Benchmarking Kudu Distributed Storage Engine on High-Performance Interconnects and Storage Devices

Nusrat Sharmin Islam, Md. Wasi-ur-Rahman, Xiaoyi Lu,

Dhabaleswar K. (DK) Panda

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

Department of Computer Science and Engineering
The Ohio State University

Outline

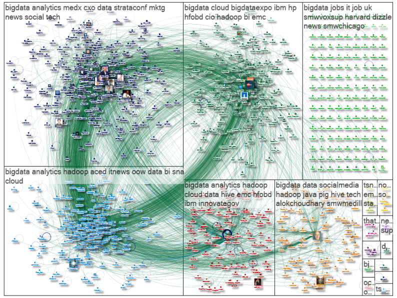
- Introduction
- Contributions
- Design
- Experimental Results
- Conclusion and Future Work

Introduction

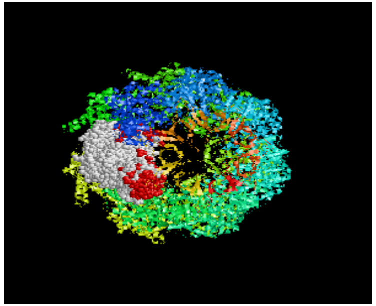
- Big Data provides groundbreaking opportunities for information management and decision making
- The amount of data is exploding; production of data in diverse fields is increasing at an astonishing rate
- IDC claims, digital universe is doubling in size every two years; will multiply 10-fold between 2013 and 2020 [*]
- Not only in internet services, scientific applications in diverse domains like Bioinformatics, Astrophysics, etc. are also dealing with Big Data problems

http://sppider.cchmc.org/sppider_doc.html

https://nodexlgraphgallery.org/Pages/Graph.aspx?graphID=1266



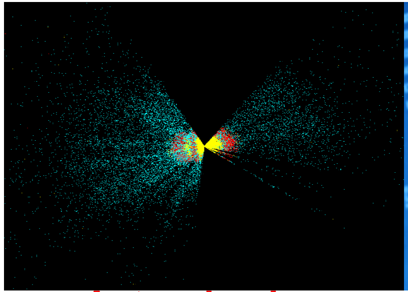
Internet Services



Bioinformatics

http://innovation.talan.fr/en/2015/02/05/trends-all-finance-will-soon-be-big-data-finance-2/

http://complex.elte.hu/astro.html



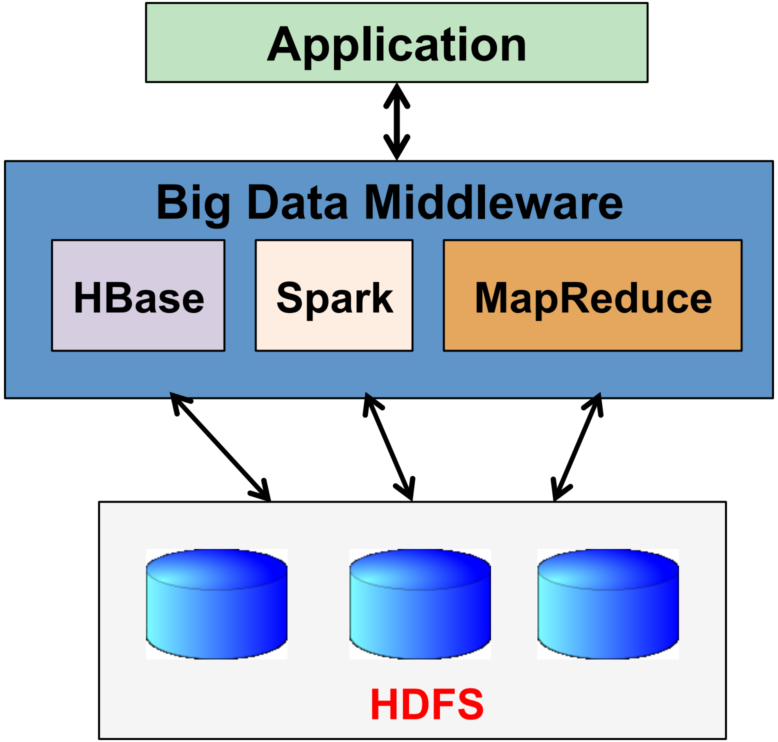
Astrophysics

Finances

[*] http://www.csc. com/insights/flxwd/78931-big data universe beginning to explode

