



# Himeno

## Performance Benchmark and Profiling

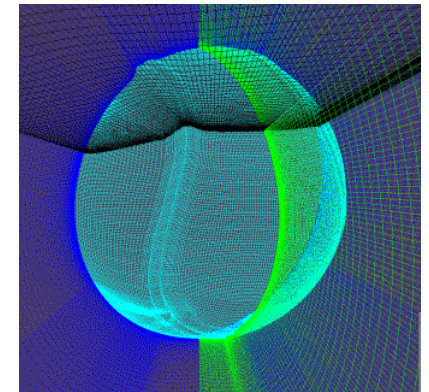
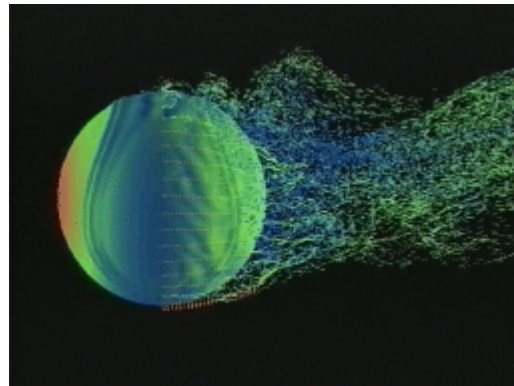
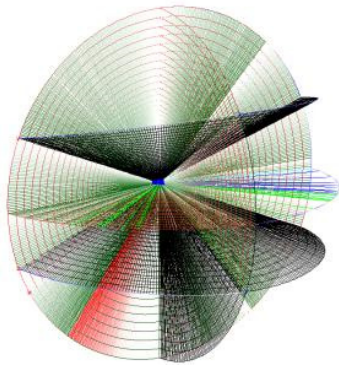
December 2010



- **The following research was performed under the HPC Advisory Council activities**
  - Participating vendors: AMD, Dell, Mellanox
  - Compute resource - HPC Advisory Council Cluster Center
- **For more info please refer to**
  - [http:// www.amd.com](http://www.amd.com)
  - [http:// www.dell.com/hpc](http://www.dell.com/hpc)
  - <http://www.mellanox.com>
  - [http:// acc.riken.jp/HPC\\_e/himenobmt\\_e.html](http://acc.riken.jp/HPC_e/himenobmt_e.html)

- **Himeno**

- Developed by Dr. Ryutaro Himeno, RIKEN, Japan
- Intends to evaluate performance of incompressible fluid analysis code
- Takes in measurements to precede major loops in solving the Poisson's equation solution using the Jacobi iteration method
- Available under the LGPL 2.0 or later



- **The following was done to provide best practices**
  - Himeno performance benchmarking
  - Interconnect performance comparisons
  - Understanding Himeno communication patterns
  - Ways to increase Himeno productivity
  - Compilers and MPI libraries comparisons
  
- **The presented results will demonstrate**
  - The scalability of the compute environment to provide nearly linear application scalability
  - The capability of Himeno to achieve scalable productivity
  - Considerations for performance optimizations

- **Dell™ PowerEdge™ R815 11-node cluster**
- **AMD™ Opteron™ 6174 (code name “Magny-Cours”) 12-cores @ 2.2 GHz CPUs**
- **4 CPUs per server node**
- **Mellanox ConnectX-2 VPI adapters for 40Gb/s QDR InfiniBand and 10Gb/s Ethernet**
- **Mellanox M3600 36-Port 40Gb/s QDR InfiniBand switch**
- **Fulcrum based 10Gb/s Ethernet switch**
- **Memory: 128GB memory per node DDR3 1333MHz**
- **OS: RHEL 5.5, MLNX-OFED 1.5.1 InfiniBand SW stack**
- **MPI: Intel MPI 4.0, Open MPI 1.5.1, Platform MPI 8.0.1**
- **Compilers: GNU Compilers 4.1.2 & 4.4, Intel Compilers 11.1, Open64 4.2.4, PGI 10.9**
- **Application: HimenoBMTxp (f77\_xp\_mpi)**
- **Benchmark Workload: “L” Grid size (512x256x256)**

