REDES DE BAJA LATENCIA

- MYRINET
- INFINIBAND
- QUADRICS

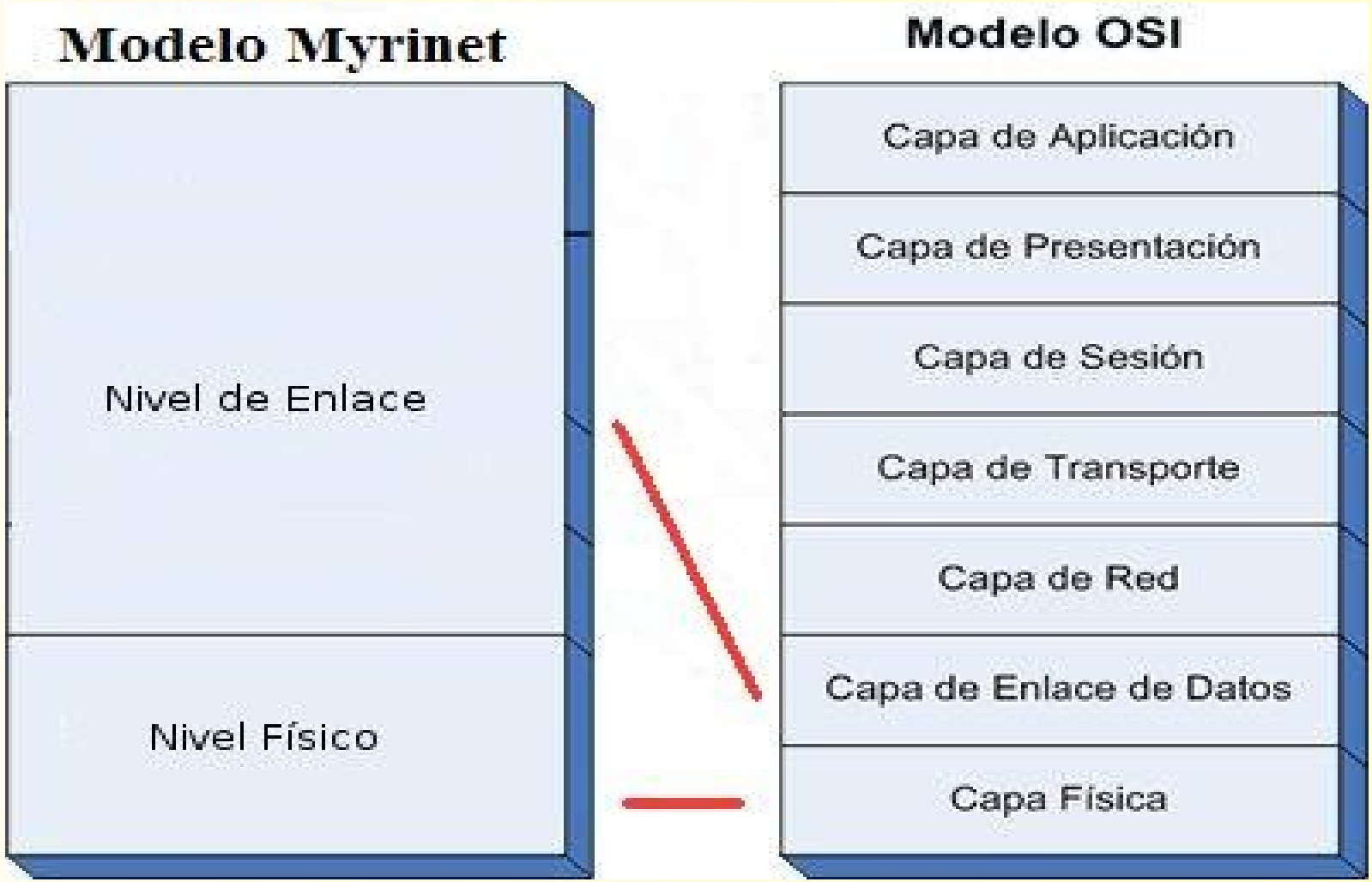
Ubay Díaz Machín Alu2238@etsii.ull.es

Myrinet

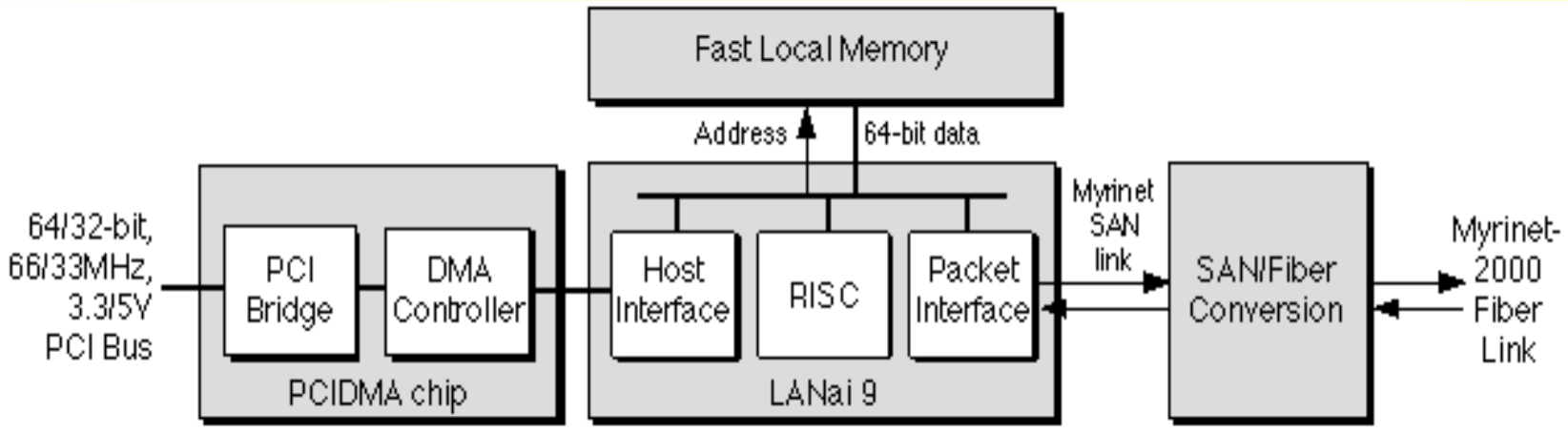
MYRINET
Características

- Baja Latencia.
- Escalabilidad-Autodetección de Topología.
- Monitorización de cada enlace.
- Software Libre.
- Tratamiento Interbloqueo (Deadlock)

MYRINET
MODELO



MYRINET
Targeta



- Se conecta directamente al bus del sistema. Y gracias a su procesador se ocupa de todas las comunicaciones.

