NETWORK-BASED COMPUTING LABORATORY

# Can High-Performance Interconnects Benefit Hadoop Distributed File System?

**Sayantan Sur** Hao Wang Jian Huang

Xiangyong Ouyang D. K. Panda

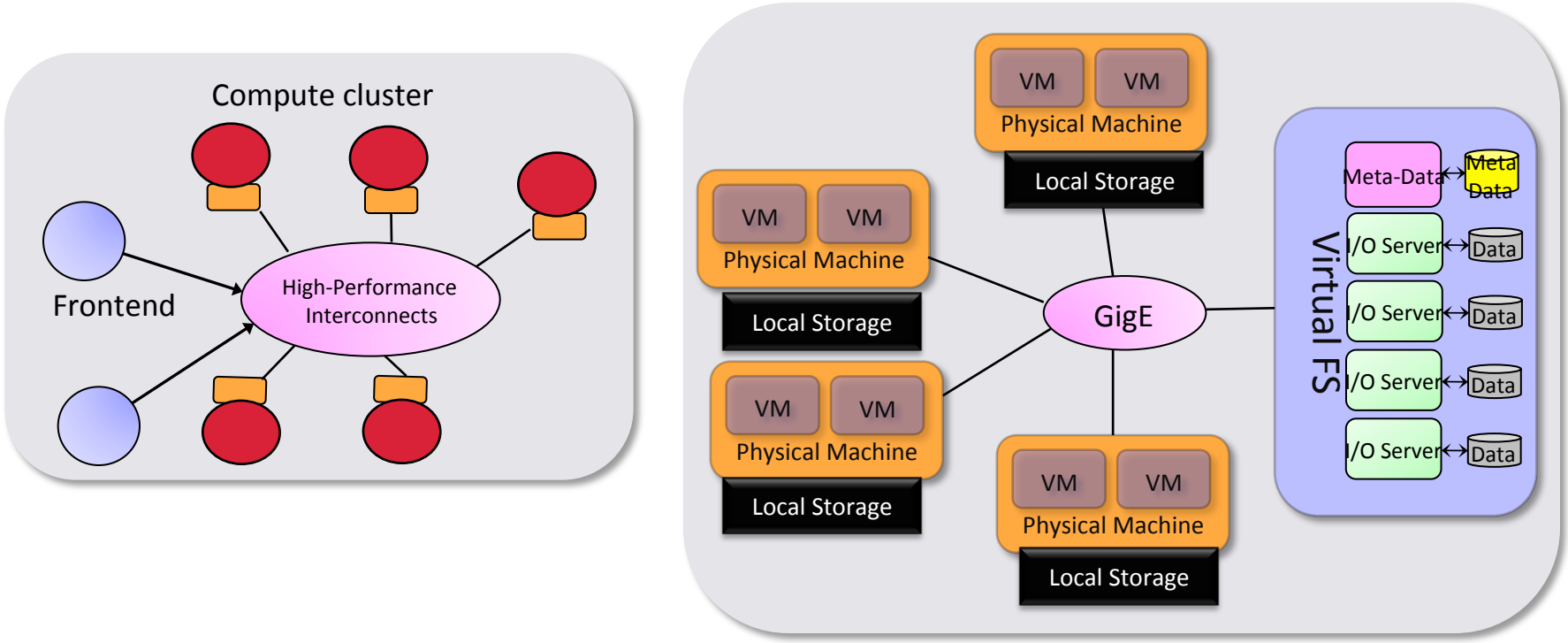
*Network-Based Computing Laboratory*
*Department of Computer Science and Engineering*
*The Ohio State University, USA*

OHIO STATE

# Introduction

- MapReduce – scalable model to process Petabytes
- Hadoop MapReduce framework widely adopted
  - Hadoop Distributed Filesystem (HDFS) provides core storage, distribution and fault-tolerance features
  - Designed with Gigabit Ethernet and Sockets in mind
- The field of High-Performance Computing (HPC) has adopted advanced interconnects
  - Low latency, High Bandwidth
  - Low CPU cycle requirement
  - InfiniBand, 10 Gigabit Ethernet are two examples
- Solid State Drives providing improved IO characteristics
- Can HDFS benefit from these two emerging technologies?

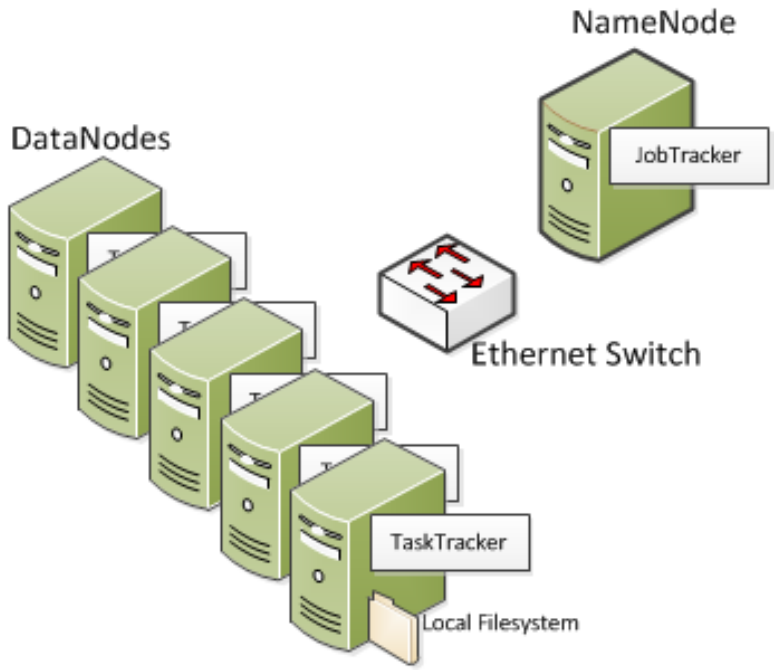
# Typical HPC and Cloud Computing Deployments




