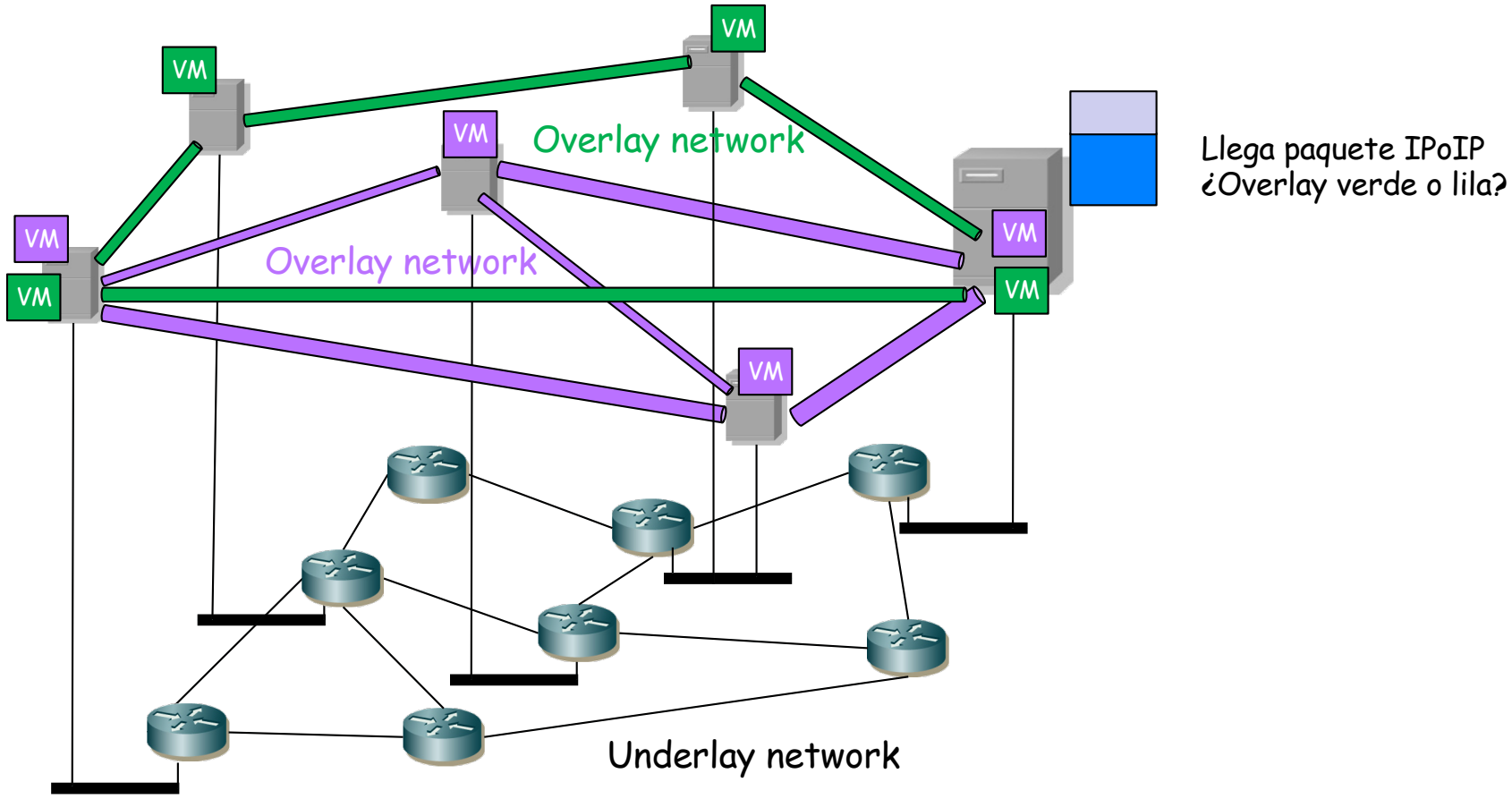
Túnel directo sobre IP

- 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?
  - Protocol = 4 : IPv4
  - Protocol = 41 : IPv6
  - Protocol = 97 : Ethernet-within-IP Encapsulation (RFC 3378), cabecera EtherIP de 2 bytes seguida de trama Ethernet



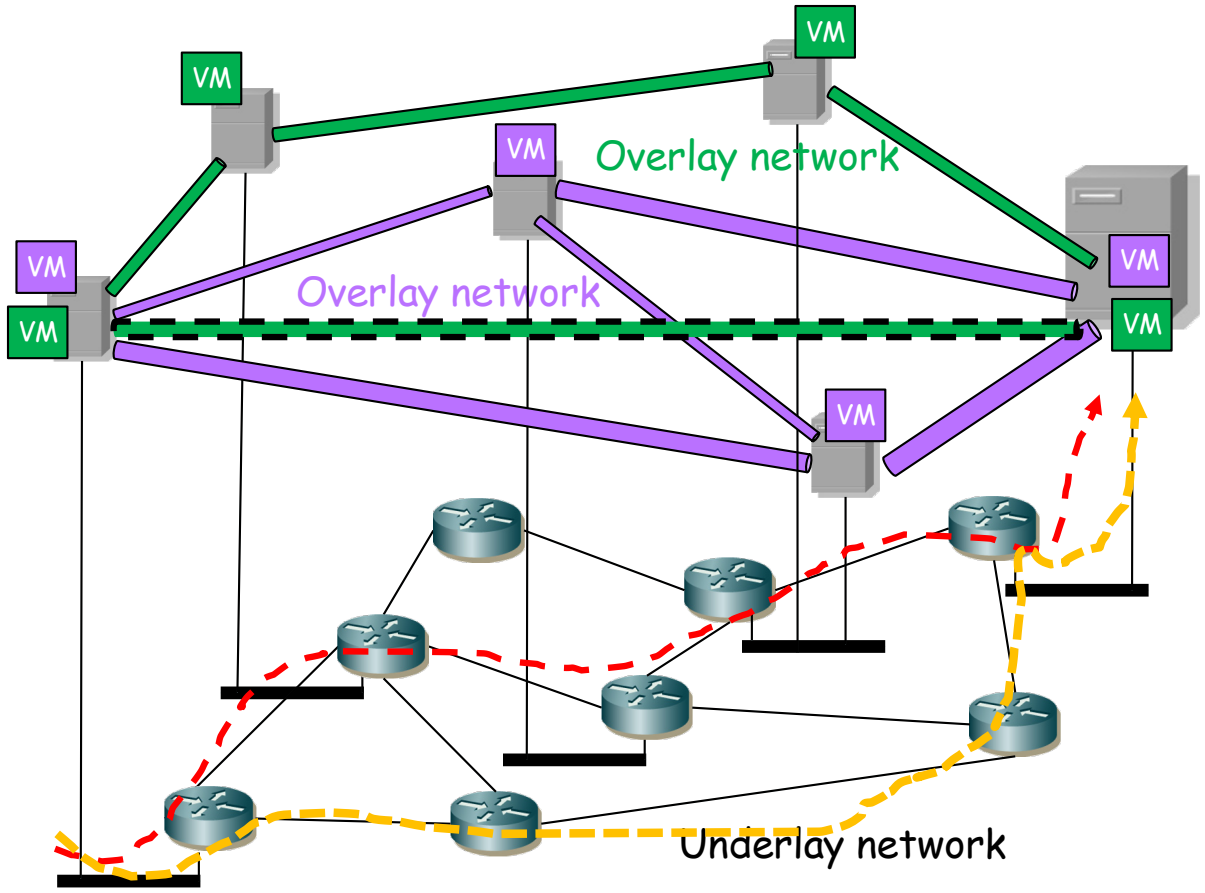
Uso en overlays

- 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



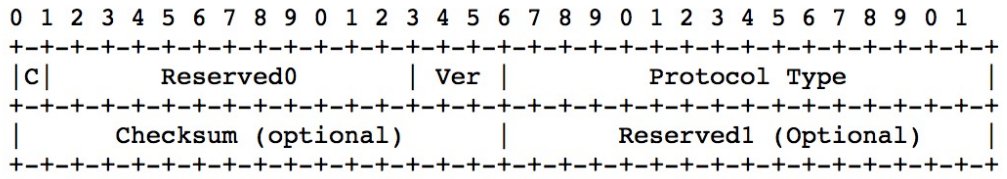
upna
Universidad Pública de Navarra
Nafarroako Unibertsitate Publikoa

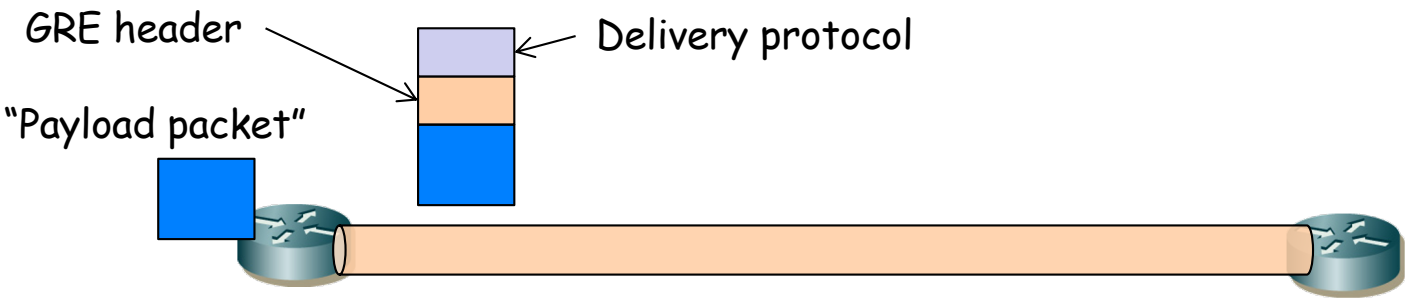
Redes de Nueva Generación
Área de Ingeniería Telemática

