HiBD
High-Performance Big Data

Accelerate Big Data Processing (Hadoop, Spark, Memcached, & TensorFlow) with HPC Technologies

Talk at Intel® HPC Developer Conference 2017 (SC '17)

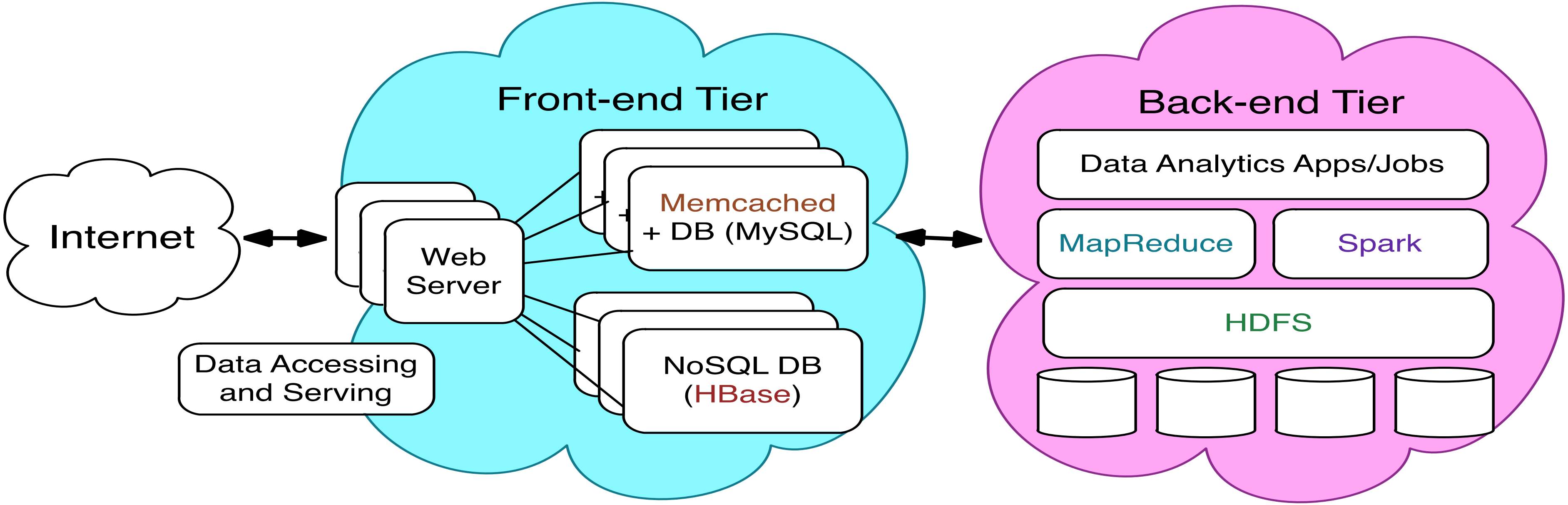
by

Dhabaleswar K. (DK) Panda
The Ohio State University
E-mail: panda@cse.ohio-state.edu
http://www.cse.ohio-state.edu/~panda

Xiaoyi Lu
The Ohio State University
E-mail: luxi@cse.ohio-state.edu
http://www.cse.ohio-state.edu/~luxi

Big Data Processing and Deep Learning on Modern Clusters

- Multiple tiers + Workflow
  - Front-end data accessing and serving (Online)
    - Memcached + DB (e.g. MySQL), HBase, etc.
  - Back-end data analytics and deep learning model training (Offline)
    - HDFS, MapReduce, Spark, TensorFlow, BigDL, Caffe, etc.



Drivers of Modern HPC Cluster Architectures

Multi-core Processors

High Performance Interconnects - InfiniBand
<1usec latency, 100Gbps Bandwidth>

Accelerators / Coprocessors
high compute density, high performance/watt
>1 TFlop DP on a chip

SSD, NVMe-SSD, NVRAM

- Multi-core/many-core technologies
- Remote Direct Memory Access (RDMA)-enabled networking (InfiniBand and RoCE)
- Solid State Drives (SSDs), Non-Volatile Random-Access Memory (NVRAM), NVMe-SSD
- Accelerators (NVIDIA GPGPUs and Intel Xeon Phi)

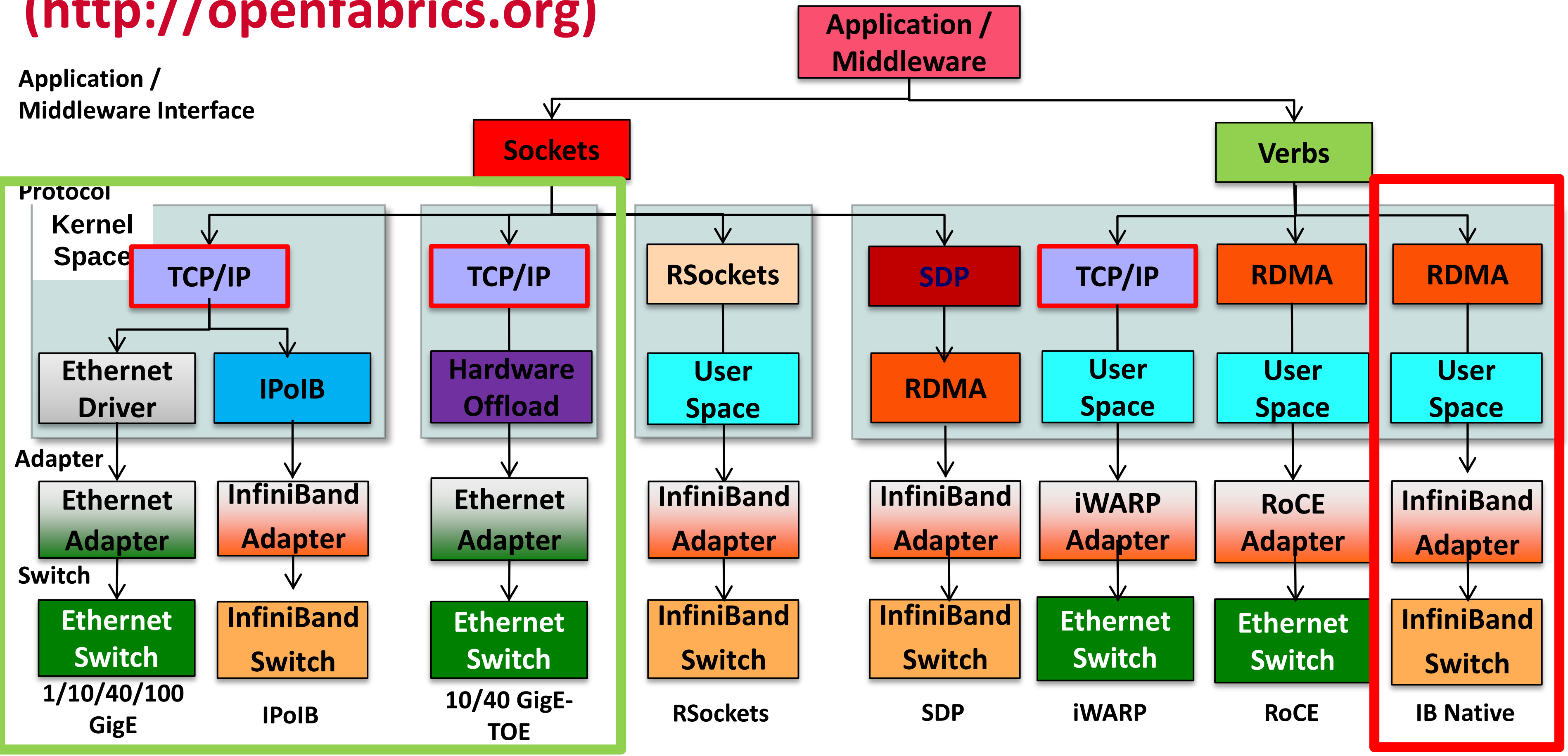
Tianhe – 2

Titan

Stampede

Tianhe – 1A

Interconnects and Protocols in OpenFabrics Stack for HPC (http://openfabrics.org)

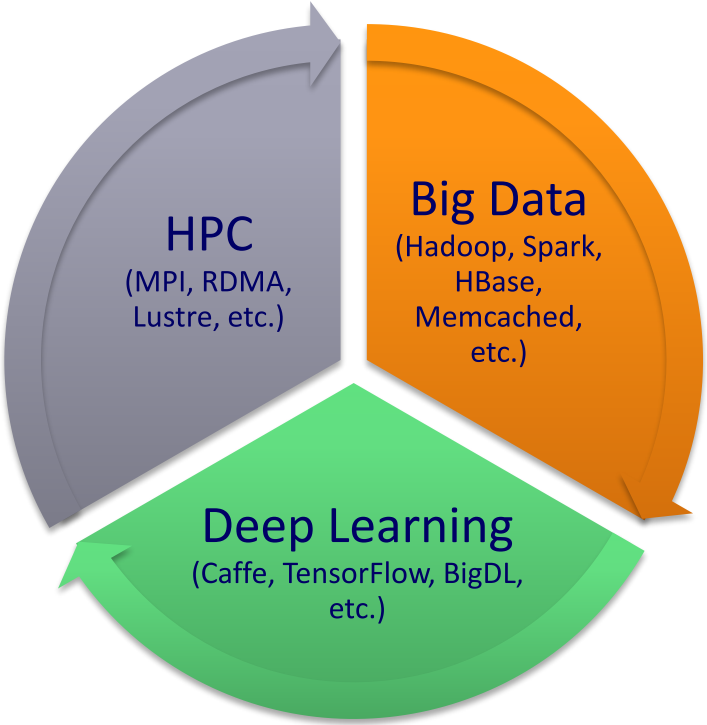


Large-scale InfiniBand Installations

- 177 IB Clusters (35%) in the Jun'17 Top500 list
  - (http://www.top500.org)
- Installations in the Top 50 (18 systems):

