Performance of Applications on Comet Nodes Utilizing MVAPICH2-GDR, Singularity, and MVAPICH2-Virt

Mahidhar Tatineni

MVAPICH User Group Meeting

August 7, 2018

UC San Diego

SDSC

SAN DIEGO SUPERCOMPUTER CENTER

INDIANA UNIVERSITY

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

Comet: System Characteristics

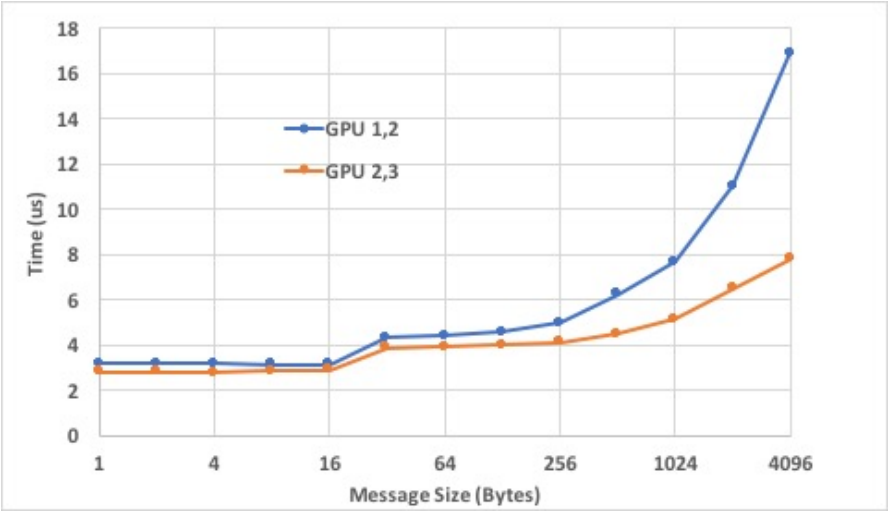
- **Total peak flops ~2.1 PF**
- **Dell primary integrator**
  - Intel Haswell processors w/ AVX2
  - Mellanox FDR InfiniBand
- **1,944 standard compute nodes (46,656 cores)**
  - Dual CPUs, each 12-core, 2.5 GHz
  - 128 GB DDR4 2133 MHz DRAM
  - 2*160GB GB SSDs (local disk)
- **72 GPU nodes**
  - 36 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**
- **Virtual Cluster Capability**
- **100 Gbps external connectivity to Internet2 & ESNet**

