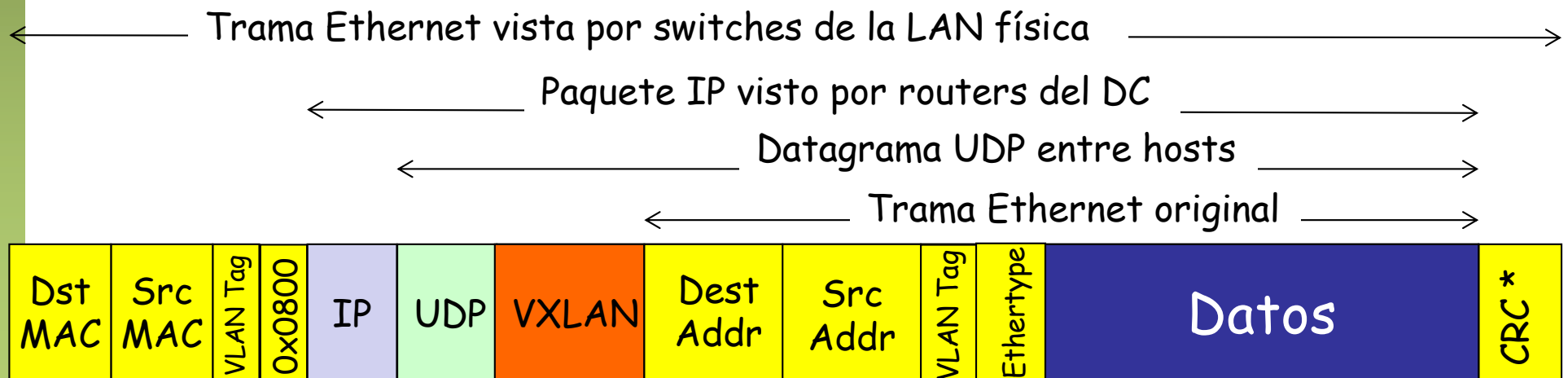


VXLAN Data Plane

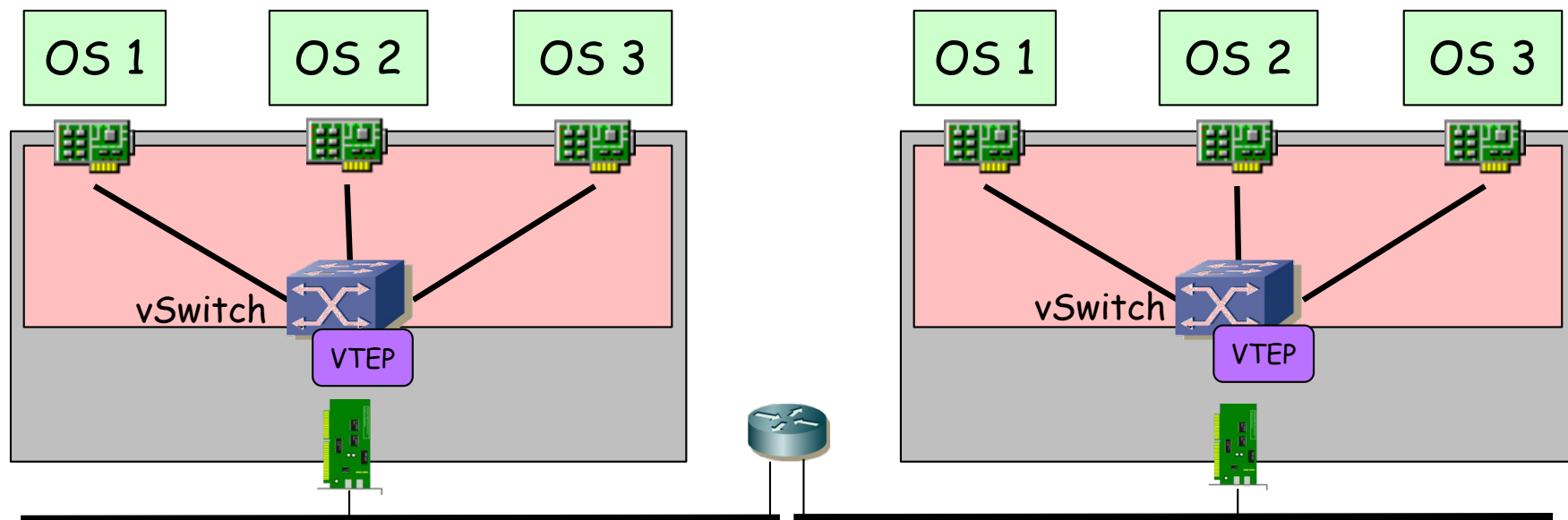
VXLAN

- 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



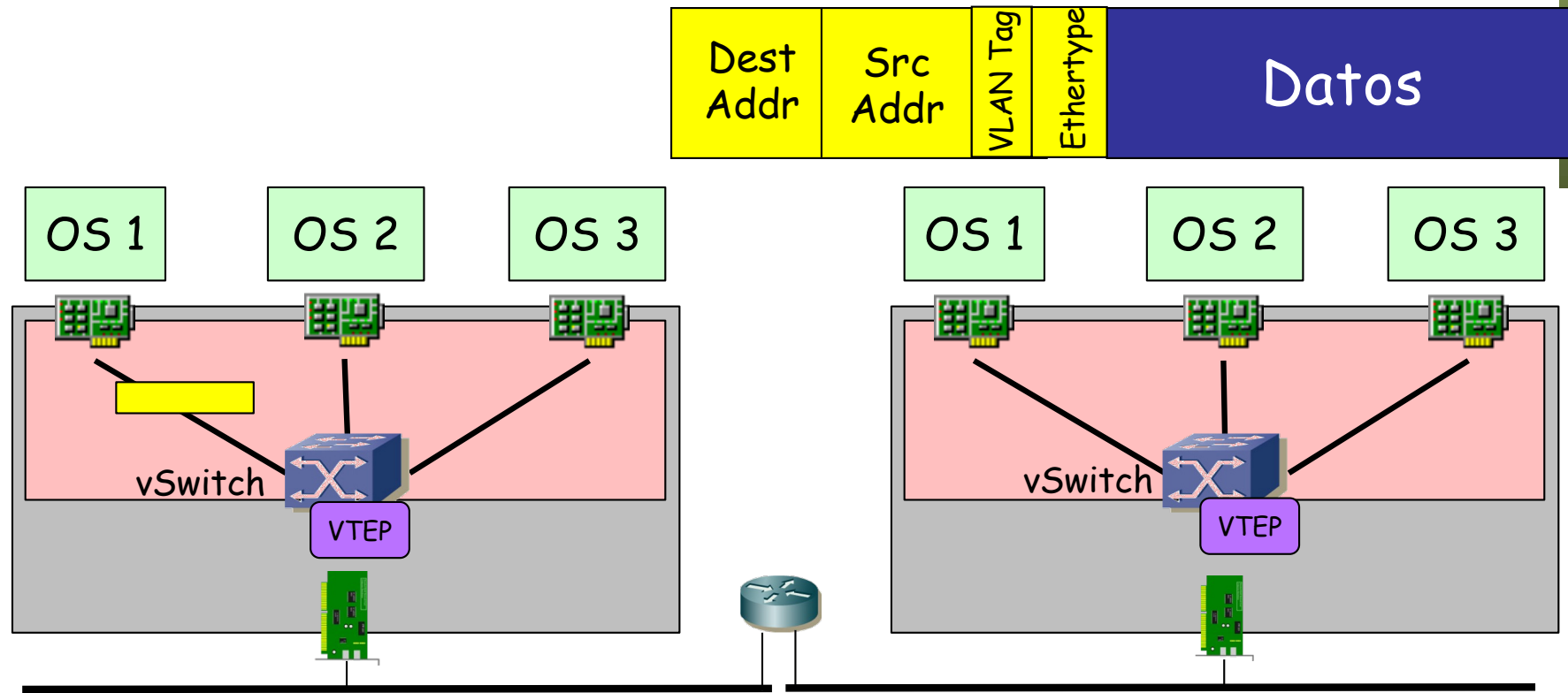
VXLAN: *Data plane*

- 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



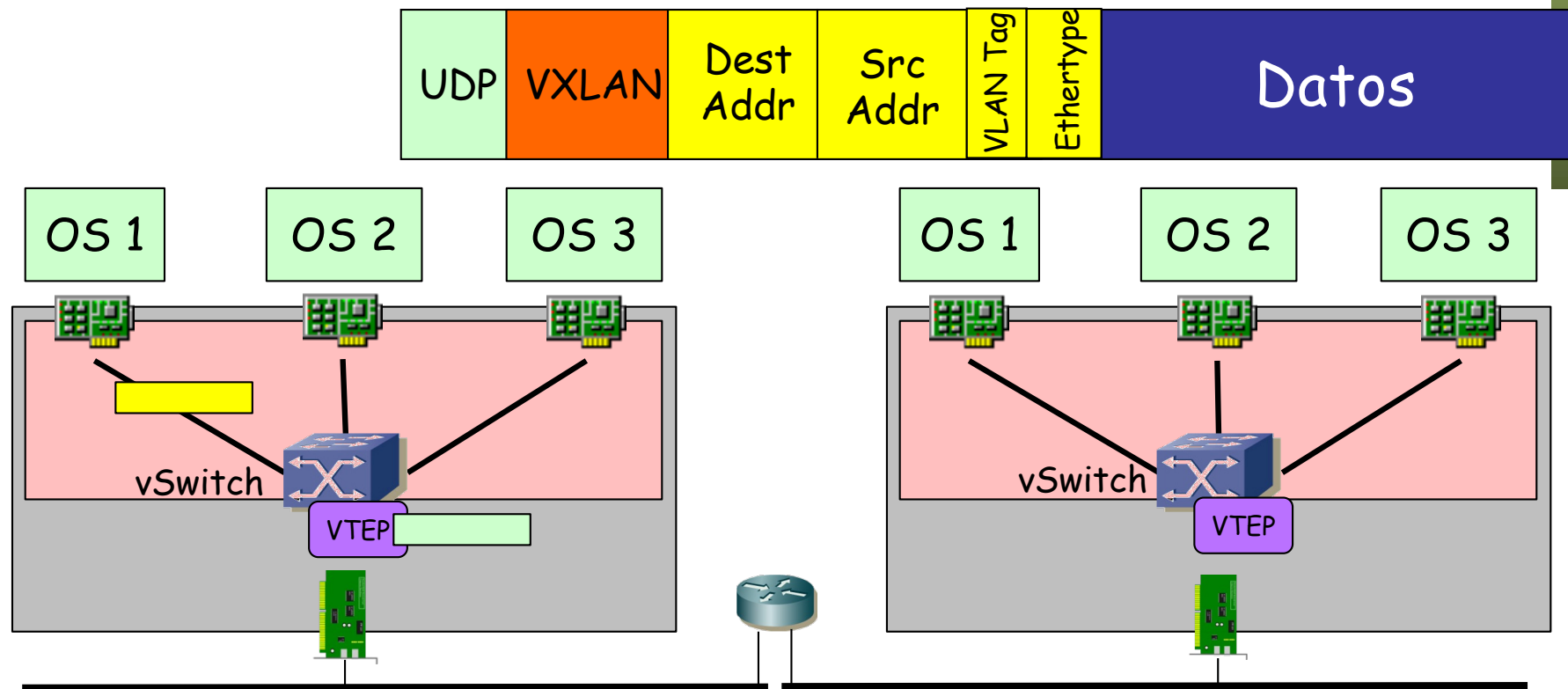