241,108 cores (Pleiades) at NASA/Ames (15th)	152,692 cores (Thunder) at AFRL/USA (36th)
220,800 cores (Pangea) in France (19th)	99,072 cores (Mistral) at DKRZ/Germany (38th)
522,080 cores (Stampede) at TACC (20th)	147,456 cores (SuperMUC) in Germany (40th)
144,900 cores (Cheyenne) at NCAR/USA (22nd)	86,016 cores (SuperMUC Phase 2) in Germany (41st)
72,800 cores Cray CS-Storm in US (27th)	74,520 cores (Tsubame 2.5) at Japan/GSIC (44th)
72,800 cores Cray CS-Storm in US (28th)	66,000 cores (HPC3) in Italy (47sth)
124,200 cores (Topaz) SGI ICE at ERDC DSRC in US (30th)	194,616 cores (Cascade) at PNNL (49th)
60,512 cores (DGX-1) at Facebook/USA (31st)	85,824 cores (Occigen2) at GENCI/CINES in France (50th)
60,512 cores (DGX SATURNV) at NVIDIA/USA (32nd)	73,902 cores (Centennial) at ARL/USA (52nd)
72,000 cores (HPC2) in Italy (33rd)	**and many more!**

Increasing Usage of HPC, Big Data and Deep Learning



Convergence of HPC, Big Data, and Deep Learning!!!

How Can HPC Clusters with High-Performance Interconnect and Storage Architectures Benefit Big Data and Deep Learning Applications?

Can the bottlenecks be alleviated with new designs by taking advantage of HPC technologies?

Can RDMA-enabled high-performance interconnects benefit Big Data processing and Deep Learning?

Can HPC Clusters with high-performance storage systems (e.g. SSD, parallel file systems) benefit Big Data and Deep Learning applications?

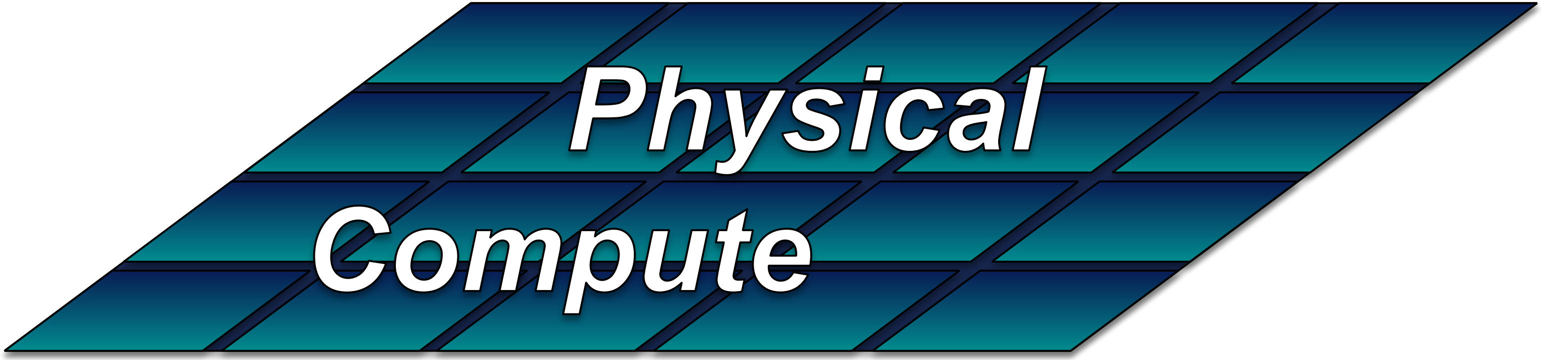
How much performance benefits can be achieved through enhanced designs?

What are the major bottlenecks in current Big Data processing and Deep Learning middleware (e.g. Hadoop, Spark)?

How to design benchmarks for evaluating the performance of Big Data and Deep Learning middleware on HPC clusters?

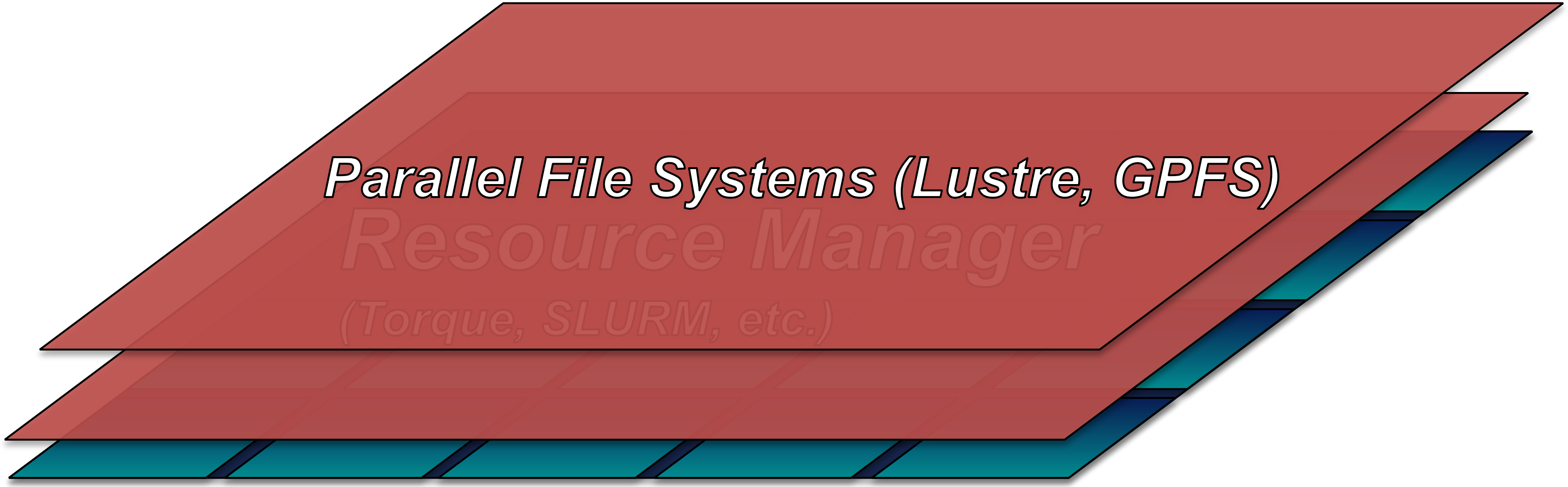
Bring HPC, Big Data processing, and Deep Learning into a “convergent trajectory”!

Can We Run Big Data and Deep Learning Jobs on Existing HPC Infrastructure?



Can We Run Big Data and Deep Learning Jobs on Existing HPC Infrastructure?

Can We Run Big Data and Deep Learning Jobs on Existing HPC Infrastructure?



Can We Run Big Data and Deep Learning Jobs on Existing HPC Infrastructure?

Designing Communication and I/O Libraries for Big Data Systems: Challenges

Applications

Benchmarks

Big Data Middleware
(HDFS, MapReduce, HBase, Spark and Memcached)

Upper level Changes?

Programming Models
(Sockets)

Other Protocols?

Communication and I/O Library

Point-to-Point Communication	Threaded Models and Synchronization	Virtualization
I/O and File Systems	QoS	Fault-Tolerance

Networking Technologies
(InfiniBand, 1/10/40/100 GigE and Intelligent NICs)

Commodity Computing System Architectures
(Multi- and Many-core architectures and accelerators)

Storage Technologies
(HDD, SSD, and NVMe-SSD)

The High-Performance Big Data (HiBD) Project

- 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 HBase
- RDMA for Memcached (RDMA-Memcached)
- RDMA for Apache Hadoop 1.x (RDMA-Hadoop)
- OSU HiBD-Benchmarks (OHB)
  - HDFS, Memcached, HBase, and Spark Micro-benchmarks
- http://hibd.cse.ohio-state.edu
- Users Base: 260 organizations from 31 countries
- More than 23,900 downloads from the project site

