Performance of Applications on Comet GPU Nodes Utilizing MVAPICH2-GDR

Mahidhar Tatineni

MVAPICH User Group Meeting

August 16, 2017

UC San Diego

SDSC

SAN DIEGO SUPERCOMPUTER CENTER

INDIANA UNIVERSITY

This work supported by the National Science Foundation, award ACI-1341698.

Comet: System Characteristics

- **Total peak flops ~2.1 PF**
- **Dell primary integrator**
  - Intel Haswell processors w/ AVX2
  - Mellanox FDR InfiniBand
- **1,944 standard compute nodes (46,656 cores)**
  - Dual CPUs, each 12-core, 2.5 GHz
  - 128 GB DDR4 2133 MHz DRAM
  - 2*160GB GB SSDs (local disk)
- **72 GPU nodes**
  - 36 nodes same as standard nodes *plus* Two NVIDIA K80 cards, each with dual Kepler3 GPUs
  - 36 nodes with 2 14-core Intel Broadwell CPUs plus 4 NVIDIA P100 GPUs
- **4 large-memory nodes**
  - 1.5 TB DDR4 1866 MHz DRAM
  - Four Haswell processors/node
  - 64 cores/node
- **Hybrid fat-tree topology**
  - FDR (56 Gbps) InfiniBand
  - Rack-level (72 nodes, 1,728 cores) full bisection bandwidth
  - 4:1 oversubscription cross-rack
- **Performance Storage (Aeon)**
  - 7.6 PB, 200 GB/s; Lustre
  - Scratch & Persistent Storage segments
- **Durable Storage (Aeon)**
  - 6 PB, 100 GB/s; Lustre
  - Automatic backups of critical data
- **Home directory storage**
- **Gateway hosting nodes**
- **100 Gbps external connectivity to Internet2 & ESNet**

Flavors of MVAPICH2 on Comet

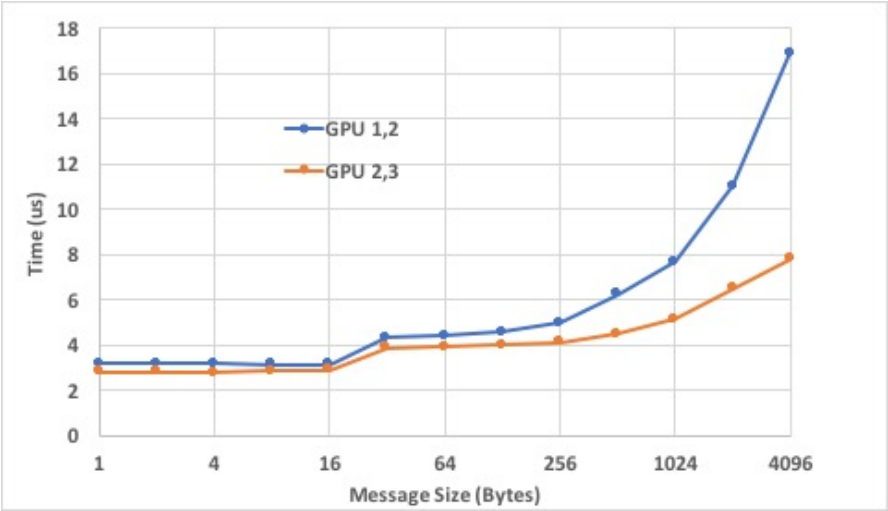
- **MVAPICH2 (v2.1) is the default MPI on Comet.**
- **MVAPICH2-X v2.2a** to provide unified high-performance runtime supporting both MPI and PGAS programming models.
- **MVAPICH2-GDR (v2.2) on the GPU nodes featuring NVIDIA K80s and P100s.**
- **RDMA-Hadoop (2x-1.1.0), RDMA-Spark (0.9.4) (from Dr. Panda's HiBD lab) also available.**

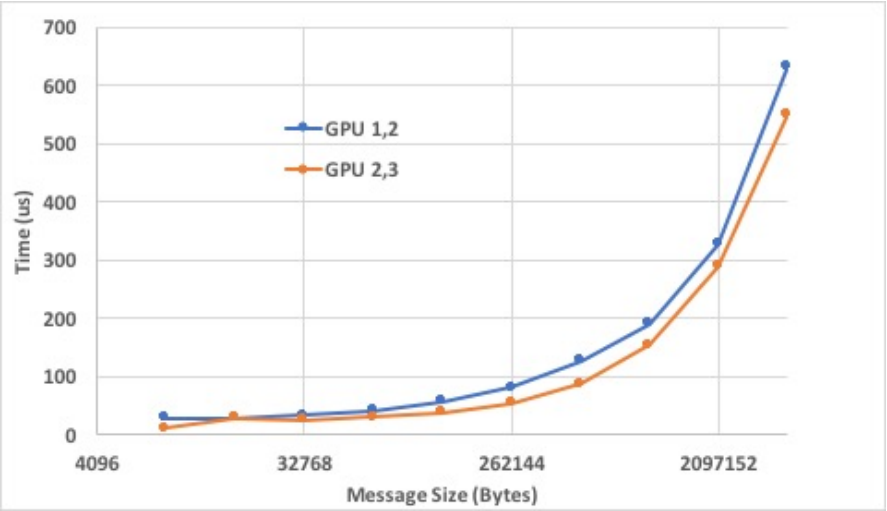
Comet K80 node architecture

```
        GPU0    GPU1    GPU2    GPU3    mlx4_0  CPU Affinity
GPU0     X      PIX     SOC     SOC     SOC     0-0,2-2,4-4,6-6,8-8,10-10,12-12,14-14,16-16,18-18,20-20,22-22
GPU1    PIX      X      SOC     SOC     SOC     0-0,2-2,4-4,6-6,8-8,10-10,12-12,14-14,16-16,18-18,20-20,22-22
GPU2    SOC     SOC      X      PIX     PHB     1-1,3-3,5-5,7-7,9-9,11-11,13-13,15-15,17-17,19-19,21-21,23-23
GPU3    SOC     SOC     PIX      X      PHB     1-1,3-3,5-5,7-7,9-9,11-11,13-13,15-15,17-17,19-19,21-21,23-23
mlx4_0  SOC     SOC     PHB     PHB      X

Legend:

  X    = Self
  SOC  = Connection traversing PCIe as well as the SMP link between CPU sockets(e.g. QPI)
  PHB  = Connection traversing PCIe as well as a PCIe Host Bridge (typically the CPU)
  PXB  = Connection traversing multiple PCIe switches (without traversing the PCIe Host Bridge)
  PIX  = Connection traversing a single PCIe switch
  NV#  = Connection traversing a bonded set of # NVLinks
```

- **4 GPUs per node**
- **GPUs (0,1) and (2,3) can do P2P communication**
- **Mellanox InfiniBand adapter associated with second socket (GPUs 2, 3)**