- HPC system design is interconnect centric
- Cloud computing environment has complex software and historically relied on sockets and ethernet

# HDFS Architecture

- HDFS is a distributed user-level file system
- Provides fault-tolerance by replicating data blocks
- Block size typically 64MB and replicated three times (possibly in different racks)
- Dedicated NameNode to store information on data blocks
- DataNodes just store blocks and schedule Map-reduce computation jobs
- Dedicated JobTracker to track jobs (and failure)



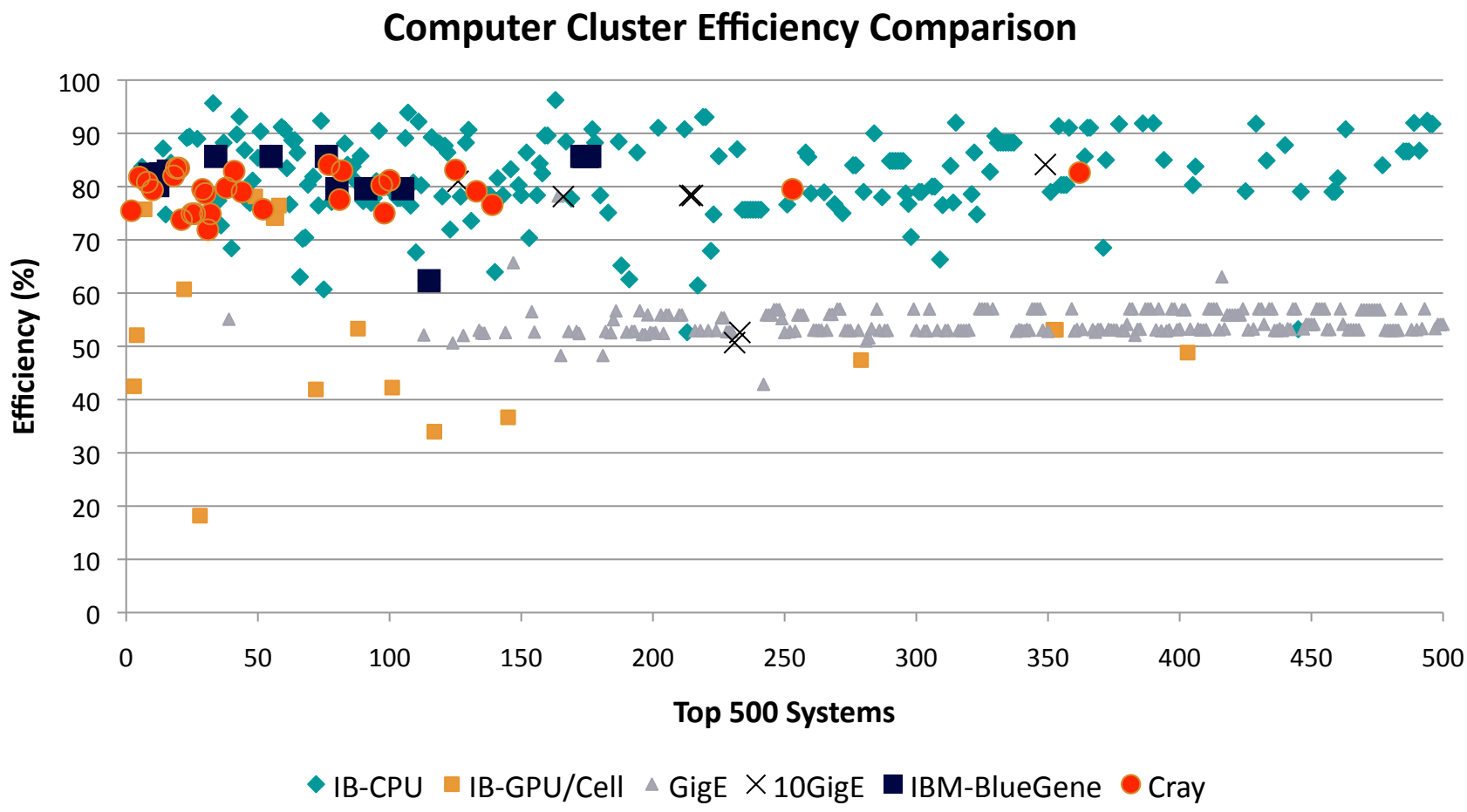
# InfiniBand and 10 Gigabit Ethernet

- InfiniBand is an industry standard packet switched network
- Has been increasingly adopted in HPC systems
- User-level networking with OS-bypass (verbs)

- 10 Gigabit Ethernet follow up to Gigabit Ethernet
- Provides user-level networking with OS-bypass (iwarp)
- Some vendors have accelerated TCP/IP by putting it on the network card (hardware offload)

- **Convergence**: possible to use both through OpenFabrics networking stack
  - Same software, different interconnects

