

GROMACS

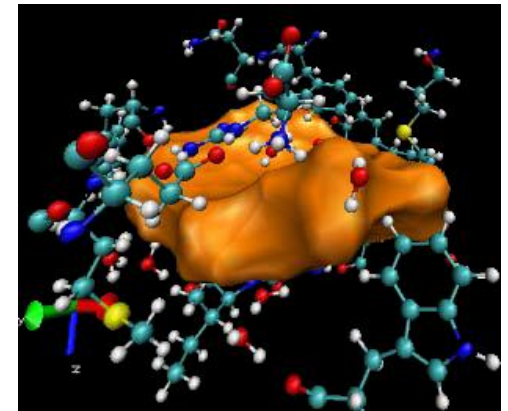
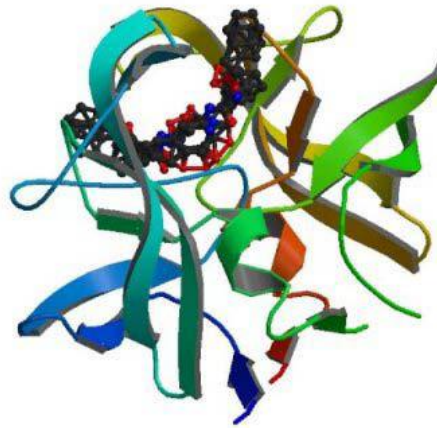
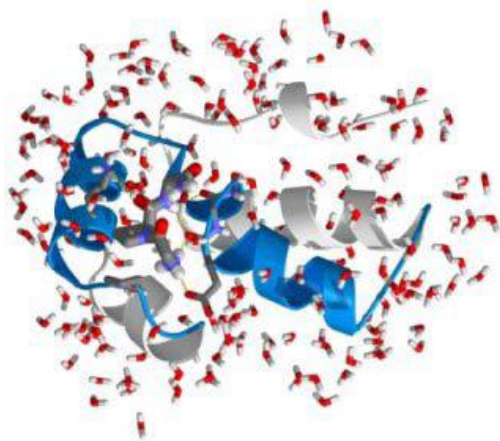
Performance Benchmark and Profiling

September 2011



- **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.amd.com)
 - [http:// www.dell.com/hpc](http://www.dell.com/hpc)
 - <http://www.mellanox.com>
 - <http://www.gromacs.org>

- **GROMACS (GRONingen MACHine for Chemical Simulation)**
 - A molecular dynamics simulation package
 - Primarily designed for biochemical molecules
 - Such as 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



- **The following was done to provide best practices**
 - GROMACS performance benchmarking
 - Understanding GROMACS communication patterns
 - Ways to increase GROMACS productivity
 - Compilers and network interconnects comparisons
- **The presented results will demonstrate**
 - The scalability of the compute environment
 - The capability of GROMACS to achieve scalable productivity
 - Considerations for performance optimizations

Test Cluster Configuration

- **Dell™ PowerEdge™ R815 11-node (528-core) cluster**
- **AMD™ Opteron™ 6174 (code name “Magny-Cours”) 12-cores @ 2.2 GHz CPUs**
- **4 CPU sockets per server node**
- **Mellanox ConnectX-2 VPI adapters for 40Gb/s QDR InfiniBand and 10Gb/s Ethernet**
- **Mellanox MTS3600Q 36-Port 40Gb/s QDR InfiniBand switch**
- **Fulcrum-based 10Gb/s Ethernet Switch**
- **Memory: 128GB memory per node DDR3 1333MHz**
- **OS: RHEL 5.5, MLNX-OFED 1.5.3 InfiniBand SW stack**
- **MPI: Platform MPI 8.1.1**
- **Compilers: GNU Compilers 4.4, PGI 11.8**
- **Libraries: AMD ACML 4.4.0, fftpack, Intel MKL 10.2 Update 5**
- **Application: GROMACS 4.5.4**
- **Benchmark workload: DPPC in Water (d.dppc) (5000 steps, 10.0 ps.)**

- **HPC Advisory Council Test-bed System**
- **New 11-node 528 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™ 6100 series platform and Mellanox ConnectX 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 48 core/32DIMMs per server – 1008 core in 42U enclosure