VXLAN: *Data plane*

- La trama Ethernet que envía una VM la recibe el vSwitch
- (...)



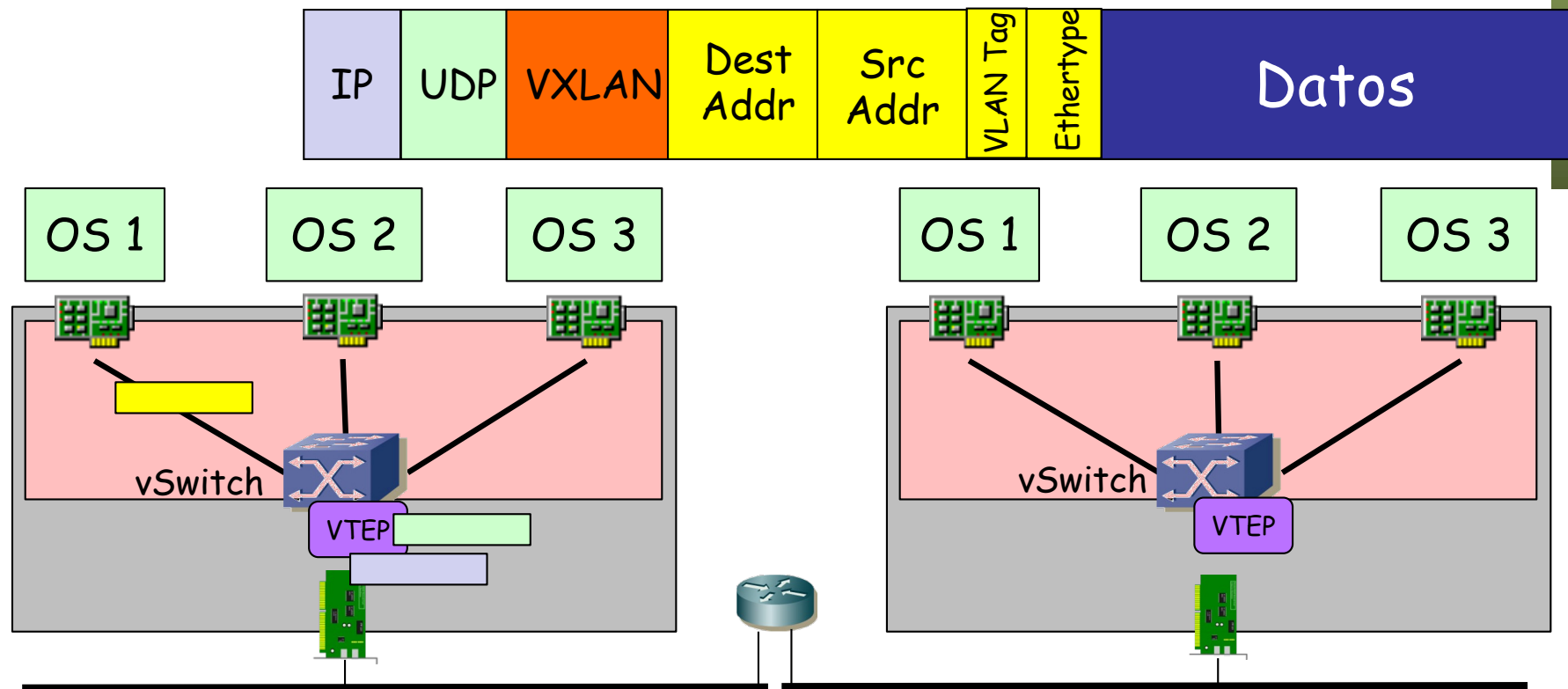
VXLAN: *Data plane*

- La trama Ethernet que envía una VM la recibe el vSwitch
- La encapsula con el VNI (configuración de la VM) en un datagrama UDP
- (...)



