HPC ADVISORY COUNCIL


# GROMACS Performance Benchmark and Profiling

October 2009

AMD The future is fusion

DELL

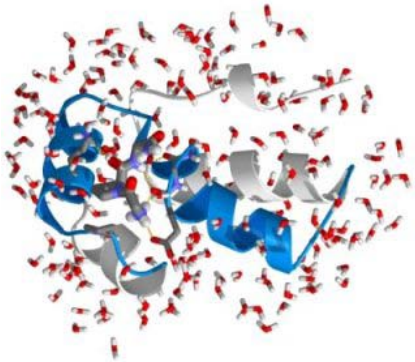
Mellanox TECHNOLOGIES

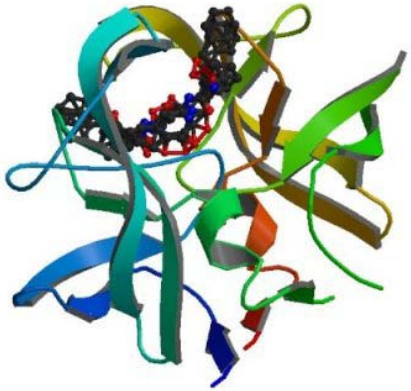
# Note

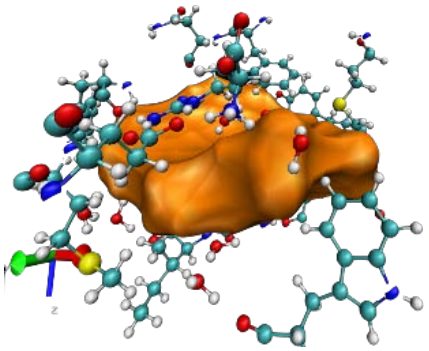
- **The following research was performed under the HPC Advisory Council activities**
  - Participating vendors: AMD, Dell, Mellanox
  - Compute resource - HPC Advisory Council Cluster Center
- **For more info please refer to**
  - www.mellanox.com, www.dell.com/hpc, www.amd.com

# GROMACS

- **GROMACS (GROningen MAchine for Chemical Simulations)**
  - 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







# Objectives

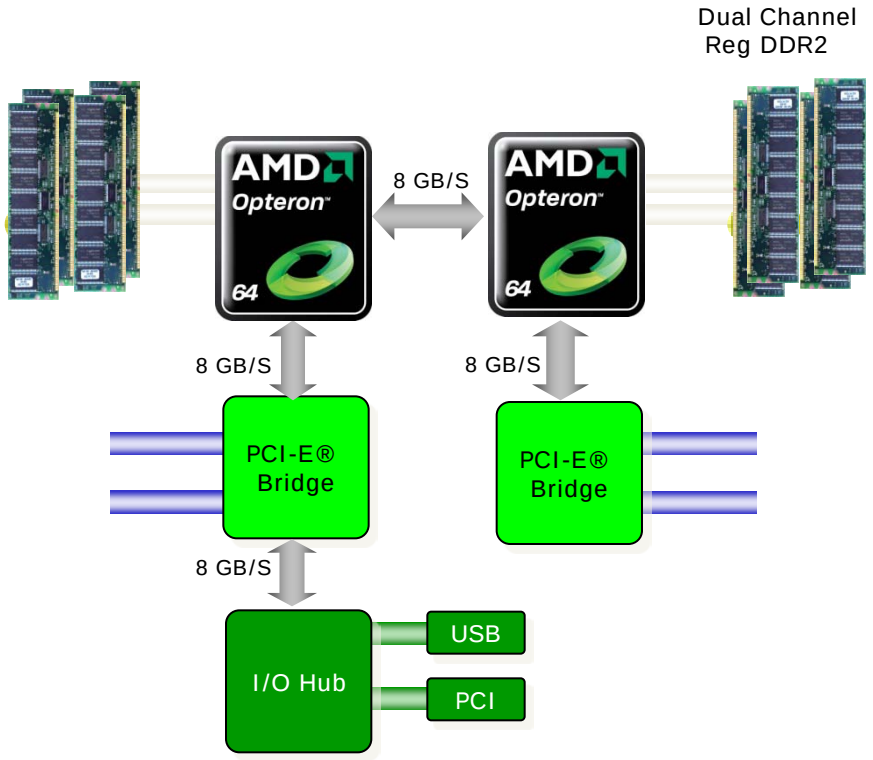
- **The presented research was done to provide best practices**
  - GROMACS performance benchmarking
  - Interconnect performance comparisons
  - MPI performance comparison
  - Power-efficient simulations
  - Understanding GROMACS communication patterns
- **The presented results will demonstrate**
  - The scalability of the compute environment to provide nearly linear application scalability
  - Considerations for power saving through balanced system configuration

