RDMA over Ethernet - A Preliminary Study

Hari Subramoni, Miao Luo, Ping Lai and Dhabaleswar. K. Panda

Computer Science & Engineering Department
The Ohio State University

HPIDC '09
Outline

• Introduction
• Problem Statement
• Approach
• Performance Evaluation and Results
• Conclusions and Future Work
Introduction

- Ethernet and InfiniBand accounts for majority of interconnects in high performance distributed computing
- End users want InfiniBand like latencies with existing Ethernet infrastructure
- Can be achieved if networks converge
- Existing options have overhead or tradeoffs in terms of performance
- No solution exists that efficiently combines the ubiquitous nature of Ethernet and the high performance offered by InfiniBand
- RDMA over Ethernet (RDMAoE) seems to provide a good option as of date

HPIDC '09
RDMAoE

- Allows running the IB transport protocol using Ethernet frames
- RDMAoE packets are standard Ethernet frames with an IEEE assigned Ethertype, a GRH, unmodified IB transport headers and payload
- InfiniBand HCA takes care of translating InfiniBand addresses to Ethernet addresses and back
- Encodes IP addresses into its GIDs and resolves MAC addresses using the host IP stack
- Use GID’s for establishing connections instead of LID’s
- No SM/SA, Ethernet management practices are used
InfiniBand Architecture & Adapters

• An industry standard for low latency, high bandwidth, System Area Networks
• Multiple features
  – Two communication types
    • Channel Semantics
    • Memory Semantics (RDMA mechanism)
  – Multiple virtual lanes
  – Quality of Service (QoS) support
• Double Data Rate (DDR) with 20 Gbps bandwidth has been there
• Quad Data Rate (QDR) with 40 Gbps bandwidth is available recently
• Multiple generations of InfiniBand adapters are available now
• The latest ConnectX DDR adapters provide support for both IB as well as RDMAoE modes

HPIDC '09
Modes of Communication using ConnectX DDR Adapter

Applications

TCP/IP
- ConnectX
  - Ethernet Switch
  - TCP/IP

IPoIB
- ConnectX
  - IB Switch
  - IPoIB

Open Fabrics
- Verbs (with RDMA)
  - ConnectX
    - IB Switch
    - Native IB
- Verbs (with RDMA)
  - ConnectX
    - Ethernet Switch
    - RDMAoE

Application Protocol

Adapter

Switch

Network Protocol

HPIDC '09
Outline

- Introduction
- Problem Statement
- Approach
- Performance Evaluation and Results
- Conclusions and Future Work
Problem Statement

• How do the different communication protocols stack up against each other as far
  – Raw sockets / verbs level performance
  – Performance for MPI applications
  – Performance for Data center applications
• Does RDMAoE bring us a step closer to the goal of network convergence
Outline

• Introduction
• Problem Statement
• Approach
• Performance Evaluation and Results
• Conclusions and Future Work

HPIDC '09
Approach

• Protocol level benchmarks to evaluate very basic performance
• MPI level benchmarks to evaluate basic MPI performance at both point to point and collective levels
• Application level benchmarks to evaluate performance of real world applications
• Evaluation using common data center applications
Outline

• Introduction
• Problem Statement
• Approach
• Performance Evaluation and Results
• Conclusions and Future Work
Experimental Testbed

• Compute Platform
  – Intel Nehalem
    • Intel Xeon E5530 Dual quad-core processors operating at 2.40 GHz
    • 12GB RAM, 8MB cache
    • PCIe 2.0 interface
• Host Channel Adapter
  – Dual port ConnectX DDR adapter
    • Configured in either RDMAoE mode or IB mode
• Network Switches
  – 24 port Mellanox IB DDR switch
  – 24 port Fulcrum Focalpoint 10GigE switch
• OFED version
  – OFED-1.4.1 for IB and IPoIB
  – Pre-release version of OFED-1.5 for RDMAoE and TCP / IP
• MPI version – MVAPICH-1.1 and MPICH-1.2.7p1

