HPC ADVISORY COUNCIL


# PFA (Pretty Fast Analysis) Performance Benchmark and Profiling

June 2011

AMD The future is fusion

DELL

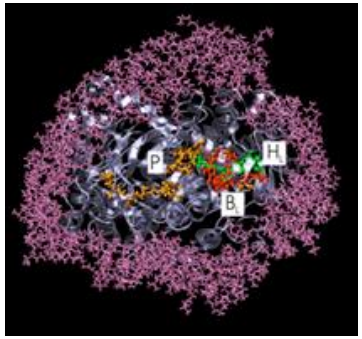
Mellanox TECHNOLOGIES

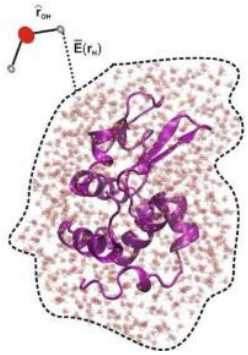
# Note

- **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.temple.edu/cst/icms/index.html

# PFA Application

- **Pretty Fast Analysis (PFA)**
  - Software for analyzing large-scale molecular dynamics (MD) simulation trajectory data
  - Reads either CHARMM or AMBER style topology/trajectory files as input, and its analysis routines can scale up to thousands of compute cores or hundreds of GPU nodes with either parallel or UNIX file I/O
  - PFA has dynamic memory management, and each code execution can perform a variety of different structural, energetic, and file manipulation operations on a single MD trajectory at once
  - The code is written in a combination of Fortan90 and C, and its GPU kernels are written with NVIDIA's CUDA API to achieve maximum GPU performance
  - PFA is produced by research staff at the Temple University Institute for Computational Molecular Science



# Objectives

- **The following was done to provide best practices**
  - PFA performance benchmarking
  - Understanding PFA communication patterns
  - Ways to increase PFA productivity
  - Compilers and MPI libraries comparisons

- **The presented results will demonstrate**
  - The scalability of the compute environment
  - The capability of PFA 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
- Memory: 128GB memory per node DDR3 1333MHz
- OS: RHEL 5.5, MLNX-OFED 1.5.2 InfiniBand SW stack
- MPI: Open MPI 1.5.3 with KNEM 0.9.6, Platform MPI 8.1.1
- Compilers: PGI 10.9, GNU Compilers 4.4
- Libraries: ACML 4.4.0
- Application: PFA
- Benchmark workload: lysozyme (1200 Frames, 10A cutoff, 100 Configurations)

# About Dell™ PowerEdge™ R815 11-node cluster

- **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

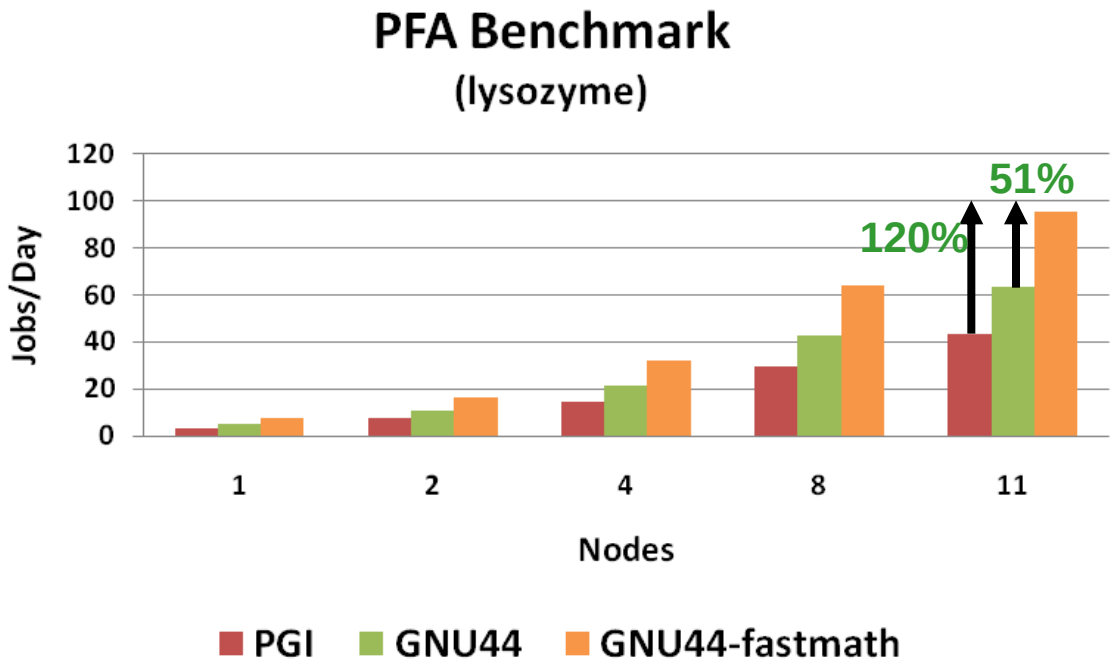
- 2x more memory and processing power in half of the space
- Energy optimized low flow fans, improved power supplies and dual SD modules

