Accelerating Big Data Processing with RDMA-Enhanced Apache Hadoop

Keynote Talk at BPOE-4, in conjunction with ASPLOS ‘14
(March 2014)

by

Dhabaleswar K. (DK) Panda
The Ohio State University
E-mail: panda@cse.ohio-state.edu
http://www.cse.ohio-state.edu/~panda
Introduction to Big Data Applications and Analytics

- Big Data has become one of the most important elements of business analytics
- Provides groundbreaking opportunities for enterprise information management and decision making
- The amount of data is exploding; companies are capturing and digitizing more information than ever
- The rate of information growth appears to be exceeding Moore’s Law
- Commonly accepted 3V’s of Big Data
  - **Volume, Velocity, Variety**
- 5V’s of Big Data – 3V + **Value, Veracity**
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
Typical Solutions or Architectures for Big Data Applications and Analytics

  - The most popular framework for Big Data Analytics
  - HDFS, MapReduce, HBase, RPC, Hive, Pig, ZooKeeper, Mahout, etc.

- Spark: [http://spark-project.org](http://spark-project.org)
  - Provides primitives for in-memory cluster computing; Jobs can load data into memory and query it repeatedly

- Shark: [http://shark.cs.berkeley.edu](http://shark.cs.berkeley.edu)
  - A fully Apache Hive-compatible data warehousing system

- Storm: [http://storm-project.net](http://storm-project.net)
  - A distributed real-time computation system for real-time analytics, online machine learning, continuous computation, etc.

- S4: [http://incubator.apache.org/s4](http://incubator.apache.org/s4)
  - A distributed system for processing continuous unbounded streams of data

- Web 2.0: RDBMS + Memcached ([http://memcached.org](http://memcached.org))
  - Memcached: A high-performance, distributed memory object caching systems
Who Are Using Hadoop?

- Focuses on large data and data analysis
- Hadoop (e.g. HDFS, MapReduce, HBase, RPC) environment is gaining a lot of momentum
- [http://wiki.apache.org/hadoop/PoweredBy](http://wiki.apache.org/hadoop/PoweredBy)
Hadoop at Yahoo!

• Scale: 42,000 nodes, 365PB HDFS, 10M daily slot hours

• A new Jim Gray’ Sort Record: 1.42 TB/min over 2,100 nodes
  • Hadoop 0.23.7, an early branch of the Hadoop 2.X
Presentation Outline

• Overview of Hadoop
• Trends in Networking Technologies and Protocols
• Challenges in Accelerating Hadoop
• Designs and Case Studies
  – HDFS
  – MapReduce
  – HBase
  – RPC
  – Combination of components
• Challenges in Big Data Benchmarking
• Other Open Challenges in Acceleration
• Conclusion and Q&A
Big Data Processing with Hadoop

• The open-source implementation of MapReduce programming model for Big Data Analytics

• Major components
  – HDFS
  – MapReduce

• Underlying Hadoop Distributed File System (HDFS) used by both MapReduce and HBase

• Model scales but high amount of communication during intermediate phases can be further optimized
Hadoop MapReduce Architecture & Operations

- Map and Reduce Tasks carry out the total job execution
- Communication in shuffle phase uses HTTP over Java Sockets
Hadoop Distributed File System (HDFS)

- Primary storage of Hadoop; highly reliable and fault-tolerant
- Adopted by many reputed organizations
  - eg: Facebook, Yahoo!
- NameNode: stores the file system namespace
- DataNode: stores data blocks
- Developed in Java for platform-independence and portability
- Uses sockets for communication!
System-Level Interaction Between Clients and Data Nodes in HDFS
Overview of Hadoop

Trends in Networking Technologies and Protocols

Challenges in Accelerating Hadoop

Designs and Case Studies
  – HDFS
  – MapReduce
  – RPC
  – Combination of components

