



# Performance Acceleration Through Hardware Protocol Processing and Remote DMA

2003 S&CS and Ongoing Computing  
PI Strategic Planning Meeting

Half Moon Bay

February 12, 2003



# Contributors and Collaborators

Helen Y. Chen, 8961

Neal Bierbaum, 8961

Frank Bielecki, 8941

Jamie Van Randwyk, 8941

Matt Leininger, 8961

Dhabaleswar K. Panda, Ohio State University

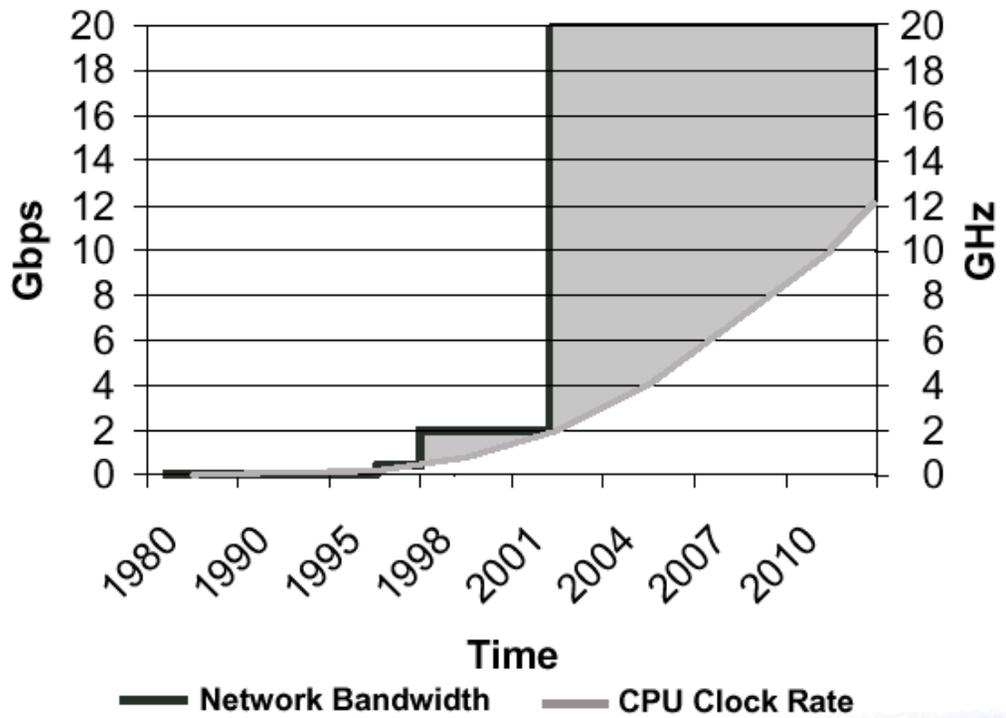
Pete Wyckoff, Ohio Supercomputing Center

Ron Minich, LANL

# Project Objective and Relevance

- The goal of this work is to fill the performance gap between networking hardware and applications
  - The disparity between physical bandwidth and “deliverable” end-to-end bandwidth continues to be one of the most challenging problems faced in building networks of ASCI components
- TCP offload Engine is essential to Red Storm in meeting its 50 Gbps network I/O requirements
- InfiniBand (IB) is the best commodity solution to delivering next generation, mid-range, capacity computing