# InfiniBand Efficiency in Top500 List



# Outline

- Introduction
- Problem Statement
- Modern Interconnects and Protocols
- Experimental Results & Analysis
- Conclusions & Future Work

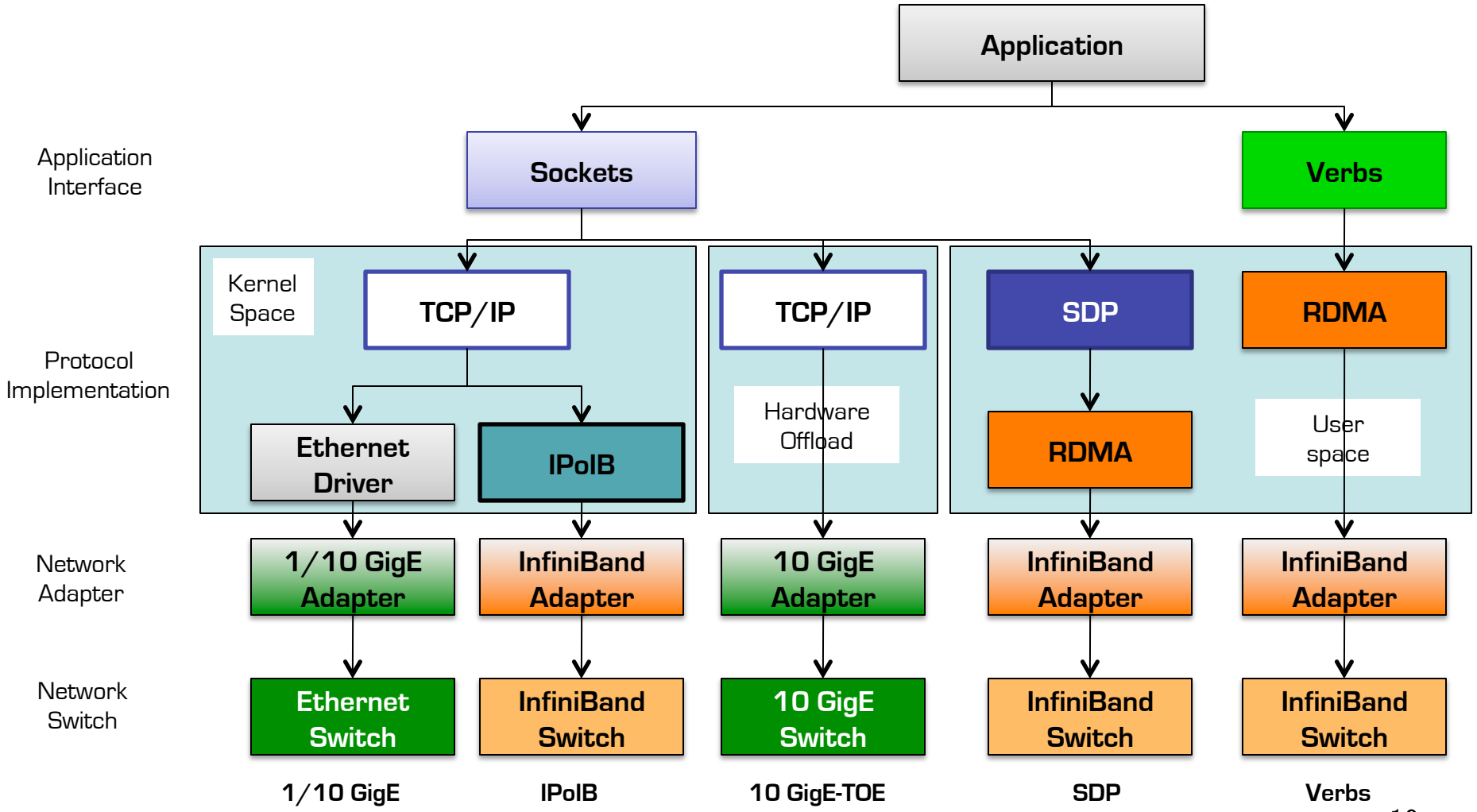
# Problem Statement

- Can High-Performance Interconnects help HDFS performance significantly?
- How much benefit is possible without any software modifications?
- Can emerging SSD technology complement performance advantages of high-performance interconnects?

# Outline

- Introduction
- Problem Statement
- Modern Interconnects and Protocols
- Experimental Results & Analysis
- Conclusions & Future Work