Big Data and Distributed File System

- Hadoop MapReduce and Spark are two popular processing frameworks for Big Data
- Hadoop Distributed File System (HDFS) is the underlying file system of Hadoop, Spark, and Hadoop database HBase
- Adopted by many reputed organizations, e.g. Facebook, Yahoo!
- HDFS, along with the upper-level middleware are being extensively used on HPC clusters
  - Enterprise is also adopting HPC technologies e.g. Oracle [*], Pivotal [#]



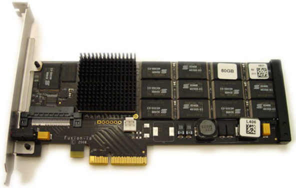
[*] https://www.oracle.com/networking/edr-infiniband-fabric/index.html
[#] http://www.gopivotal.com/solutions/analytics-workbench

Drivers of Modern HPC Cluster Architectures

Multi-core Processors

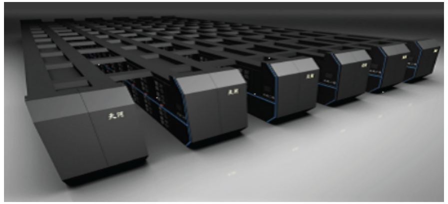
High Performance Interconnects - InfiniBand
<1usec latency, 100Gbps Bandwidth>

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



SSD, NVMe-SSD, NVRAM

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



Tianhe – 2

Titan

Stampede

Gordon

Prior Work

- In [1], an RDMA-based design for HDFS has been proposed
- In [2], a hybrid design (Triple-H) to accelerate HDFS I/O performance with heterogeneous storage and advanced placement policies has been proposed
- In [3], design to accelerate Spark and iterative applications over RDMA-Enhanced HDFS with in-memory and heterogeneous storage has been proposed

[1] 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, SC '12, November 2012

[2] N. S. 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

[3] N. S. 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

Overview of the HiBD Project and Releases

- RDMA for Apache Spark (RDMA-Spark)
- RDMA for Apache HBase (RDMA-HBase)

File System level designs support running Spark and HBase

- **RDMA for Apache Hadoop 2.x (RDMA-Hadoop-2.x)**
  - **Plugins for Apache, Hortonworks (HDP) and Cloudera (CDH) Hadoop distributions**

Installed and available on SDSC Comet

- RDMA for Apache Hadoop 1.x (RDMA-Hadoop)
- RDMA for Memcached (RDMA-Memcached)

Burst buffer Design

- OSU HiBD-Benchmarks (OHB)
- **http://hibd.cse.ohio-state.edu**
- Users Base: 215 organizations from 29 countries
- More than 21,100 downloads from project site

