Scalability Performance Analysis of OpenFOAM on Modern HPC Clustering Technologies

Pak Lui
HPC Advisory Council
Agenda

• Introduction to HPC Advisory Council
• Benchmark Configuration
• Performance Benchmark Testing and Results
• Summary
• Q&A / For More Information
The HPC Advisory Council

Mission Statement

- World-wide HPC non-profit organization (418+ members)
- Bridges the gap between HPC usage and its potential
- Provides best practices and a support/development center
- Explores future technologies and future developments
- Leading edge solutions and technology demonstrations
Centers of Excellence
Special Interest Subgroups

- **HPC|Scale Subgroup**
  - Explore usage of commodity HPC as a replacement for multi-million dollar mainframes and proprietary based supercomputers

- **HPC|Cloud Subgroup**
  - Explore usage of HPC components as part of the creation of external/public/internal/private cloud computing environments.

- **HPC|Works Subgroup**
  - Provide best practices for building balanced and scalable HPC systems, performance tuning and application guidelines.

- **HPC|Storage Subgroup**
  - Demonstrate how to build high-performance storage solutions and their affect on application performance and productivity

- **HPC|GPU Subgroup**
  - Explore usage models of GPU components as part of next generation compute environments and potential optimizations for GPU based computing

- **HPC|Music**
  - To enable HPC in music production and to develop HPC cluster solutions that further enable the future of music production
HPC music is an advanced research project about High Performance Computing and Music Production dedicated to enable HPC in music creation. Its goal is to develop HPC cluster and cloud solutions that further enable the future of music production and reproduction.
HPC Advisory Council HPC Center

Dell™ PowerEdge™ R730 32-node cluster

Dell PowerVault MD3420
Dell PowerVault MD3460

InfiniBand Storage (Lustre)

HP Proliant XL230a
Gen9 10-node cluster

HP Cluster Platform 3000SL
16-node cluster

Dell™ PowerEdge™ C6145 6-node cluster

Dell™ PowerEdge™ R815 11-node cluster

Dell™ PowerEdge™ R720xd/R720 32-node cluster

Dell™ PowerEdge™ M610 38-node cluster

Dell™ PowerEdge™ C6100 4-node cluster

White-box InfiniBand-based Storage (Lustre)
Exploring All Platforms

X86, Power, GPU, FPGA and ARM based Platforms
140+ Applications Best Practices Published

- Abaqus
- AcuSolve
- Amber
- AMG
- AMR
- ABySS
- ANSYS CFX
- ANSYS FLUENT
- ANSYS Mechanics
- BQCD
- CCSM
- CESM
- COSMO
- CP2K
- CPMD
- Dacapo
- Desmond
- DL-POLY
- Eclipse
- FLOW-3D
- GADGET-2
- GROMACS
- Himeno
- HOOMD-blue
- HYCOM
- ICON
- Lattice QCD
- LAMMPS
- LS-DYNA
- miniFE
- MILC
- MSC Nastran
- MR Bayes
- MM5
- MPQC
- NAMD
- Nekbone
- NEMO
- NWChem
- Octopus
- OpenAtom
- OpenFOAM
- MILC
- OpenMX
- PARATEC
- PFA
- PFLOTRAN
- Quantum ESPRESSO
- RADIOSS
- SPECFEM3D
- WRF

For more information, visit: http://www.hpcadvisorycouncil.com/best_practices.php
HPCAC - ISC’15

Student Cluster Competition

• University-based teams to compete and demonstrate the incredible capabilities of state-of-the-art HPC systems and applications on the International Super Computing HPC (ISC HPC) show-floor

• The Student Cluster Challenge is designed to introduce the next generation of students to the high performance computing world and community

• Submission for Participation in ISC’16 SCC opens now
ISC'15 – Student Cluster Competition Teams
2015 HPC Advisory Council Conferences

Introduction

- HPC Advisory Council (HPCAC)
  - Application best practices, case studies
  - Benchmarking center with remote access for users
  - World-wide workshops
  - Value add for your customers to stay up to date and in tune to HPC market

