



# GROMACS

## Performance Benchmark and Profiling

August 2011

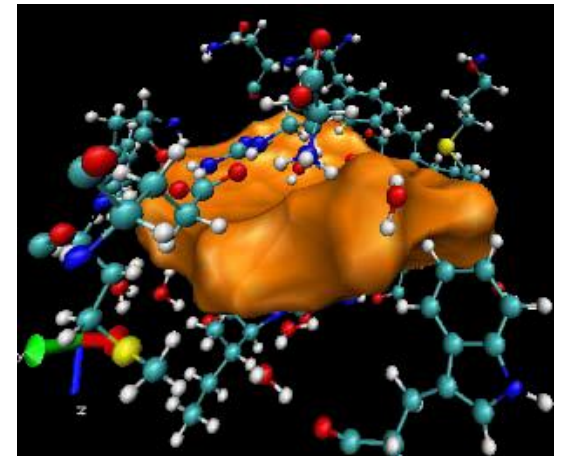
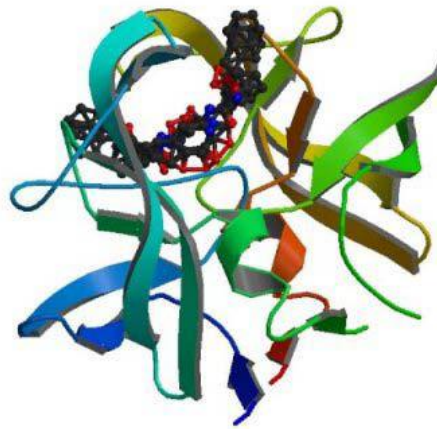
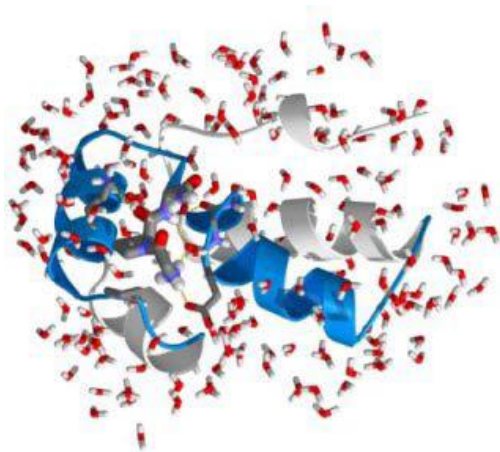


**GROMACS** FAST.  
FLEXIBLE.  
FREE.

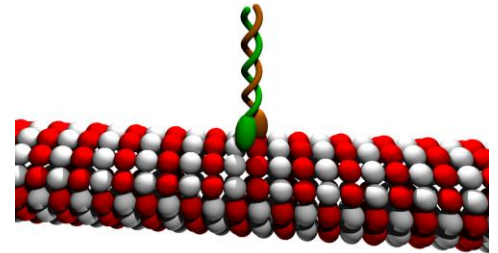


- **The following research was performed under the HPC Advisory Council activities**
  - Participating vendors: Intel, Dell, Mellanox
  - Compute resource - HPC Advisory Council Cluster Center
- **The following was done to provide best practices**
  - GROMACS performance overview
  - Understanding GROMACS communication patterns
  - Ways to increase GROMACS productivity
  - MPI libraries comparisons
- **For more info please refer to**
  - <http://www.dell.com>
  - <http://www.intel.com>
  - <http://www.mellanox.com>
  - <http://www.gromacs.org>

- **GROMACS (GRoningen MACHine for Chemical Simulation)**
  - A molecular dynamics simulation package
  - Primarily designed for biochemical molecules like proteins, lipids and nucleic acids
    - A lot of algorithmic optimizations have been introduced in the code
    - Extremely fast at calculating the nonbonded interactions
  - Ongoing development to extend GROMACS with interfaces both to Quantum Chemistry and Bioinformatics/databases
  - An open source software released under the GPL



