PC Cluster 互連網路

國家高速網路與計算中心
王順泰
(stwang@nchc.org.tw)

2006/10/04

Introduction

- *Network is the most critical part of a cluster*
- *Its capabilities and performance directly influences the applicability of the whole system for high-performance computing*
- *Starting from LAN/WAN like Fast Ethernet and ATM, to System Area Networks(SAN) like Quadrics, Myrinet, InfiniBand, etc..*

Choice of Networks

- *Ethernet (10Mbps)*
- *Fast Ethernet (100Mbps)*
- *Gigabit Ethernet (1Gbps)*
- *10G Ethernet (10Gbps)*
- *ATM (155Mbps)*
- *Dolphin (2.5Gbps)*
- *Quadrics (7.2Gbps)*
- *Myrinet (2Gbps)*
- *InfiniBand (10Gbps)*

Fast Ethernet

- *100 Mbps over UTP or fiber-optic cable*
- *IEEE 802.3 CSMA/CD frame format*
- *Switched or shared media*
- *Advantage:*
  - *Low cost*
  - *Inexpensive CAT-5 copper*
- *Disadvantage:*
  - *Bandwidth utilization is not guaranteed to be fair*
  - *Low bandwidth utilization under heavy load*

Gigabit Ethernet

- *Being standardized by IEEE 802.3z*
- *IEEE 802.3 CSMA/CD frame format*
- *High performance: 1Gb full-duplex*
- *Fully compatible with current Ethernet*
  - *Carrier Extension*
    - *1 Gbps at 200m => Slot time = 512 bytes*
    - *Net throughput for small frames is only marginally better than 100 Mbps*
  - *Frame Bursting*
    - *Three times more throughput for small frames*
- *NICs and Switches are now common*

ATM (Asynchronous Transfer Mode)

- *Connection oriented packet switching*
  - *All communication uses established connections*
  - *Utilizes a sophisticated switch network to connect nodes*
- *Designed with the assumption that link bandwidth is expensive*
- *Not so good for cluster computer interconnection*
  - *Hardware cost*
  - *Not good performance in LAN*
  - *Effective in supporting clusters over WAN*
- *Highly suitable for wide area LAN and WAN*

Dolphin (1/2)

- *ANSI/IEEE 1596-1992 Scalable Coherent Interface (SCI) standard*
  - *Defines a point-to-point interface and a set of packet protocols*
  - *An SCI interface has two unidirectional links that operate concurrently*
  - *Supports shared memory and message passing*

Dolphin (2/2)

Dolphin PCI-64/66 - PCI-SCI Adapter Card

- **Link Speeds** - 667 Mbytes/s (1333 Mbytes/s duplex)
- **Performance** - Up to 326 MBytes/s throughput (depending on PCI Host Bridge) 1.46 microsecond latency (Measured Application to application)
- **Link Standard** - ANSI/IEEE 1596-1992 Scalable Coherent Interface (SCI)
- **PCI Specification** - PCI Local Bus Specification 2.2, 64/32 bit 66/33 MHz
- **Topologies** - Ring and switch topologies
- Supports both Direct Memory Access (DMA) and programmed Remote Memory Access (RMA)

Dolphin Software

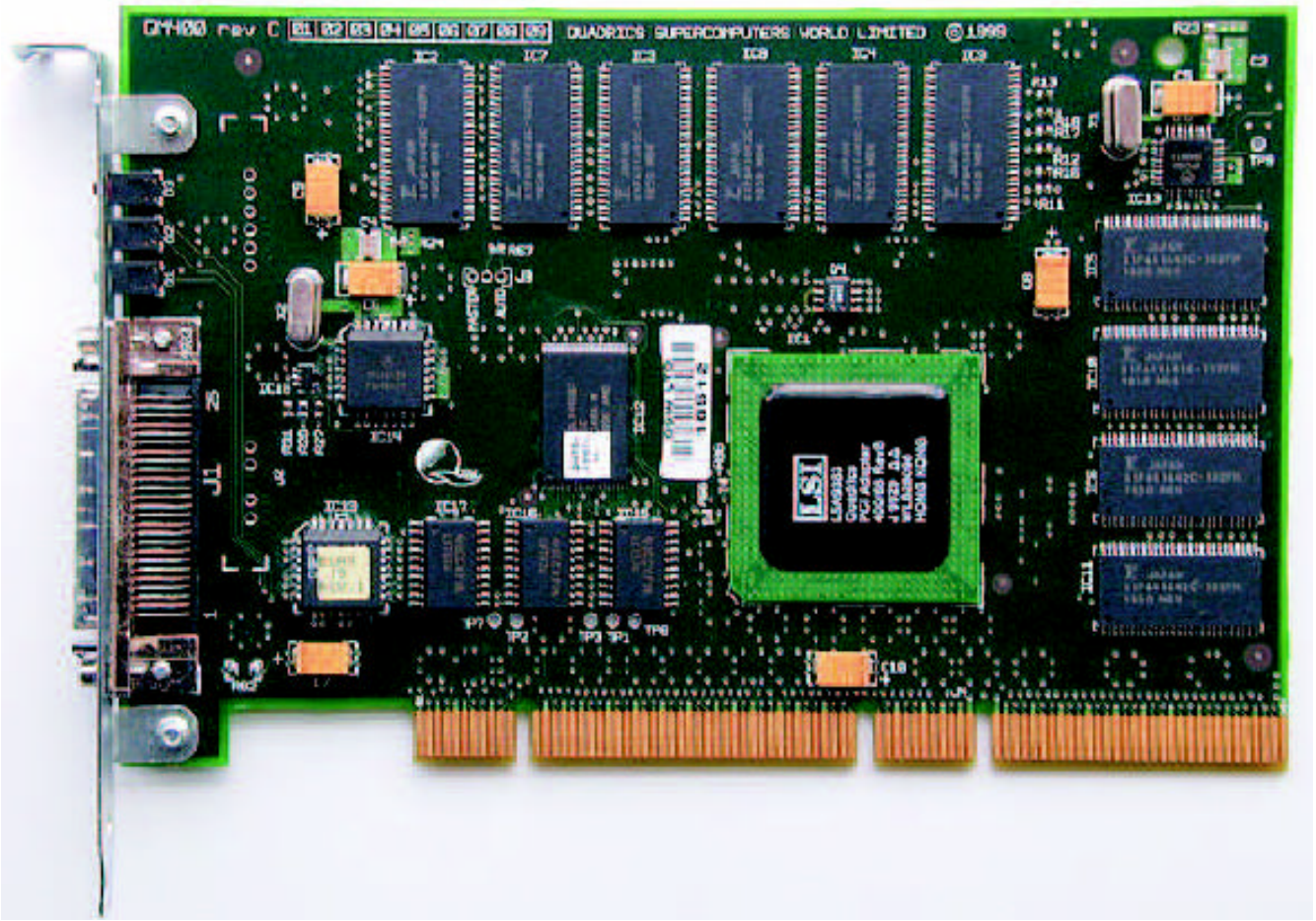
- All Dolphin SW is free open source (GPL or LGPL)
- SCI-SOCKET
  - Low Latency Socket Library
  - TCP and UDP Replacement
  - User and Kernel level support
- SCI-MPICH
  - MPICH 1.2 and some MPICH 2 features
  - New release is being prepared, beta available
- SCI Interconnect Manager
  - Automatic failover recovery.
  - No single point of failure in 2D and 3D networks.
- Other
  - SCI Reflective Memory, Scali MPI
  - Linux Labs SCI Cluster Cray-compatible shmem and Clugres PostgreSQL
  - Mandrake Soft Clustering HPC solution

Quadrics

- *High bandwidth & Low latency*
- *The Quadrics Network(QsNet) is based on two building blocks*
  - *Elan - network interface card*
  - *Elite - crossbar switch*
- *Scalable to > 4096 nodes*
- *QsNet provides:*
  - *An abstraction of distributed virtual shared memory*
  - *Network fault detection and fault tolerance*

Quadrics Components (1/3)

- QsNet host interface – Elan3

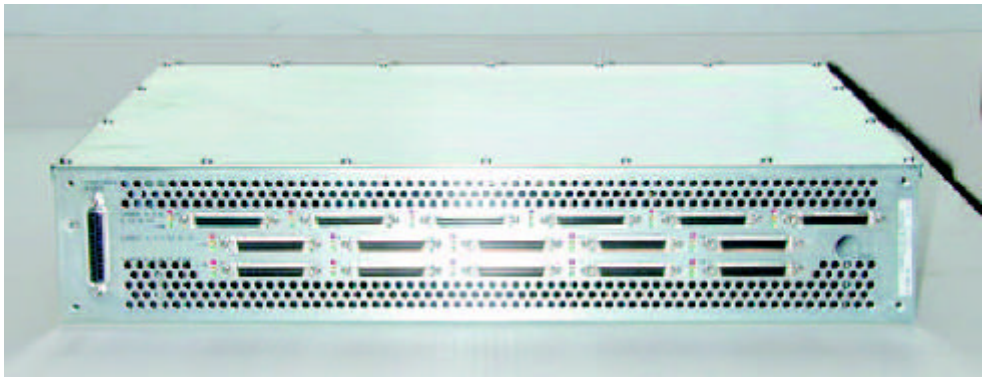


QM-400 Network Adapter PCI 64/66 interface

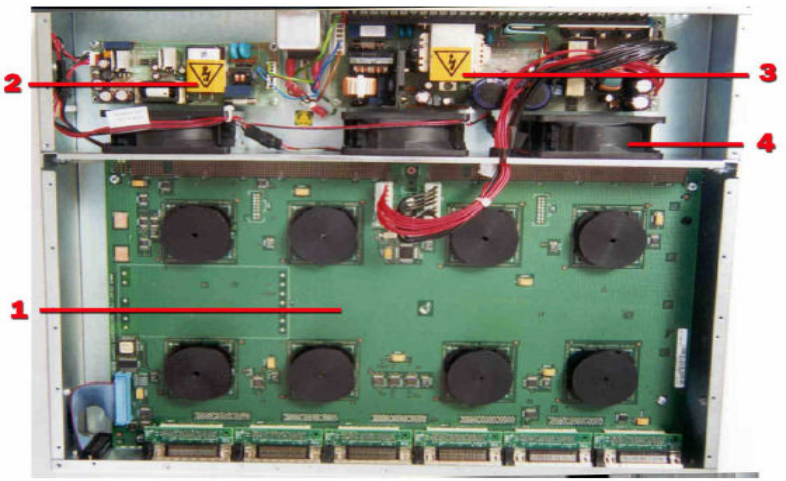
(From Quadrics)

