HiBD
High-Performance
Big Data

High-Performance Hadoop and Spark on OpenPOWER Platform

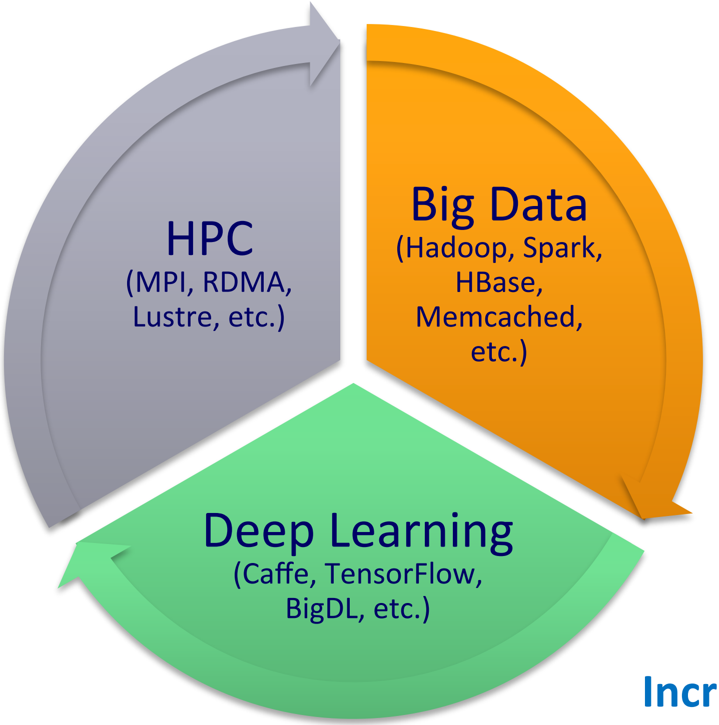
Talk at OpenPOWER Summit '18

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

Increasing Usage of HPC, Big Data and Deep Learning



HPC
(MPI, RDMA, Lustre, etc.)

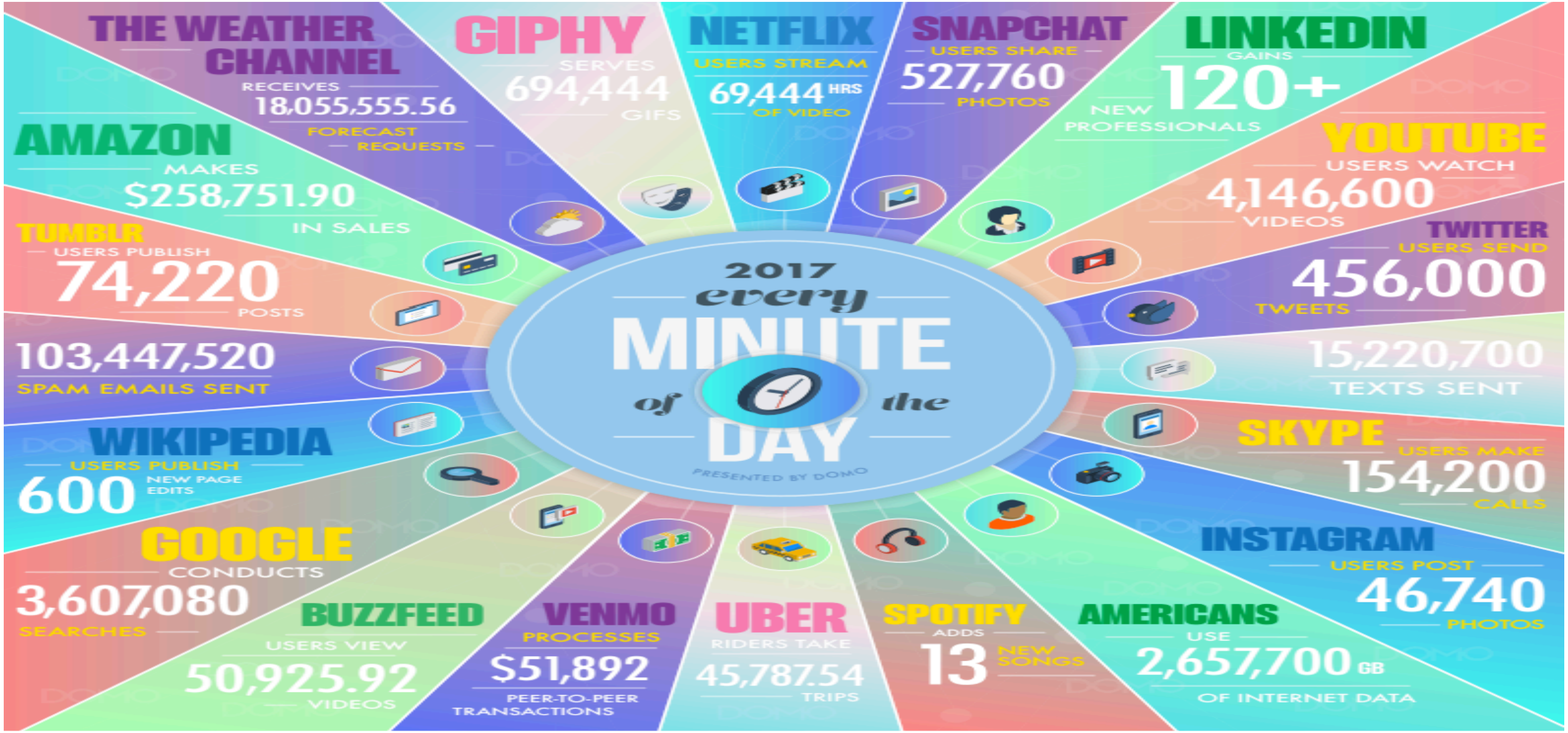
Big Data
(Hadoop, Spark, HBase, Memcached, etc.)

Deep Learning
(Caffe, TensorFlow, BigDL, etc.)

Convergence of HPC, Big Data, and Deep Learning!

Increasing Need to Run these applications on the Cloud!!

Big Velocity – How Much Data Is Generated Every Minute on the Internet?



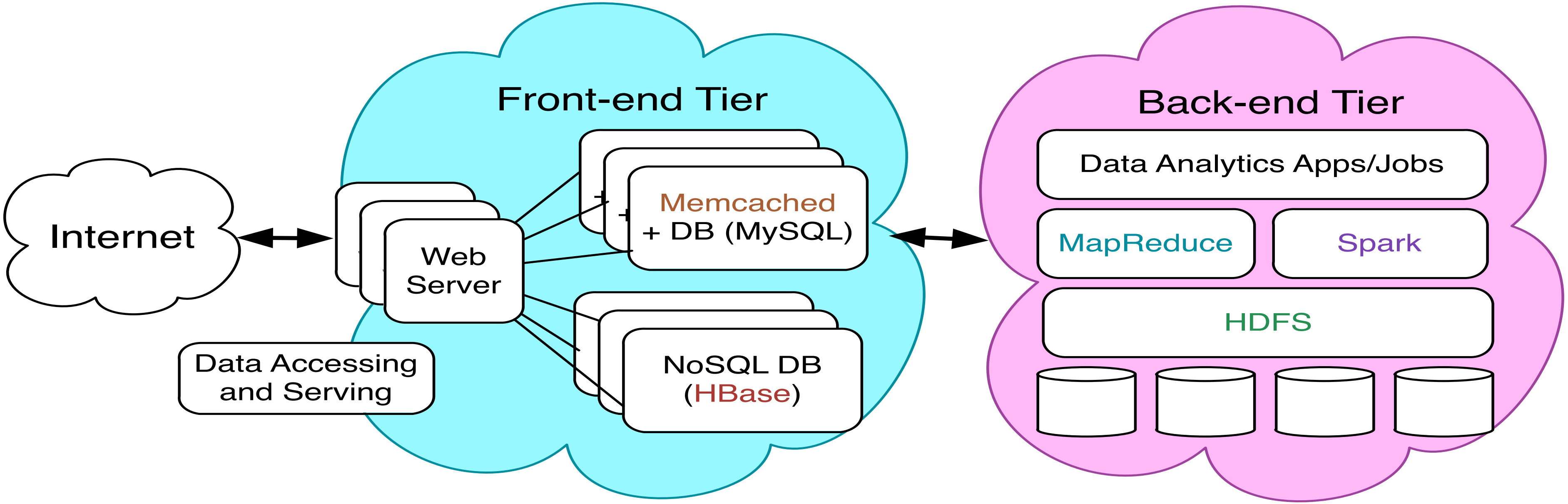
The global Internet population grew 7.5% from 2016 and now represents

3.7 Billion People.

Courtesy: https://www.domo.com/blog/data-never-sleeps-5/

Data Management and Processing on Modern Clusters

- Substantial impact on designing and utilizing data management and processing systems in multiple tiers
  - Front-end data accessing and serving (Online)
    - Memcached + DB (e.g. MySQL), HBase
  - Back-end data analytics (Offline)
    - HDFS, MapReduce, Spark