GRE

GRE

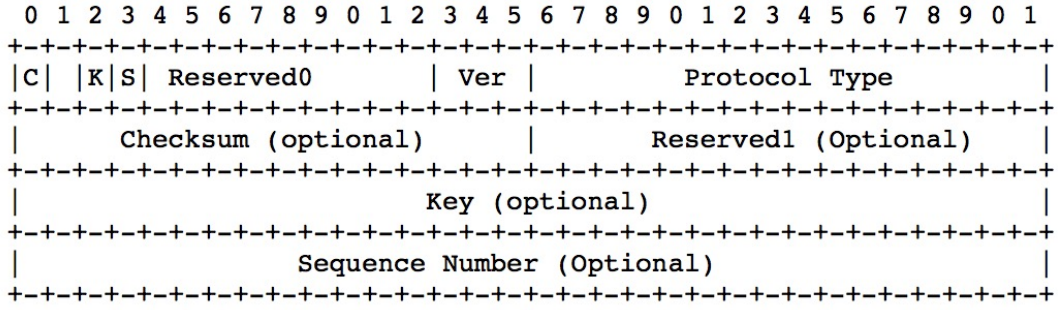
- RFC 2784 “Generic Routing Encapsulation (GRE)” (Procket Networks, Enron Communications, Cisco Systems, Juniper Networks, 2000)
- PPTP (Point-to-Point tunneling Protocol) usa algo similar a GRE
- La cabecera básica GRE ocupa 8 bytes
- Uno de los campos es un Ethertype (*Protocol Type*)
- La versión 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) (…)

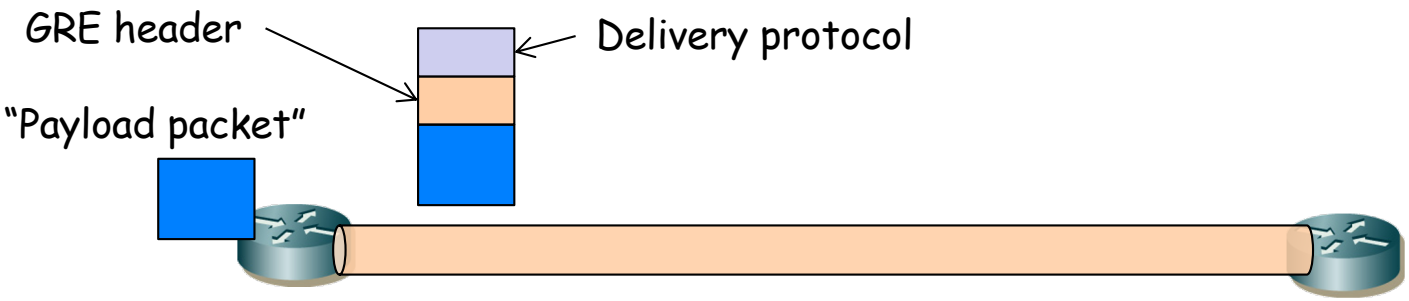




GRE

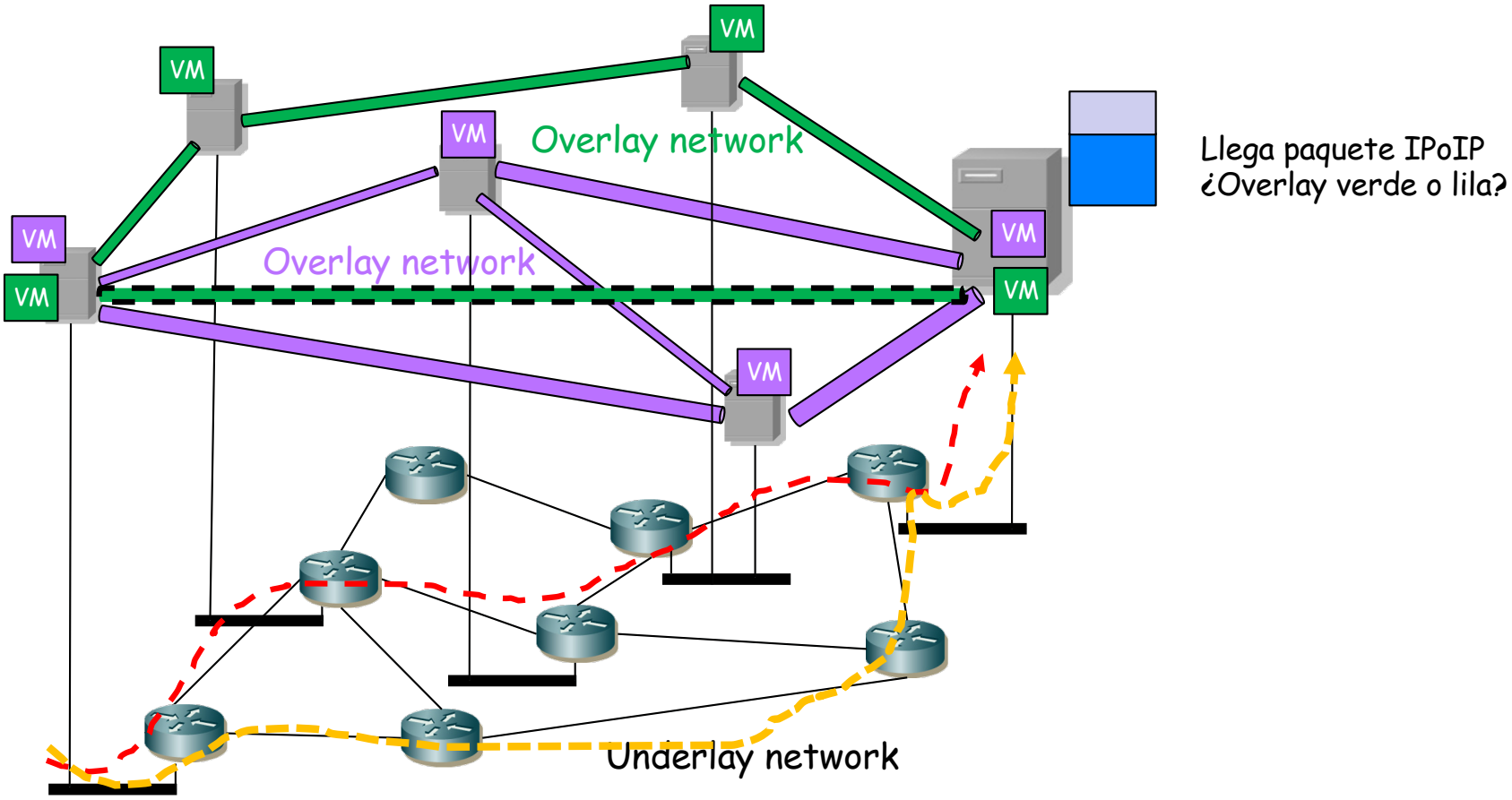
- RFC 2890 “Key and Sequence Number Extensions to GRE”
- “Key” sirve para distinguir flujos dentro del túnel
- “Sequence Number”
  - Si hay “key” entonces el número de secuencia es por “key”
  - Permite dar entrega en orden (aunque no fiable)
  - Si llega uno “anterior” lo descarta
  - Si llega uno que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





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)



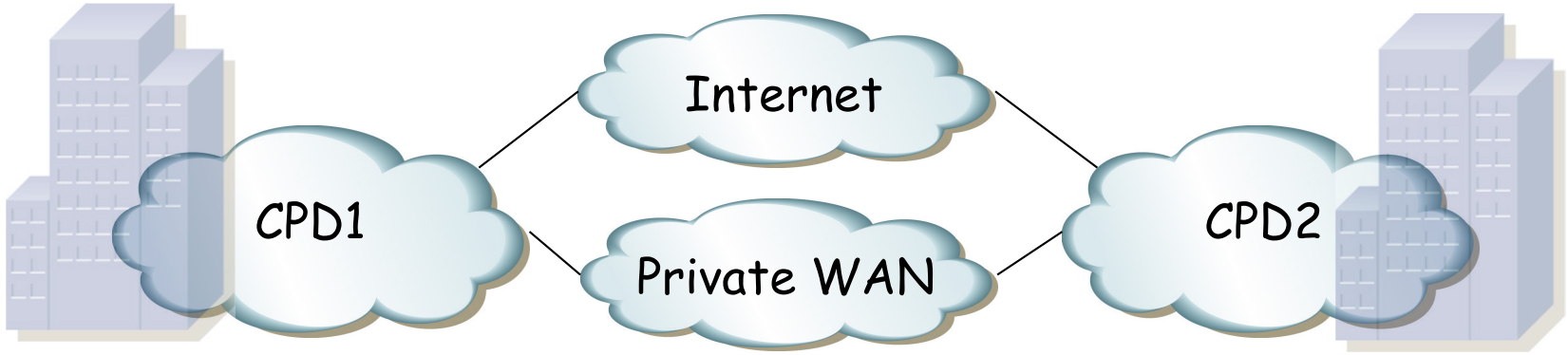
upna
Universidad Pública de Navarra
Nafarroako Unibertsitate Publikoa

