Performance Analysis and Evaluation of InfiniBand FDR and 40GigE RoCE on HPC and Cloud Computing Systems

Jerome Vienne  Jitong Chen  Md. Wasi-ur Rahman
Nusrat S. Islam  Hari Subramoni
Dhabaleswar K. (DK) Panda

Network-Based Computing Laboratory
Department of Computer Science and Engineering
The Ohio State University

IEEE Hot Interconnects
August 23, 2012
Outline

1. Introduction
2. Cloud Computing Applications
3. Performance Analysis and Evaluation
4. Conclusion
Outline

1. Introduction
2. Cloud Computing Applications
3. Performance Analysis and Evaluation
4. Conclusion
Introduction

- Commodity clusters continue to be very popular in HPC and clouds
- HPC applications and cloud computing middleware (e.g. Hadoop) have varying communication and computation characteristics
- New PCIe Gen3 interface can now deliver speeds up to 128 Gbps
- High performance interconnects are capable of delivering speeds up to 54 Gbps
  - New Mellanox’s ConnectX-3 FDR (54 Gbps)/RoCE 40 GigE
Overview of Network Protocol Stacks

- RoCE: allows the RDMA of InfiniBand to run over Ethernet.
- ConnectX-2: 10 GigE in RoCE mode or QDR (32 Gbps) in IB mode
- ConnectX-3: 40 GigE in RoCE mode or FDR (54 Gbps) in IB mode
Problem Statement

• How much benefit can the user of a HPC / Cloud installation hope to see by utilizing IB FDR / RoCE 40 GigE over IB QDR and RoCE 10 GigE interconnects, respectively?

• How does InfiniBand compare with RoCE in terms of performance?
Outline

1. Introduction
2. Cloud Computing Applications
3. Performance Analysis and Evaluation
4. Conclusion
Cloud computing economies have gained significant momentum and popularity.

Required the highest performance and reliability available.

Apache Hadoop is the most popular framework for running applications on large clusters built of commodity hardware.

Major components of Hadoop used:
  - HDFS
  - HBase
HDFS

- Hadoop Distributed File System (HDFS) is the underlying file system for Hadoop framework
- HDFS is designed for storing very large files on clusters of commodity hardware
- Two main types of nodes:
  - NameNode: responsible for storing and managing the metadata
  - DataNode: act as storage for HDFS files
- Files are usually divided into fixed-sized (64 MB) blocks and stored as independent units
- Each block is also replicated to multiple (typically three) DataNodes in order to provide fault tolerance and availability
HBase and YCSB

HBase

- Developed as part of the Apache Hadoop project
- Java-based database
- Runs on top of the Hadoop framework
- Used to host very large tables with many billions of entries
- Provides capabilities similar to Google’s BigTable

Yahoo! Cloud Serving Benchmark

- Used as our workload
- Facilitates performance comparisons of different key/value-pair and cloud data serving systems
- Defines a core set of benchmarks for four widely used systems: HBase, Cassandra, PNUTS and a simple shared MySQL implementation.
Outline

1. Introduction
2. Cloud Computing Applications
3. Performance Analysis and Evaluation
4. Conclusion
Experimental Testbed

4-node InfiniBand Linux cluster

- 16 cores/node – 2 Intel Sandy Bridge-EP 2.6 Ghz CPUs
- 32 GB main memory, 20 MB L3 shared cache
- 1 PCIe Gen3 (128 Gbps)
- Vendor modified version of OFED based on OFED-1.5.3

Network Equipment

- IB cards:
  - ConnectX-2 QDR (32 Gbps) / 10 GigE
  - ConnectX-3 FDR (54 Gbps) / 40 GigE
- 36-port Mellanox FDR switch used to connect all the nodes
Performance Results

Network Level Performance
- Latency
- Bandwidth

MPI Level Performance
- Point-to-point MPI
- MPI Collectives
- NAS Parallel Benchmarks

Impact on Cloud Computing Middlewares
- HDFS Write using TestDFSIO
- HBase Get and Put throughput
Network level latency benchmark (ib_send_lat)
- IB FDR provides best performance
- IB QDR gives better latency than 40GigE
Bandwidth

- Network level bandwidth benchmark (ib_send_bw)
- 40 GigE gives better bandwidth than IB QDR
  - Encoding: IB QDR (8/10) vs 40 GigE (64/66)
MVAPICH2 Software

