COSMO
Performance Benchmark and Profiling
January 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
  - COSMO performance overview
  - Understanding COSMO communication patterns
  - Ways to increase COSMO productivity
  - Network Interconnect comparisons

• For more info please refer to
  - http://www.dell.com
  - http://www.intel.com
  - http://www.mellanox.com
  - http://www.COSMO-model.org
COSMO

- Stands for **Consortium for Small-scale Modeling** (COSMO), which was formed in 1998
- General goal is to develop, improve and maintain a non-hydrostatic limited-area atmospheric model, to be used both for operational and for research applications by the members of the consortium
- Is a nonhydrostatic limited-area atmospheric prediction model
- has been designed for both operational numerical weather prediction (NWP) and various scientific applications on the meso-β and meso-γ scale
- Is based on the primitive thermo-hydrodynamical equations describing compressible flow in a moist atmosphere. The model equations are formulated in rotated geographical coordinates and a generalized terrain following height coordinate
Test Cluster Configuration

- **Dell™ PowerEdge™ M610 38-node (456-core) cluster**
  - Six-Core Intel X5670 @ 2.93 GHz CPUs
  - Memory: 24GB memory, DDR3 1333 MHz
  - OS: RHEL 5.5, OFED 1.5.2 InfiniBand SW stack

- **Intel Cluster Ready certified cluster**

- **Mellanox ConnectX-2 InfiniBand adapters and non-blocking switches**

- **MPI: Platform MPI 8.2**

- **Compilers and Libraries: GNU 4.6 and NetCDF 4.1.3**

- **InfiniBand-based Lustre Storage: Lustre 1.8.5**

- **Application: COSMO-RAPS 5.0**

- **Benchmark datasets: 2 global data sets in GRIB code from Oct 23, 2006, 12 UTC**
  - COSMO-EU: 24-hour forecast for a 7 km covering Europe. Domain size 665x657x40 grid points
  - COSMO-DE: 24-hour forecast for a 2.8 km covering Germany. Domain size 421x461x50 grid points
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
About Dell PowerEdge Servers

- **System Structure and Sizing Guidelines**
  - 38-node cluster build with Dell PowerEdge™ M610 blade servers
  - Servers optimized for High Performance Computing environments
  - Building Block Foundations for best price/performance and performance/watt

- **Dell HPC Solutions**
  - Scalable Architectures for High Performance and Productivity
  - Dell's comprehensive HPC services help manage the lifecycle requirements.
  - Integrated, Tested and Validated Architectures

- **Workload Modeling**
  - Optimized System Size, Configuration and Workloads
  - Test-bed Benchmarks
  - ISV Applications Characterization
  - Best Practices & Usage Analysis
**Stages in running the COSMO**

- **gme2eu:**
  - Precursory steps to interpolate GME data to the domain used for the COSMO-EU grid

- **cosmo_eu:**
  - The job runs a 24-hour forecast for a 7 km LM covering nearly the whole of Europe
  - Domain size of 665x657x40 grid points

- **eu2de:**
  - Precursory step to output of the COSMO-EU run with 7 km resolution

- **cosmo_de:**
  - The job runs a 24-hour forecast for a 2.8 km domain covering Germany
  - Domain size 421x461x50 grid points
  - This is a bigger model where a smaller time step is used compared to the 7 km runs
COSMO Performance – Network Interconnects

- **COSMO demonstrates superior scalability using InfiniBand QDR**
  - Performs closer to linear-scale as more nodes join the cluster
  - While 1GbE performance is limited after 4-node due to network traffic congestions

- **InfiniBand QDR enables higher cluster productivity**
  - Up to 127% of increased productivity over 1GbE network on cosmo_de model at 16-node

- **Both tests run on the InfiniBand-based Lustre storage**
- **Test stops at 16-node for 1GbE due to switch port limitation**

![Performance Graphs](image_url)

**COSMO Performance (cosmo_eu)**
- Performance (Jobs/Day)
- Number of Nodes: 2, 4, 8, 16, 32
- InfiniBand QDR vs. 1GbE

**COSMO Performance (cosmo_de)**
- Performance (Jobs/Day)
- Number of Nodes: 2, 4, 8, 16, 32
- InfiniBand QDR vs. 1GbE

**Graphs showing**
- 102% increase at 16-node for cosmo_eu
- 127% increase at 16-node for cosmo_de
**COSMO Profiling – Interconnects MPI/User Time**

- **Network interconnects contribution to the overall performance**
  - MPI time accounts for 28% of overall run time at 16-node for cosmo_de with IB QDR
  - MPI time accounts for 70% of overall run time at 16-node for cosmo_de with 1GbE

- **Network interconnect shows a direct impact on COSMO performance**
  - Since user time remains the same for different network interconnects
  - The use of blocking MPI calls shows performance is dependent on the network hardware

- **Both MPI and computational time reduced as cluster scales with InfiniBand QDR**

---

**COSMO Profiling**
(cosmo_de)

**MPI/User Time Ratio**

<table>
<thead>
<tr>
<th>Number of Nodes</th>
<th>Time (seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>8_QDR</td>
<td>1000</td>
</tr>
<tr>
<td>8_1GbE</td>
<td>2000</td>
</tr>
<tr>
<td>16_QDR</td>
<td>2000</td>
</tr>
<tr>
<td>16_1GbE</td>
<td>3000</td>
</tr>
</tbody>
</table>