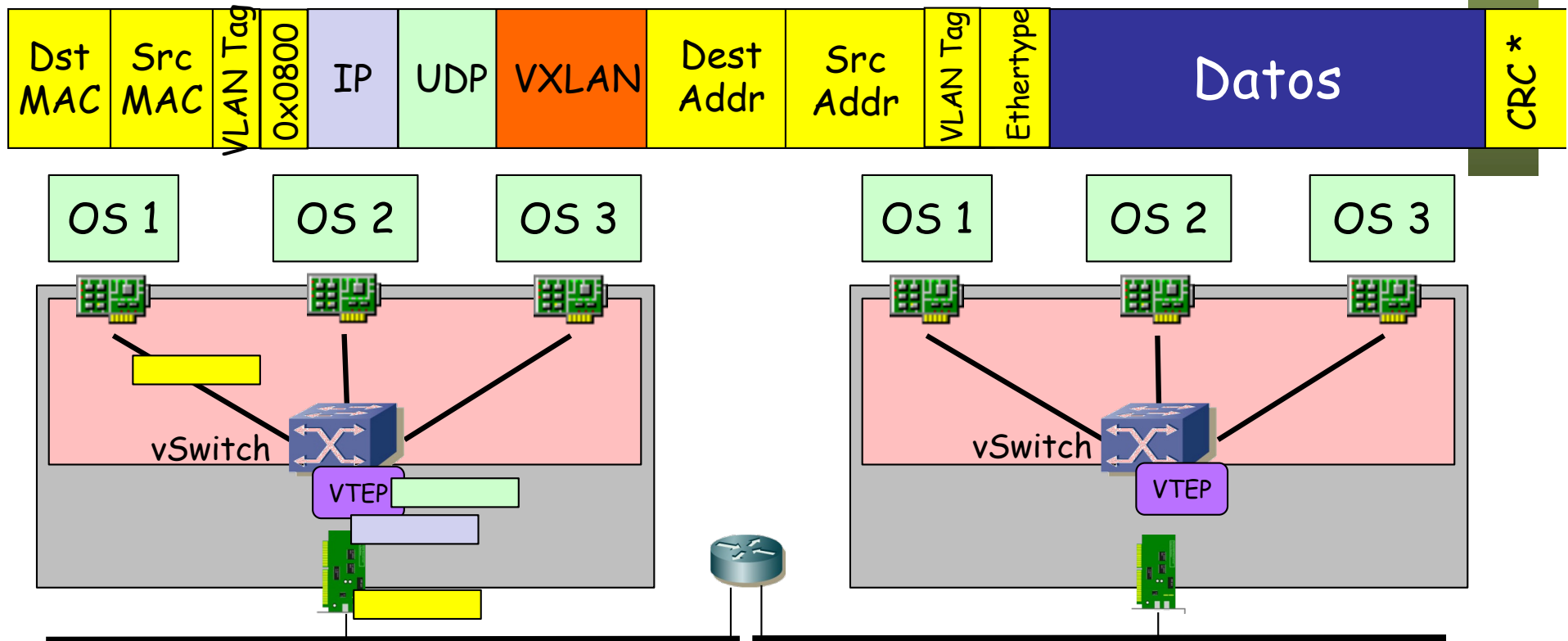
VXLAN: *Data plane*

- 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
- (...)



