HPC ADVISORY COUNCIL

LAMMPSCUDA GPU Performance

April 2011

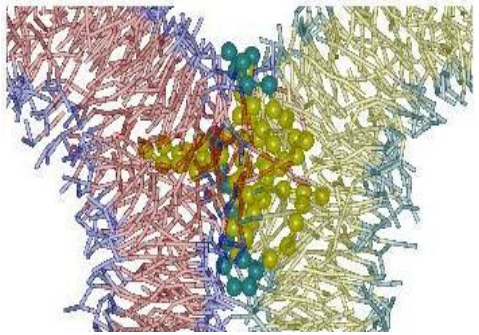
DELL intel Mellanox TECHNOLOGIES

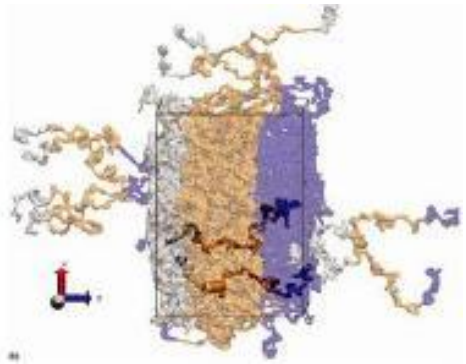
Note

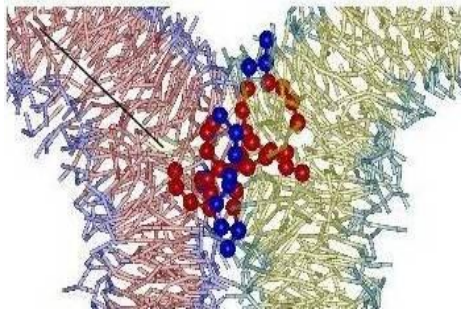
- **The following research was performed under the HPC Advisory Council activities**
  - Participating vendors: Dell, Intel, Mellanox
  - Compute resource - HPC Advisory Council Cluster Center
- **For more info please refer to**
  - http://www.dell.com
  - http://www.intel.com
  - http://www.mellanox.com
  - http://www.nvidia.com
  - http://code.google.com/p/LAMMPS
  - http://www.tu-ilmenau.de/theophys2/forschung/lammpscuda/

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 presented research was done to provide best practices**
  - LAMMPS performance benchmarking
  - Interconnect performance comparisons
  - Ways to increase LAMMPS productivity
  - GPU acceleration evaluation (GPUDirect)

