Weather Research and Forecast (WRF) Model

Performance and Profiling Analysis on Advanced Multi-core HPC Clusters

Gilad Shainer[1], Tong Liu[1], John Michalakes[2], Jacob Liberman[3],
Jeff Layton[3], Onur Celebioglu[3], Scot A. Schultz[4], Joshua Mora[4], David Cownie[4]

[1]Mellanox Technologies [2]National Center for Atmospheric Research
[3]Dell, Inc. [4]Advanced Micro Devices (AMD)

Abstract

The Weather Research and Forecast (WRF) Model is a fully functioning modeling system for atmospheric research and operational weather prediction communities. With an emphasis on efficiency, portability, maintainability, scalability and productivity, WRF has been successfully deployed over the years on a wide variety of HPC clustered compute nodes connected with high speed interconnects – currently the most used system architecture for high-performance computing. As such, understanding WRF dependency on the various clustering elements, such as the CPU, interconnects and the software libraries are crucial for enabling efficient predictions and high productivity. Our results identify WRF's communication-sensitive points and demonstrate WRF's dependency on high-speed networks and fast CPU to CPU communication. Both factors are critical to maintaining scalability and increasing productivity when adding cluster nodes. We conclude with specific recommendations for improving WRF performance, scalability, and productivity as measured in jobs per day. Because proprietary hardware and software can quickly erode cluster architecture's favorable economics, we will restrict our investigation to standards based hardware and open source software readily available to typical research institutions.

1. Introduction

Human life on Earth depends on a favorable climate and weather prediction to survive. Numerical Weather Prediction (NWP) is one of the main tools forecasters now use to help predict the weather. NWP models are comprised of mathematical equations that predict the behavior of a physical system. Since the late 1950s progress in computer technology has enabled continuing improvement in the accuracy of these forecasts such that, today, forecasts involving hundreds of billions of arithmetic operations per second are routinely run on from hundreds to thousands of CPUs in parallel communicating over high-speed networks.

One of the most used model by researchers is the Weather Research and Forecasting Model (WRF), a next-generation forecast model and assimilation system. The first operational implementation of the WRF occurred in June 2006.

2. The Weather Research and Forecasting (WRF) Model

The development of the Weather Research and Forecasting (WRF) modeling system is a multi-agency effort intended to provide a next-generation mesoscale forecast model and data assimilation system to advance both the understanding and prediction of mesoscale weather and accelerate the transfer of research advances into operations. The model was developed as a collaborative effort among the NCAR Mesoscale and Microscale Meteorology (MMM) Division, the National Oceanic and Atmospheric Administration's (NOAA) National Centers for Environmental Prediction (NCEP) and Forecast System Laboratory (FSL), the Department of Defense's Air Force Weather Agency (AFWA) and Naval Research Laboratory (NRL), the Center for Analysis and Prediction of Storms (CAPS) at the University of Oklahoma, and the Federal Aviation Administration (FAA), along with the participation of a number of university scientists.

The WRF model is designed to be an efficient massively parallel computing code to be able to take advantage of advanced high-performance computing systems. The code can be configured for both research and operations and offers numerous physics options. WRF is maintained and supported as a community model to facilitate wide use, and is suitable for use in a broad spectrum of applications across scales ranging from meters to thousands of kilometers. Such applications include research and operational NWP, data assimilation and parameterized-physics research, downscaling climate simulations, driving air quality models, atmosphere-ocean coupling, and idealized simulations.

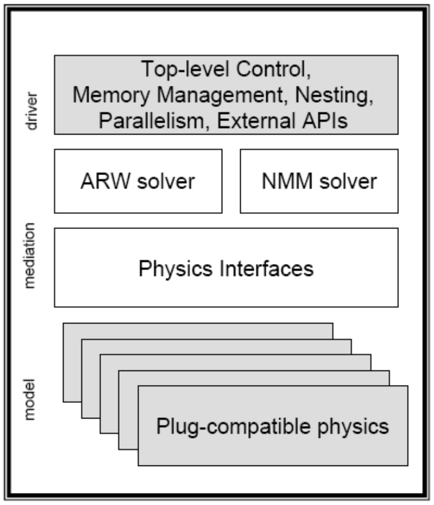

Figure 1: WRF system components

The WRF Software Framework (WSF) (Figure 1) provides the infrastructure that allows efficient use of an array of HPC systems, architectures which continue to evolve as we move into Petascale computing and beyond. The architecture accommodates multiple dynamics solvers, physics packages that plug into the solvers through a standard physics interface, programs for initialization, and the WRF variational data assimilation (WRF-Var) system. There are two dynamics solvers in the WSF: the Advanced Research WRF (ARW) solver developed primarily at NCAR, and the NMM (Nonhydrostatic Mesoscale Model) solver developed at NCEP.

