InfiniBand on Wide-Area Network

Weikuan Yu

Nageswara S.V. Rao

Jeffrey S. Vetter

Computer Science & Mathematics

Managed by UT-Battelle
for the Department of Energy

Oak Ridge National Laboratory

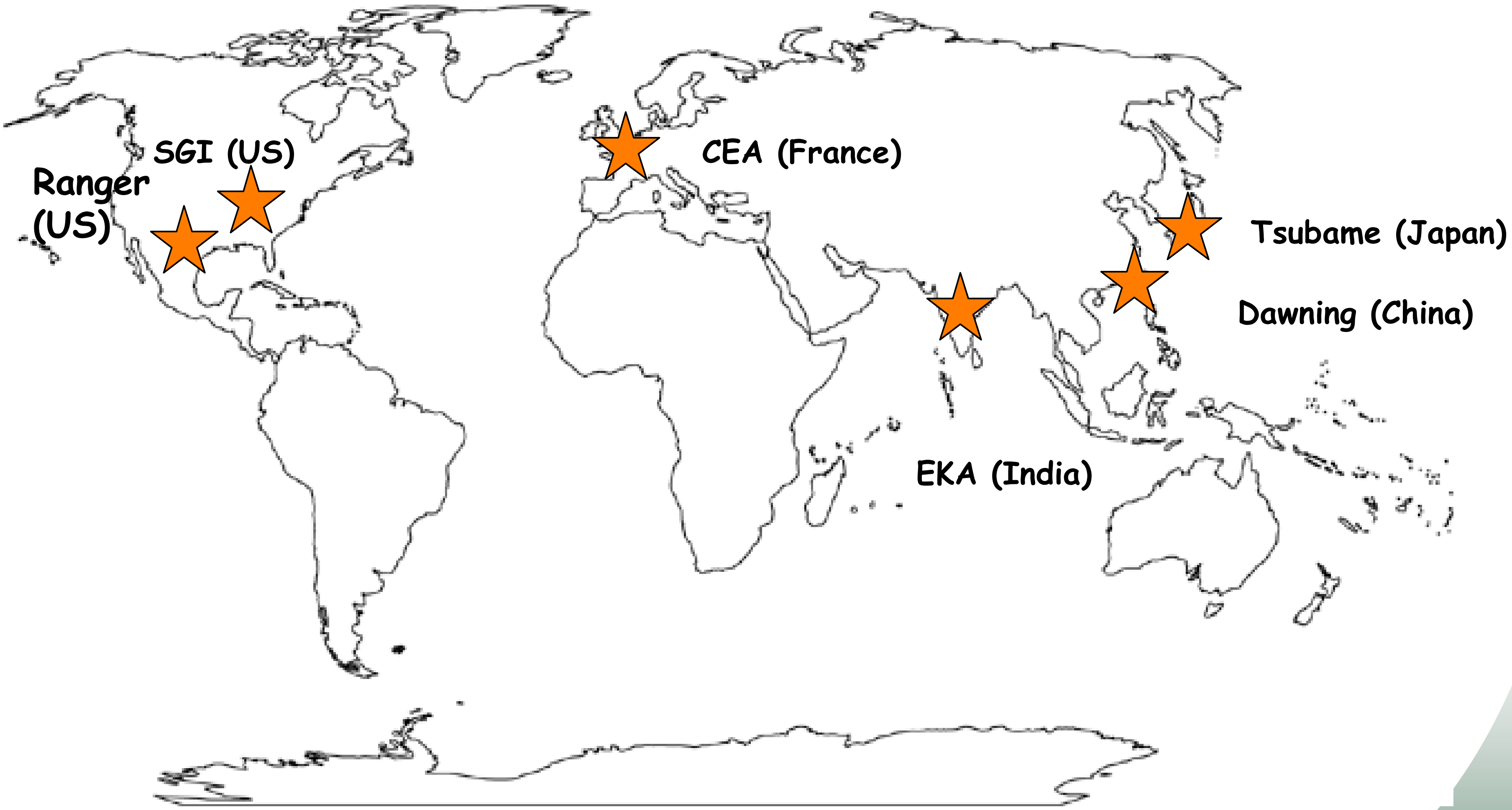
Outline

- Overview
  - Contemporary Network Technologies & InfiniBand
  - UltraScience Net at Oak Ridge National Laboratory
  - Configuration of test environment
- Performance of OFED IB on WAN
  - Network (RDMA)
  - MPI (MVAPICH)
  - Others: IPoIB, SDP, NFSoRDMA and iSER
- Perspectives

InfiniBand and Other Contemporary Network Technologies

- The race for the speed
  - SONET:
    - OC192 (10Gbps) -- OC768 (40Gbps) …
  - Ethernet:
    - 10Gbps -- 40Gbps/100Gbps
  - InfiniBand:
    - Link rates: SDR/DDR/QDR (2.5/5/10Gbps)
    - Link width of 1x/4x/12x, 20Gbps -- 40Gbps/60Gbps

