Performance of RDMA-Capable Storage Protocols on Wide-Area Network

Weikuan Yu

Nageswara S.V. Rao

Pete Wyckoff*

Jeffrey S. Vetter

Ohio Supercomputer Center*

Managed by UT-Battelle
for the Department of Energy

OAK RIDGE National Laboratory

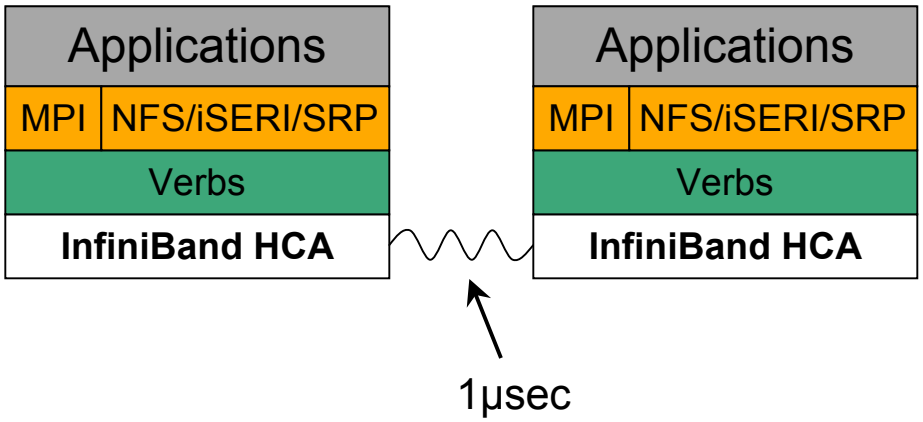
InfiniBand Clusters around the World

The Problem of Computing Islands

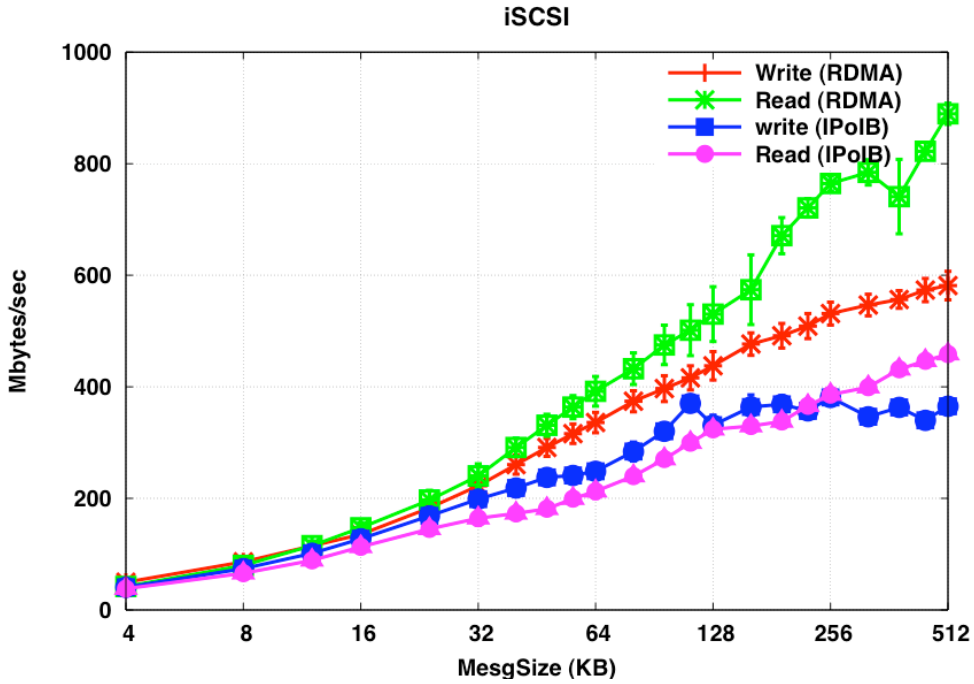
- Islands of InfiniBand (IB) clusters
  - More IB clusters are deployed
  - Some already connected, e.g. through TeraGrid
    - But only via TCP/IP protocols
- Data transfer across these islands
  - Need ever-greater data movement capabilities.
  - GridFTP, BBCP or other special storage configuration
  - TCP performance on Long Distance can be low
    - With 10GigE on UltraScience Net (no tuning)
      - 9.2 Gbps at 0.2 mile
      - 8.2 Gbps at 1400 miles
      - 2.3-2.5 Gbps at 6600+ miles