MYRINET
GAMA

Myri-10G PCI Express NIC
dual-protocol
10-Gigabit Ethernet and 10-Gigabit Myrinet

Myrinet-2000 “ Gigabit pot Canal

MYRINET
NIVELES

- **Nivel de Enlace**
  - Crea el enlace entre los sistemas.
- **Nivel Físico**
  - Control de Flujo.
  - Control de Errores.
  - Direcciona los paquetes.

MYRINET PROBLEMAS

- **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**

MYRINET
PROBLEMAS

- **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.**

MYRINET AUTOCONFIGURACIÓN

- Existe un **MAPPER** que realiza el mapa de la red.
- Para crear el mapa envía paquetes y si:
  - Responde es un terminal con de paso la ruta e identificador de este.
  - Si tras 2 envíos no recibe nada, se asume un conmutador y se envía un paquete circular para confirmar.

MYRINET
RUTEO

- **Ruteo Wormhole**

Se usa unos paquetes llamados “worn”.

Que son enviados por los switches, que lo reenvían nada más recibirlos. Sin esperar a ensamblarlos. Que puede provocar que un mismo “worm”, sea enviado a la vez por varios puertos.

En caso del que el destino no sea accesible, se devuelve por el mismo canal.

MYRINET
RUTEO

- UP/DOWN

Se basa en tomar un nodo arbitrario como nodo raíz, y a partir de este crear un árbol binario.

Asegurándonos de romper las conexiones entre los nodos del mismo nivel.

Los envíos serán subiendo por el árbol para luego bajar por este.

MYRINET
RUTEO

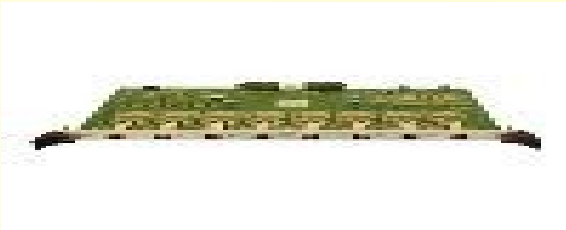
- RUTEO FUENTE

Usando un árbol “up/down”, se hace una representación de este.

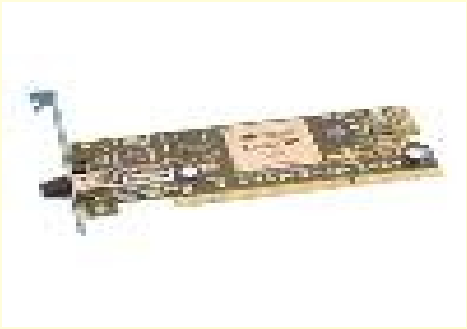
Así cuando se envía un paquete este ya tiene el camino.

MYRINET
DISPOSITIVOS

- Myricom Myrinet M3-SPINE-8F - expansion module - 8 ports = 239 $



- Myricom Myrinet network adapter = 215 $



MYRINET
USADA POR:

- Barcelona Supercomputing Center. Nº 13 (ESPAÑA).
- Caltech Nº 39 (USA).
- Indiana University Nº 42 (USA)
- University of Reading Nº 51 (UK)
- University of Southern California Nº 63 (USA)

Infiniband

Introducción

- Nace a Finales de los 90´s
- NGIO (Next Genration I/O) + FIO (Future I/O)
- En 1999 se Funda la IBTA (InfiniBand Trade Association) www.infinibandta.org
- Fue pensada como Sustituto del PCI.
- Algunos de los miembros son:

INFINIBAND™
Trade Association

Introducción

Introducción

- Alternativa Libre www.openib.org
- Tubo el fin de Crear en un inicio un Software para infiniband en Linux.
- Actualmente no sólo hacen esto sino también dan soporte a instalaciones.

INFINIBAND™
Trade Association

Four Membership Levels

- **Promoters ($5000/year, $3000 initiation)**
  - Organizations and enterprises that wish to strongly influence the process and features in software created and the accompanying promotional activities to enhance the code they use or provide
- **Adopters ($3000/year, $3000 initiation)**
  - Organizations and enterprises that wish to contribute to and participate in the processes and work of the promoters but do not feel the need to strongly affect the outcomes
- **Supporters ($1000/year, $3000 initiation)**
  - Organizations and enterprises that wish to use the OpenFabrics software, leverage the promotional activities, be tied into the work of the Alliance but not necessarily contribute
- **Consulting (Free)**
  - Organizations and individuals that the Alliance selects for honorary membership on an annual basis based on the perceived value of their membership to the Alliance

- All members agree to understand the Bylaws and Membership agreements and to work within the Alliance processes and rules described therein

Miembros de Esta Alianza

INFINIBAND℠
Trade Association

Descripción

- ***S***ystem ***A***rea ***N***etwork
  - Para comunicar Ordenadores, Sistema E/S y Dispositivos E/S.
- Plantea el uso de un único canal para la interconexión de dispositvos de un sistema.
- Muy complicado para un PC.
- Se basa en el uso de Conmutadores para las funciones de E/S.

INFINIBAND™ Trade Association

Características

- Baja Latencia
- Escalabilidad
- Open-Sourced
- Libre de Interbloqueos.
- Posibilidad de Uso de VLANs.
- Técnica de conmutación Virtual Cut-Through.
- Tolerancia Fallos.

INFINIBAND™
Trade Association

Arquitectura

- Sistema Basado en la Conmutación.
- Se basa en la conexión Punto a Punto.
- Componentes:
  - **Enlaces y repetidores**
  - **Channel adapters**
  - **Host Channel Adapter (HCA)**
  - **Target Channel Adapter (TCA)**

