Weather Research and Forecasting (WRF)
Performance Benchmark and Profiling
July 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.dell.com/hpc
  – http://www.mellanox.com
  – http://wrf-model.org
Weather Research and Forecasting (WRF)

• **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 Usage

- The WRF model includes
  - Real-data and idealized simulations
  - Various lateral boundary condition options
  - Full physics options
  - Non-hydrostatic and hydrostatic
  - One-way, two-way nesting and moving nest
  - Applications ranging from meters to thousands of kilometers
Objectives

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

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

- **Dell™ PowerEdge™ R815 11-node (704-core) cluster**
  - Memory: 128GB memory per node DDR3 1333MHz, BIOS version 2.8.2
  - 4 CPU sockets per server node

- **AMD™ Opteron™ 6276 (code name “Interlagos”) 16-core @ 2.3 GHz CPUs**

- **Mellanox ConnectX®-3 VPI Adapters**

- **Mellanox IS5030 36-Port InfiniBand switch**

- **OS: RHEL 6.2, SLES 11 SP2, MLNX-OFED 1.5.3 InfiniBand SW stack**

- **MPI: Open MPI 1.5.5, Platform MPI 8.2.1**

- **Compilers: GNU Compilers 4.7.0**

- **Application: WRF 3.4**

- **Benchmark workload:**
  - CONUS-12km - 48-hour, 12km resolution case over the Continental US from October 24, 2001
Dell™ PowerEdge™ R815 11-node cluster

- **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
Both MPIs perform at the same level for this dataset and solver
- Performance shown by the 2 MPIs are equally as good
WRF Performance – Compilers Flags

- Using Interlagos specific compiler instructions shows significant gains
  - Performance gain of 19% to 24% by compiling application using the compiler flags
- Compiler flags added for enabling AVX, FMA4 and Interlagos instructions:
  - \texttt{-march=bdver1 -mavx -mfma4}

\textbf{WRF Benchmark (conus12km)}

\begin{itemize}
  \item \texttt{GNU47}
  \item \texttt{GNU47-tuned}
\end{itemize}

Higher is better

64 Cores/Node
WRF Performance – CPU Cores Placement

- When comparing jobs running with 32 PPN versus 64 PPN (processes per node)
  - 64 PPN shows 120% better performance than jobs running with 32 PPN
  - The 32 PPN case uses 2 CPU sockets while the 64 PPN case uses 4 CPU sockets
- When comparing jobs running with 32 PPN:
  - Using 1 core in a compute unit is 31% faster than both core are active
  - Performance boosts when the idle core in a core pairs goes into sleep mode
**WRF Performance – Interconnect**

- **InfiniBand QDR** provides the network needs for running WRF efficiently
  - Shows up to 135% better performance than 40GbE at 8-node
  - Shows up to 392% better performance than 10GbE at 8-node
  - Performance stalls for Ethernet network when running for more than 2-3 nodes

- **Performance of 1GbE cannot satisfy the needs for network throughput**
  - Performance would not scale beyond 2 machines when using 1GbE

---

**WRF Benchmark**

(conus12km)

![Chart showing performance comparison between different network speeds with InfiniBand QDR providing 392% and 135% better performance than 1GbE and 40GbE, respectively.](chart.png)

*Higher is better*
WRF Profiling – MPI/User Time Ratio

- Communication time grows faster after than computation time
  - Even though both MPI and computation time would grow

![WRF Profiling](chart)

Higher is better

64 Cores/Node
WRF Profiling – Number of MPI Calls

- 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
WRF Profiling – Time Spent of MPI calls

- The most time consuming MPI function is MPI_Bcast
  - MPI_Bcast accounts for 42% of all MPI time at 11-node
- MPI Wait accounts for large percentage on small node counts
  - MPI_Wait accounts for 53% on 1 node but drops to 23% on 11 nodes
• Majority of the MPI message sizes are concentrated in the small to midrange
WRF Profiling – Data Transfer / Process

- As the cluster scales, less data is driven to each rank and each node
WRF Profiling – Aggregated Data Transfer

- 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

- **WRF shows great needs for CPU computation and network scalability**
  - Tuning WRF for both CPU and network can provide great performance improvement

- **CPU:**
  - Using system with 4 CPUs versus 2 CPUs provides 120% gain in productivity on WRF
  - Running with Turbo mode allows WRF to achieve 31% higher performance

- **Compiler:**
  - Enabling AVX, FMA4 and Interlagos instructions in compiler flags shows 22% boost

- **Interconnects:**
  - InfiniBand provides 135% better performance than 40GbE
  - InfiniBand provides 392% better performance than 10GbE
  - 1GbE performance would not scale beyond 2 machines
Thank You

HPC Advisory Council