Efficient Zero-Copy Noncontiguous I/O for Globus On InfiniBand

Weikuan Yu

Yuan Tian

Jeffrey Vetter (ORNL)

Presentation Outline

- Overview
  - Contemporary Network Technologies & InfiniBand
  - Emergence of RDMA on WAN
- Globus on InfiniBand
- Globus Zero-copy Noncontiguous I/O on IB
- Performance Evaluation
- Conclusion

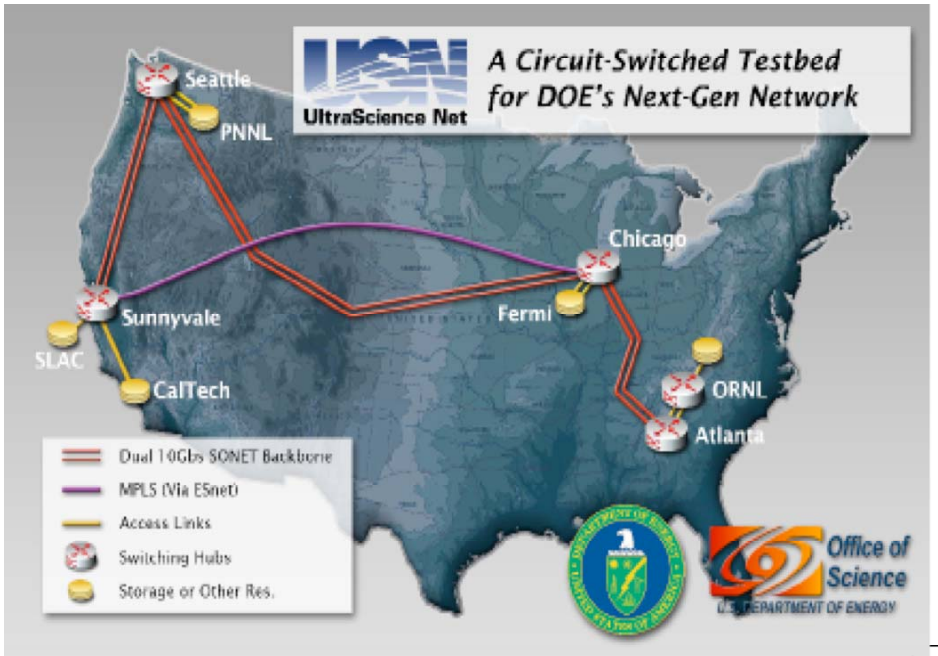
Contemporary Network Technologies On WAN

- 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 Emergence of IB/RDMA on WAN

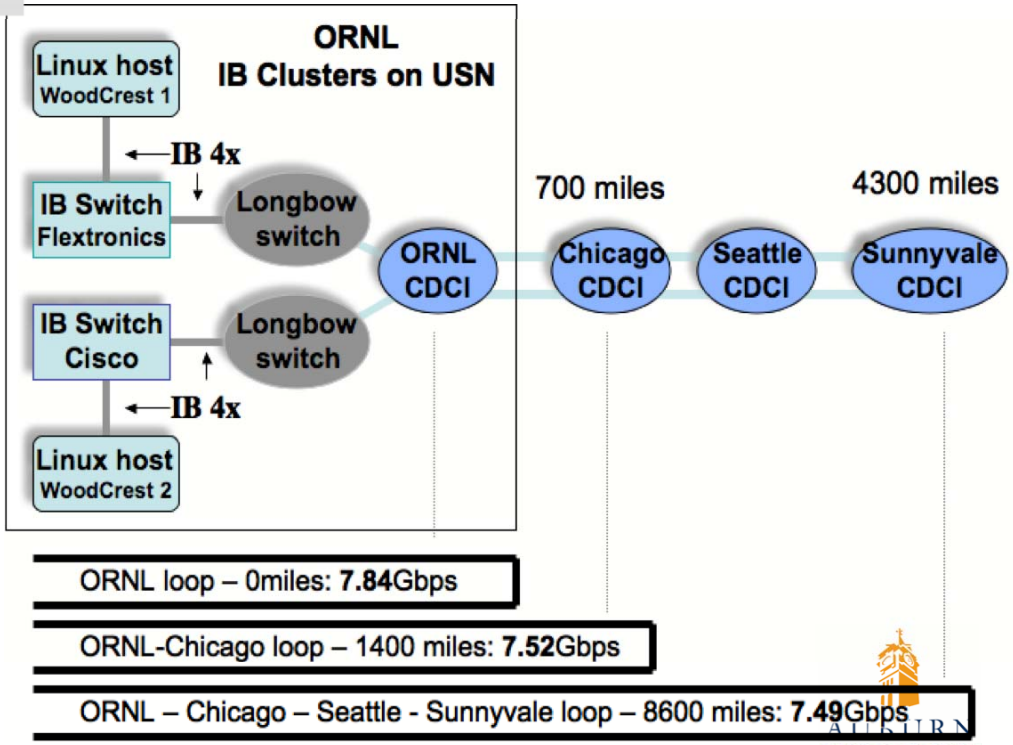
- InfiniBand Clusters around the globe
  - Many IB clusters are deployed
  - Some already connected, e.g. through TeraGrid
    - But only via TCP/IP protocols
  - Hard to tune TCP across long distance
- Range Extensions for InfiniBand on WAN
  - Obsidian Research: Longbow
  - Net.com: NX5010
- RDMA over Converged Ethernet



