NVM-aware RDMA-based Communication and I/O Schemes for High-Performance Big Data Analytics

Talk at OpenFabrics Alliance Workshop (OFAW '17)

by

Xiaoyi Lu

The Ohio State University

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

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

Dhabaleswar K. (DK) Panda

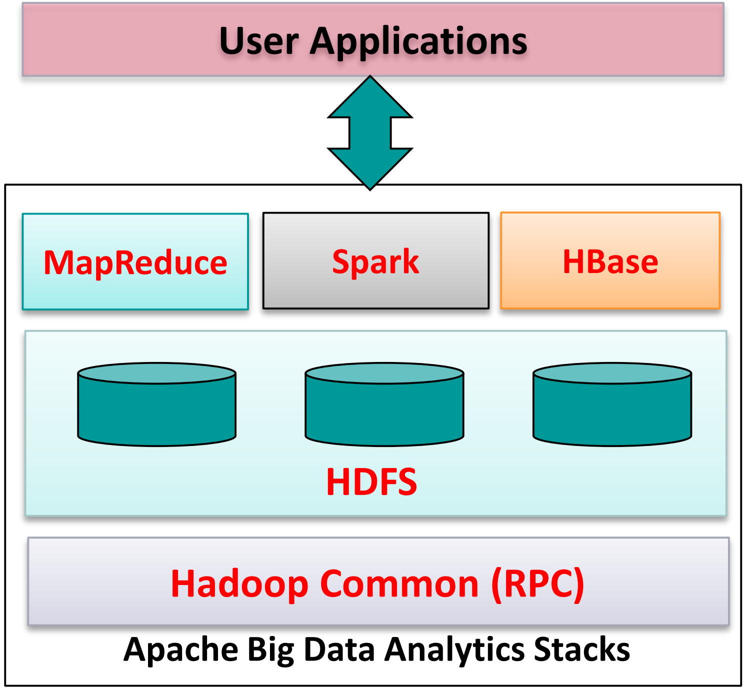
The Ohio State University

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

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

Big Data Processing with Apache Big Data Analytics Stacks

- Major components included:
  - **MapReduce** (Batch)
  - Spark (Iterative and Interactive)
  - HBase (Query)
  - **HDFS** (Storage)
  - RPC (Inter-process communication)
- Underlying Hadoop Distributed File System (HDFS) used by MapReduce, Spark, HBase, and many others
- Model scales but high amount of communication and I/O can be further optimized!



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
- Accelerators (NVIDIA GPGPUs and Intel Xeon Phi)
- Remote Direct Memory Access (RDMA)-enabled networking (InfiniBand and RoCE)
- NVRAM, SSD, Parallel Filesystems, Object Storage Clusters

SDSC Comet

TACC Stampede

Microsoft Azure

amazon web services EC2

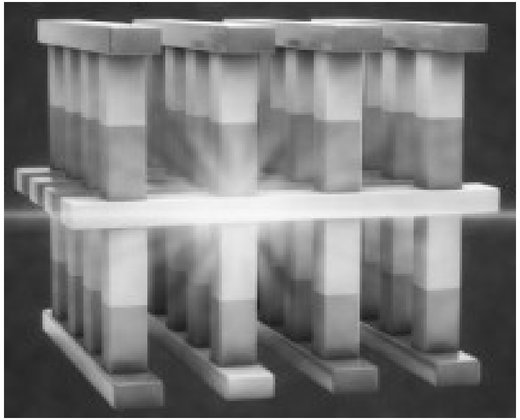
ORACLE Cloud

Chameleon Cloud

Presentation Outline

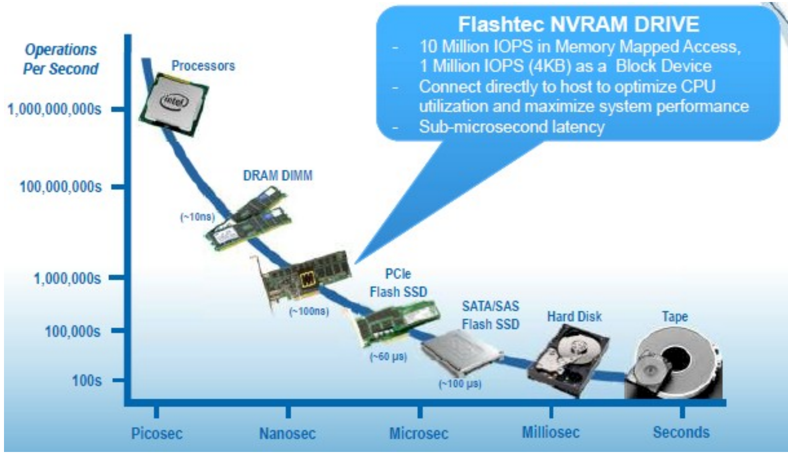
- Understanding NVRAM and RDMA
- NRCIO: NVM-aware RDMA-based Communication and I/O Schemes
- NRCIO for Big Data Analytics
- Conclusion and Q&A

Non-Volatile Memory (NVM) and NVMe-SSD



3D XPoint from Intel & Micron

Samsung NVMe SSD

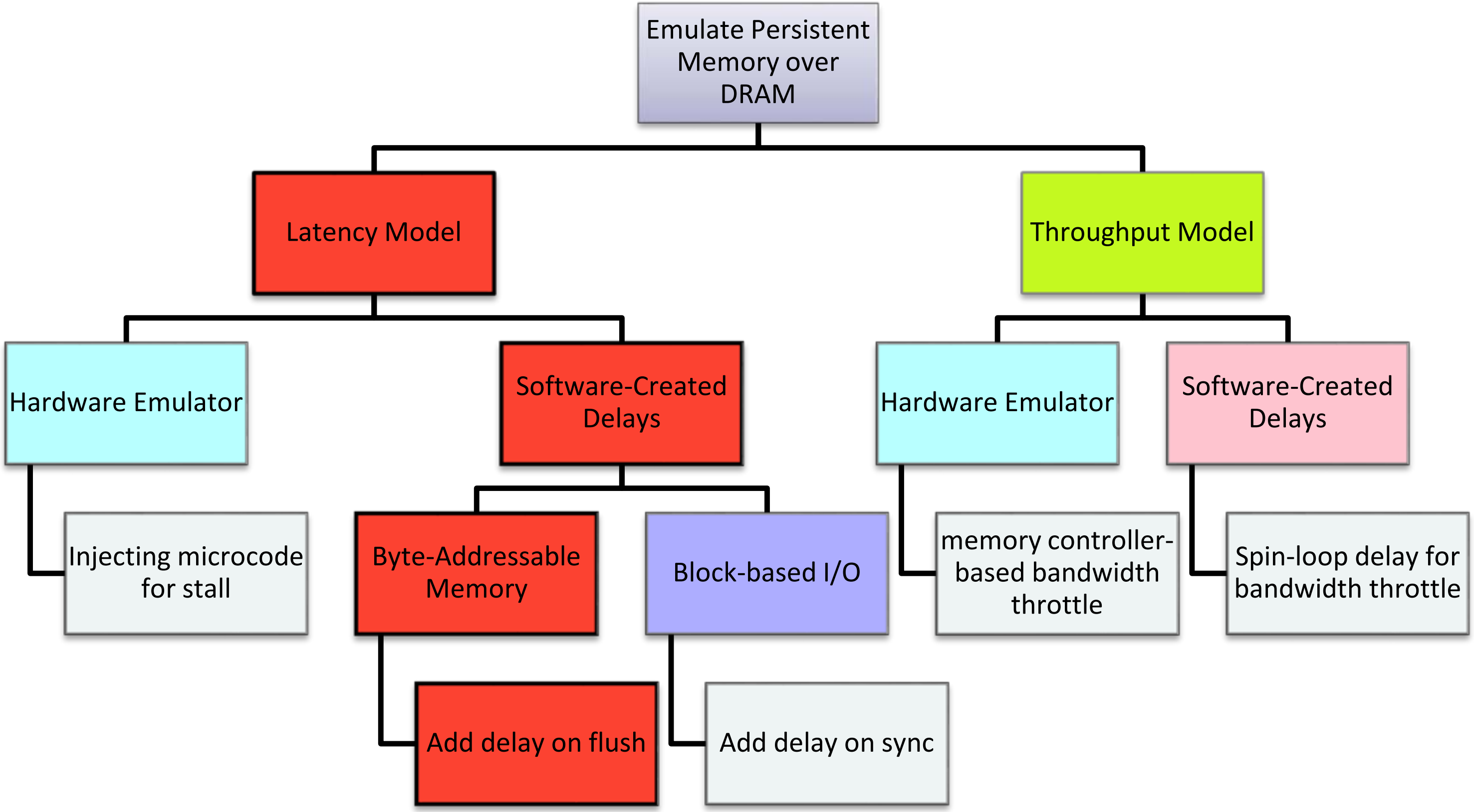


Performance of PMC Flashtec NVRAM [*]