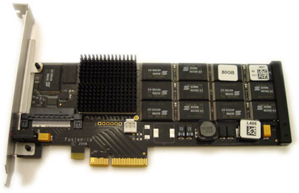


Drivers of Modern HPC Cluster and Data Center Architecture

Multi-/Many-core Processors

High Performance Interconnects – InfiniBand (with SR-IOV)
<1usec latency, 200Gbps 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)
  - Single Root I/O Virtualization (SR-IOV)
- Solid State Drives (SSDs), NVM, Parallel Filesystems, Object Storage Clusters
- Accelerators (NVIDIA GPGPUs and Intel Xeon Phi)

SDSC Comet

TACC Stampede

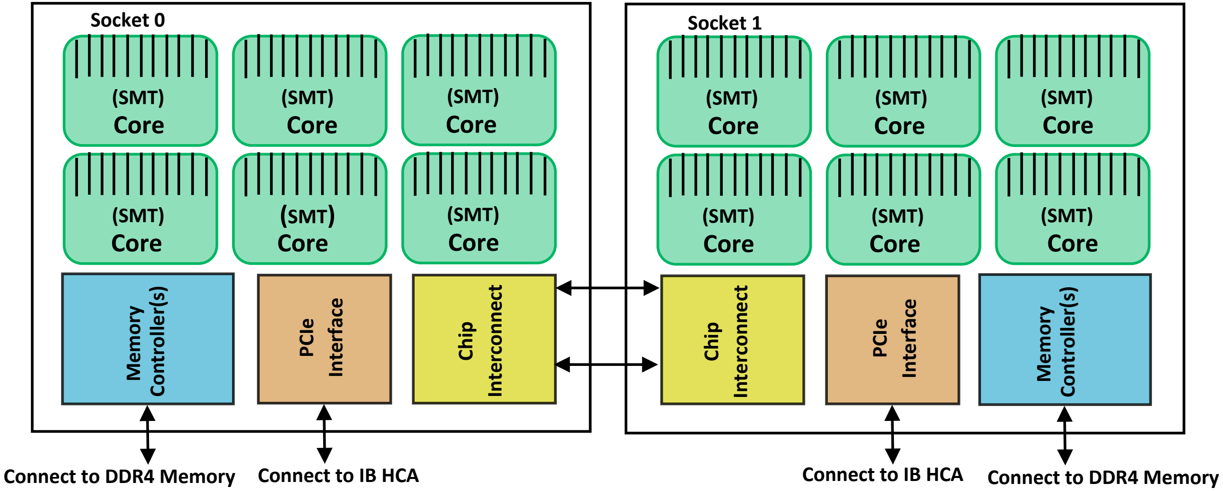
Microsoft Azure

amazon web services EC2

ORACLE Cloud

Chameleon Cloud

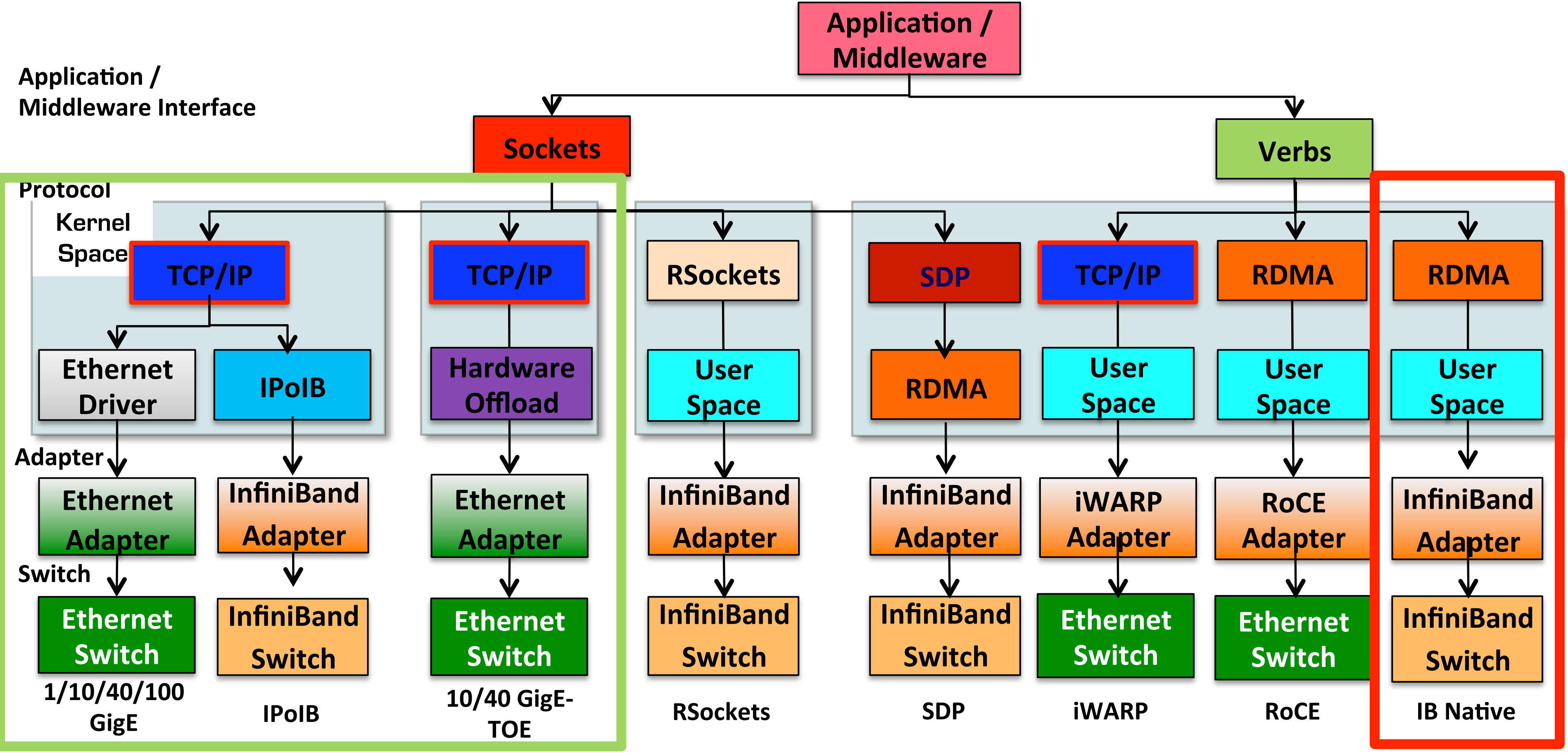
Overview of POWER Architecture



IBM POWER8

- POWER Architecture
- Dual socket (NUMA)
- Each socket with 10-12 cores (up to 5 GHz)
- 8 SMTs/core and >= 160 SMTs

Interconnects and Protocols in OpenFabrics Stack



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?

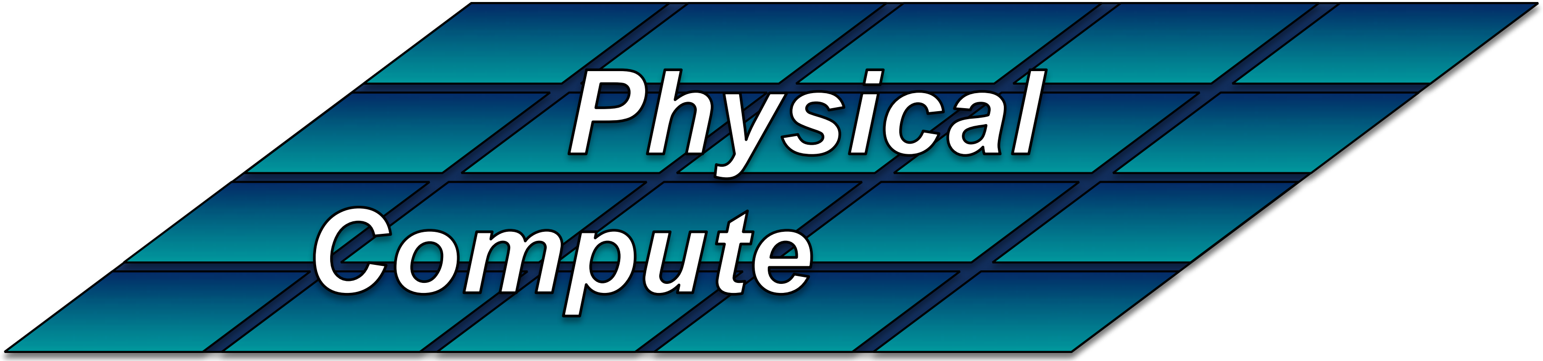
How much performance benefits can be achieved through enhanced designs?

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”!