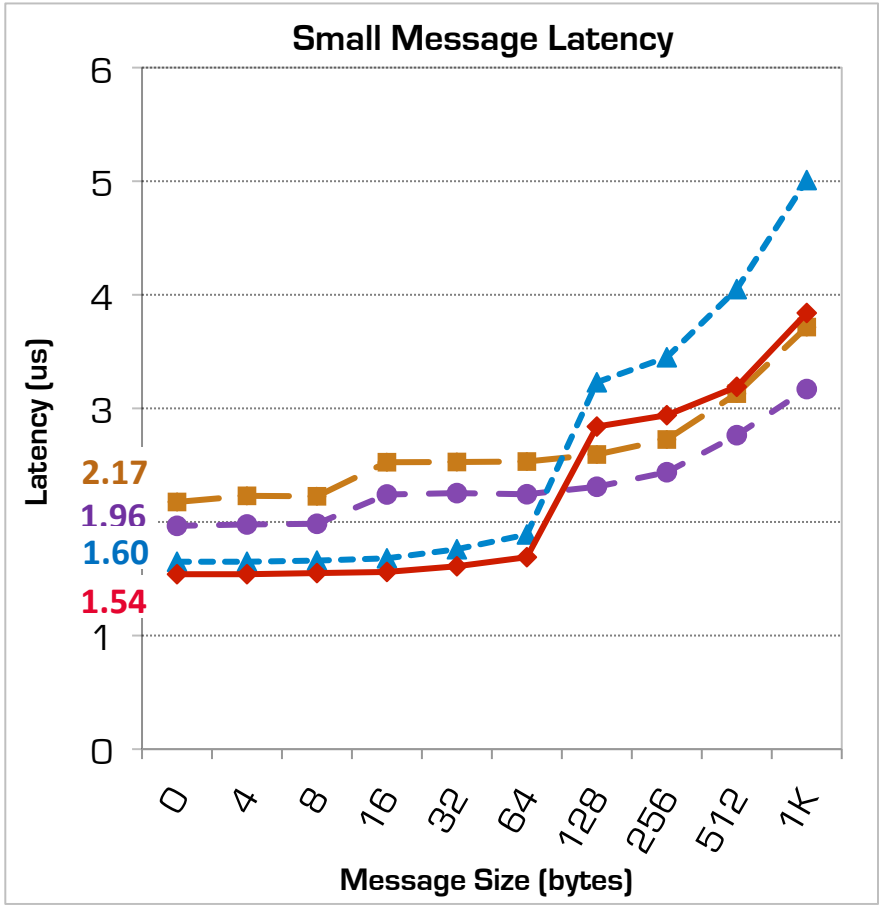
# Modern Interconnects and Protocols

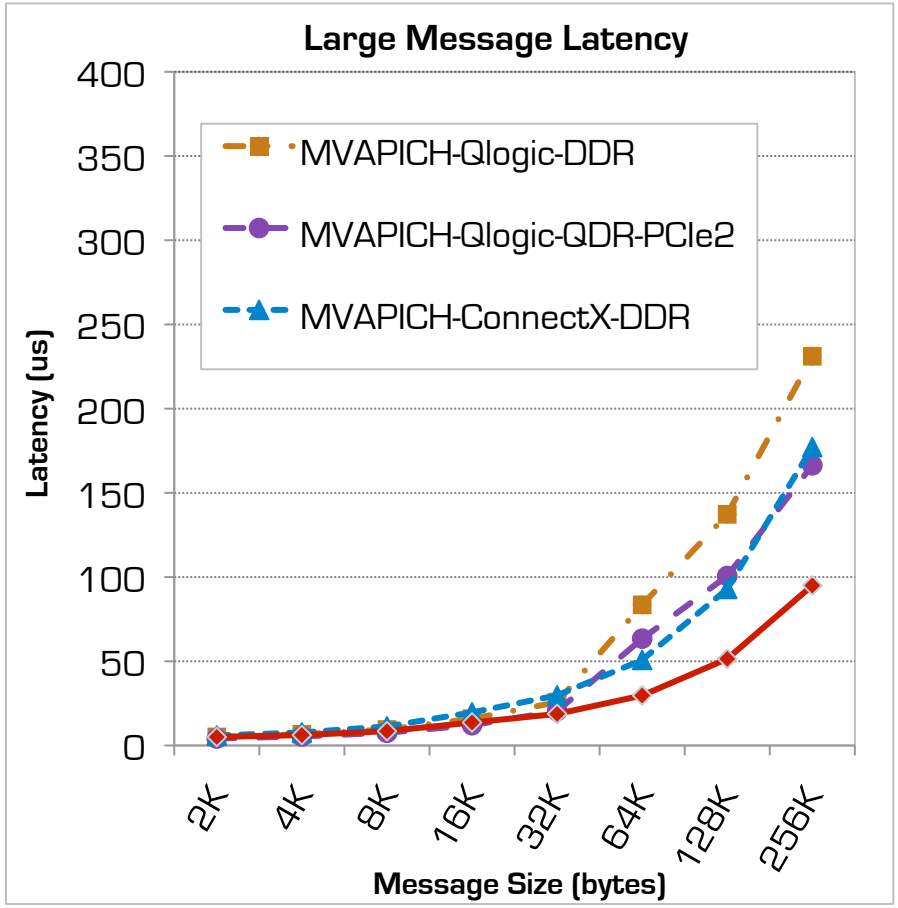


# MVAPICH and MVAPICH2 Software

- High Performance MPI Library for IB and HSE
  - MVAPICH (MPI-1) and MVAPICH2 (MPI-2.2)
  - Used by more than 1,300 organizations in 60 countries
  - More than 46,000 downloads from OSU site directly
  - Empowering many TOP500 clusters
    - 11th ranked 81,920-core cluster (Pleiades) at NASA
    - 15th ranked 62,976-core cluster (Ranger) at TACC
  - Available with software stacks of many IB, HSE and server vendors including Open Fabrics Enterprise Distribution (OFED)
  - http://mvapich.cse.ohio-state.edu