- Non-Volatile Memory (NVM) provides byte-addressability with persistence
- The huge explosion of data in diverse fields require fast analysis and storage
- NVMs provide the opportunity to build high-throughput storage systems for data-intensive applications
- Storage technology is moving rapidly towards NVM

[*] http://www.enterprisetech.com/2014/08/06/ flashtec-nvram-15-million-iops-sub-microsecond- latency/

NVRAM Emulation Techniques



Software-based NVRAM Emulation

- Latency Model
  - slower memory writes
    PCM : **clflush + delay**
    PCM-Disk : **msync + delay**
  - Delays inserted using a spin-loop (RDTSCP) or NOPS
  - E.g., Mnemosyne[1] Library, PCMSIM[2], etc.

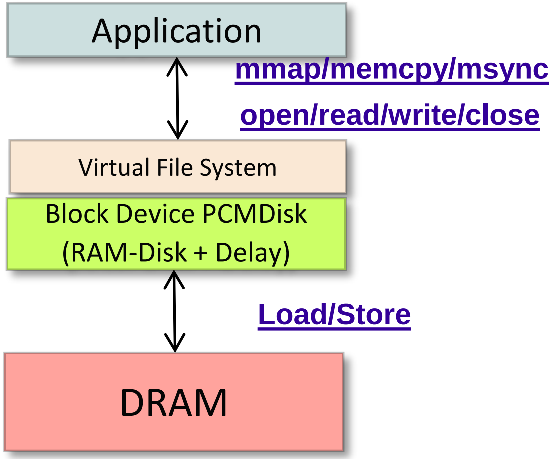
- Throughput Model
  - Throttle bandwidth by inserting delays

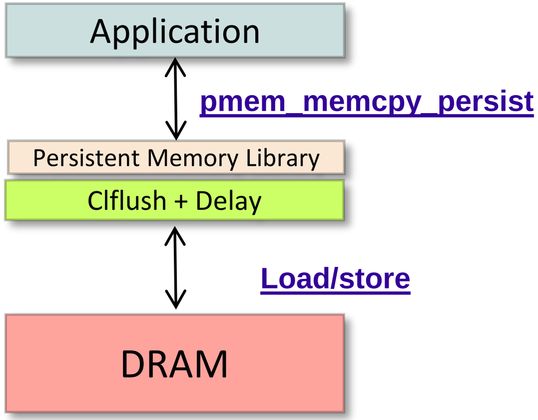
$$Latency = size * (1 - (PCM\_BW / DRAM\_BW)) / PCM\_BW$$

  - Delays inserted using a spin-loop (RDTSCP) or NOPS
  - E.g., Mnemosyne PCMDisk[1]

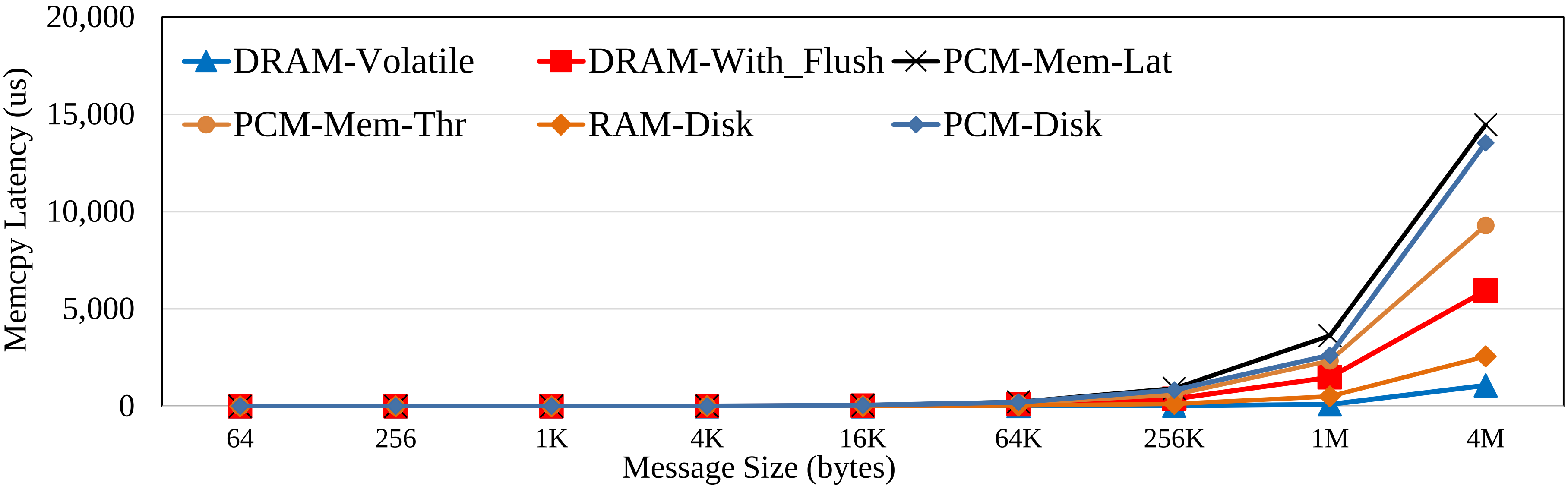
NVRAM Emulation based on DRAM

- Popular methods employed by recent works to emulate NVRAM performance model over DRAM
- Two ways:
  - Emulate byte-addressable NVRAM over DRAM
  - Emulate block-based NVM device over DRAM





Performance Evaluation of memcpy over PCM



- Latency Comparison of memcpy operations with NVRAM Latency and Throughput emulation models for PCM (PCM-write-delay = 150 ns and PCM-read-delay = 50 ns)
- PCM-Mem (pmem.io NVML library over DRAM + delay) vs. Mnemosyne PCM-Disk
- PCM-Disk/PCM-Mem-Lat models show >6x overhead over RAMDisk with sync

High-Performance Interconnects and Protocols

- High-Performance Computing (HPC) has adopted advanced interconnects and protocols
  - InfiniBand
  - 10/40/100 Gigabit Ethernet/iWARP
  - RDMA over Converged Enhanced Ethernet (RoCE)
- Very Good Performance
  - Low latency (few micro seconds)
  - High Bandwidth (100 Gb/s with EDR InfiniBand)
  - Low CPU overhead (5-10%)
- OpenFabrics software stack with IB, iWARP and RoCE interfaces are driving HPC systems

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: 215 organizations from 29 countries
- More than 21,000 downloads from the project site

Available for InfiniBand and RoCE

Significant performance improvement with 'RDMA+DRAM' compared to default Sockets-based designs;
How about RDMA+NVRAM?

