HPC ADVISORY COUNCIL

MILC
Performance Benchmark and Profiling

August 2012

AMD
The future is fusion

DELL

Mellanox
TECHNOLOGIES

Note

- **The following research was performed under the HPC Advisory Council activities**
  - Participating vendors: AMD, Dell, Mellanox
  - Compute resource - HPC Advisory Council Cluster Center
- **For more info please refer to**
  - http://www.amd.com
  - http://www.dell.com/hpc
  - http://www.mellanox.com
  - http://www.physics.utah.edu/~detar/milc/

MILC

- **MILC (MIMD Lattice Computation) QCD Code**
  - Developed by the MIMD Lattice Computation (MILC) collaboration
  - Performs large scale numerical simulations to study quantum chromodynamics (QCD)
  - QCD is the theory of the strong interactions of subatomic physics
  - Simulates 4-dimensional SU(3) lattice gauge theory on MIMD parallel machines
  - Contains a set of codes written in C.
  - Publicly available for research purposes
  - Free, open-source code, distributed under GNU General Public License

- **The MILC Collaboration**
  - Produced application codes to study several different QCD research areas
  - Is engaged in a broad research program in Quantum Chromodynamics (QCD)
  - Its research addresses fundamental questions in high energy and nuclear physics
  - Related to major experimental programs in the fields, including:
    - Studies of the mass spectrum of strongly interacting particles,
    - The weak interactions of these particles,
    - The behavior of strongly interacting matter under extreme conditions

Objectives

- **The following was done to provide best practices**
  - MILC performance benchmarking
  - Interconnect performance comparisons
  - Ways to increase MILC productivity
  - MPI libraries comparisons

- **The presented results will demonstrate**
  - The scalability of the compute environment to provide nearly linear application scalability
  - The capability of MILC to achieve scalable productivity
  - Considerations for performance optimizations

Test Cluster Configuration

- **Dell™ PowerEdge™ R815 11-node (704-core) cluster**
  - Memory: 128GB memory per node DDR3 1333MHz, BIOS version 2.8.2
  - 4 CPU sockets per server node
- **AMD™ Opteron™ 6276 (code name "Interlagos") 16-core @ 2.3 GHz CPUs**
- **Mellanox ConnectX®-3 VPI Adapters and IS5030 36-Port InfiniBand switch**
- **MLNX-OFED 1.5.3 InfiniBand SW stack**
- **OS: RHEL 6 Update 2, SLES 11 SP2**
- **MPI: Intel MPI 4 Update 3, Open MPI 1.5.5, Platform MPI 8.2.1**
- **Compilers: GNU 4.7**
- **Application: milc_qcd-7.7.8**
- **Benchmark workload: n6_256.in (Global lattice size of 32x32x32x36)**
  - Input dataset from NERSC: http://www.nersc.gov/research-and-development/benchmarking-and-workload-characterization/nersc-6-benchmarks/milc/

Dell™ PowerEdge™ R815 11-node cluster

- **HPC Advisory Council Test-bed System**

- **New 11-node 704 core cluster - featuring Dell PowerEdge™ R815 servers**
  - Replacement system for Dell PowerEdge SC1435 (192 cores) cluster system following 2 years of rigorous benchmarking and product EOL
    - System to be redirected to explore HPC in the Cloud applications

- **Workload profiling and benchmarking**
  - Characterization for HPC and compute intense environments
  - Optimization for scale, sizing and configuration and workload performance
  - Test-bed Benchmarks
    - RFPs
    - Customers/Prospects, etc
  - ISV & Industry standard application characterization
  - Best practices & usage analysis

About Dell PowerEdge™ Platform Advantages

Best of breed technologies and partners

Combination of AMD Opteron™ 6200 series platform and Mellanox ConnectX®-3 InfiniBand on Dell HPC

Solutions provide the ultimate platform for speed and scale

- Dell PowerEdge R815 system delivers 4 socket performance in dense 2U form factor
- Up to 64 core/32DIMMs per server – 1344 core in 42U enclosure

Integrated stacks designed to deliver the best price/performance/watt

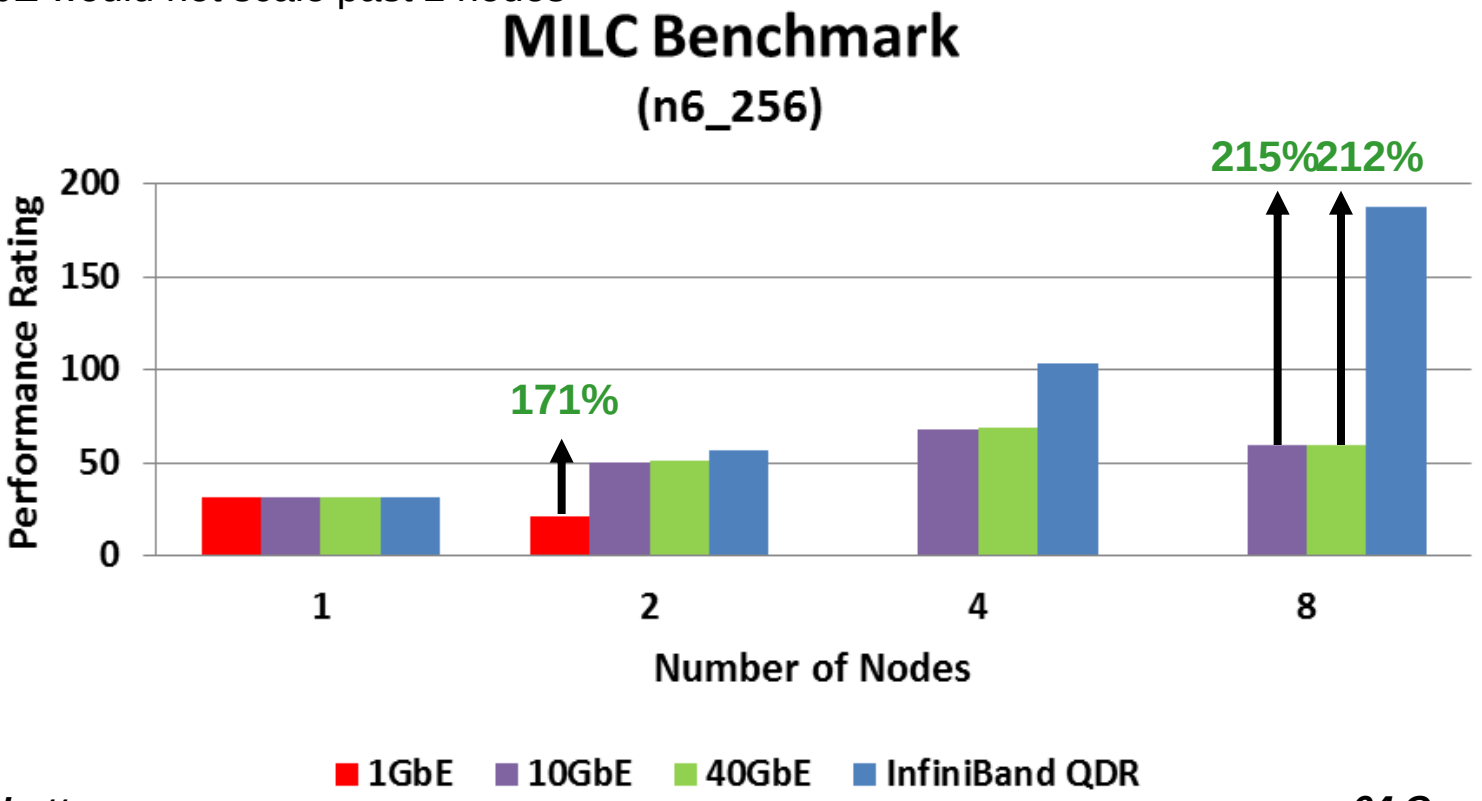
- 2x more memory and processing power in half of the space
- Energy optimized low flow fans, improved power supplies and dual SD modules