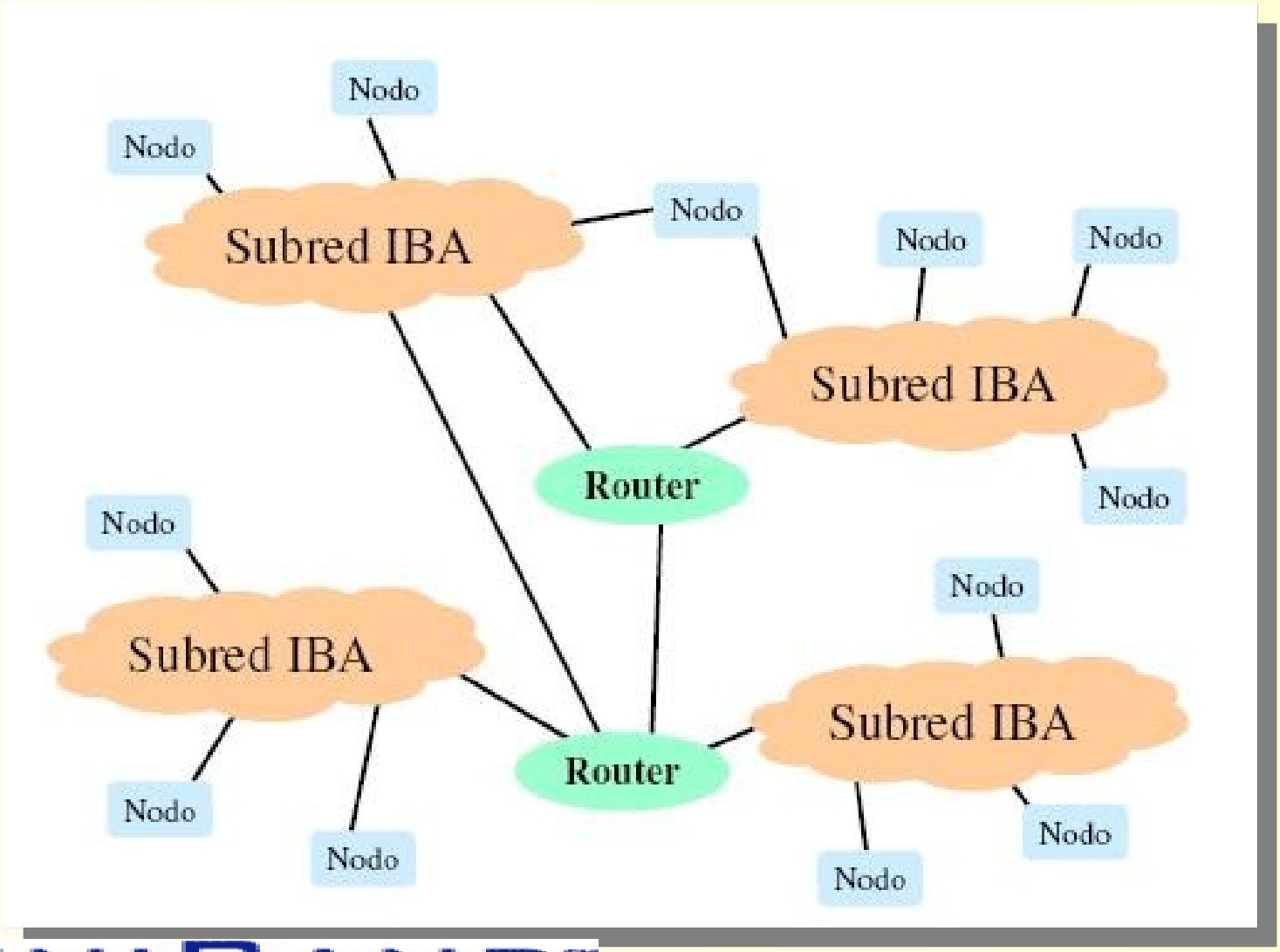
Dispositivo final que conecta un nodo con la Red.

Un HCA es para sistemas de procesamiento con gran capacidad para ejecutar distinto tipo de software.

Un TCA es para dispositivos de E/S, que por su simplicidad, no suelen tener capacidad para ejecutar software.

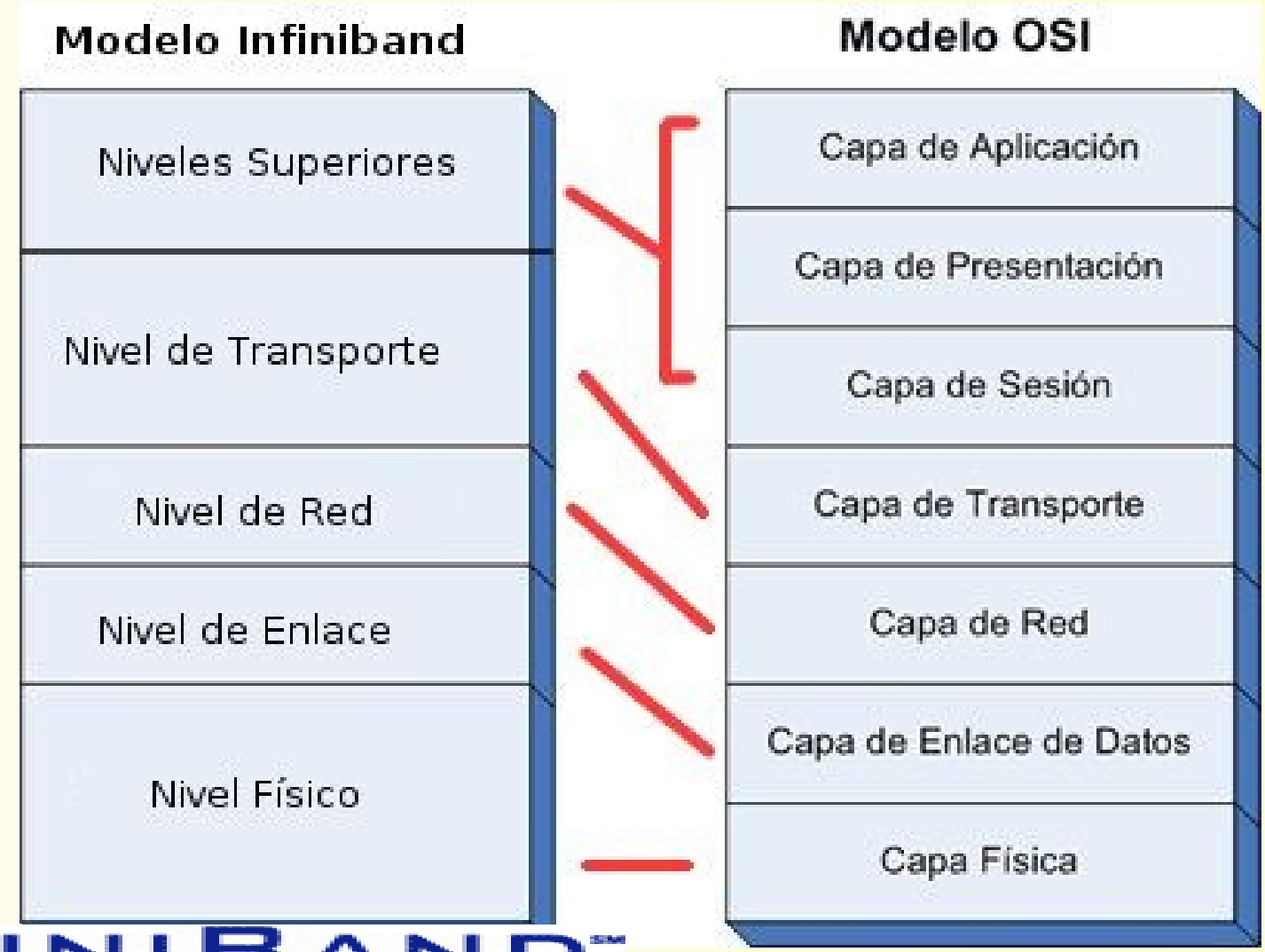
INFINIBAND™
Trade Association

Topología



INFINIBAND™
Trade Association

Modelo Infiniband



INFINIBAND™
Trade Association

Nivel Físico

- Se usa codificación 8B/10B (Mapea 8 en 10)
- Varía según el medio Físico.
  - Par trenzado
  - Fibra Óptica
  - Circuito en Placa
- Se define el carácter que marca inicio y fin de paquete.

