LAMMPS Performance Benchmark and Profiling

September 2009
Note

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

• The participating members would like to thank Lawrence Livermore National Laboratory for their guidelines

• For more info please refer to
• **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 presented research was done to provide best practices
  – LAMMPS performance benchmarking
  – Interconnect performance comparisons
  – Understanding LAMMPS communication patterns
  – Power-efficient simulations

• 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 power saving through balanced system configuration
**Test Cluster Configuration**

- **Dell™ PowerEdge™ SC 1435 24-node cluster**
- **Quad-Core AMD Opteron™ 2382 (“Shanghai”) CPUs**
- **Mellanox® InfiniBand ConnectX® 20Gb/s (DDR) HCAs**
- **Mellanox® InfiniBand DDR Switch**
- **Memory: 16GB memory, DDR2 800MHz per node**
- **OS: RHEL5U3, OFED 1.4.1 InfiniBand SW stack**
- **MPI: HP-MPI 2.3**
- **Application: LAMMPS-21Aug2009**
- **Benchmark Workload**
  - EAM - Metallic solid, Cu EAM potential with 4.95 Angstrom cutoff
  - Rhodo - Rhodopsin protein in solvated lipid bilayer, CHARMM force field with a 10 Angstrom LJ cutoff
Mellanox InfiniBand Solutions

- **Industry Standard**
  - Hardware, software, cabling, management
  - Design for clustering and storage interconnect

- **Performance**
  - 40Gb/s node-to-node
  - 120Gb/s switch-to-switch
  - 1us application latency
  - Most aggressive roadmap in the industry

- **Reliable with congestion management**

- **Efficient**
  - RDMA and Transport Offload
  - Kernel bypass
  - CPU focuses on application processing

- **Scalable for Petascale computing & beyond**

- **End-to-end quality of service**

- **Virtualization acceleration**

- **I/O consolidation Including storage**

![The InfiniBand Performance Gap is Increasing](chart.png)

InfiniBand Delivers the Lowest Latency
Quad-Core AMD Opteron™ Processor

- **Performance**
  - Quad-Core
    - Enhanced CPU IPC
    - 4x 512K L2 cache
    - 6MB L3 Cache
  - Direct Connect Architecture
    - HyperTransport™ Technology
    - Up to 24 GB/s peak per processor
  - Floating Point
    - 128-bit FPU per core
    - 4 FLOPS/clk peak per core
  - Integrated Memory Controller
    - Up to 12.8 GB/s
    - DDR2-800 MHz or DDR2-667 MHz
- **Scalability**
  - 48-bit Physical Addressing
- **Compatibility**
  - Same power/thermal envelopes as 2nd / 3rd generation AMD Opteron™ processor
Dell PowerEdge Servers helping Simplify IT

- System Structure and Sizing Guidelines
  - 24-node cluster build with Dell PowerEdge™ SC 1435 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
Dell PowerEdge™ Server Advantage

- Dell™ PowerEdge™ servers incorporate AMD Opteron™ and Mellanox ConnectX InfiniBand to provide leading edge performance and reliability
- Building Block Foundations for best price/performance and performance/watt
- Investment protection and energy efficient
- Longer term server investment value
- Faster DDR2-800 memory
- Enhanced AMD PowerNow!
- Independent Dynamic Core Technology
- AMD CoolCore™ and Smart Fetch Technology
- Mellanox InfiniBand end-to-end for highest networking performance
LAMMPS Benchmark Results

- **Input Dataset: EAM**
  - Metallic solid, Cu EAM potential with 4.95 Angstrom cutoff (45 neighbors per atom), NVE integration

- **InfiniBand provides higher utilization, performance and scalability**
  - Up to 72% higher performance versus GigE with 24 nodes configuration

Higher is better

Performance Rating = 32,000 (not 32K) × the number of cores divided by the wall-clock simulation time
LAMMPS Benchmark Results

- **Input Dataset: Rhodo**
  - Rhodopsin protein in solvated lipid bilayer, CHARMM force field with a 10 Angstrom LJ cutoff (440 neighbors per atom), particle-particle particle-mesh for long-range Coulombics, NPT integration

- **InfiniBand provides higher utilization, performance and scalability**
  - Up to 106% higher performance versus GigE with 24 nodes configuration

**LAMMPS Benchmark Result**
(Scaled-Size Rhodopsin Protein)

**Higher is better**
Performance Rating = \(32,000\) (not 32K) \(\times\) the number of cores divided by the wall-clock simulation time

8-cores per node
LAMMPS Productivity

- **Test cases**
  - Single job, run on eight cores per server
  - 2 simultaneous jobs, each runs on four cores per server

- **Both test cases produces similar productivity**
  - Dataset EAM shows nearly identical performance between two cases
  - Dataset Rhodospin shows up to 10% more jobs per day with 24 nodes

---

**LAMMPS Productivity Result (Scaled-Size EAM Metallic Solid)**

**LAMMPS Productivity Result (Scaled-Size Rhodopsin Protei)**

Higher is better

InfiniBand DDR
Power Cost Savings with Different Interconnect

- Dell economical integration of AMD CPUs and Mellanox InfiniBand saves up to $6000 in power
  - To achieve same number of application jobs enabled with Gigabit Ethernet
  - Yearly based for 24-node cluster
- As cluster size increases, more power can be saved

\[
\frac{\text{$/KWh}}{\text{KWh}} = \text{KWh} \times 0.20
\]

LAMMPS Benchmark Results Summary

- **Interconnect comparison shows**
  - InfiniBand delivers superior performance in every cluster size
  - Performance advantage extends as cluster size increases

- **Job placement**
  - Some dataset shows productivity gain when running current jobs
  - In general LAMMPS can maximized the compute resource utilization

- **Power saving**
  - InfiniBand enables up to $6000/year power savings versus GigE

- **Dell™ PowerEdge™ server blades provides**
  - Linear scalability (maximum scalability) and balanced system
    - By integrating InfiniBand interconnect and AMD processors
  - Maximum return on investment through efficiency and utilization
LAMMPS Profiling – MPI Functions

- Mostly used MPI functions
  - MPI_send, MPI_Irecv, MPI_Wait, and MPI_Allreduce

![Graph showing LAMMPS Profiling Result](image)
LAMMPS Profiling – Timing

- MPI_Send/Wait and MPI_Allreduce show the highest communication overhead

LAMMPS Profiling Result
(Scaled-Size EAM Metallic Solid)
LAMMPS Profiling – Message Transferred

- Most MPI messages are large messages
- Number of messages increases with cluster size

LAMMPS Profiling Result
(Scaled-Size EAM Metallic Solid)

Message Size (Bytes)

Number of Messages

2 Nodes
4 Nodes
8 Nodes
16 Nodes
24 Nodes
LAMMPS Profiling – Message Transferred

- Most data related MPI messages are within 16KB-256KB
- Total data transferred increases with cluster size
LAMMPS Profiling Summary

- LAMMPS were profiled to identify its communication patterns
- Frequent used message sizes
  - 16KB-256KB messages for data related communications
  - <64B for synchronizations
  - Number of messages increases with cluster size
- Interconnects effect to LAMMPS performance
  - Both interconnect latency (MPI_Allreduce) and throughput (MPI_Send/Recv) highly influence Mechanical performance
- Balanced system – CPU, memory, Interconnect that match each other capabilities, is essential for providing application efficiency
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.