RDMA (IB) in Clusters and Local Area Networks

- Sub-microsecond latency
- Superb bandwidth (32Gbps with IB QDR)
- Heavily used for clustering
- Getting popular in storage environment
  - NFS over RDMA (NFSoRDMA)
  - SCSI RDMA Protocol (SRP)
  - iSCSI over RDMA (iSER)



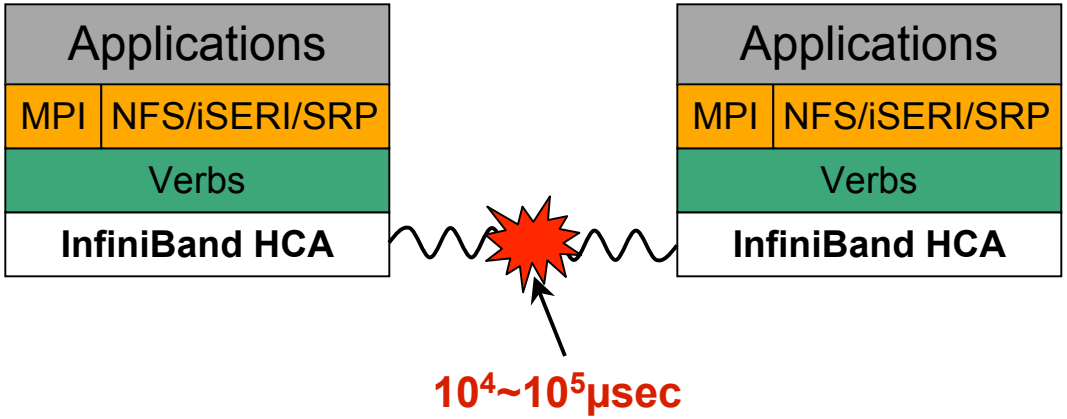
Sample Performance of RDMA-based Storage



- RDMA enables good iSCSI bandwidth within LAN
- Nearly doubled the performance for iSCSI

Feasibility of RDMA (IB) on WAN

- Long-range Extensions for InfiniBand available
  - Network Equipment Technologies (NET): NX5010
  - Obsidian Research: Longbow
- Long latency (10^4~10^5µsec)
- High bandwidth yet feasible
  - Good distance scalability and tolerance to interfering traffic
  - Good network throughput and MPI-level Performance
- Can RDMA provide a good transport protocol for storage on WAN?