- **Dell™ PowerEdge™ M610 38-node (456-core) cluster**
  - Six-Core Intel X5670 @ 2.93 GHz CPUs, Six-Core Intel X5675 @ 3.06 GHz GPUs
  - Memory: 24GB memory, DDR3 1333 MHz
  - OS: RHEL 5.5, OFED 1.5.2 InfiniBand SW stack
- **Intel Cluster Ready certified cluster**
- **Mellanox ConnectX-2 InfiniBand adapters and non-blocking switches**
- **Compiler: Intel Composer XE 2011 for Linux**
- **MPI: Intel MPI 4 Update 2, Open MPI 1.5.4 with KNEM 0.9.7, Platform MPI 8.1.1**
- **Math libraries: Intel MKL 10.3 Update 5**
- **Application: GROMACS 4.5.4**
- **Benchmark datasets:**
  - Kinesin Dimer docked to the MicroTubule (309,453 atoms, 20000 steps, 80.0 ps)



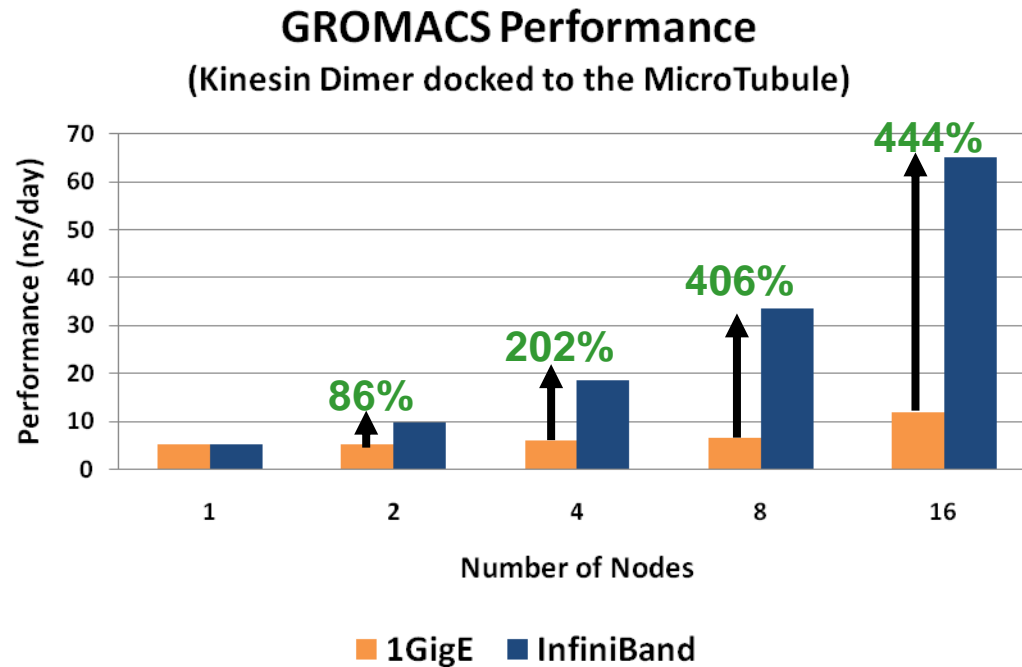
- **Intel® Cluster Ready systems make it practical to use a cluster to increase your simulation and modeling productivity**
  - Simplifies selection, deployment, and operation of a cluster
- **A single architecture platform supported by many OEMs, ISVs, cluster provisioning vendors, and interconnect providers**
  - Focus on your work productivity, spend less management time on the cluster
- **Select Intel Cluster Ready**
  - Where the cluster is delivered ready to run
  - Hardware and software are integrated and configured together
  - Applications are registered, validating execution on the Intel Cluster Ready architecture
  - Includes Intel® Cluster Checker tool, to verify functionality and periodically check cluster health



- **System Structure and Sizing Guidelines**
  - 38-node cluster build with Dell PowerEdge™ M610 blade servers
  - Servers optimized for High Performance Computing environments
  - Building Block Foundations for best price/performance and performance/watt
- **Dell HPC Solutions**
  - Scalable Architectures for High Performance and Productivity
  - Dell's comprehensive HPC services help manage the lifecycle requirements.
  - Integrated, Tested and Validated Architectures
- **Workload Modeling**
  - Optimized System Size, Configuration and Workloads
  - Test-bed Benchmarks
  - ISV Applications Characterization
  - Best Practices & Usage Analysis



