

# PFLOTRAN

## Performance Benchmark and Profiling

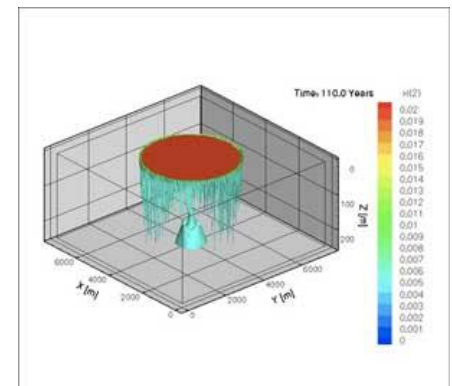
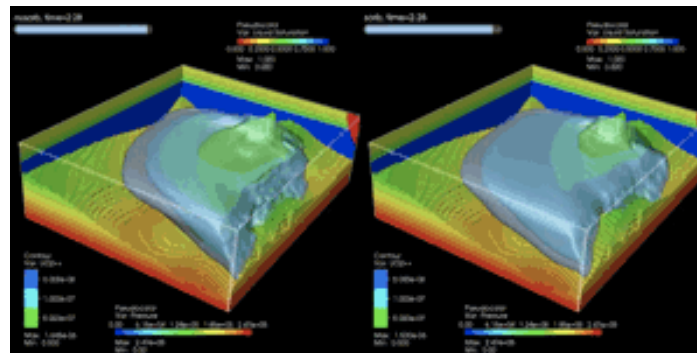
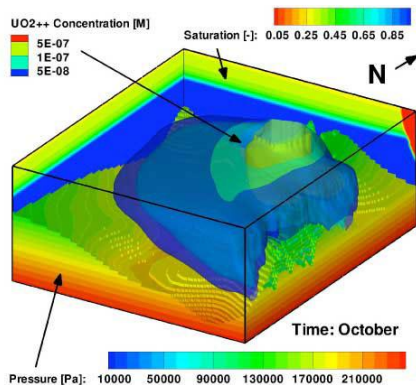
June 2011



- **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://ees.lanl.gov/source/orgs/ees/pflotran>

- **PFLOTRAN**

- Is used for modeling Multiscale-Multiphase-Multicomponent Subsurface Reactive Flows
- Builds on top of PETSc as the basis for its parallel framework, which includes:
  - Management of parallel data structures
  - Parallel solvers and preconditioners, and
  - Efficient parallel construction of Jacobian and residuals
- PFLOTRAN can be applied to a variety of important problems, such as geologic CO<sub>2</sub> sequestration, radionuclide migration and others
- PFLOTRAN is an open source application and is freely redistributable under the LGPL (Lesser Gnu Public License)



- **The following was done to provide best practices**
  - PFLOTRAN performance benchmarking
  - Interconnect performance comparisons
  - Understanding PFLOTRAN communication patterns
  - Ways to increase PFLOTRAN productivity
  - MPI libraries comparisons
  
- **The presented results will demonstrate**
  - The scalability of the compute environment
  - The capability of PFLOTRAN to achieve scalable productivity
  - Considerations for performance optimizations

- **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**
- **Mellanox Fabric Collective Accelerator™ (FCA™) version 2.1**
- **Fulcrum based 10Gb/s Ethernet 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**
- **Application: PFLOTRAN (mercurial changeset 3592:b968bf14a133)**
- **Compiler and Tools: GNU Compilers 4.4, HDF5 1.8.7, PETSc (petsc-dev: 19086:5c056871a140)**
- **Benchmark workload: (examples\_problems/100\_100\_100/calcite)**
  - 3D test problem for diffusion of tracer and reaction of calcite in a cubic domain (MAX\_STEPS=50)

- **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

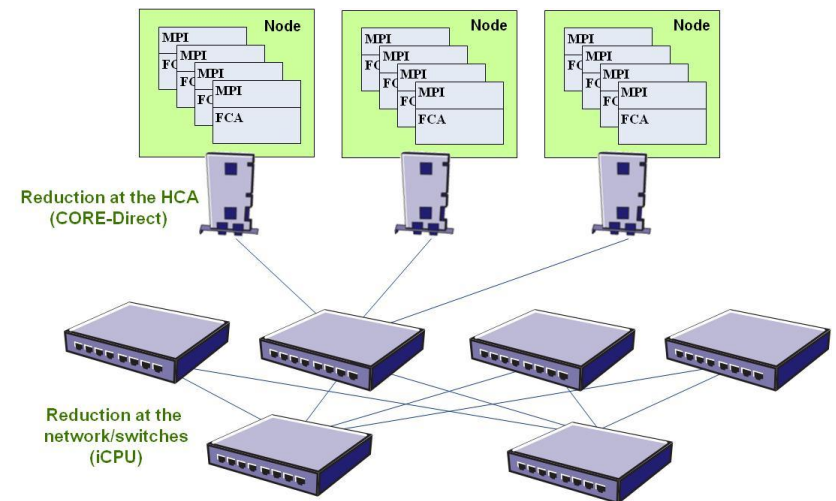
- 2x more memory and processing power in half of the space
- Energy optimized low flow fans, improved power supplies and dual SD modules

