Experimental Analysis of InfiniBand and 10GigE Technologies Over Wide-Area Connections

Nagi Rao, Bill Wing, Steve Poole, Susan Hicks, Steve Hodson
Galen Shipman, Josh Lothian
Oak Ridge National Laboratory

JET – Sept 16, 2008

Sponsored by
Department of Defense
Department of Energy
Main Question: Best ways to support high throughput data transport between remote high-performance computing, storage and analysis systems

Technical Topic Addressed:
Wide-area data transfers at 10Gbps rates at thousands of miles:
Robust end-to-end solutions: physical- through transport-layers

Significant Technical Challenges:
A. Performance of conventional TCP/IP solutions is unclear:
   1. Many TCP variants: complex to implement, analyze and test
   2. Enormous in-situ efforts are needed per connection basis to reach multiple Gbps rates
B. Novel Infiniband solutions seem promising but need to be studied:
   1. Very limited wide-area results are available
   2. Objective side-by-side comparisons with TCP/IP are not available
   3. Require capable test environments
Approach and Results

Specific Topics Discussed:
1. Capability- and capacity-based high-performance testbed:
   - Collection of high-end hosts, storage and computing systems
   - Flexible 10Gbps, 8600 mile network core
2. Throughput performance of IB over wide-area connections (SONET, 10GigE)
3. Throughput performance of 10GigE + TCP high-performance variants

Overall Task Summary:
1. Scalability and plug-in capability of IB are demonstrated over 8600 miles
2. Performance testing of recent high-performance TCP over Ethernet
3. Objective side-by-side comparison of these two technologies

Technical Results Summary: In a nutshell, over 8600 mile connection
Key Measurement – Throughput Decrease Per Mile
• IB Over SONET/10GigE: 0.02Mbps/mile
• 10GigE-HTCP: 1.3Mbps/mile
UltraScience Net: Experimental network research testbed: for advanced networking and associated technologies for high-performance

Features

- End-to-end guaranteed bandwidth channels
- Dynamic, in-advance, reservation and provisioning of fractional/full lambdas
- Secure control-plane for signaling
- Peering with ESnet, National Science Foundation CHEETAH, and other networks
USN data-plane: Node configuration

- **In the core:**
  - Two OC192 switched by Ciena CDCIs

- **At the edge:**
  - 10/1 GigE provisioning using Force10 E300s

**Data plane user connections:**

- Direct connections to
  - Core switches—SONET and 1 GigE
  - MSPP—Ethernet channels
- Utilize UltraScience Net hosts
Infiniband Over SONET: Obsidian Longbows
RDMA throughput measurements over USN

ORNL loop -0.2 mile: 7.48 Gbps
ORNL-Chicago loop – 1400 miles: 7.47 Gbps
ORNL- Chicago - Seattle loop – 6600 miles: 7.37 Gbps
ORNL – Chicago – Seattle - Sunnyvale loop – 8600 miles: 7.34 Gbps

Hosts:
dual-socket quad-core 2GHz AMD Opteron, 4GB memory
8-lane PCI-Express slot
Dual-port Voltaire 4x SDR HCA.
Performance Profiles – IB RDMA Throughputs

• Throughput Distance Profile
  – Plot throughput as a function connection length \((d)\) and message size \((s)\)
  – \(B=\text{SONET, WAN-PHY}\)
    \[ T_B(d, s) \]

• Throughput Stability Profile
  – Plot throughput as function of connection length and repetition number for fixed message size
    \[ T_B(d, s) \quad \text{and} \quad \hat{T}_B(d) \]
  – Average throughput over 10 iterations with 8M message size
    \[ \hat{T}_B(d) \]

• Throughput Decrease Per Mile (DPM): at connection length \(d_i\)
  \[ \hat{D}_B(d_i) = \frac{\hat{T}_B(d_0) - \hat{T}_B(d_i)}{d_i - d_0} \]
Distance and Stability Profiles of IB over SONET

Measurements using ib_rdma-bw – c
It uses IB CM for connection setup and management

Distance profile

\[ T_B(d, s) \]

Stability profile

8M message size

\[ \overline{T}_B(d) \]

<table>
<thead>
<tr>
<th>Connection length (miles)</th>
<th>( d_i )</th>
<th>0.2</th>
<th>1400</th>
<th>6600</th>
<th>8600</th>
</tr>
</thead>
<tbody>
<tr>
<td>Throughput (Gbps) – 8M msg</td>
<td></td>
<td>7.48</td>
<td>7.47</td>
<td>7.37</td>
<td>7.34</td>
</tr>
<tr>
<td>Std-dev (Mbps)</td>
<td></td>
<td>45.27</td>
<td>0.07</td>
<td>0.09</td>
<td>0.07</td>
</tr>
<tr>
<td>DPM (Mbps) ( D_B(d_i) )</td>
<td></td>
<td>0</td>
<td>0.012</td>
<td>0.017</td>
<td>0.016</td>
</tr>
</tbody>
</table>
Wide-Area Connections: SONET OC192 and 10GigE WAN/LAN-PHY