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

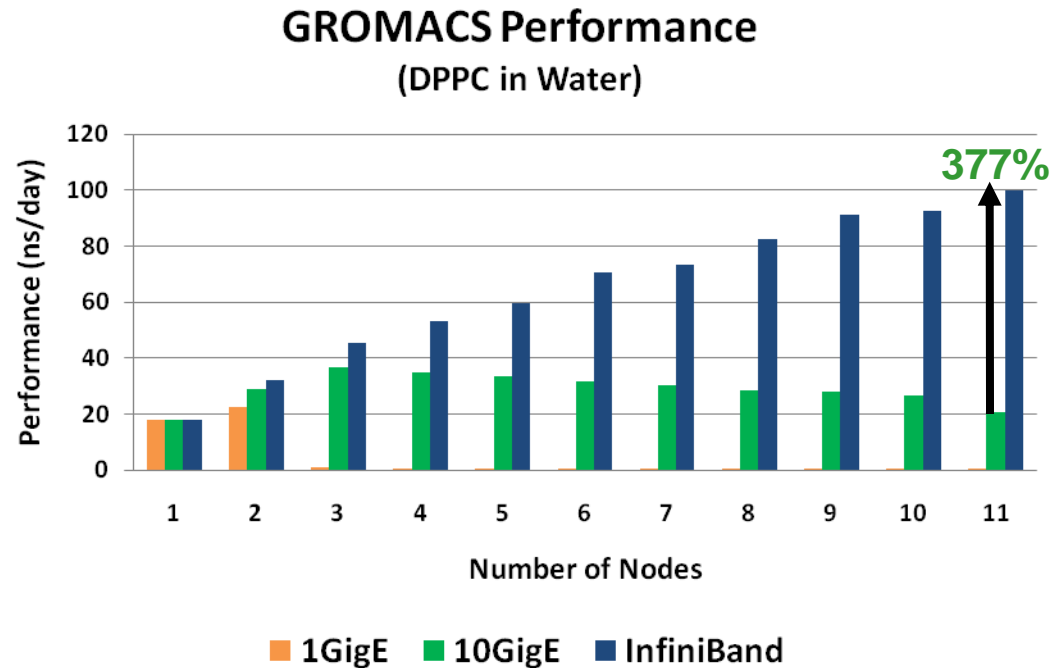
- 2x more memory and processing power in half of the space
- Energy optimized low flow fans, improved power supplies and dual SD modules

Optimized for long-term capital and operating investment protection

- System expansion
- Component upgrades and feature releases



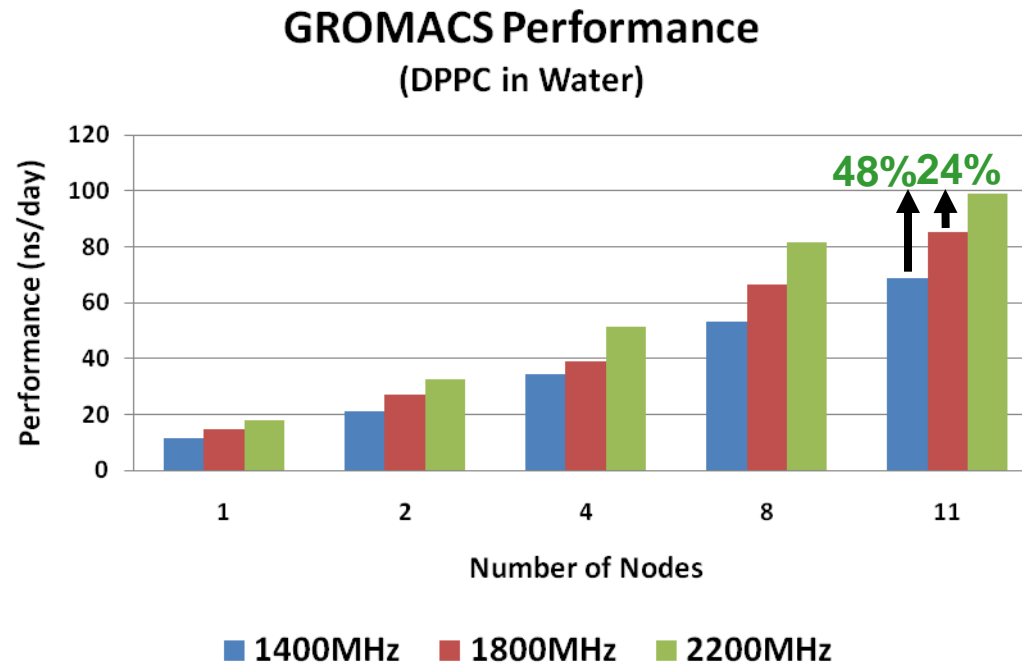
- **InfiniBand QDR delivers the best performance for GROMACS**
 - Seen up to 377% better performance than 10GigE on 11-node
- **Scalability limitation seen with Ethernet networks**
 - 10GigE performance starts to drop after 3-node
 - 1GigE performance drop takes place after 2-node