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

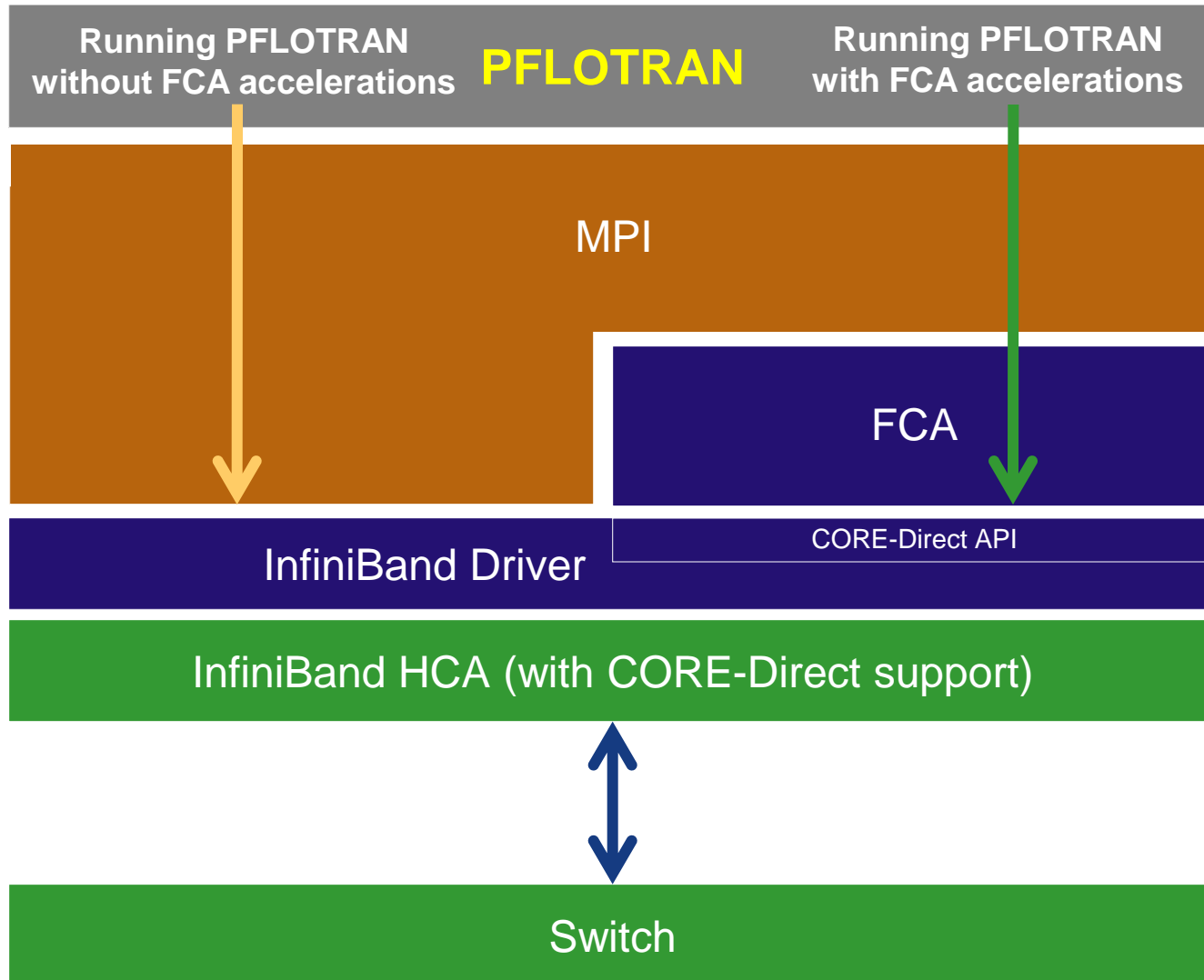
- System expansion
- Component upgrades and feature releases



- **Mellanox Fabric Collectives Accelerator (FCA)**
  - Utilized hardware accelerations on the adapter (CORE-Direct)
  - Utilized managed switches capabilities (iCPU)
  - Accelerating MPI collectives operations by offloading them to the network
  - The world first complete solution for MPI collectives offloads
  
- **FCA 2.1 supports accelerations/offloading for**
  - MPI Barrier
  - MPI Broadcast
  - MPI AllReduce and Reduce
  - MPI AllGather and AllGatherv