Redes de Nueva Generación
Área de Ingeniería Telemática

VXLAN

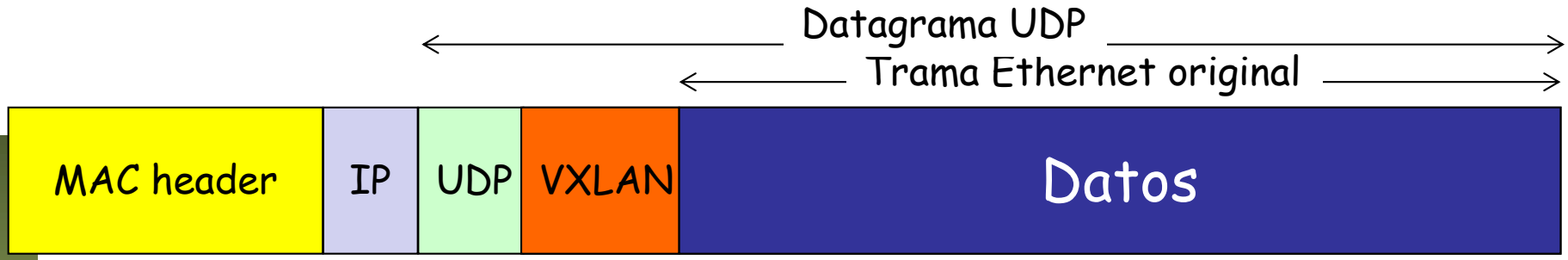
VXLAN

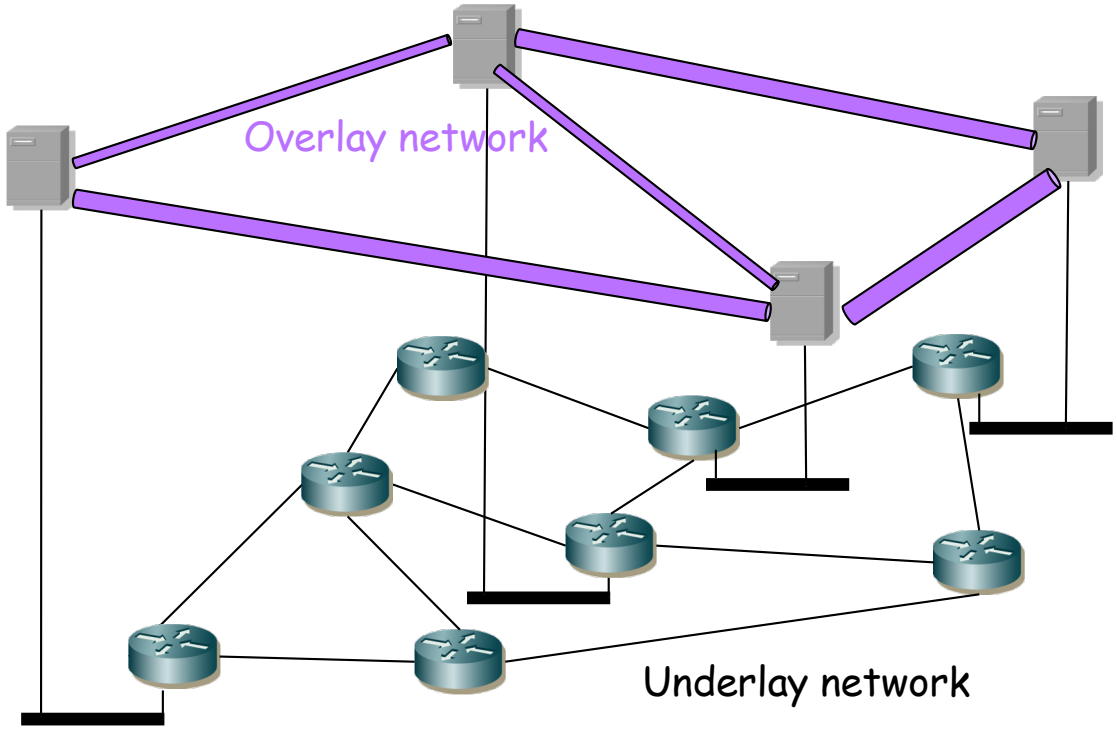
- RFC 7348 “Virtual eXtensible Local Area Network (VXLAN): A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks” (agosto 2014)
- RFC Informativa firmada por Cisco, VMware, Intel, Red Hat, Arista y Cumulus Networks
- Diseñado para un entorno de host virtualizado
- Emplea un esquema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



VXLAN

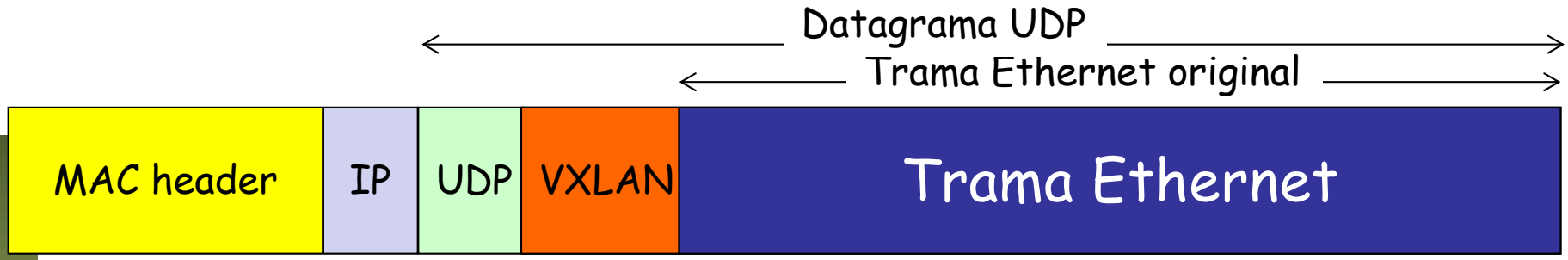
- Túnel sobre capa 4 pues hace el transporte sobre UDP
- Cabecera VXLAN con identificador de la Virtual Network (VNI)

