Note

• 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
  – Fluent performance overview
  – Understanding Fluent communication patterns
  – Ways to increase Fluent productivity
  – MPI libraries comparisons

• For more info please refer to
  – http://www.ansys.com
  – http://www.dell.com
  – http://www.intel.com
  – http://www.mellanox.com
Computational Fluid Dynamics (CFD) is a computational technology
- Enables the study of the dynamics of things that flow
- Enable better understanding of qualitative and quantitative physical phenomena in the flow which is used to improve engineering design

CFD brings together a number of different disciplines
- Fluid dynamics, mathematical theory of partial differential systems, computational geometry, numerical analysis, Computer science

ANSYS FLUENT is a leading CFD application from ANSYS
- Widely used in almost every industry sector and manufactured product
Objectives

- **The presented research was done to provide best practices**
  - Fluent performance benchmarking
    - MPI Library performance comparison
    - Interconnect performance comparison
    - CPUs comparison
    - Compilers comparison
  - **The presented results will demonstrate**
    - The scalability of the compute environment/application
    - Considerations for higher productivity and efficiency
Test Cluster Configuration

- **Dell™ PowerEdge™ R720xd 32-node (640-core) “Jupiter” cluster**
  - Dual-Socket Hexa-Core Intel E5-2680 V2 @ 2.80 GHz CPUs (Turbo mode enabled unless otherwise stated)
  - Memory: 64GB memory, DDR3 1600 MHz
  - OS: RHEL 6.2, OFED 2.3-1.0.1 InfiniBand SW stack
  - Hard Drives: 24x 250GB 7.2 RPM SATA 2.5” on RAID 0

- **Intel Cluster Ready certified cluster**

- **Mellanox Connect-IB FDR InfiniBand adapters**

- **Mellanox ConnectX-3 QDR InfiniBand and Ethernet VPI adapters**

- **Mellanox SwitchX SX6036 VPI InfiniBand and Ethernet switches**

- **MPI**: Mellanox HPC-X v1.2.0-250, (Provided): Intel MPI 4.1.030, IBM Platform MPI 9.1

- **Application**: ANSYS Fluent 15.0.7

- **Benchmarks**:
  - eddy_417k, turbo_500k, aircraft_2m, sedan_4m, truck_poly_14m, truck_14m
  - Descriptions for the test cases can be found at the [ANSYS Fluent 15.0 Benchmark](#) page
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
Fluent Performance – Interconnects

- **FDR InfiniBand enables the highest cluster productivity**
  - Surpassed other network interconnect in scalability performance
- **FDR InfiniBand tops performance among different network interconnects**
  - FDR InfiniBand outperforms QDR InfiniBand by up to 200% at 32 nodes
  - Similarly, FDR outperforms 10GbE by 16 times, and 1GbE by over 39 times

**ANSYS Fluent 15.0.7 Performance**
(eddy_417k)

![Graph showing performance comparison among different network interconnects]

- **Higher is better**
Fluent Performance – Interconnects

- FDR InfiniBand performance outperforms on other Fluent benchmarks

**ANSYS Fluent 15.0.7 Performance (turbo_500k)**

<table>
<thead>
<tr>
<th>Number of Nodes</th>
<th>1GbE</th>
<th>10GbE</th>
<th>QDR</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td>0</td>
<td>10000</td>
</tr>
<tr>
<td>2</td>
<td>5000</td>
<td></td>
<td>1000</td>
<td>15000</td>
</tr>
<tr>
<td>4</td>
<td>10000</td>
<td></td>
<td>2000</td>
<td>20000</td>
</tr>
<tr>
<td>8</td>
<td>15000</td>
<td></td>
<td>3000</td>
<td>25000</td>
</tr>
<tr>
<td>16</td>
<td>20000</td>
<td></td>
<td>4000</td>
<td>30000</td>
</tr>
<tr>
<td>32</td>
<td>25000</td>
<td></td>
<td>5000</td>
<td>35000</td>
</tr>
</tbody>
</table>