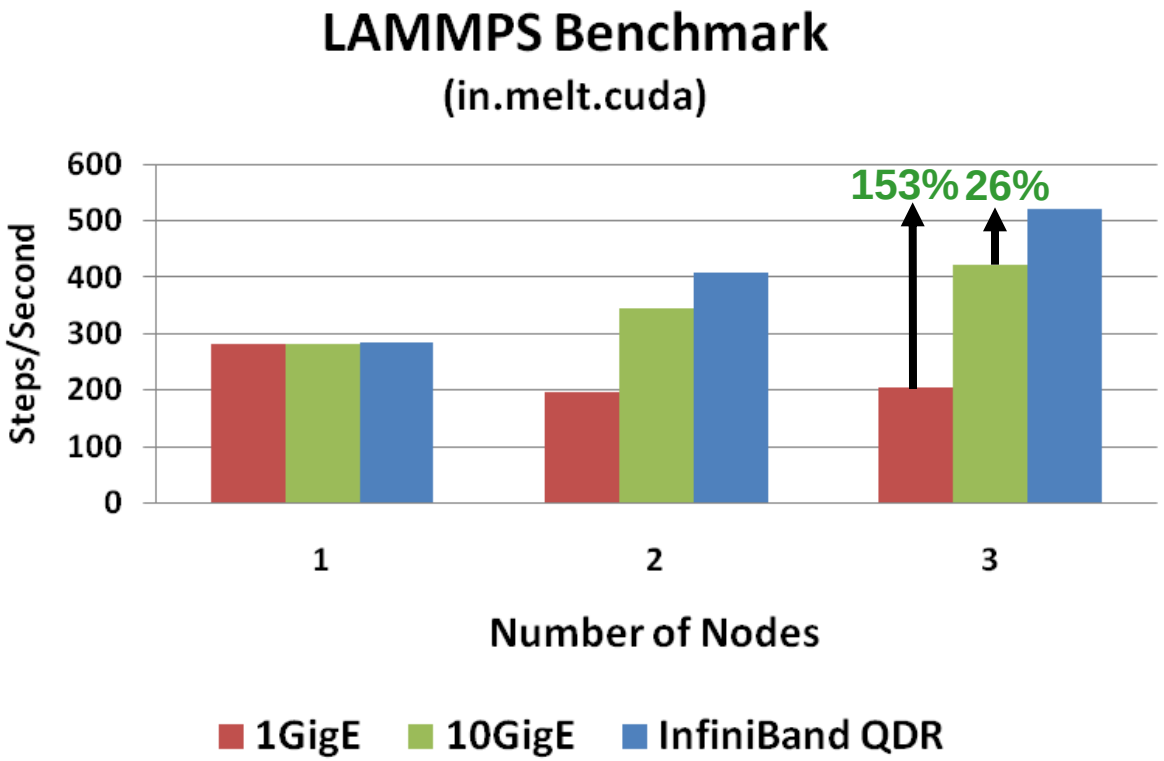
- **The presented results will demonstrate**
  - The scalability of the compute environment / application
  - Considerations for performance optimization

Test Cluster Configuration

- **Dell™ PowerEdge™ C6100 4-node cluster**
  - Six-Core Intel X5670 @ 2.93 GHz CPUs, 24GB memory per node
  - OS: RHEL5 Update 5, OFED 1.5.2 via MLNX_OFED-GPU_DIRECT: 2.1.0-1.1.0010
- **Dell™ PowerEdge™ C410x PCIe Expansion Chassis**
  - 4 NVIDIA® Tesla ™ C2050 “Fermi” GPUs (CUDA driver and runtime version 3.2)
- **Mellanox ConnectX-2 VPI Mezzanine cards for InfiniBand &10GigE**
- **Mellanox MTS3600Q 36-Port 40Gb/s QDR InfiniBand switch**
- **Mellanox – NVIDIA GPUDirect: MLNX_OFED-GPU_DIRECT 2.1.0-1.1.0010**
- **Fulcrum based 10Gb/s Ethernet switch**
- **MPI: Open MPI 1.4.2**
- **Application: gpulammps (subversion rev 627), based on LAMMPS (4 Dec 2010)**
  - LAMMPSCUDA (Version 1.1: 7th of March 2011)

LAMMPSCUDA Performance – Interconnect

- **InfiniBand provides the best data throughput**
  - Up to 153% higher performance than 1GigE at 3-node
  - Up to 26% higher performance than 10GigE at 3-node
- **1GigE provides no job gain after 1 node**
  - GPU communications between nodes become a significantly burden to the network

