

upna

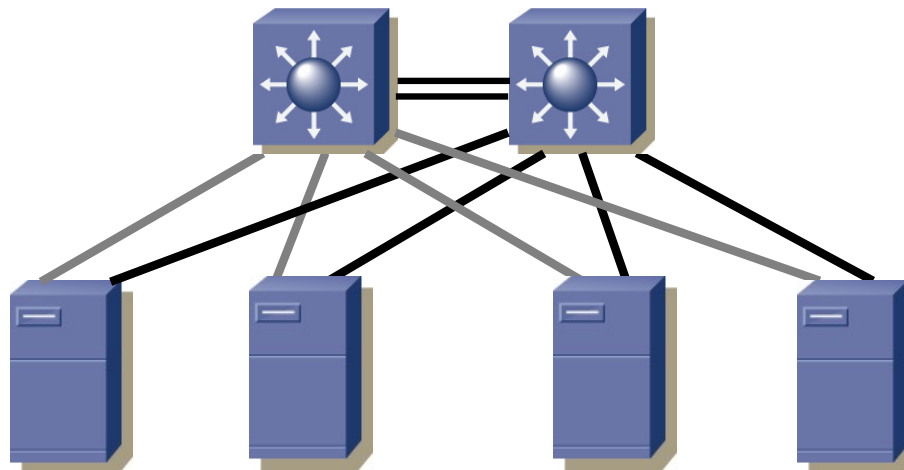
Universidad Pública de Navarra
Nafarroako Unibertsitate Publikoa

Redes de Nueva Generación
Área de Ingeniería Telemática

NICs Ethernet para servidor

Requerimientos

- Soporte de algún tipo de agregación de interfaces
 - Gestionable a nivel de usuario
 - O de sistema operativo
- Alto rendimiento
 - Arquitectura PC limitada para altas tasas de paquete/s
- Virtualización
 - Soporte eficiente de VMs en el host (más adelante)
- Almacenamiento
 - Integración con almacenamiento en red (más adelante)



upna

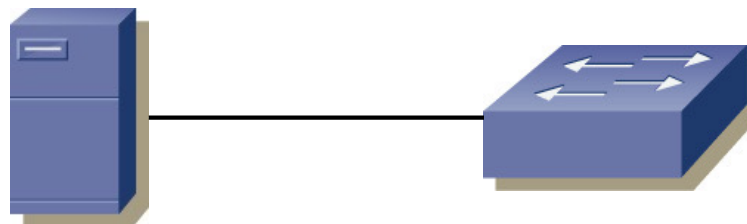
Universidad Pública de Navarra
Nafarroako Unibertsitate Publikoa

Redes de Nueva Generación
Área de Ingeniería Telemática

Server multihoming

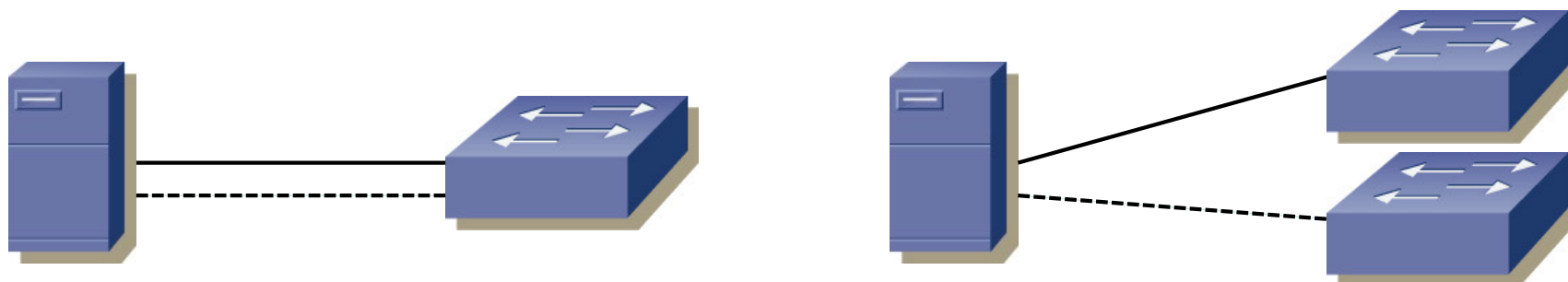
Server multihoming

- *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)
- (...)