SONET:
- Widely deployed over long-haul optical backbone networks
- OC192 over DWDM: 9.6Gbps over single wavelength - robust well-understood technology
- Utilizes Time-Division Multiplexing to provide well separated sub-lambdas – 4*OC48 or 64*OC3

10Gig Ethernet:
- Relatively recent technology: high-potential for wide deployment
- WAN-PHY – Ethernet frames are packed into SONET OC192c payload: peak 9.6 Gbps
- LAN-PHY – Natively transports Ethernet packets: full 10Gbps

A Comparison:
- SONET is widely deployed but needs more expensive infrastructure
- Sub-channel separation is more robust in SONET than 10GigE
- 10GigE more naturally transits between LAN to WAN environments
IB over 10GigE LAN-PHY and WAN-PHY

- ORNL loop -0.2 mile
- ORNL-Chicago loop – 1400 miles
- ORNL- Chicago - Seattle loop – 6600 miles
- ORNL – Chicago – Seattle - Sunnyvale loop – 8600 miles
Performance Profiles of IB Over 10GigE WAN-PHY

Results are almost the same as in SONET case

<table>
<thead>
<tr>
<th>Connection length (miles)</th>
<th>( d_i )</th>
<th>0.2</th>
<th>1400</th>
<th>6600</th>
<th>8600</th>
</tr>
</thead>
<tbody>
<tr>
<td>Throughput (Gbps) – 8M msg</td>
<td></td>
<td>7.5</td>
<td>7.49</td>
<td>7.39</td>
<td>7.36</td>
</tr>
<tr>
<td>Std-dev (Mbps)</td>
<td></td>
<td>0.07</td>
<td>0.69</td>
<td>0.00</td>
<td>0.20</td>
</tr>
<tr>
<td>DPM (Mbps) ( \hat{D}_B(d_i) )</td>
<td></td>
<td>0</td>
<td>0.012</td>
<td>0.017</td>
<td>0.016</td>
</tr>
</tbody>
</table>

\[ T_B(d, s) \]

distance profile

peak distance profile

average distance profile
Cross-Traffic Generation

- ORNL loop - 0.2 mile
- ORNL-Chicago loop – 1400 miles
- ORNL-Chicago - Seattle loop – 6600 miles
- ORNL – Chicago – Seattle - Sunnyvale loop – 8600 miles

Network Connections:
- OC192
- 10 GigE WAN-PHY
- 10 GigE LAN-PHY
- 1GigE

Hosts and Connections:
- Host 1
- Host 2
- USN
- ORNL
- Chicago
- Seattle
- Sunnyvale

Network Types:
- IB 4x
- quad-core
- dual socket
- dual-core
- single socket
- Netron10GigE
Cross-Traffic Effect of IB over 10GigE WANPHY

Competing traffic: UDP streams on WAN at 1,2,3,4 Gbps
- Distance profiles are unaffected for cross-traffic levels of up to 1Gbps
- IB throughput was drastically effected at cross-traffic level of 4 Gbps
- Effect of cross-traffic is more on large message sizes

Average throughput for 8M

<table>
<thead>
<tr>
<th>miles</th>
<th>1400</th>
<th>6600</th>
<th>8600</th>
</tr>
</thead>
<tbody>
<tr>
<td>0G</td>
<td>7.49</td>
<td>7.39</td>
<td>7.36</td>
</tr>
<tr>
<td>1G</td>
<td>7.49</td>
<td>7.39</td>
<td>7.36</td>
</tr>
<tr>
<td>2G</td>
<td>3.13</td>
<td>1.38</td>
<td>0.74</td>
</tr>
<tr>
<td>3G</td>
<td>3.25</td>
<td>1.97</td>
<td>1.02</td>
</tr>
<tr>
<td>4G</td>
<td>2.91</td>
<td>1.82</td>
<td>0.96</td>
</tr>
</tbody>
</table>
Challenges of Assessing TCP Transport

TCP or similar transport method is needed to ensure reliable data delivery over 10GigE provisioned connections

Numerous TCP variants for High-Performance:
1. Standard slow-start and Additive-Increase and Multiplicative Decrease (AIMD) are not effective for high performance operation
2. A variety of TCP variations have been proposed for high performance operation – difficult to implement, analyze and test
3. All of them require multiple streams and significant amount of tuning to provide multi-Gbps throughputs