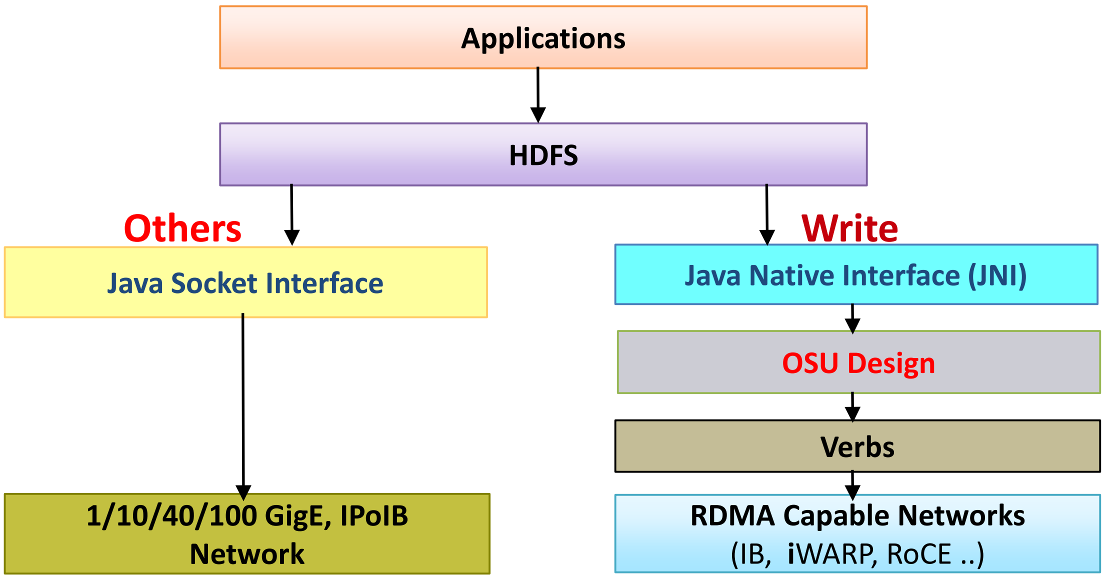
Available for InfiniBand and RoCE

Also run on Ethernet

Acceleration Case Studies and Performance Evaluation

- Basic Designs
  - Hadoop
  - Spark
  - Memcached
- Advanced Designs
  - Memcached with Hybrid Memory and Non-blocking APIs
  - Efficient Indexing with RDMA-HBase
  - TensorFlow with RDMA-gRPC
  - Deep Learning over Big Data
- BigData + HPC Cloud

Design Overview of HDFS with RDMA



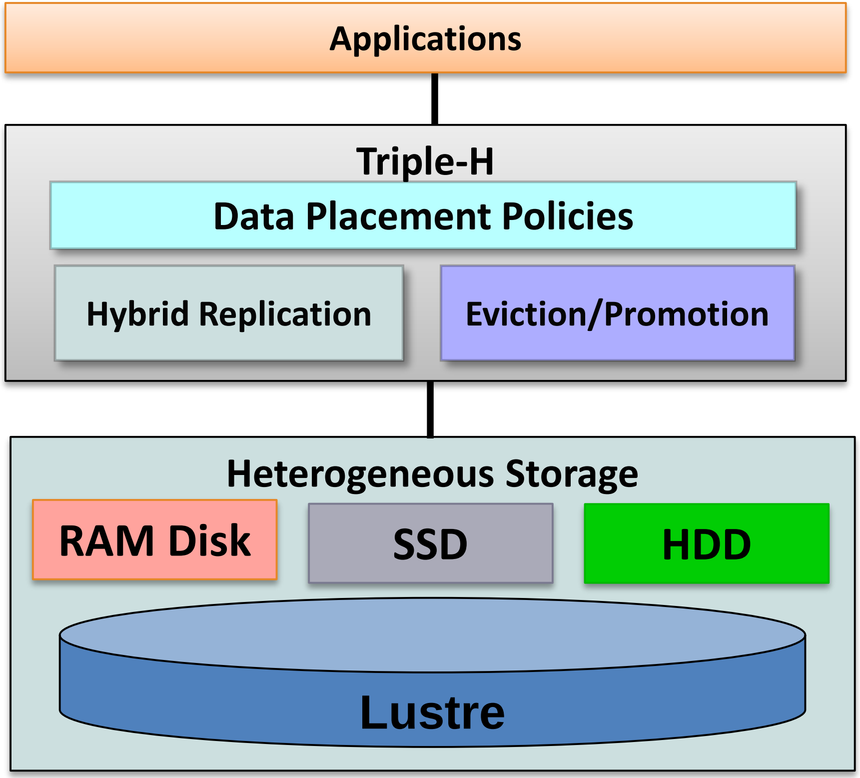
- Design Features
  - RDMA-based HDFS write
  - RDMA-based HDFS replication
  - Parallel replication support
  - On-demand connection setup
  - InfiniBand/RoCE support

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

N. S. Islam, M. W. Rahman, J. Jose, R. Rajachandrasekar, H. Wang, H. Subramoni, C. Murthy and D. K. Panda , High Performance RDMA-Based Design of HDFS over InfiniBand , Supercomputing (SC), Nov 2012

N. Islam, X. Lu, W. Rahman, and D. K. Panda, SOR-HDFS: A SEDA-based Approach to Maximize Overlapping in RDMA-Enhanced HDFS, HPDC '14, June 2014

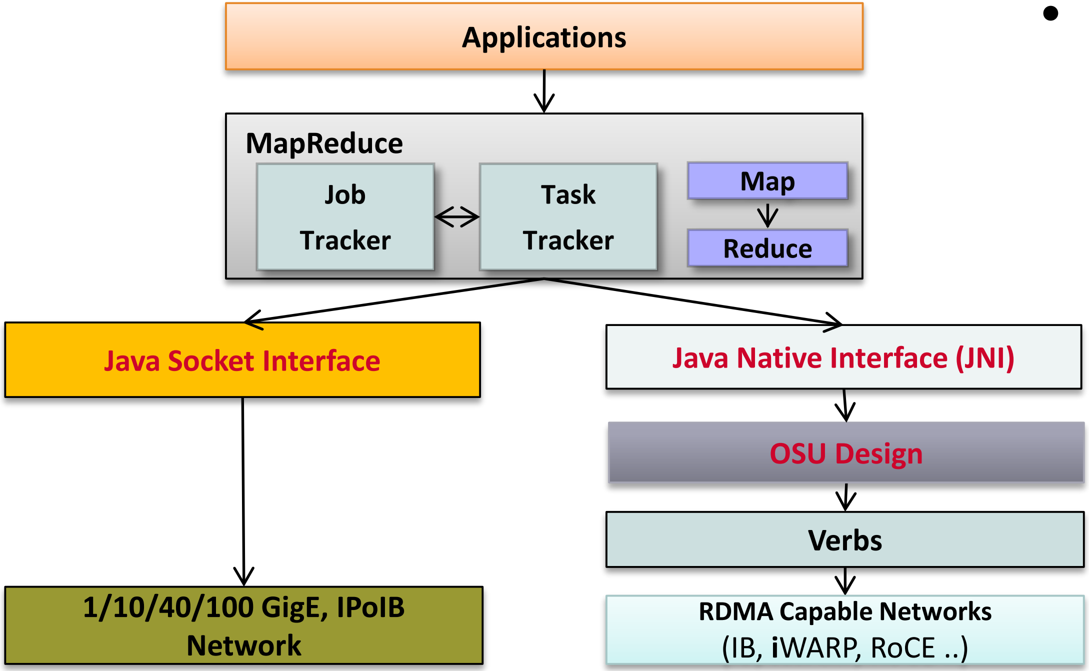
Enhanced HDFS with In-Memory and Heterogeneous Storage



- Design Features
  - Three modes
    - Default (HHH)
    - In-Memory (HHH-M)
    - Lustre-Integrated (HHH-L)
  - Policies to efficiently utilize the heterogeneous storage devices
    - RAM, SSD, HDD, Lustre
  - Eviction/Promotion based on data usage pattern
  - Hybrid Replication
  - Lustre-Integrated mode:
    - Lustre-based fault-tolerance

N. Islam, X. Lu, M. W. Rahman, D. Shankar, and D. K. Panda, Triple-H: A Hybrid Approach to Accelerate HDFS on HPC Clusters with Heterogeneous Storage Architecture, CCGrid '15, May 2015

Design Overview of MapReduce with RDMA

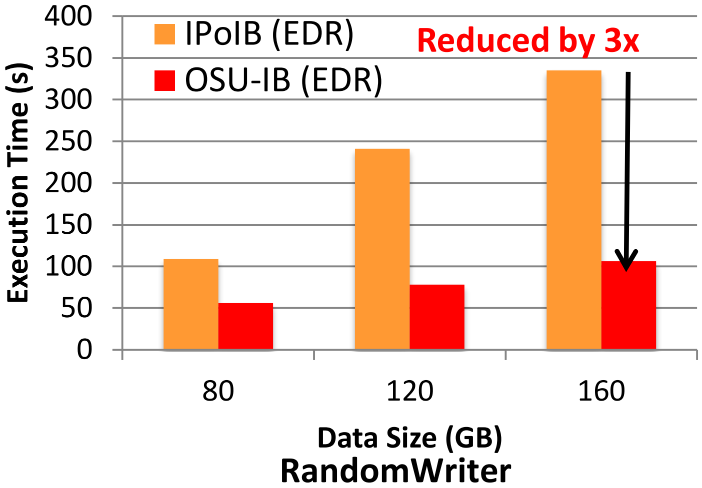


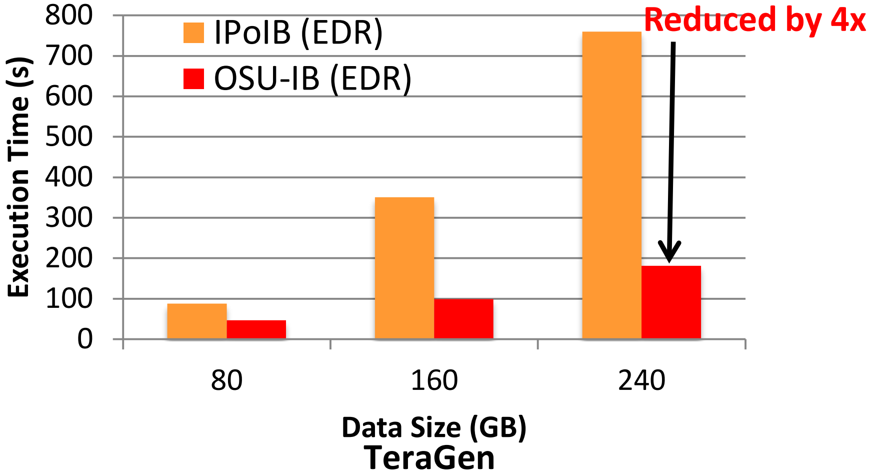
- Design Features
  - RDMA-based shuffle
  - Prefetching and caching map output
  - Efficient Shuffle Algorithms
  - In-memory merge
  - On-demand Shuffle Adjustment
  - Advanced overlapping
    - map, shuffle, and merge
    - shuffle, merge, and reduce
  - On-demand connection setup
  - InfiniBand/RoCE support