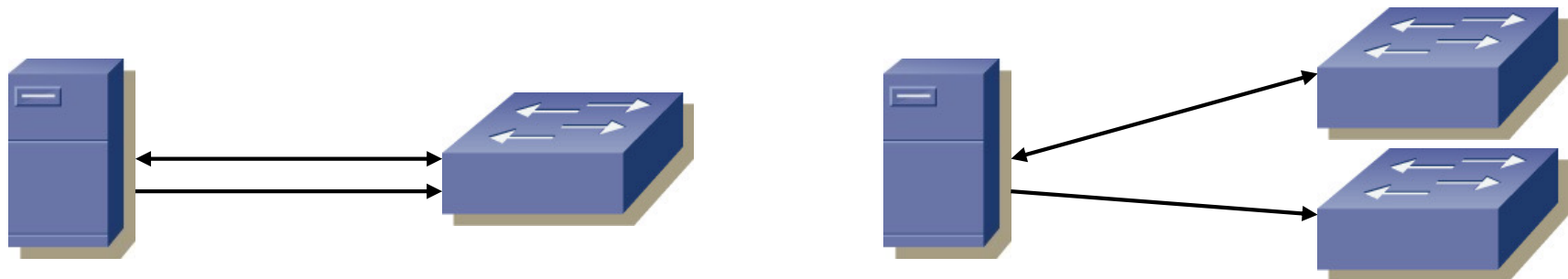
Server multihoming

- 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



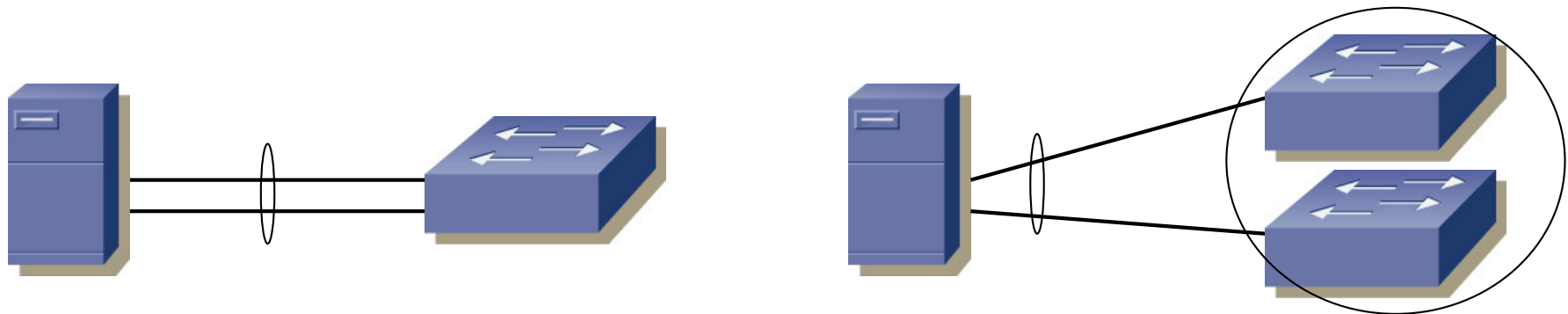
Server multihoming

- 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



upna

Universidad Pública de Navarra
Nafarroako Unibertsitate Publikoa

Redes de Nueva Generación
Área de Ingeniería Telemática

Server multihoming

upna

Universidad Pública de Navarra
Nafarroako Unibertsitate Publikoa

Redes de Nueva Generación
Área de Ingeniería Telemática

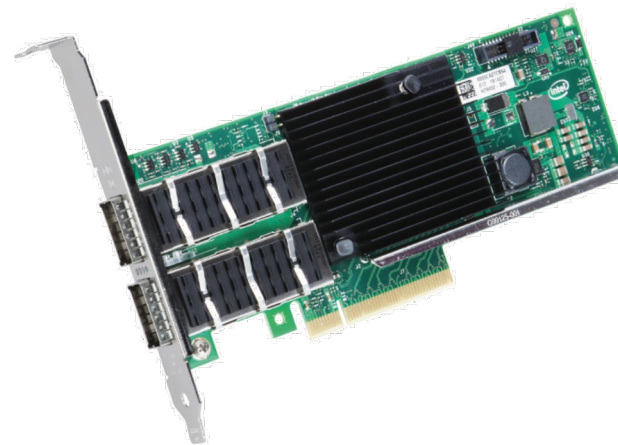


Alto rendimiento



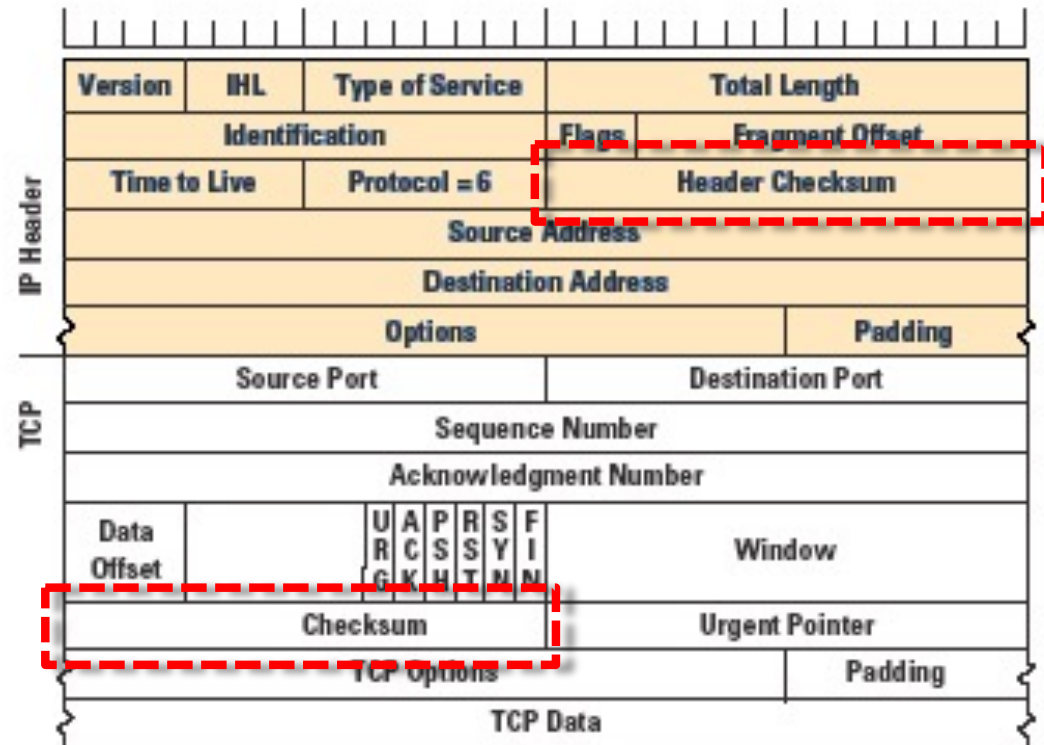
Tareas en la NIC

- 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



Checksum offload

- La NIC descarga del cálculo a la CPU
- En transmisión y recepción
- Checksum IP (v4 y v6), UDP y TCP

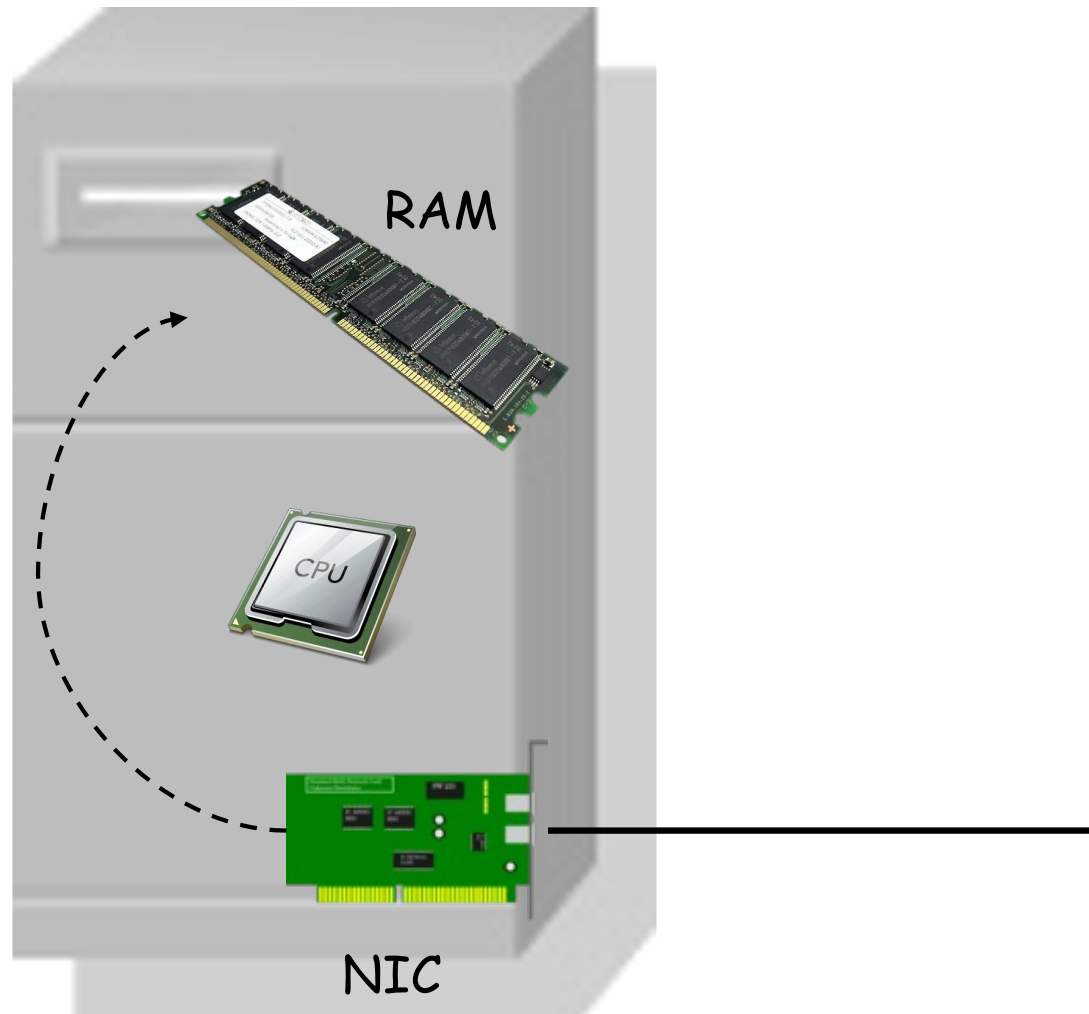


Integración en el bus

- 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
 - También puede hacer *polling* la NIC