- **HPC Advisory Council Test-bed System**
- **New 11-node 528 core cluster - featuring Dell PowerEdge™ R815 servers**
  - Replacement system for Dell PowerEdge SC1435 (192 cores) cluster system following 2 years of rigorous benchmarking and product EOL
    - System to be redirected to explore HPC in the Cloud applications
- **Workload profiling and benchmarking**
  - Characterization for HPC and compute intense environments
  - Optimization for scale, sizing and configuration and workload performance
  - Test-bed Benchmarks
    - RFPs
    - Customers/Prospects, etc
  - ISV & Industry standard application characterization
  - Best practices & usage analysis



# About Dell PowerEdge™ Platform Advantages

## Best of breed technologies and partners

Combination of AMD™ Opteron™ 6100 series platform and Mellanox ConnectX InfiniBand on Dell HPC

Solutions provide the ultimate platform for speed and scale

- Dell PowerEdge R815 system delivers 4 socket performance in dense 2U form factor
- Up to 48 core/32DIMMs per server – 1008 core in 42U enclosure

## Integrated stacks designed to deliver the best price/performance/watt

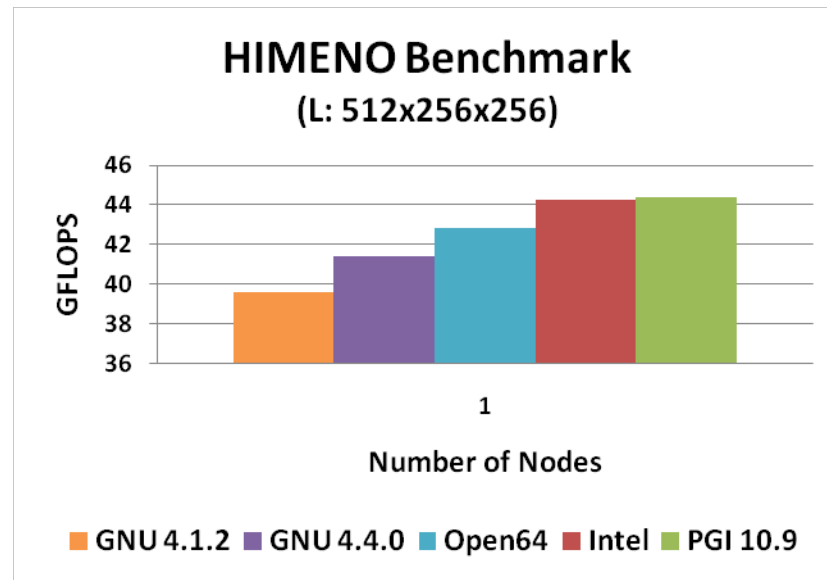
- 2x more memory and processing power in half of the space
- Energy optimized low flow fans, improved power supplies and dual SD modules

## Optimized for long-term capital and operating investment protection

- System expansion
- Component upgrades and feature releases



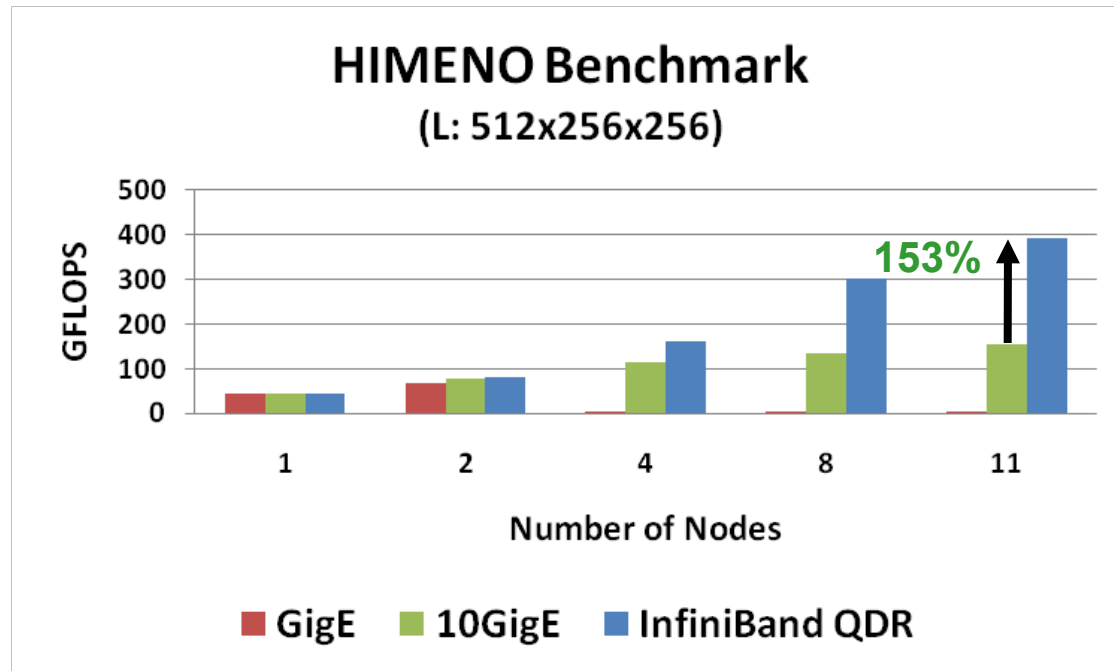
- PGI provide the best CPU utilization among the compilers tested
- Compiler flags used:
  - GNU412/GNU44: “-O3 -ffast-math -ftree-vectorize -ftree-loop-linear -funroll-loops”
  - Open64: -O3 -OPT:unroll\_level=2 -OPT:Ofast -ipa -ftz -Ofast -OPT:keep\_ext=on -msse3 -mso -HP -INLINE -LNO -ffast-math
  - Intel “-O3 -ip -xSSE2 -w -ftz -align all -fno-alias -fp-model fast=1 -convert big\_endian”
  - PGI: “-fastsse -Mipa=fast,inline -Mconcur”