**ANSYS Fluent 15.0.7 Performance (aircraft_2m)**

<table>
<thead>
<tr>
<th>Number of Nodes</th>
<th>1GbE</th>
<th>10GbE</th>
<th>QDR</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td>0</td>
<td>1000</td>
</tr>
<tr>
<td>2</td>
<td>500</td>
<td></td>
<td>100</td>
<td>2000</td>
</tr>
<tr>
<td>4</td>
<td>1000</td>
<td></td>
<td>200</td>
<td>3000</td>
</tr>
<tr>
<td>8</td>
<td>1500</td>
<td></td>
<td>300</td>
<td>4000</td>
</tr>
<tr>
<td>16</td>
<td>2000</td>
<td></td>
<td>400</td>
<td>5000</td>
</tr>
<tr>
<td>32</td>
<td>2500</td>
<td></td>
<td>500</td>
<td>6000</td>
</tr>
</tbody>
</table>

**ANSYS Fluent 15.0.7 Performance (sedan_4m)**

<table>
<thead>
<tr>
<th>Number of Nodes</th>
<th>1GbE</th>
<th>10GbE</th>
<th>QDR</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>50</td>
<td></td>
<td>10</td>
<td>200</td>
</tr>
<tr>
<td>4</td>
<td>100</td>
<td></td>
<td>20</td>
<td>300</td>
</tr>
<tr>
<td>8</td>
<td>150</td>
<td></td>
<td>30</td>
<td>400</td>
</tr>
<tr>
<td>16</td>
<td>200</td>
<td></td>
<td>40</td>
<td>500</td>
</tr>
<tr>
<td>32</td>
<td>250</td>
<td></td>
<td>50</td>
<td>600</td>
</tr>
</tbody>
</table>

**ANSYS Fluent 15.0.7 Performance (truck_poly_14m)**

<table>
<thead>
<tr>
<th>Number of Nodes</th>
<th>1GbE</th>
<th>10GbE</th>
<th>QDR</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td></td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
<td></td>
<td>20</td>
<td>30</td>
</tr>
<tr>
<td>8</td>
<td>15</td>
<td></td>
<td>30</td>
<td>40</td>
</tr>
<tr>
<td>16</td>
<td>20</td>
<td></td>
<td>40</td>
<td>50</td>
</tr>
<tr>
<td>32</td>
<td>25</td>
<td></td>
<td>50</td>
<td>60</td>
</tr>
</tbody>
</table>
Fluent Performance – MPI Implementations

- HPC-X delivers higher scalability performance than other MPIs compared
  - HPC-X outperforms over the default Platform MPI by 10%, and Intel MPI by 19%
- Support of HPC-X on Fluent is based on the support of Open MPI on Fluent
- The new “yalla” pml reduces the overhead. Flags used for HPC-X:
  - -mca coll_fca_enable 1 -mca coll_fca_np 0 -mca pml yalla -map-by node -mca mtl mxm -mca mtl_mxm_np 0 -x MXM_TLS=self,shm,ud --bind-to core

ANSYS Fluent 15.0.7 Performance
(eddy_417k)

Higher is better
Fluent Performance – MPI Implementations

- HPC-X outperforms other MPIs on other benchmark data

**ANSYS Fluent 15.0.7 Performance**
(turbo_500k)

- Solver Rating vs. Number of Nodes for Intel MPI, Platform MPI, and HPC-X.

**ANSYS Fluent 15.0.7 Performance**
(aircraft_2m)

- Solver Rating vs. Number of Nodes for Intel MPI, Platform MPI, and HPC-X.

**ANSYS Fluent 15.0.7 Performance**
(sedan_4m)

- Solver Rating vs. Number of Nodes for Intel MPI, Platform MPI, and HPC-X.