Higher is better

48 Cores/Node

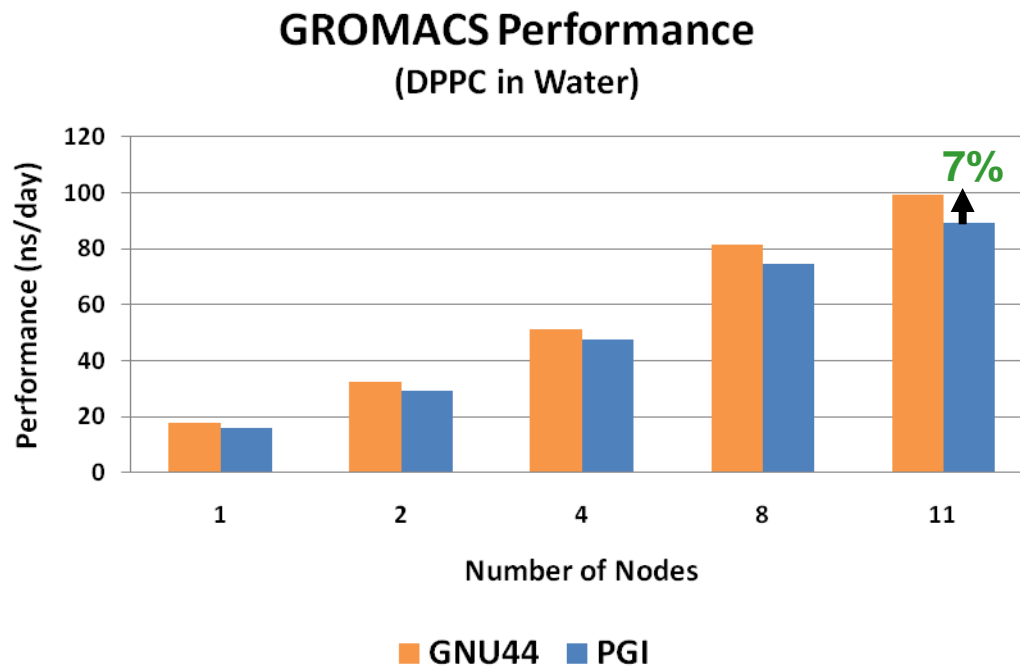
- **Higher CPU core frequency enables higher job performance**
 - Up to 24% better job performance between 2200MHz vs 1800MHz
 - Up to 48% better job performance between 2200MHz vs 1400MHz
 - The increase in CPU core frequencies can directly improve the overall job performance



Higher is better

48 Cores/Node

- **Executable generated by GNU compilers runs faster**
 - Up to 7% faster than with PGI compilers
- **Using the default optimization and linker flags:**
 - **GNU:** “-O3 -msse2 -fomit-frame-pointer -finline-functions -Wall -Wno-unused -funroll-all-loops -std=gnu99”
 - **PGI:** “-tp istanbul -O4 -fastsse -Msmartalloc=huge -Mconcur -Mipa=fast,inline -Mvect=prefetch -Munroll -fPIC”