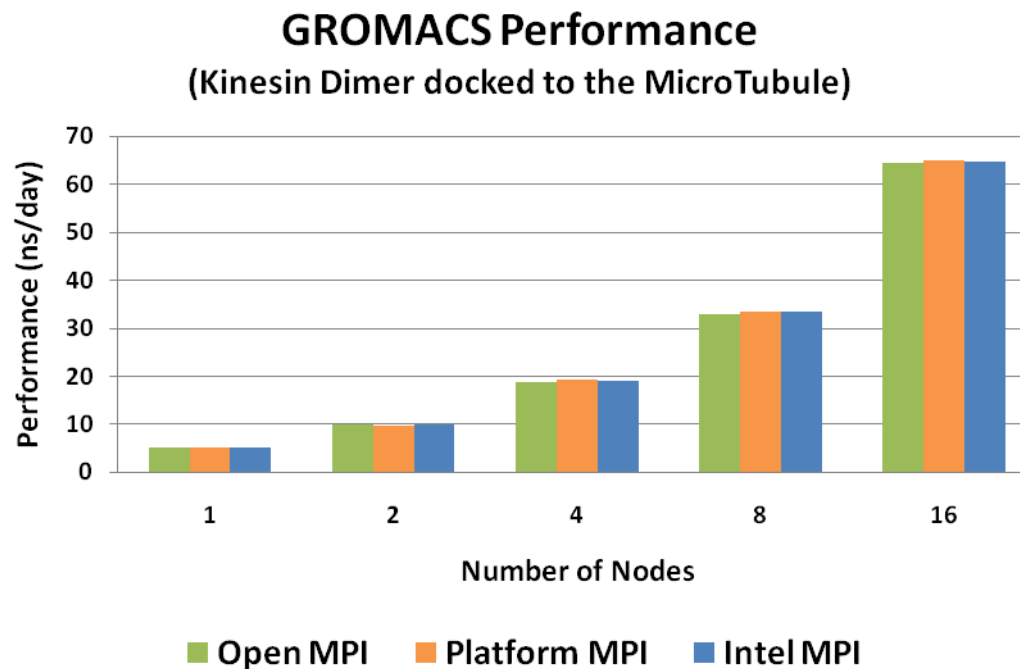
- **InfiniBand enables higher cluster productivity**
  - Increasing the performance by up to 444% over 1GigE
  - Performance benefits begins with 2 nodes
  - 1GigE shows little gain in performance despite more systems are used



*Higher is better*

*InfiniBand QDR*

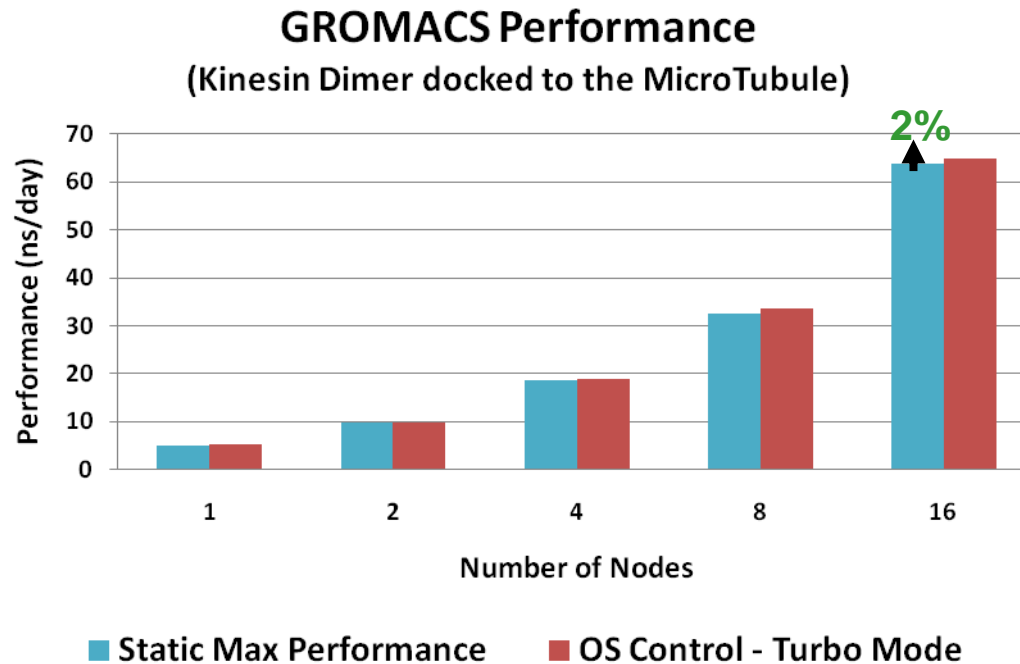
- **All MPI performs similarly in performance**
  - Reflects that each MPI implementation handles efficiently for the MPI data transfers



