Merge or Split: Mutual Influence between Big Data and HPC Techniques?

Panel at HPBDC '16

by

Dhabaleswar K. (DK) Panda

The Ohio State University

E-mail: panda@cse.ohio-state.edu

http://www.cse.ohio-state.edu/~panda

Q1: What is Impact of Big Data Techniques on HPC?

- Many HPC applications are generating huge amount of data
- MPI has been the dominant programming model for HPC to process the data
- MPI stack has been optimized on all different HPC hardware for the last 25 years
- Multiple things are happening:
  - To carry out Hadoop and Spark-like work using MPI
  - Hybrid programming (Spark and MPI together, Hadoop and MPI together)
  - New fault-tolerant techniques in parallel file systems (based on HDFS)

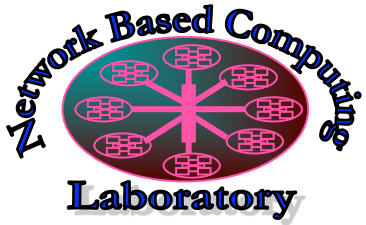
Q2: What is the Impact of HPC Techniques on Big Data?

- Already happening at multiple places
  - Using RDMA-enabled HPC networking technologies like InfiniBand and RoCE
  - Using Parallel File Systems (such as Lustre) for MapReduce applications
  - Exploitation of high-performance storage (such as SSD/ NVRAM) for Big Data middleware (e.g. Key-Value store)
- Hardware and software stacks are being enabled by multiple organizations
  - TACC, Cray, Intel, Mellanox, Cray, IBM, OSU, ..

Overview of the OSU High-Performance Big Data (HiBD) Project and Releases

- RDMA for Apache Spark
- RDMA for Apache Hadoop 2.x (RDMA-Hadoop-2.x)
  - Plugins for Apache, Hortonworks (HDP) and Cloudera (CDH) Hadoop distributions
- RDMA for Apache Hadoop 1.x (RDMA-Hadoop)
- RDMA for Memcached (RDMA-Memcached)
- OSU HiBD-Benchmarks (OHB)
  - HDFS and Memcached Micro-benchmarks
- http://hibd.cse.ohio-state.edu
- Users Base: 165 organizations from 22 countries
- More than 16,500 downloads from the project site
- RDMA for Apache HBase and Impala (upcoming)

Available for InfiniBand and RoCE



RDMA for Apache Spark Distribution

- High-Performance Design of Spark over RDMA-enabled Interconnects
  - High performance RDMA-enhanced design with native InfiniBand and RoCE support at the verbs-level for Spark
  - RDMA-based data shuffle and SEDA-based shuffle architecture
  - Non-blocking and chunk-based data transfer
  - Easily configurable for different protocols (native InfiniBand, RoCE, and IPoIB)
- Current release: 0.9.1
  - Based on Apache Spark 1.5.1
  - Tested with
    - Mellanox InfiniBand adapters (DDR, QDR and FDR)
    - RoCE support with Mellanox adapters
    - Various multi-core platforms
    - RAM disks, SSDs, and HDD
  - http://hibd.cse.ohio-state.edu

