2021 OFA Virtual Workshop

RDMA SPARK MEETS OSU INAM: PERFORMANCE ENGINEERING BIG DATA APPLICATIONS ON HPC CLUSTERS

Mansa Kedia, Pouya Kousha, Hari Subramoni, Aamir Shafi, and Dhabaleswar K. (DK) Panda

The Ohio State University

E-mail: shafi.16@osu.edu

https://people.engineering.osu.edu/people/shafi.16

OUTLINE

- Motivation
- Introduction to RDMA Spark and OSU INAM
- Analyzing RDMA Spark Shuffle Stage using OSU INAM
- Usage Scenario:
  - MPI
  - RDMA Spark
- Conclusions and Future Work

BACKGROUND

- Big Data computing frameworks, including Spark and Dask, enable high-performance data science on HPC clusters built with InfiniBand/RoCE interconnects:
  - Recent trend to support accelerators
  - Adding support for low-latency and high-throughput networks like InfiniBand and RoCE

- The programming model for these frameworks enables programmer productivity
  - Hiding low-level communication and I/O details

- It is challenging to analyze the performance of their applications at finer granularity:
  - Such knowledge, in the context of communication, is critical to maximize utilization of underlying network and the application

MOTIVATION

- One popular Big Data computing framework is the Spark framework
  - Java, Scala, and Python APIs to transform and act upon Resilient Distributed Datasets (RDDs)
  - Components include Spark SQL, MLlib, Spark Steaming, GraphX, and others

- Spark web user interface enables monitoring the resource consumption of the Spark cluster. This, however, lacks:
  - Network topology details and in-depth network communication information

- The core idea of this work is to utilize the OSU INAM* framework for monitoring and analyzing network resource usage of the Spark framework on HPC systems

* Visualize and Analyze your Network Activities using OSU INAM, Presented at OFAWS 2020, https://www.openfabrics.org/wp-content/uploads/2020-workshop-presentations/305.-OFA-Virtual-Workshop-2020-PPT-Template.pdf

OUTLINE

- Motivation
- Introduction to RDMA Spark and OSU INAM
- Analyzing RDMA Spark Shuffle Stage using OSU INAM
- Usage Scenario:
  - MPI
  - RDMA Spark
- Conclusions and Future Work

THE HIGH-PERFORMANCE BIG DATA (HIBD) PROJECT

- Since 2013
- **RDMA for Apache Spark**
- MPI4Dask 0.1 (MVAPICH2-GDR based Dask Distributed Library)
- RDMA for Apache Hadoop 3.x (RDMA-Hadoop-3.x)
- RDMA for Apache Hadoop 2.x (RDMA-Hadoop-2.x)
  - Plugins for Apache, Hortonworks (HDP) and Cloudera (CDH) Hadoop distributions
- RDMA for Apache Kafka
- RDMA for Apache HBase
- RDMA for Memcached (RDMA-Memcached)
- RDMA for Apache Hadoop 1.x (RDMA-Hadoop)
- OSU HiBD-Benchmarks (OHB)
- http://hibd.cse.ohio-state.edu
- **Users Base: 340 organizations from 37 countries**
- **More than 39,250 downloads from the project site**

