ECLIPSE 2012
Performance Benchmark and Profiling

August 2012
• The following research was performed under the HPC Advisory Council activities
  – Participating vendors: Intel, Dell, Mellanox
  – Compute resource - HPC Advisory Council Cluster Center

• The following was done to provide best practices
  – ECLIPSE performance overview
  – Understanding ECLIPSE communication patterns
  – Ways to increase ECLIPSE productivity
  – MPI libraries comparisons

• For more info please refer to
  – http://www.dell.com
  – http://www.intel.com
  – http://www.mellanox.com
Schlumberger ECLIPSE

- **Oil and gas reservoir simulation software**
  - Developed by Schlumberger

- **Offers multiple choices of numerical simulation techniques for accurate and fast simulation for**
  - Black-oil
  - Compositional
  - Thermal
  - Streamline
  - Others

- **ECLIPSE support MPI to achieve high performance and scalability**
Objectives

- The following was done to provide best practices
  - ECLIPSE performance benchmarking
  - Interconnect performance comparisons
  - MPI performance comparison
  - Understanding ECLIPSE communication patterns

- The presented results will demonstrate
  - The scalability of the compute environment to provide nearly linear application scalability
  - The capability of ECLIPSE to achieve scalable productivity
Test Cluster Configuration

- **Dell™ PowerEdge™ R720xd 16-node (256-core) “Jupiter” cluster**
  - Dual-Socket Eight-Core Intel E5-2680 @ 2.70 GHz CPUs (Static max Perf in BIOS)
  - Memory: 64GB memory, DDR3 1600 MHz
  - OS: RHEL 6.2, OFED 1.5.3 InfiniBand SW stack
  - Hard Drives: 24x 250GB 7.2 RPM SATA 2.5” on RAID 0

- **Intel Cluster Ready certified cluster**

- **Mellanox ConnectX-3 FDR InfiniBand VPI adapters**

- **Mellanox SwitchX SX6036 InfiniBand switch**

- **MPI: Intel MPI 4 Update 3, Platform MPI 8.1.2**

- **Application: Schlumberger ECLIPSE 2012.1**

- **Benchmarks:**
  - Four million cell model (FOURMILL.DATA)
About Intel® Cluster Ready

- Intel® Cluster Ready systems make it practical to use a cluster to increase your simulation and modeling productivity
  - Simplifies selection, deployment, and operation of a cluster

- A single architecture platform supported by many OEMs, ISVs, cluster provisioning vendors, and interconnect providers
  - Focus on your work productivity, spend less management time on the cluster

- Select Intel Cluster Ready
  - Where the cluster is delivered ready to run
  - Hardware and software are integrated and configured together
  - Applications are registered, validating execution on the Intel Cluster Ready architecture
  - Includes Intel® Cluster Checker tool, to verify functionality and periodically check cluster health
PowerEdge R720xd
Massive flexibility for data intensive operations

- **Performance and efficiency**
  - Intelligent hardware-driven systems management with extensive power management features
  - Innovative tools including automation for parts replacement and lifecycle manageability
  - Broad choice of networking technologies from GigE to IB
  - Built in redundancy with hot plug and swappable PSU, HDDs and fans

- **Benefits**
  - Designed for performance workloads
    - from big data analytics, distributed storage or distributed computing where local storage is key to classic HPC and large scale hosting environments
    - High performance scale-out compute and low cost dense storage in one package

- **Hardware Capabilities**
  - Flexible compute platform with dense storage capacity
    - 2S/2U server, 6 PCIe slots
  - Large memory footprint (Up to 768GB / 24 DIMMs)
  - High I/O performance and optional storage configurations
    - HDD options: 12 x 3.5” or 24 x 2.5 + 2x 2.5 HDDs in rear of server
    - Up to 26 HDDs with 2 hot plug drives in rear of server for boot or scratch
ECLIPSE Performance – Processors

- Intel E5-2680 (Sandy Bridge) cluster outperforms prior generations
  - Performs 83% better than X5670 cluster at 16 nodes
- **System components used:**
  - Jupiter: 2-socket Intel E5-2680 @ 2.7GHz, 1600MHz DIMMs, FDR IB, 24 disks
  - Janus: 2-socket Intel X5670 @ 2.93GHz, 1333MHz DIMMs, QDR IB, 1 disk

ECLIPSE 2012 Performance (FOURMILL)

![Chart showing performance comparison between Jupiter and Janus across different node counts.