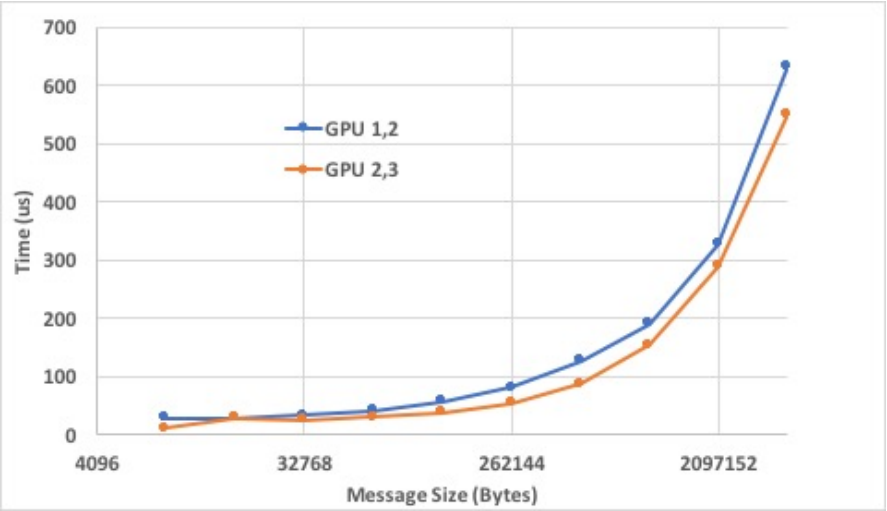
Comet K80 node architecture

```
        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)**

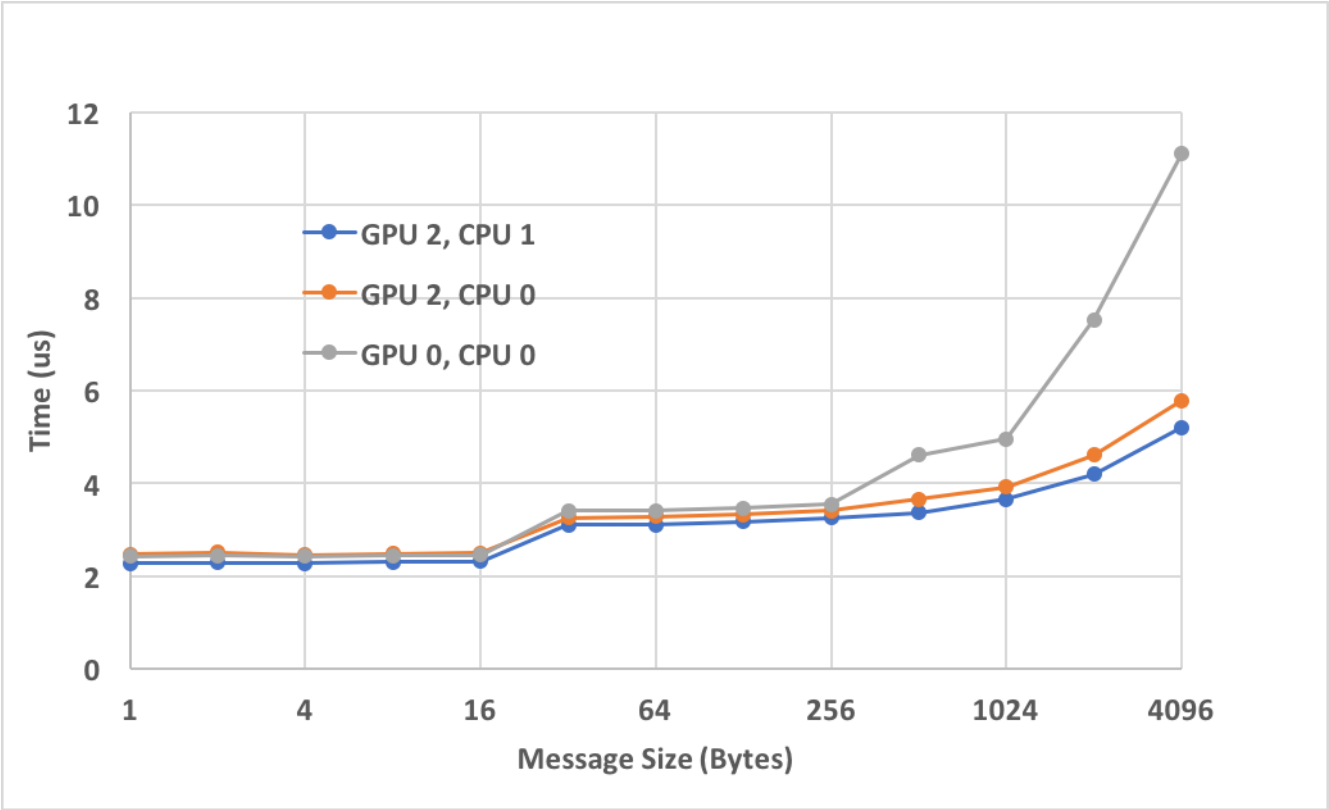
OSU Latency (osu_latency) Benchmark
Intra-node, K80 nodes





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

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