Available for InfiniBand and RoCE

Also runs on Ethernet

Available for x86 and OpenPOWER

Support for Singularity and Docker

HiBD

High-Performance Big Data

The Ohio State University

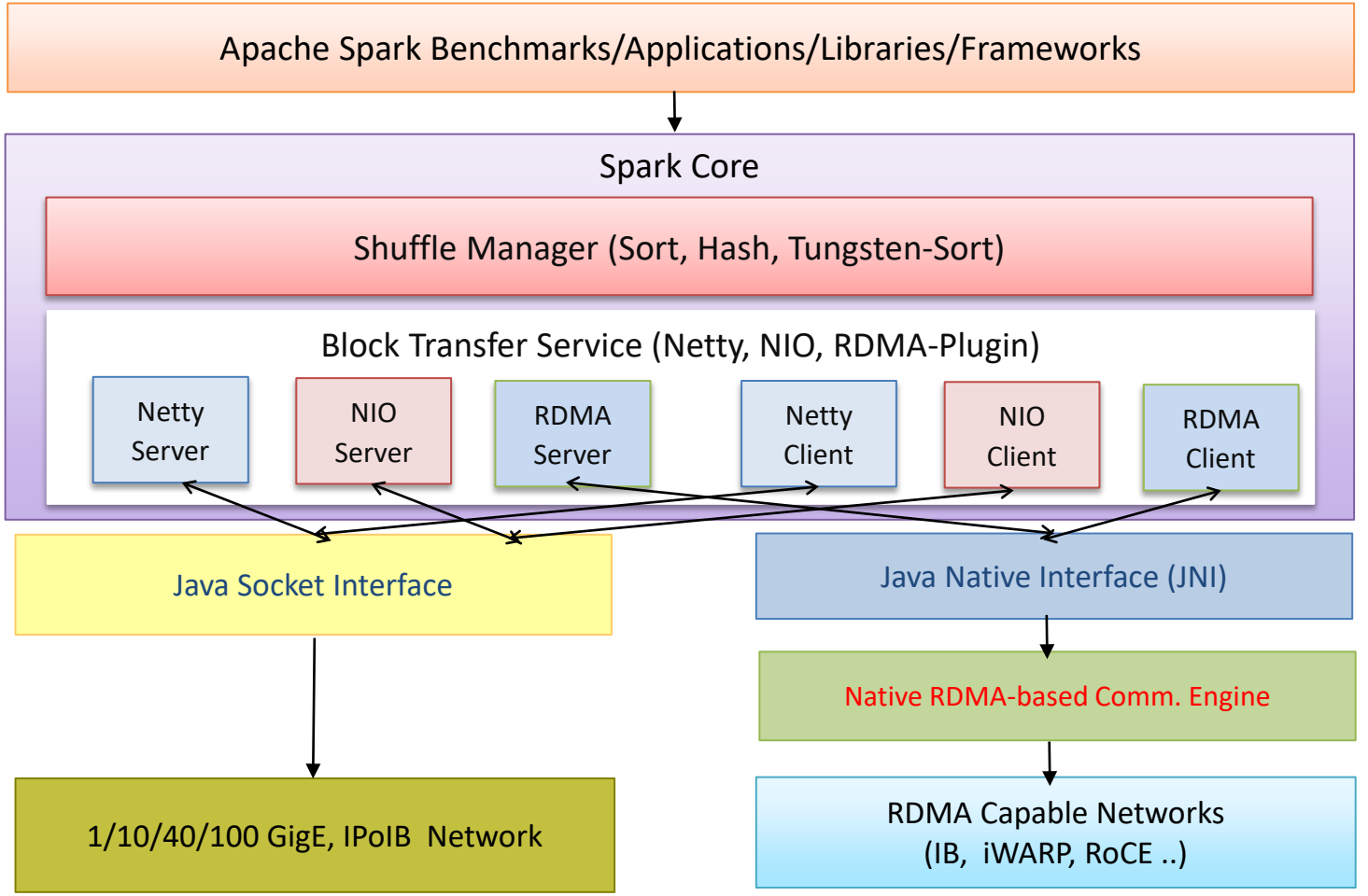
Network Based Computing Laboratory

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, 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**

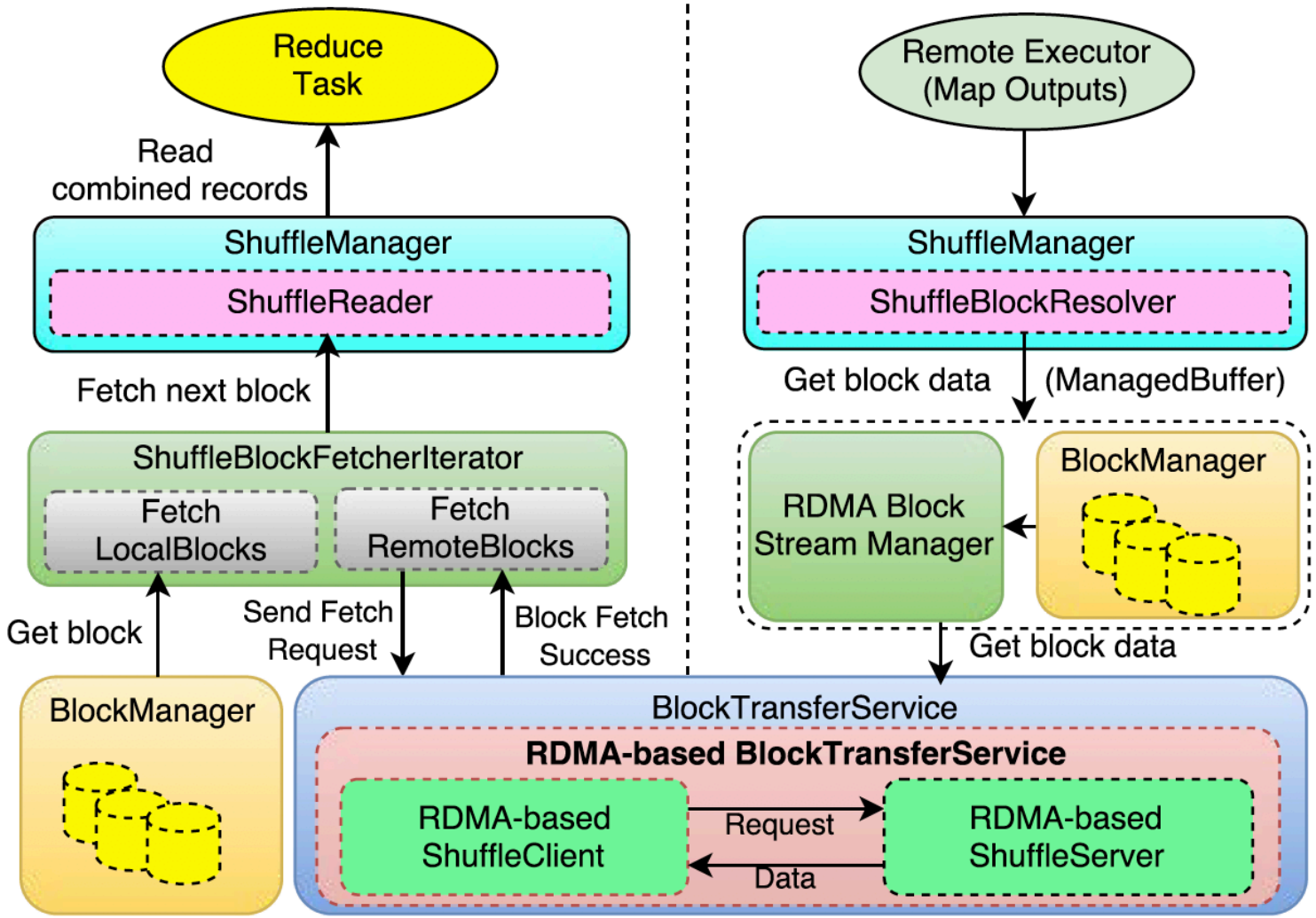
DESIGN OVERVIEW OF SPARK WITH RDMA

- Design Features
  - Enables high performance RDMA communication, while supporting traditional socket interface
  - JNI Layer bridges Scala based Spark with communication library written in native code
  - RDMA based shuffle plugin
  - SEDA-based architecture
  - Dynamic connection management and sharing
  - Non-blocking data transfer
  - Off-JVM-heap buffer management
  - InfiniBand/RoCE support



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.