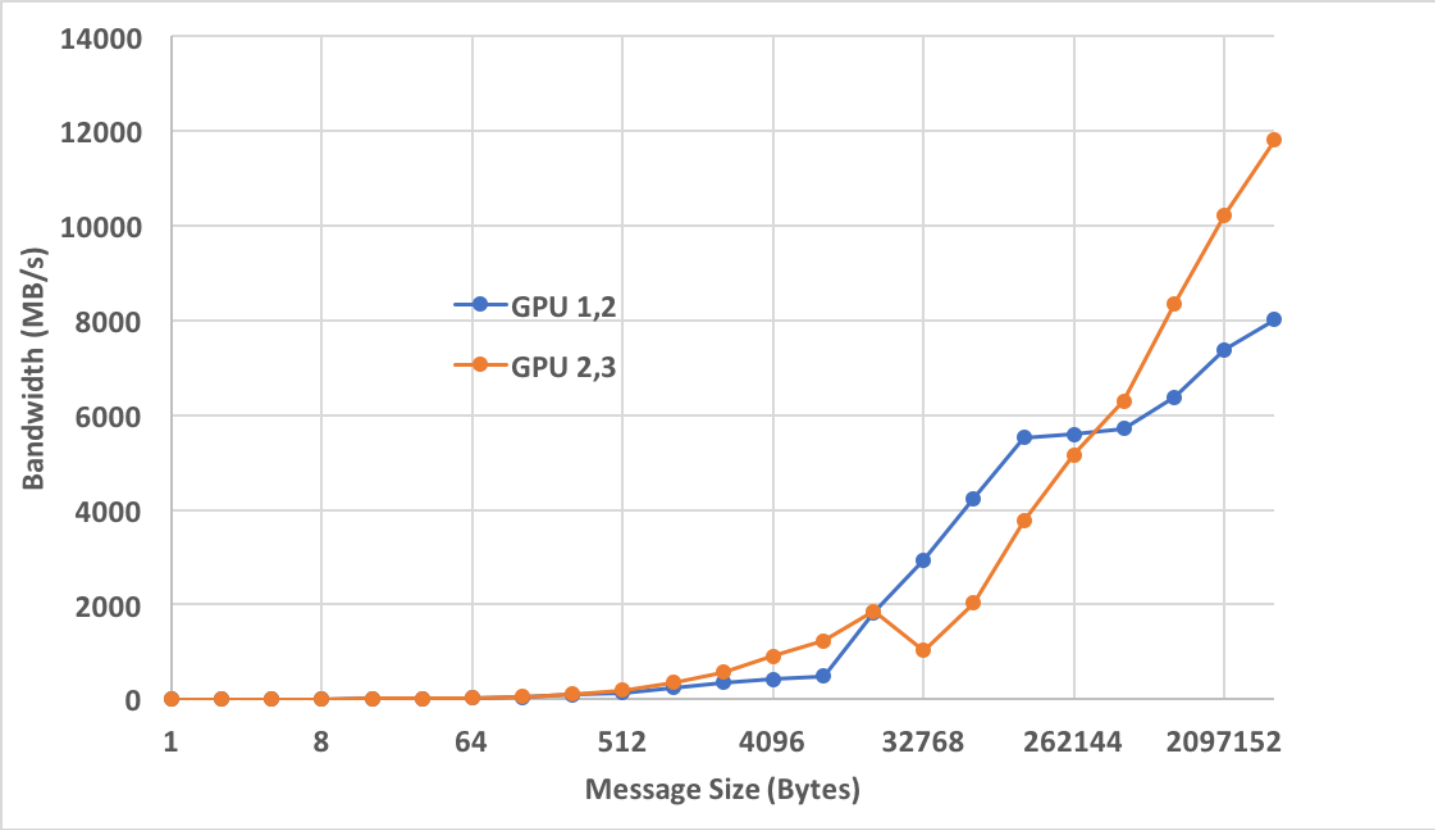
OSU Latency (osu_latency) Benchmark
Intra-node, K80 nodes



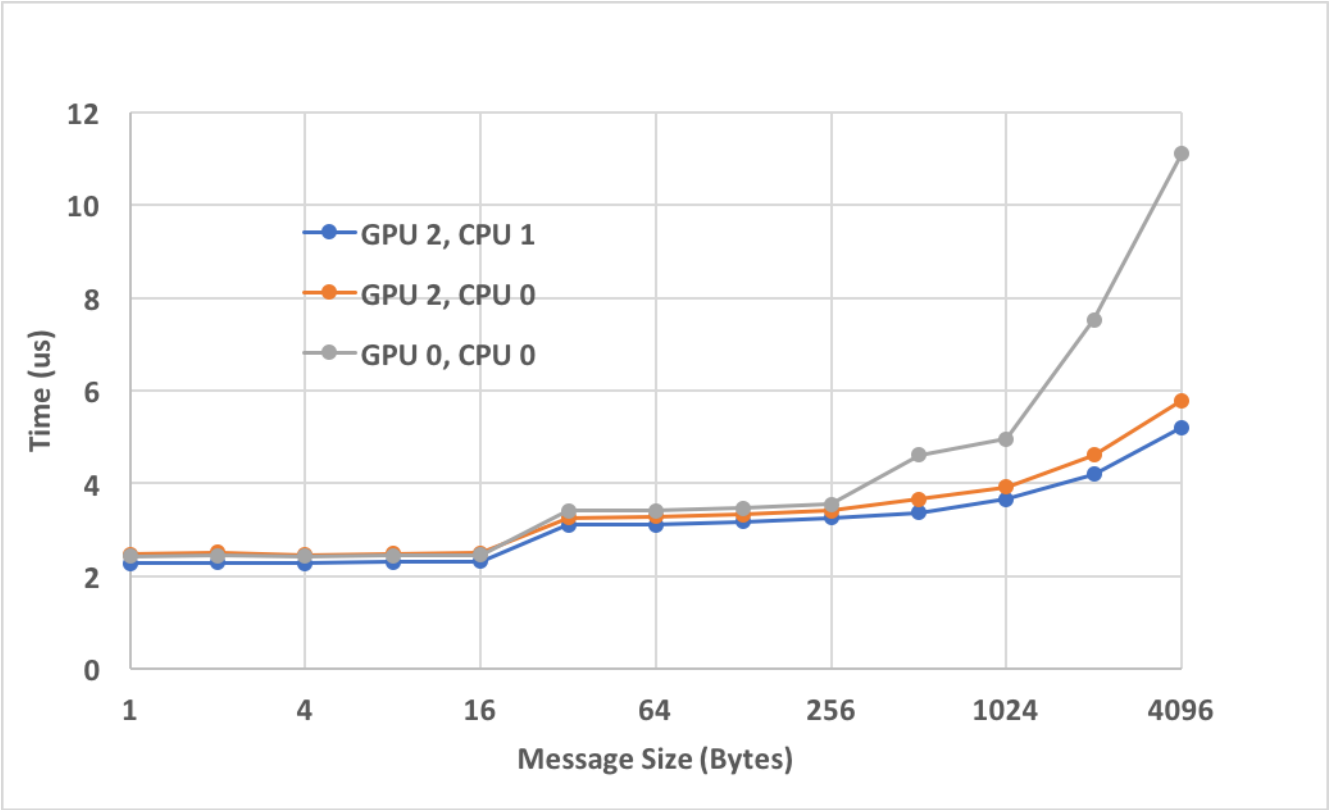


- **Latency between GPU 2 , GPU 3: 2.82 μs**
- **Latency between GPU 1 , GPU 2: 3.18 μs**