Challenges in Big Data Benchmarking

Other Open Challenges in Acceleration

Conclusion and Q&A
Trends for Commodity Computing Clusters in the Top 500 List (http://www.top500.org)
HPC and Advanced Interconnects

- High-Performance Computing (HPC) has adopted advanced interconnects and protocols
  - InfiniBand
  - 10 Gigabit Ethernet/iWARP
  - RDMA over Converged Enhanced Ethernet (RoCE)

- Very Good Performance
  - Low latency (few micro seconds)
  - High Bandwidth (100 Gb/s with dual FDR InfiniBand)
  - Low CPU overhead (5-10%)

- OpenFabrics software stack with IB, iWARP and RoCE interfaces are driving HPC systems

- Many such systems in Top500 list
Trends of Networking Technologies in TOP500 Systems

Percentage share of InfiniBand is steadily increasing

Interconnect Family – Systems Share
Open Standard InfiniBand Networking Technology

- Introduced in Oct 2000
- High Performance Data Transfer
  - Interprocessor communication and I/O
  - Low latency (<1.0 microsec), High bandwidth (up to 12.5 GigaBytes/sec -> 100Gbps), and low CPU utilization (5-10%)
- Flexibility for LAN and WAN communication
- Multiple Transport Services
  - Reliable Connection (RC), Unreliable Connection (UC), Reliable Datagram (RD), Unreliable Datagram (UD), and Raw Datagram
  - Provides flexibility to develop upper layers
- Multiple Operations
  - Send/Recv
  - RDMA Read/Write
  - Atomic Operations (very unique)
    - high performance and scalable implementations of distributed locks, semaphores, collective communication operations
- Leading to big changes in designing HPC clusters, file systems, cloud computing systems, grid computing systems, ....
Use of High-Performance Networks for Scientific Computing

- Message Passing Interface (MPI)
  - Most commonly used programming model for HPC applications
- Parallel File Systems
- Storage Systems
- Almost 13 years of Research and Development since InfiniBand was introduced in October 2000
- Other Programming Models are emerging to take advantage of High-Performance Networks
  - UPC
  - OpenSHMEM
  - Hybrid MPI+PGAS (OpenSHMEM and UPC)
MVAPICH2/MVAPICH2-X Software

- High Performance open-source MPI Library for InfiniBand, 10Gig/iWARP, and RDMA over Converged Enhanced Ethernet (RoCE)
  - MVAPICH (MPI-1), MVAPICH2 (MPI-2.2 and MPI-3.0), Available since 2002
  - MAPICH2-X (MPI + PGAS), Available since 2012
  - Support for GPGPUs and MIC
  - Used by more than 2,100 organizations (HPC Centers, Industry and Universities) in 71 countries
  - More than 203,000 downloads from OSU site directly
- Empowering many TOP500 clusters
  - 7th ranked 462,462-core cluster (Stampede) at TACC
  - 11th ranked 74,358-core cluster (Tsubame 2.5) at Tokyo Institute of Technology
  - 16th ranked 96,192-core cluster (Pleiades) at NASA and many others
- Available with software stacks of many IB, HSE, and server vendors including Linux Distros (RedHat and SuSE)
  - http://mvapich.cse.ohio-state.edu
- Partner in the U.S. NSF-TACC Stampede System
One-way Latency: MPI over IB with MVAPICH2

**Small Message Latency**

- **Message Size (bytes)**: 0, 4, 8, 16, 32, 64, 128, 256, 512, 1K
- **Latency (us)**: 0.99, 1.09, 1.12, 1.56, 1.64, 1.66, 1.82

**Large Message Latency**

- **Message Size (bytes)**: 2K, 4K, 8K, 16K, 32K, 64K, 128K, 256K
- **Latency (us)**: 0.00, 5.00, 10.00, 15.00, 20.00, 25.00, 30.00, 35.00, 40.00