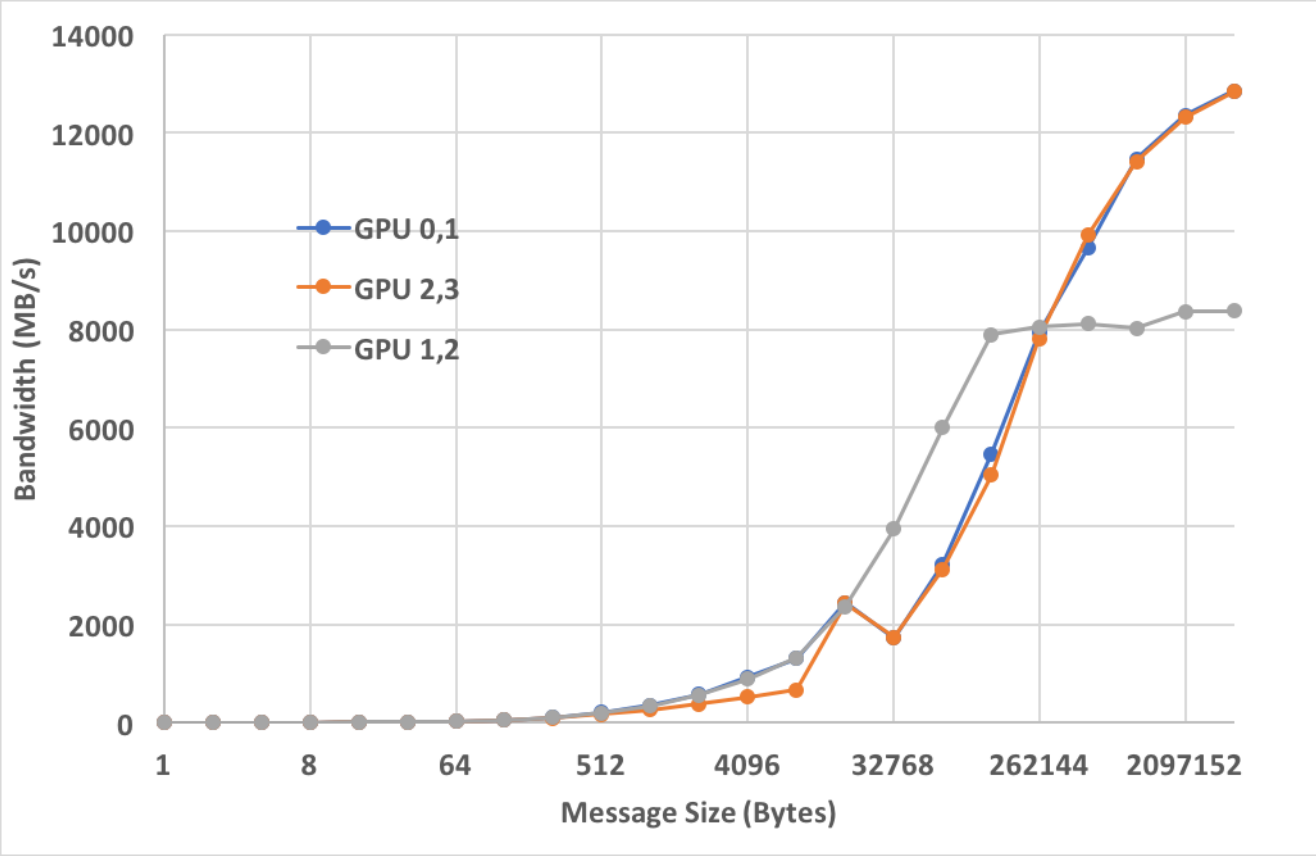
- **Latency between GPU 2 , process bound to CPU 1 on both nodes: 2.27 *μ*s**
- **Latency between GPU 2 , process bound to CPU 0 on both nodes: 2.47 *μ*s**
- **Latency between GPU 0 , process bound to CPU 0 on both nodes: 2.43 *μ*s**

Comet P100 node architecture

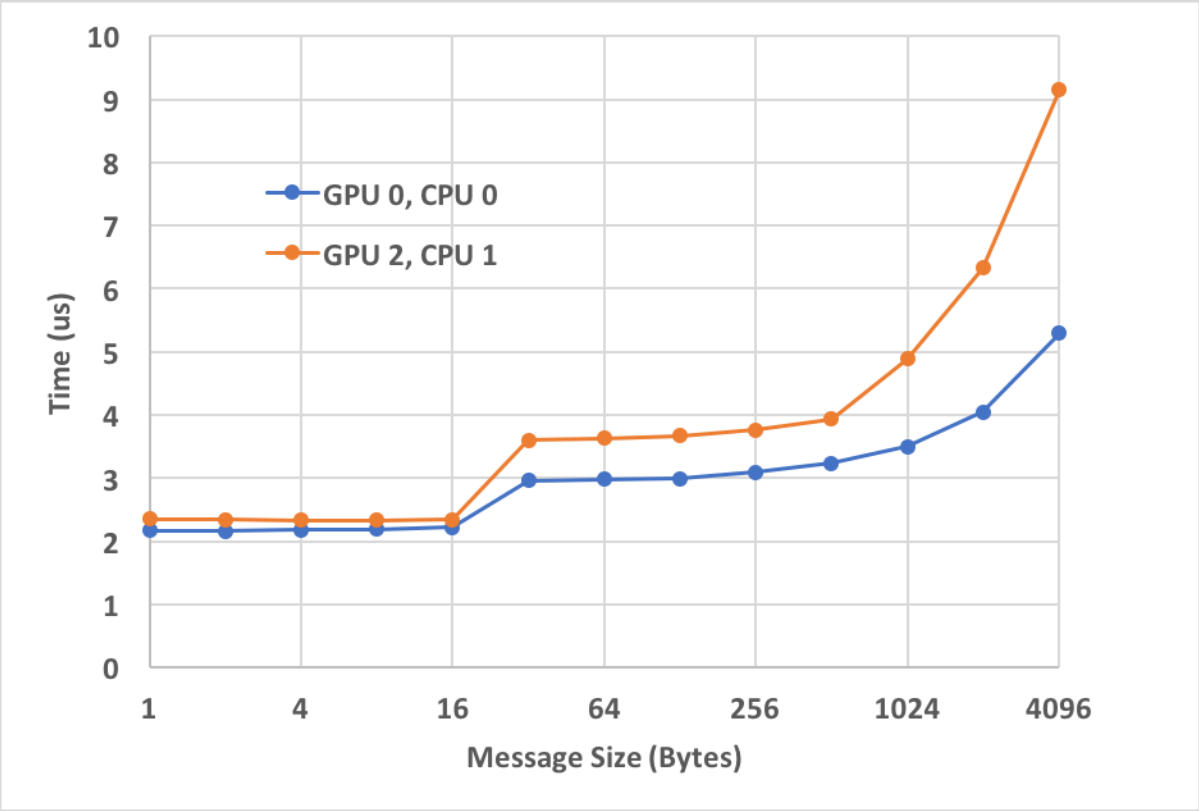
```
        GPU0    GPU1    GPU2    GPU3    mlx4_0  CPU Affinity
GPU0     X      PIX     SOC     SOC     PHB     0-0,2-2,4-4,6-6,8-8,10-10,12-12,14-14,16-16,18-18,20-20,22-22,24-24,26-26
GPU1    PIX      X      SOC     SOC     PHB     0-0,2-2,4-4,6-6,8-8,10-10,12-12,14-14,16-16,18-18,20-20,22-22,24-24,26-26
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 Bandwidth (osu_bw) Benchmark Intra-node, P100 nodes