# Test Cluster Configuration

- Dell™ PowerEdge™ SC 1435 24-node cluster
- Quad-Core AMD Opteron™ 2382 (“Shanghai”) CPUs
- Mellanox® InfiniBand ConnectX® 20Gb/s (DDR) HCAs
- Mellanox® InfiniBand DDR Switch
- Memory: 16GB memory, DDR2 800MHz per node
- OS: RHEL5U3, OFED 1.4.1 InfiniBand SW stack
- MPI: Open MPI-1.3.3, MVAPICH-1.1.0
- Application: GROMACS 4.0.5
- Benchmark Workload
  - D.DPPC (A phospholipid membrane,121,856 atoms)

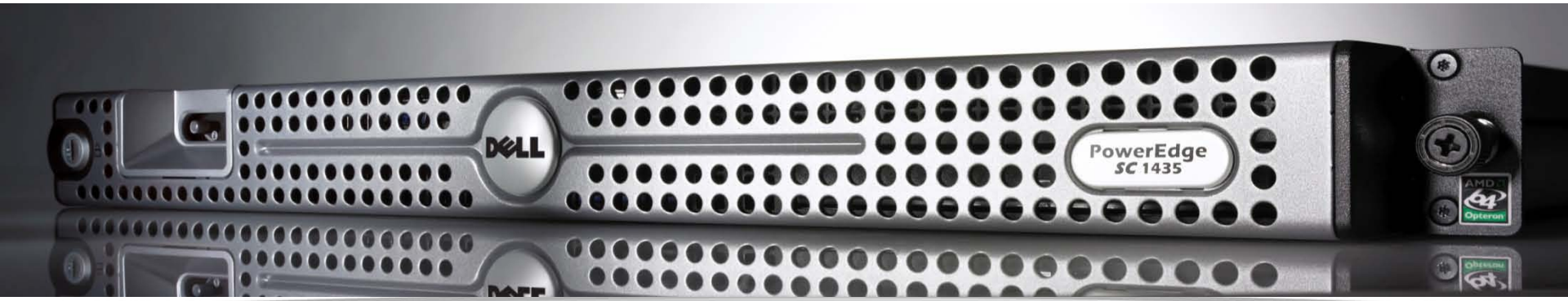
# About Quad-Core AMD Opteron™ Processor

- **Performance**
  - Quad-Core
    - Enhanced CPU IPC
    - 4x 512K L2 cache
    - 6MB L3 Cache
  - Direct Connect Architecture
    - HyperTransport™ Technology
    - Up to 24 GB/s peak per processor
  - Floating Point
    - 128-bit FPU per core
    - 4 FLOPS/clk peak per core
  - Integrated Memory Controller
    - Up to 12.8 GB/s
    - DDR2-800 MHz or DDR2-667 MHz
- **Scalability**
  - 48-bit Physical Addressing
- **Compatibility**
  - Same power/thermal envelopes as 2nd / 3rd generation AMD Opteron™ processor