Some InfiniBand Clusters around the World

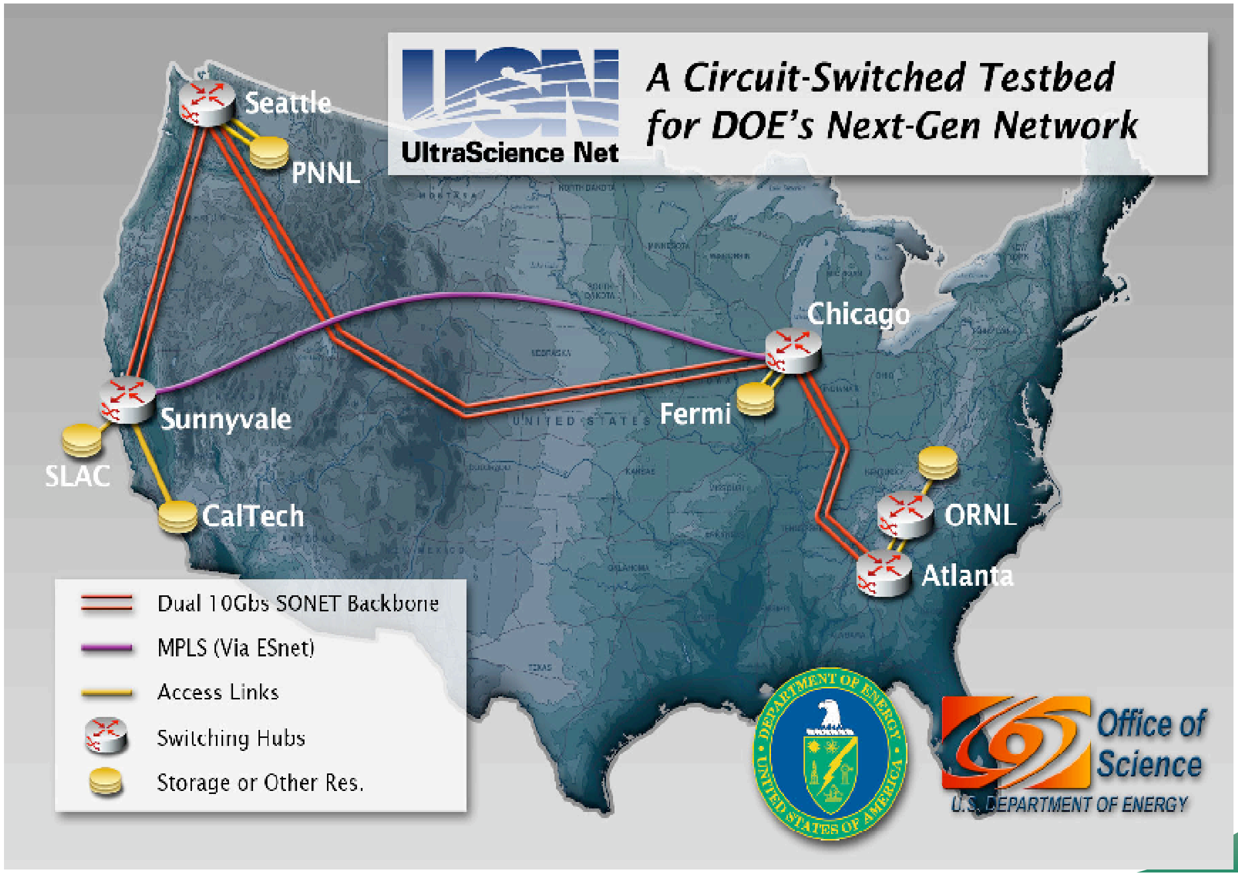


The need of IB on WAN

- InfiniBand Clusters around the globe
  - Many IB clusters are deployed
  - Some already connected, e.g. through TeraGrid
    - But only via TCP/IP protocols
  - TCP performance on Long Distance may be low
    - With 10GigE on USN (no tuning)
      - 9.2 Gbps at 0.2 miles
      - 8.2 Gbps at 1400 miles
      - 2.3-2.5 Gbps at 6600+ miles
- Range Extensions for InfiniBand on WAN
  - Obsidian Research: Longbow
  - Net.com: NX5010