*Higher is better*

*InfiniBand QDR*

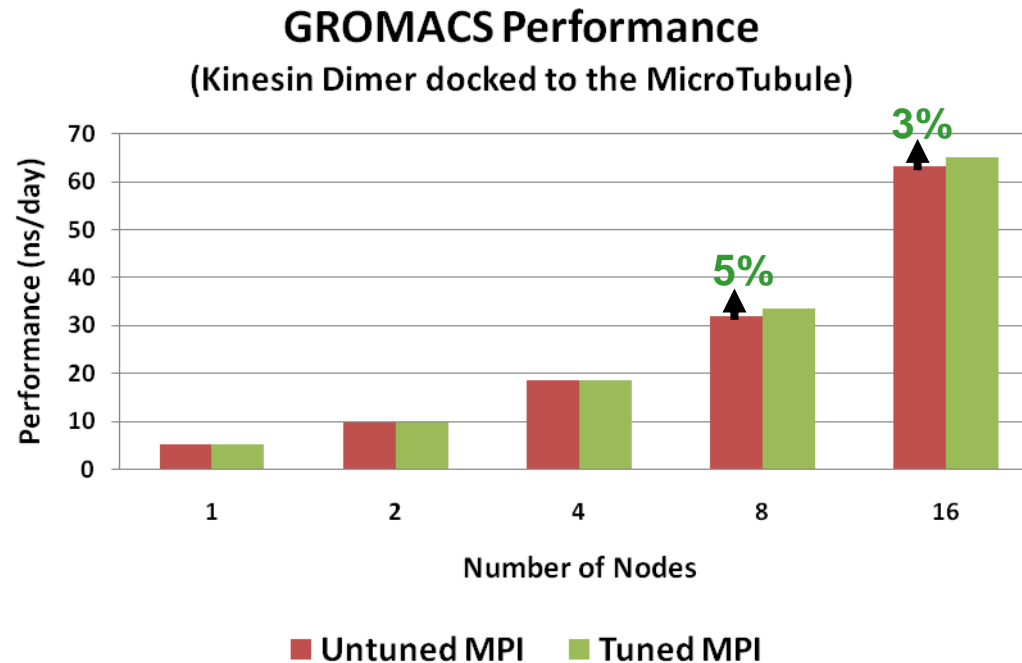
- **Setting Turbo Mode would allow the most optimal performance**
  - Seen an advantage of around 2% over the static Maximum Performance setting
  - OS controls CPU frequencies through the ACPI power management



