Himeno
Performance Benchmark and Profiling

December 2010
• The following research was performed under the HPC Advisory Council activities
  – Participating vendors: Intel, Dell, Mellanox
  – Compute resource - HPC Advisory Council Cluster Center

• For more info please refer to
  – http://www.dell.com
  – http://www.intel.com
  – http://www.mellanox.com
  – http://accc.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
Objectives

• **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
Test Cluster Configuration

- **Dell™ PowerEdge™ M610 14-node cluster**
  - Six-Core Intel X5670 @ 2.93 GHz CPUs
  - Memory: 24GB memory, DDR3 1333 MHz
  - OS: CentOS5U4, OFED 1.5.1 InfiniBand SW stack

- **Intel Cluster Ready certified cluster**

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

- **MPI: Intel MPI 4.0 U1, Open MPI 1.5, Platform MPI 8.0.1**

- **Compilers: GNU Compilers 4.1.2 and 4.4, Intel Compilers 12.0.0**

- **Application: HimenoBMTxp (f77_xp_mpi)**

- **Benchmark dataset: “XL” Grid size (1024x512x512) and “L” Grid size (512x256x256)**
• **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
Dell PowerEdge Servers helping Simplify IT

• **System Structure and Sizing Guidelines**
  – 14-node cluster build with Dell PowerEdge™ M610 blades server
  – 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
Himeno Performance – Compilers

- Intel compilers provide better CPU cores utilization
- Compiler flags used:
  - Intel "-O3 -ip -xSSE4.2 -w -ftz -align all -fno-alias -fp-model fast=1 -convert big_endian"
  - GNU: “-O3 -ffast-math -ftree-vectorize -ftree-loop-linear -funroll-loops”

**HIMENO Benchmark**
(XL: 1024x512x512)

Higher is better
• All MPI implementations performs generally the same
  – Intel MPI shows slightly better performance as the cluster scales
• **InfiniBand enables higher scalability**
  – Up to 86% higher performance than Ethernet at 14-node with the L dataset
  – Up to 51% higher performance than Ethernet at 14-node with the XL dataset
Himeno Performance – Scalability

- **XL dataset shows better scalability than L dataset**
  - More work can be processed with XL by increasing the node count

- **L dataset is more network dependent than XL dataset**
  - XL presumably involves more CPU computation thus less reliant on inter-nodal communications

**HIMENO Benchmark**
(L: 512x256x256)

**HIMENO Benchmark**
(XL: 1024x512x512)

*Higher is better*
• The MPI/User time ratio shows no bottleneck by MPI over IB network
  – Both dataset shows more than 80% of the time spent on user code
  – A small time percentage is spent for communications between the MPI ranks
• The share of the computational time becomes larger for the XL dataset
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
- The number of MPI calls dropped by 86% from L to XL dataset
  - While the ratio of the MPI calls remains the same
  - Reflects that more computation for XL rather than communications
Himeno Profiling – Time Spent of by MPI Calls

- MPI_Init becomes the surprise leader in MPI time consumer
  - Since MPI time becomes insignificant compared to actual computation time
Himeno Profiling – MPI Message Sizes

- Messages increase accelerates with the node count increases
- The XL involves less communications but larger messages
- 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
Summary

- The MPI/User time ratio shows no bottleneck by MPI over IB network
  - Both dataset shows more than 80% of the time spent on user code
  - A small time percentage is spent for communications between the MPI ranks
- The share of the computational time becomes larger for the XL dataset
- The number of MPI calls dropped by 86% from L to XL dataset
- Since MPI becomes insignificant, thus all MPI performs generally the same
- InfiniBand enables higher scalability
  - Up to 86% higher performance than Ethernet at 14-node with the L dataset
  - Up to 51% higher performance than Ethernet at 14-node with the XL dataset
- The most used MPI functions are MPI_Isend and MPI_Irecv
  - Each accounted for 38% of all MPI functions on a 14-node job
- MPI_Init is the leader in MPI time consumer
  - Since MPI time becomes insignificant compared to actual computation time
- Majority of the MPI message sizes are in the range from 16KB to 64KB
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.