Computational Fluid Dynamics (CFD) Simulations at Scale
OpenFOAM open source applications

HPC|Scale Working Group, Sep 2010
P. Lui**, G. Shainer*, B. Sparks*, T. Liu**, N. Eicker***
*HPC Advisory Council, USA, **Mellanox Technologies, USA, ***Jülich Supercomputing Centre, Germany
The following research was performed under the HPC Advisory Council HPC|Scale working group:

- CFD simulation performance at scale
  - Using OpenFOAM application
- The effect of non-blocking vs blocking network configurations
- Scalability comparison among networking technologies
  - InfiniBand DDR
  - InfiniBand QDR non-blocking configuration
  - InfiniBand QDR 2:1 blocking configuration
  - Cray XT5
HPC Advisory Council Mission

• World-wide HPC organization (170+ members)
• Bridges the gap between HPC usage and its potential
• Provides best practices and a support/development center
• Explores future technologies and future developments
• Explores advanced topics – HPC|Cloud, HPC|Scale, HPC|GPU, HPC|Storage
• Leading edge solutions and technology demonstrations
• For more info: http://www.hpcadvisorycouncil.com
HPC Advisory Council Members
HPC Advisory Council HPC Center

Vesta
- Dell PowerEdge R815 6-node cluster
- Four processors AMD Opteron 6172 (Magny-Cours), 48 Cores per node
- Dual Mellanox ConnectX-5e 400Gb/s InfiniBand adapters per node
- Mellanox 36-Port 40Gb/s InfiniBand Switch
- Memory 128 GB, 1333 MHz memory per node

Messa
- Dell PowerEdge M610 14-node cluster
- Six-Core Intel® Xeon® processor X5670 @ 2.93 GHz
- Mellanox ConnectX-5e-2 40Gb/s InfiniBand adapters
- Mellanox 26-Port 40Gb/s InfiniBand Switch
- Memory: 240GB memory per node

Plutus
- HP Cluster Platform 3000SL
- 16 nodes HP ProLiant SL2Lz170z scalable servers
- Six-Core Intel® Xeon® processor X5670 @ 2.93 GHz
- Mellanox ConnectX-5e-2 40Gb/s InfiniBand adapters
- Mellanox 36-Port 40Gb/s InfiniBand Switch
- Memory: 240GB memory per node

Venus
- SUN 2200 8-node duster
- Intel Xeon Quad-core X5472 2.93GHz
- Mellanox InfiniBand ConnectX® 20Gb/s InfiniBand adapter
- Mellanox 20Gb/s InfiniBand Switch
- Memory: 220GB

Janus
- Dell PowerEdge M610 16-node cluster
- Quad-Core Intel® Xeon® processor X5570 @ 2.93 GHz
- Intel Cluster Ready certified cluster
- Mellanox ConnectX® MCQH29-IXC 4X 400Gb/s InfiniBand mezzanine card
- Mellanox M361Q 36-Port 40Gb/s InfiniBand Switch
- Memory: 240GB memory per node

Hedion
- Dell PowerEdge™ M803 11-node cluster
- Quad-Core AMD Opteron™ 2380 ("Shanghai") CPUs
- Mellanox ConnectX® 20Gb/s InfiniBand mezz card
- Mellanox 20Gb/s InfiniBand Switch Module
- Memory: 1GB memory, DDR2 800MHz per node

Helios
- Mellanox ConnectX® 20Gb/s InfiniBand technology
- 32 Rackable Systems c1000 DC powered rack-mount servers
- 04 Quad-Core Intel® Xeon® 3300 SeriesProcessors
- Light weight 30 AWG InfiniBand cables from W. L. Gore & Associates, Inc.
- 16Gb Clusterware™ HPC cluster management
- 8GB FB-DIM memory from WeiTeC Industries

Osiris
- Dell PowerEdge SC 1435 24-node cluster
- Quad-Core AMD Opteron™ 2382 SE CPUs
- Mellanox InfiniBand ConnectX® HCAs
- Mellanox InfiniBand 36-port switch
- Memory: 16GB memory, DDR2 800MHz per node
Special Interest Subgroups

**HPC|Scale Subgroup**

This group’s mission is to explore usage of commodity HPC as a replacement for multi-million dollar mainframes and proprietary based supercomputers with networks and clusters of microcomputers acting in unison to deliver high-end computing services.

