HiBD
High-Performance
Big Data

Architectural Convergence of Big Data and Extreme-Scale Computing: Marriage of Convenience or Conviction

Panel at Charm++ (April ‘18)

by

Dhabaleswar K. (DK) Panda

The Ohio State University

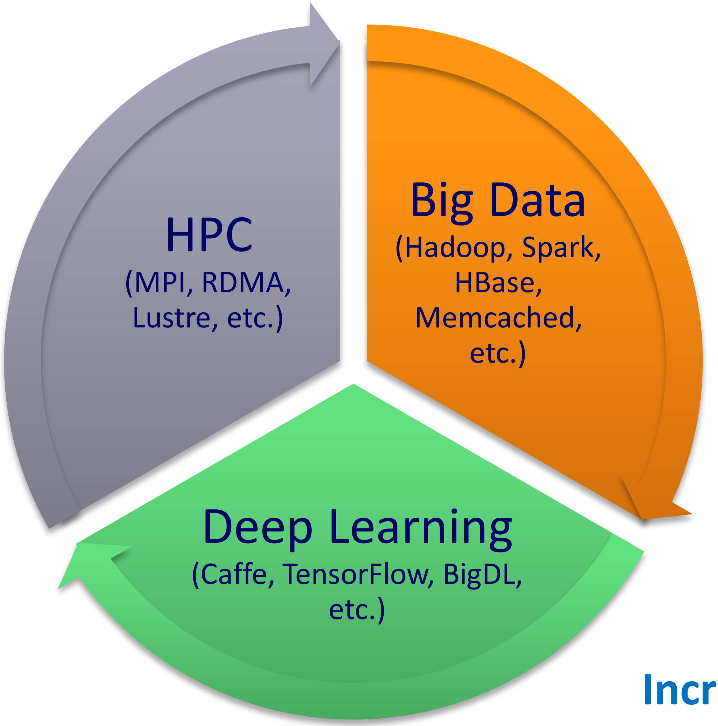
E-mail: panda@cse.ohio-state.edu

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

My Answer

Conviction

Increasing Usage of HPC, Big Data and Deep Learning


HPC
(MPI, RDMA, Lustre, etc.)

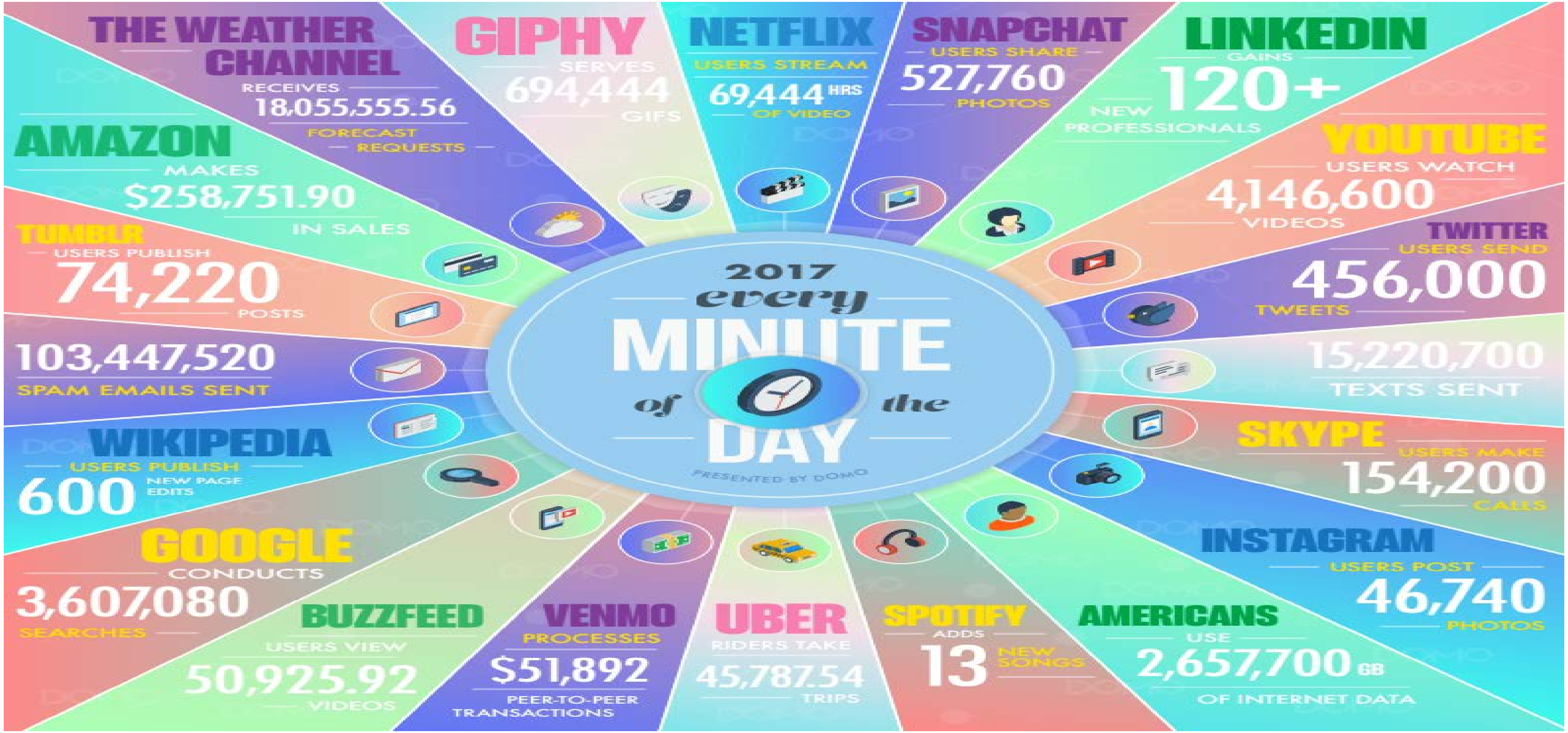
Big Data
(Hadoop, Spark, HBase, Memcached, etc.)