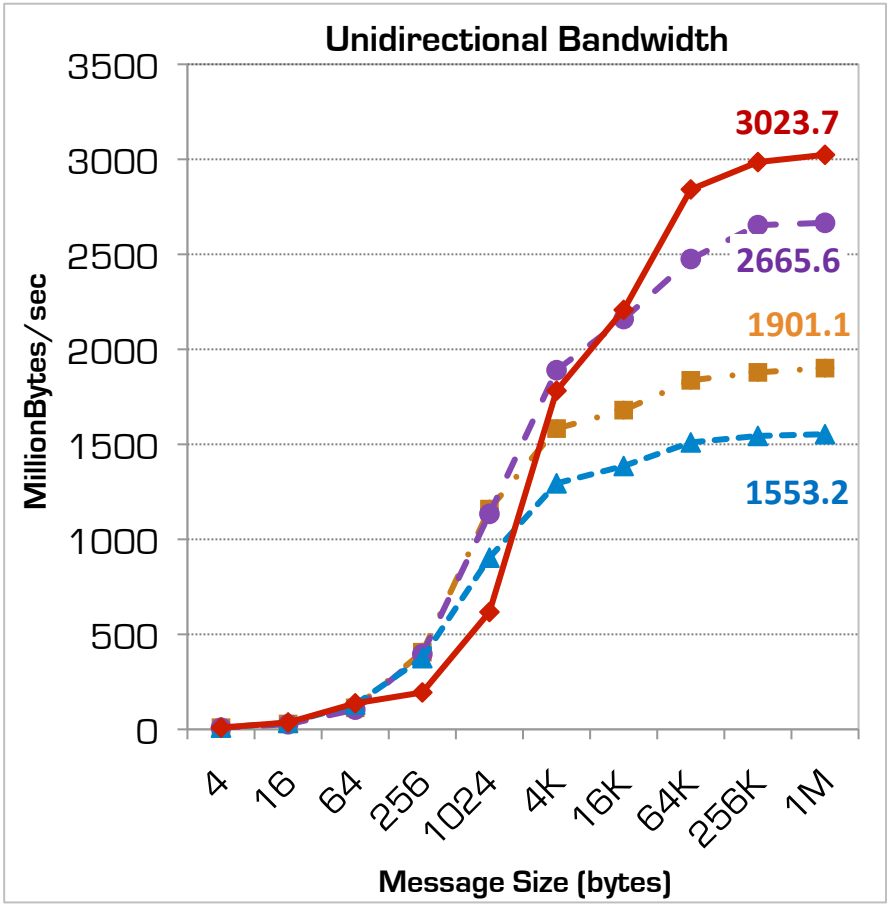
# One-way Latency: MPI over IB

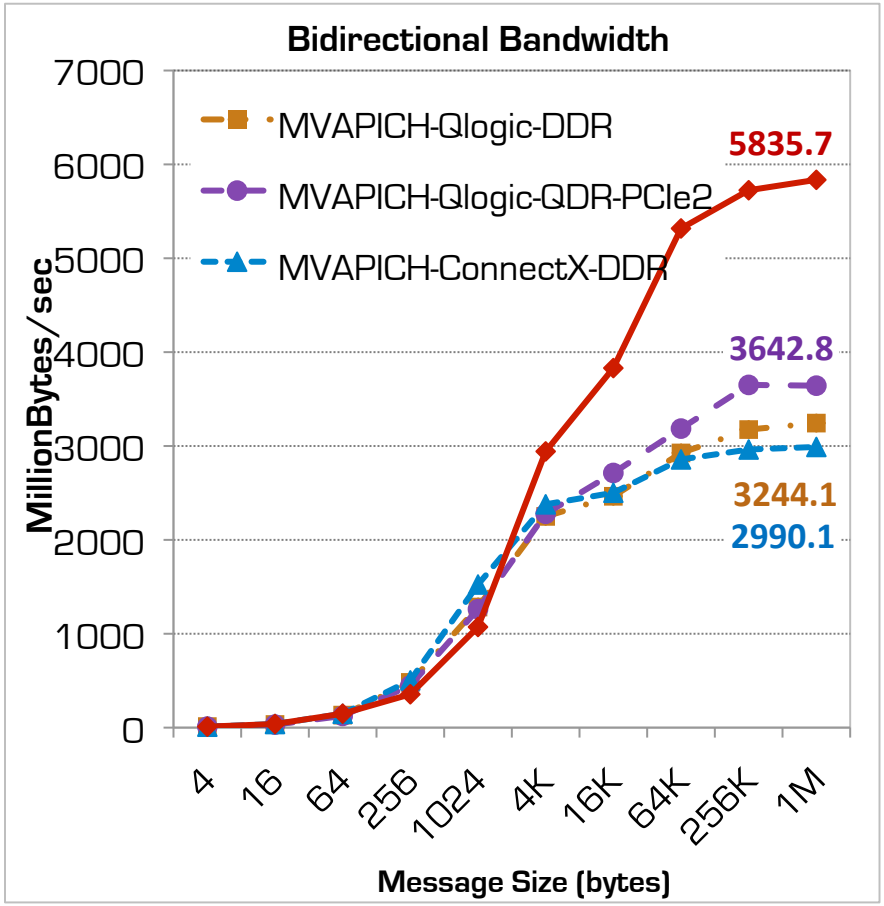



All numbers taken on 2.4 GHz Quad-core (Nehalem) Intel with IB switch

# Bandwidth: MPI over IB




All numbers taken on 2.4 GHz Quad-core (Nehalem) Intel with IB switch

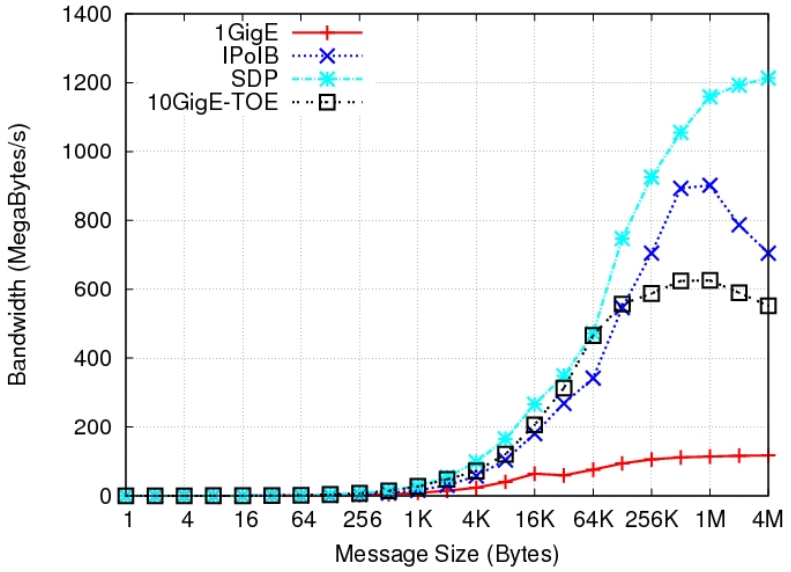
# Outline

- Introduction
- Problem Statement
- Modern Interconnects and Protocols