*Higher is better*

*Open MPI 1.5  
12 Cores/Node*

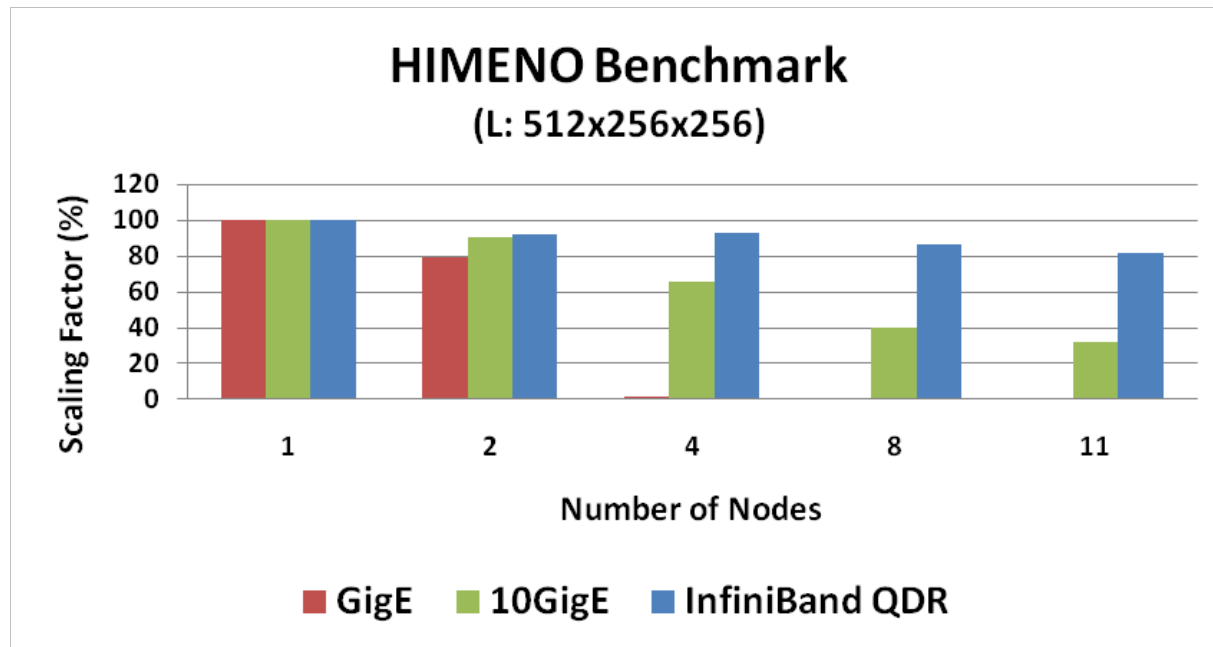
- **InfiniBand enables higher scalability**
  - Up to 153% gain over 10GigE at 11-node (528 processes) with the L dataset
- **The performance of GigE plummets after 2 nodes**
  - The effect of MPI saturating the Ethernet network