*Higher is better*

*InfiniBand QDR*

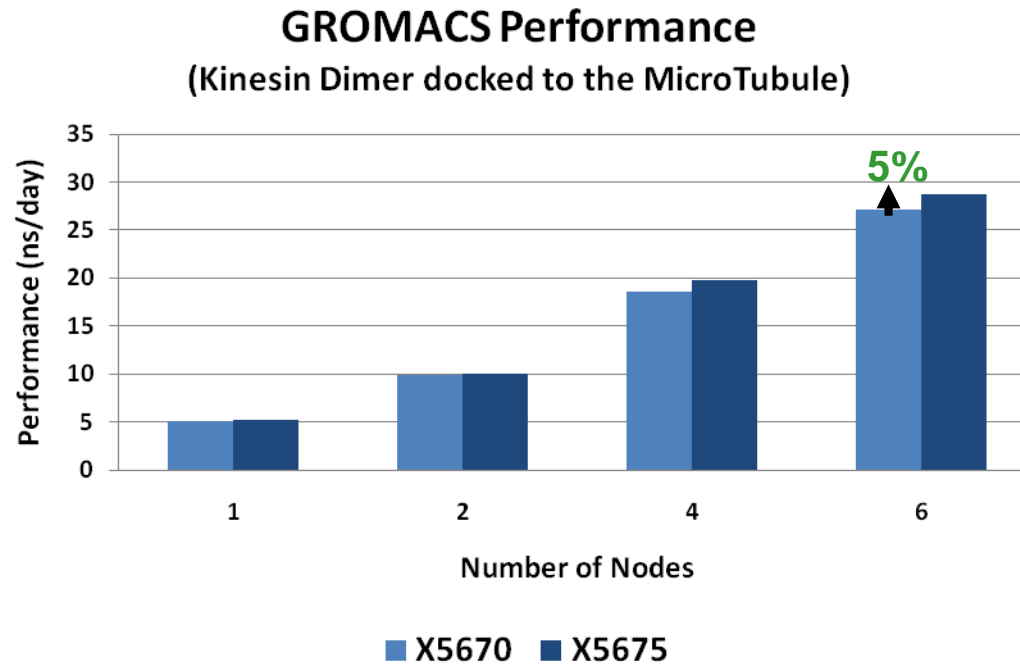
- **Tuning MPI would provide marginally better in speedup**
  - Seen a speedup gain of 3% to 5%
- **Tuning parameters used:**
  - I\_MPI\_SPIN\_COUNT=1 I\_MPI\_RDMA\_TRANSLATION\_CACHE=1  
I\_MPI\_RDMA\_RNDV\_BUF\_ALIGN=65536 I\_MPI\_SPIN\_COUNT=121  
I\_MPI\_DAPL\_DIRECT\_COPY\_THRESHOLD=65536



*Higher is better*

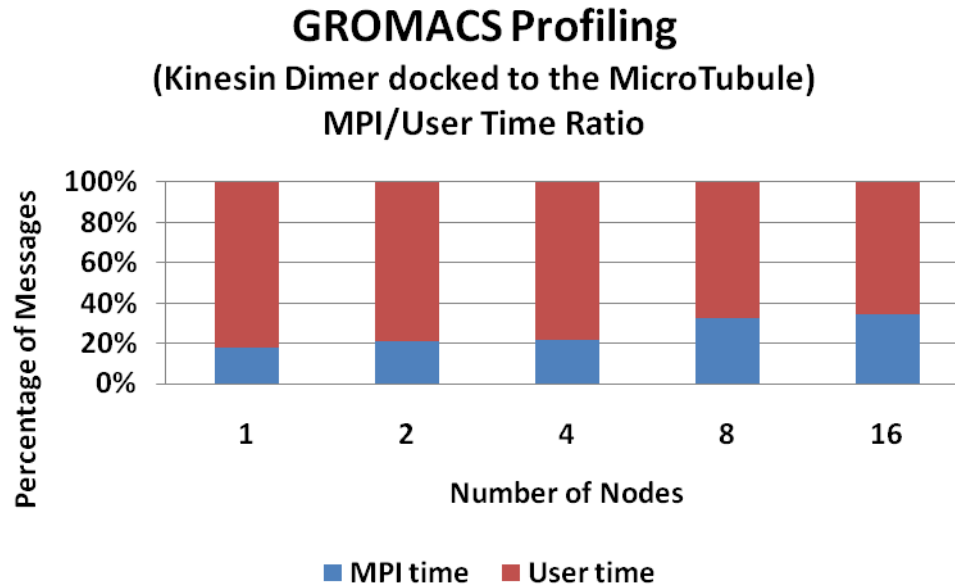
*InfiniBand QDR*

- **MPI communication percentage increases as the cluster scales**
  - Seen a 5% gain by using faster CPU at 6-node, comparing X5670 versus X5675
  - Accounts for the difference in CPU speeds (2.93GHz versus 3.06GHz)