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?



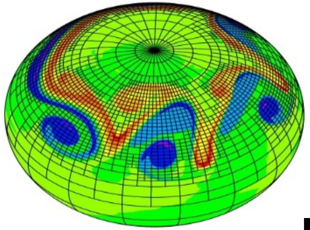
The global Internet population grew 7.5% from 2016 and now represents

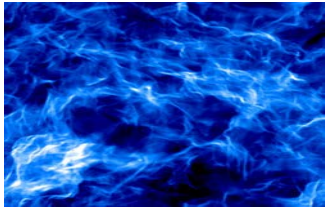
3.7 Billion People.

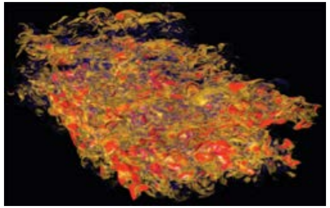
Courtesy: https://www.domo.com/blog/data-never-sleeps-5/

Not Only in Internet Services - Big Data in Scientific Domains

- Scientific Data Management, Analysis, and Visualization
- Applications examples
  - Climate modeling
  - Combustion
  - Fusion
  - Astrophysics
  - Bioinformatics
- Data Intensive Tasks
  - Runs large-scale simulations on supercomputers
  - Dump data on parallel storage systems
  - Collect experimental / observational data
  - Move experimental / observational data to analysis sites
  - Visual analytics – help understand data visually

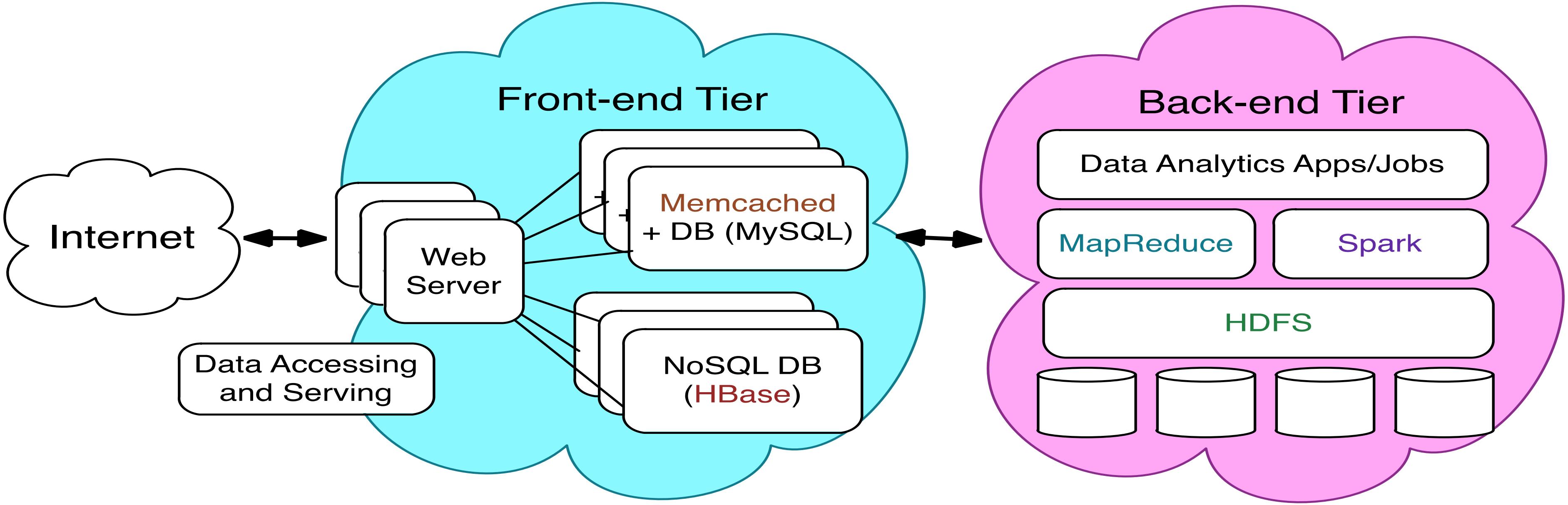






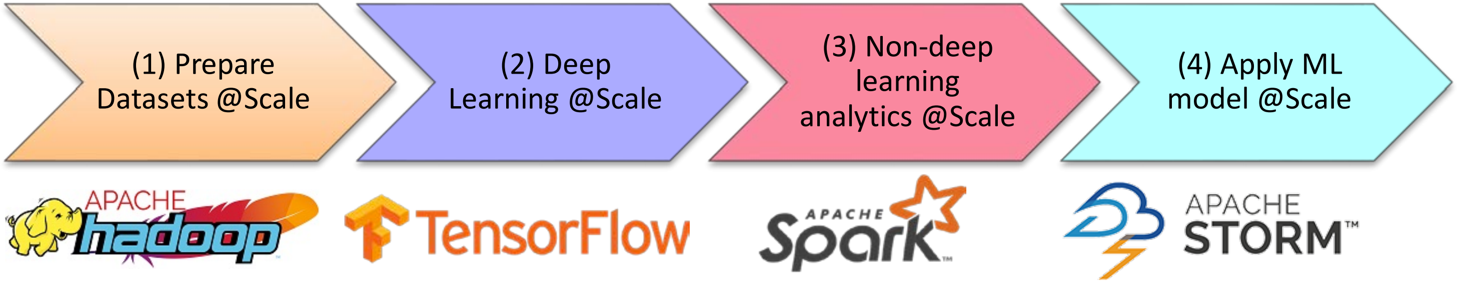
Big Data Management and Processing on Modern Clusters