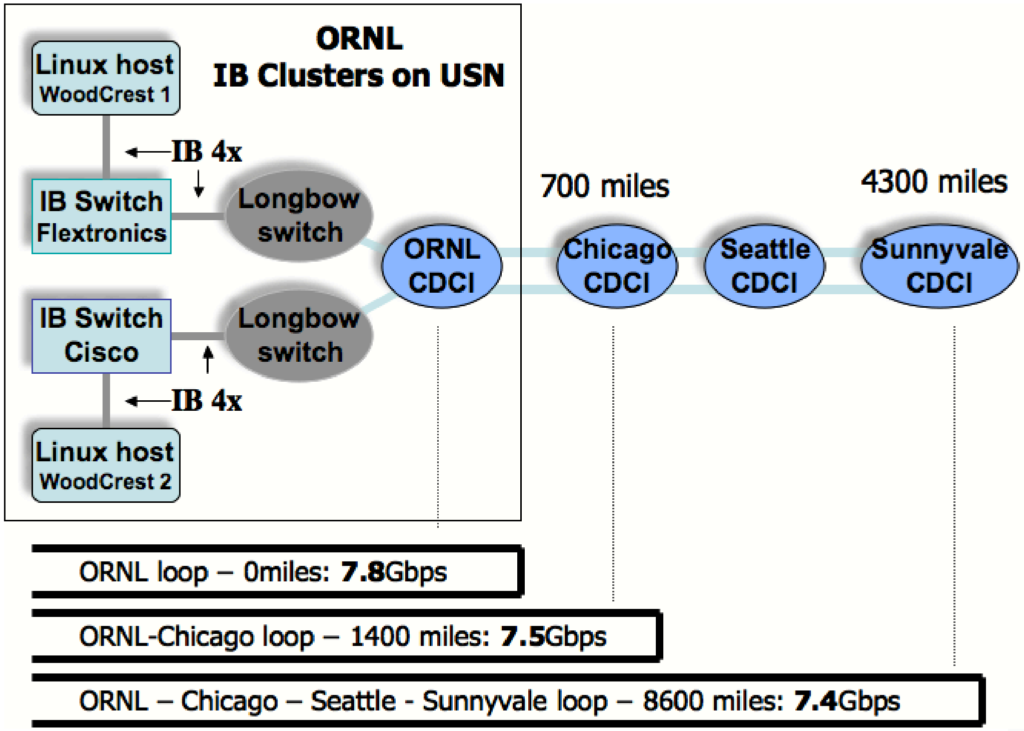
UltraScience Net at ORNL

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



Configuration of Test Environment

- Hardware
  - UltraScience Net
  - Longbow switches
  - Mellanox PCI-Express 4x DDR InfiniHost III HCAs
  - Two Clusters each running its own subnet manger
- Software
  - OFED-1.2.5.4 and OFED-1.3
  - MVAPICH/MVAPICH2

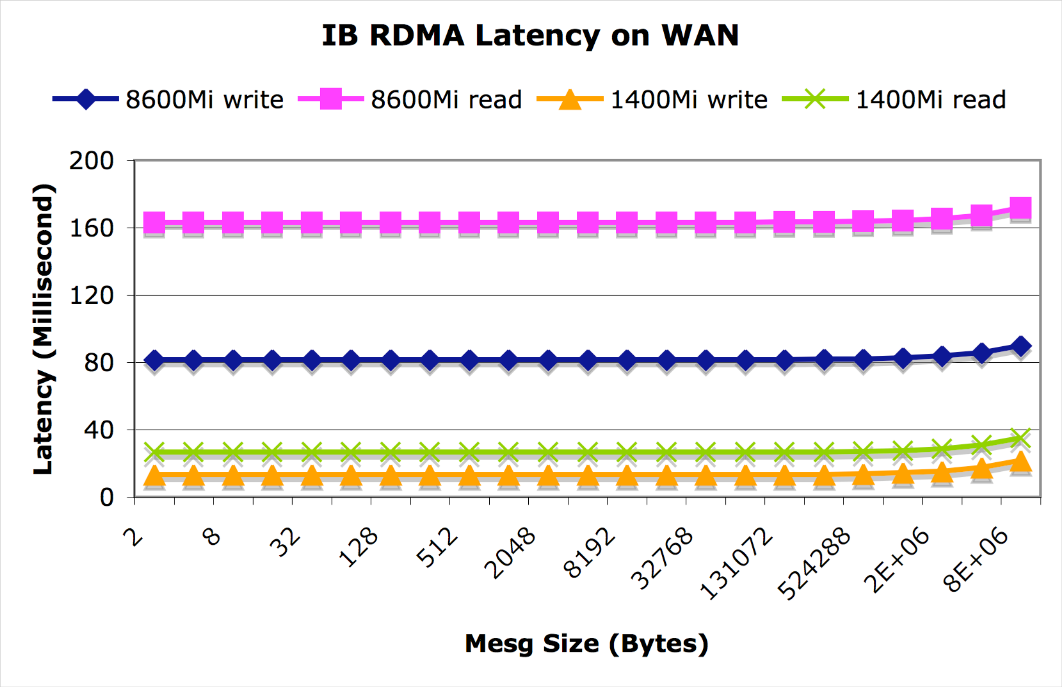


ORNL loop – 0miles: **7.8**Gbps

ORNL-Chicago loop – 1400 miles: **7.5**Gbps

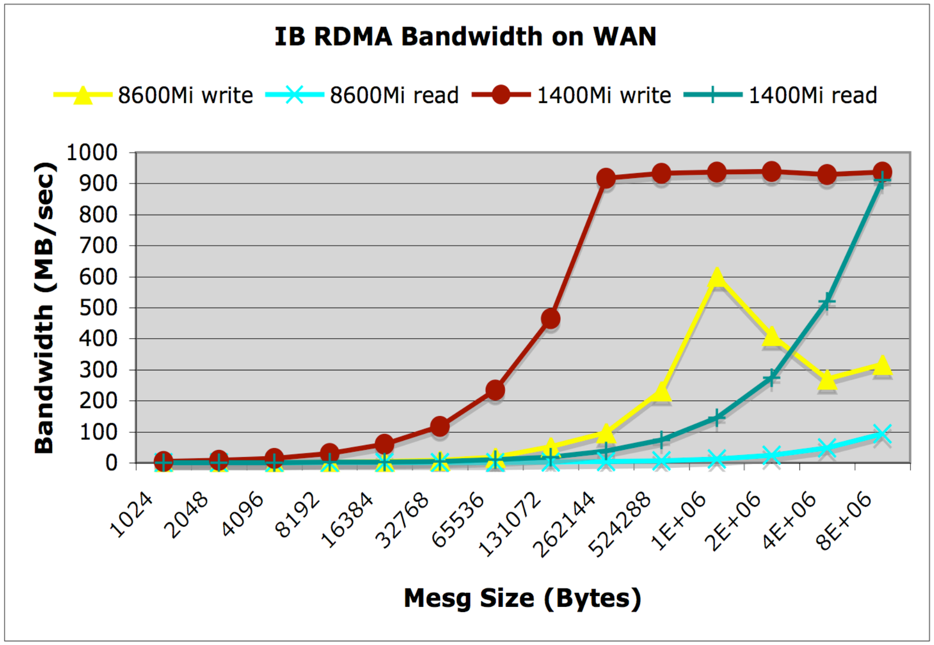
ORNL – Chicago – Seattle - Sunnyvale loop – 8600 miles: **7.4**Gbps

