HPC ADVISORY COUNCIL


# ICON
# Performance Benchmark and Profiling

September 2013

DELL intel Mellanox TECHNOLOGIES

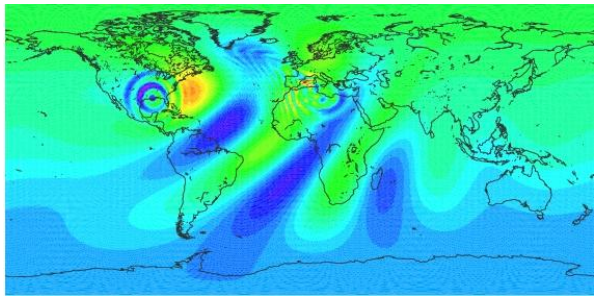
# Note

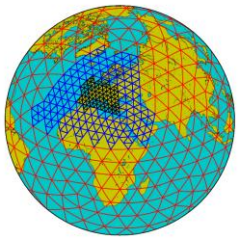
- **The following research was performed under the HPC Advisory Council activities**
  - Participating vendors: Intel, Dell, Mellanox
  - Compute resource - HPC Advisory Council Cluster Center

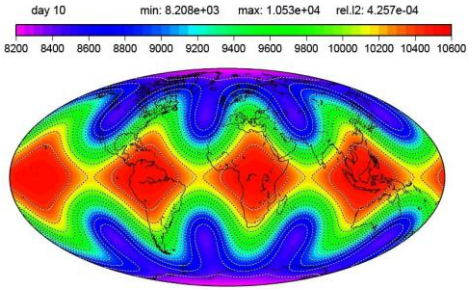
- **The following was done to provide best practices**
  - ICON performance overview
  - Understanding ICON communication patterns
  - Ways to increase ICON productivity
  - Network Interconnect comparisons

- **For more info please refer to**
  - http://www.dell.com
  - http://www.intel.com
  - http://www.mellanox.com
  - http://www.mpimet.mpg.de/en/science/models/icon.html

- **ICON**
  - ICON GCM: ICOsahedral Non-hydrostatic General Circulation Model
  - The ICON dynamical core is the 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™ 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 2.0-3.0.0 InfiniBand SW stack
  - Hard Drives: 24x 250GB 7.2 RPM SATA 2.5" on RAID 0
- **Mellanox Connect-IB FDR InfiniBand adapters and ConnectX-3 Ethernet adapters**
- **Mellanox SwitchX SX6036 InfiniBand VPI switch**
- **MPI: Platform MPI 9.1, Open MPI 1.6.5 (w/ Mellanox FCA 2.5)**
- **Compilers & Libraries: GNU Compilers 4.4.6, NetCDF 4.1.3, HDF5 1.8.7 (szip & zlib)**
- **Application: ICON (based on ICON_RAPS_1.1 rev 8144, ICON_RAPS_2.0)**
- **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

# PowerEdge R720xd

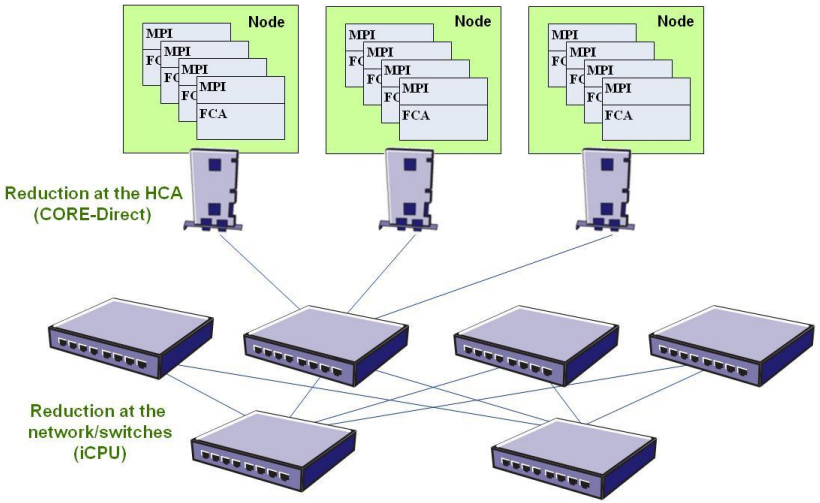
## Massive flexibility for data intensive operations