Optimized for long-term capital and operating investment protection

- System expansion
- Component upgrades and feature releases

MILC Performance – Interconnects

- **InfiniBand QDR delivers the highest network performance and scalability**
  - InfiniBand QDR outperforms 1GbE by 171% at 2 nodes
  - InfiniBand QDR outperforms 10GbE and 40GbE by over 212% at 8 nodes
  - 10GbE and 40GbE scalability is limited after 4 nodes
  - 1GbE would not scale past 2 nodes

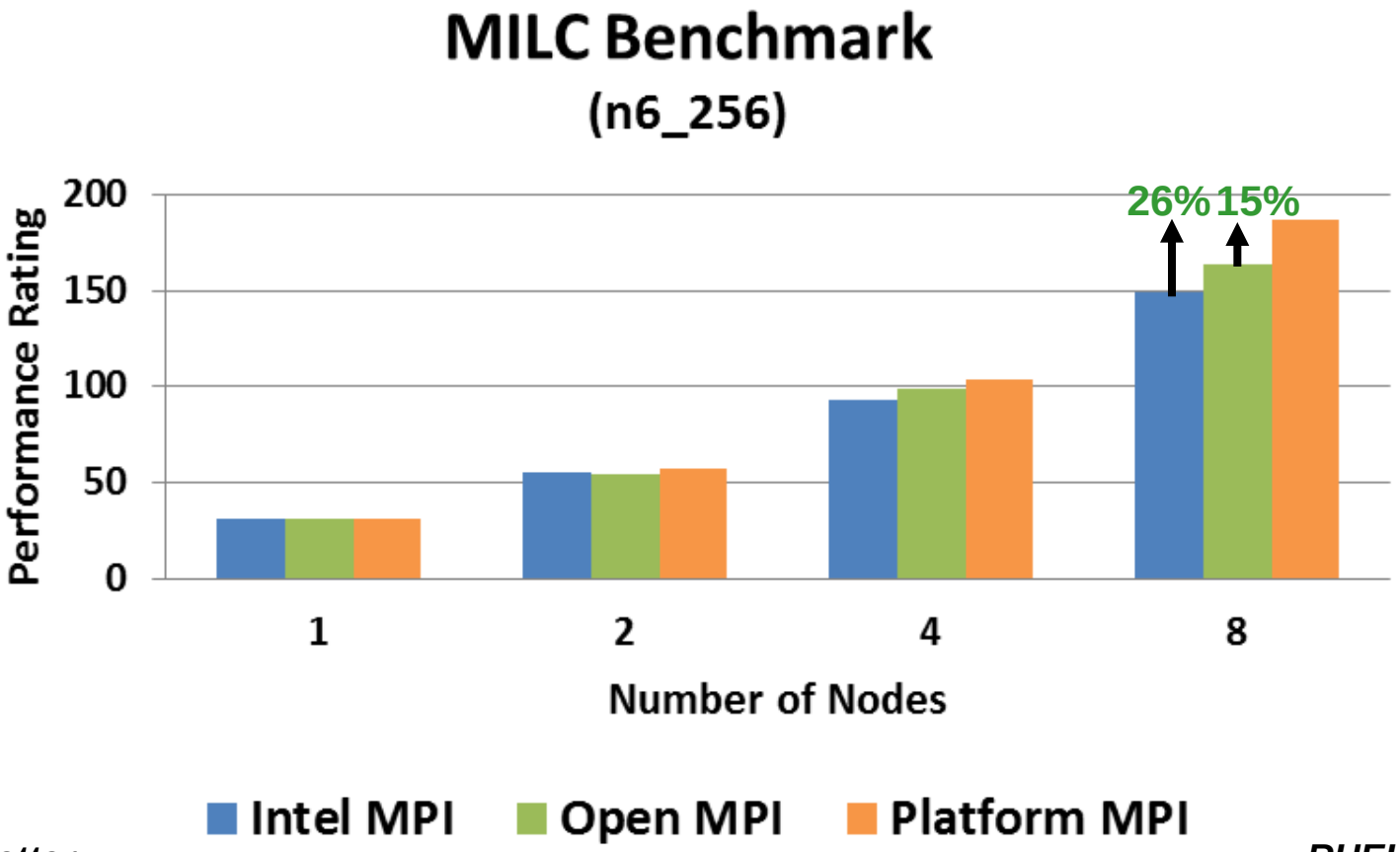


Higher is better

64 Cores/Node

MILC Performance – MPI

- **Platform MPI performs the best among the MPI tested**
  - Up to 26% higher performance than Intel MPI at 8 nodes (512 processes)
  - Up to 15% higher performance than Open MPI at 8 nodes (512 processes)