HiBD High-Performance Big Data

Network Based Computing Laboratory

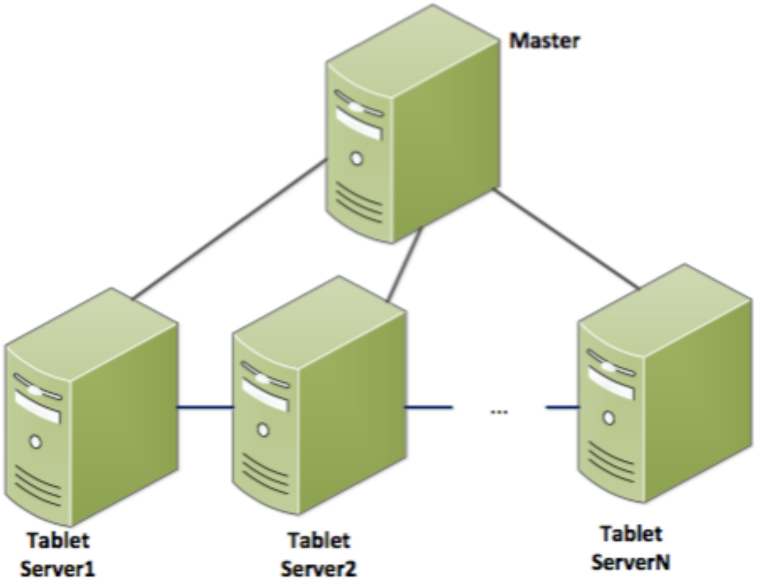
The Ohio State University

Motivation

- Data in HDFS is static; does not support random write or update operation
- Storage engines for fast analytics such as Kudu supports low-latency random access and in-place update
  - Closes the gap between fast sequential write of HDFS and random write of HBase
  - Brings the best of the two storage systems into a single platform
  - Batch as well as OLAP applications can benefit from Kudu

Motivation

- Like HDFS, Kudu replicates data across multiple Tablet servers
  - Data transfer over the network
  - Can high-performance interconnects help?
- Data stored in the local storage devices on the Table servers
  - In Tables, each Table divided into multiple Tablets
  - Can high-performance storage devices help?
- No micro-benchmarks to evaluate the impact of network and storage on Kudu operations



Outline

- Introduction
- Contributions
- Design
- Experimental Results
- Conclusion and Future Work

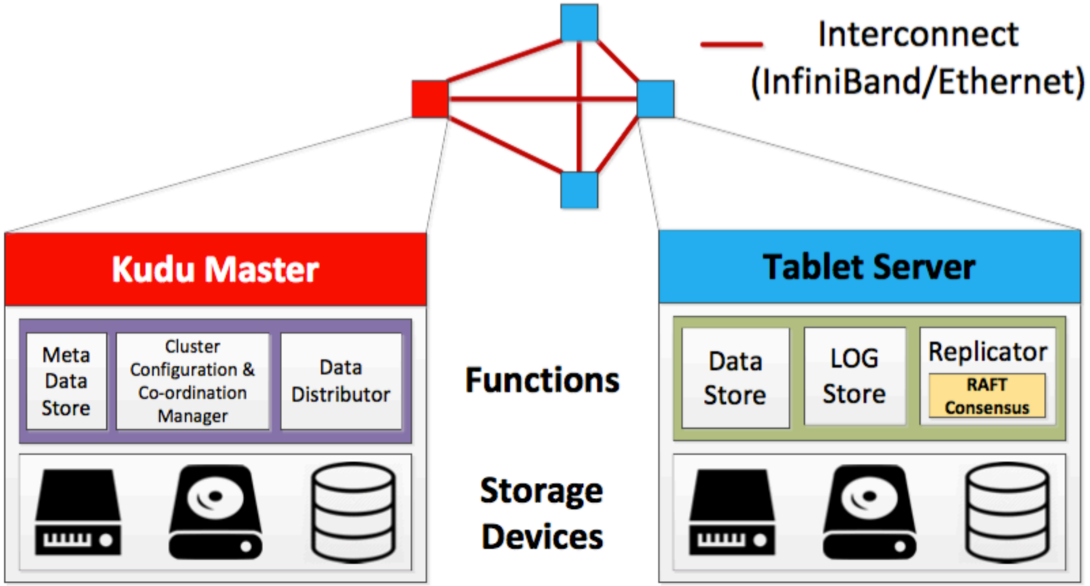
Contributions

- A complete methodology to evaluate Kudu on HPC clusters
- A micro-benchmark for evaluating Kudu (standalone) insert, update, and read operations for both single and multiple clients
- The impact of high-performance network and storage devices on different Kudu operations

Outline

- Introduction
- Contributions
- Design
- Experimental Results
- Conclusion and Future Work