Quadrics Components (2/3)

- QsNet switch - Elite



16-port QM-S16 16-port Switch (From Quadrics)



QM-S16 – Internal View (From Quadrics)

Quadrics Components (3/3)

Quadrics QsNetII Network Adapter. (Elan4)

QsNetII QS8A-AA 8-ports Stand-Alone Switch

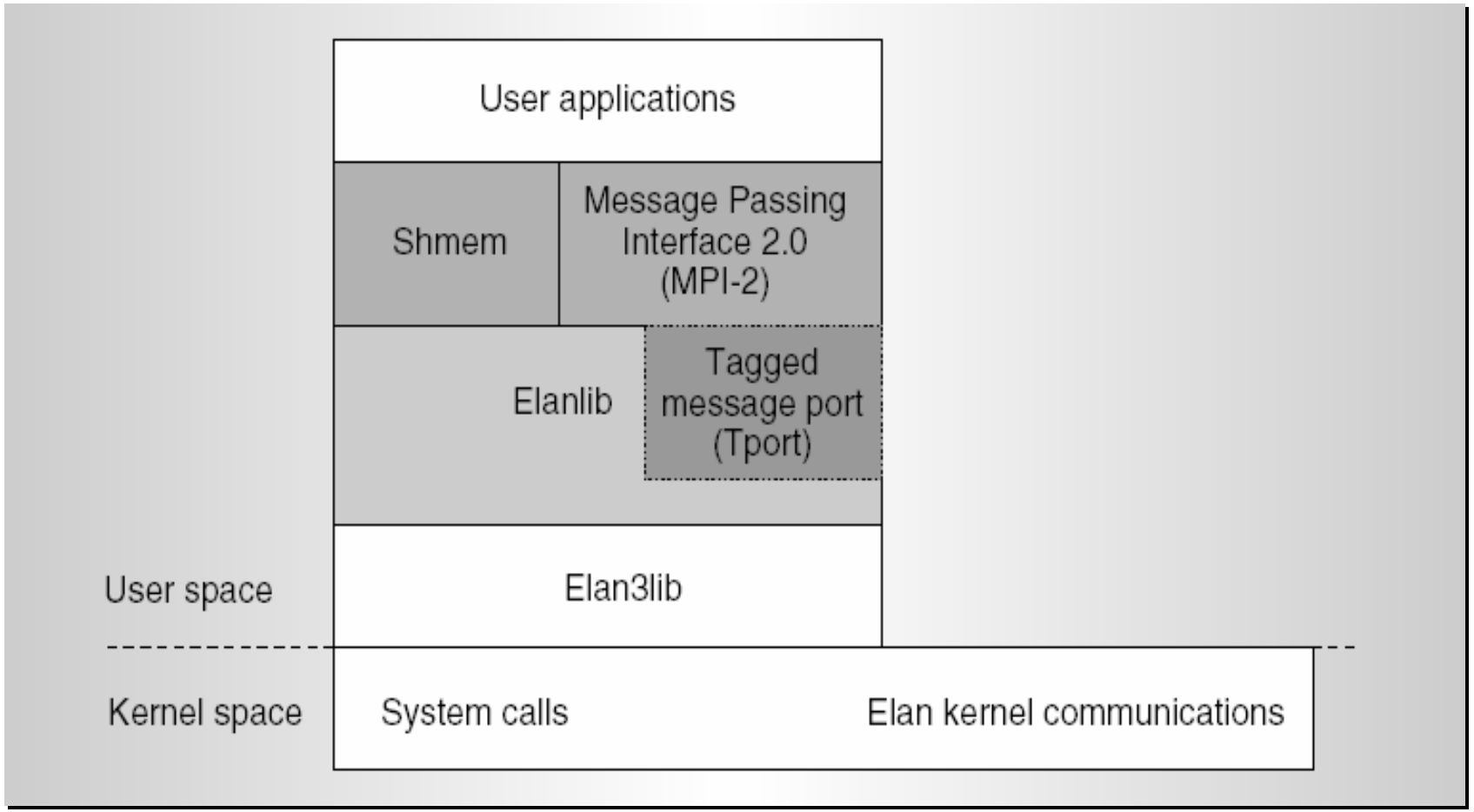
QsNetII QS32A-CA-CR 32-ports Stand-Alone Switch

QsNetII QS5A-LF-LA 128-ports Stand-Alone Switch

Quadrics Software

- Elan specific software
  - Device drivers & OS mods
  - IP over Elan, kernel messaging
  - Switch monitoring test and control
  - Programming libraries
    - Elan3lib
    - Elanlib and Tports
    - Quadrics MPI: MPICH 1.2.4 base
    - Quadrics Shmem
- System software
  - RMS resource management system

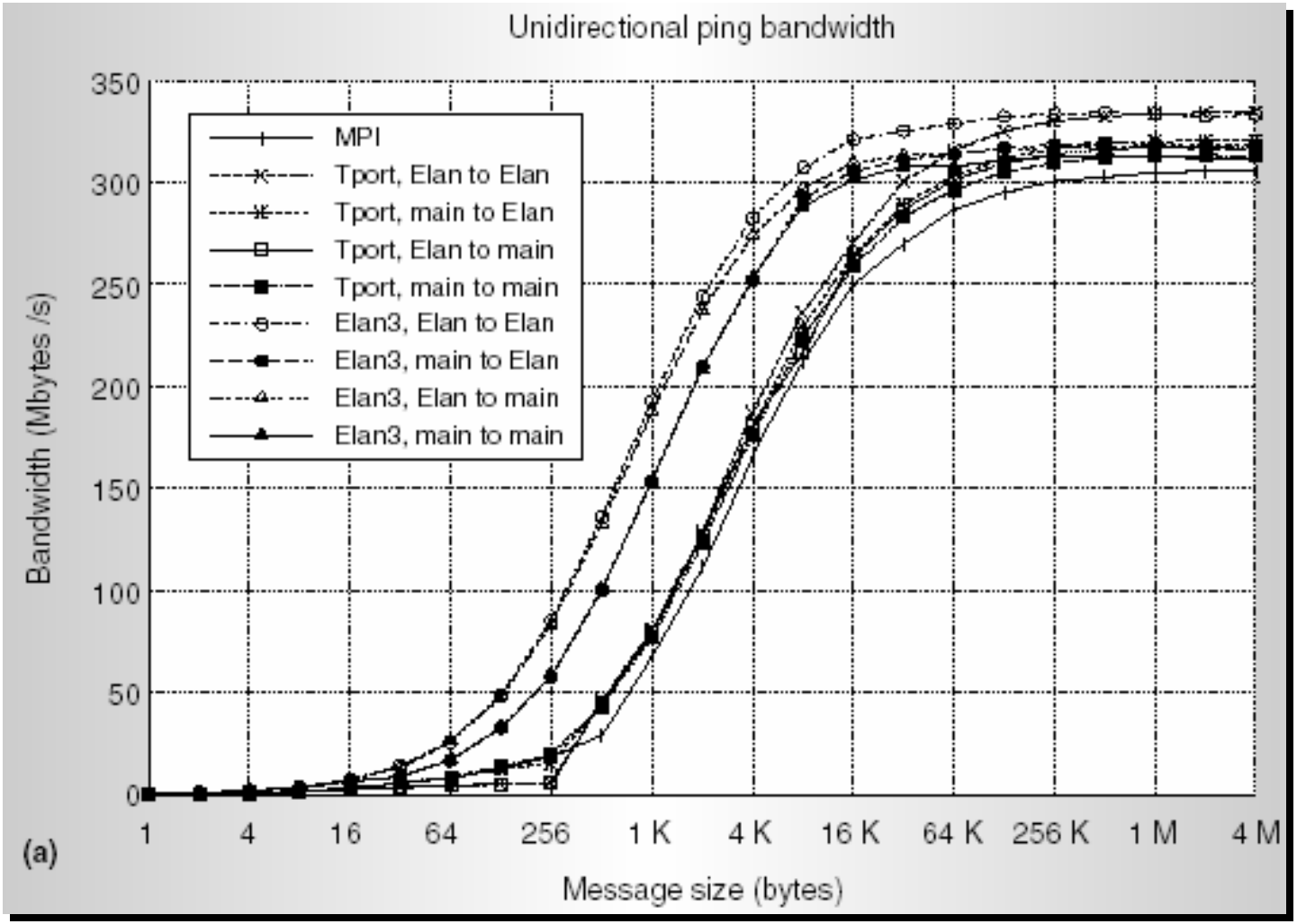
Programmable Libraries Hierarchy



Quadrics Communication Performance (1/3)

