# Redes de Baja Latencia

**Myrinet**

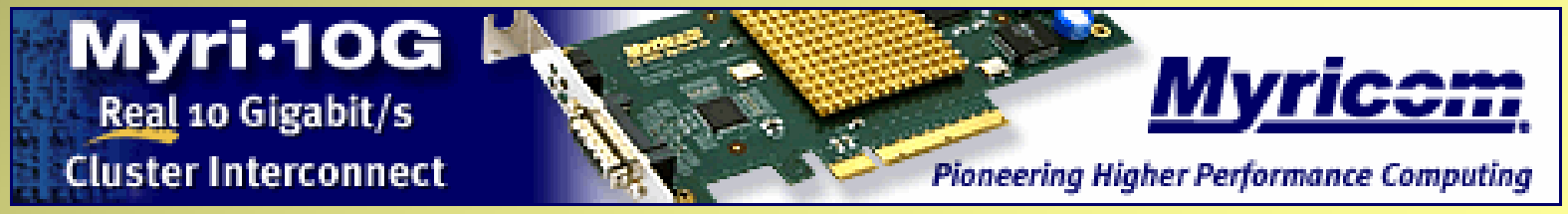
**Infiniband**

**Quadrics**

# Finalidad de las Redes de Baja Latencia

- Incremento de prestaciones.
- Procesamiento a nivel de cluster.

# Myrinet


Myri·10G
Real 10 Gigabit/s
Cluster Interconnect
Myricom
Pioneering Higher Performance Computing

# Myrinet

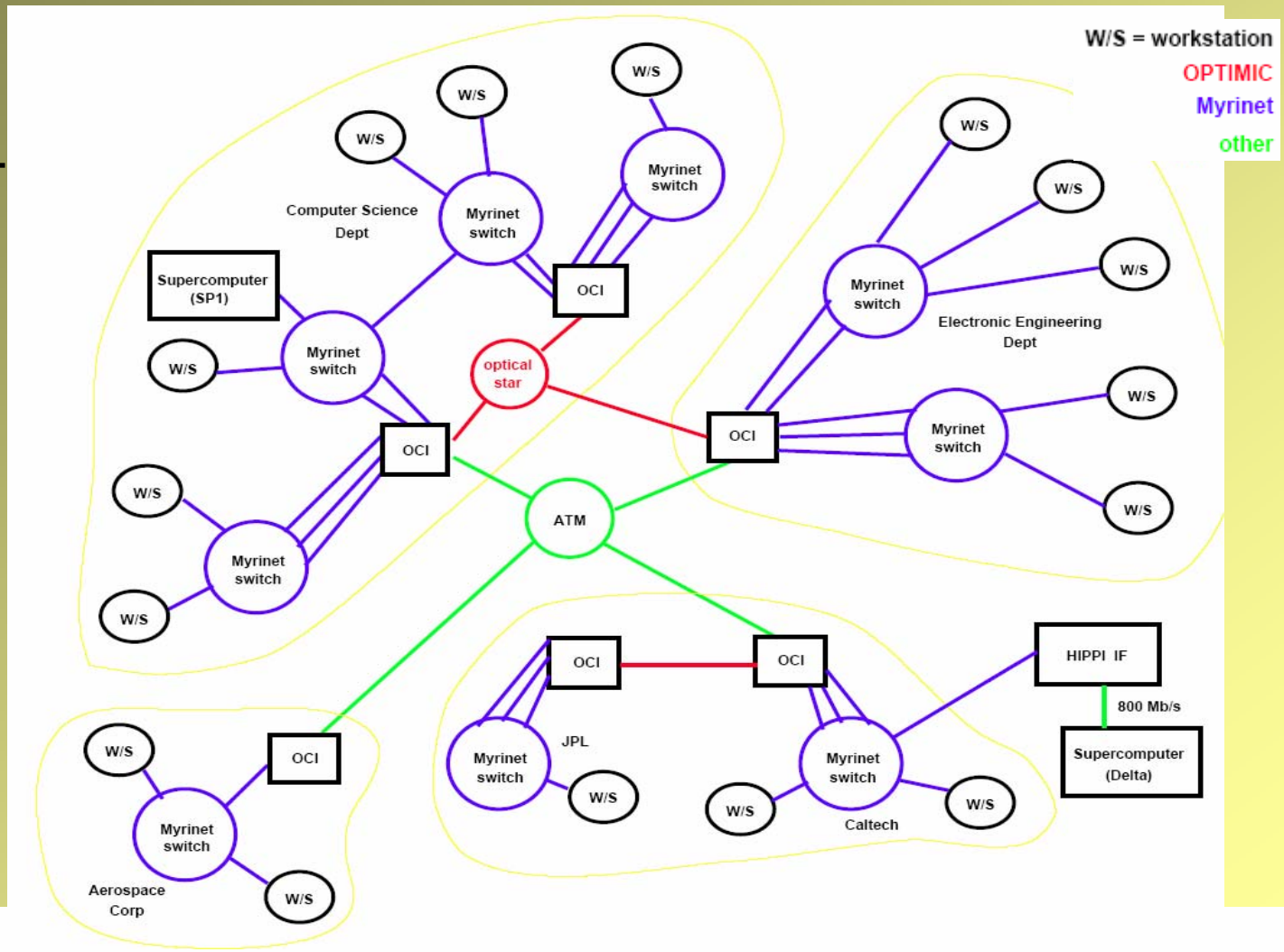
- Desarrollado por Myricom 1994
- Gama de productos
  - Myrinet-2000.
  - Myri-10G.
- www.myri.com

# Características Principales de Myrinet

- Baja Latencia: 2.3µs–3.2µs.
- Escalabilidad-Autodetección de Topología.
- Monitorización de cada enlace.
- Open-sourced.
- Tratamiento Interbloqueo (Deadlock).
- Redes virtuales VLAN (Myri-10G)

