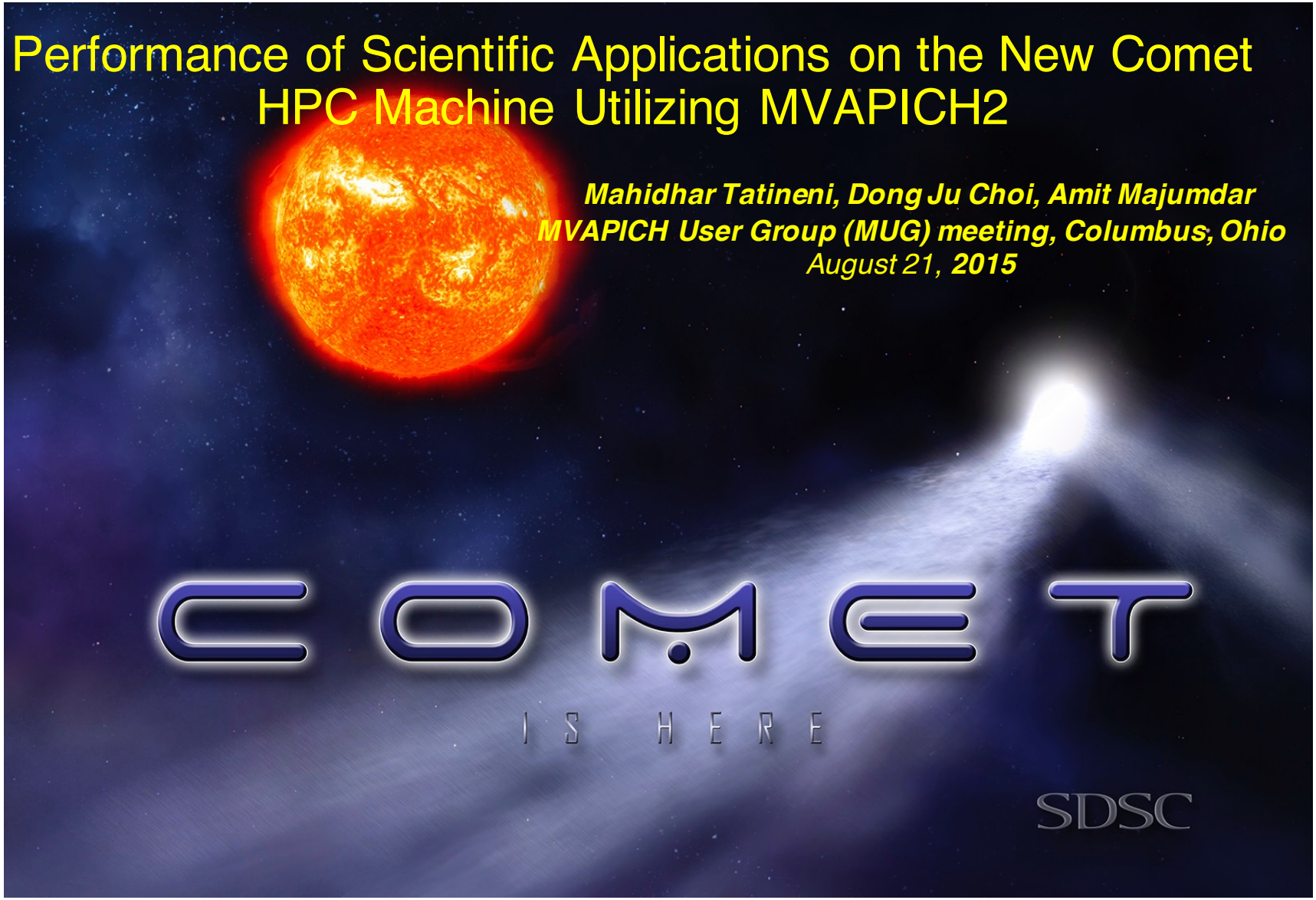
Performance of Scientific Applications on the New Comet HPC Machine Utilizing MVAPICH2

