Battle of the Network Stars!

Panel at SC '03

Dhabaleswar K. Panda
Department of Computer and Info. Science
The Ohio State University

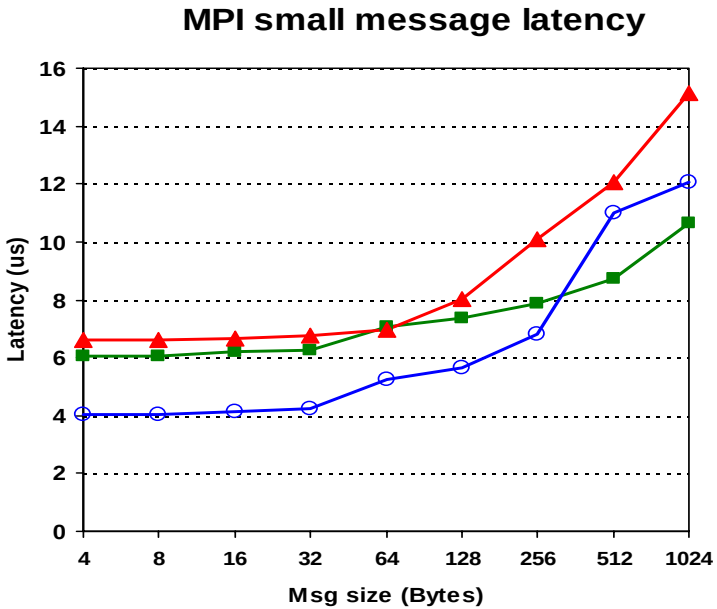
E-mail: panda@cis.ohio-state.edu
http://www.cis.ohio-state.edu/~panda

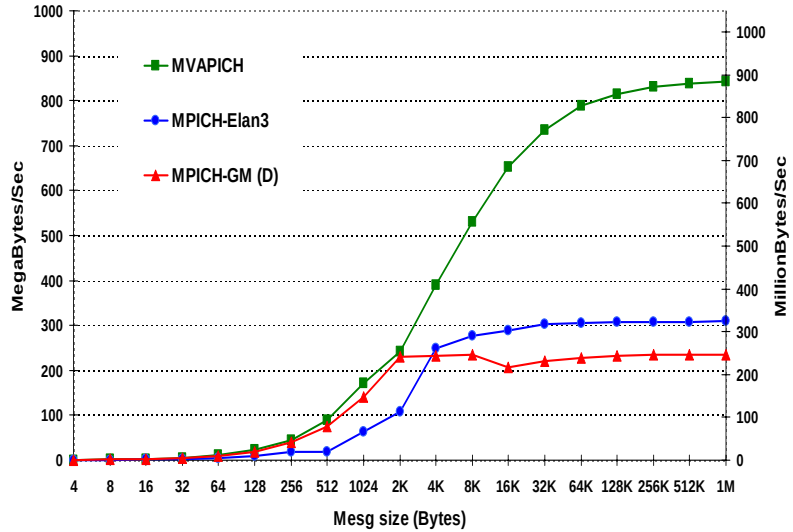
Which Interconnect is the best for High-Performance Computing and Why?

- Performance (latency, bandwidth, and CPU utilization) and Cost
  - High Performance with High Cost (HPHC)
  - High Performance with Low Cost (HPLC)
  - Low Performance with Low Cost (LPLC)
  - Low Performance with High Cost (LPHC)
- Candidates
  - Gigabit Ethernet (1/10)
  - InfiniBand
  - Myrinet
  - Quadrics

	HC	LC
HP	Quadrics	InfiniBand (Lowest) Myrinet (Lower)
LP	10.0 GigE	1.0 GigE

MPI-level Latency and Bandwidth Comparisons(3.0 GHz Xeon)





- MPI/IBA (MVAPICH 0.9.2) gives **6.0** us latency and **843** MBytes/sec
- MPICH/GM **6.3** us latency and **250** MBytes/sec (**500** MBytes/sec with dual-port E-cards)
- MPI/Quadrics (Elan3) **4.0** us latency
- MPI/Quadrics (Elan4) around **2.4** us latency and around **908** MBytes/sec(Fabrizio)
- 10.0 GigE, at the sockets level (**10** us latency and around **916** MBytes/sec (Wu))

Cost

- Actual cost is very difficult to determine. List prices can be used for comparison
- Adapters and Switch ports
  - InfiniBand
    - around $700/adapter and $350-400/port
  - Myricom
    - $1195 for the latest E-cards, per port cost depends on the system configuration
  - Quadrics
    - $1200/Elan4 adapter, $1.5-2K per port (small systems), $2.5K per port (larger systems)
  - 10.0 GigE
    - around $5.2K/adapter, $10K per module (1-2 ports)
- Putting together large-scale clusters with InfiniBand
  - 1100-node (2200 processors) VT cluster with $5.5 M – 10.28 TFlops
  - 192-node (384 processors) TeraFlop (TOTS) demo with less than $1M

Future Trends (next 2-3 years)

- Quadrics will remain in the High Performance High Cost category
- InfiniBand will remain in the High Performance Lowest Cost Category
  - 12X (30.0 Gbps), PCI-Express are coming up
  - Commodity and open standard will continue to reduce the price further
- Myrinet (with MX software) will be in the High Performance Lower Cost Category
- If 10.0 GigE with good off-load engine can be designed, it will enter the High Performance Lower Cost category

Other Questions

- Evaluation of interconnects
  - Latency and bandwidth is not sufficient
  - Many other measures (overhead, overlap, buffer reuse, …) needs to be done
  - SC '03 and Hot Interconnect '03 paper from OSU
- RDMA/TCP/10GigE
  - Has good potential
  - By the time solutions are ready, InfiniBand would have moved to 12X (30 Gbps)
- PCI-X vs. PCI-Express
  - Will have impact
  - High-end servers will move to PCI-Express (provided it delivers all its promises), low-end systems will have PCI-X
  - All network adapters will quickly move
  - Mellanox has already announced that InfiniBand HCAs with PCI-Express will be available in 1Q 2004
- Sockets interface
  - Good for datacenters, Not for HPC