

# Overview of Common Interconnects for Commodity Clusters

Gelato ICE Conference  
San Jose, Ca  
April 26, 2006  
Doug Johnson  
djohnson@osc.edu



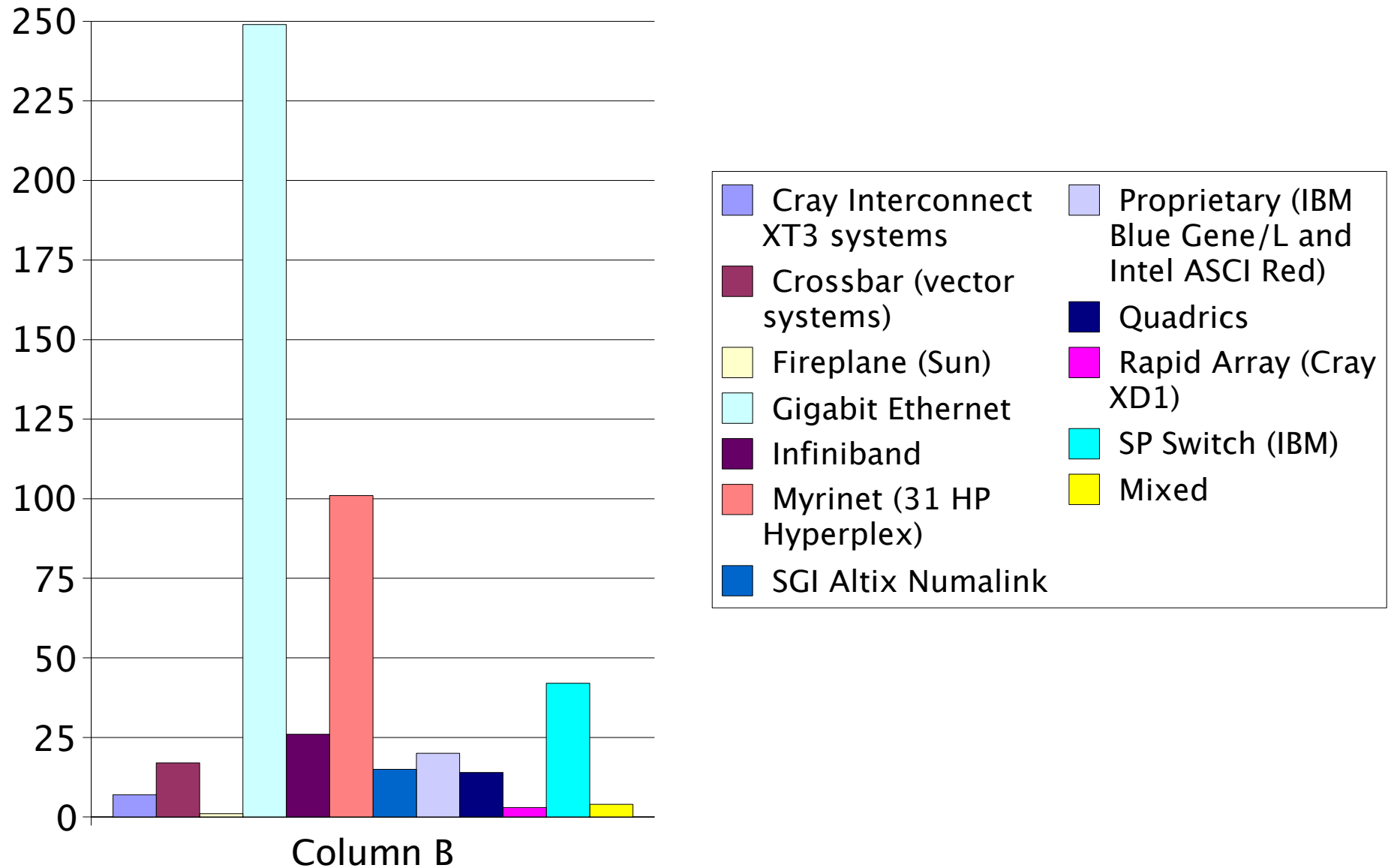
# Introduction

- Preliminaries
  - What is needed in a high speed interconnect for clusters?
    - Will constrain discussion to one facet of performance
    - Software is important (more than MPI)
- Hardware
  - Network Interfaces
  - Switch Fabrics
- Software
- Qualitative performance comparisons
- Conclusions