Mahidhar Tatineni, Dong Ju Choi, Amit Majumdar

MVAPICH User Group (MUG) meeting, Columbus, Ohio

August 21, ***2015***



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

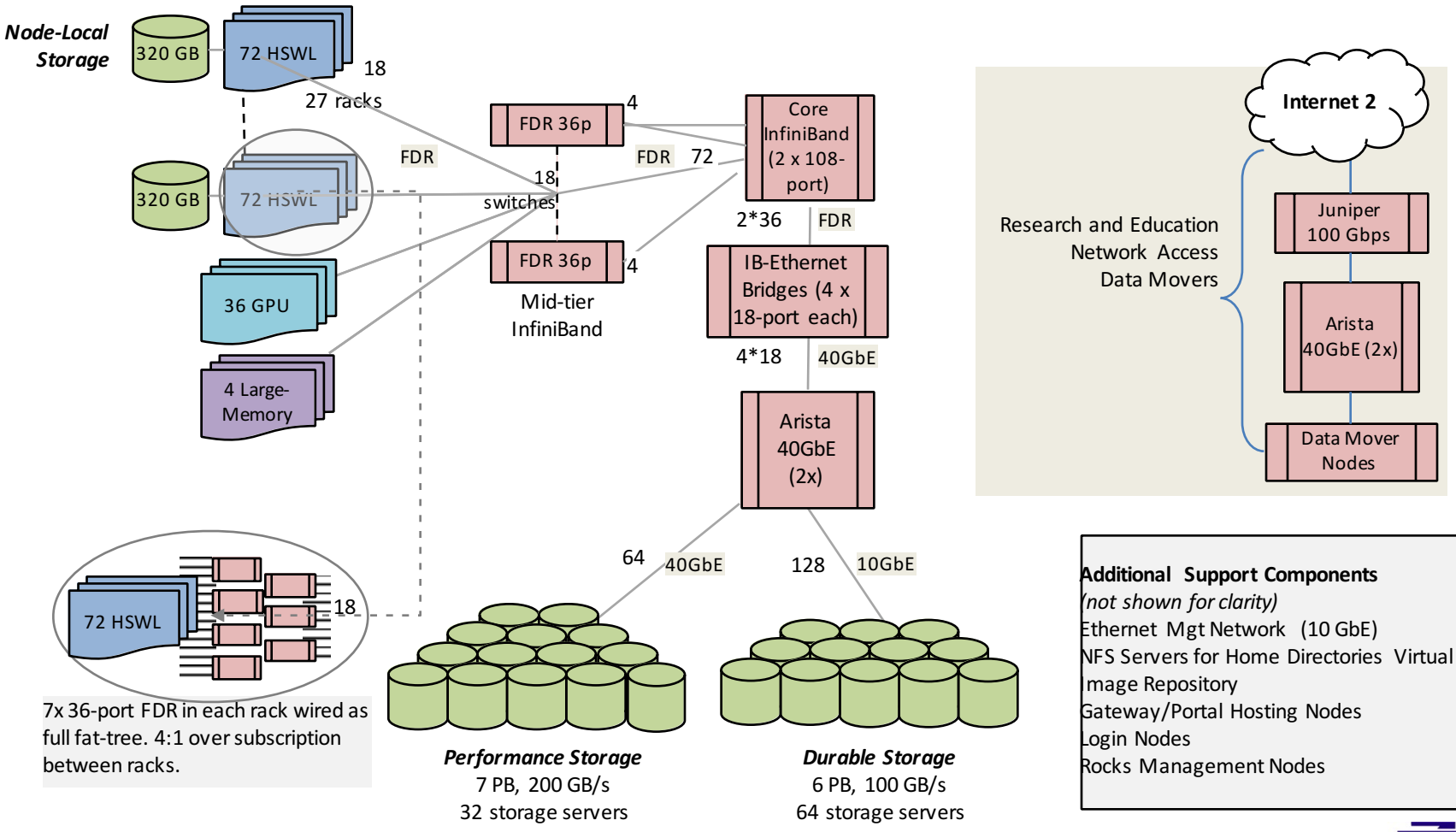
Comet
"HPC for the long tail of science"

iPhone panorama photograph of 1 of 2 server rows

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)
- **36 GPU nodes**
  - Same as standard nodes *plus*
  - Two **NVIDIA K80 cards**, each with dual Kepler3 GPUs
- **4 large-memory nodes (**August **2015)**
  - 1.5 TB DDR4 1866 MHz DRAM
  - Four Haswell processors/node
- **High Performance Virtualization leveraging Single Root IO Virtualization (SR-IOV)**
- **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**
- **Virtual image repository**
- **100 Gbps external connectivity to Internet2 & ESNet**

Comet Network Architecture
InfiniBand compute, Ethernet Storage



Gordon – A Data Intensive Supercomputer

- Designed to accelerate access to massive amounts of data in areas of genomics, earth science, engineering, medicine, and others
- Emphasizes memory and IO over FLOPS.
- Appro integrated 1,024 node Sandy Bridge cluster
- 300 TB of high performance Intel flash
- Large memory supernodes via vSMP Foundation from ScaleMP
- 3D torus interconnect from Mellanox
- In production operation since February 2012
- Funded by the NSF and available through the NSF Extreme Science and Engineering Discovery Environment program (XSEDE)