**Details**:

- DDR, QDR - 2.4 GHz Quad-core (Westmere) Intel PCI Gen2 with IB switch
- FDR - 2.6 GHz Octa-core (SandyBridge) Intel PCI Gen3 with IB switch
- ConnectIB-Dual FDR - 2.6 GHz Octa-core (SandyBridge) Intel PCI Gen3 with IB switch
- ConnectIB-Dual FDR - 2.8 GHz Deca-core (IvyBridge) Intel PCI Gen3 with IB switch
Bandwidth: MPI over IB with MVAPICH2

**Unidirectional Bandwidth**

- Qlogic-DDR
- Qlogic-QDR
- ConnectX-DDR
- ConnectX2-PCIe2-QDR
- ConnectX3-PCIe3-FDR
- Sandy-ConnectIB-DualFDR
- Ivy-ConnectIB-DualFDR

**Bidirectional Bandwidth**

- DDR, QDR - 2.4 GHz Quad-core (Westmere) Intel PCI Gen2 with IB switch
- FDR - 2.6 GHz Octa-core (SandyBridge) Intel PCI Gen3 with IB switch
- ConnectIB-Dual FDR - 2.6 GHz Octa-core (SandyBridge) Intel PCI Gen3 with IB switch
- ConnectIB-Dual FDR - 2.8 GHz Deca-core (IvyBridge) Intel PCI Gen3 with IB switch
Presentation Outline

- Overview of Hadoop
- Trends in Networking Technologies and Protocols
- Challenges in Accelerating Hadoop
- Designs and Case Studies
  - HDFS
  - MapReduce
  - RPC
  - Combination of components
- Challenges in Big Data Benchmarking
- Other Open Challenges in Acceleration
- Conclusion and Q&A
Can High-Performance Interconnects Benefit Hadoop?

- Most of the current Hadoop environments use Ethernet Infrastructure with Sockets
- Concerns for performance and scalability
- Usage of High-Performance Networks is beginning to draw interest from many companies
- What are the challenges?
- Where do the bottlenecks lie?
  - Can these bottlenecks be alleviated with new designs (similar to the designs adopted for MPI)?
  - Can HPC Clusters with High-Performance networks be used for Big Data applications using Hadoop?
## Designing Communication and I/O Libraries for Big Data Systems: Challenges

<table>
<thead>
<tr>
<th>Applications</th>
<th>Benchmarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Big Data Middleware (HDFS, MapReduce, HBase and Memcached)</td>
<td>Other Protocols?</td>
</tr>
<tr>
<td>Programming Models (Sockets)</td>
<td></td>
</tr>
<tr>
<td>Communication and I/O Library</td>
<td></td>
</tr>
<tr>
<td>Point-to-Point Communication</td>
<td>Threaded Models and Synchronization</td>
</tr>
<tr>
<td>I/O and File Systems</td>
<td>QoS</td>
</tr>
<tr>
<td>Networking Technologies (InfiniBand, 1/10/40GigE and Intelligent NICs)</td>
<td>Commodity Computing System Architectures (Multi- and Many-core architectures and accelerators)</td>
</tr>
<tr>
<td>Virtualization</td>
<td>Fault-Tolerance</td>
</tr>
<tr>
<td>Storage Technologies (HDD and SSD)</td>
<td></td>
</tr>
</tbody>
</table>

**BPOE-4, ASPLOS '14**
Common Protocols using Open Fabrics

Application /Middleware Interface

- Application /Middleware
  - Sockets
  - RDMA
  - User space
  - User space
  - User space
  - User space
  - User space
  - User space

Protocol

- TCP/IP
  - Ethernet Driver
  - IPoIB
  - InfiniBand Adapter
  - Ethernet Adapter
  - InfiniBand Switch
  - IPoIB
  - Ethernet Adapter
  - IPoIB Switch

- RSockets
  - User space
  - User space
  - User space
  - User space
  - User space
  - User space
  - RSockets