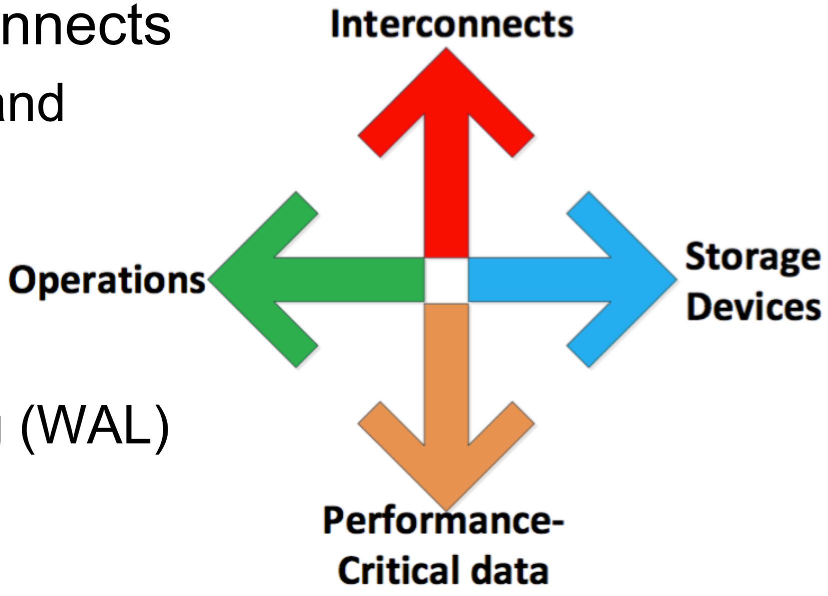
Kudu



- Storage engine for structured data
- Master manages cluster configuration and coordination; stores metadata
- The Tablet serves store the actual data and logs
- Uses RAFT consensus algorithm for data replication

Evaluation Methodology

- Different High-Performance Interconnects
  - 40GigE Ethernet, 100Gbps InfiniBand
- Different types of Storage Devices
  - SSD, HDD
- Performance Critical Data
  - Table data and/or Write Ahead Log (WAL)
- Different Kudu operations
  - Insert, Update, and Read



Micro-benchmark

Supported Operations	Performance Metrics
Insert	Latency, Throughput
Update	Latency, Throughput
Read	Latency, Throughput

- The benchmark supports both single and multi-client modes
- The number of records can be passed as a parameter to the benchmark

Outline

- Introduction
- Problem Statement
- Design
- Experimental Results
- Conclusion and Future Work

Experimental Setup

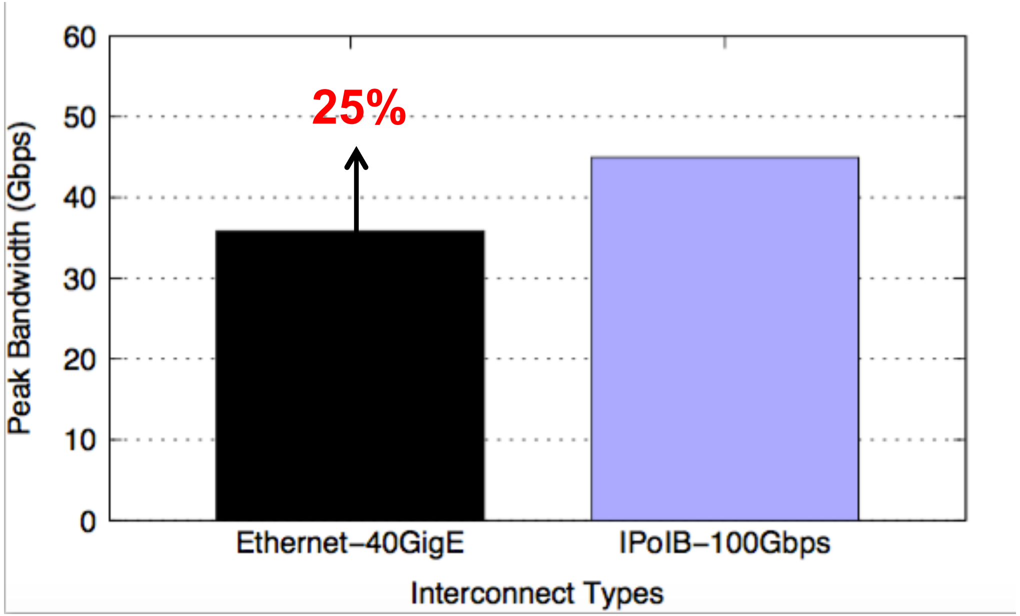
Hardware:

OSU RI2
❖5 nodes (408 cores) ❖Intel Xeon E5-2680 dual 14-core v4 2.4GHz processors ❖40GigE and InfiniBand-EDR (100Gbps) ❖512GB RAM, 380GB NVMe-SSD, two 1TB HDDs

Software:

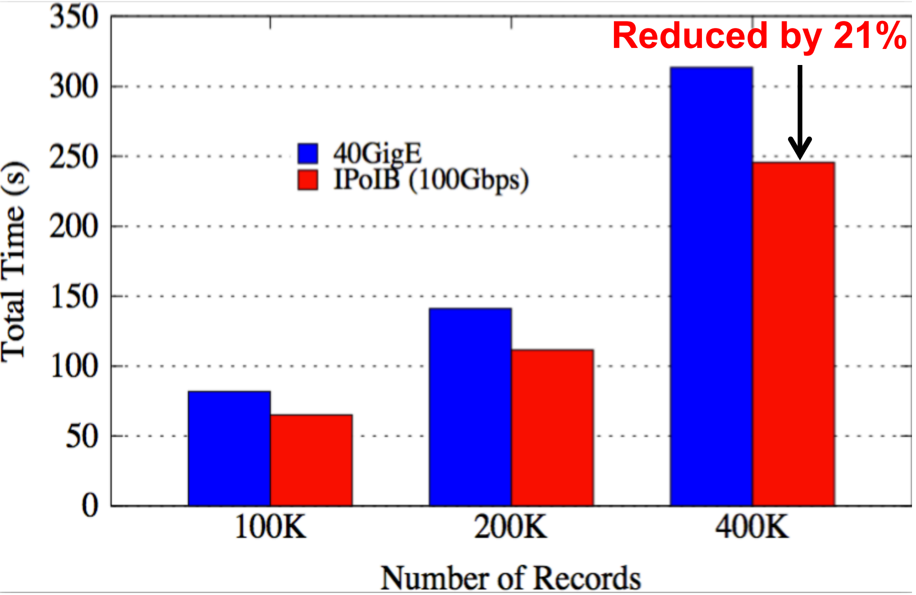
Kudu 1.0.1

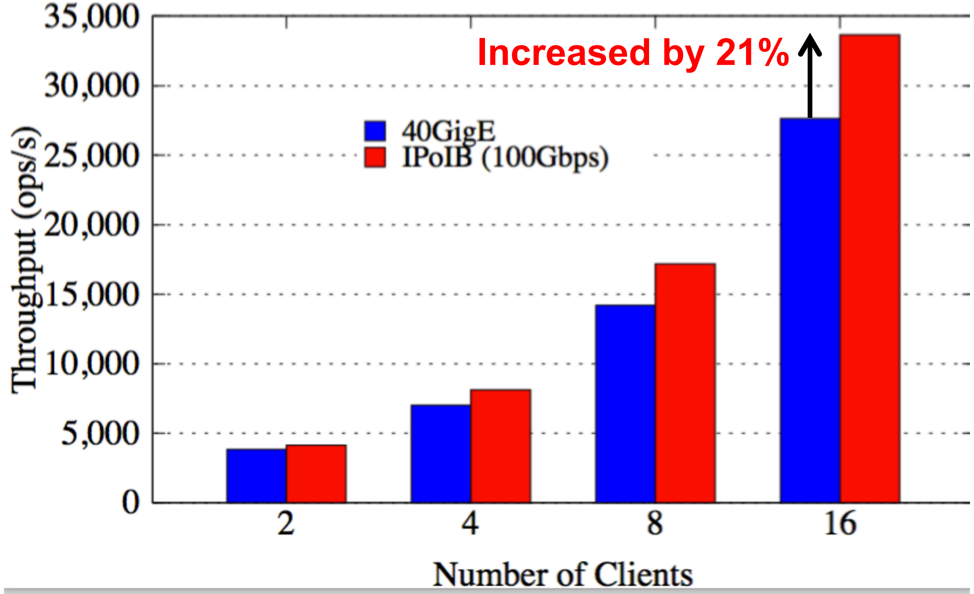
Basic Performance of Interconnects



- Peak Bandwidth
  - *The performance of IPoIB (100Gbps) is **25%** higher than 40GigE Ethernet*