SDSC

APPRO

NSF

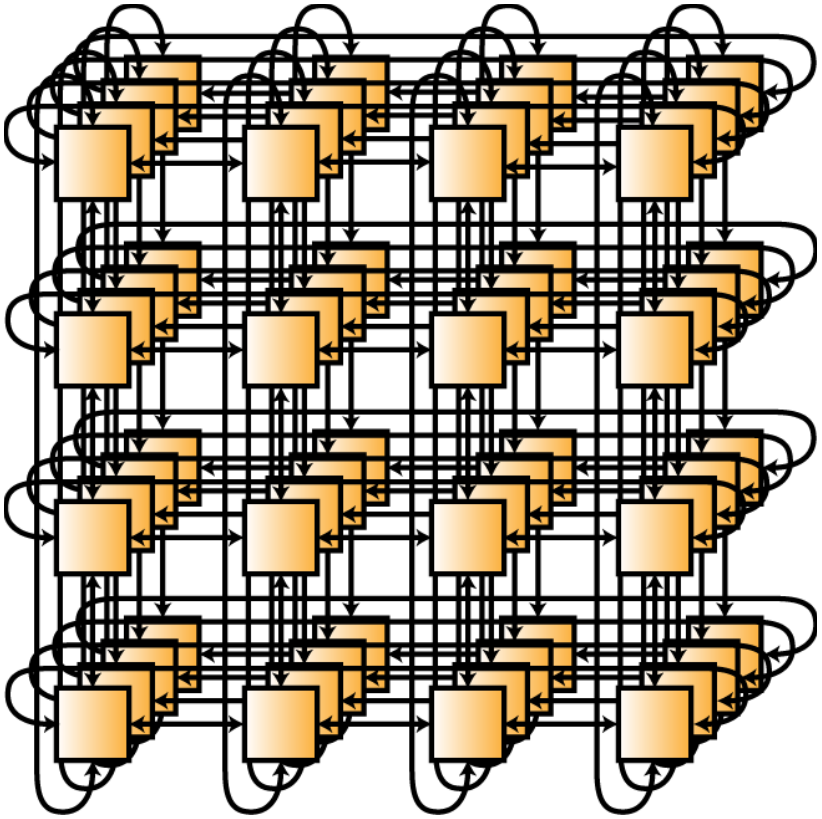
intel

Mellanox TECHNOLOGIES

ScaleMP

XSEDE
Extreme Science and Engineering Discovery Environment

3D Torus of Switches



- Linearly expandable
- Simple wiring pattern
- Short Cables- Fiber Optic cables generally not required
- Lower Cost :40% as many switches, 25% to 50% fewer cables
- Works well for localized communication
- Fault Tolerant within the mesh with 2QoS Alternate Routing
- Fault Tolerant with Dual-Rails for all routing algorithms

3rd dimension wrap-around not shown for clarity

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 (v2.1 and v2.1rc2) on the GPU nodes (featuring NVIDIA K80s)**
- ***Coming Soon* … MVAPICH2-Virt, to leverage SR-IOV to support virtualized HPC clusters within Comet.**
- **RDMA-Hadoop (from Dr. Panda's HiBD lab) also available.**

Applications benchmarked using MVAPICH2

- **NEURON -** Simulation environment for modeling individual neurons and networks of neurons. Benchmark used large-scale model of olfactory bulb, 10,500 cells, for 40,000 time steps.

- **OpenFOAM** - Open source software for computational fluid dynamics (CFD). Benchmark involves LES of flow over backward-facing step, 8M mesh points, 200 time steps.

- **Quantum ESPRESSO** - Integrated suite of Open-Source computer codes for electronic-structure calculations and materials modeling at the nanoscale. Gold surface covered by thiols and water, 4 kpoints, 586 atoms, 2552 electrons, 5 iterations.

- **RAxML** - Phylogenetic Analysis tool. Comprehensive analysis, 218 taxa, 2294 characters, 1846 patterns, 100 bootstraps.

Results: NEURON benchmark

Benchmark used large-scale model of olfactory bulb, 10,500 cells, for 40,000 time steps. Results compared with runs on Gordon.

Cores	Gordon (with OpenMPI v1.6.5) Time (s)	Gordon (with MVAPICH2 1.9) Time (s)	Comet (with MVAPICH2 2.1) Time (s)
48	1947	1429	1207
96	1016	732	590
192	485	374	304
384	249	183	144
768	163	104	84

Results: OpenFOAM benchmark

Benchmark involves LES of flow over backward-facing step, 8M mesh points, 200 time steps.

Cores	Gordon (Using OpenMPI 1.6.5) Time (s)	Gordon (Using MVAPICH2 1.9) Time (s)	Comet (Using MVAPICH2 2.1) Time (s)
96	1962	1621	1368
144	1169	928	787
192	786	665	544
384	310	284	207

Results: QE Benchmark

Gold surface covered by thiols and water, 4 kpoints, 586 atoms, 2552 electrons, 5 iterations.

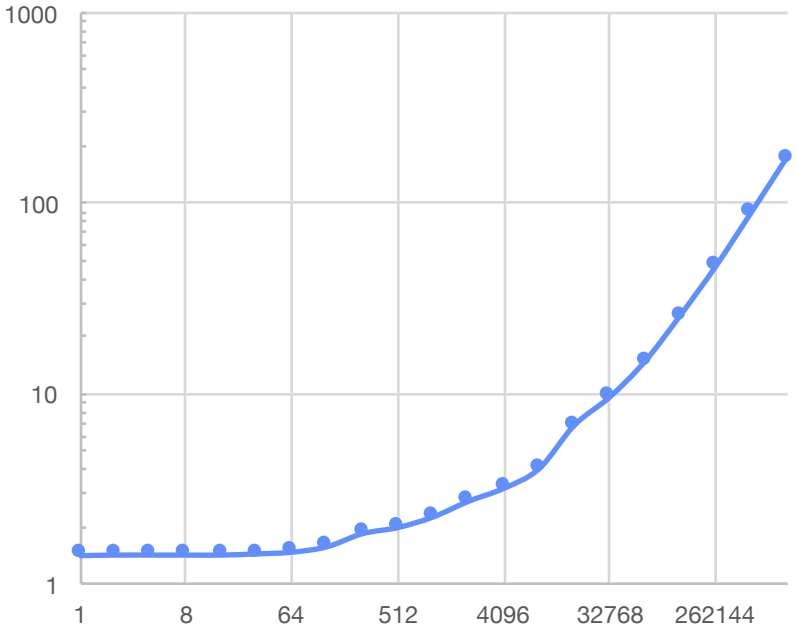
Cores	Gordon (Using OpenMPI 1.6.5) Time (s)	Gordon (Using MVAPICH2 1.9) Time (s)	Comet (Using MVAPICH2 2.1) Time (s)
96	3046	2564	1893
192	1754	1333	940
384	1025	778	543
768	784	486	345