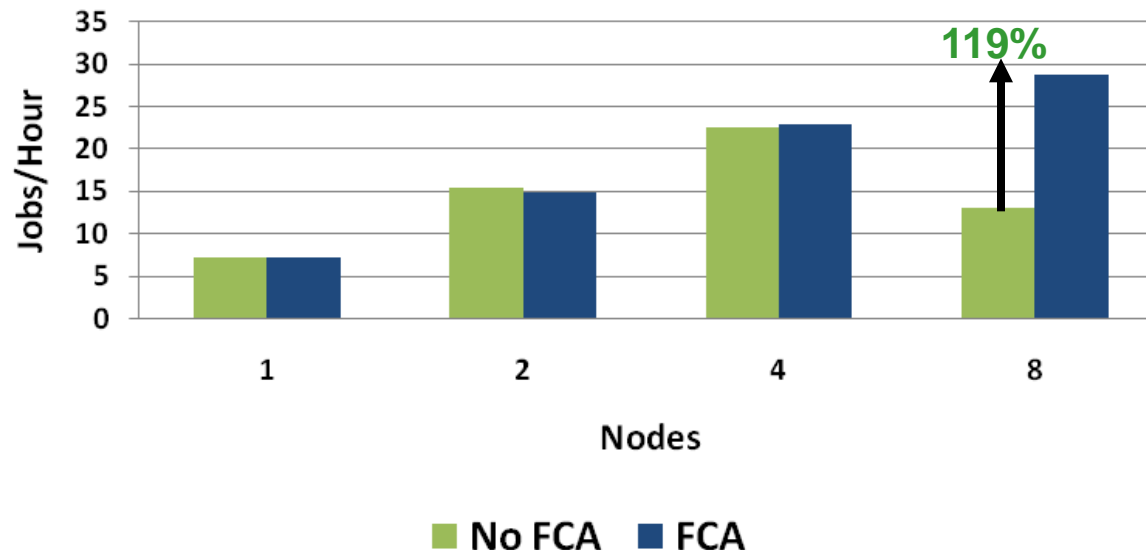


# Software Layers Overview



- **FCA enables nearly 119% performance gain at 8 nodes / 384 cores**
  - Bigger advantage expected at higher node count / core count
- **Open MPI Flags used:**
  - `-mca btl_openib_if_include mlx4_0,mlx4_1 -mca btl_sm_use_knem 1 --bind-to-core -mca coll_fca_enable 1`

**PFLOTRAN Benchmark**  
(100\_100\_100\_calcite)

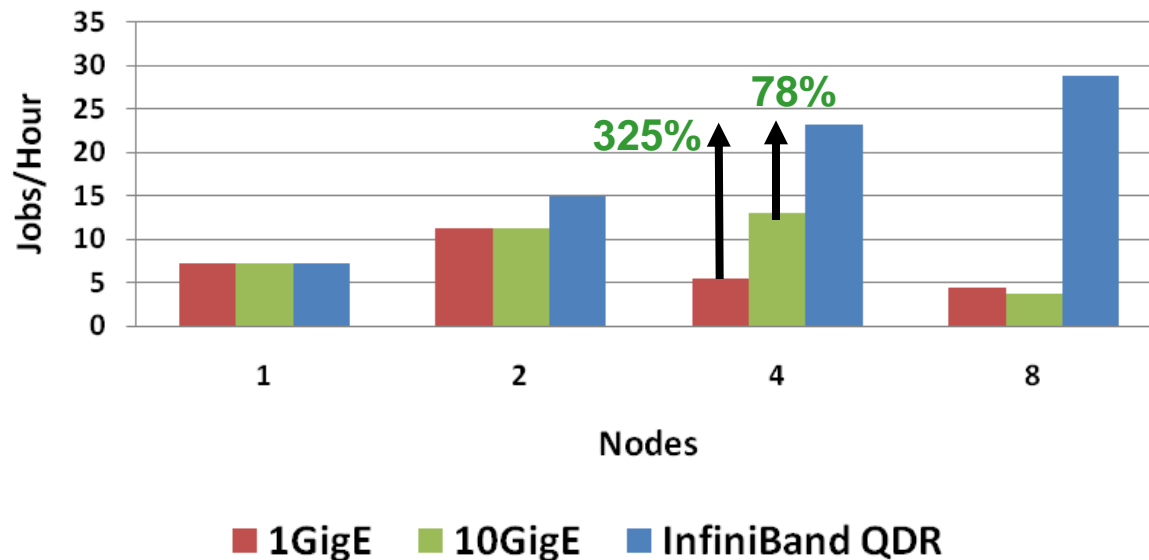


*Higher is better*

*Open MPI  
48 Cores/Node*

- **InfiniBand allows scalability for PFLOTRAN**
  - Up to 325% higher performance than 1GigE at 4-node
  - Up to 78% higher performance than 10GigE at 4-node
- **Ethernet performance does not scale beyond 2 node**
  - Ethernet performance drops significantly due to significant demand in the network