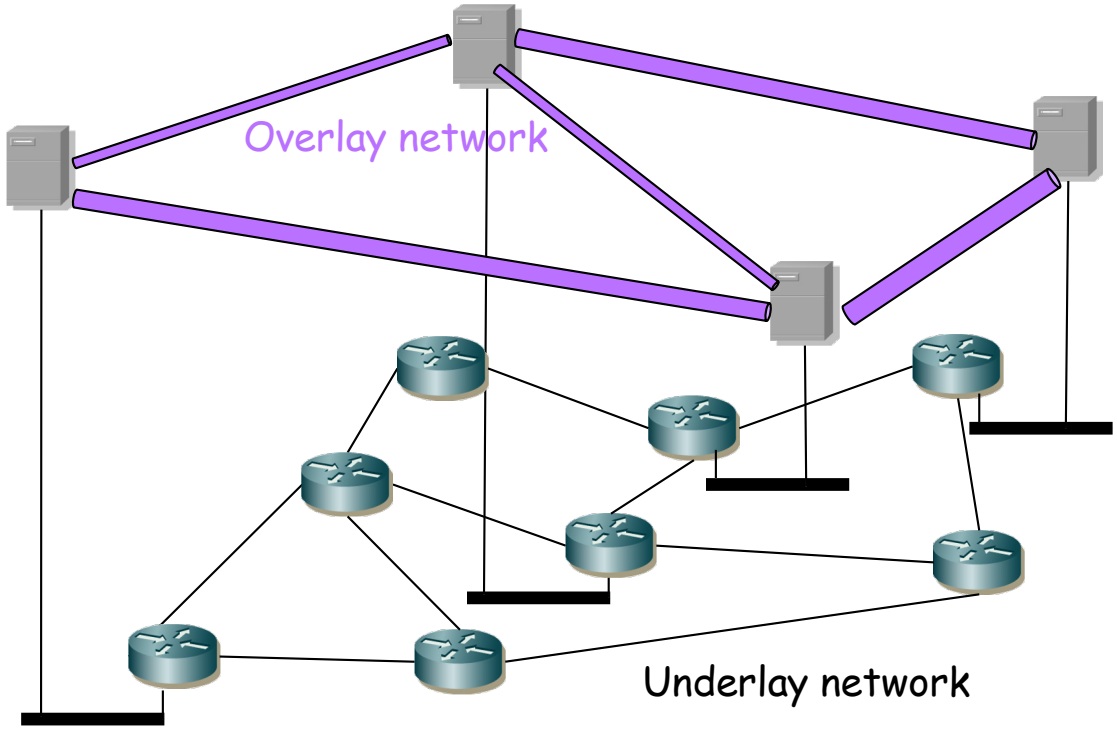




VXLAN

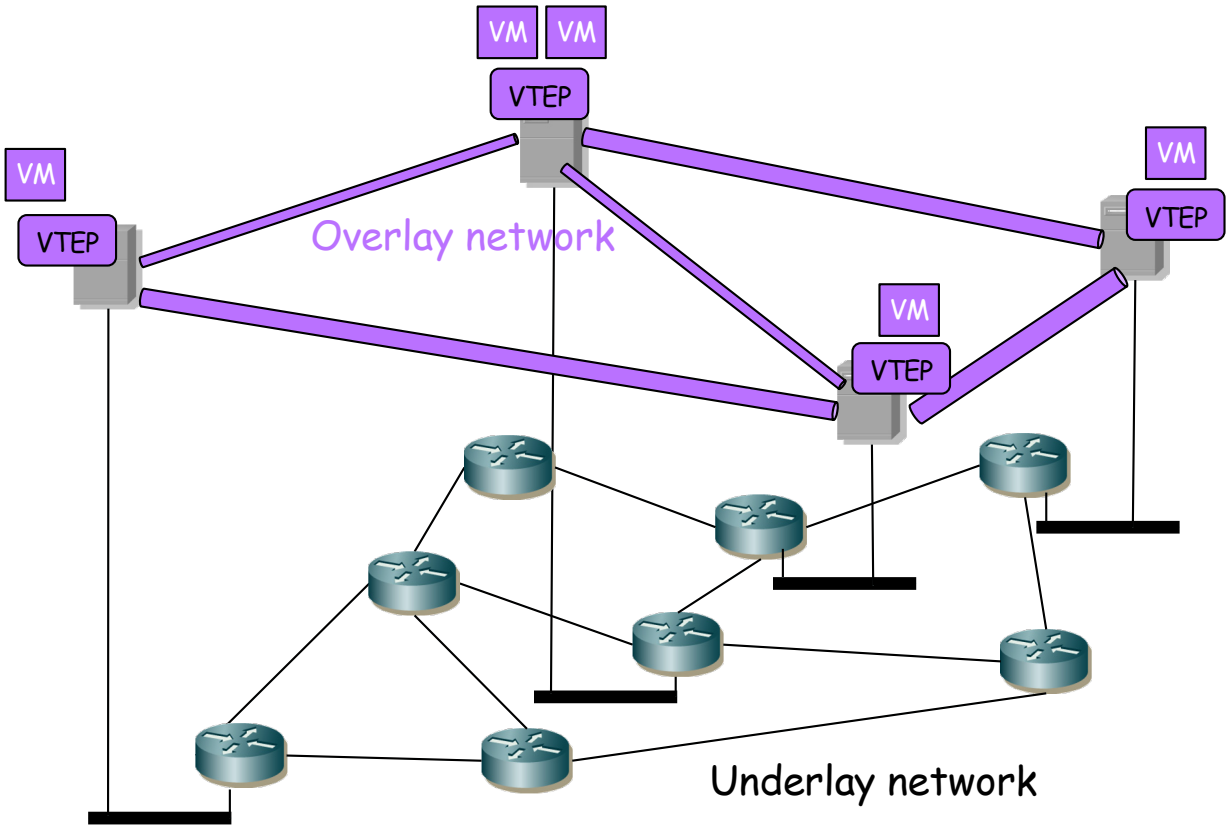
- Contenido: trama Ethernet entregada por la VM





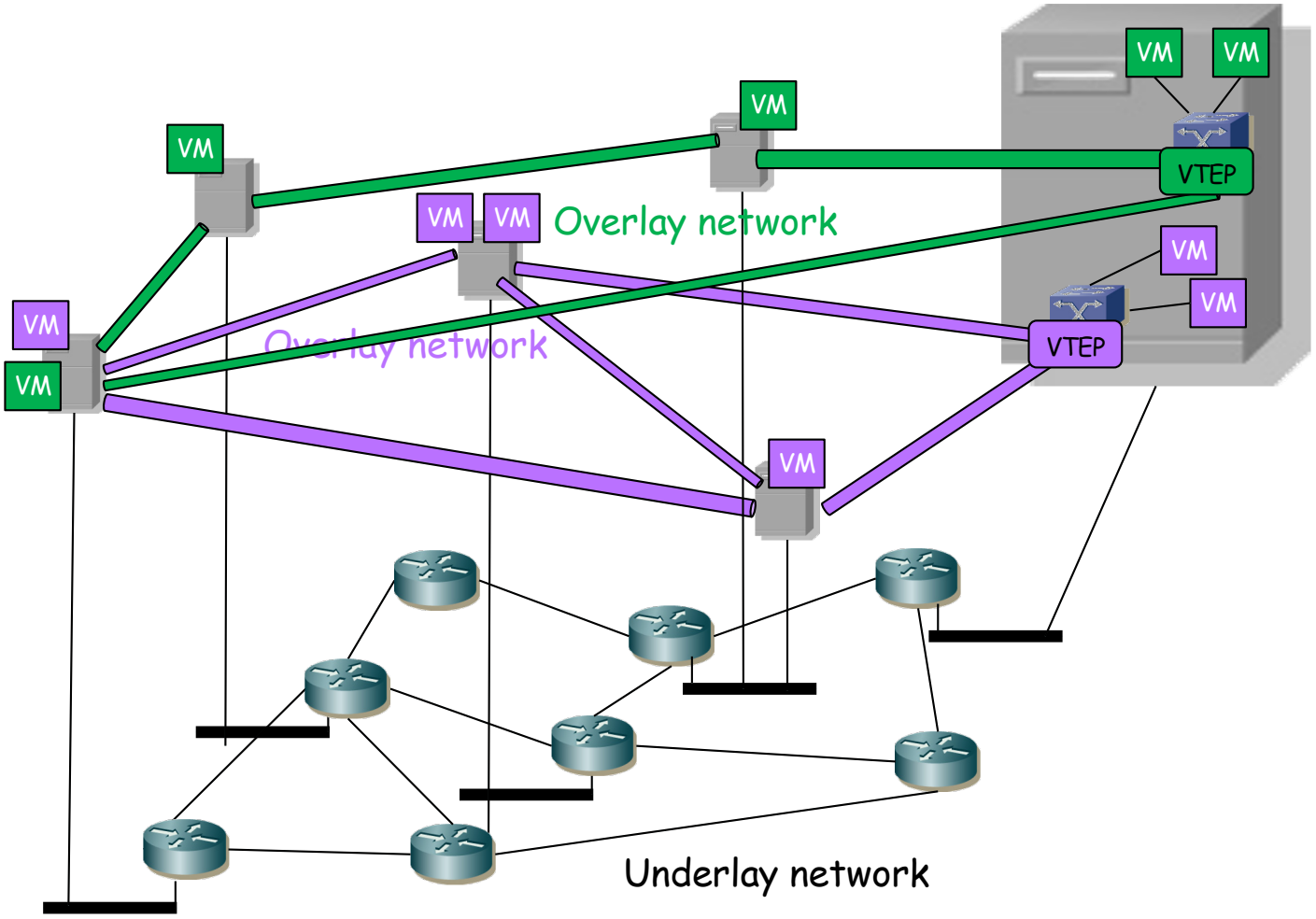
VXLAN

- 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

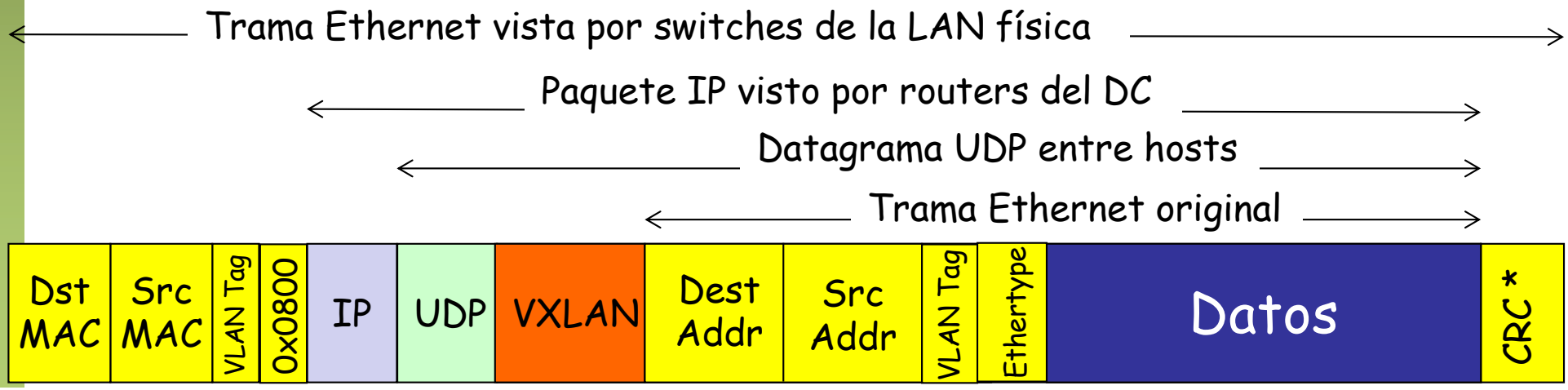


upna
Universidad Pública de Navarra
Nafarroako Unibertsitate Publikoa
Redes de Nueva Generación
Área de Ingeniería Telemática

VXLAN Data Plane

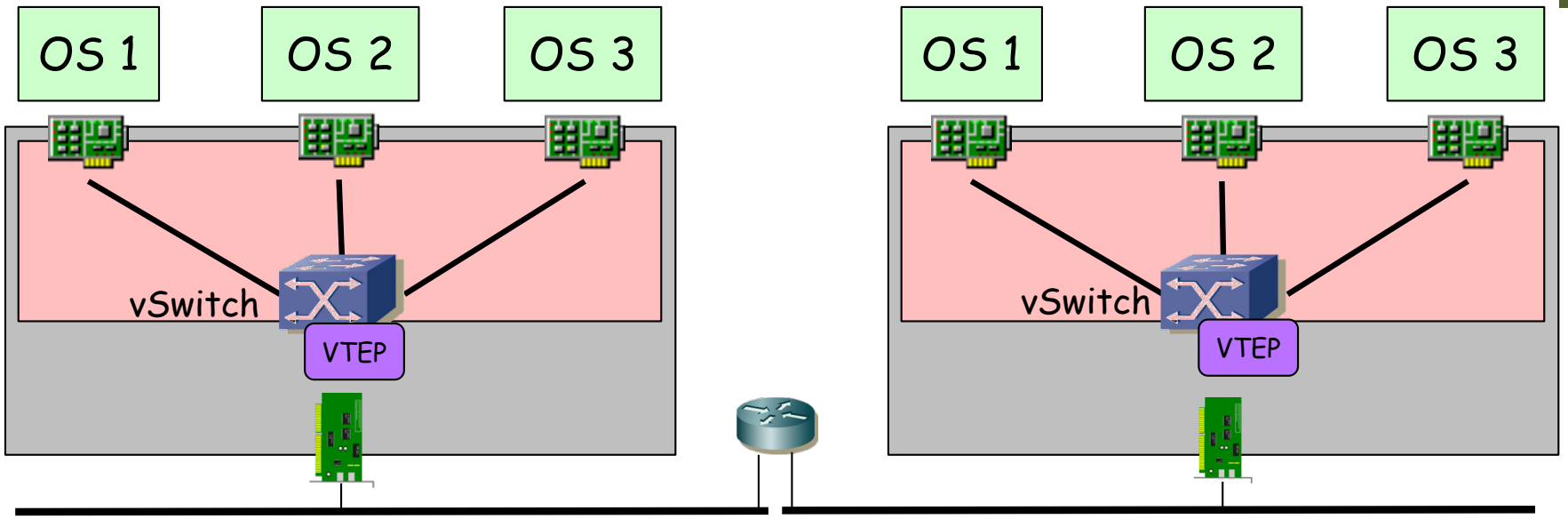
VXLAN

- Puerto destino 4789, puerto origen se recomienda un hash de campos de la trama original para facilitar el balanceo de flujos en la red IP
- La cabecera VXLAN es de 8 bytes y fundamentalmente contiene el VNI
- VNI = *VXLAN Network Identifier* (de 24 bits)
- En un entorno de DC con múltiples usuarios permite separar más de los 4094 que permitiría una etiqueta de VLAN
- Los VLAN Tags (trama externa e interna) son opcionales
- Trama interna sin CRC
- Para las máquinas virtuales es transparente