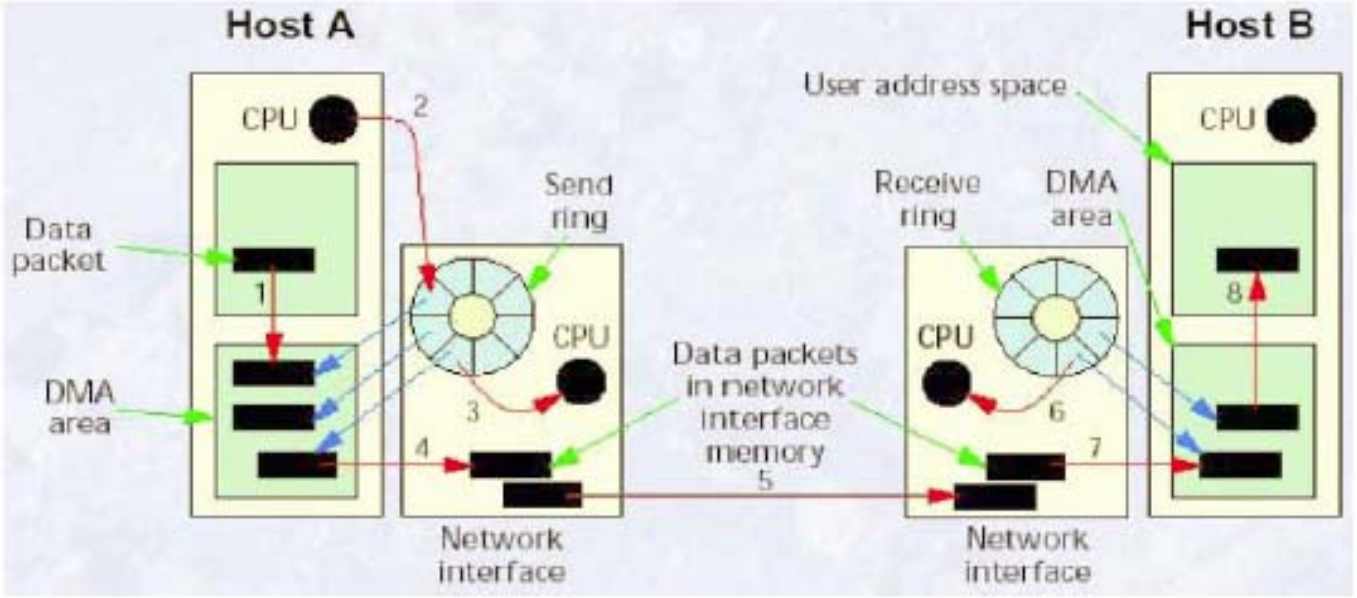
# Ejemplo de Myrinet

Myrinet en SUPERNET




Myricom

# Protocolo red Myrinet

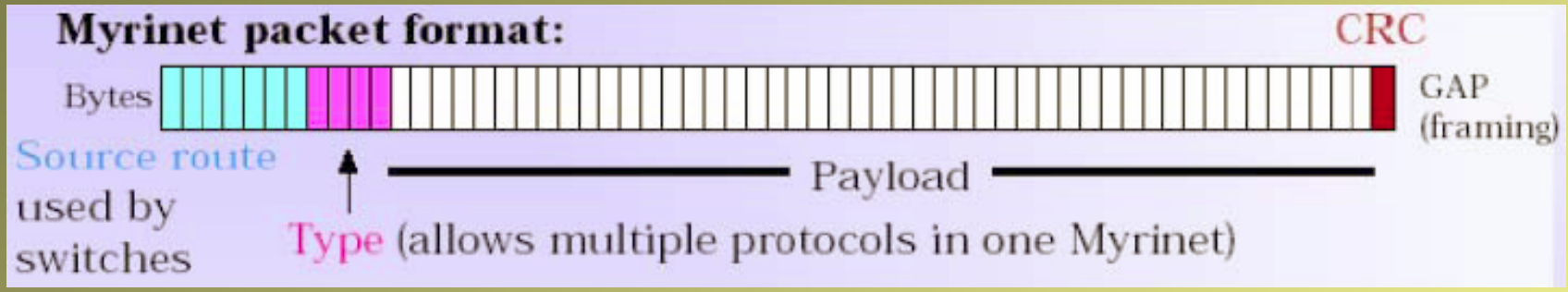



(1) Host A copia datos de usuario en el área DMA del CIR
(2) Host A escribe un descriptor de paquete en el anillo de envío
(3) El interfaz de red lee el descriptor y copia el paquete a su memoria usando DMA (4)
(5) El paquete se transfiere por la red
(6) El interfaz de red receptor lee el anillo de recepción para encontrar un buffer libre en el área DMA
(7) Se copia el paquete recibido en esa área DMA
(8) Host B copia el paquete desde el área DMA a la memoria de usuario

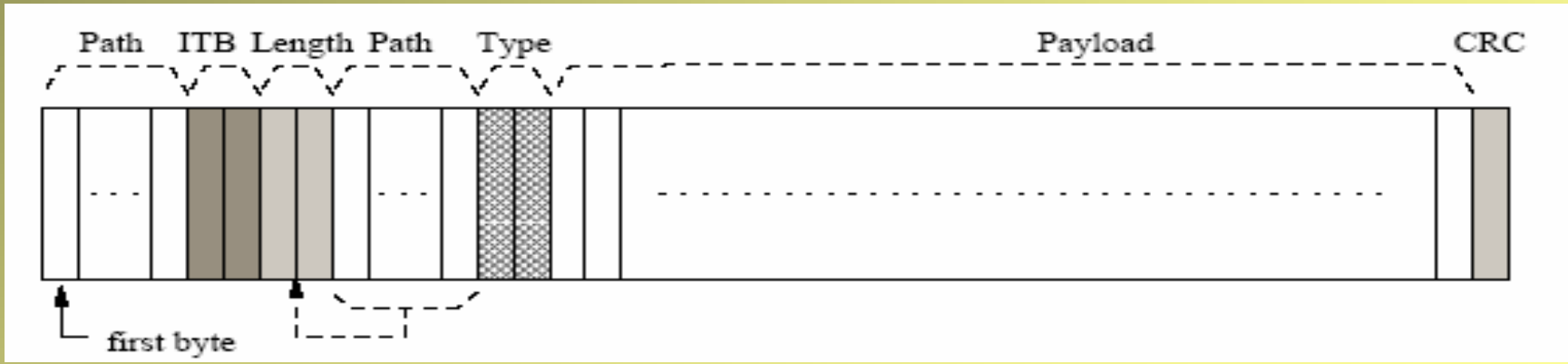
# Problemas del protocolo

- **Todos los mensajes pasan por un área DMA: Esto introduce una copia de memoria de usuario a dicha área.**
- **Falta de protección: Los usuarios acceden directamente al interfaz de red, pudiendo modificar datos de otros usuarios**
- **Uso de polling para control de transferencia: Polling frecuente es costoso mientras que polling poco frecuente es lento en su respuesta. La alternativa de las interrupciones es costosa**
- **Protocolo no es fiable: Si el emisor es más rápido que el receptor, se perderán paquetes**
- **Protocolo sólo soporte mensajes punto-a-punto: No hay soporte directo de comunicaciones colectivas.**