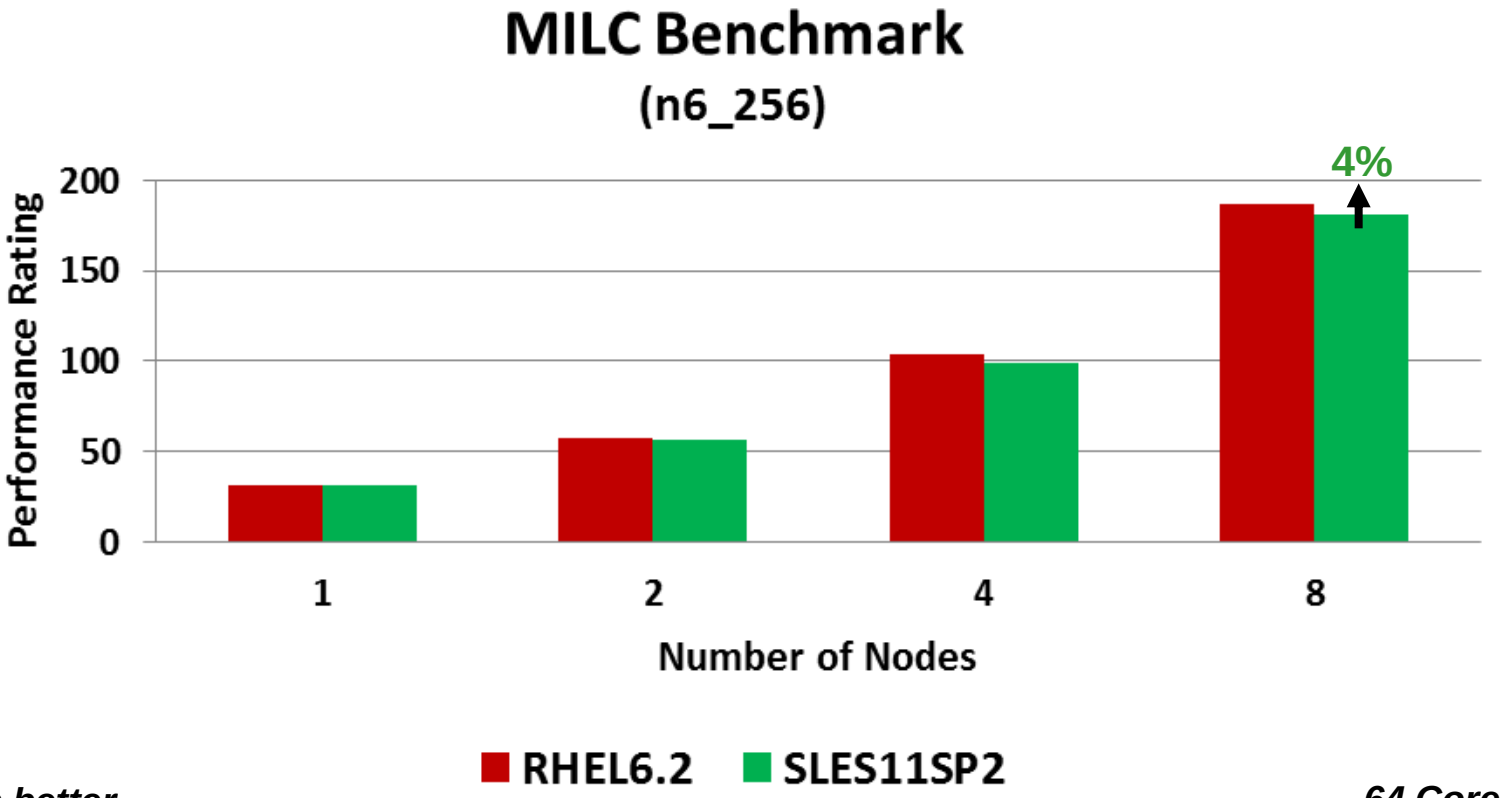


Higher is better

RHEL 6 U2

MILC Performance – Operating Systems

- **RHEL performs slightly better than SLES**
  - RHEL 6 Update 2 performs about 4% better than SLES 11 SP2

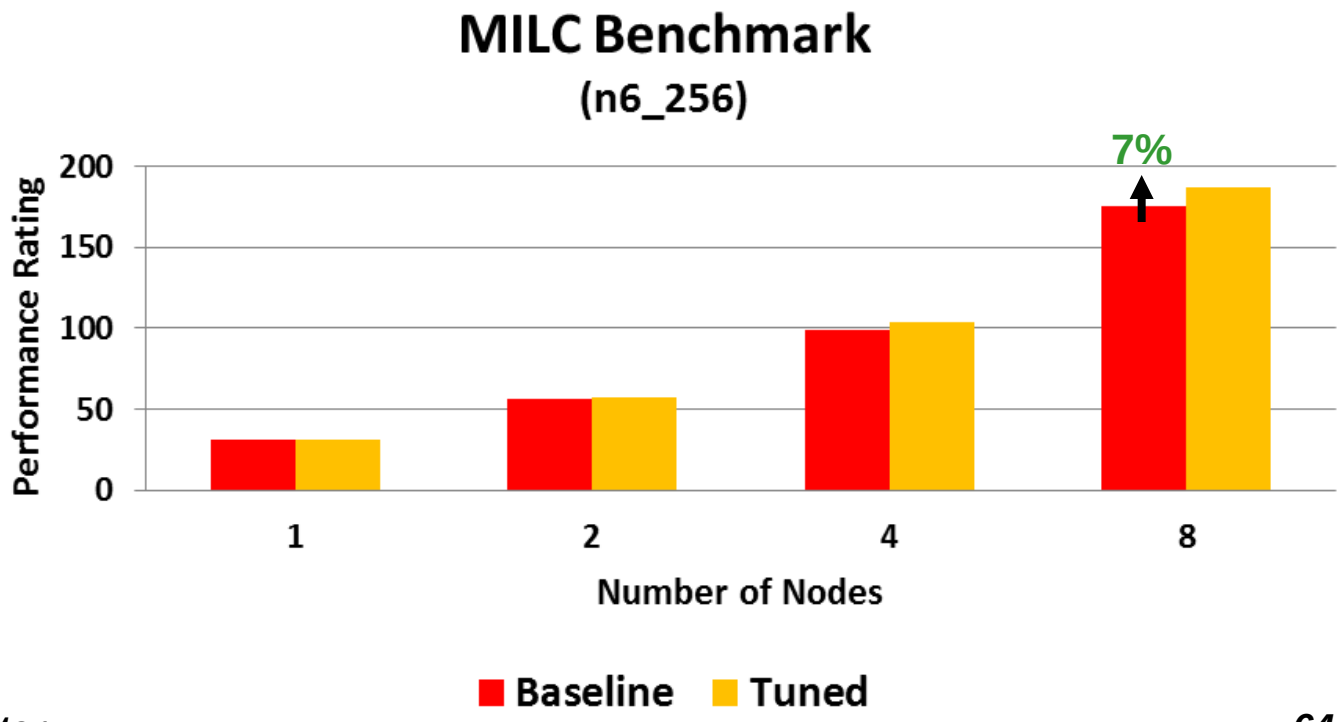


Higher is better

64 Cores/Node

MILC Performance – Compiler Flags

- **Improvement in performance with additional compiler flags**
  - Seen up to 7% better performance with Bulldozer specific flags specified
  - Baseline: OCFLAGS=-O3 -m64
  - Tuned: OCFLAGS=-O3 -m64 **-march=bdver1 -mavx -mfma4**

