HPC ADVISORY COUNCIL

COSMO-Model
Performance Benchmark and Profiling

August 2013

CONSORTIUM FOR SMALL SCALE MODELING
COSMO

DELL

intel

Mellanox
TECHNOLOGIES

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**
  - 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

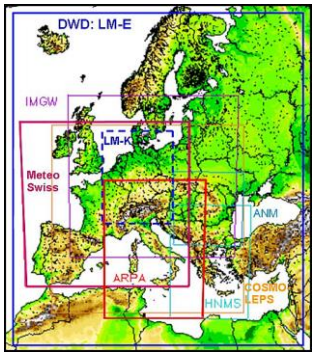
Objectives

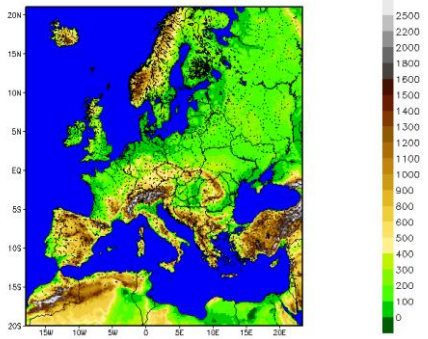
- **The following was done to provide best practices**
  - COSMO-Model performance benchmarking
  - Interconnect performance comparisons
  - Processor generation performance comparison
  - Understanding COSMO-Model communication patterns

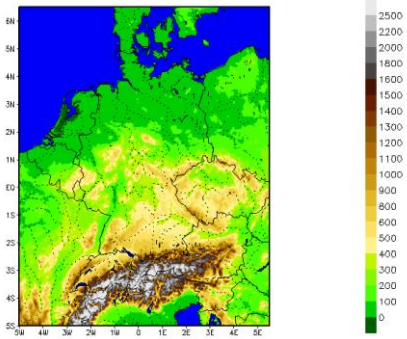
- **The presented results will demonstrate**
  - The scalability of the compute environment to provide nearly linear application scalability
  - The capability of COSMO-Model to achieve scalable productivity

COSMO-Model

- **COSMO**
  - Stands for **Co**nsortium for **S**mall-scale **Mo**deling (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
- **COSMO-Model**
  - Is the limited area forecast model of the COSMO 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

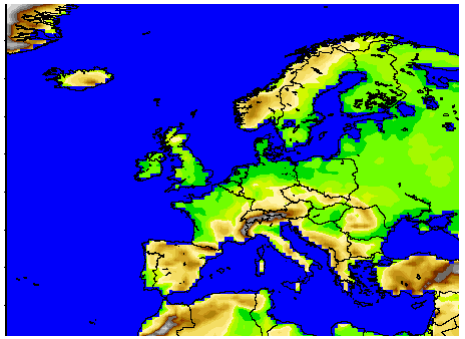




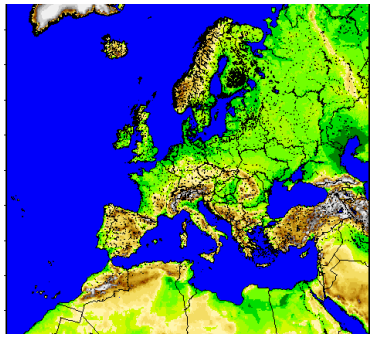


Test Cluster Configuration

- **Dell™ PowerEdge™ R720xd 32-node "Jupiter" cluster**
  - 16-node Dual-Socket Ten-Core Intel E5-2680 V2 @ 2.80 GHz CPUs
  - 16-node Dual-Socket Eight-Core Intel E5-2680 @ 2.70 GHz CPUs
  - Memory: 64GB memory, DDR3 1600 MHz
  - OS: RHEL 6.2, OFED 2.0 InfiniBand SW stack
  - Hard Drives: 24x 250GB 7.2 RPM SATA 2.5" on RAID 0



- **Mellanox ConnectX-3 FDR InfiniBand VPI adapters**
- **Mellanox SwitchX SX6036 InfiniBand VPI switch**
- **Intel Cluster Ready certified cluster**
- **Compilers and Libraries: Intel Composer XE 2013.0.079**



- **Application and Benchmark Datasets: COSMO-RAPS 5.1**
  - 2 global data sets in GRIB: GME (Jan 11, 2012 12 UTC), IFS (Oct 23, 2006, 12 UTC)
  - COSMO-EU: 24-hour forecast for a 7 km LM covering Europe. Domain size 665x657x40 grid points
  - COSMO-DE: 12-hour forecast for a 2.8 km covering Germany. Domain size 724x780x65 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