HiBD High-Performance Big Data

Network Based Computing Laboratory

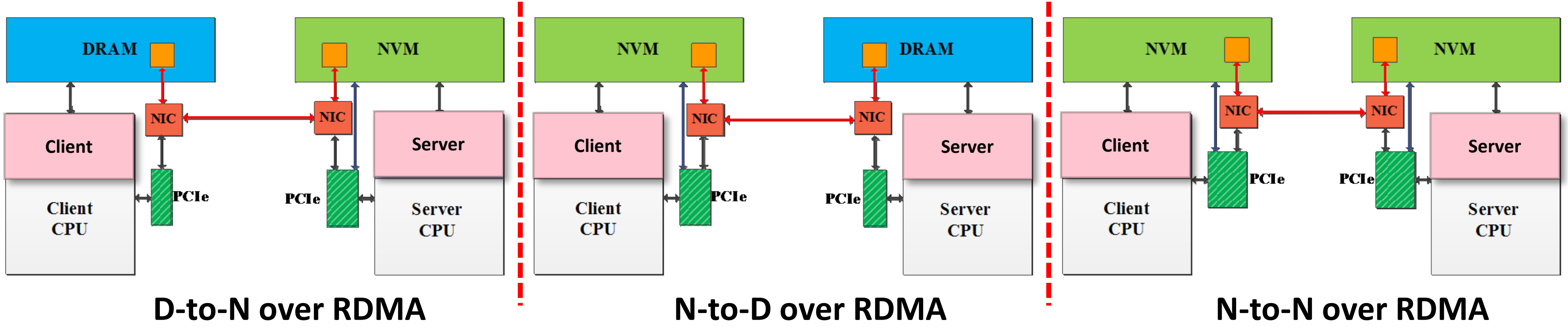
THE OHIO STATE UNIVERSITY

Presentation Outline

- Understanding NVRAM and RDMA
- NRCIO: NVM-aware RDMA-based Communication and I/O Schemes
- NRCIO for Big Data Analytics
- Conclusion and Q&A

Design Scope (NVM for RDMA)

D-to-D over RDMA: Communication buffers for client and server are allocated in DRAM (Common)



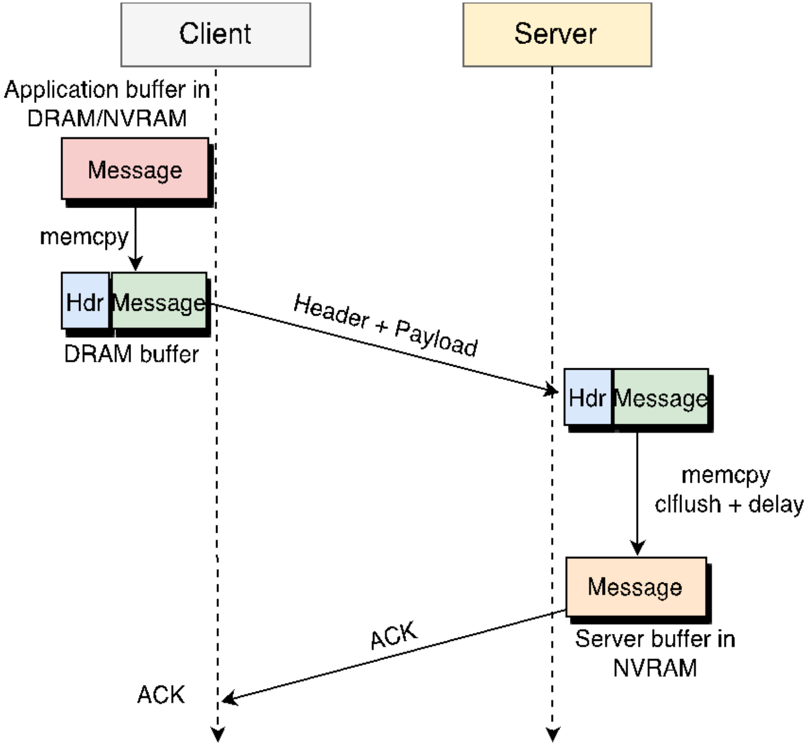
D-to-N over RDMA: Communication buffers for client are allocated in DRAM; Server uses NVM

N-to-D over RDMA: Communication buffers for client are allocated in NVM; Server uses DRAM

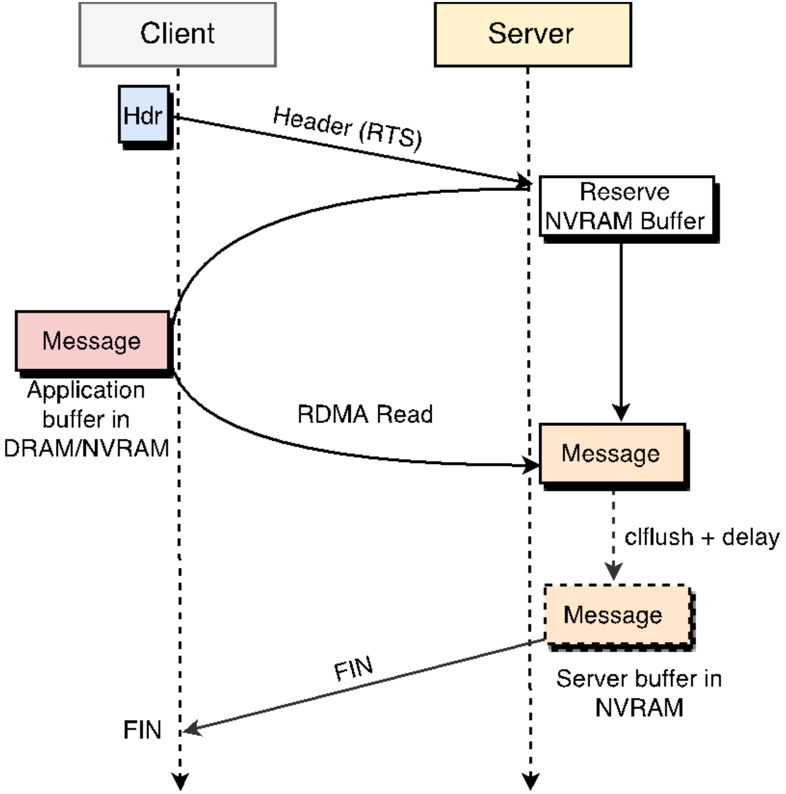
N-to-N over RDMA: Communication buffers for client and server are allocated in NVM