- Experimental Results & Analysis
- Conclusions & Future Work

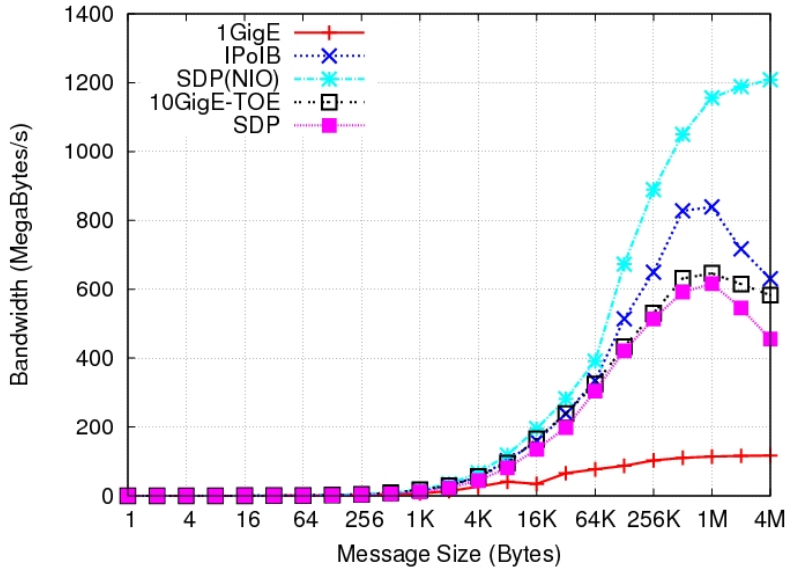
# Experimental Setup

- Hadoop 0.20.2
- Sun/Oracle Java 1.6.0
- Intel Xeon 2.33GHz Quad Core CPUs
- Main memory 6GB, 250GB Hard disk
- Intel X-25E 64GB SSD
- Mellanox MT25208 DDR (16Gbps) InfiniBand
- Chelsio T320 (10GbE)
- We dedicate one node as NameServer another as JobTracker
- We vary the DataNode from 2, 4 and 8