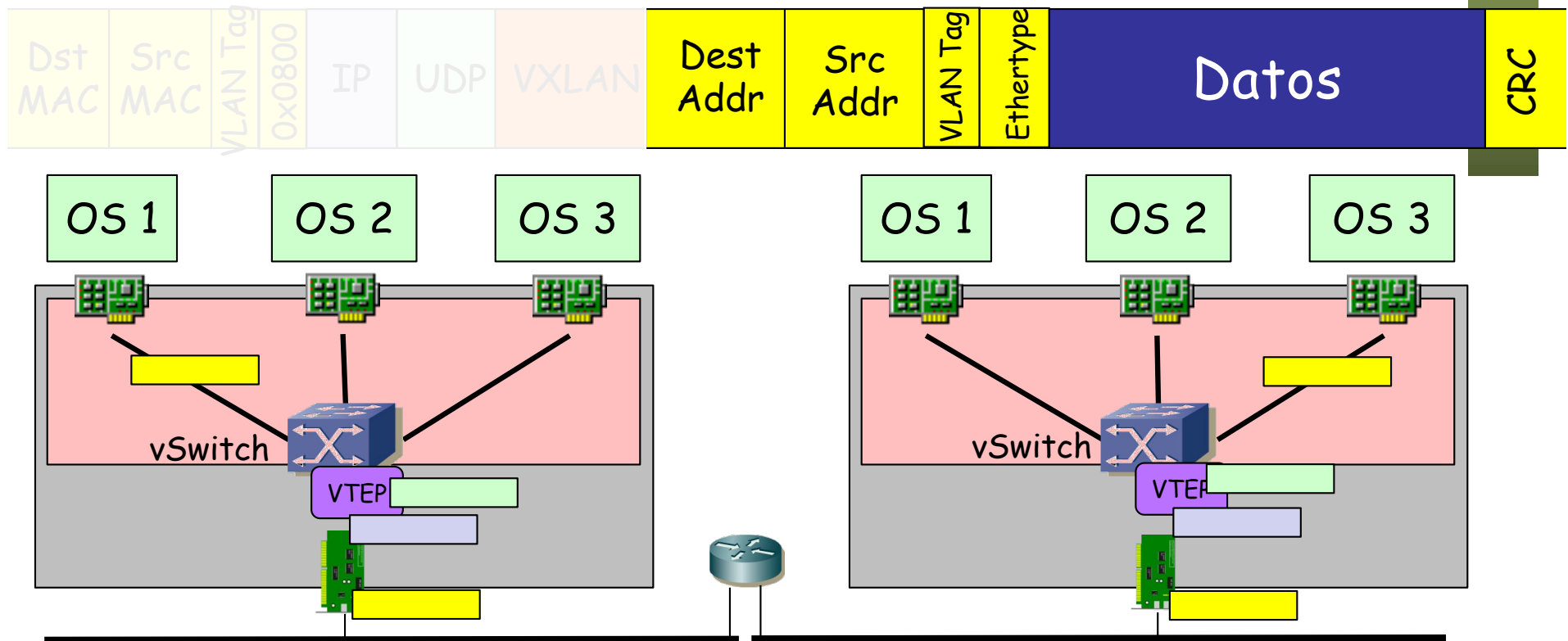
VXLAN: *Data plane*

- 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



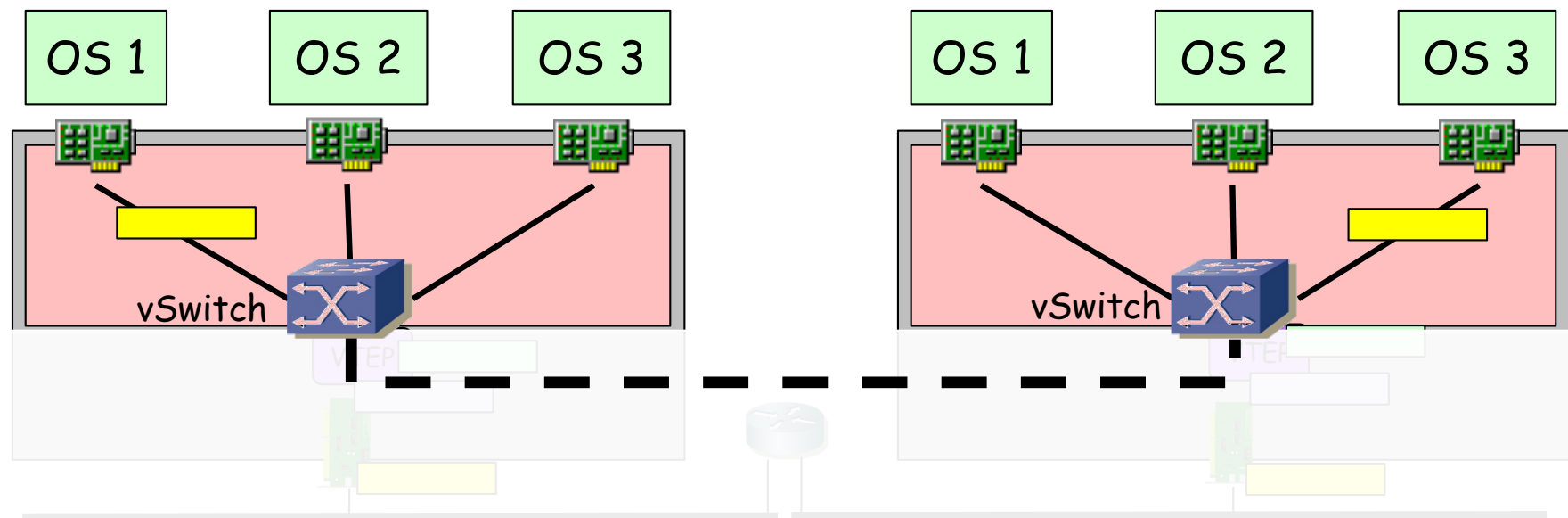
VXLAN: *Data plane*

- 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



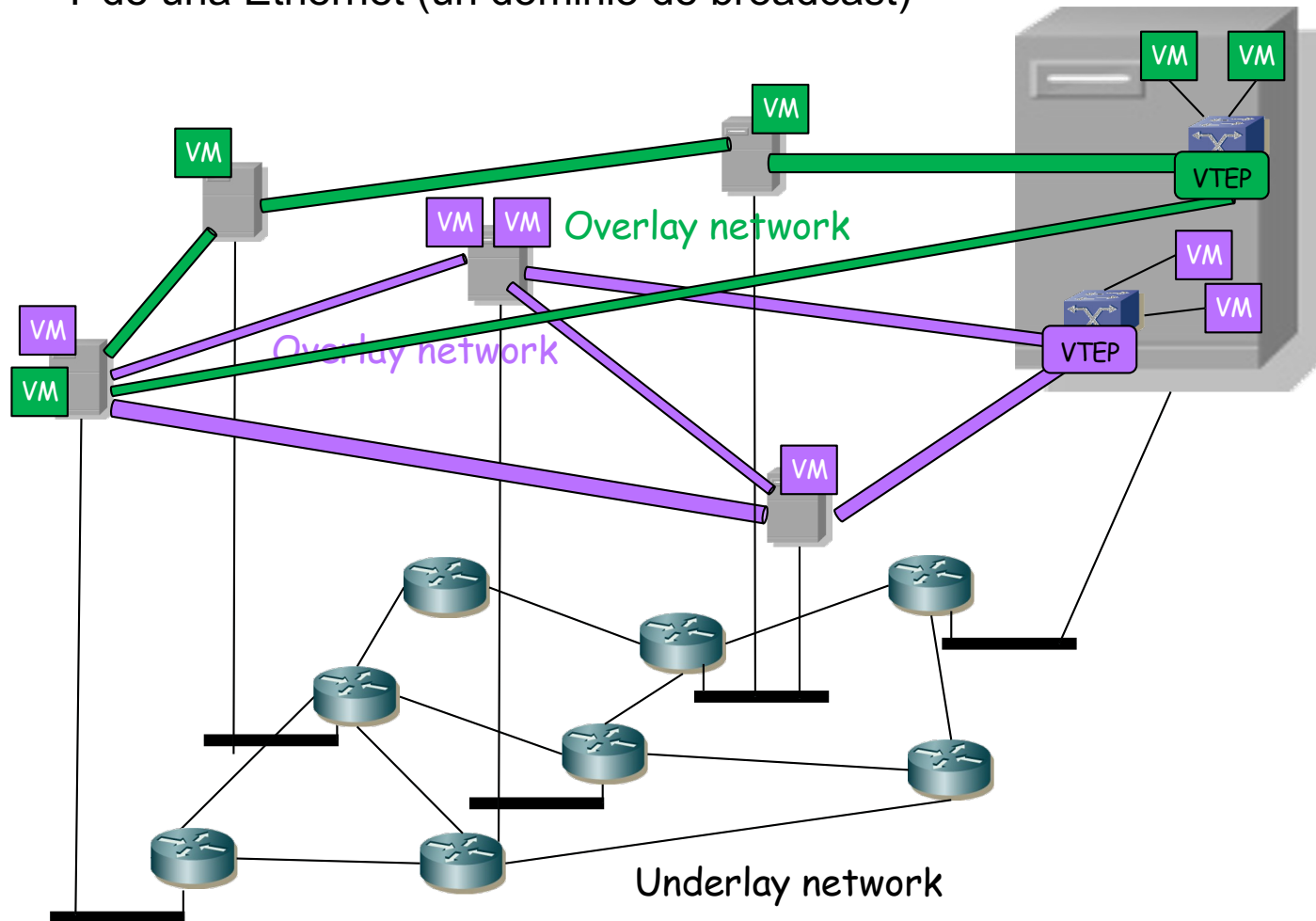
VXLAN: *Data plane*

- 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)