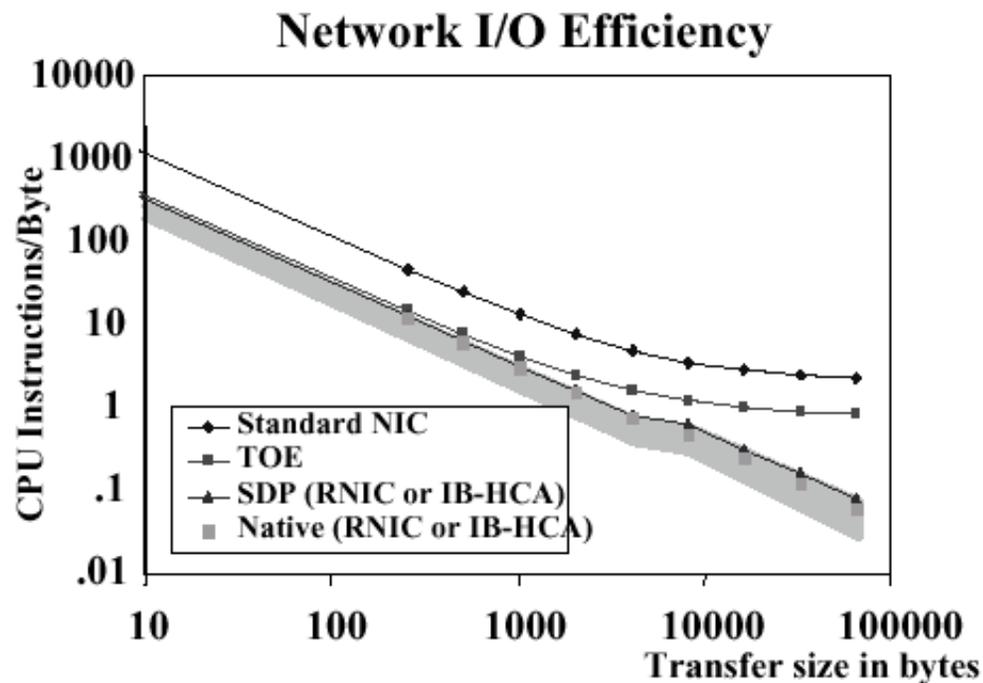
# WHY TCP Offload ?



Courtesy: IBM Server I/O team leader

- TCP can saturate 1 GHz of CPU per 1 Gbps line speed
- Networking speeds outpacing server and storage speeds
- TCP/ IP offload essential at 10 Gbps

# 10 GigE also need RDMA



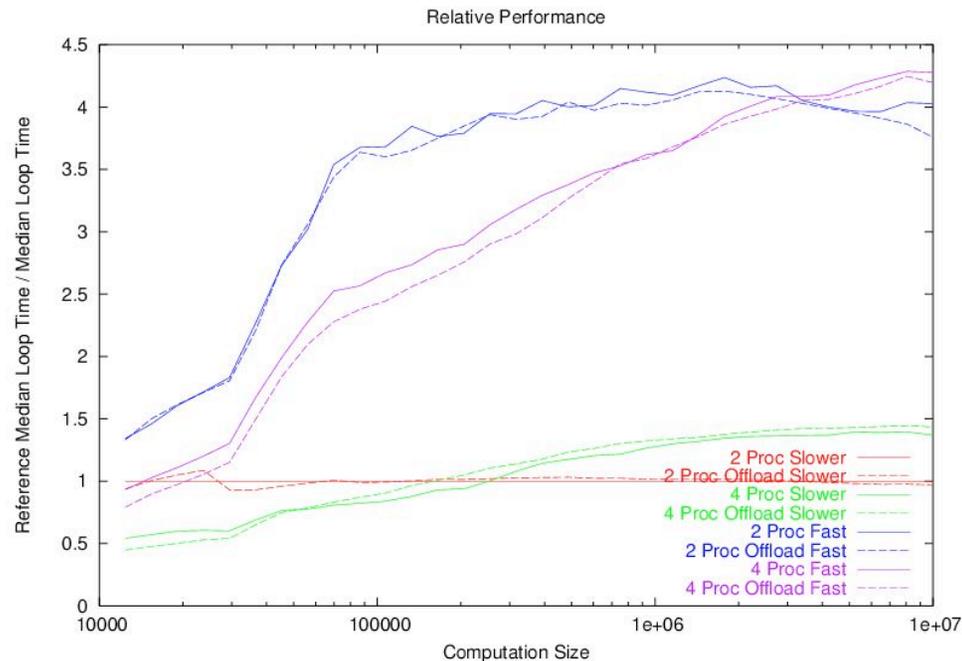
- Network stack processing consumes memory bandwidth (3x on receives)
  - 3 GB/s required to support 10 GigE per copy
- TOE removes all but the receive side copy

Courtesy:, IBM Server I/O team leader

## Early TOE Experience

- Transparent, fully Embedded TCP/IP is hard
  - Must retain kernel stack for other external and internal interfaces
  - TOE architecture is critical to scalability
- Many vendors are targeting higher level protocol interfaces such as iSCSI