OSU Bandwidth (osu_bw) Benchmark Intra-node, K80 nodes

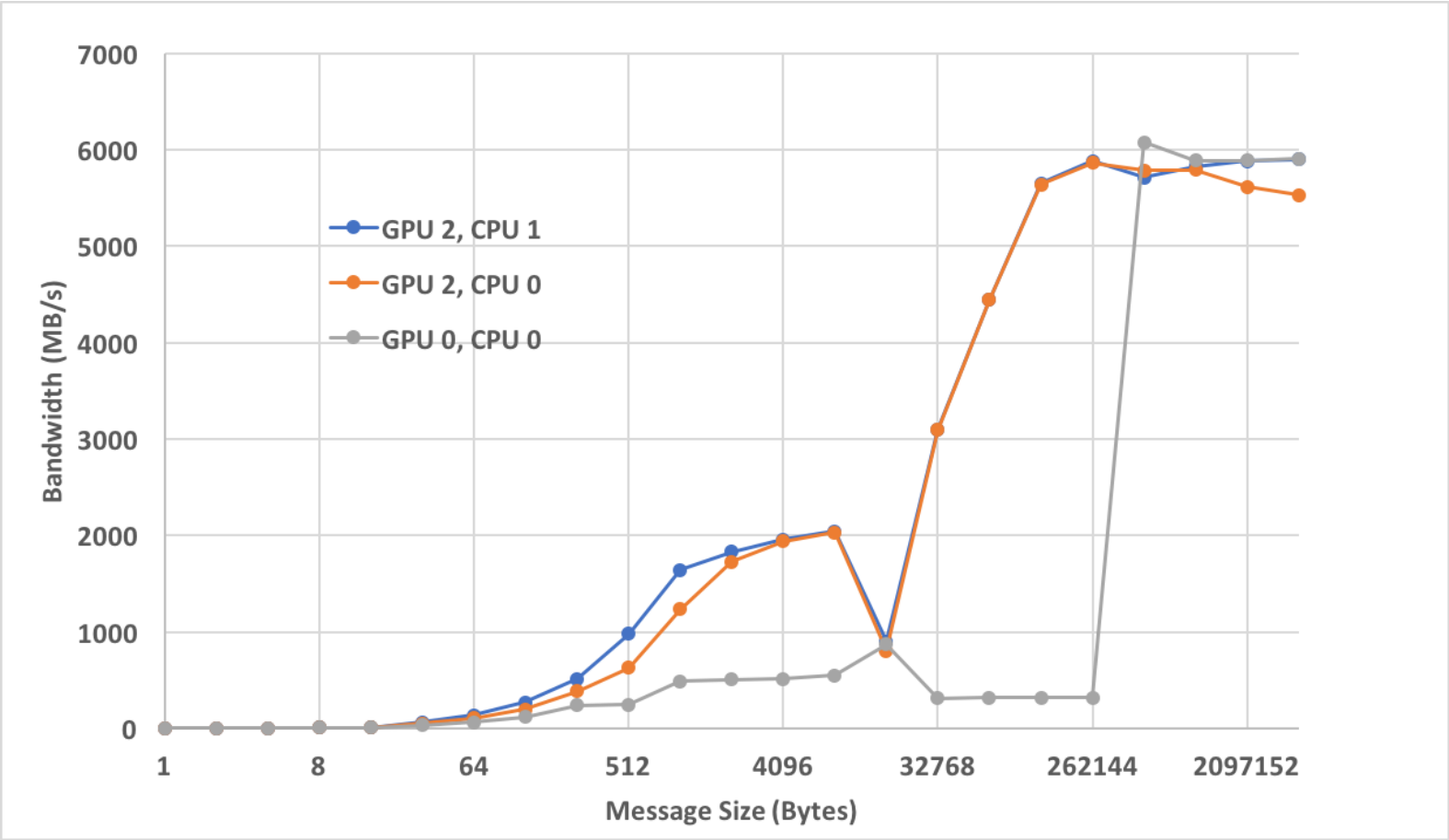


OSU Latency (osu_latency) Benchmark Inter-node, K80 nodes



- **Latency between GPU 2 , process bound to CPU 1 on both nodes: 2.27 *μ*s**
- **Latency between GPU 2 , process bound to CPU 0 on both nodes: 2.47 *μ*s**
- **Latency between GPU 0 , process bound to CPU 0 on both nodes: 2.43 *μ*s**