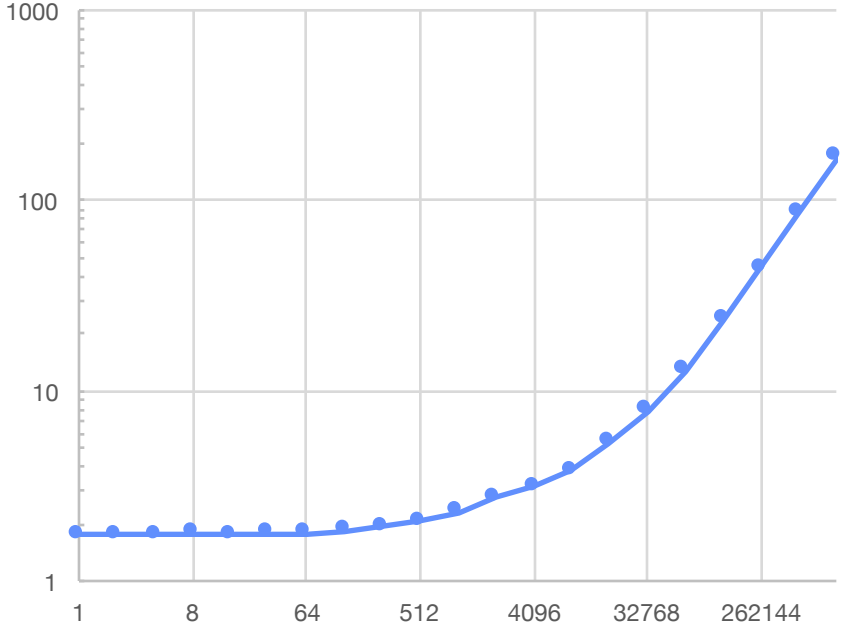
Results: RAxML Benchmark

Phylogenetic Analysis tool. Comprehensive analysis, 218 taxa, 2294 characters, 1846 patterns, 100 bootstraps.

Cores	Gordon (Using MVAPICH2 1.9) Time (s)	Comet (Using MVAPICH2 2.1) Time (s)
16	664	528
24	525	417
32	458	362
48	323	263

MVAPICH2-X: OSU Benchmark results on Comet





- **OSU OpenSHMEM Put and Get Tests (v5.0)**
- **Put latency – 1.42*μ*s, Get latency – 1.76*μ*s.**

MVAPICH2-X Results
OSU Atomics Benchmark

# Operation	Million ops/s	Latency (us)
shmem_int_fadd	0.37	2.68
shmem_int_finc	0.38	2.60
shmem_int_add	0.40	2.51
shmem_int_inc	0.40	2.48
shmem_int_cswap	0.38	2.66
shmem_int_swap	0.39	2.58