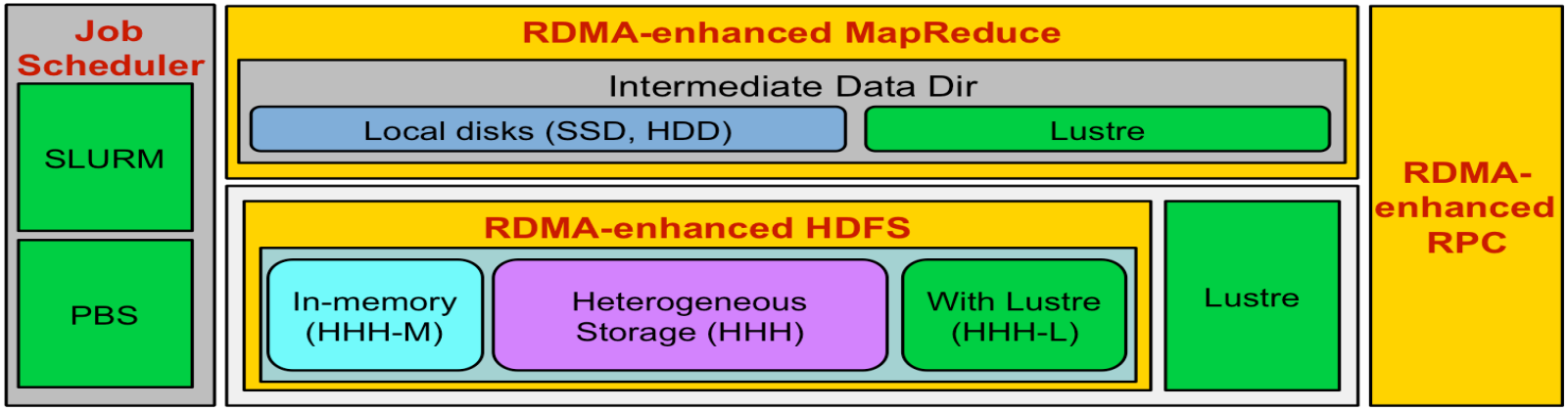
RDMA for Apache Hadoop 2.x Distribution

- High-Performance Design of Hadoop over RDMA-enabled Interconnects
  - High performance RDMA-enhanced design with native InfiniBand and RoCE support at the verbs-level for HDFS, MapReduce, and RPC components
  - Enhanced HDFS with in-memory and heterogeneous storage
  - High performance design of MapReduce over Lustre
  - Plugin-based architecture supporting RDMA-based designs for Apache Hadoop, CDH and HDP
  - Easily configurable for different running modes (HHH, HHH-M, HHH-L, and MapReduce over Lustre) and different protocols (native InfiniBand, RoCE, and IPoIB)
- Current release: 0.9.9
  - Based on Apache Hadoop 2.7.1
  - Compliant with Apache Hadoop 2.7.1, HDP 2.3.0.0 and CDH 5.6.0 APIs and applications
  - Tested with
    - Mellanox InfiniBand adapters (DDR, QDR and FDR)
    - RoCE support with Mellanox adapters
    - Various multi-core platforms
    - Different file systems with disks and SSDs and Lustre
  - http://hibd.cse.ohio-state.edu

RDMA for Memcached Distribution

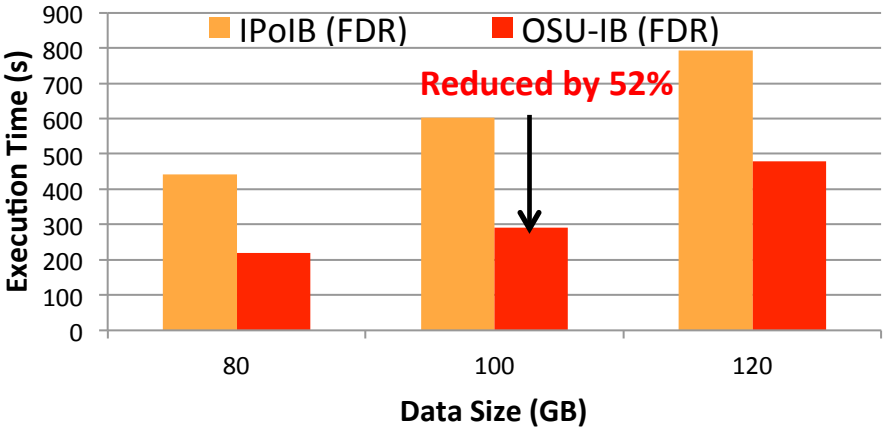
- High-Performance Design of Memcached over RDMA-enabled Interconnects
  - High performance RDMA-enhanced design with native InfiniBand and RoCE support at the verbs-level for Memcached and libMemcached components
  - High performance design of SSD-Assisted Hybrid Memory
  - Easily configurable for native InfiniBand, RoCE and the traditional sockets-based support (Ethernet and InfiniBand with IPoIB)
- Current release: 0.9.4
  - Based on Memcached 1.4.24 and libMemcached 1.0.18
  - Compliant with libMemcached APIs and applications
  - Tested with
    - Mellanox InfiniBand adapters (DDR, QDR and FDR)
    - RoCE support with Mellanox adapters
    - Various multi-core platforms
    - SSD
  - **http://hibd.cse.ohio-state.edu**