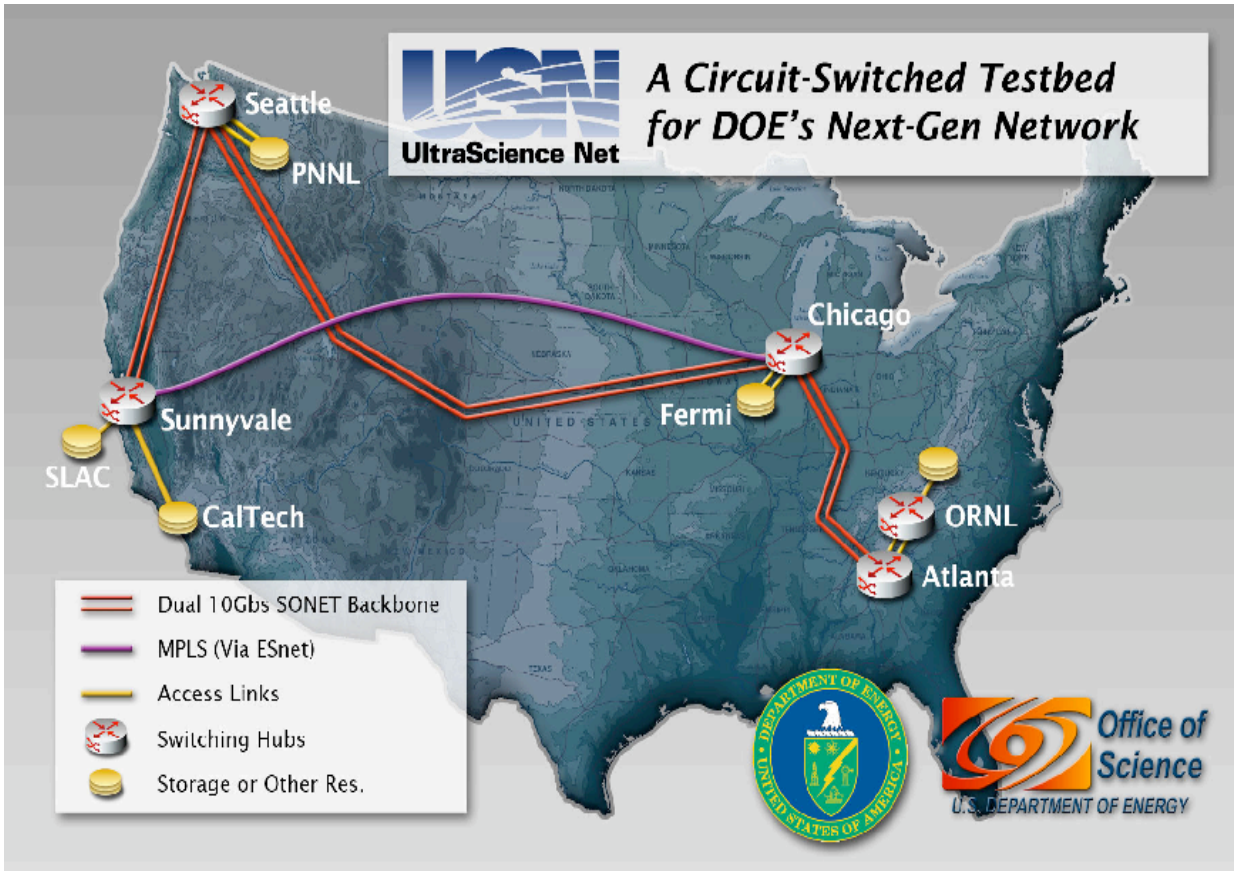


Experimental Environment

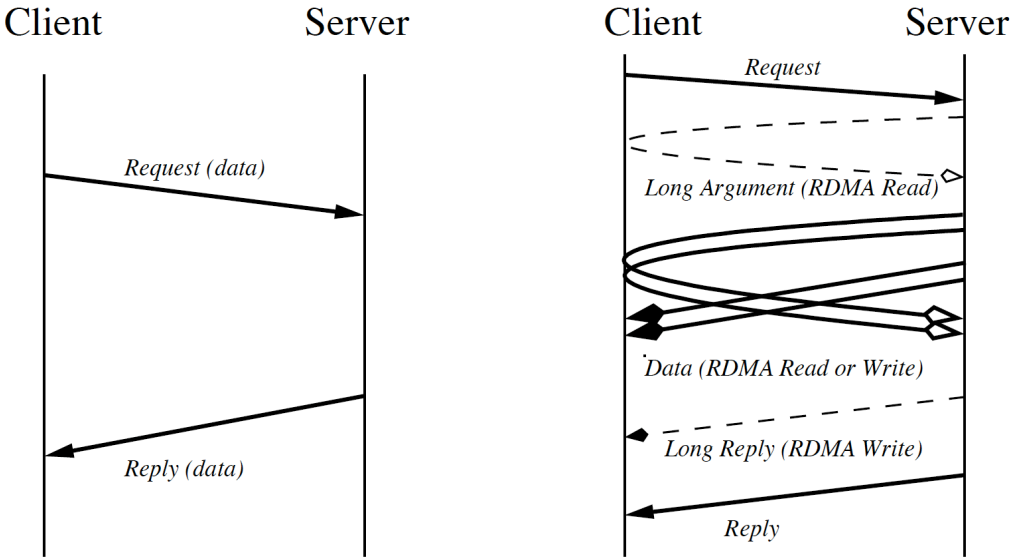
- Hardware
  - Long-range IB extension devices from NET (Network Equipment Technologies, Inc)
  - Mellanox PCI-Express 4x DDR HCAs (InfiniHost-III and Connect-X)
- Software Packages
  - OFED-1.3 from openfabrics.org
  - Linux-2.6.25 with NFSoRDMA and iSER support
- Performance of RDMA-based Storage Protocols on WAN
  - NFS over RDMA
  - iSCSI over RDMA