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

M. W. Rahman, X. Lu, N. S. Islam, and D. K. Panda, HOMR: A Hybrid Approach to Exploit Maximum Overlapping in MapReduce over High Performance Interconnects, ICS, June 2014

Performance Numbers of RDMA for Apache Hadoop 2.x – RandomWriter & TeraGen in OSU-RI2 (EDR)

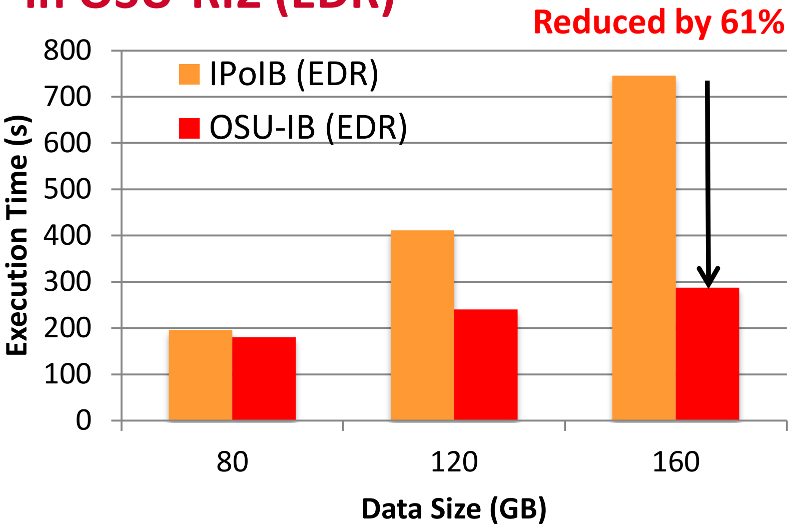




Cluster with 8 Nodes with a total of 64 maps

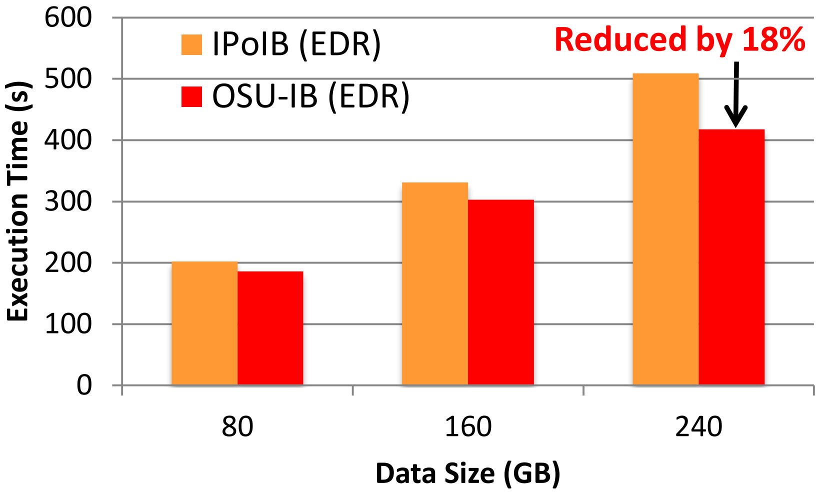
- RandomWriter
  - **3x** improvement over IPoIB for 80-160 GB file size
- TeraGen
  - **4x** improvement over IPoIB for 80-240 GB file size

Performance Numbers of RDMA for Apache Hadoop 2.x – Sort & TeraSort in OSU-RI2 (EDR)



Sort

Cluster with 8 Nodes with a total of 64 maps and 14 reduces

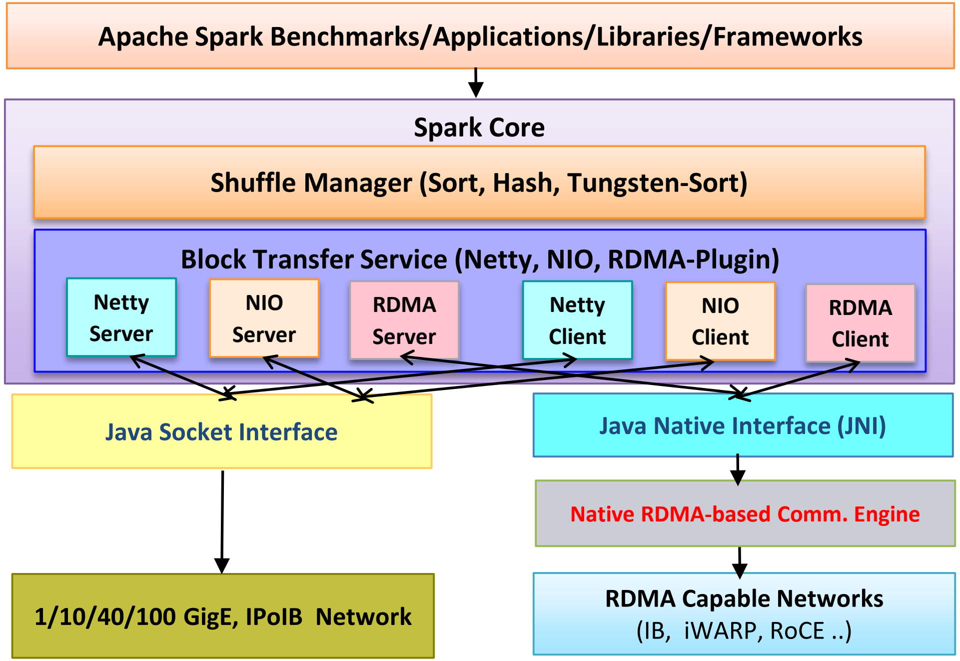


TeraSort

Cluster with 8 Nodes with a total of 64 maps and 32 reduces

- Sort
  - **61%** improvement over IPoIB for 80-160 GB data
- TeraSort
  - **18%** improvement over IPoIB for 80-240 GB data

Design Overview of Spark with RDMA



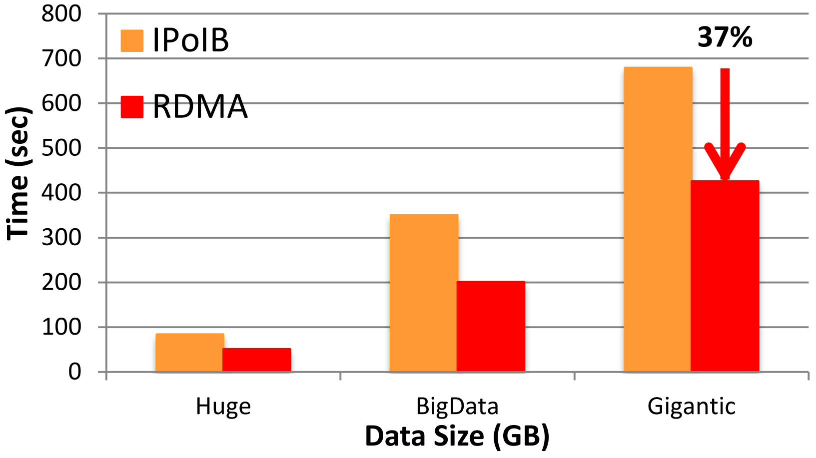
- Design Features
  - RDMA based shuffle plugin
  - SEDA-based architecture
  - Dynamic connection management and sharing
  - Non-blocking data transfer
  - Off-JVM-heap buffer management
  - InfiniBand/RoCE support

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

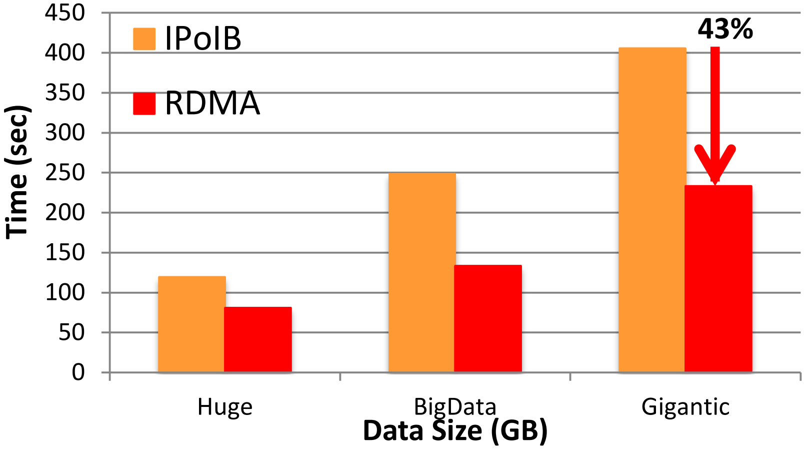
X. Lu, M. W. Rahman, N. Islam, D. Shankar, and D. K. Panda, Accelerating Spark with RDMA for Big Data Processing: Early Experiences, Int'l Symposium on High Performance Interconnects (HotI'14), August 2014

X. Lu, D. Shankar, S. Gugnani, and D. K. Panda, High-Performance Design of Apache Spark with RDMA and Its Benefits on Various Workloads, IEEE BigData '16, Dec. 2016.

