LAMMPS-KOKKOS
Performance Benchmark and Profiling

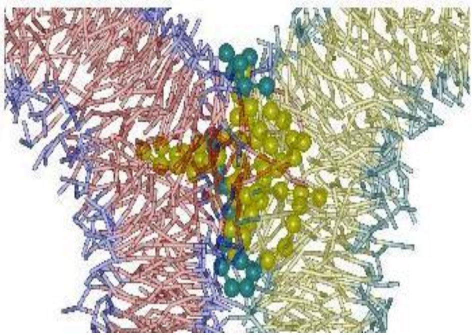
September 2015

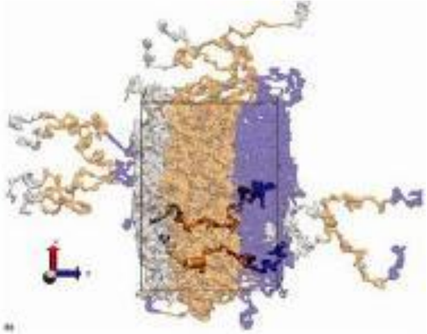
Note

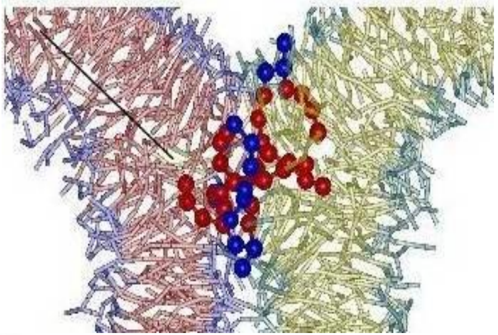
- **The following research was performed under the HPC Advisory Council activities**
  - Participating vendors: Intel, Dell, Mellanox, NVIDIA
  - Compute resource - HPC Advisory Council Cluster Center
- **The following was done to provide best practices**
  - LAMMPS performance overview
  - Understanding LAMMPS communication patterns
  - Ways to increase LAMMPS productivity
- **For more info please refer to**
  - http://lammps.sandia.gov
  - http://www.dell.com
  - http://www.intel.com
  - http://www.mellanox.com
  - http://www.nvidia.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-KOKKOS package contains**
  - Versions of pair, fix, and atom styles that use data structures and macros provided by the Kokkos library
- **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
    - MPI Library performance comparison
    - Interconnect performance comparison
    - CPUs comparison
    - Optimization tuning
- **The presented results will demonstrate**
  - The scalability of the compute environment/application
  - Considerations for higher productivity and efficiency

Test Cluster Configuration

- **Dell PowerEdge R730 32-node (896-core) “Thor” cluster**
  - Dual-Socket 14-Core Intel E5-2697v3 @ 2.60 GHz CPUs (BIOS: Maximum Performance, Home Snoop)
  - Memory: 64GB memory, DDR4 2133 MHz, Memory Snoop Mode in BIOS sets to Home Snoop
  - OS: RHEL 6.5, MLNX_OFED_LINUX-3.0-1.0.1 InfiniBand SW stack
  - Hard Drives: 2x 1TB 7.2 RPM SATA 2.5” on RAID 1
- **Mellanox ConnectX-4 EDR 100Gb/s InfiniBand Adapters**
- **Mellanox Switch-IB SB7700 36-port EDR 100Gb/s InfiniBand Switch**
- **Mellanox ConnectX-3 FDR VPI InfiniBand and 40Gb/s Ethernet Adapters**
- **Mellanox SwitchX-2 SX6036 36-port 56Gb/s FDR InfiniBand / VPI Ethernet Switch**
- **Dell InfiniBand-Based Lustre Storage based on Dell PowerVault MD3460 and Dell PowerVault MD3420**
- **NVIDIA Tesla K40 (on 8 Nodes) and NVIDIA Tesla K80 GPUs (on 2 Nodes); 1 GPU per node**
- **MPI: Mellanox HPC-X v1.3 (based on Open MPI 1.8.7) with CUDA 6.5 and 7.0 support**
- **Application: LAMMPS 15May15**
- **Benchmarks: Input data with embedded-atom method (in.eam)**

