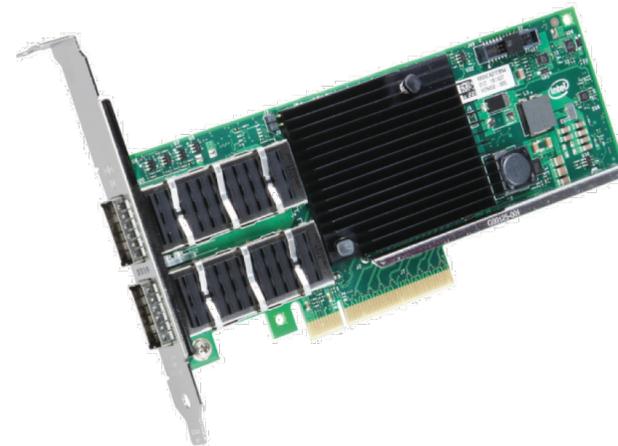


NICs Ethernet para servidor

Tareas en la NIC

- Por un enlace 10GE pueden llegar en 1 segundo más de 14 millones de tramas de 64 bytes
- Eso da a la CPU unos 67ns para procesar cada una
- Las CPUs tienen serios problemas para procesar en ese tiempo cabeceras TCP/IP
- Una NIC puede incluir electrónica para llevar a cabo ciertas tareas de TCP/IP descargando a la CPU
- La NIC puede incluir ASICs, Network Processors o un procesador con un sistema operativo de tiempo real
- A 400Gbps una trama cada 1,67ns lo cual está en el rango de los mejores tiempos de acceso a memoria

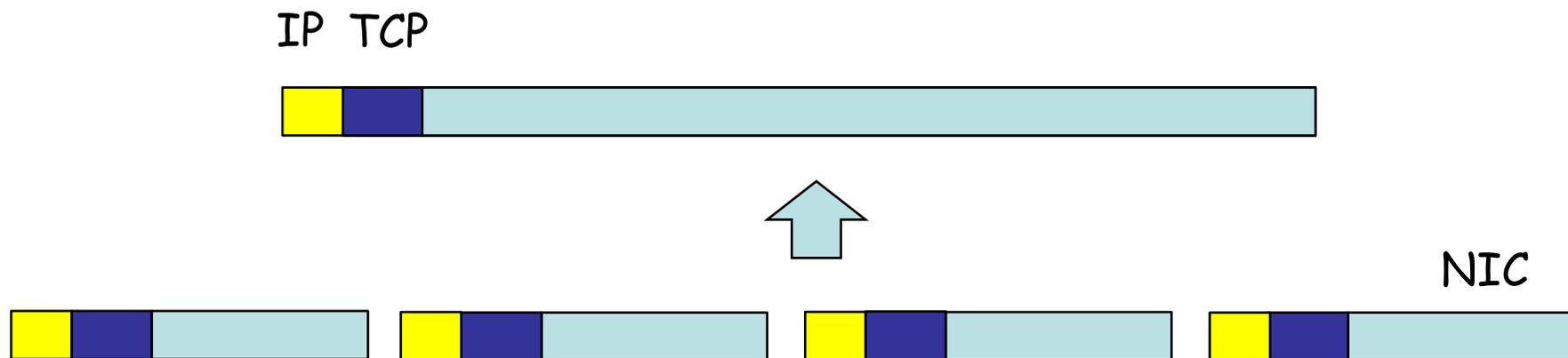


Integración en el bus

- DMA
 - *Direct Memory Access*
 - Transferencia desde la NIC a memoria sin requerir a la CPU
- Coalescencia de interrupciones
 - Las NICs solían generar una interrupción por paquete
 - Alto coste para la CPU
 - Por ejemplo los mainframes tienen CPUs dedicadas a atender I/O
 - La coalescencia hace que la NIC genere una interrupción para un grupo de paquetes en vez de por cada uno