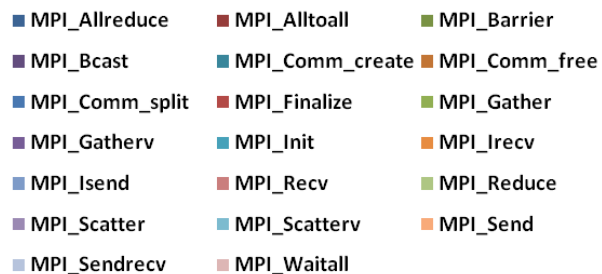
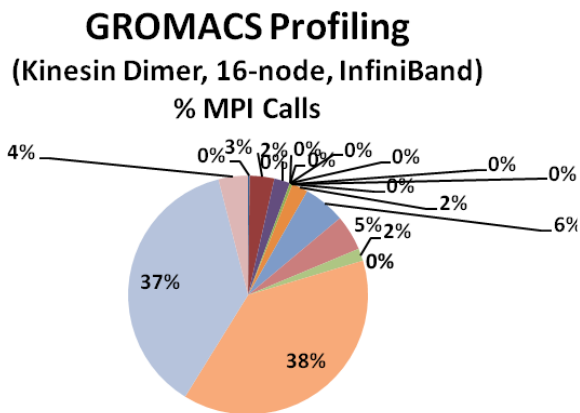


*InfiniBand QDR*

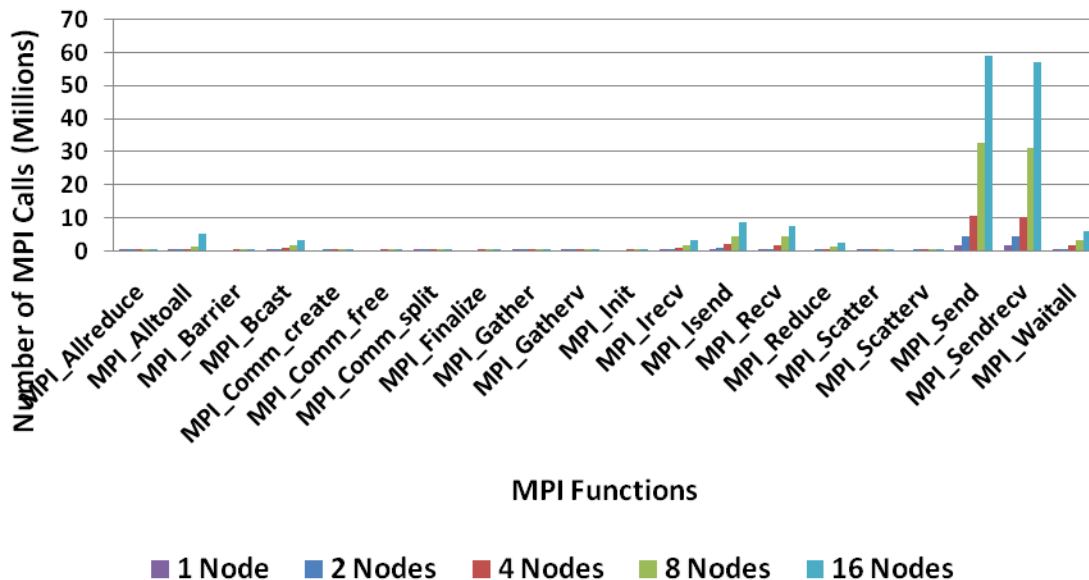
- **Computation time is dominant compared to MPI communication time**
  - MPI time only accounts for around 30% at 16-node
  - MPI communication ratio increases as the cluster scales
  - Means tuning for CPU or computation performance could yield better results



- **MPI\_Send and MPI\_Sendrecv are the most used MPI calls**
  - Each is accounted for around 38% of the MPI function calls on a 16-node job
- **GROMACS uses MPI functions for data transfers**
  - Seen a large volume of MPI calls for transferring data
  - Inferring the application is network bandwidth-bound