- RDMA on WAN Network at ORNL
  - UltraScience Net across Oak Ridge, Atlanta, Chicago, Seattle, and Sunnyvale
  - OC192 backbone connections
  - 4300 miles one way, 8600 miles loop-back

Our work on InfiniBand/WAN

- Evaluating IB vs 10 GigE (SC08)
- MPI (NAS'08)
- NFS and iSCSI (PDSW'08)
- GridFTP over RDMA (P2S2'10)

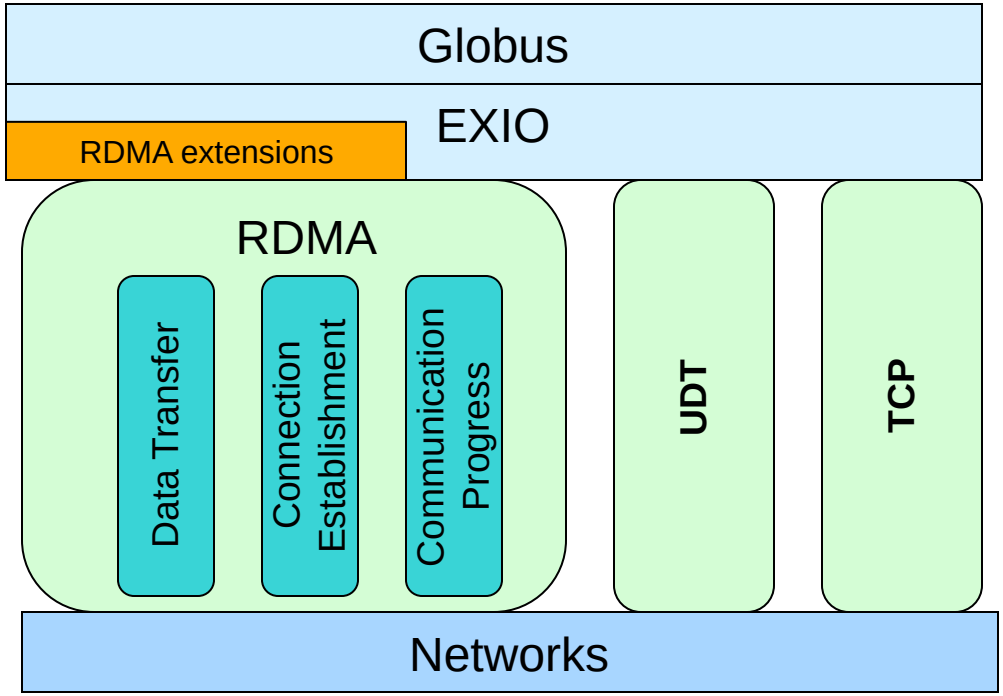


Presentation Outline

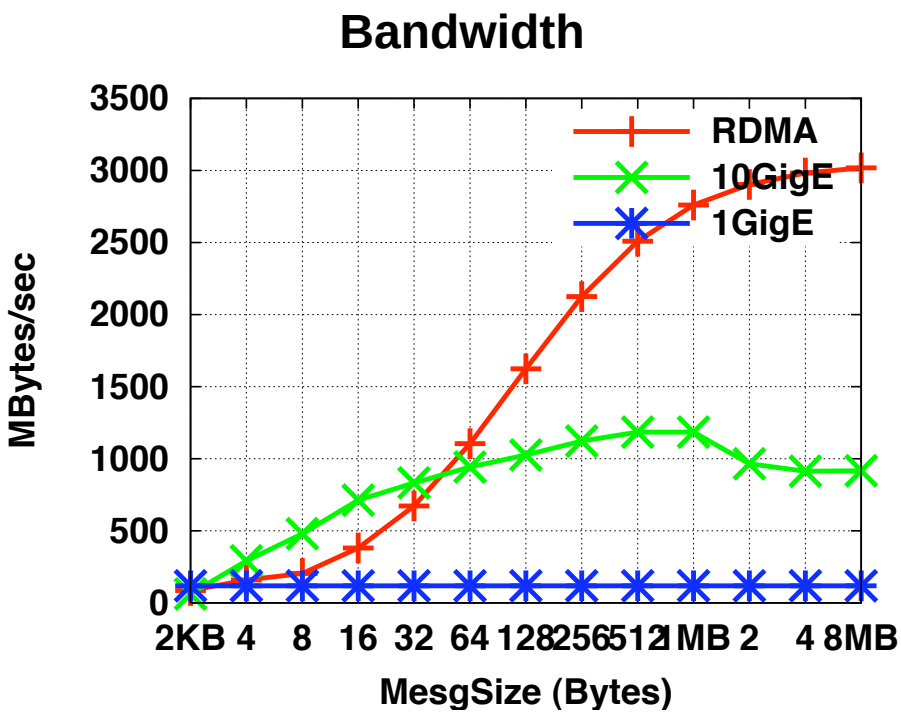
- Overview