---

**COSMO Profiling**
(cosmo_de)

**MPI/User Time Ratio**

<table>
<thead>
<tr>
<th>Number of Nodes</th>
<th>Time (seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>8_QDR</td>
<td>1000</td>
</tr>
<tr>
<td>8_1GbE</td>
<td>2000</td>
</tr>
<tr>
<td>16_1GbE</td>
<td>3000</td>
</tr>
</tbody>
</table>

---

**Network of Expertise**
**COSMO Profiling – Disk IO Usage**

- **Pre- and post-cursor steps of COSMO shows large disk writes**
  - **gme2eu**: Precursory step for interpolating GME data to the domain used by cosmo_eu
  - **eu2de**: The step for outputting of the cosmo_eu run with 7 km resolution
  - About 150MB write and 350MB read are the limits by the current storage device
The actual simulation of models show periodic reads and writes

- Bursts of reads and writes occurred periodically throughout the simulations
- Shows significantly more writes than reads occurred
- Test is configured to read and writes from the first rank
- Accounts for reads and writes of the GRIB files during each of the 24 hours in the weather forecast simulation

**COSMO Profiling**
(cosmo_eu)

**COSMO Profiling**
(cosmo_de)

**InfiniBand QDR**
COSMO Profiling – MPI/User Time Ratio

- Both MPI and computational time reduced in scale as cluster scales
  - Shows close to linear scaling as the number of compute nodes doubled
  - Infers that COSMO is designed run as scale with the use of InfiniBand QDR
• **MPI_Sendrecv is the most used MPI calls**
  – The cosmo_de model has around 5 times as much send and receive than cosmo_eu
  – Accounted for 59% of the MPI function calls on a 32-node job for cosmo_eu
  – Accounted for 88% of the MPI function calls on a 32-node job for cosmo_de

• **COSMO has a large percent of MPI calls for blocking data transfers**
  – The blocking MPI APIs requires the data transfer to complete before progressing
COSMO Profiling – Time Spent by MPI Calls

- **Majority of the MPI time is spent on MPI_Sendrecv**
  - cosmo_eu: MPI_Sendrecv(39%), MPI_Allreduce(19%), MPI_Wait(13%) on 32-node
  - cosmo_de: MPI_Sendrecv(51%), MPI_Allreduce(19%), MPI_Wait(10%) on 32-node
- **MPI communication time is reduced as the cluster scales**
  - As more nodes take on the computational work, the job completes faster
  - Which reduces the communication time for each MPI calls
**COSMO Profiling – MPI Message Sizes**

- **Different message sizes pattern seen for both models**
  - `cosmo_eu`: Majority of message sizes are in the 0-64 byte range
  - `cosmo_de`: A large percentage of message sizes in 257-1K byte range
- **Significantly larger volume of small messages in `cosmo_eu`**
  - Small messages typically used for synchronization, tends to be more latency sensitive

---

**COSMO Profiling**

(cosmo_eu)

**MPI Message Sizes**

- Y-axis: Number of Messages (Millions)
- X-axis: Message Sizes
- Legend: 2 Nodes, 4 Nodes, 8 Nodes, 16 Nodes, 32 Nodes

**COSMO Profiling**

(cosmo_de)

**MPI Message Sizes**

- Y-axis: Number of Messages (Millions)
- X-axis: Message Sizes
- Legend: 2 Nodes, 4 Nodes, 8 Nodes, 16 Nodes, 32 Nodes

*InfiniBand QDR*
COSMO Profiling – Aggregated Transfer

- **Aggregated data transfer refers to:**
  - Total amount of data being transferred in the network between all MPI ranks collectively
- **Significant amount of data transfer takes place in COSMO**
  - Around 8TB of data being exchanged between the nodes at 32-node for cosmo_eu model
- **More data is transferred as the node count increases**
  - Jump in network traffic from 4.5TB to 7.5TB between 16- and 32-node

**COSMO-Model Profiling**
(cosmo_eu)
Aggregated Data Transferred

**COSMO-Model Profiling**
(cosmo_de)
Aggregated Data Transferred

*InfiniBand QDR*
• Substantial less data transfers between MPI processes as cluster scales
  – Reducing data communications from 20-30GB an single node simulation
• Each process sends and receives generally the same amount of data
  – But the cosmo_eu model communicates half as much than the cosmo_de model
COSMO – Summary

• **COSMO delivers superior linear scalability and performance**
  – COSMO can take advantage of additional compute power by using InfiniBand QDR

• **Superior network is needed for COSMO to run at the most efficient rate**
  – InfiniBand QDR delivers up to 127% of increased productivity over 1GbE at 16-node
  – 1GbE performance hinders the scalability of COSMO starting at 4-node

• **Disk IO is also a crucial component in COSMO**
  – Shows periodic reads and writes in various stages of the COSMO simulations

• **Profiling**
  – MPI_Sendrecv is the most used and the most time-consuming MPI function
  – Significant data transfers take place in both cosmo_eu and cosmo_de models
Thank You

HPC Advisory Council