3. Architecture of multi-core HPC clusters for WRF

WRF simulations are typically carried out on high-performance computing (HPC) clusters as they require an effective compute resource that can handle complex and parallel simulations. HPC clusters are scalable performance compute solutions based on industry standard hardware connected by a private system high speed network. The main benefits of clusters are affordability, flexibility, availability, high performance and scalability. A cluster uses the aggregated power of compute server nodes to form a high-performance solution for parallel applications such as the WRF model. When more compute power is needed, it can be simply achieved by adding more server nodes to the cluster.

The way HPC clusters are architected (i.e. multi-core, multi-processor based HPC servers with high speed interconnects) has a great influence on the overall application performance and productivity. In order to meet the demand of more powerful HPC servers, more execution cores (e.g. dual, quad-core) are being integrated into each processor and more processors are being tightly connected (e.g. 2,4,8 processors connected through HyperTransport^{TM} technology, a packet-based, high-bandwidth, scalable, low latency point-to-point technology that links processors to each other, processors to coprocessors and processors to I/O and peripheral controllers). There are important challenges in this strategy (e.g. larger scale integration, reduction of voltages and core frequencies) in order to keep the power consumption low while increasing the computational capabilities of the HPC servers.

The cluster interconnect is very critical to deliver efficiency and scalability for the applications as it needs to handle the networking requirements of each CPU core without imposing additional networking overhead. In a multi-core multi-socket HPC server based cluster, the driving factors of performance and scalability for WRF have shifted from the frequency and cache size per core to the memory and interconnect throughput per core. The memory bottleneck can be solved by using interconnects that support Direct Memory Access (DMA), Remote DMA and zero-copy transactions.

4. Dell HPC clustering

Over the past decade, commodity clusters have largely replaced proprietary supercomputers for high performance computing applications. According to the June 2008 Top500 [1] list of the world's fastest supercomputers, cluster architectures are used in almost 50% of the top 100 systems. This is primarily due to the highly competitive price for performance they can achieve. Dell designs, integrates, and tests HPC clusters built from industry-standard servers, leading open source and commercial software, and high speed interconnects. These clusters combine the performance of proprietary systems with the simplicity, value, and flexibility of standards based hardware.

This study was conducted on a test cluster comprised of 24 Dell PowerEdge SC1435 servers. The SC1435 is an AMD-based 1U 2-socket server that can support up to 32GB of DDR-2 memory in 8 DIMM sockets. It has one 8x PCI-Express expansion slot and two integrated Gigabit Ethernet Network Interface Cards. When used in conjunction with Mellanox InfiniBand HCAs and 3rd generation Quad-Core AMD Opteron™ processors, the SC1435 provides an ideal building block for HPC clusters due to the rack density, energy efficiency, and price/performance it can deliver.

For this study, each cluster server was equipped with a Mellanox ConnectX DDR InfiniBand HCA for inter-node communication. The servers were also deployed and configured via a Gigabit Ethernet management fabric. The cluster topology is depicted in Figure 2.

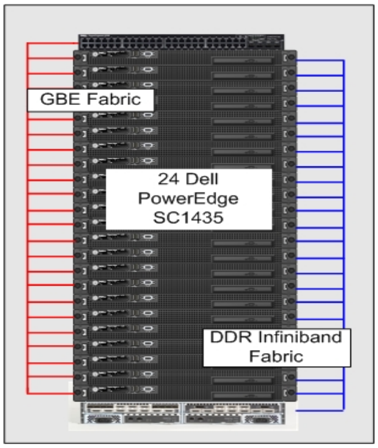

Figure 2 – cluster topology

5. Quad-Core AMD Opteron™ processors and tools for HPC servers

The HPC server configuration utilized in the WRF performance study is based on the latest available AMD processor architecture. AMD's third generation AMD Opteron™ processors with Direct Connect Architecture are designed for advanced scaling linearity in systems with up to 32 cores (i.e. 8 Quad-Core processors). AMD's Direct Connect Architecture helps eliminate the bottlenecks inherent in a front-side bus by directly connecting the processors, the memory controller, and the I/O to the central processor unit to enable improved overall system performance and efficiency in applications such as WRF.

AMD technology offers an integrated DDR2 DRAM Memory Controller with AMD Memory Optimizer Technology which allows for lower cost High-bandwidth, energy-efficient DDR2 memory. With third-generation Quad-Core AMD Opteron processors, the instruction fetch bandwidth, data cache bandwidth, and memory controller to cache bandwidth have all been doubled over the previous generation technology to help keep the 128-bit floating-point pipeline full.

Deployment of clusters can become both an energy consumption and cost challenge. With AMD's enhanced AMD PowerNow!™ and AMD CoolCore™ Technologies, today's clusters can deliver performance on demand while minimizing power consumption. Second generation AMD Opteron based platforms can be upgraded to AMD's third generation processors in the same thermal envelope, allowing for increased computational capacity without altering datacenter power and cooling infrastructures.

AMD also offers additional tools such as the AMD Core Math library (ACML), AMD Performance Primitives (APP), and AMD CodeAnalyst for extensive profiling single and multithreaded applications. AMD software tools can be downloaded for free from http://developer.amd.com.