- **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

# About Mellanox FCA

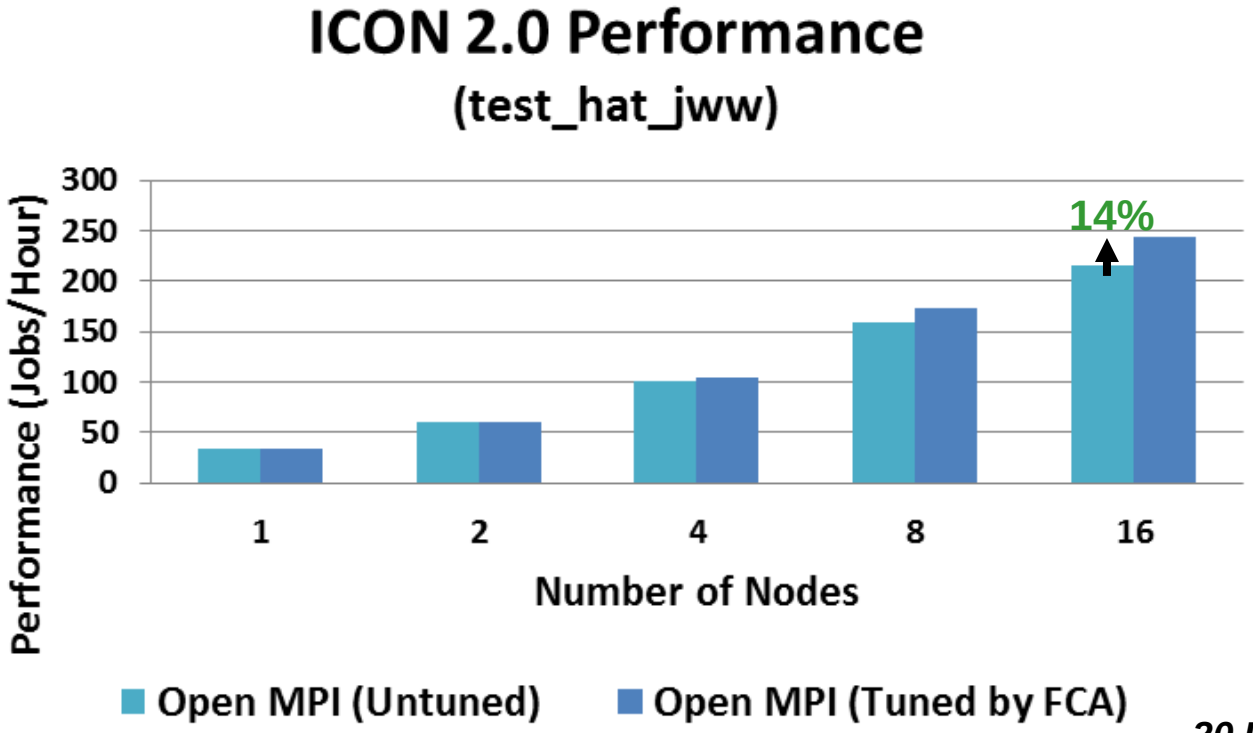
- **Mellanox Fabric Collectives Accelerator (FCA)**
  - Utilized hardware accelerations on the adapter (CORE-Direct)
  - Utilized managed switches capabilities (iCPU)
  - Accelerating MPI collectives operations
  - The world first complete solution for MPI collectives offloads


- **FCA 2.1 supports accelerations/offloading for**
  - MPI Barrier
  - MPI Broadcast
  - MPI AllReduce and Reduce
  - MPI AllGather and AllGatherv



# ICON Performance – Open MPI Tuning

- **Mellanox FCA accelerates ICON performance up to 14%**
  - At 16 nodes, 320 MPI processes
  - Performance benefit increases with cluster size
- **Flags used for enabling FCA:**
  - "-mca coll_fca_enable 1"

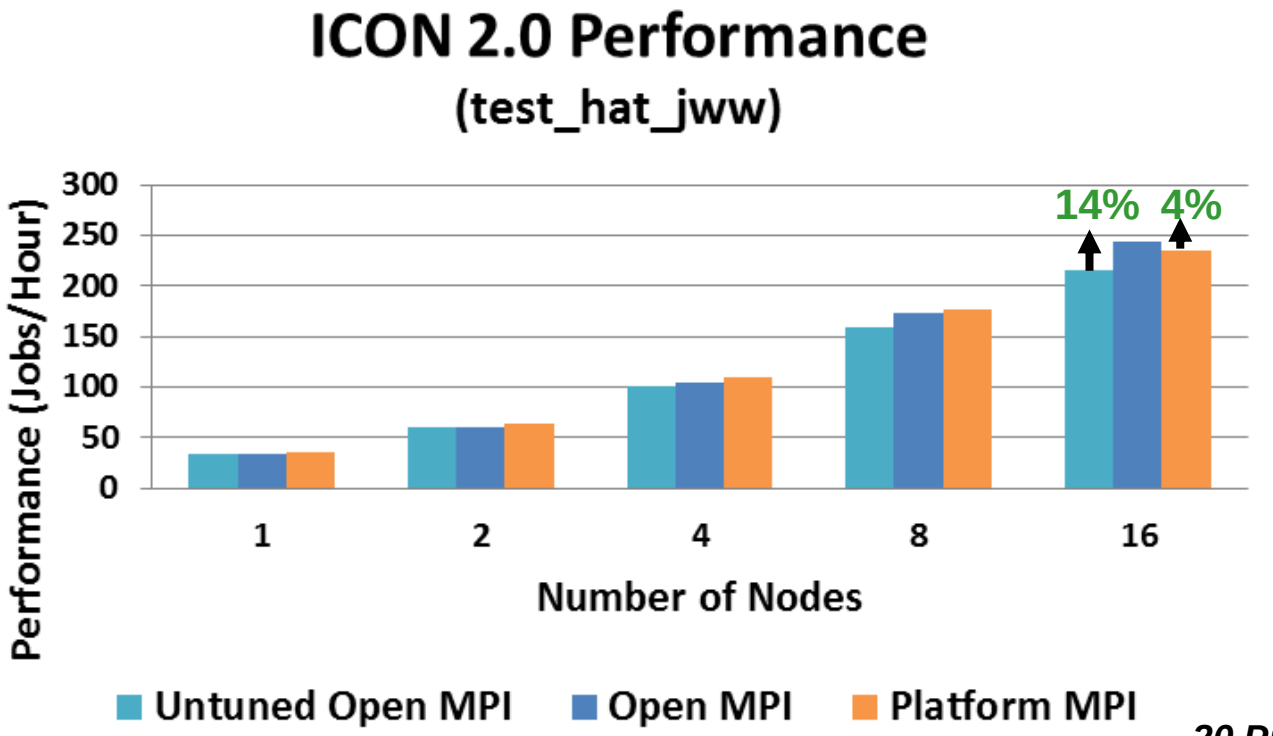



*Intel Ivy Bridge*

*20 Processes/Node*

# ICON Performance – Mellanox FCA

- **FCA enabled Open MPI provides better scalability performance**
  - Over Platform MPI by 4%
- **No extra flags were used for both cases except for enabling processor binding**