- SDP
  - User space
  - User space
  - User space
  - User space
  - User space
  - User space

- iWARP
  - User space
  - User space
  - User space
  - User space
  - User space
  - User space

- RDMA
  - User space
  - User space
  - User space
  - User space
  - User space
  - User space

- RoCE
  - User space
  - User space
  - User space
  - User space
  - User space
  - User space

- IB Verbs
  - IPoIB Switch
  - InfiniBand Switch
  - IPoIB Adapter
  - Ethernet Adapter
  - Ethernet Switch
  - InfiniBand Adapter
  - Ethernet Adapter
  - Ethernet Switch
  - RSocket Adapter
  - Ethernet Adapter
  - Ethernet Switch
  - Ethernet Adapter
  - Ethernet Switch

- Hardware Offload
  - InfiniBand Adapter
  - InfiniBand Switch
  - 10/40 GigE-TOE

- IPoIB
  - Ethernet Adapter
  - IPoIB Switch

- Ethernet Adapter
  - Ethernet Adapter
  - Ethernet Adapter
  - Ethernet Switch
  - 1/10/40 GigE

- InfiniBand Adapter
  - InfiniBand Adapter
  - InfiniBand Switch
  - IPoIB

- Ethernet Switch
  - Ethernet Adapter
  - Ethernet Switch
  - Ethernet Switch
  - Ethernet Switch
  - IPoIB Switch

- IPoIB
  - Ethernet Adapter
  - IPoIB Switch
Can Hadoop be Accelerated with High-Performance Networks and Protocols?

- **Current Design**
  - Application
  - Sockets
  - 1/10 GigE Network

- **Enhanced Designs**
  - Application
  - Accelerated Sockets
  - Verbs / Hardware Offload
  - 10 GigE or InfiniBand

- **Our Approach**
  - Application
  - OSU Design
  - Verbs Interface
  - 10 GigE or InfiniBand

- Sockets not designed for high-performance
  - Stream semantics often mismatch for upper layers (Hadoop and HBase)
  - Zero-copy not available for non-blocking sockets
Presentation Outline

• Overview of Hadoop
• Trends in Networking Technologies and Protocols
• Challenges in Accelerating Hadoop
• Designs and Case Studies
  – HDFS
  – MapReduce
  – RPC
  – Combination of components
• Challenges in Big Data Benchmarking
• Other Open Challenges in Acceleration
• Conclusion and Q&A
RDMA for Apache Hadoop Project

- High-Performance Design of Hadoop over RDMA-enabled Interconnects
  - High performance design with native InfiniBand and RoCE support at the verbs-level for HDFS, MapReduce, and RPC components
  - Easily configurable for native InfiniBand, RoCE and the traditional sockets-based support (Ethernet and InfiniBand with IPoIB)

- Current release: 0.9.8
  - Based on Apache Hadoop 1.2.1
  - Compliant with Apache Hadoop 1.2.1 APIs and applications
  - Tested with
    - Mellanox InfiniBand adapters (DDR, QDR and FDR)
    - RoCE
    - Various multi-core platforms
    - Different file systems with disks and SSDs
  - [http://hadoop-rdma.cse.ohio-state.edu](http://hadoop-rdma.cse.ohio-state.edu)
Design Overview of HDFS with RDMA

Enables high performance RDMA communication, while supporting traditional socket interface

- JNI Layer bridges Java based HDFS with communication library written in native code
- Only the communication part of HDFS Write is modified; No change in HDFS architecture

Design Features
- RDMA-based HDFS write
- RDMA-based HDFS replication
- InfiniBand/RoCE support

Applications

HDFS

Write

Java Native Interface (JNI)

OSU Design

Verbs

RDMA Capable Networks (IB, 10GE/ iWARP, RoCE ..)

