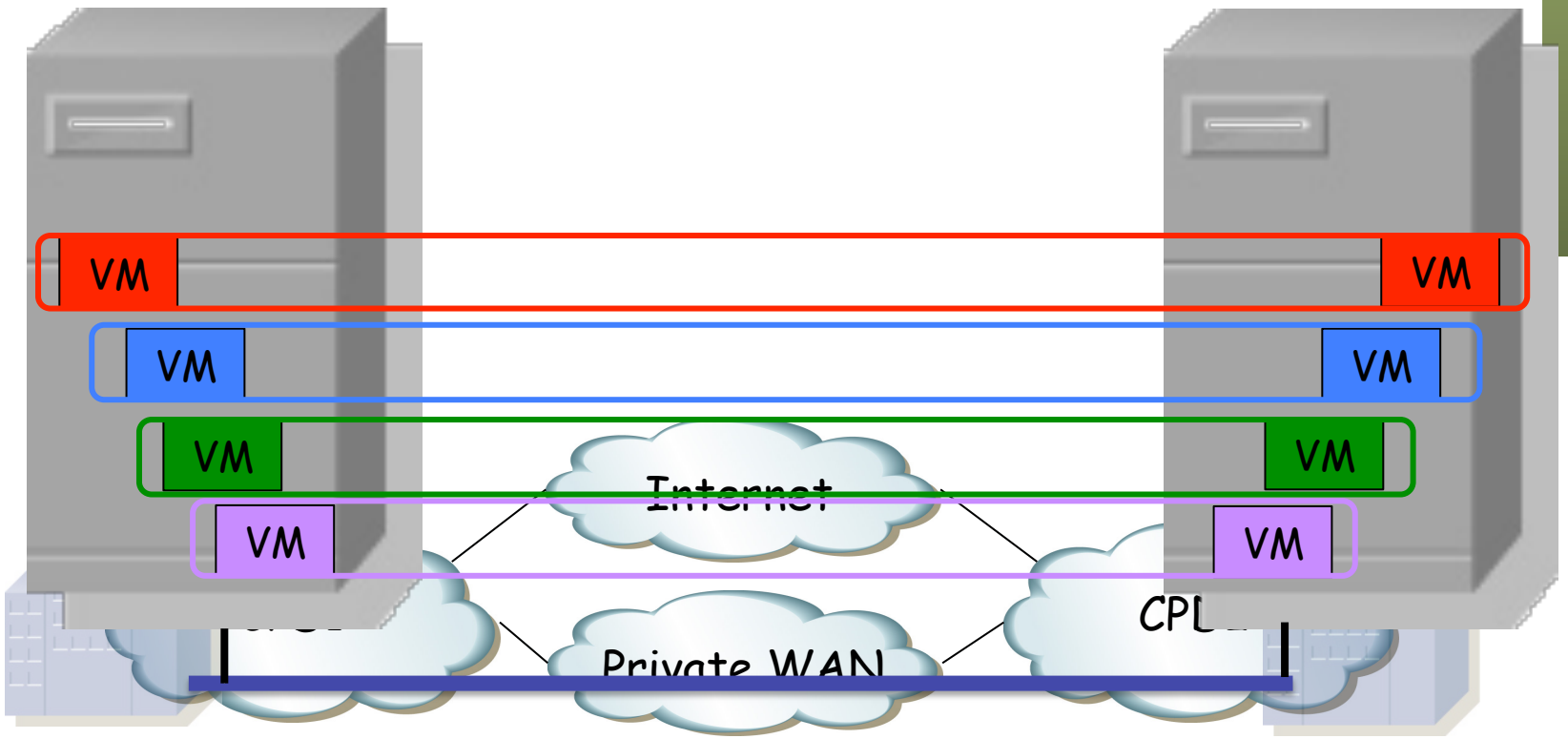
VXLAN

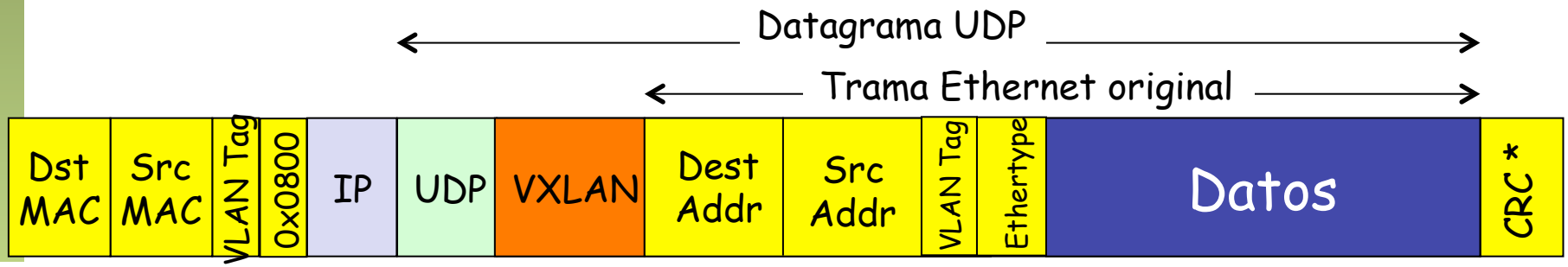
VXLAN

- 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



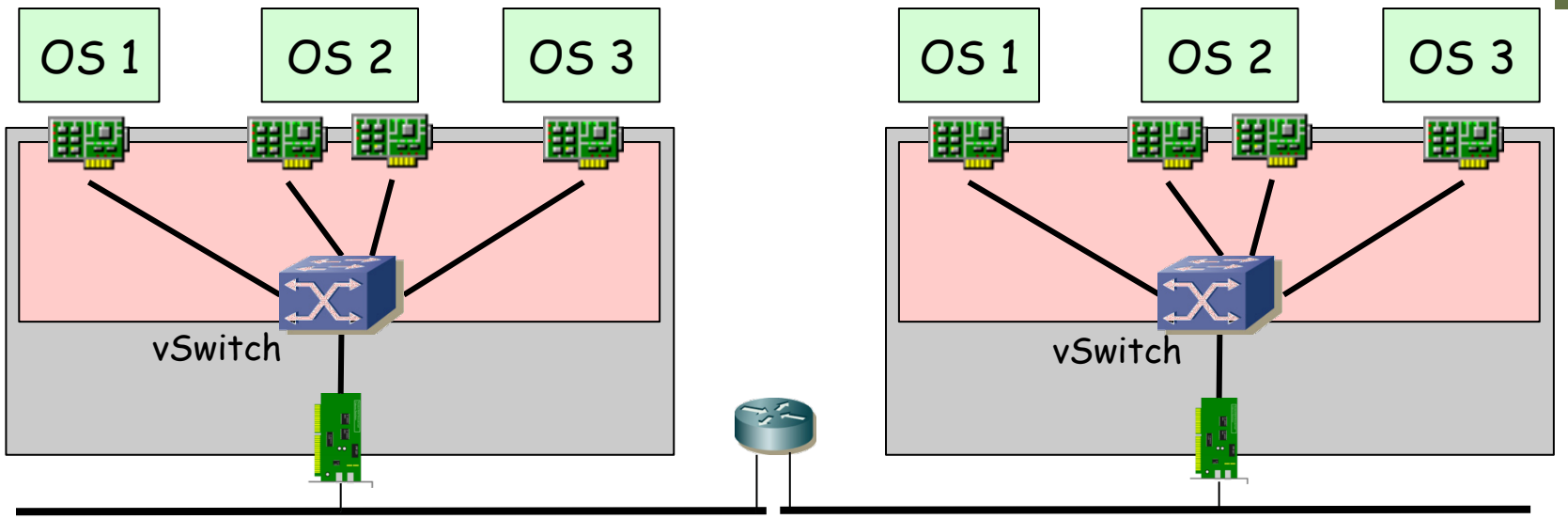
VXLAN

- Emplea un esquema de overlay de capa 2 sobre capa 3 (o sea, un túnel), en el mismo data center o en otro
- En realidad sobre capa 4 pues hace el transporte sobre UDP
- 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
- Para las máquinas virtuales es transparente