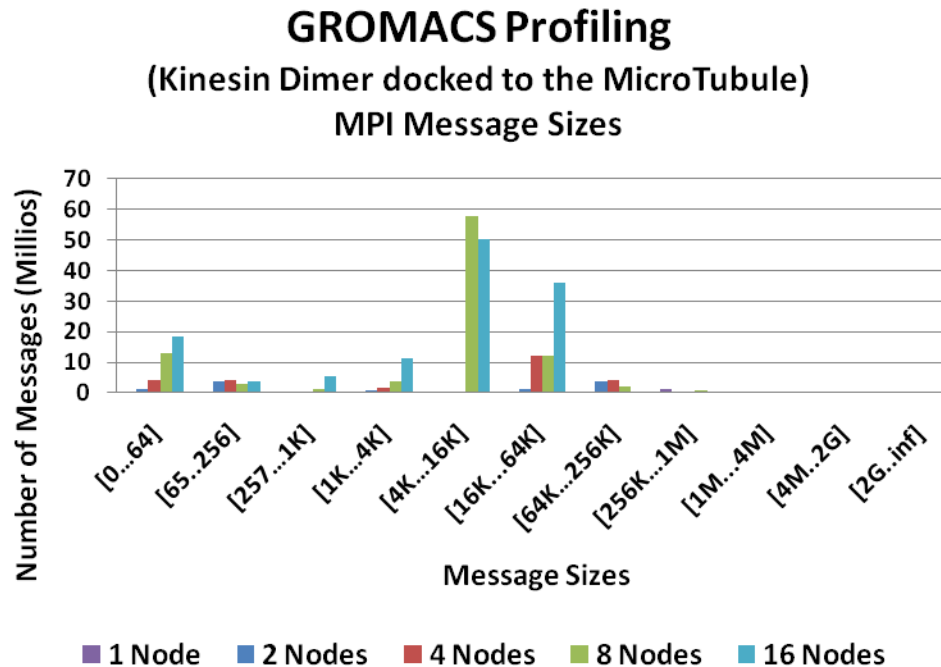


**GROMACS Profiling**  
(Kinesin Dimer docked to the MicroTubule)  
Number of MPI Calls



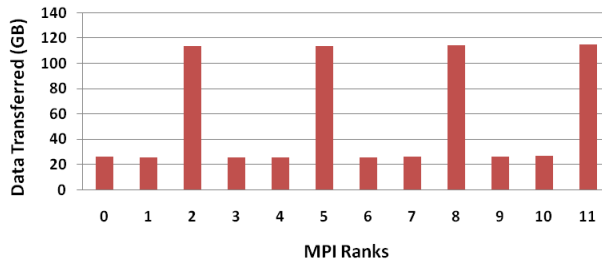


- **Most of the MPI messages are in the medium sizes**
  - Most message sizes are between 4KB to 16KB, and 16KB to 64KB

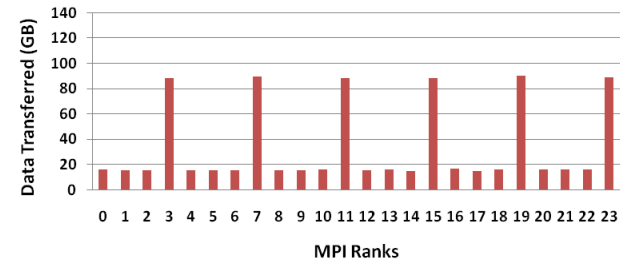


- **As the cluster grows, substantial less data transfers between MPI processes**
  - Reducing data communications from 115GB an single node simulation
  - To around 15GB for a 16-node simulation

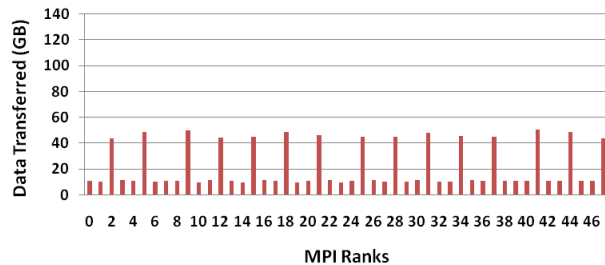
**GROMACS Profiling**  
(Kinesin Dimer, 1-node)  
Data Transferred by Ranks



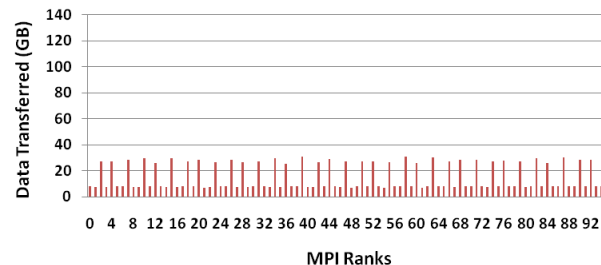
**GROMACS Profiling**  
(Kinesin Dimer, 2-node)  
Data Transferred by Ranks