# Introduction (cont.)

- Need to constrain the list of commodity interconnects
  - Limited by interconnects found in systems on the Top500 list.
  - Not a measure of interconnect quality by any means.
  - Reflection of what is being deployed in large systems.
  - Leaves 4 commodity interconnects

# Top 500 List System Interconnects



# Preliminaries

- What are the important characteristics of Interconnects (for the purposes of this presentation)
  - Performance
    - LogP a model for parallel computation
      - L: latency
      - o: overhead
      - g: gap
      - P: number of processors
    - LogGP updated to incorporate large messages (throughput)
      - G: gap between bytes
    - Overhead will be emphasized
  - Software support

# Hardware

- Network Interface Cards
  - Gigabit Ethernet
  - Infiniband
  - Myrinet
  - Quadrics
- Network Fabrics

# Gigabit Ethernet NICs

- Gigabit Ethernet NICs are ‘free’
  - Typical server or workstation system board has two on-board NICs.
- Linux includes broad support for the major vendor’s chipsets.
  - Intel e1000 family (many models)
  - Broadcom
  - Syskonnect
  - National Semiconductor
  - SiS

# Gigabit Ethernet NICs (cont.)

- New transport offload Ethernet cards
  - Support RDMA/iWARP
    - Supported by OpenIB/OpenFabrics software stack
  - Gigabit and 10Gb adapters
    - Chelsio
    - Ammasso
  - Uses standard Ethernet switches

# Myrinet NICs

- High-speed, OS bypass, and low latency network cards.
- Programmable, uses a custom VLSI processor.
  - Flexibility allows the support of different protocols
- Two families of interfaces available
  - Myrinet 2000, 2Gb link speed
  - Myinet-10G, 10Gb link speed

# Myrinet NICs (cont.)

- Myrinet 2000 NIC types
  - “D”, 225MHz processor, 1 link (PCI-X)
    - Latency  $\sim 3.5\mu\text{s}$  (MX software)
  - “F”, 333MHz processor, 1 link (PCI-X)
    - Latency  $\sim 2.7\mu\text{s}$  (MX software)
  - “E”, 333MHz processor, 2 links (PCI-X)
    - Latency  $\sim 2.6\mu\text{s}$  (MX software)
- Near line-rate throughput (even with bidirectional MPI\_Sendrecv)

# Myrinet NICs (cont.)

- Myrinet-10G
  - New generation of Myrinet NICs
    - Uses standard 10Gb Ethernet as physical layer
    - PCI Express card
    - $\sim 2\mu\text{s}$  MPI latency
  - Dual protocol NIC
    - 10Gb Ethernet
    - MX (myrinet messaging primitives)
  - In 10Gb Ethernet mode, uses standard 10Gb switches
  - MX mode requires Myrinet-10Gb switches
    - MX does not use Ethernet as the network layer protocol

# Infiniband NICs

- First order network for connecting computing to I/O devices
- The NICs (host channel adapters, HCAs) come in different speeds which are multiples of 2.5Gb/s
  - 802.3x 8B/10B coding, or  $2.0 \times 10^9$  bits/s
    - 10Gb Infiniband actually 8Gb/s
  - 1, 4, and 12x HCAs are available
  - SDR and DDR for the 4x HCAs
  - Single and dual port 4x HCAs

# Infiniband NICs

- Same physical layer as 10GB Ethernet (different rate)
  - PCI-x, PCI Express, HTX and GX HCAs
    - GX is memory bus connector for IBM Power systems
    - HTX is a HyperTransport expansion slot
  - Guaranteed in-order delivery
  - Vendors to choose from
    - Mellanox
    - Topsping
    - Voltaire
    - Silverstorm (Infinicon)
    - Pathscale\*
    - Cray Rapid Array\*
- \* Rely on host CPU more than other IB NICs

# Quadrics NICs

- QsNet II
  - PCI-X
  - 10bits, 1.333Gbaud or 900MBytes/s
  - 200MHz 64 bit processor
  - 64MB memory (2.7GB/s mem bw)
- Copy of TLB is maintained on NIC
  - Removes need to ‘pin’ memory
  - Require OS patch
- Latest generation of NICs have  $<2\mu\text{s}$  MPI latency.