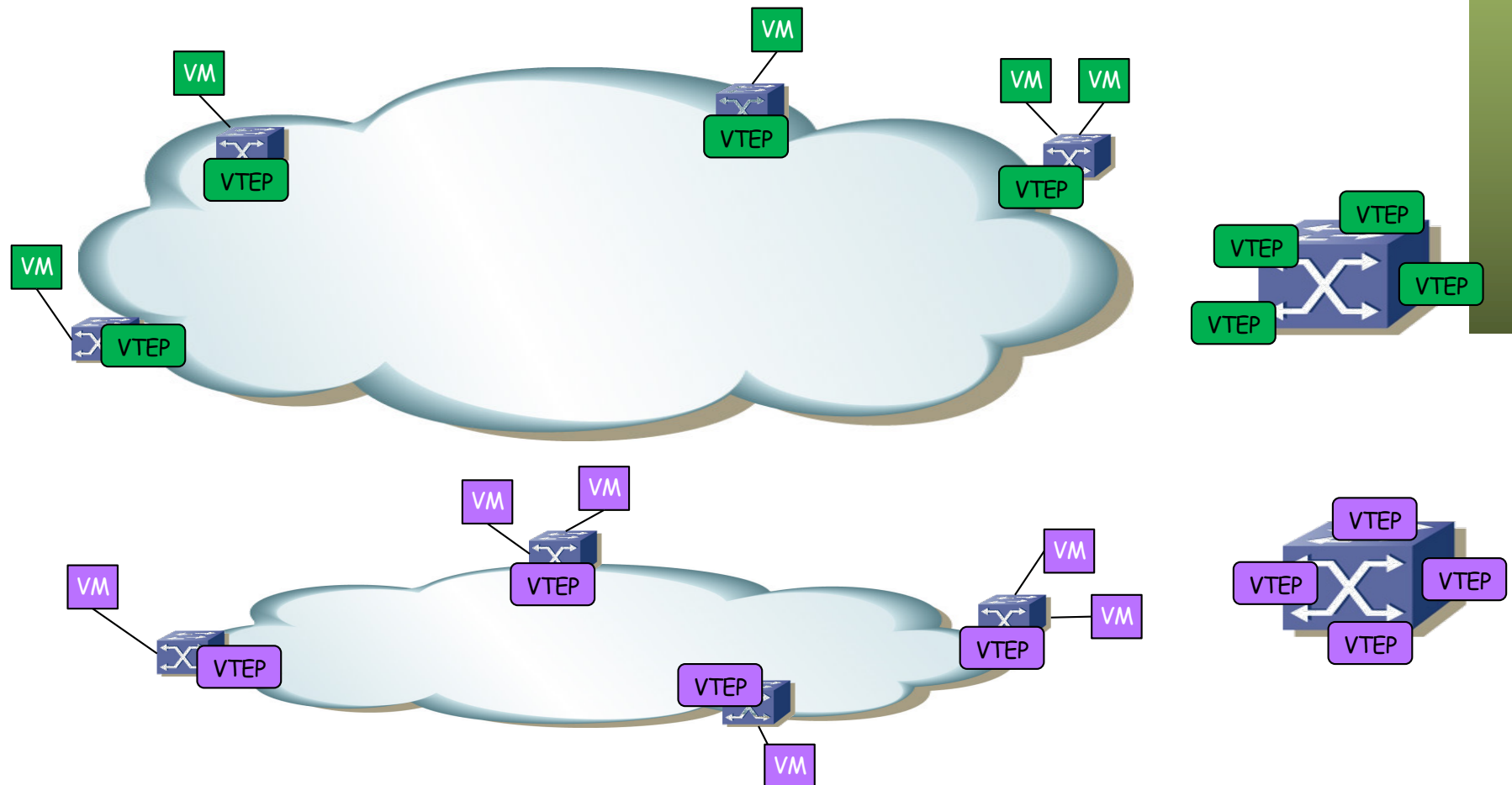
Uso en overlays

- 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)



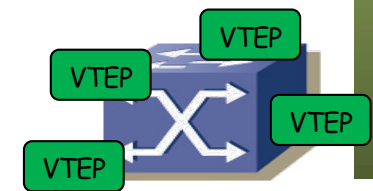
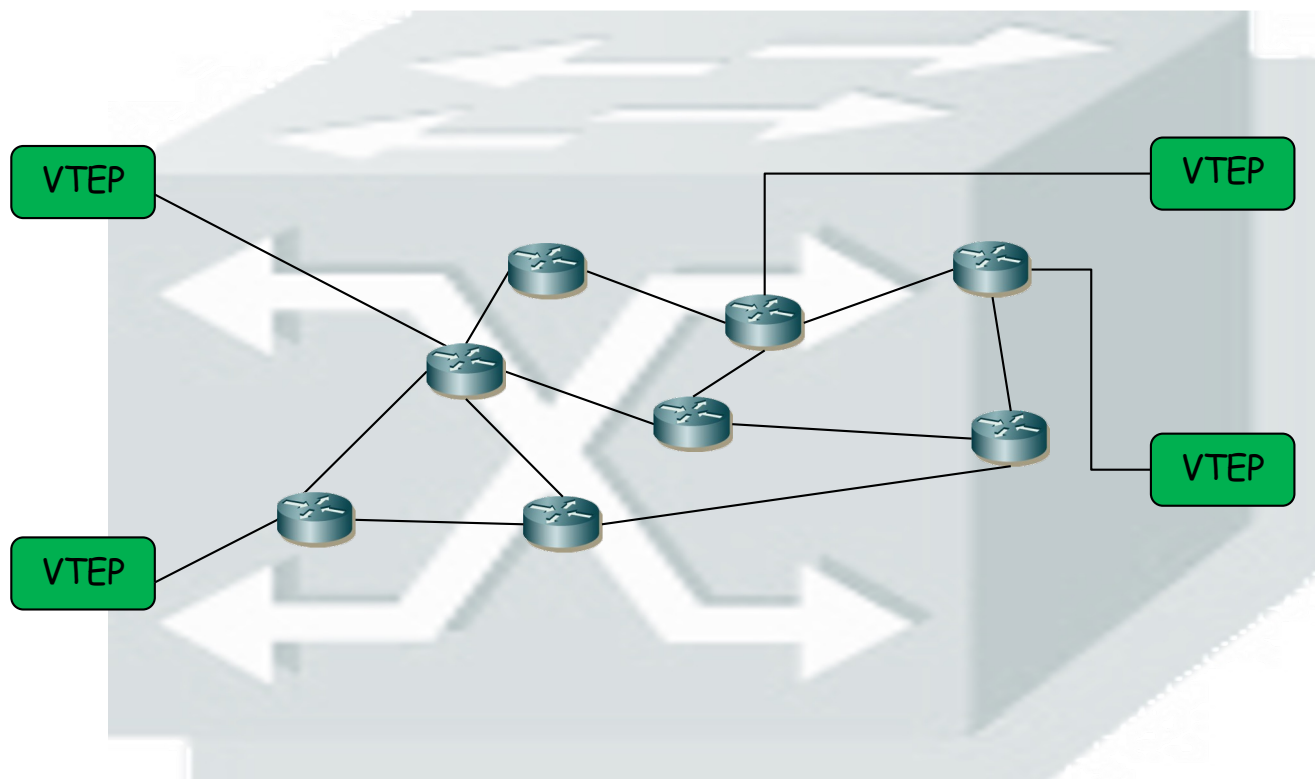
Abstracción de la underlay

- 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



VXLAN-GPE

- draft-ietf-nvo3-vxlan-gpe-12 (2021-09-22, Cisco, Arrcus, Intel)
- <https://datatracker.ietf.org/doc/draft-ietf-nvo3-vxlan-gpe/>
- GPE = Generic Protocol Extension
- Soporte de
 - Encapsulación de múltiples protocolos
 - OAM
 - Ingress-replicated BUM traffic
 - Versiones
- Puerto UDP destino 4790
- Puerto UDP origen para entropía