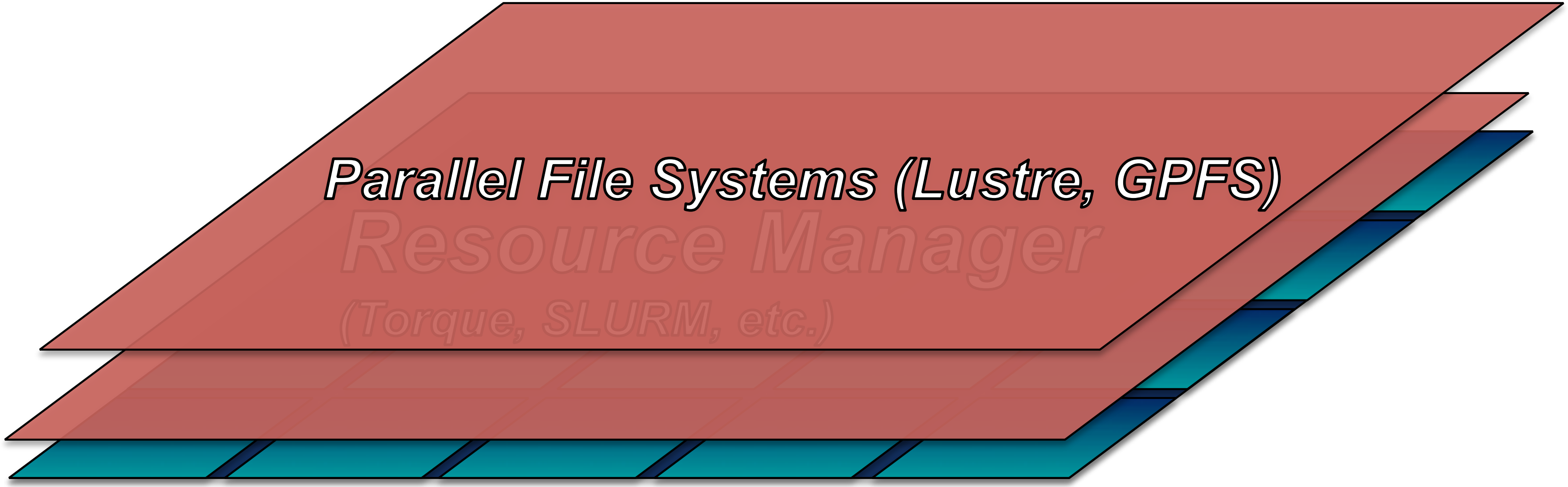
Can We Run Big Data and Deep Learning Jobs on Existing HPC Infrastructure?



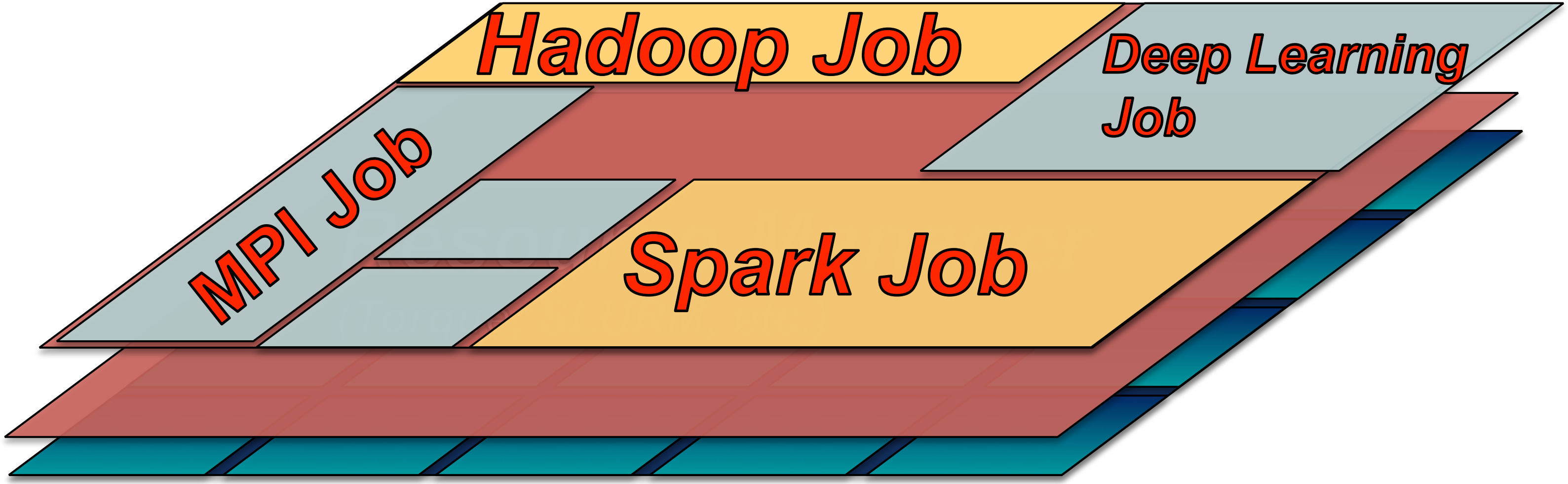
Can We Run Big Data and Deep Learning Jobs on Existing HPC Infrastructure?

Resource Manager
(Torque, SLURM, etc.)

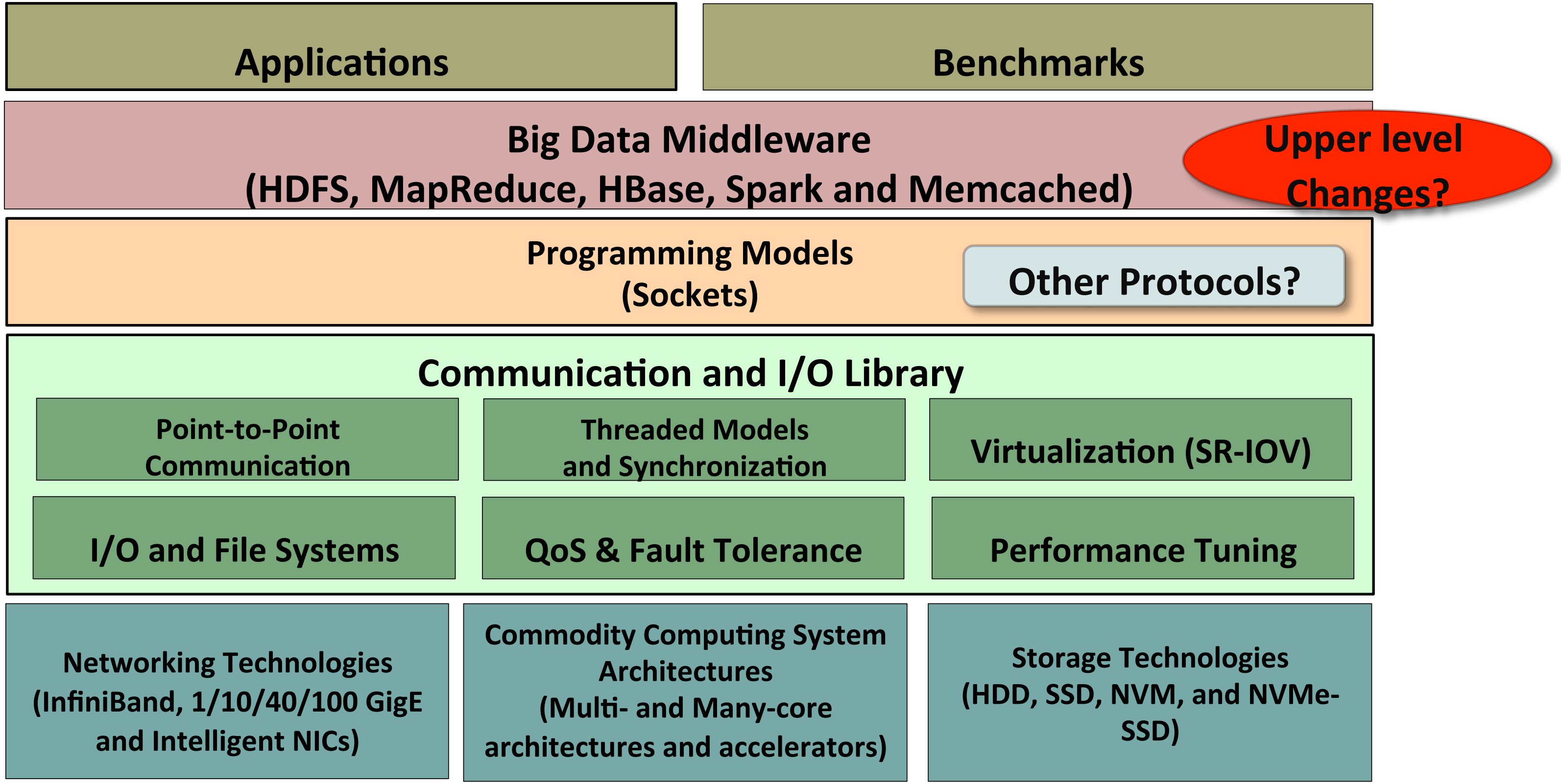
Can We Run Big Data and Deep Learning Jobs on Existing HPC Infrastructure?