  - Contemporary Network Technologies & InfiniBand
  - Emergence of RDMA on WAN
- Globus on InfiniBand
- Globus Zero-copy Noncontiguous I/O on IB
- Performance Evaluation
- Conclusion

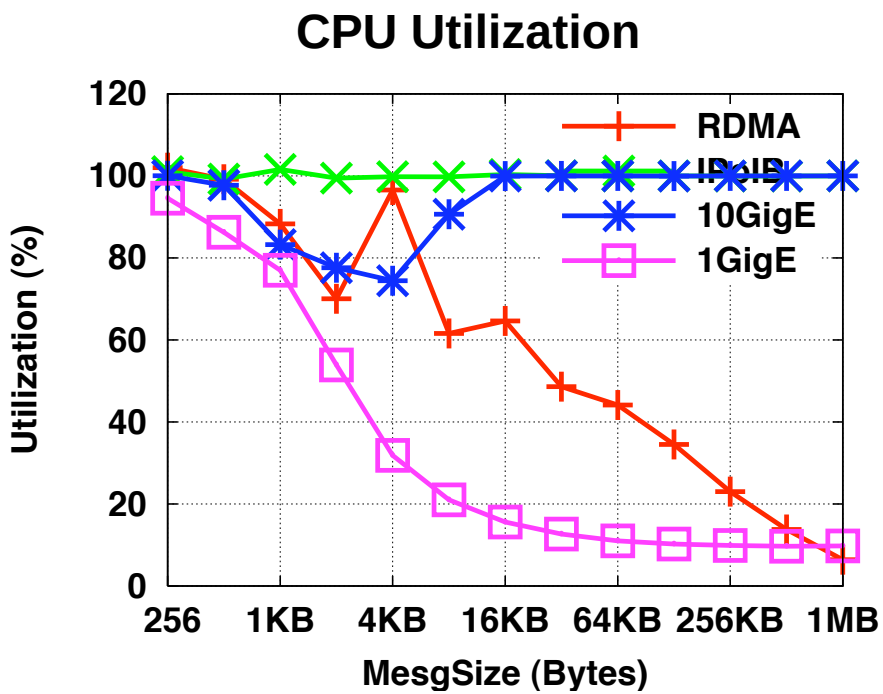
Globus on InfiniBand

- Design Globus with InfiniBand RDMA support
  - Maintain authentication and security protocols
  - Extending XIO with RDMA (EXIO)
- Main Issues
  - Connection management
  - Message transmission protocols