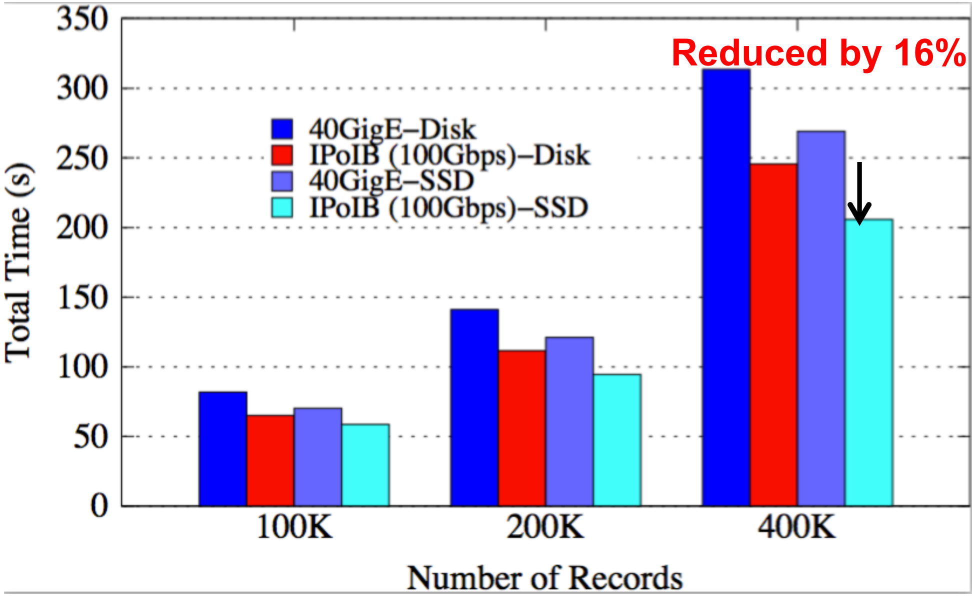
Kudu Insert Operation Performance

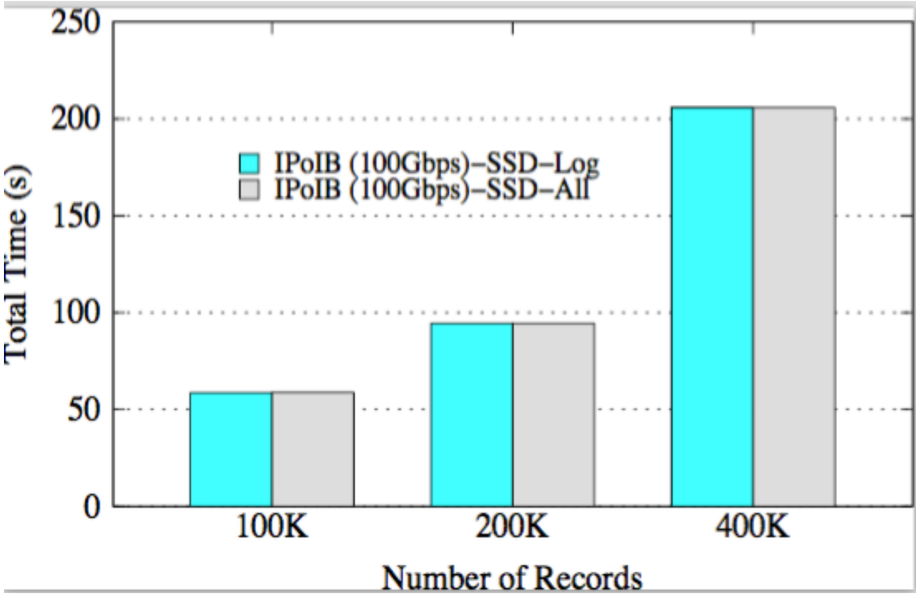




- Insert Latency
  - *IPoIB (100Gbps) is **21%** better than 40GigE Ethernet*
- Insert Throughput
  - *IPoIB (100Gbps) is **21%** better than 40GigE Ethernet*