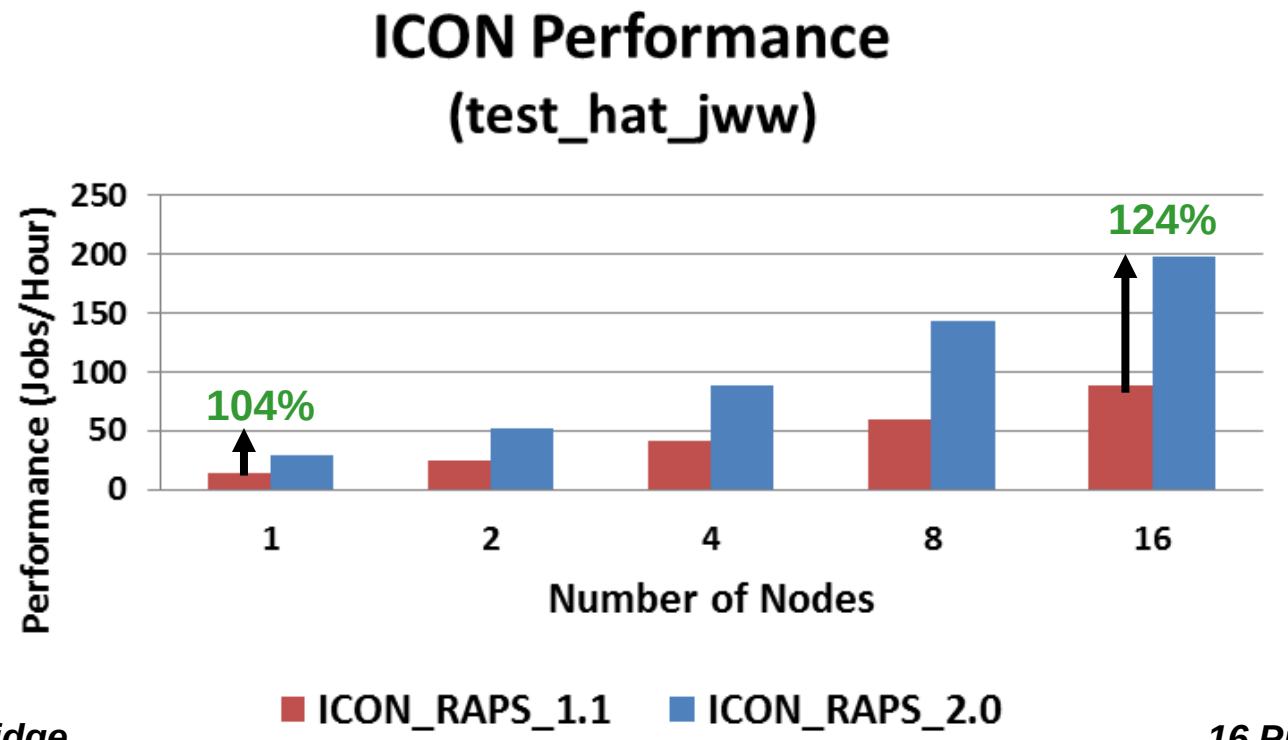



*Intel Ivy Bridge*

*20 Processes/Node*

# ICON Performance – Software Version

- **FDR InfiniBand enables higher cluster productivity**
  - Up to 282-284% of increased productivity over 10/40GbE network for a 16-node job
  - Over 15 times 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