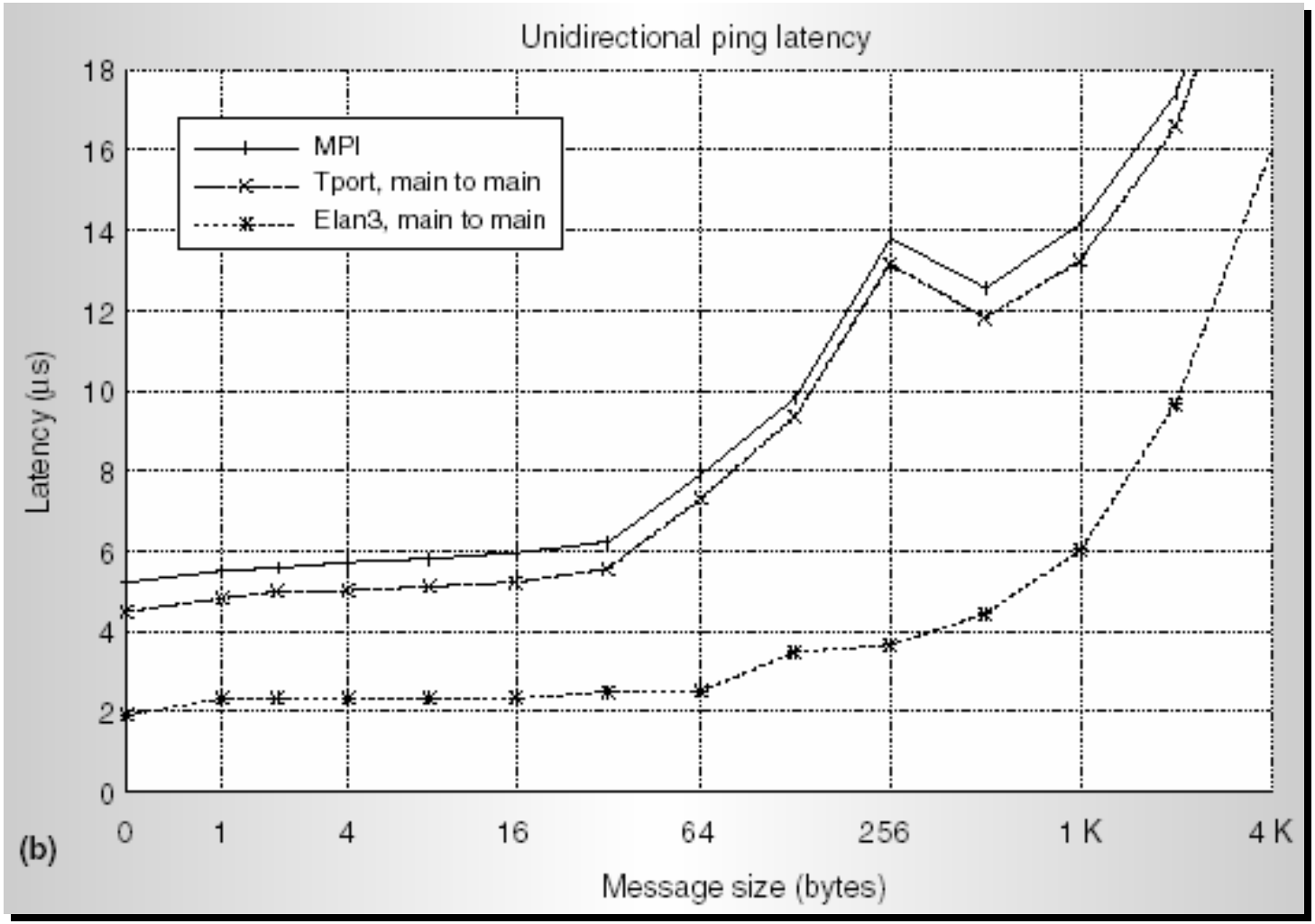
	QsNet	QsNet^{II}
Bandwidth (Sustainable transfer rate)	350 MB/s (2.8 Gb/s)	900 MB/s (7.2 Gb/s)
Latency	< 5 µs	< 3 µs

Quadrics Communication Performance (2/3)



(From IEEE MICRO-2002 P.54)

Quadrics Communication Performance (3/3)



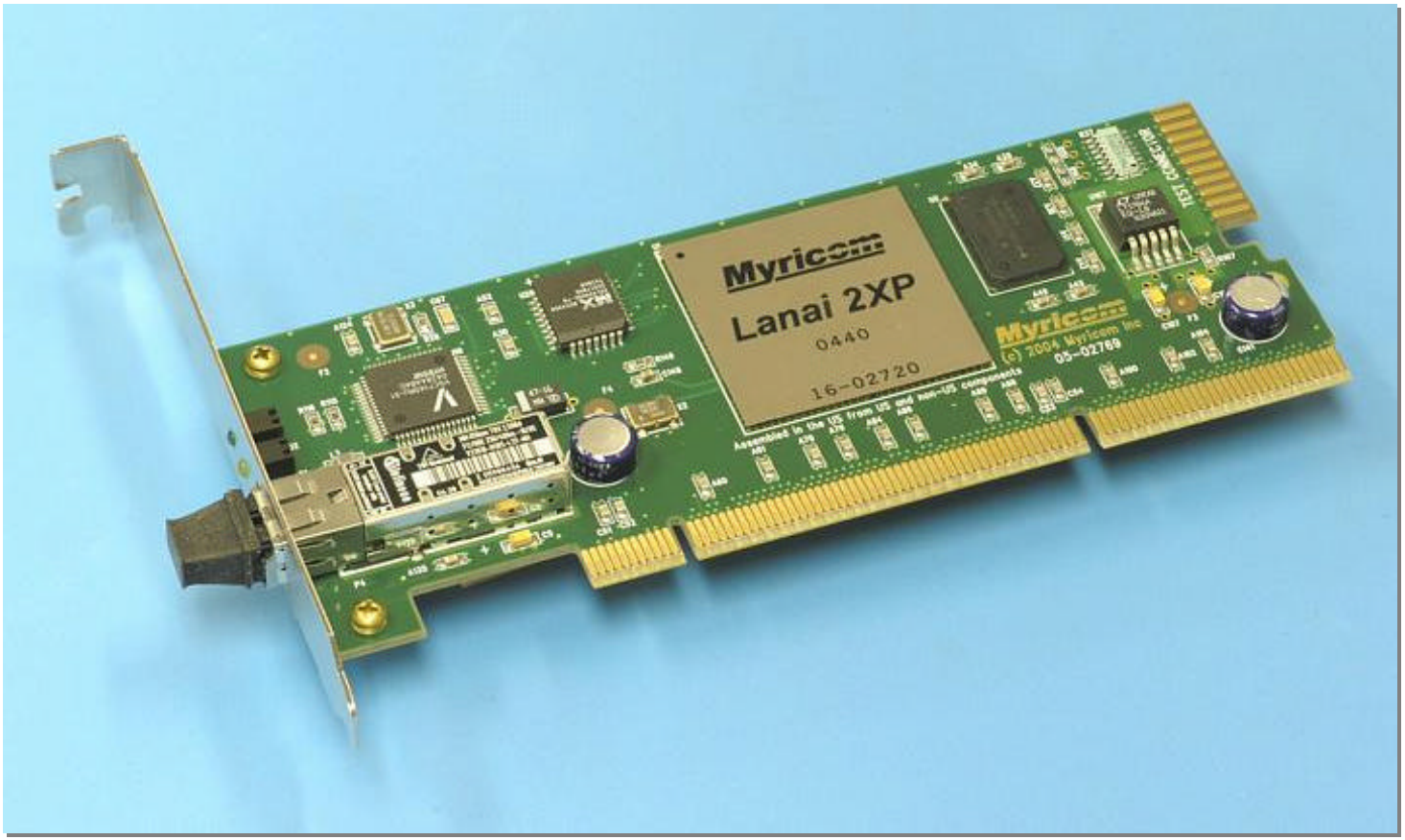
(From IEEE MICRO-2002 P.54)

Myrinet

- *A main product of Myricom (founded in 1994)*
- *A SAN evolved from supercomputer technology*
  - *High Performance: high-data-rate (2 Gbps), low-latency (less than 5 μs)*
  - *High Availability: by detecting and isolating faults, and using alternative communication paths*
  - *High Reliability: the MTBF of Myrinet switches and interfaces exceeds several million hours*
  - *Variable size frames: Myrinet packets may be of any length and it can encapsulate other types of packets*
- *Quite popular in the research community*
  - *All HW & SW specifications are open & public*

Myrinet Components (1/2)