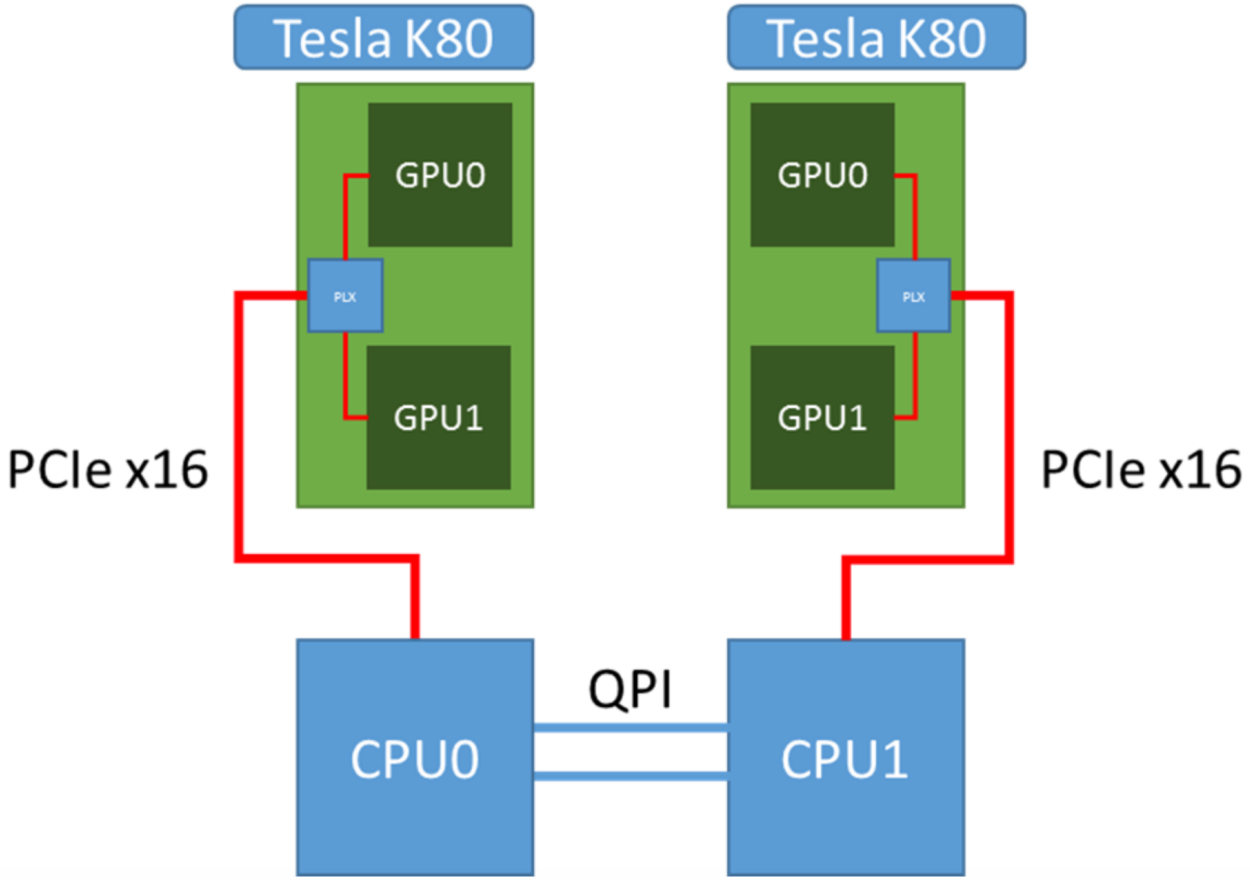
GPU Nodes – MVAPICH2-GDR

- **MVAPICH2-GDR takes advantage of GPUDirect RDMA technology on NVIDIA GPUs nodes with Mellanox InfiniBand interconnect.**

- **Available via "module load mvapich2-gdr" on the GPU nodes.**

- **Info on features at:**
  - http://mvapich.cse.ohio-state.edu/features/#mv2gdr

GPU Node Architecture



MVAPICH2-GDR – Sample Run

```
mpirun_rsh -np 2 –hostfile $HFILE MV2_USE_CUDA=1 MV2_CPU_MAPPING=13
MV2_USE_GPUDIRECT_GDRCOPY=1
MV2_GPUDIRECT_GDRCOPY_LIB=/opt/nvidia/gdrcopy/lib/libgdrapi.so  ./osu_latency D D
# OSU MPI-CUDA Latency Test
# Send Buffer on DEVICE (D) and Receive Buffer on DEVICE (D)
# Size          Latency (us)
0                    1.37
1                    2.86
2                    2.88
4                    2.85
8                    2.86
…
….

1048576              267.51
2097152              443.15
4194304              802.74
```

HOOMD-blue Benchmark using MVAPICH2-GDR

- HOOMD-blue is a *general-purpose* particle simulation toolkit
- **Lennard-Jones liquid benchmark with** *N*=64000, *rho*=0.382, *r_cut=3.0.*

References:

- HOOMD-blue web page: http://codeblue.umich.edu/hoomd-blue
- 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 Benchmark results (preliminary)

Lennard-Jones liquid benchmark with *N*=64000, *rho*=0.382, *r_cut=3.0.*

Number of GPUs	Particle timesteps/s
2 (On same node)	68.74 million
4 (1 per node)	78.93 million
4 (All on one node)	80.39 million

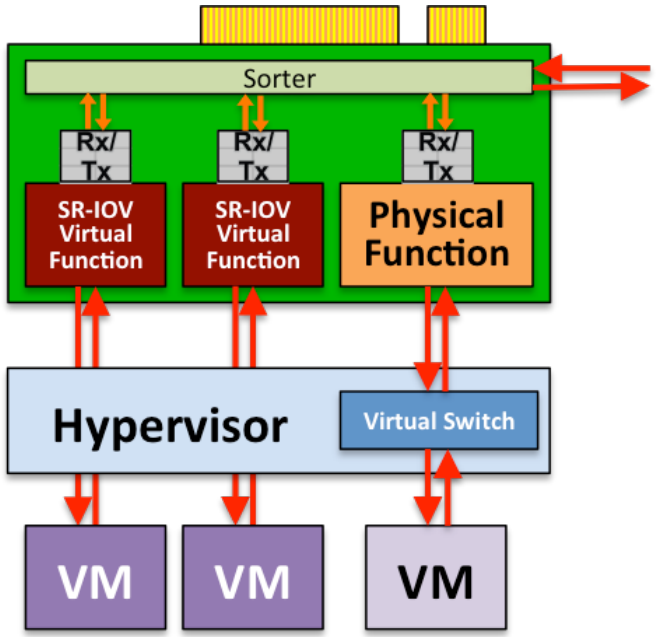
Hadoop-RDMA

Network-Based Computing Lab (Ohio State University)

- **HDFS, MapReduce, and RPC over native InfiniBand and RDMA over Converged Ethernet (RoCE).**
- **Based on Apache Hadoop.**
- **Works with the SLURM scheduler.**
- **Version RDMA-Apache-Hadoop-2.x 0.9.7 available on Comet:**
  - **/share/apps/compute/hadoop**
- **More details :**
  - http://hibd.cse.ohio-state.edu/