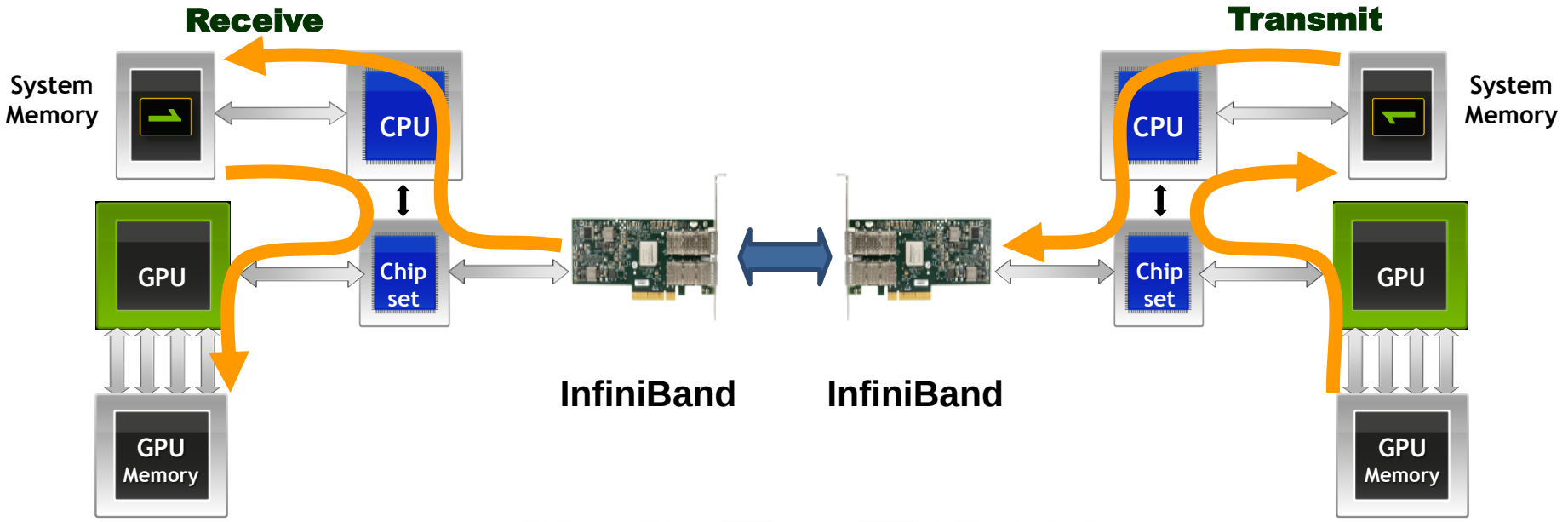


Higher is better

12 Cores/Node

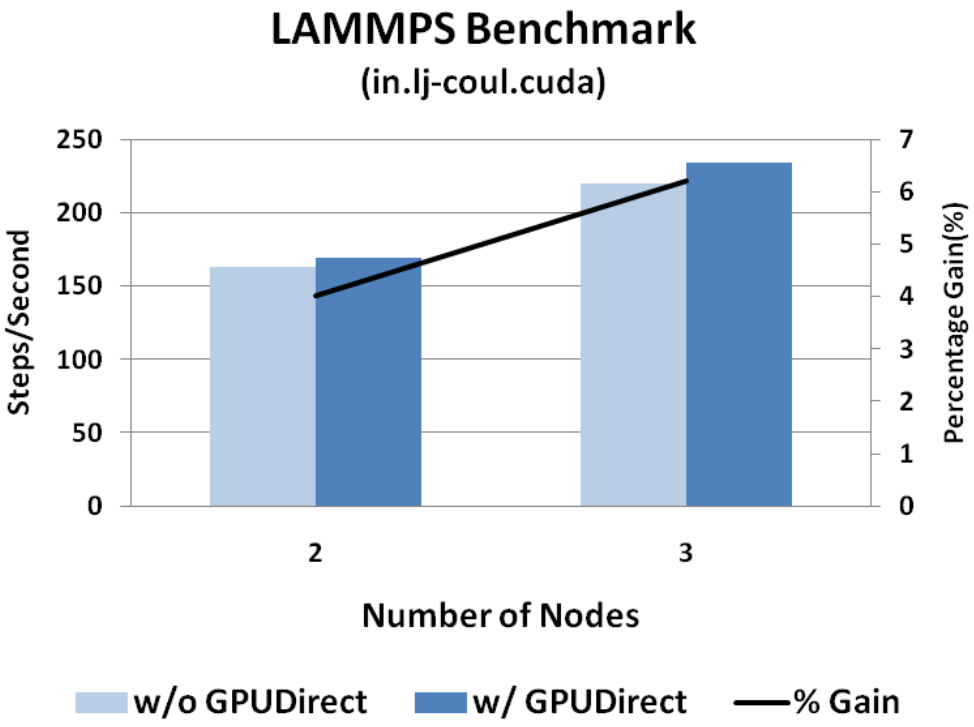
LAMMPSCUDA – GPUDirect

- **Benefit of GPUDirect**
  - Allows GPU to directly access the system memory w/o CPU involvement
  - Utilizes native RDMA for efficient data transfer over InfiniBand
  - Reduce latency by 30% for GPU communications
- **Open MPI – needs to set MCA parameter to enable GPUDirect**
  - --mca btl_openib_flags 310
- **LAMMPSCUDA input data file needs to have pinned memory enabled**
  - "accelerator cuda gpu/node 1 **pinned 1**"



LAMMPSCUDA – in.lj-coul.cuda (1GPU/node)

- **Dataset: in.lj-coul.cuda**