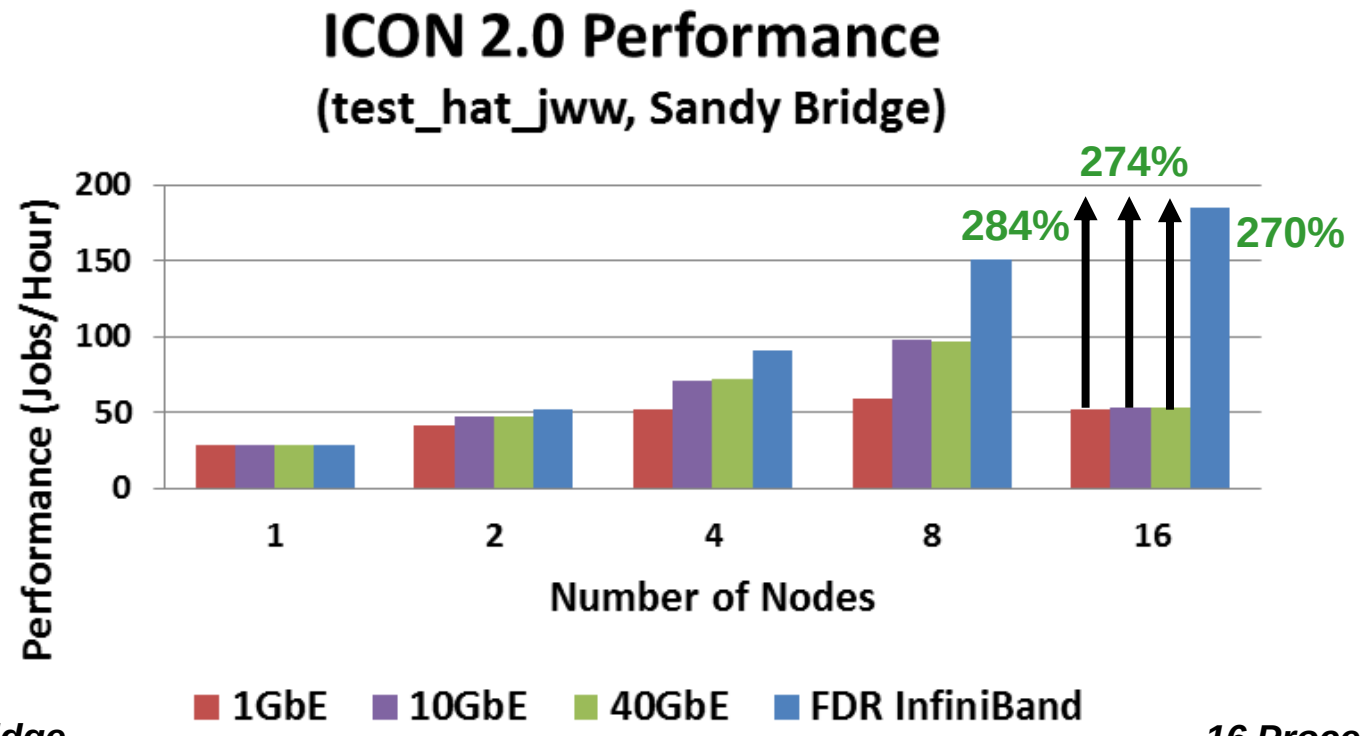



*Intel Sandy Bridge*

*16 Processes/Node*

# ICON Performance – Network Interconnects

- **FDR InfiniBand enables higher cluster productivity**
  - Up to 270-284% of increased productivity over 1/10/40GbE network for a 16-node job
- **ICON demonstrates good scalability using InfiniBand**
  - Performance gain for 1GbE performance is limited after 4-node due to network congestion

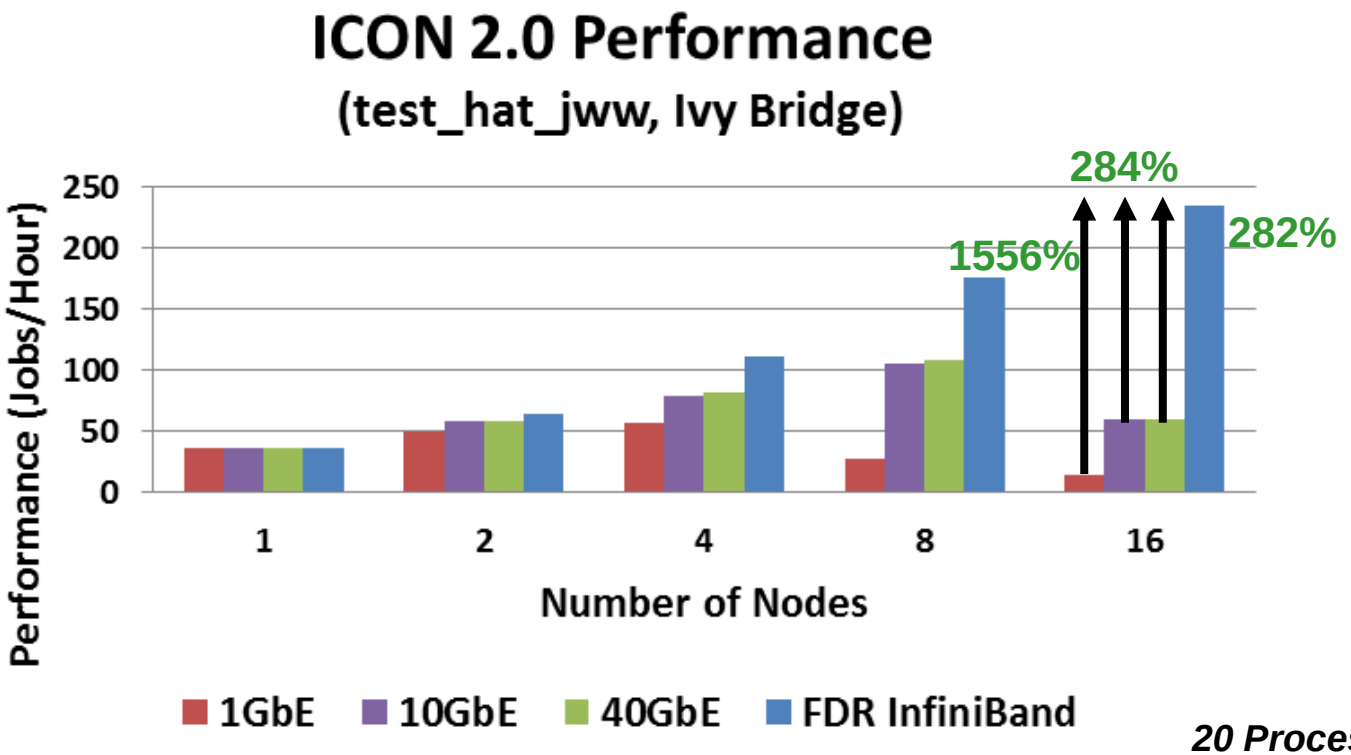



*Intel Sandy Bridge*

*16 Processes/Node*

# ICON Performance – Network Interconnects

- **FDR InfiniBand enables higher cluster productivity**
  - Up to 282-284% of increased productivity over 10/40GbE network for a 16-node job
  - Over 15 times 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