INFINIBAND℠
Trade Association

Nivel de Enlace

- Describe los formatos de paquete a usar y protocolos.
- Controla el Flujo y el encaminamiento en la subred.
- Hay 2 tipos de Paquetes:
  - Paquetes de Gestión
  - Paquetes de Datos.

Nivel de Red

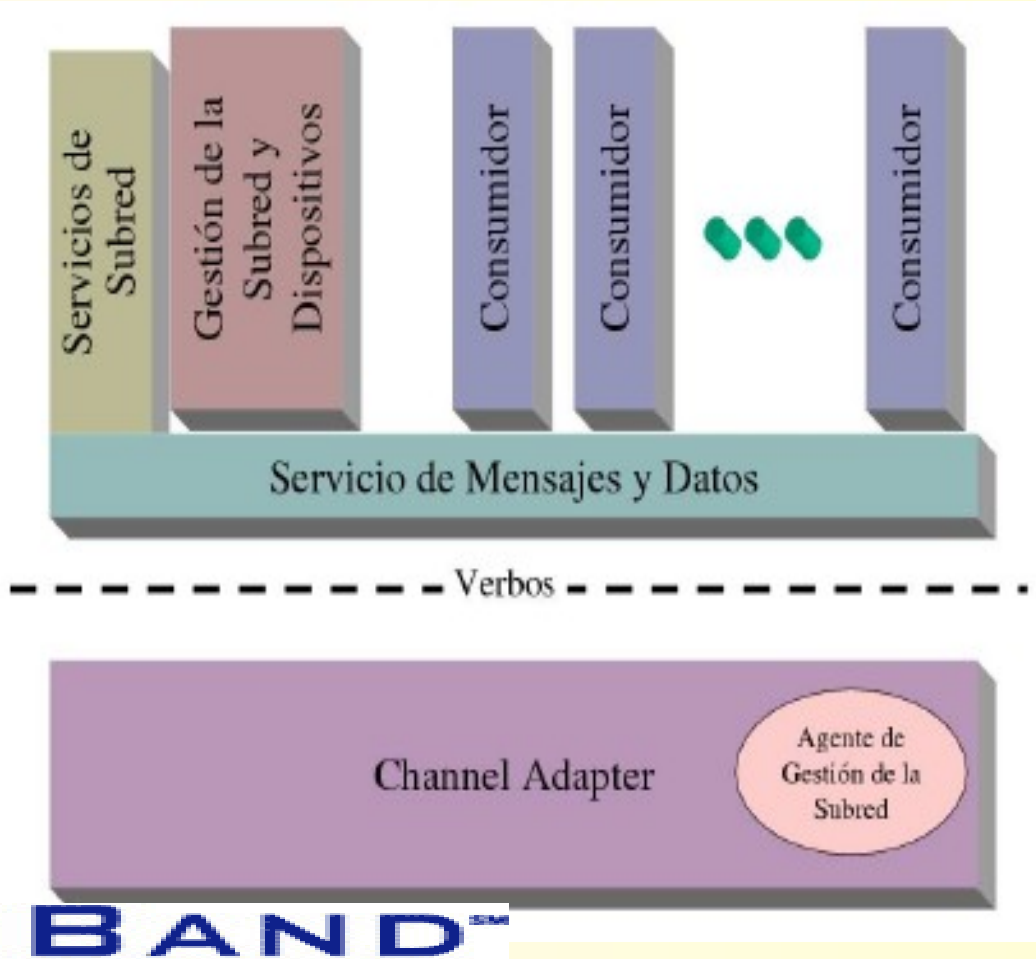
- Encamina entre diferentes Subredes
- Usa direcciones con el formato Ipv6
- 2001:0db8:85a3:08d3:1319:8a2e:0370:7334

Nivel de Transporte

- Se encarga de asegurar que llega la información y que la información que llegue sea la correcta.

Niveles Superiores

- Encontramos aquí cualquier protocolo.



INFINIBAND™
Trade Association

Velocidades

- Usa dos canales, uno de envío otro de recepción. Con una capacidad de 2.5 gbps.

Caudal de Infiniband, bruto / eficaz

	SDR	DDR	QDR
1X	2.5 / 2 gbps	5 / 4 gbps	10 / 8 gbps
4X	10 / 8 gbps	20 / 16 gbps	40 / 32 gbps
12X	30 / 24 gbps	60 / 48 gbps	120 / 96 gbps

INFINIBAND™
Trade Association

Dispositivos

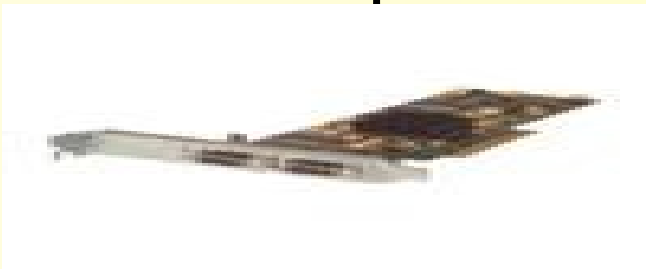
HP 4X DDR InfiniBand Mezzanine
HCA - adaptador de red
Precio: 472,28 €

Topspin InfiniBand Switch
Module - switch - 14 ports
Precio: $7,702.99

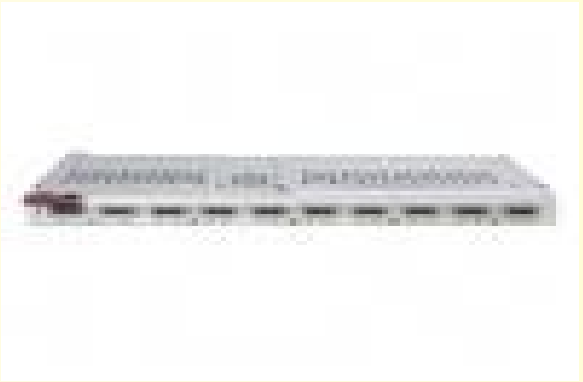
INFINIBAND™
Trade Association

Dispositivos

- Topspin InfiniBand Host Channel Adapter - network adapter = 1200 $



- Visionman Infiniband Switch / Designed for SuperBl = 7200 $



INFINIBAND℠
Trade Association

Usada Por:

- NCSA (USA) Nº 8
- NNSA/Sandia National Laboratories (USA) Nº 11
- GSIC Center, Tokyo Institute of Technology (Japón) Nº 14
- Texas Advanced Computing Center/Univ. of Texas (USA) Nº 15

INFINIBAND℠
Trade Association

QUADRICS

QUADRICS
INTRODUCCIÓN

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

- Tecnología Europea.
- **Gama de productos:**
  - QsNet
  - QsNet II
  - QsTenG

QUADRICS
INTRODUCCIÓN

- En 2004 BULL, seleccionó Quadrics para la creación del TERA10 (Nº 19), Francia.
- Hace poco fue seleccionada por HP, para la actualización del SHARCNET (Nº 491) , Canada.
- En Agosto de 2005 Firmó un acuerdo con STMicroelectronics , para el diseño de una nueva generación de redes.
- En Noviembre de 2005, anuncia una tecnología basada en la tecnologia 10 GigaBit.