Performance Benefits of RXIO





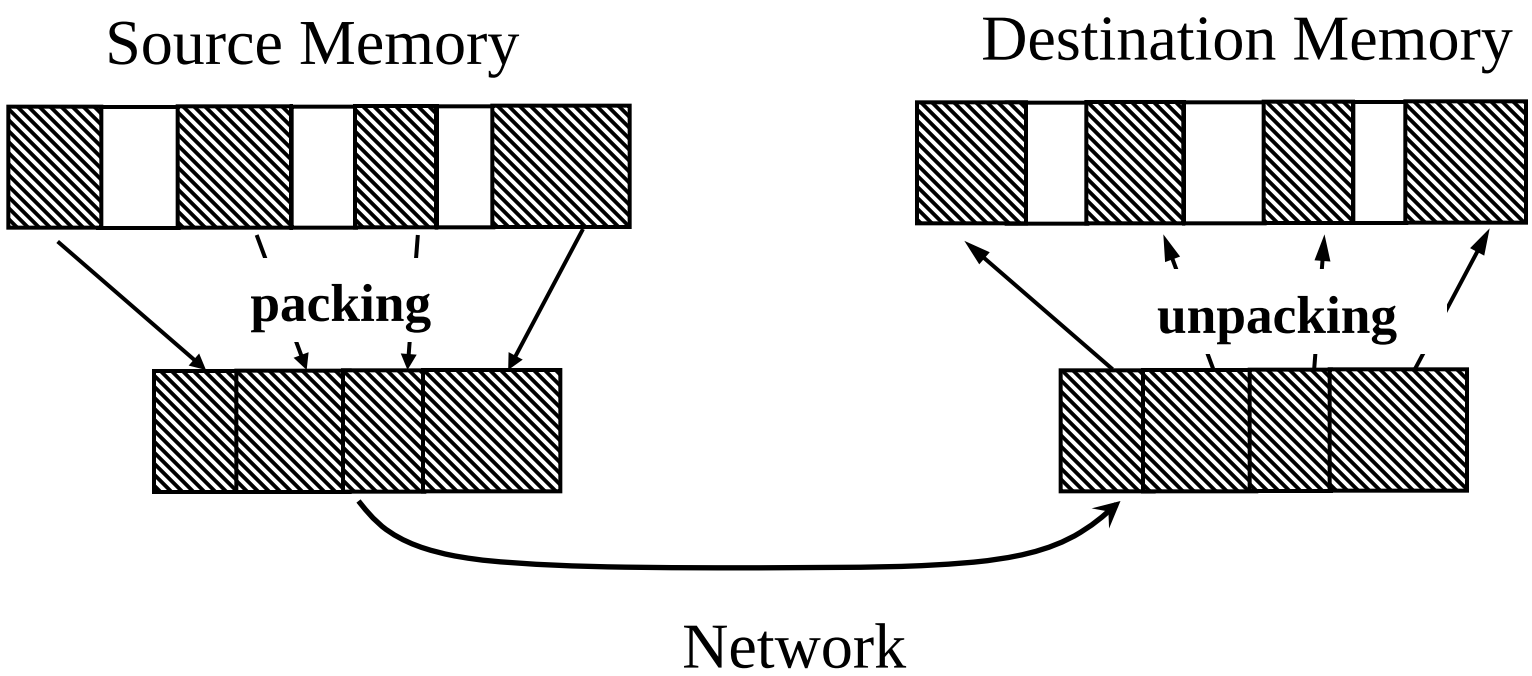
- Improve GridFTP bandwidth by three times compared to 10GigE
- At the same time, dramatically reduce the CPU utilization

Presentation Outline

- Overview
  - Contemporary Network Technologies & InfiniBand
  - Emergence of RDMA on WAN
- Globus on InfiniBand
- Globus Zero-copy Noncontiguous I/O on IB
- Performance Evaluation
- Conclusion

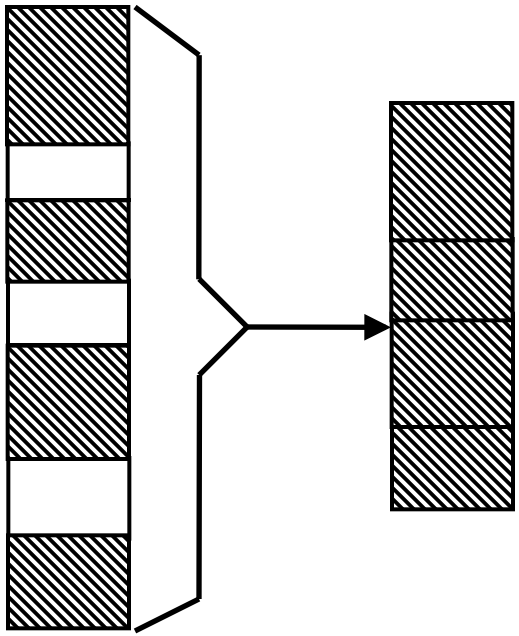
Packing and Unpacking

- Data packing on the sender side
- Unpacking on the receiver side
- Consume memory bandwidth and CPU