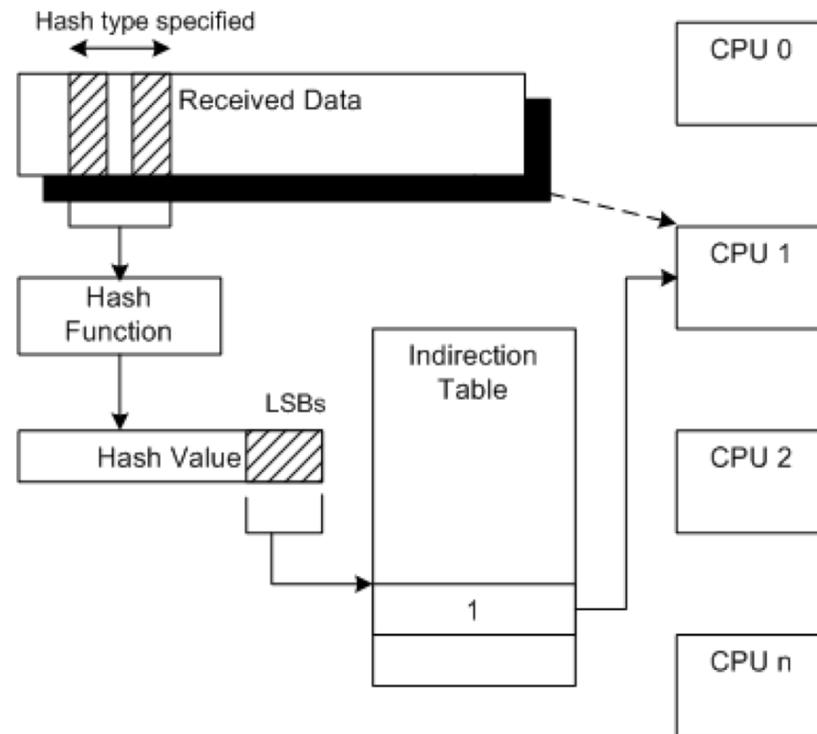
LRO

- *Large Receive Offload*
- La NIC une varios segmentos TCP en uno solo
- Crea unas cabeceras TCP e IP para ese nuevo segmento
- Reduce el número de interrupciones y procesamiento de cabeceras en el kernel



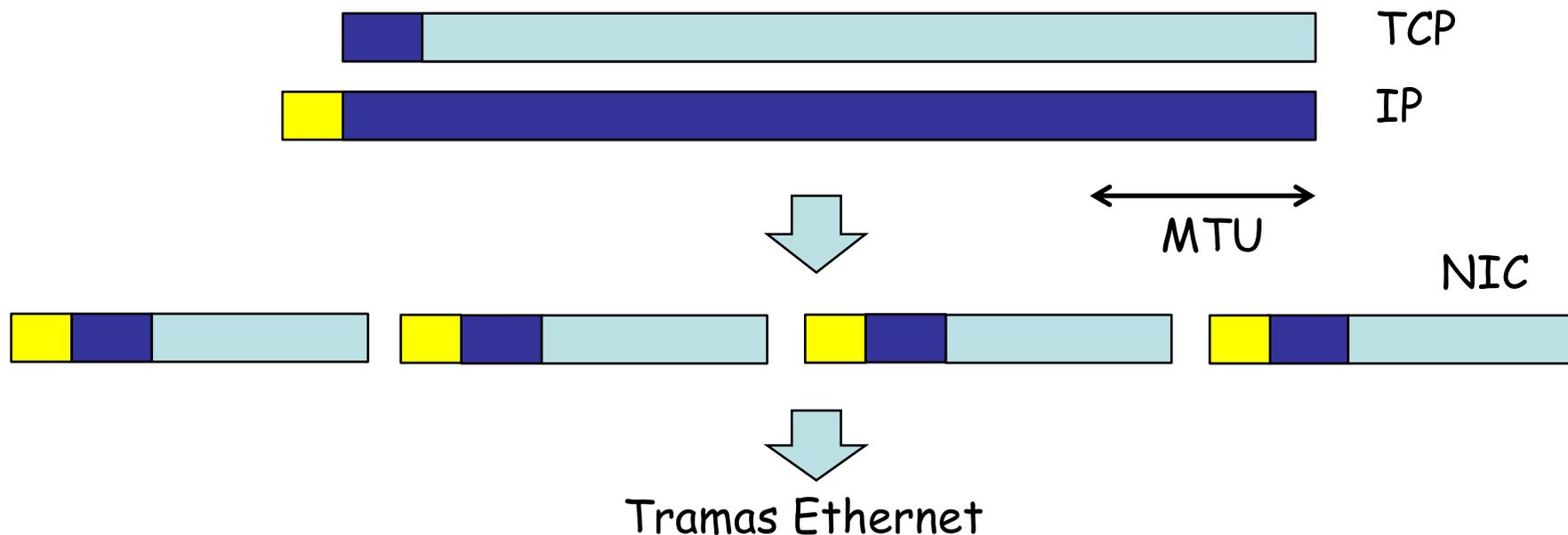
RSS

- *Receive Side Scaling*
- NIC calcula un hash sobre el paquete recibido y con él decide a qué CPU manda la interrupción
- Permite paralelizar entre varias CPUs el procesamiento del tráfico recibido