NVRAM-aware Communication in NRCIO

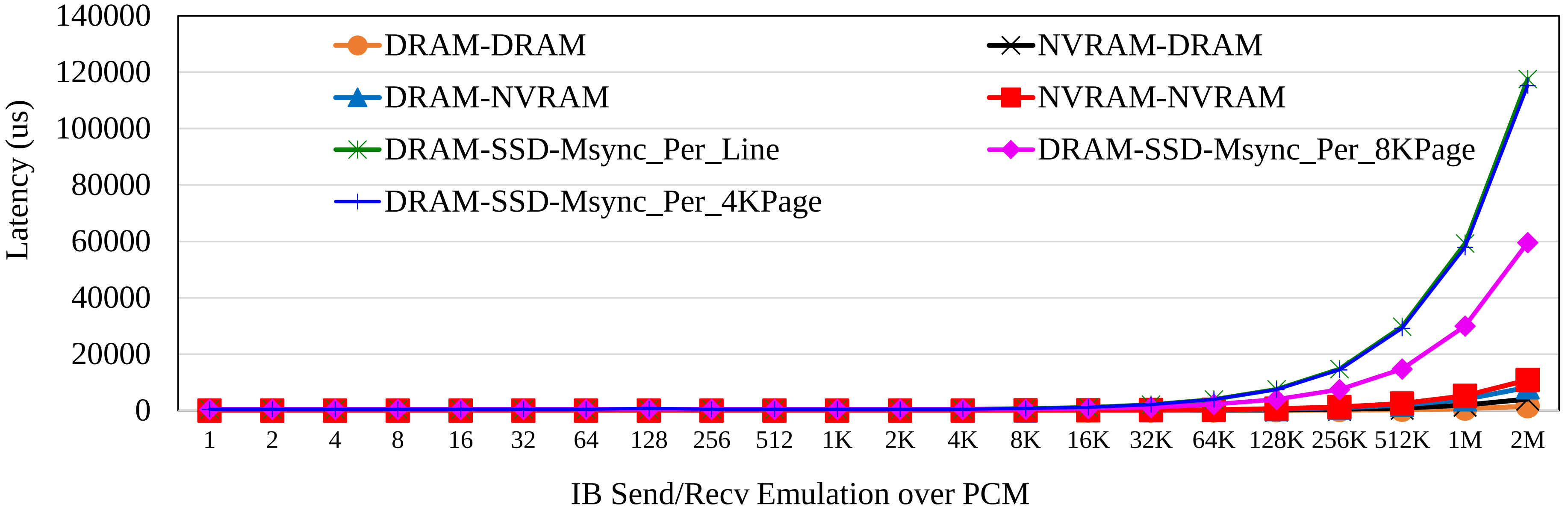
NRCIO Send/Recv over NVRAM



NRCIO RDMA_Read over NVRAM

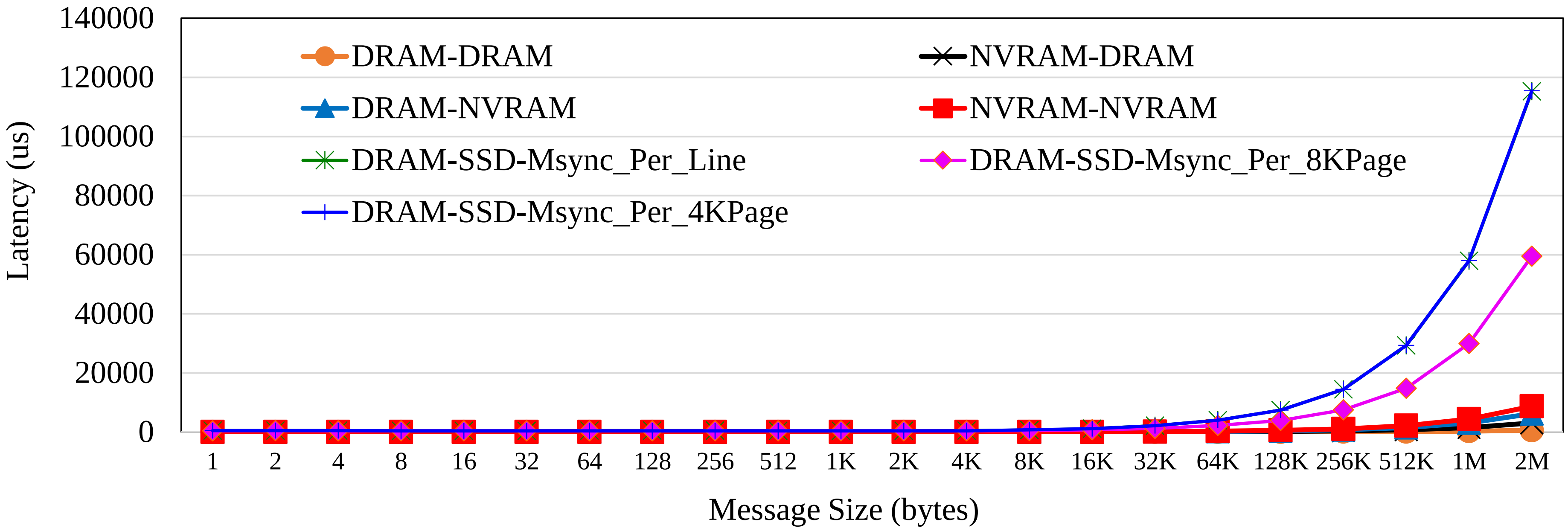


NRCIO Send/Recv Emulation over PCM



- Comparison of communication latency using NRCIO send/receive semantics over InfiniBand QDR network and PCM memory
- High communication latencies due to slower writes to non-volatile persistent memory
  - NVRAM-to-Remote-NVRAM (NVRAM-NVRAM) => ~10x overhead vs. DRAM-DRAM
  - DRAM-to-Remote-NVRAM (DRAM-NVRAM) => ~8x overhead vs. DRAM-DRAM
  - DRAM-to-Remote-NVRAM (DRAM-NVRAM) => ~4x overhead vs. DRAM-DRAM

NRCIO RDMA-Read Emulation over PCM



IB RDMA_READ Emulation over PCM

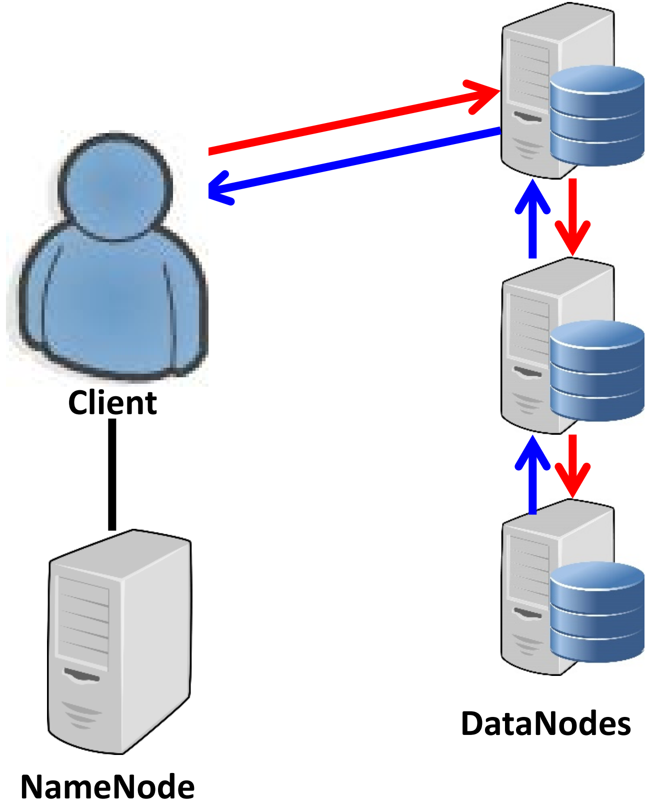
- Communication latency with NRCIO RDMA-Read over InfiniBand QDR + PCM memory
- Communication overheads for large messages due to slower writes into NVRAM from remote memory; similar to Send/Receive
- RDMA-Read outperforms Send/Receive for large messages; as observed for DRAM-DRAM

Presentation Outline

- Understanding NVRAM and RDMA
- NRCIO: NVM-aware RDMA-based Communication and I/O Schemes
- NRCIO for Big Data Analytics
- Conclusion and Q&A

Opportunities of Using NVRAM+RDMA in HDFS

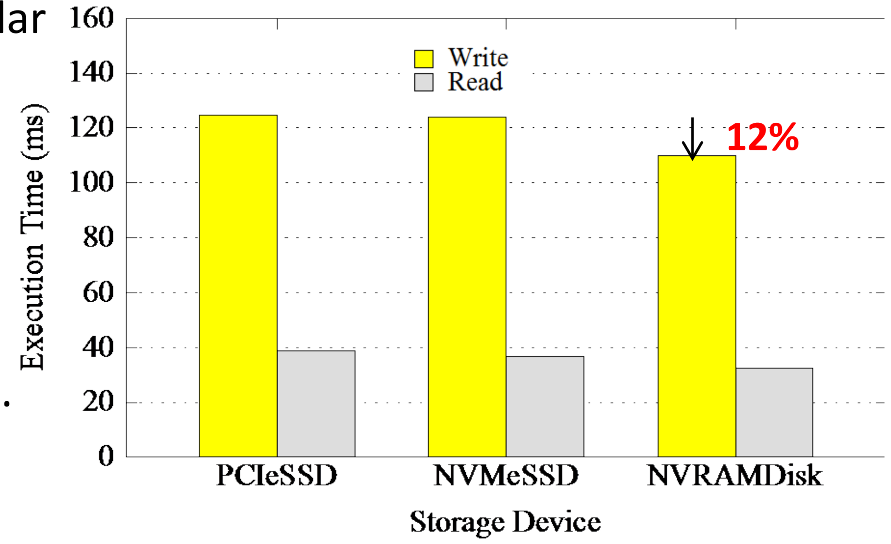
- Files are divided into fixed sized blocks
  - Blocks divided into packets
- NameNode: stores the file system namespace
- DataNode: stores data blocks in local storage devices
- Uses block replication for fault tolerance
  - Replication enhances data-locality and read throughput
- Communication and I/O intensive
- Java Sockets based communication
- Data needs to be persistent, typically on SSD/HDD



Can HDFS be benefited by fully exploiting the byte-addressability of NVM?