Quadrics

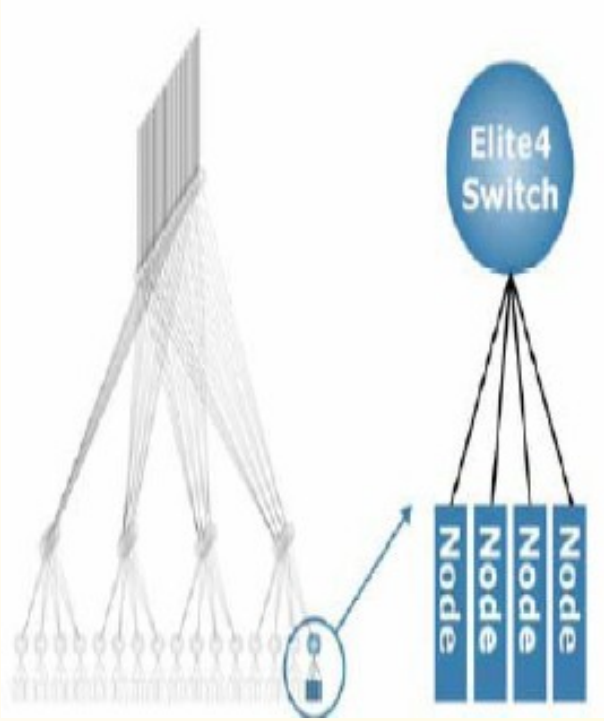
QUADRICS CARACTERÍSTICAS

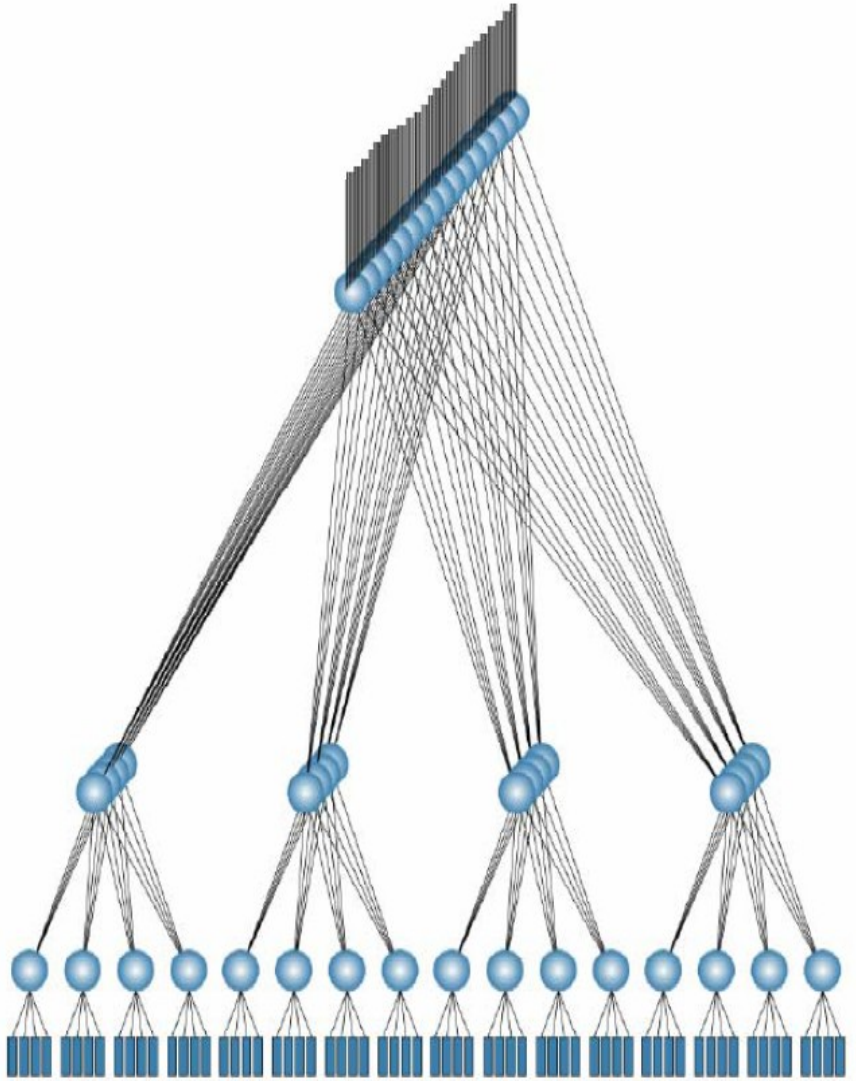
- Baja Latencia: 1.26 µs- 5 µs.
- Escalabilidad
- Libre de Interbloqueo.
- Soporte Redes Virtuales (VLAN)
- Limite de Plataformas (Unix/Linux)
- Ofrecen soporte Técnico (Pago del servicio).
- Conmutación Wormhole.

Quadrics

QUADRICS
TOPOLOGIA

- Usa Arboles tipo Fat-Tree





Quadrics

QUADRICS
PRODUCTOS

- QsNet I - Quadrics interconnect based around the elan3/elite3 ASICs (350MBs @ 5us MPI latency)
- QsNet II - Quadrics interconnect based around the elan4/elite4 ASICs (912MB/s on SR1400 EM64T and 1.26us MPI latency on HP DL145G2)
- QsNet II E-Series- a range of small-medium configurations (8-128-way) at less than USD 1,800 price/per port
- QsTenG - 10 Gigabit Ethernet switch for up to 96 ports.

QUADRICS
DISPOSITIVOS

PRICING FOR QsNetII E-SERIES

	8	2 - 8 way: configured with Standalone 2U 8-way switch (QS8-A)		List Price
Part Number	**Qty**	**Product ID and Description**	**Unit Price**	**Extended Price**
3X-CM500-BA	8	QM500-BA Network Adapter, 64 Mbytes local SDRAM.	$997	$7,976
3X-CS8A0-AA	1	QS8A Switch, 8 Way, Copper Interconnect	$3,764	$3,764
3X-CM581-03	8	QM581-03 EOP Link Cable, 3M long	$135	$1,080
3X-CM574-AA	0	QM574 EMI Port Shield (1 off)	$16	$0
3X-CM585-AA	1	QM585 Rack Mount Kit QS8A	$253	$253
			TOTAL	**$13,073**
			PRICE PER PORT	**$1,634**

Quadrics

QUADRICS
USADA POR

- Commissariat a l'Energie Atomique (CEA) Nº 19 (Francia).
- Government Classified Nª 44 (USA).