PowerEdge R720xd

Massive flexibility for data intensive operations

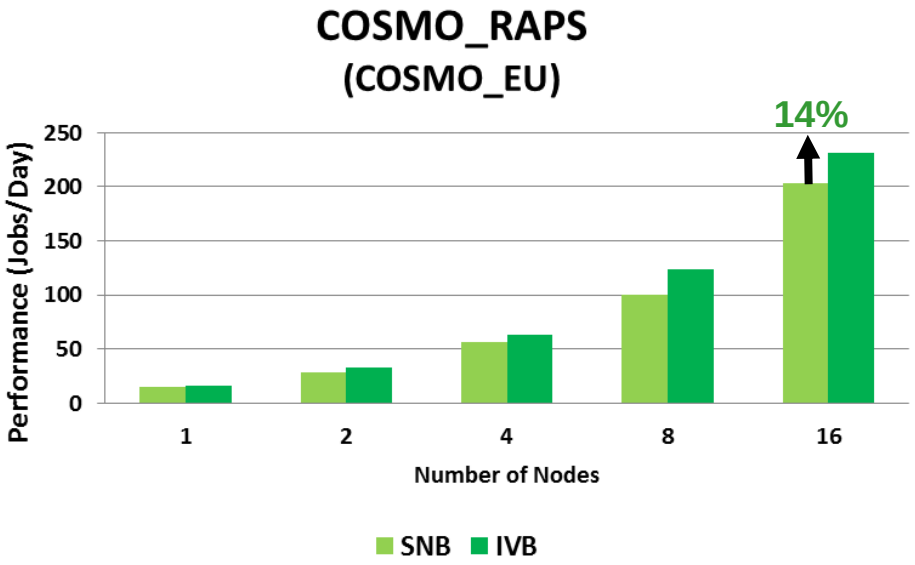
- **Performance and efficiency**
  - Intelligent hardware-driven systems management with extensive power management features
  - Innovative tools including automation for parts replacement and lifecycle manageability
  - Broad choice of networking technologies from GbE to IB
  - Built in redundancy with 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

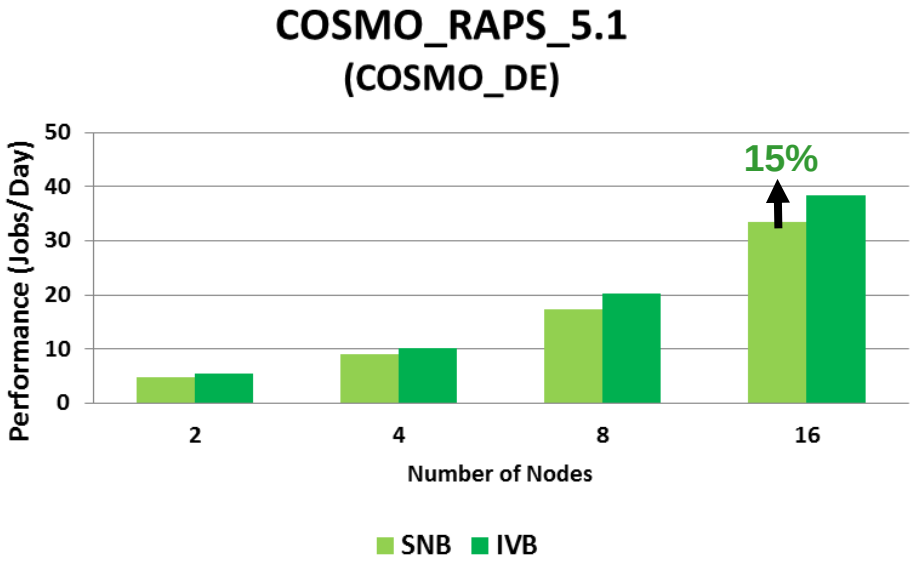
COSMO Performance – Stages

- **Stages in running COSMO Model**
  - **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 Central Europe
    - Domain size 724x780x65 grid points
    - This is a bigger model where a smaller time step is used compared to the 7 km runs

COSMO Performance – Processors

- **E5-2680 V2 (Ivy Bridge) cluster outperforms prior generation**
  - Performs up to 15% better than E5-2680 cluster (Sandy Bridge) at 16 nodes
- **System components used:**
  - IVB: 2-socket 10-core E5-2680 V2 @2.8GHz,1600MHz DIMMs, FDR IB, 24 HDDs
  - SNB: 2-socket 8-core E5-2680 @ 2.7GHz,1600MHz DIMMs, FDR IB, 24 HDDs