Different Modes of RDMA for Apache Hadoop 2.x



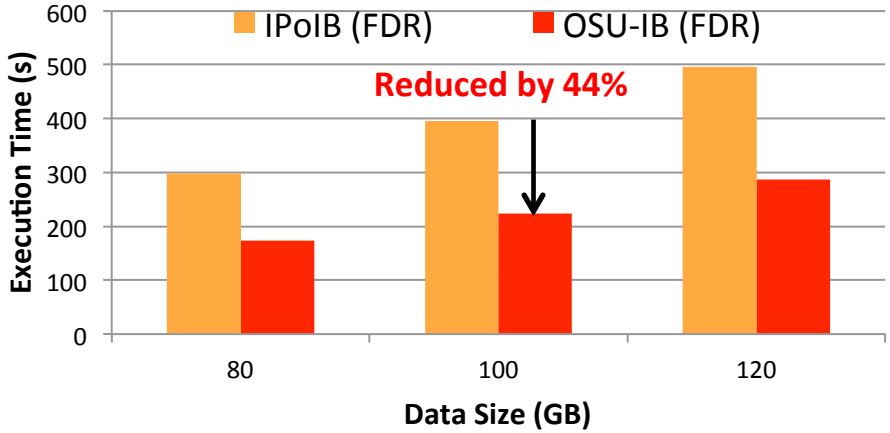
- **HHH**: Heterogeneous storage devices with hybrid replication schemes are supported in this mode of operation to have better fault-tolerance as well as performance. This mode is enabled by **default** in the package.
- **HHH-M**: A high-performance in-memory based setup has been introduced in this package that can be utilized to perform all I/O operations in-memory and obtain as much performance benefit as possible.
- **HHH-L**: With parallel file systems integrated, HHH-L mode can take advantage of the Lustre available in the cluster.
- **MapReduce over Lustre, with/without local disks**: Besides, HDFS based solutions, this package also provides support to run MapReduce jobs on top of Lustre alone. Here, two different modes are introduced: with local disks and without local disks.
- **Running with Slurm and PBS**: Supports deploying RDMA for Apache Hadoop 2.x with Slurm and PBS in different running modes (HHH, HHH-M, HHH-L, and MapReduce over Lustre).

Performance Numbers of RDMA for Apache Hadoop 2.x – Sort & TeraSort in TACC-Stampede



Cluster with 32 Nodes with a total of 128 maps and 57 reduces

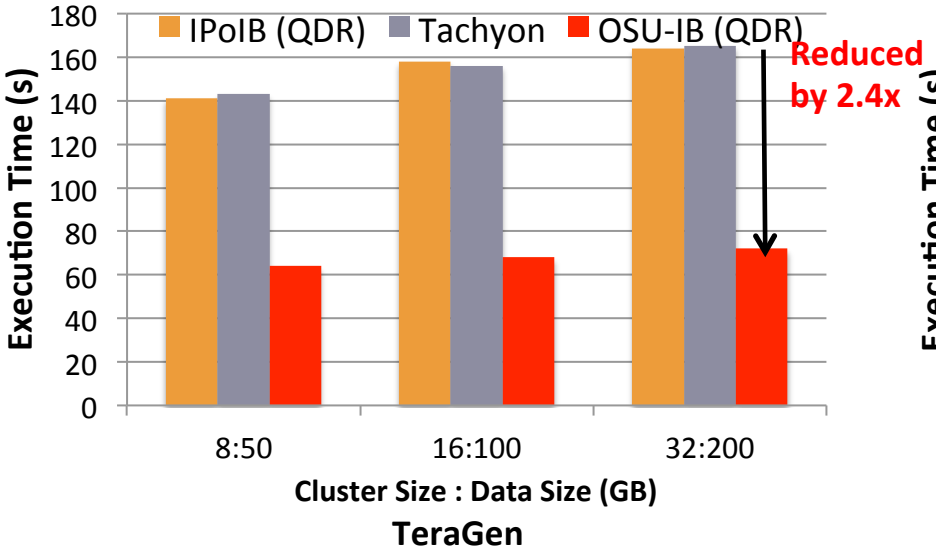
- Sort with single HDD per node
  - **40-52%** improvement over IPoIB for 80-120 GB data