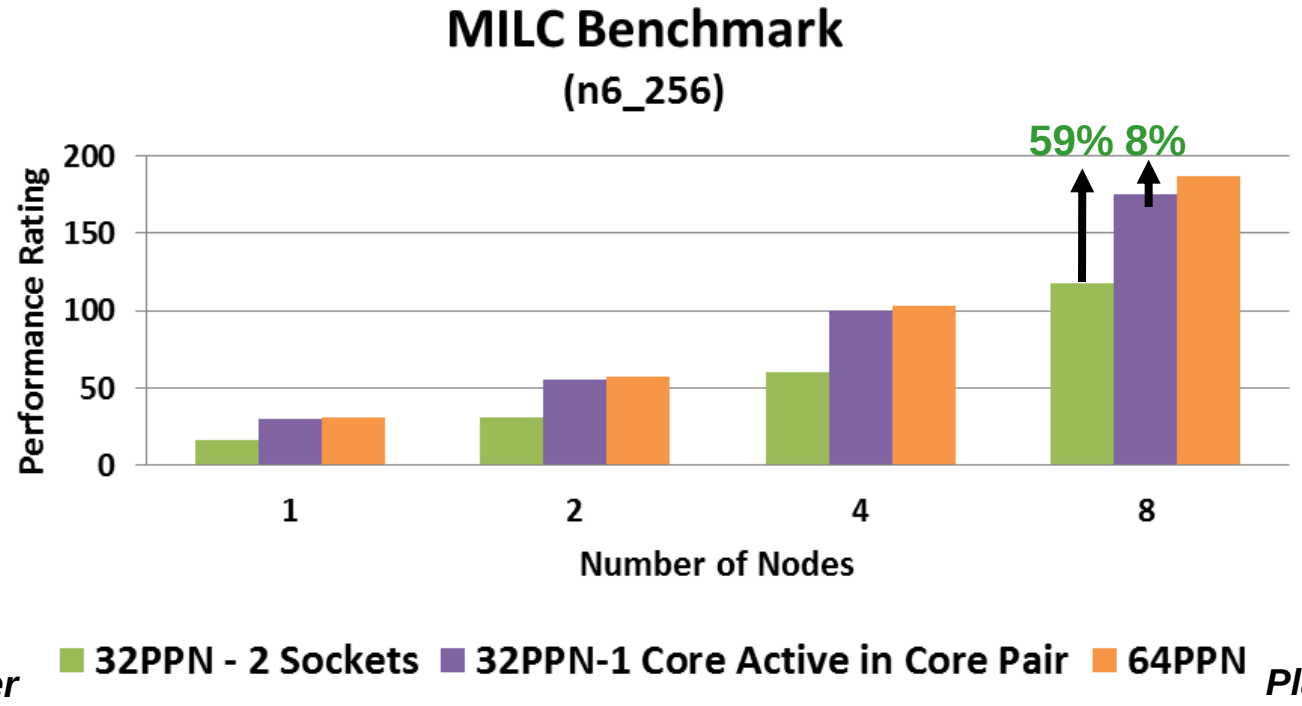


Higher is better

64 Cores/Node

MILC Performance – CPU Cores Placement

- **Comparing jobs running with 32 PPN versus 64 PPN (processes per node)**
  - Running with 4 CPUs (64PPN) is 59% faster than jobs running with 2 CPUs (32 PPN)
  - The 32 PPN case uses 2 CPU sockets while the 64 PPN case uses 4 CPU sockets
- **CPU core clock frequency increases when only 1 core in each core pair is active**
  - While the non-active core is in sleep mode
  - Running with both cores is 8% faster than running with only 1 active core in a core pair

