Weather Research and Forecast (WRF) Model

Performance and Profiling Analysis on Advanced Multi-core HPC Clusters

March 2009
The following research was performed under the HPC Advisory Council activities:

- HPC Advisory Council Cluster Center

Special thanks to AMD, Dell, Mellanox:

- In particular to
  - Joshua Mora (AMD)
  - Jacob Liberman (Dell)
  - Gilad Shainer and Tong Liu (Mellanox Technologies)

Special thanks to John Michalakes for his support and guidelines.

For more info please refer to:

- [http://hpcadvisorycouncil.mellanox.com/](http://hpcadvisorycouncil.mellanox.com/)
The HPC Advisory Council

**Distinguished HPC alliance (OEMs, IHVs, ISVs, end-users)**

- More than 60 members world wide
- 1st tier OEMs (AMD, Dell, HP, Intel, Sun) and systems integrators across the world
- Leading OSVs and ISVs (LSTC, Microsoft, Schlumberger, Wolfram etc.)
- Strategic end-users (Total, Fermi Lab, ORNL, OSU, VPAC etc.)
- Storage vendors (DDN, IBRIX, LSI, Panasas etc)

**The Council mission**

- Bridge the gap between high-performance computing (HPC) use and its potential
- Bring the beneficial capabilities of HPC to new users for better research, education, innovation and product manufacturing
- Bring users the expertise needed to operate HPC systems
- Provide application designers with the tools needed to enable parallel computing
- Strengthen the qualification and integration of HPC system products
HPC Advisory Council Activities

HPC Advisory Council

Network of Expertise

Best Practices
- Oil and Gas
- Automotive
- Bioscience
- Weather
- CFD
- Quantum Chemistry
- and more....

HPC Outreach and Education

Cluster Center
End-user applications benchmarking center

HPC Technology Demonstration

40Gb/s InfiniBand Distributed Visualization over Scinet

Network of Expertise
The WRF Model

• The Weather Research and Forecasting (WRF) Model
  – Numerical weather prediction system
  – Designed for operational forecasting and atmospheric research

• WRF developed by
  – National Center for Atmospheric Research (NCAR),
  – The National Centers for Environmental Prediction (NCEP)
  – Forecast Systems Laboratory (FSL)
  – Air Force Weather Agency (AFWA)
  – Naval Research Laboratory
  – Oklahoma University
  – Federal Aviation Administration (FAA)
WRF Best Practices Objectives

• Ways to improve performance, productivity, efficiency
  – Knowledge, expertise, usage models

• The following presentation aims to review
  – WRF performance benchmarking
  – Interconnect performance comparisons
  – WRF productivity
  – Understanding WRF communication patterns
  – MPI libraries comparison
Test Cluster Configuration

- **Dell™ PowerEdge™ SC 1435 24-node cluster**
  - Dual socket 1-U rack server supports 8x PCIe and 800MHz DDR2 DIMMs
  - Energy efficient building block for scalable and productive high performance computing

- **Quad-Core AMD Opteron™ Model 2382 processors (“Shanghai”), 2358 (“Barcelona”)**
  - Industry leading technology that delivers performance and power efficiency
  - Up to 21GB/sec memory throughput per compute node

- **Mellanox® InfiniBand ConnectX® DDR HCAs and Mellanox InfiniBand DDR Switch**
  - High performance I/O consolidation interconnect – transport offload, HW reliability, QoS
  - Supports up to 40Gb/s, 1usec latency, high message rate, low CPU overhead

- **MPI: Open MPI 1.3, MVAPICH 1.1, HP MPI 2.2.7**

- **Application: WRF V3, 12km CONUS benchmark case**

- **Compiler: Gfortran v4.2**
  - Flags: FCOPTIM= -O3 -ffast-math -ftree-vectorize -ftree-loop-linear -funroll-loops
  - Out of the box experience
WRF Performance Results - Interconnect

- Comparison between clustering interconnects - InfiniBand and GigE
- Opteron “Barcelona” CPUs
- InfiniBand high speed interconnect enables almost linear scaling
  - Maximized system performance and enable faster simulations
- Gigabit Ethernet limits WRF performance and slow down simulations