  - GI-System, Debye length 2.5, 7.0A cutoff
  - 64584 atoms, 2000 steps, charge atom style, pair force: lj/cut/coul/debye 2.5 7.0
- **GPUDirect demonstrates an increase in job productivity**
  - Up to 4-6% increase in performance with on 2 and 3 nodes



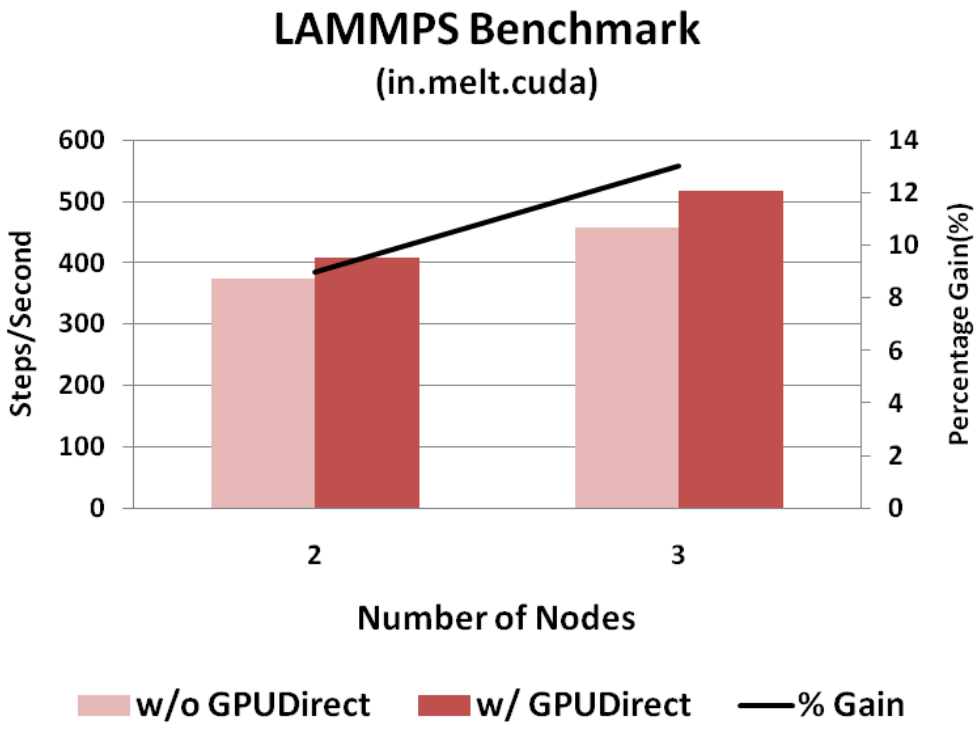
Higher is better

InfiniBand QDR

12 Cores/Node

LAMMPSCUDA – in.melt.cuda (1GPU/node)

- **Dataset: in.melt.cuda**
  - 3d Lennard-Jones melt, 2.5A cutoff
  - 108000 atoms, 2000 steps, atomic atom style, pair force: lj/cut 2.5
- **GPUDirect demonstrates an increase in job productivity**
  - Up to 9-13% increase in performance with GPUDirect on 2 and 3 nodes