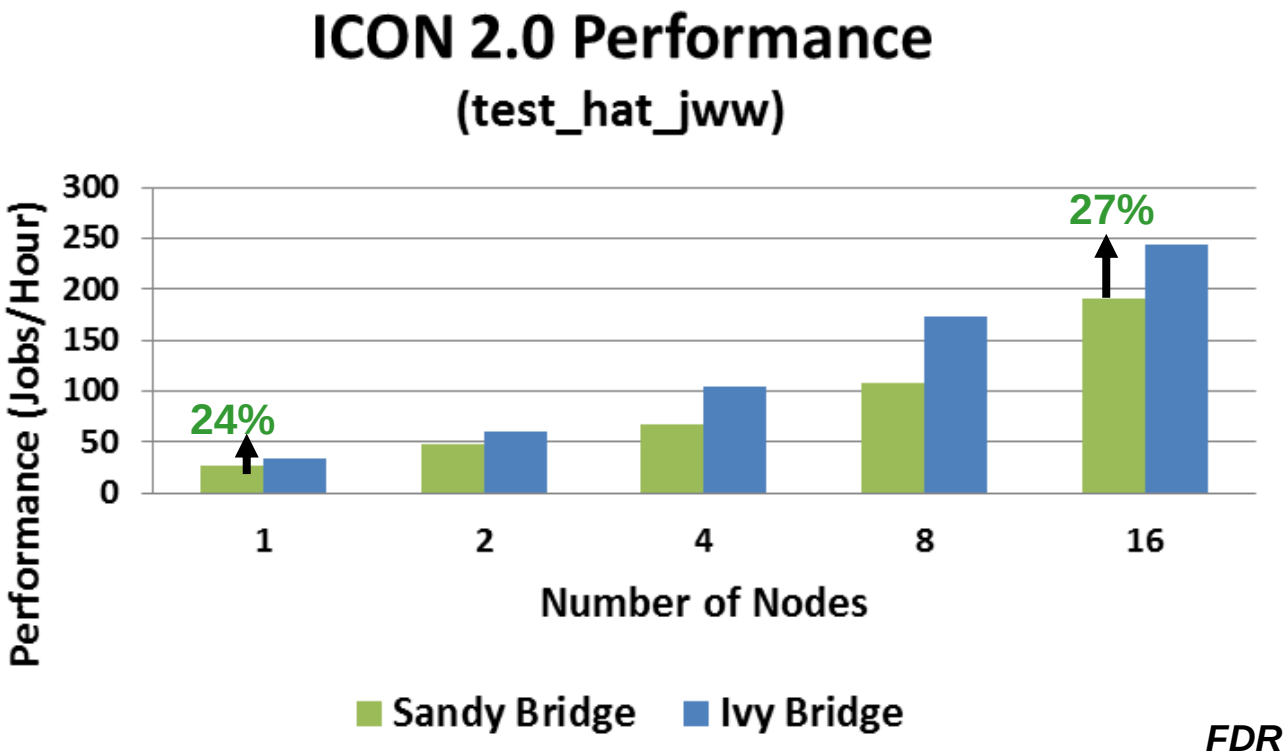



*Intel Ivy Bridge*

*20 Processes/Node*

# ICON Performance – CPU

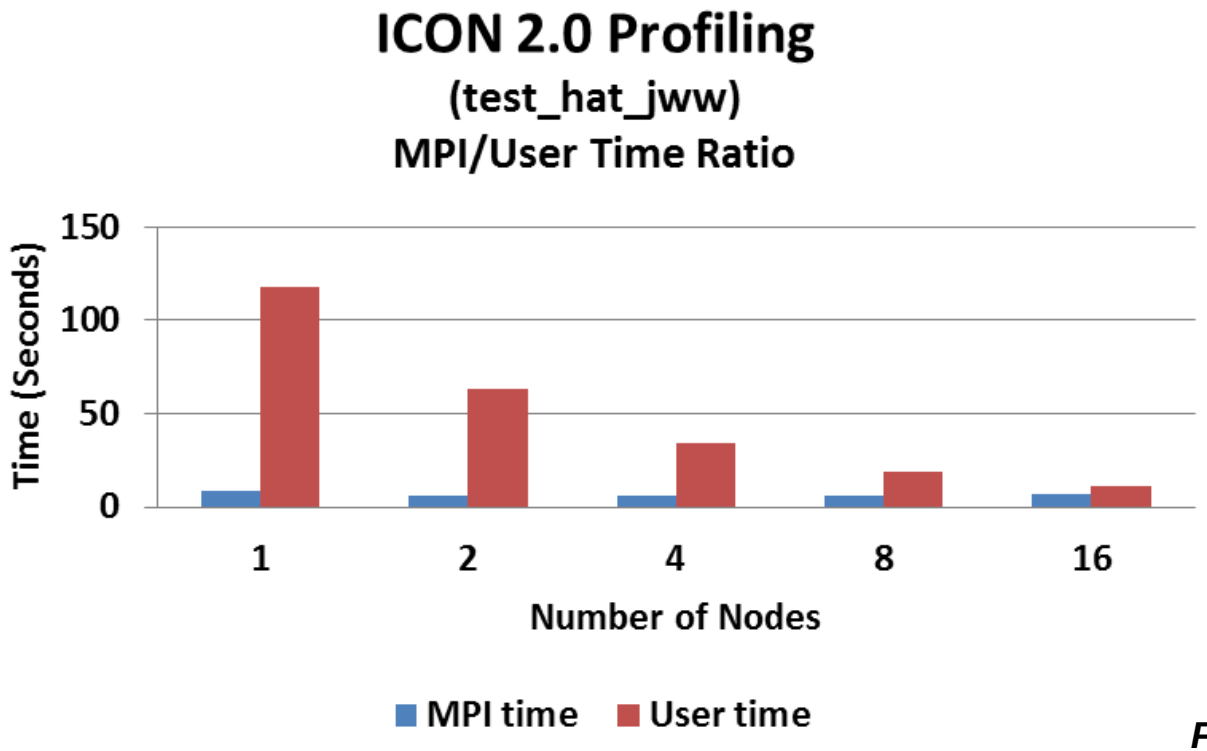
- **Intel Ivy Bridge outperfoms previous generations**
  - Delivers up to 27% higher performance