- 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



Deep Learning over Big Data (DLoBD)

- Deep Learning over Big Data (**DLoBD**) is one of the most efficient analyzing paradigms
- More and more deep learning tools or libraries (e.g., Caffe, TensorFlow) start running over big data stacks, such as Apache Hadoop and Spark
- **Benefits** of the DLoBD approach
  - Easily build a powerful data analytics **pipeline**
    - E.g., Flickr DL/ML Pipeline, *"How Deep Learning Powers Flickr"*, http://bit.ly/1KIDfof



(1) Prepare Datasets @Scale

(2) Deep Learning @Scale

(3) Non-deep learning analytics @Scale

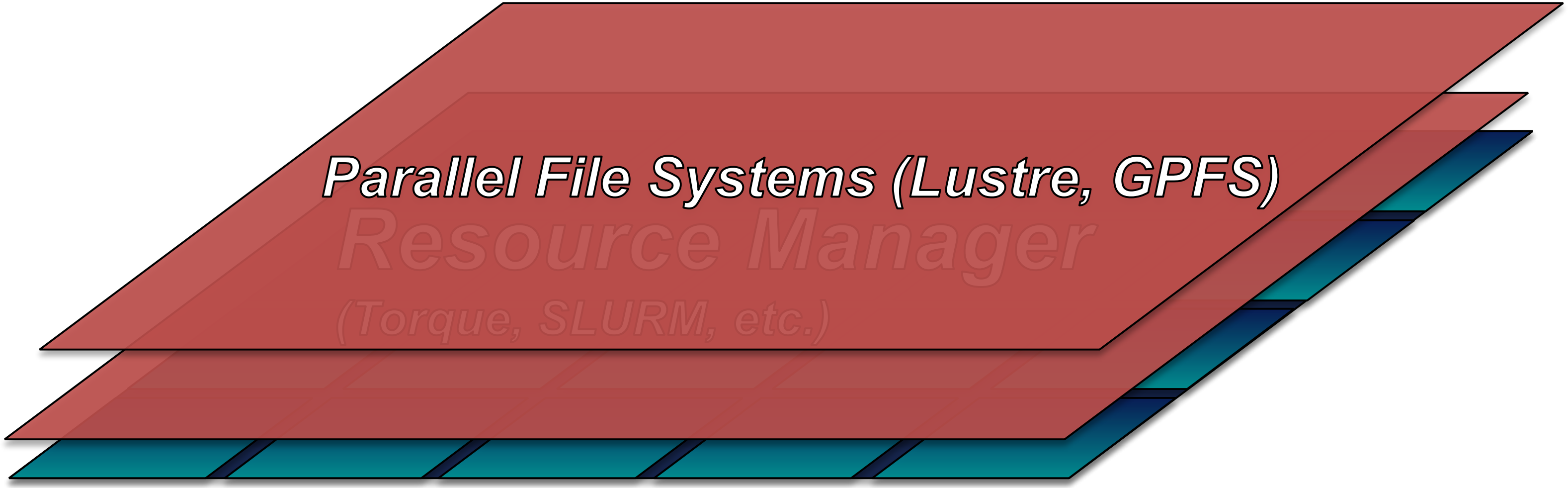
(4) Apply ML model @Scale

  - Better data **locality**
  - Efficient resource sharing and **cost effective**

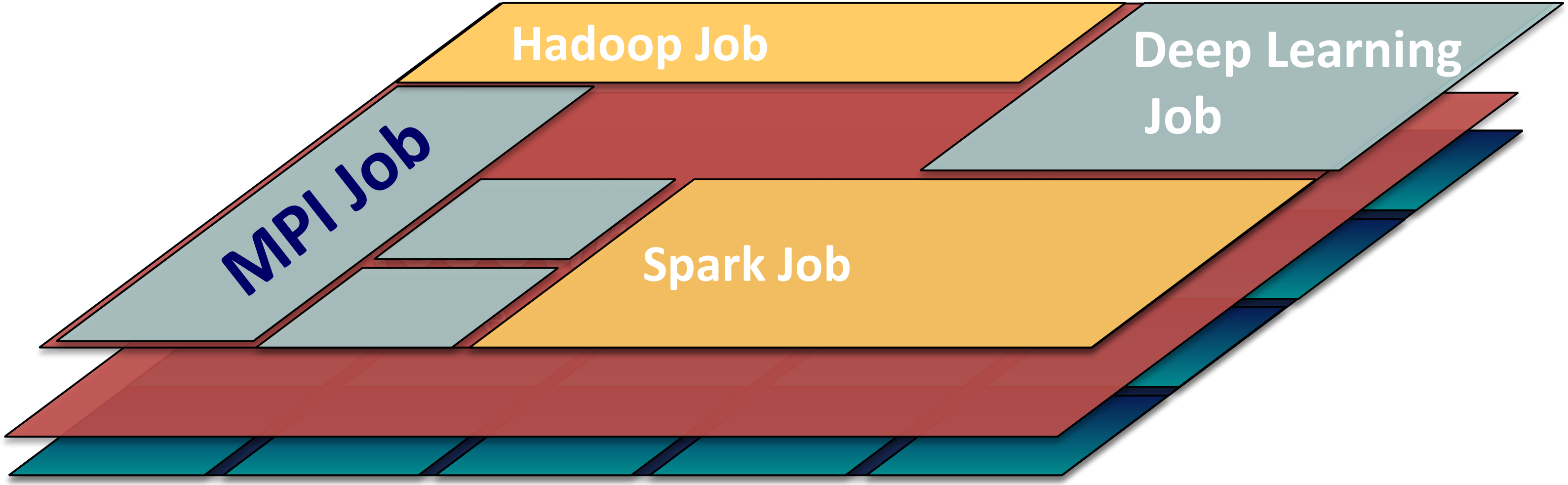
Need to Run Big Data and Deep Learning Jobs on Existing HPC Infrastructure?