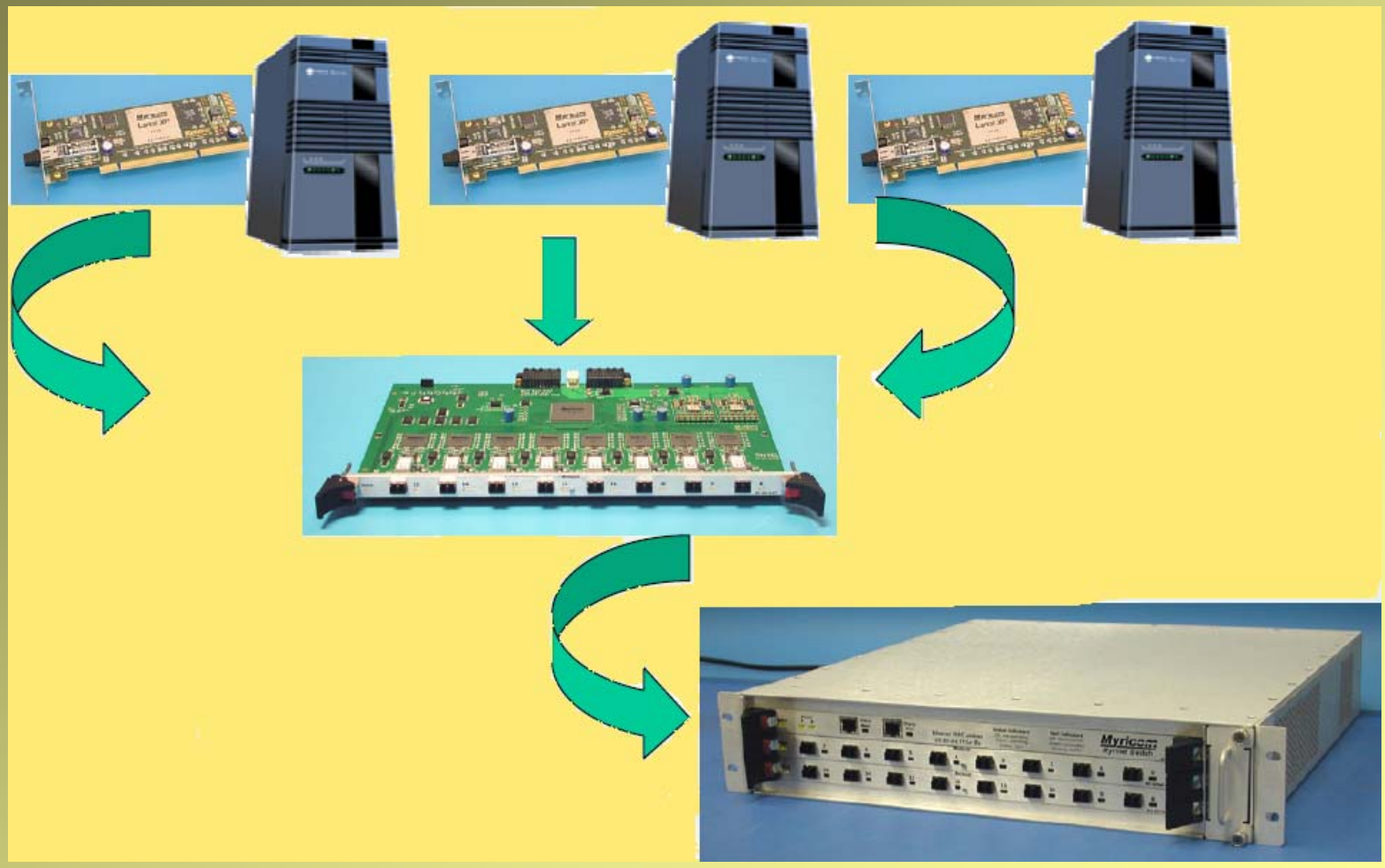
# Paquetes Myrinet




- Inlusión de ITB



# Arquitectura de red Myrinet



# Myrinet-2000 vs Myri-10G

| Dispositivos/Características | Myrinet-2000 | Myri-10G |
|---|---|---|
| Envío Full-duplex de las MAC y enlaces de la red | 2+2 Gigabits/s | 10+10 Gigabits/s |
| Enlaces de los cables | Conector LC duplex para fibras de hasta 200m | 10-Gigabit Ethernet cables de cobre o fibra |
| NIC slot | Simple y doble puerto de la PCI-X | Puerto simple PCI-Express, soportadando los protocolos 10G Myrinet y 10G Ethernet |
| Switches | Basados en 16 y 32 puertos. | Basados en 16 puertos. |
| Switch networks | Se tienen hasta 256 puertos como máximo en un único armario de la red. Se llega hasta diez mil combinando componentes. | Se tienen hasta 128 puertos como máximo en un único armario de la red. Se llega hasta diez mil combinando componentes. |
| Opera con otros | Gigabit Ethernet | 10-Gigabit Ethernet |
| Soporte software | Myrinet Express (MX-2G) o GM-2 | Myrinet Express (MX-10G) |
| Latencia MX o MPI | 2.6µs (D-card NICs) 3.2µs (E- or F-card NICs) | 2µs |
| Ratio de envío unidireccional con MX | 247 MBytes/s (MAC con un puerto) 495 MBytes/s (MAC de 2 puertos) | 1.2 GBytes/s |
| Ratio de envío TCP/IP (Emulando Ethernet con MX) | 1.98 Gbits/s (MAC con un puerto) 3.95 Gbits/s (MAC de 2 puertos) | 9.6 Gbits/s |

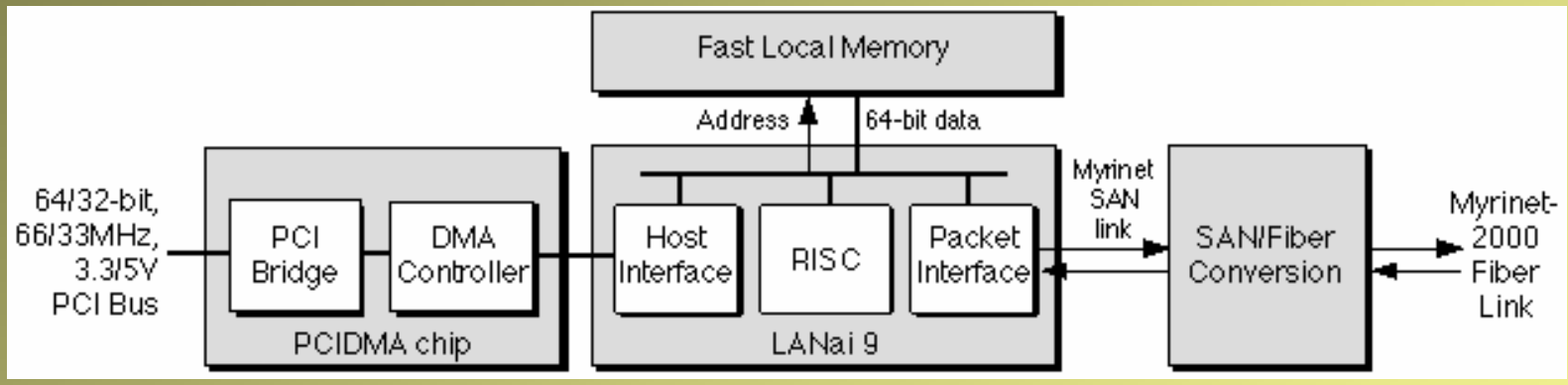
# Middleware

- Myrinet 2000

| Firmware/ Driver/ API | Middleware | | | | | |
|---|---|---|---|---|---|---|
| | MPI-1 | VIA | PVM | Sockets | DAPL | ClusterTools |
| MX (MX-2G) | MPICH-MX | n/a | n/a | Sockets-MX | En desarrollo | En desarrollo |
| GM | MPICH-GM | VI-GM | PVM-GM | Sockets-GM | DAPL-GM | ClusterTools |

# Tarjetas de Red Myrinet

- Arquitectura básica




Myricom

# Switch o Conmutadores

- Modularidad

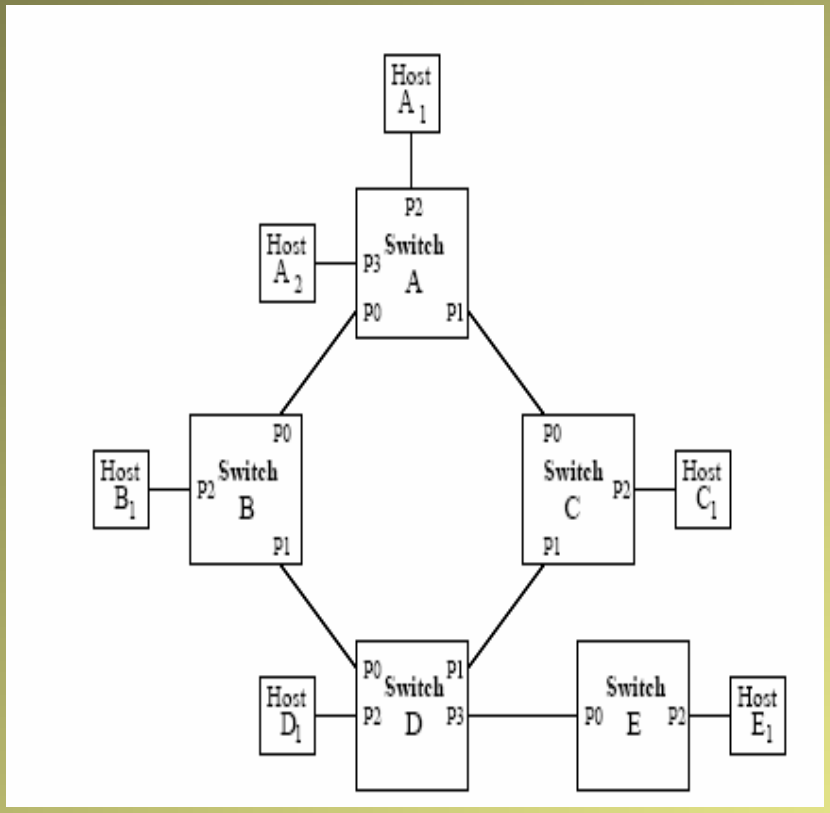

Myricom

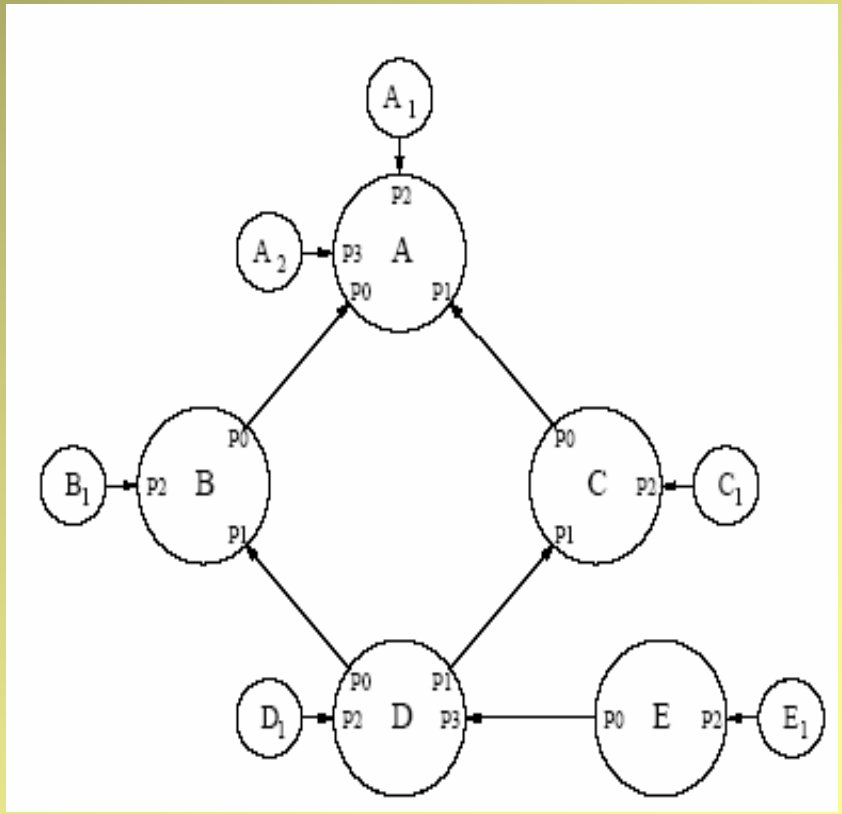
# Protocolos de Ruteo

**Típico protocolos sobre Infiniband**

- **Ruteo Wormhole**
- **Protocolo Up/Down**
- **Ruteo Fuente (source)**

# Ruteo: Protocolo Up/Down




Myricom

# Autodetección

- **Mapper: Mecanismo de exploración de red.**
- **No se considera la última detección.**
- **Se envían paquetes que se han de reenviar por los host.**
- **Los conmutadores no tienen capacidad de procesar paquetes. El maper indica el camino de ida y vuelta.**

# Máquinas con Myrinet

- MareNostrum.
- MACH-5.
- Big Red.
- Blade Center.
- HPC.
- Tungsten.
- CITerra.