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 (SR-IOV)

I/O and File Systems

QoS & Fault Tolerance

Performance Tuning

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, NVM, and NVMe-SSD)

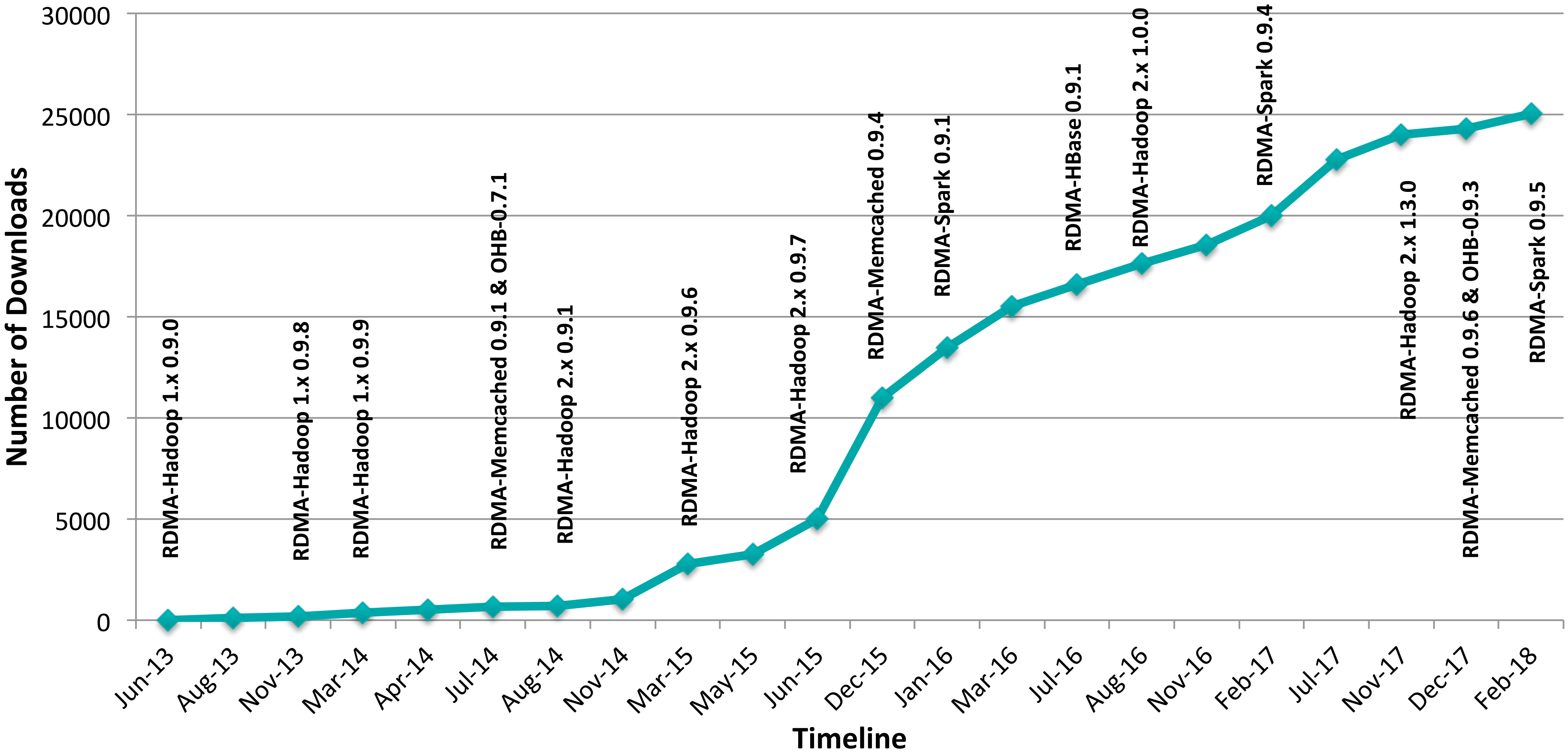
The High-Performance Big Data (HiBD) Project

- 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: 275 organizations from 34 countries
- More than 25,450 downloads from the project site

Available for InfiniBand and RoCE

Also run on Ethernet

HiBD Release Timeline and Downloads

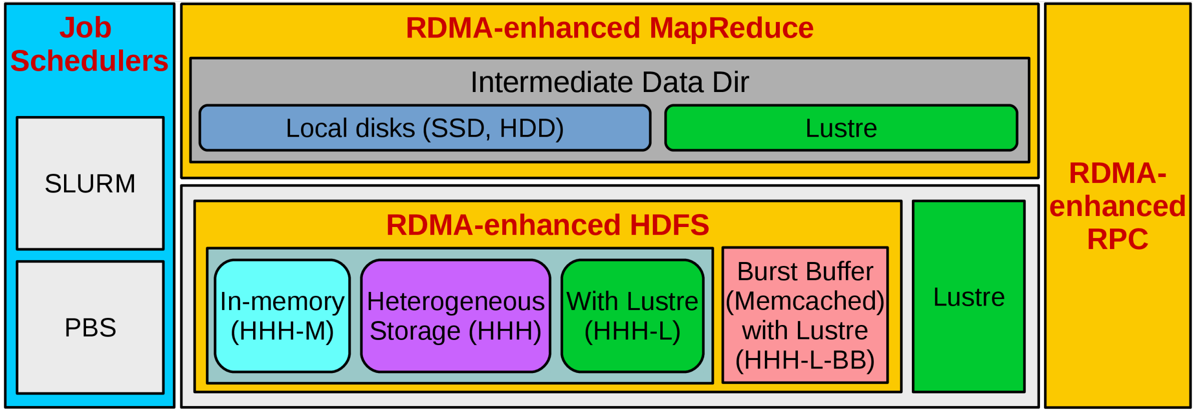


RDMA for Apache Hadoop 2.x Distribution

- 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
  - Memcached-based burst buffer for MapReduce over Lustre-integrated HDFS (HHH-L-BB mode)
  - Plugin-based architecture supporting RDMA-based designs for Apache Hadoop, CDH and HDP
  - Support for OpenPOWER
  - Easily configurable for different protocols (native InfiniBand, RoCE, Ethernet, and IPoIB)
- Current release: 1.3.0
  - Based on Apache Hadoop 2.8.0
  - Compliant with Apache Hadoop 2.8.0, HDP 2.5.0.3 and CDH 5.8.2 APIs and applications
  - Tested with
    - Mellanox InfiniBand adapters (DDR, QDR, FDR, and EDR)
    - RoCE support with Mellanox adapters
    - Various multi-core platforms (x86, POWER)
    - Different file systems with disks and SSDs and Lustre

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.
- **HHH-L-BB**: This mode deploys a Memcached-based burst buffer system to reduce the bandwidth bottleneck of shared file system access. The burst buffer design is hosted by Memcached servers, each of which has a local SSD.
- **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).

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
  - Off-JVM-heap buffer management
  - Support for OpenPOWER
  - Easily configurable for different protocols (native InfiniBand, RoCE, Ethernet, and IPoIB)
- Current release: 0.9.5
  - Based on Apache Spark 2.1.0
  - Tested with
    - Mellanox InfiniBand adapters (DDR, QDR, FDR, and EDR)
    - RoCE support with Mellanox adapters
    - Various multi-core platforms (x86, POWER)
    - RAM disks, SSDs, and HDD
  - http://hibd.cse.ohio-state.edu

RDMA for Apache HBase Distribution

- High-Performance Design of HBase over RDMA-enabled Interconnects
  - High performance RDMA-enhanced design with native InfiniBand and RoCE support at the verbs-level for HBase
  - Compliant with Apache HBase 1.1.2 APIs and applications
  - On-demand connection setup
  - OpenPOWER support is being worked out and will be available in future
  - Easily configurable for different protocols (native InfiniBand, RoCE, Ethernet, and IPoIB)
- Current release: 0.9.1
  - Based on Apache HBase 1.1.2
  - Tested with
    - Mellanox InfiniBand adapters (DDR, QDR, FDR, and EDR)
    - RoCE support with Mellanox adapters
    - Various multi-core platforms
  - http://hibd.cse.ohio-state.edu