- In-memory storage in HDFS is becoming popular
  - Data persistence is challenging
- HDFS stores files in local storage
  - Competition for physical memory
- HPC clusters usually equipped with high performance interconnects and protocols (e.g. RDMA) and parallel file systems like Lustre
- NVM is also emerging and making its way into HPC systems
  - Non-volatile and byte-addressable

NVMs becoming popular in NVMe SSD

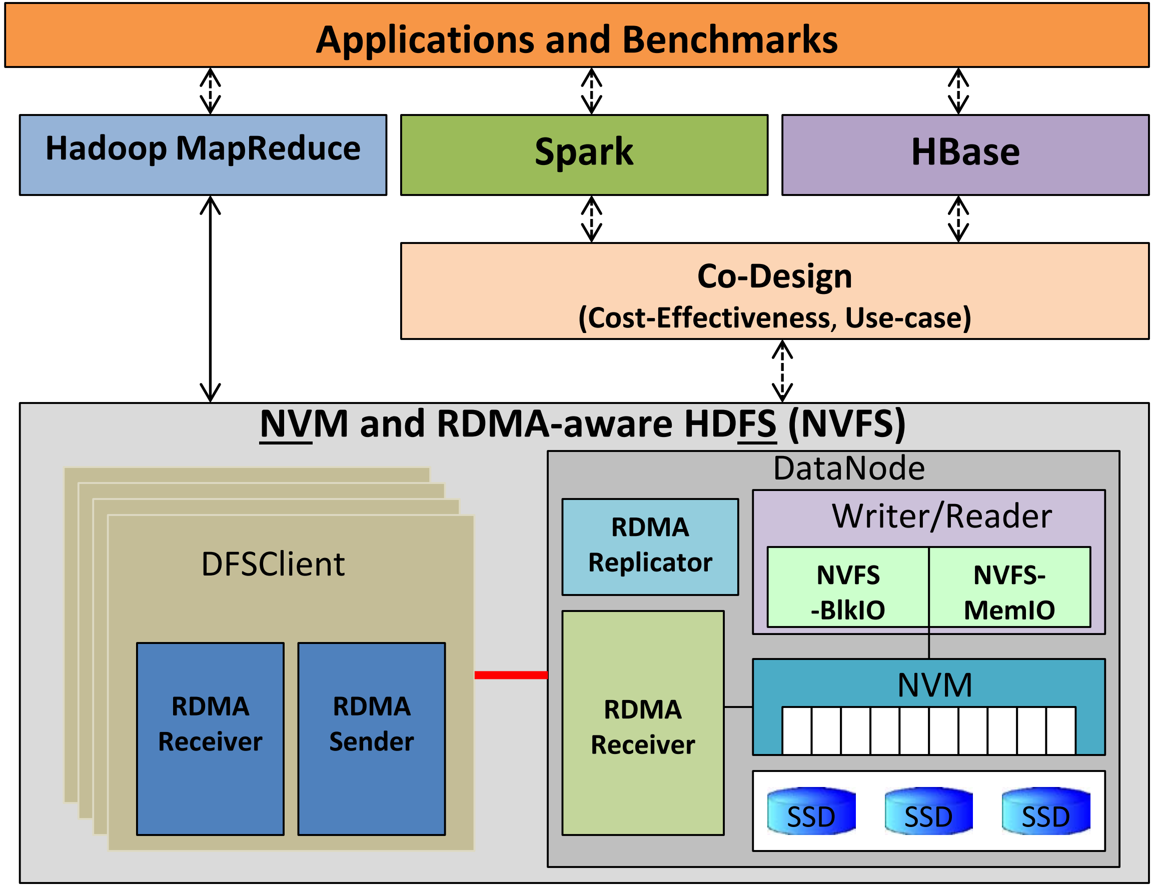


HDFS block I/O time

NVRAMDisk = RAMDisk backed by NVM

Requires re-assessment of the design choices for HDFS!

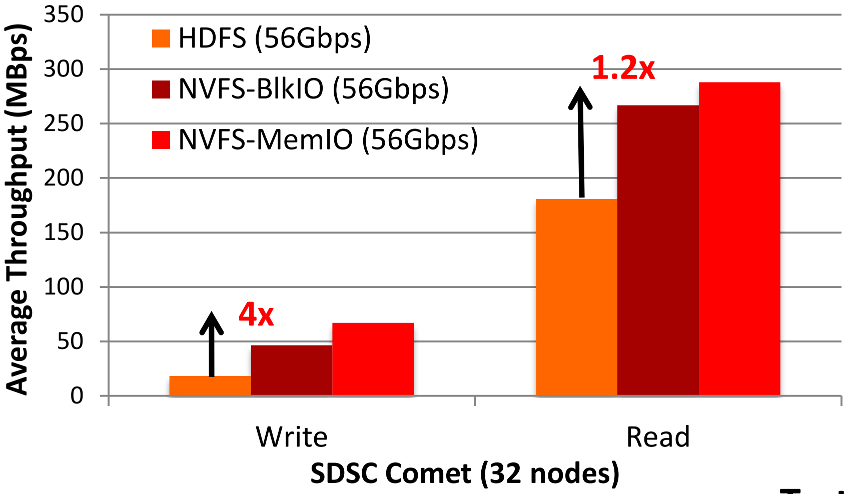
Design Overview of NVM and RDMA-aware HDFS (NVFS)

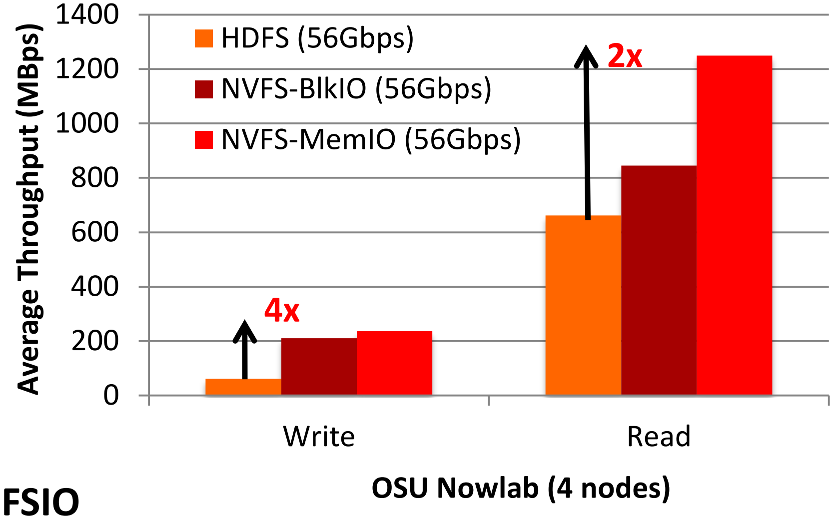


- **Design Features**
  - RDMA over NVM
  - HDFS I/O with NVM
    - Block Access
    - Memory Access
  - Hybrid design
    - NVM with SSD as a hybrid storage for HDFS I/O
  - Co-Design with Spark and HBase
    - Cost-effectiveness
    - Use-case

N. S. Islam, M. W. Rahman , X. Lu, and D. K. Panda, High Performance Design for HDFS with Byte-Addressability of NVM and RDMA, 24th International Conference on Supercomputing (ICS), June 2016