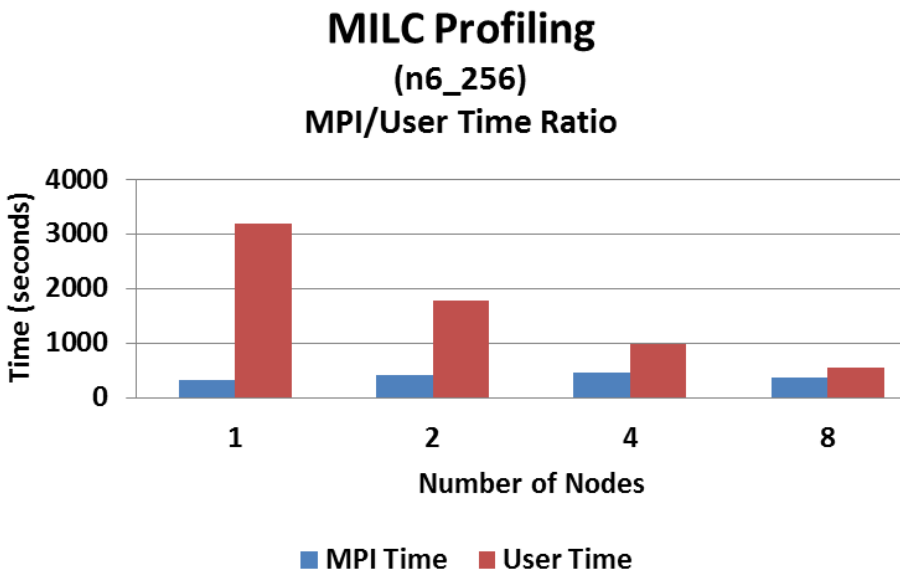


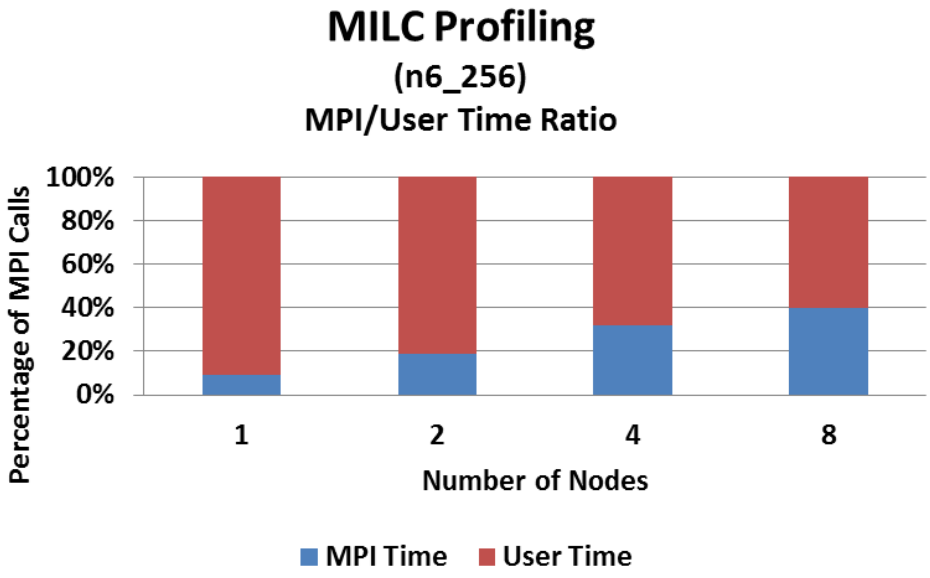
Higher is better

Platform MPI

MILC Profiling – MPI/User Time Ratio

- **Computation time halves as compute node doubles**
  - With the communication time stays roughly constant
  - Good scalability is seen as compute workload is roughly halved without adding much MPI time



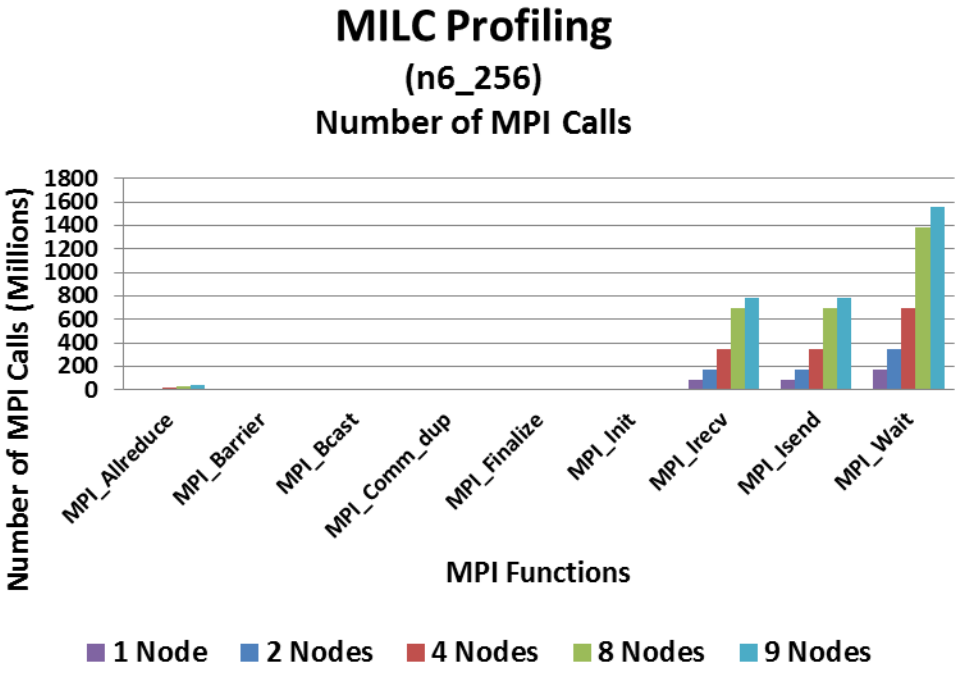


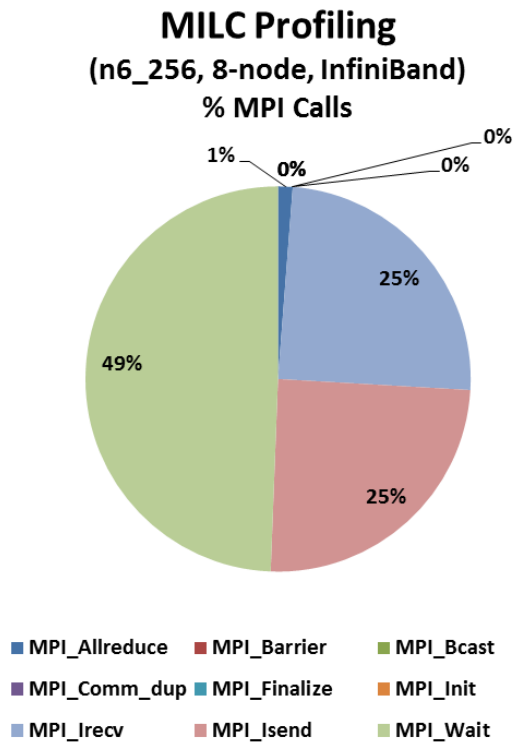
Higher is better

64 Cores/Node

MILC Profiling – Number of MPI Calls

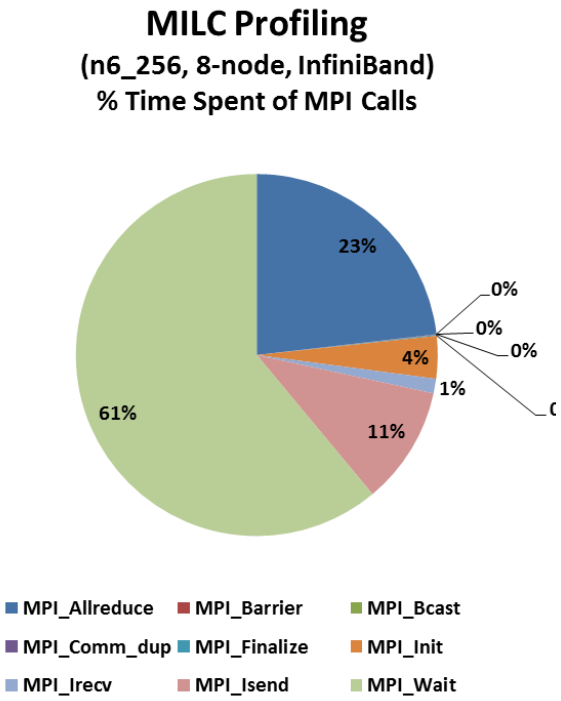
- **The most used MPI function are MPI_Wait, MPI_Isend, MPI_Irecv**
  - MPI_Wait accounts for 49% of all the MPI calls made
  - MPI_Isend and MPI_Irecv accounts for 25% of all MPI calls made
- **Non-blocking MPI communications are used heavily**