*Higher is better*

**12 Cores/Node**

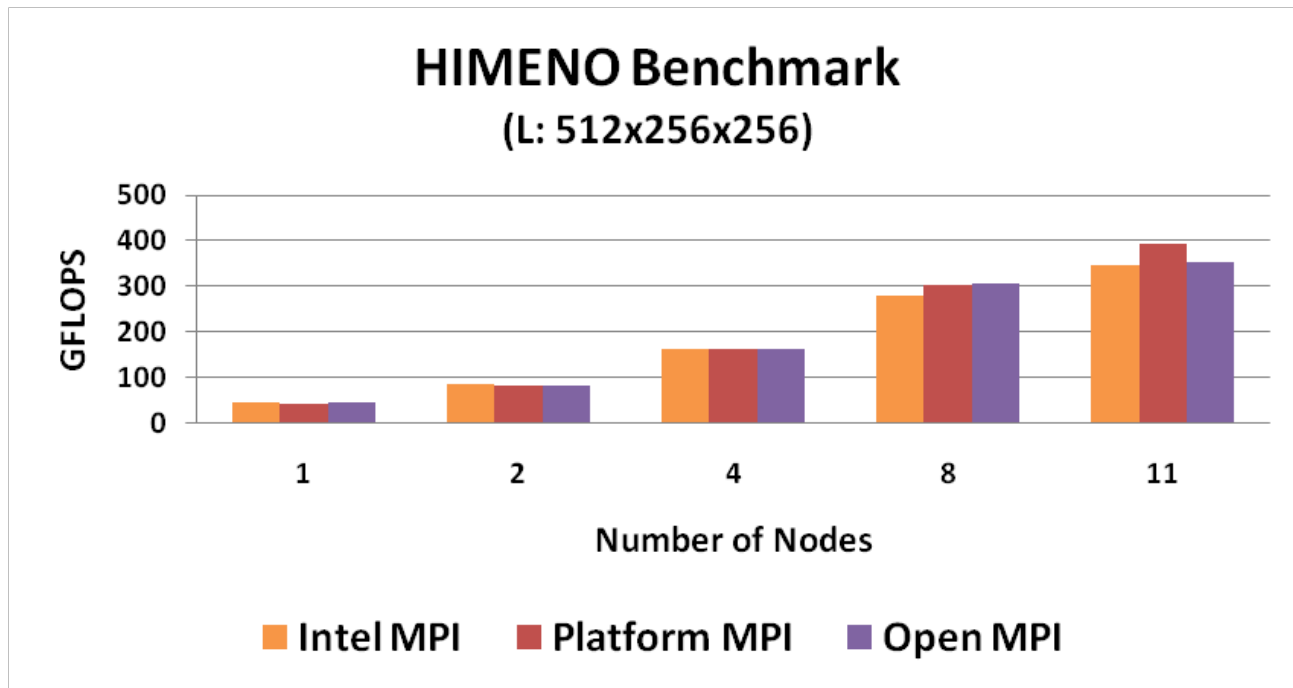
- L dataset demonstrates network dependency
- Scalability of 10 GigE dropped down to 32% at 11-node
  - While scalability of InfiniBand QDR maintains above 80% throughout



*Higher is better*

**12 Cores/Node**

- **Platform MPI performs better at 528 cores (or 11-node)**
  - Both Platform MPI and Open MPI perform slightly performance as the cluster scales