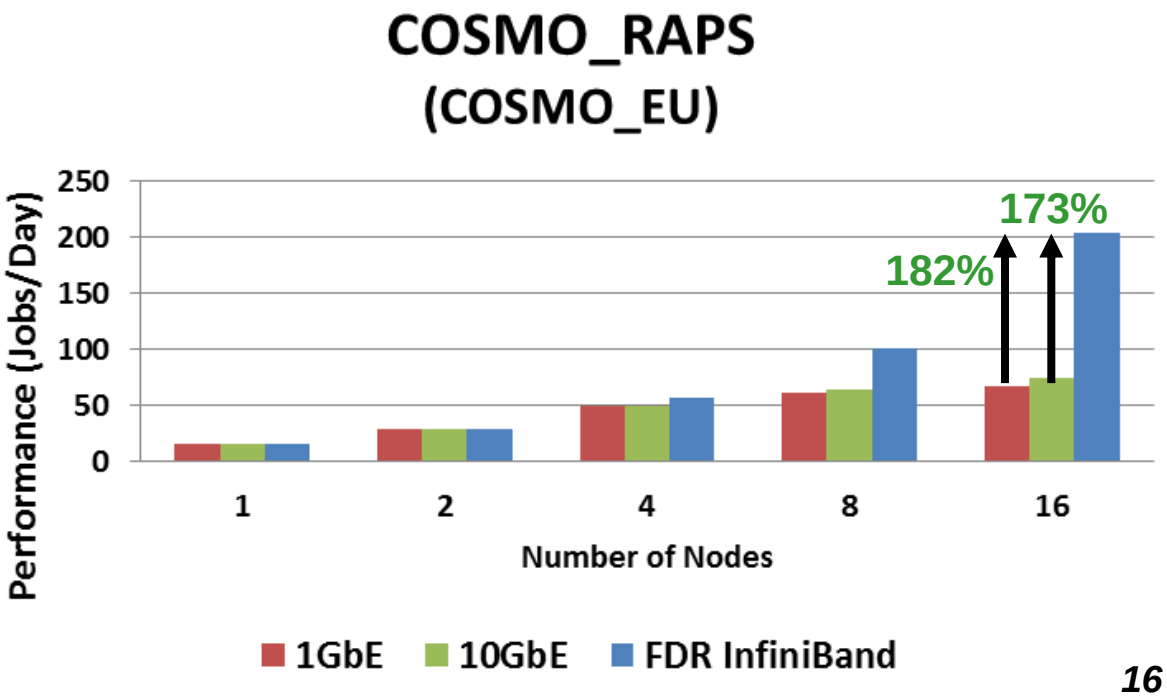


Higher is better

FDR InfiniBand

COSMO Performance – Network Interconnects

- **COSMO demonstrates superior scalability using FDR InfiniBand**
  - Performs closer to linear-scale as more nodes join the cluster
  - While Ethernet performance is limited after 4-node due to network traffic congestions
- **FDR InfiniBand enables higher cluster productivity**
  - Up to 182% of increased productivity over 1GbE network at 16-node
  - Up to 173% of increased productivity over 10GbE network at 16-node