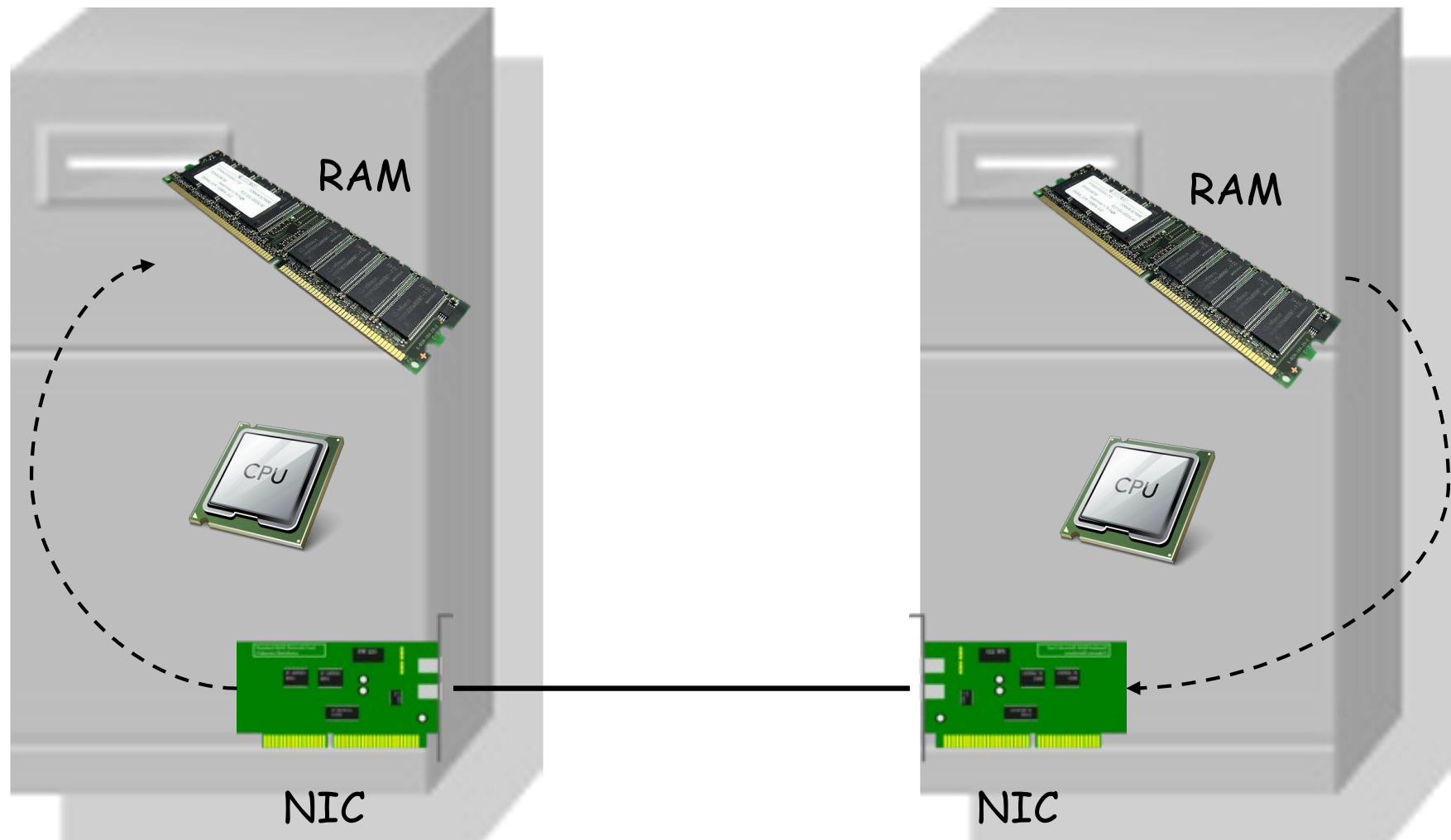
Integración en el bus

- DMA
 - *Direct Memory Access*
 - Transferencia desde la NIC a memoria sin requerir a la CPU



RDMA

- Remote Direct Memory Access
- Copias entre RAM de hosts diferentes sin involucrar a la CPU
- Latencia de pocos microsegundos



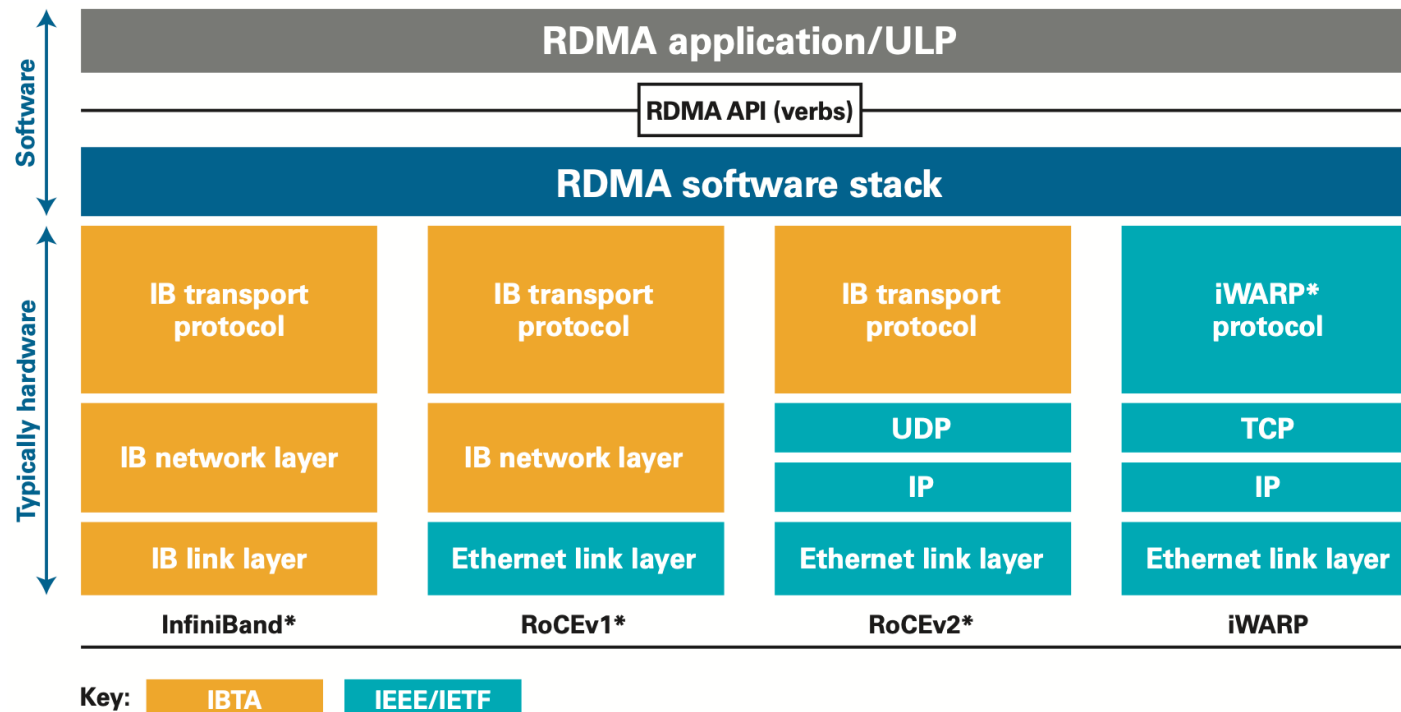
RDMA

iWARP

- RFCs 5040, 5041, 5044
- Sobre TCP o SCTP

RoCE

- RDMA over Converged Ethernet (DCB, Data Center Bridging)
- RoCE v1 sobre Ethernet, v2 sobre UDP
- RoCE v1 mecanismos de control de flujo y congestión de DCB
- RoCE v2 emplea control de congestión basado en ECN



Jumbo frames

- Tramas Ethernet con MTU superior a 1500bytes
- 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 (...)



Jumbo frames

- Tramas Ethernet con MTU superior a 1500bytes
- 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



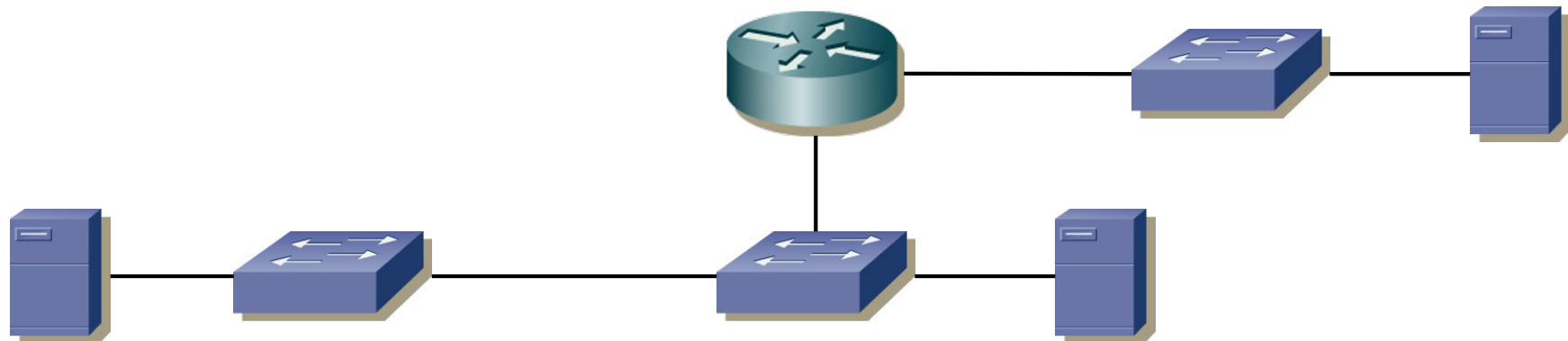
Jumbo frames

- 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?
 - (...)