LSO

- *Large Segment Offload, TCP Segmentation Offload*
- TCP entrega a la NIC paquetes más grandes que la MTU
- La propia NIC hace la segmentación de nivel TCP
- Eso le obliga a crear nuevas cabeceras TCP e IP, descargando de ello a la CPU
- Requiere que la NIC sepa segmentar el protocolo (solo TCP)
- Problemas con encriptación (IPSec)
- Genera ráfagas de tráfico

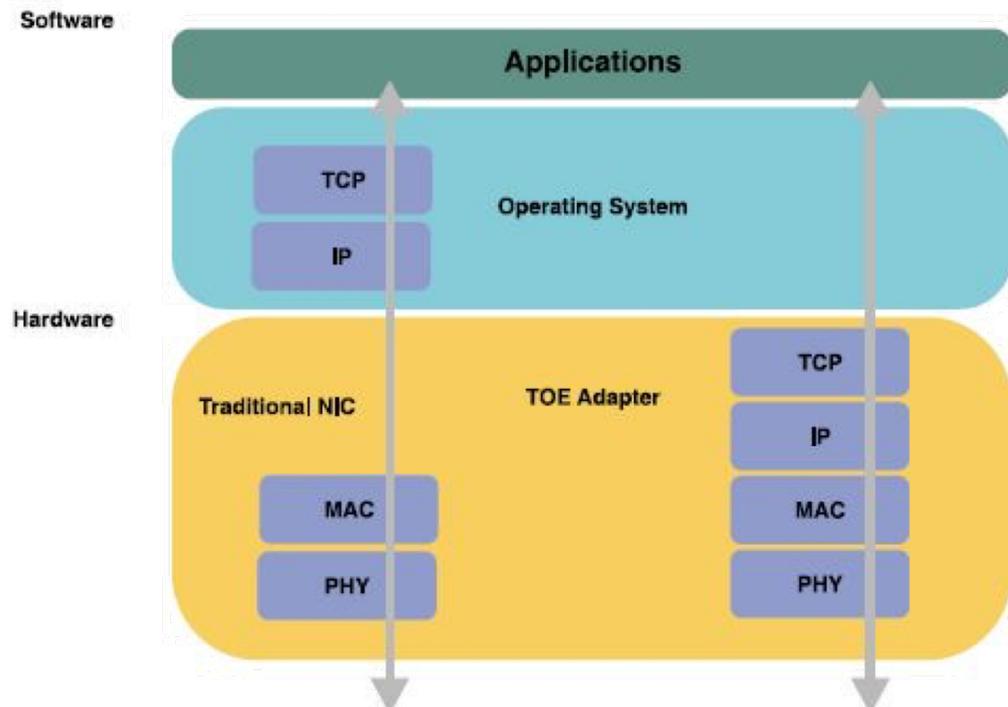


TOE

- *TCP/IP Offload Engine*
- Los datos pueden pasar directamente de la aplicación a la NIC
- La NIC puede emplearse para todas las tareas de la fase de transferencia y emplear la CPU para el establecimiento y terminación
- O se puede emplear la NIC para todo
- Requiere soporte del sistema operativo