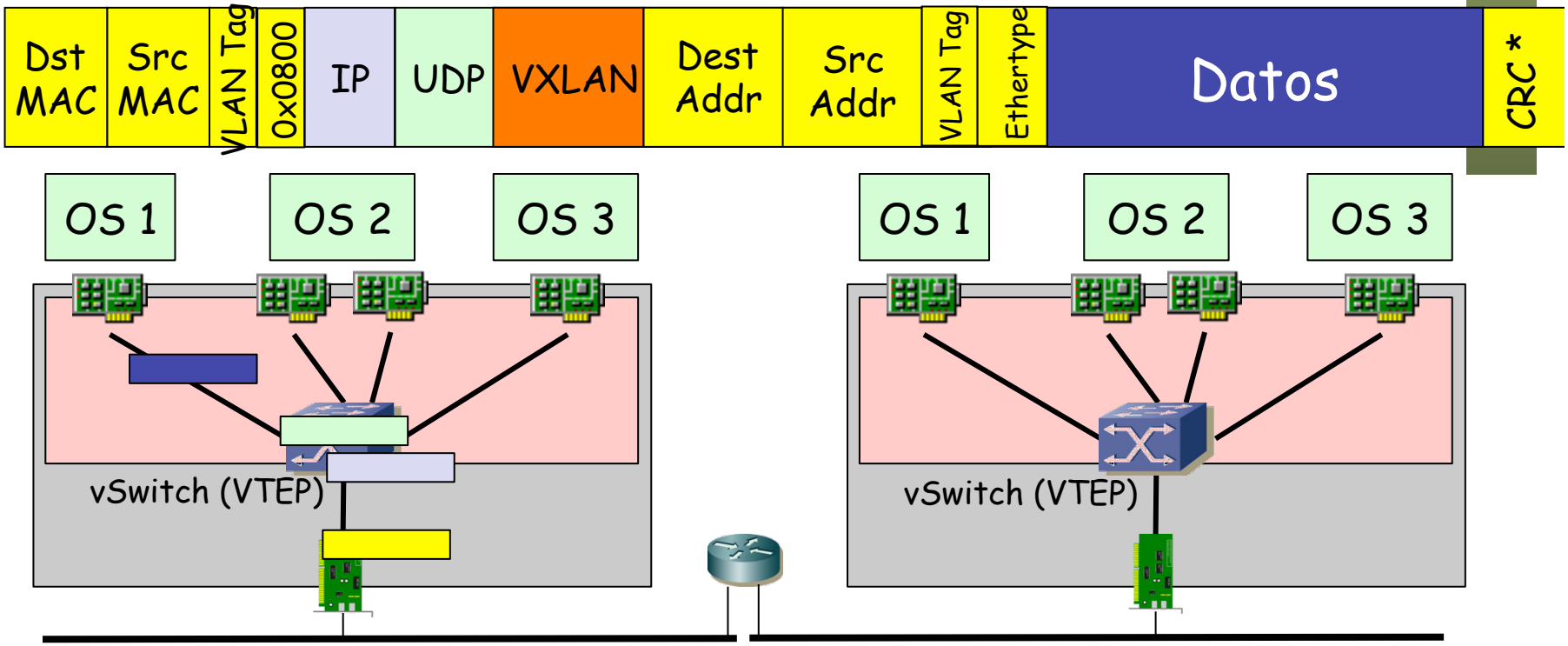
VXLAN: *Data plane*

- 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 extremo que encapsula la trama original se llama el VTEP (VXLAN Tunnel End Point)
- El VTEP se suele encontrar en el hypervisor (transparente para la VM)
- Podría estar si no en un ToR switch