# Infiniband

# Inicio de Infiniband

- Se inició en 1999
- Fusión de los proyectos Future I/O y Next Generation I/O
- www.infinibandta.org
- Inicialmente desarrollado por:









# Miembros actuales

























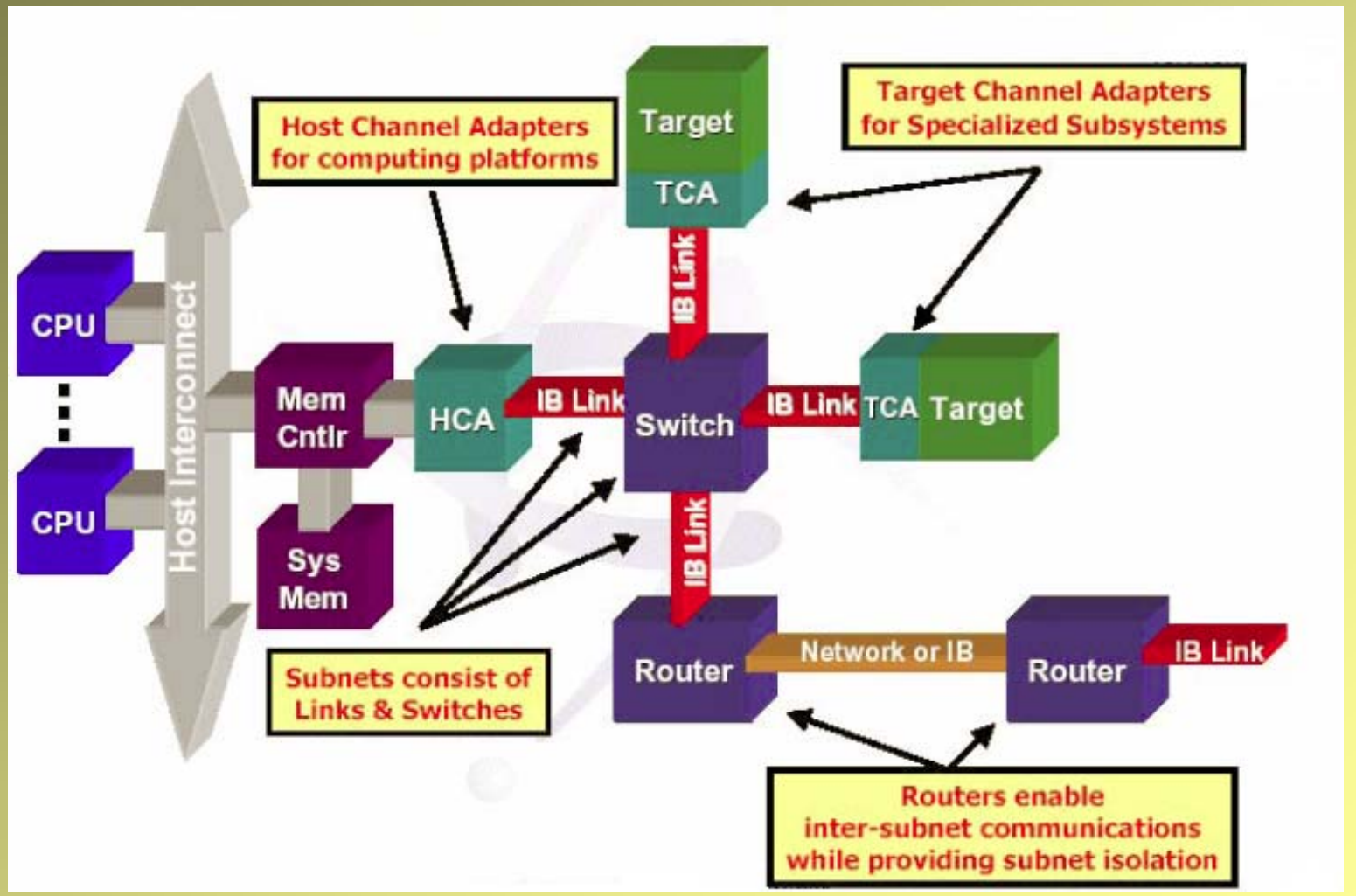
# Nacimiento de Infiniband

- Se ideó como un sustituto del bus PCI.
  - Alta complejidad para uso a nivel local.
- Define un nuevo estándar de interconexión entre sistemas.
- Sustituye el bus E/S tradicional por una red de conmutadores basados en canales, que interconecta unidades de procesamiento con dispositivos de E/S.

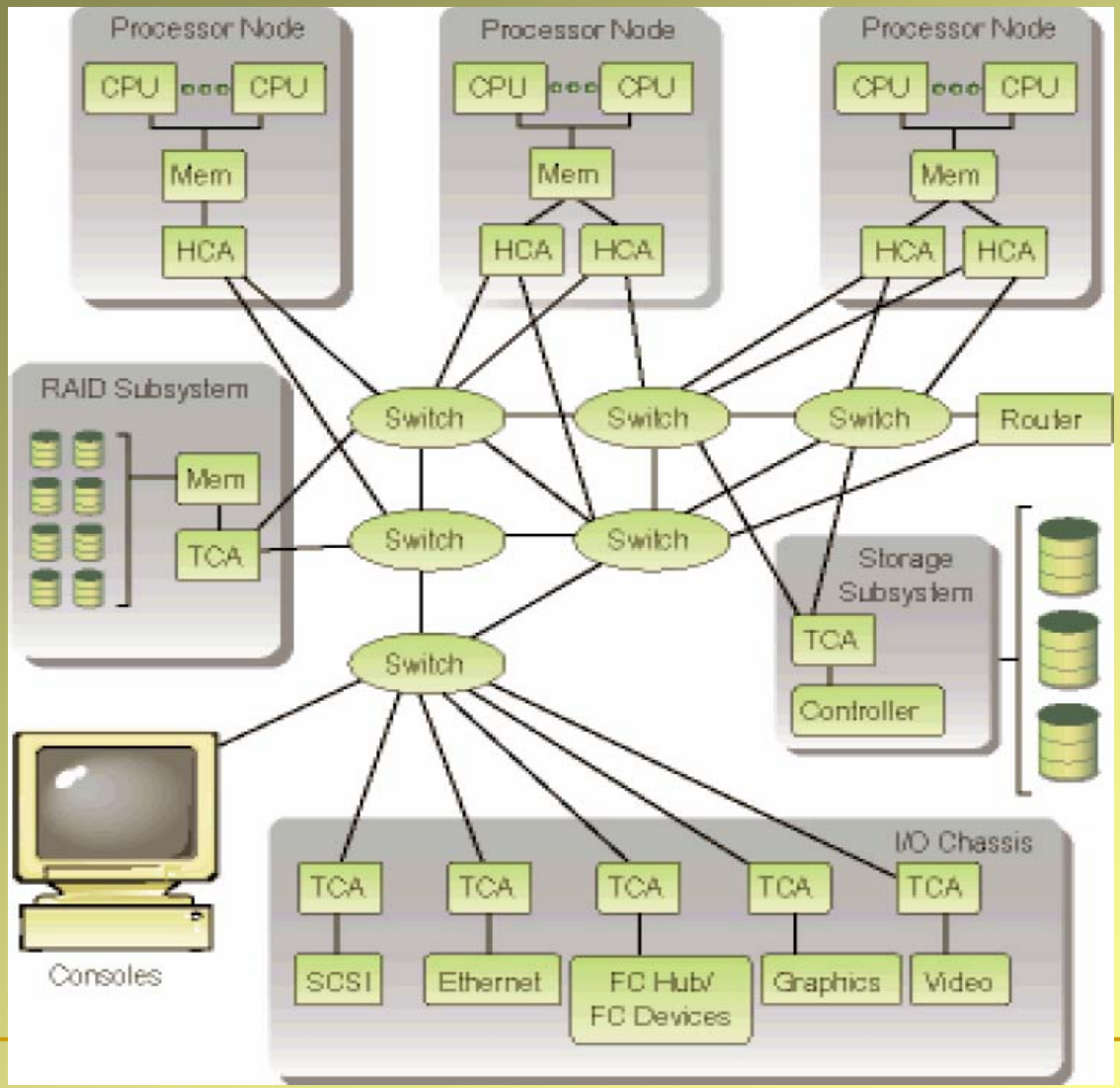
# Características Principales de Infiniband