PowerEdge R730
Massive flexibility for data intensive operations

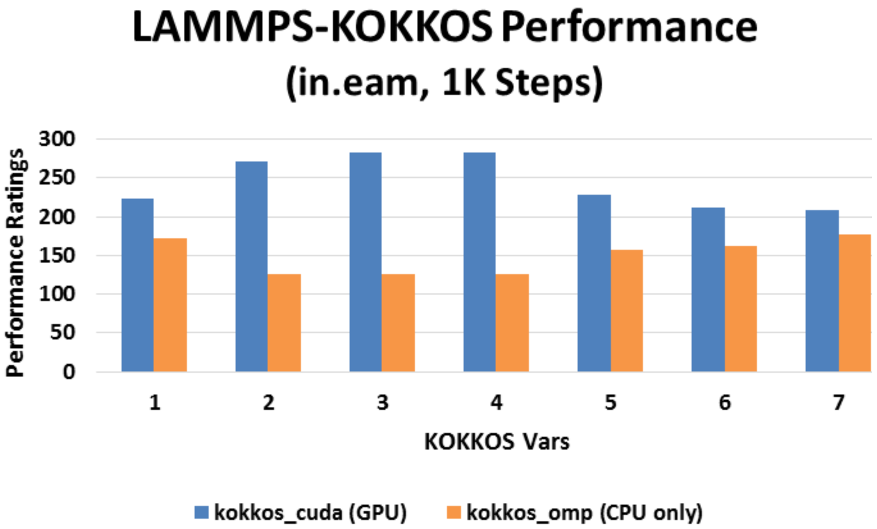
- **Performance and efficiency**
  - Intelligent hardware-driven systems management with extensive power management features
  - Innovative tools including automation for parts replacement and lifecycle manageability
  - Broad choice of networking technologies from Ethernet to InfiniBand
  - Built in redundancy with hot plug and swappable PSU, HDDs and fans

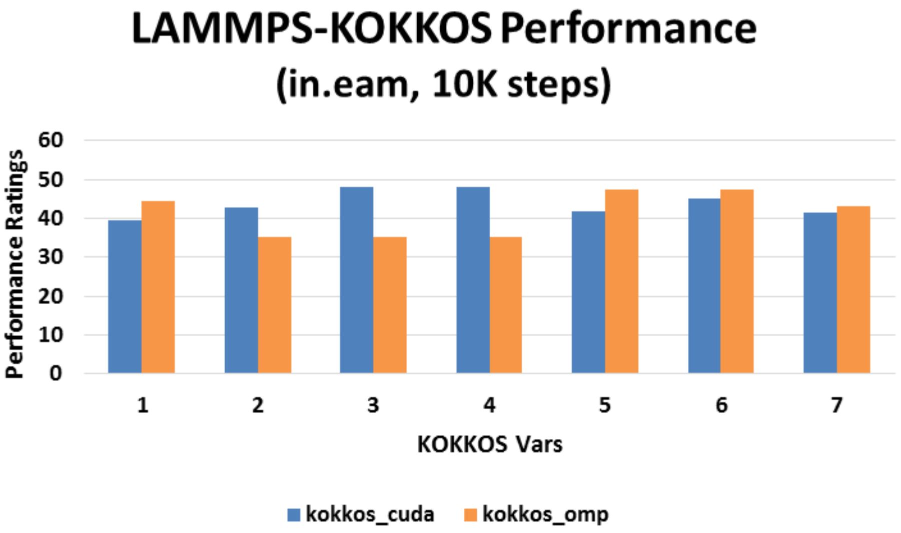
- **Benefits**
  - Designed for performance workloads
  - High performance scale-out compute and low cost dense storage in one package

- **Hardware Capabilities**
  - Flexible compute platform with dense storage capacity
    - 2S/2U server, 6 PCIe slots
  - Large memory footprint (Up to 768GB / 24 DIMMs)
  - High I/O performance and optional storage configurations
    - HDD options: 12 x 3.5” - or - 24 x 2.5 + 2x 2.5 HDDs in rear of server
    - Up to 26 HDDs with 2 hot plug drives in rear of server for boot or scratch

LAMMPS Performance – KOKKOS Vars

- **Kokkos variable determines the communication models between host and device**
  - The best of kokkos vars used for CPU and GPU tests are different
  - The most favorite kokkos vars for GPU appears to be among #2, 3, or 4
  - The most favorite kokkos vars for CPU appears to be among #1, 5, 6, or 7