Evaluation with Hadoop MapReduce

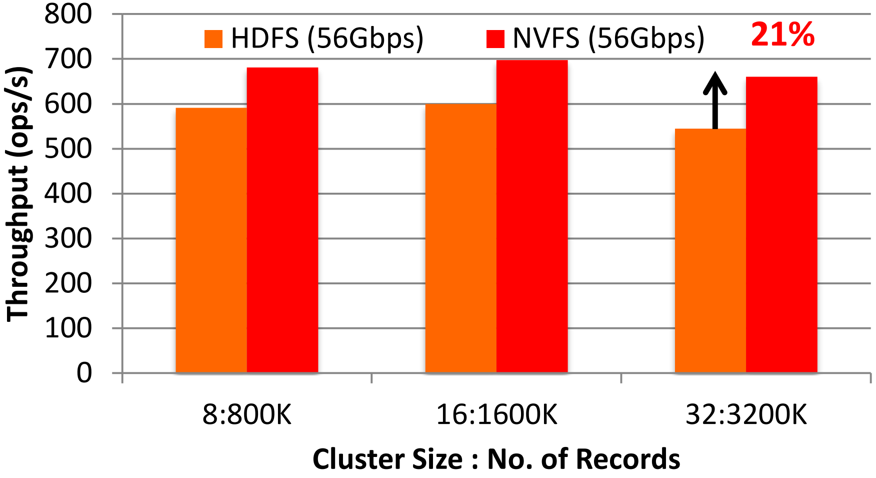




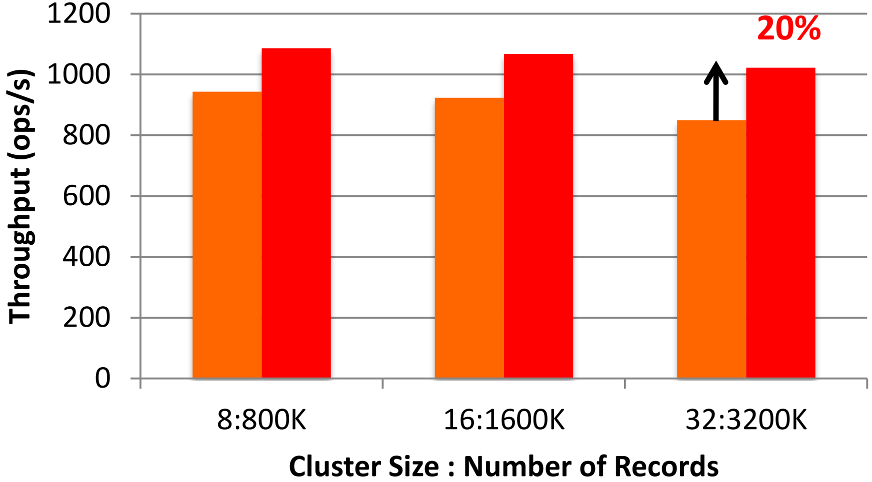
TestDFSIO

- TestDFSIO on SDSC Comet (32 nodes)
  - Write: NVFS-MemIO gains by **4x** over HDFS
  - Read: NVFS-MemIO gains by **1.2x** over HDFS
- TestDFSIO on OSU Nowlab (4 nodes)
  - Write: NVFS-MemIO gains by **4x** over HDFS
  - Read: NVFS-MemIO gains by **2x** over HDFS

Evaluation with HBase



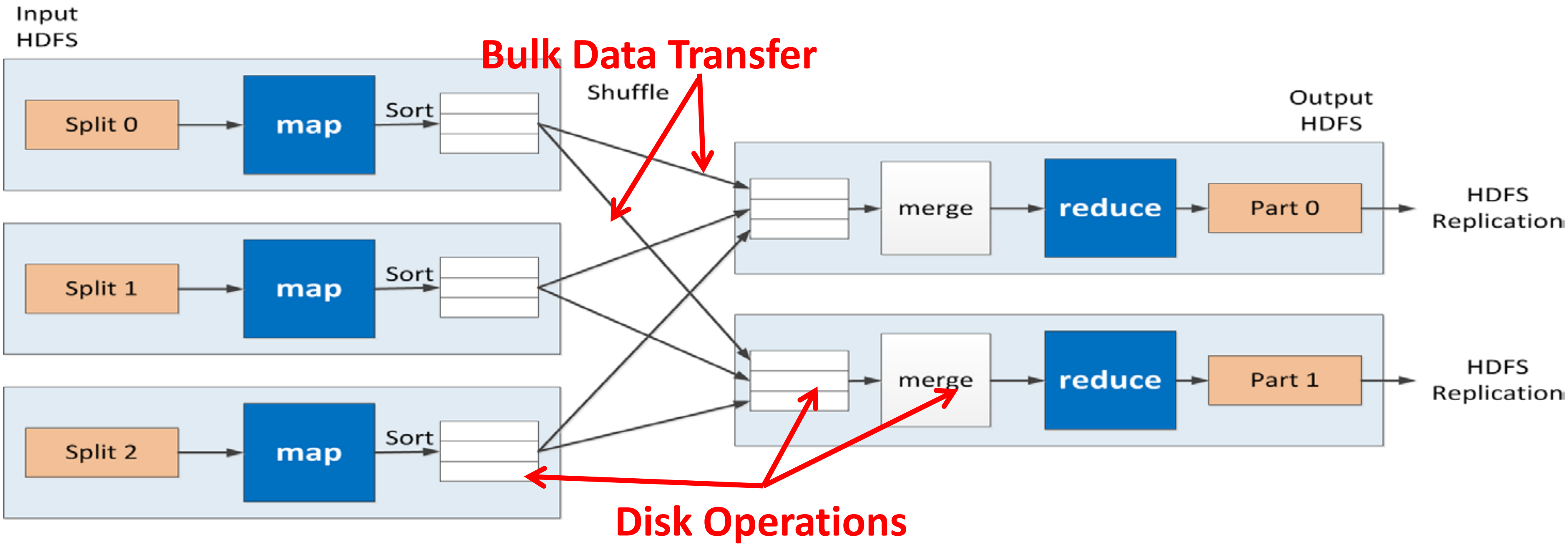
HBase 100% insert



HBase 50% read, 50% update

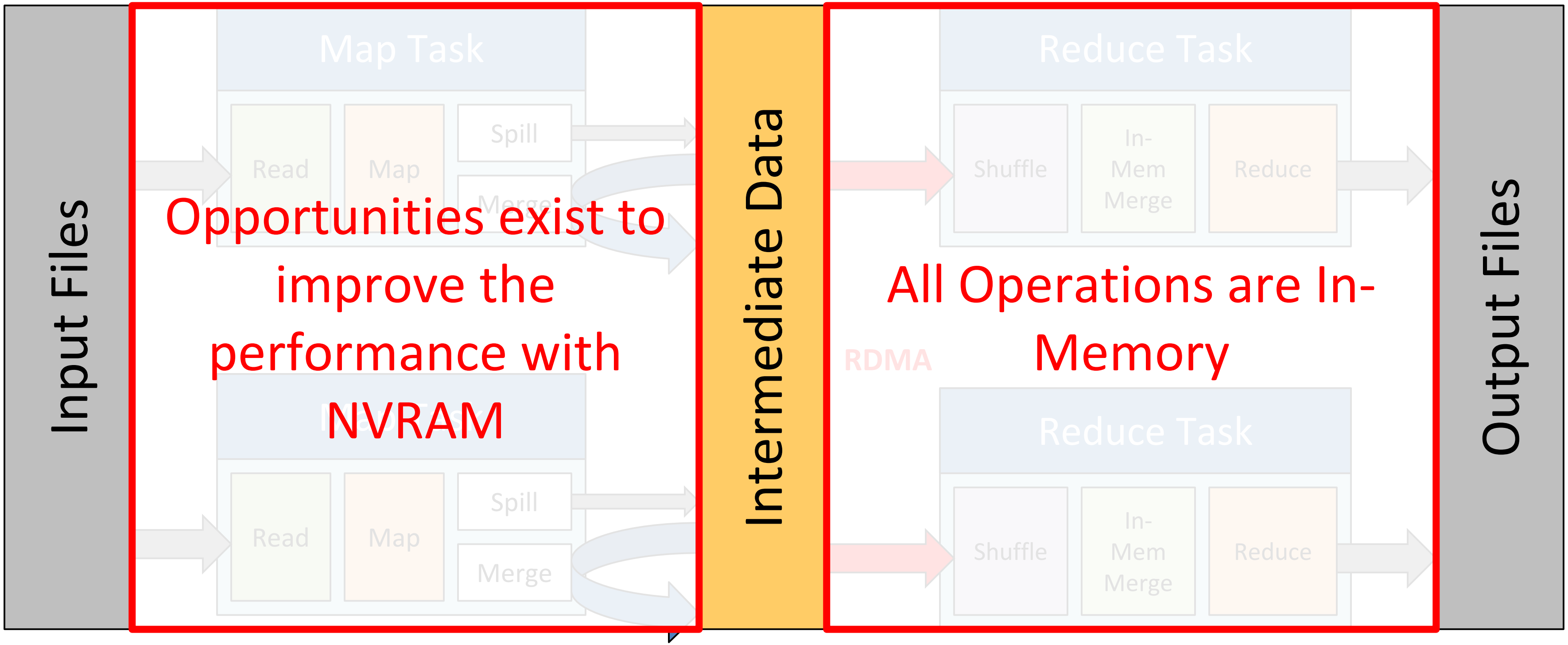
- YCSB 100% Insert on SDSC Comet (32 nodes)
  - NVFS-BlkIO gains by **21%** by storing only WALs to NVM
- YCSB 50% Read, 50% Update on SDSC Comet (32 nodes)
  - NVFS-BlkIO gains by **20%** by storing only WALs to NVM