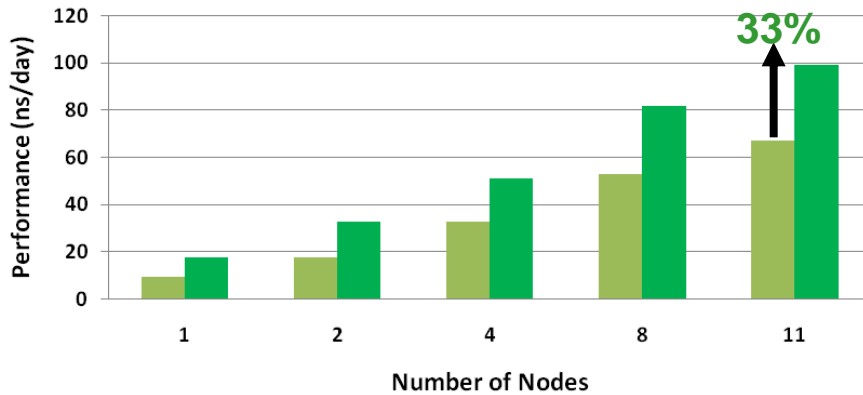


Higher is better

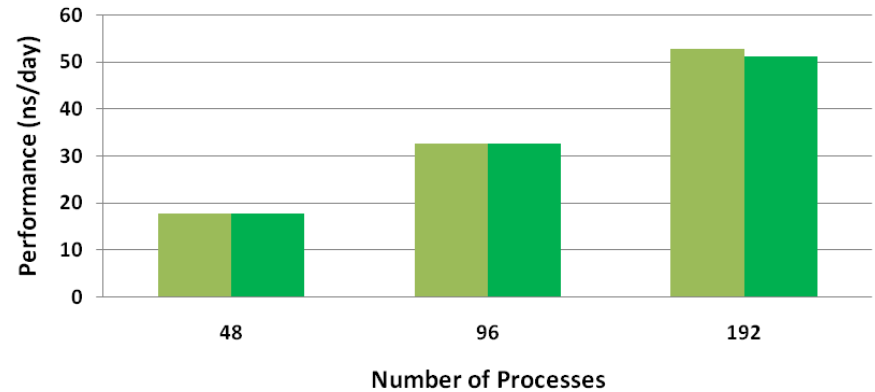
Platform MPI
48 Cores/Node

- **Allocating more processes per node can yield higher system utilization**
 - Seen 33% in performance gain by running with 48 PPN versus 24 PPN at 11-node
- **Using 48 PPN achieves the same performance as using 24 PPN**
 - GROMACS can fully utilize all CPU cores available in a system
 - GROMACS can benefit by reducing hardware footprint with high core-count CPUs
- **No loss in performance by spreading the workload to more nodes**
 - Only InfiniBand allows lossless performance by spreading the workload over the network

GROMACS Performance
(DPPC in Water)



GROMACS Performance
(DPPC in Water)