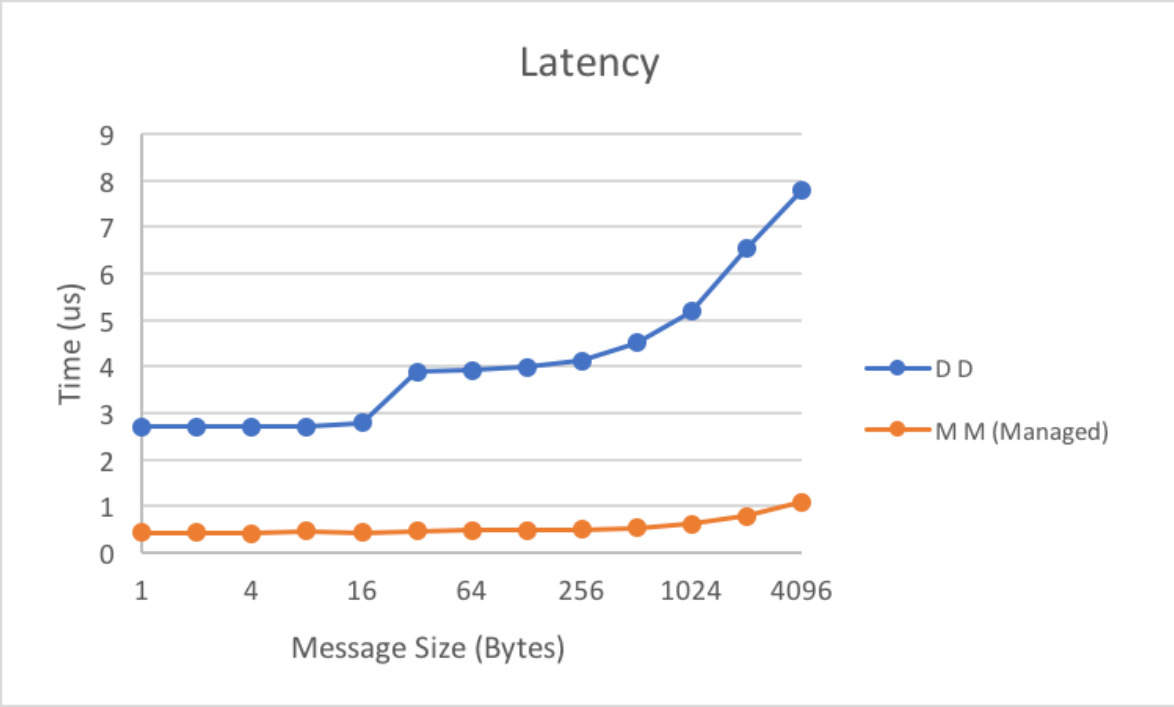


OSU Latency (osu_latency) Benchmark Inter-node, P100 nodes



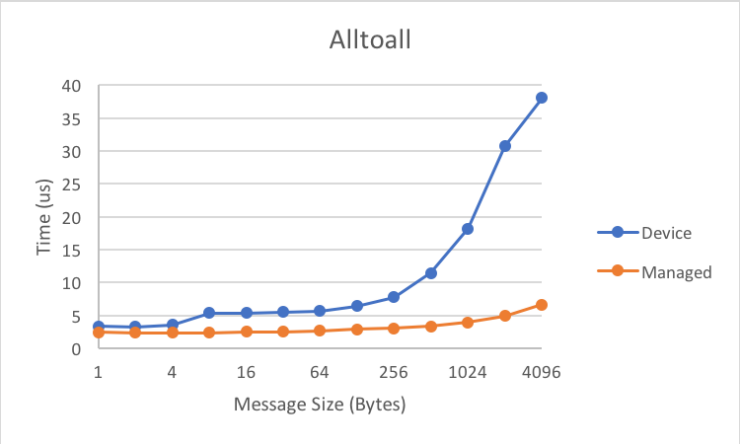
- **Latency between GPU 0 , process bound to CPU 0 on both nodes: 2.17 *μ*s**
- **Latency between GPU 2 , process bound to CPU 1 on both nodes: 2.35 *μ*s**

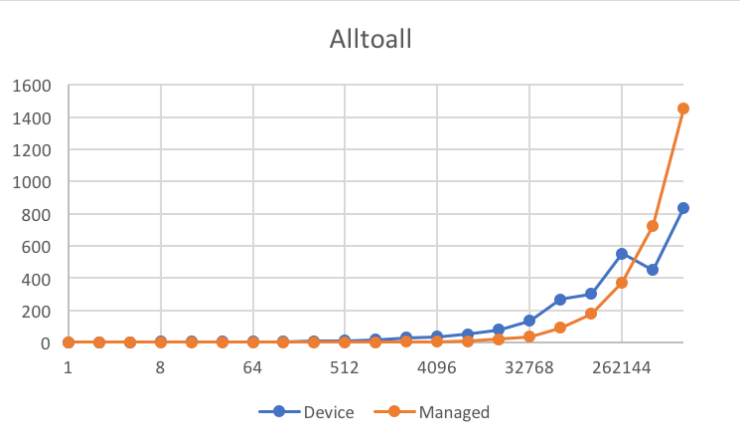
OSU Latency (osu_latency) Benchmark Intra-node, P100 nodes