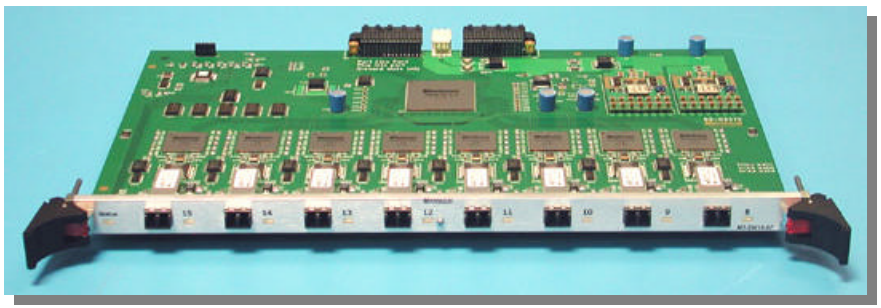
- *Myrinet host interface*

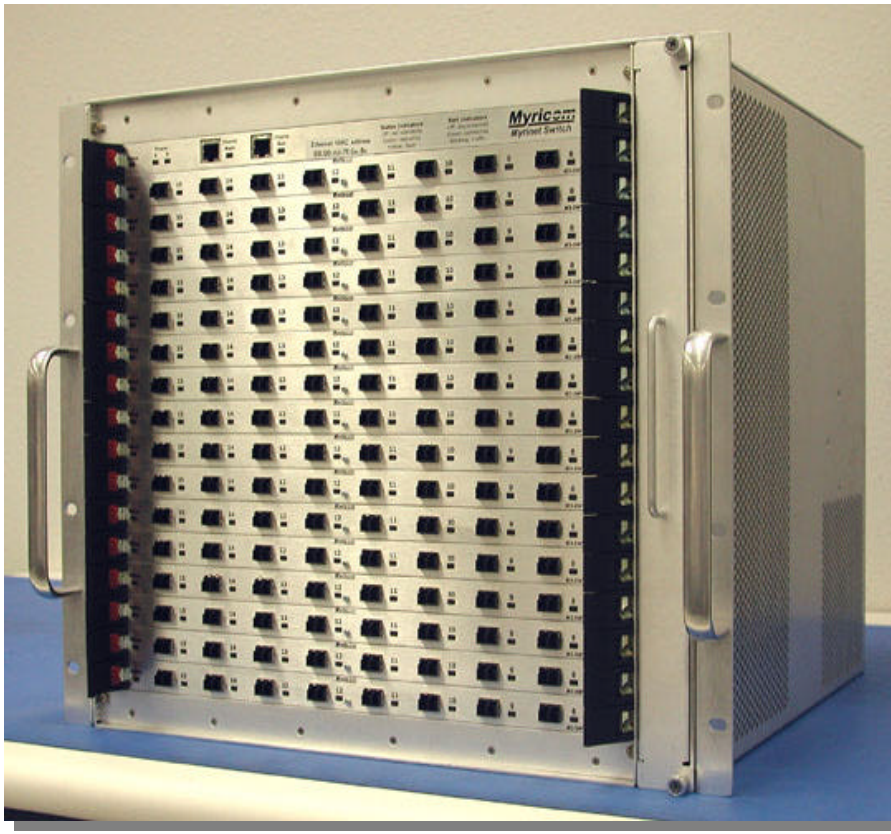


M3F-PCIXD-2 Myrinet-Fiber/PCI-X NIC with a standard PCI faceplate
(From Myricom)

Myrinet Components (2/2)

- *Myrinet switches*





Myrinet switch with Fiber ports

(From Myricom)

Myrinet Software

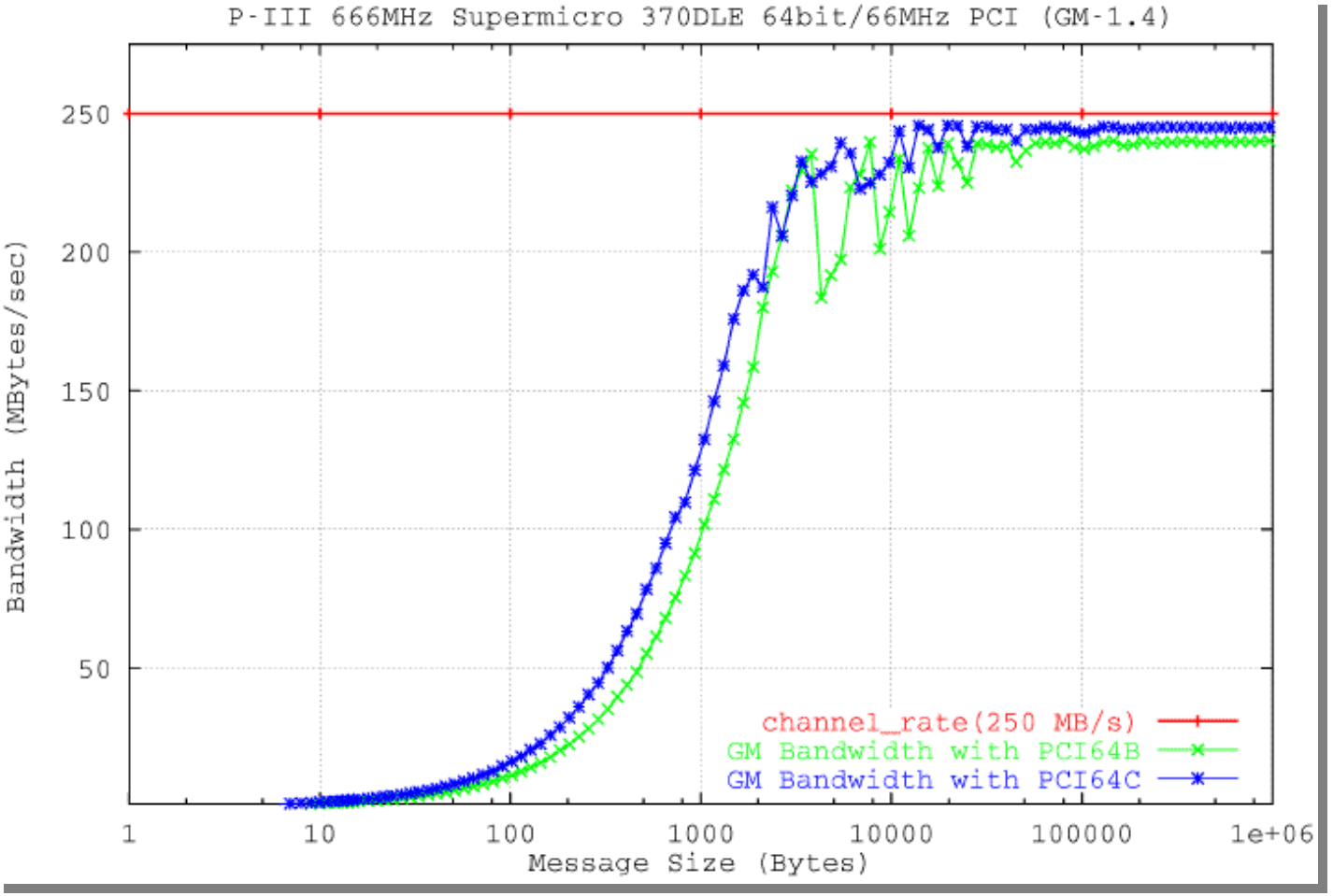
- *Myrinet GM - the low-level message-passing system for Myrinet networks*
  - *Concurrent, protected, user-level access to the Myrinet interface*
  - *Automatic mapping and route computation*
  - *Automatic recovery from transient network problems*
  - *Scalability to thousands of nodes (> 10000)*
  - *Reliable, ordered delivery of messages*
- *Middleware over GM*
  - *MPICH-GM*
  - *PVM over GM*
  - *VI-GM*
  - *IP-GM*
  - *Sockets-GM*

Myrinet Communication Performance (1/3)

- *The three principal performance metrics of cluster interconnect are:*
  - *Sustained data rate for large messages*
  - *Latency for short messages*
  - *Host-CPU utilization per message*

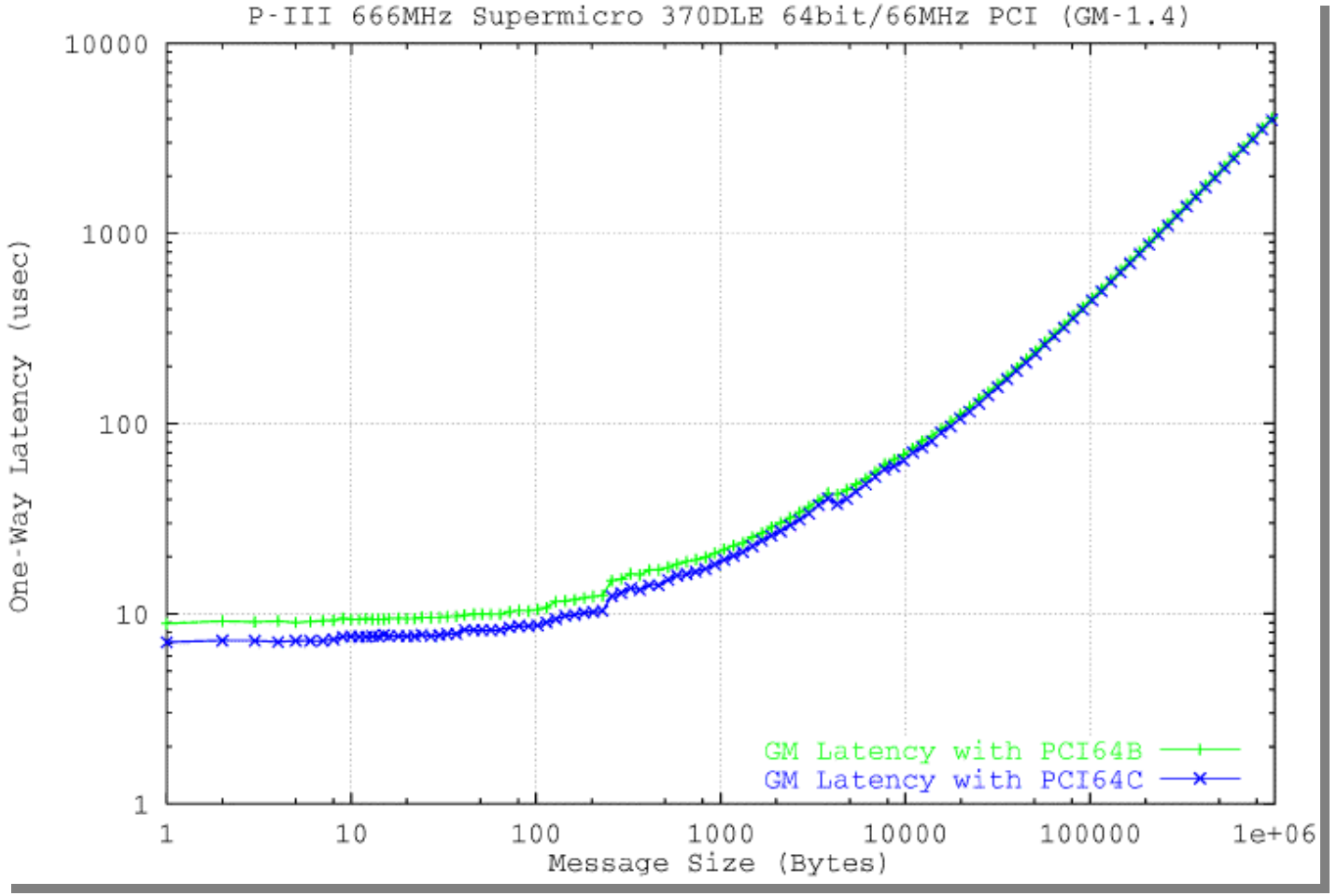
	64-bit/66MHz PCI slots	64-bit/33MHz PCI slots	32-bit PCI slots
one-way data rate	~240 MB/s (~1.9 Gb/s)	~200-240 MB/s (1.6-1.9 Gb/s)	~100-130 MB/s (0.8-1.0 Gb/s)
short-message latency	~7-9 μs	~9-11 μs	~11-16 μs
Host-CPU utilization per message send/receive	~0.3 μs / 0.75 μs	~0.3 μs / 0.75 μs	~0.3 μs / 0.75 μs