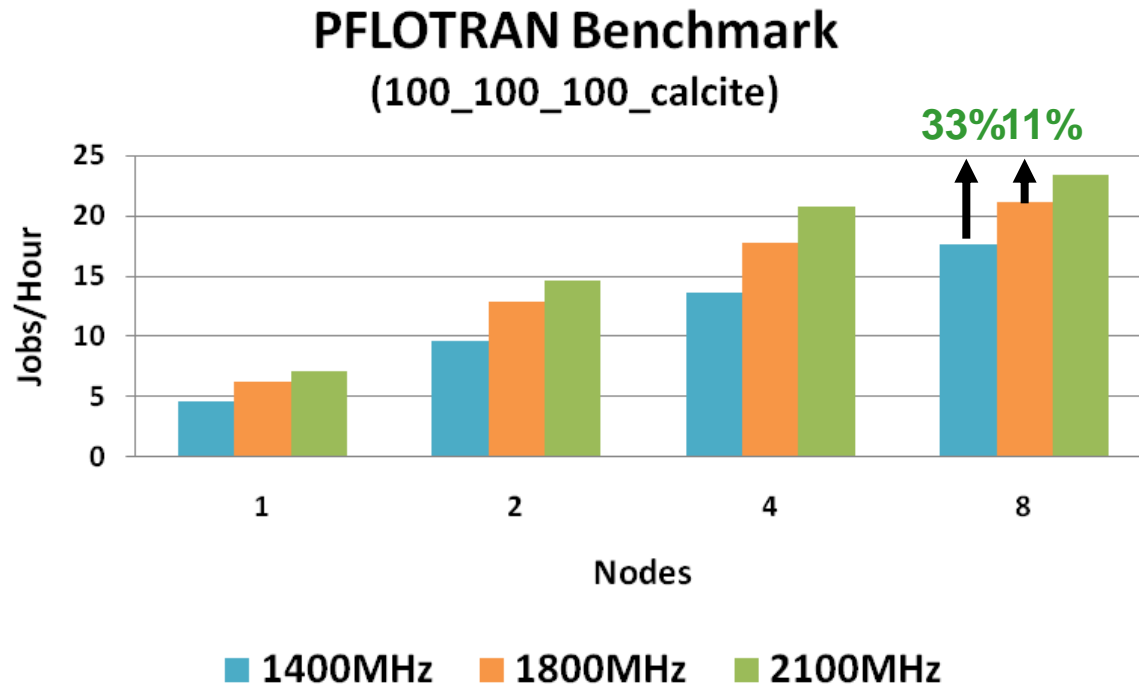
**PFLOTRAN Benchmark**  
(100\_100\_100\_calcite)



*Higher is better*

*Open MPI  
48 Cores/Node*

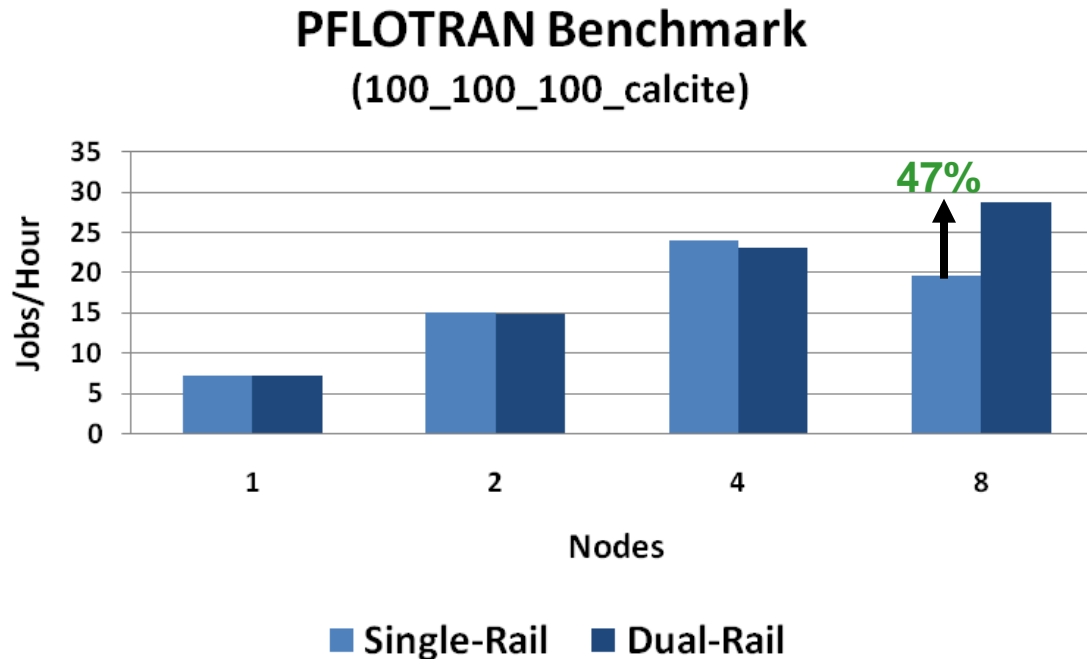
- **Increasing CPU core frequency enables higher job efficiency**
  - Up to 11% better job performance between 2200MHz vs 1800MHz on 8-node
  - Up to 33% better job performance between 2200MHz vs 1400MHz on 8-node



*Higher is better*

*QDR Dual-Rail  
48 Cores/Node*

- **Dual-rail (Dual InfiniBand cards) enables better performance than single-rail**
  - Up to 47% better job performance when equipped with 2 InfiniBand cards per node
  - Allows round robins of small messages between the InfiniBand cards