- 2015 Conferences
  - USA (Stanford University) – February, 2015
  - Switzerland (CSCS) – March 31, 2015
  - Brazil (University of São Paulo) – August 26, 2015
  - Spain (BSC) – September 22, 2015
  - China (HPC China) – November 9, 2015
  - South Africa – December 2015

- For more information
  - [www.hpcadvisorycouncil.com](http://www.hpcadvisorycouncil.com)
  - info@hpcadvisorycouncil.com
HPCAC Conferences 2015

Conferences

HPC Advisory Council
Brazil Conference 2015
August 26, 2015
Laboratorio Nacional de Computação Científica (LNCC)
Av. dos Escravos, 888
13083-000, São Paulo, Brazil
www.esi-group.com

HPC Advisory Council
Spain Conference 2015
September 22, 2015
Campus Diagonal Nord
Barcelona, Spain

HPC Advisory Council
Switzerland Conference 2015
March 22-25, 2015
Lugano Convention Center
Lugano, Switzerland

HPC Advisory Council
Stanford Conference 2015
February 2-3, 2015
Munger Conference Center and Center Hall
Stanford University
Second RDMA Competition

China (July – October)

• 28 universities participated
• Programming Applications to utilize RDMA technology
• Opportunity to influence the future of data centers
• Create research opportunities for the universities
2015 HPC Advisory Council Conferences

• USA (Stanford University) – February 2015
• Switzerland – March 2015
• Brazil – August 2015
• Spain – September 2015
• China (HPC China) – November 2015
• South Africa – December 2015
Note

• The following research was performed under the HPC Advisory Council activities
  ‣ Participating vendors: Intel, Dell, Mellanox
  ‣ Compute resource - HPC Advisory Council Cluster Center
• The following was done to provide best practices
  ‣ OpenFOAM performance overview
  ‣ Understanding OpenFOAM communication patterns
  ‣ Ways to increase OpenFOAM productivity
  ‣ MPI libraries comparisons
• For more info please refer to
  ‣ http://www.dell.com
  ‣ http://www.intel.com
  ‣ http://www.mellanox.com
  ‣ http://www.openfoam.org
Objectives

• The following was done to provide best practices
  ‣ OpenFOAM performance benchmarking
  ‣ Interconnect performance comparisons
  ‣ MPI performance comparison
  ‣ Understanding OpenFOAM communication patterns

• 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
Test Cluster Configuration 1 - “Ivy Bridge” cluster

Dell™ PowerEdge™ R720xd 32-node (640-core) “Jupiter” cluster

- Dell™ PowerEdge™ R720xd 32-node “Jupiter” cluster
  - Dual-Socket 10-core Intel E5-2680v2 @ 2.80 GHz CPUs
  - Memory: 64GB memory, DDR3 1600 MHz
  - OS: RHEL 6.2, MLNX_OFED 2.3-1.0.1 InfiniBand SW stack
  - Hard Drives: 24x 250GB 7.2 RPM SATA 2.5” on RAID 0
- Intel Cluster Ready certified cluster
- Mellanox Connect-IB FDR InfiniBand adapters
- Mellanox ConnectX-3 FDR InfiniBand and Ethernet adapters
- Mellanox SwitchX SX6036 InfiniBand and Ethernet VPI switch
- MPI: HPC-X v1.2.0-250 (based on Open MPI 1.8.2rc6) with FCA 2.5 and MXM 3.2
- Application: OpenFOAM 2.2.2
- Benchmark datasets:
  - Lid Driven Cavity Flow - 1 Million elements, DP, 2D icoFoam solver for laminar, isothermal, incompressible flow
  - Automotive data – 670 Million elements, DP, pimpleFoam solver
OpenFOAM Performance

Setup

• Modify input files
  ‣ Constant/polyMesh/blockMeshDict
  ‣ System/decomposeParDict -> set ‘numberOfSubdomains’

• Create mesh
  ‣ blockMesh
  ‣ decomposePar

• Run solver
  ‣ mpirun -np XYZ <optional list of arguments> <Solver> -parallel

• Common causes for scalability woes
  ‣ Problem size too small, not enough work for the processes to be efficiently used
  ‣ Bottleneck in systems (Interconnect, file system being too slow)