Others

Java Socket Interface

Others

1/10 GigE, IPoIB Network

1/10 GigE, IPoIB Network
Communication Time in HDFS

- Cluster with HDD DataNodes
  - 30% improvement in communication time over IPoIB (QDR)
  - 56% improvement in communication time over 10GigE
- Similar improvements are obtained for SSD DataNodes

Evaluations using TestDFSIO

- Cluster with 8 HDD DataNodes, single disk per node
  - 24% improvement over IPoIB (QDR) for 20GB file size
- Cluster with 4 SSD DataNodes, single SSD per node
  - 61% improvement over IPoIB (QDR) for 20GB file size
Evaluations using TestDFSIO

- Cluster with 4 DataNodes, **1 HDD** per node
  - 24% improvement over IPoIB (QDR) for 20GB file size

- Cluster with 4 DataNodes, **2 HDD** per node
  - 31% improvement over IPoIB (QDR) for 20GB file size

- 2 HDD vs 1 HDD
  - 76% improvement for OSU-IB (QDR)
  - 66% improvement for IPoIB (QDR)
Evaluations using Enhanced DFSIO of Intel HiBench on SDSC-Gordon

- Cluster with 32 DataNodes, single SSD per node
  - 47% improvement in throughput over IPoIB (QDR) for 128GB file size
  - 16% improvement in latency over IPoIB (QDR) for 128GB file size
Evaluations using Enhanced DFSIO of Intel HiBench on TACC-Stampede

- Cluster with 64 DataNodes, single HDD per node
  - 64% improvement in throughput over IPoIB (FDR) for 256GB file size
  - 37% improvement in latency over IPoIB (FDR) for 256GB file size
Presentation Outline

• Overview of Hadoop
• Trends in Networking Technologies and Protocols
• Challenges in Accelerating Hadoop
• Designs and Case Studies
  – HDFS
  – MapReduce
  – RPC
  – Combination of components
• Challenges in Big Data Benchmarking
• Other Open Challenges in Acceleration
• Conclusion and Q&A
Design Overview of MapReduce with RDMA

- Enables high performance RDMA communication, while supporting traditional socket interface
- JNI Layer bridges Java based MapReduce with communication library written in native code
- InfiniBand/RoCE support
Evaluations using Sort

- With 8 HDD DataNodes for 40GB sort
  - 41% improvement over IPoIB (QDR)
  - 39% improvement over UDA-IB (QDR)

- With 8 SSD DataNodes for 40GB sort
  - 52% improvement over IPoIB (QDR)
  - 44% improvement over UDA-IB (QDR)

Evaluations using TeraSort

- 100 GB TeraSort with 8 DataNodes with 2 HDD per node
  - 51% improvement over IPoIB (QDR)
  - 41% improvement over UDA-IB (QDR)
Performance Evaluation on Larger Clusters

- For 240GB Sort in 64 nodes
  - 40% improvement over IPoIB (QDR) with HDD used for HDFS

- For 320GB TeraSort in 64 nodes
  - 38% improvement over IPoIB (FDR) with HDD used for HDFS
Evaluations using PUMA Workload

- 50% improvement in Self Join over IPoIB (QDR) for 80 GB data size
- 49% improvement in Sequence Count over IPoIB (QDR) for 30 GB data size
• 50 small MapReduce jobs executed in a cluster size of 4
• Maximum performance benefit 24% over IPoIB (QDR)
• Average performance benefit 12.89% over IPoIB (QDR)
Presentation Outline

• Overview of Hadoop
• Trends in Networking Technologies and Protocols
• Challenges in Accelerating Hadoop
• Designs and Case Studies
  – HDFS
  – MapReduce
  – RPC
  – Combination of components
• Challenges in Big Data Benchmarking
• Other Open Challenges in Acceleration
• Conclusion and Q&A
Design Overview of Hadoop RPC with RDMA

Applications

Hadoop RPC

Default

Java Socket Interface

1/10 GigE, IPoIB Network

Our Design

Java Native Interface (JNI)

OSU Design

Verbs

RDMA Capable Networks (IB, 10GE/ iWARP, RoCE ..)