OSU Bandwidth (osu_bw) Benchmark
Inter-node, K80 nodes

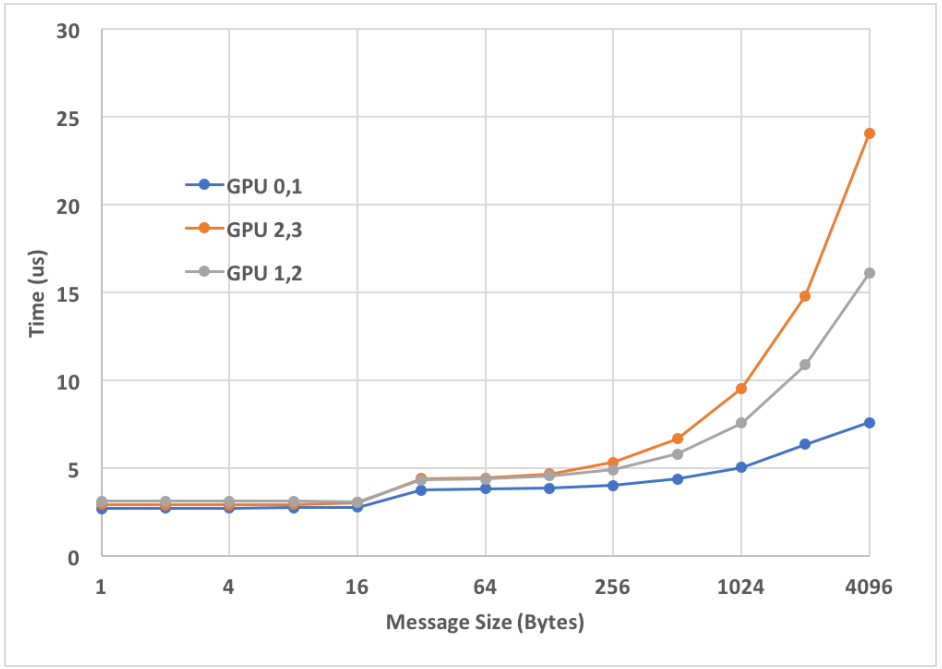


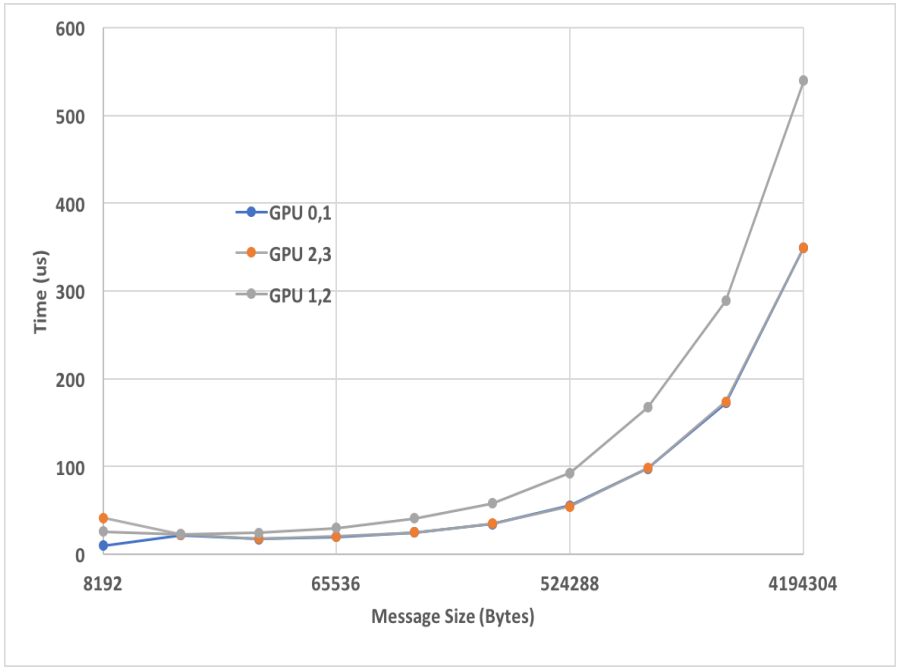
Comet P100 node architecture

```
        GPU0    GPU1    GPU2    GPU3    mlx4_0  CPU Affinity
GPU0     X      PIX     SOC     SOC     PHB     0-0,2-2,4-4,6-6,8-8,10-10,12-12,14-14,16-16,18-18,20-20,22-22,24-24,26-26
GPU1    PIX      X      SOC     SOC     PHB     0-0,2-2,4-4,6-6,8-8,10-10,12-12,14-14,16-16,18-18,20-20,22-22,24-24,26-26
GPU2    SOC     SOC      X      PIX     SOC     1-1,3-3,5-5,7-7,9-9,11-11,13-13,15-15,17-17,19-19,21-21,23-23,25-25,27-27
GPU3    SOC     SOC     PIX      X      SOC     1-1,3-3,5-5,7-7,9-9,11-11,13-13,15-15,17-17,19-19,21-21,23-23,25-25,27-27
mlx4_0  PHB     PHB     SOC     SOC      X

Legend:

  X    = Self
  SOC  = Connection traversing PCIe as well as the SMP link between CPU sockets(e.g. QPI)
  PHB  = Connection traversing PCIe as well as a PCIe Host Bridge (typically the CPU)
  PXB  = Connection traversing multiple PCIe switches (without traversing the PCIe Host Bridge)
  PIX  = Connection traversing a single PCIe switch
  NV#  = Connection traversing a bonded set of # NVLinks
```

- **4 GPUs per node**
- **GPUs (0,1) and (2,3) can do P2P communication**
- **Mellanox InfiniBand adapter associated with first socket (GPUs 0, 1)**

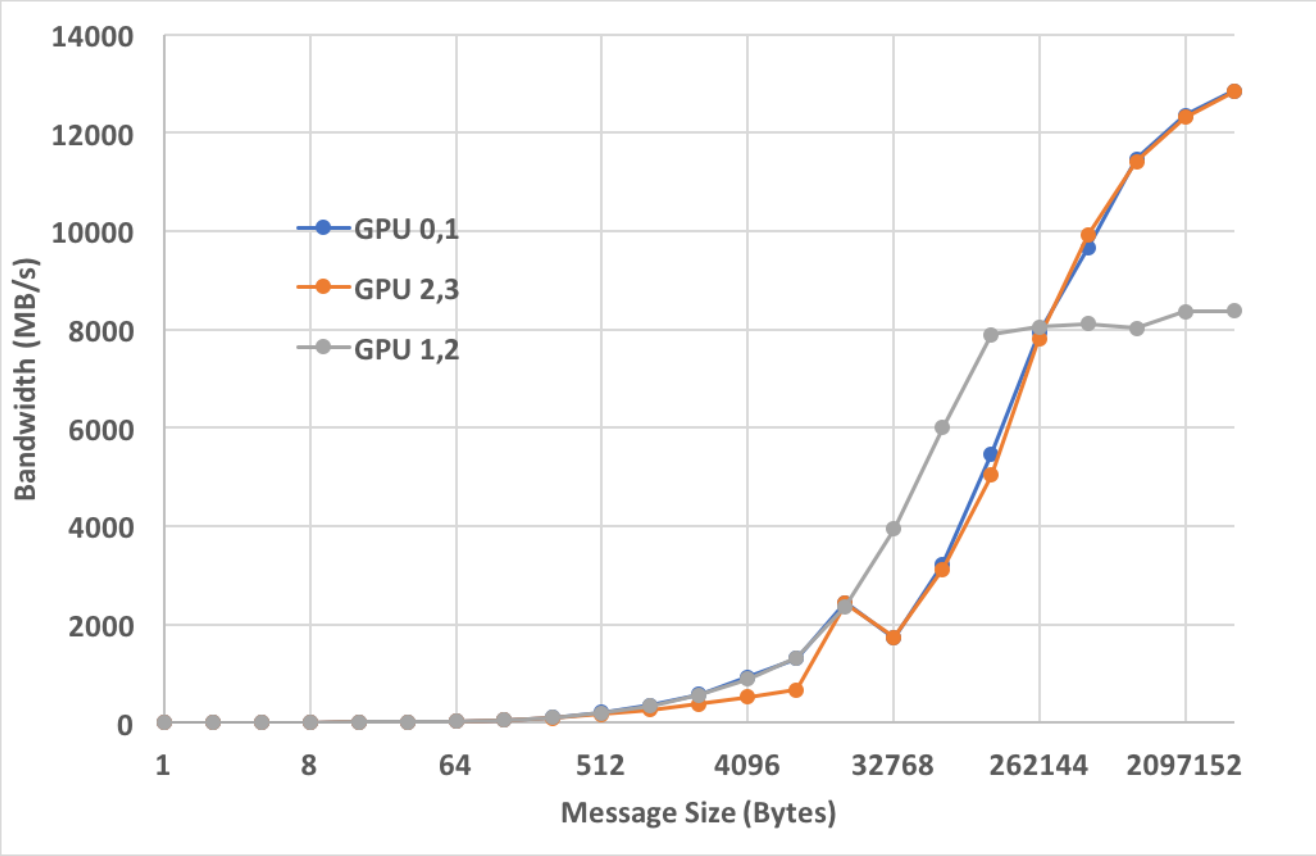
OSU Latency (osu_latency) Benchmark
Intra-node, P100 nodes