# TOE and MPI Performance



Published at IEEE LCN workshop Nov, 2002

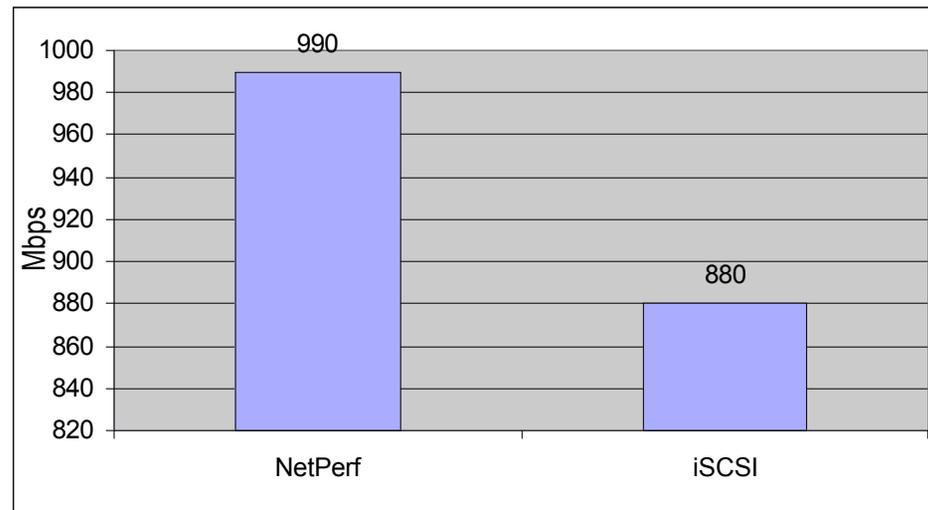
- The Alacritech TOE
  - Data movement in ASIC on adapter
  - Control and exception handling using host stack
- No performance gain on dual 1GHz processor compute nodes
  - MPIpro calculations on a 2-d array with overlapping compute and communication
- Benchmark tools available
  - [nrbierb@sandia.gov](mailto:nrbierb@sandia.gov)

# Why iSCSI

- A better cost/performance alternative to FC SAN (economy of scale)
- Distance advantage suitable to ASCI DISCOM applications?
- Well known management and security software

# iSCSI Protocol Performance

- Mismatch in driver support precluded TOE evaluation
  - Syskonnet NIC's and 1GHz Pentium processors
  - Intel user level drivers for initiator (SCSI client) and target (SCSI server)
  - Cisco kernel driver for initiator



# TOE FY03-04 Plans

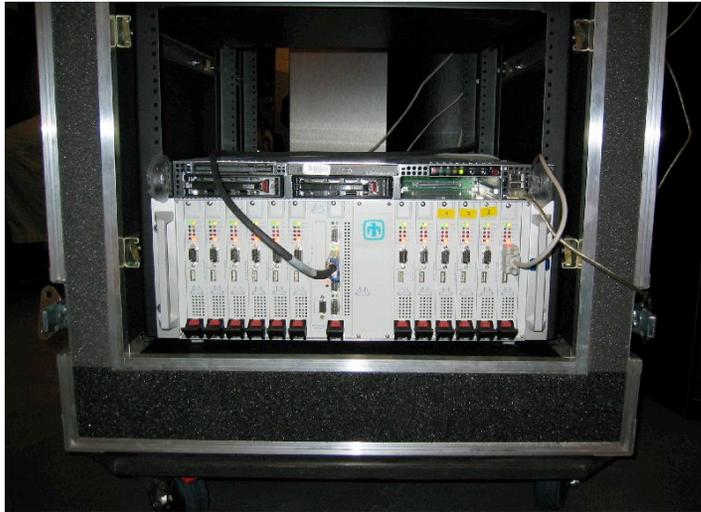
- Adaptec collaboration
  - TOE NIC, iSCSI HBA, IPsec
  - 10 GigE TOE R&D involvement
- TOE and RDMA Applications
  - IPC (MPI)
  - Storage I/O
    - iSCSI driver, Linux file system, and SCSI class tuning for TCP/IP transport
    - Security
  - File system
    - Network file system
    - InfiniArray parallel file system (SC02 demo)

