

Fall 2009 Reading Group: Presentation #4

“Maximizing MPI Point-to-Point Communication Performance on RDMA-enabled Clusters with Customized Protocols”

“Efficient High Performance Collective Communication for Cell Blade”

Shanyuan Gao

University of North Carolina at Charlotte

September 29, 2009

Paper 1

Maximizing MPI Point-to-Point Communication Performance on RDMA-enabled Clusters with Customized Protocols

- ▶ Matthew Small, Xin Yuan
- ▶ Florida State University

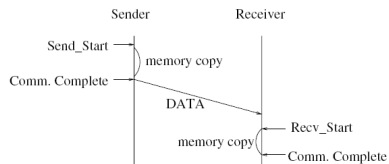
Paper 1 Introduction

- ▶ MPI point-to-point communications are realized with two internal protocols — *eager* protocol and *rendezvous* protocol
- ▶ Traditional *sender-initiated* protocols are simple, but slow
- ▶ Modern networks such as InfiniBand, and Myrinet can drastically reduce the communication cost
- ▶ Complex protocols can take advantage of modern networks and improve the performance of point-to-point communication

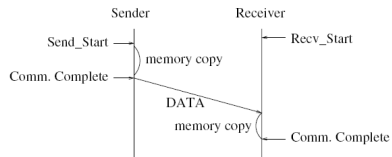
Paper 1 Contribution

Integration of four protocols:

- ▶ Eager protocol



(a) Receive posted after data arrived

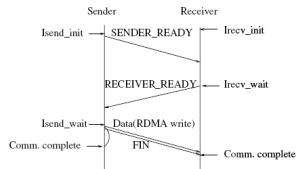


(b) Receive posted before data arrived

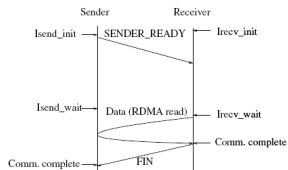
Paper 1 Contribution

Integration of four protocols:

- ▶ Eager protocol
- ▶ Sender-initiated rendezvous protocol



(a) RDMA write based protocol

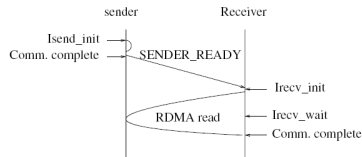


(b) RDMA read based protocol

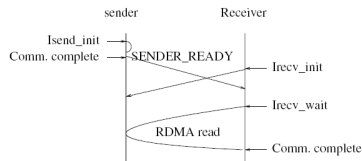
Paper 1 Contribution

Integration of four protocols:

- ▶ Eager protocol
- ▶ Sender-initiated rendezvous protocol
- ▶ Hybrid protocol



(a) hybrid protocol (sender early)

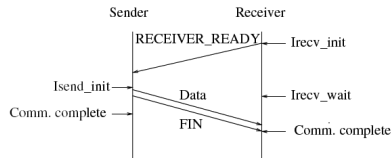


(b) hybrid protocol (both)

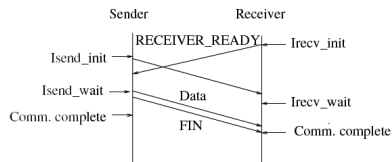
Paper 1 Contribution

Integration of four protocols:

- ▶ Eager protocol
- ▶ Sender-initiated rendezvous protocol
- ▶ Hybrid protocol
- ▶ Receiver-initiated rendezvous protocol



(a) Receiver-initiated rendezvous(receiver early)



(b) Receiver-initiated rendezvous(both)

Paper 1 Results

- ▶ Pingpong benchmark: less communication overheads and higher performance than traditional rendezvous protocol
- ▶ Progress benchmark: the proposed protocol can overlap the computation and communication
- ▶ NPB2.4 benchmark: significant performance gain for BT and SP

Paper 1 Implications

- ▶ The feature of modern network, such as RDMA can facilitate traditional MPI operations
- ▶ A complex protocol can efficiently overlap the computation and communication
- ▶ We have AIREN network and DMA. What can we do?

Paper 2

Efficient High Performance Collective Communication for the Cell Blade

- ▶ Qasim Ali, Samuel P. Midkiff, Vijay S. Pai
- ▶ Purdue University

Paper 2 Introduction

- ▶ Collective communication are important and time-consuming
- ▶ The hardware feature of Cell processor can benefit the communication and computation

Paper 2 Contribution

- ▶ Implemented five new algorithms targeting common collective communication operations: barrier, reduce, broadcast, all-gather and all-reduce
- ▶ Measured the performance of the new algorithms, compared the results with Cell Messaging Library (CML) and Buffered Mode MPI (BMM)
- ▶ The testing results showed that the proposed algorithms are up to 19.21 times faster than previously-presented algorithms

Paper 2 Implications

- ▶ Barrier on Cell: 0.3 μ s for 8 PPE's on one Cell, 0.5 μ s for 16 PPE's on two Cell (connected via BIF (20GB/s))
- ▶ Hardware MPI_Barrier on our cluster, 29 μ s for 8 nodes, 31 μ s for 16 nodes (offchip, AIREN network (4Gb/s))
- ▶ Similiar collective communications onchip? Barrier for Blast core?