- **Latency between GPU 0,1 ; MPI tasks pinned to cores 0:2**
- **M M: CUDA Managed (or Unified) memory allowing a common memory allocation for GPU, CPU**
- **MVAPICH2 to perform communications directly from managed buffers**

osu_alltoall Benchmark
Intra-node, P100 nodes





- **Alltoall with 4 tasks (cores 0:2:1:3), 4 GPUs**
- **CUDA Managed (or Unified) memory allowing a common memory allocation for GPU, CPU**
- **MVAPICH2 to perform communications directly from managed buffers**

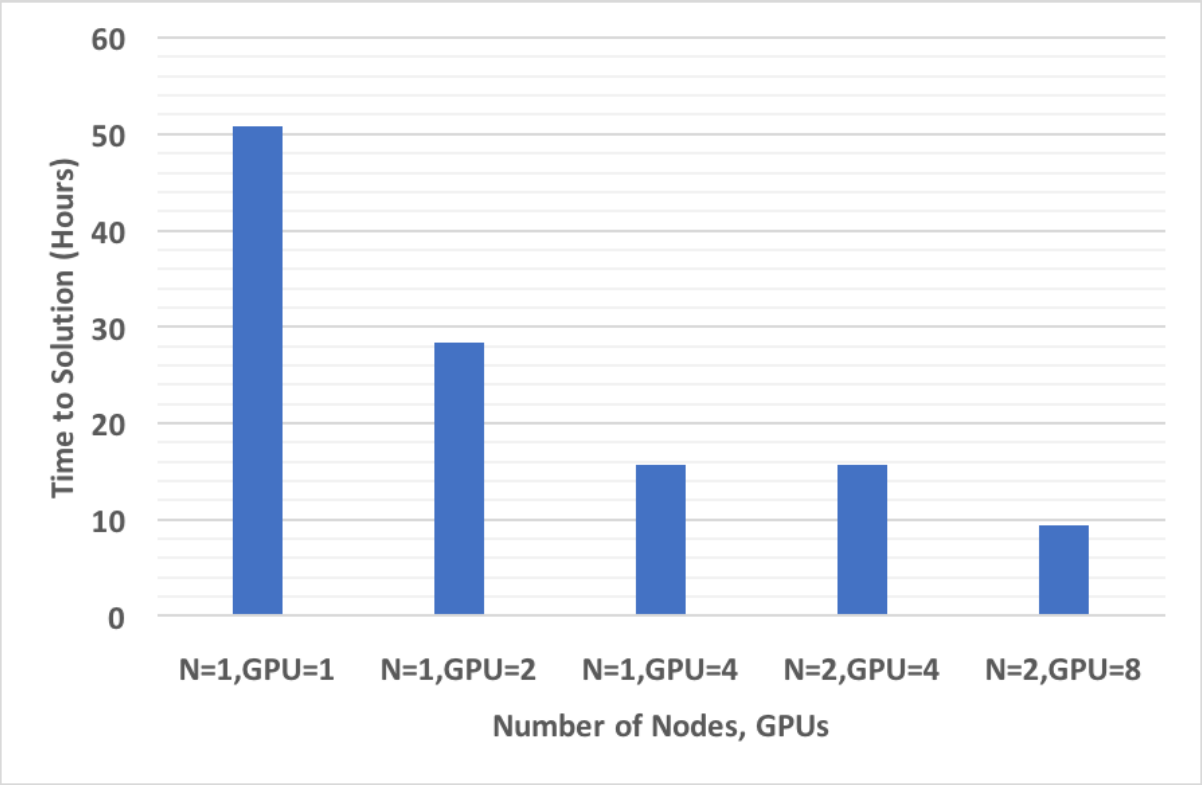
HOOMD-blue Benchmarks using MVAPICH2-GDR

- HOOMD-blue is a *general-purpose* particle simulation toolkit
- **Results for Hexagon benchmark**.