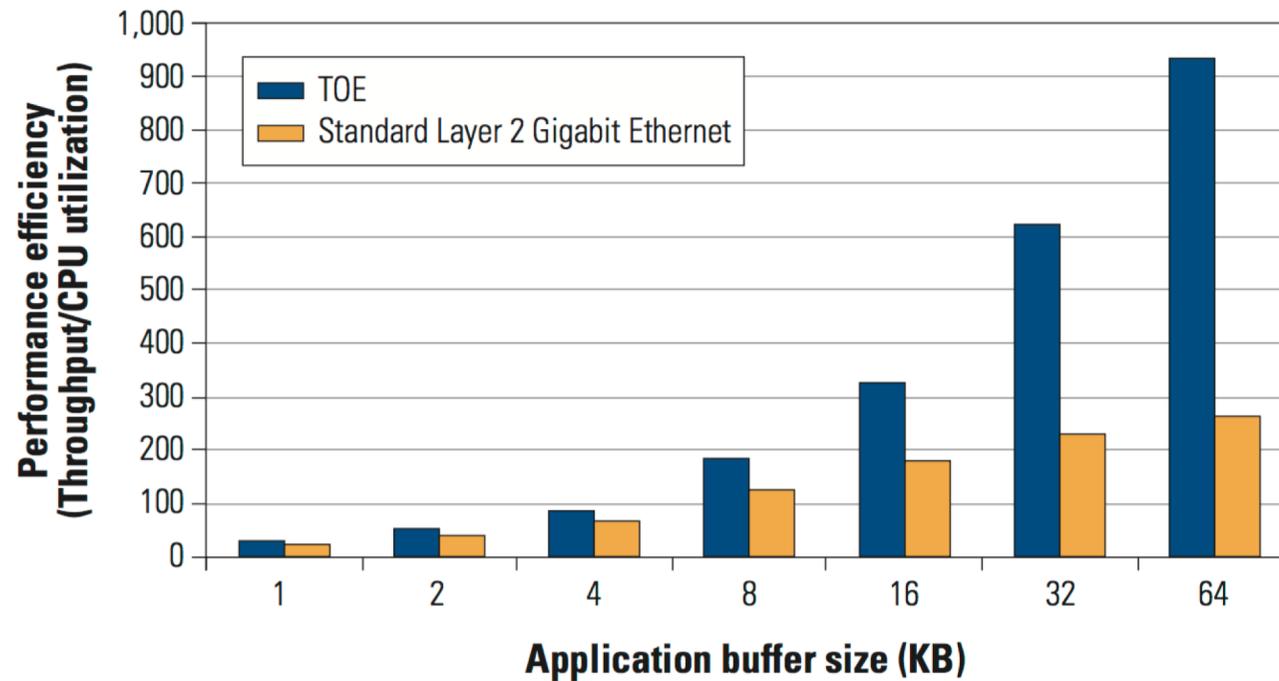


TCP/IP Offload Engine, TOE



TOE

- Puede mejorar el throughput
- Reduce la carga sobre la CPU



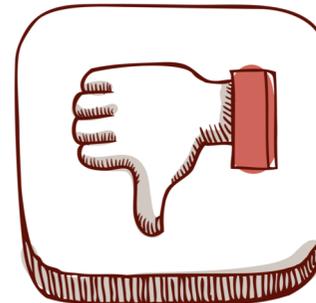
Jumbo frames

- No están estandarizadas, la MTU estándar sigue siendo de 1500bytes
- Motivos para limitarlo
 - NICs tenían memoria limitada
 - Se quería limitar el tiempo que una estación tenía capturado el medio transmitiendo
 - El CRC es menos efectivo cuanto más grande es la trama
- Hoy en día no son problemas reales:
 - Decenas o centenares de Megabytes en la NIC
 - No tenemos medio compartido (ni coaxial ni hubs)
 - El CRC de Ethernet soporta más de 11 Kbytes de trama



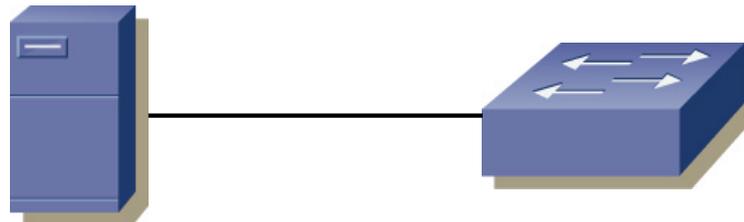
Jumbo frames

- Diversos estándares han ido aumentando el tamaño máximo de la trama (802.1Q, 802.1ad, MPLS, FCoE, etc)
- A estas últimas en ocasiones se las llama “Baby Giant”
- Jumbo frames suelen estar cerca de los 9 Kbytes (que se puedan transportar bloques de datos de 8Kbytes + encapsulados varios)
- ¿Positivo?
 - Cuanto más grandes menor ratio de cabeceras y menos interrupciones
 - Menos carga de procesamiento de cabeceras en equipos de red y hosts
- ¿Negativo?
 - Todos los equipos del camino deben soportarlas
 - Posibles problemas con implementaciones que esperan 1500 bytes
 - Mayores tramas sufren mayor retardo así que no son adecuadas para todos los servicios
 - Mayores tramas pueden llenar antes los buffers de los conmutadores