6. InfiniBand high-speed interconnect technology

Choosing the right interconnect technology is essential for maximizing HPC system efficiency. Slow interconnects delay data transfers between servers, causing poor utilization of the compute resources and slow execution of simulations. An interconnect that requires CPU cycles as part of the networking process will decrease the compute resources available to the application and therefore will slow down and limit the number of simulations that can be executed on a given cluster. Furthermore, unnecessary overhead on the CPU increases the system jitter which in return limits the cluster's scalability.

Interconnect flexibility is another requirement for multi-core systems. As various cores can perform different tasks, it is necessary to provide Remote Direct Memory Access (RDMA) along with the traditional semantics of Send/Receive. RDMA and Send/Receive in the same network provides the user with a variety of tools that are crucial for achieving the best application performance and the ability to utilize the same network for multiple tasks. Moreover, in multi-core multi-processor environments, it is essential to have a interconnect that provides the same low latency for each process/core (zero scalable latency), regardless of the number of cores and processors that operate simultaneously, in order to guarantee linear application scalability.

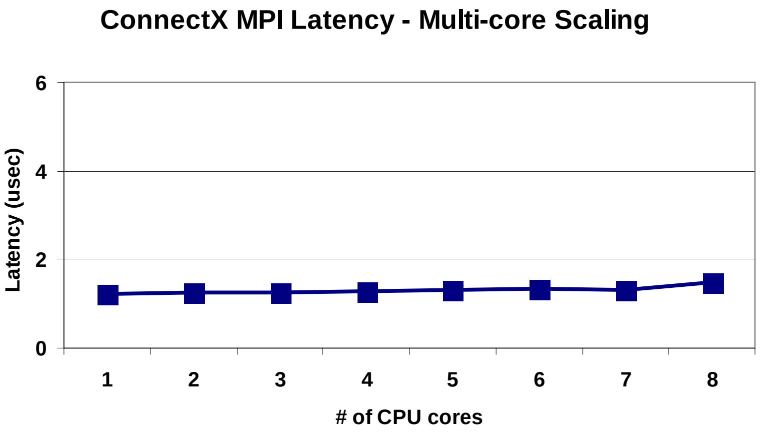

Figure 3 – MPI multi-core latency with Mellanox ConnectX InfiniBand HCA

By providing low-latency, high-bandwidth and extremely low CPU overhead, InfiniBand [5] has become the most deployed high-speed interconnect, replacing proprietary or low-performance solutions. The InfiniBand Architecture (IBA) is an industry-standard fabric designed to provide high bandwidth, low-latency computing, scalability for ten-thousand nodes and multiple CPU cores per server platform and efficient utilization of compute processing resources. Mellanox ConnectX InfiniBand adapters and InfiniScale IV-based switches are the leading edge InfiniBand solutions that have been designed for HPC clustering technology. ConnectX and InfiniScale IV deliver up to 40Gb/s of bandwidth between servers and up to 120Gb/s between switches. This high-performance bandwidth is matched with ultra-low application latency of 1µsec, and switch latencies under 100ns that enable efficient scale out of compute systems. For multi-core systems, it is essential to provide zero scalable latency, which means to provide the same low latency regardless of how many process are running between cluster nodes. Figure 3 shows the MPI multi-core latency benchmark results between two 8-core servers using Mellanox ConnectX InfiniBand adapters. The benchmark measured the latency of eight different cases – from a single process running between the two systems using a single core per system, up to eight parallel processes running between the two systems using all the available cores. According to the results, ConnectX adapters enable zero scalable latency that guarantees the same low latency for each of the CPU cores, regardless on how many core communicate at the same time. Moreover, InfiniBand was designed to be fully offloaded, meaning all the communications are being handled within the interconnect without involving the CPU. This further enables the ability to scale up with linear performance by reducing the system jitter and enabling efficient synchronizations between the execution cores.

7. WRF performance scalability and profiling analysis

We selected one of the standard benchmark cases provided by the WRF developers[1]. The 12km CONUS Case is a 48-hour, 12km resolution case over the Continental U.S. (CONUS) domain October 24, 2001 that uses the Eulerian Mass (EM) dynamics. The computational cost for this domain is about 28.5 billion floating point operations per average time step (72 seconds). To measure cost over a fully spun-up (all moisture fields and processes fully activated) forecast interval, the benchmark period is hours 25-27 (3 hours), starting from a restart file from the end of hour 24.

The performance metric is the model speed, directly measured as the average cost per time step over a representative period of model integration, ignoring I/O and initialization cost. Results are presented as normalized floating-point rate and as simulation speed. Floating-point rate provides a measure of efficiency relative to the theoretical peak capability of the computing system. It is the average number of floating-point operations per time step divided by the average number of seconds per time step. Simulation speed is the measure of wall clock time required to complete the simulation.