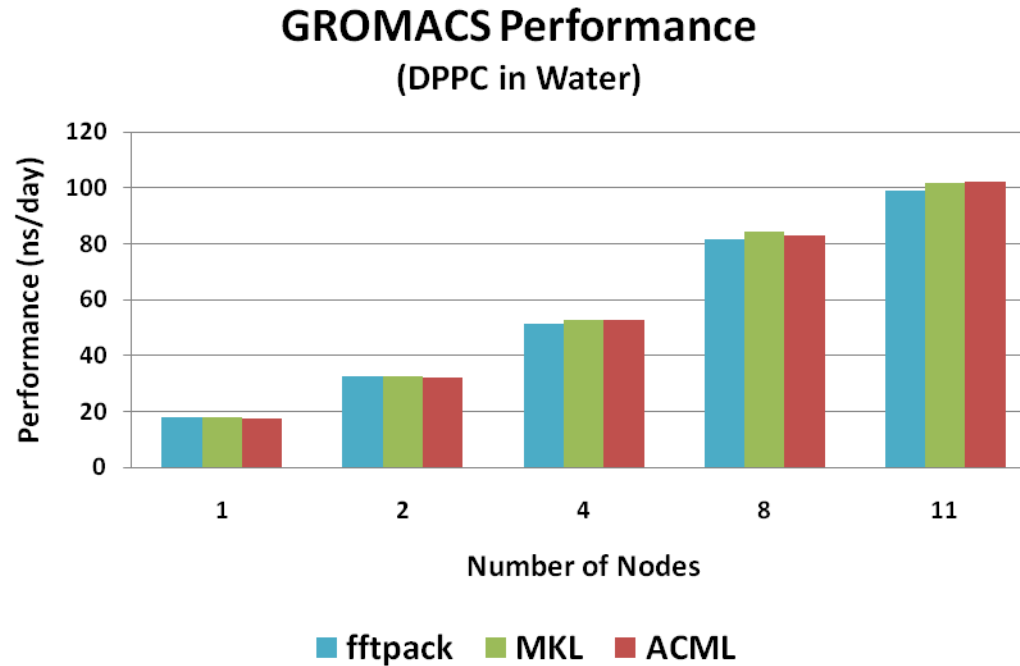
Higher is better

■ 24 PPN ■ 48 PPN

■ 24 PPN ■ 48 PPN

InfiniBand QDR

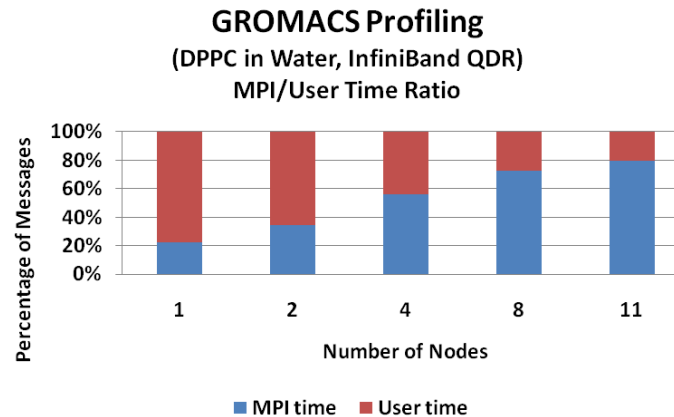
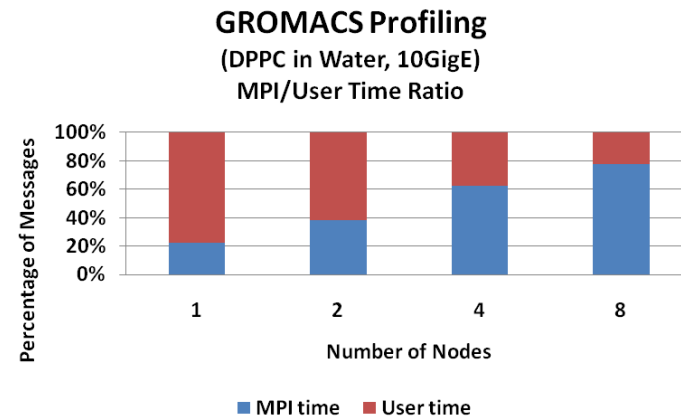
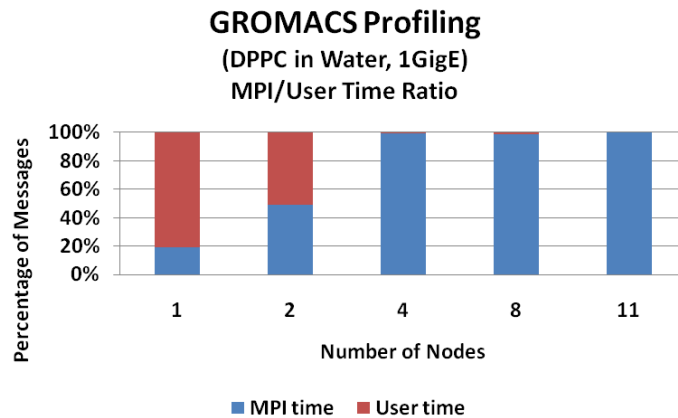
- **ACML and MKL deliver a slightly better performance than fftpack**
 - ACML and MKL are 1-2% better performance than fftpack



Higher is better

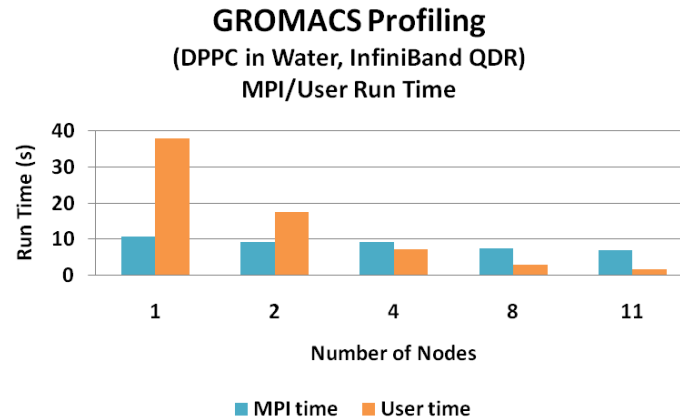
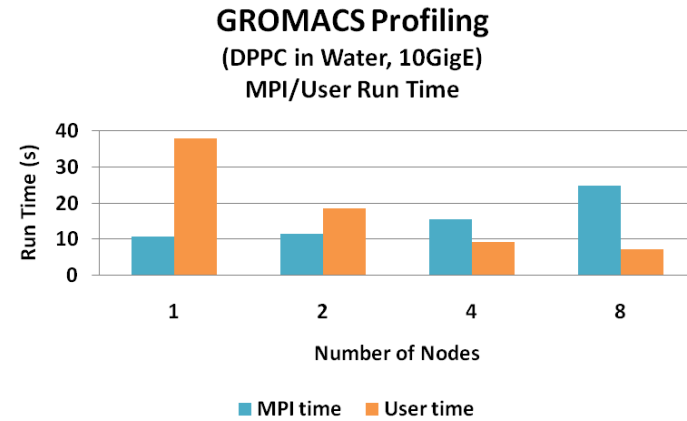
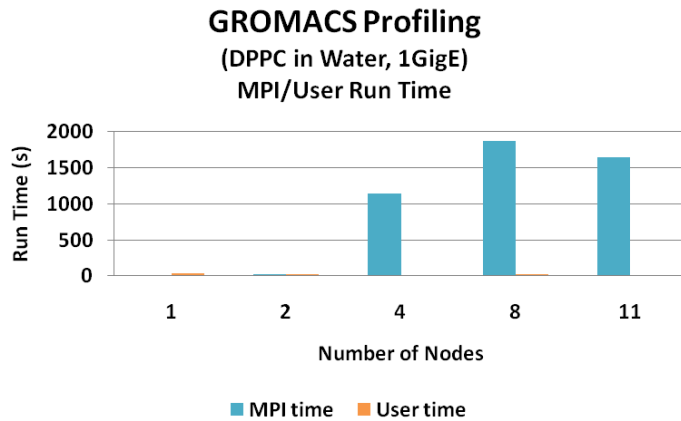
48 Cores/Node