# About Dell PowerEdge™ Server Advantage

- **Dell™ PowerEdge™ servers incorporate AMD Opteron™ and Mellanox ConnectX InfiniBand to provide leading edge performance and reliability**
- **Building Block Foundations for best price/performance and performance/watt**
- **Investment protection and energy efficient**
- **Longer term server investment value**
- **Faster DDR2-800 memory**
- **Enhanced AMD PowerNow!**
- **Independent Dynamic Core Technology**
- **AMD CoolCore™ and Smart Fetch Technology**
- **Mellanox InfiniBand end-to-end for highest networking performance**

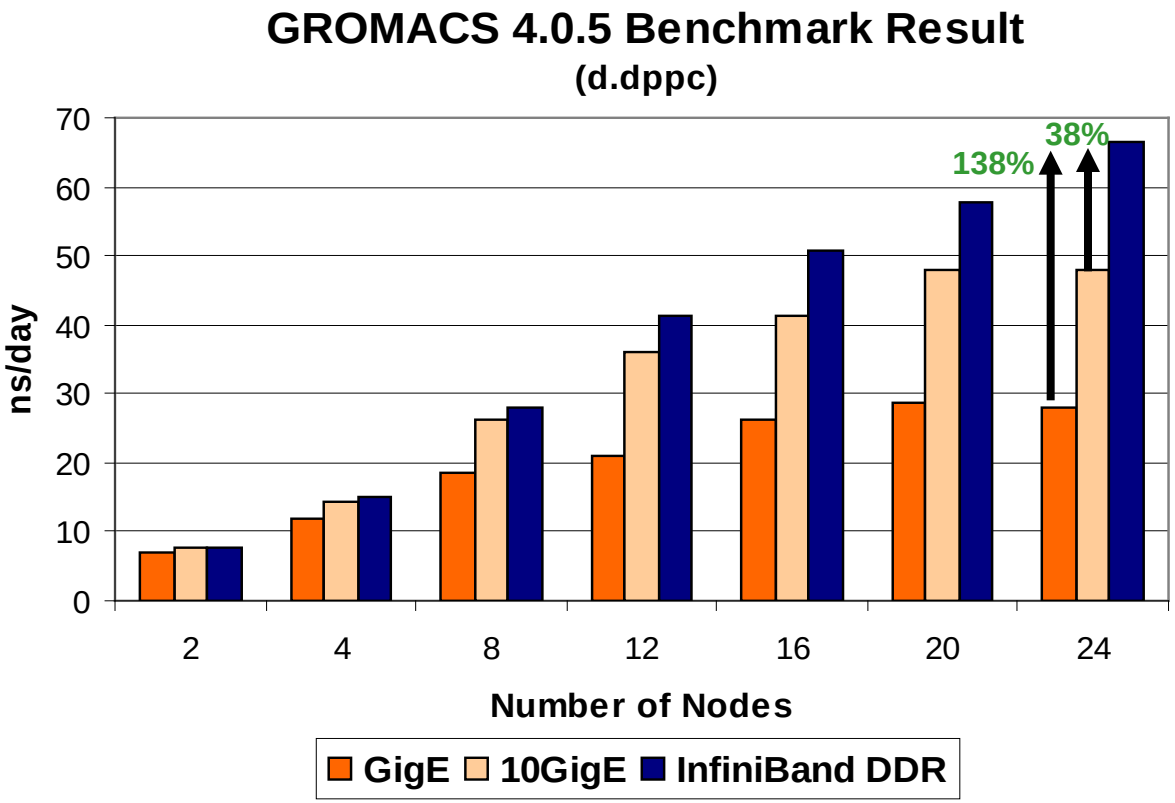


# GROMACS Application Summary

- **GROMACS is a simulation software for molecular dynamics**
  - Support bonded interactions (biochemical) and non-bonded interactions (non-biological )
- **Profiling results shows the scaling capabilities of GROMACS**
  - Good scaling was demonstrated to 24 server nodes
  - No limitations found to hold scaling beyond that size
  - CPUs and memory bandwidth provide the needed capabilities for the continuous increase in performance
- **Beyond 20 server nodes, GROMACS requires InfiniBand capabilities**
  - Beyond 10Gb/s bandwidth, lowest latency for MPI collectives operations
  - Networking optimizations for collectives operation, in particular AllReduce and Broadcast, expected to greatly increase performance and efficiency
    - Hardware capabilities to handle MPI collectives