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?
- ¿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



upna

Universidad Pública de Navarra
Nafarroako Unibertsitate Publikoa

Redes de Nueva Generación
Área de Ingeniería Telemática

Segmentación/Coalescencia

LRO

- *Large Receive Offload, Receive Segment Coalescing*
- La NIC une varios segmentos TCP en uno solo
- (...)

Sistema operativo

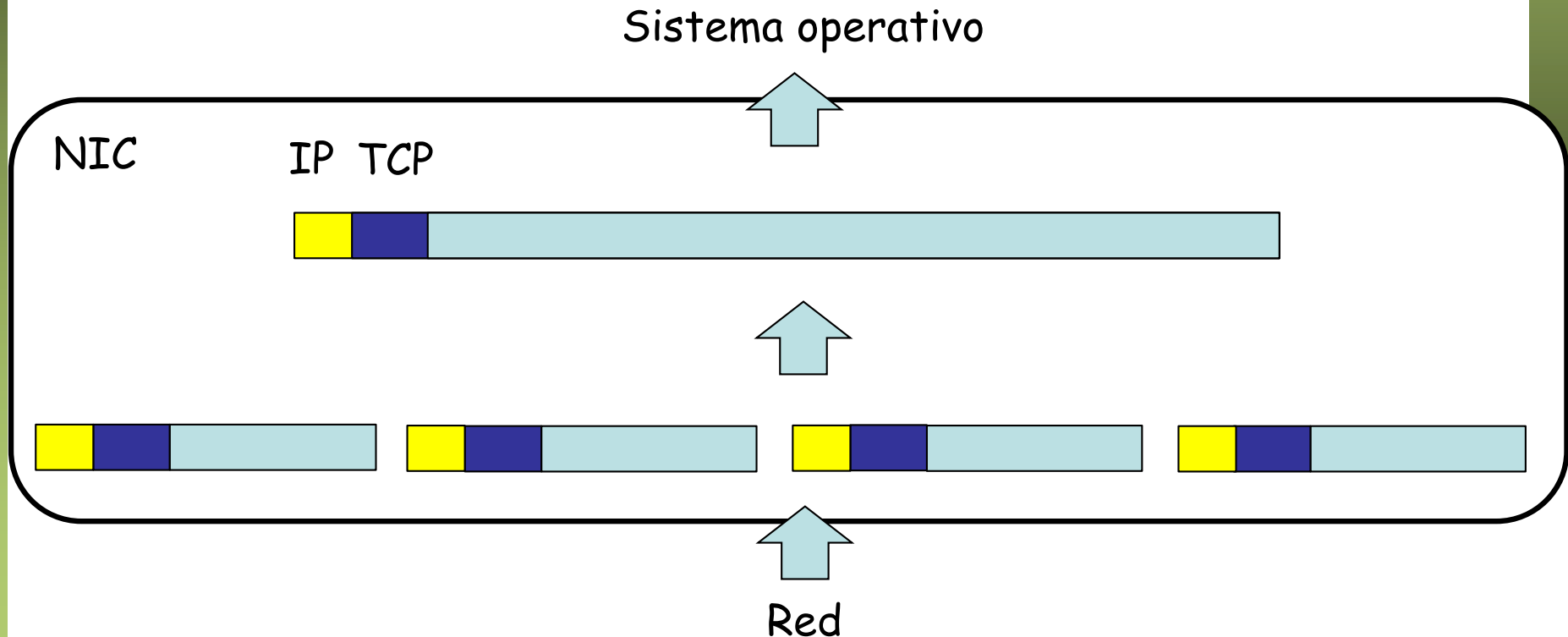
NIC



Red

LRO

- *Large Receive Offload, Receive Segment Coalescing*
- 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



LRO: Ejemplo

The screenshot shows a Wireshark interface with a packet list and a packet details pane. The packet list shows several packets, with packet 26 highlighted. The packet details pane shows the structure of packet 26, including Ethernet II, Internet Protocol Version 4, and Transmission Control Protocol. The 'Total Length' field in the IP section is highlighted in blue.

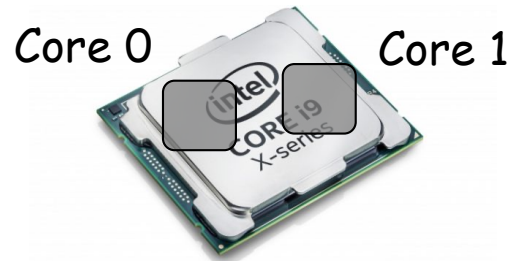
No.	Time	Source	Destination	tcp.len	frame.len	Info
24	1612366802.319898	192.168.1.3	192.168.1.2	158	224	Application Data
25	1612366802.319977	192.168.1.2	192.168.1.3	0	66	60260 → 443 [ACK] Seq=1077 Ack=29
26	1612366802.320132	192.168.1.3	192.168.1.2	2896	2962	Application Data, Application Data
27	1612366802.320133	192.168.1.3	192.168.1.2	1448	1514	Application Data [TCP segment of
28	1612366802.320191	192.168.1.3	192.168.1.2	944	1010	Application Data
29	1612366802.320193	192.168.1.3	192.168.1.2	2896	2962	Application Data, Application Data
30	1612366802.320194	192.168.1.2	192.168.1.3	0	66	60260 → 443 [ACK] Seq=1077 Ack=50
31	1612366802.320194	192.168.1.2	192.168.1.3	0	66	60260 → 443 [ACK] Seq=1077 Ack=70
32	1612366802.320196	192.168.1.3	192.168.1.2	2392	2458	Application Data, Application Data

▶ Frame 26: 2962 bytes on wire (23696 bits), 2962 bytes captured (23696 bits)
▶ Ethernet II, Src: Universa_2c:dc:32 (00:1e:37:2c:dc:32), Dst: Universa_2c:dc:6c (00:1e:37:2c:dc:6c)
▼ Internet Protocol Version 4, Src: 192.168.1.3, Dst: 192.168.1.2
 0100 = Version: 4
 0101 = Header Length: 20 bytes (5)
▶ Differentiated Services Field: 0x00 (DSCP: CS0, ECN: Not-ECT)
 Total Length: 2948
 Identification: 0x12b6 (4790)
▶ Flags: 0x40, Don't fragment
 Fragment Offset: 0
 Time to Live: 64
 Protocol: TCP (6)
 Header Checksum: 0x9968 [validation disabled]
 [Header checksum status: Unverified]
 Source Address: 192.168.1.3
 Destination Address: 192.168.1.2
▶ Transmission Control Protocol, Src Port: 443, Dst Port: 60260, Seq: 2978, Ack: 1077, Len: 2896
▼ Transport Layer Security
 ▼ TLSv1.3 Record Layer: Application Data Protocol: http-over-tls
 Opaque Type: Application Data (23)
 Version: TLS 1.2 (0x0303)
 Length: 1317