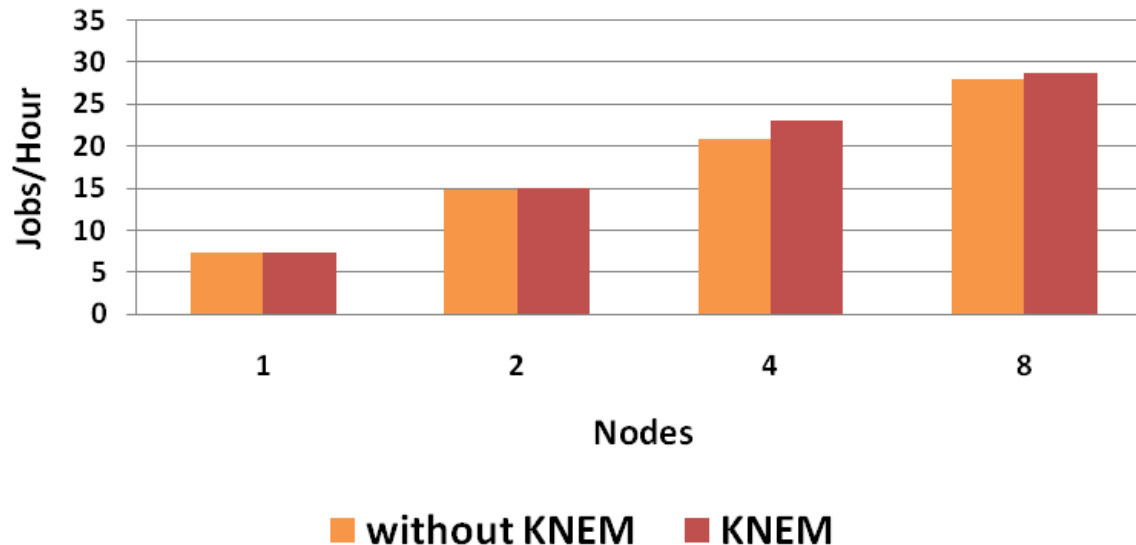


*Higher is better*

*Open MPI+FCA  
48 Cores/Node*

- **KNEM improves shared memory communications**
  - by using RDMA for intra-node communications for large messages
- **Observed only a slight improvement with running with KNEM**
  - PFLOTRAN does not benefit much from memory copy offloading for large messages
  - Profiling shows majority of MPI messages are small messages

**PFLOTRAN Benchmark**  
(100\_100\_100\_calcite)

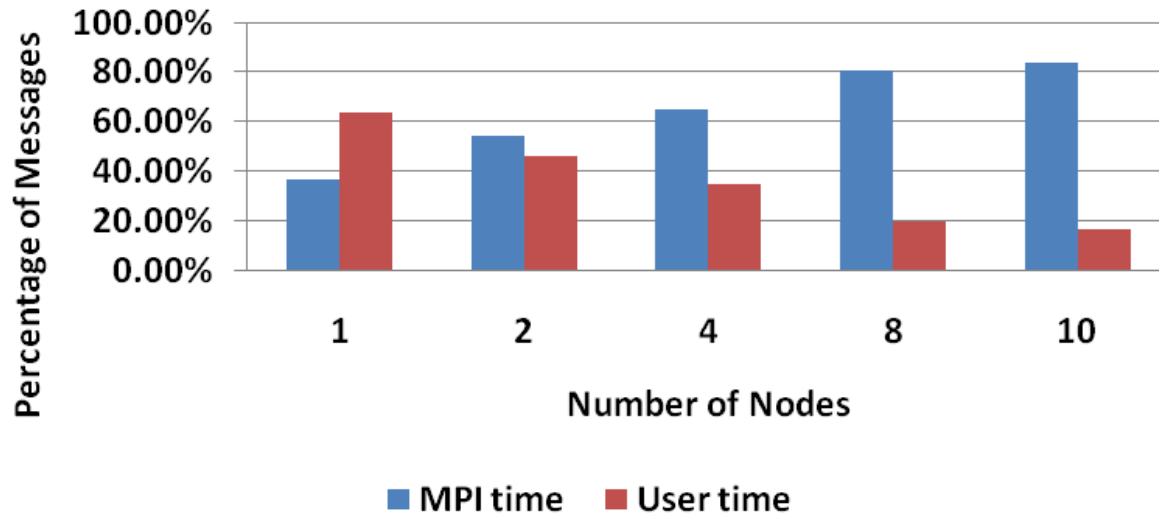


*Higher is better*

*Open MPI+FCA  
48 Cores/Node*

- **Significant increase in communications time as more nodes are added**
  - Compute time crosses over communication time at 2 nodes