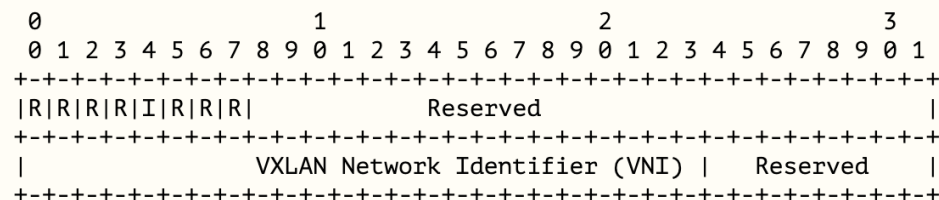


Figure 1: VXLAN Header

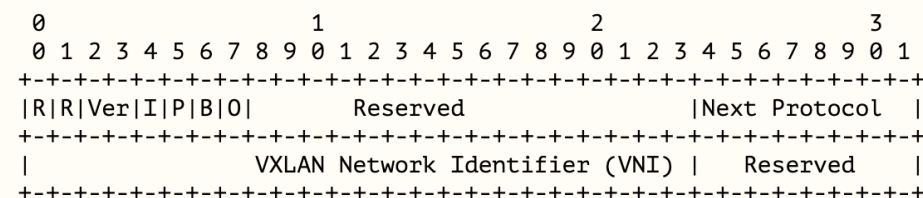


Figure 2: VXLAN-GPE Header

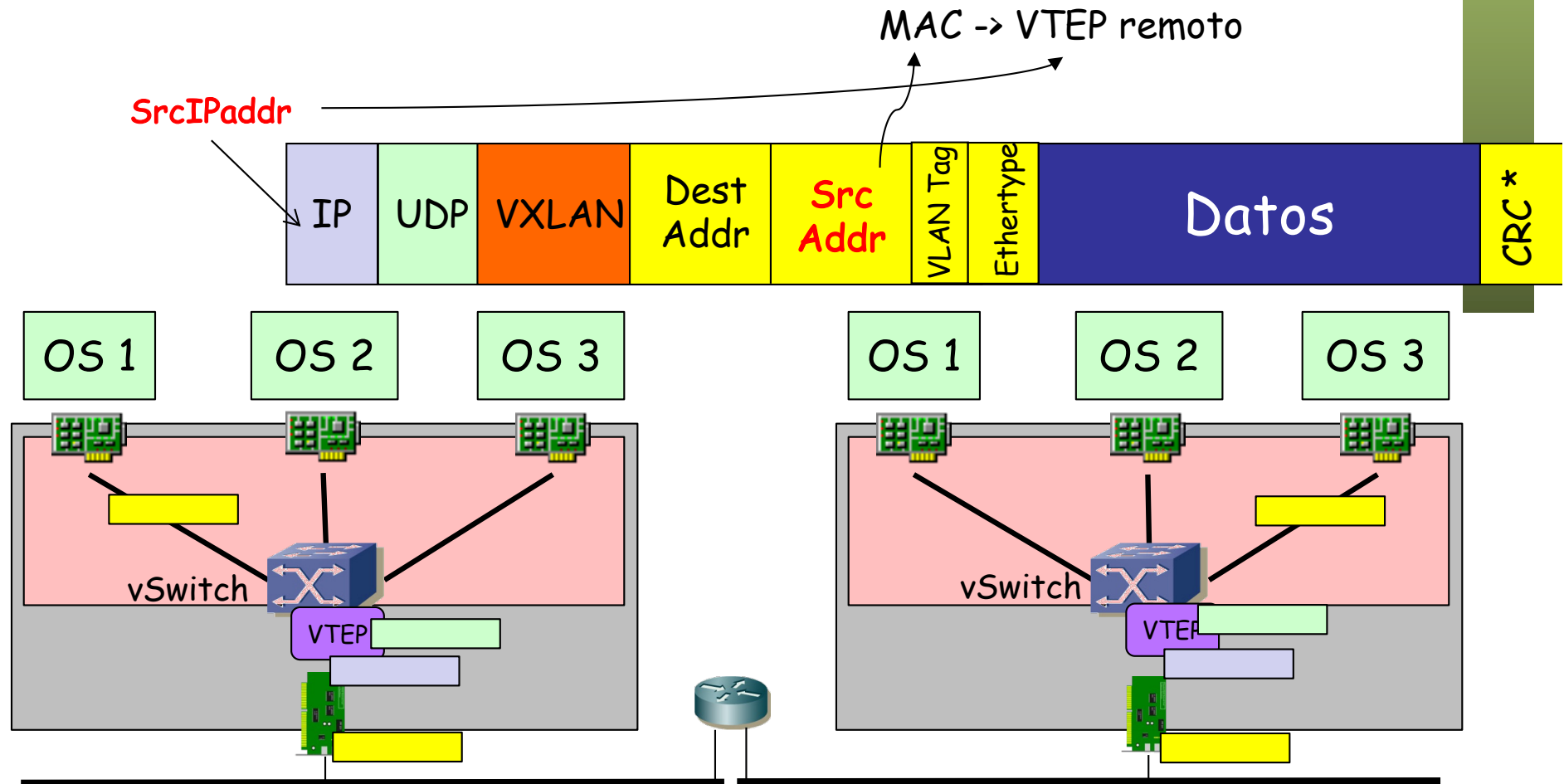
- This draft defines the following Next Protocol values:
 - 0x00 : Reserved
 - 0x01 : IPv4
 - 0x02 : IPv6
 - 0x03 : Ethernet
 - 0x04 : Network Service Header [RFC8300]
 - 0x05 to 0x7D: Unassigned
 - 0x7E, 0x7F: Experimentation and testing
 - 0x80 to 0xFD: Unassigned (shim headers)
 - 0xFE, 0xFF: Experimentation and testing (shim headers)

VXLAN Data Plane

VXLAN Control Plane

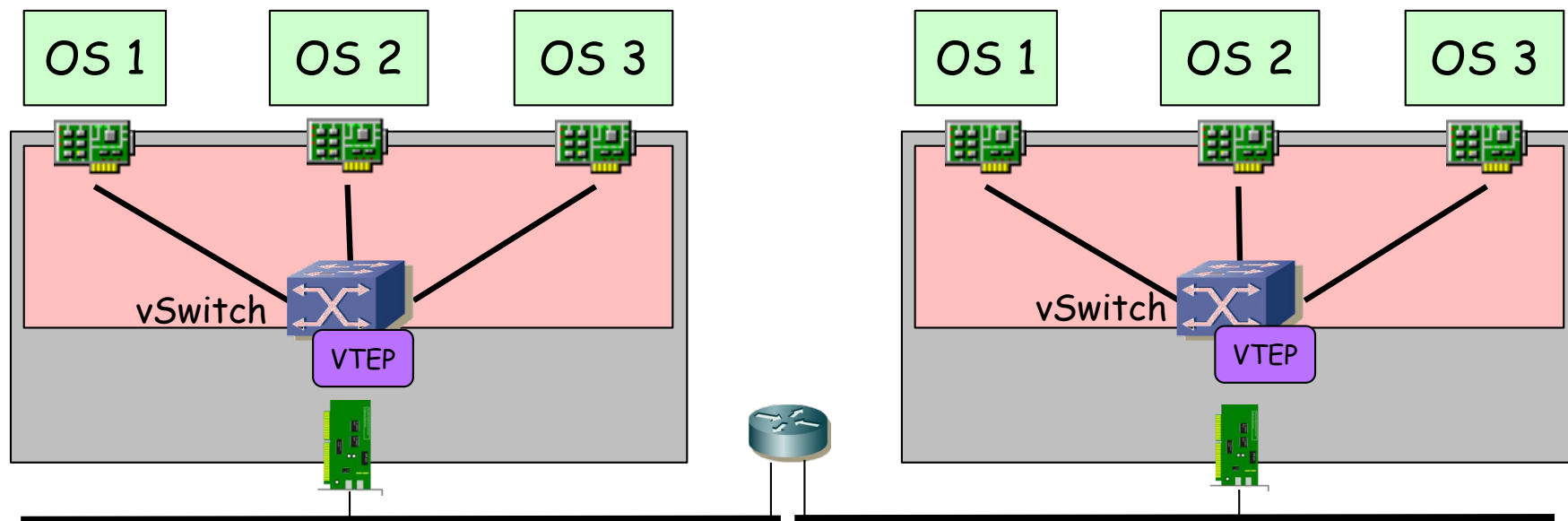
VXLAN: *Control plane*

- Los vSwitch deben aprender la dirección IP del host de la VM
- Aprende con información del plano de datos: al recibir un paquete 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