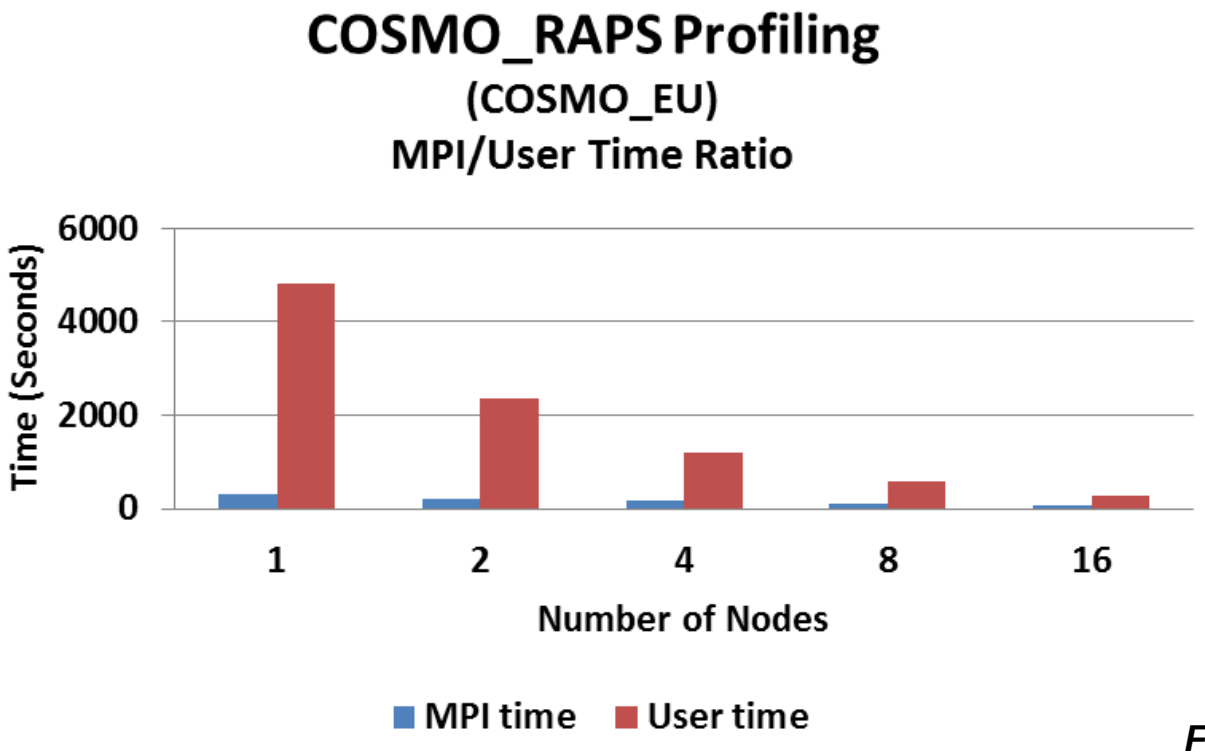


Higher is better

16 Processes/Node

COSMO Profiling – MPI/User Time Ratio

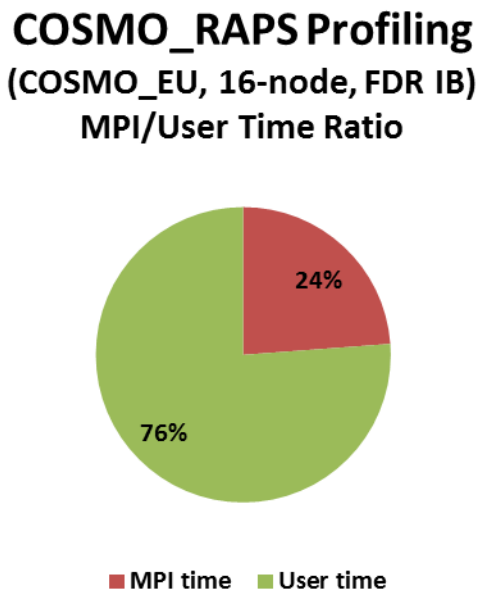
- **Both MPI and computational time reduced as cluster scales with FDR InfiniBand**
  - Shows close to linear scaling as the number of compute nodes doubled
- **FDR InfiniBand enables COSMO-Model to scale**
  - Data exchange between processes are offloaded to the InfiniBand hardware using RDMA
  - Thus CPU can concentrate on computation, thus providing best scalability performance

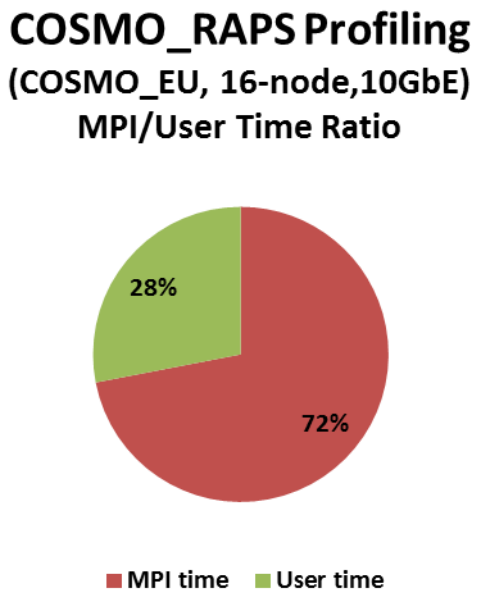


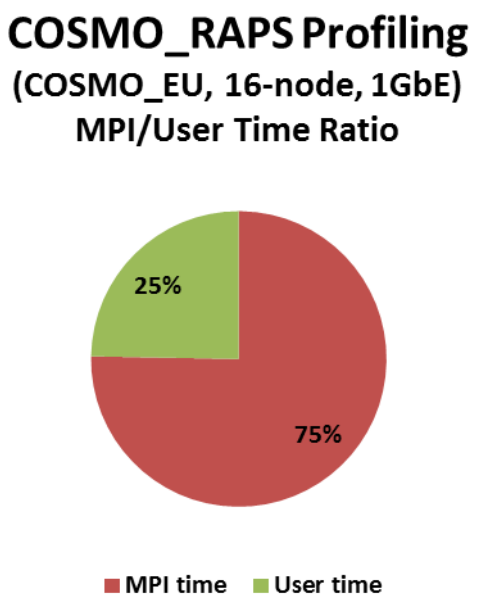
FDR InfiniBand

COSMO Profiling – MPI/User Time Ratio

- **Network interconnects contribution to the overall performance**
  - Time on FDR InfiniBand accounts for 24% of overall run time
  - Time on 10GbE accounts for 72% of overall run time
  - Time on 1GbE accounts for 75% of overall run time
- **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

