LAMMPS
Performance Benchmark and Profiling

July 2012
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://www.mellanox.com)
LAMMPS

- Large-scale Atomic/Molecular Massively Parallel Simulator
  - Classical molecular dynamics code which can model:
    - Atomic
    - Polymeric
    - Biological
    - Metallic
    - Granular, and coarse-grained systems

- LAMMPS runs efficiently in parallel using message-passing techniques
  - Developed at Sandia National Laboratories
  - An open-source code, distributed under GNU Public License
Objectives

• The following was done to provide best practices
  – LAMMPS performance benchmarking
  – Interconnect performance comparisons
  – Ways to increase LAMMPS productivity
  – MPI libraries comparisons

• The presented results will demonstrate
  – The scalability of the compute environment to provide nearly linear application scalability
  – The capability of LAMMPS to achieve scalable productivity
  – Considerations for performance optimizations
Test Cluster Configuration

- **Dell™ PowerEdge™ R815 11-node (704-core) cluster**
  - Memory: 128GB memory per node DDR3 1333MHz, BIOS version 2.8.2
  - 4 CPU sockets per server node
- **AMD™ Opteron™ 6276 (code name “Interlagos”) 16-core @ 2.3 GHz CPUs**
- Mellanox ConnectX®-3 VPI Adapters and IS5030 36-Port InfiniBand switch
- OS: SLES 11 SP2, MLNX-OFED 1.5.3 InfiniBand SW stack
- MPI: Open MPI 1.5.5, Platform MPI 8.2.1
- Compilers: Open64 4.5.1
- Libraries: ACML 5.1.0, FFTW 2.1.5
- Application: LAMMPS-4Jul12
- Benchmark workload:
  - Rhodo -Rhodopsin protein in solvated lipid bilayer, CHARMM force field with a 10 Angstrom LJ cutoff
Dell™ PowerEdge™ R815 11-node cluster

• **HPC Advisory Council Test-bed System**

• **New 11-node 704 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™ 6200 series platform and Mellanox ConnectX®-3 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 64 core/32DIMMs per server – 1344 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
Both MPIs perform at the same level for this dataset
- Performance shown by the 2 MPIs are equally as good
- Both MPI allows LAMMPS to efficiently scale to many systems

LAMMPS Benchmark
(Scaled-size Rhodopsin Protein)

Higher is better

Performance Rating = 32,000 (not 32K) × the number of cores divided by the wall-clock simulation time for 100 steps
• XRC and SRQ enhance scalability Infiniband performance at high node count
  – XRC boosts performance by 16% at 11-node
  – SRQ boosts performance by 6% at 11 node
• No difference in performance is seen between SLES11 SP2 over RHEL6 U2
  – No performance gain is seen by using one over the other operating system
LAMMPS Performance – CPU Cores Placement

- Comparing jobs running with 32 PPN versus 64 PPN (processes per node)
  - Running with 4 CPUs (64PPN) is 69% faster than jobs running with 2 CPUs (32 PPN)
  - The 32 PPN case uses 2 CPU sockets while the 64 PPN case uses 4 CPU sockets

- CPU core frequency jumps when only 1 core in each core pair is active
  - While the non-active core is in sleep mode
  - Running with both cores is 41% faster than running with only 1 active core in a core pair

LAMMPS Benchmark
(Scaled-size Rhodopsin Protein)

Higher is better

Platform MPI
• **AMD “Interlagos”** provides higher scalability than previous generations
  
  - Improved by 142% vs “Shanghai”
  - AMD Opteron 2382 “Shanghai” with InfiniBand DDR and PCIe Gen1
  - AMD Opteron 6276 “Interlagos” with InfiniBand QDR and PCIe Gen2
LAMMPS Performance – Compiler and Libraries

- Both ACML and FFTW shows equally good performance
  - No difference is seen between either of the 2 math libraries
- Compiled using compiler flags for AVX, FMA4 and Interlagos instructions:
  - `-march=bdver1 -mavx -mfma4`

LAMMPS Benchmark
(Scaled-size Rhodopsin Protein)

Higher is better

- FFTW 2.1.5 + Open64 4.5.1
- ACML 5.1.0 + Open64 4.5.1

64 Cores/Node
LAMMPS Profiling – MPI/User Time Ratio

- Communication time share grows steadily as more nodes are used
  - The scaled-size data problem causes more computation needs to take place

LAMMPS Profiling
(Scaled-size Rhodopsin Protein)
MPI/User Time Ratio

Higher is better

64 Cores/Node
LAMMPS Profiling – Number of MPI Calls

- The most used MPI function are MPI_Send, MPI_Wait, and MPI_Irecv
  - Each accounts for 27% of all the MPI calls made
- Point-to-point sends and receives are called heavily

![LAMMPS Profiling Diagram](image-url)
The most time consuming MPI function is MPI_Send
- MPI_Send accounts for 41% of all MPI time at 11-node
LAMMPS Profiling – MPI Message Sizes

- Majority of the MPI message sizes are concentrated in the midrange
  - Spike between 4KB to 16KB
  - The rest of the concentrations are in 16KB to 64KB
LAMMPS Profiling – Data Transfer / Process

- As the cluster scales, more data is driven to each rank and each node
  - Due to the scaled-size nature of the dataset, it causes more data to be generated
LAMMPS Profiling – Aggregated Data Transfer

- Aggregated data transfer refers to:
  - Total amount of data being transferred in the network between all MPI ranks collectively
- The total data transfer increases as the cluster scales
- The larger the dataset is, more data will be sent to the network
  - Exponential growth of data exchanges that takes place on the network transfer

**LAMMPS Profiling**
(Scaled-size Rhodopsin Protein)
Aggregated Data Transferred

![Data Transfer Chart]

- 1 Node
- 2 Nodes
- 4 Nodes
- 8 Nodes
- 11 Nodes
Summary

• Balanced hardware allows LAMMPS to achieve good performance and scalability

• CPU:
  – Using system with 4 CPUs versus 2 CPUs provides 69% gain in productivity on LAMMPS

• OS:
  – Running jobs in the SLES and RHEL provides similar system productivity for LAMMPS

• Compiler:
  – Both ACML and FFTW shows equally good performance
  – Compiler flags for AVX, FMA4 and Interlagos instructions: (-march=bdver1 -mavx -mfma4)

• InfiniBand:
  – XRC boosts performance by 16% at 11-node
  – SRQ boosts performance by 6% at 11 node
Thank You
HPC Advisory Council