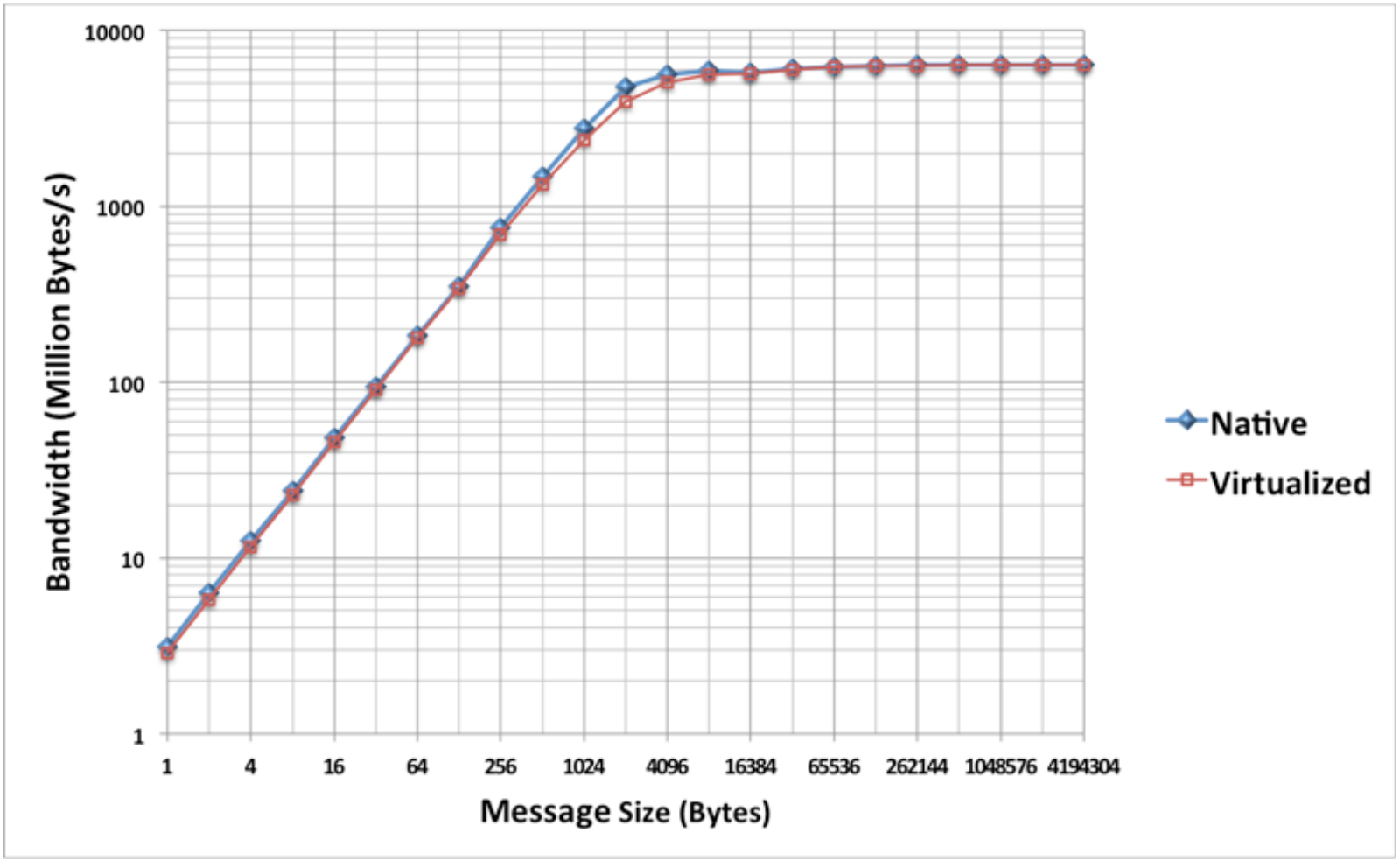
Single Root I/O Virtualization in HPC

- **Problem**: Virtualization generally has resulted in significant I/O performance degradation (e.g., excessive DMA interrupts)
- **Solution**: SR-IOV and Mellanox ConnectX-3 InfiniBand host channel adapters
  - One physical function → multiple virtual functions, each light weight but with its own DMA streams, memory space, interrupts
  - Allows DMA to bypass hypervisor to VMs
- ***SRIOV enables virtual HPC cluster w/ near-native InfiniBand latency/bandwidth and minimal overhead***