High Performance MPI Library for IB and 10/40GE

- Used by more than 1,930 organizations in 68 countries
- More than 124,000 direct downloads from OSU site
- Empowering many TOP500 clusters
  - 11th ranked 81,920-core cluster (Pleiades) at NASA
  - 14th ranked 73,278-core (Tsubame 2.0) at Tokyo Institute of Technology
  - 40th ranked 62,976-core cluster (Ranger) at TACC
- Available with software stacks of many IB, 10/40GE and server vendors including Open Fabrics Enterprise Distribution (OFED)
- Also supports uDAPL device (for networks supporting uDAPL)
- http://mvapich.cse.ohio-state.edu/
• MPI level latency benchmark (OMB: osu_latency)
• IB FDR provides best performance
• IB QDR gives better latency than 10/40GigE
• MPI level bandwidth benchmark (OMB: osu_bw)
• 40 GigE gives better bandwidth than IB QDR
  • Encoding: IB QDR (8/10) vs RoCE 40 GigE (64/66)
MPI Collective: Scatter

- MPI level collective benchmark (OMB: osu_scatter)
- IB FDR provides best performance
NAS Parallel Benchmarks Class C

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>IB (QDR)</th>
<th>IB (FDR)</th>
<th>RoCE (10 GigE)</th>
<th>RoCE (40 GigE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FT</td>
<td>9.96s</td>
<td>8.80s</td>
<td>14.39s</td>
<td>9.71s</td>
</tr>
<tr>
<td>IS</td>
<td>0.80s</td>
<td>0.64s</td>
<td>1.32s</td>
<td>0.71s</td>
</tr>
<tr>
<td>MG</td>
<td>2.02s</td>
<td>1.98s</td>
<td>2.20s</td>
<td>1.99s</td>
</tr>
<tr>
<td>BT</td>
<td>24.79s</td>
<td>24.74s</td>
<td>26.23s</td>
<td>24.83s</td>
</tr>
</tbody>
</table>

- Design to mimic computation and data movement in CFD applications
- FT, IS: Communication bound
- MG, BT: Computation bound
TestDFSIO

- File system benchmark that measures the I/O performance of HDFS
- *HDFS Write* is more network sensitive compared to *HDFS Read* (occurs locally in a node in most of the cases)
- In sequential write, each map task opens a file and writes specific amount of data to the file.
- A single reduce task aggregates the results of all the map tasks

**Protocol**

- We start two map tasks each writing a file to three DataNodes
- We vary the file size from 1 GB to 10 GB
- We measure the throughput of sequential write reported by TestDFSIO
Due to the higher bandwidth of IPoIB (FDR) system, sequential write provides better throughput for all the file sizes compared to IPoIB (QDR).

- Up to 19% benefit for IPoIB (FDR) over IPoIB (QDR)
- The throughput of sequential write is improved by 31% over Sockets (40 GigE) compared to Sockets (10 GigE)
HBase evaluation

- Using YCSB as our workload, we perform 100% Get, 100% Put and a 50% Get and Put Mix operations.
  - HBase Get operation requires less network communication.
  - HBase Put creates more network traffic (all the data are written to both MemStore and HDFS)
  - Mix Get and Put generates network traffic (some old data are replaced in MemStore by the new ones each time)
- Three region servers are used.
- The region servers communicate with the master (HDFS NameNode) and the HBase client through the underlying interconnect.
- Usually region servers are configured to reside in the same nodes as HDFS DataNodes, to improve data locality.
- For these workloads, we have used 320,000 records to be inserted to and read from HBase.
HBase *Get* and *Put* throughput

- 9% benefit for IPoIB (FDR) with Get-Put-Mix
- Up to 10% benefit for IPoIB (FDR) with 100% Put
- Overall, IPoIB (FDR) 25% better than Sockets (40GigE)
Performance Characterization

Network Level Performance

IB FDR > RoCE 40 GigE > IB QDR > RoCE 10 GigE

Performance of HPC Applications

IB FDR > RoCE 40 GigE > IB QDR > RoCE 10 GigE

Performance of Cloud Computing Middlewares

IPoIB FDR > IPoIB QDR > Sockets 40 GigE > Sockets 10 GigE
Outline

1. Introduction
2. Cloud Computing Applications
3. Performance Analysis and Evaluation
4. Conclusion
Conclusion

- Carried out a comprehensive performance evaluation of four possible modes of communication
- Latest InfiniBand FDR interconnect gives the best performance
- Network level evaluations and for HPC applications: RoCE 40 GigE performance better than IB QDR
- Cloud computing middleware: IPoIB QDR performance better than RoCE 40 GigE
Thanks for your attention
Questions?

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

MVAPICH Web Page
http://mvapich.cse.ohio-state.edu/