• Design Features
  – JVM-bypassed buffer management
  – On-demand connection setup
  – RDMA or send/recv based adaptive communication
  – InfiniBand/RoCE support

Enables high performance RDMA communication, while supporting traditional socket interface
• Hadoop RPC over IB PingPong Latency
  – 1 byte: 39 us; 4 KB: 52 us
  – 42%-49% and 46%-50% improvements compared with the performance on 10 GigE and IPoIB (32Gbps), respectively

• Hadoop RPC over IB Throughput
  – 512 bytes & 48 clients: 135.22 Kops/sec
  – 82% and 64% improvements compared with the peak performance on 10 GigE and IPoIB (32Gbps), respectively
Presentation Outline

- Overview of Hadoop
- Trends in Networking Technologies and Protocols
- Challenges in Accelerating Hadoop
- Designs and Case Studies
  - HDFS
  - MapReduce
  - RPC
  - Combination of components
- Challenges in Big Data Benchmarking
- Other Open Challenges in Acceleration
- Conclusion and Q&A
Performance Benefits - TestDFSIO

- For 5GB,
  - 53.4% improvement over IPoIB, 126.9% improvement over 10GigE with HDD used for HDFS
  - 57.8% improvement over IPoIB, 138.1% improvement over 10GigE with SSD used for HDFS

- For 20GB,
  - 10.7% improvement over IPoIB, 46.8% improvement over 10GigE with HDD used for HDFS
  - 73.3% improvement over IPoIB, 141.3% improvement over 10GigE with SSD used for HDFS
• For 80GB Sort in a cluster size of 8,
  – 40.7% improvement over IPoIB, 32.2% improvement over 10GigE with HDD used for HDFS
  – 43.6% improvement over IPoIB, 42.9% improvement over 10GigE with SSD used for HDFS
Performance Benefits - TeraSort

For 100GB TeraSort in a cluster size of 8,

- 45% improvement over IPoIB, 39% improvement over 10GigE with HDD used for HDFS
- 46% improvement over IPoIB, 40% improvement over 10GigE with SSD used for HDFS
Performance Benefits – RandomWriter & Sort in SDSC-Gordon

- RandomWriter in a cluster of single SSD per node
  - 16% improvement over IPoIB (QDR) for 50GB in a cluster of 16
  - 20% improvement over IPoIB (QDR) for 300GB in a cluster of 64

- Sort in a cluster of single SSD per node
  - 20% improvement over IPoIB (QDR) for 50GB in a cluster of 16
  - 36% improvement over IPoIB (QDR) for 300GB in a cluster of 64
Performance Evaluation with BigDataBench

- Better benefit for communication intensive benchmarks, like Sort
- Smaller benefit for CPU intensive benchmarks, like WordCount
Presentation Outline

• Overview of Hadoop
• Trends in Networking Technologies and Protocols
• Challenges in Accelerating Hadoop
• Designs and Case Studies
  – HDFS
  – MapReduce
  – RPC
  – Combination of components
• Challenges in Big Data Benchmarking
• Other Open Challenges in Acceleration
• Conclusion and Q&A
Are the Current Benchmarks Sufficient for Big Data?

- The current benchmarks provide some performance behavior.
- However, do not provide any information to the designer/developer on:
  - What is happening at the lower-layer?
  - Where the benefits are coming from?
  - Which design is leading to benefits or bottlenecks?
  - Which component in the design needs to be changed and what will be its impact?
  - Can performance gain/loss at the lower-layer be correlated to the performance gain/loss observed at the upper layer?
OSU MPI Micro-Benchmarks (OMB) Suite