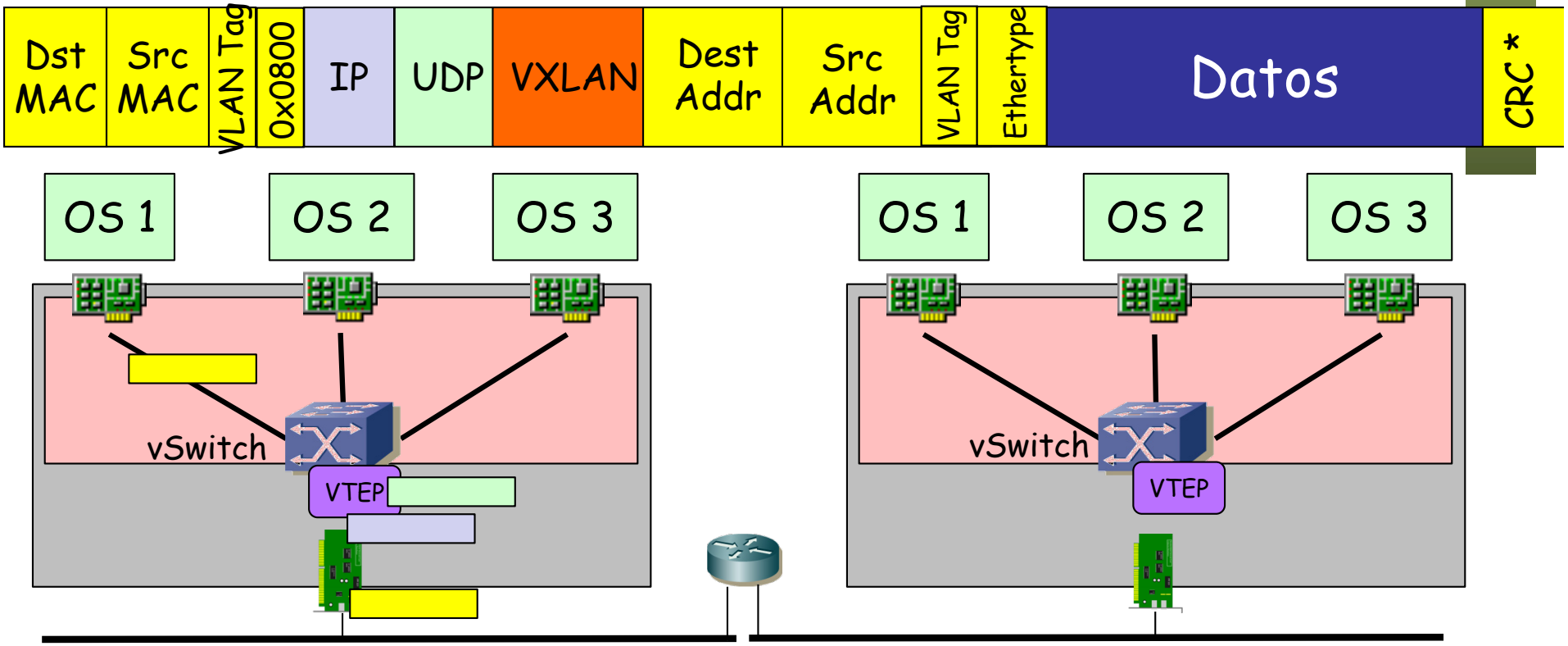
VXLAN: *Data plane*

- Cada overlay se conoce como un “segmento VXLAN”
- Los hosts (VMs) de un segmento VXLAN solo pueden comunicarse entre ellos
- Se pueden repetir las direcciones MAC en distintos segmentos
- El VTEP se suele encontrar en el hypervisor (transparente para la VM)
- Podría estar en un ToR switch



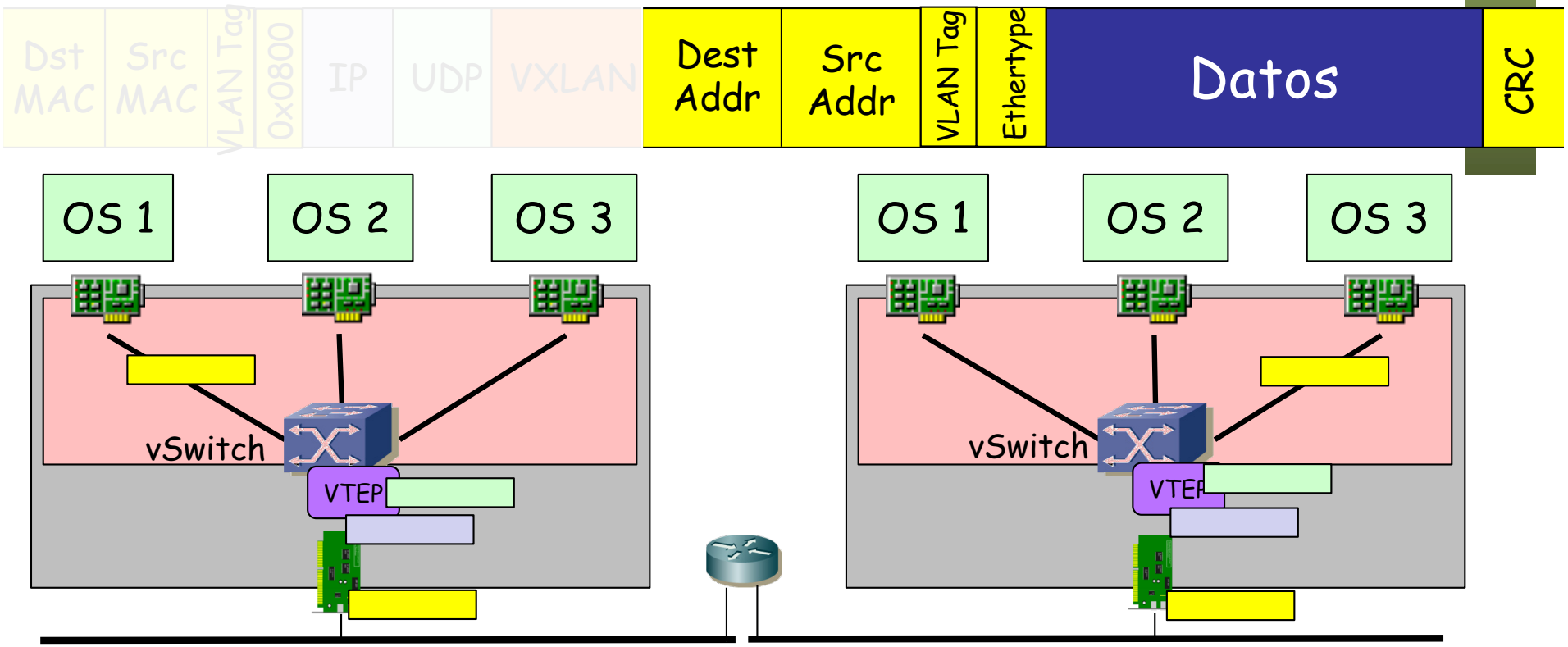
VXLAN: *Data plane*

- La trama Ethernet que envía una VM la recibe el vSwitch
- La encapsula con el VNI (configuración de la VM) en un datagrama UDP
- Averigua la dirección IP del host que contiene la VM con esa MAC destino
- Le envía el paquete IP que contiene la trama
- Por supuesto en una trama Ethernet