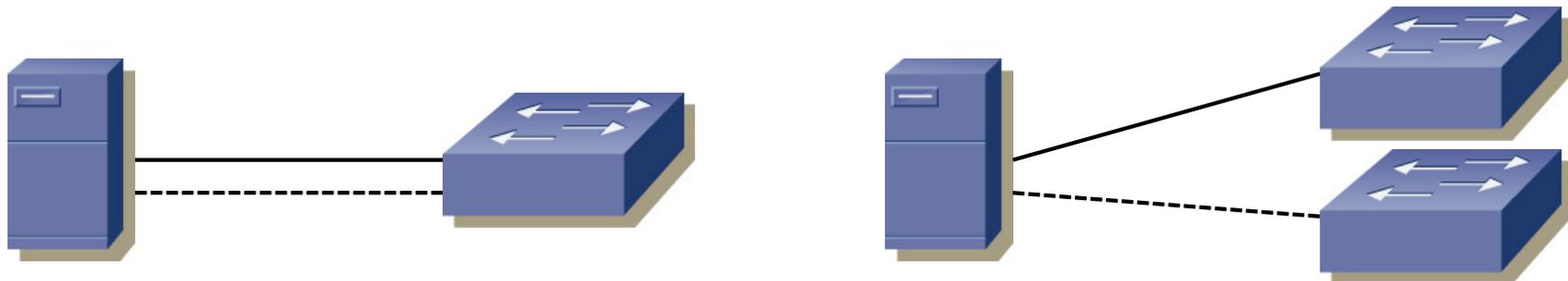
Server multihoming

- *NIC teaming / bonding / aggregation*
- Un servidor conectado a un conmutador presenta puntos únicos de fallo: la NIC, el cable, el conmutador
- Estas soluciones requieren colaboración del driver y normalmente también del sistema operativo
- Tenemos varias mejoras posibles (con una segunda o más NICs)
- (...)



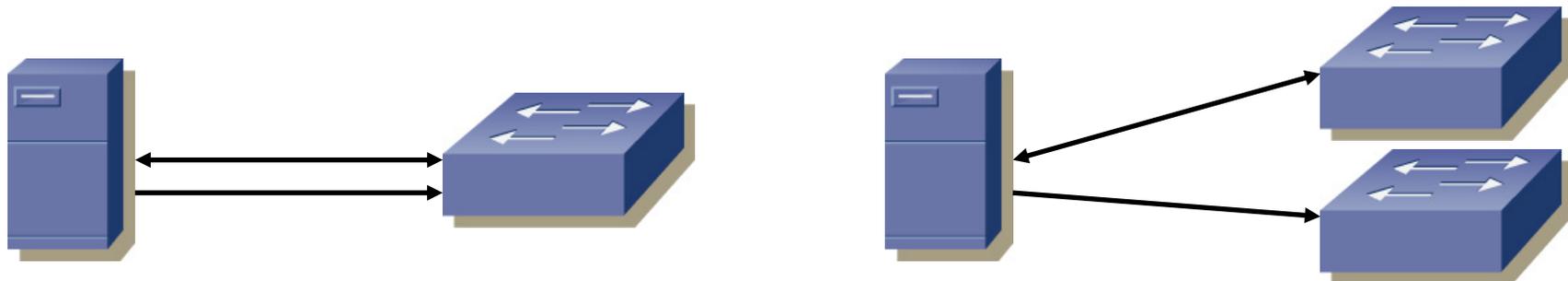
Server multihoming

- Un segundo enlace, modo activo-pasivo
 - Si falla el primero (la NIC, el conmutador o el cable) se activa el segundo con la misma dirección MAC e IP
 - Se desaprovecha el segundo enlace



Server multihoming

- Un segundo enlace, modo activo-pasivo
- O se usan los dos enlaces para transmitir pero solo se recibe por uno
- Cada interfaz suele enviar con diferente dirección MAC origen para no tener *MAC flapping* en el conmutador



Server multihoming

- Un segundo enlace, modo activo-pasivo
- O se usan los dos enlaces para transmitir pero solo se recibe por uno
- O se forma un LAG (802.3ad / 802.1AX)
 - Permite usar la capacidad de ambos enlaces
 - Normalmente requiere colaboración por parte del switch
 - Si se quiere redundancia de switch hay que hacer una agregación en la que un extremo son 2 conmutadores