Opportunities to Use NVRAM+RDMA in MapReduce

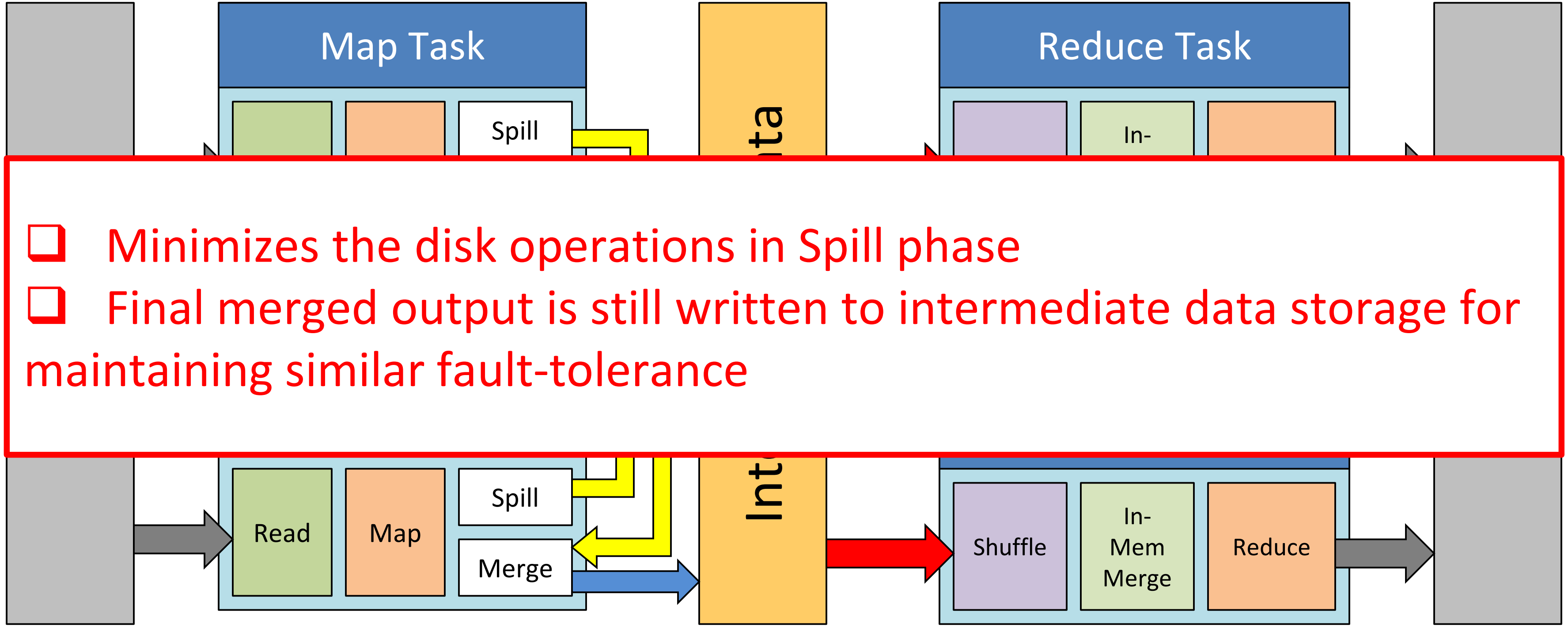


- Map and Reduce Tasks carry out the total job execution
  - Map tasks read from HDFS, operate on it, and write the intermediate data to local disk (persistent)
  - Reduce tasks get these data by shuffle from NodeManagers, operate on it and write to HDFS (persistent)
- Communication and I/O intensive; Shuffle phase uses HTTP over Java Sockets; I/O operations take place in SSD/HDD typically

Opportunities to Use NVRAM in MapReduce-RDMA Design



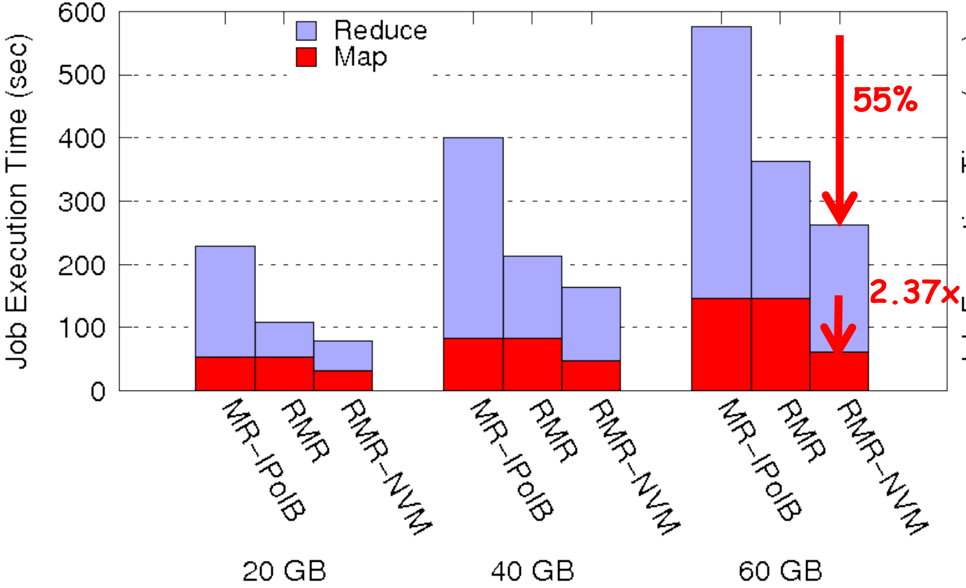
NVRAM-Assisted Map Spilling in MapReduce-RDMA

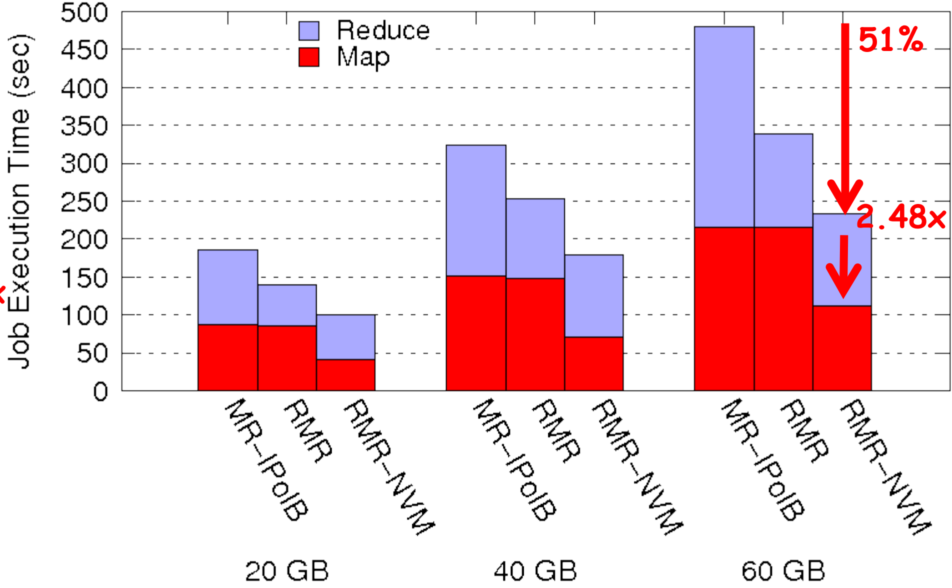


- ❑ Minimizes the disk operations in Spill phase
- ❑ Final merged output is still written to intermediate data storage for maintaining similar fault-tolerance

M. W. Rahman, N. S. Islam, X. Lu, and D. K. Panda, Can Non-Volatile Memory Benefit MapReduce Applications on HPC Clusters? PDSW-DISCS, with SC 2016.