# GROMACS Benchmark Results - Interconnect

- **InfiniBand provides higher utilization, performance and scalability**
  - Up to 38% faster than 10 GigE and 138% than GigE with 24 nodes configuration
  - Both GigE and 10GigE stop scaling after 20 nodes



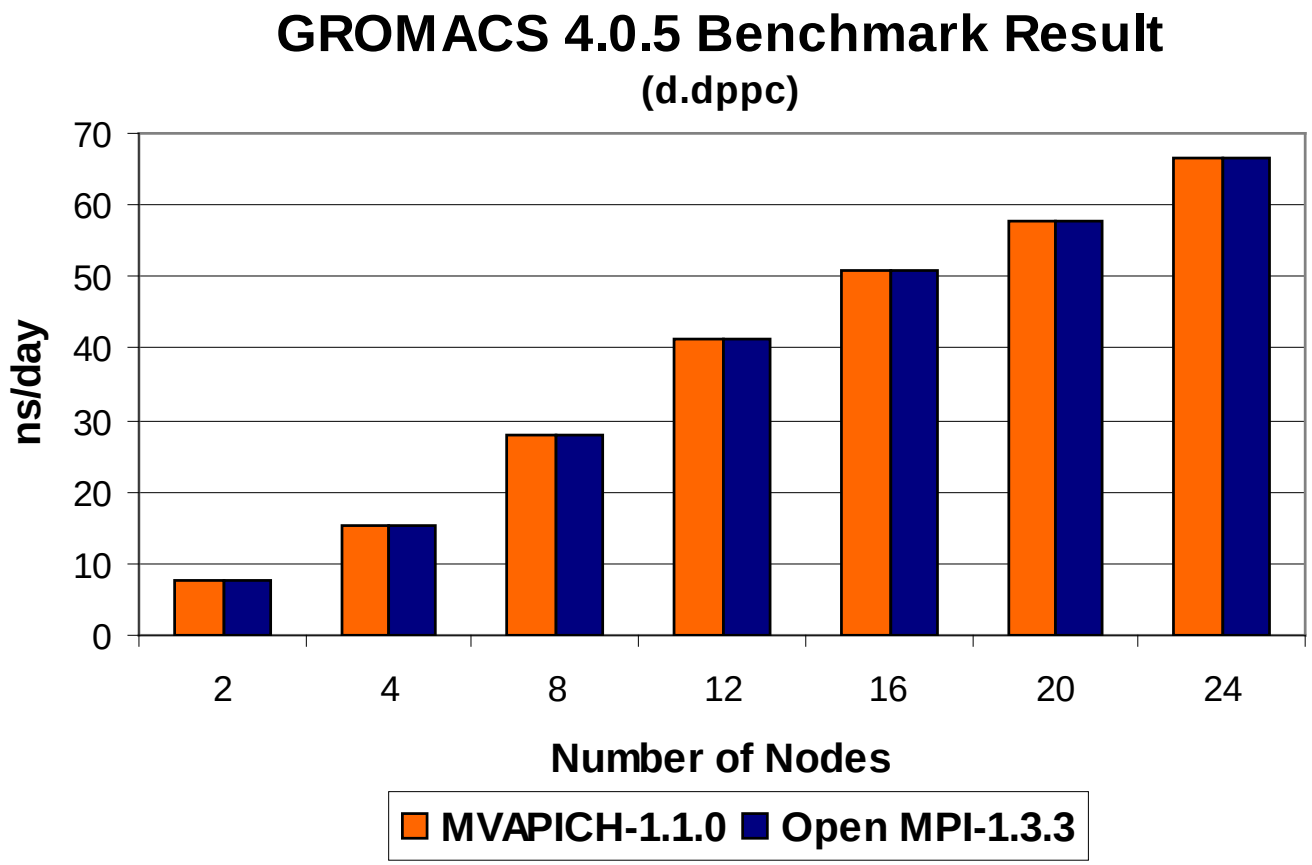

*Higher is better*