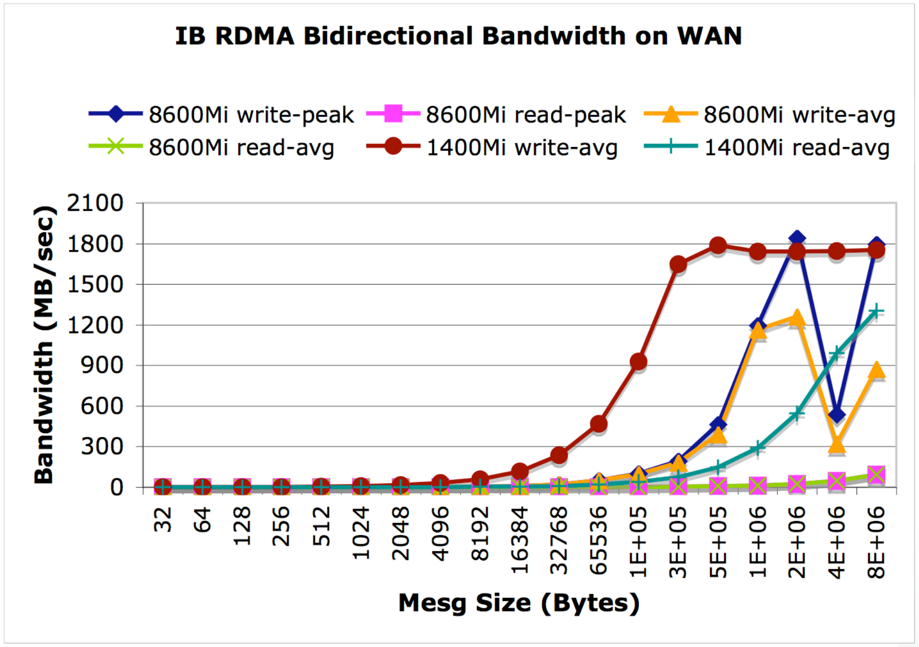
RDMA Latency (Longbow)



- Latency is determined by distance
- Latency of RDMA read is twice as long

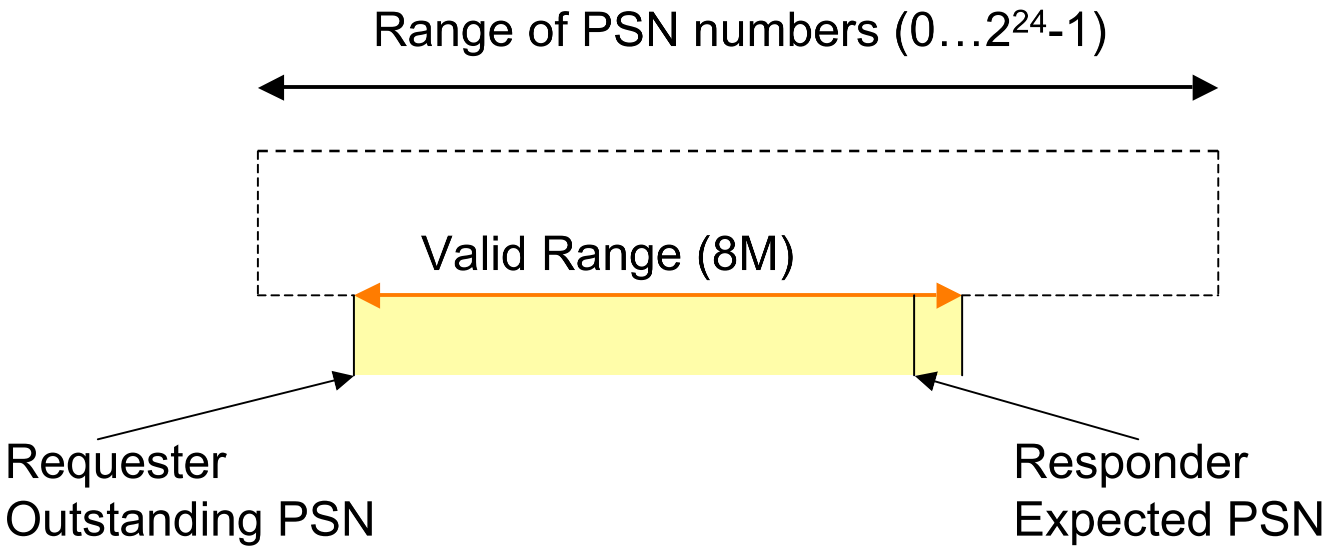
RDMA Bandwidth (RC)





- 7.5Gbps for 1400 miles, 7.2Gbps for 8600 miles
- At long distances, bandwidth is low for messages (< 1MB).
- The performance of RDMA read is particularly low

PSN Flow Control for a RC-based QP



Range of PSN numbers (0…2^{24}-1)