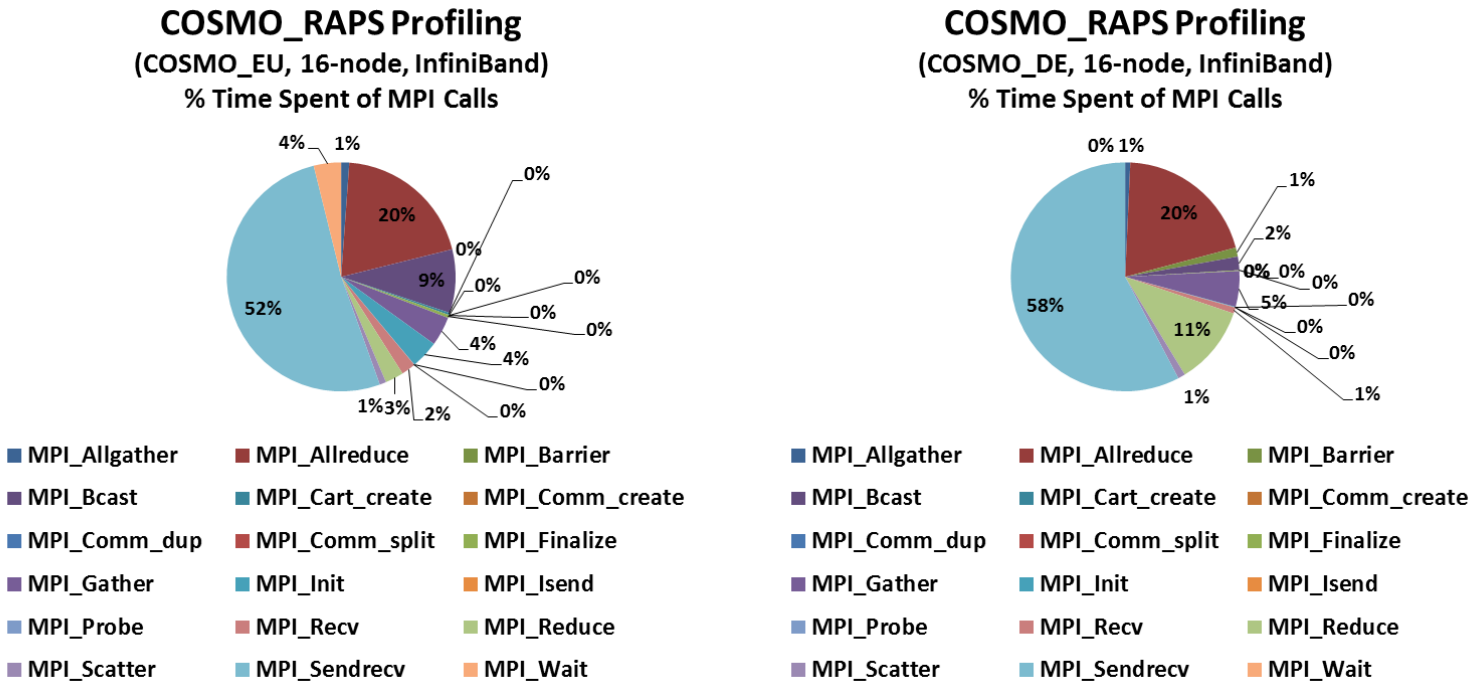






COSMO Profiling – MPI Time Distribution

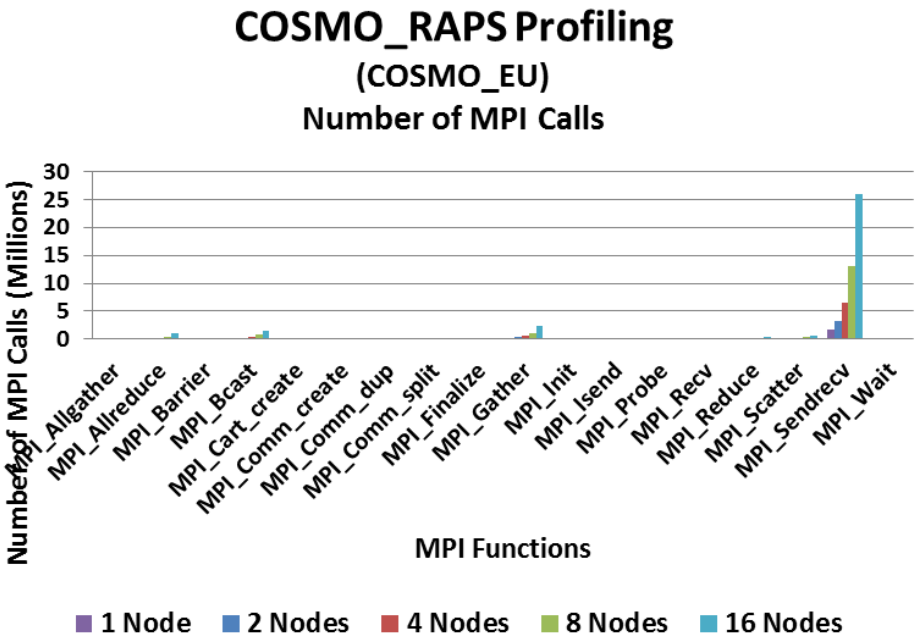
- **Majority of the MPI time is spent on MPI_Sendrecv**
  - cosmo_eu: MPI_Sendrecv(52%), MPI_Allreduce(20%), MPI_Bcast(9%)
  - cosmo_de: MPI_Sendrecv(58%), MPI_Allreduce(20%), MPI_Reduce(11%)
- **MPI communication time is reduced as the cluster scales**
  - As more nodes take on the computational work, the job completes faster
  - Which reduces the communication time for each MPI calls