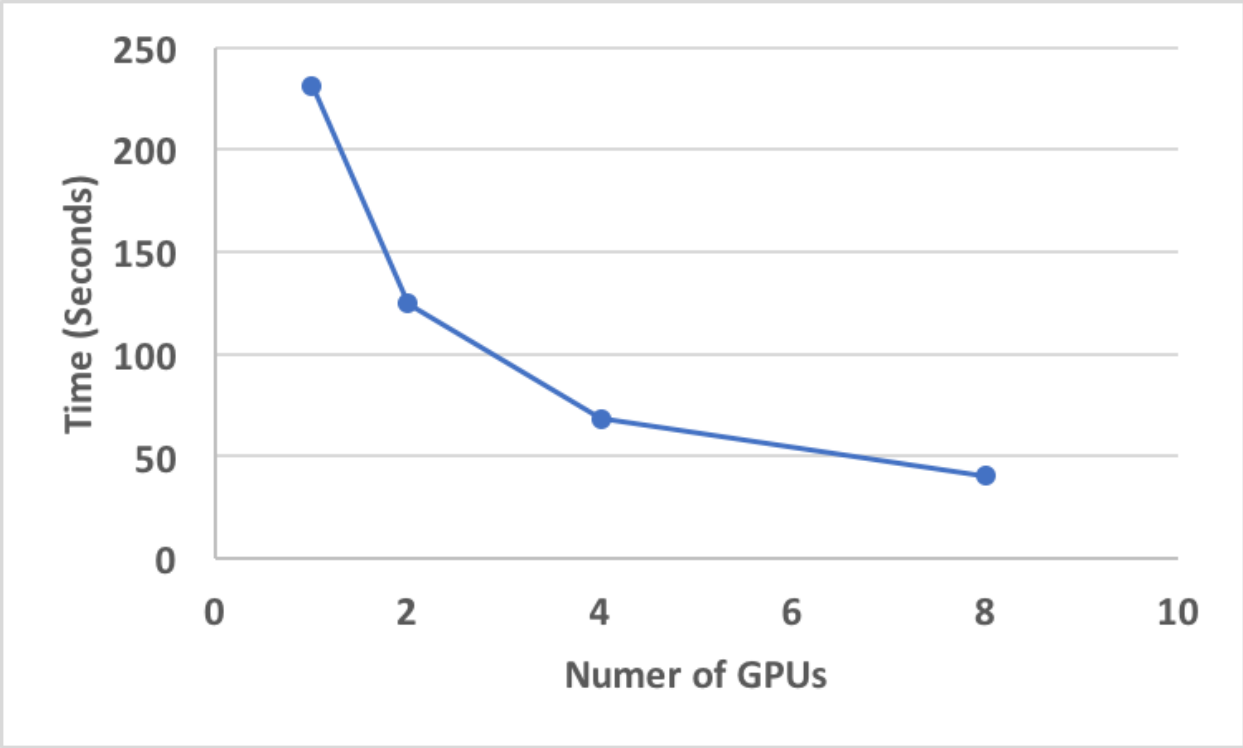
References:

- 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: Hexagon Benchmark
Strong scaling on K80 nodes



OSU-Caffe, CIFAR10 Quick on K80 nodes MVAPICH2-GDR/2.2, CUDA/7.5



Virtualization on Comet

- **Containers using Singularity (http://singularity.lbl.gov)**
  - Migrate complex software stacks from their campus to Comet.
  - Singularity runs in user space, and requires very little special support – in fact it actually reduces it in some cases.
  - Applications include: Tensorflow, Torch, Fenics, and custom user applications.
  - Docker images can be imported into Singularity
  - Currently used by ~20 research groups on Comet.
- **Comet Virtual Clusters**
  - KVM based full virtualization with SRIOV support.
  - Full root access, PXE install, persistent disk images, near native InfiniBand
  - Nucleus Rest API and Cloudmesh (Indiana University) management
  - Backends to scheduled jobs consuming XSEDE allocations.