# Network Fabrics

- Large clusters have stressed the ability to scale networks
  - Clusters of 1000s of nodes not uncommon
- Present several difficult choices
  - Blocking or non-blocking
  - Density of switch host ports
  - Redundancy for larger fabrics
  - Which vendor (In case of Ethernet)

# Network Fabrics (cont.)

- Ethernet switches
  - Significant number of vendors to choose from
  - Latencies and performance vary greatly
    - Switches may be internally oversubscribed or packet rate limited
    - Range is  $<500\text{ns}$  – on the order of  $10\mu\text{s}$ .
  - Wide range of port densities
    - 1U 24–48 gige ports ( $< \$100/\text{port}$ )
    - 1260 ports, full bisectional bandwidth
  - Large port count and redundant fabrics can be designed

# Network Fabrics (cont.)

- Ethernet (cont.)
  - 10Gb becoming common
    - 1–4, 10Gb uplink ports in 48 gige port switches
    - Inexpensive 8, 24 10Gb port switches
    - Large chassis support 28–224 10Gb ports
    - Cost is still an issue
      - CX4 is cheapest, 15M length limit
      - Optics range from 26M on MM to ~100km on SM
  - 10Gb switches from non-traditional vendors
    - Quadrics QsTenG 96 port switch
    - Future Myrinet switches will support 10Gb Ethernet ports

# Network Fabrics (cont.)

- Myrinet
  - Switches from 8 – 256 host ports (128 ports for Myri10G)
  - Up to 4096 port networks can be constructed
  - Switch latency
    - ~500ns small Myrinet-2000 switch
    - ~1.1 $\mu$ s large Myrinet-2000 switch
    - ~200ns Myri-10G switch

# Network Fabrics (cont.)

- Infiniband
  - Switches from 10 – 288 host ports
  - Fat tree topologies
  - > 4000 port networks have been constructed
  - Protocol routing switches are available
    - Ethernet and FC network integration
  - Switch latencies
    - 288 port switch – 420ns
    - 24 port switches – 140ns

# Network Fabrics (cont.)

- Protocol routing switches are available
  - Ethernet and FC network integration
- Switch latencies
  - 288 port switch – 420ns
  - 24 port switches – 140ns
- Quadrics
  - Switches and fabrics are fat-tree
  - Switches from 8 – 128 host ports
  - 4096 host port networks possible

# Software

- Low level software
- MPI
- SDP
- SRP
- iSER/iSCSI
- NFS over RDMA
- Others

# Software (cont.)

- Low level software
  - Ethernet
    - TCP,UDP/IP
  - Myrinet
    - MX
  - Infiniband
    - VERBS
      - Sometimes this is an API (VAPI), other times not
  - Quadrics
    - QsNet libs

# Software (cont.)

- Comments about Ethernet and Linux
  - Driver and kernel must be involved
    - Context switches and extra copies need to be minimized
      - sendfile, Van Jacobsen channels, NAPI
  - Need to ensure the chipset for your system supports the following
    - Zero-copy (scatter gather)
    - Jumbo frames
    - CRC offload
    - TCP segmentation offload (TSO)
  - Sufficient TX and RX descriptors

# Software (cont.)

- MPI1 or MPI2
  - Vendors typically support subset of MPI2 such as parallel I/O and one-sided communications
  - Not always possible to map efficiently onto network layer protocol
    - Collectives
    - Asynchronous sends/receives
  - Matching on the NIC for MPI\_Isend and MPI\_Irecv
    - Without matching an extra process must poll (or driver is entered)

# Software (cont.)

- Vendors with ‘true’ asynchronous MPI send and receive for all message sizes
  - Myricom (MX)
  - Quadrics
- IP over high-speed interconnect
  - Myricom, Infiniband, Quadrics
  - Must be tuned in the same fashion as traditional Ethernet adapters

# Software (cont.)

- SDP (Socket Direct Protocol)
  - Allows socket applications to bypass OS stack
    - Recompile or LD\_LIBRARY\_PATH
  - Supported on Myrinet, Infiniband, and Quadrics
- SRP (SCSI Remote Direct Protocol)
  - Maps SCSI onto RDMA for Infiniband
  - Direct Infiniband connections to storage controllers
    - DataDirect Networks
    - Engenio

# Software (cont.)

- iSER/iSCSI
  - Allows SCSI to be layered over TCP/IP or RDMA
  - May be more prevalent in the future than SRP
- uDAPL
  - Common API for RDMA interfaces. Intends to be used by upper layer protocols such as MPI, SRP, etc..
  - Supported by OpenIB (Infiniband, and iWARP) and Myrinet.
  - Used by Intel MPI
- NFS over RDMA
  - Layered on uDAPL