## Optimized for long-term capital and operating investment protection

- System expansion
- Component upgrades and feature releases

# PFA Performance – Compilers

- **Tuning with compiler flags can optimize runtime performance**
  - GNU shows better performance than PGI when similar optimizations are used
- **Optimization and linker flags used:**
  - PGI: "-O3 -fpic -fastsse -fast -tp istanbul-64 -lacml -llapack –lblas -lpgftnrtl"
  - GNU44: "-O3 -fpic -funroll-loops -mfpmath=sse -march=barcelona -lacml -llapack -lblas –lgomp"
  - GNU44-fastmath: same as GNU44 above and with "--fast-math" (**NOTE**: faster but may lose precision)



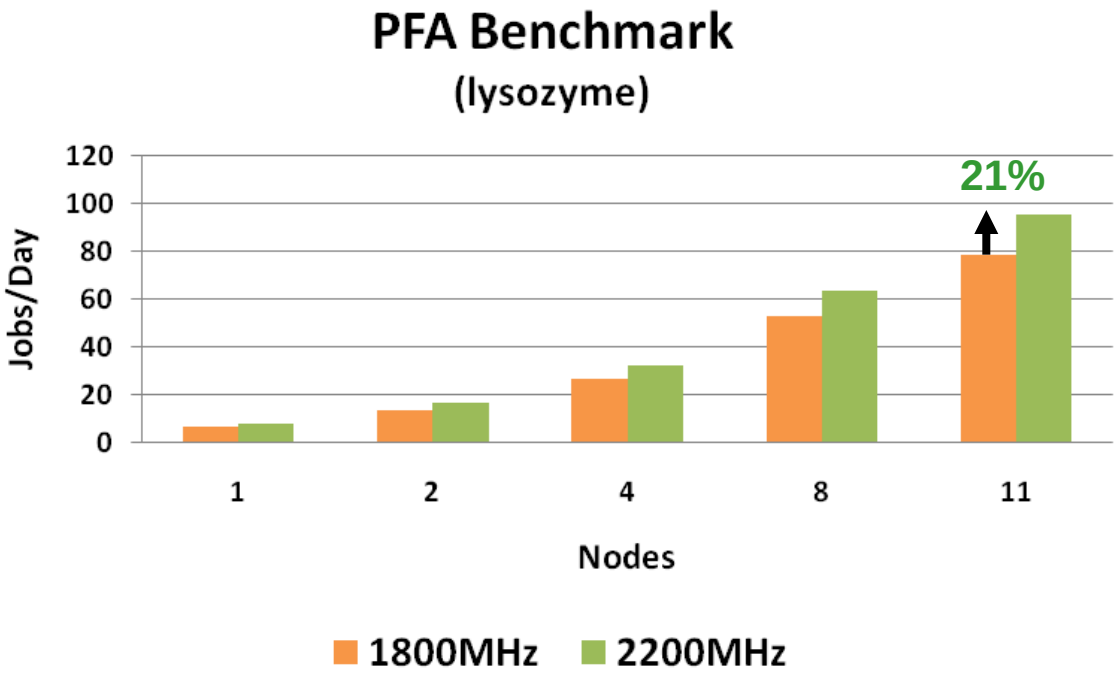

***Higher is better***

***Open MPI***

***48 Cores/Node***

# PFA Performance – CPU Frequency

- **Higher CPU core frequency enables higher job performance**
  - Up to 21% better job performance between 2200MHz vs 1800MHz on 11-node
  - As typical for a compute bound application, performance is affected by CPU frequency