8-cores per node

*Open MPI*

# GROMACS Benchmark Results - MPI

- **MVAPICH and Open MPI provide similar performance**

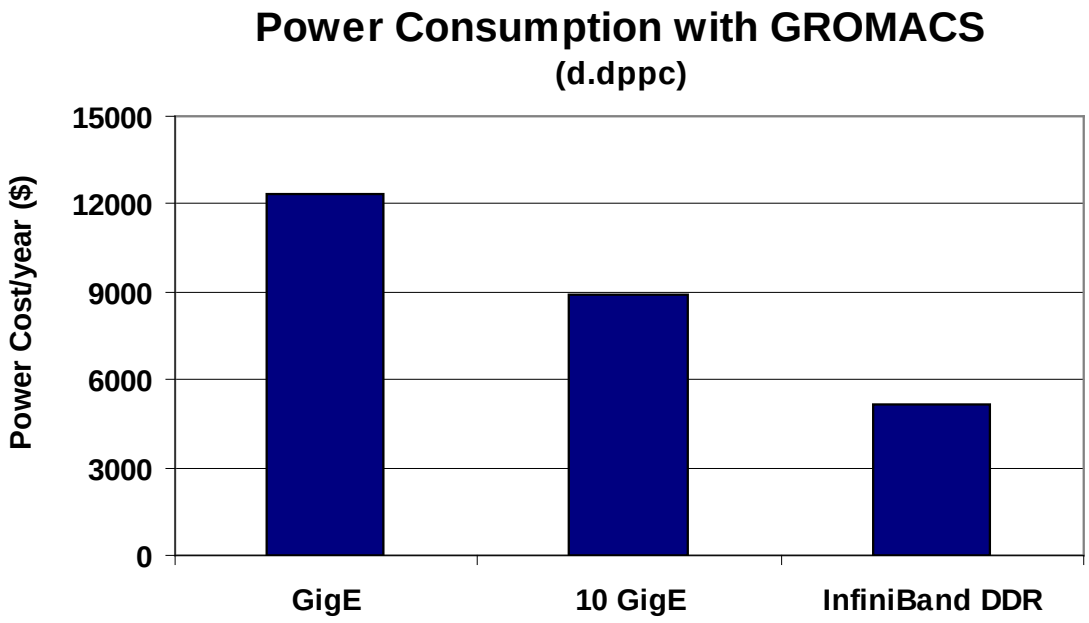



*Higher is better*

*InfiniBand DDR*

# Power Cost Savings with Different Interconnect

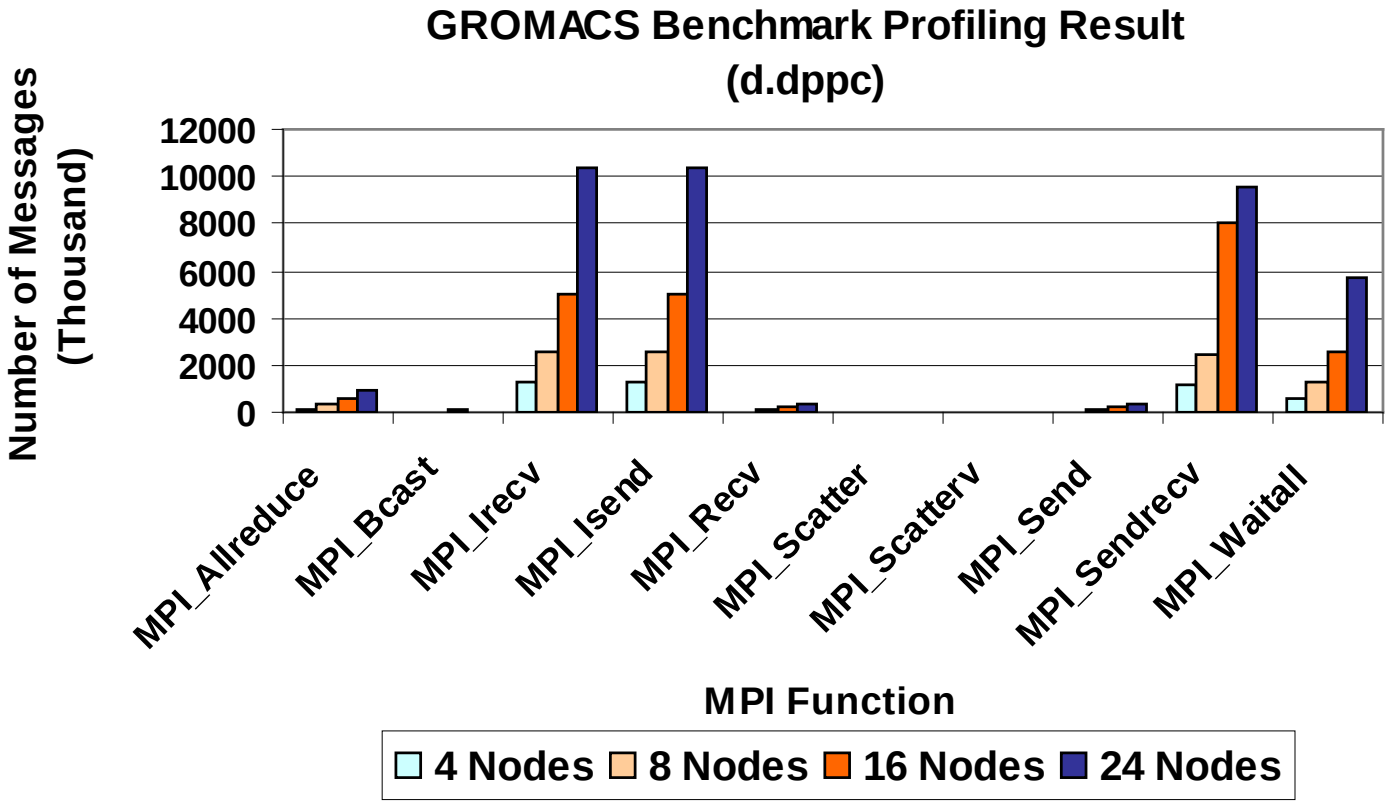
- **Dell integration saves up to $7000 in power**
  - To achieve same number of application jobs enabled with Gigabit Ethernet
  - Yearly based for 24-node cluster
- **As cluster size increases, more power can be saved**




*$/KWh = KWh * $0.20*
*For more information - http://enterprise.amd.com/Downloads/svrpwrusecompletefinal.pdf*