Comparison with Sort and TeraSort

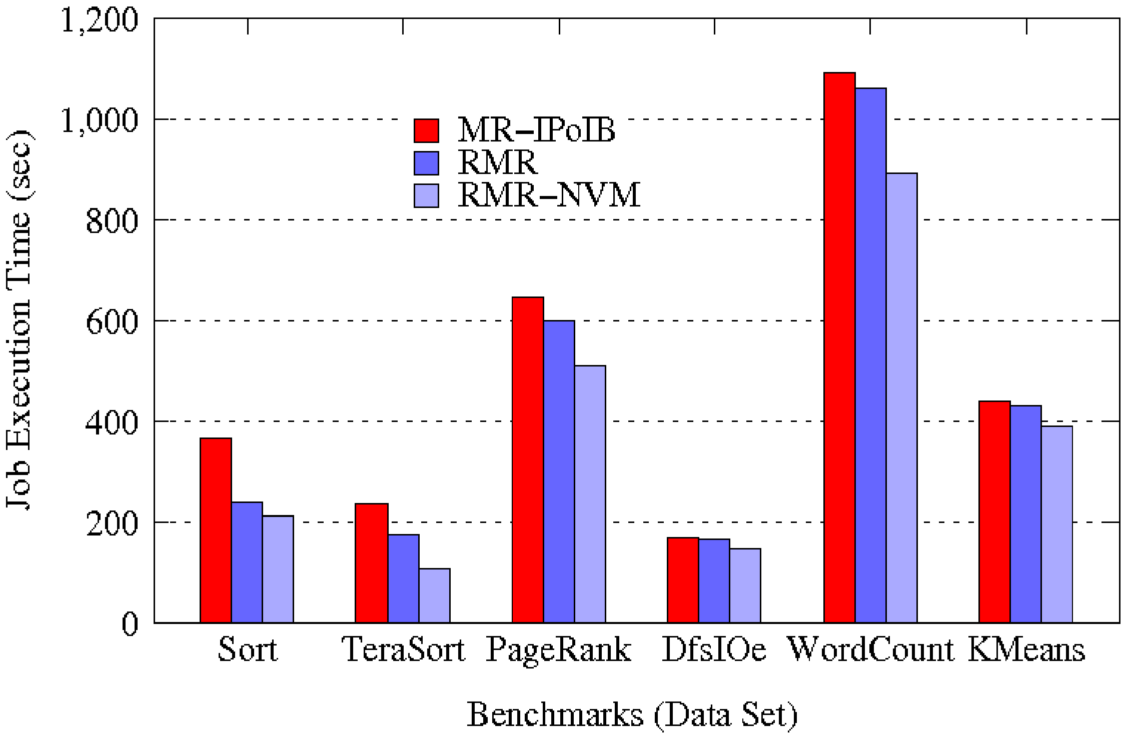




- RMR-NVM achieves **2.37x** benefit for Map phase compared to RMR and MR-IPoIB; overall benefit **55%** compared to MR-IPoIB, **28%** compared to RMR
- RMR-NVM achieves **2.48x** benefit for Map phase compared to RMR and MR-IPoIB; overall benefit **51%** compared to MR-IPoIB, **31%** compared to RMR

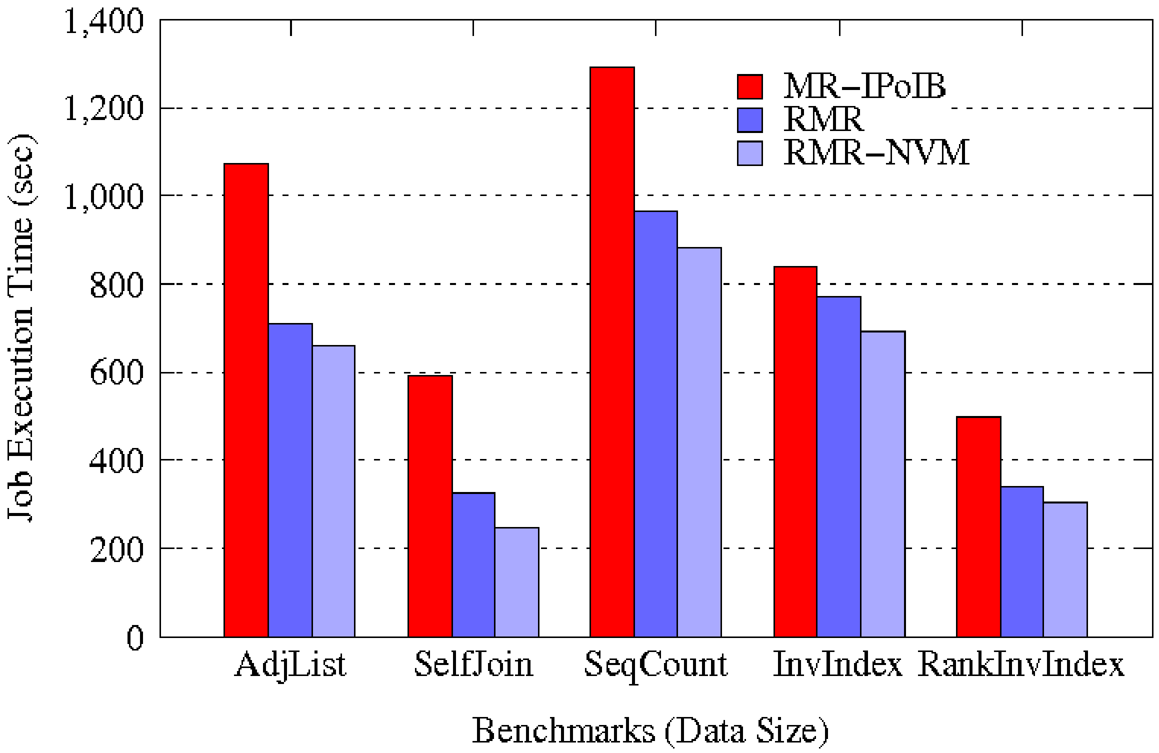
Evaluation of Intel HiBench Workloads

- We evaluate different HiBench workloads with Huge data sets on 8 nodes
- Performance benefits for Shuffle-intensive workloads compared to MR-IPoIB:
  - Sort: **42%** (25 GB)
  - TeraSort: **39%** (32 GB)
  - PageRank: **21%** (5 million pages)
- Other workloads:
  - WordCount: **18%** (25 GB)
  - KMeans: **11%** (100 million samples)



Evaluation of PUMA Workloads

- We evaluate different PUMA workloads on 8 nodes with 30GB data size
- Performance benefits for Shuffle-intensive workloads compared to MR-IPoIB :
  - AdjList: **39%**
  - SelfJoin: **58%**
  - RankedInvIndex: **39%**
- Other workloads:
  - SeqCount: **32%**
  - InvIndex: **18%**



Presentation Outline

- Understanding NVRAM and RDMA
- NRCIO: NVM-aware RDMA-based Communication and I/O Schemes
- NRCIO for Big Data Analytics
- Conclusion and Q&A

Conclusion and Future Work

- Exploring NVM-aware RDMA-based Communication and I/O Schemes for Big Data Analytics
- Proposed a new library, **NRCIO** (work-in-progress)
- Re-design HDFS storage architecture with NVRAM
- Re-design RDMA-MapReduce with NVRAM
- Results are promising
- Further optimizations in NRCIO
- Co-design with more Big Data analytics frameworks

The 3rd International Workshop on High-Performance Big Data Computing (HPBDC)

HPBDC 2017 will be held with IEEE International Parallel and Distributed Processing Symposium (IPDPS 2017), Orlando, Florida USA, May, 2017

Keynote Speaker: Prof. Satoshi Matsuoka, Tokyo Institute of Technology, Japan

Panel Moderator: Prof. Jianfeng Zhan (ICT/CAS)
Panel Topic: Sunrise or Sunset: Exploring the Design Space of Big Data Software Stack
Panel Members (Confirmed so far): Prof. Geoffrey C. Fox (Indiana University Bloomington); Dr. Raghunath Nambiar (Cisco); Prof. D. K. Panda (The Ohio State University)

Six Regular Research Papers and One Short Research Papers
Session I: High-Performance Graph Processing
Session II: Benchmarking and Performance Analysis

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

Thank You!

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

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

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

Network-Based Computing Laboratory

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

The High-Performance Big Data Project

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