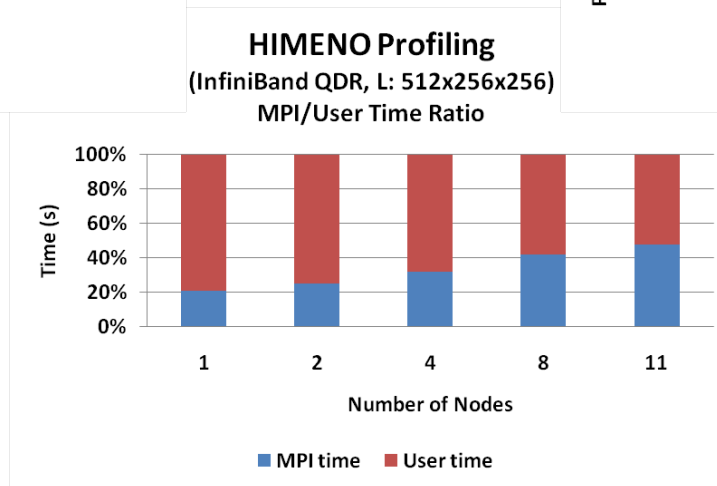
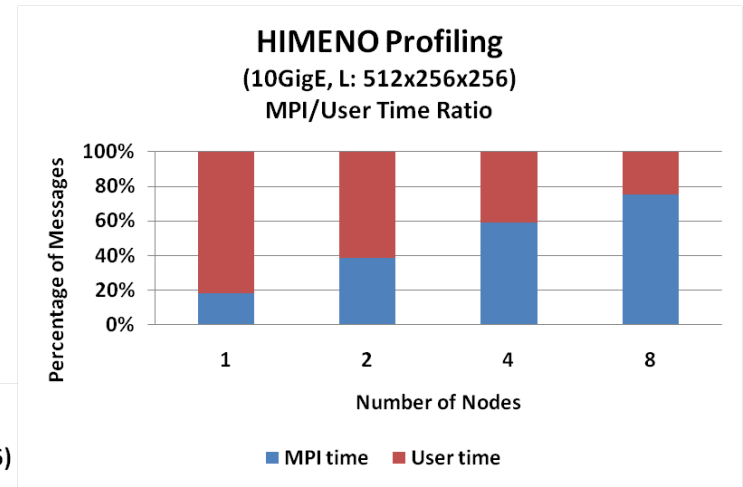
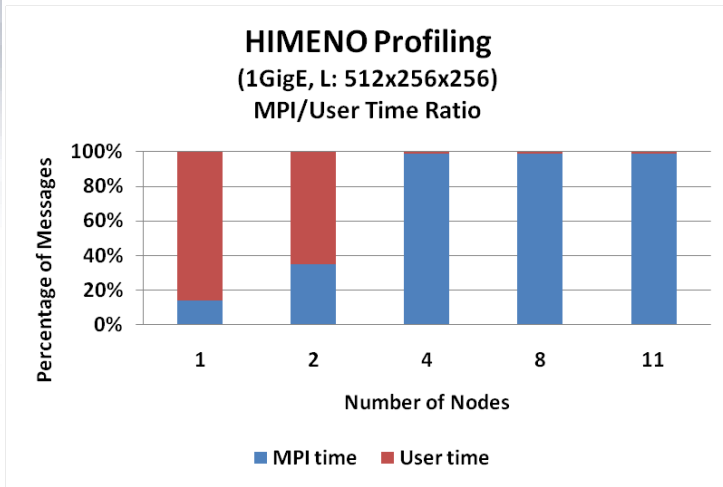


*Higher is better*

**12 Cores/Node**

# Himeno Profiling – MPI/User Time Ratio

- Shows InfiniBand maintain its low consumption by MPI Communications
- 1GigE is overtaken by MPI communications after 2-node (96 processes)



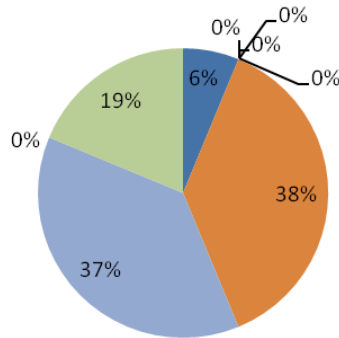
*Higher is better*

**12 Cores/Node**

# Himeno Profiling – Number of MPI Calls

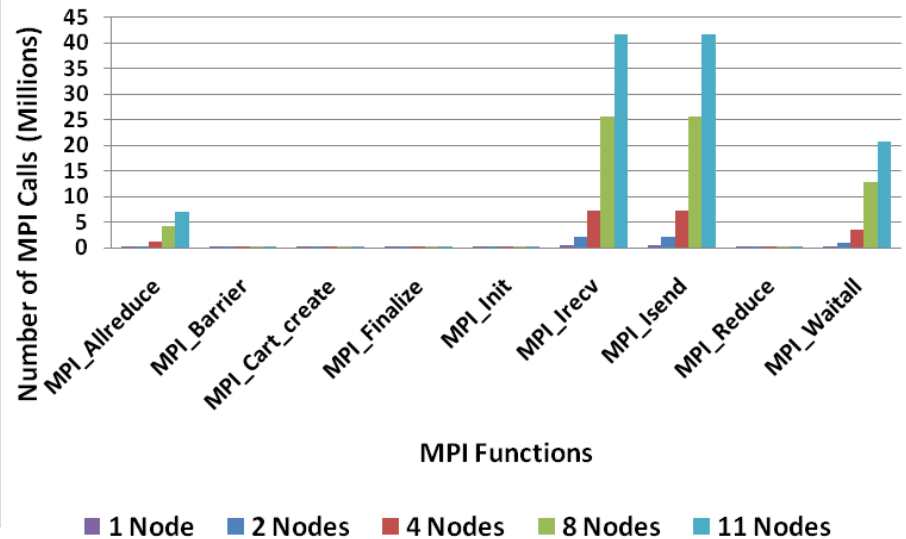
- The most used MPI functions are **MPI\_Isend** and **MPI\_Irecv**
  - Each accounted for 38% of all MPI functions on a 14-node job