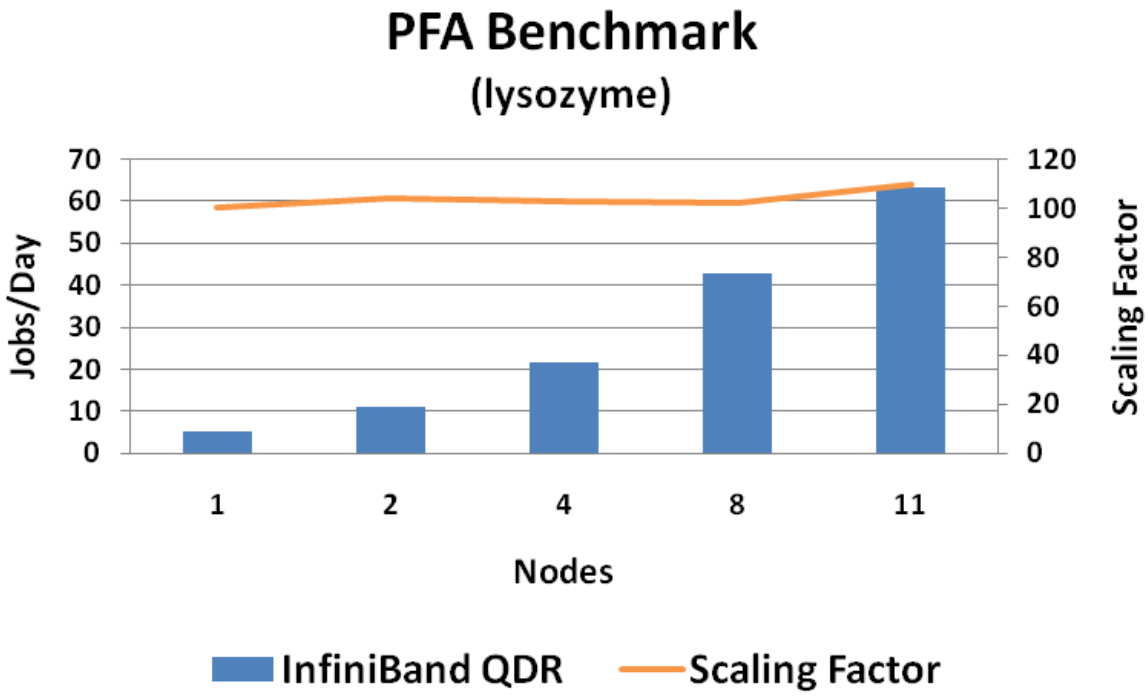



*Higher is better*

*48 Cores/Node*

# PFA Performance – Scalability

- **PFA demonstrates good scalability and system utilization**
  - As more compute nodes are added into the cluster, the performance doubles
  - Can fully benefit by adding more machines to reduce the overall job runtime