- A comprehensive suite of benchmarks to
  - Compare performance of different MPI libraries on various networks and systems
  - Validate low-level functionalities
  - Provide insights to the underlying MPI-level designs
- Started with basic send-recv (MPI-1) micro-benchmarks for latency, bandwidth and bi-directional bandwidth
- Extended later to
  - MPI-2 one-sided
  - Collectives
  - GPU-aware data movement
  - OpenSHMEM (point-to-point and collectives)
  - UPC
- Has become an industry standard
- Extensively used for design/development of MPI libraries, performance comparison of MPI libraries and even in procurement of large-scale systems
- Available from [http://mvapich.cse.ohio-state.edu/benchmarks](http://mvapich.cse.ohio-state.edu/benchmarks)
- Available in an integrated manner with MVAPICH2 stack
Challenges in Benchmarking of RDMA-based Designs

Big Data Middleware
(HDFS, MapReduce, HBase and Memcached)

Programming Models
(Sockets)

Applications

Benchmarks

Communication and I/O Library

Point-to-Point Communication

I/O and File Systems

Virtualization

Threaded Models and Synchronization

QoS

Fault-Tolerance

Networking Technologies
(InfiniBand, 1/10/40GigE and Intelligent NICs)

Commodity Computing System Architectures
(Multi- and Many-core architectures and accelerators)

Storage Technologies
(HDD and SSD)

Current Benchmarks

Correlation?

No Benchmarks

BPOE-4, ASPLOS ’14
Iterative Process – Requires Deeper Investigation and Design for Benchmarking Next Generation Big Data Systems and Applications
Presentation Outline

• Overview of Hadoop
• Trends in Networking Technologies and Protocols
• Challenges in Accelerating Hadoop
• Designs and Case Studies
  – HDFS
  – MapReduce
  – RPC
  – Combination of components
• Challenges in Big Data Benchmarking
• Other Open Challenges in Acceleration
• Conclusion and Q&A
Designing Communication and I/O Libraries for Big Data Systems: Solved a Few Initial Challenges

Applications

Big Data Middleware
(HDFS, MapReduce, HBase and Memcached)

Networking Technologies
(InfiniBand, 1/10/40GigE and Intelligent NICs)

Storage Technologies
(HDD and SSD)

Programming Models
(Sockets)

Point-to-Point Communication

Threaded Models and Synchronization

Communication and I/O Library

Virtualization

Fault-Tolerance

I/O and File Systems

QoS

Commodity Computing System Architectures
(Multi- and Many-core architectures and accelerators)

Networking Technologies

RDMA Protocol

Upper level Changes?

BPOE-4, ASPLOS '14
Open Challenges

• Multi-threading and Synchronization
  – Multi-threaded model exploration
  – Fine-grained synchronization and lock-free design
  – Unified helper threads for different components
  – Multi-endpoint design to support multi-threading communications

• QoS and Virtualization
  – Network virtualization and locality-aware communication for Big Data middleware
  – Hardware-level virtualization support for End-to-End QoS
  – I/O scheduling and storage virtualization
  – Live migration
Open Challenges (Cont’d)

• Support of Accelerators
  – Efficient designs for Big Data middleware to take advantage of NVIDIA GPGPUs and Intel MICs
  – Offload computation-intensive workload to accelerators
  – Explore maximum overlapping between communication and offloaded computation

• Fault Tolerance Enhancements
  – Novel data replication schemes for high-efficiency fault tolerance
  – Exploration of light-weight fault tolerance mechanisms for Big Data

• Support of Parallel File Systems
  – Optimize Big Data middleware over parallel file systems (e.g. Lustre) on modern HPC clusters
Case Study 1 - HDFS Replication: Pipelined vs Parallel

- Basic mechanism of HDFS fault tolerance is Data Replication
  - Replicates each block to multiple DataNodes

Can Parallel Replication benefit HDFS and its applications when HDFS runs on different kinds of High Performance Networks?
Parallel Replication Evaluations on Different Networks using TestDFSIO