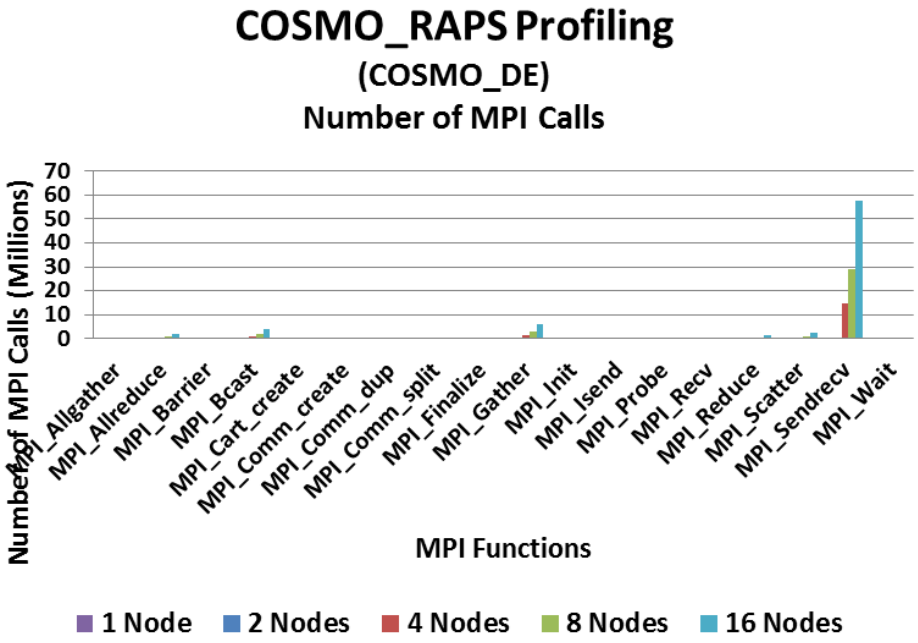


FDR InfiniBand

COSMO Profiling – MPI Calls

- **MPI_Sendrecv is the most used MPI calls**
  - The cosmo_de model has around 2.4 times as much send and receive than cosmo_eu
  - Accounted for 80-82% of the MPI function calls on a 16-node job
- **COSMO has a large percent of MPI calls for blocking data transfers**
  - The blocking MPI APIs requires the data transfer to complete before progressing