RSS

- *Receive Side Scaling*
- Multi-CPU o CPU multi-core

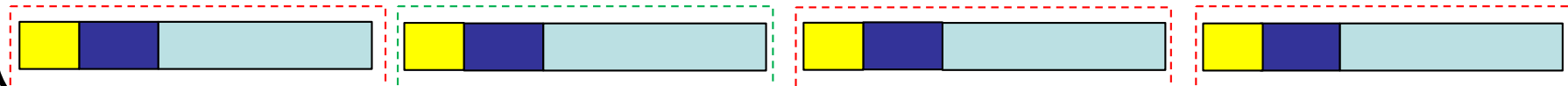
Sistema operativo



NIC

Flujo 1

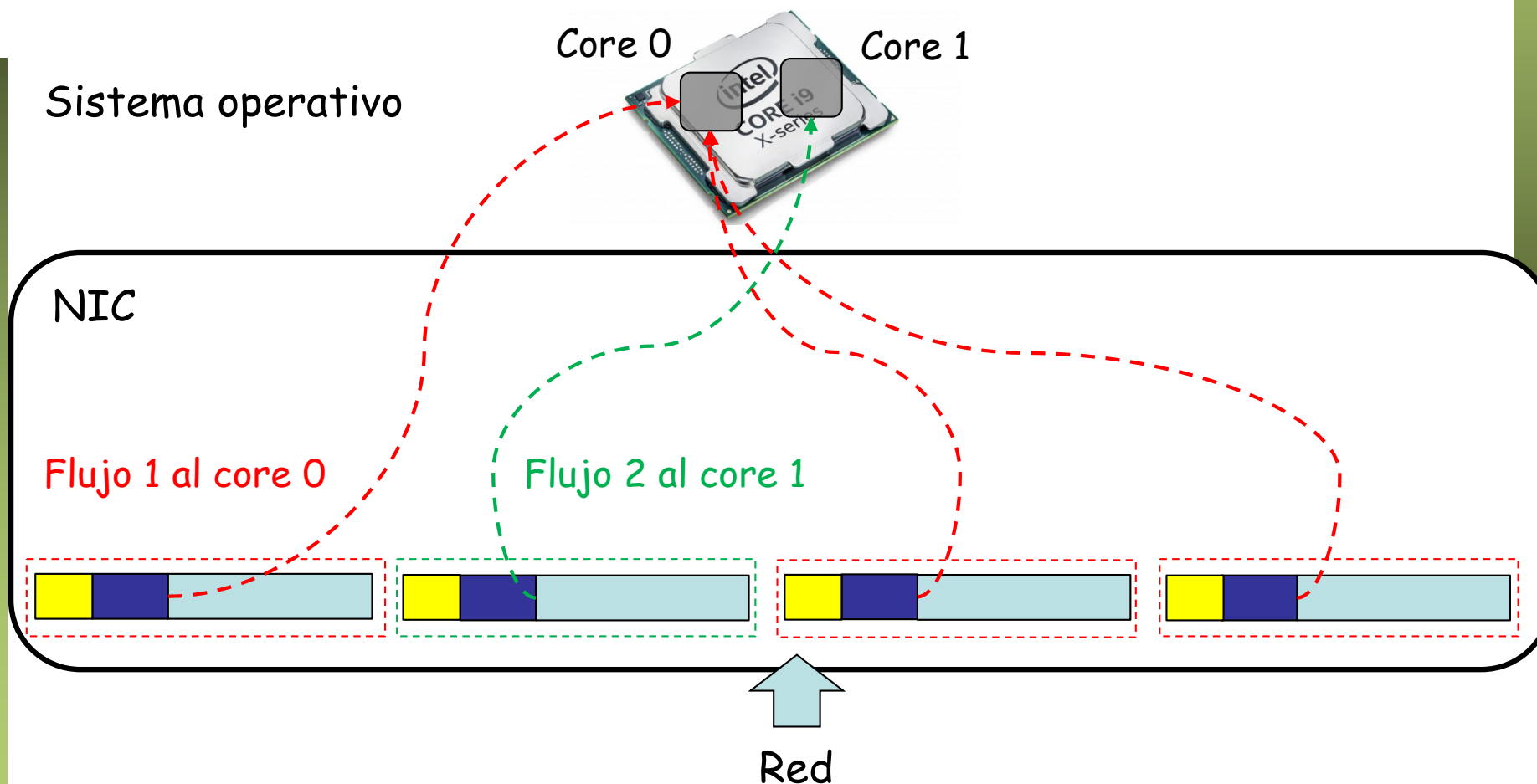
Flujo 2



Red

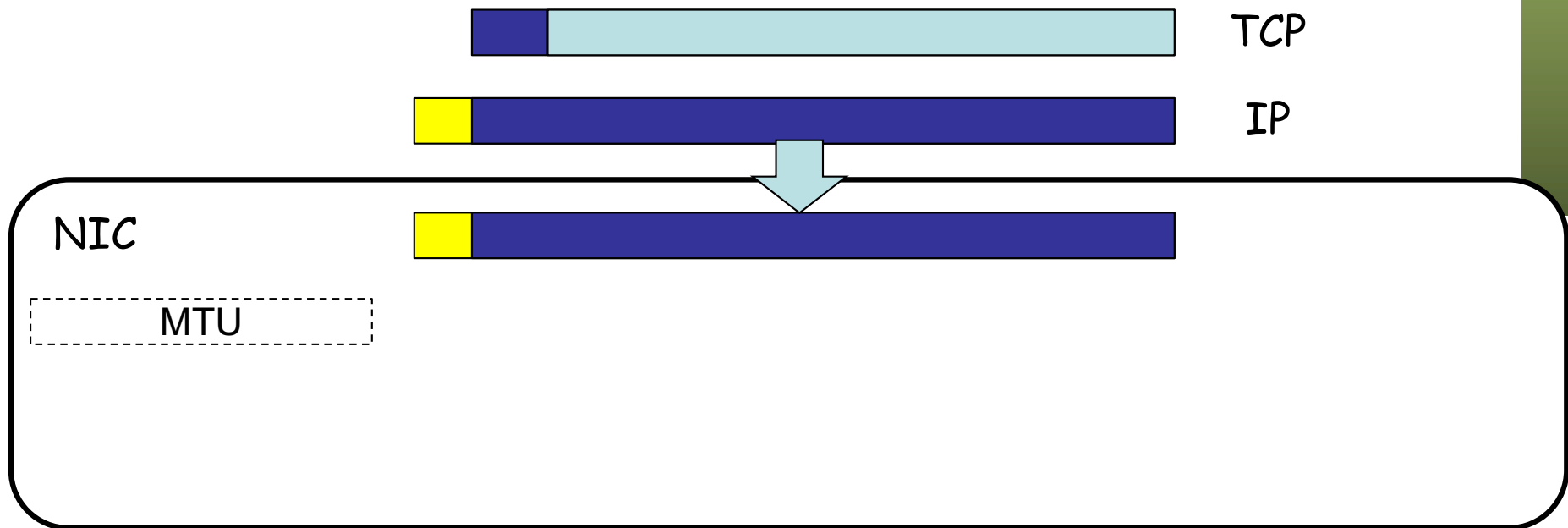
RSS

- 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

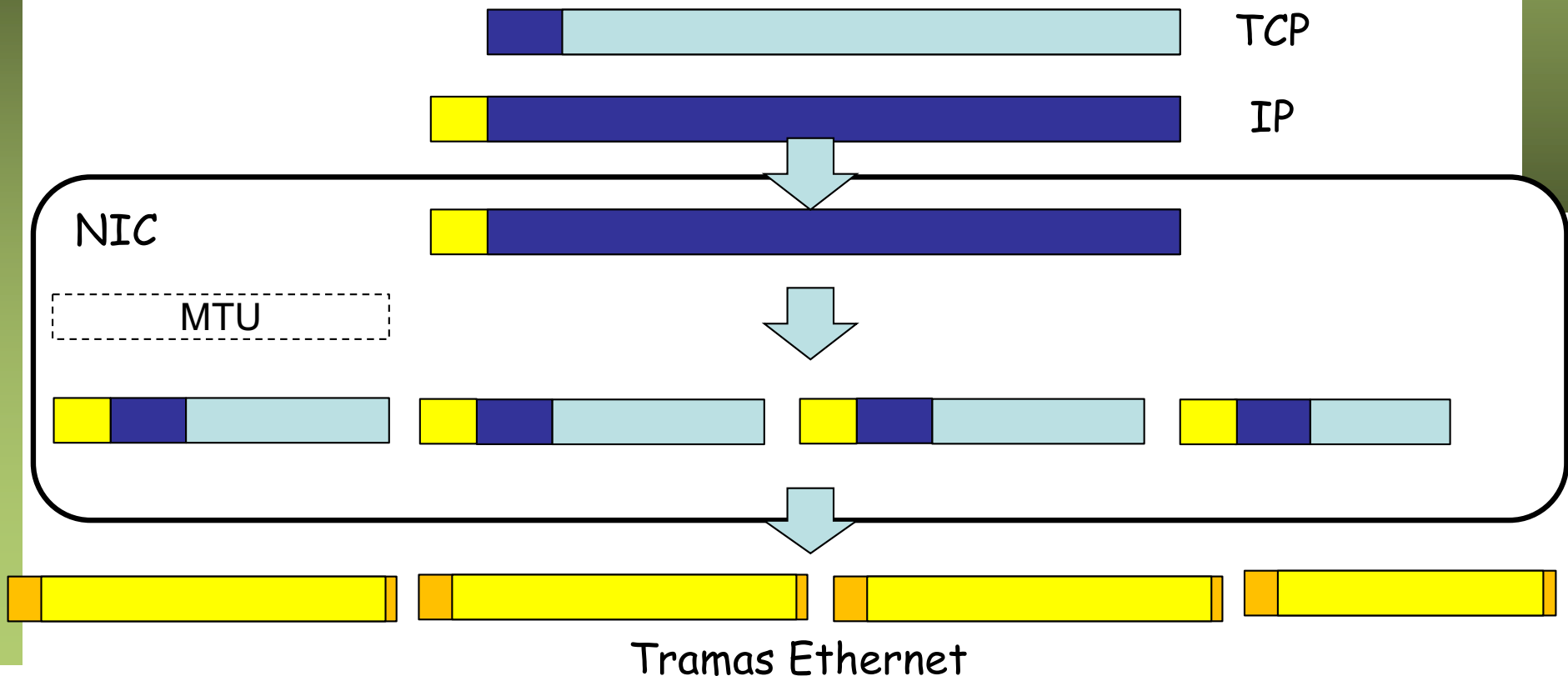
- *Large Segment Offload, TCP Segmentation Offload*
- Busca reducir carga de trabajo a la CPU en transmisión
- TCP entrega a la NIC paquetes más grandes que la MTU
- (...)