TOP SECRET

- Lawrence Livermore National Laboratory Nº 47 (USA).
- Los Alamos National Laboratory Nº 91 (USA).

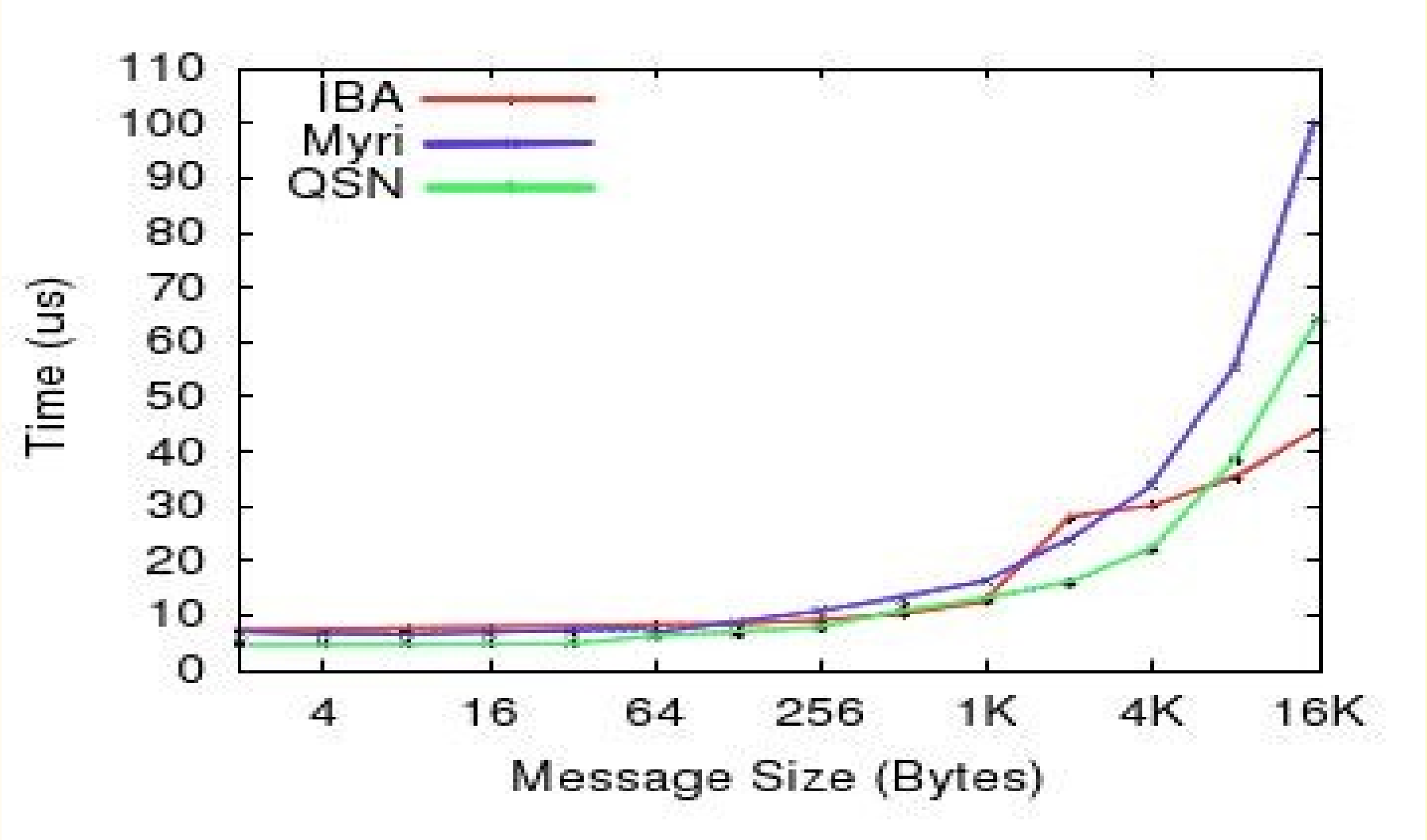
Comparativas

- Myrinet
- Infiniband
- Quadrics

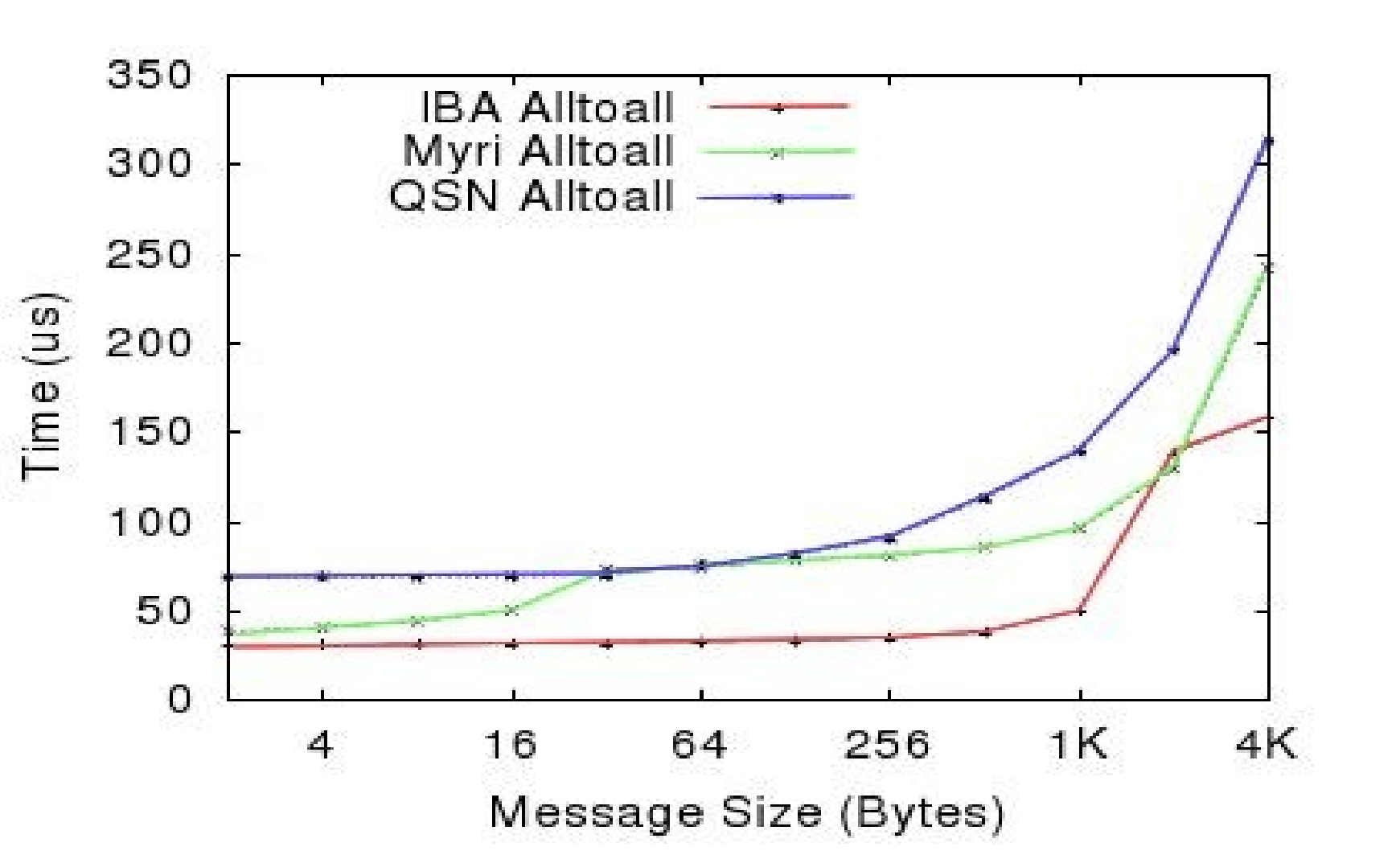
TABLA COMPARATIVA

RED	Velocidad Maxima	Latencia	$ Precio/Puesto	Relación Precio
Quadrics	10 Gigabits	1,26	1634	0,43
Myrinet	10 Gigabits	2,3	244	0,064
Infiniband	120/96 Gigabits	2,6	1922	0,506