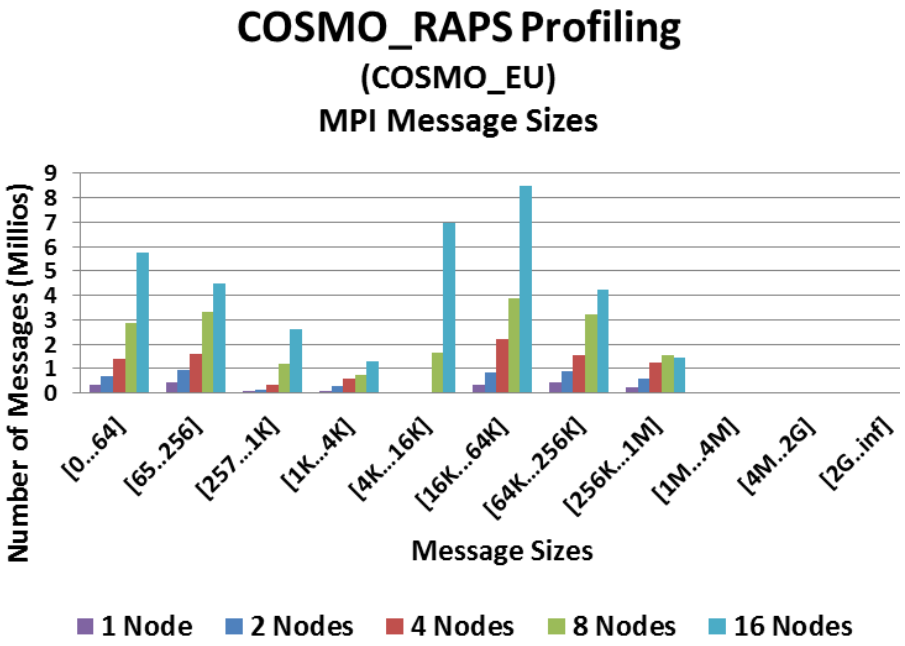


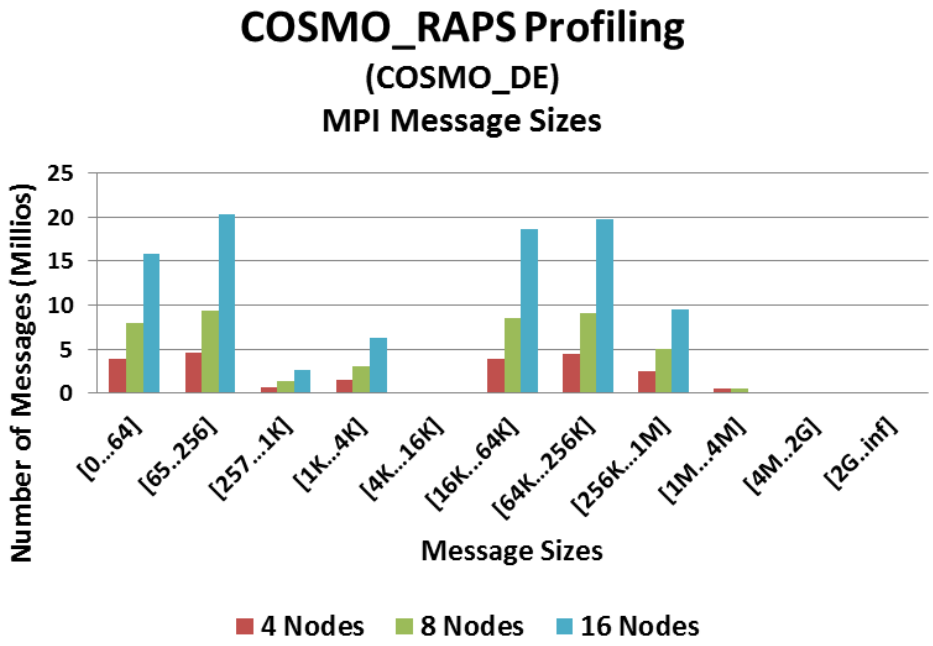


FDR InfiniBand

COSMO Profiling – MPI Message Sizes

- **A range of message sizes seen used for both models**
  - cosmo_eu: Majority of message sizes are in the 16K-64K byte range
  - cosmo_de: A large percentrage of message sizes in 65-256, 16K to 256K byte range

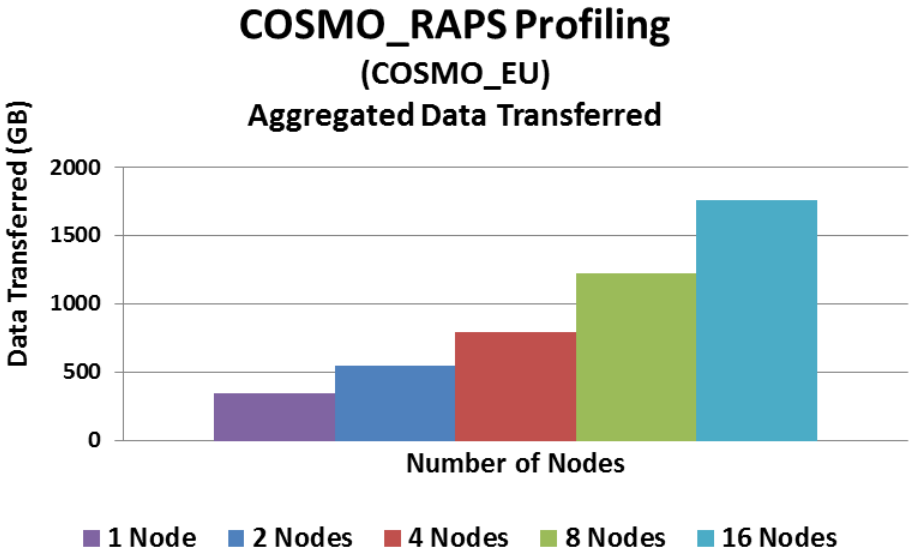


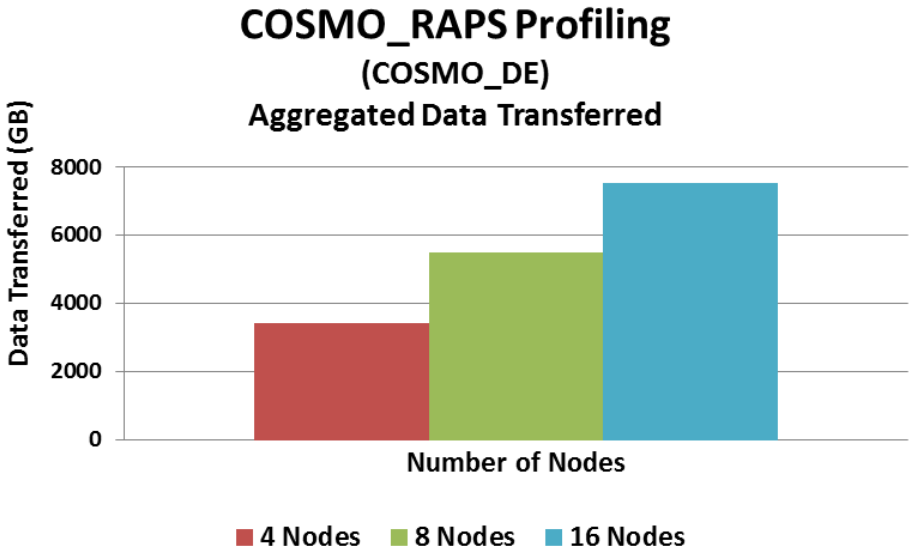


FDR InfiniBand

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 16-node for cosmo_eu model
- **More data is transferred as the node count increases**
  - Jump in network traffic from 5.7TB to 7.8TB between 8-node and 16-node on cosmo_de