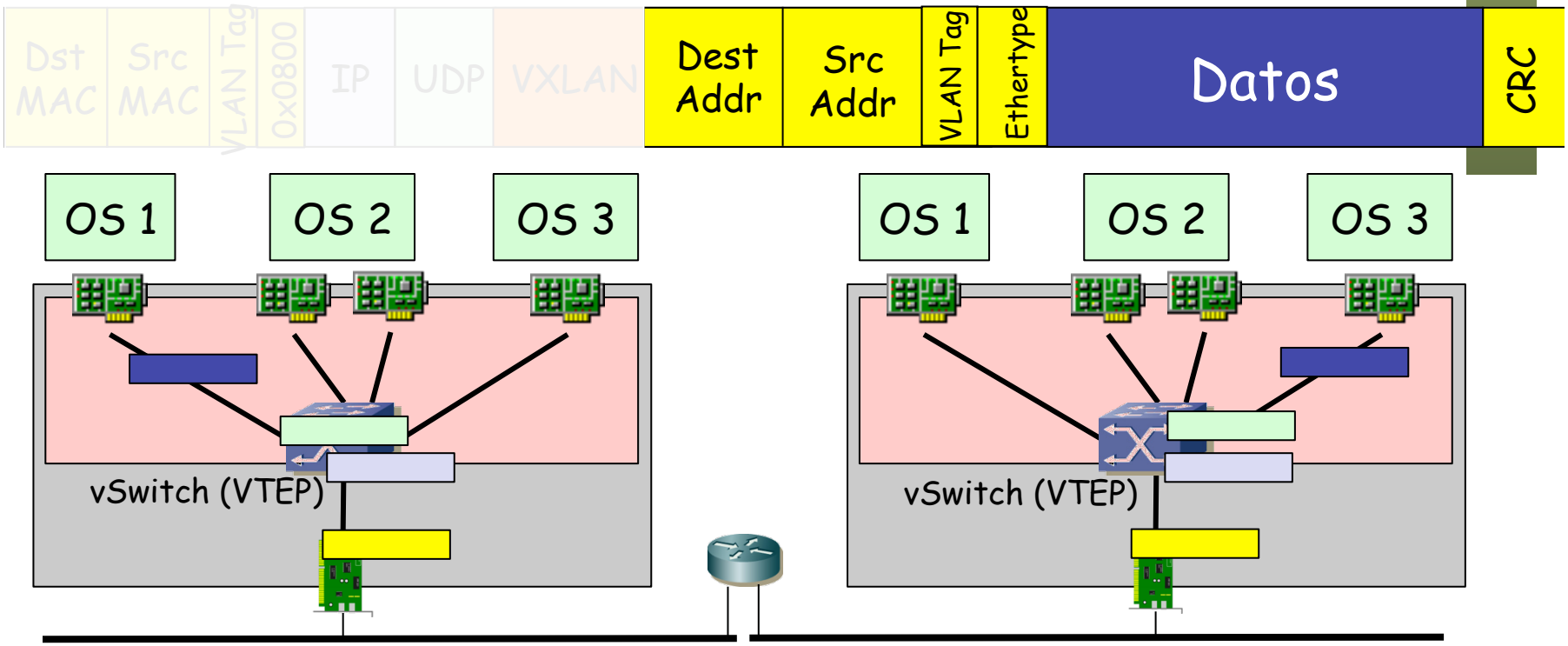
VXLAN: *Data plane*

- 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



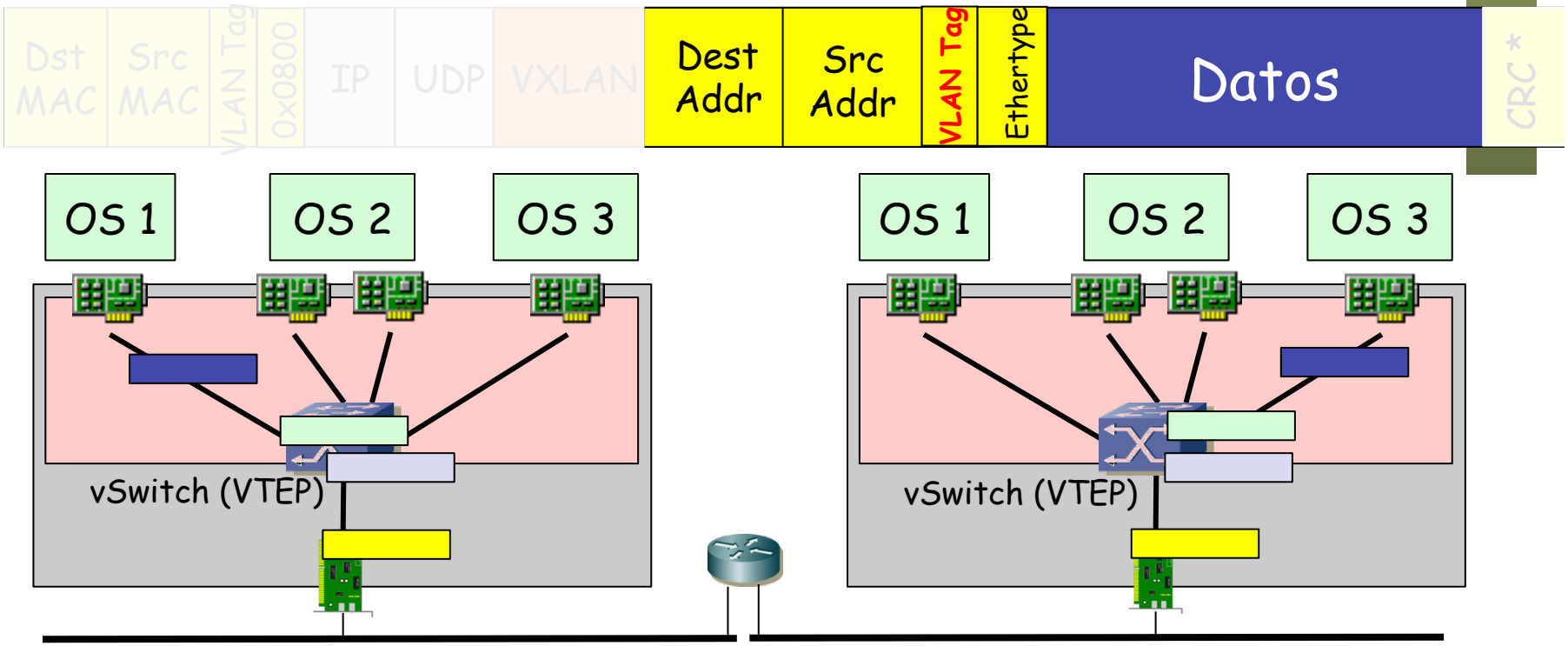
VXLAN: *Data plane*

- Si hay LAGs los switches que repartan flujos en función de capa 3+ pueden repartir estos flujos por los enlaces (si lo hacen por MAC peor)
- 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



