OpenFOAM
Performance Benchmark and Profiling

April 2012
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.mellanox.com)
OpenFOAM Application

• OpenFOAM® (Open Field Operation and Manipulation) CFD Toolbox can simulate
  − Complex fluid flows involving
    • Chemical reactions
    • Turbulence
    • Heat transfer
  − Solid dynamics
  − Electromagnetics
  − The pricing of financial options

• OpenFOAM is Open source, produced by OpenCFD Ltd
Objectives

• The following was done to provide best practices
  – OpenFOAM performance benchmarking
  – Interconnect performance comparisons
  – Ways to increase OpenFOAM productivity
  – MPI libraries comparisons

• The presented results will demonstrate
  – The scalability of the compute environment to provide nearly linear application scalability
  – The capability of OpenFOAM to achieve scalable productivity
  – Considerations for performance optimizations
Test Cluster Configuration

- Dell™ PowerEdge™ R815 11-node (704-core) cluster
- AMD™ Opteron™ 6276 (code name “Interlagos”) 16-core @ 2.3 GHz CPUs
- 4 CPU sockets per server node
- Mellanox ConnectX®-3 InfiniBand Adapters
- Mellanox SwitchX™ 6036 36-Port InfiniBand switch
- Memory: 128GB memory per node DDR3 1333MHz
- OS: RHEL 6.2, MLNX-OFED 1.5.3 InfiniBand SW stack, FCA version 2.1
- MPI: Open MPI 1.5.5, Platform MPI 8.2
- Compilers: GNU Compilers 4.6.3
- Application: OpenFOAM 2.1.0
- Benchmark workload:
  - Datasets with 46 and 95 million cells using the simpleFoam (Steady-state solver for incompressible, turbulent flow)
• HPC Advisory Council Test-bed System

• New 11-node 704 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™ 6200 series platform and Mellanox ConnectX®-3 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 64 core/32DIMMs per server – 1344 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 allows OpenFOAM to scale at the highest rate
- Showing unlimited continuous gain to 8 nodes

Pure Ethernet protocol shows limited scalability
- The performance of 1GbE plummet after 2 nodes (128 processes)
- InfiniBand QDR provides 35% higher productivity than 10GbE for 46MIL dataset
- InfiniBand QDR provides 29% higher productivity than 10GbE for 95MIL dataset
OpenFOAM Performance – MPI

- Both MPIs perform at the same level for this dataset and solver
  - Performance shown by the 2 MPIs are equally as good
  - MPI profiling shows the solver based heavily on pure send and receive
  - Reflects that both MPI implementations performs those heavily used calls
  - Processor binding are enabled when running the job (No special tuning flags are used)

Higher is better

- **OpenFOAM Profiling**
  - (simpleFoam, 46MIL)
  - Performance (Jobs/Hour)

- **OpenFOAM Profiling**
  - (simpleFoam, 95MIL)
  - Performance (Jobs/Hour)

- **Number of Nodes**
  - 4
  - 8

- **MPI Functions**
  - 1
  - 2
  - 4
  - 8

- **Platforms**
  - Platform MPI 8.2
  - Open MPI 1.5.5

64 Cores/Node
OpenFOAM Performance – FCA

- FCA improvement based on the amount of time spent on MPI Collective ops
  - FCA shows ~6% gain for the simpleFoam solver
- The dataset and solver is based heavily on MPI sends and receives
  - Therefore the FCA gain is limited by the time spent on MPI collective ops
  - More gain is expected when more nodes are in use

**OpenFOAM Profiling**
(simpleFoam, 95MIL, Open MPI)

Performance (Jobs/Hour)

Number of Nodes

- 4 Nodes: Baseline: 5.0, FCA: 5.5
- 8 Nodes: Baseline: 8.0, FCA: 8.6

Higher is better

64 Cores/Node
OpenFOAM Performance – Processor Speeds

• Productivity gain is seen at higher CPU core frequency
  – Up to 7.6% gain in productivity for core speed at 2300MHz versus at 2100MHz

OpenFOAM Profiling
(simpleFoam, 95MIL)

Higher is better

64 Cores/Node
OpenFOAM Performance – FP Precision

- OpenFOAM allows configuring for either SP and DP for floating point precision
- Running at SP is shown to be faster than running at DP
  - Seen around 46% faster running at SP (Single Precision) versus DP (Double Precision)
  - All other slides are running using Single Precision

OpenFOAM Profiling
(simpleFoam, 95MIL)

Higher is better
OpenFOAM Profiling – MPI/User Time Ratio

- Computation time grows after than communication time
  - Even though both MPI and computation time would grow
  - This explains why computation time has a higher percentage for 95MIL case vs 46MIL case

Higher is better

64 Cores/Node
The most used MPI function are MPI_Irecv and MPI_Isend
- Each accounts for 46% of all the MPI calls made

The simpleFoam solver uses the non-blocking sends and receives heavily
- Purely point-to-point sends and receives are seen
- The non-blocking communication calls allows overlapping computation and communication
OpenFOAM Profiling – Time Spent of MPI calls

- **The most used MPI function is MPI_Waitall**
  - Accounts for 31% of the time spent in MPI for the 46MIL dataset
  - Accounts for 33% of the time spent in MPI for the 95MIL dataset

- **MPI Collective operations accounts for around 8-9% on a 8-node**
  - That amount of time can be the potential gain by FCA for collective acceleration
OpenFOAM Profiling – MPI Message Sizes

- Majority of the MPI message sizes are concentrated in the small to midrange
  - Highest in the range from 0B to 64B
- The larger dataset shows more messages are transferred

![OpenFOAM Profiling](image1)

![OpenFOAM Profiling](image2)
OpenFOAM Profiling – Data Transfer / Process

- Data transferred to each MPI rank shows high variance
  - No patterns can be seen with from the traffic
- As the cluster scales, less data is driven to each rank and each node
  - 1GB-5GB per rank in 1-node job versus 1.5GB per rank in a 8-node job
Aggregated data transfer refers to:
- Total amount of data being transferred in the network between all MPI ranks collectively
- The total data transfer increases as the cluster scales
- The larger the dataset is, more data will be sent to the network
Summary

- **InfiniBand allows OpenFOAM to scale at the highest rate**
  - Showing unlimited continuous gain to 8 nodes
  - InfiniBand QDR provides 35% higher productivity than 10GbE for 46MIL dataset

- **Both Open MPI or Platform MPI shows good performance**
  - No apparent difference in performance gain seen from one over another

- **FCA shows gain based on the amount of MPI Collective ops used**
  - Shows around 6% gain at 8 nodes, more gain expected on more nodes

- **Higher CPU core frequency enables higher performance**
  - Up to 7.6% gain in productivity for 2300MHz versus 2100MHz

- **Both CPU and MPI time would grow as the cells in the dataset grows larger**
  - The computation time grows faster than the communication time

- **MPI Communication type are mainly non-blocking for the simpleFoam solver**
  - Purely non-blocking point-to-point data send and receives are seen
Thank You

HPC Advisory Council