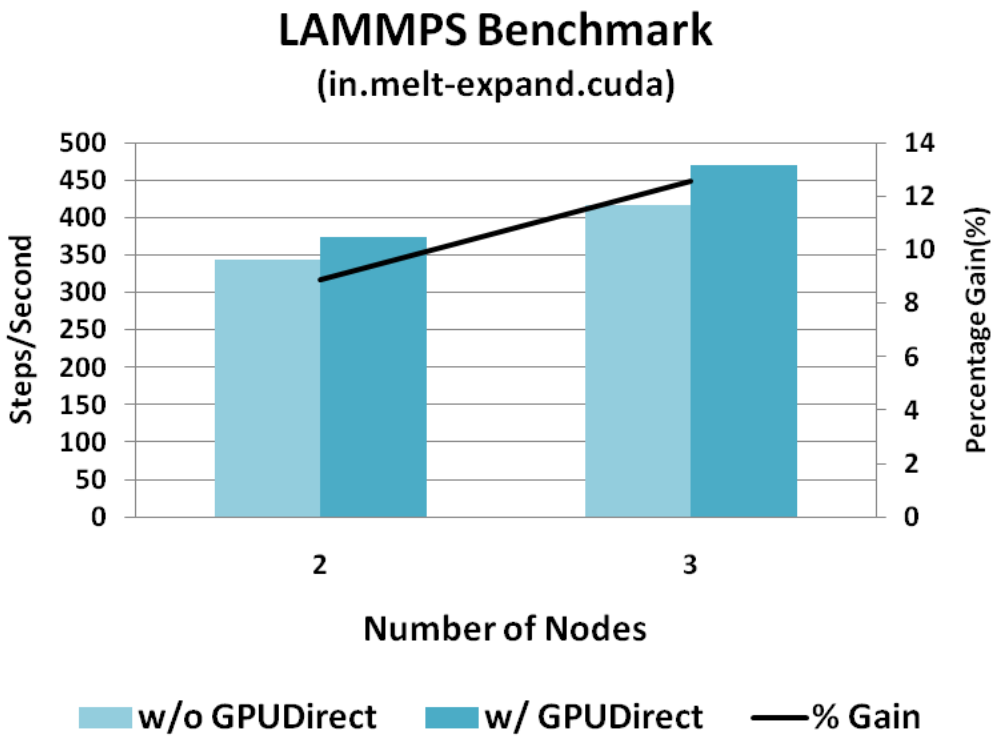
Higher is better

InfiniBand QDR

12 Cores/Node

LAMMPSCUDA – in.melt-expand.cuda (1GPU/node)

- **Dataset: in.melt-expand.cuda**
  - 3d Lennard-Jones melt, 2.5A cutoff
  - 108000 atoms, 2000 steps, atomic atom style, pair force: lj/expand 2.5
- **GPUDirect demonstrates an increase in job productivity**
  - Up to 9-13% increase in performance with GPUDirect on 2 and 3 nodes