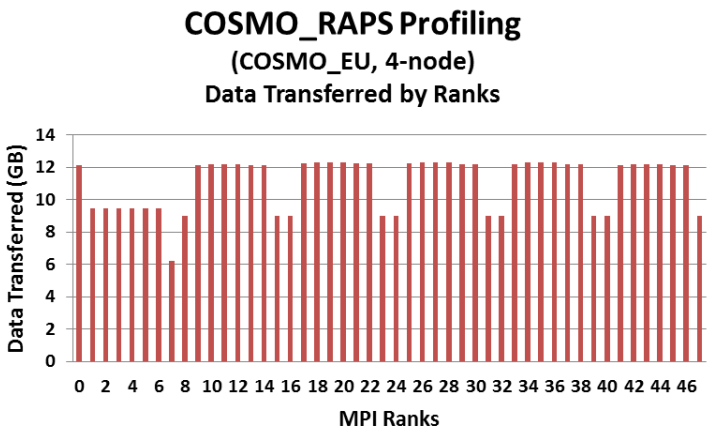


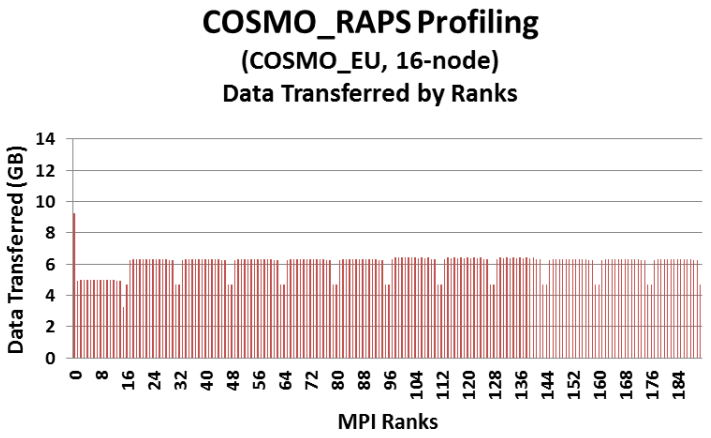


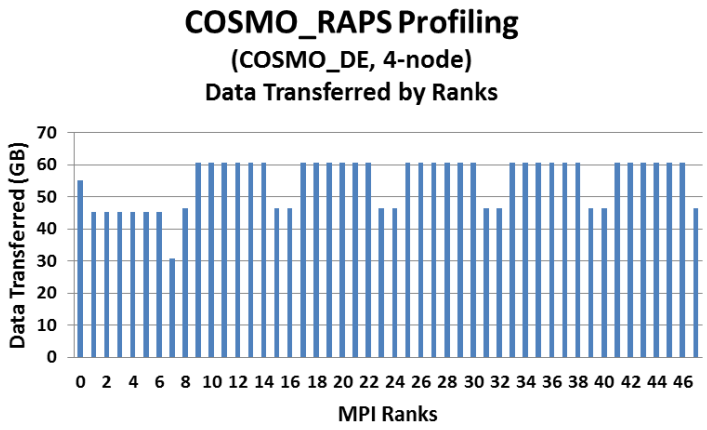
FDR InfiniBand

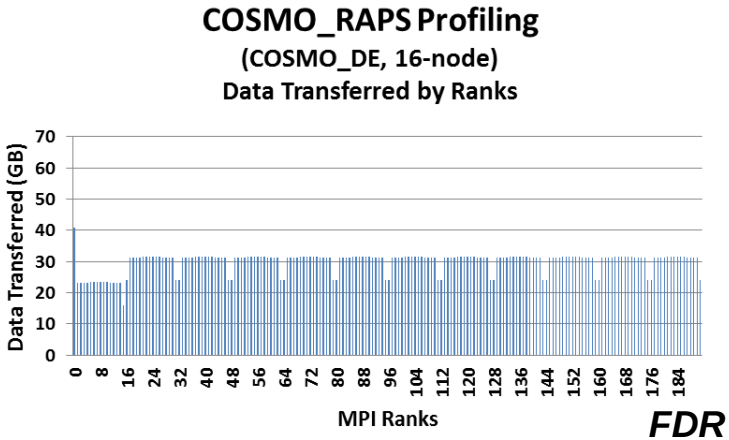
COSMO Profiling – MPI Data Transfer

- **Substantial less data transfers between MPI processes as cluster scales**
  - Data communications reduced from 12GB (4-node) to 6GB (16-node) simulation
- **Each process sends and receives generally the same amount of data**
  - cosmo_de model communicates 5 times as much as the cosmo_eu model









FDR InfiniBand

COSMO – Summary

- **COSMO delivers superior linear scalability and performance**
  - COSMO can take advantage of additional compute power by using FDR InfiniBand
- **Superior network productivity needed for COSMO to run efficiently**
  - FDR InfiniBand delivers up to 182% of increased productivity over 1GbE at 16-node
  - FDR InfiniBand delivers up to 173% of increased productivity over 10GbE at 16-node
  - Ethernet performance hinders the scalability of COSMO starting at 4-node
- **Intel Ivy Bridge-EP series and FDR InfiniBand enable COSMO-Model to scale**
  - The E5-2680 V2 (IVB) cluster outperforms E5-2680 (SNB) cluster by 15% at 16 nodes
- **MPI Profiling**
  - Both COSMO_EU and COSMO_DE show similar communication characteristics
  - 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