**PFLOTRAN Profiling**  
(100\_100\_100\_calcite)  
MPI/User Time Ratio

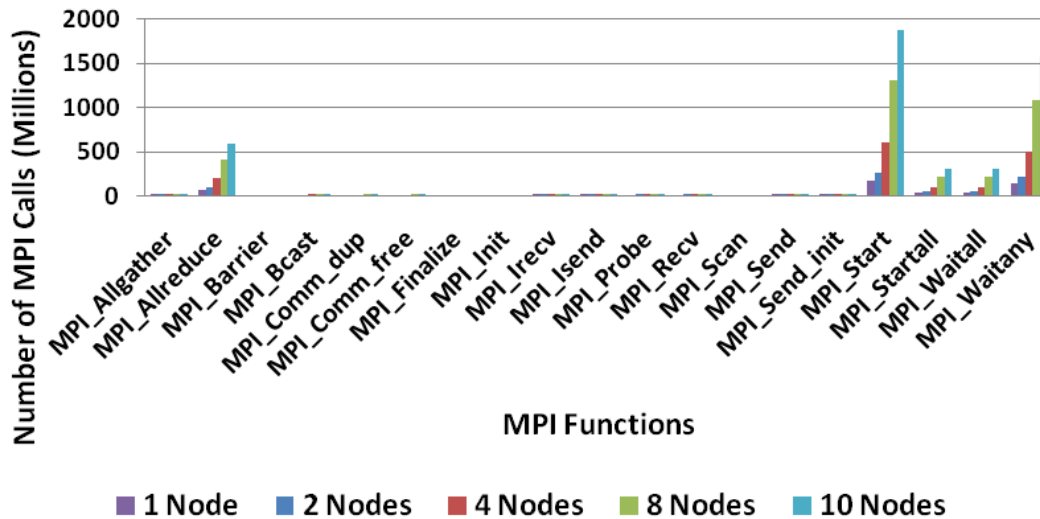


**48 Cores/Node**

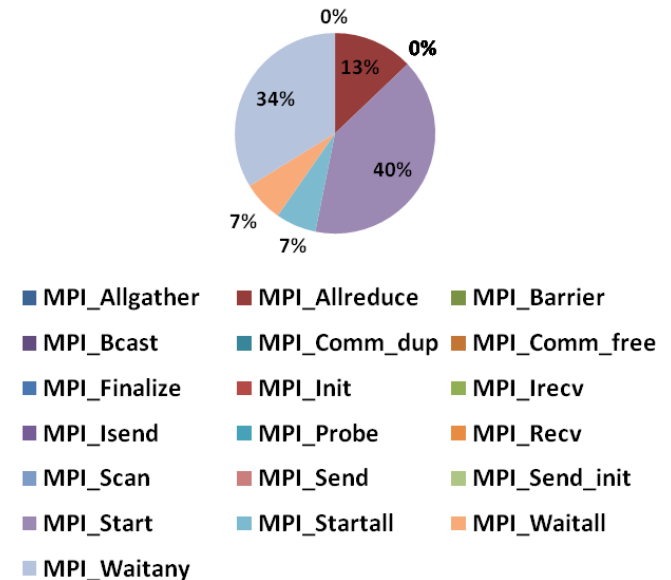
# PFLOTRAN Profiling – Number of MPI Calls

- **The most used MPI function is MPI\_Start**
  - MPI\_Start initiates a communication with a persistent request handle
  - Represents 40% of MPI calls used for 8-node
- **The top data communications call is MPI\_Allreduce**
  - Accounted for 13% of calls

**PFLOTRAN Profiling**  
(hanford\_infiltration)  
Number of MPI Calls



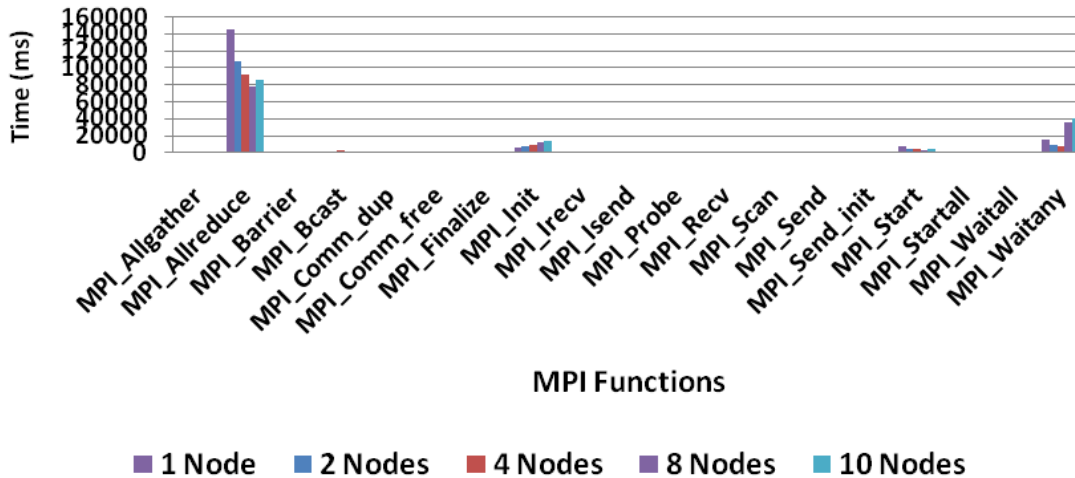
**PFLOTRAN Profiling**  
(100\_100\_100\_calcite, 8-node, InfiniBand)  
% MPI Calls