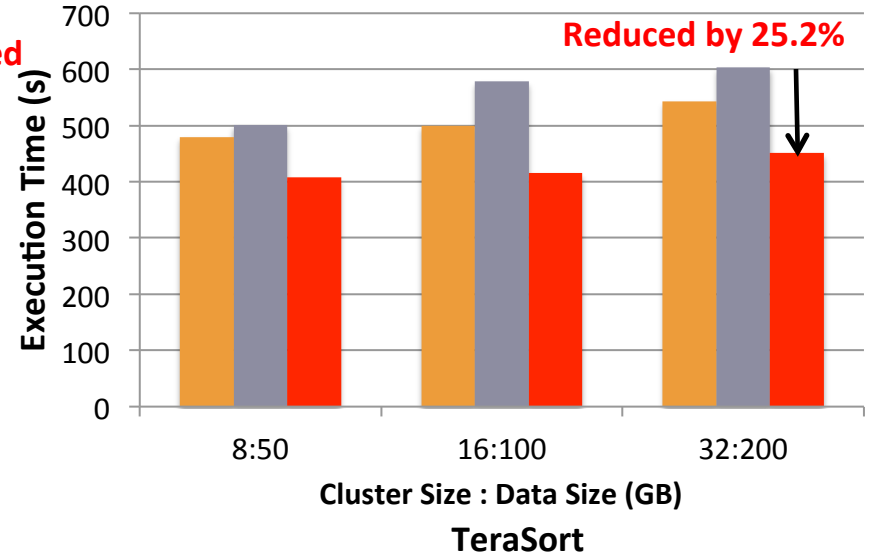


Cluster with 32 Nodes with a total of 128 maps and 64 reduces

- TeraSort with single HDD per node
  - **42-44%** improvement over IPoIB for 80-120 GB data

Evaluation with Spark on SDSC Gordon (HHH vs. Tachyon/Alluxio)

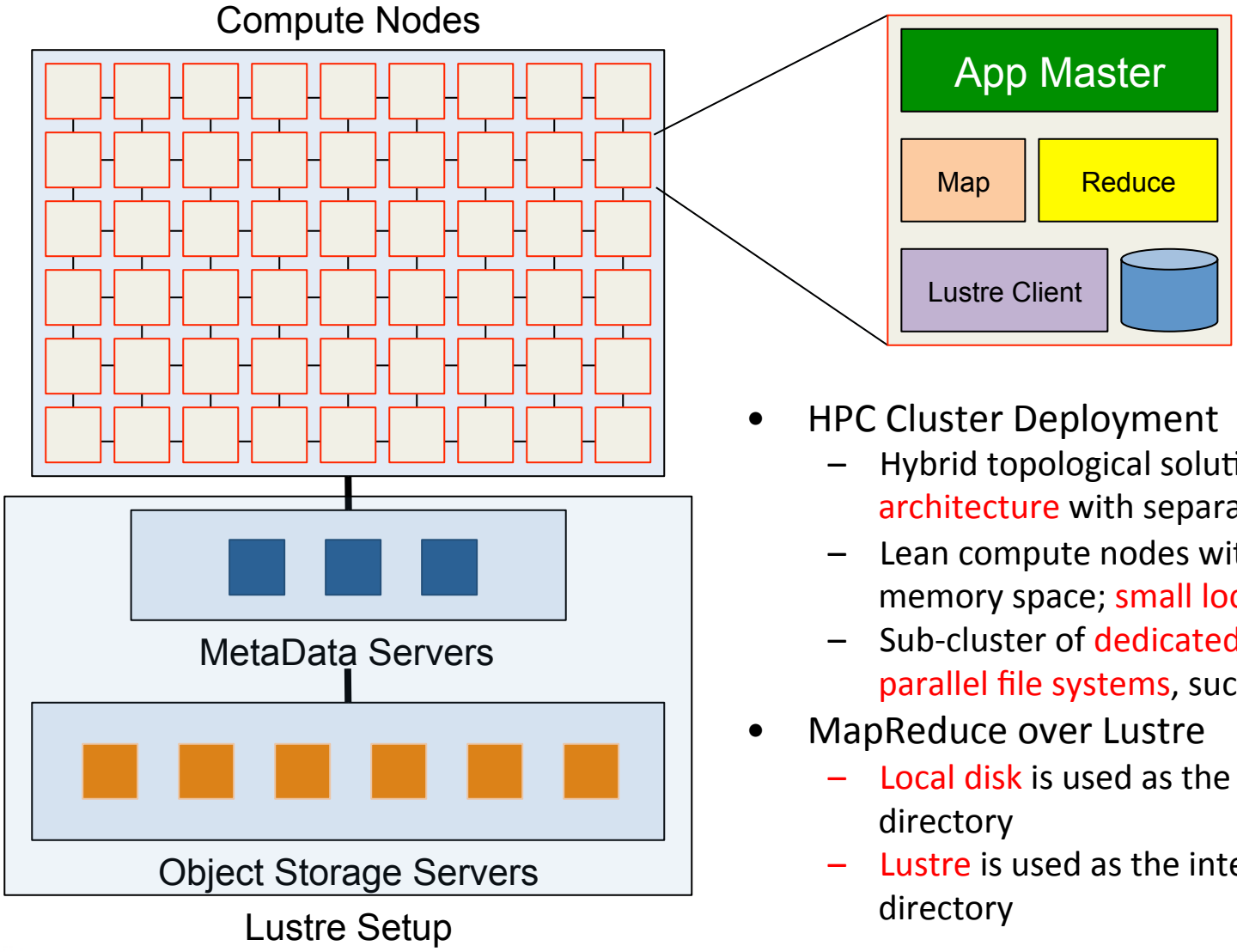




- For 200GB TeraGen on 32 nodes
  - Spark-TeraGen: HHH has 2.4x improvement over Tachyon; 2.3x over HDFS-IPoIB (QDR)
  - Spark-TeraSort: HHH has 25.2% improvement over Tachyon; 17% over HDFS-IPoIB (QDR)