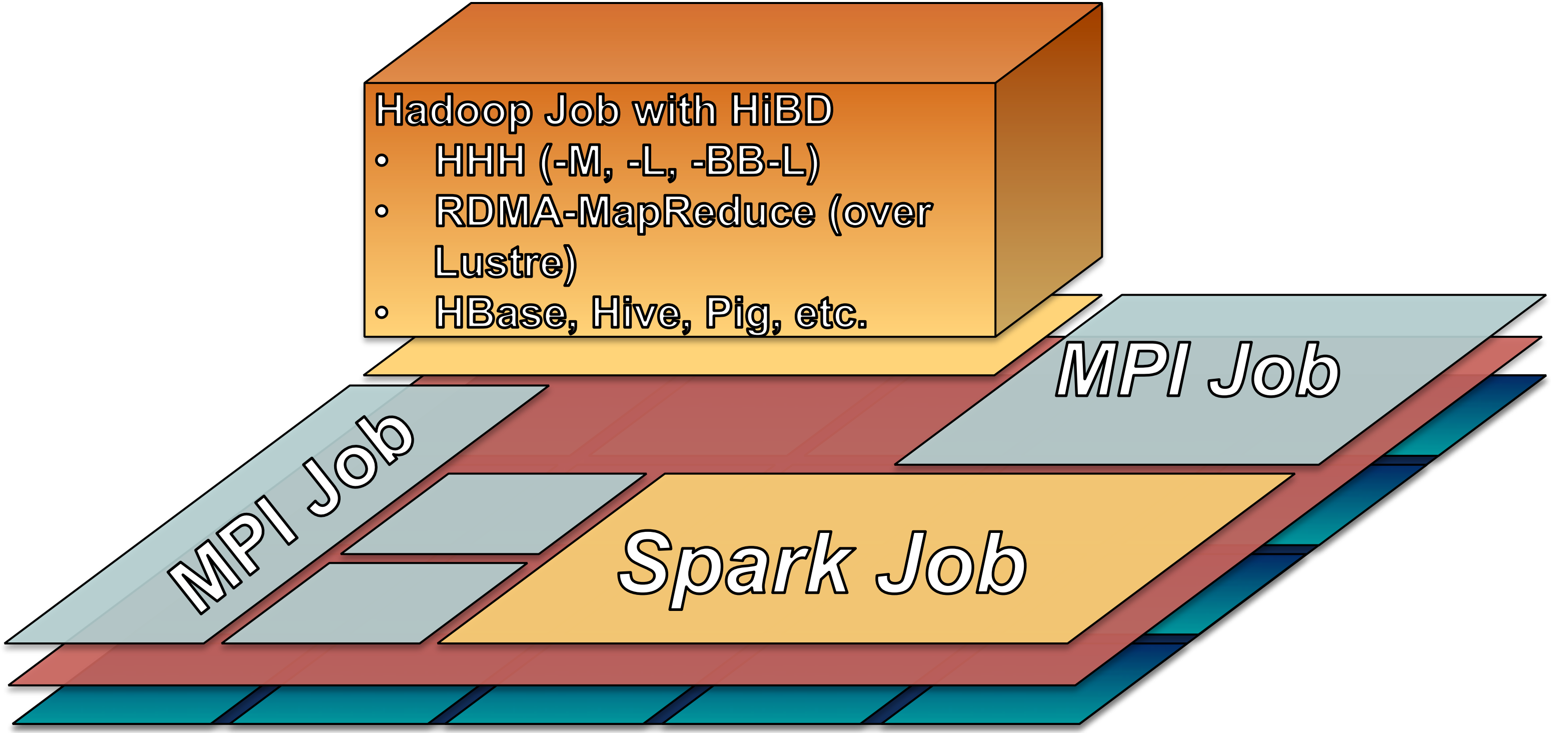
RDMA for Memcached Distribution

- 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
  - Non-Blocking Libmemcached Set/Get API extensions
  - Support for burst-buffer mode in Lustre-integrated design of HDFS in RDMA for Apache Hadoop-2.x
  - OpenPOWER support is being worked out and will be available in future
  - Easily configurable for different protocols (native InfiniBand, RoCE, Ethernet, and IPoIB)
- Current release: 0.9.6
  - Based on Memcached 1.5.3 and libMemcached 1.0.18
  - Compliant with libMemcached APIs and applications
  - Tested with
    - Mellanox InfiniBand adapters (DDR, QDR, FDR, and EDR)
    - RoCE support with Mellanox adapters
    - Various multi-core platforms
    - SSD
  - http://hibd.cse.ohio-state.edu

OSU HiBD Micro-Benchmark (OHB) Suite – HDFS, Memcached, HBase, and Spark

- Micro-benchmarks for Hadoop Distributed File System (HDFS)
  - Sequential Write Latency (**SWL**) Benchmark, Sequential Read Latency (**SRL**) Benchmark, Random Read Latency (**RRL**) Benchmark, Sequential Write Throughput (**SWT**) Benchmark, Sequential Read Throughput (**SRT**) Benchmark
  - Support benchmarking of
    - Apache Hadoop 1.x and 2.x HDFS, Hortonworks Data Platform (HDP) HDFS, Cloudera Distribution of Hadoop (CDH) HDFS
- Micro-benchmarks for Memcached
  - **Get** Benchmark, **Set** Benchmark, and **Mixed** Get/Set Benchmark, **Non-Blocking API** Latency Benchmark**, Hybrid Memory** Latency Benchmark
  - Yahoo! Cloud Serving Benchmark (YCSB) Extension for RDMA-Memcached
- Micro-benchmarks for HBase
  - **Get** Latency Benchmark, **Put** Latency Benchmark
- Micro-benchmarks for Spark
  - GroupBy, SortBy
- Current release: 0.9.3
- **http://hibd.cse.ohio-state.edu**

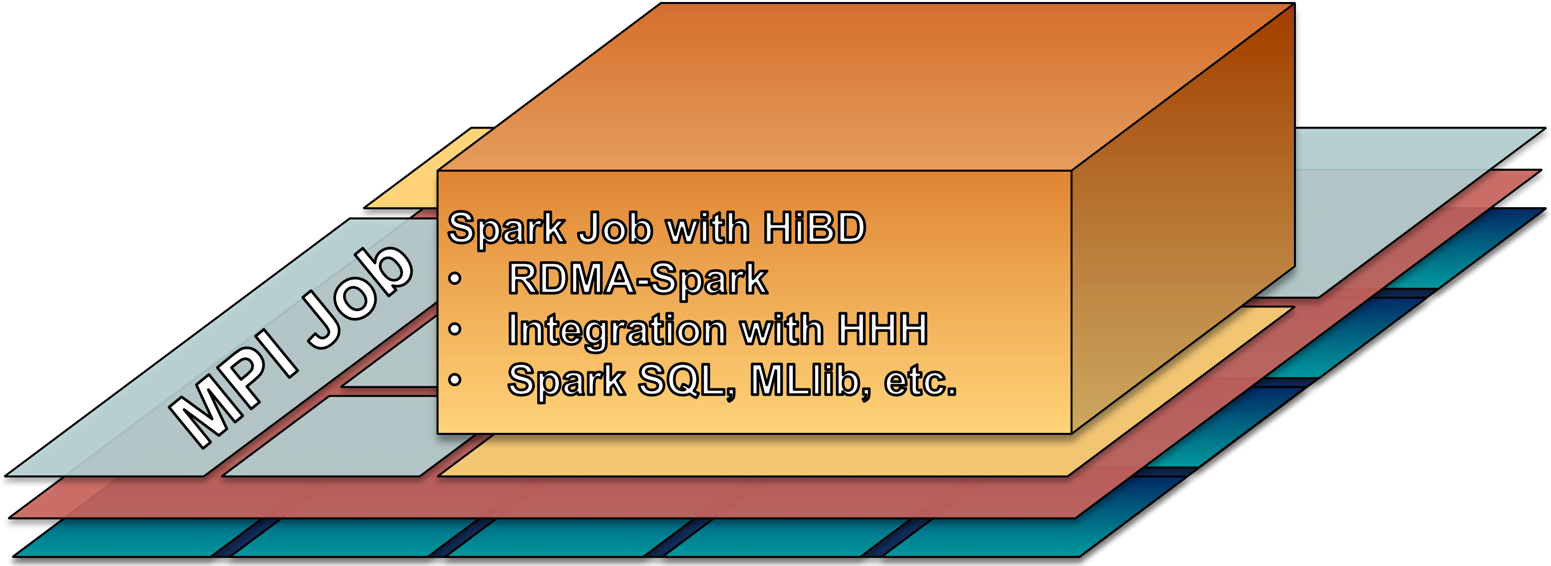
Using HiBD Packages on Existing HPC Infrastructure



Hadoop Job with HiBD

- HHH (-M, -L, -BB-L)
- RDMA-MapReduce (over Lustre)
- HBase, Hive, Pig, etc.

Using HiBD Packages on Existing HPC Infrastructure



Spark Job with HiBD

- RDMA-Spark
- Integration with HHH
- Spark SQL, MLlib, etc.

HiBD Packages on SDSC Comet and Chameleon Cloud

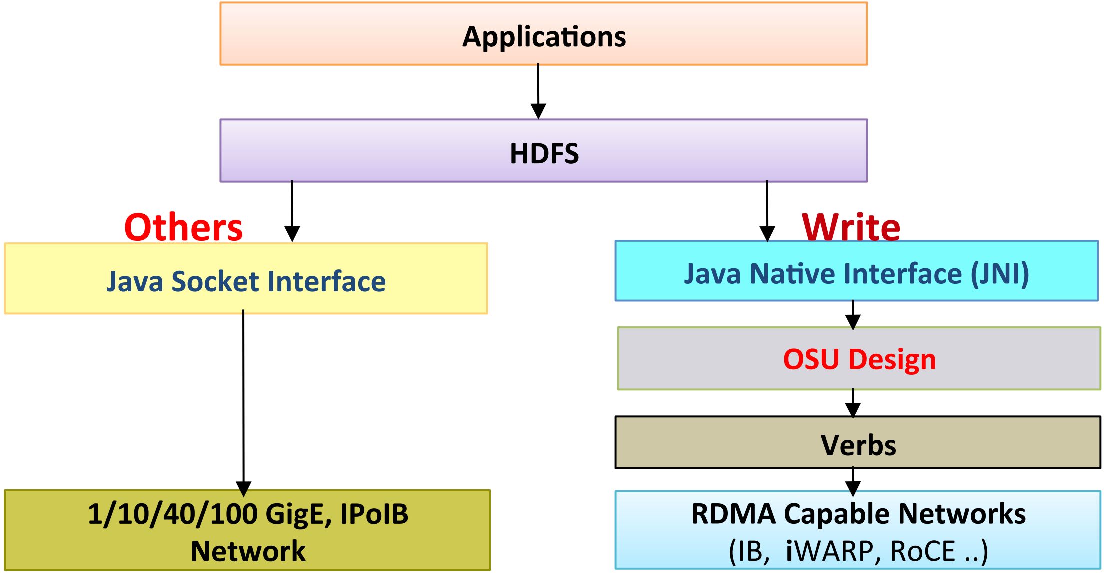
- RDMA for Apache Hadoop 2.x and RDMA for Apache Spark are installed and available on SDSC Comet.
  - Examples for various modes of usage are available in:
    - RDMA for Apache Hadoop 2.x: /share/apps/examples/HADOOP
    - RDMA for Apache Spark: /share/apps/examples/SPARK/
  - Please email help@xsede.org (reference Comet as the machine, and SDSC as the site) if you have any further questions about usage and configuration.
- RDMA for Apache Hadoop is also available on Chameleon Cloud as an appliance
  - https://www.chameleoncloud.org/appliances/17/