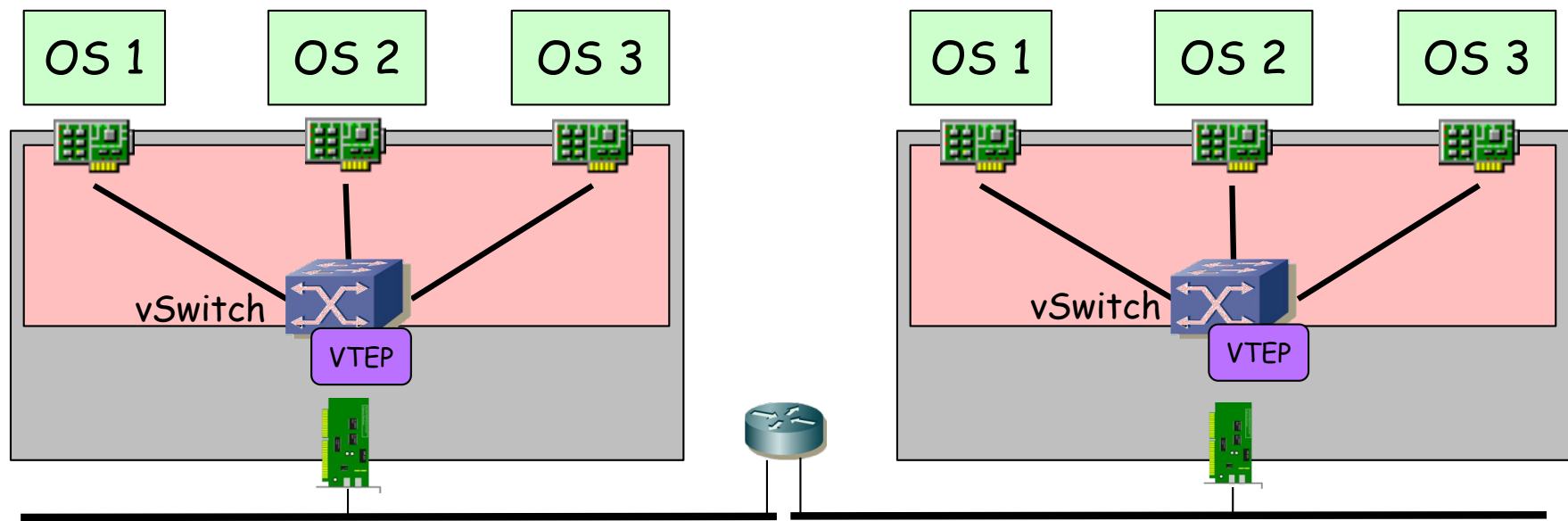
VXLAN: *Control plane*

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



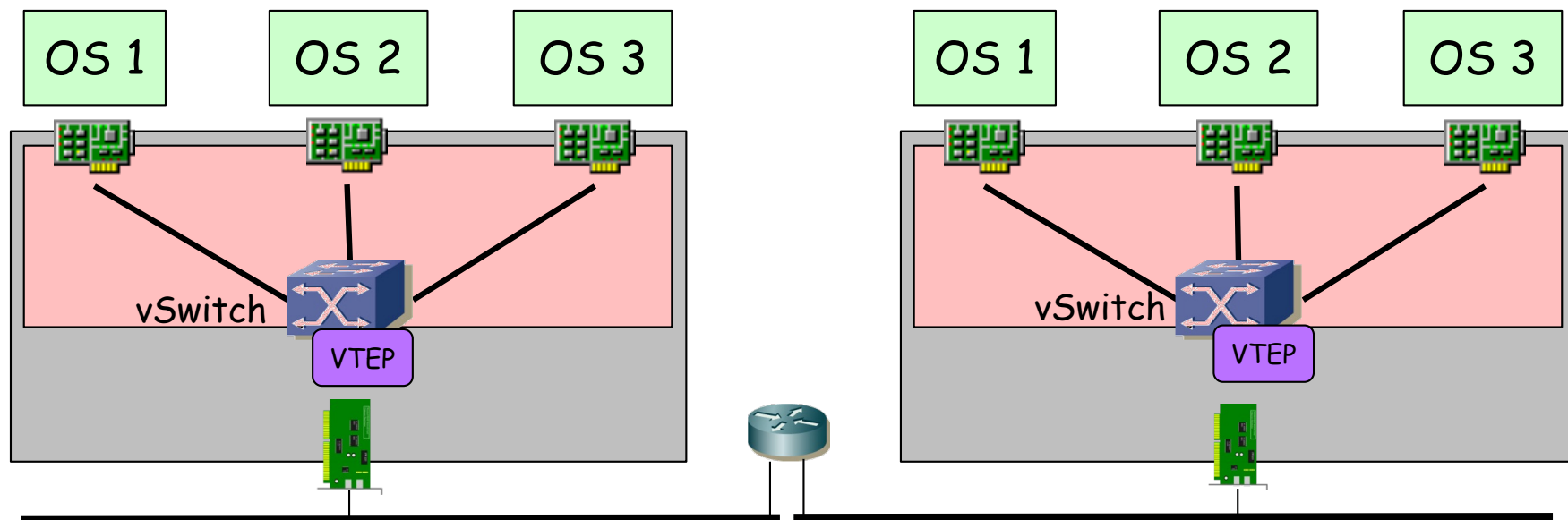
VXLAN: *Control plane*

- ¿Y el BUM? Por ejemplo los ARP
- (...)



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
- Replicación unicast en origen es posible pero requiere que cada VTEP conozca a priori la dirección de cada otro