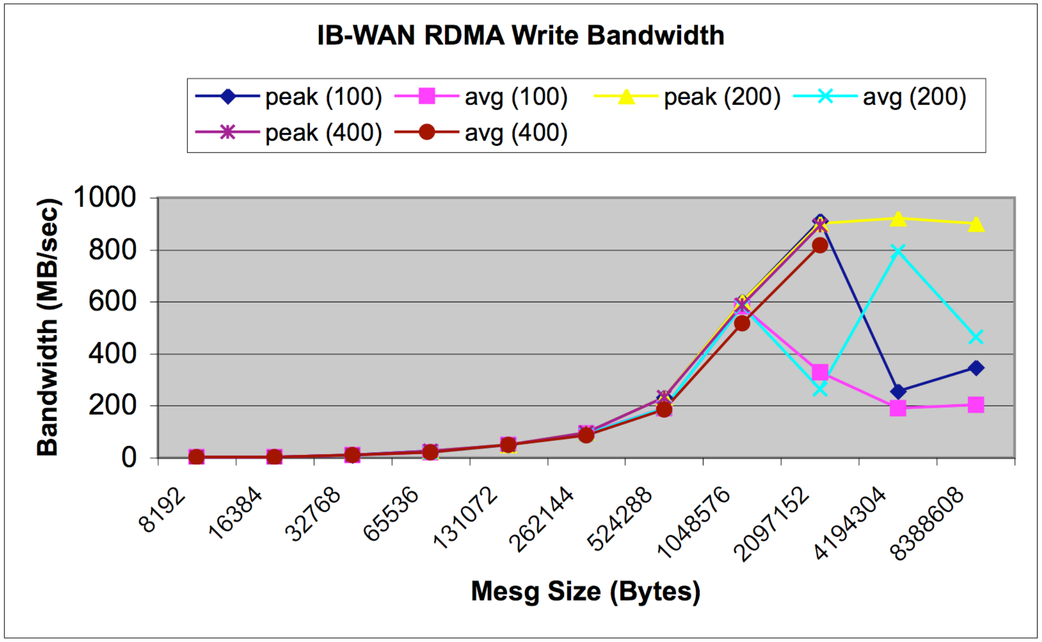
Valid Range (8M)

Requester
Outstanding PSN

Responder
Expected PSN

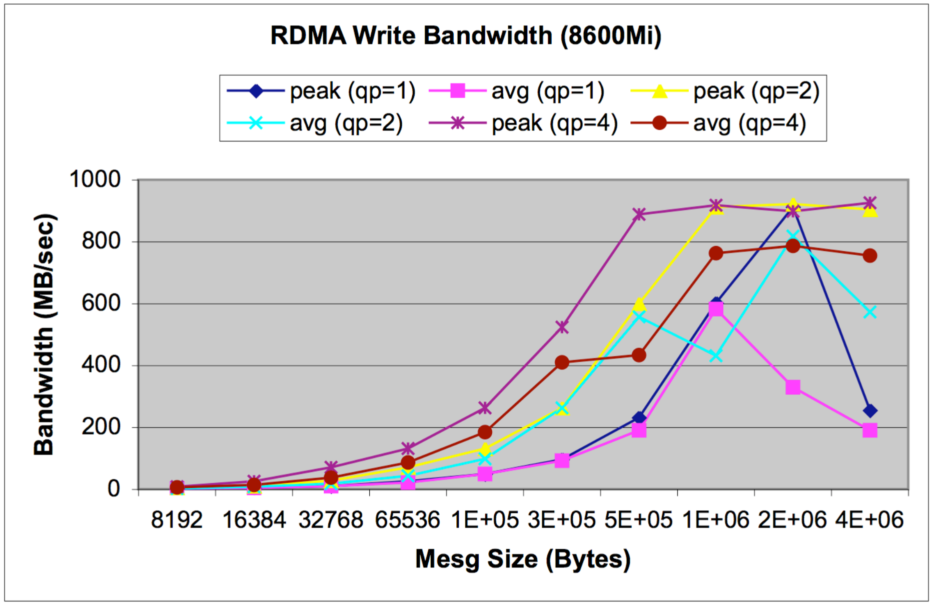
- InfiniBand uses a go-back N protocol for RC
  - Bandwidth = (effective window size) * PMTU / RTT
- To improve throughput
  - Inject more packets into a single QP
  - Employ more concurrent QPs
  - Increase the maximum number of RDMA Read operations per QP

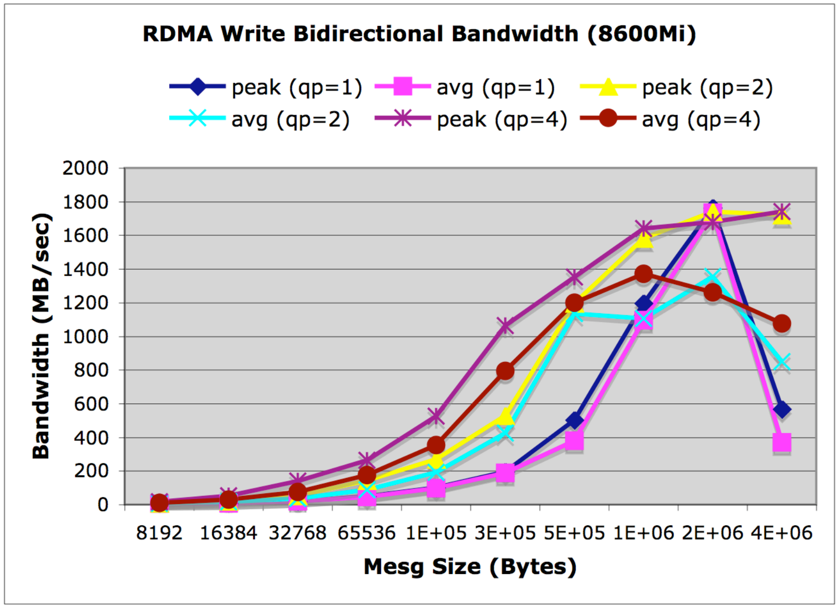
Increased Queue Depths (RC) - 8600 miles