M. Tatineni, X. Lu, D. J. Choi, A. Majumdar, and D. K. Panda, Experiences and Benefits of Running RDMA Hadoop and Spark on SDSC Comet, XSEDE'16, July 2016

Acceleration Case Studies and Performance Evaluation

- HDFS and MapReduce
- Spark

Design Overview of HDFS with RDMA



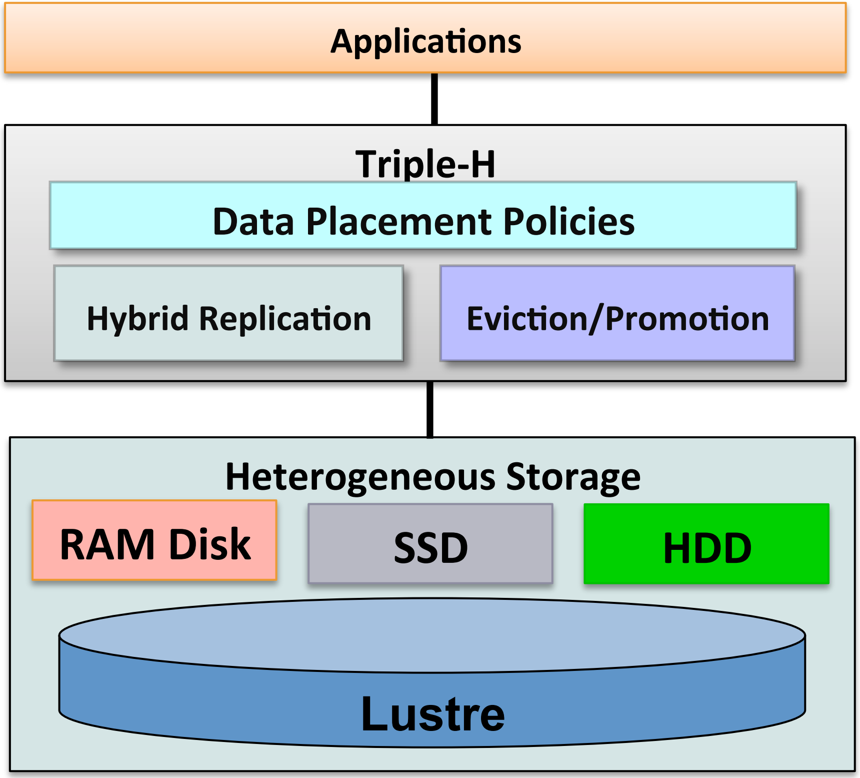
- Design Features
  - RDMA-based HDFS write
  - RDMA-based HDFS replication
  - Parallel replication support
  - On-demand connection setup
  - InfiniBand/RoCE support

- 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

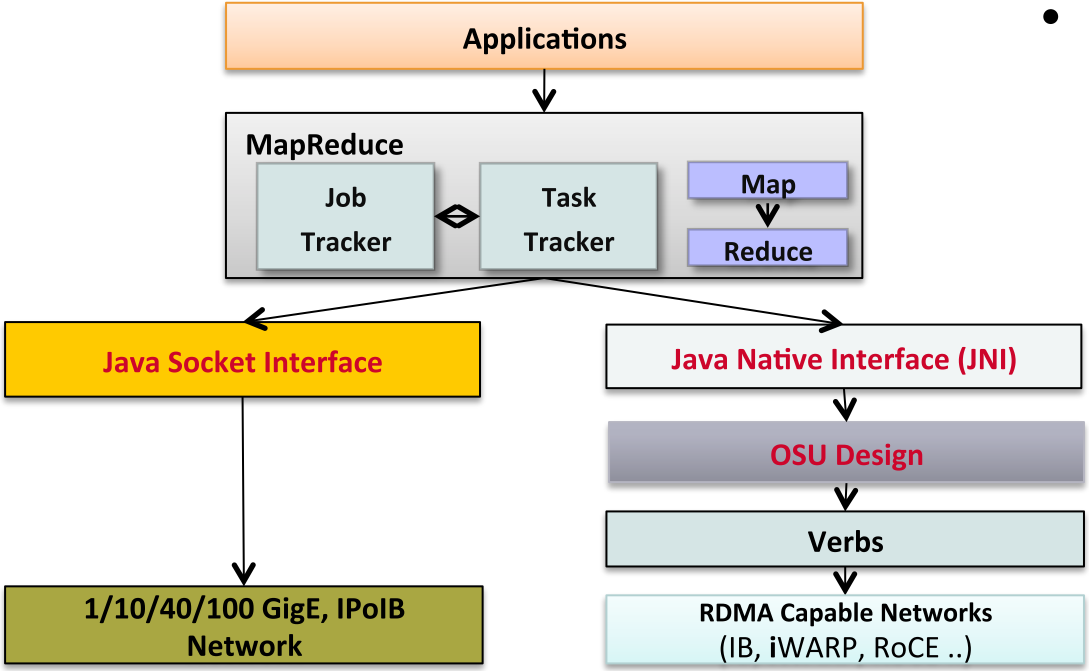
Enhanced HDFS with In-Memory and Heterogeneous Storage



- 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

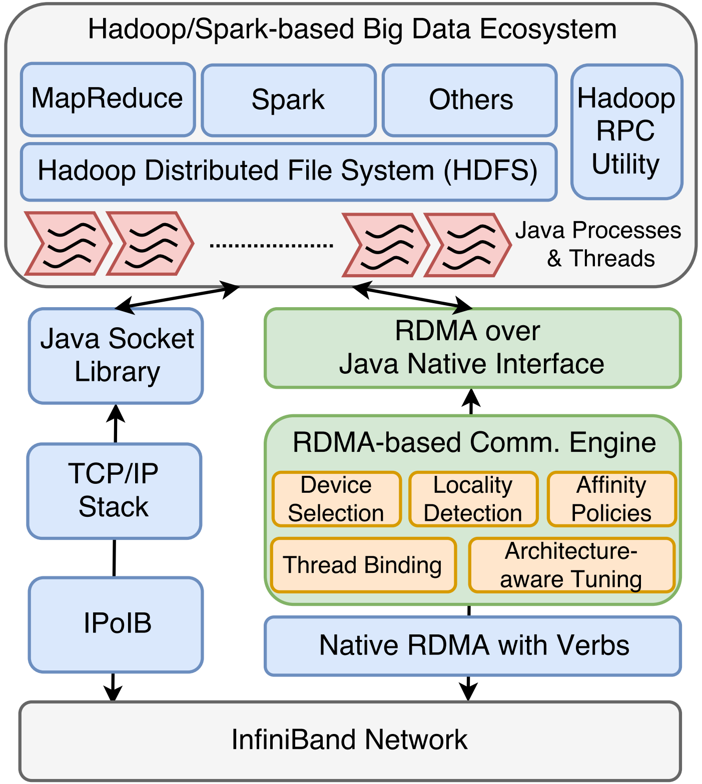

Applications
MapReduce
Job Tracker
Task Tracker
Map
Reduce
Java Socket Interface
Java Native Interface (JNI)
OSU Design
Verbs
1/10/40/100 GigE, IPoIB Network
RDMA Capable Networks (IB, iWARP, RoCE ..)