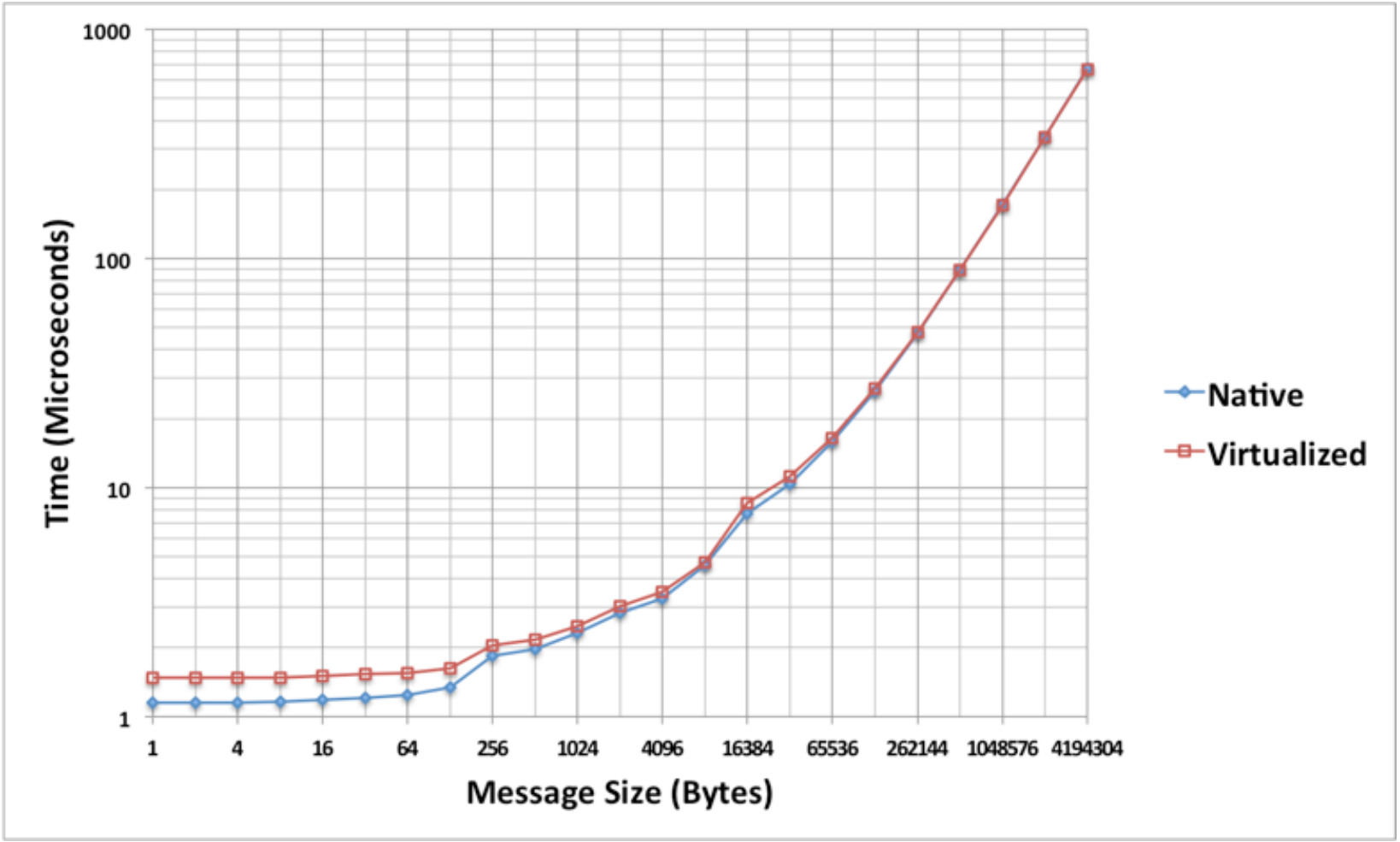


Mellanox TECHNOLOGIES

Comet: MPI bandwidth slowdown from SR-IOV is at most 1.21 for medium-sized messages & negligible for small & large ones



Comet: MPI latency slowdown from SR-IOV is at most 1.32 for small messages & negligible for large ones