N. Islam, M. W. Rahman, X. Lu, D. Shankar, and D. K. Panda, Performance Characterization and Acceleration of In-Memory File Systems for Hadoop and Spark Applications on HPC Clusters, IEEE BigData '15, October 2015

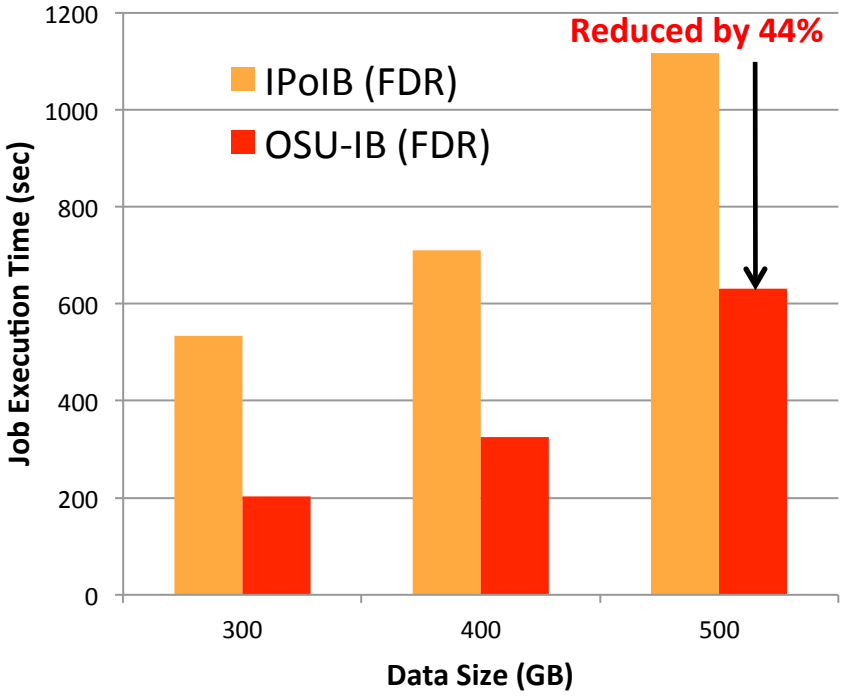
Optimize Hadoop YARN MapReduce over Parallel File Systems