RDMA-SPARK PROPOSED DESIGN CHANGE – DATA COLLECTION (CONT.)



OVERVIEW OF OSU INAM

- INAM - InfiniBand Network Analysis and Monitoring tool that can analyze traffic on the InfiniBand network with inputs from the MPI runtime
- Remotely monitors IB clusters in real time by querying various subnet management entities and gathering input from the MPI runtimes
- Capability to analyze and profile node-level, job-level and process-level activities for MPI communication
- Visualize the data transfer happening in a “live” or “historical” fashion for entire network, job or set of nodes
- OSU INAM has been downloaded more than 4,400 times directly from the OSU site (http://mvapich.cse.ohio-state.edu/tools/osu-inam/)

OUTLINE

- Motivation
- Introduction to RDMA Spark and OSU INAM
- Analyzing RDMA Spark Shuffle Stage using OSU INAM
- Usage Scenario:
  - MPI
  - RDMA Spark
- Conclusions and Future Work

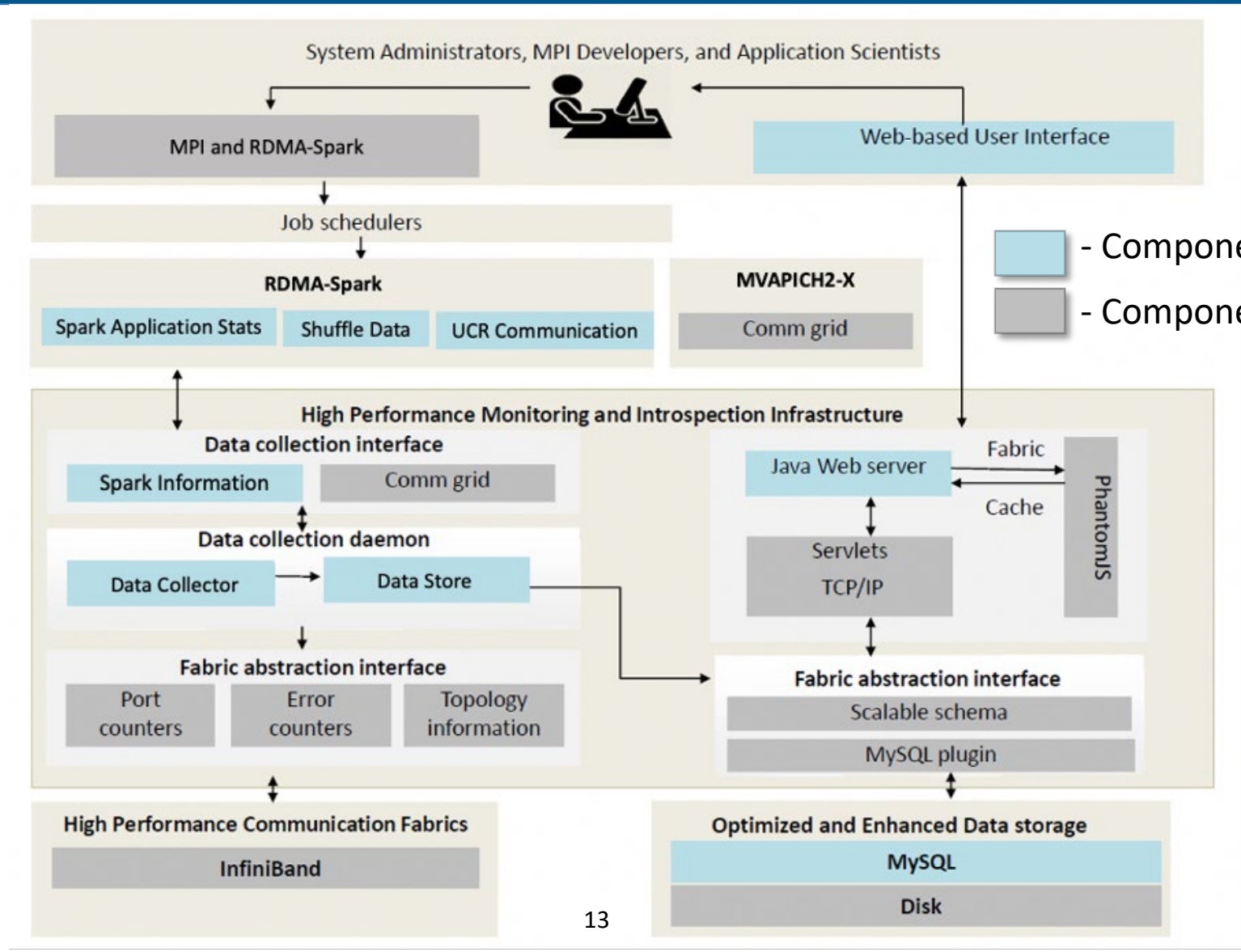
RDMA SPARK MEETS INAM

- OSU INAM is enhanced to allow monitoring and analyzing RDMA Spark jobs on HPC systems

- The vanilla Spark WebUI provides basic communication information like the volume of data sent and received by Spark workers:
  - The UI has no knowledge of the native C code for RDMA communication

- RDMA Spark in conjunction with OSU INAM can provide the following information:
  - Summary of all jobs and stages
  - Overview of jobs
  - Detailed Shuffle information:
    - Amount of data exchange between each pair of executors
    - Time taken for each exchange
  - Task-level information