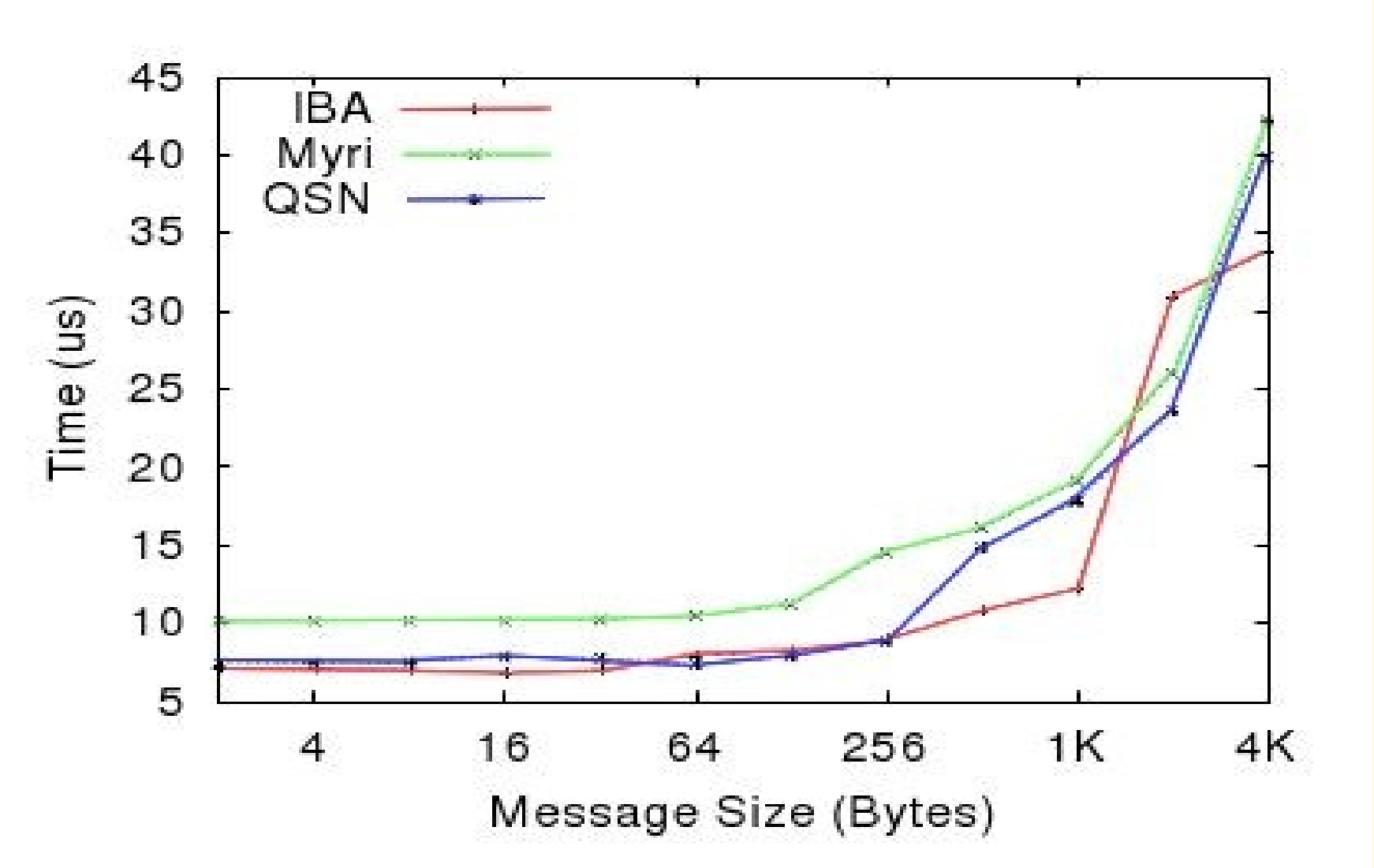
Latencia en MPI



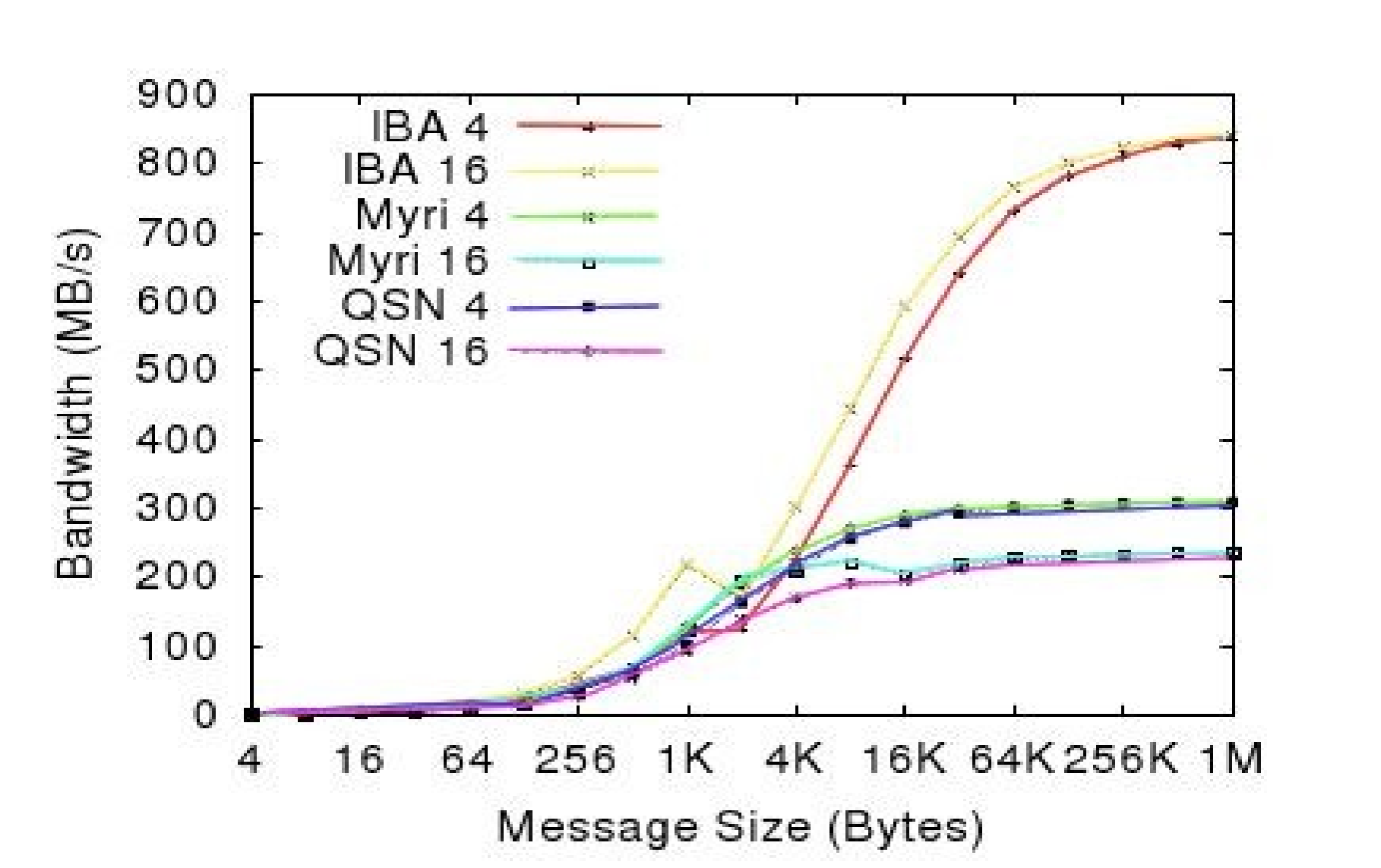
Comunicación entre todos los Nodos MPI



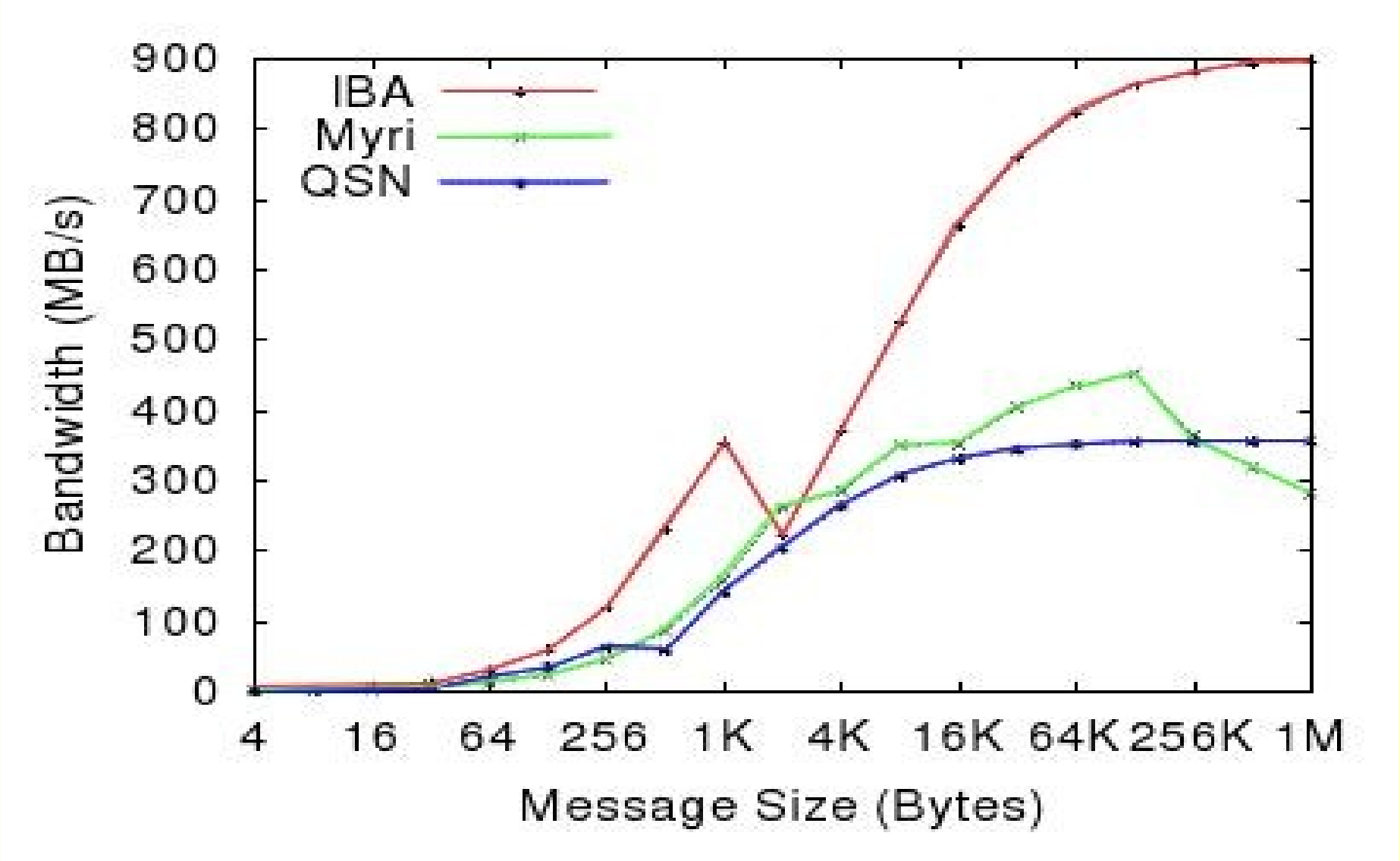
Tiempo de Un Reduce MPI



Ancho de Banda Usado en MPI



Ancho de Banda en Comunicación Bidireccional MPI



Comparativas de Precios
32 Puestos

- **Myrinet, $** 7863 y $ 244 por puerto.
- **Infiniband,** $ 61.506 y $ 1922 por puerto.
- **QsNet II E-Series,** $ 52.292 y $ 1634 por puerto.

Tendencias de Uso de Redes