Performance Evaluation on SDSC Comet – HiBench PageRank



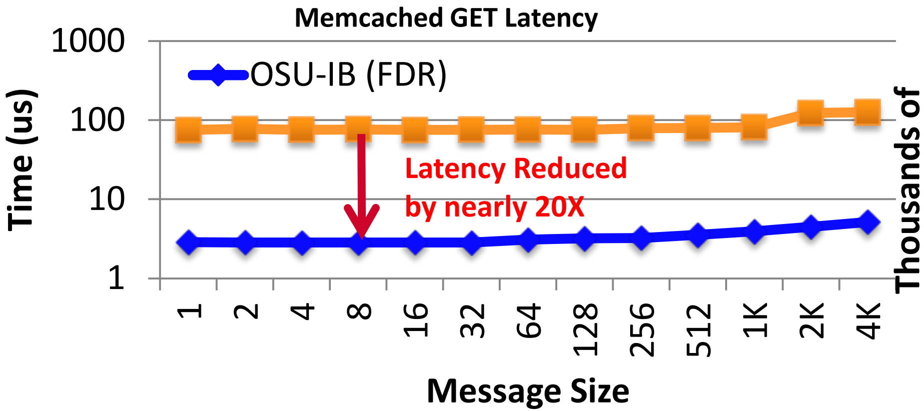
32 Worker Nodes, 768 cores, PageRank Total Time

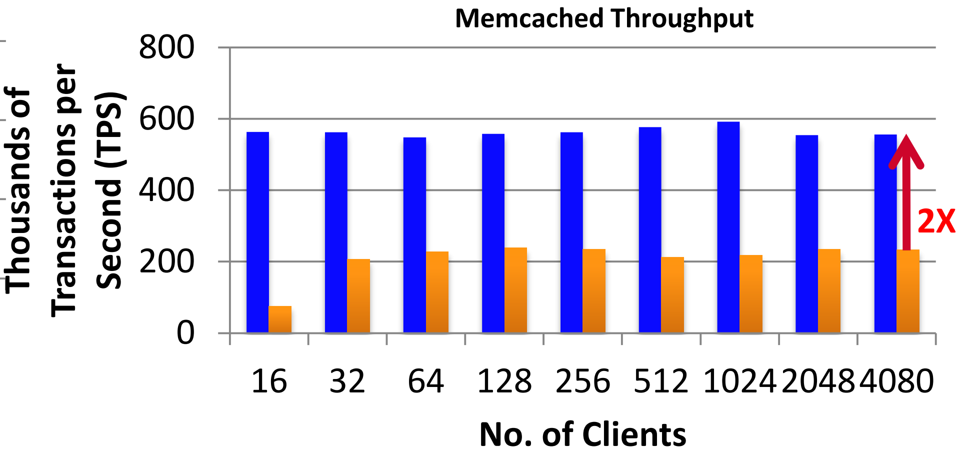


64 Worker Nodes, 1536 cores, PageRank Total Time

- InfiniBand FDR, SSD, 32/64 Worker Nodes, 768/1536 Cores, (768/1536M 768/1536R)
- RDMA-based design for Spark 1.5.1
- RDMA vs. IPoIB with 768/1536 concurrent tasks, single SSD per node.
  - 32 nodes/768 cores: Total time reduced by 37% over IPoIB (56Gbps)
  - 64 nodes/1536 cores: Total time reduced by 43% over IPoIB (56Gbps)

Memcached 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

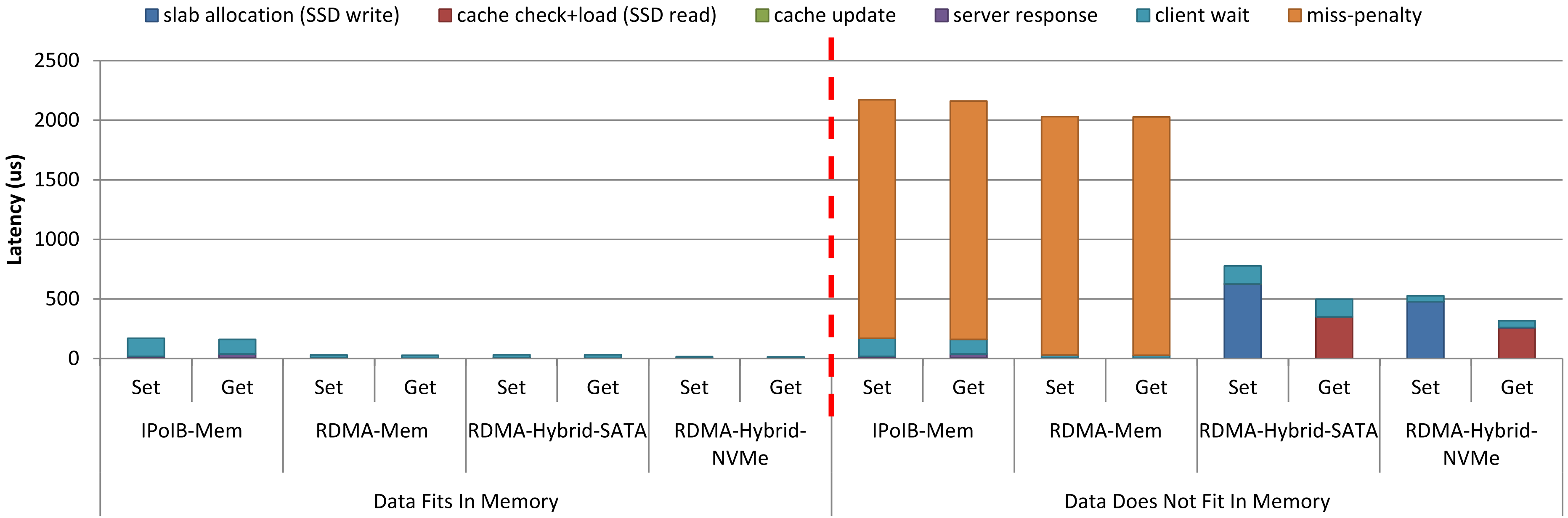
J. Jose, H. Subramoni, M. Luo, M. Zhang, J. Huang, M. W. Rahman, N. Islam, X. Ouyang, H. Wang, S. Sur and D. K. Panda, Memcached Design on High Performance RDMA Capable Interconnects, ICPP'11

J. Jose, H. Subramoni, K. Kandalla, M. W. Rahman, H. Wang, S. Narravula, and D. K. Panda, Scalable Memcached design for InfiniBand Clusters using Hybrid Transport, CCGrid'12

Acceleration Case Studies and Performance Evaluation

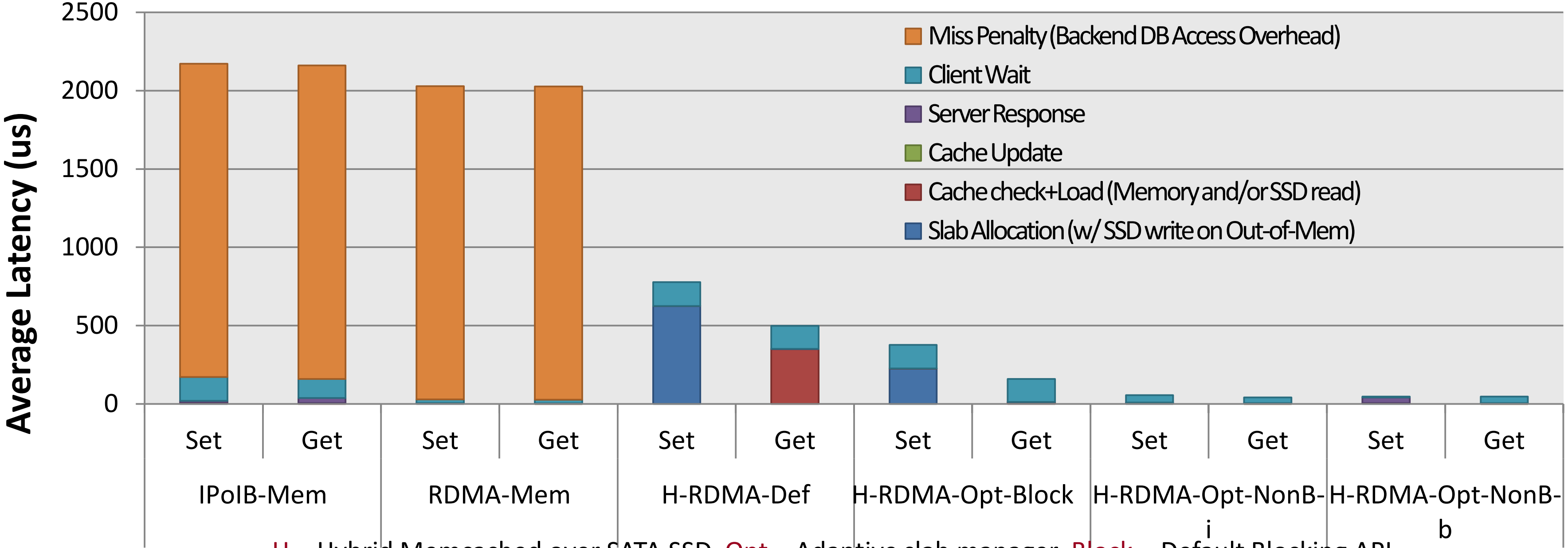
- Basic Designs
  - Hadoop
  - Spark
  - Memcached
- Advanced Designs
  - Memcached with Hybrid Memory and Non-blocking APIs
  - Efficient Indexing with RDMA-HBase
  - TensorFlow with RDMA-gRPC
  - Deep Learning over Big Data
- BigData + HPC Cloud