- 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

Accelerations for OpenPOWER Platform

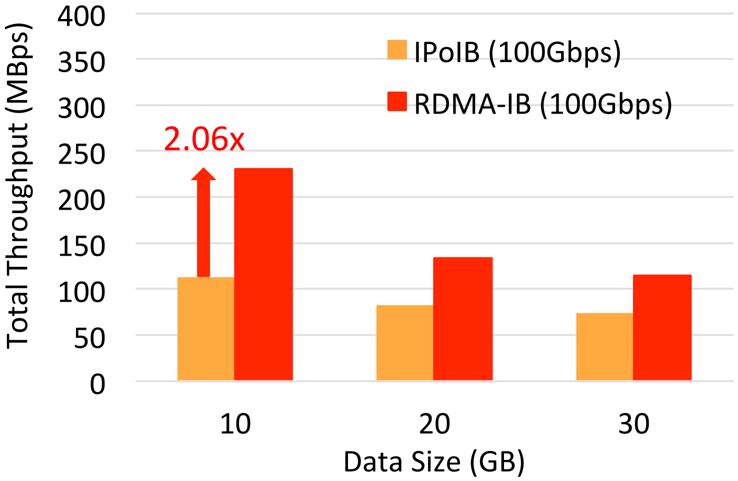


- RDMA Device Selection and Locality Detection
- Affinity Policies and Enforcement
  - One Communication Thread to N SMT Threads on the Closer-to-HCA Socket
    - One Communication Thread to All Physical Cores on the Closer-to-HCA Socket (e.g., N=80)
    - One Communication Thread to One Physical Core on the Closer-to-HCA Socket (e.g., N=8)
  - One Communication Thread to N SMT Threads on a CPU
    - One Communication Thread to One Physical Core on the CPU (e.g., N=8)
- POWER architecture aware tuning
- Support both Hadoop and Spark

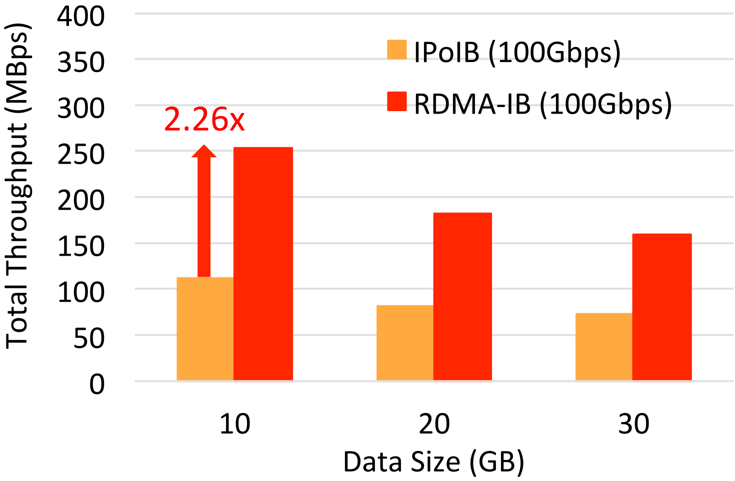
X. Lu, H. Shi, D. Shankar, and D. K. Panda, Performance Characterization and Acceleration of Big Data Workloads on OpenPOWER System, IEEE BigData '17, Dec. 2017.

Performance of RDMA-Hadoop on OpenPOWER

TestDFSIO Throughput



HHH Mode

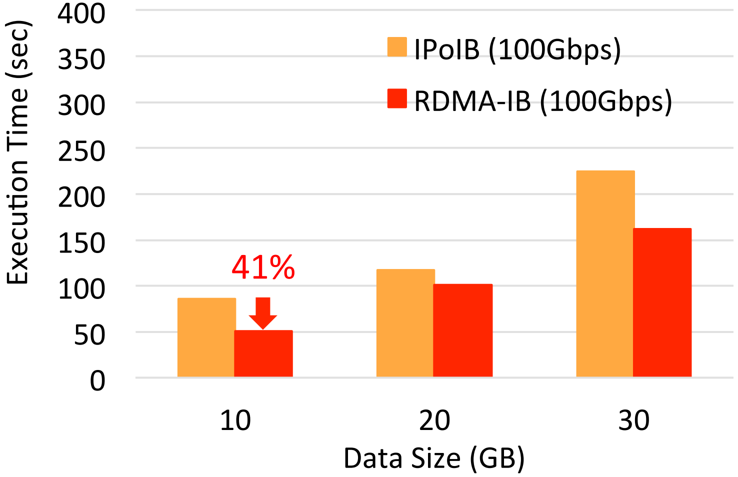


HHH-M Mode

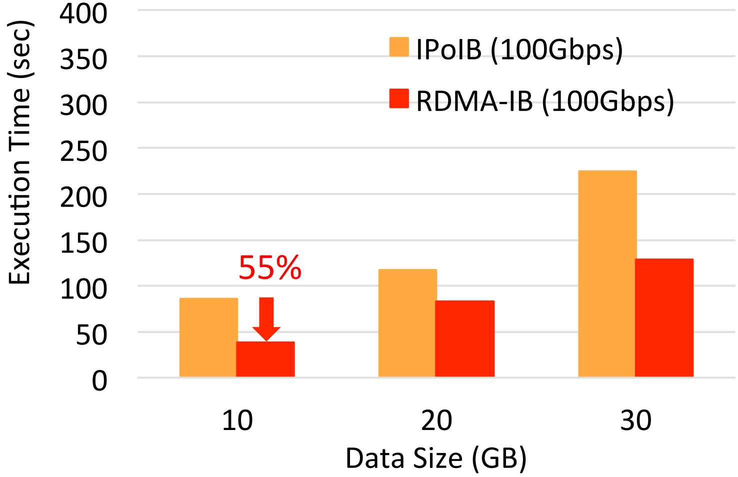
- For TestDFSIO throughput experiment, RDMA-IB design of HHH mode has an improvement of 1.57x - 2.06x compared to IPoIB (100Gbps).
- In HHH-M mode, the improvement goes up to 2.18x - 2.26x compared to IPoIB (100Gbps).

Performance of RDMA-Hadoop on OpenPOWER

Sort Execution Time



HHH Mode

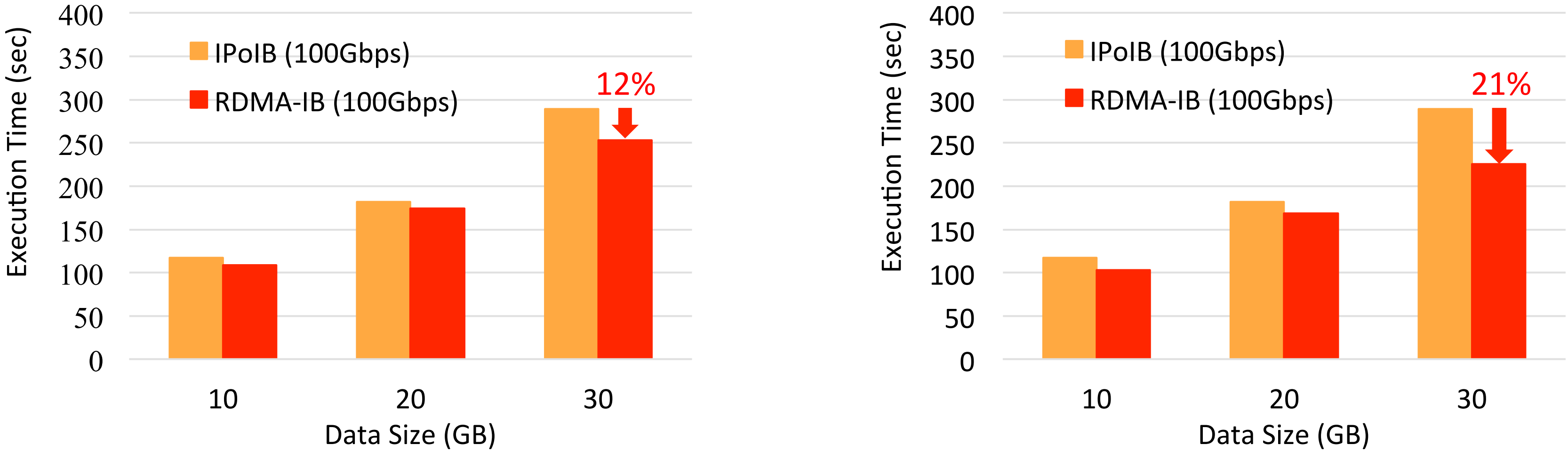


HHH-M Mode

- The RDMA-IB design of HHH mode reduces the job execution time of Sort by a maximum of 41% compared to IPoIB (100Gbps).
- The HHH-M design reduces the execution time by a maximum of 55%.

Performance of RDMA-Hadoop on OpenPOWER

TeraSort Execution Time



HHH Mode

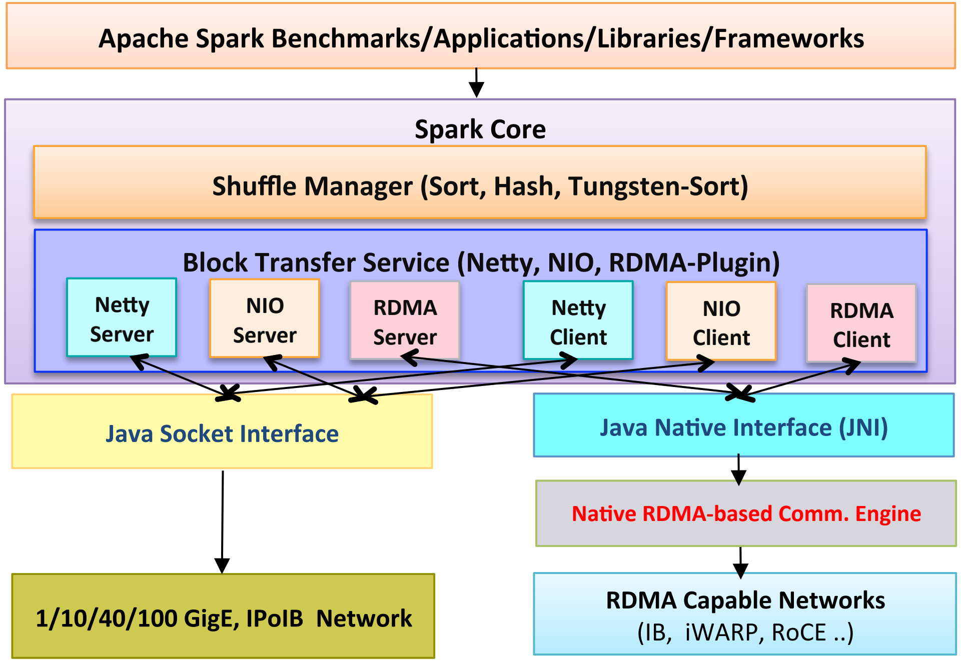
HHH-M Mode

- The RDMA-IB design of HHH mode reduces the job execution time of TeraSort by a maximum of 12% compared to IPoIB (100Gbps).
- In HHH-M mode, the execution time of TeraSort is reduced by a maximum of 21% compared to IPoIB (100Gbps).