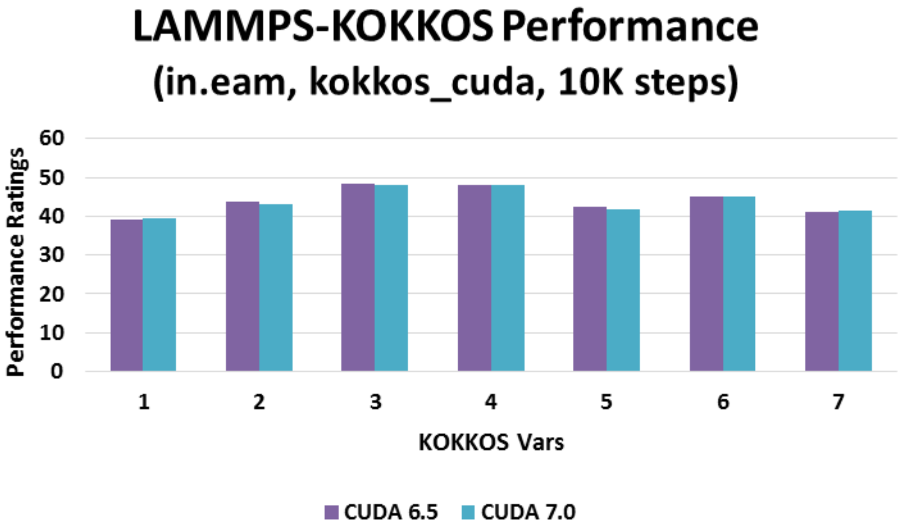




Higher is better

LAMMPS Performance – CUDA Versions

- **Both CUDA 6.5 and 7.0 versions perform similarly**
  - Using the given input data and workload

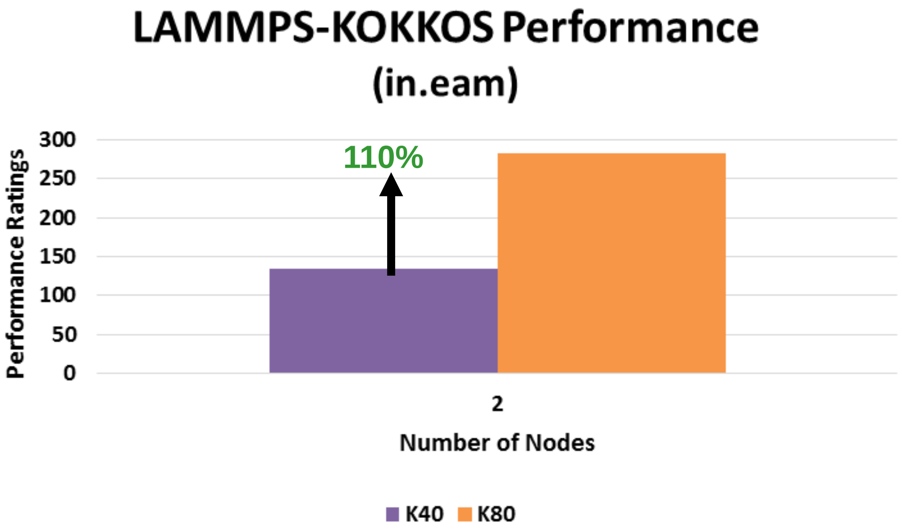


Higher is better

8 Nodes; 1x K40 / Node

LAMMPS Performance – Tesla K40 vs K80

- **Tesla K80 doubles the performance of K40 using with LAMMPS**
  - Demonstrates 110% increase in performance on 2 nodes with 1 GPU per node