*FDR InfiniBand*

# ICON Profiling – Interconnects MPI/User Time

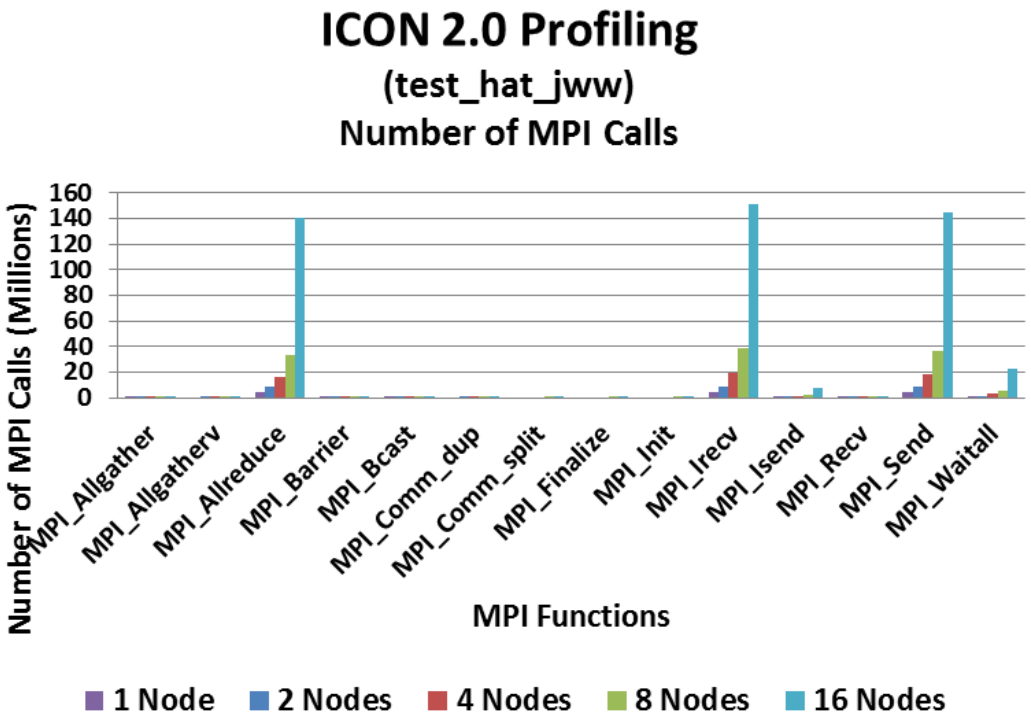
- **InfiniBand reduces the overall runtime of ICON**
  - Time for communication remains the same while user (compute) time reduces as more nodes are added to the cluster
- **InfiniBand allows more system runtime for the actual computation for a job**
  - Network communication accounts for 26% of overall runtime at 8-node w/ FDR InfiniBand