- Baja Latencia.
- Escalabilidad.
- Open-sourced.
- Libre de Interbloqueo (Deadlock).
- Permite redes virtuales (VLAN).
- Técnica de conmutación Virtual Cut-Through.
- Tolerancia Fallos.

# Aquitectura de una red Infiniband



# Aquitectura de una red Infiniband



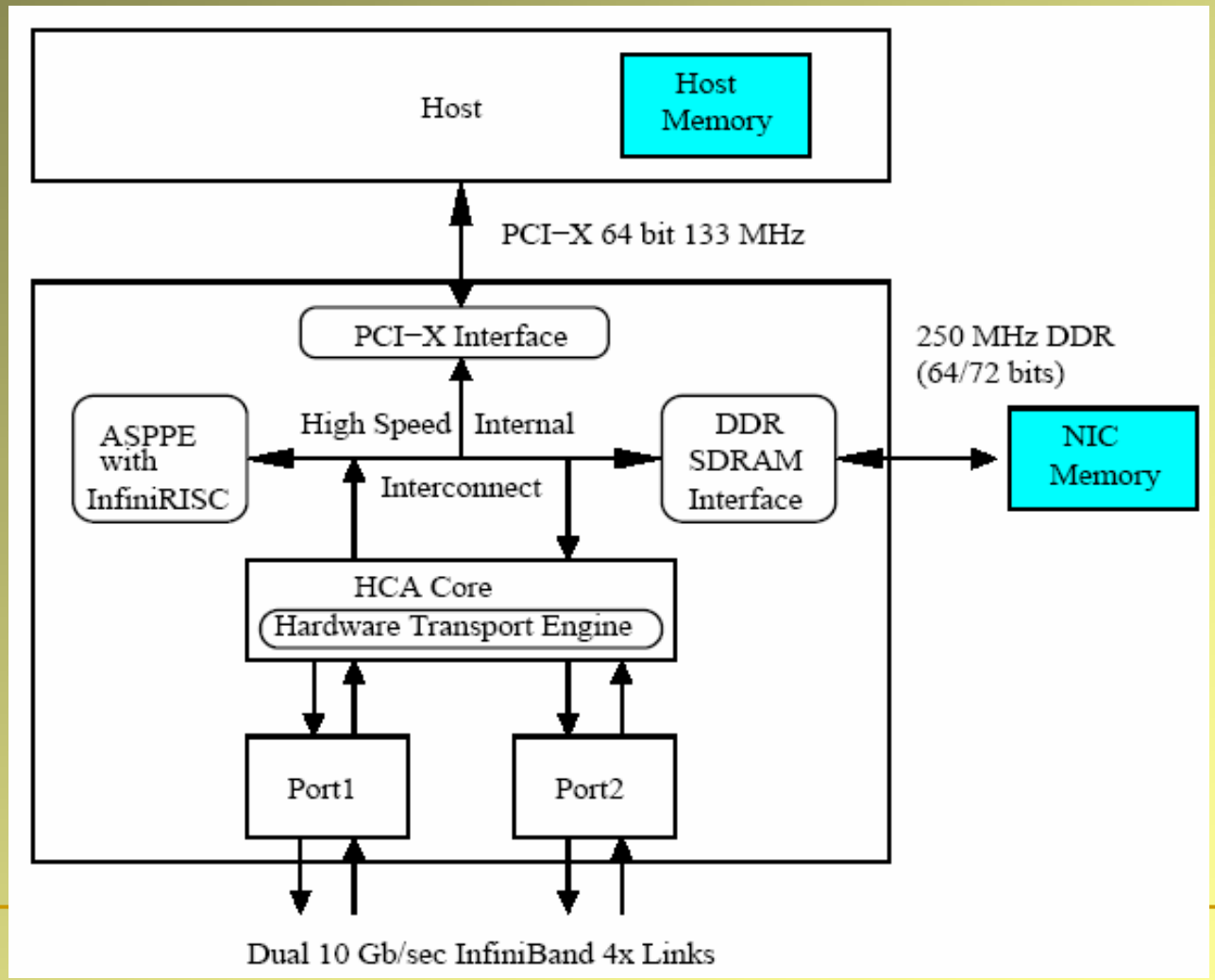
# Componentes hardware de Infiniband: **TCA**

- TCA: Target Channel Adapter, adaptador del canal destino.
  - Realiza las conversiones necesarias de la información para que sean comprensibles por el dispositivo E/S.
  - Incluyen un controlador específico para el I/O.
  - Es necesario tener al menos un TCA por dispositivo.

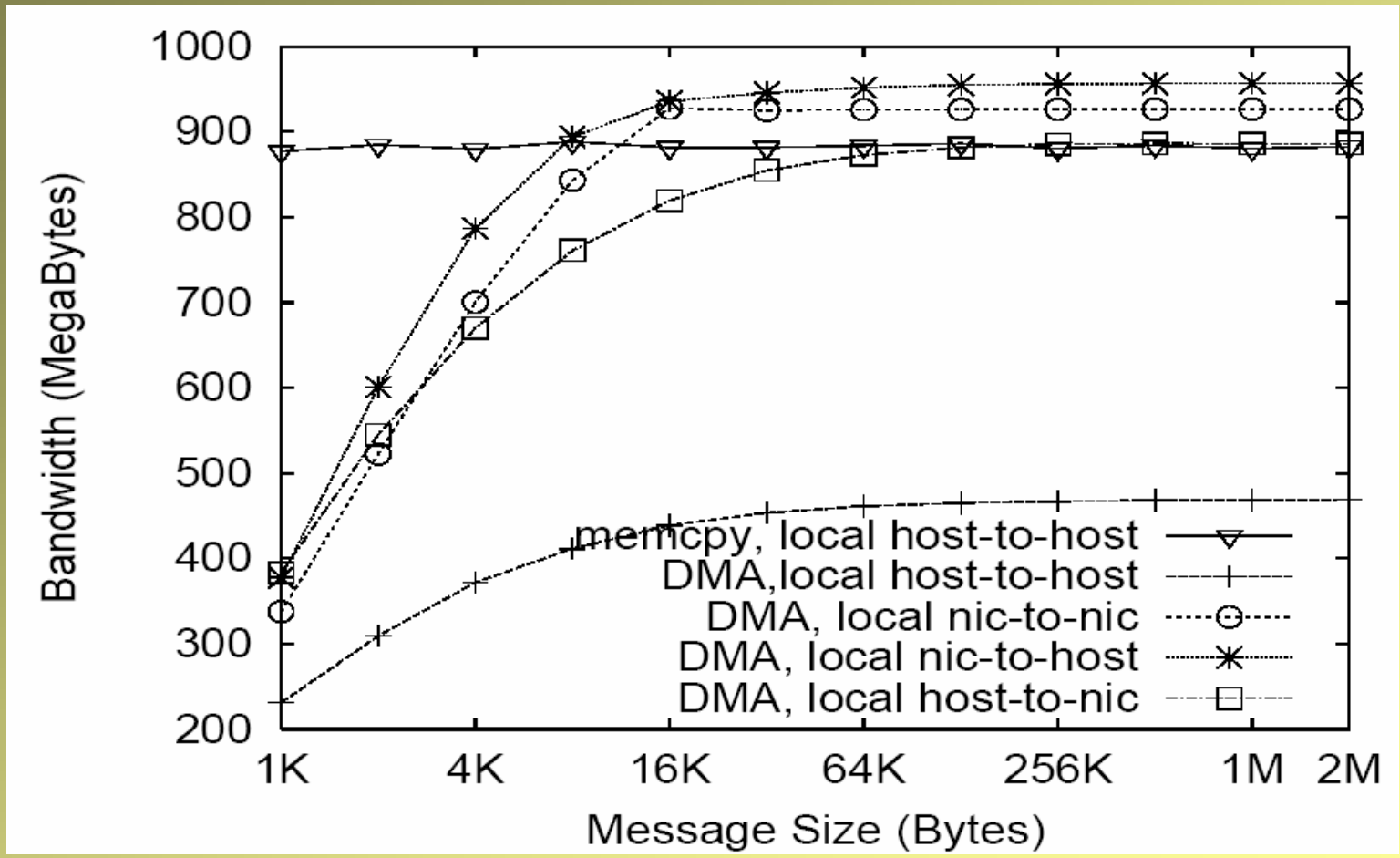
# Componentes hardware de Infiniband: **HCA**