InfiniBand RDMA Scatter/Gather

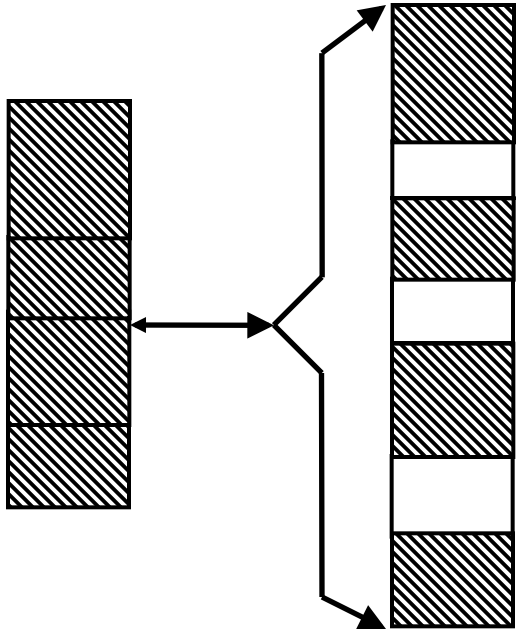
RDMA Write



source
(**local**)

destination

RDMA Read

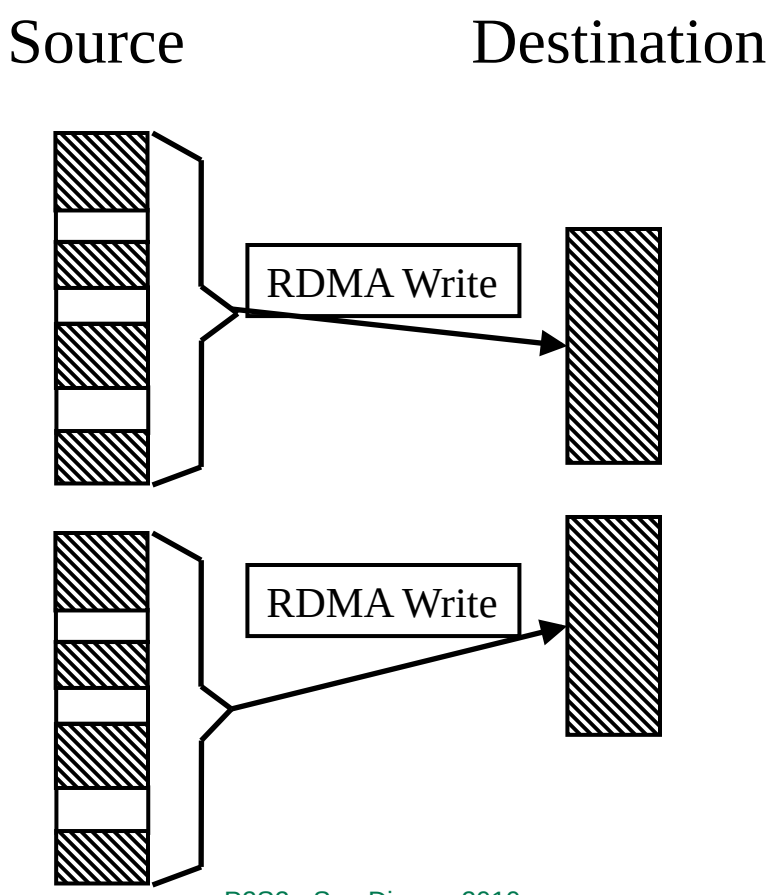


Source

Destination
(**local**)

Enabling Globus Noncontiguous I/O

- Leverage RDMA write for this prototype
- For each contiguous region at the destination, gather enough fragments from the source memory