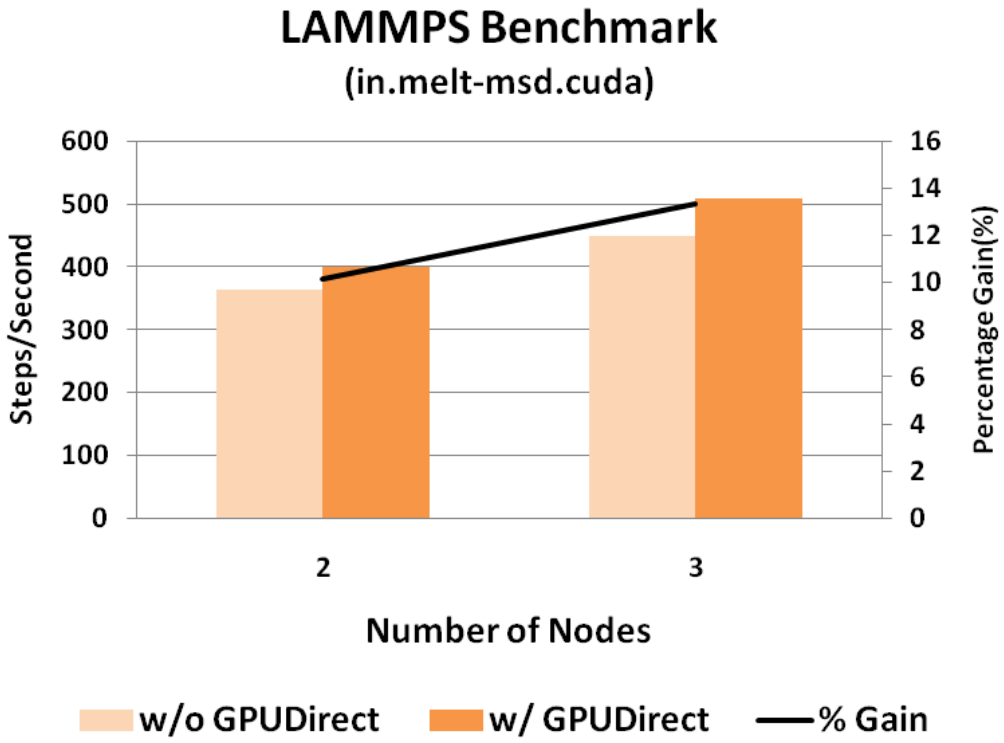
Higher is better

InfiniBand QDR

12 Cores/Node

LAMMPSCUDA – in.melt-msd.cuda (1GPU/node)

- **Dataset: in.melt-msd.cuda**
  - 3d Lennard-Jones melt (Mean Square Displacement), 2.5A cutoff
  - 108000 atoms, 2000 steps, atomic atom style, pair force: lj/cut 2.5
- **GPUDirect demonstrates an increase in job productivity**
  - Up to 10-13% increase in performance with GPUDirect on 2 and 3 nodes