# Software (cont.)

- Filesystems
  - Lustre
    - Ethernet, Myrinet, Infiniband, Quadrics
  - PVFS
    - Infiniband – RDMA
    - Myrinet, Quadrics – TCP/IP
  - Others?
    - Commercial NAS

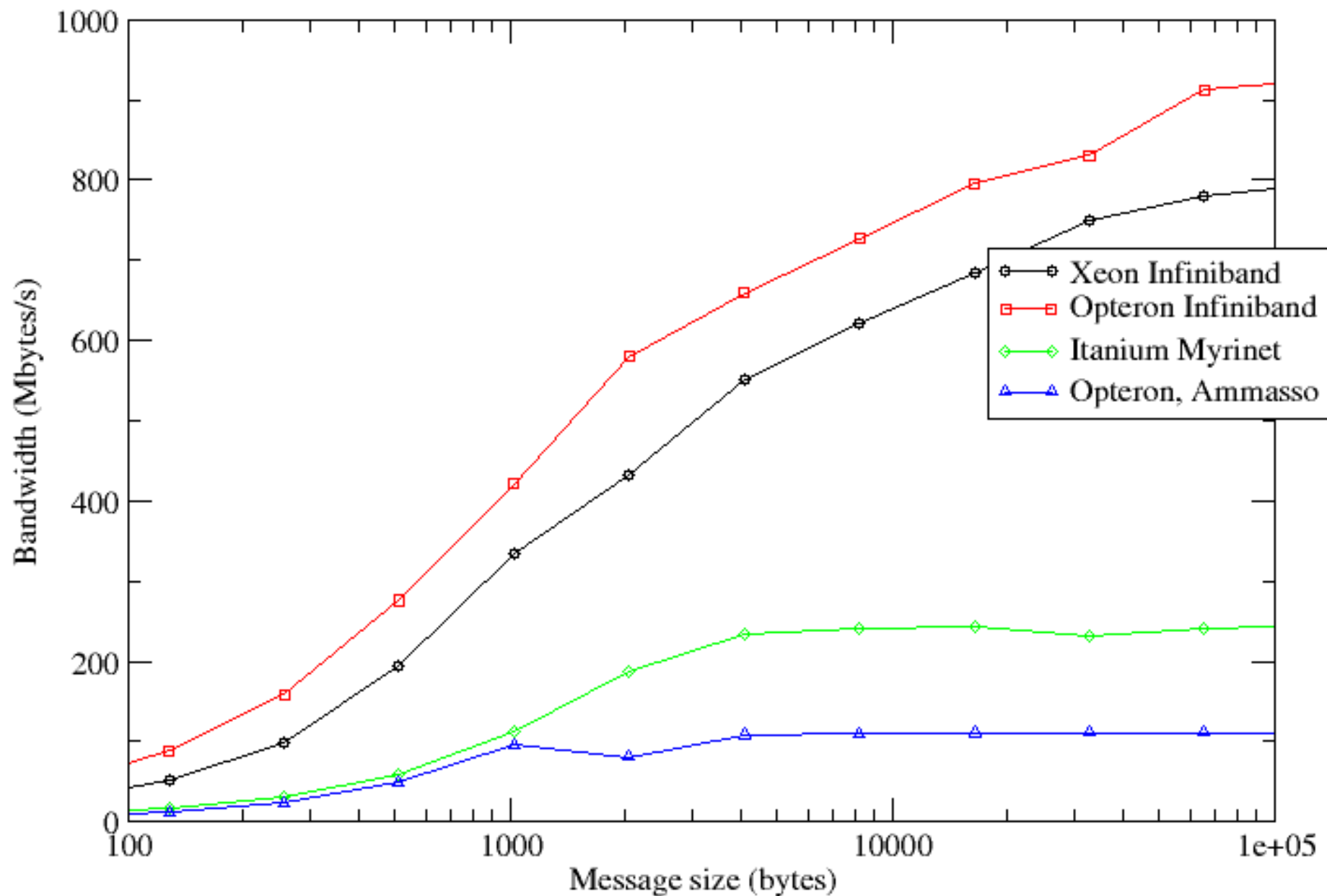
# Qualitative Performance Comparisons

- Four interconnects compared
  - Intel Xeon with Infiniband
    - Mellanox Technologies MT23108 InfiniHost (rev a1) PCI-x 4x sdr
    - Topspin 540 switch
  - AMD Opteron with Infiniband, Ammasso
    - Mellanox Technologies MT25204 [InfiniHost III Lx HCA] (rev a0)
    - Mellanox MTS 14400-144 – 8, 12port line cards
      - IB-Gold-1.8.0