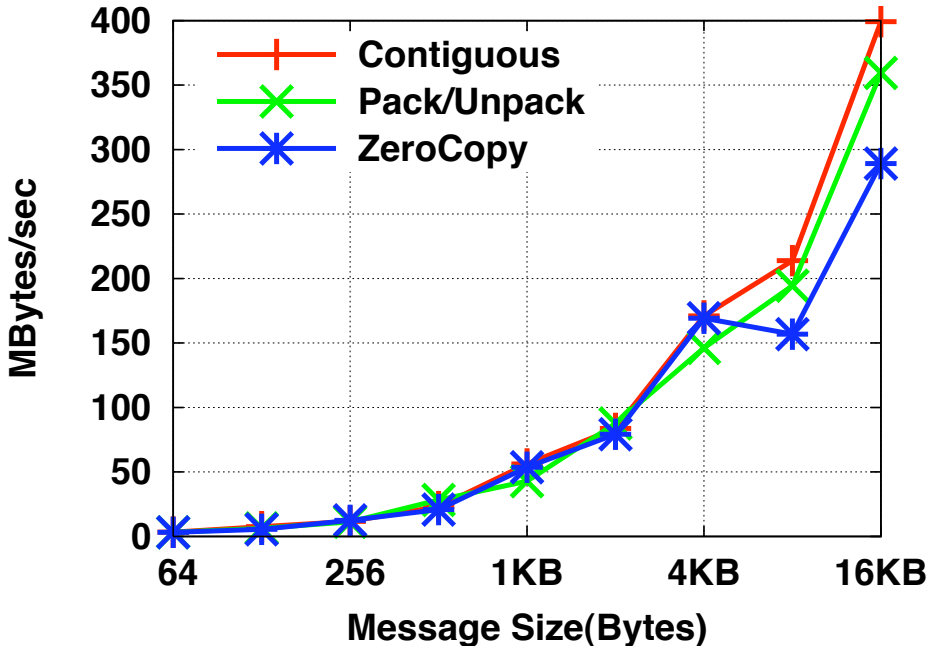
Presentation Outline

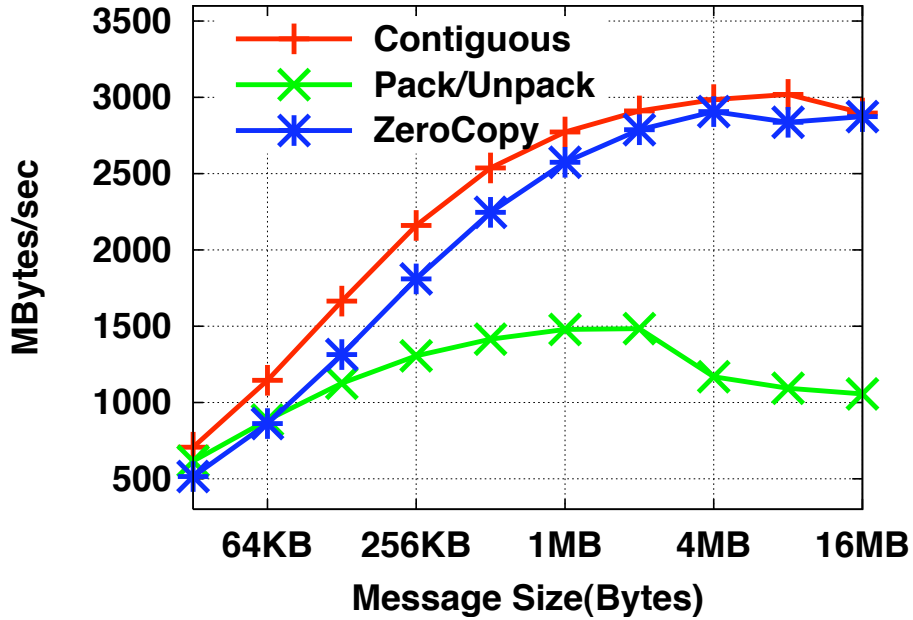
- Overview
  - Contemporary Network Technologies & InfiniBand
  - Emergence of RDMA on WAN
- Globus on InfiniBand
- Globus Zero-copy Noncontiguous I/O on IB
- Performance Evaluation
- Conclusion

Configuration of Experimental System

- Dual-socket, Quad-core Supermicro servers
- QDR InfiniBand Connect-X adaptors
  - 36-port QDR switch
- Myrinet 10-GigE cards
- Centos-5.2 with Linux-2.6.18
- OFED-1.4.2 software stack
- Globus-1.5 release