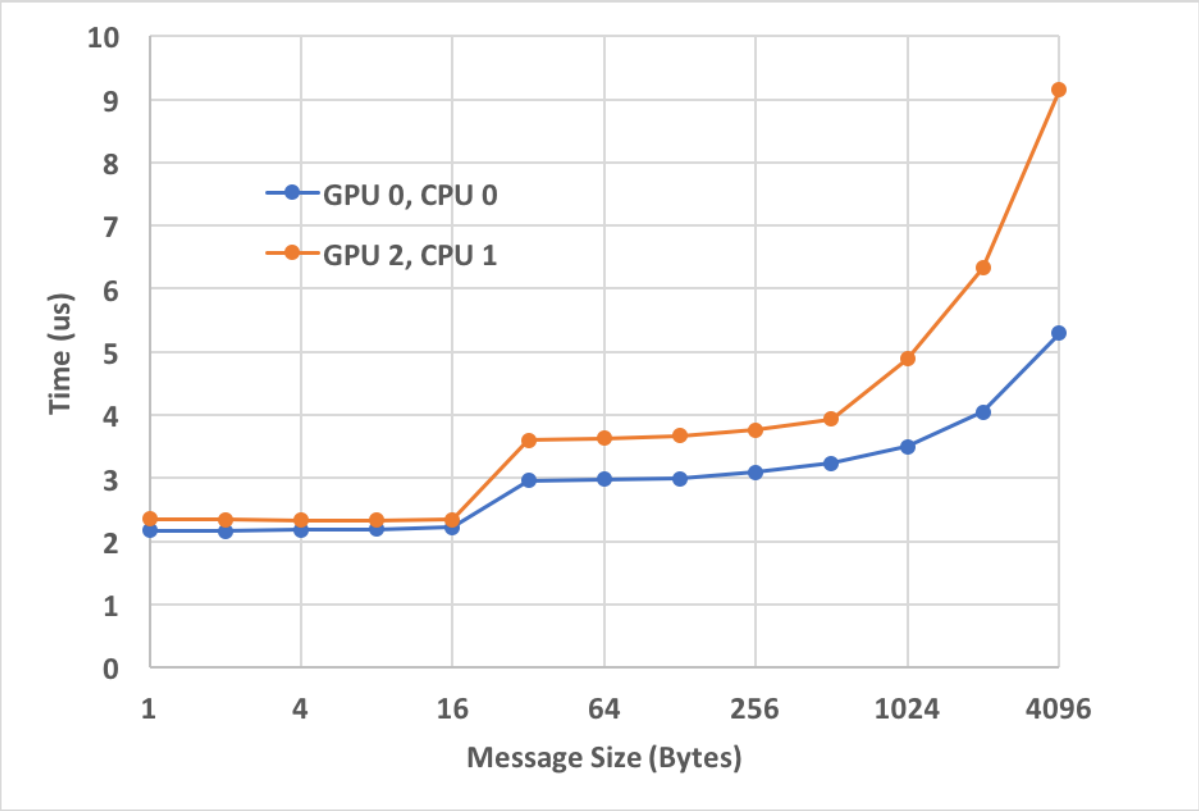


- **Latency between GPU 0 , GPU 1: 2.73 *μ*s**
- **Latency between GPU 2 , GPU 3: 2.95 *μ*s**
- **Latency between GPU 1 , GPU 2: 3.13 *μ*s**