# Why Infiniband?

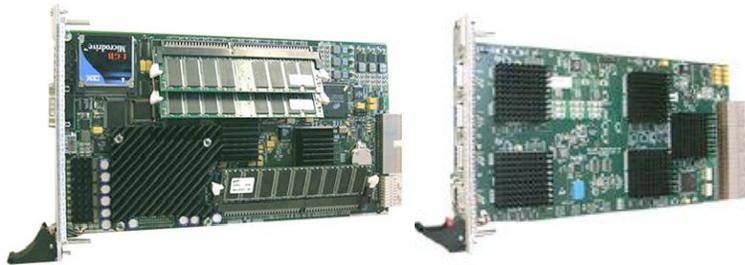
IB is the only standard-based, full featured IPC interconnect for HPC

- Switched-based interconnect fabric
- High-speed (2.5, 10, 30 Gb/s)
- Provides reliable and unreliable transport services
- Hardware support for remote DMA operations
- Hardware support for atomic operations over the network
- Multicast support for switches and host channel adapters
- Features for fault tolerance, QoS mgmt., reconfigure for RAS

# Prototype Mellanox IB Blade Cluster



- . Single proc. Xeon 2.2 GHz
- . 1 GB RAM
- . 4X Infiniband backplane (10 Gb/s)
- . 12 CPU blades in 4U chassis
- . Four 4X ports out-of-box
- . RedHat Linux 7.2
- . High density computing power
- . High reliability