- No consistent performance improvement with different transmit queue depths

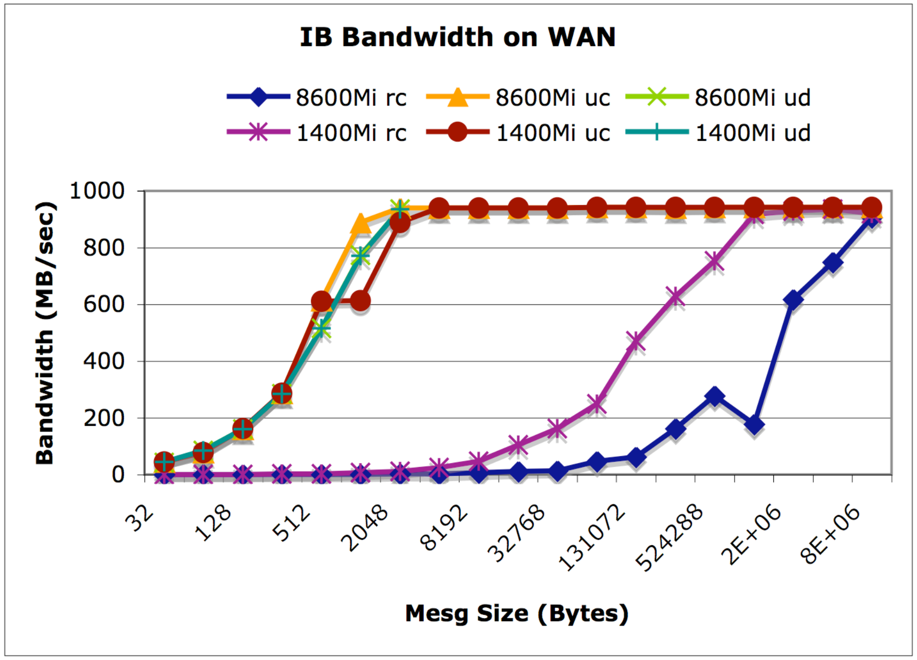
Multiple Connections - 8600 miles

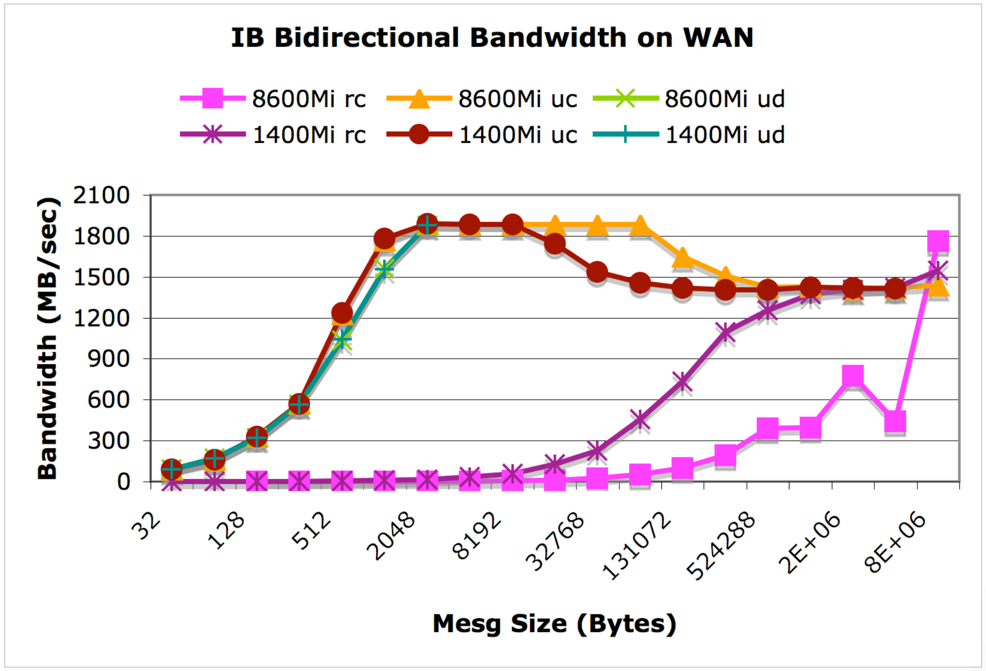




- With multiple connections (QPs)
  - Better throughput for all mid-size message
  - Sustained bandwidth of 7.4Gbps at 8600 miles

Bandwidth (UC & UD)