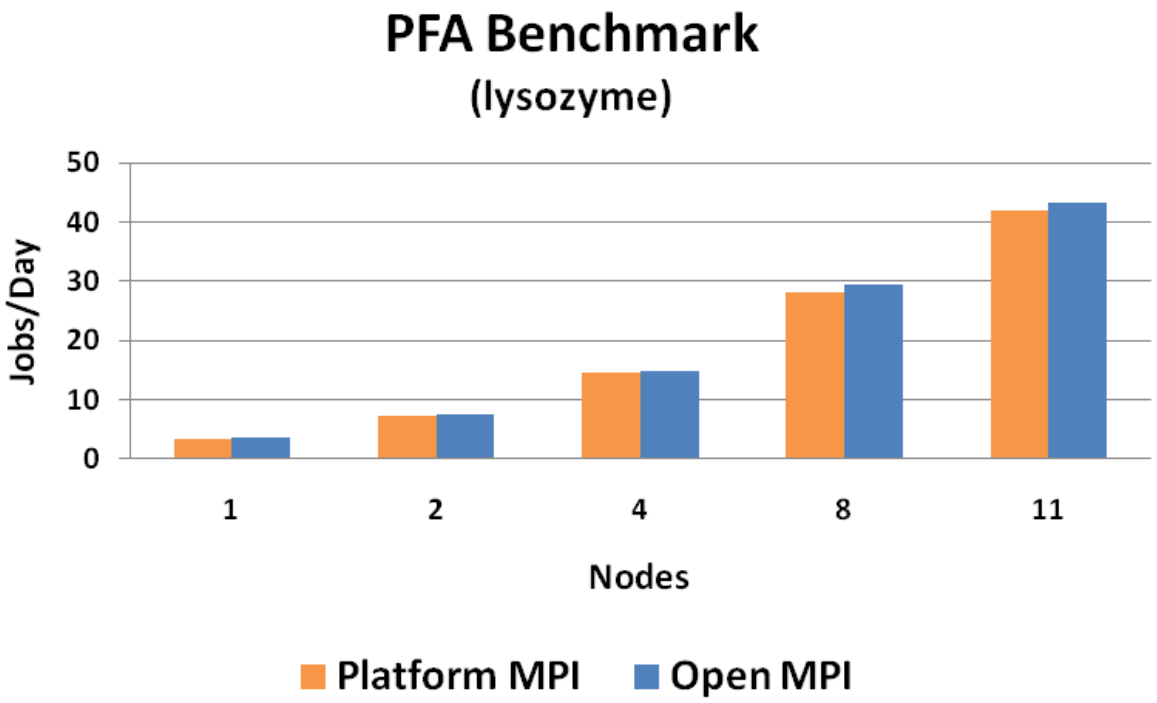



***Higher is better***

***48 Cores/Node***

# PFA Performance – MPI

- **Open MPI performs slightly better**
- **Profiling shows limited MPI calls take place**
  - Which explains little difference is seen when comparing the 2 MPI implementations