Using the CONUS 12km benchmark, we performed the following analysis:

- Compared the performance of different interconnect technologies, namely InfiniBand and gigabit Ethernet
- Measured WRF scalability at increasing core counts
- Identified methods for increasing WRF productivity through job placement
- Profiled WRF network utilization in order to identify points with the greatest performance impact.

---

[1] http://www.mmm.ucar.edu/wrf/WG2/bench

The performance and profiling analysis was carried out as part of the HPC Advisory Council [6] research activities, using the HPC Advisory Council Cluster Center. The cluster configuration is summarized in table 1 below.

Table 1. Cluster benchmark configuration

Application	WRF V3.0
Servers	24 Dell PowerEdge SC1435 servers
Processors	2 Quad-Core AMD Opteron™ 2358 processors at 2.4 GHz per node
Memory	8 x 2 GB, 667 MHz Registered DDR-2 DIMMs per node
OS	Red Hat® Enterprise Linux® 5 Update 1 OS
Compilers	gFortran 4.1.2 FCOPTIM = -O3 -ftree-vectorize -ftree-loop-linear -funroll-loops
Message Passing Interface (MPI)	• Open MPI 1.3 • MVAPICH 1.1 • HP MPI 2.2.7
Interconnect	Mellanox MT25408 ConnectX DDR InfiniBand OpenFabrics Enterprise Distribution (OFED) 1.3 software stack

7.1 WRF Performance results and the interconnect effect

The following analysis was targeted as an out-of-the-box experience. As such, we have used an open source compiler and various MPI libraries. The compiler chosen was the gfortran with the flags described before for maximizing CONUS 12km benchmark performance. For the MPI library the open source MVAPICH [7] and Open MPI [8] were tested. Since both of those MPI libraries showed similar performance, we have chosen to provide the Open MPI based results only.

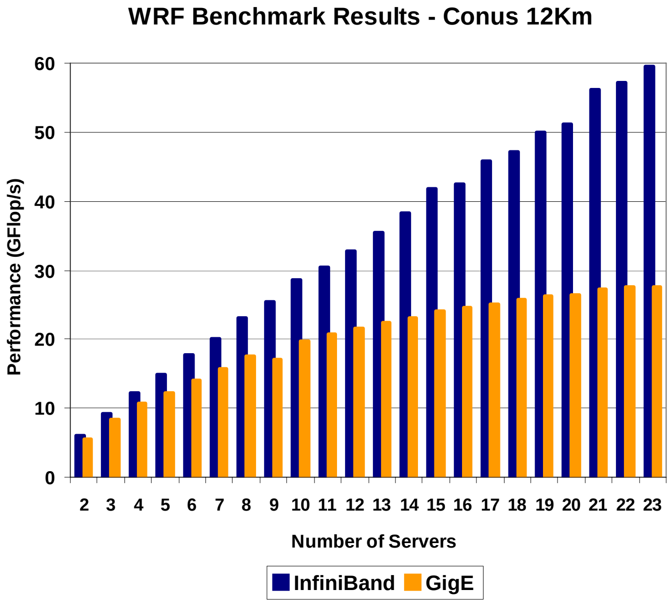

Figure 4: WRF performance results and interconnect comparison

Figure 4 shows the CONUS 12km performance results achieved per cluster size from two to twenty four server nodes using all 8 cores on each server for full node timing. When using InfiniBand as the cluster interconnect, WRF showed linear scalability, i.e. with the addition of more server nodes, the total performance of WRF increased accordingly. Comparing to other public databases [4], the

performance achieved with the out-of-the-box experience have matched or outperformed other results achieved with commercial compilers or MPI libraries.
When using gigabit Ethernet as the cluster interconnect, we measured a degradation in the WRF capabilities in every clusters size as compared to InfiniBand. InfiniBand has outperformed gigabit Ethernet, achieving from 8% higher performance with a 2-node cluster to 115% higher performance with a 24-node cluster. Moreover, gigabit Ethernet failed to show linear scalability and it limited performance gain beyond 20 nodes. A 10-node InfiniBand cluster met the performance level achieved by a 24-node gigabit Ethernet cluster running the particular workload. Therefore using InfiniBand gives the cluster administrator the flexibility to either use 2.4x fewer nodes to achieve a given performance level in order to save energy costs or to increase productivity when utilizing a larger cluster. Looking at the asymptotic behavior of the gigabit Ethernet based results, one could deduct that adding more nodes to the 24-node Ethernet based configuration will not result in significant increased performance.

Since InfiniBand interconnect delivers lower latency and higher bandwidth than gigabit Ethernet, the application's communication overhead through the InfiniBand fabric is much lower and therefore the scaling of the application is better on InfiniBand than on gigabit Ethernet for a given amount of servers. Quantification of interconnect latency and bandwidth on WRF performance is provided throughout subsequent sections of the paper.