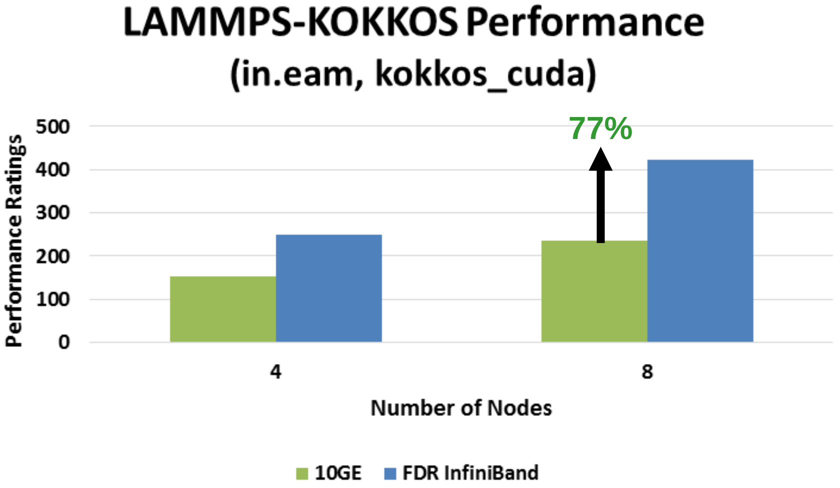


Higher is better

1 GPU / Node

LAMMPS Performance – Network Interconnects

- **EDR InfiniBand delivers superior scalability in application performance**
  - InfiniBand delivers 77% higher performance than 10GbE on 8 nodes
- **Performance of 4 IB nodes outperforms 8 Ethernet (10GbE) nodes**
  - Benefits of InfiniBand over Ethernet expect to increase as cluster scales
  - Scalability for Ethernet stops beyond 4 nodes; while InfiniBand continue to scale

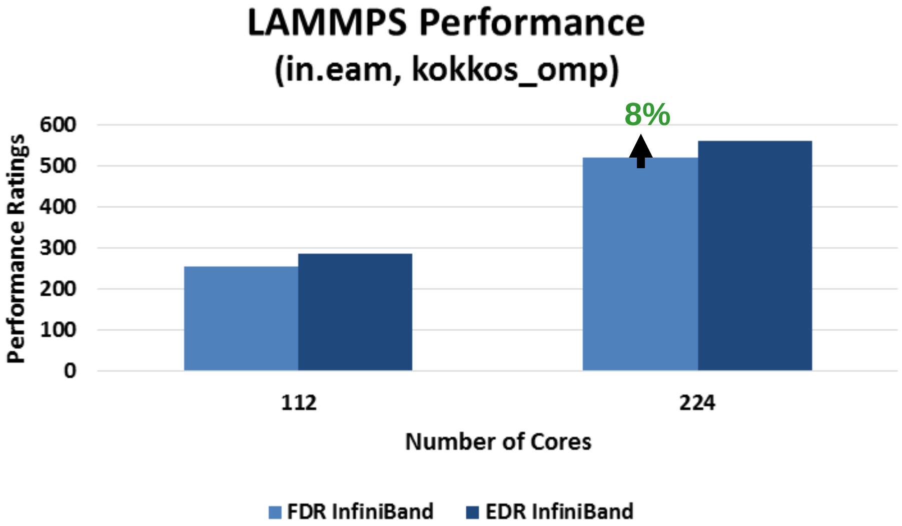


Higher is better

GPU: 1 K40 / Node

LAMMPS Performance – Network Interconnects

- **EDR InfiniBand delivers superior scalability in application performance**
  - EDR IB demonstrates an 8% increase on 8 Nodes / 224 Cores
  - Performance gap between FDR and EDR expect to increase as cluster scales