Performance Evaluation on IB FDR + SATA/NVMe SSDs (Hybrid Memory)



- Memcached latency test with Zipf distribution, server with 1 GB memory, 32 KB key-value pair size, total size of data accessed is 1 GB (when data fits in memory) and 1.5 GB (when data does not fit in memory)
- When data fits in memory: RDMA-Mem/Hybrid gives 5x improvement over IPoIB-Mem
- When data does not fit in memory: RDMA-Hybrid gives 2x-2.5x over IPoIB/RDMA-Mem

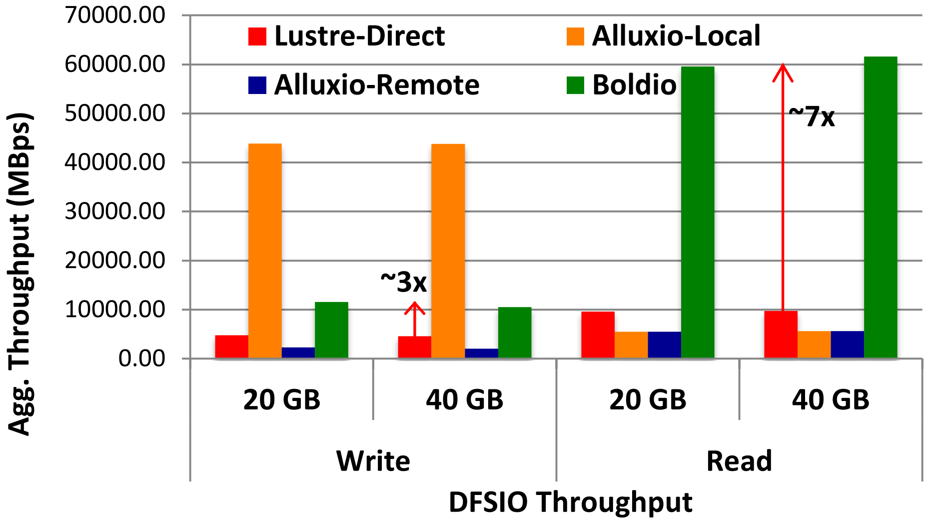
Performance Evaluation with Non-Blocking Memcached API

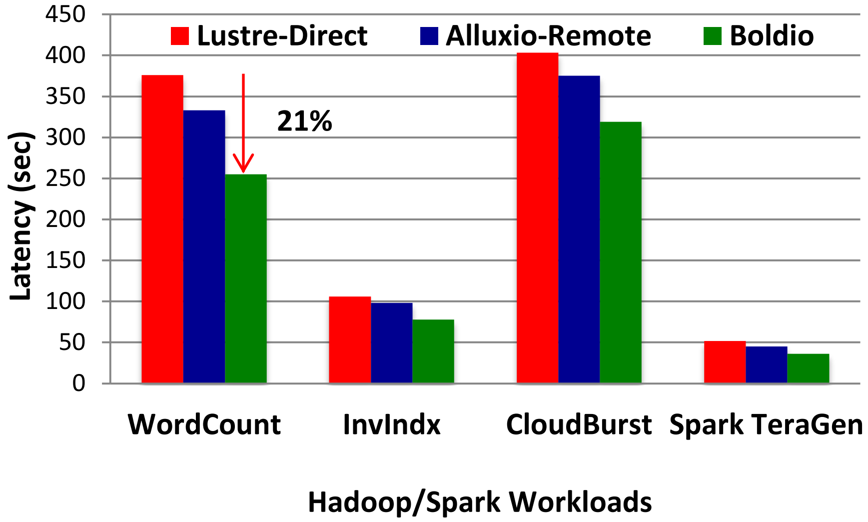


H = Hybrid Memcached over SATA SSD Opt = Adaptive slab manager Block = Default Blocking API
NonB-i = Non-blocking iset/iget API NonB-b = Non-blocking bset/bget API w/ buffer re-use guarantee

- Data does not fit in memory: Non-blocking Memcached Set/Get API Extensions can achieve
  - >16x latency improvement vs. blocking API over RDMA-Hybrid/RDMA-Mem w/ penalty
  - >2.5x throughput improvement vs. blocking API over default/optimized RDMA-Hybrid
- Data fits in memory: Non-blocking Extensions perform similar to RDMA-Mem/RDMA-Hybrid and >3.6x improvement over IPoIB-Mem

Performance Evaluation with Boldio for Lustre + Burst-Buffer





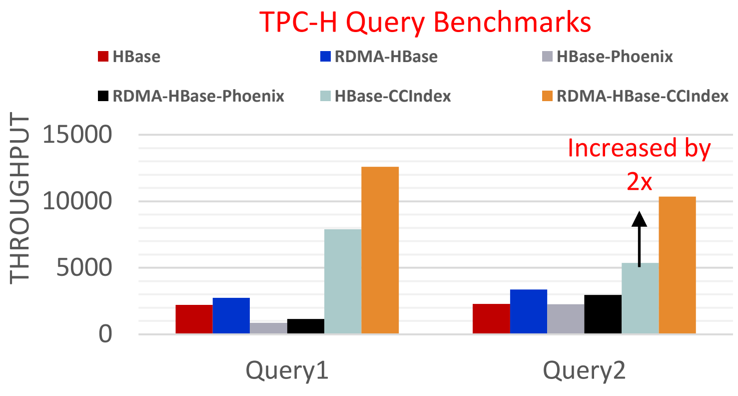
- InfiniBand QDR, 24GB RAM + PCIe-SSDs, 12 nodes, 32/48 Map/Reduce Tasks, 4-node Memcached cluster
- Boldio can improve
  - throughput over Lustre by about **3x** for write throughput and **7x** for read throughput
  - execution time of Hadoop benchmarks over Lustre, e.g. Wordcount, Cloudburst by **>21%**
- Contrasting with Alluxio (formerly Tachyon)
  - Performance degrades about 15x when Alluxio cannot leverage local storage (Alluxio-Local vs. Alluxio-Remote)
  - Boldio can improve throughput over Alluxio with all remote workers by about 3.5x - 8 .8x (Alluxio-Remote vs. Boldio)

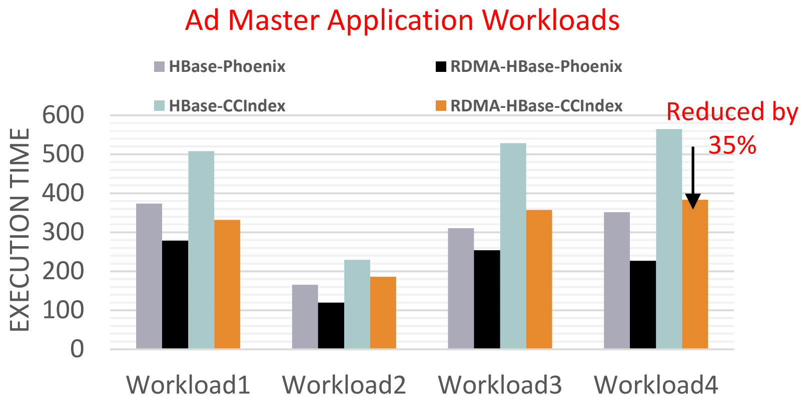
D. Shankar, X. Lu, D. K. Panda, Boldio: A Hybrid and Resilient Burst-Buffer over Lustre for Accelerating Big Data I/O, IEEE Big Data 2016.