7.2 Utilizing job placement for higher productivity

Productivity is measured by the number of application jobs that can be achieved in a given time, usually one day. A higher number of jobs per day equals higher productivity. But productivity by itself can not be the sole indicator for optimal job placement. It should be reviewed together with the application job run time and the user requirements for run time. For example, achieving the maximum number of jobs per day, while each job run time is 24 hours, will provide very poor non real-time weather prediction.

With the increased complexity of high-performance applications, a single job consuming all the cluster resources might create bottlenecks within the CPU to CPU or CPU to memory communications. We have shown near-linear scaling when using the InfiniBand interconnect between nodes. Therefore, the question becomes how much scalability will we see if a single job is limited

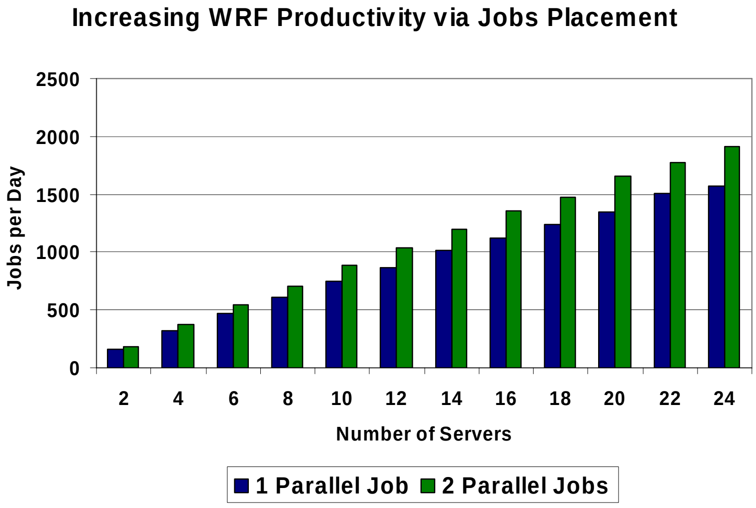

Figure 5: WRF productivity results

to running within each cluster node? These bottlenecks and productivity capabilities can be explored by comparing a single job run on the entire cluster versus multiple jobs run in parallel using job placement to a specific CPU socket.

Figure 5 shows the productivity results comparing a single job run to two parallel jobs. Each of the two parallel jobs used a single socket per server, on the entire cluster, therefore half of the system CPU cores. In the parallel test, the CPU to CPU communication has been reduced. For each of the cluster size cases, from two servers to twenty-four, the parallel job approach with CPU affinity increased the cluster's total productivity from 15% to 20%. The productivity increase was due to the fact that the cluster completed two jobs in only 30-40% more time than it took to complete one job without affinity. Therefore using CPU affinity needs to be reviewed and decided per user case.

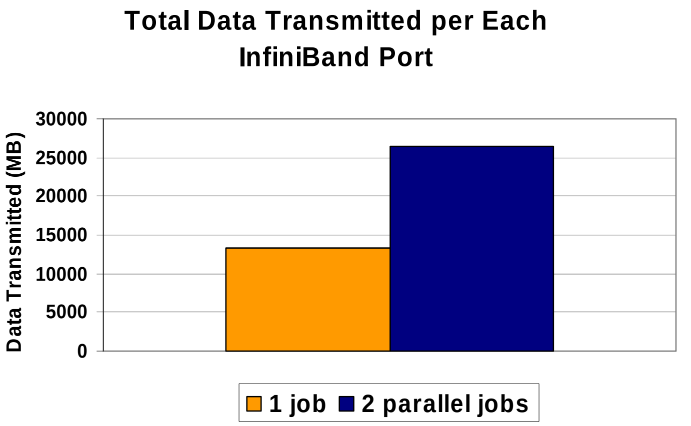

Figure 6: Interconnect activity for single and parallel WRF job run

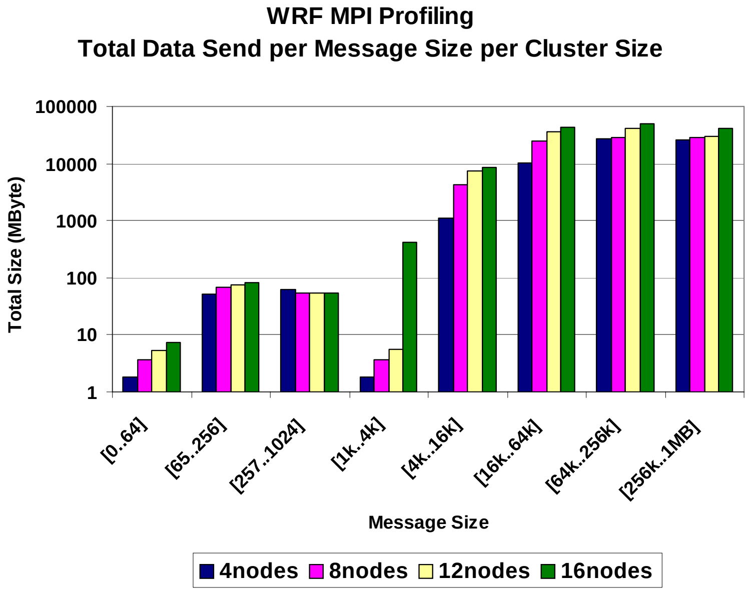

Figure 7: WRF profiling – total communication data send

Figure 6 compares the interconnect activity of the single job cases with the two parallel jobs for a single server and a single InfiniBand port. As expected, the two fold increase in the number of jobs causes two fold increase in the interconnect activity. The productivity increase over the single InfiniBand port is enabled by moving the bottlenecked traffic between the CPUs and CPU-memory to the interconnect, better utilizing the interconnect's high throughput capabilities.