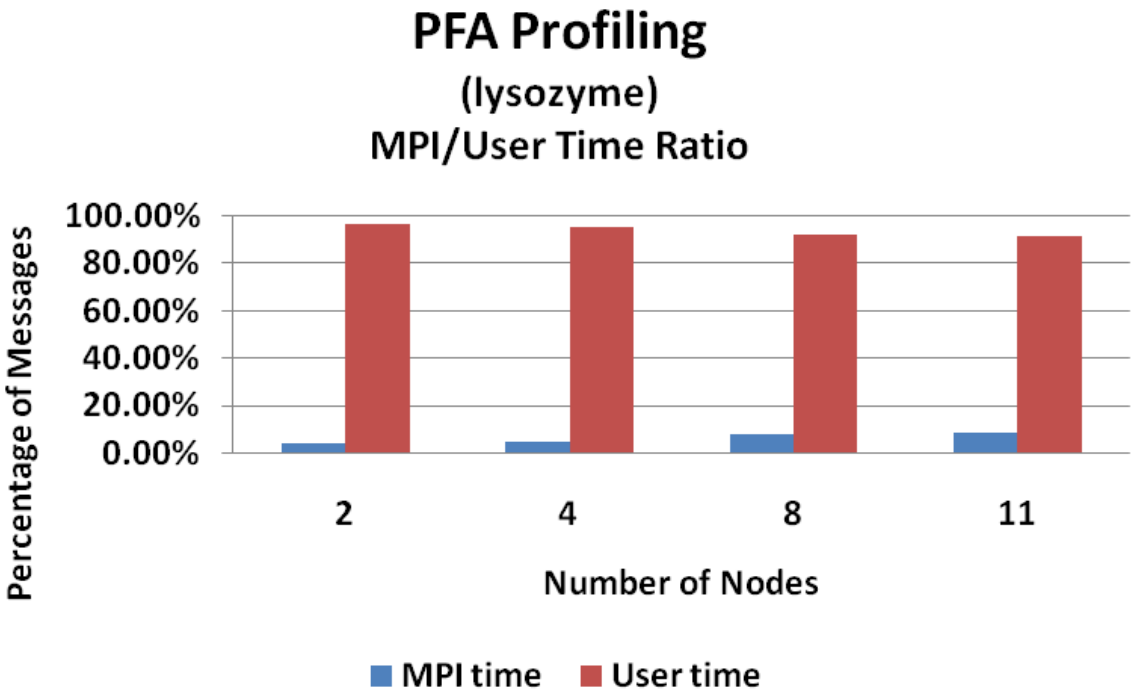

***Higher is better***

***PGI Compilers***

***48 Cores/Node***

# PFA Profiling – MPI/User Time Ratio

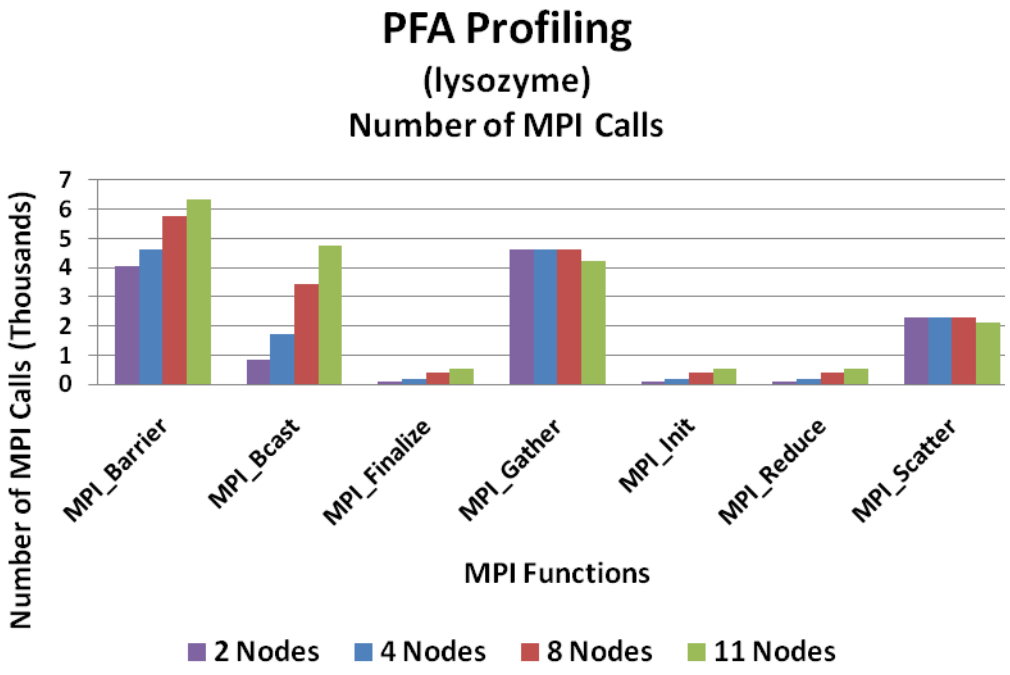
- **Slight increase in communications time as more nodes are added**
  - Less than 10% of time the job is spent on communications
  - Shows very limit communications (or dependencies) between parallel tasks
  - Typically seen in applications with embarassingly parallel workload