- **InfiniBand QDR reduces the amount of time for MPI communications**
 - 1GigE becomes efficient to handle the MPI communications after 2-node
 - 10GigE has a percentage of communications than InfiniBand QDR



48 Cores/Node

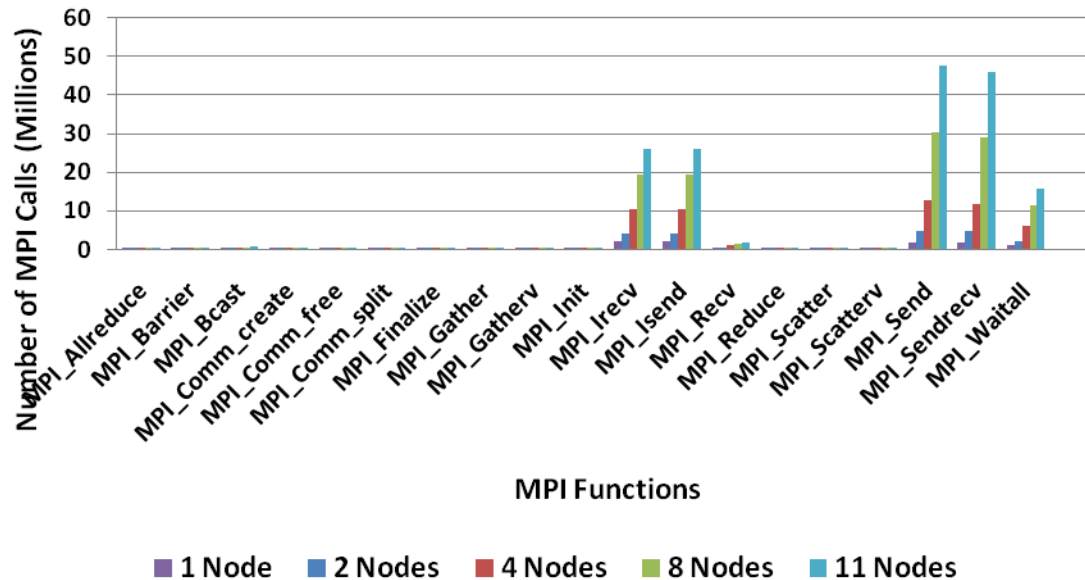
- **Communication time for InfiniBand stays flat as cluster grows**
 - MPI run times for InfiniBand remains constant as the node number increases
 - MPI run times for 10GigE and 1GigE increase as the node number increases