Tramas Ethernet

LSO

- 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



upna

Universidad Pública de Navarra
Nafarroako Unibertsitate Publikoa

Redes de Nueva Generación
Área de Ingeniería Telemática

Segmentación/Coalescencia

upna

Universidad Pública de Navarra
Nafarroako Unibertsitate Publikoa

Redes de Nueva Generación
Área de Ingeniería Telemática

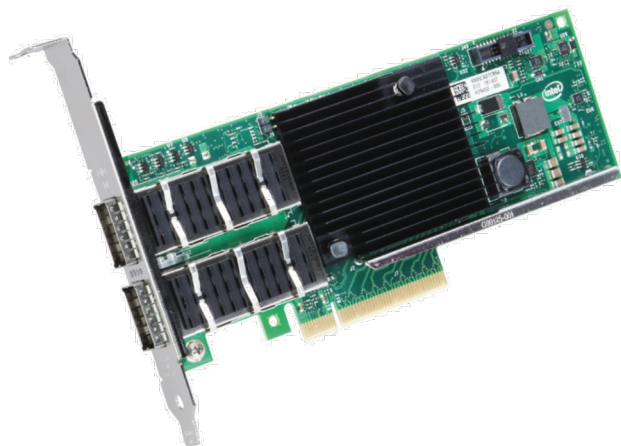


TOE

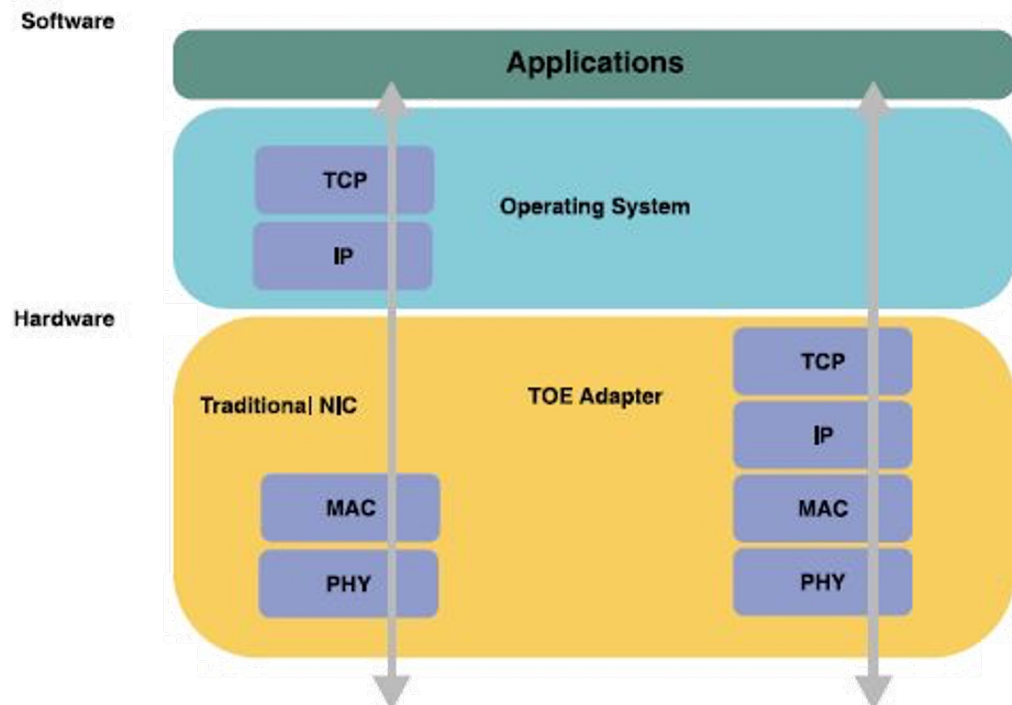


TOE

- *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

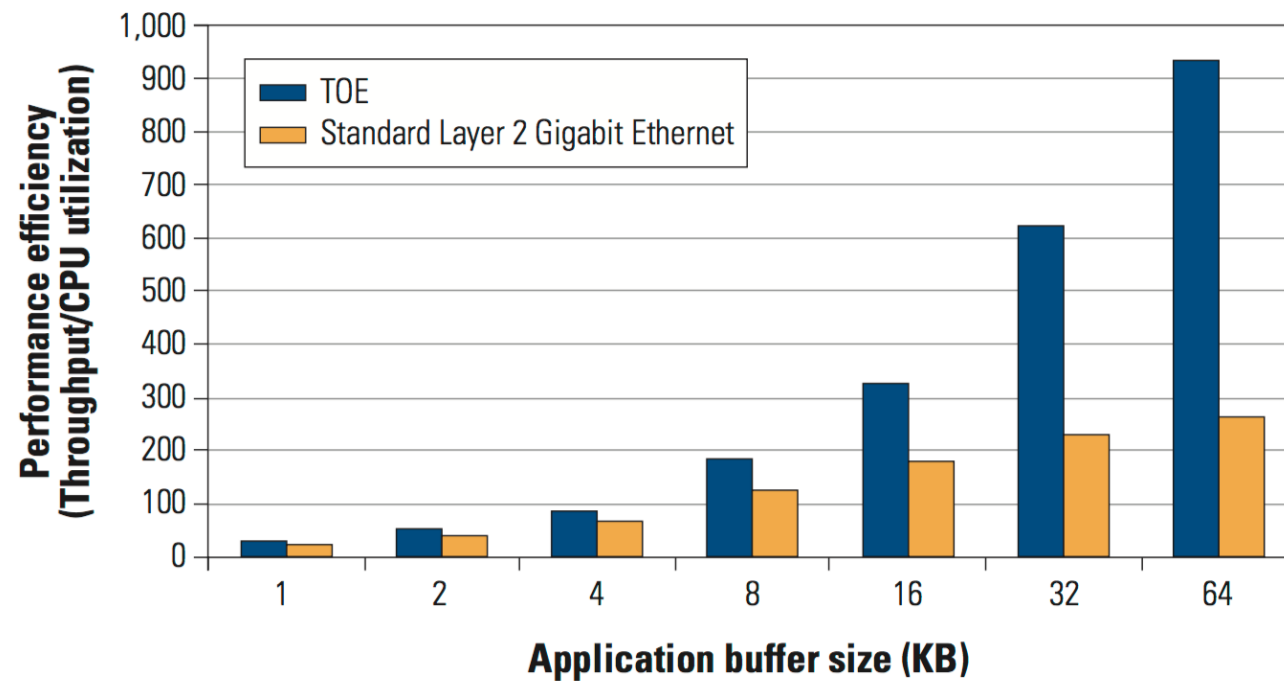


TCP/IP Offload Engine, TOE



TOE

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



<http://www.dell.com/downloads/global/power/ps3q06-20060132-Broadcom.pdf>

upna

Universidad Pública de Navarra
Nafarroako Unibertsitate Publikoa

Redes de Nueva Generación
Área de Ingeniería Telemática

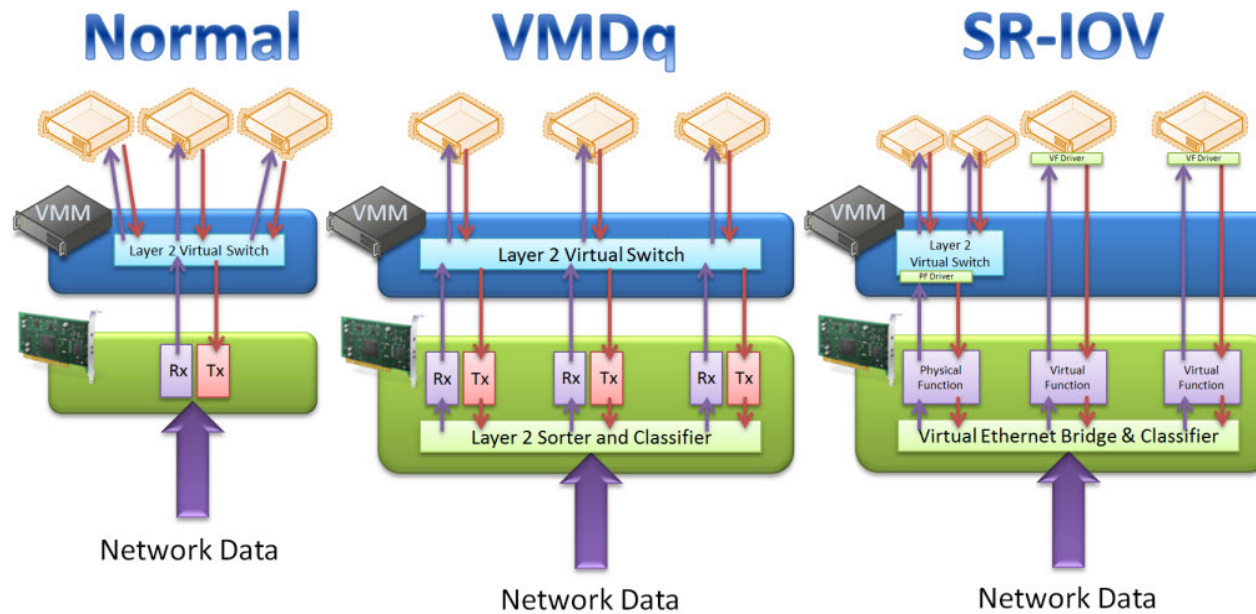


TOE



Otras funcionalidades