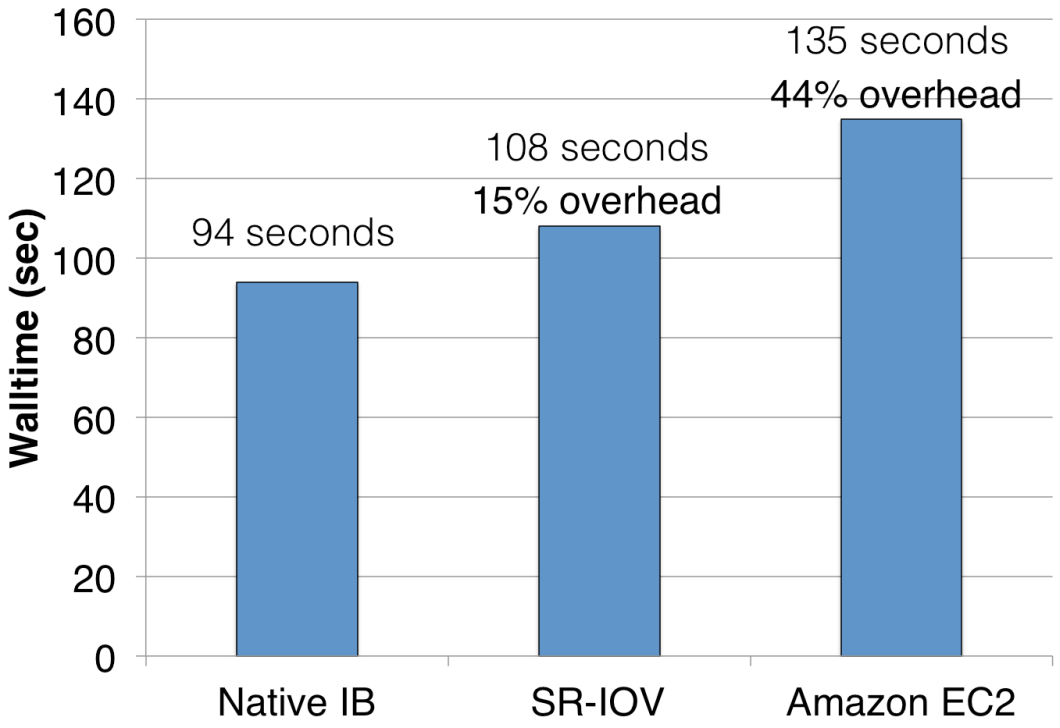


Application Benchmarks: WRF

- **WRF CONUS-12km benchmark. The domain is 12 KM horizontal resolution on a 425 by 300 grid with 35 vertical levels, with a time step of 72 seconds.**
- **Originally run using six nodes (96 cores) over QDR4X InfiniBand virtualized with SR-IOV. Now have Comet results.**
- **SR-IOV test cluster has 2.2 GHz Intel(R) Xeon E5-2660 processors.**
- **Amazon instances were using 2.6 GHz Intel(R) Xeon E5-2670 processors.**

Weather Modeling (WRF)

- 96-core (6-node) calculation
- Nearest-neighbor communication
- Scalable algorithms
- SR-IOV incurs modest (15%) performance hit
- …but still still 20% faster*** than Amazon
- Comet (new results): The overhead for this case is 2%!

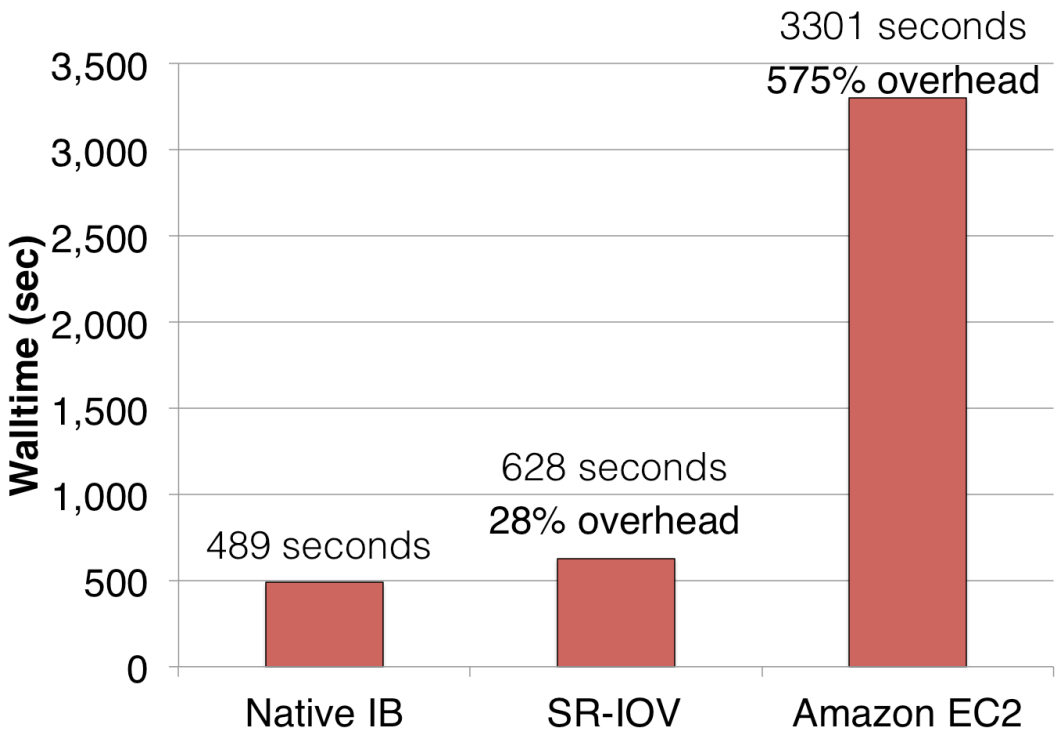


WRF 3.4.1 – 3hr forecast

*** 20% faster despite SR-IOV cluster having 20% slower CPUs

Quantum ESPRESSO: 5x Faster than EC2

- 48-core, 3 node calc
- CG matrix inversion (irregular comm.)
- 3D FFT matrix transposes (All-to-all communication)
- 28% slower w/ SR-IOV
- SR-IOV still > 500% faster*** than EC2
- Comet (new results): SR-IOV is only 8% slower than the native result!



Quantum Espresso 5.0.2 – DEISA AUSURF112 benchmark
*** 20% faster despite SR-IOV cluster having 20% slower CPUs

Summary

- **Comet architecture geared to support a wide range of applications and diverse user base.**
- **MVAPICH2, MVAPICH2-X, MVAPICH2-GDR, and *MVAPICH2-Virt* will enable users to leverage the various features with efficiency.**
- **Good performance achieved on a diverse array of applications.**

Thanks!
Questions: Email mahidhar@sdsc.edu