OSU Bandwidth (osu_bw) Benchmark Intra-node, P100 nodes

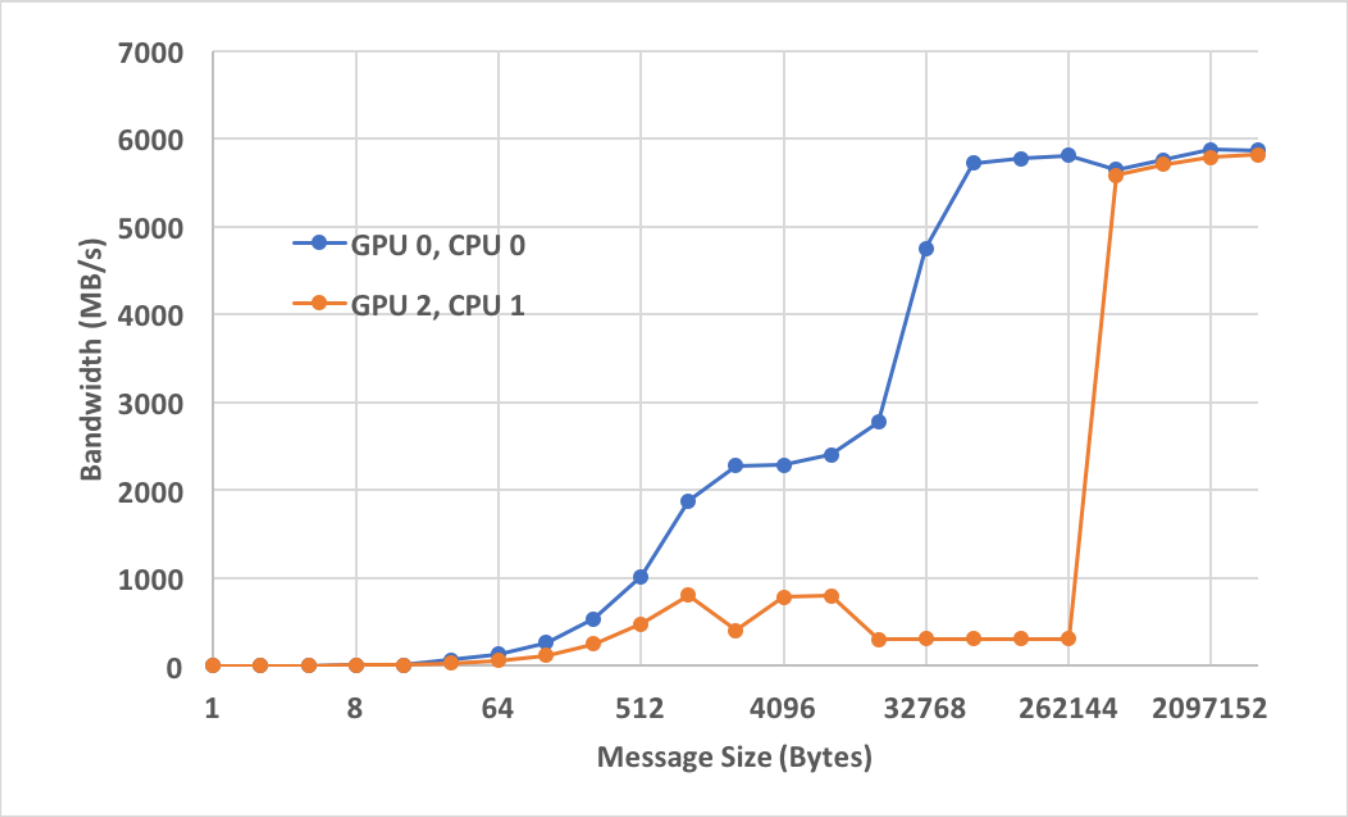


OSU Latency (osu_latency) Benchmark Inter-node, P100 nodes



- **Latency between GPU 0 , process bound to CPU 0 on both nodes: 2.17 *μ*s**
- **Latency between GPU 2 , process bound to CPU 1 on both nodes: 2.35 *μ*s**

OSU Bandwidth (osu_bw) Benchmark Inter-node, P100 nodes



HOOMD-blue Benchmarks using MVAPICH2-GDR

- HOOMD-blue is a *general-purpose* particle simulation toolkit
- **Results for Lennard-Jones liquid, Dodecahedron, Hexagon benchmarks**.

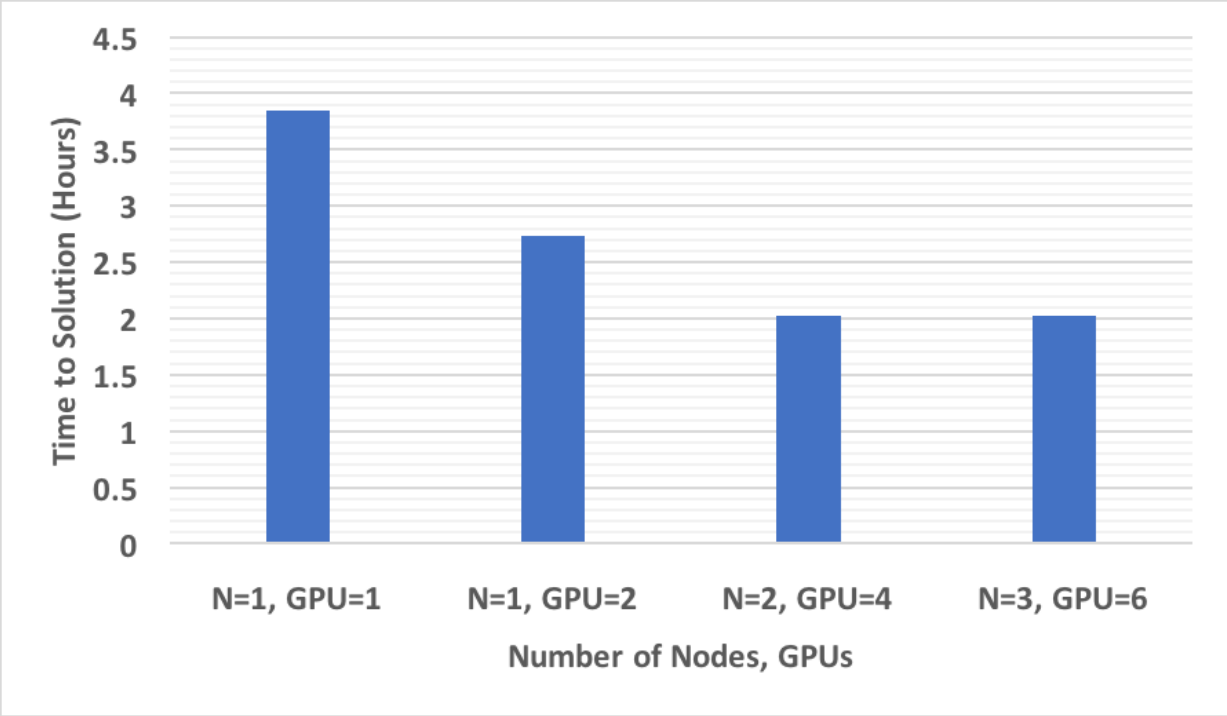
References:

- HOOMD-blue web page: http://glotzerlab.engin.umich.edu/hoomd-blue/
- HOOMD-blue Benchmarks page: http://glotzerlab.engin.umich.edu/hoomd-blue/benchmarks.html
- J. A. Anderson, C. D. Lorenz, and A. Travesset. General purpose molecular dynamics simulations fully implemented on graphics processing units *Journal of Computational Physics* 227(10): 5342-5359, May 2008. 10.1016/j.jcp.2008.01.047
- J. Glaser, T. D. Nguyen, J. A. Anderson, P. Liu, F. Spiga, J. A. Millan, D. C. Morse, S. C. Glotzer. Strong scaling of general-purpose molecular dynamics simulations on GPUs *Computer Physics Communications* 192: 97-107, July 2015. 10.1016/j.cpc.2015.02.028

HOOMD-Blue: lj-liquid benchmark

- N=64000, ρ=0.382
- Lennard-Jones pair force
  rcut=3.0
  ϵ=1.0
  σ=1.0
  δt=0.005
- Integration: Nosé-Hoover NVT
  T=1.2
  τ=0.5