Myrinet Communication Performance (2/3)



(From Myricom)

Myrinet Communication Performance (3/3)



(From Myricom)

InfiniBand

- *The New Standard for Clustering & Storage*
  - *InfiniBand enables industry standard servers to provide computing & storage power*
  - *HPC clustering adoption is underway*
  - *Oracle and IBM databases with InfiniBand support*
- *Industry Standard*
  - *The highest performance interconnect through time*
  - *30 Gb/sec switches now shipping*
  - *Applications in many markets: Clustering, Storage, Embedded, etc*
  - *Enables affordable 10 Gb/sec technology*
- *PCI Express balanced 20 Gb/sec bandwidth makes InfiniBand even better*

Available InfiniBand Products

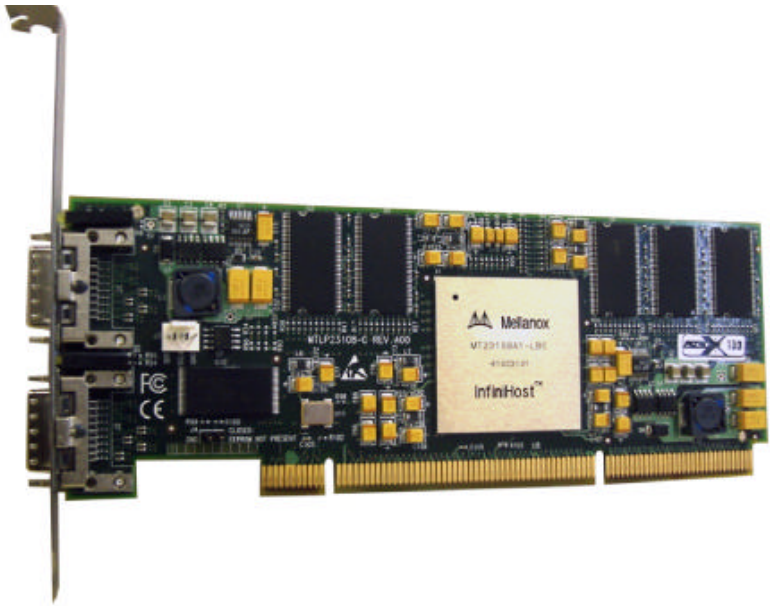
- *InfiniCon:*

  http://www.infinicon.com/
- *Mellanox*

  http://www.mellanox.com/
- *TopSpin*

  http://www.topspin.com/
- *Voltaire*
  - *http://www.voltaire.com/*

InfiniBand Components - Mellanox



Mellanox Dual-Port 10Gb/s InfiniBand HCA Card with PCI-X
(From Mellanox)

Mellanox 8-port 4X Switch
(From Mellanox)

Mellanox 144-port 4X Switch
(From Mellanox)

InfiniBand Software Stack

- *Lower-level*
  - *VAPI (Verbs-Level API) from Mellanox*
  - *Modified and customized VAPI from other vendors*
  - *IBAL (IB Access layer) with UVP (User Verb Provider) from Mellanox and Intel*

    open source: http://infiniband.sourceforge.net
  - *A new initiative of OpenIB for Linux+IBA community*

    http://www.openib.org

InfiniBand Software Stack

- *Upper-level Software stack*
  - *MPI (Message Passing Interface)*
  - *SDP (Sockets Direct Protocol)*
  - *IPoIB (IP over IB)*
  - *SRP (SCSI RDMA Protocol)*
  - *uDAPL (user-level Direct Access Provider Library)*
  - *kDAPL (kernel-level Direct Access Provider Library)*

InfiniBand Subnet Manager

- *Subnet Manager*
  - *From different vendors with different functionalities and features*
  - *MiniSM (original) from Mellanox*
  - *VFM (Voltaire Fabric Manager) from Voltaire*
  - *Fabric Manager from Topspin*
  - *InfiniView Fabric Manager from InfiniCon*
  - *OpenSM from Mellanox and Intel*
  - *Open-source initiative*

    http://infiniband.sourceforge.net

InfiniBand – openib.org

- *A new open source organization*
  - *Focusing on effort for Open Source IB support for Linux community*
  - *Design of complete software stack with best of breed components*
  - *Users should be able to download the entire stack and run without any problem*

InfiniBand V.S Proprietary Interconnects

	Interface		
	InfiniBand	Myrinet	Quadrics
Small Packet (MPI)	~5.5 μs	~7.3 μs	~5.0 μs

		Interface		
		InfiniBand	Myrinet	Quadrics
Fabric Size	8	110 ns (1 hops)	500 ns (1 hops)	225 ns (1 hops)
	128	550 ns (5 hops)	1500 ns (3 hops)	1225 ns (5 hops)
	256	770 ns (7 hops)	3500 ns (7 hops)	2475 ns (11 hops)

From: InfiniCon http://www.infinicon.com/

Network Interfaces Overview

- *Ethernet family (CSMA/CD):*
  - *Ethernet*
  - *Fast Ethernet*
  - *Gigabit Ethernet*
  - *10G Ethernet*
- *OS Kernel bypass:*
  - *Dolphin*
  - *Quadrics*
  - *Myrinet*
  - *InfiniBand*

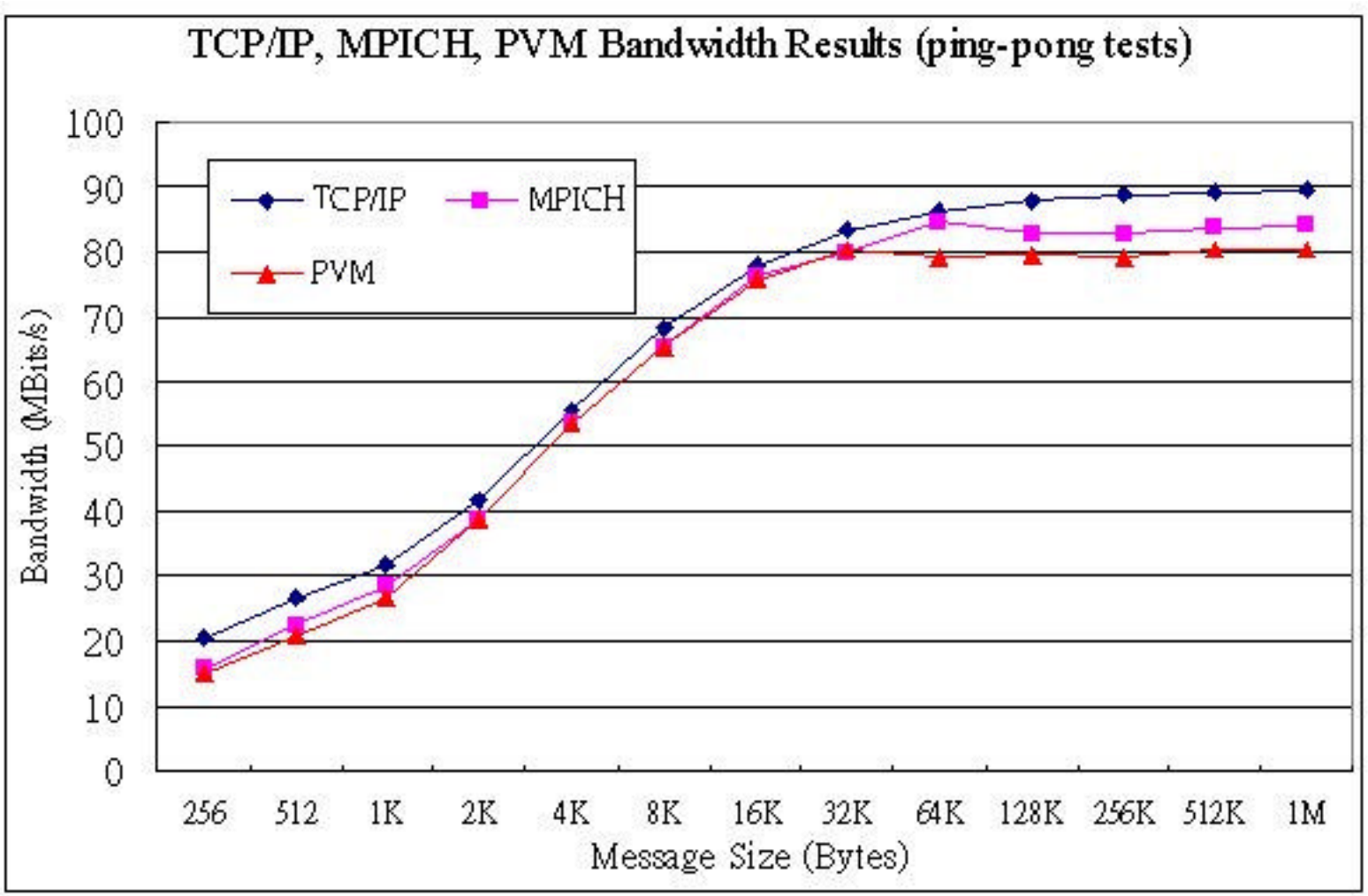
Communication Software

- *Traditional OS supported facilities (heavy weight due to protocol processing)*
  - *Sockets (TCP/IP), etc.*
- *Light weight protocols (User Level)*
  - *Myrinet GM (Myricom)*
  - *Elanlib (Quadrics)*
  - *Active Messages (Berkeley)*
  - *Fast Messages (Illinois)*
  - *U-net (Cornell)*
  - *XTP (Virginia)*
- *System software can be built on top of the above protocols*