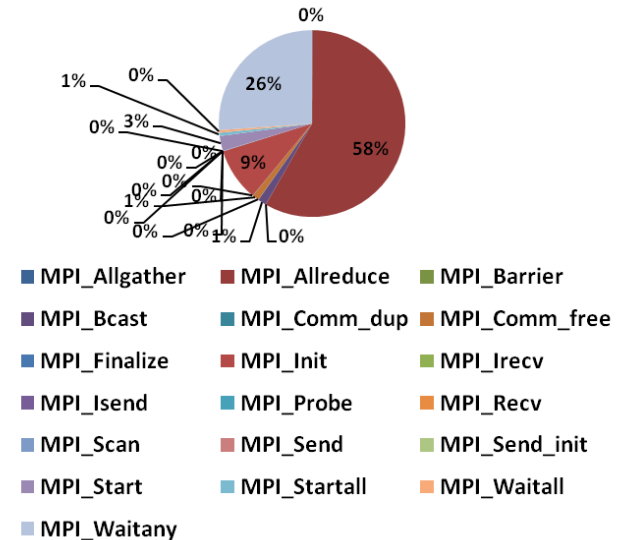
# PFLOTRAN Profiling – Time Spent of MPI Calls

- **The largest time consumer is MPI\_Allreduce for data communications**
  - Occupies 58% of all MPI time at 8-node
  - MPI\_Allreduce is a MPI collective operation that runtime can be reduced by FCA
- **The next largest time consumer are MPI\_Waitany**
  - Occupies 26% of all MPI time at 8-node

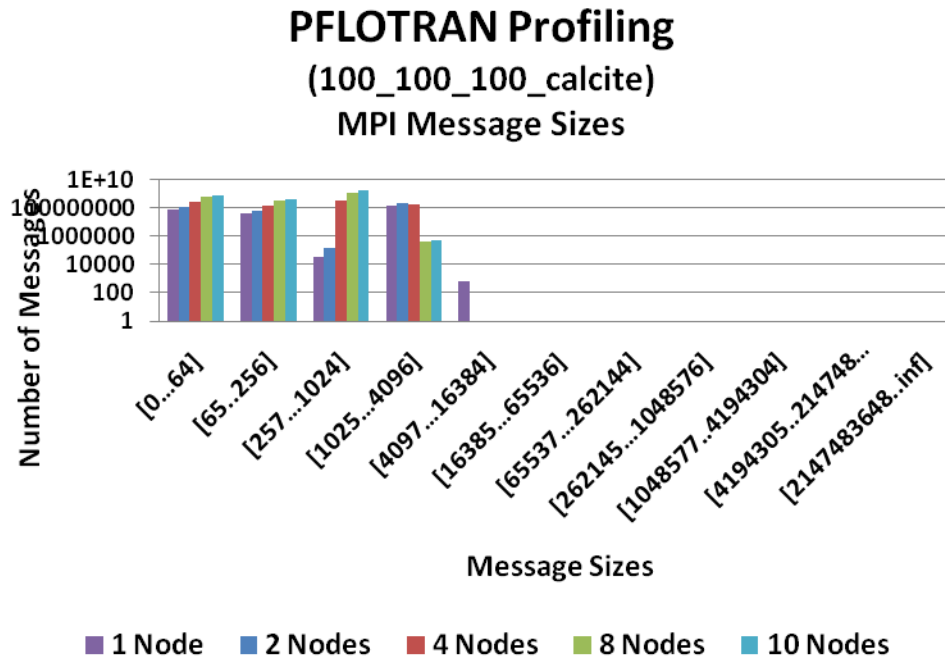
**PFLOTRAN Profiling**  
(100\_100\_100\_calcite)  
Time Spent of MPI Calls



**PFLOTRAN Profiling**  
(100\_100\_100\_calcite, 8-node)  
% Time Spent of MPI Calls

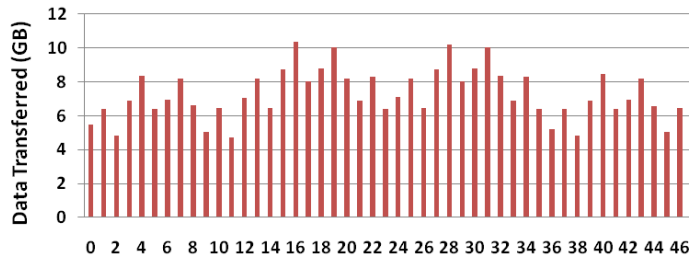


- **Majority of the MPI message sizes are small messages**
  - In the range of less than 4K
  - Small messages are typical used for synchronization, implies that app is latency sensitive