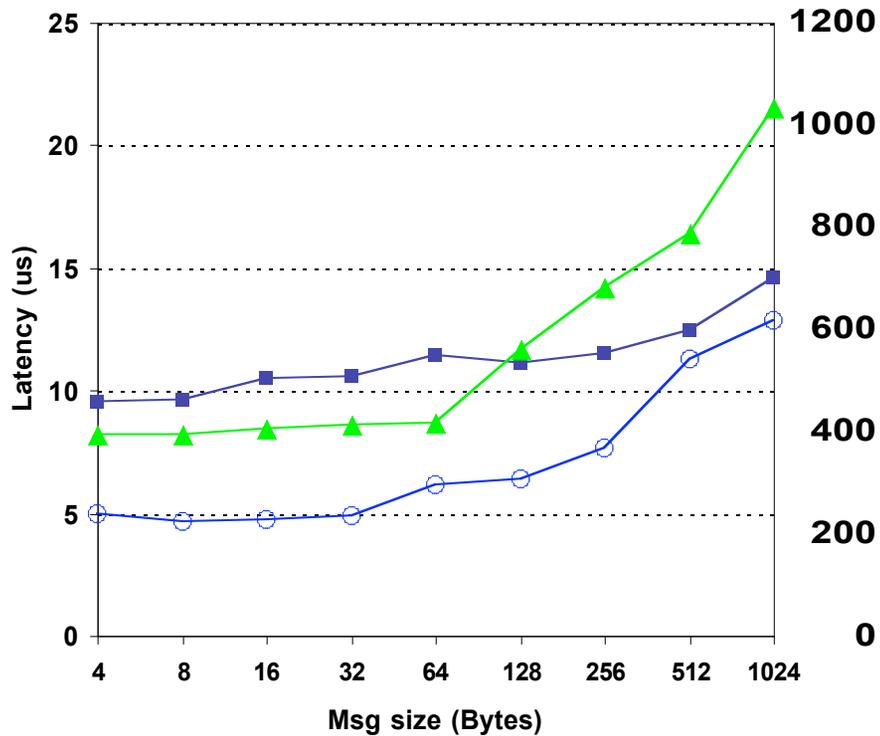


# MVAPICH Performance Demonstrated at SC02

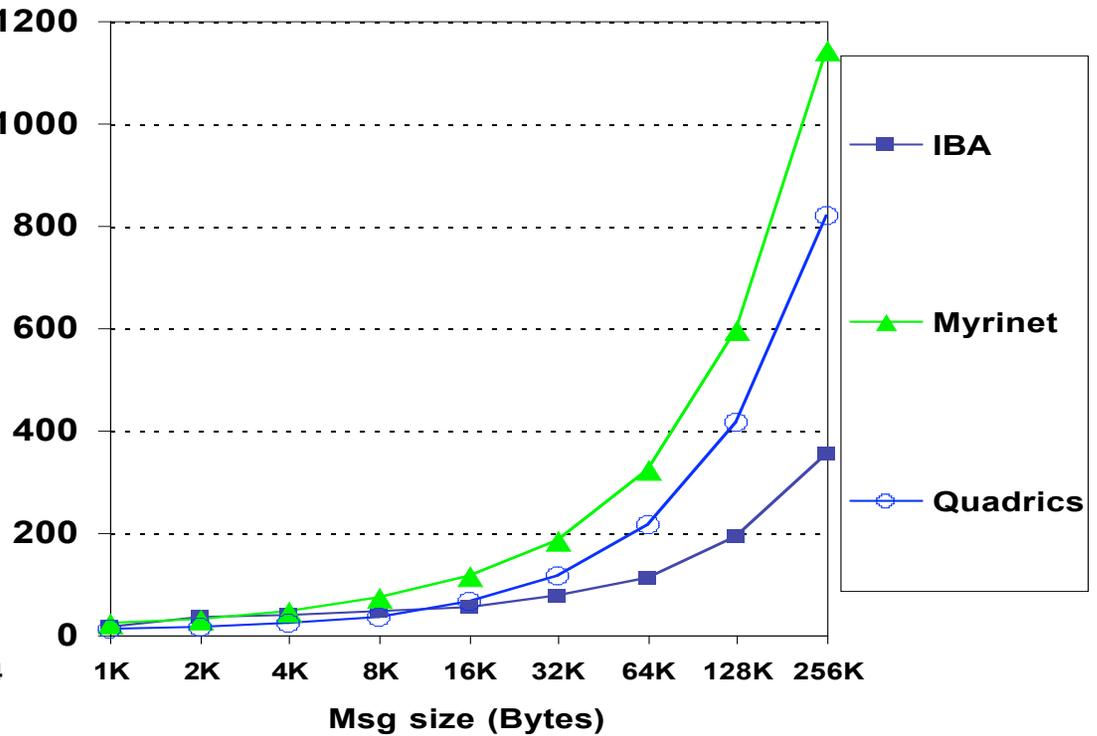
- Based on MVICH (Lawrence Berkeley)
  - .VIA implementation for MPICH-1.2.2.2 (Argonne)
- RDMA read/write operation for Rendezvous protocol
- Send/receive used for control messages
- Infiniband Software Development Kits are undergoing rapid improvement
- Performance numbers are from Mellanox SDK's released in mid-November
- <http://nowlab.cis.ohio-state.edu/projects/mpi-iba/>