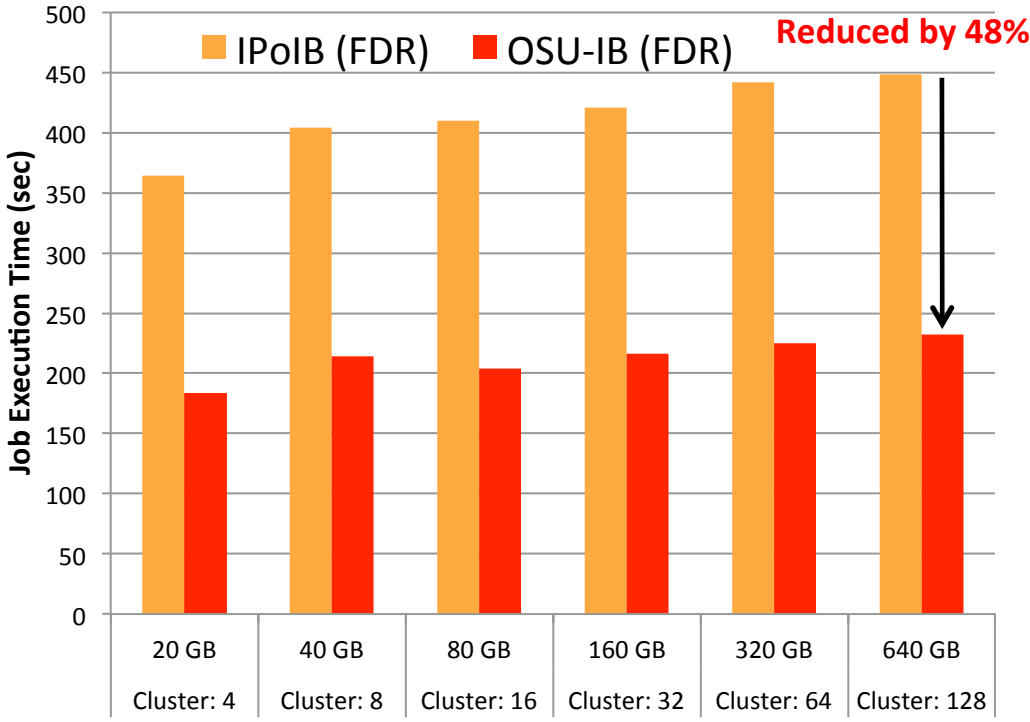


- HPC Cluster Deployment
  - Hybrid topological solution of Beowulf architecture with separate I/O nodes
  - Lean compute nodes with light OS; more memory space; small local storage
  - Sub-cluster of dedicated I/O nodes with parallel file systems, such as Lustre
- MapReduce over Lustre
  - Local disk is used as the intermediate data directory
  - Lustre is used as the intermediate data directory

Performance Improvement of MapReduce over Lustre on TACC-Stampede

- Local disk is used as the intermediate data directory

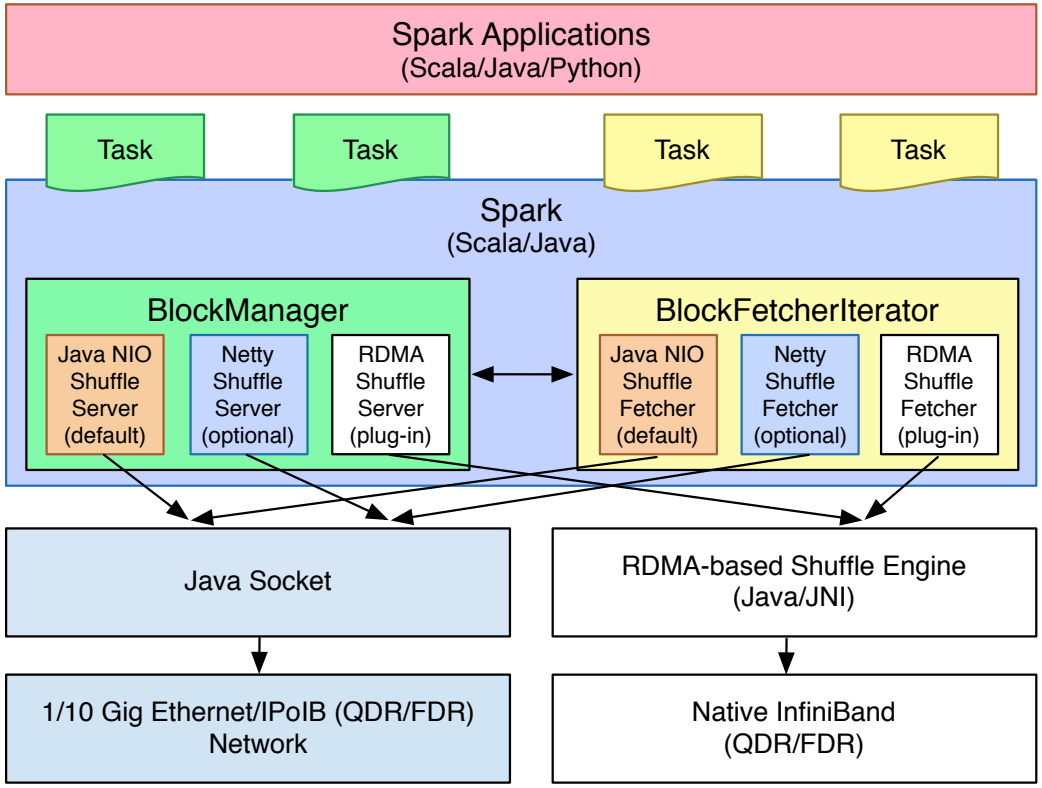




- For 500GB Sort in 64 nodes
  - 44% improvement over IPoIB (FDR)
- For 640GB Sort in 128 nodes
  - 48% improvement over IPoIB (FDR)

M. W. Rahman, X. Lu, N. S. Islam, R. Rajachandrasekar, and D. K. Panda, MapReduce over Lustre: Can RDMA-based Approach Benefit?, Euro-Par, August 2014.