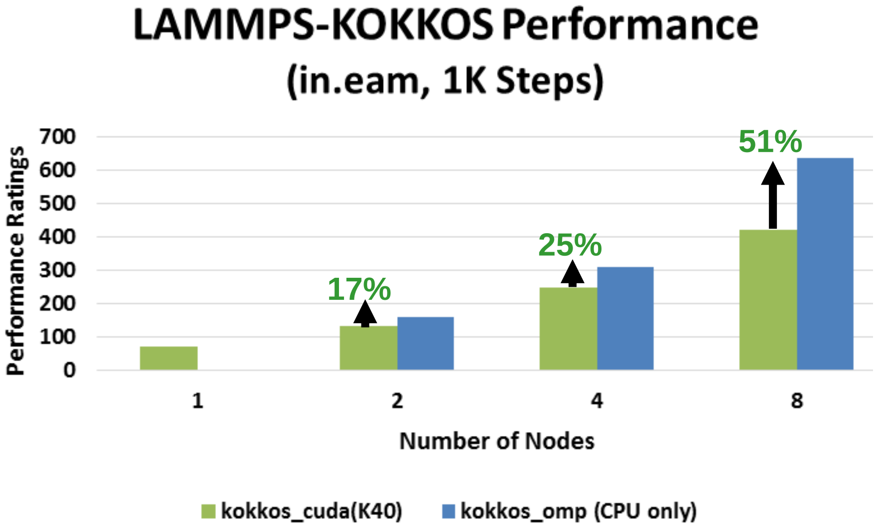


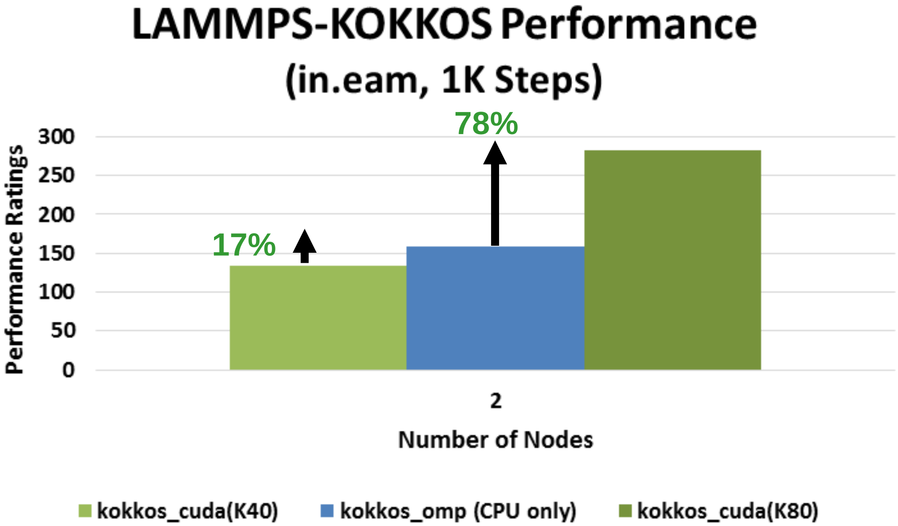
Higher is better

28 MPI Processes / Node

LAMMPS Performance – Scalability

- **kokkos_omp demonstrates higher speed up and performance than kokkos_cuda**
  - CPU performance outpaces Tesla K40 performance using 28 cores/node
  - With the availability of the Tesla K80, it outperforms CPU; performance gap expect to grow





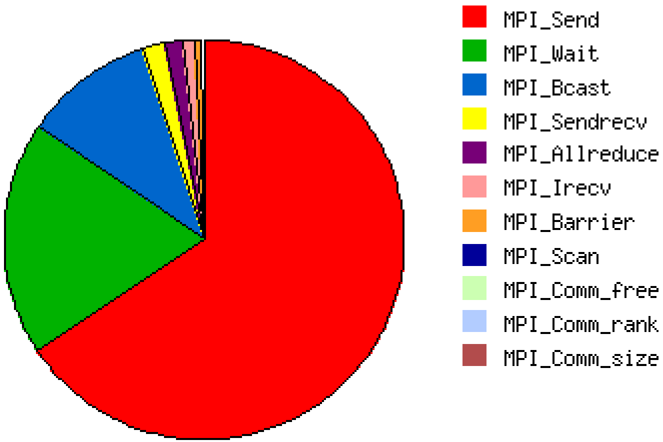
CPU: 28 Processes/Node
GPU: 1 GPU / Node

Higher is better

LAMMPS Profiling – % of MPI Calls

- **The most time consuming MPI calls for LAMMPS-KOKKOS (cuda):**
  - MPI_Send: 67% MPI / 8% Wall
  - MPI_Wait: 18% MPI / 2% Wall
  - MPI_Bcast: 11% MPI / 1% Wall