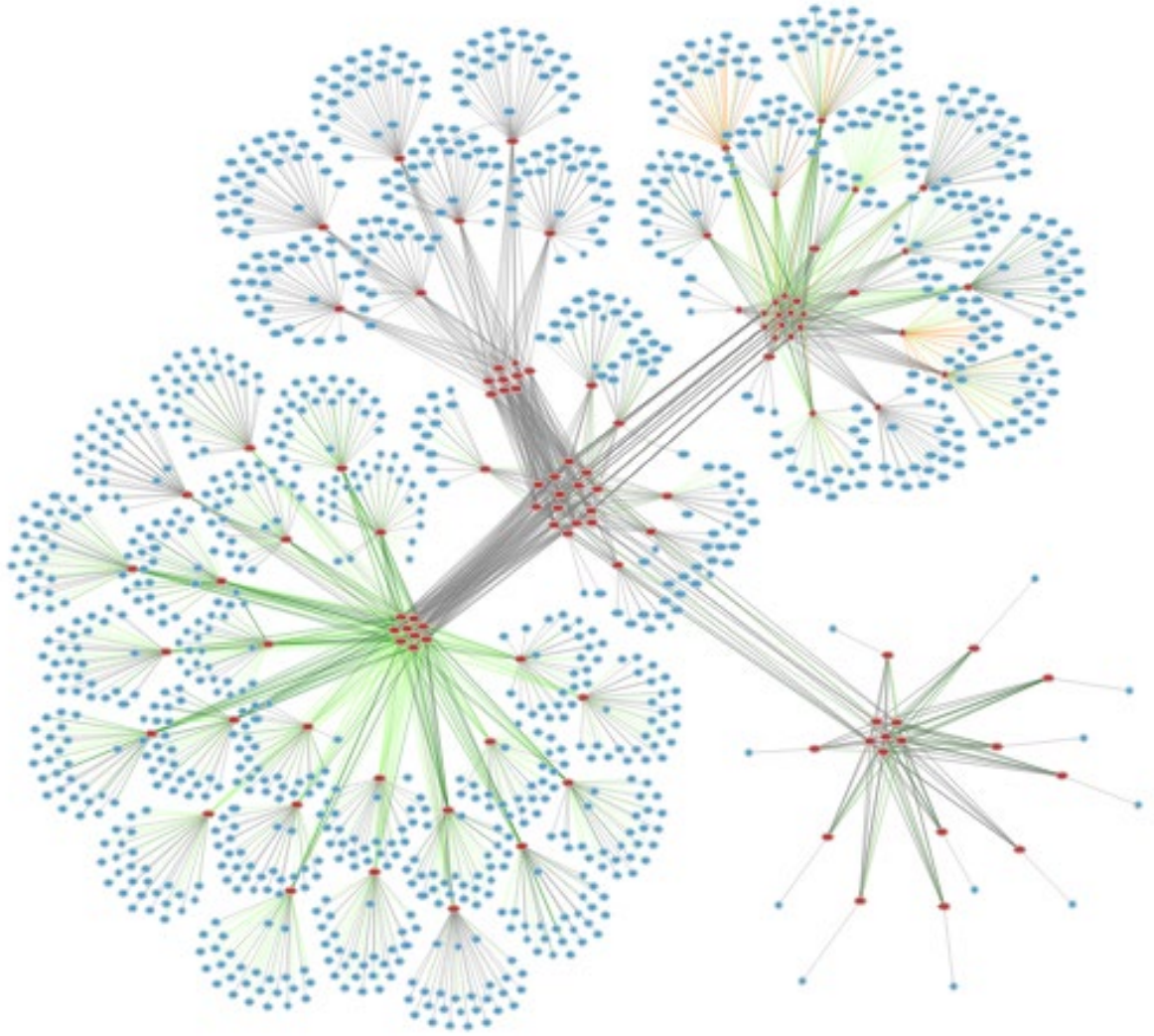
PROPOSED ARCHITECTURE



OUTLINE

- Motivation
- Introduction to RDMA Spark and OSU INAM
- Analyzing RDMA Spark Shuffle Stage using OSU INAM
- Usage Scenario:
  - MPI
  - RDMA Spark
- Conclusions and Future Work

VIEWING TOPOLOGY USING OSU INAM

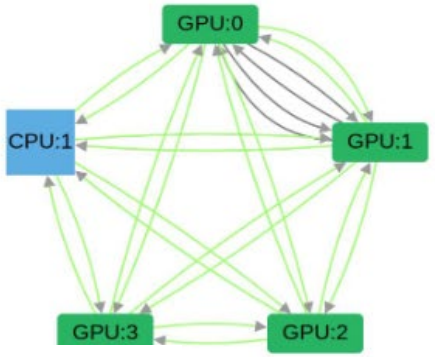


- Network view showing 3 heterogeneous clusters at OSC
- Clusters are connected to the same InfiniBand Fabric:
  - 114 switches
  - 1,428 compute nodes
  - 3,402 links

NETWORK AND LIVE JOBS VIEW GENERATION TIMING ON OSC WITH 1K JOBS

View	Average	Min	Max	STDEV.p
Network View	196.15 ms	187 ms	206.09 ms	5.75 ms
Live Jobs View	18.17 ms	16 ms	20 ms	1 ms

MONITORING MPI JOBS USING OSU INAM



Topology within a node

Rank Communication Grid

	0	1	2	3	4	5	6	7
0	758.78 MB	759.04 MB	758.76 MB	758.78 MB	758.87 MB	759.04 MB	758.78 MB	0.00 bytes
1	759.04 MB	758.78 MB	0.00 bytes	759.04 MB	758.78 MB	758.91 MB	759.04 MB	759.04 MB
2	0.00 bytes	758.78 MB	758.76 MB	759.04 MB	758.78 MB	758.78 MB	759.04 MB	759.01 MB
3	758.78 MB	758.97 MB	758.76 MB	0.00 bytes	758.78 MB	759.04 MB	758.78 MB	759.04 MB
4	758.78 MB	759.04 MB	758.81 MB	758.78 MB	759.04 MB	0.00 bytes	758.78 MB	758.78 MB
5	759.04 MB	758.78 MB	759.04 MB	758.78 MB	0.00 bytes	758.78 MB	758.91 MB	758.78 MB
6	758.78 MB	758.78 MB	758.78 MB	759.04 MB	758.78 MB	759.04 MB	0.00 bytes	759.04 MB
7	758.78 MB	0.00 bytes	758.84 MB	758.78 MB	759.04 MB	758.78 MB	758.78 MB	758.78 MB