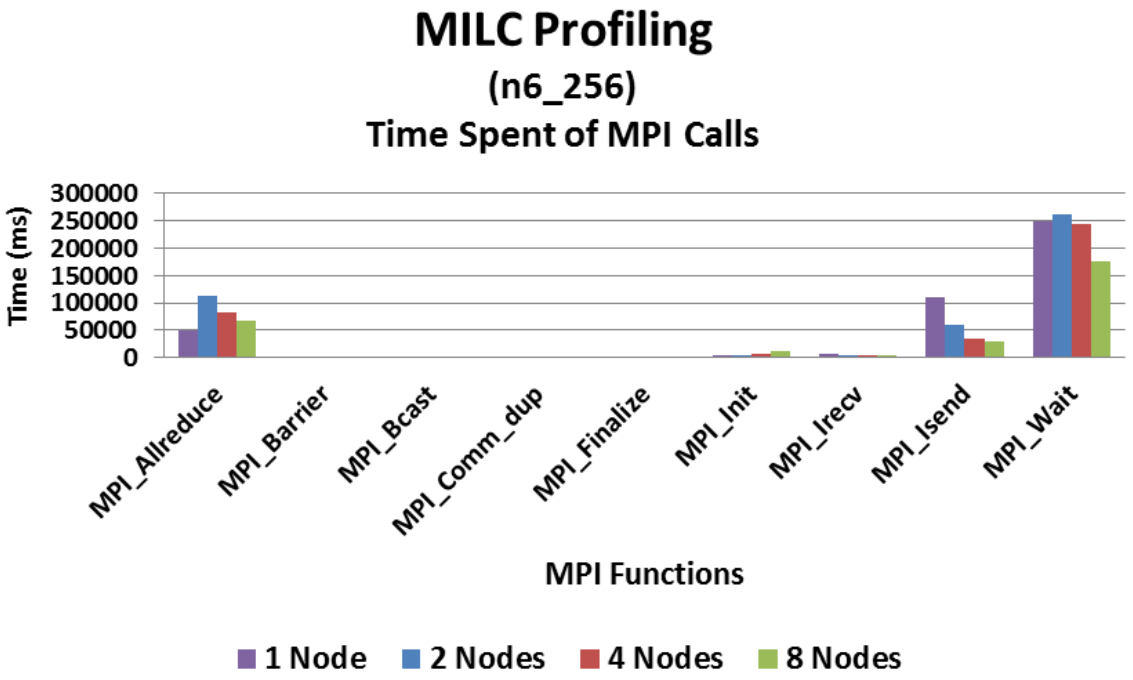




MILC Profiling – Time Spent of MPI calls

- **The most time consuming MPI function is MPI_Wait**
  - MPI_Wait accounts for 61% of all MPI time at 8 nodes
  - MPI_Allreduce accounts for 23% of all MPI time at 8 nodes
  - MPI_Isend accounts for 11% of MPI time at 8 nodes

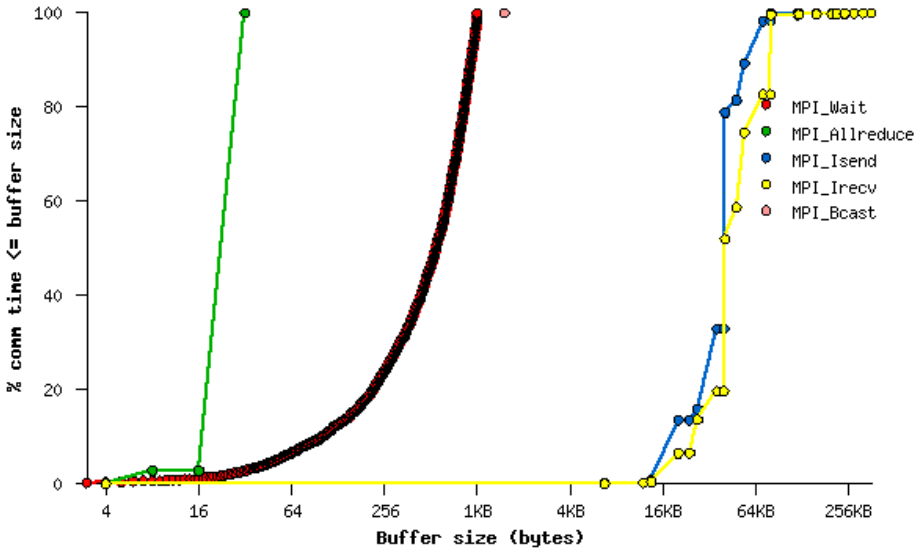




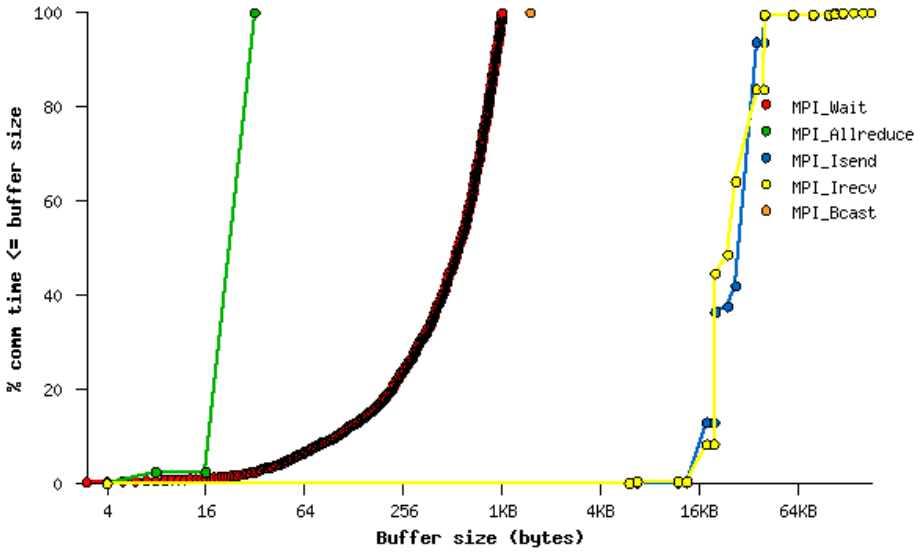
MILC Profiling –MPI Message Sizes

- **Majority of the MPI message sizes are concentrated in the midrange**
  - Message sizes between 256B to 1KB for MPI_Wait
  - Midrange data communications shift down from ~64KB to ~16KB as cluster scales