Testing TCP variants:
1. Congestion control dynamics are very complex and not amenable to simple analytical treatment – need complementary experiments
2. Dynamically loadable congestion control modules in linux 2.6.18 kernels: auto-tuning is effective in buffer management BIC, CUBIC, Hamilton TCP (HTCP), Scalable TCP, Highspeed TCP, TCP Vegas
3. Measurements show high variance – robust measurement and analysis methods are needed
10GigE Connections

- **ORNL loop** - 0.2 mile
- **ORNL-Chicago loop** - 1400 miles
- **ORNL-Chicago - Seattle loop** - 6600 miles
- **ORNL – Chicago – Seattle - Sunnyvale loop** - 8600 miles

- **10 GigE WAN-PHY**
- **10 GigE LAN-PHY**
- **OC192**

**ORNL**

- **Quad-core Dual socket**
- **host**
- **Myrinet 10 GigE NICS Over PCI-Express**

- **ORNL E300**
- **ORNL CDCI**
- **Chicago CDCI**
- **Seattle CDCI**
- **Sunnyvale CDCI**
- **Chicago E300**
- **Sunnyvale E300**
Performance Profiles – TCP Throughputs

BIC and Hamilton TCP – pluggable Linux modules

- **Throughput Distance Profile**
  - Plot throughput as a function connection length \((d)\) and number of streams \((s)\)
  - \(A=\text{BIC,HTCP}\)
  \[
  T_A(d,n)
  \]

- **Throughput Stability Profile**
  - Plot throughput as function of connection length and repetition number of streams
  - Average throughput over repetitions and range of number of streams 15-20
  \[
  \hat{T}_B(d)
  \]

- **Throughput Decrease Per Mile**
  \[
  \hat{D}_A(d_i) = \frac{\hat{T}_A(d_0) - \hat{T}_A(d_i)}{d_i - d_0}
  \]
Performance of TCP over 10GigE BIC with Linux auto-tuning

<table>
<thead>
<tr>
<th>Connection length (miles)</th>
<th>$d_i$</th>
<th>0.2</th>
<th>1400</th>
<th>6600</th>
<th>8600</th>
</tr>
</thead>
<tbody>
<tr>
<td>Throughput (Gbps) – 8M msg</td>
<td></td>
<td>9.12</td>
<td>6.69</td>
<td>0.76</td>
<td>0.50</td>
</tr>
<tr>
<td>Std-dev (Mbps)</td>
<td></td>
<td>64.11</td>
<td>70.08</td>
<td>24.96</td>
<td>21.08</td>
</tr>
<tr>
<td>DPM (Mbps) $\hat{D}_B(d_i)$</td>
<td></td>
<td>0</td>
<td>1.74</td>
<td>1.27</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Better than IB for local connections
Performance of TCP over 10GigE
Hamilton TCP with Linux auto-tuning

<table>
<thead>
<tr>
<th>Connection length (miles) $d_i$</th>
<th>0.2</th>
<th>1400</th>
<th>6600</th>
<th>8600</th>
</tr>
</thead>
<tbody>
<tr>
<td>Throughput (Gbps) – 8M msg</td>
<td>9.21</td>
<td>6.71</td>
<td>1.22</td>
<td>1.79</td>
</tr>
<tr>
<td>Std-dev (Mbps)</td>
<td>12.25</td>
<td>37.42</td>
<td>18.96</td>
<td>128.15</td>
</tr>
<tr>
<td>DPM (Mbps) $\hat{D}_B(d_i)$</td>
<td>0</td>
<td>1.79</td>
<td>1.21</td>
<td>0.87</td>
</tr>
</tbody>
</table>
Comparative Performance of BIC and Hamilton TCP

Multiple streams are needed to get high throughputs

We also tested
  Highspeed TCP
  Scalable-TCP
  TCP-Vegas
not as good as BIC and HCTP
Conclusions and Results

We conducted structured experiments to assess:
1. Throughput performance of IB RDMA over wide-area connections
   SONET and 10GigE
2. Throughput performance of 10GigE + TCP high performance versions

Experimental Results:
1. Scalability and plug-in capability of IB ~7.3 Gbps over 8600 miles
2. Performance testing of recent high-performance TCP over Ethernet:
   best performance is ~2Gbps at 8600 miles, but above 9Gbps locally
3. Side-by-side comparison over 8600 mile connection:
   • IB Over SONET/10GigE: 0.02Mbps/mile
   • 10GigE-HTCP: 1.3Mbps/mile
4. Cross-traffic effects:
   IB performance is robust with upto1Gbps cross-traffic either on WAN
   connection or using their own 1GigE ports.
   TCP is less prone to such effects but more testing is needed.
Thank you