Resource Manager
(Torque, SLURM, etc.)

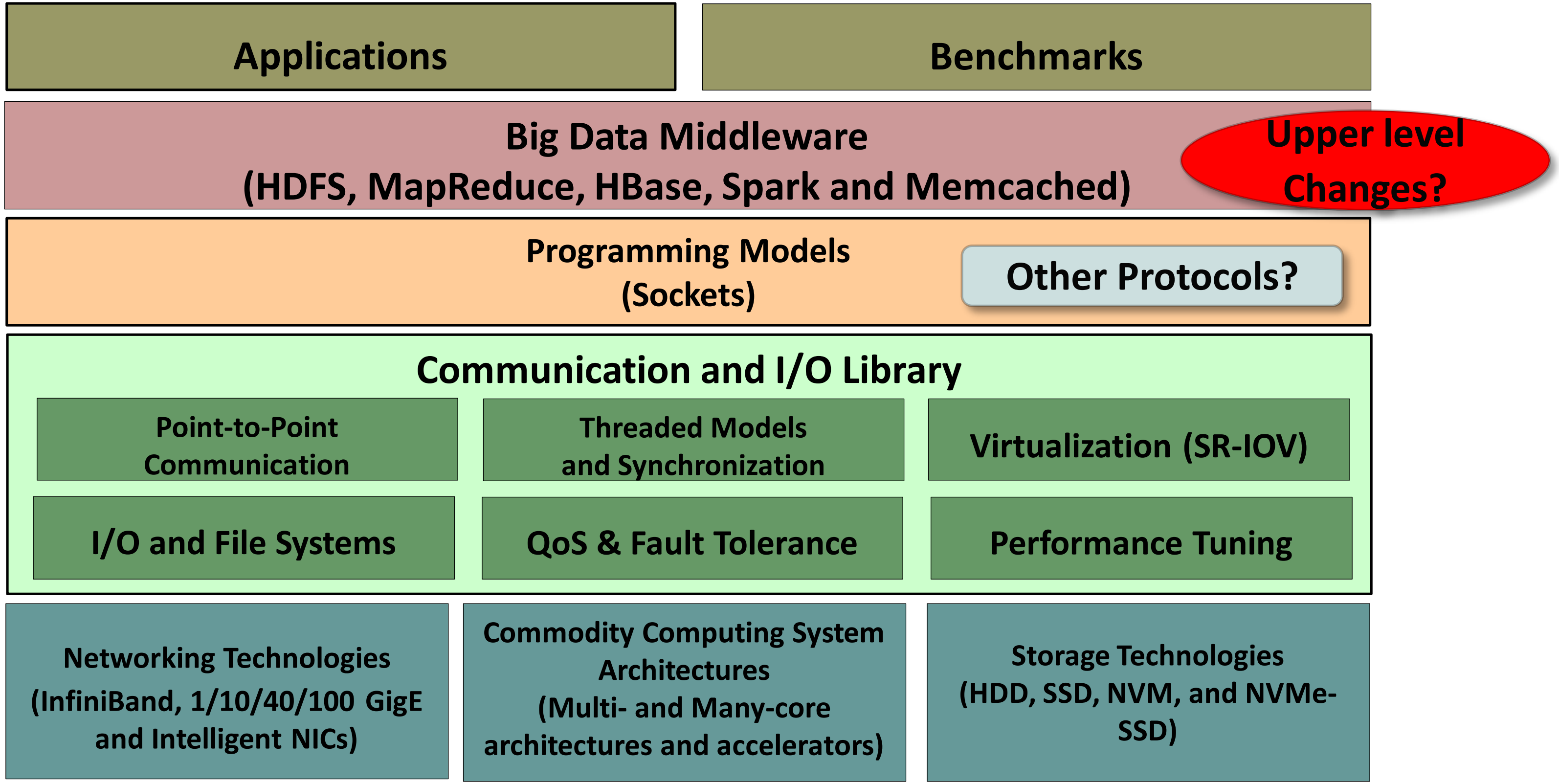
Need to Run Big Data and Deep Learning Jobs on Existing HPC Infrastructure?