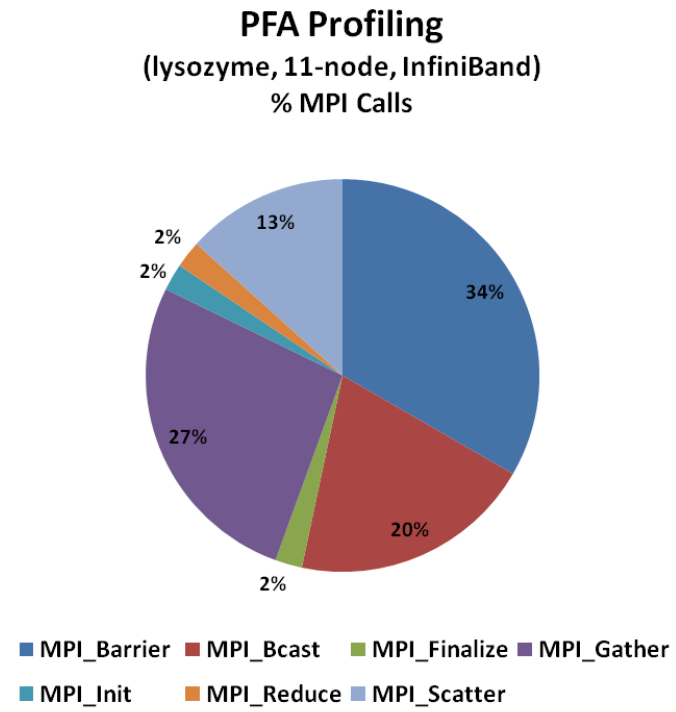



***48 Cores/Node***

# PFA Profiling – Number of MPI Calls

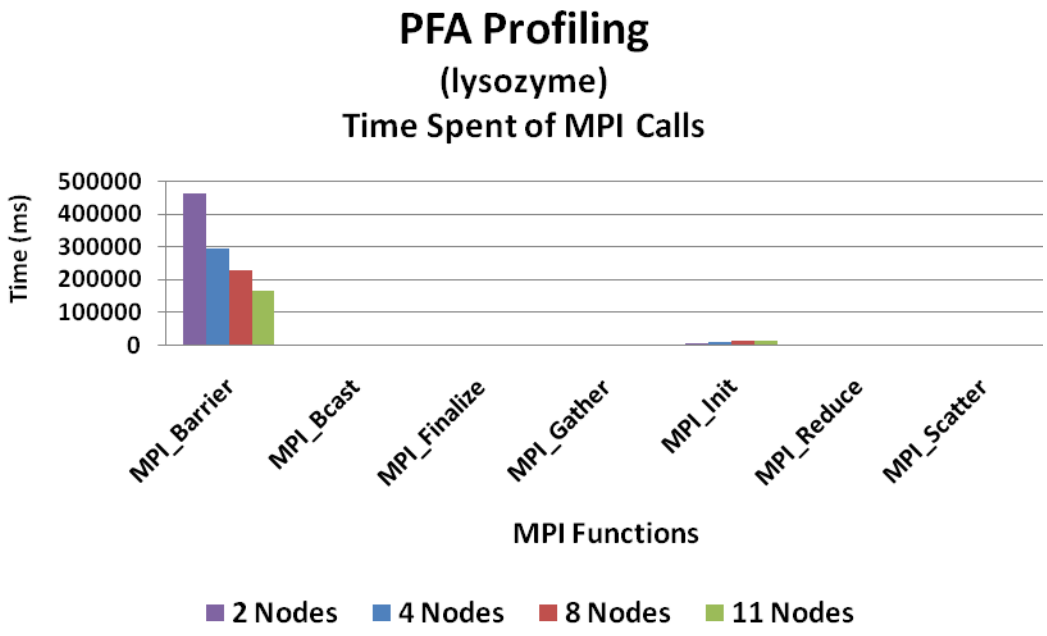
- **The most used MPI function is MPI_Barrier**
  - Represents 34% of MPI calls used for 8-node
- **The number of MPI_Gather and Scatter calls stay flat**
  - While the number of MPI_Bcast calls increases at a faster pace as the cluster scales