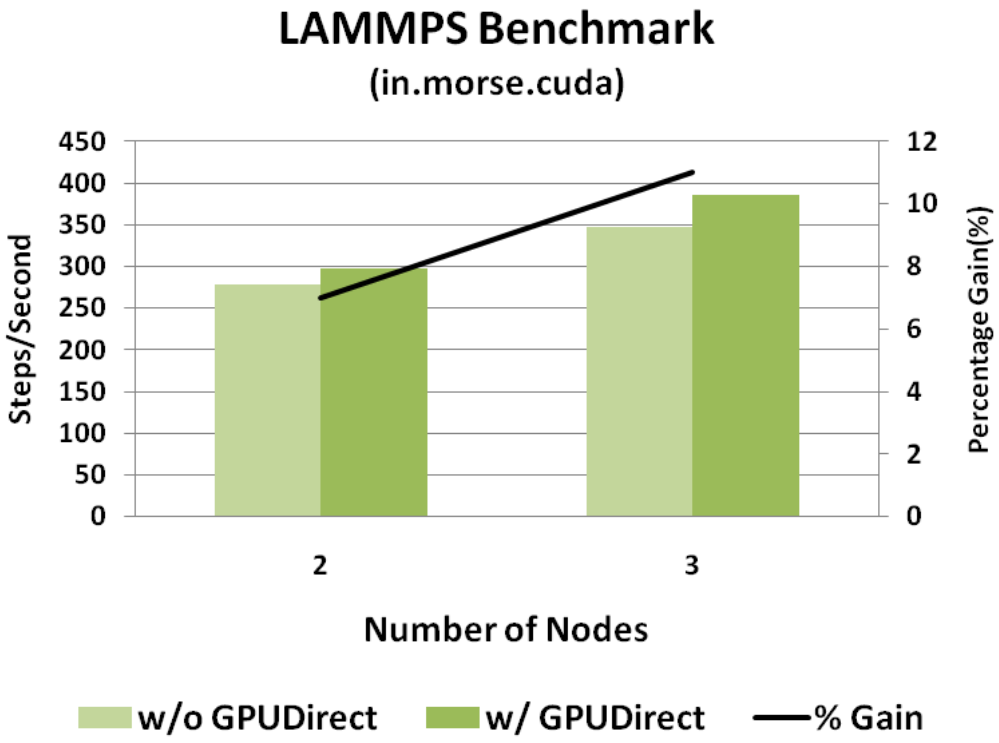
InfiniBand QDR

12 Cores/Node

Higher is better

LAMMPSCUDA – in.melt-morse.cuda(1GPU/node)

- **Dataset: in.morse.cuda**
  - Morse interaction, 3.0A cutoff
  - 108000 atoms, 2000 steps, atomic atom style, pair force: morse 3.0, box of 30x30x30
- **GPUDirect demonstrates an increase in job productivity**
  - Up to 7-11% increase in performance with GPUDirect on 2 and 3 nodes



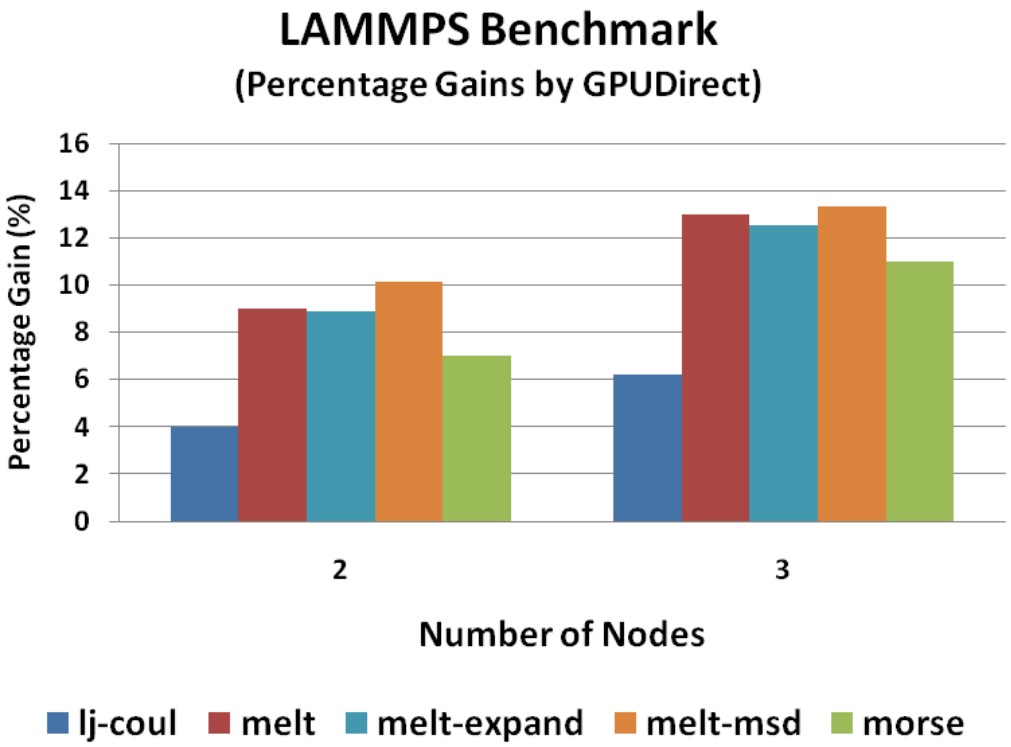
Higher is better

InfiniBand QDR

12 Cores/Node

LAMMPSCUDA – GPUDirect (1GPU/node)

- **GPUDirect enables faster speedup with 1 GPU per node on multiple nodes**
  - Speed up of 4% to 10% for a 2-node job
  - Speed up of 6% to 13% for a 3-node job
- **The productivity gained will increase as the cluster size increases**