- HCA: Host Channel Adapter, adaptador del canal al host.
  - Se encarga del tránsito de la información entre los dispositivos I/O y la memoria.
  - Adapta la información al formato de la red Infiniband.
  - Descarga de trabajo al procesador gestionando independientemente la E/S a través del DMA.
- Existen dos versiones DDR y SDR.

# Arquitectura e interfase de la HCA MT23108. Memoria DDR.



# Velocidad de transmisión de la HCA MT23108 con DDR.



# Paquetes Infiniband

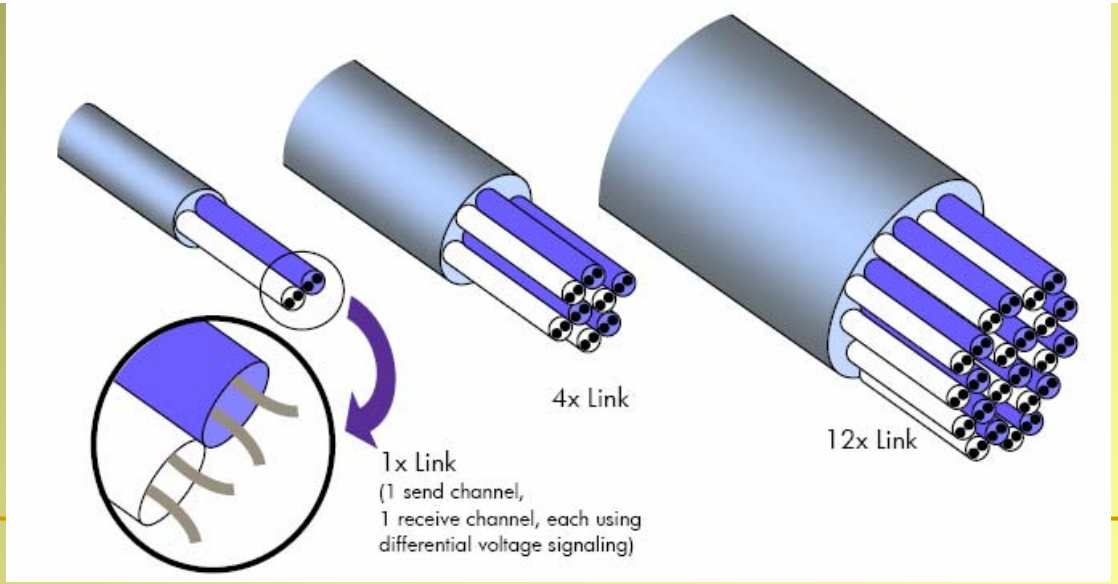

- Start-of-Frame [SOF]
- End-of-Frame [EOF]
- Header: se compone de una parte para la capa de ruteo y otra para el transporte.
- Payload: Carga útil.
- Fill: Separador entre paquetes.
- ICRC y VCRC: Invariante CRC y Variante CRC.

# Ruteo en Infiniband

- Up/Down.
- DFS (mejora del Up/Down).
- Smart-routing.
- Adaptive-trail.
- Minimal adaptive.

# Enlaces físicos en Infiniband (Myri-10G)

| InfiniBand Link | Signal Pairs | Signaling Rate | Data Rate (Full Duplex) |
|---|---|---|---|
| 1X-SDR | 2 | 2.5 Gbps | 2.0 Gbps |
| 4X-SDR | 8 | 10 Gbps (4 x 2.5 Gbps) | 8 Gbps (4 x 2 Gbps) |
| 12X-SDR | 24 | 30 Gbps (12 x 2.5 Gbps) | 24 Gbps (12 x 2 Gbps) |
| 1X-DDR | 2 | 5 Gbps | 4.0 Gbps |
| 4X-DDR | 8 | 20 Gbps (4 x 5 Gbps) | 16 Gbps (4 x 4 Gbps) |
| 12X-DDR | 24 | 60 Gbps (12 x 5 Gbps) | 48 Gbps (12 x 4 Gbps) |
| 1X-QDR | 2 | 10 Gbps | 8.0 Gbps |
| 4X-QDR | 8 | 40 Gbps (4 x 5 Gbps) | 32 Gbps (4 x 8 Gbps) |
| 12XQDDR | 24 | 1200 Gbps (12 x 5 Gbps) | 96 Gbps (12 x 8 Gbps) |



# Máquinas con Infiniband

- Thunderbird.
- Columbia.
- TSUBAME Grid Cluster.
- Jaws.
- Lonestar.
- Darwin.
- Lincoln.
- TSUBAME Grid Cluster.

# Quadrics

# Quadrics

- Aparece en 1996
- Subsidiaria de Alenia Aeronáutica, y parte del grupo Fimmeccanica .



- Tecnología Europea.
- Gama de productos:
  - QsNet
  - QsNet II (2004)
  - QsTenG

# Características principales de Quadrics

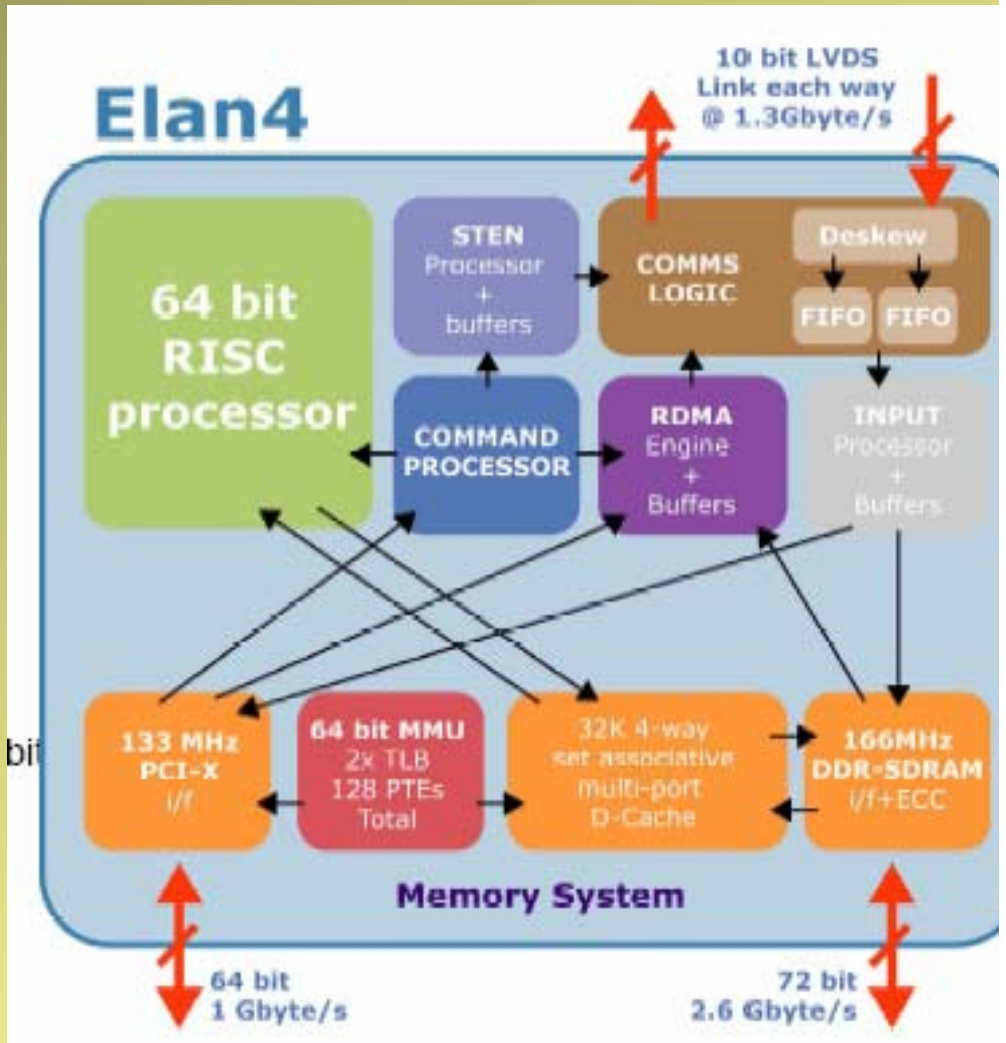
- Baja Latencia: 1.26 µs- 5 µs.
- Escalabilidad.
- Open-sourced.
- Libre de Interbloqueo (Deadlock).
- Soporta redes virtuales (VLAN).
- Limite plataformas donde opera (Linux, Unix).
- Ofrecen soporte Técnico (Pago del servicio).
- Conmutación Wormhole.