# MPI-level Latency and Comparisons

## MPI small message latency

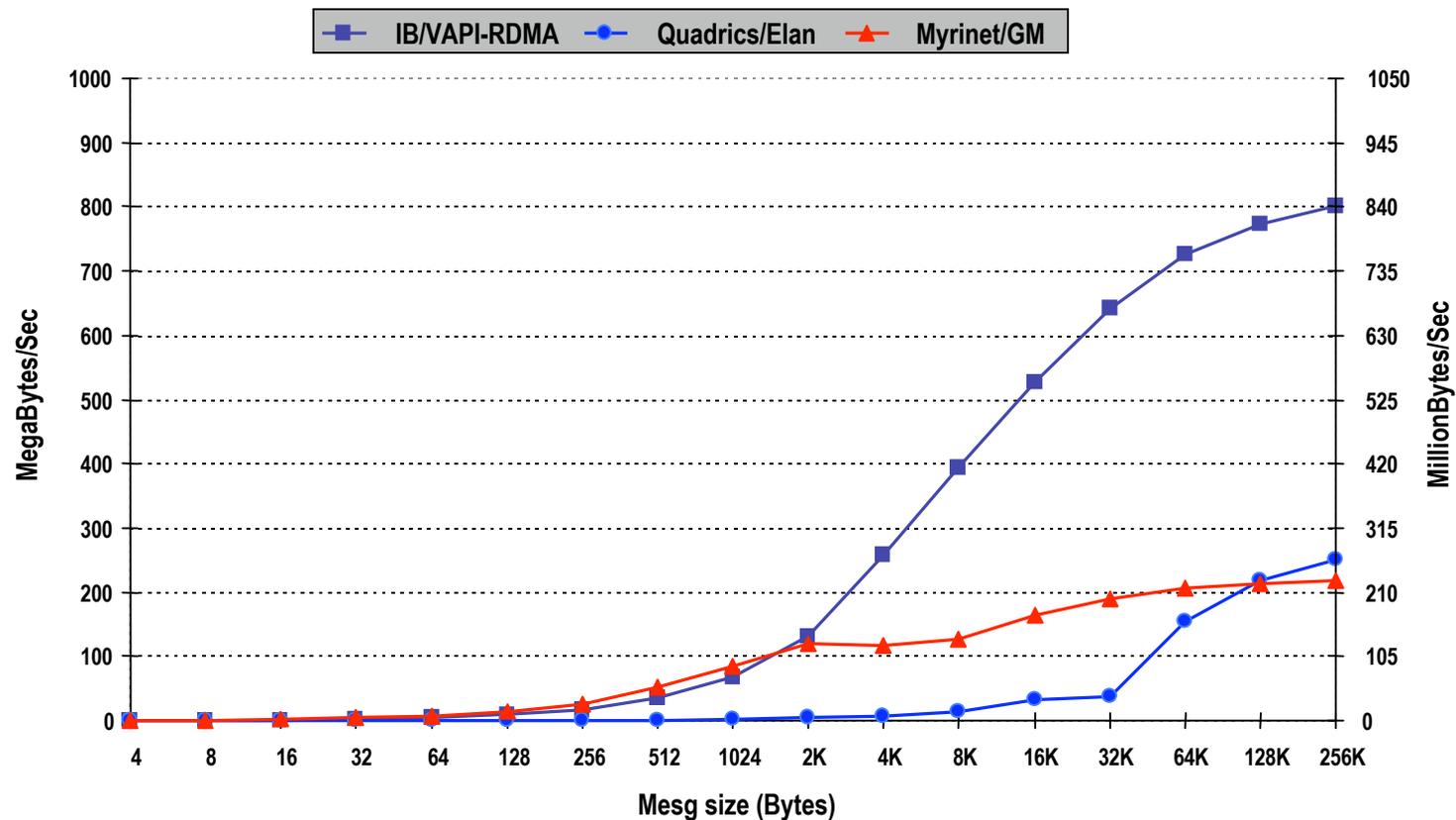


## MPI large message latency



- MPI/IBA gives 9.5 microsec latency for short messages
- MPI/IBA performs better than MPI/GM/Myrinet for messages > 128 bytes
- Better than MPI /Elan3/Quadrics for messages > 8K bytes

# MPI-level Bandwidth and Comparisons



- MPI/IBA delivers Bandwidth up to 800 MegaBytes/sec (840 MillionBytes/sec)

# IBA Future Plans

- Port BProc for dynamic reconfiguration through process migration
- Embed Linux kernel in system BIOS for fast boot
- Scale up cluster size using larger 64-128 port switches with 12X chips
- Evaluate one or more relevant I/O options
  - SRP to Fibre Channel Storage via gateway
  - DAPL to iSCSI storage via gateway
  - SRP to IB storage
  - DAFS/DAPL over IB
  - Lustre/Portal over IB
  - PVFS/VIPL
- Promote open source and high performance Linux drivers

# Major Accomplishment

- Prototyped IP Storage testbed
  - iFCP and iSCSI are scalable to WAN applications
  - iFCP performance demonstrated at SC01
- Evaluated performance of Alacritech TOE as the MPI interconnect
  - Result published at Nov. 02 IEEE LCN workshop
- Prototyped 4x IB blade cluster
  - Fastest MPI performance demonstrated at SC02

# Major Issues

- 10 Gbps I/O (IB and 10 GigaE) strains Server architecture
  - Need 2.5 GB/s of bus bandwidth (PCI Express or Hyper Transport)
  - Memory bandwidth needs to be several times of I/O bandwidth (send-and-receive from network, to-and-from CPU and/or storage)
- Slow IB adoption by storage vendors
  - Require gateway solutions to address storage and network I/O
  - Need open source, high performance Linux drivers
- TCP Offload poses significant technical challenge
  - Lack of standard driver interface for major OS's
  - Complete ASCI solutions can be expensive
  - Integrated ASCI implementation may not perform as well
- RDMAP for TCP/IP requires modification of Socket API