Higher is better

Performance Rating = Jobs/Day
OpenFOAM Performance

icoFoam Solver - Interconnects

- FDR InfiniBand provides better scalability performance than Ethernet
- 544% better performance than 10GbE at 16 nodes
- 179% better performance than 1GbE at 16 nodes
- Ethernet would have limited scalability;
  - 4 IB connected nodes shows higher performance as 8 Ethernet connected nodes
OpenFOAM Performance

icoFoam Solver – QDR InfiniBand vs FDR InfiniBand

• FDR runs at 56Gbps while QDR runs at 40Gbps
• FDR InfiniBand delivers better application performance
  ‣ Up to 27% better performance than InfiniBand QDR
  ‣ Using Mellanox ConnectX-3 PCIe Gen3 in FDR mode and QDR mode
OpenFOAM Profiling

icoFoam Solver - % MPI Time

- MPI profiling clearly shows large time usage in MPI collective op’ns
- MPI_Allreduce accounts for 79% to 85% of all MPI time for icoFoam
- Tuning MPI libraries for MPI collective offloading related to collective op’ns
  ‣ Will greatly influence the system performance for the icoFoam solver
OpenFOAM Performance

icoFoam Solver - Open MPI Tuning with Collective Offload

• FCA enables nearly 51% performance gain at 16 nodes
• More advantage expected at higher node count / core count
  ‣ Normally FCA is enabled for >64 cores; FCA is enabled for all processes shown below
• FCA supports accelerations/offloading
  ‣ MPI_BARRIER, MPI_BROADCAST, MPI_ALLREDUCE, MPI_REDUCE, MPI_ALLGATHER and MPI_ALLGATHERV
  ‣ To enable FCA at runtime: “--mca coll_fca_enable 1 --mca coll_fca_np 0”
  ‣ Both cases at runtime: “--bind-to-core -mca btl openib,sm,self”

![OpenFOAM Performance](chart)

Higher is better

Jupiter Cluster
OpenFOAM Performance

icoFoam Solver - MPI Comparisons

- Tuned Open MPI has significant advantage on CPU larger core count
- Provided 51% of greater scalability than the un-tuned Open MPI at 16-node
- Outperformed other commercial MPIs by have MPI collective offloads
  - Due to offloading the MPI_Reduce to hardware

![OpenFOAM Performance Chart](attachment:image.png)

*Higher is better*
Test Cluster Configuration 2 - “Haswell cluster”

Dell PowerEdge R730 32-node (896-core) “Thor” cluster

- Dell PowerEdge R730 32-node (896-core) “Thor” cluster
  - Dual-Socket 14-Core Intel E5-2697v3 @ 2.60 GHz CPUs
  - BIOS: Maximum Performance, Home Snoop
  - Memory: 64GB memory, DDR4 2133 MHz
  - OS: RHEL 6.5, MLNX_OFED_LINUX-3.0-2.0.1 InfiniBand SW stack
  - Hard Drives: 2x 1TB 7.2 RPM SATA 2.5” on RAID 1
- Mellanox ConnectX-4 EDR 100Gb/s InfiniBand Adapters
- Mellanox Switch-IB SB7700 36-port EDR 100Gb/s InfiniBand Switch
- Mellanox ConnectX-3 FDR VPI InfiniBand and 40Gb/s Ethernet Adapters
- Mellanox SwitchX-2 SX6036 36-port 56Gb/s FDR InfiniBand / VPI Ethernet Switch
- Dell InfiniBand-Based Lustre Storage based on Dell PowerVault MD3460 and MD3420
  - Based on Intel Enterprise Edition for Lustre (IEEL) software version 2.1
- NVIDIA Tesla K40 (on 8 Nodes) and NVIDIA Tesla K80 GPUs (on 2 Nodes); 1 GPU per node
- MPI: Mellanox HPC-X v1.3 (based on Open MPI 1.8.8) with CUDA 6.5 and 7.0 support
- Application: OpenFOAM 2.2.2
- Benchmark: Automotive data – 670 Million elements, DP, pimpleFoam solver
OpenFOAM Profiling

pimpleFoam Solver - Number of MPI Calls

- The pimpleFOAM solver is the transient solver for incompressible flow
- Mostly in non-blocking pt-to-pt communication, plus other MPI operations
- MPI_Irecv, MPI_Isend and MPI_Waitall are almost used exclusively
- MPI_Irecv, MPI_Isend (45% each), MPI_Allreduce (6%) at 896 cores
OpenFOAM Profiling