- VMDq, SR-IOV, etc, asociadas a la presencia de máquinas virtuales



upna

Universidad Pública de Navarra
Nafarroako Unibertsitate Publikoa

Redes de Nueva Generación
Área de Ingeniería Telemática



Alto rendimiento



upna

Universidad Pública de Navarra
Nafarroako Unibertsitate Publikoa

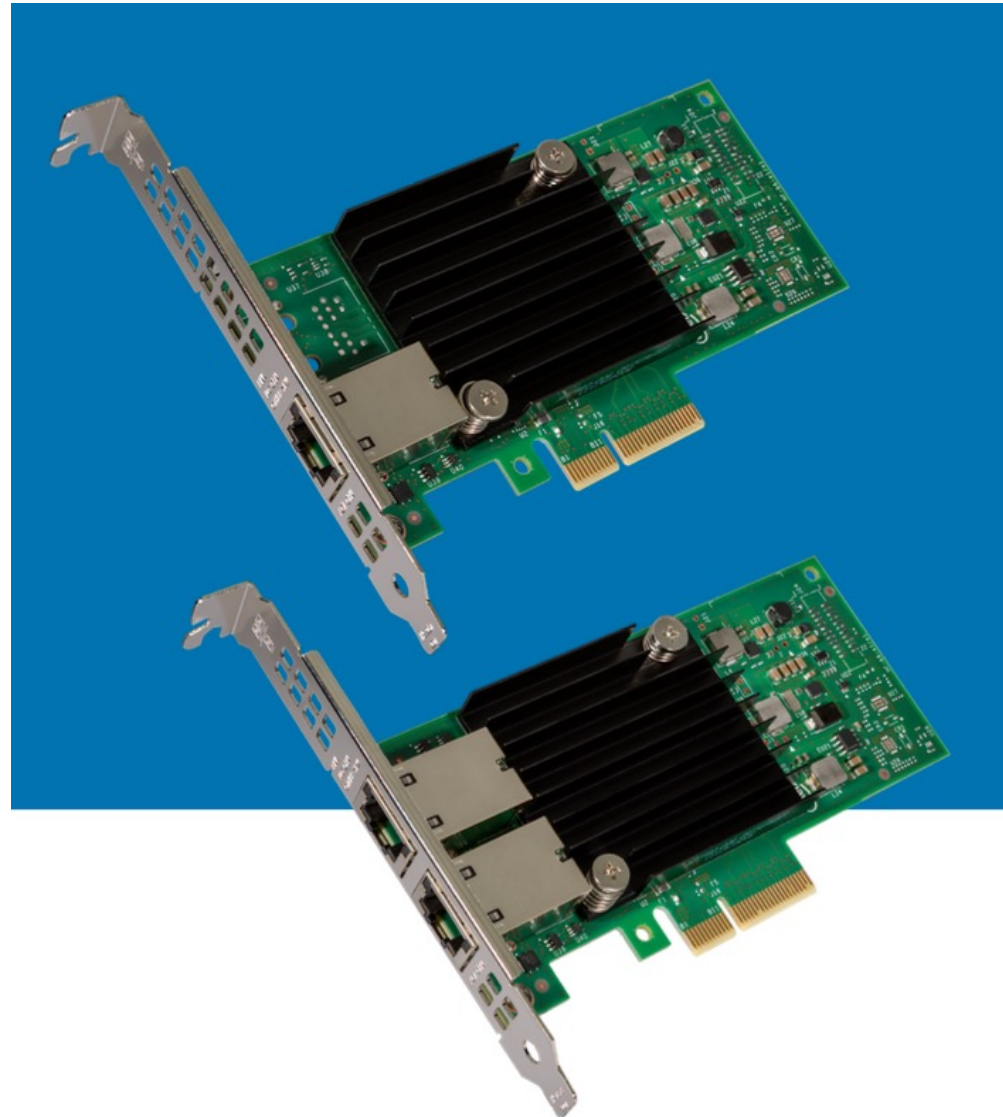
Redes de Nueva Generación
Área de Ingeniería Telemática

Ejemplos

Intel® Ethernet X550-T2

Key Features

- Backward compatible with existing 1000BASE-T networks
- Supports NBASE-T technology (2.5 and 5GbE over CAT5e)
- Standard CAT6a cabling with RJ45 connectors
- Low cost, low power, 10GbE performance for the entire data center
- Flexible I/O virtualization for port partitioning and quality of service (QoS) of up to 64 virtual ports
- Single-chip solution with integrated MAC + PHY
- PCIe 3.0 with up to 8.0GT/s



<https://www.intel.es/content/www/es/es/products/sku/88209/intel-ethernet-converged-network-adapter-x550t2/specifications.html>