Need to 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: 280 organizations from 34 countries
- More than 25,750 downloads from the project site

Available for InfiniBand and RoCE

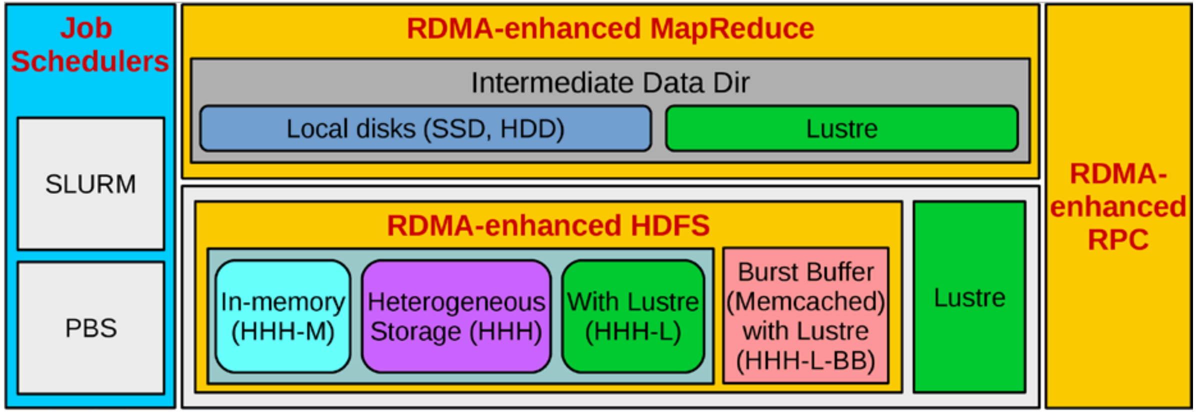
Also run on Ethernet

Available for x86 and OpenPOWER

Upcoming Release will have support

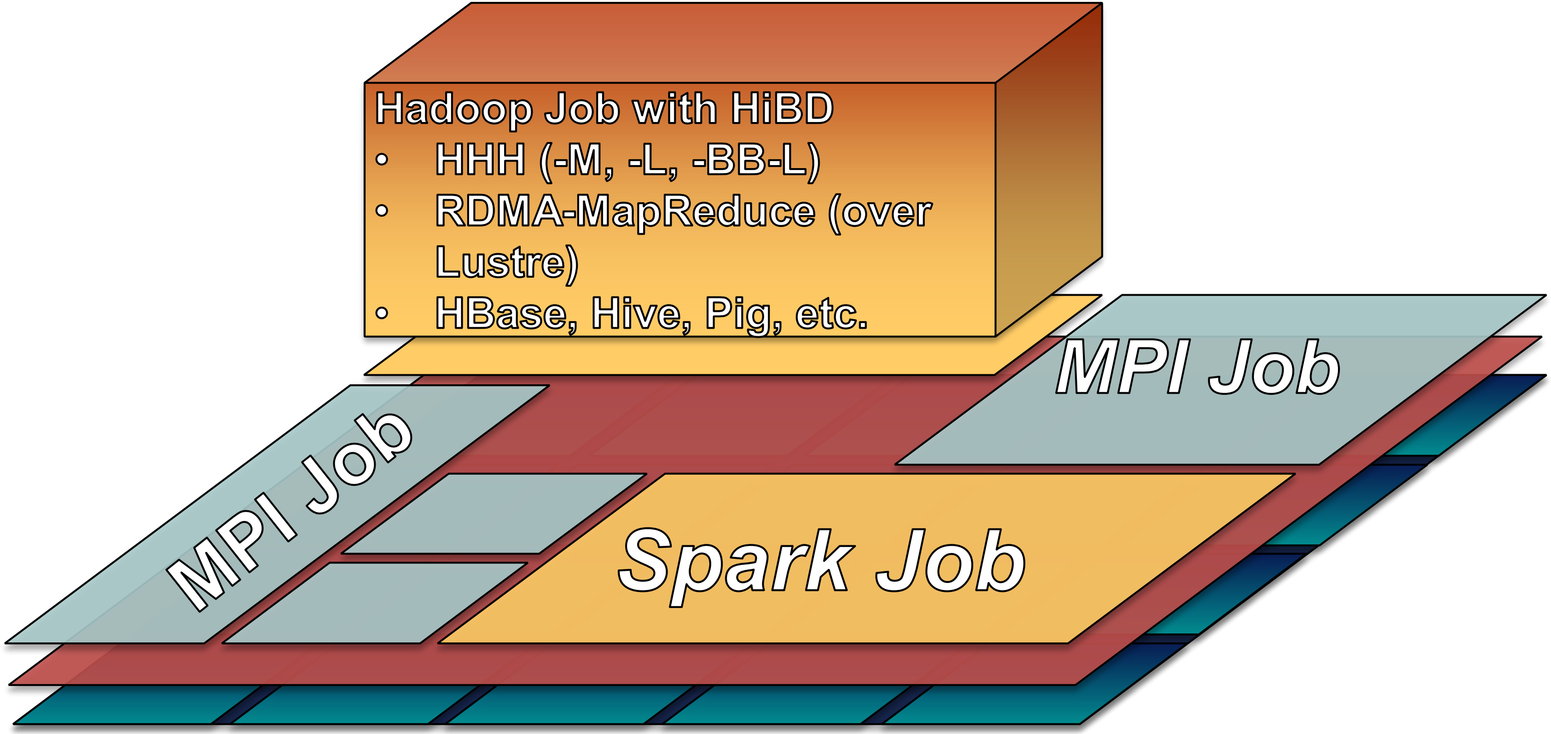
For Singularity and Docker

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).