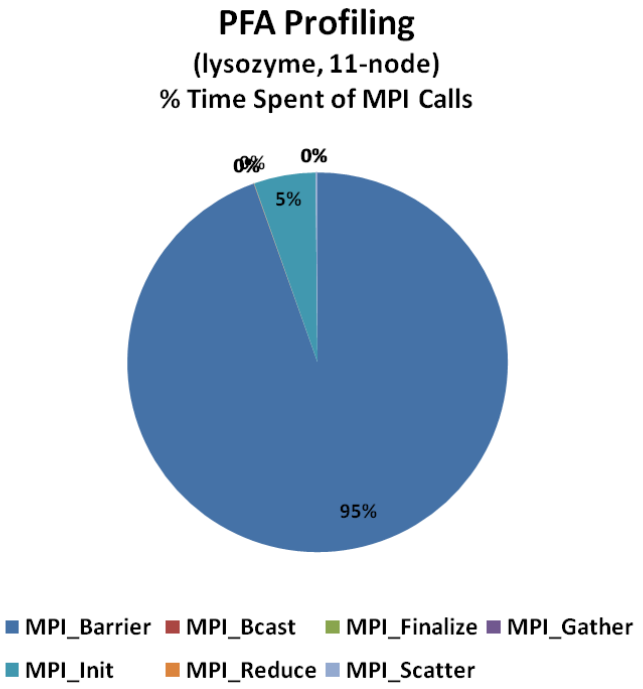


# PFA Profiling – Time Spent of MPI Calls

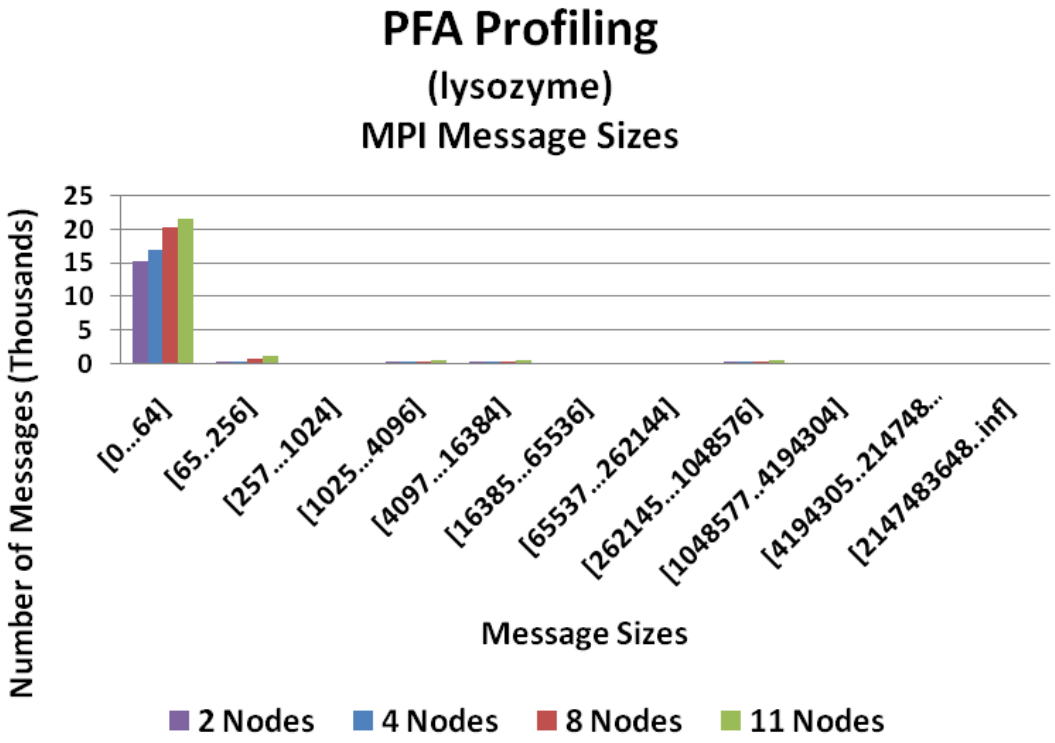
- **The largest time consumer is MPI_Barrier for data communications**
  - Occupies 95% of all MPI time at 8-node
  - Besides computation, the rest of application time spends on MPI_Barrier



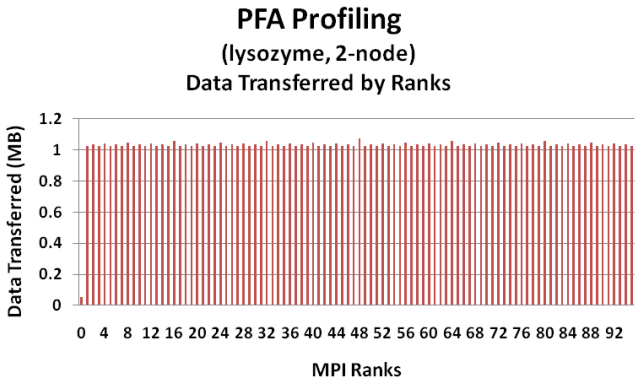
# PFA Profiling – MPI Message Sizes

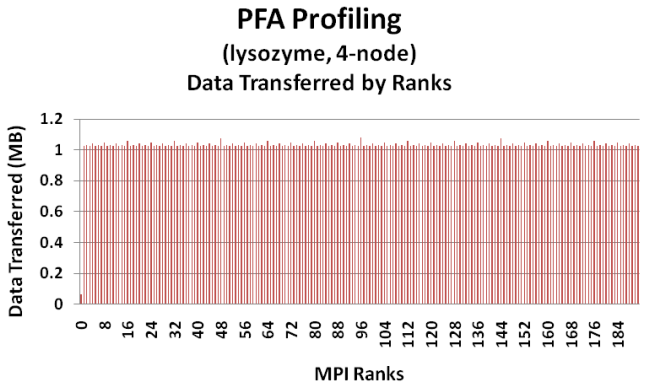
- **Majority of the MPI message sizes are small messages**
  - In the range of less than 64 bytes
  - Small messages are typical used for synchronization