Kokkos_cuda, 8 nodes / 224 processes

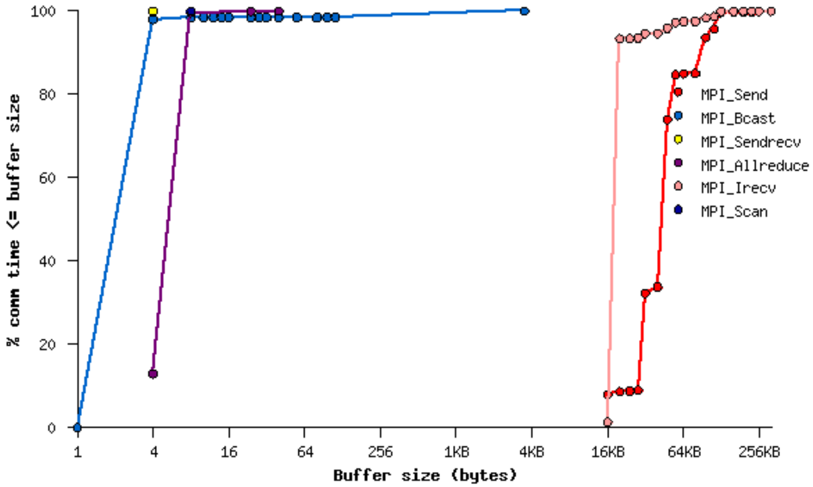


EDR InfiniBand

LAMMPS Profiling – MPI Message Size Distribution

- **For the most time consuming MPI calls**
  - MPI_Send: 48KB (26% MPI) time, 32KB (15%), 56KB (7%)
  - MPI_Wait: 0B (18% MPI time)
  - MPI_Bcast: 4B (10% MPI time)