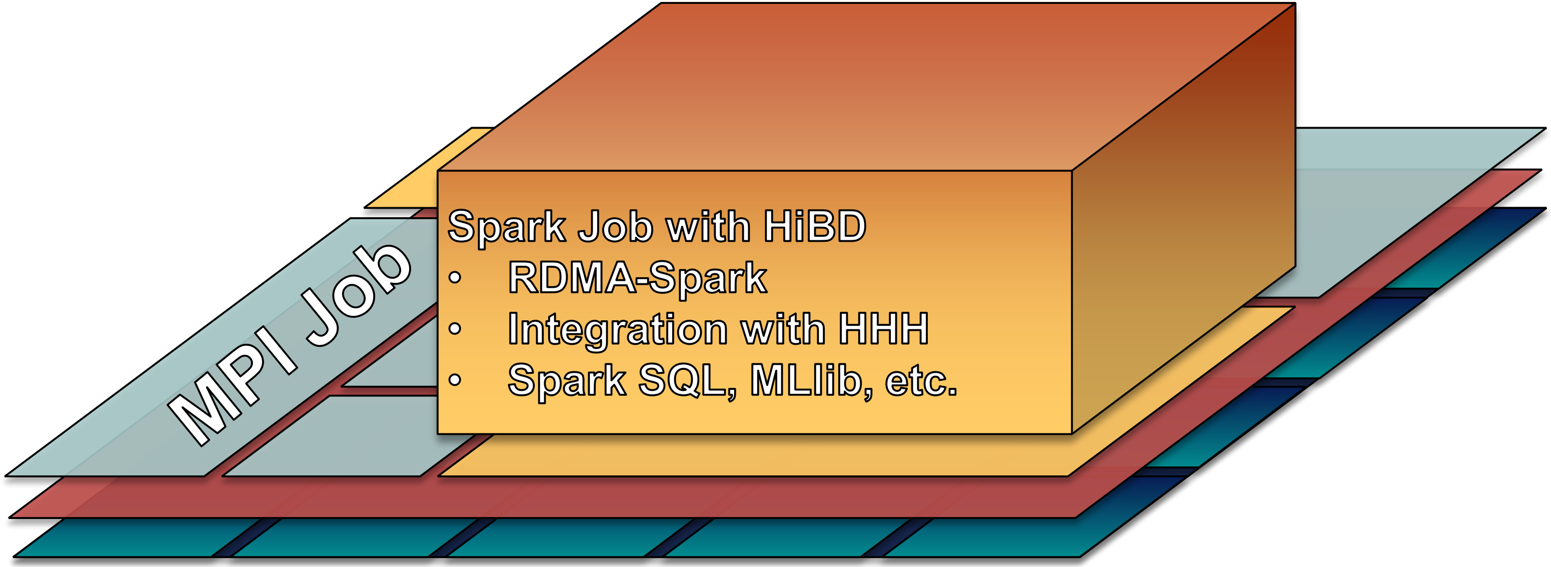
Using HiBD Packages on Existing HPC Infrastructure



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

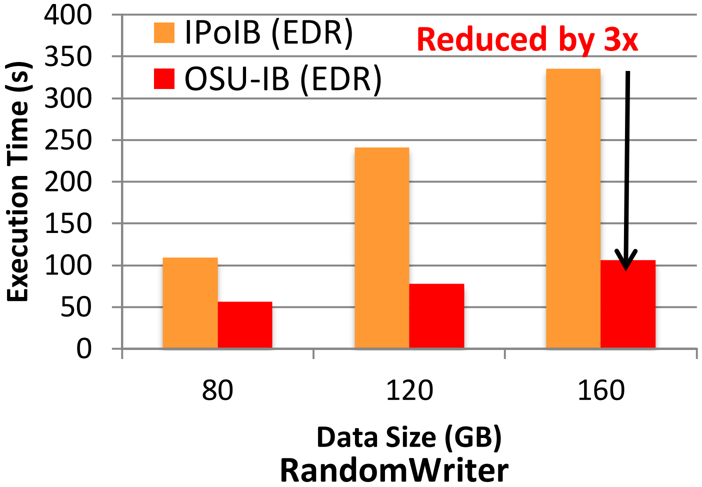
- RDMA-Spark
- Integration with HHH
- Spark SQL, MLlib, etc.

HiBD Packages on SDSC Comet and Chameleon Cloud

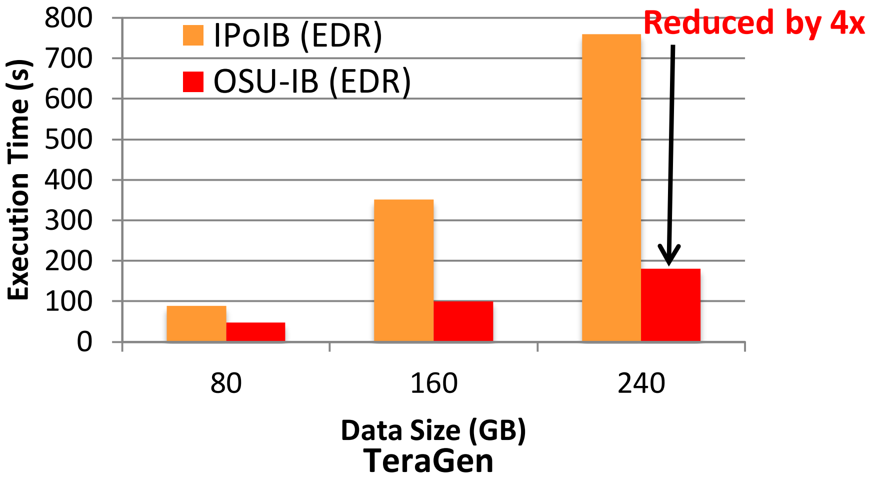
- RDMA for Apache Hadoop 2.x and RDMA for Apache Spark are installed and available on SDSC Comet.
  - Examples for various modes of usage are available in:
    - RDMA for Apache Hadoop 2.x: /share/apps/examples/HADOOP
    - RDMA for Apache Spark: /share/apps/examples/SPARK/
  - Please email help@xsede.org (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