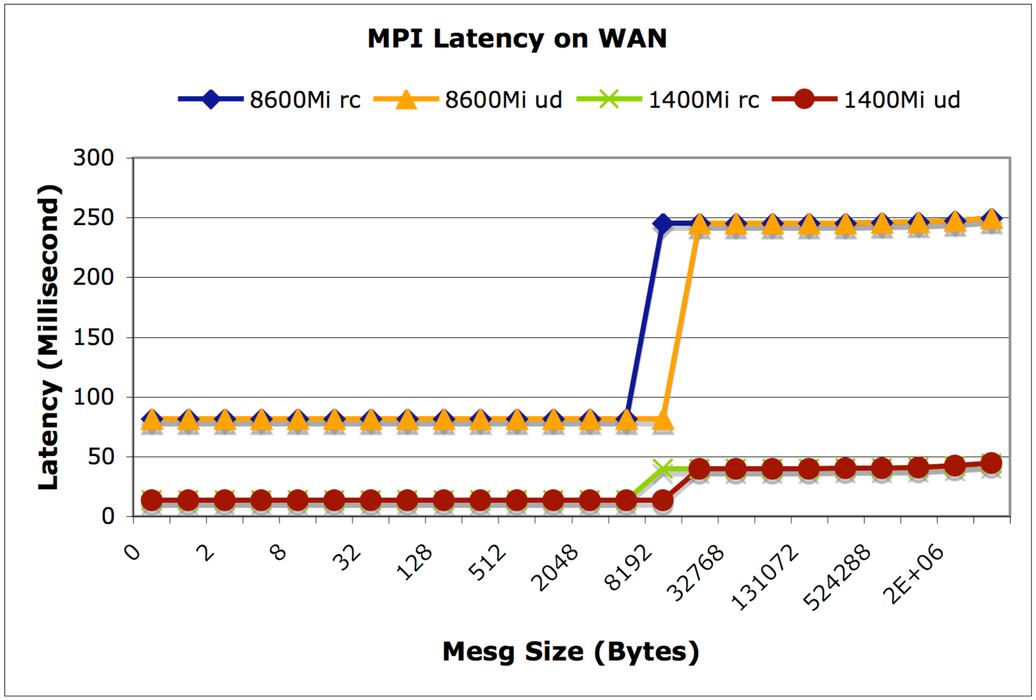


- Instant injection of all IB packets on the wire
- Very rare message loss at long distances or when there is a big burst of messages
- Peak bandwidth of 7.5Gbps

Outline

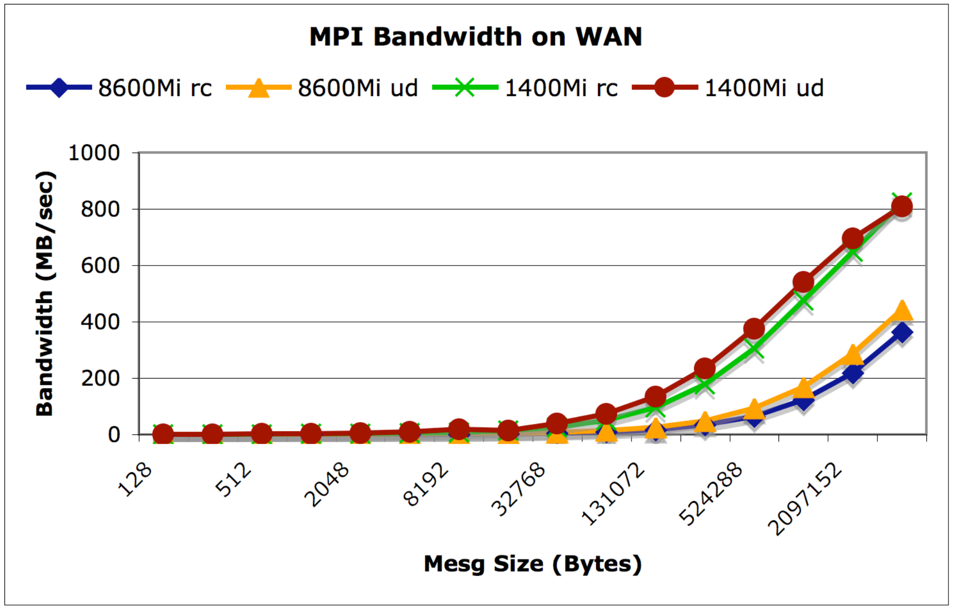
- Overview
  - Contemporary Network Technologies & InfiniBand
  - UltraScience Net at Oak Ridge National Laboratory
  - Configuration of test environment
- Performance of OFED IB on WAN
  - Network (send/receive, RDMA)
  - MPI (MVAPICH)
  - Other Protocols
- Perspectives

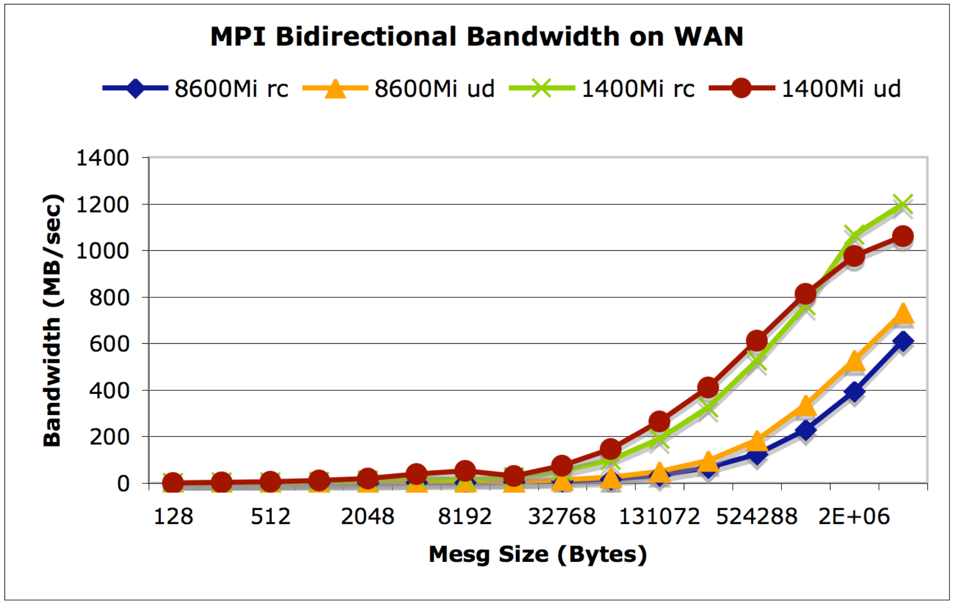
MPI Latency - RC and UD



- Latency determined by distances
- Latency triples for large messages bigger than rendezvous thresholds

