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

Application / Middleware

Application / Middleware Interface

Sockets

TCP/IP

IPoIB

Hardware Offload

TCP/IP

RSockets

Verbs

TCP/IP

RDMA

RDMA

User Space

IPoIB

Switch

Ethernet Adapter

1/10/40/100 GigE

InfiniBand Adapter

Adapter

iWARP

RoCE

RDMA

User Space

Ethernet Adapter

TCP/IP

IPoIB

InfiniBand Adapter

10/40 GigE-TOE

Ethernet Adapter

InfiniBand Adapter

IPoIB

Ethernet Switch

InfiniBand Switch

RDMA

iWARP Adapter

RoCE Adapter

IB Native

User Space

RSockets

SDP

TCP/IP

User Space

User Space

User Space

User Space

InfiniBand Adapter

Ethernet Switch

InfiniBand Switch

InfiniBand Switch

Ethernet Switch

Network Based Computing Laboratory

Intel® HPC Developer Conference 2017
Large-scale InfiniBand Installations

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

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

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

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?

Can HPC clusters with high-performance interconnects benefit Big Data and Deep Learning?

How much performance benefits can be achieved through enhanced designs?

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?

Resource Manager
(Torque, SLURM, etc.)
Can We Run Big Data and Deep Learning Jobs on Existing HPC Infrastructure?

Parallel File Systems (Lustre, GPFS)
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

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

Benchmark

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)

Upper level Changes?
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
- Available for InfiniBand and RoCE
  Also run on Ethernet
- \url{http://hibd.cse.ohio-state.edu}
- Users Base: 260 organizations from 31 countries
- More than 23,900 downloads from the project site
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

---


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

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

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

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

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

Cluster with 8 Nodes with a total of 64 maps
Performance Numbers of RDMA for Apache Hadoop 2.x – Sort & TeraSort in OSU-RI2 (EDR)

- **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


Performance Evaluation on SDSC Comet – HiBench PageRank

• 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


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
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

- 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

Network Based Computing Laboratory Intel® HPC Developer Conference 2017
• 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)

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

Overview of RDMA-gRPC with TensorFlow

Worker services communicate among each other using RDMA-gRPC
Performance Benefit for TensorFlow

- TensorFlow performance evaluation on RI2
  - Up to 17% performance speedup over IPoIB for sigmoid net (20 epochs).
  - Up to 23% performance speedup over IPoIB for resnet50 (batch size 4).
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

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

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**

[Logos of various sponsors]

**Equipment Support by**

[Logos of various equipment providers]
Personnel Acknowledgments

Current Students
- A. Awan (Ph.D.)
- M. Bayatpour (Ph.D.)
- S. Chakraborty (Ph.D.)
- C.-H. Chu (Ph.D.)
- S. Gusanani (Ph.D.)
- M. Rahman (Ph.D.)

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. Rajachandrsekar (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.)

Current Research Scientists
- X. Lu
- H. Subramoni

Current Research Specialist
- J. Smith
- M. Arnold

Past Research Scientist
- K. Hamidouche
- S. Sur

Past Programmers
- D. Bureddy
- J. Perkins

Current Post-doc
- A. Ruhela

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

Past Post-Docs
- J. Lin
- M. Luo
- E. Mancini
- S. Marcarelli
- J. Vienne
- H. Wang

Network Based Computing Laboratory
Intel® HPC Developer Conference 2017
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/