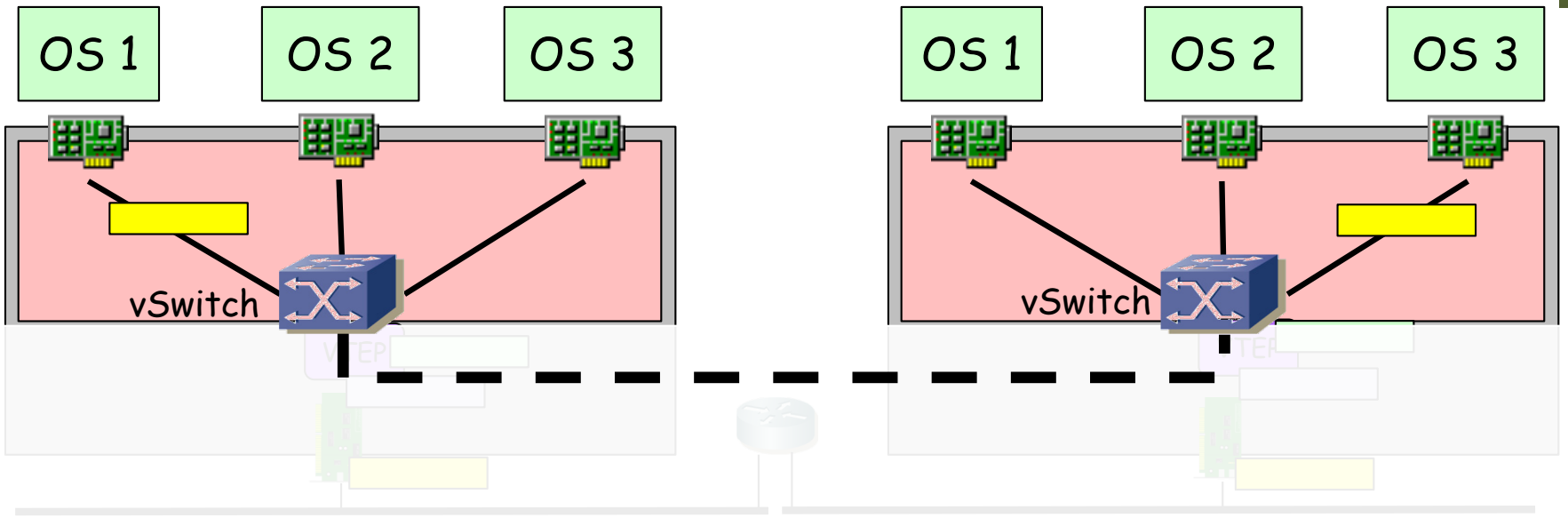
VXLAN: *Data plane*

- Si hay LAGs o ECMP los switches que repartan flujos en función de capa 3+ pueden repartir estos flujos
- En el receptor el proceso es el inverso
- La VM destino nunca ve el paquete VXLAN
- Recibe directamente la trama que envió la VM origen



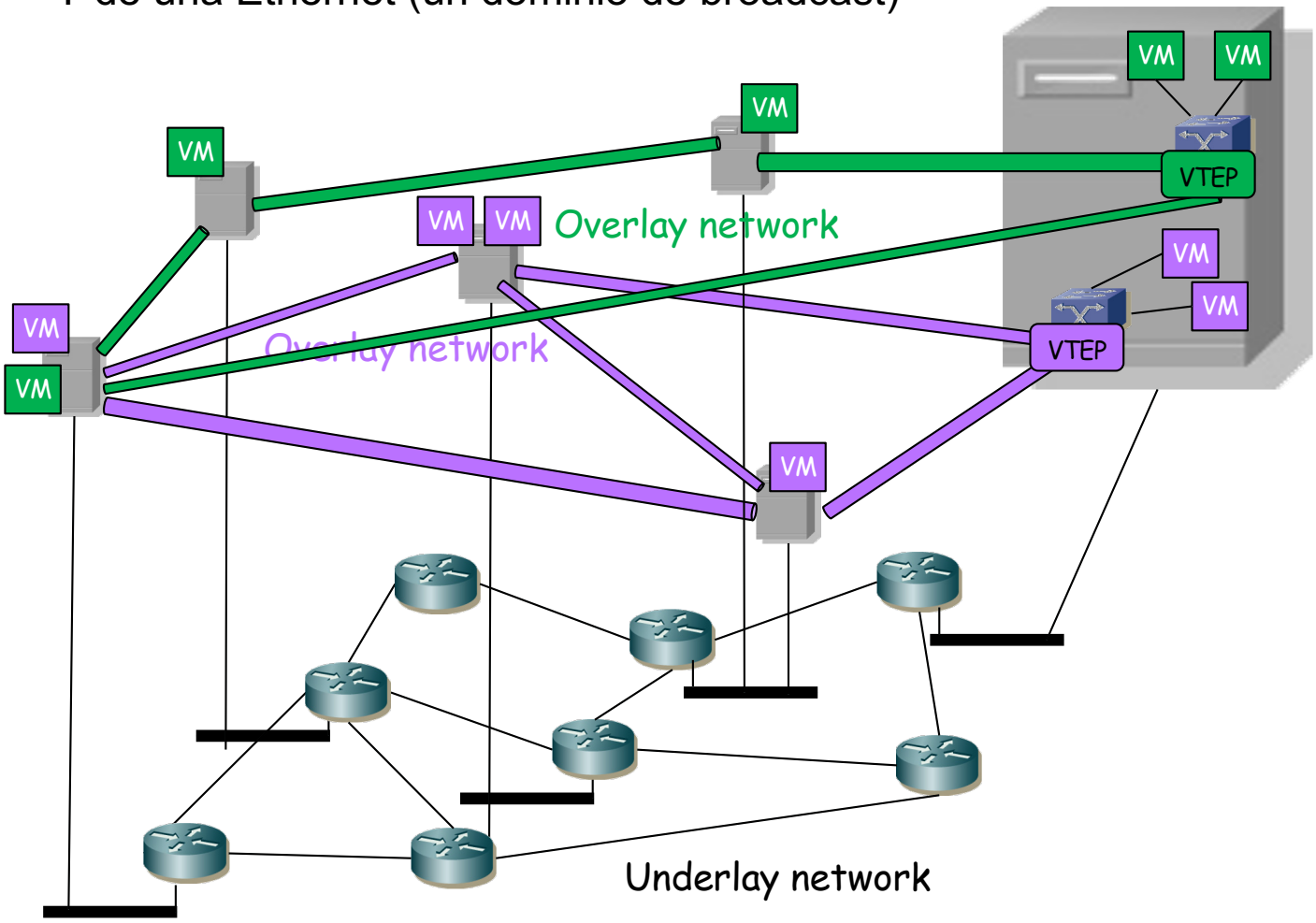
VXLAN: *Data plane*

- El transporte entre las VMs es de las tramas Ethernet
- Se comportan como si estuvieran en la misma VLAN
- ¿O en varias VLANs? A fin de cuentas transporta el V-Tag
- La RFC no lo deja claro y parece más inclinada a retirar esa etiqueta (sección 6.1)



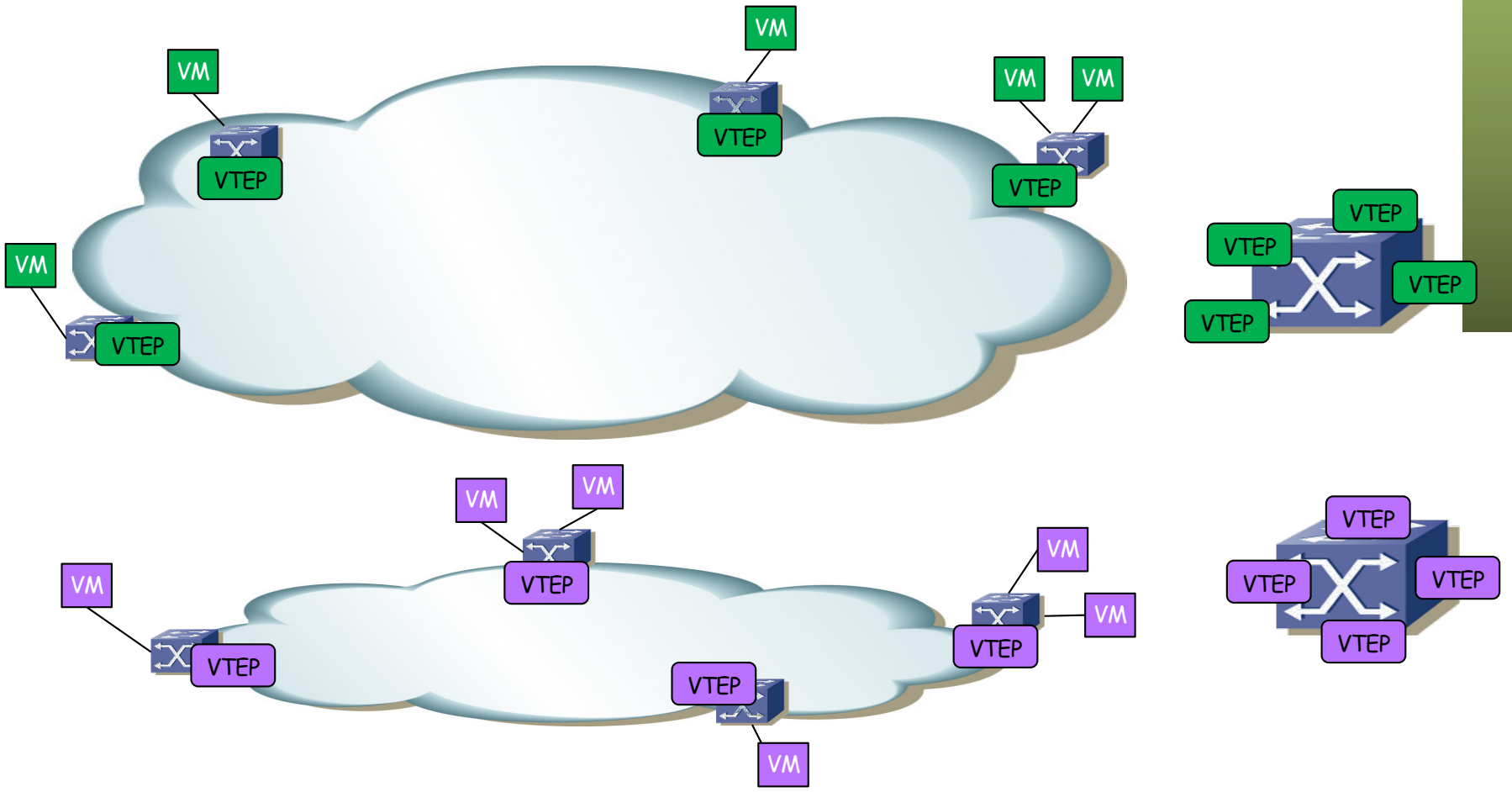
Uso en overlays

- La underlay ve paquetes IP entre los hosts
- Ve que transportan paquetes UDP si lo necesita para el ECMP
- El resto es transparente
- Tenemos las ventajas de una red IP (routing, multipath, no STP)
- Y de una Ethernet (un dominio de broadcast)