UltraScience Net at ORNL

- Experimental WAN Network
  - Oak Ridge, Atlanta, Chicago, Seattle, and Sunnyvale
  - OC192 backbone connections
  - 4300 miles one way, 8600 miles loop-back



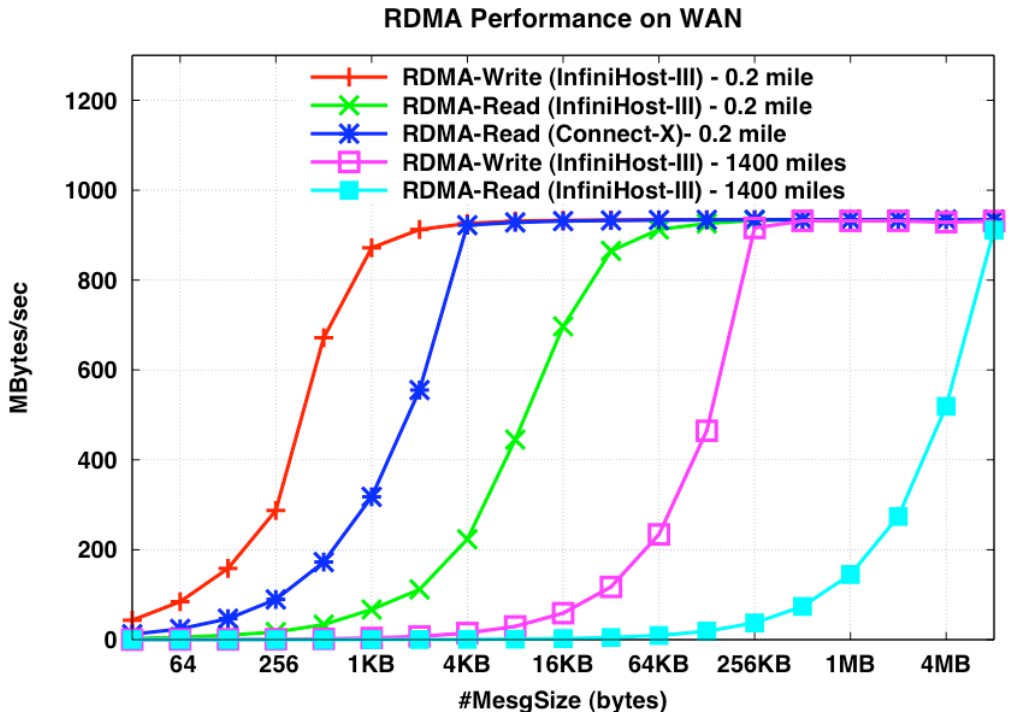
RDMA-based Transport



Non–RDMA Transport RDMA–Based Transport

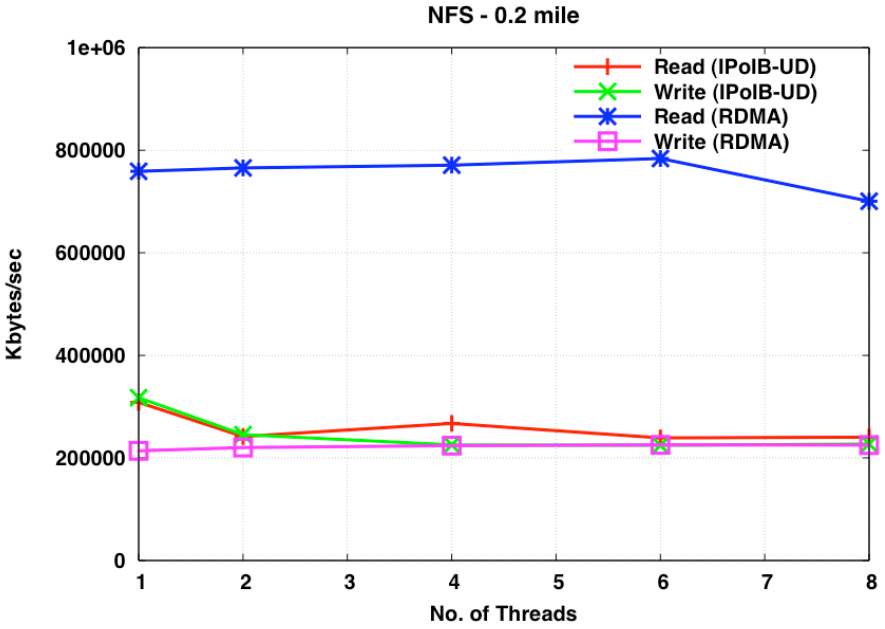
- Request and request becomes pure control messages, and have to travel long distance on WAN
- Use of RDMA read (round-trip operations) for clients to write data
- Possible additional control messages for NFSoRDMA for long arguments
- Further fragmentation due to the use of page-based operations