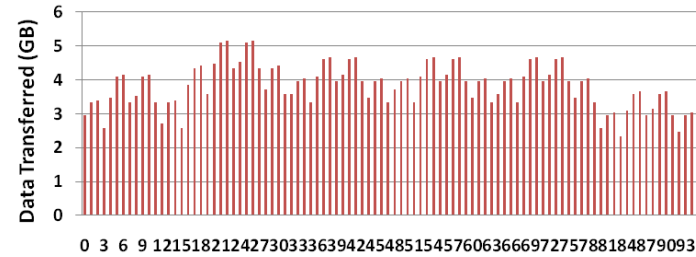


- **Data transferred to each MPI rank drops as cluster scales**
  - From a maximum of 10GB per rank at 1-node, drops to 5GB, 3.5GB, and eventually to 2.5GB
  - The amount of data transferred is spread across the cluster
- **As the cluster scales, less data is transferred to each rank**

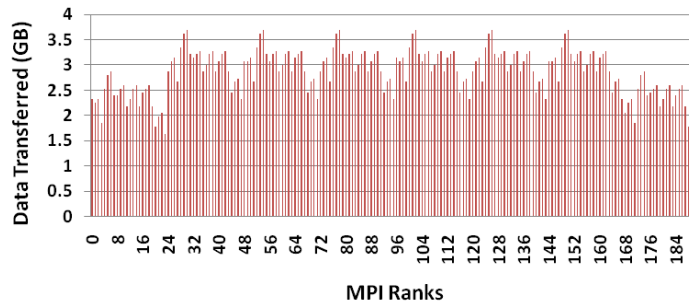
**PFLOTRAN Profiling**  
(100\_100\_100\_calcite, 1-node)  
Data Transferred by Ranks



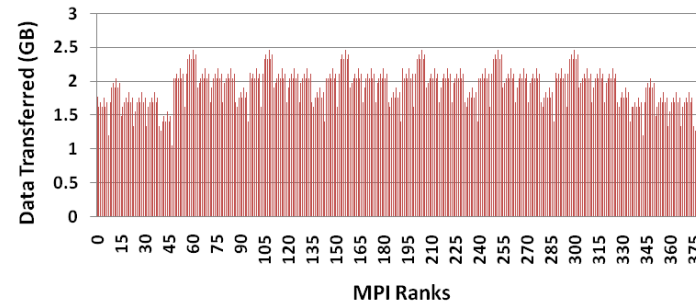
**PFLOTRAN Profiling**  
(100\_100\_100\_calcite, 2-node)  
Data Transferred by Ranks



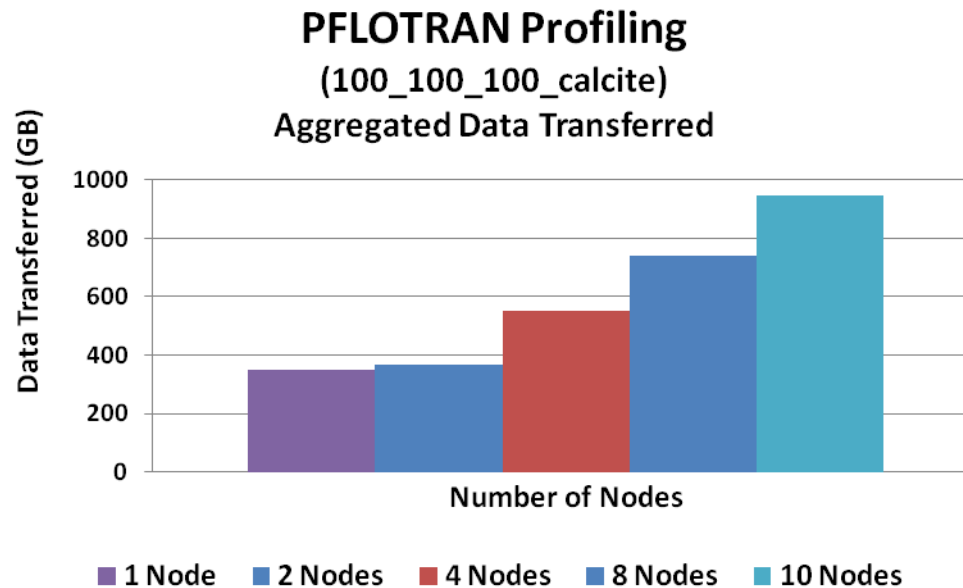
**PFLOTRAN Profiling**  
(100\_100\_100\_calcite, 4-node)  
Data Transferred by Ranks



**PFLOTRAN Profiling**  
(100\_100\_100\_calcite, 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 steadily increases as the cluster scales**
  - As a compute node being added, more data communications will happen



*InfiniBand QDR*

- **Networking:**
  - InfiniBand QDR is the only low latency interconnect that allows PFLOTRAN to scale
  - Shows high sensitivity to network latency
- **CPU:**
  - Shows gains in job productivity by using higher CPU frequency
- **MPI**
  - Using FCA with Open MPI to allow PFLOTRAN to offload MPI communication messages to the networking hardware
- **Data transfer on the network**
  - Majority of MPI messages are small messages less than 4KB
  - Significantly more data being transferred as the number of compute node increases
  - MPI\_Allreduce is a large share of communication which can be offloaded by FCA

# 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