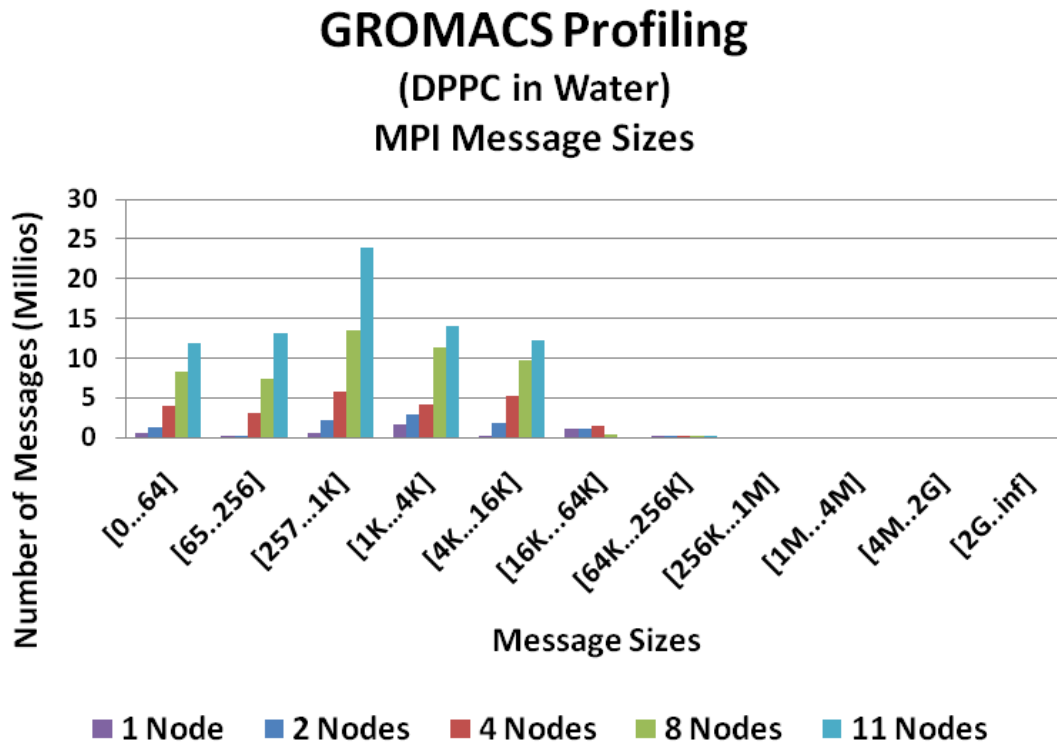
48 Cores/Node

- **The most used MPI functions are for data transfers**
 - MPI_Send
 - MPI_Sendrecv
 - MPI_Isend
 - Reflects that GROMACS requires good network throughput
- **The number of calls increases proportionally as the cluster scales**

GROMACS Profiling
(DPPC in Water)
Number of MPI Calls

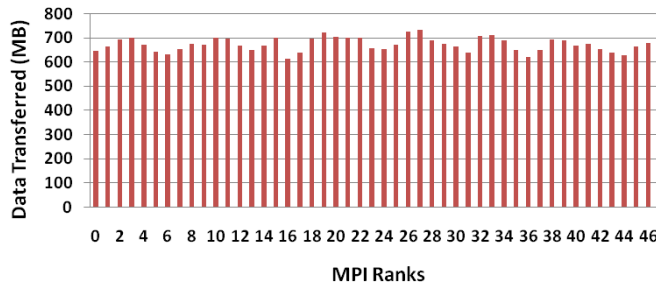


- **Majority of the MPI messages are small to median message sizes**
 - In the ranges of between 257B and 1KB
 - All of the MPI messages are in the sizes less than 256KB
- **Low network latency requires for good small MPI message performance**

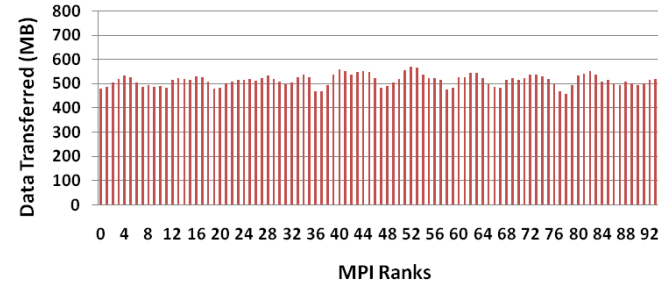


- **Data transferred to each MPI rank is generate constant for all MPI processes**
 - Amount of data transfer to each rank is reduced as more nodes are in the job
 - From around 650MB per rank on 1-node down to around 300MB per rank for 8-node

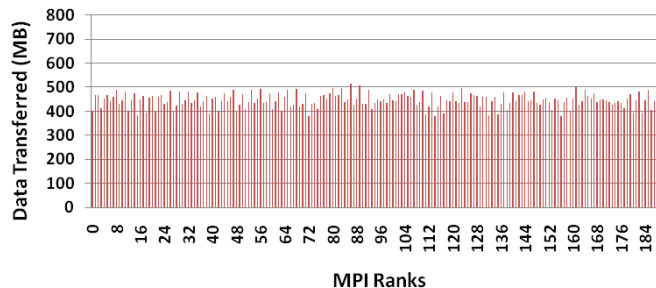
GROMACS Profiling
(DPPC in Water, 1-node)
Data Transferred by Ranks