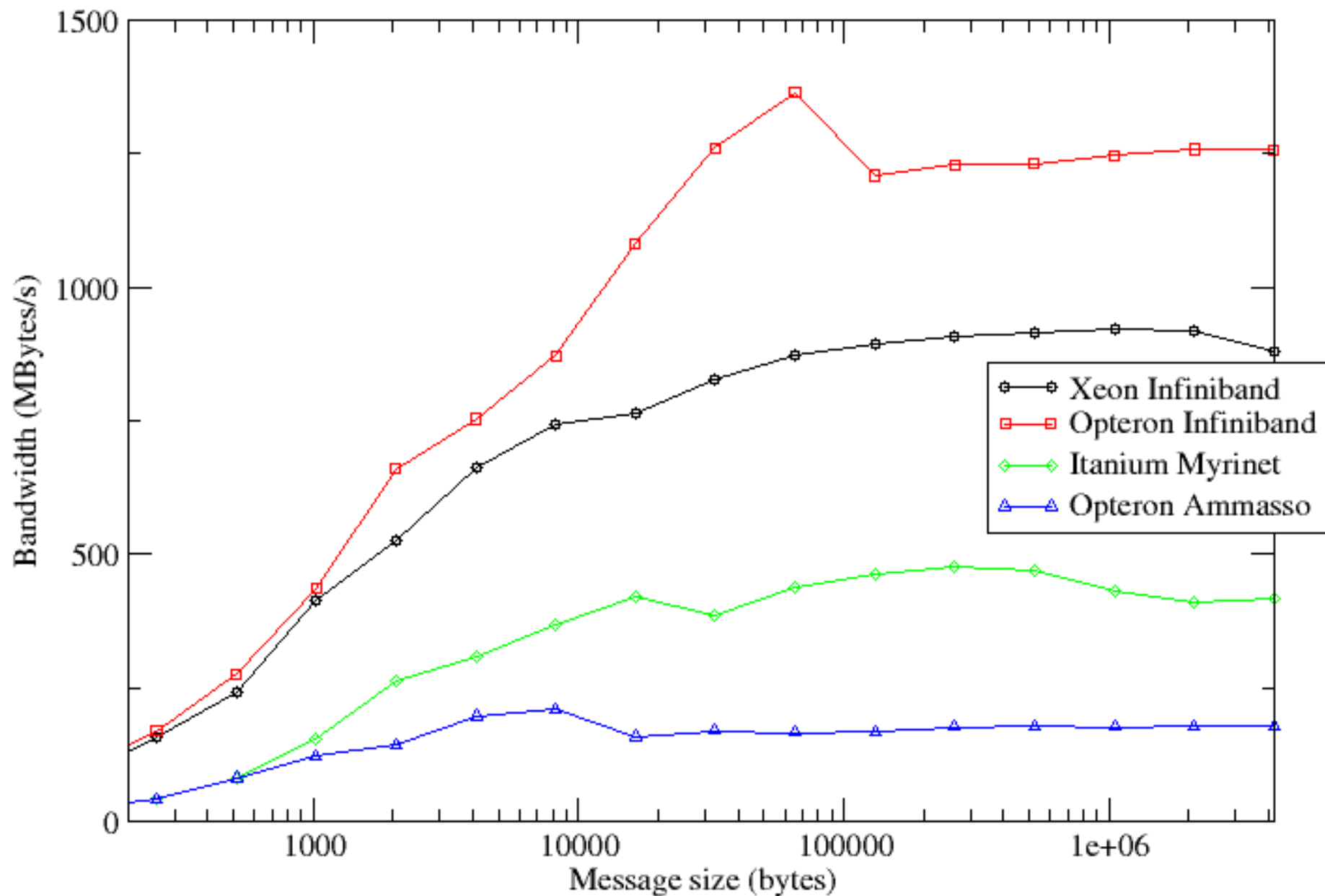
# Qualitative Performance Comparisons (cont.)