Absolute

279280

Link utilization by MPI processes

MF0;ibswitch:MTS3610/L11/U1 -- > node008 HCA-1

Communication grid for message exchange between processes

Overview of MPI Usages

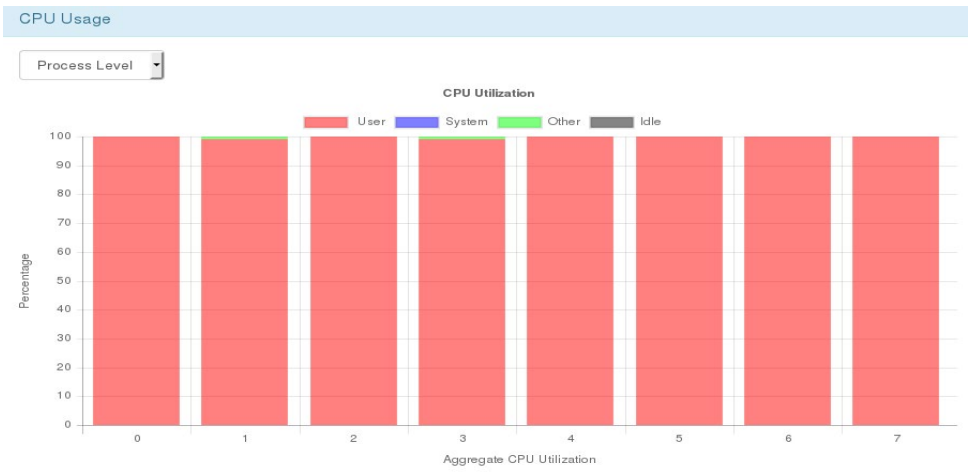
MONITORING MPI JOBS USING OSU INAM (CONT.)

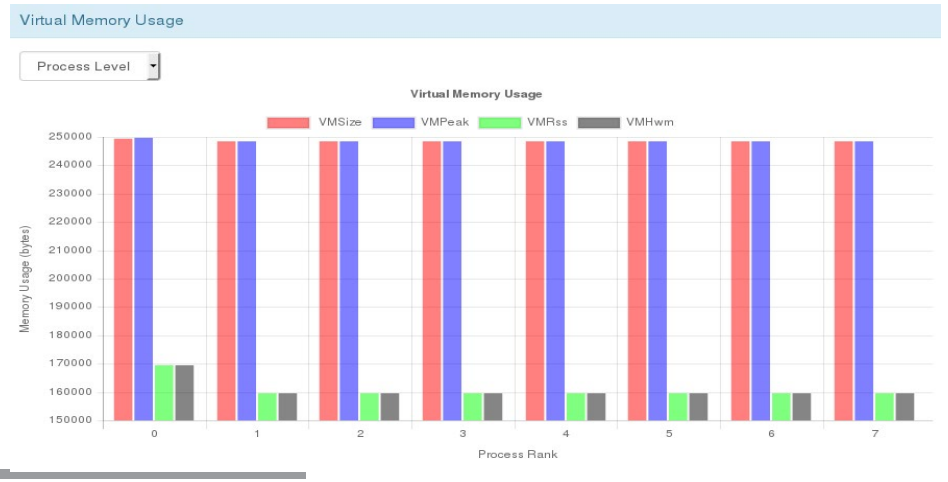
Monitoring Jobs Based on Various Metrics

Overview of MPI Usages

Job ID	CPU User Usage	Virtual Memory Size	Total Communication	Total Inter Node	Total Intra Node	Total Collective	RMA Sent
270747	99	8.19 Mb	92.35 Gb	36.69 Gb	55.66 Gb	64.46 Gb	0.00 bytes
270748	99	15.12 Mb	149.98 Gb	58.23 Gb	91.76 Gb	102.78 Gb	0.00 bytes
270749	99	30.39 Mb	151.23 Gb	58.35 Gb	92.88 Gb	100.34 Gb	0.00 bytes
270759	99	17.99 Mb	58.71 Gb	37.29 Gb	21.43 Gb	303.73 Kb	0.00 bytes
270765	99	9.42 Mb	32.52 Gb	23.19 Gb	9.33 Gb	0.00 bytes	0.00 bytes

Profiling and Reporting Performance Metrics at Different Granularities





OUTLINE

- Motivation
- Introduction to RDMA Spark and OSU INAM
- Analyzing RDMA Spark Shuffle Stage using OSU INAM
- Usage Scenario:
  - MPI
  - RDMA Spark
- Conclusions and Future Work

RDMA SPARK MEETS INAM: JOBS AND STAGES LEVEL SUMMARY

Jobs Details

Jobs Level Summary for RDMA Spark as provided by OSU INAM

Search

Job Id	Job Name	Submission Time	Completion Time	Status	Number of tasks	Number of active tasks	Number of completed tasks	Number of failed tasks	Number of active stages	Number of completed stages	Number of failed stages
2	count at SortByTest.scala:47	2020-11-12T15:56:56.975GMT	2020-11-12T15:57:00.596GMT	SUCCEEDED	48	0	48	0	0	2	0
1	sortByKey at SortByTest.scala:47	2020-11-12T15:56:55.318GMT	2020-11-12T15:56:56.949GMT	SUCCEEDED	16	0	16	0	0	1	0
0	count at SortByTest.scala:45	2020-11-12T15:56:52.320GMT	2020-11-12T15:56:55.291GMT	SUCCEEDED	16	0	16	0	0	1	0

Showing 1 to 3 of 3 rows

Stages Details

Spark jobs have stages. Stages Level Summary for RDMA Spark Jobs as provided by OSU INAM