# Microbenchmark Level Evaluation



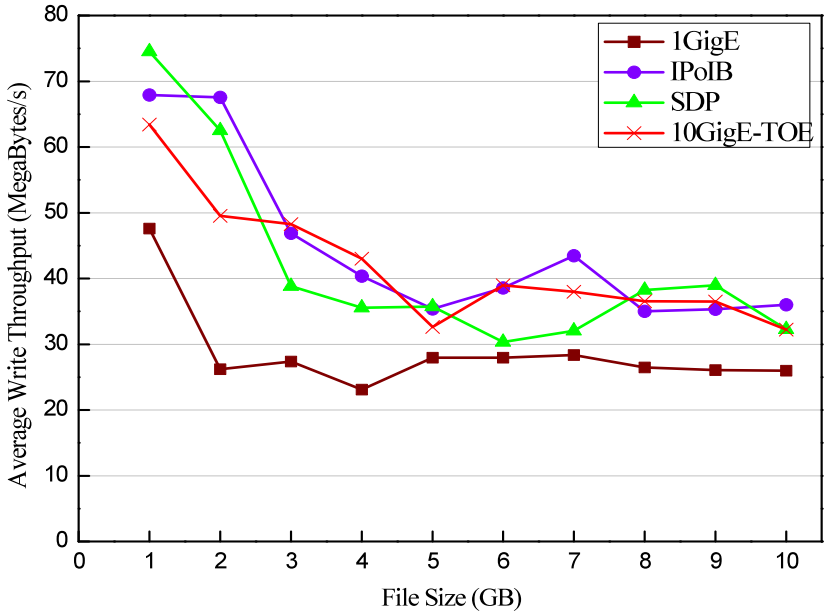

**Bandwidth with C version**

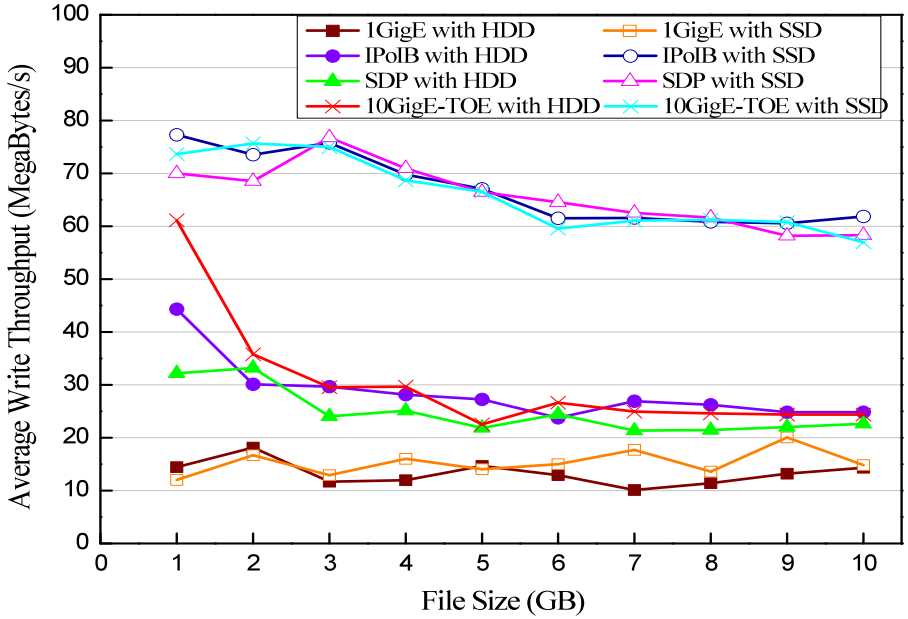



**Bandwidth with Java version**

- Sockets level ping-pong bandwidth test
- Client sends data to server; server receives data; sends an ack back
- Java performance depends on usage of NIO (allocateDirect)
- C and Java versions of the benchmark have similar performance
- HDFS does not use direct allocated blocks or NIO on DataNode

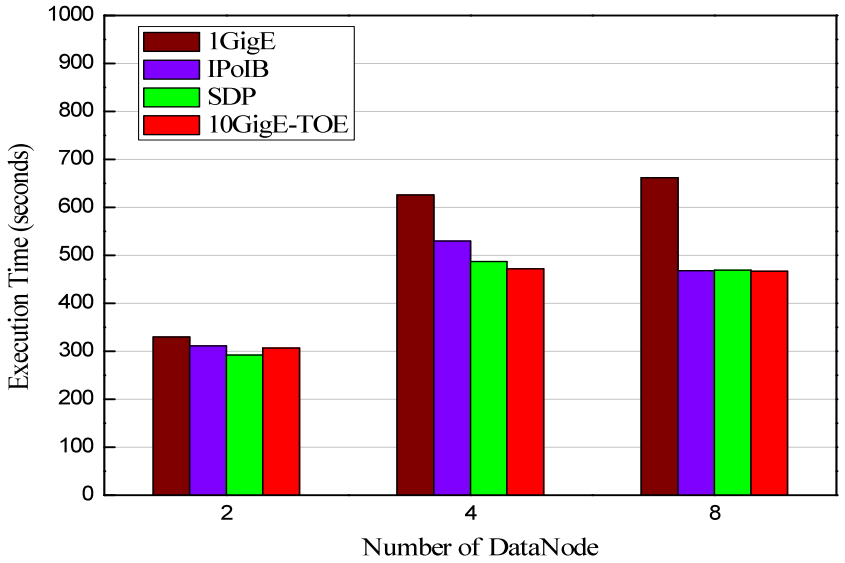
# DFS IO Write Performance

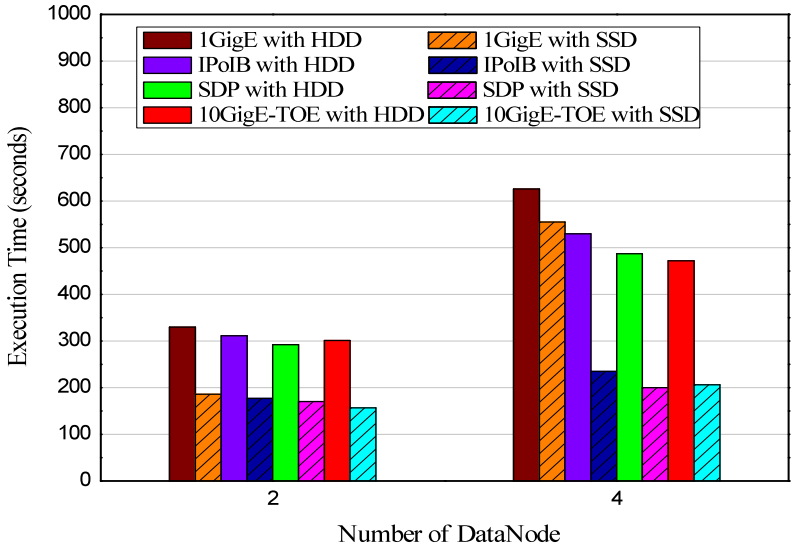



- DFS IO included in Hadoop, measures sequential access throughput
- We have two map tasks each writing to a file of increasing size (1-10GB)
- Eight data nodes for HDD (left) four data nodes for SSD (right)
- Significant improvement with IPoIB, SDP and 10GigE
- With SSD, performance improvement is almost seven or eight fold!
- SSD benefits not seen without using high-performance interconnect!