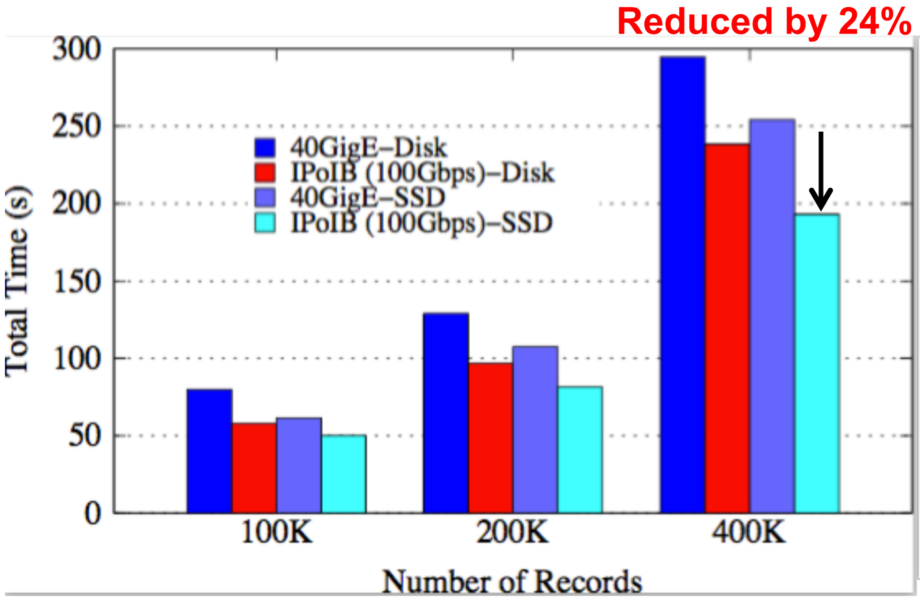
Kudu Insert Operation Performance

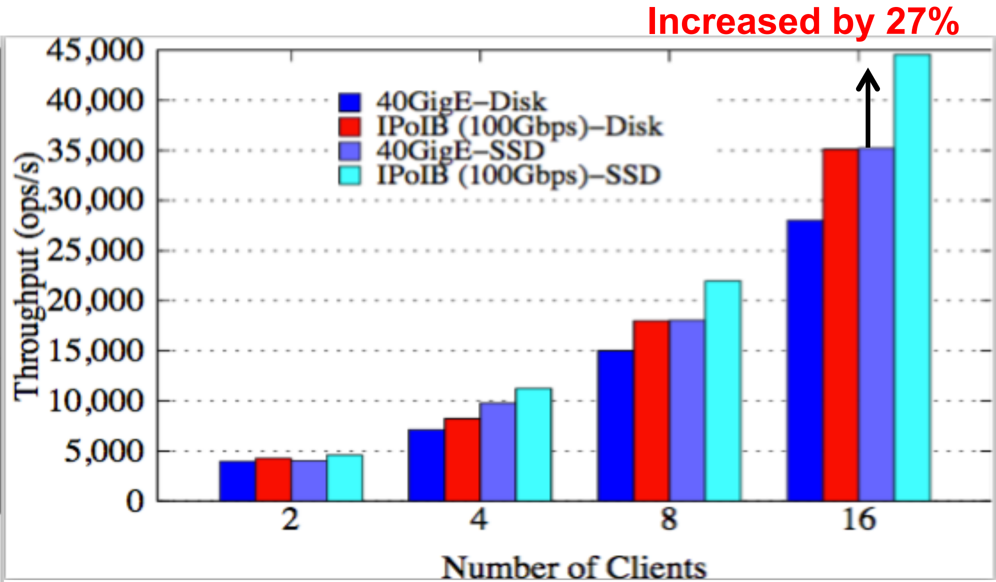




- Insert Latency
  - *SSD reduces the latency by* ***16%*** over IPoIB (100Gbps), **14%** over 40GigE Ethernet
  - *Storing all data to SSD is as good as storing the WALS (only) to SSD*

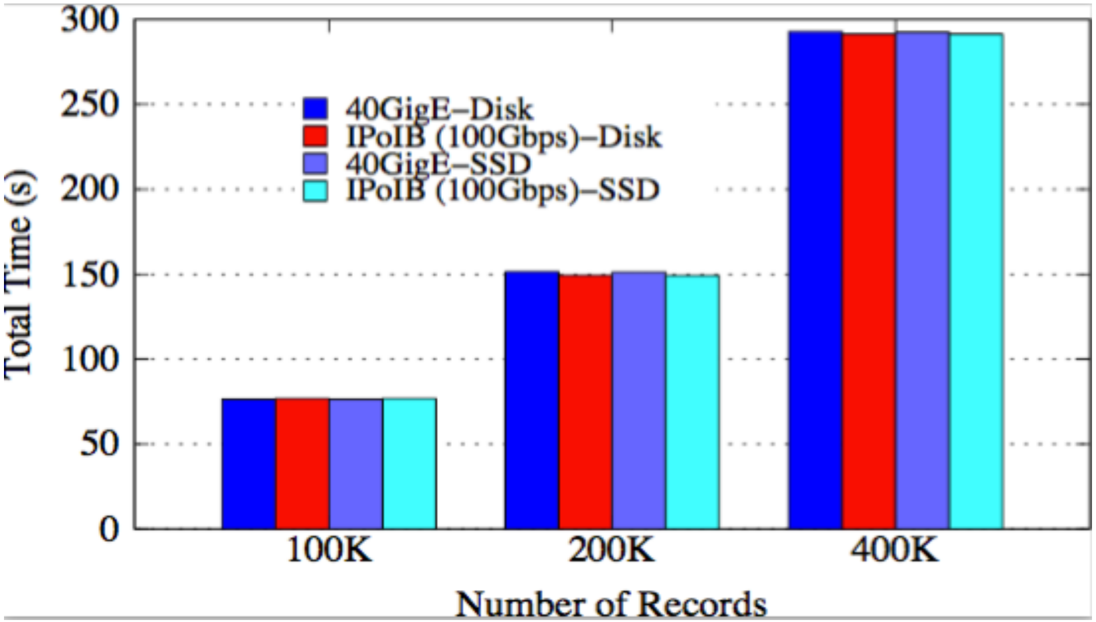
Kudu Update Operation Performance





- Update Latency (WALS on SSD)
  - *IPoIB (100Gbps) is 24% better than 40GigE Ethernet*
- Update Throughput
  - *IPoIB (100Gbps) is 27% better than 40GigE Ethernet*

Kudu Read Operation Performance



- Read Latency
  - Local reads; not much difference over different interconnects
  - In-memory data read; not much difference for SSD

Outline

- Introduction
- Contributions
- Design
- Experimental Results
- Conclusion and Future Work

Conclusion and Future Work

- Proposed an evaluation methodology and a micro-benchmark to evaluate Kudu on HPC platforms
  - High-performance interconnects and storage impact Kudu performance

- RDMA-based design of Kudu
- NVM-aware Kudu

Thank You!

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

Network-Based Computing Laboratory

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

The High-Performance Big Data Project

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