- Intel Itanium with Myrinet
  - Myrinet-2000 “C” NICs (202MHz lanai)
  - gm-2, mpich-1.2.5..10
    - Software on production system, over 1.5 years since upgrade
- AMD Opteron with Ammasso NIC
  - Ammasso 1100 NIC
    - OpenIB
  - SMC 2624T Gigabit Ethernet switch
- Tests use MPI\_Isend, MPI\_Irecv
  - <http://nowlab.cse.ohio-state.edu/>

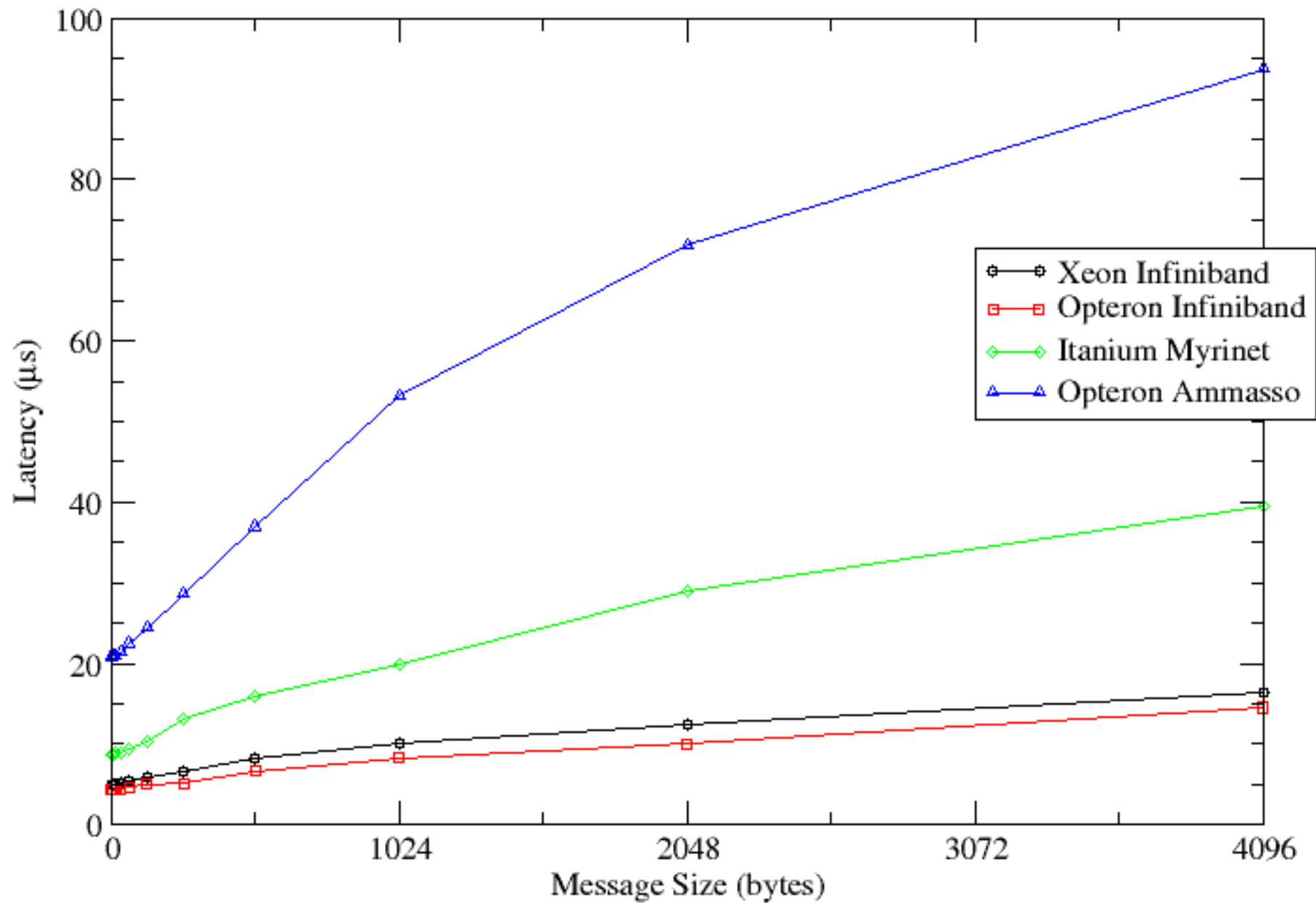
# Point-to-Point Bandwidth



# Bidirectional Bandwidth



# Point-to-Point Latency



# Conclusion

- High-speed interconnects for clusters should support several concurrent upper layer protocols
  - Overhead likely important
- Network fabrics should support needs of system area network
  - More than compute node interprocess communication
  - Connectivity to other networks or devices will be useful

References in 'notes' sections of online copy of slides.