2 Nodes (128 processes)



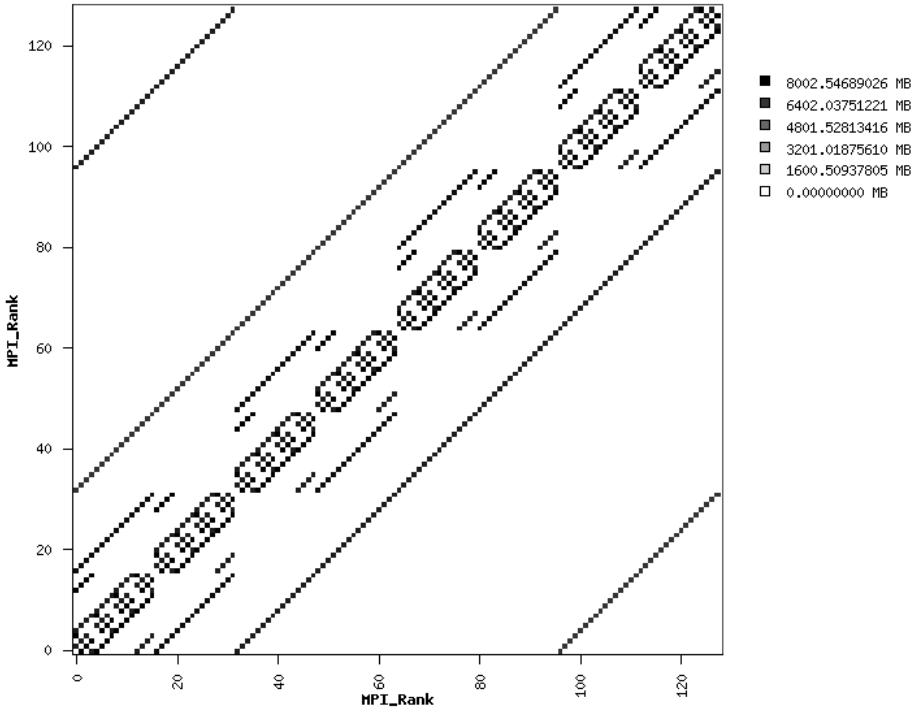
4 Nodes (256 Processes)



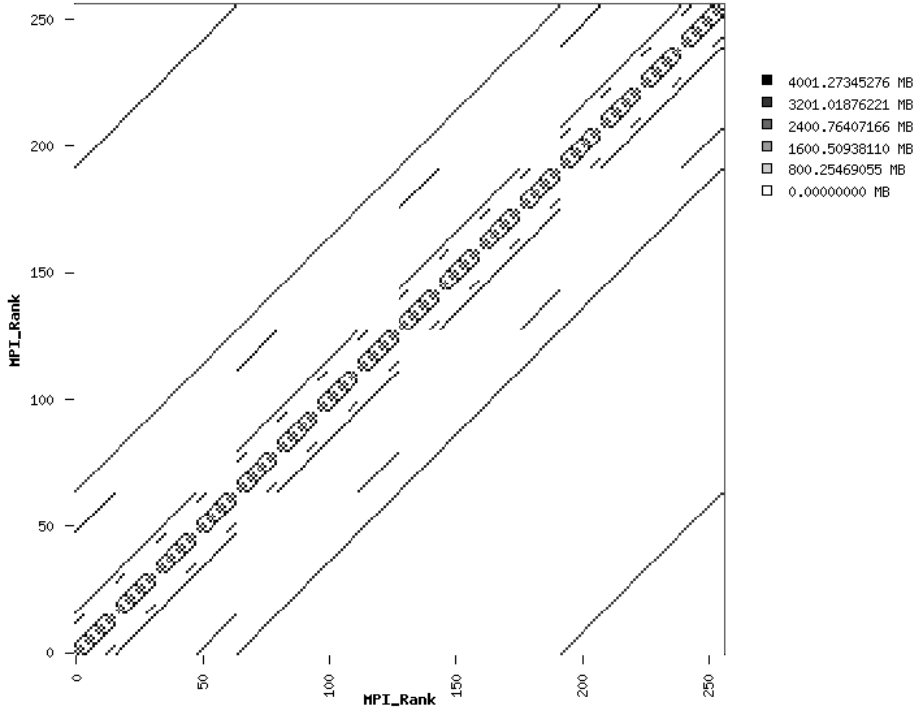
MILC Profiling – Data Transfer / Process

- **As the cluster scales, less data is driven to each rank and each node**
  - Drops from maximum of 8GB per rank to 4GB per rank