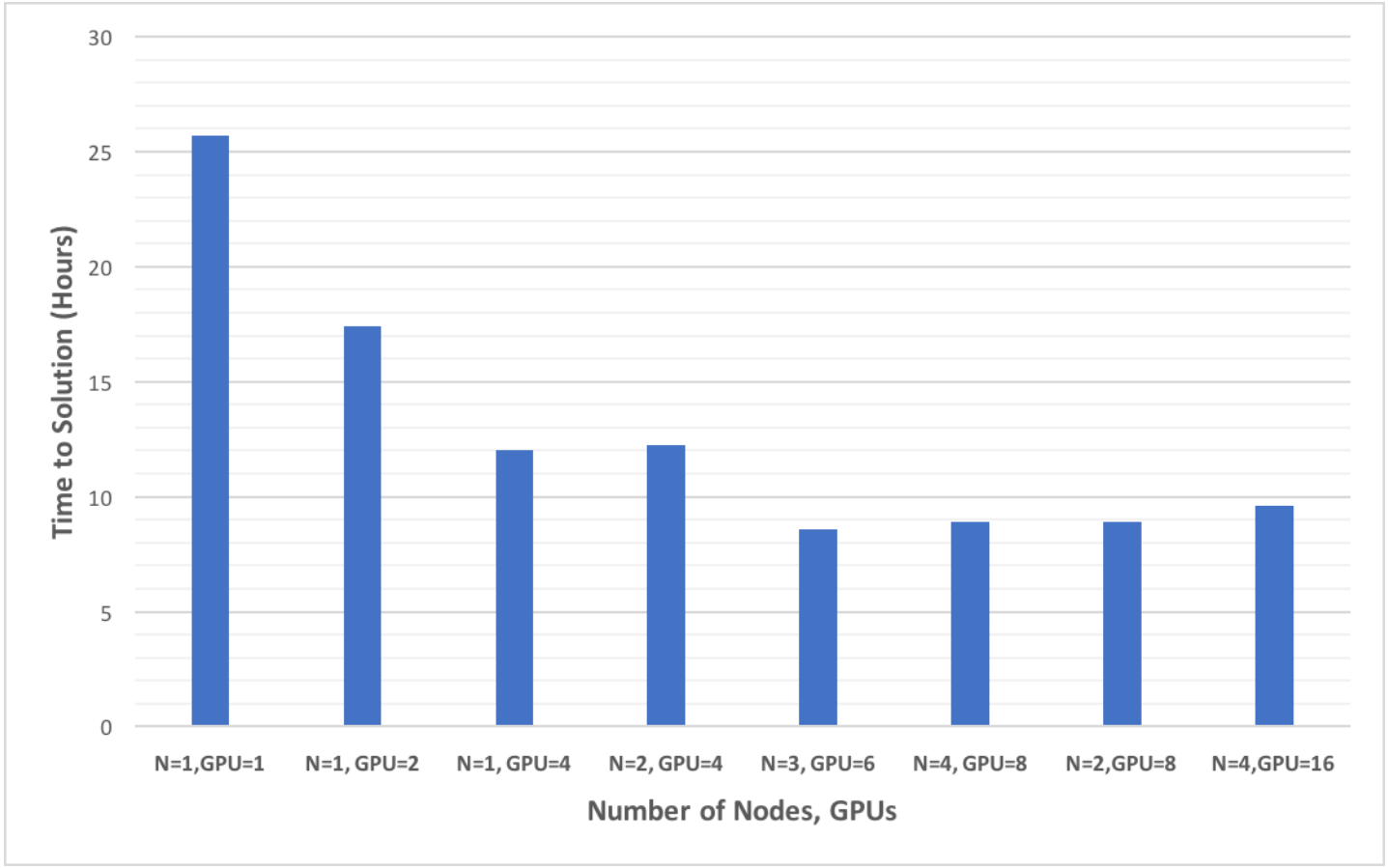
HOOMD-Blue : lj-liquid benchmark
Strong scaling on K80 nodes



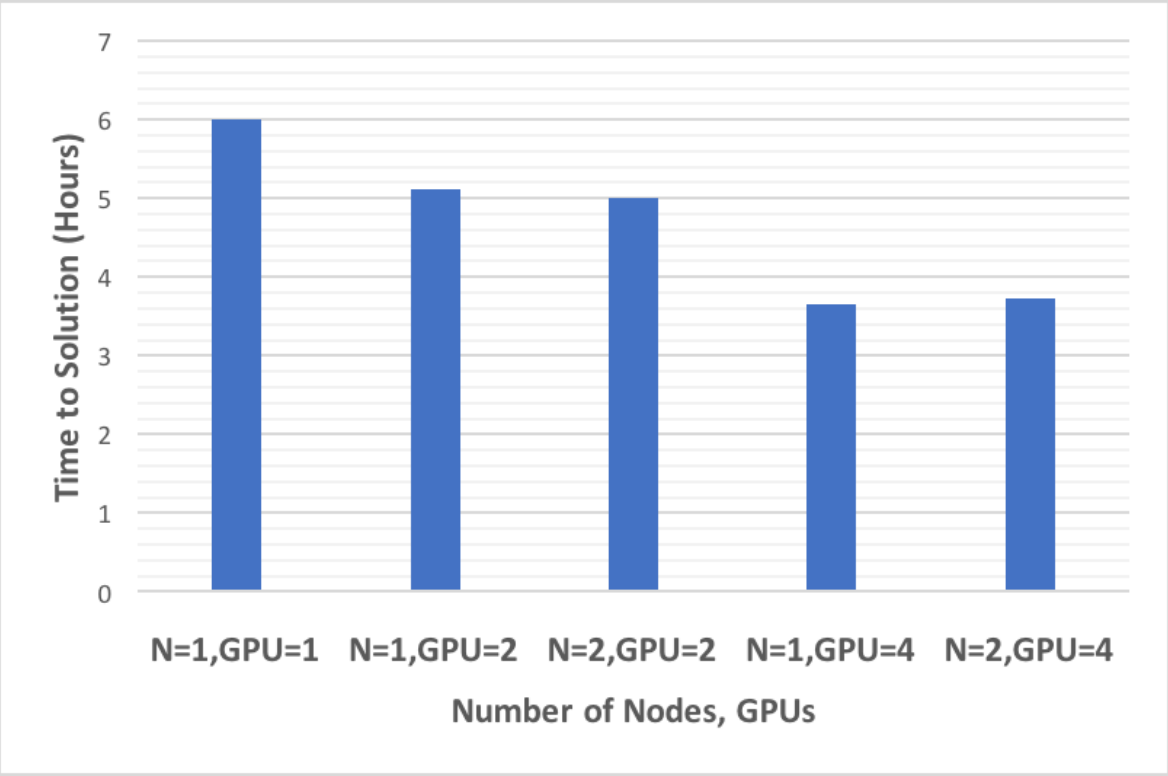
HOOMD-Blue: Dodecahdron Benchmark

- N=131,072
- Hard particle Monte Carlo
  - Vertices: *details in dodecahdron/bmark.py*
  - d=0.3
  - a=0.26
  - nselect=4
- Synthetic benchmark of 3D convex polyhedra performance.