**HIMENO Profiling**  
(InfiniBand QDR, 11-node, L: 512x256x256)  
% of MPI Calls

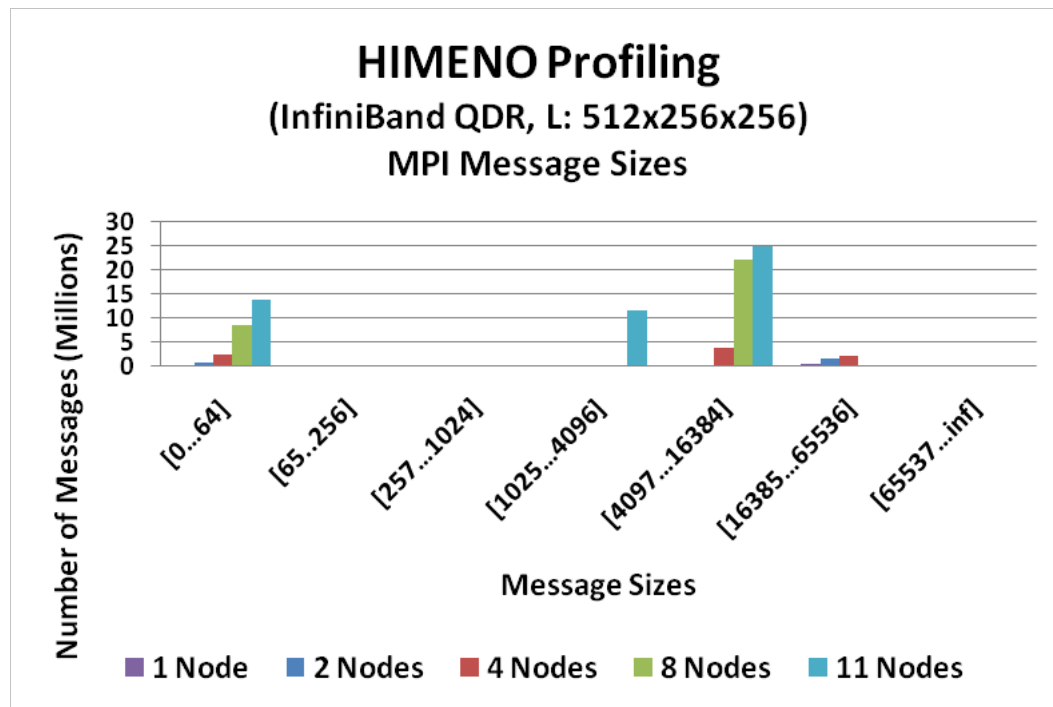


■ MPI\_Allreduce ■ MPI\_Barrier ■ MPI\_Cart\_create  
■ MPI\_Finalize ■ MPI\_Init ■ MPI\_Irecv  
■ MPI\_Isend ■ MPI\_Reduce ■ MPI\_Waitall

**HIMENO Profiling**  
(InfiniBand QDR, L: 512x256x256)  
Number of MPI Calls



- **Messages increase accelerates with the node count increases**
- **Majority of the MPI message sizes are**
  - in the range from 16KB to 64KB for the L dataset
  - In the range from 64B to 256B and beyond 64KB for XL dataset



- **PGI provide the best CPU utilization among the compilers tested**
- **Platform MPI performs better at 528 cores (or 11-node)**
- **L dataset is network dependent**
- **Scalability of 10GigE dropped off to 32% at 11-node**
  - While scalability of InfiniBand QDR maintains above 80% throughout

# Thank You

## HPC Advisory Council



All trademarks are property of their respective owners. All information is provided "As-Is" without any kind of warranty. The HPC Advisory Council makes no representation to the accuracy and completeness of the information contained herein. HPC Advisory Council Mellanox undertakes no duty and assumes no obligation to update or correct any information presented herein