Design Overview of Spark with RDMA

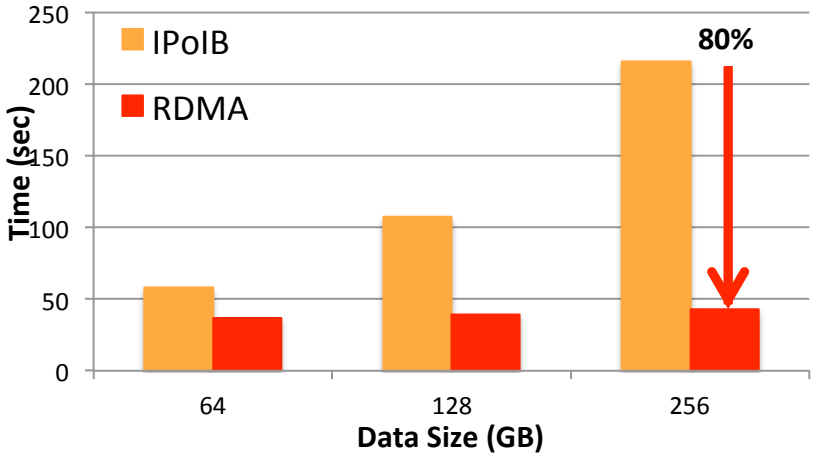


- Design Features
  - RDMA based shuffle
  - SEDA-based plugins
  - Dynamic connection management and sharing
  - Non-blocking and out-of-order data transfer
  - Off-JVM-heap buffer management
  - InfiniBand/RoCE support

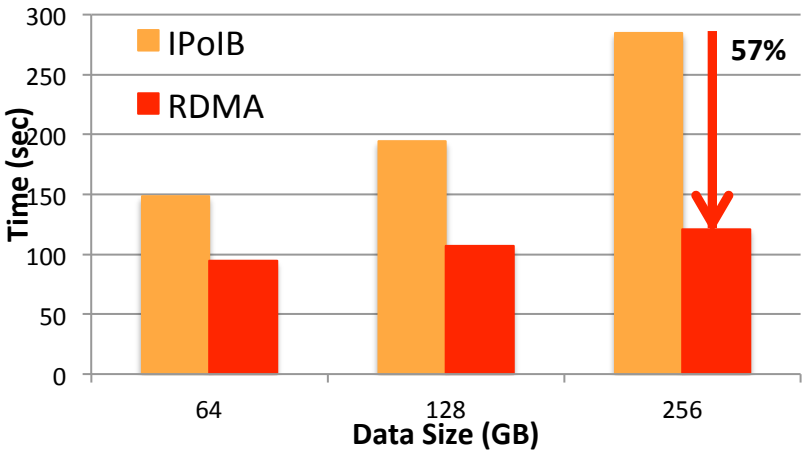
- Enables high performance RDMA communication, while supporting traditional socket interface
- JNI Layer bridges Scala based Spark with communication library written in native code

X. Lu, M. W. Rahman, N. Islam, D. Shankar, and D. K. Panda, Accelerating Spark with RDMA for Big Data Processing: Early Experiences, Int'l Symposium on High Performance Interconnects (HotI'14), August 2014

Performance Evaluation on SDSC Comet – SortBy/GroupBy



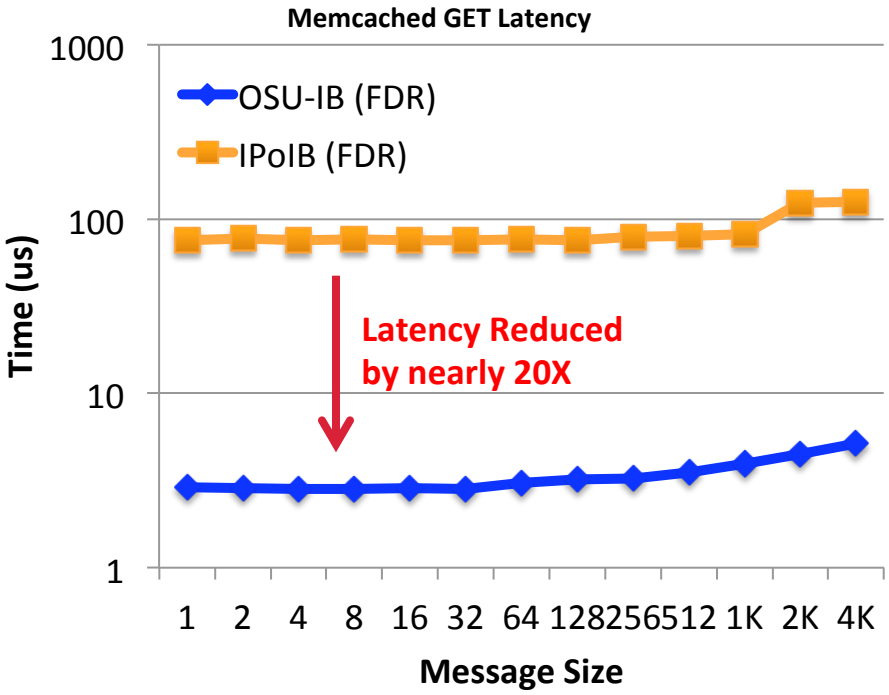
64 Worker Nodes, 1536 cores, SortByTest Total Time