Comet VC Use Cases

- **CAIDA Hackathon**
  - Root access to nodes for custom OS and software stack.
  - Full control of network stack inside virtual compute nodes by attendees and easy ‘repair’ by CAIDA admins
  - Full isolation of virtual cluster from production resources and filesystems
- **Open Science Grid**
  - Simple install using existing management infrastructure (PXE, Foreman, Puppet)
  - Multiple XSEDE allocations consuming SUs via OSG VC with no effort from allocated projects
  - Largest OSG provider of resources (> 2x) for last LIGO run

Application example using Singularity and MVAPICH2

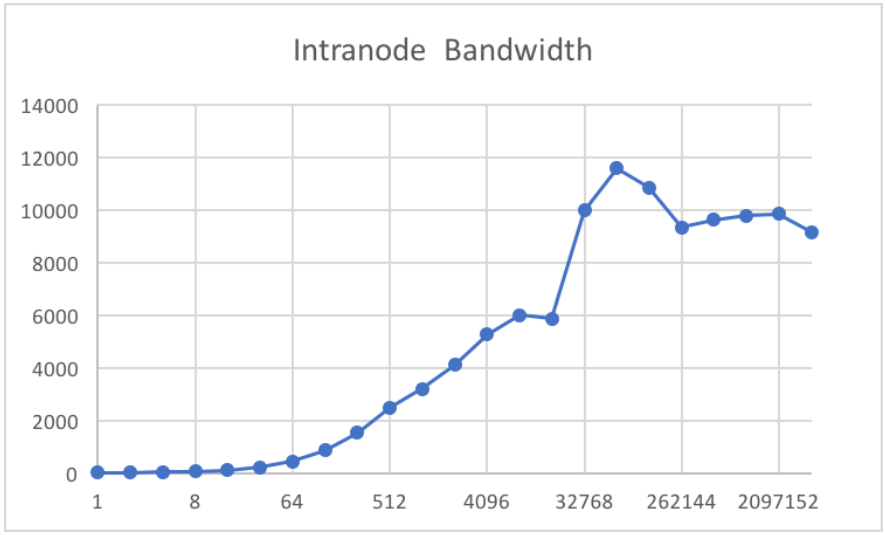
- **Neuron YuEtAl2012 benchmark, compared the same build options using gnu+MVAPICH2 compilers via singularity.**

Cores	Time (seconds)
192	373
384	188
768	107

MVAPICH2-Virt on Comet Virtual Cluster

- **Preliminary tests as the QEMU version is old. Upgrade upcoming that will enable latest Virt version.**

```
[mahidhar@compute-0-2 osu-micro-benchmarks-4.4.1]$ mpirun_rsh -export-all -np 2
 compute-0-2 compute-0-2 MV2_VIRT_USE_IVSHMEM=1 ./mpi/pt2pt/osu_bw
Failed to find IVShmem device, will fallback to SR-IOV
Failed to find IVShmem device, will fallback to SR-IOV
# OSU MPI Bandwidth Test v4.4.1
# Size      Bandwidth (MB/s)
1                       7.07
2                      14.66
4                      29.33
8                      58.20
16                    116.68
32                    228.76
64                    458.74
128                   902.30
256                  1624.19
512                  2839.06
1024                 4041.73
2048                 5621.80
4096                 7257.50
8192                 8540.89
16384                9339.66
32768               10566.51
65536               11864.57
```



PSDNS Benchmark on Virtual Cluster

- FFT based application; Communication intensive, mainly alltoallv – bisection bandwidth limited.

Cores (Nodes)	Time per Step (s)
32(2)	101.51
64(4)	67.03
128(8)	33.99

Summary

- **OSU benchmarks show expected results using MVAPICH2-GDR**
- **Comet offers several MVAPICH2 flavors - MVAPICH2, MVAPICH2-GDR, MVAPICH2-Virt, MVAPICH2-X.**
- **Upcoming upgrades on Comet**
  - OS will be upgraded to current CentOS 7 version
  - Will enable XPMEM support
  - GPU drivers will be upgraded to 396.26, will enable CUDA 9
  - Defaults for MVAPICH2 will be upgraded to latest release versions
- **Thanks to the MVAPICH group for excellent support for the various MVAPICH installations on SDSC machines!**

Thanks!

Questions: Email mahidhar@sdsc.edu