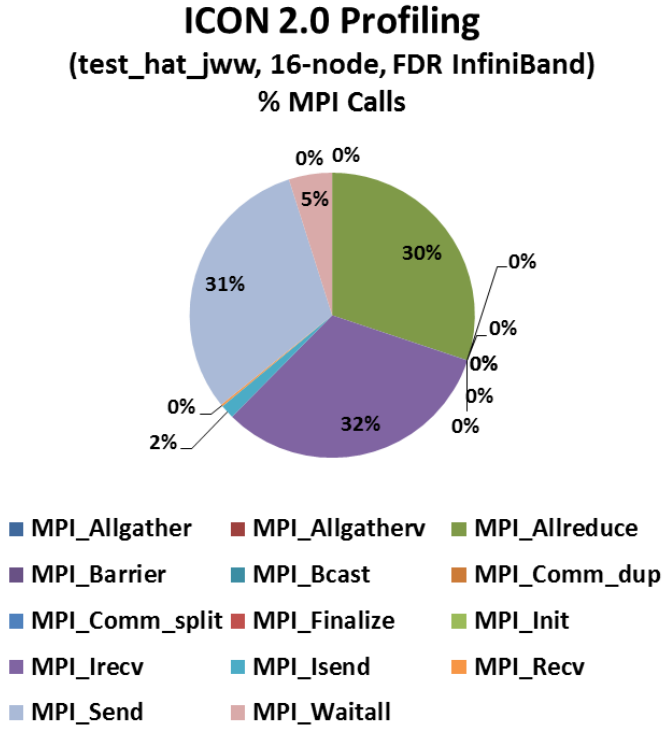



***FDR InfiniBand***

# ICON Profiling – MPI Calls

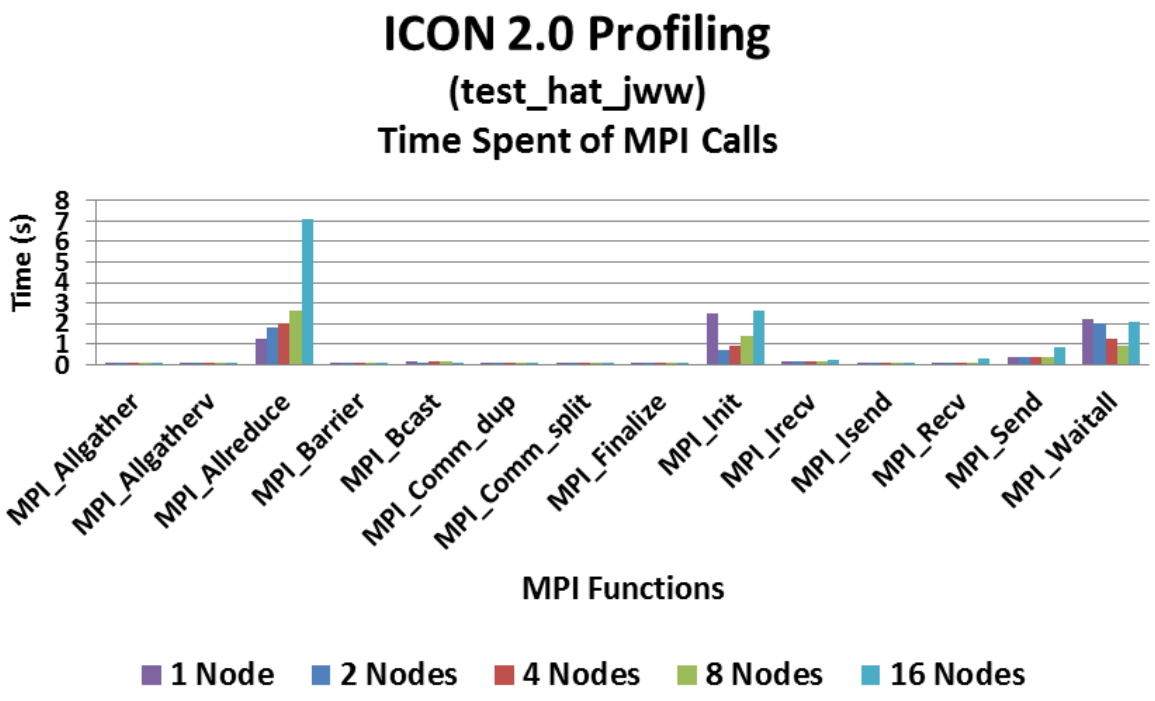
- **MPI_Allreduce, MPI_Irecv and MPI_Send are the most used MPI calls**
  - MPI_Irecv is accounted for 32% of the MPI function calls on a 16-node run
  - MPI_Send is accounted for 31% of the MPI function calls on a 16-node run
  - MPI_Allreduce is accounted for 30% of the MPI function calls on a 16-node run

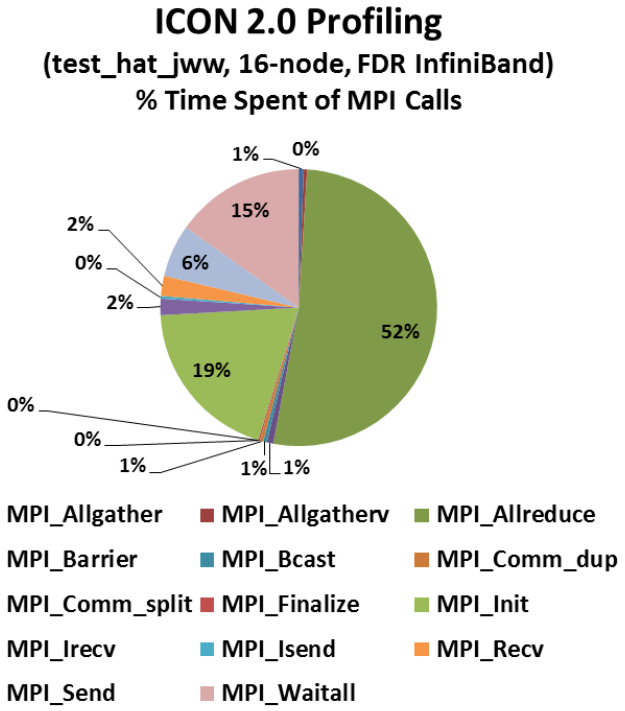


# ICON Profiling – Time Spent by MPI Calls

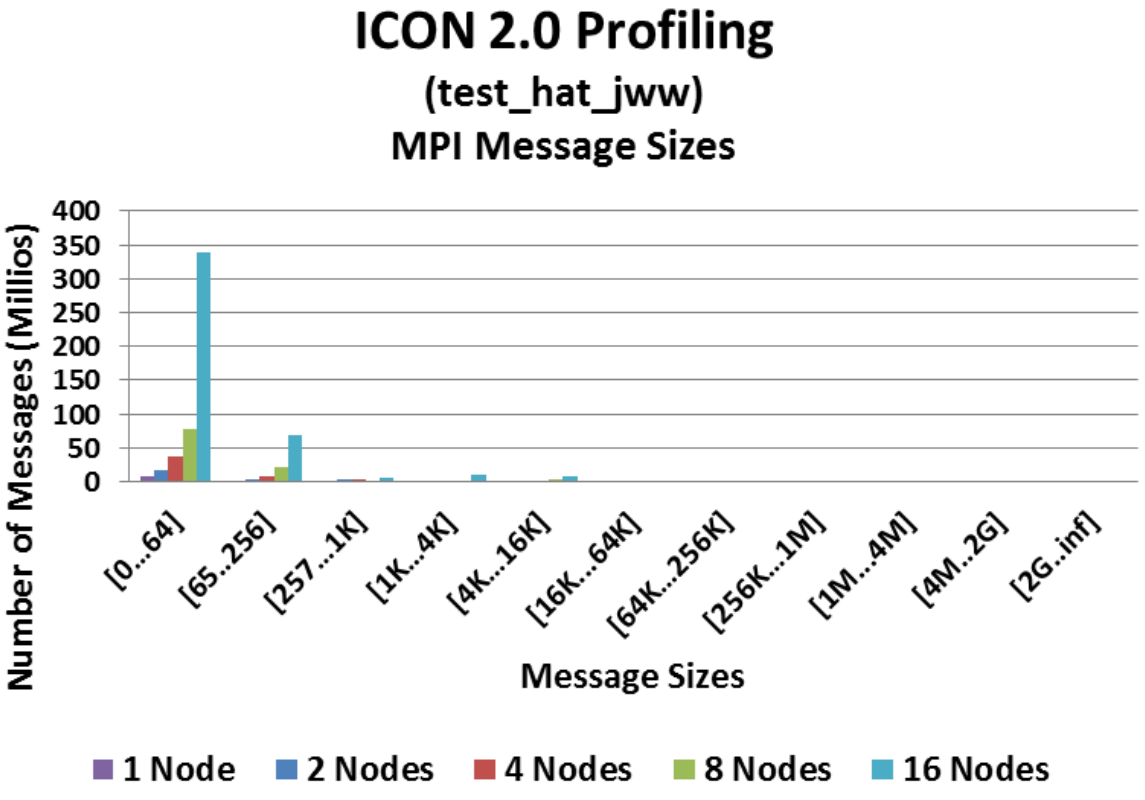
- **Majority of the MPI time is spent on MPI_Sendrecv**
  - MPI_Allreduce(52%), MPI_Init(19%), MPI_Waitall(15%) on 16-node
- **MPI_Allreduce takes more time to complete as the cluster grows**
  - Mellanox FCA can reduce Allreduce time by offloading collective operations to IB HW