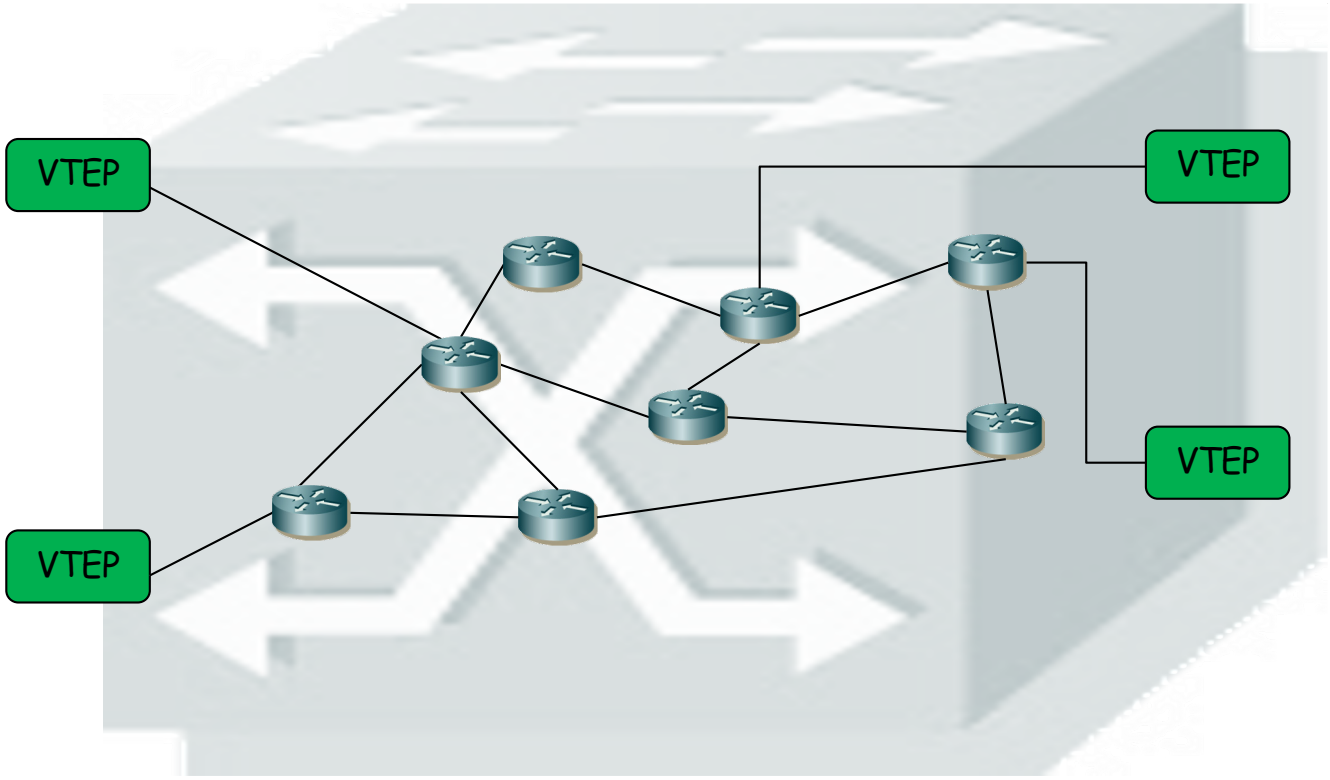
Abstracción de la underlay

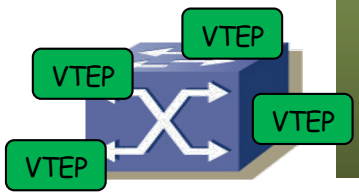
- La underlay da un servicio de interconexión a cada overlay
- Se comporta como un switch cuyo backplane se base en IP para mover las tramas entre los interfaces, que son los VTEPs



Abstracción de la underlay

- La underlay da un servicio de interconexión a cada overlay
- Se comporta como un switch cuyo backplane se base en IP para mover las tramas entre los interfaces, que son los VTEPs



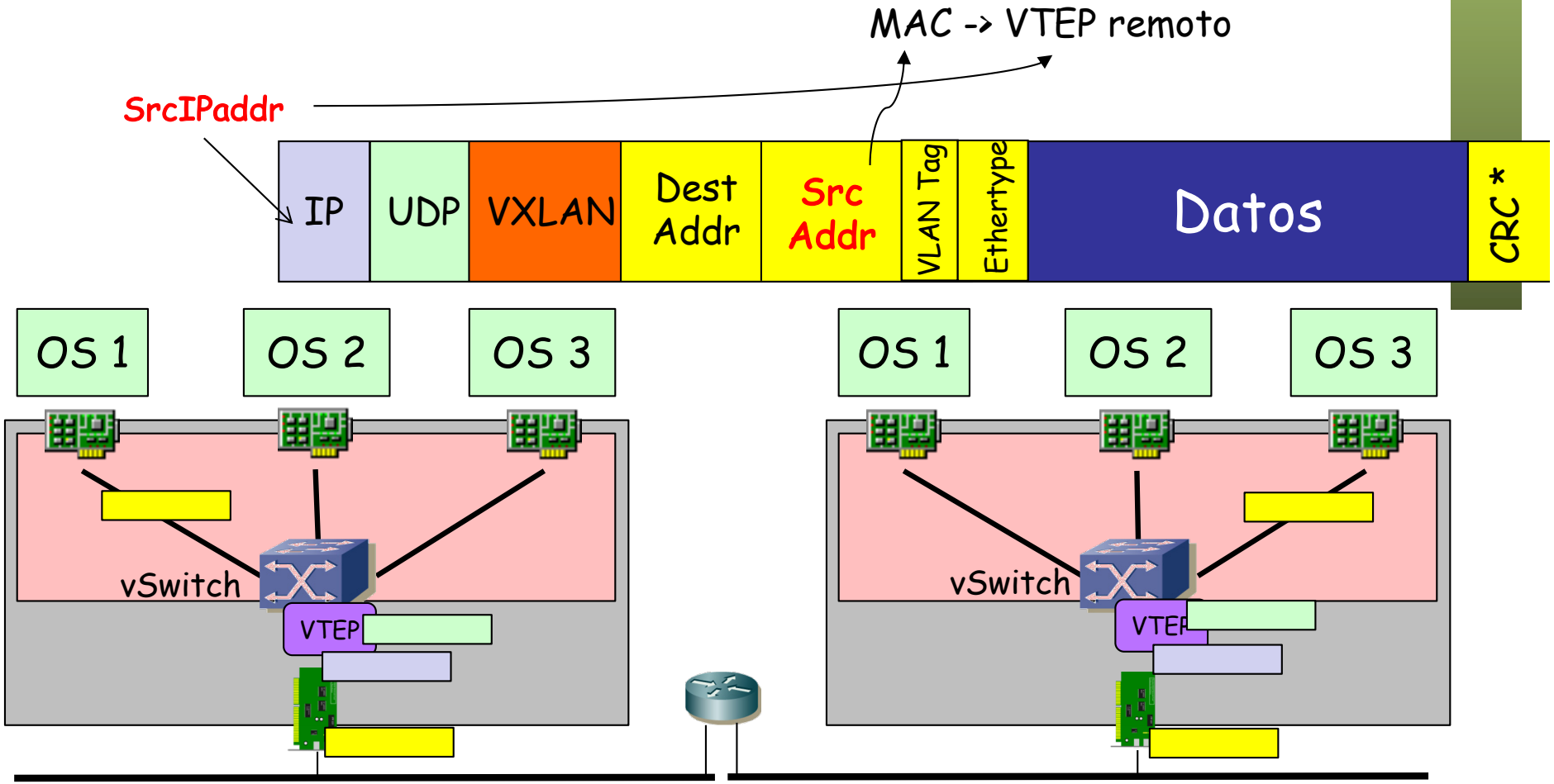


upna
Universidad Pública de Navarra
Nafarroako Unibertsitate Publikoa
Redes de Nueva Generación
Área de Ingeniería Telemática