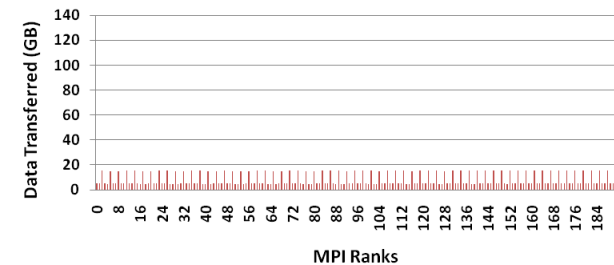
**GROMACS Profiling**  
(Kinesin Dimer, 4-node)  
Data Transferred by Ranks



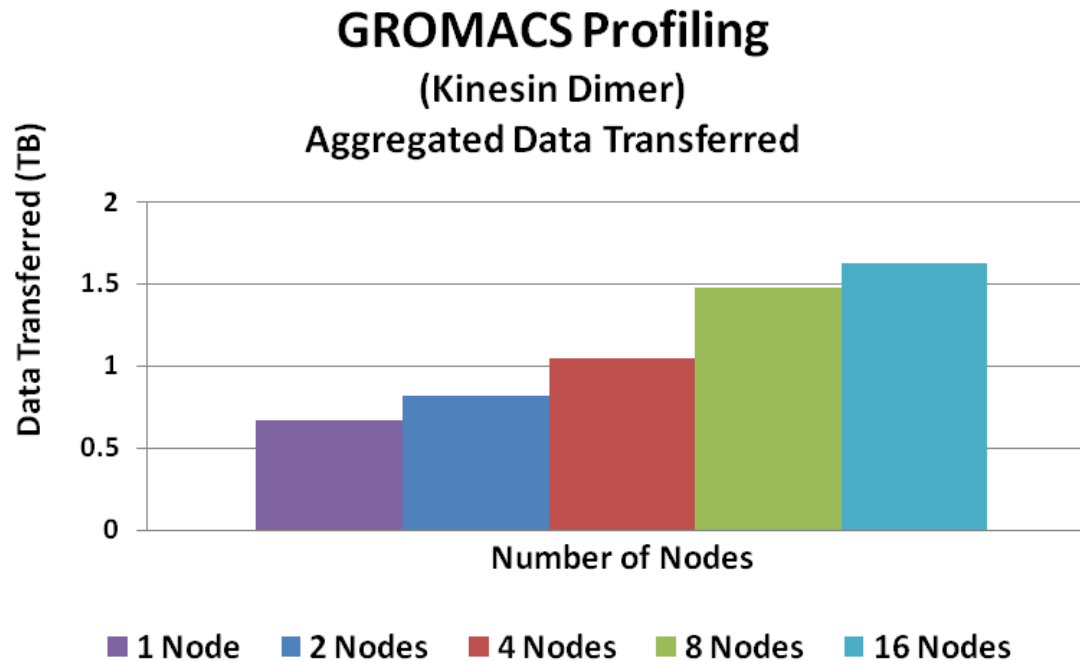
**GROMACS Profiling**  
(Kinesin Dimer, 8-node)  
Data Transferred by Ranks



**GROMACS Profiling**  
(Kinesin Dimer, 16-node)  
Data Transferred by Ranks



- **Aggregated data transfer refers to:**
  - Total amount of data being transferred in the network between all MPI ranks collectively
- **Large data transfer takes place in GROMACS**
  - Seen around 750GB to 1.5TB of data being exchanged between the nodes



*InfiniBand QDR*

- **Performance**

- InfiniBand allows GROMACS to run at the most efficient rate
- InfiniBand enables highest network throughput to allow GROMACS to scale
- Ethernet would not allow scale, ended up wasting valuable system resources

- **Tuning**

- MPI parameters tuning can provide some benefits around 3-5%
- Tuning for CPU or computation can deliver higher performance
- As the CPU/MPI time ratio shows significantly more computation is taken place
- Using faster CPU clock speed would yield 5% higher performance (2.93GHz vs 3.06GHz)
- Enabling Turbo Mode allows additional 2% in CPU utilization over static Max Performance
- Spreading the computational workload to more nodes can get job done faster

- **Profiling**

- MPI\_Send and MPI\_Sendrecv are the most used MPI functions
- MPI\_Allreduce is the most dominant MPI function call

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