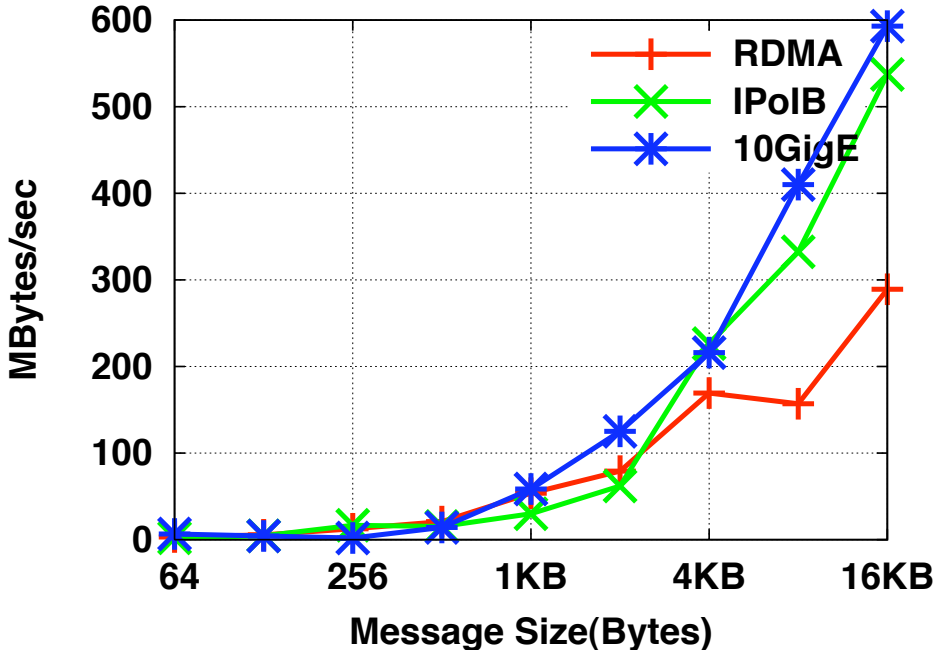
Comparison of Alternative Implementations of Noncontiguous I/O

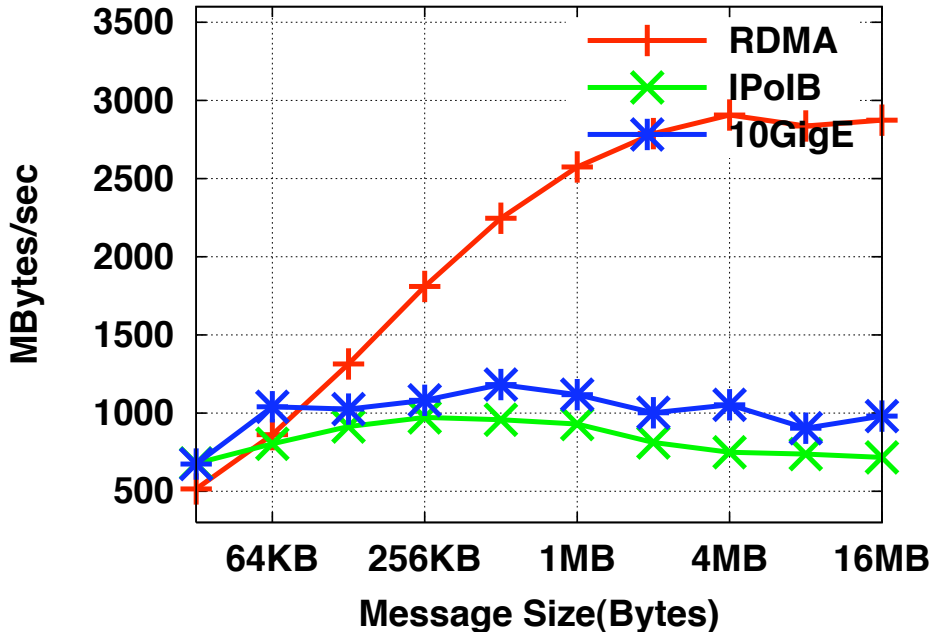




- Significant performance advantages for large messages
- Closely match that of contiguous I/O