pimpleFoam - % MPI Time

- MPI profiling reports large consumption in large messages
  - MPI_Probe (blocking test) accounts for 48% of all MPI time
  - MPI_Recv consumes 26% of MPI time, follows by MPI_Waitall (20%)
- MPI_Probe calls are concentrated in large messages
- Interconnect with higher large-message bandwidth would be advantageous for the pimpleFoam performance
OpenFOAM Performance
pimpleFoam Solver – EDR vs FDR InfiniBand

• EDR IB runs at 100Gbps on a single link while FDR IB runs at 56Gbps
• EDR InfiniBand delivers better application performance
  ‣ EDR InfiniBand Up to 13% better performance than EDR InfiniBand
  ‣ Due to the large concentration of large messages of MPI_Probe calls
OpenFOAM Performance

pimpleFoam Solver - MPI Tuning

- Transport method for MPI provide significant advantage on large core counts
- Mellanox HPC-X shows 13% higher performance than Intel MPI at 896 cores
  - HPC-X uses MXM UD transport and hcoll for collective offload
  - Intel MPI uses OFA and DAPL as the transport

![OpenFOAM Performance Diagram](chart.png)

Higher is better

Thor Cluster
OpenFOAM Performance
pimpleFoam Solver - MPI Tuning

• Cluster with Intel E5-2697v3 (Haswell) outperforms prior cluster generation
  ‣ 2 configurations (Jupiter – Ivy Bridge based) and (Thor – Haswell based) are compared
• Thor cluster performs 41% better than the Jupiter cluster at 32 nodes
• Performance difference likely comes from the increase of CPU cores (by 40%)
  ‣ InfiniBand is designed to handle communication traffic due to additional cores
OpenFOAM Performance

pimpleFoam Solver - MPI Tuning

• Tuning MPI streamlines IB messaging, thus accelerates performance
• MXM accelerate the pimpleFoam workload by 24% on a 32-node/640-core simulation
• Slashed 1.5 hours from a 7.3 hours simulation by just enabling messaging acceleration
• Benefit of FCA depends on the usage of MPI collective operations in solvers
• To enable MXM at runtime: “-mca pml yalla -mca mtl mxm -mca mtl_mxm_np 0”
OpenFOAM Summary

- OpenFOAM scales well to ~1000 of CPU cores with a suitable hardware setup
  - OpenFOAM scales well for both simple tutorial cases and for complex industrial cases
- InfiniBand delivers the best application performance for OpenFOAM
  - icoFoam: FDR IB delivers ~27% higher performance than QDR InfiniBand at 16 nodes
  - Ethernet solution would not scale beyond a few nodes
  - pimpleFoam: EDR IB outperforms FDR IB by 13% at 32 nodes/896 cores due to concentration of large msg
- OpenFOAM MPI profiling
  - icoFoam: Time used by MPI accounts for 50% of total runtime at 16 nodes
  - icoFoam: MPI_Allreduce accounts for 79% to 85% of all MPI time
  - pimpleFoam: MPI_Irecv, MPI_Isend (45% each), MPI_Allreduce (6%) at 896 cores
- MPI tuning packages (such as FCA and MXM) are proven to accelerate OpenFOAM applications
  - icoFoam: FCA provided nearly 51% faster runtime at 16 nodes for OpenFOAM with Open MPI
  - pimpleFoam: MXM streamlined InfiniBand messaging, accelerated workload by 24% on 32 nodes
  - These performance improvement are expected to boost at larger scale
  - Other OpenFOAM solvers can benefit by utilizing these MPI tuning packages
Contact Us

- Web: [www.hpcadvisorycouncil.com](http://www.hpcadvisorycouncil.com)
- Email: [info@hpcadvisorycouncil.com](mailto:info@hpcadvisorycouncil.com)
- Twitter: [www.twitter.com/hpccouncil](http://www.twitter.com/hpccouncil)
- YouTube: [www.youtube.com/user/hpcadvisorycouncil](http://www.youtube.com/user/hpcadvisorycouncil)