Higher is better
Performance Enhancement - Interconnect

- InfiniBand 20Gb/s vs 10GigE vs GigE
- AMD Opteron “Shanghai” CPUs
- InfiniBand outperforms GigE and 10GigE in all test scenarios
- WRF demonstrates performance decrease with 10GigE and GigE beyond 20 nodes
  - Not seen before with “Barcelona” CPUs

**WRF Benchmark Results - Conus 12Km (Shanghai CPU)**

- Higher is better
- 113%
• Opteron “Shanghai” CPU increases WRF performance by almost 20%
  – Compared to Opteron “Barcelona”

WRF Benchmark Results - Conus 12Km (InfiniBand)

Higher is better
Utilizing CPU affinity for higher productivity

Two cases
- Single job over the entire systems
- Two jobs, each utilized single CPU in every server (CPU affinity)

CPU affinity enables up to 20% more jobs per day
WRF Profiling – Data Transferred

WRF MPI Profiling
Total Data Send per Message Size per Cluster Size

Message Size

<table>
<thead>
<tr>
<th>4nodes</th>
<th>8nodes</th>
<th>12nodes</th>
<th>16nodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>[0..64]</td>
<td>[65..256]</td>
<td>[257..1024]</td>
<td>[1k..4k]</td>
</tr>
<tr>
<td>[4k..16k]</td>
<td>[16k..64k]</td>
<td>[64k..256k]</td>
<td>[256k..1MB]</td>
</tr>
</tbody>
</table>

Total Size (MByte)
WRF Profiling – Message Sizes

WRF MPI Profiling
Total Number of Messages per Cluster Size

Message Size

- [0..64]
- [65..256]
- [257..1024]
- [1k..4k]
- [4k..16k]
- [16k..64k]
- [64k..256k]
- [256k..1MB]

Number of Messages (Millions)

- 4nodes
- 8nodes
- 12nodes
- 16nodes
WRF Profiling – Message Distribution

WRF MPI Profiling
Message Distribution

Number of Servers

- [0..64]
- [65..256]
- [257..1024]
- [1k..4k]
- [4k..16k]
- [16k..64k]
- [64k..256k]
- [256k..1MB]
• WRF model was profiled to determine dependency networking capabilities
• Majority of data transferred between compute nodes
  – Done with 16KB-1MB message size
  – Data transferred increases with cluster size in those message sizes
• Most used message sizes
  – <64B messages  – mainly synchronizations
  – 16KB-64KB – mainly compute related
• Message size distribution
  – With cluster size, there is increase in both small and larger messages
  – From the total number of messages
    • The percentage of large messages increases on behalf of small messages
• WRF shows dependency on both clustering latency and throughput
  – Latency – synchronizations
  – Throughput (interconnect bandwidth) – compute
MPI Low Level Performance Comparison

**MPI Latency**

- OpenMPI
- MVAPICH
- HP-MPI

**MPI Bi-Dir Bandwidth**

- OpenMPI
- MVAPICH
- HP-MPI

**MPI Uni-Dir Bandwidth**

- OpenMPI
- MVAPICH
- HP-MPI

- ↓ 30%
WRF Benchmark Results - Conus 12Km

Performance (GFlop/s)

Number of Servers

OpenMPI

HP-MPI
WRF Profiling Summary – MPI Library

• **WRF model shows dependency on latency**
  – Mainly for <64B messages

• **WRF model shows dependency on throughput**
  – Mainly for 16KB-64KB messages

• **MPI comparison**
  – MPI libraries tested – Open MPI, MVAPICH, HP-MPI
  – All show same latency up to 128B
    • Beyond that MVAPICH and Open MPI show better latency
  – MVAPICH and Open MPI show higher bi-directional throughput

• **WRF mode results**
  – MVAPICH and Open MPI show similar performance results
  – HP-MPI shows average of 10% lower performance results
    • Due to lower bandwidth and higher latency
Conclusions and Future Work

• **WRF is the next generation model for weather forecasting**
  – Critical tool for severe storms prediction and alerts
  – Operational since 2006, one of the most used models nowadays

• **Efficient WRF Model usage requires HPC systems**
  – Real-time, accurate and large scale weather analysis

• **WRF Model profiling analysis proves the needs for**
  – High throughput and low latency interconnect solution
  – NUMA aware application for fast access to memory
  – Expert integration and the right choice of MPI library

• **Future work**
  – Power-aware simulations, large memory pages effect
  – Optimized MPI collective operations and collectives offload
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.