Search

Stage Id	Stage Name	Submission Time	Completion Time	Status	Input Bytes	Input Records	Output Bytes	Output Records	Shuffle Read Bytes	Shuffle Read Records	Shuffle Write Bytes	Shuffle Write Records	Memory Bytes Spilled
3	count at SortByTest.scala:47	12 Nov 2020 15:56:59:58 GMT	12 Nov 2020 15:57:00:96 GMT	COMPLETE	0 B	0	0 B	0	4.30 GB	1048576	0 B	0	0 B
2	flatMap at SortByTest.scala:34	12 Nov 2020 15:56:56:91 GMT	12 Nov 2020 15:56:59:35 GMT	COMPLETE	2.18 GB	524288	0 B	0	0 B	0	4.30 GB	1048576	0 B
1	sortByKey at SortByTest.scala:47	12 Nov 2020 15:56:55:26 GMT	12 Nov 2020 15:56:56:49 GMT	COMPLETE	0 B	0	0 B	0	0 B	0	0 B	0	0 B
0	count at SortByTest.scala:45	12 Nov 2020 15:56:52:05 GMT	12 Nov 2020 15:56:55:87 GMT	COMPLETE	0 B	0	0 B	0	0 B	0	0 B	0	0 B

Showing 1 to 4 of 4 rows

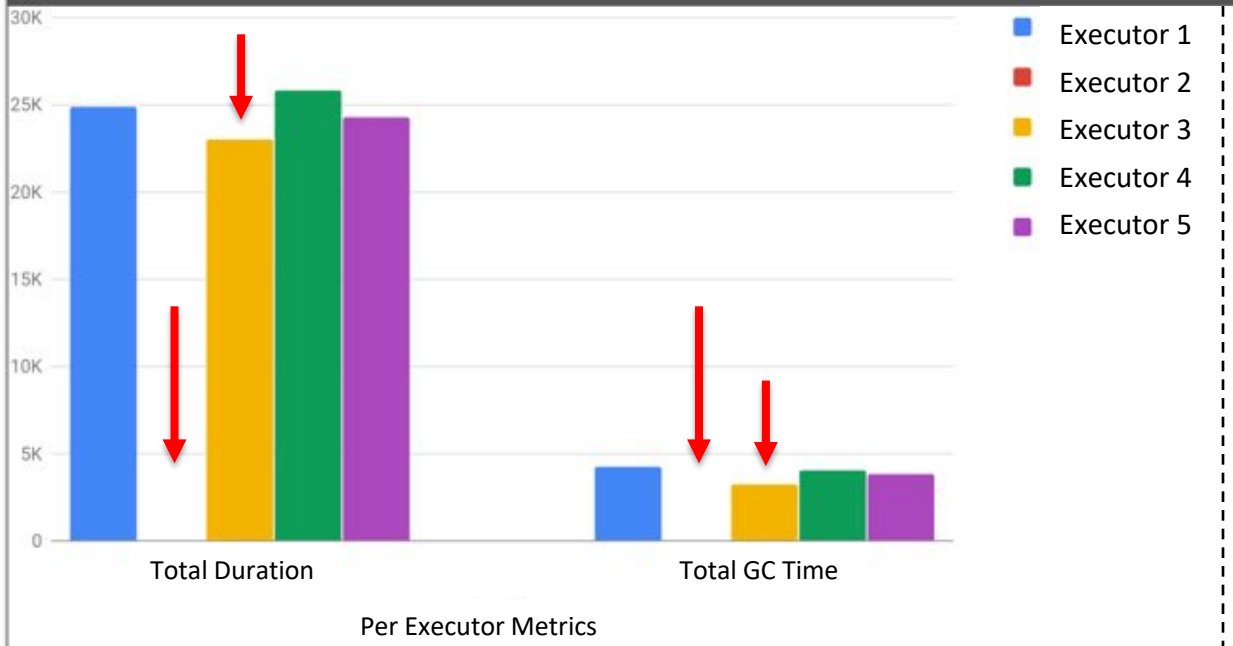
RDMA SPARK MEETS INAM: SINGLE JOB SUMMARY

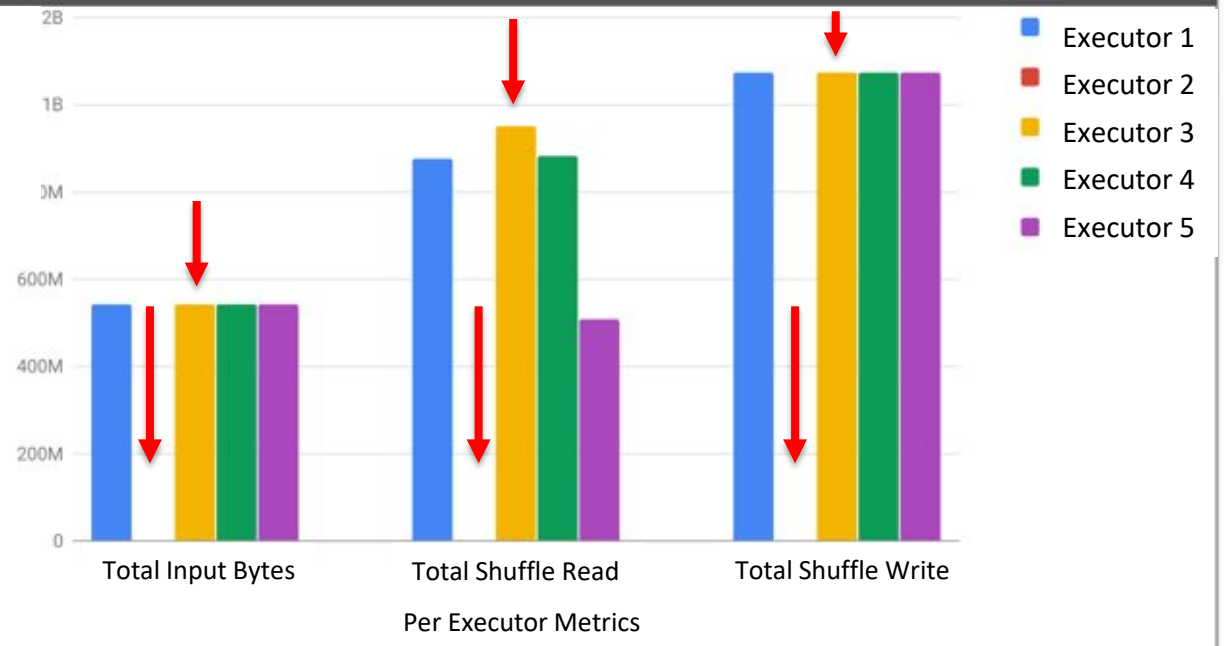
Job Information

Spark Application id: app-20201112105651-0006

Overview of a single RDMA Spark Job as provided by OSU INAM