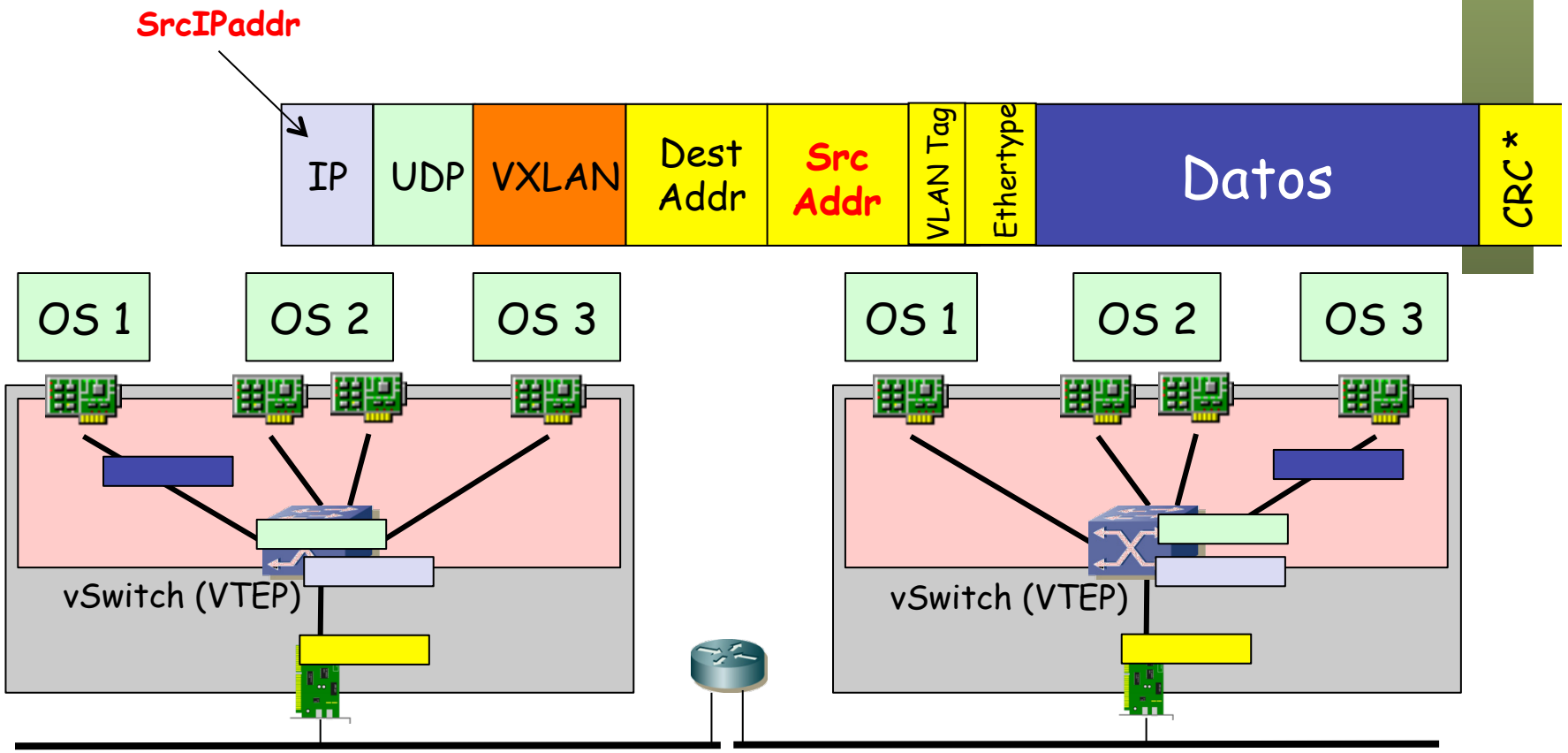
VXLAN: *Data plane*

- 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)