MPI Bandwidth - RC and UD



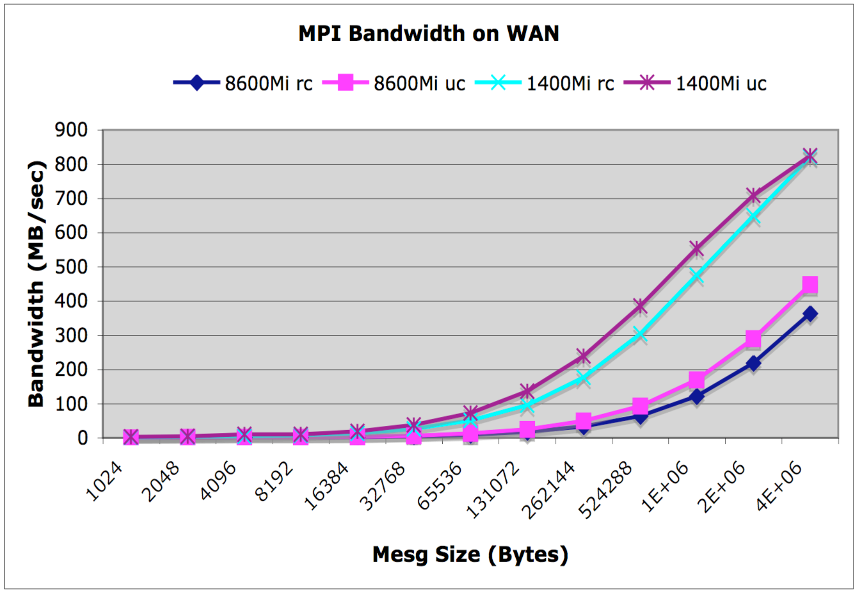


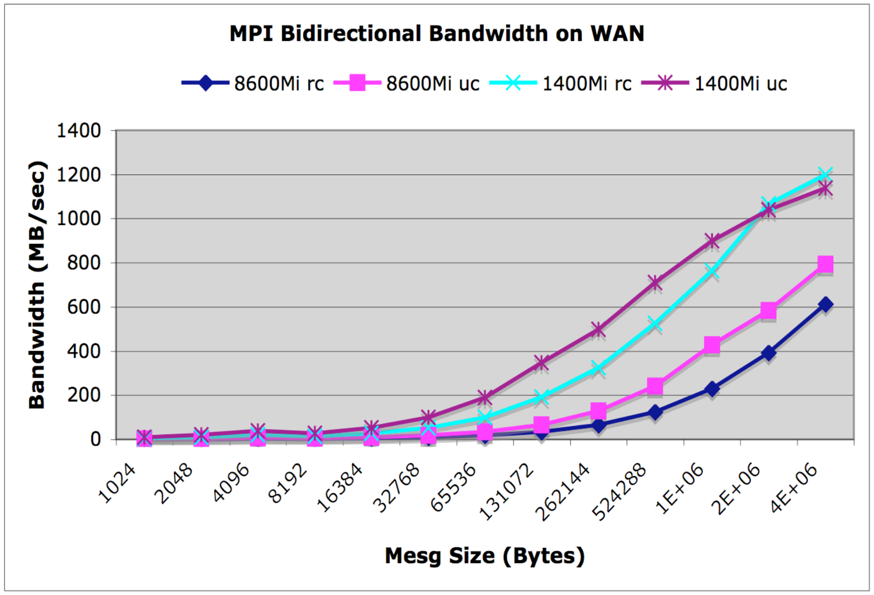
- Longer distance requires larger window sizes
- 443MB/sec achievable with MVAPICH/UD at 8600miles

IB services for Distance Scalability

- UD
  - PMTU
  - Message fragmentation/reassembly
  - Reliability + Ordering
  - Send/Receive, No RDMA
  - Connection Scalability
- UC
  - Arbitrary message length
  - Reliability + Ordering
  - Send/Receive + RDMA
  - Distance Scalability

MPI with UC





- MPI-UC on WAN that takes advantage of UC and rare message losses
- Improve the sustained bandwidth, compared to RC

Other Protocols

- IPoIB and SDP
  - Both performed poorly at long distances (BIC only)
    - 2Gbps at 1400 miles, 400 Mbps at 8600miles
  - Use 10GigE for applications that require TCP-based legacy protocols
  - IPoIB are enabled for occasional use, for example, when needed for management purposes
- NFSoRDMA
  - Initial evaluation at 0.2 and 1400 miles
- iSCSI over RDMA
  - Initial evaluation at 0.2, 1400 and 8600 miles
  - 300MB/sec for writes and 500MB/sec reads (0.2miles)

Perspectives

- Alternative to long-range networking
  - SONET (IB)
  - 10GigE
    - IB --> 10GigE
    - TCP --> 10GigE
    - iWARP --> 10GigE
- MPI over IB
  - MVAPICH/UD already available
  - MVAPICH/UC
    - Reliability implemented
    - Additional work on message ordering, congestion control
  - Latency-oriented optimizations no longer as important

Perspectives - continued

- File and Storage Protocols
  - Tune and optimize iSER and NFSoRDMA for WAN
  - Continue to use RDMA Read for NFSoRDMA and iSER??
- Enable Grid-oriented Protocols over InfiniBand
  - bbcp/gridFTP
  - SRB/SRM

Acknowledgment

- Obsidian Research
- Net.com
- Pete Wyckoff @ OSC
- Thank you!