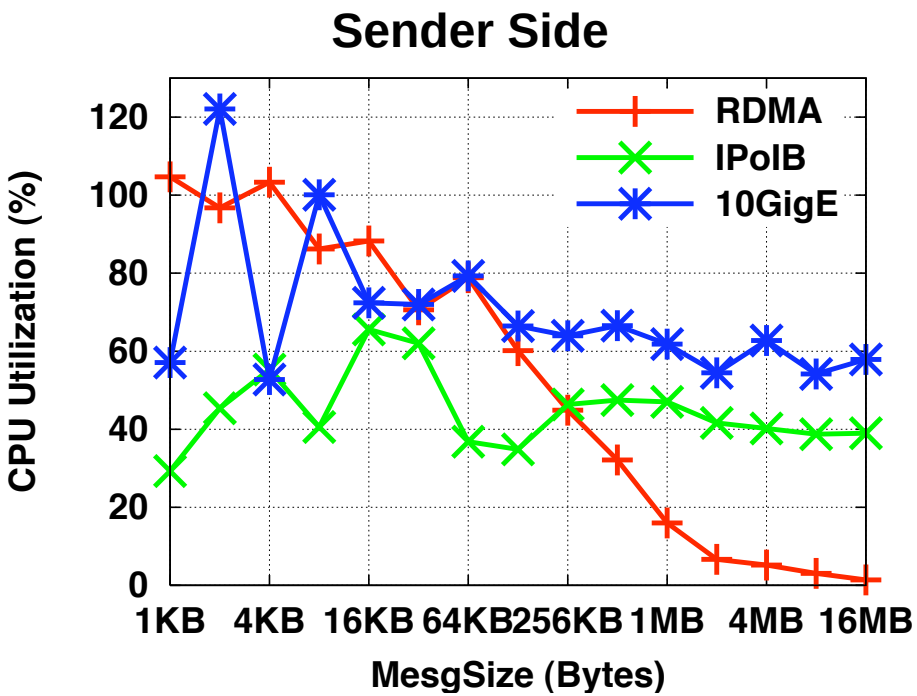
Comparison of Different Networks

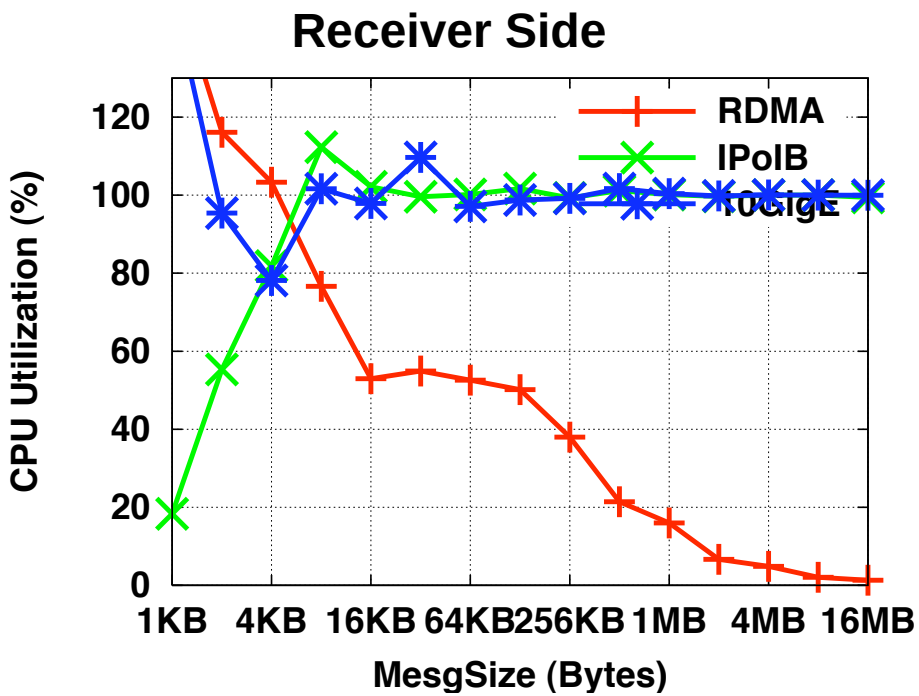




- Significant Improvement for large messages
- Medium messages need more turning

Benefits in CPU Utilization





- Achieve better bandwidth for small messages without comparable CPU utilization
- Dramatically reduce the CPU utilization for large messages

Conclusion and Future Perspectives

- Demonstrated that IB/RDMA can significantly improve the bandwidth of Globus
- Demonstrated that noncontiguous RDMA scatter/gather is beneficial to Globus
- While increasing the bandwidth of Globus, IB/RDMA significantly cuts down the CPU utilization
- Plan to identify real-world applications to showcase the benefits of IB/RDMA
- Need to secure a RDMA-capable long distance network

Acknowledgement

- Parallel Architecture and System Laboratory
  - More: http://pasl.eng.auburn.edu/
- Faculty
  - Weikuan Yu
  - Bongen Gu
    - Visiting Professor
- Students
  - Yuan Tian
  - Xinyu Que
  - Yandong Wang
  - Bin Wang
  - Cong Xu
  - Adarsh Jain