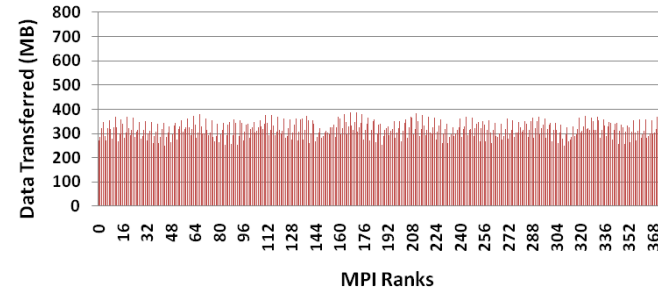
GROMACS Profiling
(DPPC in Water, 2-node)
Data Transferred by Ranks



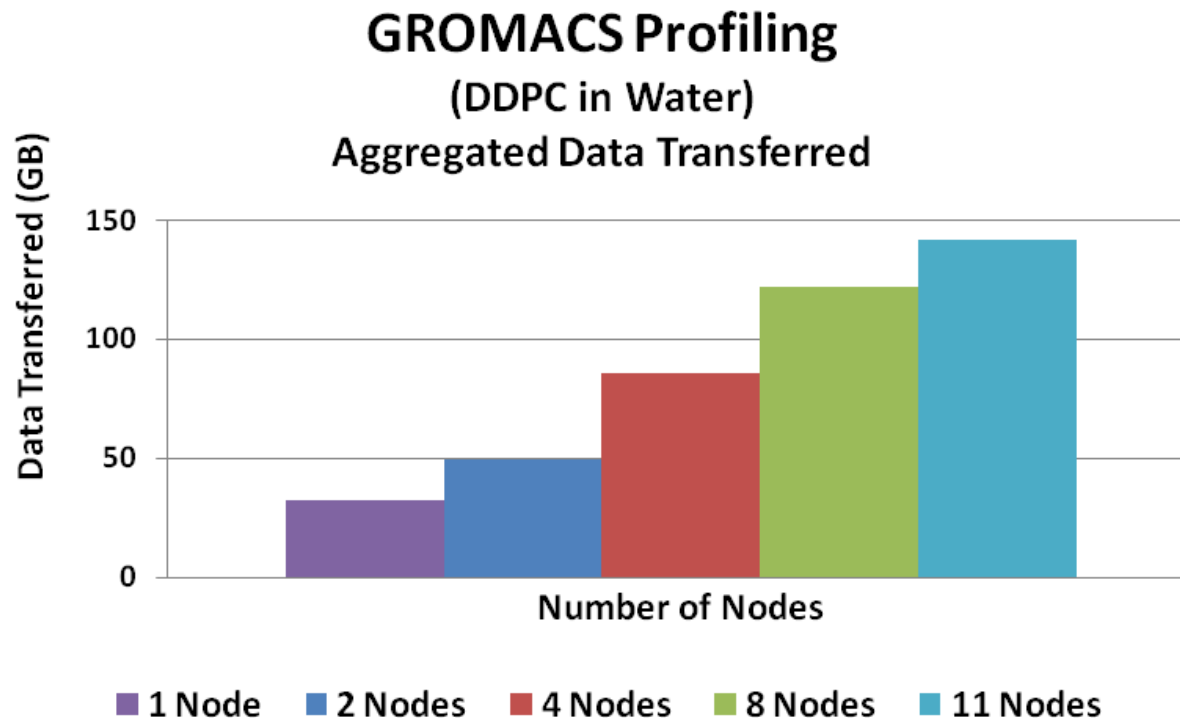
GROMACS Profiling
(DPPC in Water, 4-node)
Data Transferred by Ranks



GROMACS Profiling
(DPPC in Water, 8-node)
Data Transferred by Ranks



- **Aggregated data transfer refers to:**
 - Total amount of data being transferred in the network between all MPI ranks collectively
- **The total data transfer increases steadily as the cluster scales**
 - For this dataset, a good amount of data being sent and received across the network
 - As a compute node being added, more data communications will take place



InfiniBand QDR

- **GROMACS is a compute and network sensitive application**
 - GROMACS has a high demand for CPU utilization and network interconnect performance
- **CPU:**
 - Higher CPU core frequency allows GROMACS to achieve higher performance
 - GROMACS can benefit by using high core-count CPUs, thus reducing hardware footprint
- **Interconnects:**
 - InfiniBand QDR can deliver good scalability for GROMACS
 - 10GigE and 1GigE would not scale and become inefficient to run beyond 2-3 nodes
- **Math Libraries:**
 - ACML and MKL has a slight advantage over fftpack
- **Compilers:**
 - GNU compilers shows higher CPU performance than PGI compilers

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