- Higher is better

InfiniBand FDR
ECLIPSE Performance – MPI

• **Platform MPI outperforms Intel MPI at larger scale**
  – Up to 33% higher performance than Intel MPI at 16-node
  – No change in work done with Intel MPI between 8 and 16 nodes

• **CPU binding optimization flag used in all cases shown**
  – No other optimization flags are used
ECLIPSE Performance – Ethernet Interconnect

- **40GbE-RoCE provides better scalability performance than Ethernet**
  - provides up to 170% better performance than 10GbE at 16-node
  - provides up to 65% better performance than 40GbE at 16-node
  - provides up to 43% better performance than 10GbE-RoCE at 16-node

*Higher is better* 16 Processes/Node
ECLIPSE Performance – InfiniBand Interconnect

- InfiniBand FDR delivers better application performance
  - Up to 32% better performance than InfiniBand QDR

**ECLIPSE 2012 Performance**
*(FOURMILL, Platform MPI)*

**Relative Performance (Jobs/Day)**

<table>
<thead>
<tr>
<th>Number of Nodes</th>
<th>Relative Base</th>
<th>InfiniBand QDR</th>
<th>InfiniBand FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
<td>0.6</td>
<td>0.75</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0.8</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>16</td>
<td>1</td>
<td>3.2</td>
<td>4</td>
</tr>
</tbody>
</table>

32%

*Higher is better*
InfiniBand FDR allows larger percentage time spent on computation
- InfiniBand FDR provides faster data transfers, higher CPU utilization
- Other interconnects would waste system time on data transfers and utilization
The most used MPI calls is MPI_Waitany
- The next most used calls are point-to-point APIs:
- For example: MPI_Isend, MPI_Irecv and MPI_Recv

Higher is better
ECLIPSE Profiling – MPI Comm. Time

- Majority of MPI communication time is spent on MPI_Recv
  - MPI_Recv (76%), MPI_Waitany (19%), MPI_Bcast (3%)
ECLIPSE Profiling – Message Sizes

- **Majority of messages are midrange messages**
  - Messages between 16KB to 64KB are mostly used
  - Some concentration in small messages around 0 to 256B
- **Number of messages increases as node count increases**
ECLIPSE Profiling – MPI Data Transfer

- As the cluster grows, more data transfers between MPI processes
  - Increase from 6GB average per process at 1-node to 7GB average per rank at 16-node
ECLIPSE Profiling – Aggregated Transfer

- **Aggregated data transfer refers to:**
  - Total amount of data being transferred in the network between all MPI ranks collectively

- **Very large increase of data transfer takes place in ECLIPSE at scale**
  - High network throughput is required for delivering the network bandwidth
  - Increase from 90GB of data at 1 node to 1.8TB of data at 16 nodes

---

ECLIPSE Profiling (FOURMILL)

**Aggregated Data Transferred**

<table>
<thead>
<tr>
<th>Number of Nodes</th>
<th>Data Transferred (TB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Node</td>
<td>0</td>
</tr>
<tr>
<td>2 Nodes</td>
<td>0.5</td>
</tr>
<tr>
<td>4 Nodes</td>
<td>1.0</td>
</tr>
<tr>
<td>8 Nodes</td>
<td>1.5</td>
</tr>
<tr>
<td>16 Nodes</td>
<td>2.0</td>
</tr>
</tbody>
</table>

*InfiniBand FDR*
ECLIPSE – Summary

• Performance
  – Intel Xeon E5-2600 series and InfiniBand FDR enable ECLIPSE to scale with 16 nodes
  – The E5-2680 cluster outperforms X5670 cluster by 83% at 16 nodes

• Network Interconnect
  – InfiniBand FDR provides highest performance for ECLIPSE users
  – InfiniBand FDR provides up to 32% of performance gain over InfiniBand QDR
  – Platform MPI scales better than Intel MPI at large node counts (>8 nodes)
  – RoCE provides best network scalability performance than Ethernet
    • 40GbE-RoCE provides up to 170% better performance than 10GbE at 16-node
    • 40GbE-RoCE provides up to 65% better performance than 40GbE at 16-node
  – 1GbE does not scale beyond 2 nodes

• Profiling
  – Network throughput is essential for delivering the 1.8TB of aggregated data for 256 ranks
  – Large percentage of MPI messages are in the midrange between 16KB to 64KB
Thank You

HPC Advisory Council