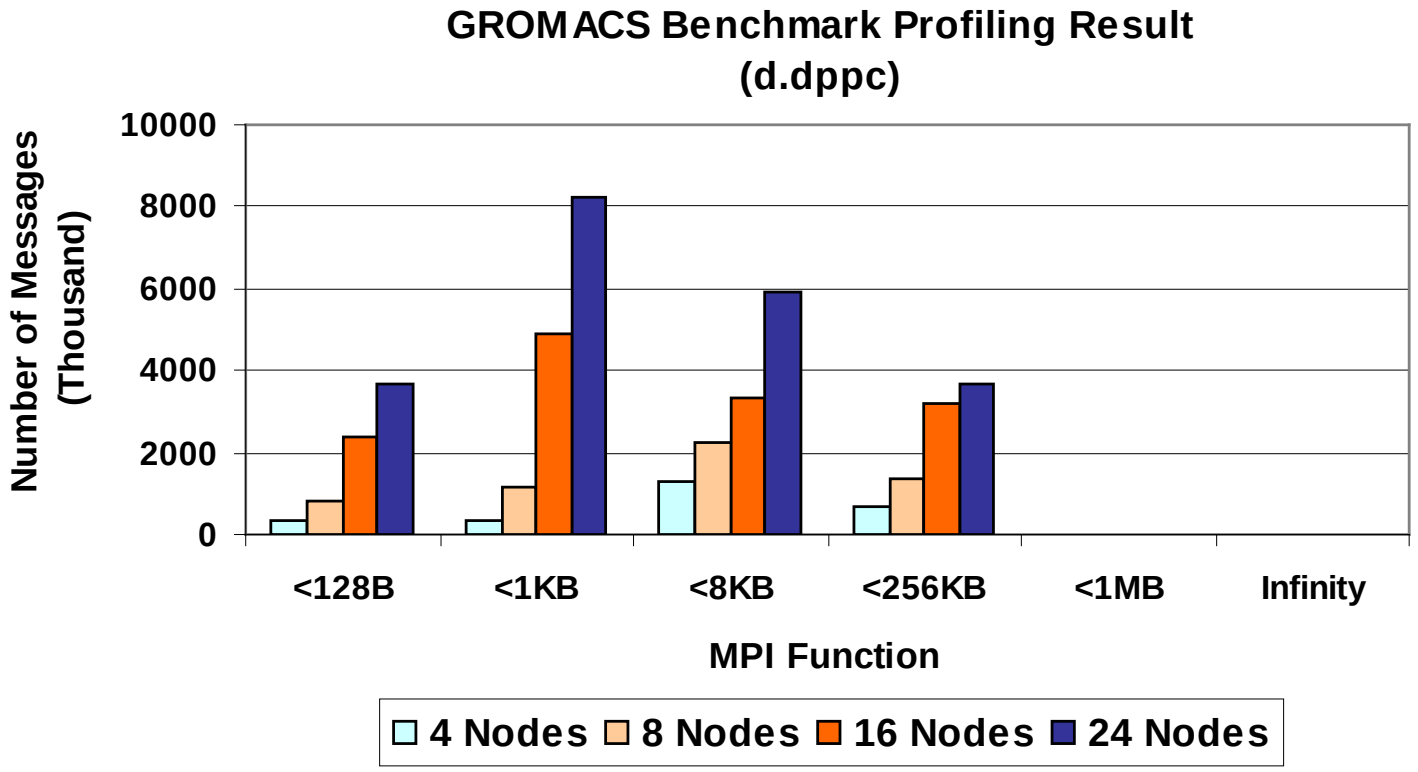
# GROMACS Profiling – MPI Functions

- **Most often used MPI functions**
  - MPI_Irecv, MPI_Isend, MPI_Sendrecv, and MPI_Waitall



# GROMACS Profiling – Message Size

- Both small and large messages are transferred between ranks
- Number of messages increases with cluster size
- High bandwidth interconnect (>=10Gb/s starting at 8 nodes) is required
- Highest bandwidth interconnect (>10Gb/s starting at 16 nodes) is required



# GROMACS Profiling Summary

- **GROMACS was profiled to identify its communication patterns**
- **Frequently used message sizes**
  - 1KB-256KB messages for data related communications
  - <128B for synchronizations
  - Number of messages increases with cluster size
- **Interconnect effect on GROMACS performance**
  - Interconnect bandwidth (MPI_Sendrecv) and latency (MPI_Allreduce) highly influence GROMACS performance

# Thank You

## HPC Advisory Council