7.3 WRF Communications Profiling

Profiling application characteristics is essential to understanding the application's performance dependency on the various cluster subsystems. In particular, communication pattern profiling can help in choosing the most efficient interconnect and MPI library, and identifying the critical communication sensitivity points that greatly influence on the application’s performance, scalability and productivity.

WRF profiling data are presented in figures 7 and 8. Figure 7 represents the total data sent per interconnect message size. Most of the communication data sent between the compute systems are carried by very large message sizes. Most of the data is sent via 16K Byte to 64K Byte message sizes, but the large portion of it also in larger sizes of 64K Byte – 1M Byte. This shows the need for fast interconnect technology that can deliver the highest throughput.

Figure 8 shows the number of messages that were sent between the cluster nodes. Two main peaks were observed – one of messages between 0 to 64B and the second between 16K Byte and 64K Byte. The 0-64B area represents mostly the synchronization or control messages and the 16KB-64KB the compute messages. Those areas will have the greatest impact on WRF’s overall performance and are therefore considered the sensitivity points.

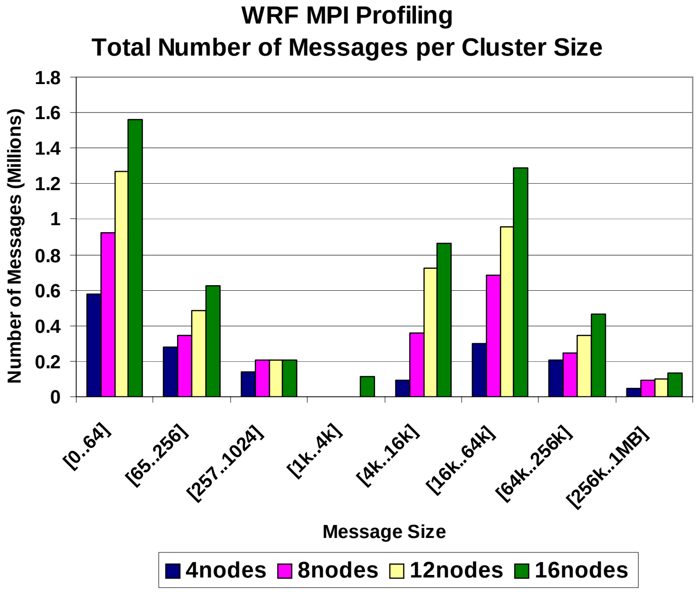

Figure 8: WRF profiling – number of network messages

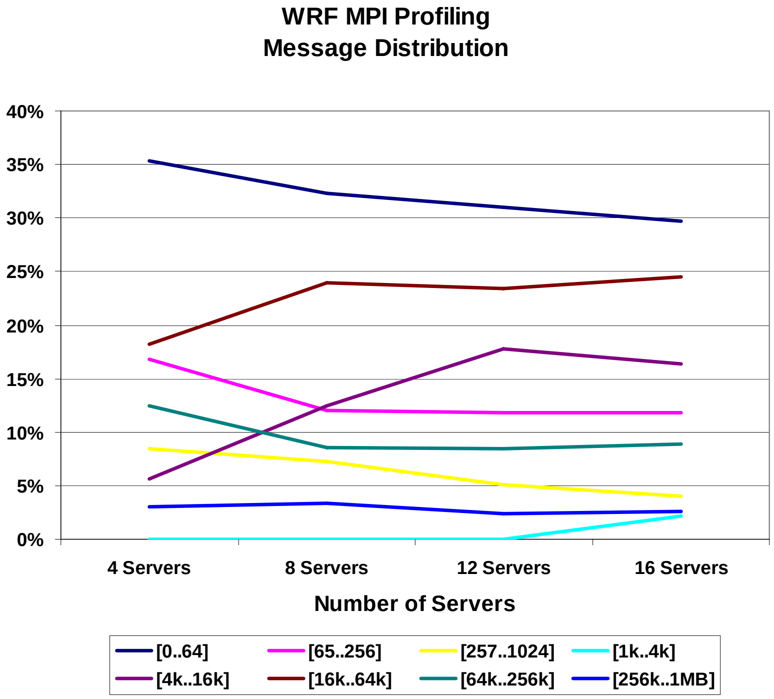

Figure 9: WRF profiling – message distribution

We have also noticed that the number of messages grow in every size category as the size of the cluster increases. In the small size area, due to more synchronization required, and in the large size areas since there are more compute cycles. This indicates the need for lowest latency in order to avoid creating a bottleneck for the fast synchronizations, especially in the 64B range, and for highest throughput in order to minimize or eliminate the CPU idle cycles while waiting to receive the compute information.