**HPC|Cloud Subgroup**

This group’s mission is to explore usage of HPC components as part of the creation of external/public/internal/private cloud computing environments.

**HPC|Works Subgroup**

This group’s mission is to provide best practices for building balanced and scalable HPC systems, performance tuning and application guidelines.

**HPC|Storage Subgroup**

This group’s mission is to demonstrate how to build high-performance storage solutions and their affect on application performance and productivity. One of the main interests of the HPC|Storage subgroup is to explore Lustre based solutions, and to expose more users to the potential of Lustre over high-speed networks.

**HPC|GPU Subgroup**

This group’s mission is to explore usage models of GPU components as part of next generation compute environments and potential optimizations for GPU based computing.
University Award Program

- **University award program**
  - Universities are encouraged to submit proposals for advanced research around high-performance computing
  - Twice a year, the HPC Advisory Council will select a few proposals

- **Selected proposal will be provided with:**
  - Exclusive computation time on the HPC Advisory Council’s Compute Center
  - Invitation to present the research results in one of the HPC Advisory Council’s worldwide workshops, including sponsorship of travel expenses (according to the Council Award Program rules)
  - Publication of the research results on the HPC Advisory Council website and related publications
  - Publication of the research results and a demonstration if applicable within HPC Advisory Council world-wide technology demonstration activities

- **Proposals for the 2011 HPC Advisory Council University Award Program can be submitted between August 1, 2010 through October 30, 2010. The selected proposal(s) will be determined by November 15th and the winner(s) will be notified.**
Joining the HPC Advisory Council

www.hpcadvisorycouncil.com

info@hpcadvisorycouncil.com
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
Low Scale Testing

- **HP ProLiant SL2x170z G6 16-node cluster**
  - Six-Core Intel X5670 @ 2.93 GHz CPUs
  - Memory: 24GB per node
  - OS: CentOS5U4, OFED 1.5.1 InfiniBand SW stack
- **Mellanox ConnectX-2 VPI adapters for 40Gbps InfiniBand and 10GigE**
- **Mellanox InfiniBand switch**
- **Fulcrum based 10GigE switch**
- **MPI: Open MPI 1.4.1, Platform MPI 7.1**
- **Application: OpenFOAM 1.7.1**
- **Benchmark Workload**
  - Lid-driven cavity flow (2000x2000)
OpenFOAM Performance - Interconnect

- **Input Dataset**
  - Lid-driven cavity flow (4 million cells)

- **InfiniBand QDR enables linear scalability**
  - 83% higher performance than GigE at 16 nodes
  - 44% higher performance than 10GigE at 8 nodes
  - Performance advantage extends as cluster size increases

- **InfiniBand reduces power/job**
  - by 40% or more compared to GigE

---

OpenFOAM Benchmark
(Lid-driven cavity flow)

![OpenFOAM Benchmark Graph](chart)

- **Jobs/day**
  - Higher is better

- **12 Cores/Node**
OpenFOAM Performance – MPI Libraries

- **Input Dataset**
  - Lid-driven cavity flow (4 million cells)

- **Performance comparison**
  - Platform MPI enables 10% higher performance versus Open MPI at 16 nodes

**OpenFOAM Benchmark**

*(Lid-driven cavity flow)*

```
Number of Nodes

<table>
<thead>
<tr>
<th>Jobs/day</th>
<th>Open MPI</th>
<th>Platform MPI</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>10%</td>
<td></td>
</tr>
</tbody>
</table>
```