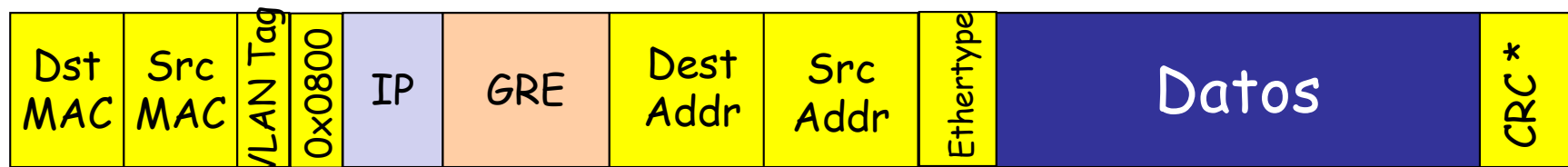
VXLAN Control Plane

VXLAN

NVGRE

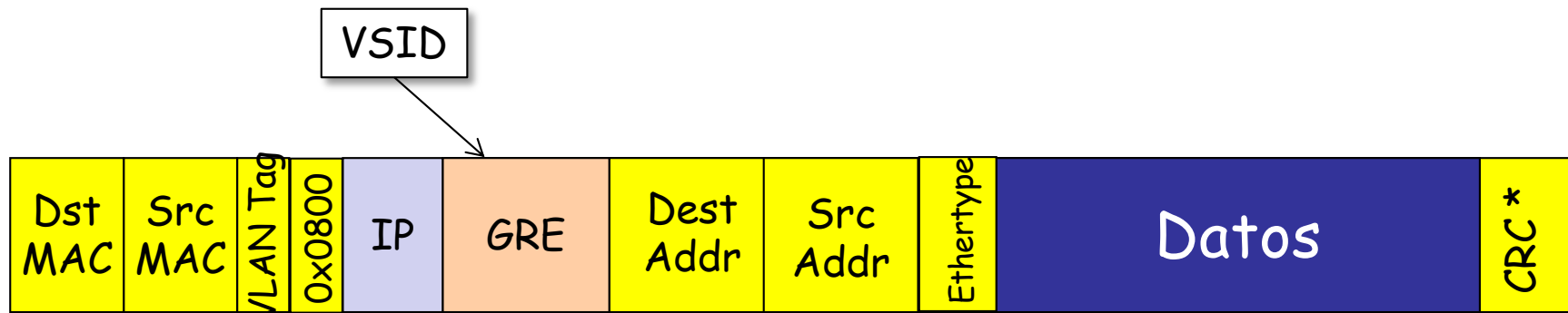
NVGRE

- 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



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

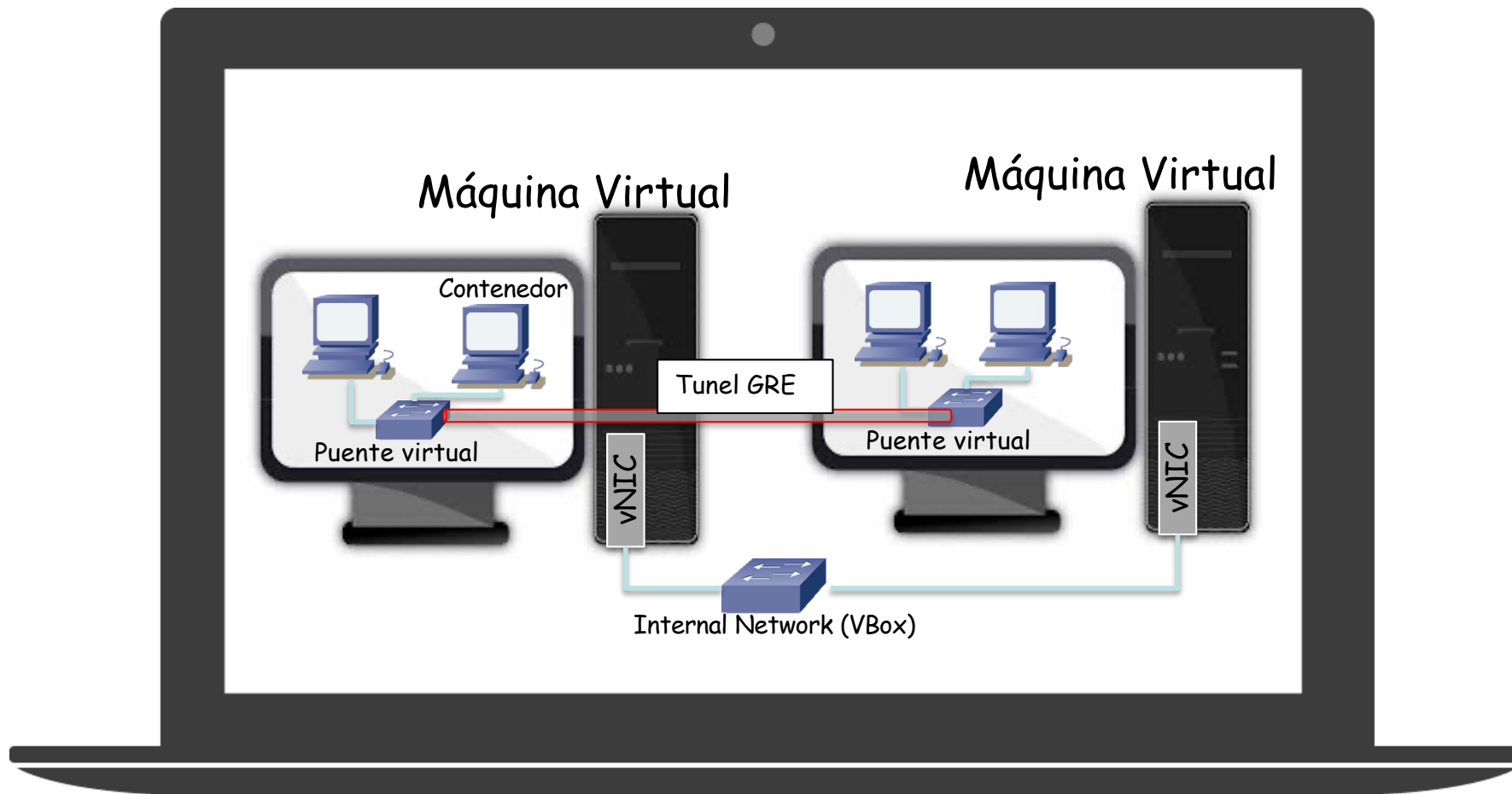


Geneve

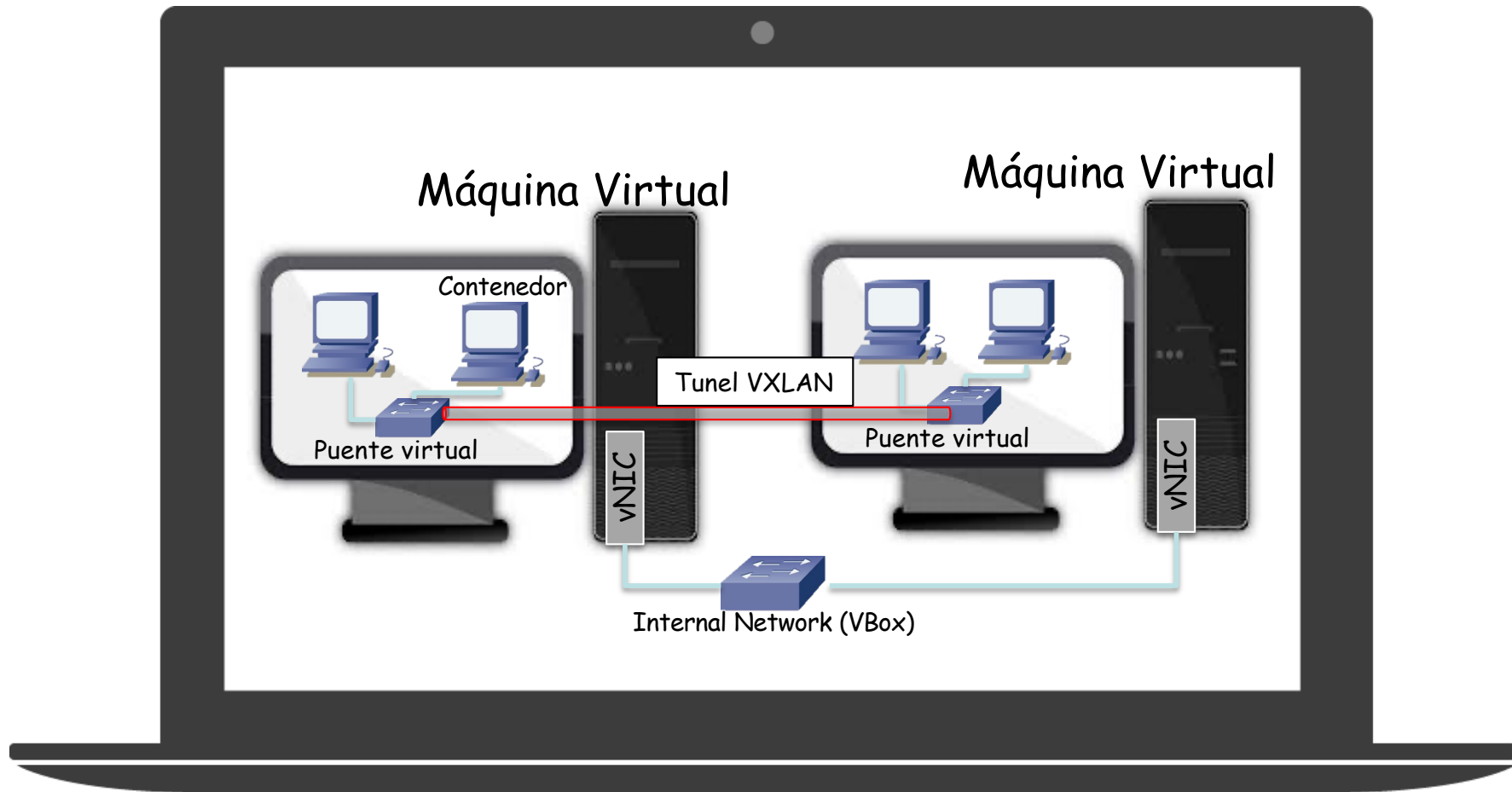
Overlays en el data center

Práctica

Túnel GRE



Túnel VXLAN



Práctica