The message distribution per message size is presented in figure 9. While the number of messages grows in every message size category, the biggest growth with cluster size is in the 16K-64K Byte range. The percentage of small synchronization messages shrinks as the cluster size grows. The percentage of the largest messages remains constant, as they serve as the problem data distribution, and then the compute data is being sent via the 16K-64K Byte range. These message distribution results also emphasize the need for the highest interconnect throughput for achieving high efficiency and scalability.

7.4 MPI Libraries Comparison

We have identified two critical sensitivity points in the WRF Model – latency for 0-64B messages, and throughput for 16K-64KB messages. Of course the latency for other small messages and the throughput for other large message size are important but not as critical. In order to provide good efficiency and scalability, the chosen interconnect needs to provide the lowest latency for the 0-64B message range and the highest throughput for 16K-64B message rage. The same is applicable for the MPI libraries.

Figures 10 and 11 demonstrate the performance differences between three MPI libraries, OSU MVAPICH, Open MPI and HP-MPI for the InfiniBand interconnect. We compared the performance of these MPI libraries with the OSU MPI bandwidth and latency benchmarks.

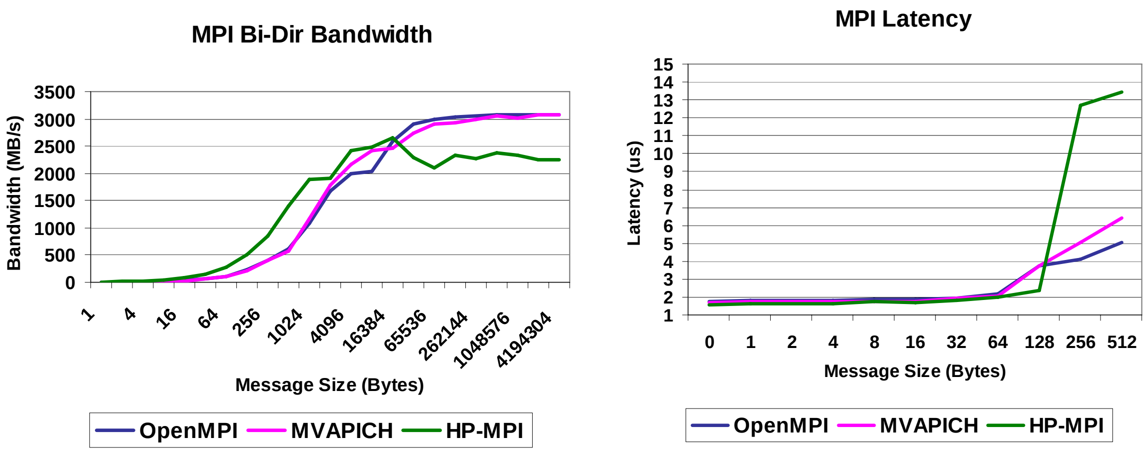

Figure 10, 11 – MPI bi-directional bandwidth and latency comparison

While MVAPICH and Open MPI provide the same performance, HP MPI shows lower bandwidth above 16KB message size and higher latency above 64B message size. In particular, at the WRF sensitivity points – 64B latency and 16K-64KB bandwidth, all tested MPI libraries show the same latency, but MVAPICH and Open MPI show on average 30% higher throughput than HP MPI.

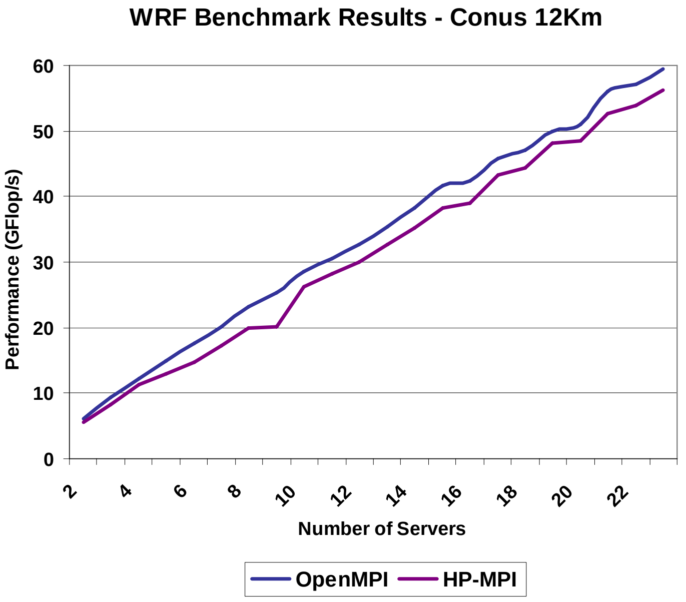

Figure 12 – WRF performance with different MPI libraries

The influence of the performance difference between the MPI versions at the sensitivity points is reflected in the WRF model performance as showed in figure 12. Due to the bandwidth advantage of MVAPICH and Open MPI, the usage of those libraries results in 10% higher performance or GFlops compared to HP MPI. Future optimizations in the MPI libraries, in particular around the WRF sensitivity points, are expected to boost WRF efficiency and to increase productivity.