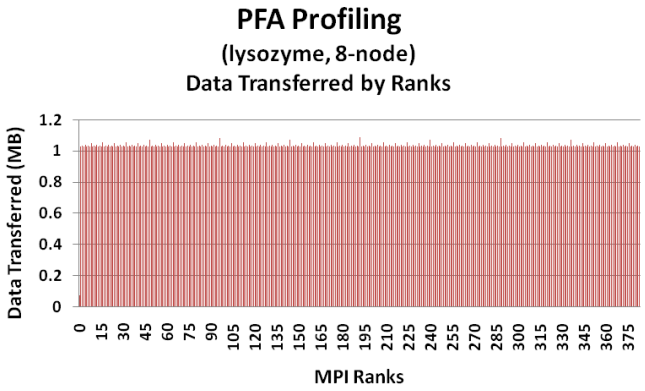


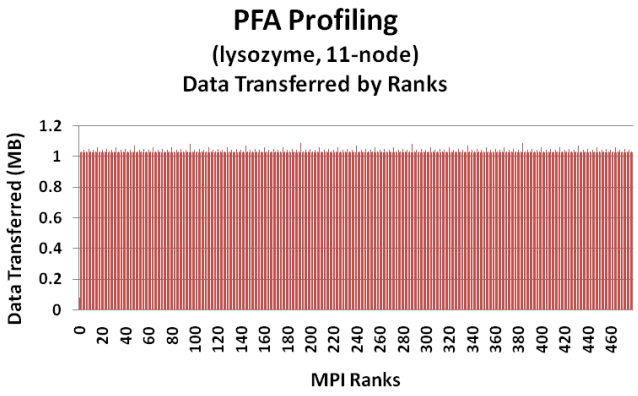
# PFA Profiling – Data Transfer By Process

- **Data transferred to each MPI rank is consistent for any number of processes**
  - Shows very little data transfers happened



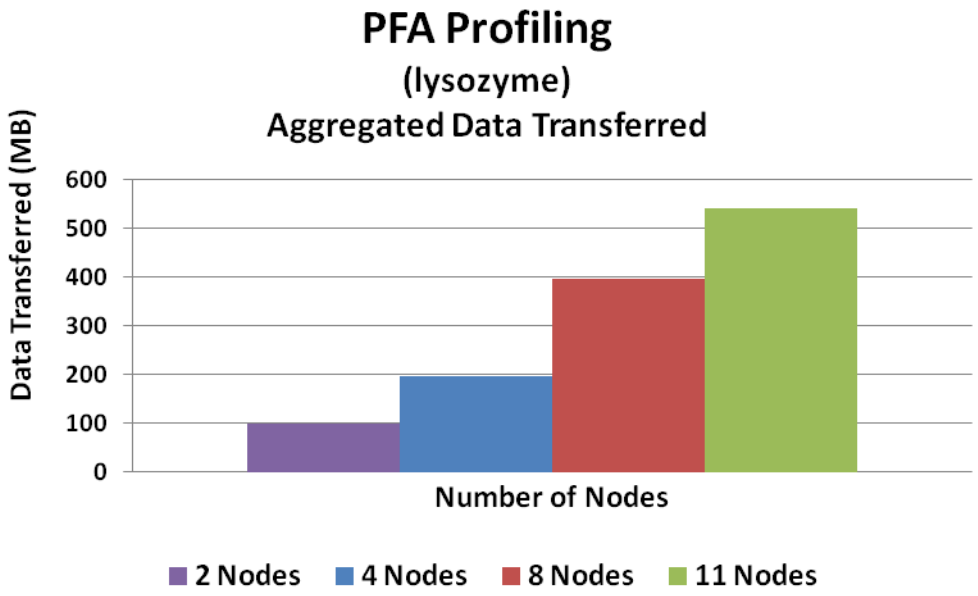



# PFA Profiling – Aggregated Data Xfer

- **Aggregated data transfer refers to:**
  - Total amount of data being transferred in the network between all MPI ranks collectively
- **The total data transfer steadily increases as the cluster scales**
  - As a compute node being added, more data communications will happen




***InfiniBand QDR***

# Summary

- **PFA is a compute intensive application that has high demand for CPU power**
  - Increasing the number of compute nodes in the job has a direct impact on job performance
  - Good scalability that allows spreading workload by utilizing additional compute nodes
- **CPU:**
  - Shows higher job productivity when using CPU with higher core frequency
- **Compilers:**
  - Compilers and tuning compiler flags have an impact on better job performance
  - Can benefit by having higher CPU frequency
  - Shows high sensitivity to network latency
- **MPI Communications:**
  - Over 90% of the time spend are in computation for a 11-node (528 procs) jobs
  - Majority of the MPI communications time happens in MPI_Barrier

# Thank You

## HPC Advisory Council