# ICON Profiling – MPI Message Sizes

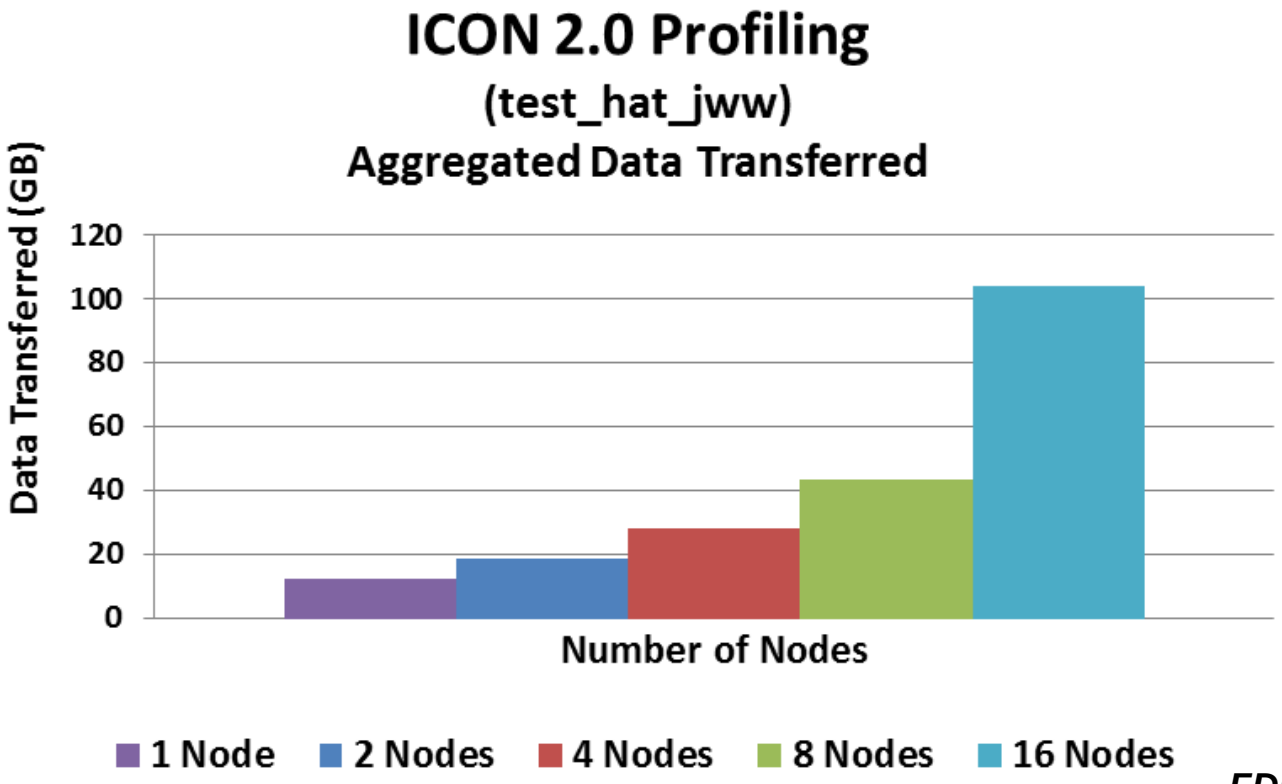
- **Small message sizes are the most dominant**
  - Majority of message sizes are in the 0-64 byte range
  - Large volume of small messages typically means the application is latency sensitive




***FDR InfiniBand***

# ICON Profiling – Aggregated Transfer

- **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 4GB




***FDR InfiniBand***

# ICON – Summary

- **ICON delivers good scalability and performance**
  - ICON can take advantage of additional compute power by using FDR InfiniBand
- **Mellanox FCA delivers better performance for Open MPI**
  - Enabling FCA shows 14% improvement on a 16-node run in Open MPI
  - FCA-enabled Open MPI also shows 4% gain over Platform MPI at 16-node
  - Performance benefit increases with cluster size
- **InfiniBand is needed for ICON to run at the most efficient rate at scale**
  - FDR InfiniBand delivers ~282%-284% of higher performance over 10/40GbE at 16-node
  - 1GbE performance suffers at high scale, FDR IB performs 15 times better 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