Executor Details

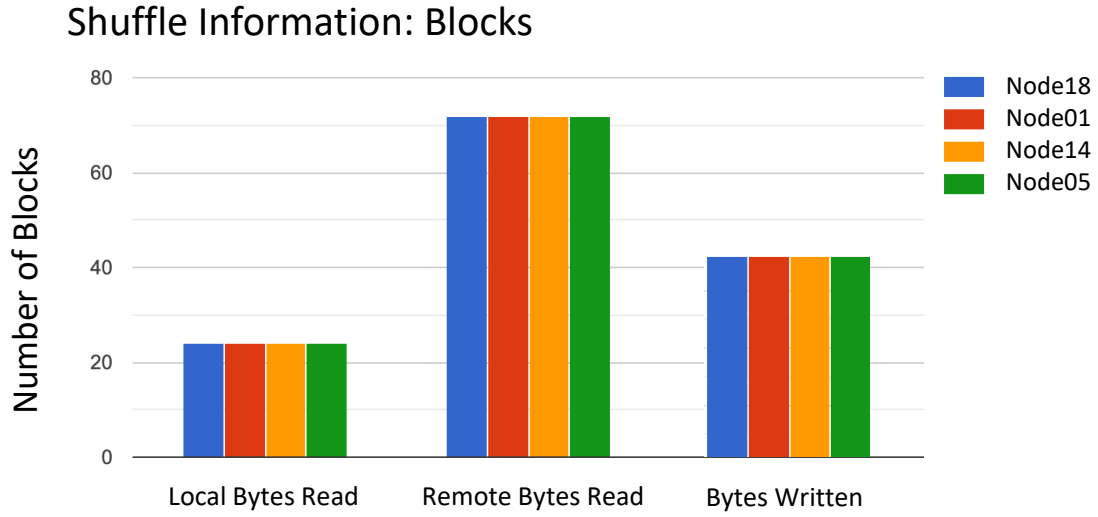


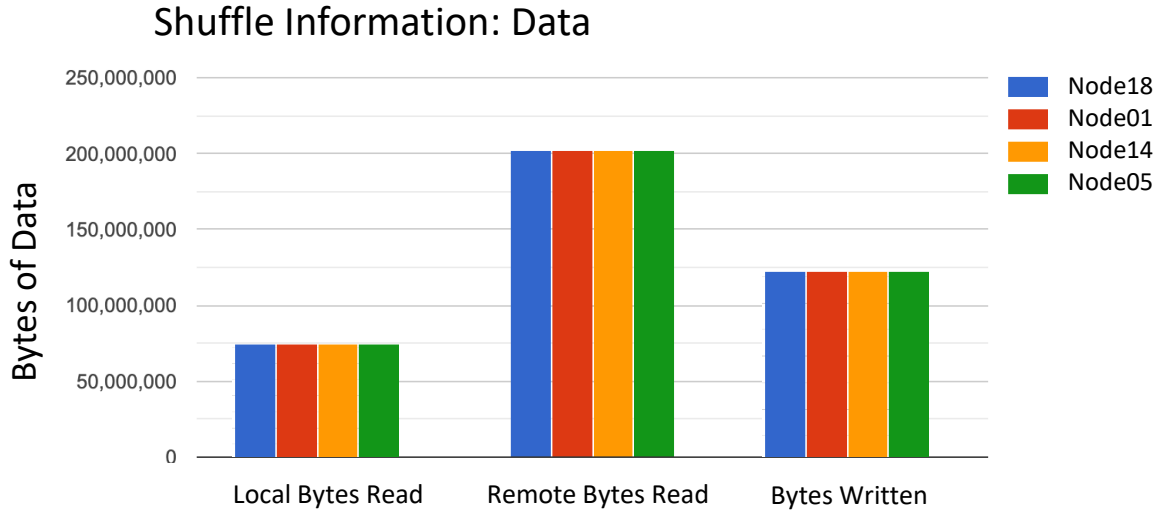


Search

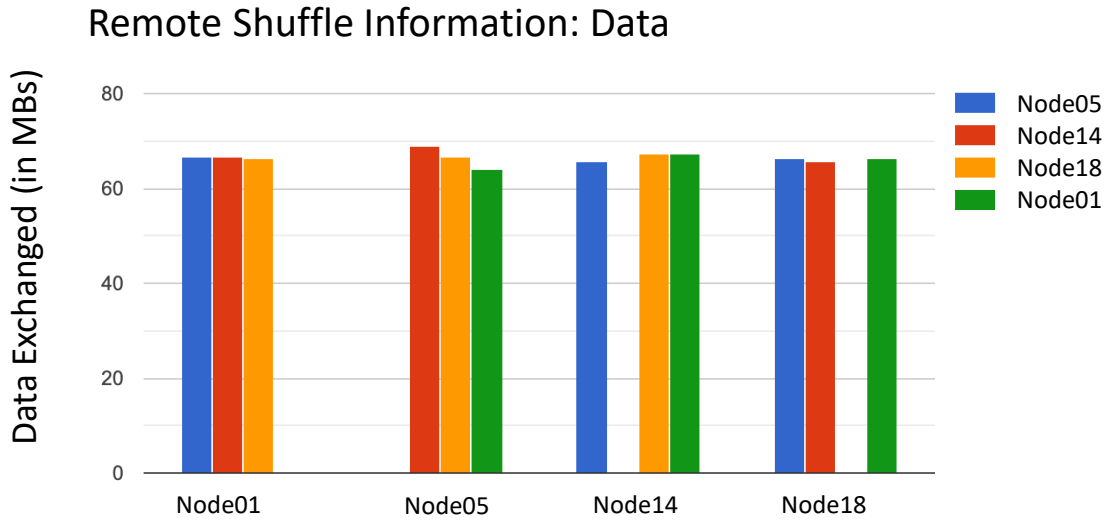
Host Port	Memory Used	Storage Memory	Disk Used	Total Cores	Failed Tasks	Completed tasks	Total tasks	Total Duration	Total GC time	Input Bytes	Shuffle Read	Shuffle Write
Executor 2	0 B	2.05 GB	0 B	0	0	0	0	0 ms	0 ms	0 B	0 B	0 B
Executor 3	0 B	1.02 GB	0 B	4	0	20	20	24.93 s	4.30 s	544.44 MB	878.53 MB	1.08 GB

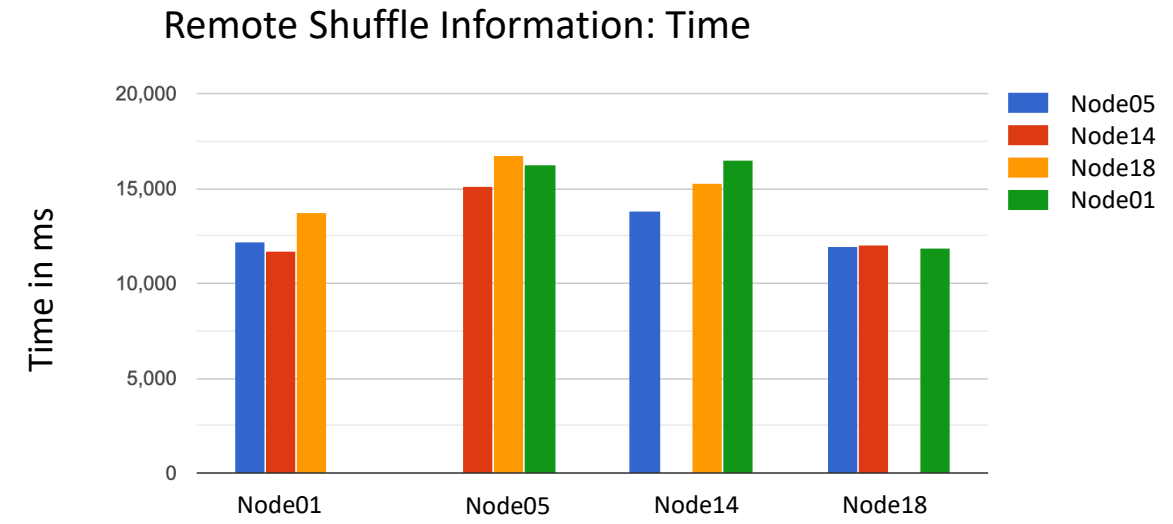
RDMA SPARK MEETS INAM: NETWORK USAGE FOR SHUFFLE STAGE





Per Worker Statistics for the Shuffle Stage





RDMA SPARK MEETS INAM: TASK-LEVEL INFORMATION

- Task-level information:
  - the launch time
  - executor run time
  - executor deserialize time
  - executor deserialize time
  - executor deserialize CPU time
  - result serialization time
  - GC time
  - executor CPU time

Task-level Information for Spark Jobs

OUTLINE

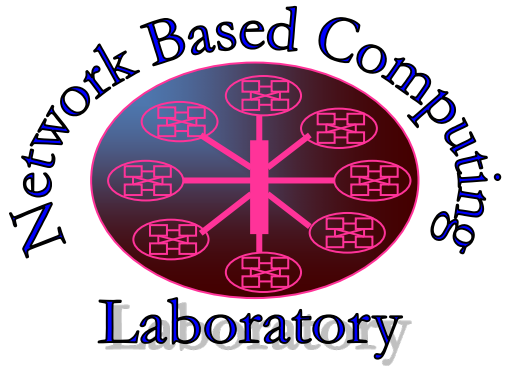
- Motivation
- Introduction to RDMA Spark and OSU INAM
- Analyzing RDMA Spark Shuffle Stage using OSU INAM
- Usage Scenario:
  - MPI
  - RDMA Spark
- Conclusions and Future Work

CONCLUSIONS AND FUTURE WORK

- Enhanced RDMA Spark and OSU INAM to monitor and analyze the communication traffic on the InfiniBand and RoCE networks
- The main contribution is to provide a detailed view of the network traffic incurred by the shuffle operation of Spark jobs
- Users include Big Data applications users, developers, and HPC system administrators
- Future enhancements:
  - Collect and visualize data transfer during read/write from the underlying HDFS filesystem
  - Allow users to set notifications for any overlap of spark data transfer with other I/O or MPI traffic
- We plan to publicly release RDMA Spark and OSU INAM enhancements

THANK YOU!

kousha.2@osu.edu, subramon@cse.ohio-state.edu, shafi.16@osu.edu, panda@cse.ohio-state.edu



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

The High-Performance MPI/PGAS Project
http://mvapich.cse.ohio-state.edu/

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