Communication Performance (1/7)

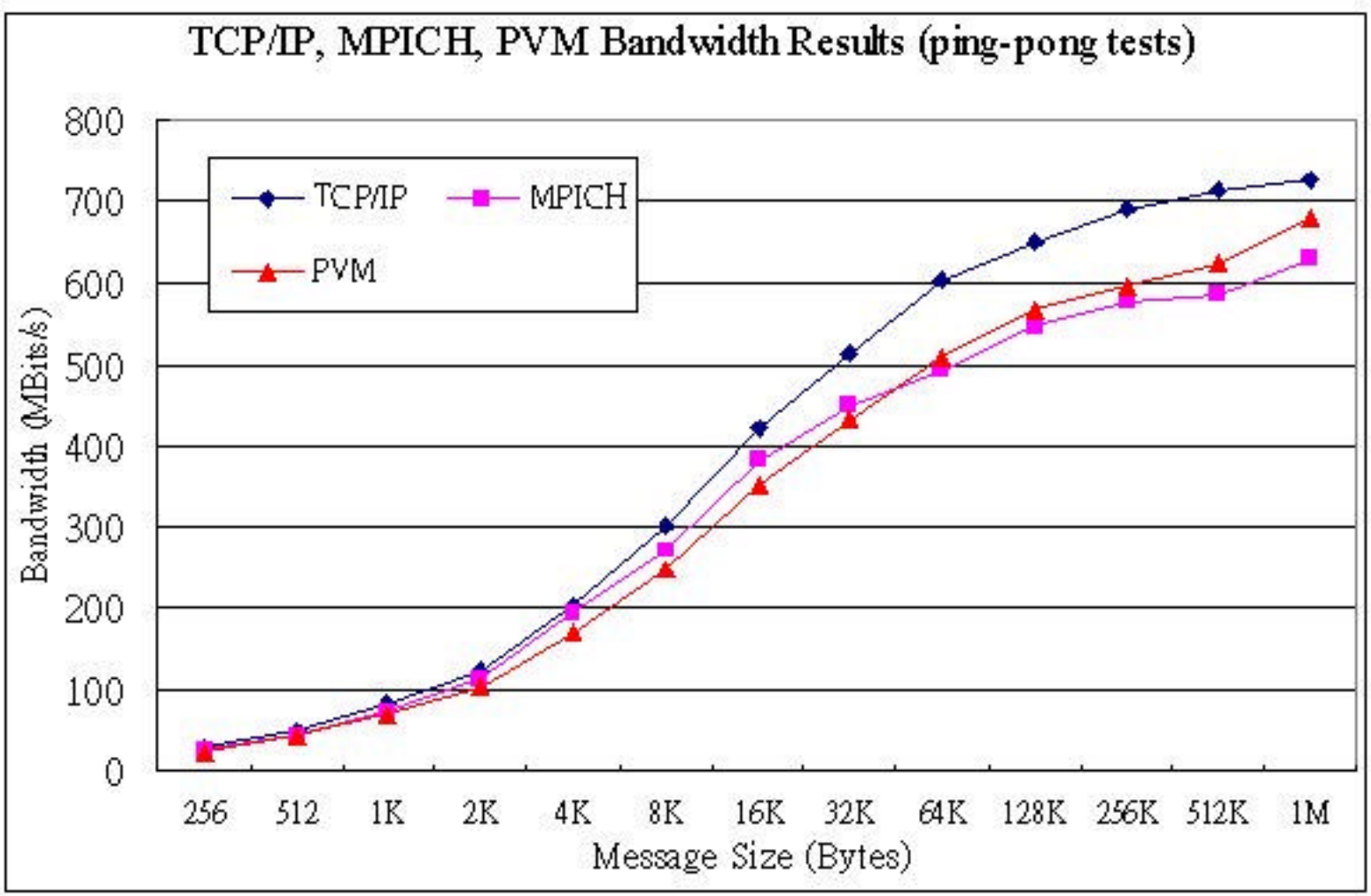
- *NetPIPE - Network Protocol Independent Performance Evaluator*
  - *Clearly shows the overhead associated with different protocol layers*
  - *Provided with protocol-specific shims for TCP/IP, PVM, TCGMSG, SHMEM, GM, MPI-2, GPSHMEM, ARMCI, and LAPI*
- *PMB - Pallas MPI Benchmarks*
  - *Compare the performance of various computing platforms or MPI implementations*

Communication Performance (2/7)



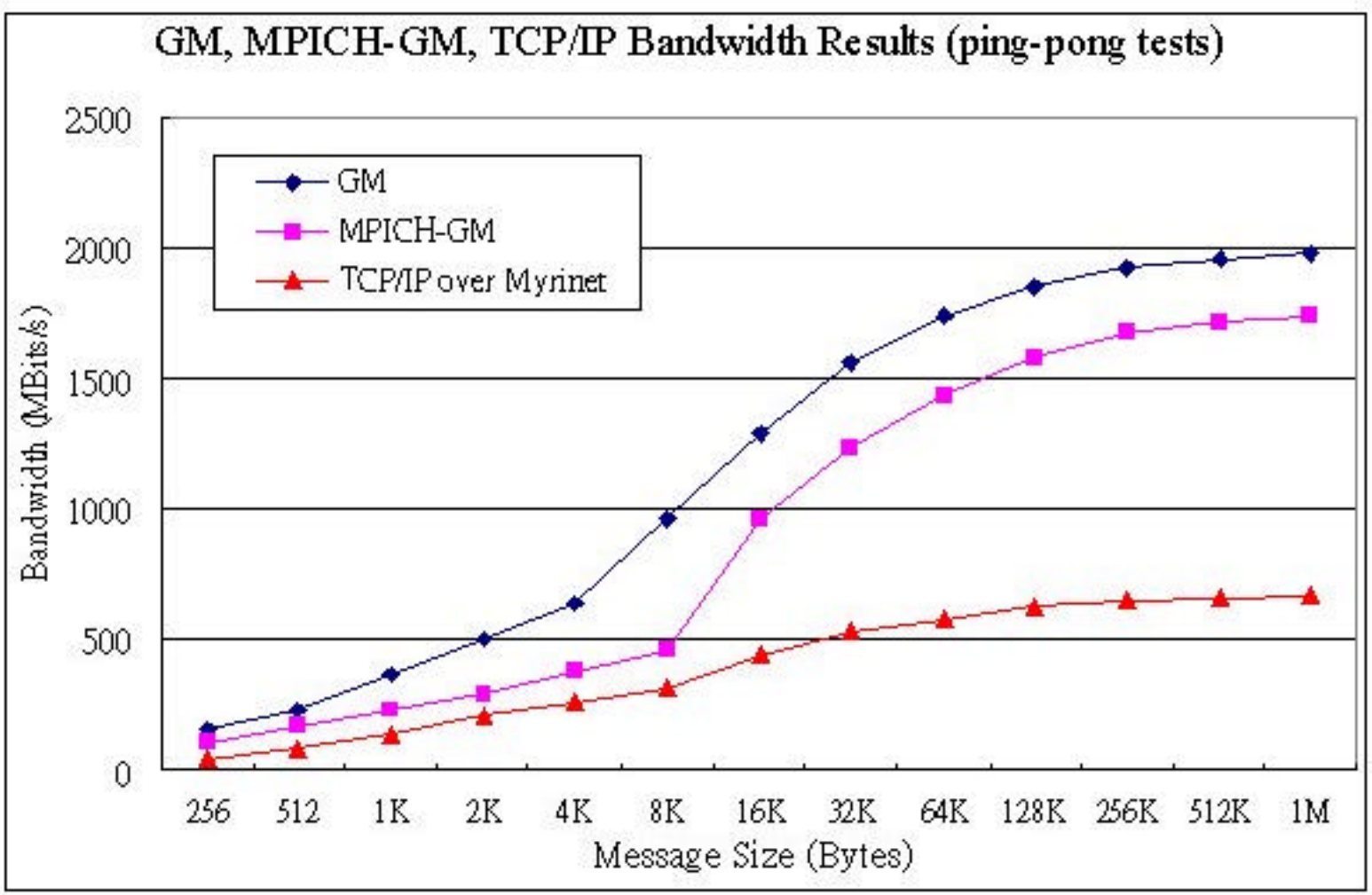
TCP/IP, MPICH, PVM Bandwidth over Fast Ethernet

Communication Performance (3/7)



