• 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
  – ICON performance overview
  – Understanding ICON communication patterns
  – Ways to increase ICON productivity
  – Network Interconnect comparisons

• For more info please refer to
  – http://www.dell.com
  – http://www.intel.com
  – http://www.mellanox.com
  – http://icon.enes.org
ICON

- ICON GCM: ICOsahedral Non-hydrostatic General Circulation Model
- The ICON dynamical core is a new development initiated by the Max Planck Institute for Meteorology (MPI-M) and the Deutscher Wetterdienst (DWD)
- The goal of ICON is to develop a new generation of general circulation models for the atmosphere and the ocean in a unified framework.
- The ICON dynamical core solves the fully compressible non-hydrostatic equations of motion for simulations at very high horizontal resolution.
- The discretization of the continuity and tracer transport equations will be consistent so that mass of air and its constituents are conserved, which is a requirement for atmospheric chemistry.
- Furthermore, the vector invariant form of the momentum equation will be used, and thus, vorticity dynamics will emphasized.
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: Open MPI 1.5.4, Platform MPI 8.2**

• **Compilers and Libraries: GNU 4.6 and NetCDF 4.1.3, HDF5 1.8.8**

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

• **Application: ICON revision 3272 (ICON_RAPS_1.1)**

• **Benchmark dataset:**
  - exp.test_hat_jww.run: hydrostatic atmosphere on a triangular R2B04 grid with initial condition for the Jablonowski Williamson baroclinic wave test
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
ICON Performance – MPI

- **Platform MPI provides better scalability performance over Open MPI**
  - Up to 72% of increased productivity over Open MPI on a 32-node job
  - Up to 21% of increased productivity over Open MPI on a 16-node job
- **Scalability of Open MPI is limited to around 16 nodes**
- **No extra flags were used for both cases except for enabling processor binding**

![ICON Performance chart](exp.test_hat_jww.run)
• InfiniBand QDR enables higher cluster productivity
  – Up to 149% of increased productivity over 1GbE network for a 16-node job
  – Up to 116% of increased productivity over 1GbE network on a 8-node job
• ICON demonstrates good scalability using InfiniBand
  – Performance gain for 1GbE performance is limited after 4-node due to network congestion
• Test stops at 16-node for 1GbE due to switch port limitation
• **InfiniBand reduces the overall runtime of ICON**
  – Time for communication remains the same while user (compute) time reduces as more nodes are added to the cluster

• **InfiniBand allows more system runtime for the actual computation for a job**
  – Network communication accounts for 27% of overall runtime at 8-node w/ InfiniBand QDR
  – Network communication accounts for 58% of overall run time at 8-node w/ 1GbE
**ICON Profiling – MPI Calls**

- **MPI_Allreduce**, **MPI_Irecv** and **MPI_Send** are the most used MPI calls
  - MPI_Irecv is accounted for 33% of the MPI function calls on a 32-node run
  - MPI_Send is accounted for 31% of the MPI function calls on a 32-node run
  - MPI_Allreduce is accounted for 30% of the MPI function calls on a 32-node run

![ICON Profiling](exp.test_hat_jww.run) (32-node, InfiniBand) % MPI Calls

![ICON Profiling](exp.test_hat_jww.run) (32-node, InfiniBand) % MPI Calls

- **MPI_Allreduce**
- **MPI_Barrier**
- **MPI_Bcast**
- **MPI_Comm_dup**
- **MPI_Comm_split**
- **MPI_Finalize**
- **MPI_Init**
- **MPI_Irecv**
- **MPI_Isend**
- **MPI_Recv**
- **MPI_Send**
- **MPI_Waitall**

![InfiniBand QDR](image)
**ICON Profiling – Time Spent by MPI Calls**

- **Majority of the MPI time is spent on MPI_Sendrecv**
  - MPI_Allreduce(46%), MPI_Send(24%), MPI_Waitall(19%) on 16-node
- **MPI_Allreduce takes more time to complete as the cluster grows**
  - While the time for MPI_Send is reduced

**ICON Profiling**
(exp.test_hat_jww.run)
Time Spent of MPI Calls

**ICON Profiling**
(exp.test_hat_jww.run, 16-node, InfiniBand)
% Time Spent of MPI Calls

- MPI_Allreduce: 46%
- MPI_Sendrecv: 24%
- MPI_Send: 24%
- MPI_Waitall: 19%

**MPI Functions**

1 Node | 2 Nodes | 4 Nodes | 8 Nodes | 16 Nodes | 32 Nodes

**InfiniBand QDR**
**ICON Profiling – MPI Message Sizes**

- **Small message sizes are the most dominant**
  - Majority of message sizes are in the 0-64 byte range
  - Large volume of small messages typically means the application is latency sensitive

---

**Graph**

**ICON Profiling**
(exp.test_hat_jww.run)

**MPI Message Sizes**

![Graph of MPI Message Sizes]

- **Number of Messages (Millions)**
- **Message Sizes**
  - 0-64
  - 65-256
  - 257-1K
  - 1K-4K
  - 4K-16K
  - 16K-64K
  - 64K-256K
  - 256K-1M
  - 1M-4M
  - 4M-2G
  - 2G-inf

- **Legend**
  - 1 Node
  - 2 Nodes
  - 4 Nodes
  - 8 Nodes
  - 16 Nodes
  - 32 Nodes

*InfiniBand QDR*
ICON Profiling – MPI Data Transfer

- **The amount of data transfers increases gradually as the cluster scales**
  - Each additional process would add around 1GB to the network messaging
- **Each process transfers roughly the same amount of data**
  - Except for the 1st MPI rank takes on the most network communication, at roughly 2.5GB
• **Aggregated data transfer refers to:**
  - Total amount of data being transferred in the network between all MPI ranks collectively

• **The amount of data transfer increases steadily as the node count increases**
  - As previously shown, each MPI rank increases the overall data transfer by roughly 1GB
ICON – Summary

• **ICON delivers good scalability and performance**
  – ICON can take advantage of additional compute power by using InfiniBand QDR

• **Platform MPI delivers better performance than Open MPI**
  – Shows around 72% improvement over Open MPI for a 32-node run

• **InfiniBand is needed for ICON to run at the most efficient rate at scale**
  – InfiniBand QDR delivers up to 149% of better performance over 1GbE at 16-node
  – With the RDMA capability, InfiniBand frees up the system for the actual computation

• **Profiling**
  – Majority of MPI messages falls in the small messages range (of 0-64 byte)
  – Typically small message means the application is network latency sensitive
  – MPI_Allreduce is the most time-consuming MPI function
Thank You
HPC Advisory Council