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



RandomWriter

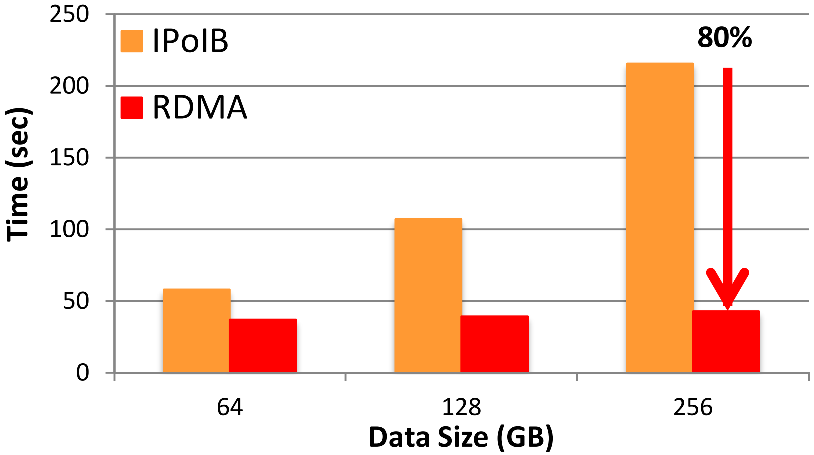


TeraGen

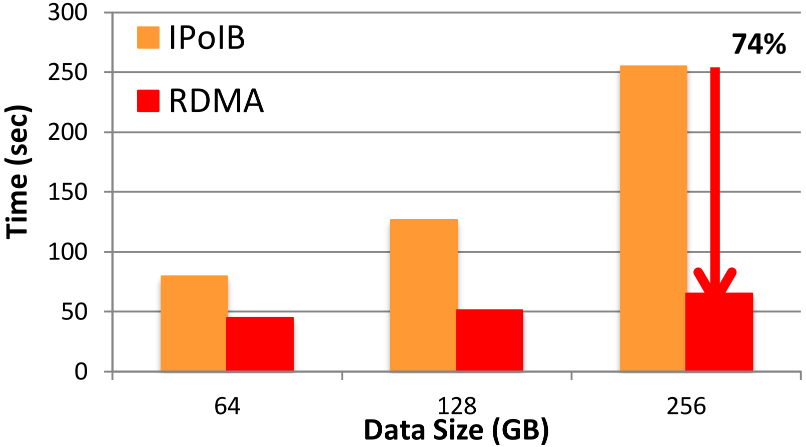
Cluster with 8 Nodes with a total of 64 maps

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

Performance Evaluation of RDMA-Spark on SDSC Comet – SortBy/GroupBy



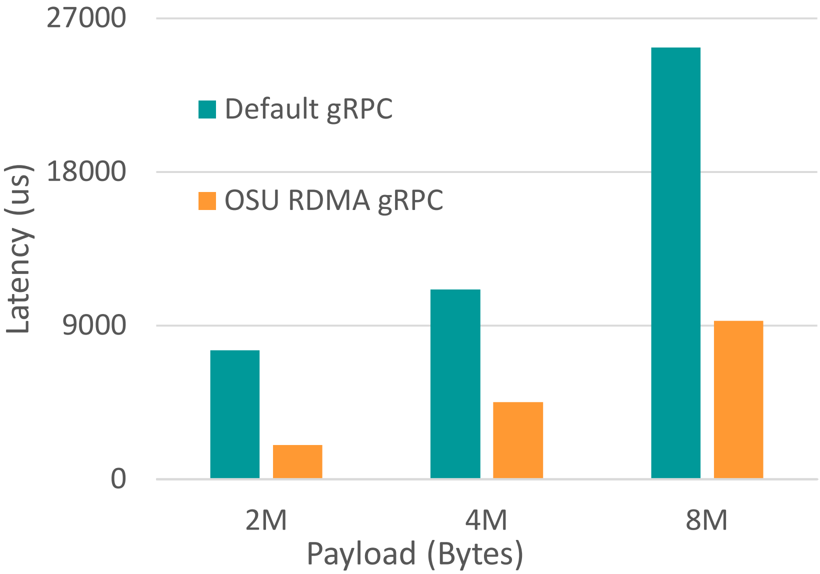
64 Worker Nodes, 1536 cores, SortByTest Total Time