HPIDC '09
MVAPICH / MVAPICH2 Software

- High Performance MPI Library for IB and 10GE
  - MVAPICH (MPI-1) and MVAPICH2 (MPI-2)
  - Used by more than 960 organizations in 51 countries
  - More than 32,000 downloads from OSU site directly
  - Empowering many TOP500 clusters
    - 8th ranked 62,976-core cluster (Ranger) at TACC
  - Available with software stacks of many IB, 10GE and server vendors including Open Fabrics Enterprise Distribution (OFED)
  - Also supports uDAPL device to work with any network supporting uDAPL
    - http://mvapich.cse.ohio-state.edu/

HPIDC '09
List of Benchmarks

- OSU Microbenchmarks (OMB)
  - Version 3.1.1
  - [http://mvapich.cse.ohio-state.edu/benchmarks/](http://mvapich.cse.ohio-state.edu/benchmarks/)

- Intel Collective Microbenchmarks (IMB)
  - Version 3.2

- NAS Parallel Benchmarks (NPB)
  - Version 3.3
  - [http://www.nas.nasa.gov/](http://www.nas.nasa.gov/)
Verbs Level Evaluation
Inter-Node Latency

<table>
<thead>
<tr>
<th>Message Size (Bytes)</th>
<th>Native IB</th>
<th>RDMAoE</th>
<th>TCP/IP</th>
<th>IPoIB</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>25.5</td>
<td>3.03</td>
<td>1.66</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>29.9</td>
<td>2.91</td>
<td>1.79</td>
<td></td>
</tr>
<tr>
<td>64</td>
<td>31.7</td>
<td>2.97</td>
<td>1.84</td>
<td></td>
</tr>
<tr>
<td>256</td>
<td>33.7</td>
<td>3.03</td>
<td>1.91</td>
<td></td>
</tr>
<tr>
<td>1K</td>
<td>35.0</td>
<td>3.09</td>
<td>1.99</td>
<td></td>
</tr>
</tbody>
</table>

- For small messages
  - Native IB verbs offers best latency of **1.66 us**
  - RDMAoE comes very close to this at **3.03 us**

HPIDC '09
MPI Level Evaluation
Inter-Node Latency

- For small messages
  - Native IB verbs offers best latency of 1.8 \textit{us}
  - RDMAoE comes very close to this at 3.6 \textit{us}

HPIDC '09
Inter-Node Multipair Latency

- 4 pairs of processes communicating simultaneously
- For small messages
  - Native IB verbs offers best latency of 1.66 $us$
  - RDMAoE comes very close to this at 3.51 $us$

HPIDC '09
Collective Performance
Allgater Latency (32-cores)

• For small messages
  • Native IB verbs offers best latency of 12.97 \text{us}
  • RDMAoE comes very close to this at 22.71 \text{us}

HPIDC '09
For small messages

- Native IB verbs offers best latency of 10.05 us
- RDMAoE comes very close to this at 12.78 us
Performance of NAS Benchmarks

- 32 process, Class C
- Numbers normalized to Native-IB
- Performance of Native IB and RDMAoE are very close with Native IB giving the best performance

HPIDC '09
Evaluation of Data Center Applications

- We evaluate FTP, a common data center application
- We use our own version of FTP over native IB verbs (FTP-ADTS [2]) to evaluate RDMAoE and Native IB
- GridFTP [1] is used to evaluate performance of TCP/IP and IPoIB
- RDMAoE shows performance comparable to Native IB


HPIDC ‘09
Outline

- Introduction
- Problem Statement
- Approach
- Performance Evaluation and Results
- Conclusions and Future Work
Conclusions & Future Work

• Perform comprehensive evaluation of all possible modes of communication (Native IB, RDMAoE, TCP/IP, IPoIB) using
  – Verbs
  – MPI
  – Application and,
  – Data center level experiments

• Native IB gives the best performance followed by RDMAoE

• RDMAoE provides a high performance solution to the problem of network convergence

• As part of future work, we plan to
  – Perform large scale evaluations including studies into the effect of network contention on the performance of these protocols
  – Study these protocols in a comprehensive manner for file systems

HPIDC ’09
Thank you!

{subramon, laipi, luom, panda}@cse.ohio-state.edu

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