# Arquitectura Quadrics

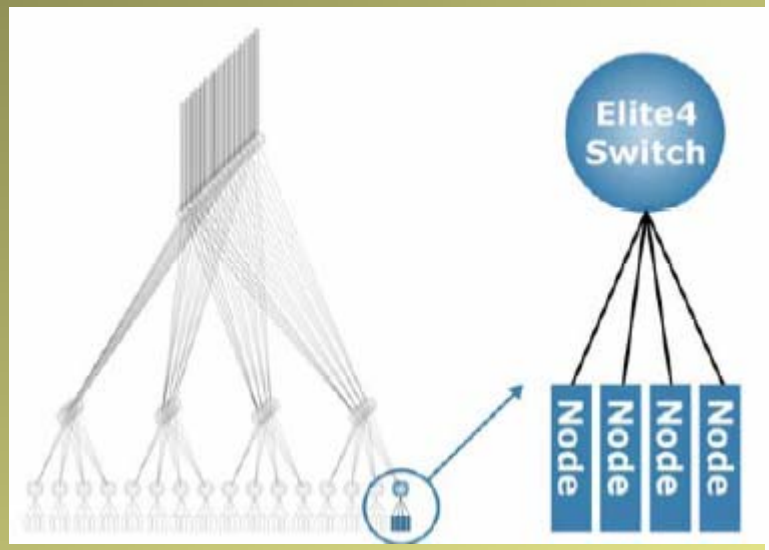
- QsNet: Compuesto por dos bloques hardware:
  - Elan: Interfaz de red programable.
  - Elite: Se encarga de la comunicación (Switch).
    - 400MB/s.
    - Detección de errores de paquetes (CRC)
    - Dos niveles de prioridad.
    - Soporte hardware para broadcast
    - Ruteo adaptativo
    - Los switch se conectan a través de la topología quaternary fat-tree.

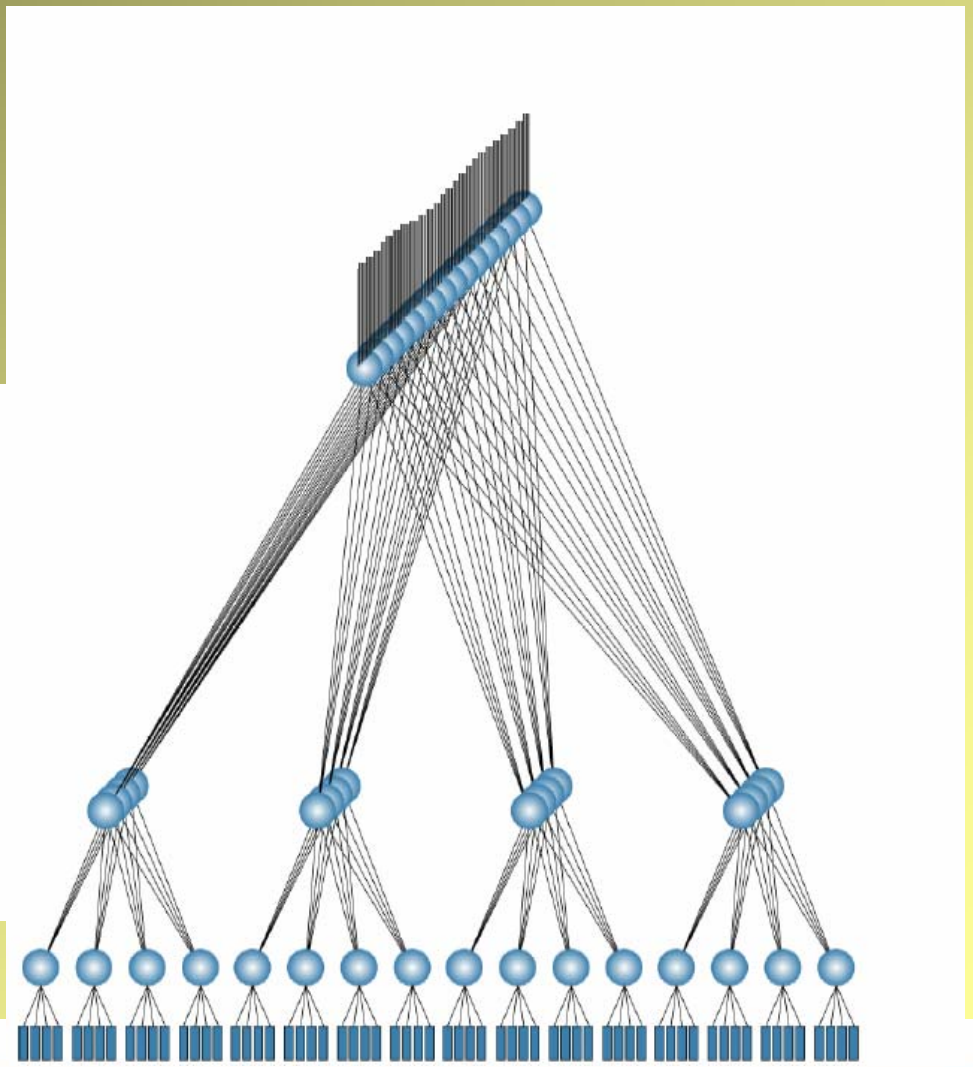
# Tarjeta de red Elan

- PCI-X.
- 10bits, 1.333Gbaud o 900MBytes/s.
- Procesador 64 bits a 200MHz.
- 64MB memory (2.7GB/s transferencia con la mem).
- Se mantiene una copia de la TLB en la MAC.
- Latencia menor a 2µs sobre MPI.