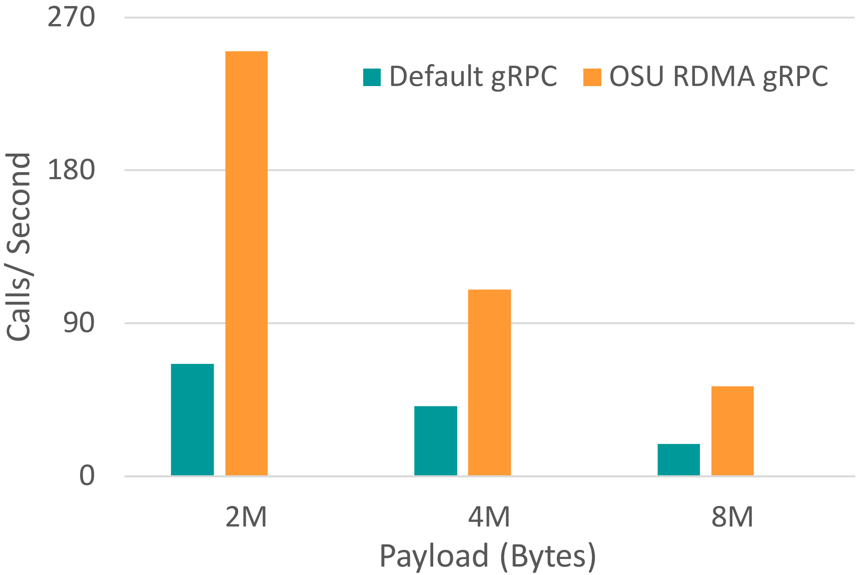


64 Worker Nodes, 1536 cores, GroupByTest Total Time

- InfiniBand FDR, SSD, 64 Worker Nodes, 1536 Cores, (1536M 1536R)
- RDMA 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 74% over IPoIB (56Gbps)

Performance Benefits for RDMA-gRPC with TensorFlow Communication Mimic Benchmark



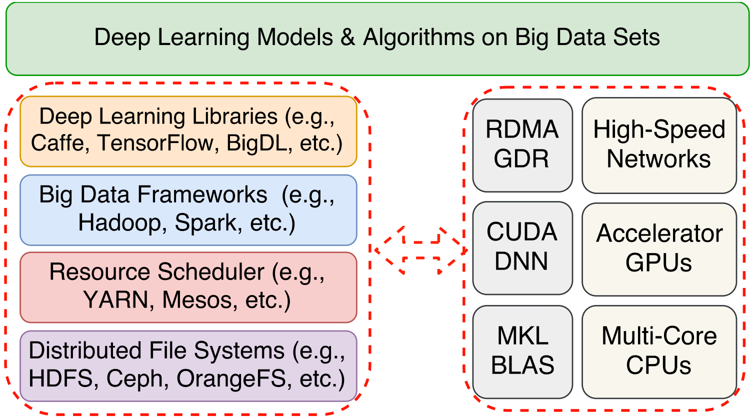


TensorFlow communication Pattern mimic benchmark

- **TensorFlow communication pattern mimic on SDSC-Comet-FDR**
  - Single process spawns both gRPC server and gRPC client
  - **Up to 3.6x** performance speedup over IPoIB for Latency
  - **Up to 3.7x** throughput improvement over IPoIB

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

- Benefits of Deep Learning over Big Data (DLoBD)
  - Easily integrate deep learning components into Big Data processing workflow
  - Easily access the stored data in Big Data systems
  - No need to set up new dedicated deep learning clusters; Reuse existing big data analytics clusters
- Challenges
  - 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



Deep Learning Models & Algorithms on Big Data Sets

Deep Learning Libraries (e.g., Caffe, TensorFlow, BigDL, etc.)

Big Data Frameworks (e.g., Hadoop, Spark, etc.)

Resource Scheduler (e.g., YARN, Mesos, etc.)

Distributed File Systems (e.g., HDFS, Ceph, OrangeFS, etc.)

RDMA GDR

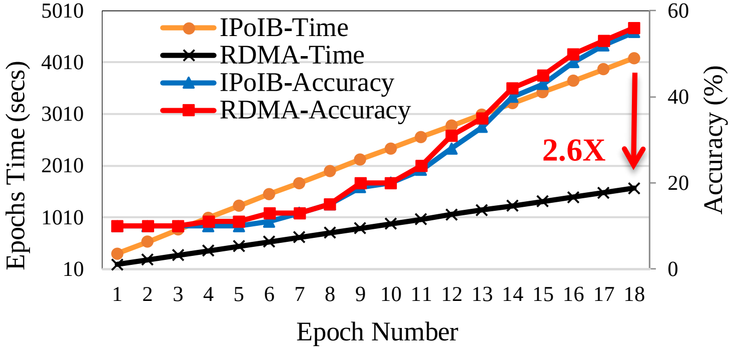
High-Speed Networks

CUDA DNN

Accelerator GPUs

MKL BLAS

Multi-Core CPUs



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.

Concluding Remarks

- Software architecture convergence between HPC and BigData is already happening
- Next Phase: Convergence between

HPC +Big Data + Deep Learning

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

Workshop Date: May 21st, 2018

Keynote Talk: Prof. Geoffrey Fox, Twister2: A High-Performance Big Data Programming Environment

Six Regular Research Papers and Two Short Research Papers

Panel Topic: Which Framework is the Best for High-Performance Deep Learning: Big Data Framework or HPC Framework?

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

HPBDC 2017 was held in conjunction with IPDPS'17

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

HPBDC 2016 was held in conjunction with IPDPS'16

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

Thank You!

panda@cse.ohio-state.edu

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

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/