<https://cdrdv2.intel.com/v1/dl/getcontent/333369>

Intel® Ethernet X550-T2

Features	Description
General	
RJ45 connections over CAT6A cabling	• Ensures compatibility with cable length up to 100 meters.
RoHS-compliant, lead-free technology	• Complies with the European Union (EU) directives to reduce the use of hazardous materials.
I/O Features for Multi-core Processor Servers	
MSI-X support	• DMA engine – Enhances data acceleration across the platform (network, chipset, processor), lowering CPU usage.
Low latency	• Based on the sensitivity of the incoming data, the adapter can bypass the automatic moderation of time intervals between interrupts.
Header splits and replication in receive	• Helps the software device driver focus on the relevant part of the packet without the need to parse it.
Multiple queues – 64 Tx and Rx per port	• Network packet handling without waiting or buffer overflow providing efficient packet prioritization.
Tx/Rx IP, SCTP, TCP, and UDP checksum offloading (IPv4, IPv6) capabilities	• Checksum and segmentation capability extended to a new standard packet type.
Tx TCP segmentation offload (IPv4, IPv6)	• Increased throughput and lower processor usage. • Compatible with large-send offload feature (in Microsoft Windows Server operating systems).
IPsec	• Offloads IPsec capability onto the adapter instead of software to significantly improve throughput and CPU usage.
Compatible with x4, x8 and x16 standard and Low-profile PCIe slots	• Enables each PCIe slot port to operate without interfering or competing with other PCIe slot port.
Receive Side Scaling for Windows Environment and Scalable I/O for Linux Environments (IPv4, IPv6 and TCP/ UDP)	• Enables the direction of the interrupts to the processor cores in order to improve CPU use rate.

Specifications	
General	
Connections	RJ45 copper
Cable Distance	10GBASE-T: 100 m using CAT6A, 55 m using CAT6 1000BASE-T: 100 m using CAT5e, CAT6 or CAT6A

Intel® Ethernet X550-T2

Virtualization Features

Multi-mode I/O virtualization operations	<ul style="list-style-type: none">• Supports two modes of operation of virtualized environments:<ul style="list-style-type: none">- Direct assignment of part of the port resources to different guest operating systems using the PCI SIG SR-IOV standard (also known as native mode or pass-through mode)- Central management of the networking resources by hypervisor (also known as software switch acceleration mode)• A hybrid model, where some of the VMs are assigned a dedicated share of the port and the rest are serviced by a hypervisor is also supported
VxLAN stateless offloads	<ul style="list-style-type: none">• A framework for overlaying virtualized layer 2 networks over layer 3 networks. VxLAN enables users to create a logical network for VMs across different networks.
NVGRE stateless offloads	<ul style="list-style-type: none">• Network Virtualization using Generic Routing Encapsulation. The encapsulation of an Ethernet layer 2 Frame in IP that enables the creation of virtualized layer 2 subnets that can span physical layer 3 IP networks.
Virtual Machine Device Queues (VMDq)	<ul style="list-style-type: none">• Offloads data sorting from the hypervisor to silicon, improving data throughput and CPU usage.• QoS feature for Tx data by providing round-robin servicing and preventing head-of-line blocking.• Sorting based on MAC addresses and VLAN tags.
64 Transmit (Tx) and receive (Rx) Queue pairs per port	<ul style="list-style-type: none">• Supports VMware NetQueue and Microsoft VMQ.• MAC/VLAN filtering for pool selection and either DCB or RSS for the queue in pool selection.
FPP – 64 VFs per port	<ul style="list-style-type: none">• VFs appear as Ethernet controllers in Linux operating systems that can be assigned to VMs, Kernel processes or teamed using the Linux bonding drivers.
Support for PCI-SIG SR-IOV specification	<ul style="list-style-type: none">• Up to 64 VFs per port.
IEEE 802.1Q VLAN support with VLAN tag insertion, Stripping and packet filtering for up to 4096 VLAN tags	<ul style="list-style-type: none">• Ability to create multiple VLAN segments.• Filtering packets belonging to certain VLANs.

Broadcom 5720 Quad-Port 1GbE

Features

- Quad-port GbE Network Daughter Card for Dell PowerEdge 12G rack servers
- Two x1 PCI Express® (PCIe™) v2.0 (5 GT/s)
- Energy Efficient Ethernet (EEE)
- Full line-rate performance across all ports
- Broad OS and hypervisor support
- iSCSI remote boot support
- Preboot eXecution Environment (PXE) support
- Support for VMware® NetQueue™ and Microsoft® VMQ
- Link aggregation and automatic load balancing
- Wake-on-LAN support
- MSI and MSI-X support
- IPv4 and IPv6 offloads
- Stateless offload
- TCP, UDP, and IP checksum
- Large Send Offload (LSO)
- TCP Segmentation Offload (TSO)
- Receive Side Scaling (RSS)
- Transmit Side Scaling (TSS)
- VLAN support with VLAN tagging
- Jumbo frame support for frames larger than 1500 bytes
- Precision Time Protocol (PTP)
- Broadcom Advanced Control Suite (BACS) management application and integration into Dell's embedded management framework (iDRAC7 and Lifecycle Controller)

OS Support

Microsoft
Linux

VMware
Citrix

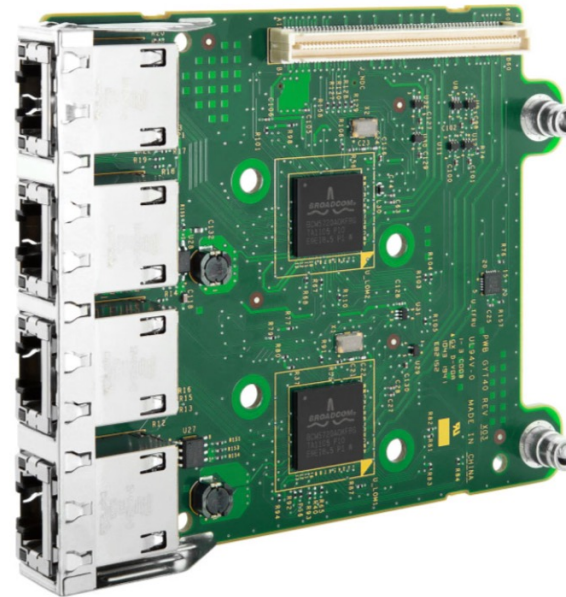
Windows Server® 2008, 2008 R2, all editions

Red Hat® Enterprise Linux (RHEL) 5.7/5.8, 6.1/6.2

Novell® SUSE® Linux Enterprise Server (SLES) 10 SP4, 11 SP2

vSphere™ 4.1 and 5.0

XenServer 6.0



IEEE 802.3x—Flow control

IEEE 802.3 (Clause 30)—Statistics for SNMP MIB II, Ethernet-like MIB, and Ethernet MIB

IPv4 and IPv6 offload

Teaming support

Intel Ethernet Controller XL710-BM1/BM2

Performance

40Gb throughput	Wire-rate down to 128 bytes
10Gb throughput	Wire-rate down to 64 bytes
Standard Linux Stack Latency	~8 μ s

Additional Features

Enhanced Transmission Selection (draft IEEE 802.1az)
Priority Flow Control (draft IEEE 802.1Qbb)
Data Center Bridging (DCB/DCB-X) Support; up to eight traffic classes
Jumbo Frame Support—Up to 9.5 KB (9728 Bytes)
VLAN Support

TCP/IP/L2 Features

Receive Side Scaling (RSS) for TCP and UDP traffic
Large Send Off-load (LSO) / Generic Send Off-load (GSO) including encapsulated traffic
TCP/UDP/IP/SCTP Checksum Off-load including encapsulated traffic
IPv4, IPv6

Virtualization Interface Features

Features	Implementation
Emulated Support	Driver Optimizations and VMDQ enablement
Direct Assignment Support	PF and VF assignment with SR-IOV
Virtual Bridging Support	VEPA/802.1Qbg
Virtual Functions	Up to 128 per device
Network Virtualization	VxLAN, MACinUDP, NVGRE, IPinGRE



upna

Universidad Pública de Navarra
Nafarroako Unibertsitate Publikoa

Redes de Nueva Generación
Área de Ingeniería Telemática

NICs Ethernet para servidor

upna

Universidad Pública de Navarra
Nafarroako Unibertsitate Publikoa

Redes de Nueva Generación
Área de Ingeniería Telemática

NICs Ethernet para servidor