2 Nodes

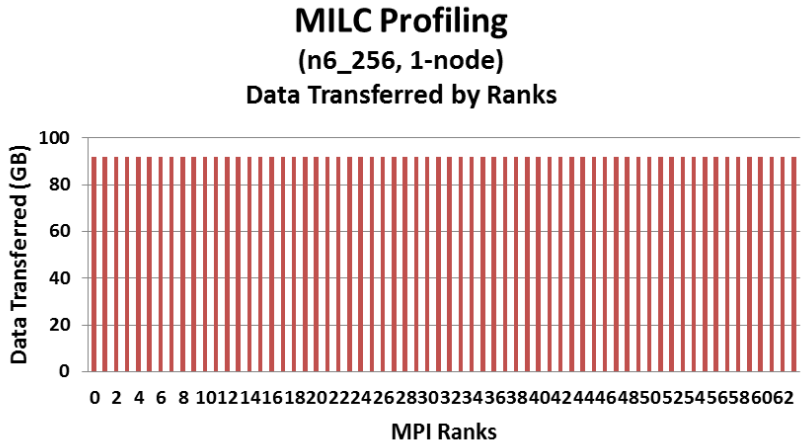


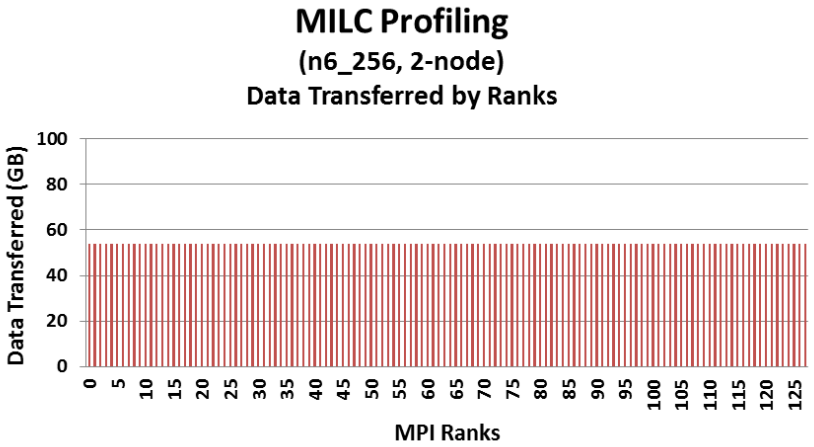
4 Nodes

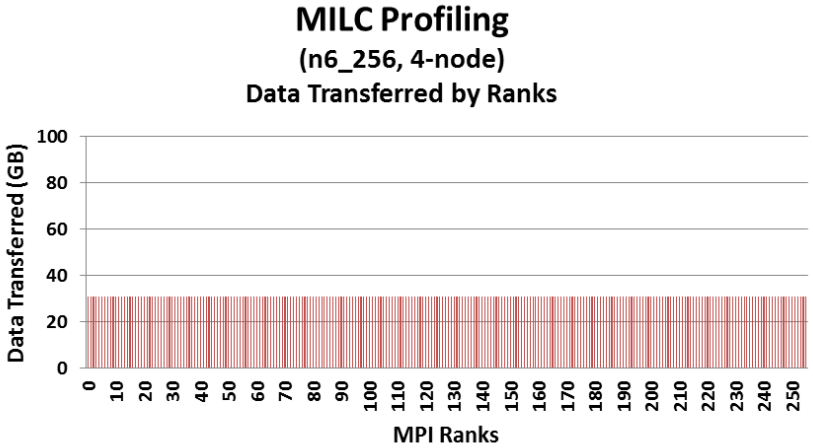


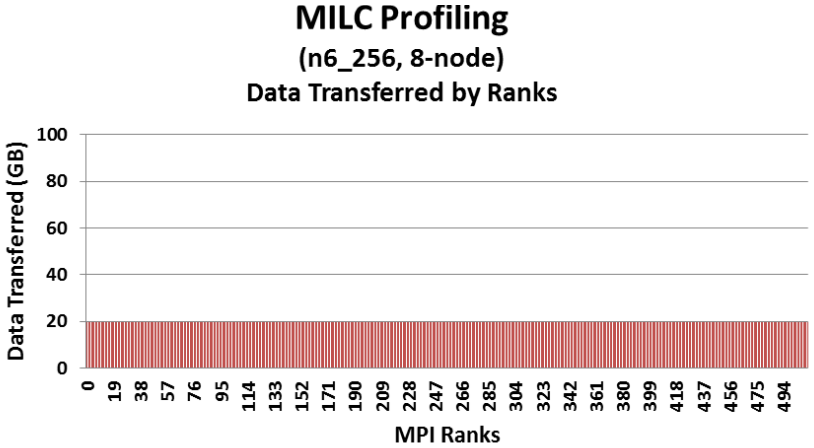
MILC Profiling – Data Transfer / Process

- **Data exchanges per rank is nearly halved as cluster size doubles**
  - The total amount of data stays close to constant as cluster scales



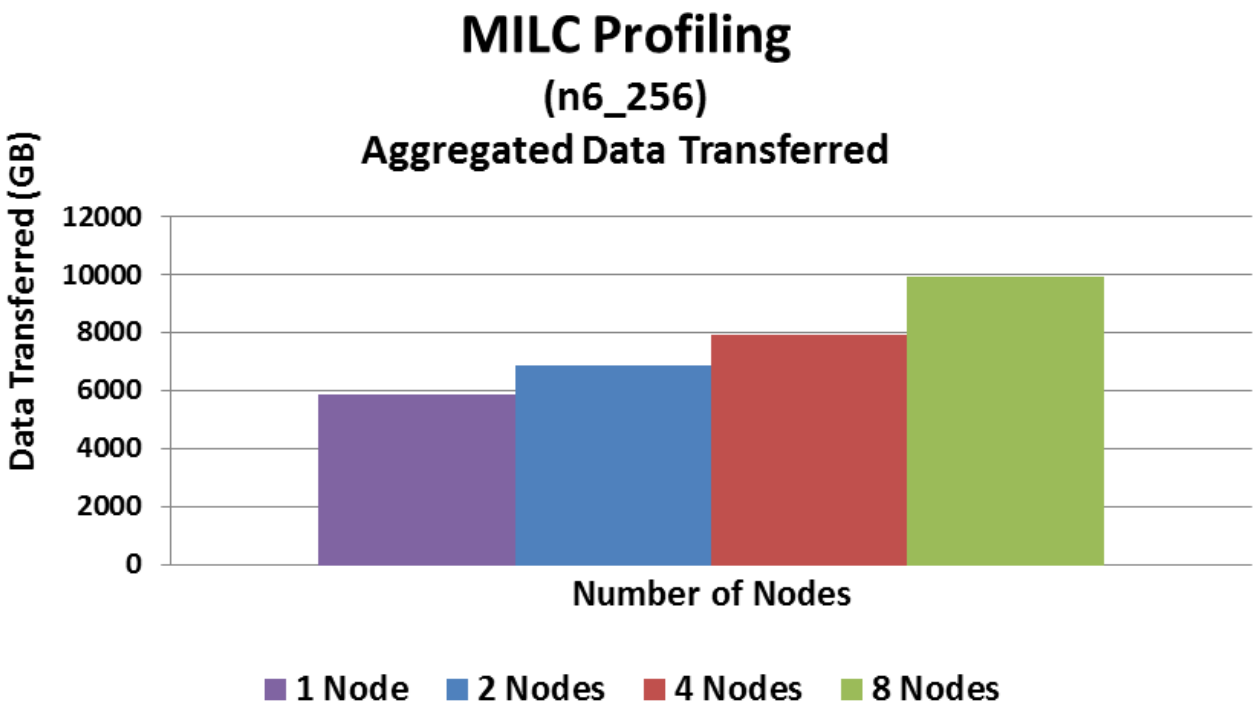






MILC Profiling – Aggregated Data Transfer

- **Aggregated data transfer refers to:**
  - Total amount of data being transferred in the network between all MPI ranks collectively
- **Small growth in data exchanges as the cluster scales**
  - Scalability can be achieved easier without the burden of introducing additional data



Summary

- **MILC demonstrates impressive scalability performance with balanced hardware**
- **CPU:**
  - Using system with 4 CPUs versus 2 CPUs provides 59% gain in productivity at 8 nodes
- **Network:**
  - InfiniBand QDR delivers the highest network performance and scalability for MILC
  - Outperforms over 212% than 10GbE and 40GbE at 8 nodes
  - 10GbE and 40GbE scalability limits after 4 nodes, and 1GbE would not scale past 2 nodes
- **OS:**
  - Running jobs in the RHEL provides slightly (4%) better performance than SLES
- **MPI:**
  - Platform MPI provides 26% performance than Intel MPI and 15% better than Open MPI
- **Compiler:**
  - Tuned flags for Bullozer architecture provides additional of 7% gain at 8 nodes
  - Compiler flags for AVX, FMA4 and Interlagos instructions: (-march=bdver1 -mavx –mfma4)

Thank You

HPC Advisory Council