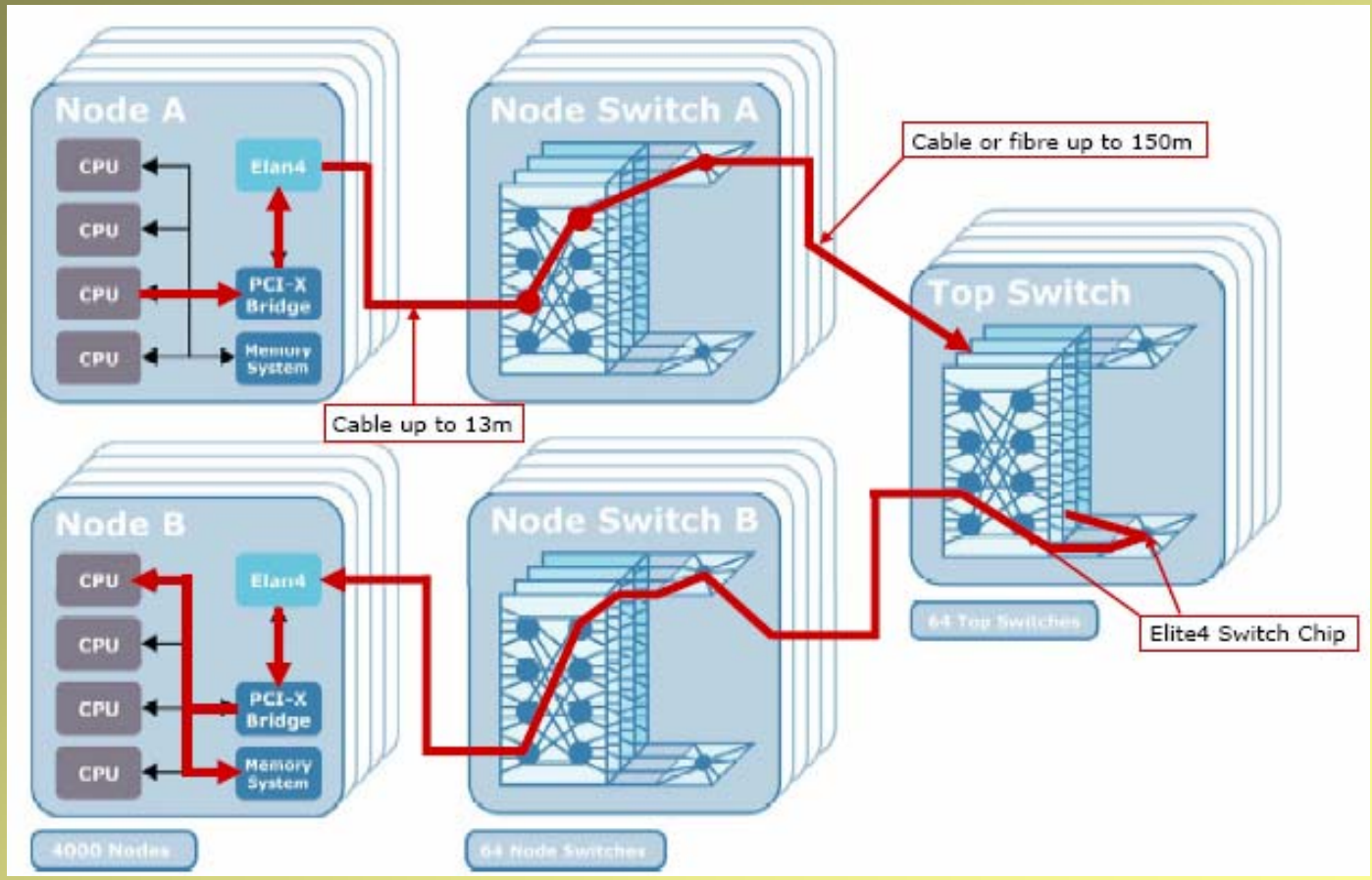


# Topología Quaternary Fat-Tree



# Mensajes a través de Quaternary Fat-Tree



# QsNet vs QsNetII

| | QsNet | QsNet$^{II}$ |
|---|---|---|
| **Bus interface** | PCI 2.1 | PCI-X 1.0 |
| **Peak bus bandwidth** | 528 Mbytes/s | 1064Mbytes/s |
| **QsNet link width** | 10 bits | 10 bits |
| **QsNet line rate** | 400Mbaud | 1.333Gbaud |
| **Sustainable transfer rate** | 350Mbytes/s | 900Mbytes/s |
| **On chip cache** | 4Kbytes unified | 32kbytes D + 16Kbytes I |
| **Local Memory** | 64Mbytes ECC SDRAM | 64Mbytes ECC DDR SDRAM |
| **Peak Memory Bandwidth** | 800Mbytes/s | 2.67Gbyte/s |
| **IO processor** | 100Mhz 32 bit | 200MHz 64 bit |
| **Physical Addressing** | 48 bits | 52 bits |
| **Virtual Address** | 32 bit VA, 4K contexts | 64 bit VA, 4K/16K contexts |
| **MMU** | 16 entry TLB + table walk | 2 x 64 entry TLB + hash table |

# QsTenG

- Cumple el estándar de la tarjetas10 Gigabit Ethernet.
- Letecia en torno al 10 µs.
- 12 slots para tarjetas en línea.
- Trabaja con switches Gigabit Ethernet.

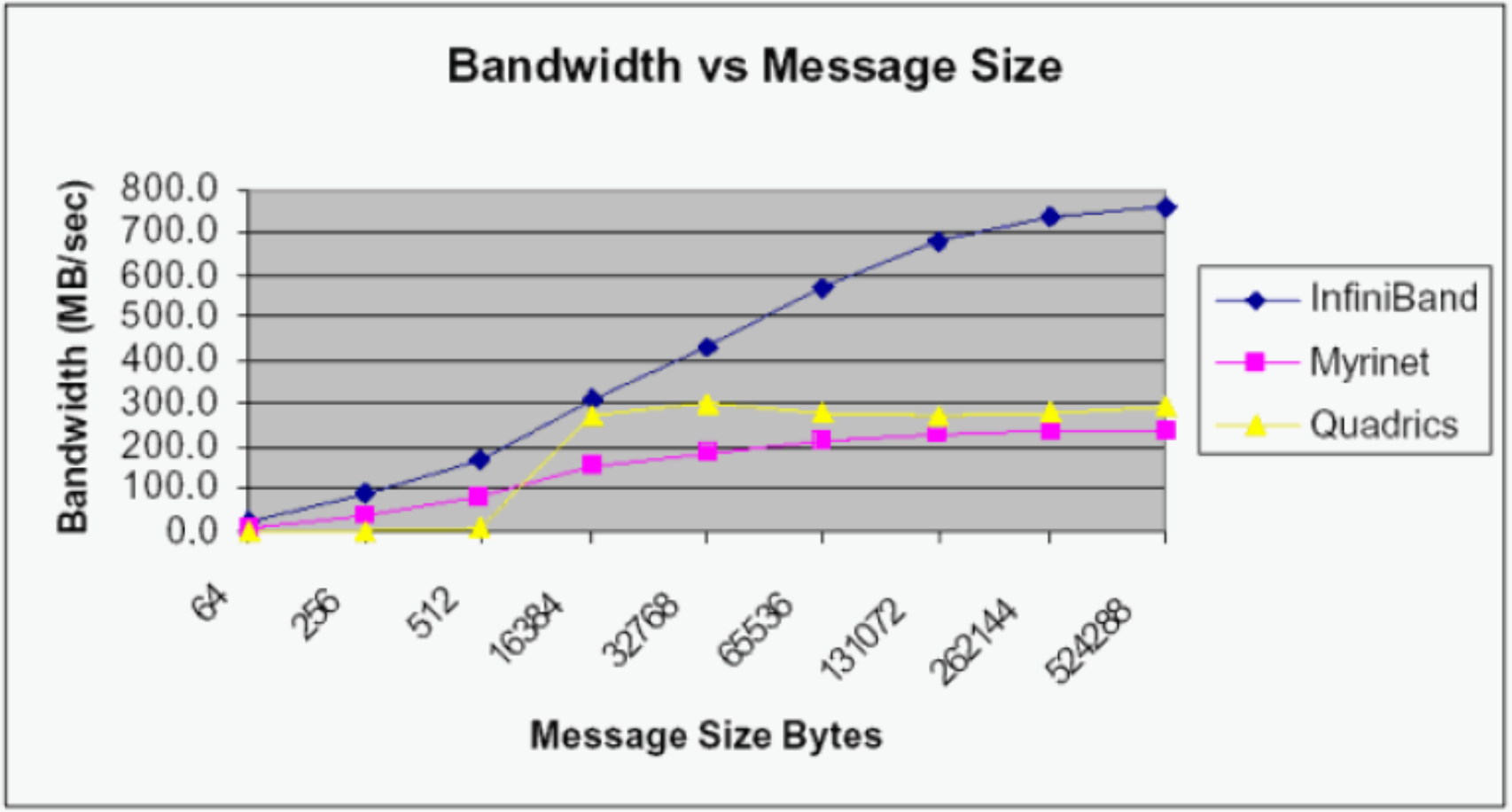
# Máquinas con Quadrics

- Tera-10.
- Thunder.
- Albert.
- Mpp2.
- MCR.

# Comparativas de las tecnologías

| Technology | Vendor | MPI latency, usec | Bandwidth per link [unidirectional, MB/s] |
|---|---|---|---|
| Myrinet XP2 | Myricom | 5.7 | 495 |
| QsNet II | Quadrics | 2 | 900 |
| Infiniband | Voltaire | 3.5 | 830 |

# Ancho de banda




*Source: Ohio State University, Xeon 2.2 GHz up processor platform*

# Precios por Puertos

- Infiniband
  - Para puertos de cable 4X/8X 761,40 € por puerto.
  - Para cableado de menor capacidad entre 392,02€ y 505,10€ por puerto.
- Myrinet 829,28 €/puerto.
- Quadrics 2223,98 €/puerto.