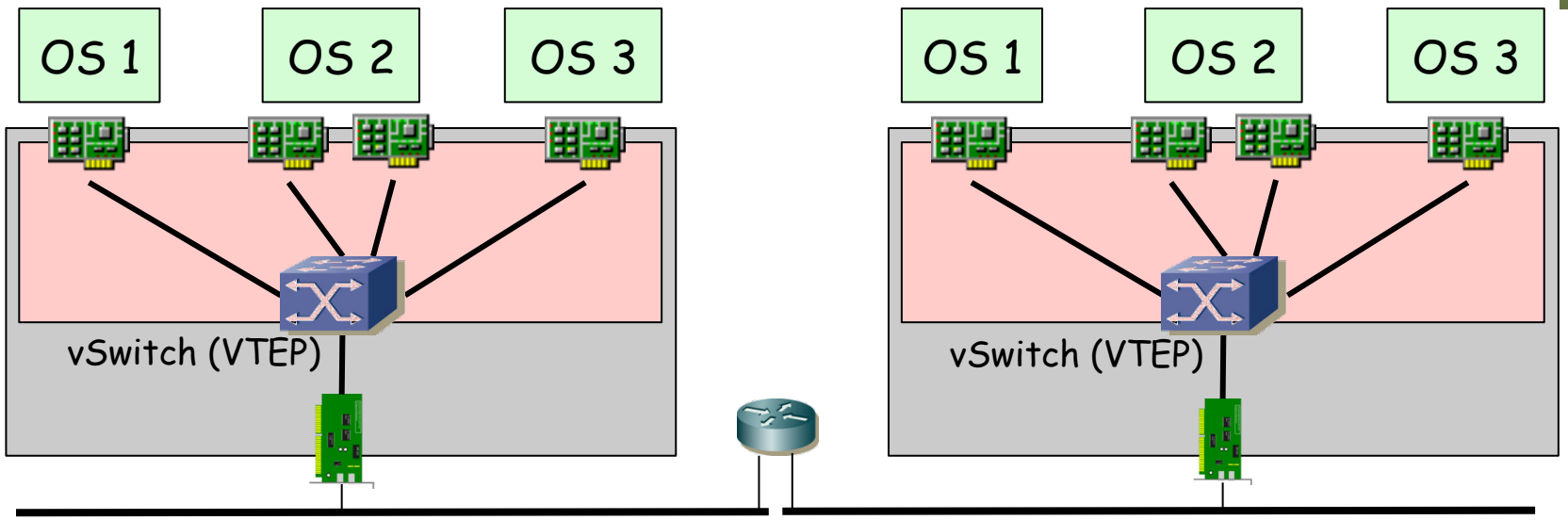
VXLAN: *Control plane*

- Los vSwitch deben aprender la dirección IP del host que hospeda una VM
- En este caso, al recibir un paquete de datos
- Aprende que la dirección MAC origen en el contenido es de un host en la máquina con dirección IP la origen en el continente



VXLAN: *Control plane*

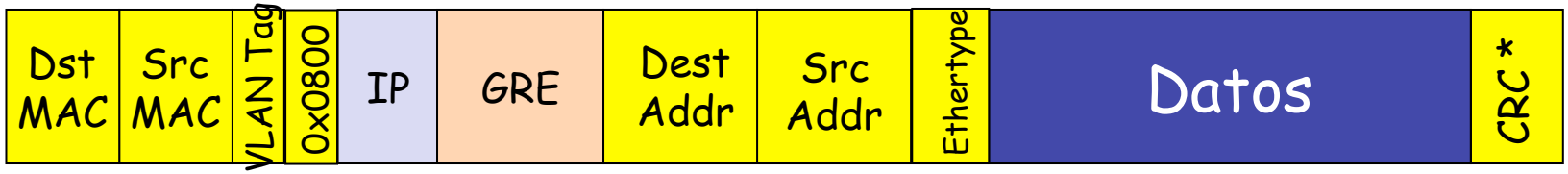
- ¿Y el BUM? Por ejemplo los ARP
- Se envía a un grupo multicast IP (uno por segmento VXLAN)
- Todos los hosts del segmento VXLAN pertenecen a ese grupo
- Esto implica routing multicast en la red IP (algo como PIM-SM)
- El número de grupos multicast soportados por la red puede ser limitado, lo cual llevaría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



NVGRE

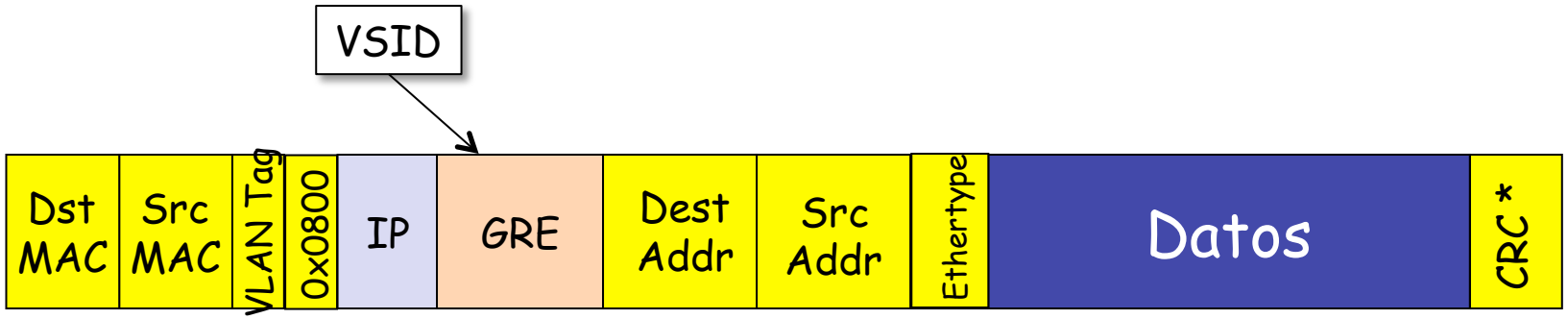
NVGRE

- 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