**ANSYS Fluent 15.0.7 Performance**
(truck_poly_14m)

- Solver Rating vs. Number of Nodes for Intel MPI, Platform MPI, and HPC-X.
Fluent Performance – Turbo Mode and Clock

- Advantages are seen with running higher clock rate with Fluent
  - Either by enabling Turbo mode or higher CPU clock frequency
- Boosting CPU clock rate yields higher performance at lower cost
  - Increasing to 2800MHz (from 2200MHz) run 42% faster, 18% of increased power
- Running turbo mode also yields higher performance but at higher cost
  - Increase of 13% of performance at a expense of a 25% of increased power usage

**ANSYS Fluent 15.0.7 Performance (eddy_417k)**

- Power (Watts)
  - Number of Nodes: 1
  - 2000MHz, No Turbo: 25%
  - 2800MHz, Turbo: New data

**ANSYS Fluent 15.0.7 Performance (eddy_417k)**

- Solver Rating
  - Number of Nodes: 1, 2, 4, 8, 16, 32
  - 2000MHz, No Turbo
  - 2800MHz, No Turbo
  - 2800MHz, Turbo

*Higher is better*
• Results demonstrated by HPCAC outperforms the previous best record
  – The ANSYS Fluent 15.0 Benchmark publishes ANSYS Fluent performance results
  – HPCAC achieved 26.36% higher performance than the best published results (as of 9/22/2014), despite slower CPUs are used on the Jupiter cluster by the HPCAC
  – The 32-node/640-core result beats previous record of 96-node/1920-core by 8.53%
  – Performance is expected to climb on the Jupiter cluster if more nodes are available

ANSYS Fluent 15.0.7 Performance
(eddy_417k)
**Fluent Profiling – I/O Profiling**

- **Minor disk I/O activities take place on all MPI ranks for this workload**
  - Majority of the read activities are disk appeared at the beginning of the job run

**ANSYS Fluent 15.0.7**
- **(Rank 0 Node)**
- **(non-Rank 0 Node)**

*InfiniBand FDR*
Fluent Profiling – Point-to-point dataflow

- Communication seems to be limited to MPI ranks that is closer to self
  - Heavy communications seen between first and last ranks
- Communication pattern does not change as the cluster scales
  - However, the amount of data being transferred is reduced as the node scales

InfiniBand FDR

2 nodes

32 nodes
Fluent Profiling – Time Spent by MPI Calls

- Majority of the MPI time is spent on MPI_Waitall
  - Accounts for 30% Wall time
  - MPI_Allreduce – 20%
  - MPI_Recv – 11%

- Some load imbalances in network are observed
  - Some ranks spent more time MPI_Waitall and MPI_Allreduce
  - Might be related to how workload is distributed among the MPI ranks
Fluent Profiling – MPI Message Sizes

- Majority of data transfer messages are small to medium sizes
  - MPI_Allreduce: Large concentration of 4-byte msg (~18% wall time)
  - MPI_Wait: Large concentration of 16-byte msg (~11% wall time)

```
eddy_417k, 32 nodes
```
**Fluent – Summary**

- **Performance**
  - Jupiter cluster outperforms other system architectures on Fluent
    - FDR InfiniBand delivers higher performance against QDR InfiniBand by 200%
    - FDR IB outperforms 10GbE by up to 11 times at 32 nodes / 640 cores
  - FDR InfiniBand enable Fluent to break previous performance record
    - Outperforms previously set record by 26.35% at 640 cores/ 32 nodes
    - Outperforms previously set record by 8.52% at 1920 cores/ 96 nodes
  - HPC-X MPI delivers higher performance against other MPI Implementation
    - HPC-X outperforms Platform MPI by 10%, outperforms Intel MPI by 19%

- **CPU**
  - Higher CPU clock rate and Turbo mode yields higher performance for Fluent
    - Bumping CPU clock (from 2200MHz to 2800MHz) yields 42% faster perf at 18% of increased power
    - Enabling turbo mode translates to 13% of increase performance at a 25% of additional power usage

- **Profiling**
  - Heavy usage in small msg in MPI_Waitall, MPI_Allreduce, MPI_Recv communications
Thank You

HPC Advisory Council