Kokkos_cuda, 8 nodes / 224 processes

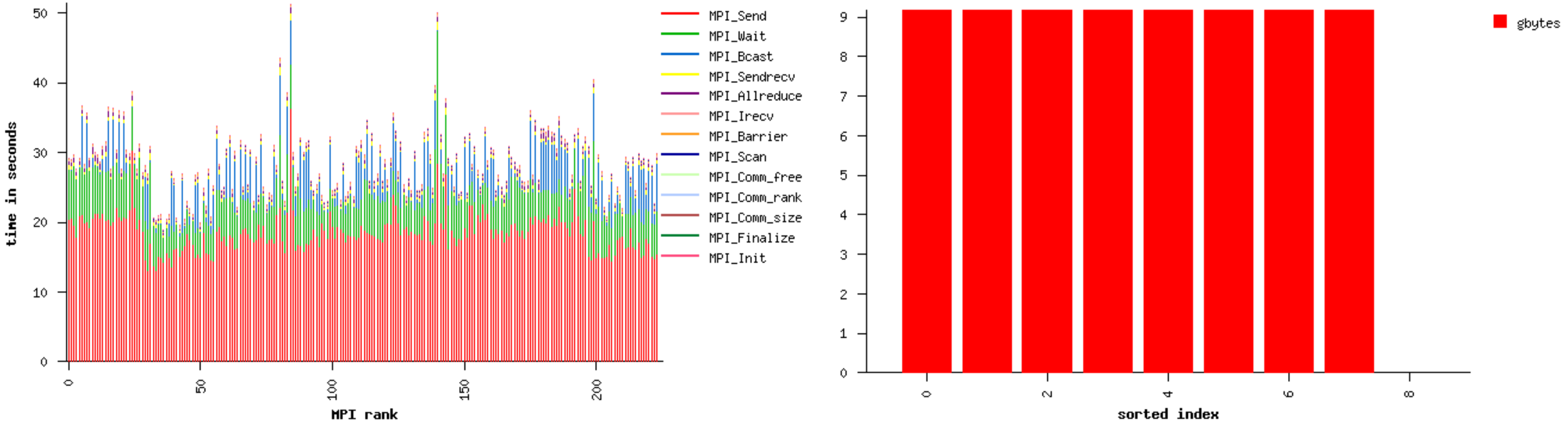


EDR InfiniBand

LAMMPS Profiling – MPI Memory Consumption

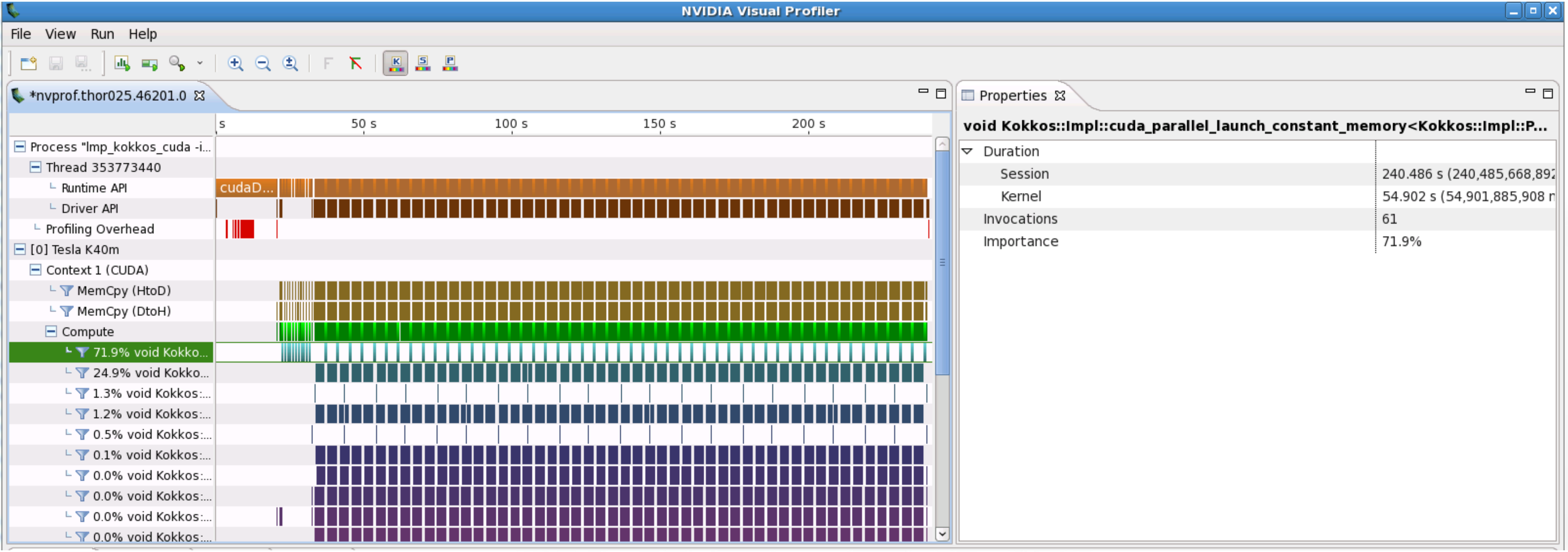
- **Some load imbalance is seen on the processes**
- **Memory consumption:**
  - About 9GB of memory is used on each compute node for this input data

Kokkos_cuda, 8 nodes / 224 processes



LAMMPS Profiling – CUDA Profiler

- **NVIDIA Visual Profiler and nvprof: Profilers for GPUs**
  – Shows many Memcpy occurs between host and device throughout the run



void Kokkos::Impl::cuda_parallel_launch_constant_memory<Kokkos::Impl::P...

Duration	
Session	240.486 s (240,485,668,892
Kernel	54.902 s (54,901,885,908 n
Invocations	61
Importance	71.9%

LAMMPS Summary

- **Performance**
  - Compute: cluster of the current generation outperforms system architecture of previous generations
    - Tesla K80 doubles the performance of K40 using with LAMMPS-KOKKOS
    - The KOKKOS vars can be a significant performance implication to the performance of LAMMPS
    - CUDA versions 6.5 and 7.0 performs similarly
  - Network: EDR InfiniBand demonstrates superior scalability in LAMMPS performance
    - EDR IB provides higher performance by 77% over 10GbE, 86% higher over 1GbE on 8 nodes
    - Performance of 4 IB nodes outperforms 8 Ethernet (10GbE) nodes
    - Benefits of InfiniBand over Ethernet expect to increase as cluster scales
    - Scalability for Ethernet stops beyond 4 nodes; while InfiniBand continue to scale
  - MPI Profiles
    - The most time consuming MPI calls for LAMMPS-KOKKOS (cuda):
    - MPI_Send: 67% MPI / 8% Wall
    - MPI_Wait: 18% MPI / 2% Wall
    - MPI_Bcast: 11% MPI / 1% Wall

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 undertakes no duty and assumes no obligation to update or correct any information presented herein