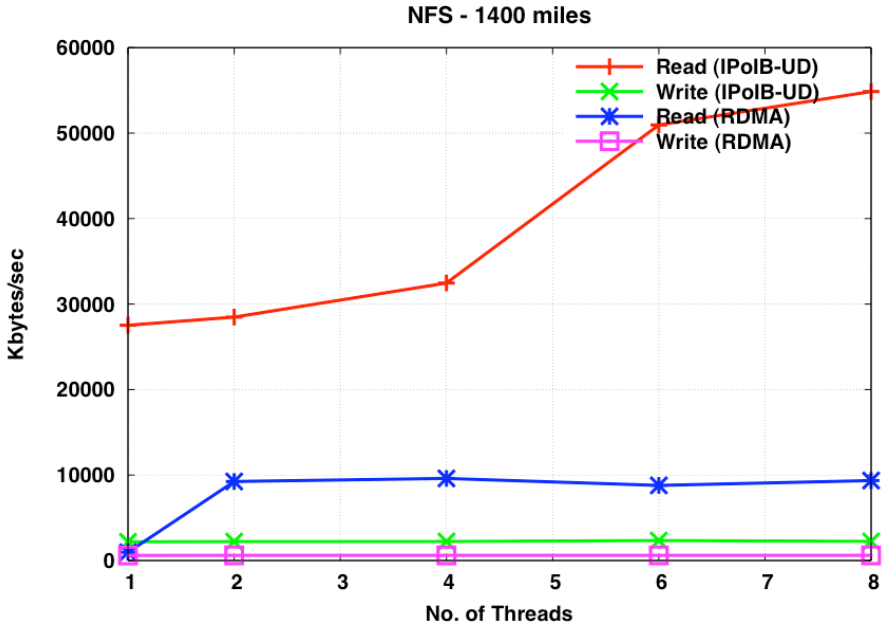
RDMA on WAN



- RDMA has good network-level performance within short distance WAN
- High bandwidth at long distance is only possible for large messages
- Low RDMA-read performance for page-based messages (4KB), even at 0.2 mile when using InfiniHost-III HCAs