8. Conclusions and Future Work

Numerical weather prediction models are critical tools for forecasters. WRF is designed to provide real-time, extremely accurate and sophisticated weather analysis. For efficient analysis, WRF requires high-performance computing systems. Commodity clusters have become very important for high performance computing due to the price for performance, flexibility and scalability they can deliver. In this paper we analyzed WRF performance on HPC clusters in order to identify best practices for researchers interested in maximizing WRF's potential. Because high performance computing systems offer a complex array of hardware and software intended to improve performance, we elected to limit our investigation to the maximum performance that can be achieved on commodity clusters comprised of open source software and standards based hardware. As such, we have used open source MPI libraries and InfiniBand drivers, and free compilers. We believe that this approach benefits researchers in three ways. First, this mirrors the out-of-box experience (first use after system install) a researcher might expect in a typical research environment. Second, this approach simplifies the cluster design and acquisition process. Finally, it enhances value by demonstrating that performance gains can be made without using costly software components that can quickly erode the favorable economics of commodity clusters. The results of this study demonstrate that simplicity and value do not come at the expense of performance.

The results of our study show that a high speed, low latency clustering interconnect is essential for high-performance and scalability. The InfiniBand interconnect outperformed gigabit Ethernet by up to 115%. Furthermore, InfiniBand showed better scalability than gigabit Ethernet. With InfiniBand, WRF performance improved as additional compute nodes were added, while gigabit Ethernet showed little performance gain after 20 nodes. Faster WRF run times translate into improved performance/watt, optimizing power/performance criteria for power-aware simulations.

While the first conclusion is well known from other papers in the literature, we have investigated WRF network, or core-to-core communication during the model time-stepping. The communication profiling results identified the WRF Model's sensitivity points which greatly effect WRF efficiency. Our results show that WRF depends on very low latency in particular for communication messages smaller than 64B and for high throughput for communications messages between 16K Byte and 64K Byte. Although the growth of smaller sized messages was expected as the number of cluster nodes increased due to synchronization overhead, it is important to note that the number of larger messages increased at the same rate. This indicates that interconnect latency and throughput carry equal weight in improved WRF performance. The MPI library comparison showed the importance of providing the most efficient function implementation to address those sensitivity points. The lower latency and higher throughput afforded by Open MPI and MVAPICH resulted in improved performance over HP MPI.

Although interconnect type was the greatest determinant in improving WRF scalability as the size of the cluster increased, it was also observed that overall cluster productivity could be improved by up to 20% by running simultaneous jobs on the cluster rather than allocating the entire cluster to a single job. Increased productivity is the result of two factors. First, the AMD platform is a Non-Uniform Memory Addressing (NUMA) architecture which means that each multi-core processor enjoys faster access to its local memory than to remote memory. Through core and memory affinity, the parallel WRF jobs eliminate the remote memory access penalties plus increases cache hits, and thereby increase overall application performance. Second, parallel jobs with smaller core counts reduce the synchronization overhead for each application which also benefits WRF efficiency.

From this study there are many avenues for future research including measuring the performance impact of using large memory pages, the effect of MPI collective operations and the influence of offloaded collective operations. Additionally, this study could be expanded to include measuring energy efficiency while maximizing productivity.

References

[1] TOP500 – http://www.top500.org

[2] Skamarock, W. C., J. B. Klemp, J. Dudhia, D. O. Gill, D. M. Barker, W. Wang, and J. G.Powers. A description of the advanced research WRF version 2. Technical Report NCAR/TN-468+STR, National Center for Atmospheric Research, January 2007.

[3] J. Dudhia, D. Gill, T. Henderson, J. Klemp, W. Skamarock, and W. Wang: The Weather Research and Forecast Model: Software Architecture and Performance. Proceedings of the Eleventh ECMWF Workshop on the Use of High Performance Computing in Meteorology. Eds. Walter Zwieflhofer and George Mozdzynski. World Scientific, 2005, pp 156-168

[4] Michalakes, J.,: Weather Research and Forecast (WRF) Model Port to Windows: Preliminary Report

[5] InfiniBand Trade Association – http://www.infinibandta.org

[6] HPC Advisory Council - http://www.mellanox.com/partners/HPC_Advisory_Council.php

[7] MVAPICH MPI - http://mvapich.cse.ohio-state.edu/

[8] Open MPI - http://www.open-mpi.org/