12 Cores/Node

Higher is better
OpenFOAM MPI Profiling – MPI Time

- **MPI_Allreduce** generates most communication overhead

12 Cores/Node
Large Scale Testing Environments

- **JuRoPa supercomputer at the Jülich Supercomputing Centre**
  - Dual socket Intel Xeon X5570 quad-core @ 2.93 GHz
  - 24 GB memory (DDR3, 1066 MHz)
  - Mellanox InfiniBand QDR, non-blocking network configuration
  - SUSE SLES 11
  - ParTec MPI

- **Itasca supercomputer at the Minnesota Supercomputing Institute**
  - HP ProLiant BL280c G6 blade servers
  - Dual socket Intel Xeon X5560 quad-core @ 2.80 GHz
  - Mellanox InfiniBand QDR, 2:1 blocking network configuration
  - SUSE SLES 11
  - OpenMPI-1.4.2, Platform MPI 7.1

- **Application**
  - OpenFOAM 1.7.1
  - Benchmark Dataset: Laminar Cavity Flow (2D, 16Million Cells)
Non-Blocking vs Blocking Networks

- **Non-Blocking Network**
  - Enables full wire speed or full bandwidth from the network adapter between any server nodes
  - Provides equal (or higher) throughput within the network as a sum of throughput capability from each node

- **Blocking Network**
  - Oversubscribed networks reduce the maximum available throughput in the network by the oversubscribed factor
  - In a 2:1 oversubscription, the throughput in the network is half the sum of the throughput from all the nodes
Non-Blocking Network Configuration

- None of the server nodes can “block” traffic of other nodes
  - When all the nodes are sending data at full wire speed
2:1 Blocking Network Configuration

- **Network throughput will be half**
  - If all nodes will try to send data in full wire speed
- **Oversubscribed fabrics sometimes can provide a lower cost solution**
  - When the target application do not require the highest throughput

![Diagram of 2:1 Blocking Network Configuration](image-url)
- Non-blocking InfiniBand network

OpenFOAM 1.7.1 Performance (JuRoPa)
(Laminar Cavity Flow Benchmark)
• 2:1 Blocking InfiniBand network

OpenFOAM 1.7.1 Performance (Itasca)
(Laminar Cavity Flow Benchmark)
Performance Comparison – Network Blocking

- **Non-blocking InfiniBand network provides ~27% higher performance versus 2:1 blocking network configuration.**

![OpenFOAM 1.7.1 Performance](chart)

**OpenFOAM 1.7.1 Performance**  
(Laminar Cavity Flow Benchmark)
Reference Environments

• **Murska supercomputer at CSC**
  – HP CP4000 BL ProLiant supercluster
  – Dual socket dual-core AMD Opteron 64-bit CPUs @ 2.6GHz
  – Mellanox InfiniBand DDR, non-blocking network

• **Louhi supercomputer at CSC**
  – Cray XT5 Massively Parallel Processor (MPP) supercomputer
  – Dual socket quad-core AMD Opteron 64-bit CPUs @ 2.3GHz
  – RHEL 4 Linux operating system

• **The reference environment are only used for scalability comparisons, and not for performance comparison**
Scalability Comparison

- InfiniBand QDR delivers highest scalability
  - Versus Cray XT5, and versus InfiniBand DDR
  - Murska – IB DDR, JuRoPa and Itasca – IB QDR
Conclusions

• **OpenFoam demonstrates good scaling capabilities**
  – Testing includes systems configuration up to 1K cores

• **For OpenFOAM, non blocking network delivers higher performance compared to 2:1 blocking configuration**
  – 27% higher performance in average

• **Commondity-based InfiniBand QDR demonstrates highest scalability over**
  – the proprietary-based Cray XT5 and
  – InfiniBand DDR
Thank You
HPC Advisory Council
www.hpcadvisorycouncil.com