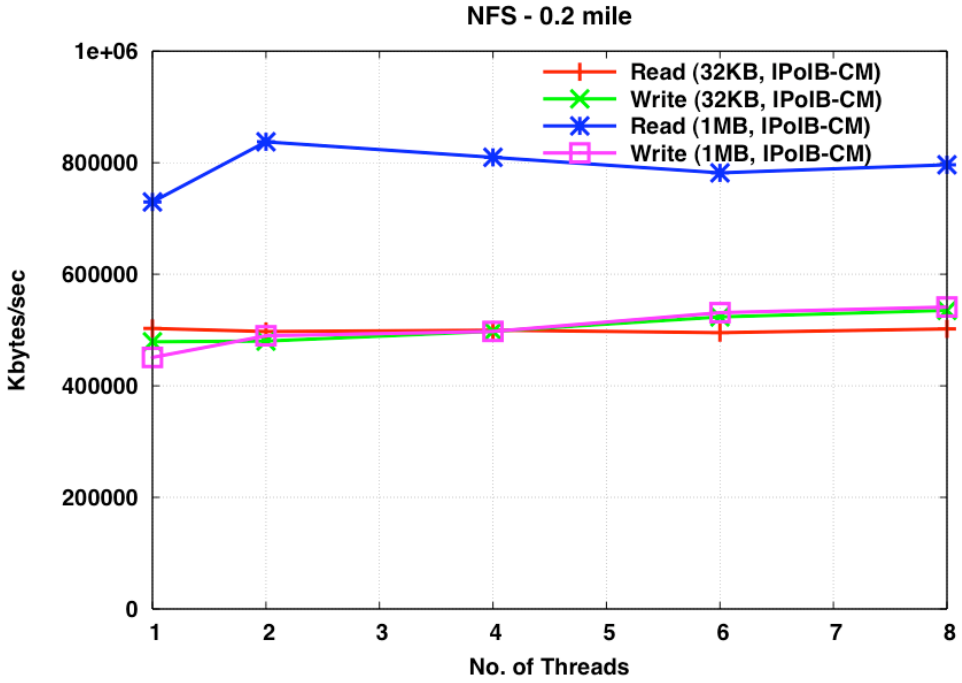
NFS over RDMA





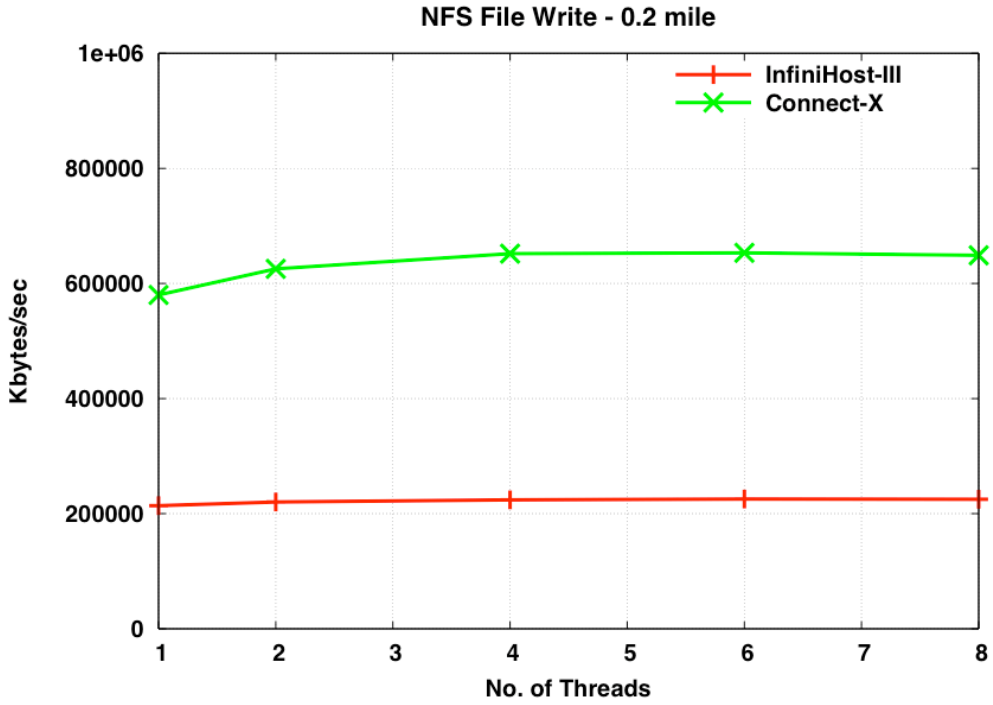
- NFS over RDMA achieves good bandwidth within short distance
- But significant optimizations are needed for long distance