Cluster with 8 HDD DataNodes
- 11% improvement over 10GigE
- 10% improvement over IPoIB (32Gbps)
- 12% improvement over OSU-IB (32Gbps)

Cluster with 32 HDD DataNodes
- 8% improvement over IPoIB (32Gbps)
- 9% improvement over OSU-IB (32Gbps)

Case Study 2 - Performance Improvement Potential of MapReduce over Lustre on HPC Clusters (Initial Study)

- For 500GB Sort in 64 nodes
  - 44% improvement over IPoIB (FDR)

- For 100GB Sort in 16 nodes
  - 33% improvement over IPoIB (QDR)

- Can more optimizations be achieved by leveraging more features of Lustre?
Future Plans of OSU Big Data Project

• Upcoming RDMA for Apache Hadoop Releases will support
  – HDFS Parallel Replication
  – Advanced MapReduce
  – HBase
  – Hadoop 2.0.x

• Memcached-RDMA

• Exploration of other components (Threading models, QoS, Virtualization, Accelerators, etc.)

• Advanced designs with upper-level changes and optimizations

• Comprehensive set of Benchmarks

• More details at http://hadoop-rdma.cse.ohio-state.edu
More Details on Acceleration of Hadoop..

Tutorial today afternoon (1:30-5:00pm), Canyon C

Accelerating Big Data Processing with Hadoop and Memcached on Datacenters with Modern Networking and Storage Architecture
Concluding Remarks

- Presented an overview of Big Data, Hadoop and Memcached
- Provided an overview of Networking Technologies
- Discussed challenges in accelerating Hadoop
- Presented initial designs to take advantage of InfiniBand/RDMA for HDFS, MapReduce, RPC and HBase
- Results are promising
- Many other open issues need to be solved
- Will enable Big Data community to take advantage of modern HPC technologies to carry out their analytics in a fast and scalable manner
Personnel Acknowledgments

**Current Students**
- S. Chakraborthy (Ph.D.)
- N. Islam (Ph.D.)
- J. Jose (Ph.D.)
- M. Li (Ph.D.)
- S. Potluri (Ph.D.)
- R. Rajachandrasekhar (Ph.D.)

**Past Students**
- P. Balaji (Ph.D.)
- D. Buntinas (Ph.D.)
- S. Bhagvat (M.S.)
- L. Chai (Ph.D.)
- B. Chandrasekharan (M.S.)
- N. Dandapanthula (M.S.)
- V. Dhanraj (M.S.)
- T. Gangadharappa (M.S.)
- K. Gopalakrishnan (M.S.)

**Past Post-Docs**
- H. Wang
- X. Besseron
- H.-W. Jin
- M. Luo

**Past Programmers**
- D. Bureddy

**Current Senior Research Associate**
- H. Subramoni

**Current Post-Docs**
- X. Lu
- K. Hamidouche

**Current Programmers**
- M. Arnold
- J. Perkins

**Past Research Scientist**
- S. Sur

**Past Research Scientist**
- E. Mancini
- S. Marcarelli
- J. Vienne

**Past Post-Docs**
- M. Luo (Ph.D.)
- W. Huang (Ph.D.)
- W. Jiang (M.S.)
- S. Kini (M.S.)
- M. Koop (Ph.D.)
- R. Kumar (M.S.)
- S. Krishnamoorthy (M.S.)
- K. Kandalla (Ph.D.)
- P. Lai (M.S.)
- J. Liu (Ph.D.)

**Past Programmers**
- G. Santhanaraman (Ph.D.)
- A. Singh (Ph.D.)
- J. Sridhar (M.S.)
- S. Sur (Ph.D.)
- H. Subramoni (Ph.D.)
- K. Vaidyanathan (Ph.D.)
- A. Vishnu (Ph.D.)
- J. Wu (Ph.D.)
- W. Yu (Ph.D.)
Pointers

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

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

http://hadoop-rdma.cse.ohio-state.edu

panda@cse.ohio-state.edu

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