Acceleration Case Studies and Performance Evaluation

- HDFS and MapReduce
- Spark

Design Overview of Spark with RDMA


Apache Spark Benchmarks/Applications/Libraries/Frameworks
Spark Core
Shuffle Manager (Sort, Hash, Tungsten-Sort)
Block Transfer Service (Netty, NIO, RDMA-Plugin)
Netty Server
NIO Server
RDMA Server
Netty Client
NIO Client
RDMA Client
Java Socket Interface
Java Native Interface (JNI)
Native RDMA-based Comm. Engine
1/10/40/100 GigE, IPoIB Network
RDMA Capable Networks
(IB, iWARP, RoCE ..)

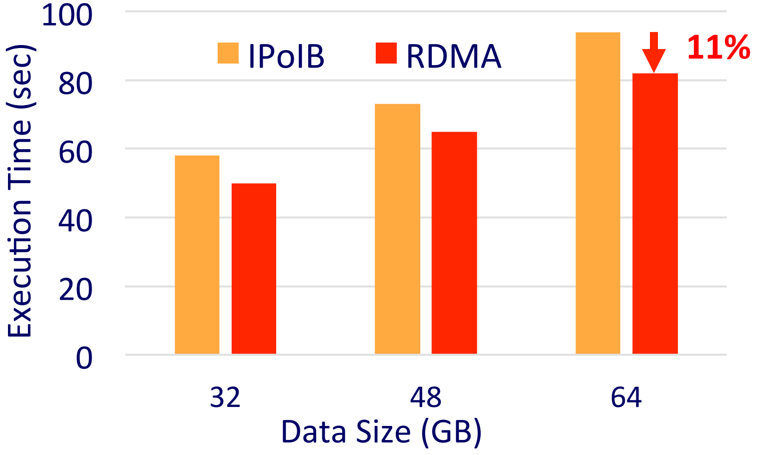
- Design Features
  - RDMA based shuffle plugin
  - SEDA-based architecture
  - Dynamic connection management and sharing
  - Non-blocking data transfer
  - Off-JVM-heap buffer management
  - InfiniBand/RoCE support

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

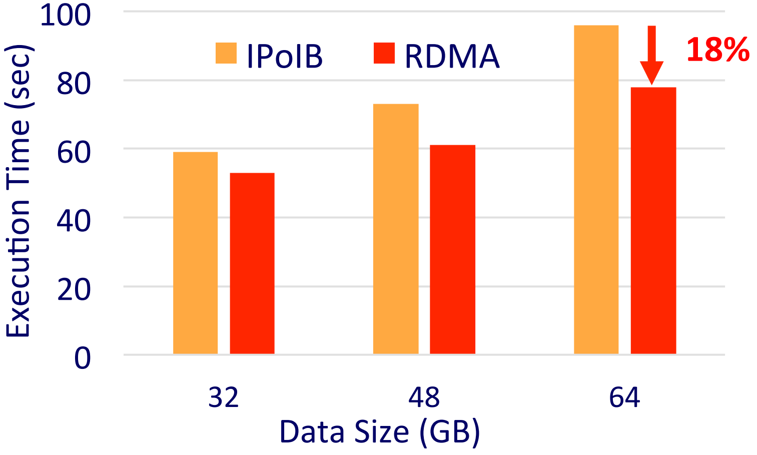
X. Lu, M. W. Rahman, N. Islam, D. Shankar, and D. K. Panda, Accelerating Spark with RDMA for Big Data Processing: Early Experiences, Int'l Symposium on High Performance Interconnects (HotI'14), August 2014

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 of RDMA-Spark on OpenPOWER



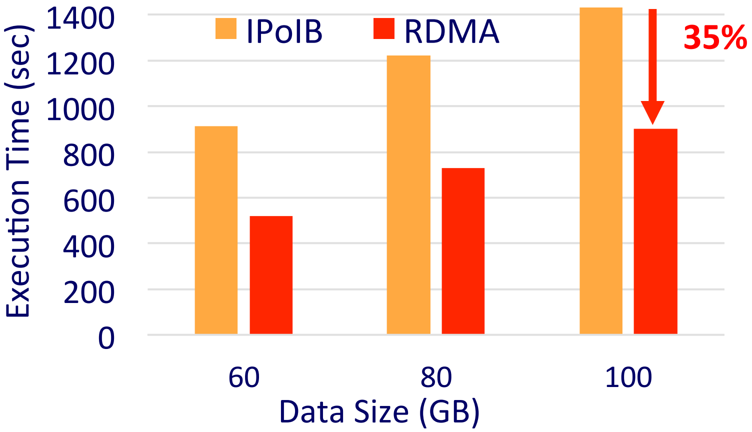
GroupBy



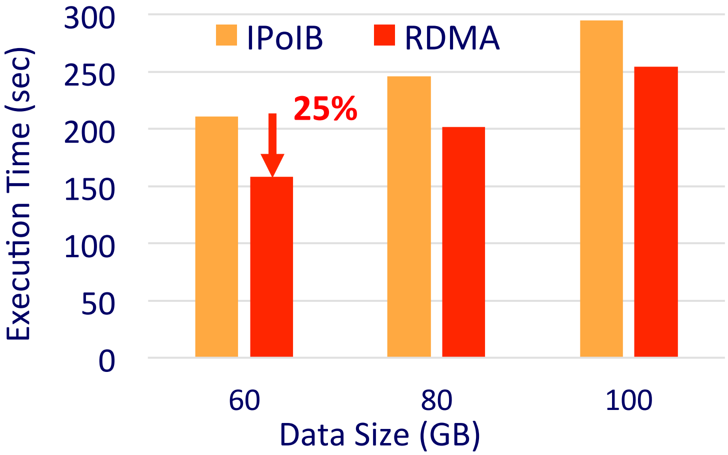
SortBy

- GroupBy: RDMA design outperforms IPoIB by a maximum of 11%
- SortBy: RDMA design outperforms IPoIB by a maximum of 18%

Performance of RDMA-Spark on OpenPOWER



TeraSort



Sort

- TeraSort: RDMA design outperforms IPoIB by a maximum of 35%
- Sort: RDMA design outperforms IPoIB by a maximum of 25%

Ongoing R&D Activities and Plans

- OpenPOWER support and additional accelerations for other Big Data Analytics and Deep Learning Stacks
  - Apache HBase, Memcached
  - TensorFlow, Caffe
  - Many others
- "GPGPU + POWER + InfiniBand" for Big Data and Deep Learning applications
- Exploring acceleration opportunities on POWER9
- Virtualization on OpenPOWER
  - KVM, Xen, etc.
  - Docker, Singularity, etc.

Concluding Remarks

- Discussed challenges in accelerating Big Data Stacks with HPC technologies
- Presented basic and advanced designs to take advantage of InfiniBand/RDMA for HDFS, MapReduce, and Spark on HPC clusters
- ‘OpenPOWER with InfiniBand’ platform is becoming available in many HPC clusters and datacenters
- Performance characterization and acceleration for RDMA-based communication engine on POWER8 architecture
- Results are promising for Big Data processing tools
- Will enable Big Data community to take advantage of modern HPC technologies and POWER architecture to carry out their analytics in a fast and scalable manner

The 4th International Workshop on High-Performance Big Data Computing (HPBDC)

HPBDC 2018 will be held with IEEE International Parallel and Distributed Processing Symposium (IPDPS 2018), Vancouver, British Columbia CANADA, May, 2018

http://web.cse.ohio-state.edu/~luxi/hpbdc2018

HPBDC 2017 was held in conjunction with IPDPS'17
Keynote Talk: Prof. Satoshi Matsuoka, Converging HPC and Big Data / AI Infrastructures at Scale with BYTES-Oriented Architectures
Panel Topic: Sunrise or Sunset: Exploring the Design Space of Big Data Software Stack
Six Regular Research Papers and One Short Research Papers

http://web.cse.ohio-state.edu/~luxi/hpbdc2017

HPBDC 2016 was held in conjunction with IPDPS'16
Keynote Talk: Dr. Chaitanya Baru, High Performance Computing and Which Big Data?
Panel Topic: Merge or Split: Mutual Influence between Big Data and HPC Techniques
Six Regular Research Papers and Two Short Research Papers

http://web.cse.ohio-state.edu/~luxi/hpbdc2016

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

HiBD
High-Performance
Big Data

Network-Based Computing Laboratory
http://nowlab.cse.ohio-state.edu/

The High-Performance Big Data Project
http://hibd.cse.ohio-state.edu/