NFS - Large block size



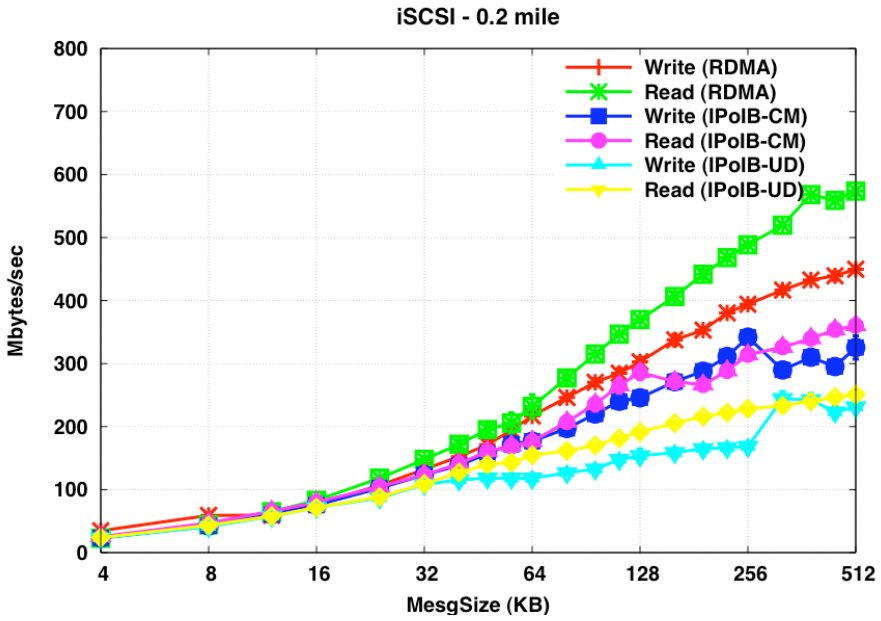
- NFS over IPoIB-CM benefits from large block size
- NFS over RDMA needs to support large block size for better fit on long-distance WAN

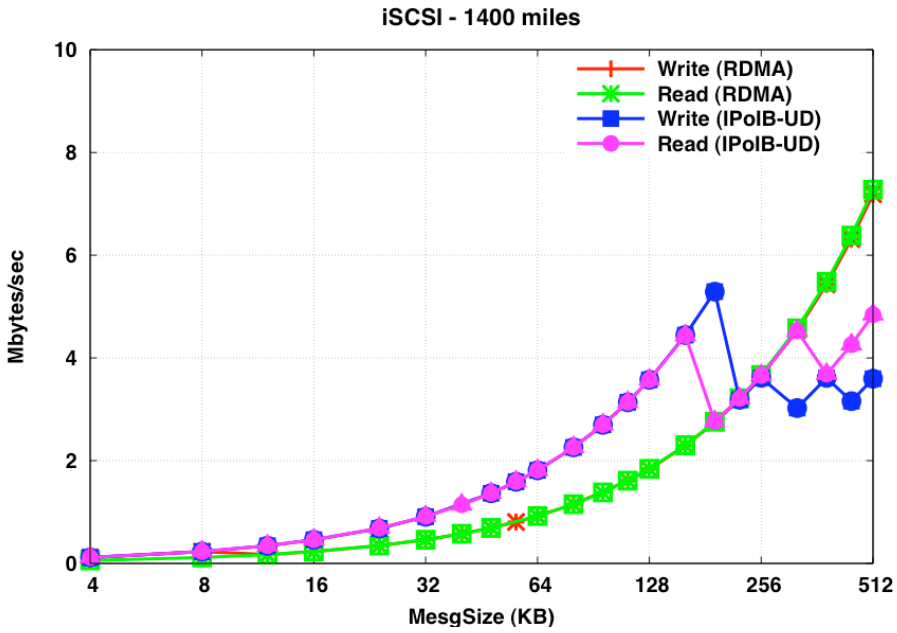
NFS over RDMA - using Connect-X



- Better RDMA read in connect-X improves the performance of file write for NFS over RDMA
- Performance at long distance is yet to determine

iSCSI over RDMA (iSER)





- RDMA enables high-performance iSCSI within short distance
- RDMA has good promise over long distance as shown with large messages

Perspectives

- Long-range InfiniBand
  - InfiniBand over SONET is promising
  - Storage protocols are not yet exploiting the bandwidth potential of RDMA at long distance
- RDMA-based Storage on WAN
  - Need to enable large block sizes
  - Need to avoid page-based RDMA operations in NFS
    - Utilize IB FRMR support to avoid small RDMA operations
  - Need to allow more concurrent RDMA read operations

Acknowledgment

- Network Equipment Technologies, Inc
  - Andrew DiSilvestre
  - Rich Erikson
  - Brad Chalker
- ORNL
  - Susan Hicks
  - Philip Roth
- Mellanox