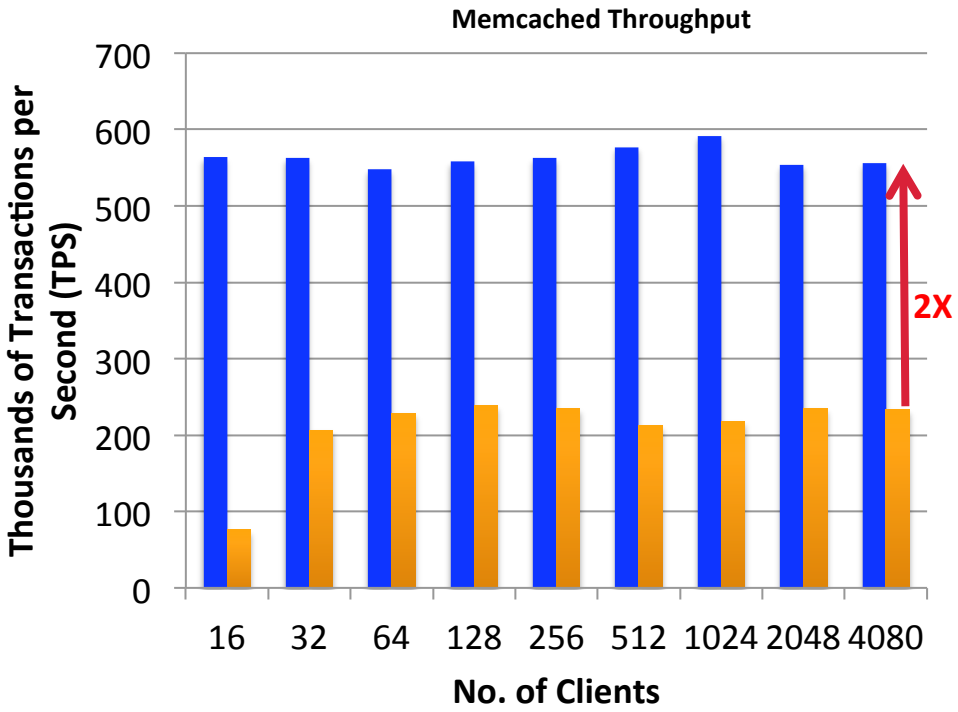


64 Worker Nodes, 1536 cores, GroupByTest Total Time

- InfiniBand FDR, SSD, 64 Worker Nodes, 1536 Cores, (1536M 1536R)
- RDMA-based design for Spark 1.5.1
- RDMA vs. IPoIB with 1536 concurrent tasks, single SSD per node.
  - SortBy: Total time reduced by up to 80% over IPoIB (56Gbps)
  - GroupBy: Total time reduced by up to 57% over IPoIB (56Gbps)

Memcached-RDMA Performance (FDR Interconnect)





Experiments on TACC Stampede (Intel SandyBridge Cluster, IB: FDR)

- Memcached Get latency
  - 4 bytes OSU-IB: 2.84 us; IPoIB: 75.53 us
  - 2K bytes OSU-IB: 4.49 us; IPoIB: 123.42 us
- Memcached Throughput (4bytes)
  - 4080 clients OSU-IB: 556 Kops/sec, IPoIB: 233 Kops/s
  - Nearly 2X improvement in throughput

Q3: Future Mutual Influence between HPC and Big Data Techniques?

- Upcoming Deep Learning Domain with Big Data
  - Need HPC in a big manner for training with huge amount of data
  - Environments like Caffe, CNTK, etc. are already using MPI and GPUs in an extensive manner
  - Need to also process Big Data during the operational phase of the model
- Hybrid solutions with techniques from both HPC and Big Data are required to handle these applications
- Some such solutions are coming out … more will be coming out in future