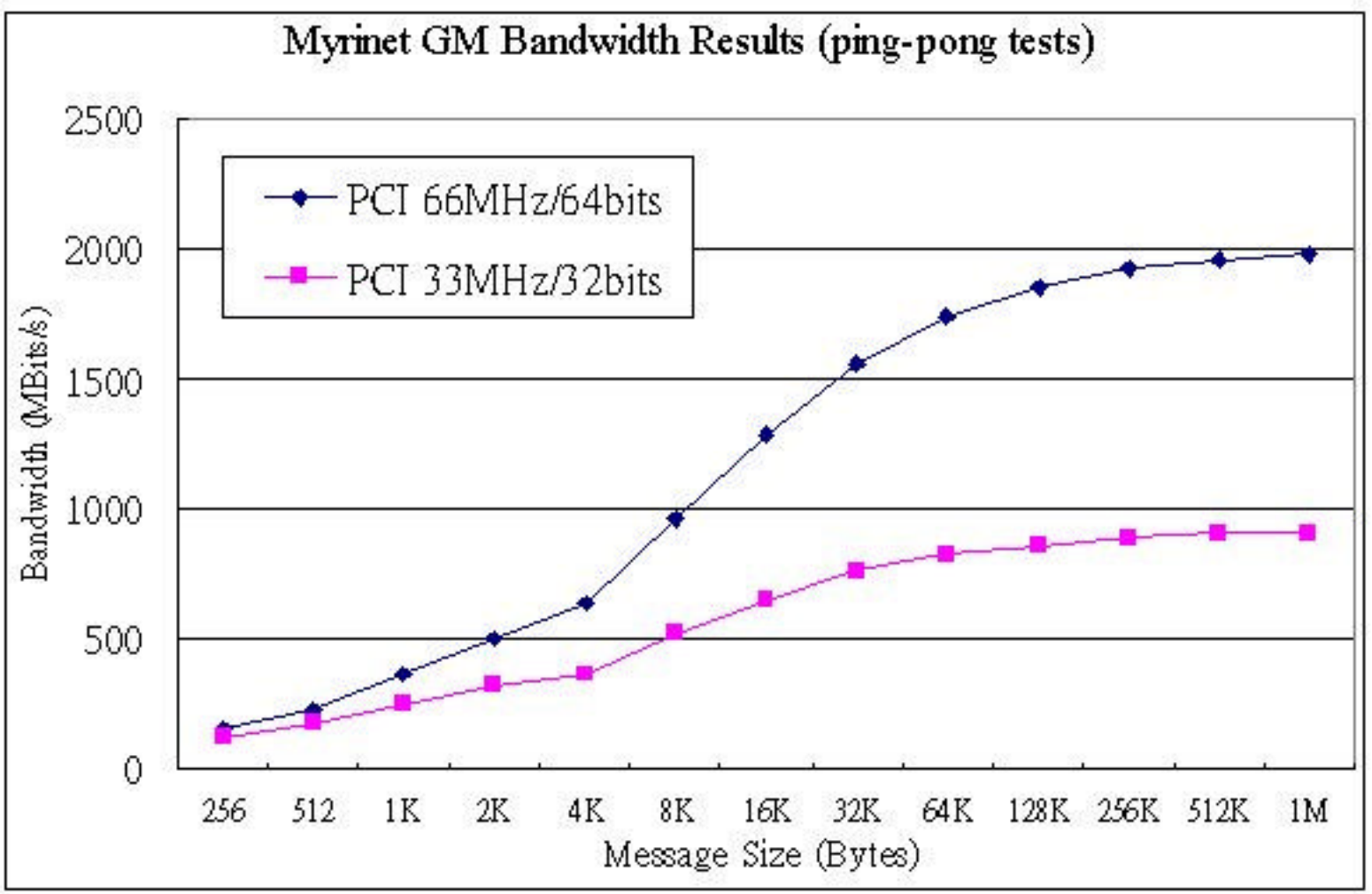
TCP/IP, MPICH, PVM Bandwidth over Gigabit Ethernet

Communication Performance (4/7)



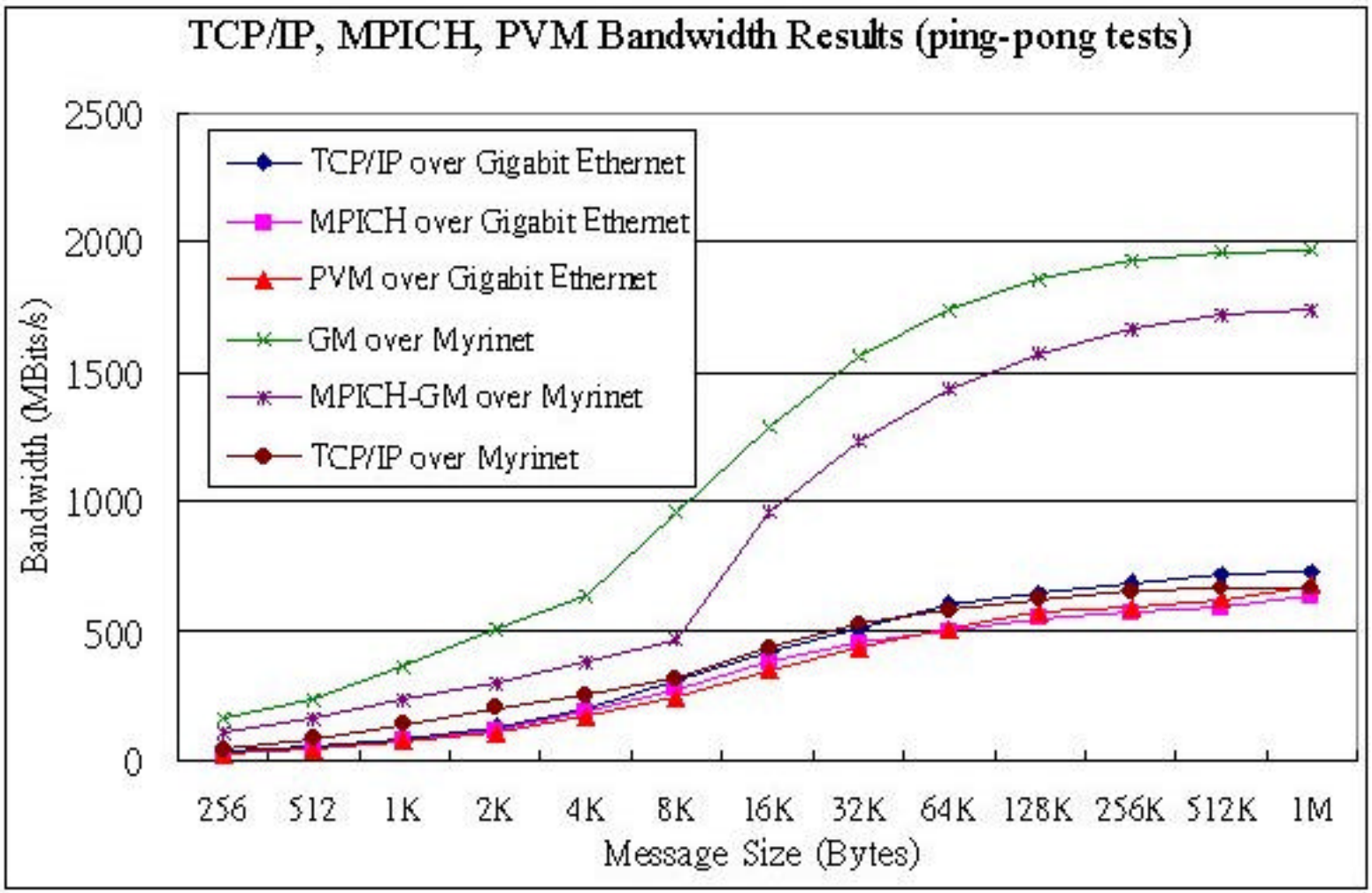
GM, MPICH-GM, TCP/IP Bandwidth over Myrinet

Communication Performance (5/7)



Myrinet GM Bandwidth – PCI 66/64 and PCI 33/32

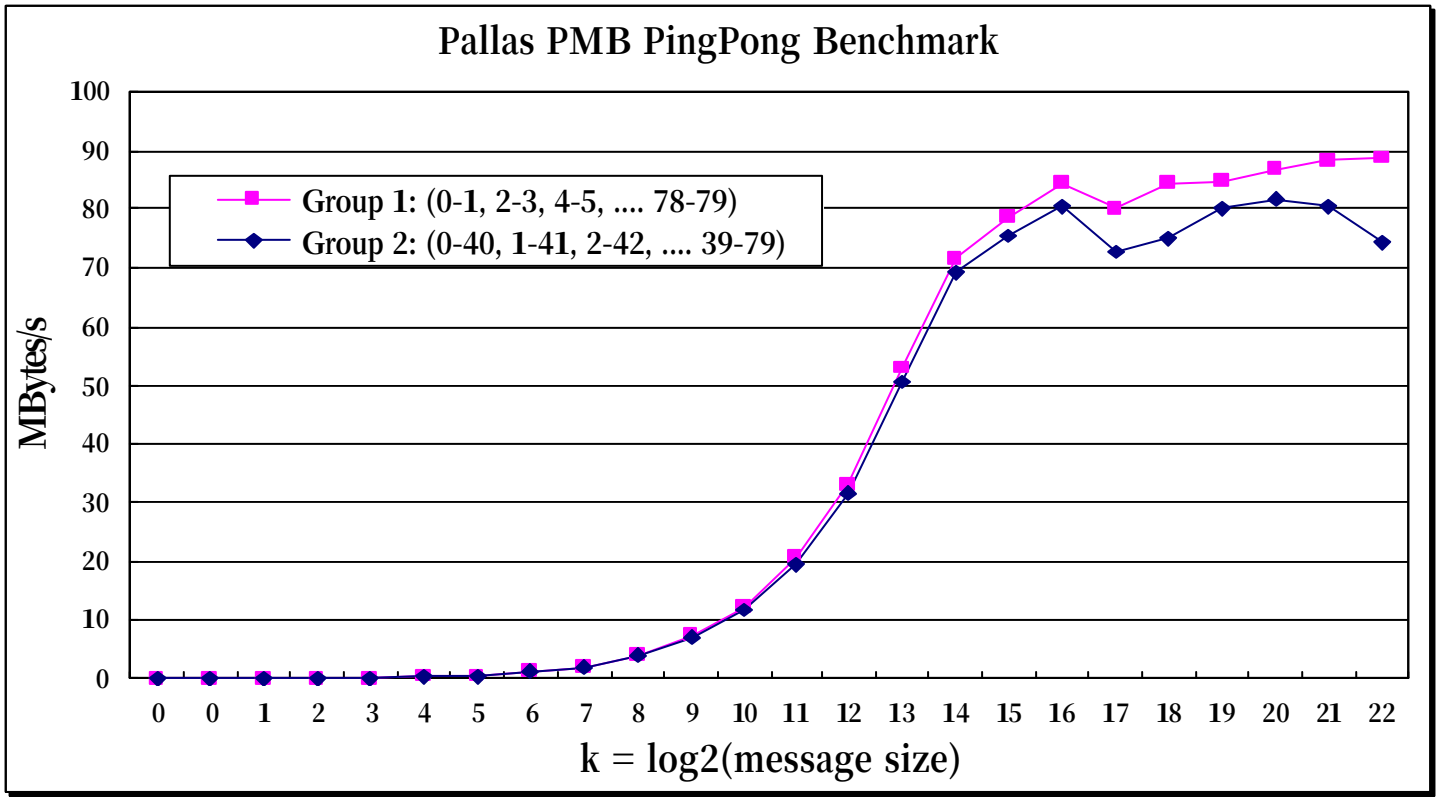
Communication Performance (6/7)