VXLAN Control Plane

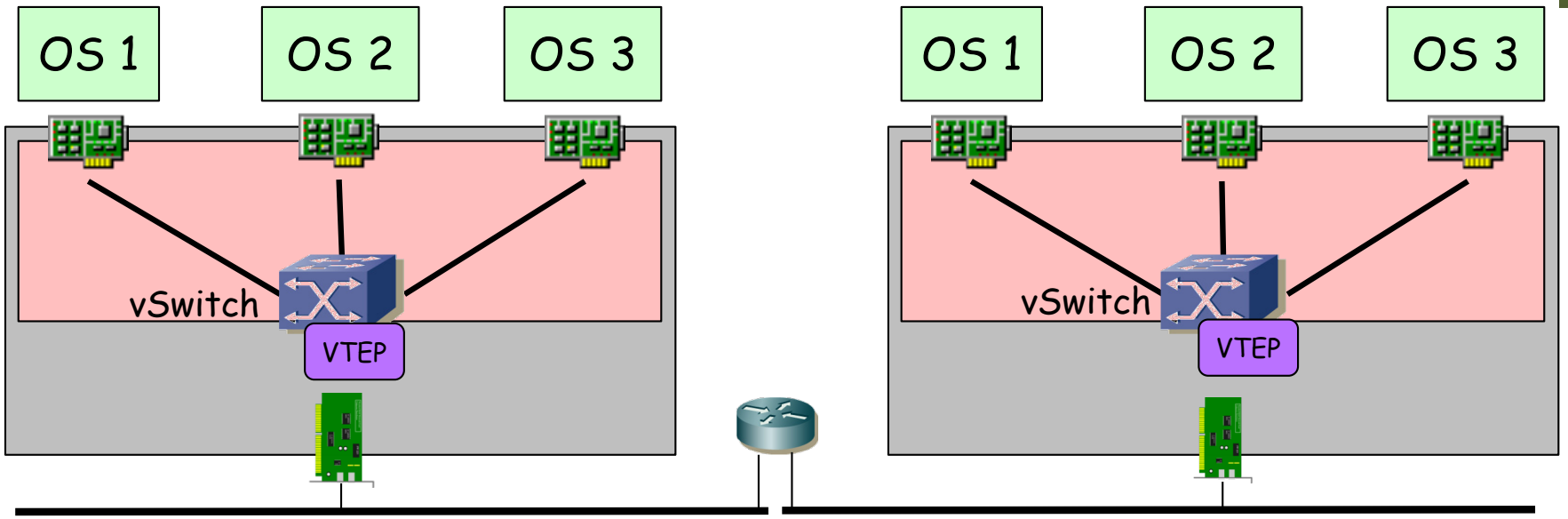
VXLAN: *Control plane*

- Los vSwitch deben aprender la dirección IP del host de la VM
- Aprende con información del plano de datos: al recibir un paquetes de datos
- Aprende que la dirección MAC origen en la trama contenido es de un host en el VTEP con dirección IP la origen en el continente
- Cuando tenga una trama para esa MAC sabe a qué VTEP enviarla



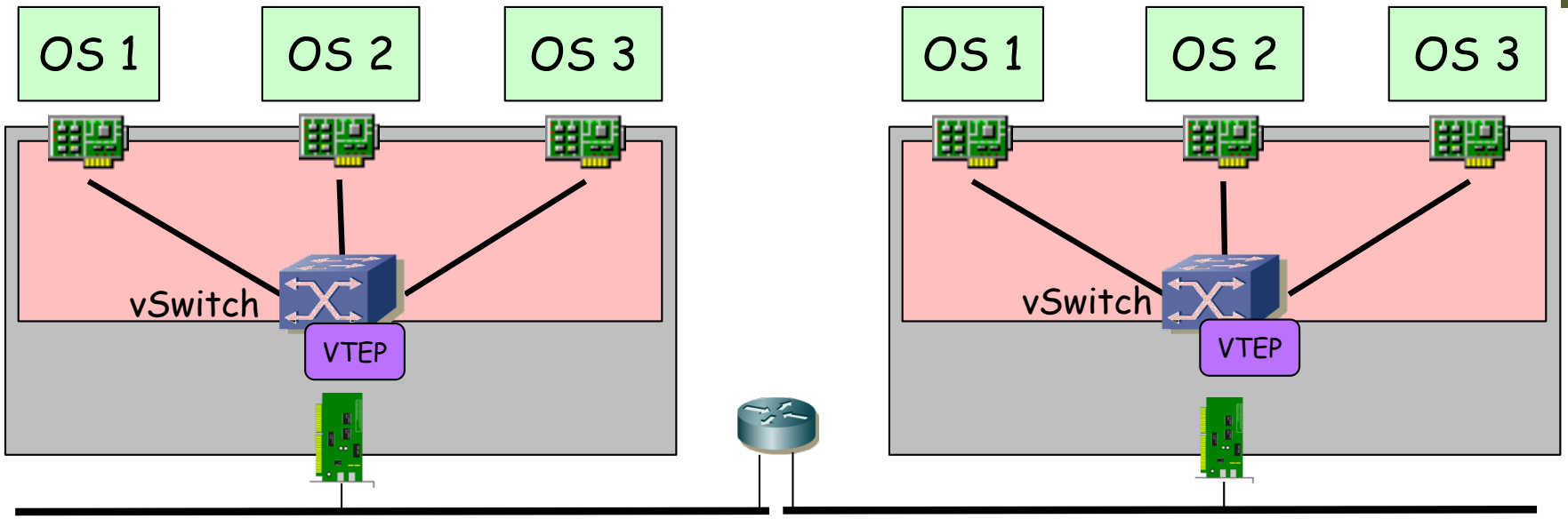
VXLAN: *Control plane*

- ¿Y el BUM?
  - Broadcast
  - Unknown unicast
  - Multicast



VXLAN: *Control plane*

- ¿Y el BUM? Por ejemplo los ARP
- Se envía a un grupo multicast IP (uno por segmento VXLAN)
- Todos los hosts del segmento VXLAN pertenecen a ese grupo
- Esto implica routing multicast en la red IP (algo como PIM-SM)
- El número de grupos multicast soportados por la red puede ser limitado, lo cual llevaría a compartirlos para varios segmentos VXLAN
- Hay soluciones unicast e híbridas, propietarias, mediante algún tipo de controlador o empleando MP-BGP

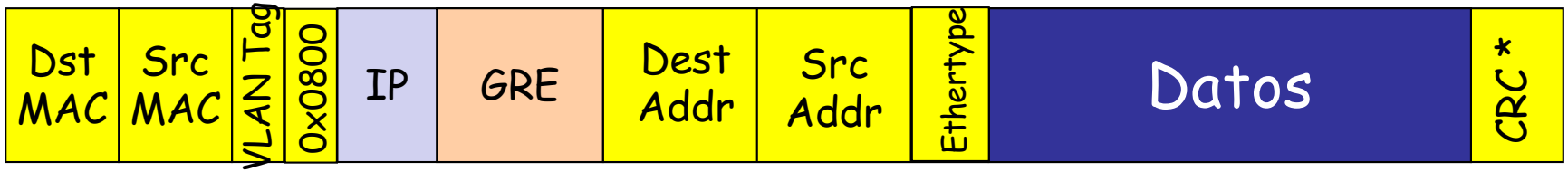


upna
Universidad Pública de Navarra
Nafarroako Unibertsitate Publikoa
Redes de Nueva Generación
Área de Ingeniería Telemática

NVGRE

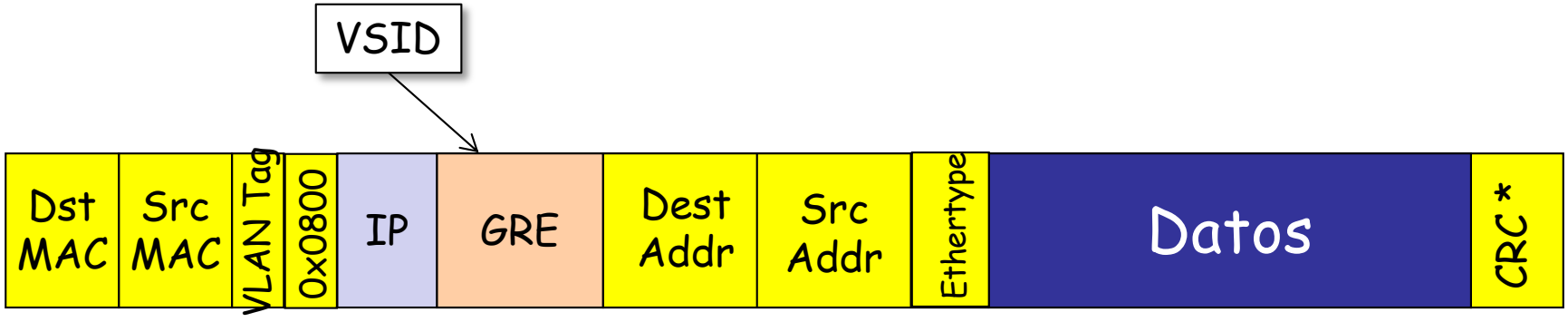
NVGRE

- RFC 7637 "*NVGRE: Network Virtualization using Generic Routing Encapsulation*"
- RFC Informativa (Sept.2015) firmada por Microsoft
- Crea una topología capa 2 virtual sobre una red capa 3
- La trama (sin V-TAG) es encapsulada en el extremo (host, switch virtual, etc) en un paquete GRE y en un paquete IP (protocolo 0x2F)



NVGRE

- El extremo se llama el NVGRE Endpoint
- La cabecera GRE contiene un Virtual Subnet ID (VSID)
  - De 24 bits (parte del campo *key* de GRE)
  - Los 8 bits restantes de la clave se usan para distinguir flujos y poder hacer reparto de carga en routers que entiendan GRE
  - Permite identificar un dominio broadcast capa 2 en un entorno multi-tenant



NVGRE

- La RFC no detalla cómo el Endpoint conoce la dirección del destino al que mandar el paquete IP
- Broadcast y multicast
  - Se puede emplean encaminamiento multicast IP con una o más direcciones multicast por VSID
  - Se puede implementar con N-way unicast
- Lo soporta Hyper-V (draft propuesto por Microsoft)
- NICs pueden soportar *offloading* del encapsulado NVGRE

Geneve

Geneve

- RFC 8926 (Noviembre 2020): “Geneve: Generic Network Virtualization Encapsulation”, firmada por Intel y Vmware
- Formato estandarizado para el data plane
- Sobre UDP
- Puerto destino 6081, origen un hash de header encapsulado
- Protocol Type es el Ethertype del contenido (0x6558 = Ethernet)
- VNI de 24 bits
- 0 o más opciones en formato TLV
- Deja indeterminado el plano de control

https://datatracker.ietf.org/wg/nvo3/documents/