Accelerating Indexing Techniques on HBase with RDMA

- Challenges
  - Operations on Distributed Ordered Table (DOT) with indexing techniques are network intensive
  - Additional overhead of creating and maintaining secondary indices
  - Can RDMA benefit indexing techniques (Apache Phoenix and CCIndex) on HBase?
- Results
  - Evaluation with Apache Phoenix and CCIndex
  - Up to 2x improvement in query throughput
  - Up to 35% reduction in application workload execution time

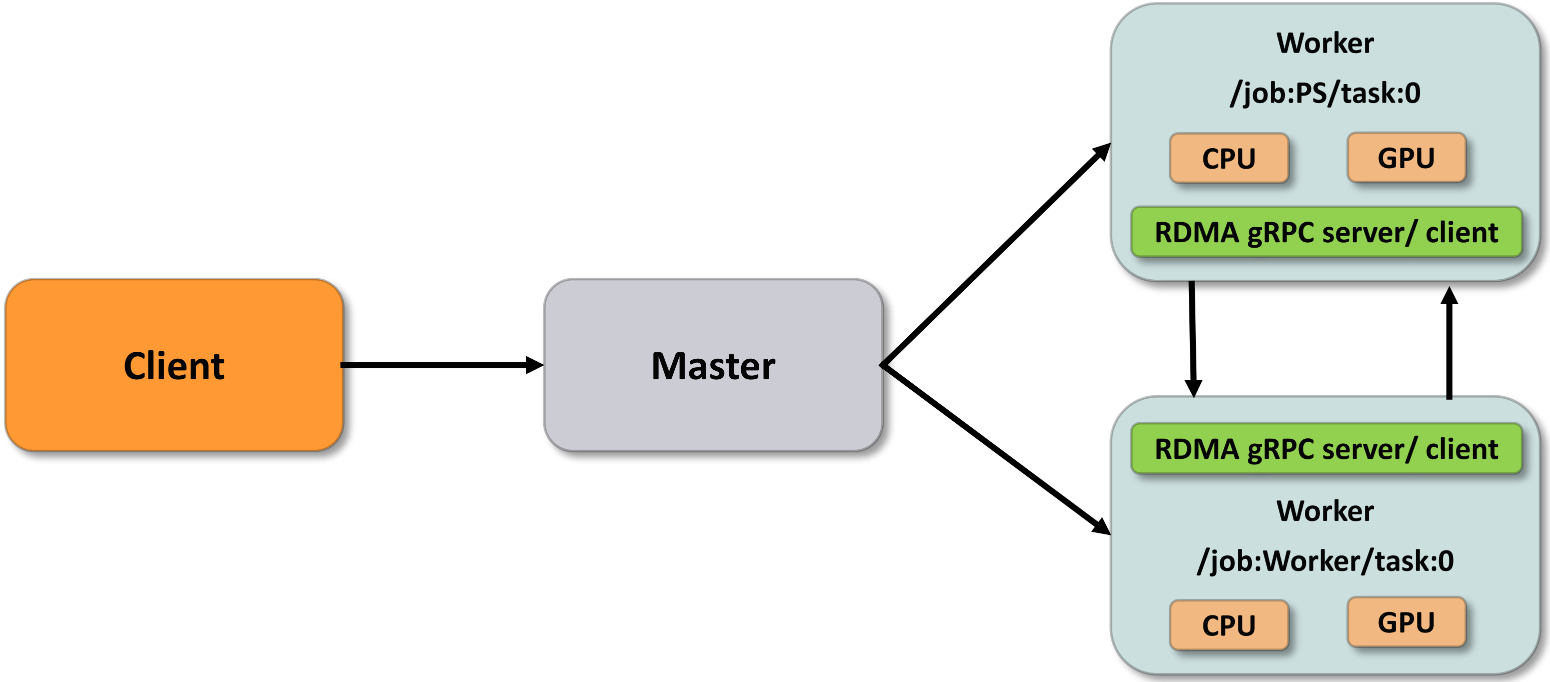
Collaboration with Institute of Computing Technology, Chinese Academy of Sciences





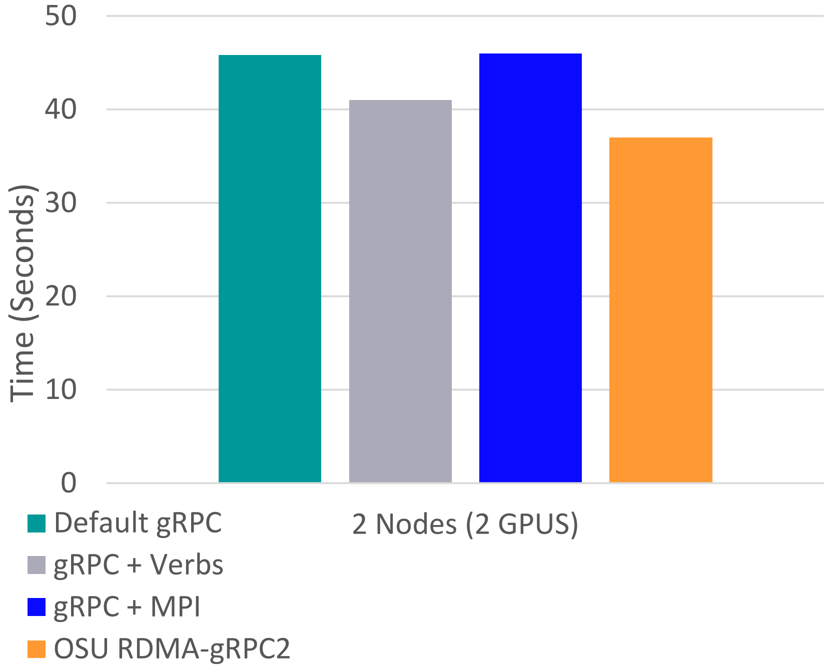
S. Gugnani, X. Lu, L. Zha, and D. K. Panda, Characterizing and Accelerating Indexing Techniques on Distributed Ordered Tables, IEEE BigData, 2017.

Overview of RDMA-gRPC with TensorFlow

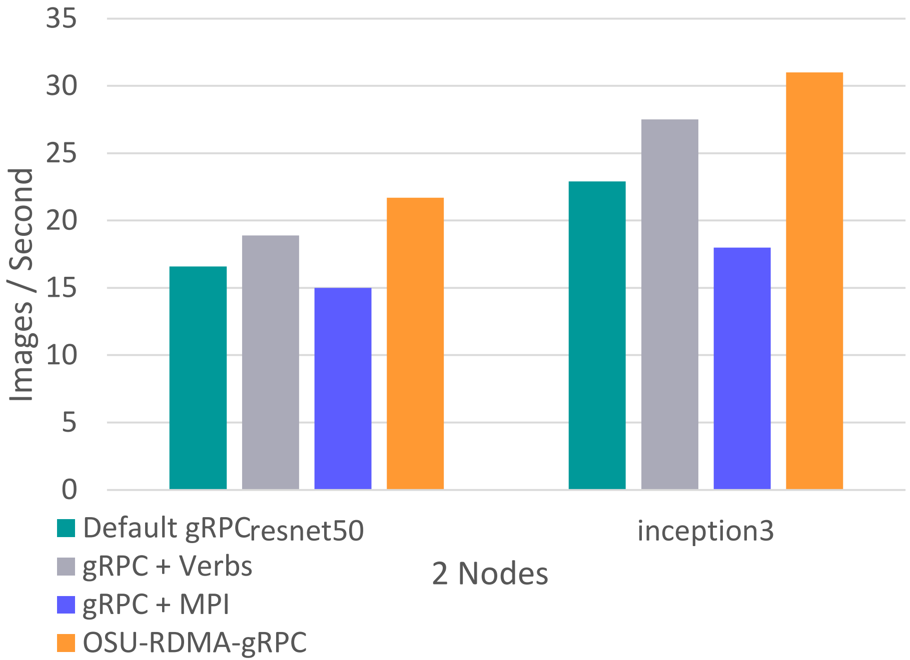


<u>Worker services communicate among each other using RDMA-gRPC</u>

Performance Benefit for TensorFlow



sigmoid net



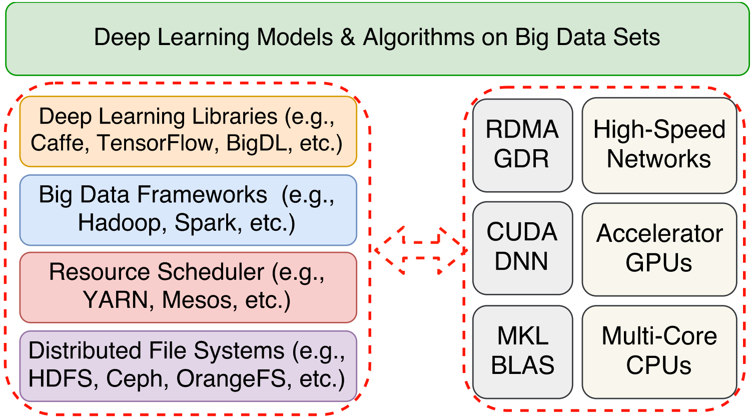
CNN

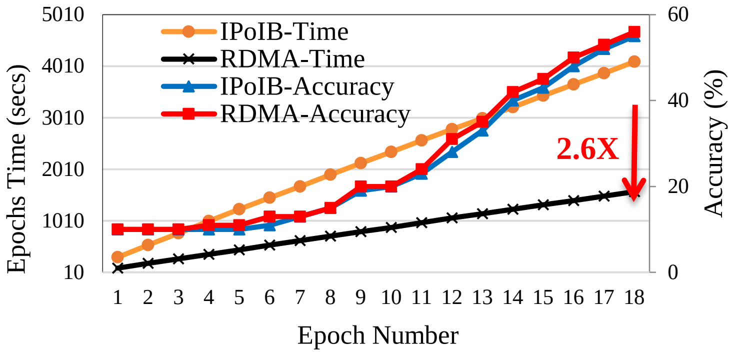
- TensorFlow performance evaluation on RI2
  - Up to 19% performance speedup over IPoIB for Sigmoid net (20 epochs).
  - Up to 35% and 30% performance speedup over IPoIB for resnet50 and Inception3 (batch size 8).