GM, TCP/IP, MPICH, PVM Bandwidth over Myrinet and Gigabit Ethernet

Communication Performance (7/7)

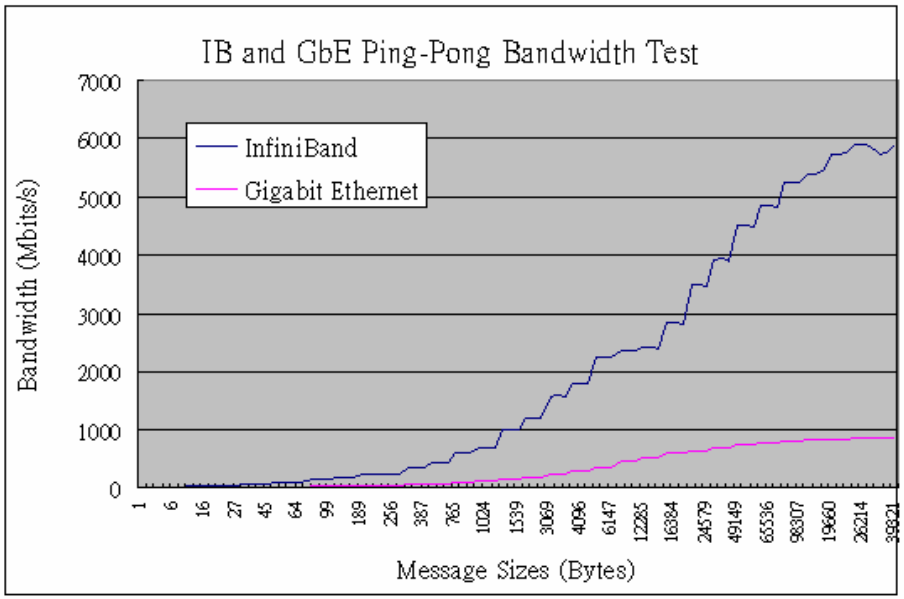
- *Simultaneously performing the PMB ping-pong between 40 pairs of processors*
  - *Group 1: (0-1, 2-3, 4-5, ... 78-79)*
  - *Group 2: (0-40, 1-41, 2-42, ... 39-79)*

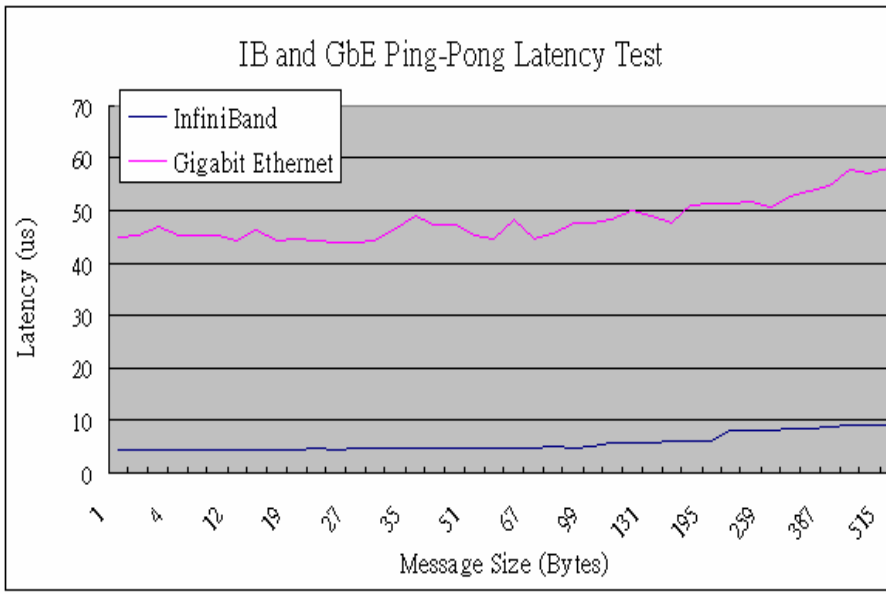


Performance in Mbytes/s for the MPI ping-pong benchmark

InfiniBand Performance (1/2)

- Platform – Mellanox PCI-X IB card on AMD64 cluster

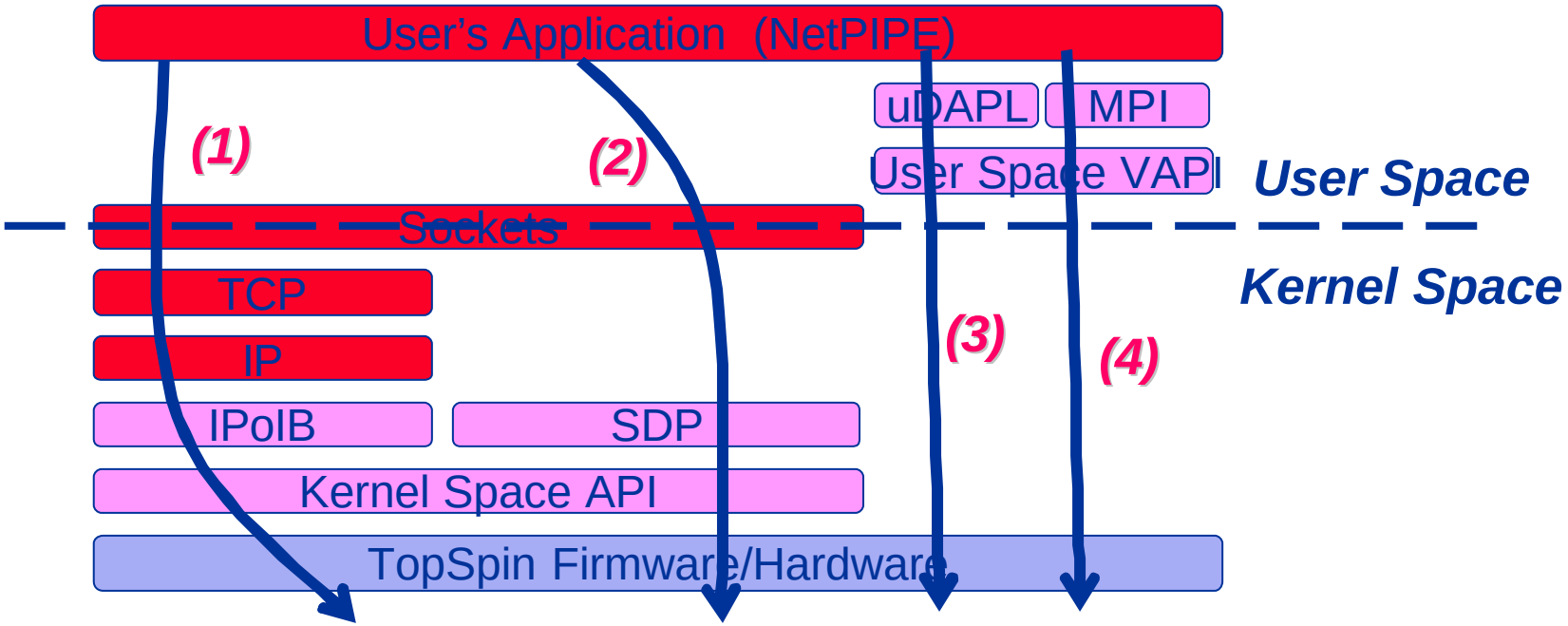




InfiniBand and Gigabit Ethernet Bandwidth & Latency

InfiniBand Performance (2/2)

- *Platform – TopSpin PCI-X IB card on EM64T cluster*



Path	(1)	(2)	(3)	(4)
Bandwidth (Mbits/s)	1828	2468	4594	4998
Latency (us)	32.0	21.5	14.2	5.3

Summary

- *Gigabit Ethernet*
  - *Fully compatible with current Ethernet.*
  - *NICs and Switches are now common.*
- *Dolphin*
  - *SCI standard.*
  - *356 MB/s, less than 4 μs.*
- *Quadrics*
  - *QsNet I & QsNet II*
  - *900 MB/s, less than 3 μs.*
- *Myrinet*
  - *495 MB/s, less than 5 μs.*
  - *Programmable microcontroller.*
  - *Quite popular in the research community.*
- *InfiniBand*
  - *Industry standard.*
  - *PCI Express is available.*

Resources

1) Dolphin website - http://www.dolphinics.com/
2) Quadrics website - http://www.quadrics.com/
3) Quadrics Linux home page / download - http://www.quadrics.com/website/pages/03LinuxSoftware.html
4) Myrinet website- http://www.myri.com/
5) Myrinet software and documentation download page - http://www.myri.com/scs/index.html/
6) NetPIPE - http://www.scl.ameslab.org/netpipe/
7) PMB (Pallas MPI Benchmarks) - http://www.pallas.com/e/products/pmb/
8) InfiniCon - http://www.infinicon.com/
9) Mellanox - http://www.mellanox.com/
10) TopSpin - http://www.topspin.com/
11) Voltaire - http://www.voltaire.com/