HOOMD-Blue: Dodecahdron Benchmark
Strong scaling on K80 nodes



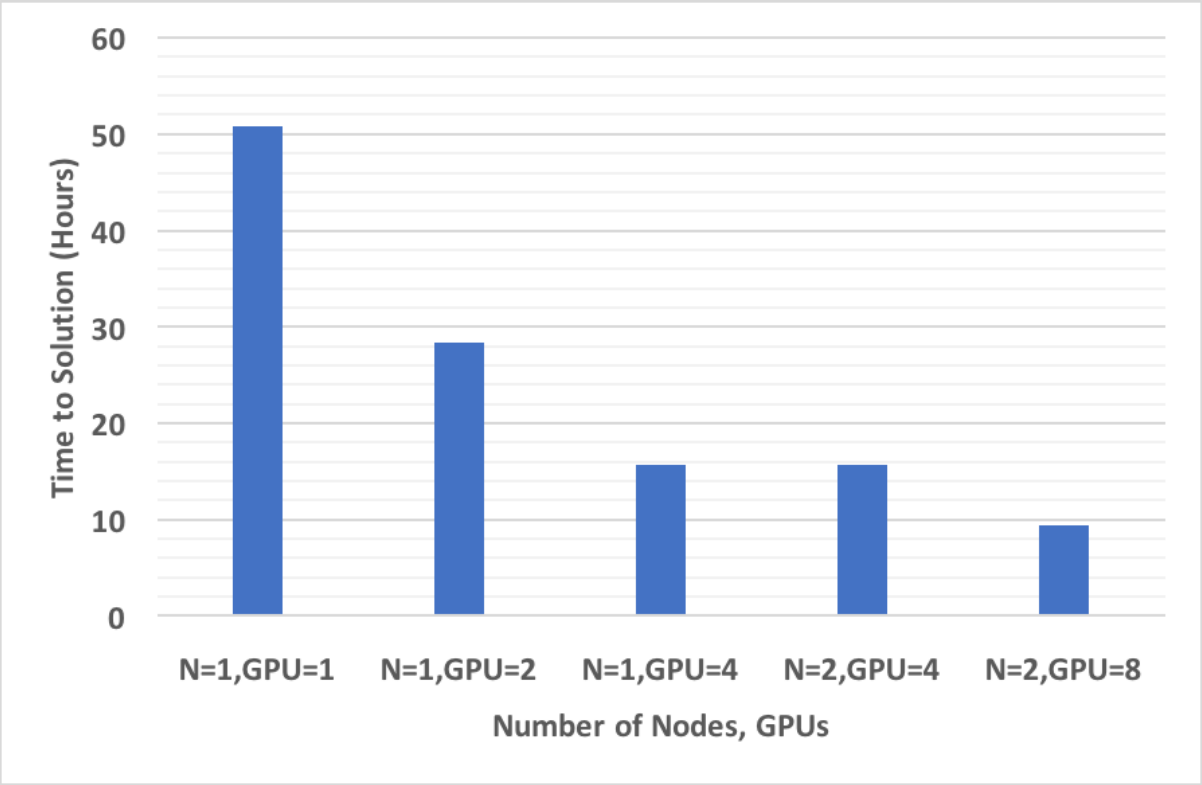
HOOMD-Blue: Dodecahdron Benchmark
Strong scaling on P100 nodes



HOOMD-Blue: Hexagon Benchmark

- N=1,048,576N=1,048,576
- Hard particle Monte Carlo
  - Vertices: [[0.5,0],[0.25,0.433012701892219],[-0.25,0.433012701892219],[-0.5,0],[-0.25,-0.433012701892219],[0.25,-0.433012701892219]]
  - d=0.17010672166874857
  - a=1.047197551196597 6
  - nselect=4
- Log file period: 10000 time steps
- SDF analysis
  - xmax==0.02
  - $\delta x = 10-4$
  - period: 50 time steps
  - navg=2000
- DCD dump period: 100000

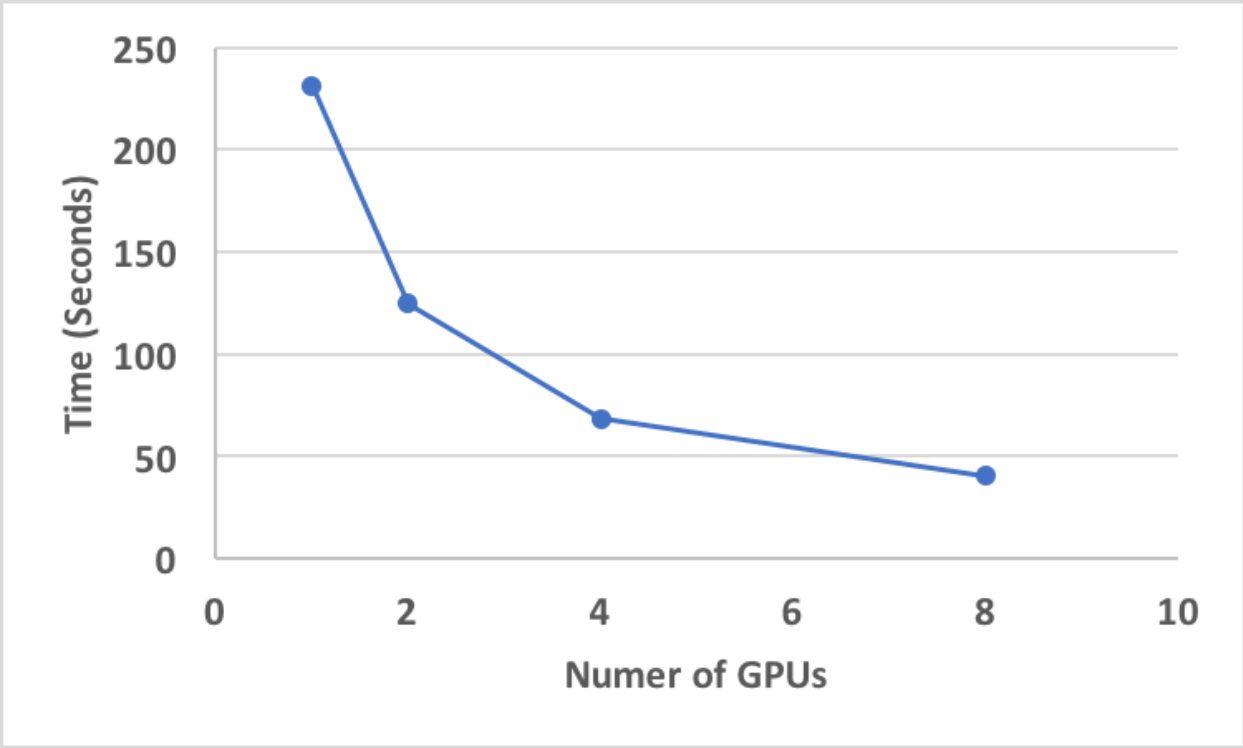
HOOMD-Blue: Hexagon Benchmark
Strong scaling on K80 nodes



OSU-Caffe, CIFAR10 Quick on K80 nodes

- Preliminary results with K80 nodes.
- Current runs with data in Lustre filesystem (/oasis/scratch/comet)
- All Comet GPU nodes have 280GB of SSD based local scratch space. Future tests with larger test cases planned to evaluate performance advantages of using the SSDs.

OSU-Caffe, CIFAR10 Quick on K80 nodes



Summary

- OSU benchmarks show expected results using MVAPICH2-GDR for GPUs associated with same socket, with HCA on same socket.
- Strong scaling results for HOOMD-Blue on several benchmarks: Performance from runs using all GPUs on a node is similar to runs splitting the same number of GPUs on multiple nodes.
- Preliminary results for OSU-Caffe with CIFAR10 benchmark show good scaling. All Comet GPU nodes have 280GB of SSD based local scratch. Future tests with larger test cases planned to evaluate performance advantages of using the SSDs.
- Thanks to the MVAPICH group for excellent support for the various MVAPICH installations on SDSC machines!

Thanks!

Questions: Email mahidhar@sdsc.edu