R. Biswas, X. Lu, and D. K. Panda, Accelerating gRPC and TensorFlow with RDMA for High-Performance Deep Learning over InfiniBand, Under Review.

High-Performance Deep Learning over Big Data (DLoBD) Stacks

- Challenges of Deep Learning over Big Data (DLoBD)
  - Can RDMA-based designs in DLoBD stacks improve performance, scalability, and resource utilization on high-performance interconnects, GPUs, and multi-core CPUs?
  - What are the performance characteristics of representative DLoBD stacks on RDMA networks?
- Characterization on DLoBD Stacks
  - CaffeOnSpark, TensorFlowOnSpark, and BigDL
  - IPoIB vs. RDMA; In-band communication vs. Out-of-band communication; CPU vs. GPU; etc.
  - Performance, accuracy, scalability, and resource utilization
  - RDMA-based DLoBD stacks (e.g., BigDL over RDMA-Spark) can achieve 2.6x speedup compared to the IPoIB based scheme, while maintain similar accuracy





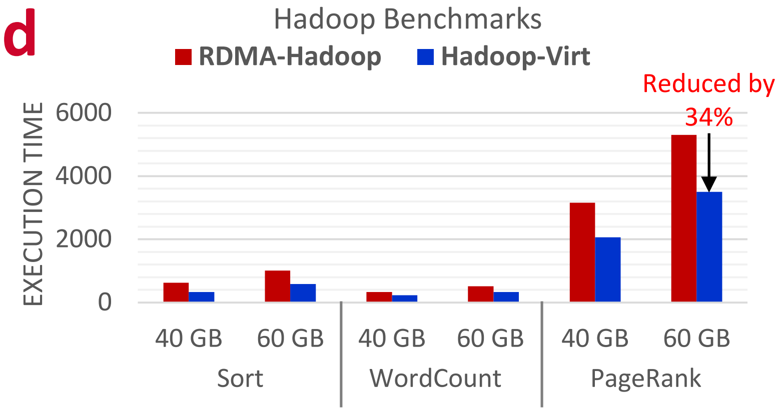
X. Lu, H. Shi, M. H. Javed, R. Biswas, and D. K. Panda, Characterizing Deep Learning over Big Data (DLoBD) Stacks on RDMA-capable Networks, HotI 2017.

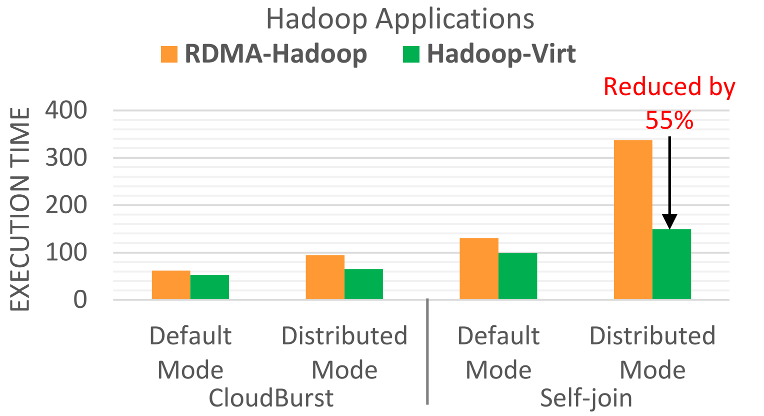
Acceleration Case Studies and Performance Evaluation

- Basic Designs
  - Hadoop
  - Spark
  - Memcached
- Advanced Designs
  - Memcached with Hybrid Memory and Non-blocking APIs
  - Efficient Indexing with RDMA-HBase
  - TensorFlow with RDMA-gRPC
  - Deep Learning over Big Data
- BigData + HPC Cloud

Virtualization-aware and Automatic Topology Detection Schemes in Hadoop on InfiniBand

- Challenges
  - Existing designs in Hadoop not virtualization-aware
  - No support for automatic topology detection
- Design
  - Automatic Topology Detection using MapReduce-based utility
    - Requires no user input
    - Can detect topology changes during runtime without affecting running jobs
  - Virtualization and topology-aware communication through map task scheduling and YARN container allocation policy extensions





S. Gugnani, X. Lu, and D. K. Panda, Designing Virtualization-aware and Automatic Topology Detection Schemes for Accelerating Hadoop on SR-IOV-enabled Clouds, CloudCom'16, December 2016

Concluding Remarks

- Discussed challenges in accelerating Big Data middleware with HPC technologies
- Presented basic and advanced designs to take advantage of InfiniBand/RDMA for HDFS, MapReduce, RPC, HBase, Memcached, Spark, gRPC, and TensorFlow
- Results are promising
- Many other open issues need to be solved
- Will enable Big Data community to take advantage of modern HPC technologies to carry out their analytics in a fast and scalable manner
- Looking forward to collaboration with the community

OSU Participating at Multiple Events on BigData Acceleration

- Tutorial
  - Big Data Meets HPC: Exploiting HPC Technologies for Accelerating Big Data Processing and Management (Sunday, 1:30-5:00 pm, Room #201)
- BoF
  - BigData and Deep Learning (Tuesday, 5:15-6:45pm, Room #702)
  - SigHPC Big Data BoF (Wednesday, 12:15-1:15pm, Room #603)
  - Clouds for HPC, Big Data, and Deep Learning (Wednesday, 5:15-7:00pm, Room #701)
- Booth Talks
  - OSU Booth (Tuesday, 10:00-11:00am, Booth #1875)
  - Mellanox Theater (Wednesday, 3:00-3:30pm, Booth #653)
  - OSU Booth (Thursday, 1:00-2:00pm, Booth #1875)
- Student Poster Presentation
  - Accelerating Big Data processing in Cloud (Tuesday, 5:15-7:00pm, Four Seasons Ballroom)
- Details at http://hibd.cse.ohio-state.edu

Funding Acknowledgments

Funding Support by

Office of Science U.S. Department of Energy

NSF

Mellanox Technologies

Microsoft

QLogic

Cray The Supercomputer Company

intel

Linux Networx

NVIDIA

NetApp

Sun microsystems

Cisco Systems

Equipment Support by

Mellanox Technologies

intel

AMD

Sun microsystems

Appro

advanced clustering technologies, inc.

Microway Technology you can count on

NVIDIA

QLogic

Personnel Acknowledgments

Current Students

- A. Awan (Ph.D.)
- M. Bayatpour (Ph.D.)
- S. Chakraborthy (Ph.D.)
- C.-H. Chu (Ph.D.)
- S. Guganani (Ph.D.)
- J. Hashmi (Ph.D.)
- N. Islam (Ph.D.)
- M. Li (Ph.D.)
- M. Rahman (Ph.D.)
- D. Shankar (Ph.D.)
- A. Venkatesh (Ph.D.)
- J. Zhang (Ph.D.)

Current Research Scientists

- X. Lu
- H. Subramoni

Current Post-doc

- A. Ruhela

Current Research Specialist

- J. Smith
- M. Arnold

Past Students

- A. Augustine (M.S.)
- P. Balaji (Ph.D.)
- S. Bhagvat (M.S.)
- A. Bhat (M.S.)
- D. Buntinas (Ph.D.)
- L. Chai (Ph.D.)
- B. Chandrasekharan (M.S.)
- N. Dandapanthula (M.S.)
- V. Dhanraj (M.S.)
- T. Gangadharappa (M.S.)
- K. Gopalakrishnan (M.S.)
- W. Huang (Ph.D.)
- W. Jiang (M.S.)
- J. Jose (Ph.D.)
- S. Kini (M.S.)
- M. Koop (Ph.D.)
- K. Kulkarni (M.S.)
- R. Kumar (M.S.)
- S. Krishnamoorthy (M.S.)
- K. Kandalla (Ph.D.)
- P. Lai (M.S.)
- J. Liu (Ph.D.)
- M. Luo (Ph.D.)
- A. Mamidala (Ph.D.)
- G. Marsh (M.S.)
- V. Meshram (M.S.)
- A. Moody (M.S.)
- S. Naravula (Ph.D.)
- R. Noronha (Ph.D.)
- X. Ouyang (Ph.D.)
- S. Pai (M.S.)
- S. Potluri (Ph.D.)
- R. Rajachandrasekar (Ph.D.)
- G. Santhanaraman (Ph.D.)
- A. Singh (Ph.D.)
- J. Sridhar (M.S.)
- S. Sur (Ph.D.)
- H. Subramoni (Ph.D.)
- K. Vaidyanathan (Ph.D.)
- A. Vishnu (Ph.D.)
- J. Wu (Ph.D.)
- W. Yu (Ph.D.)

Past Research Scientist

- K. Hamidouche
- S. Sur

Past Programmers

- D. Bureddy
- J. Perkins

Past Post-Docs

- D. Banerjee
- X. Besseron
- H.-W. Jin
- J. Lin
- M. Luo
- E. Mancini
- S. Marcarelli
- J. Vienne
- H. Wang

Thank You!

{panda, luxi}@cse.ohio-state.edu

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

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

Network-Based Computing Laboratory

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

The High-Performance Big Data Project

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