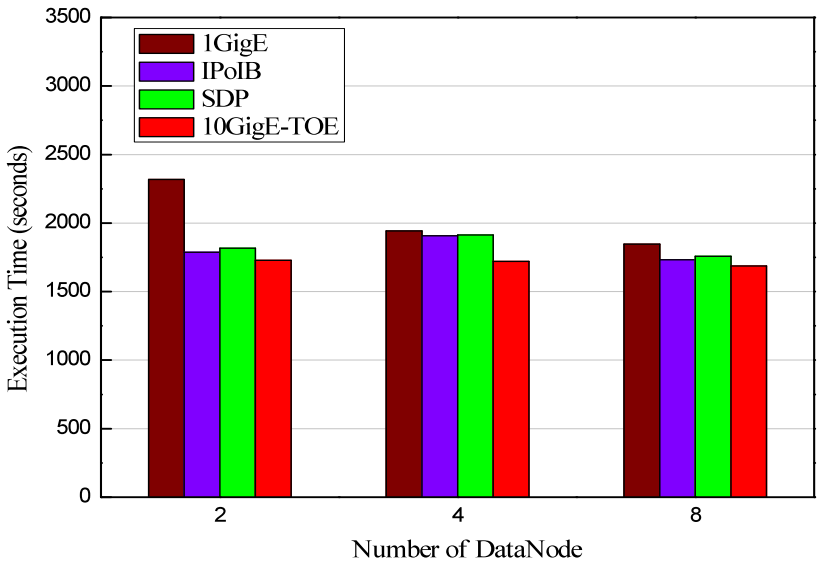
# RandomWriter Performance

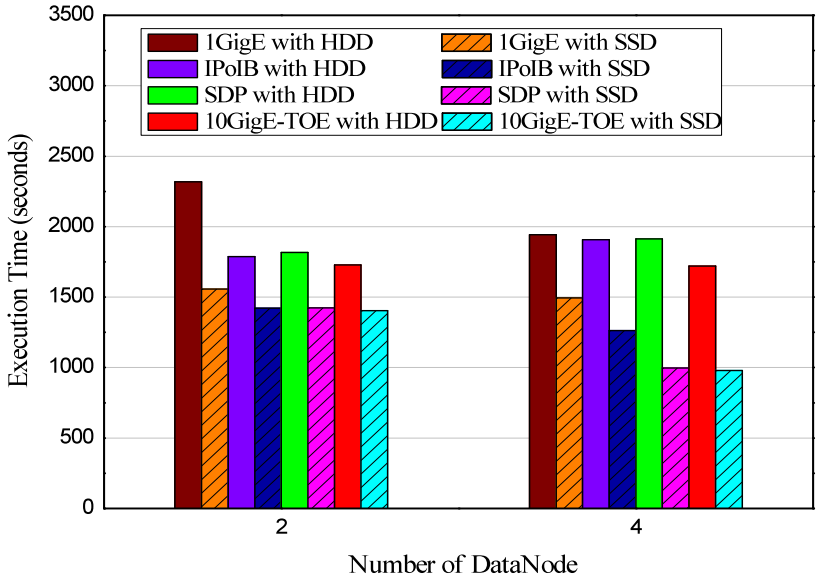



- Each map generates 1GB of random binary data and writes to HDFS
- 30% improvement for HDD using 8 DataNodes (IPoIB, SDP, 10GigE)
- SSD improves execution time by 50% with 1GigE for two DataNodes
- However, when using four DataNodes, unless IPoIB, SDP or 10GigE is used, benefits are not observed
- IPoIB, SDP and 10GigE can improve performance by 59% on four DataNodes

# Sort Benchmark



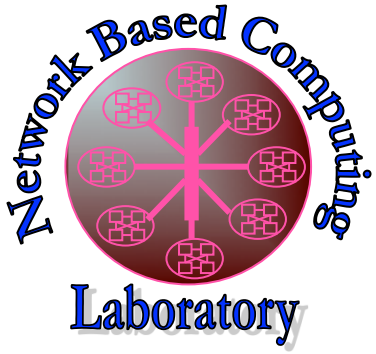

- Sort: baseline benchmark for Hadoop
- Bound by disk IO bandwidth for sort phase, but network performance bound for reduce phase
- SSD improves performance by 28% using 1GigE with two DataNodes
- Benefit of 48% on four DataNodes using SDP, IPoIB or 10GigE

# Conclusions and Future Work

- High-Performance interconnects can be used to boost performance of HDFS workloads
- Benefits are observed when SSD is used instead of Hard disk in combination with fast network
  - Disk IO is still a bottleneck in HDFS
  - Using SSDs alone cannot improve performance
  - Must couple High-Performance interconnect with SSDs
- HDFS design can be improved to use lower level communication (verbs) to further leverage advanced networks
- We are currently looking at more workloads and HBase performance

# Thank You!

{surs, wangh, huangjia, ouyangx, panda}@cse.ohio-state.edu





Network-Based Computing Laboratory

http://nowlab.cse.ohio-state.edu/

MVAPICH Web Page

http://mvapich.cse.ohio-state.edu/