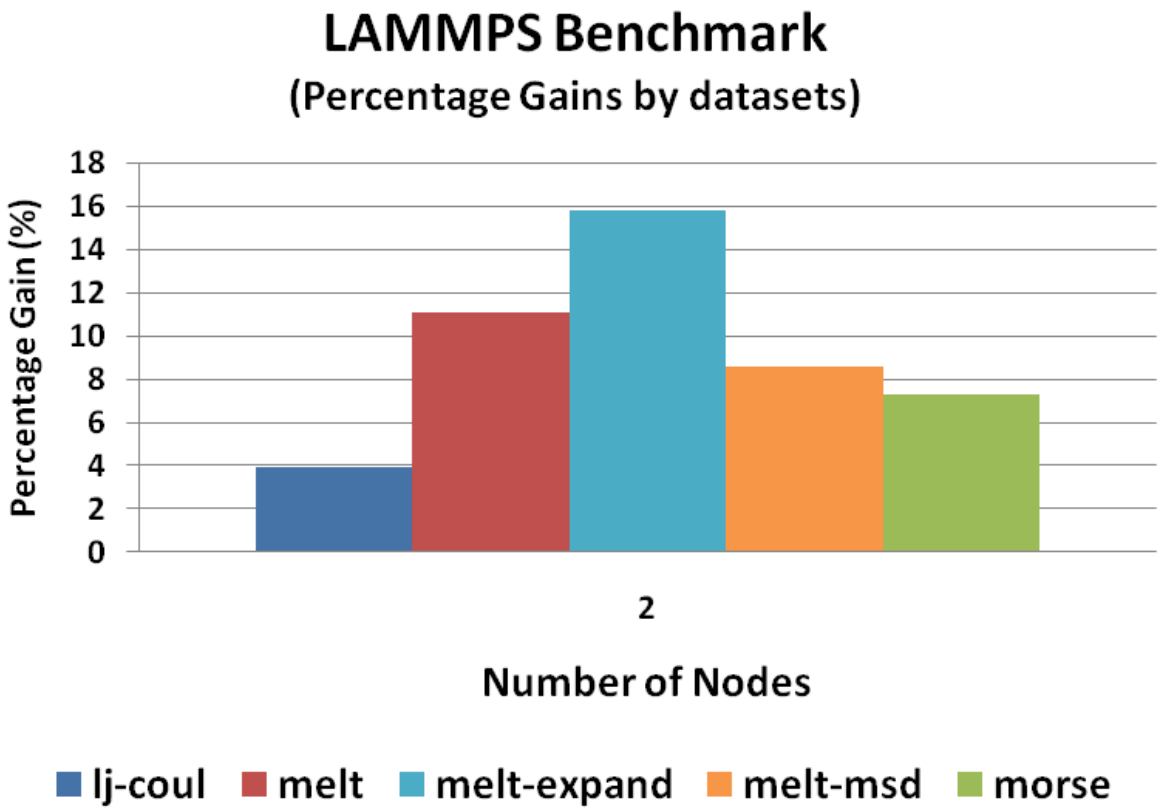
Higher is better

InfiniBand QDR

Open MPI

LAMMPSCUDA – GPUDirect (2GPUs/node)

- **GPUDirect enables faster speedup with 2 GPUs per node on 2 nodes**
  - Speed up of 16% for a 2-node job with the melt-expand dataset
- **As the cluster scales, more speed up will be seen**



Higher is better

InfiniBand QDR

Open MPI

Summary

- **LAMMPS performance with GPUDirect**
  - GPUDirect enables a substantial performance speedup
  - Results demonstrated on small scale enivornment
  - Bigger performance advantage is expected at larger system size
  - Performance boost depends on the benchmark – dataset

- **Interconnects effect to LAMMPS performance**
  - InfiniBand enables the highest performance and best GPU cluster scalability
  - Mainly due to low CPU overhead, native RDMA and low latency

Thank You

HPC Advisory Council