provided by UNT Digital Library

INL/EXT-08-14212

High Performance Computing CFRD Final Technical Report

Kelly J. Knight Jon Berkoe Pam Johnson Hope Forsmann Kurt Hamman Tom Foust

January 2003

The INL is a U.S. Department of Energy National Laboratory operated by Battelle Energy Alliance

High Performance Computing CFRD Final Technical Report

Kelly J. Knight¹ Jon Berkoe1 Pam Johnson2 Hope Forsmann2 Kurt Hamman2 Tom Foust²

1 Bechtel National Inc. Research and Development 2 Idaho National Engineering and Environmental Laboratory

January 2003

Idaho National Laboratory Idaho Falls, Idaho 83415

http://www.inl.gov

Prepared for the U.S. Department of Energy Under DOE Idaho Operations Office Contract DE-AC07-05ID14517

CONTENTS

FIGURES

ACRONYMS

WTP Waste Treatment Project

High Performance Computing CFRD Final Technical Report

1. INTRODUCTION

The Bechtel Waste Treatment Project (WTP), located in Richland, WA, is comprised of many processes containing complex physics. Accurate analyses of the underlying physics of these processes are needed to reduce the amount of added costs, during and after construction, that are due to unknown process behavior. The WTP will have tight operating margins in order to complete the treatment of waste on schedule. The combination of tight operating constraints, coupled with complex physical processes, requires analysis methods that are more accurate than traditional approaches. This study is focused specifically on multidimensional, computer-aided solutions.

Many skills and tools are required to solve engineering problems. Many physical processes are governed by nonlinear, partial differential equations. These governing equations have few, if any, closed-form solutions. Past and present solution methods require assumptions to reduce these equations to solvable forms. Computational methods take the governing equations and solve them directly on a computational grid. This ability to approach the equations in their exact form reduces the number of assumptions that must be made. This approach increases the accuracy of the solution and its applicability to the problem at hand. Recent advances in computer technology have allowed computer simulations to become an essential tool for problem solving.

In order to perform computer simulations as quickly and accurately as possible, both hardware and software must be evaluated. With regards to hardware, the average consumer personal computers are not configured for optimal scientific use. Only a few vendors create high-performance computers to satisfy engineering needs. Software must be optimized for quick and accurate execution. Operating systems must utilize the hardware efficiently, while supplying the software with seamless access to the computer's resources.

From the perspective of Bechtel Corporation and the Idaho National Engineering and Environmental Laboratory (INEEL), it is crucial to know the capabilities of a software package's shared memory processor (SMP) version or cluster (distributed memory) version. Of utmost importance is the knowledge of a software package's cost and implementation challenges. Additionally, it is important to determine the hardware performance of a computing workstation. The level of performance of software is inextricably tied to the computer hardware upon which it is run. **Bechtel can do more for its clients in the same amount of time and/or solve more complex problems if computer workstations and associated software are optimized.** As a Bechtel Management and Operations Facility, INEEL engineers and scientists find solutions to problems important to Bechtel. Both INEEL engineers and managers must be informed and educated in high-performance computing (HPC) techniques and issues to better accomplish their research.

One mission of the WTP is to treat and prepare nuclear waste for long-term storage. This process involves complex procedures and physics. The computer simulation of the pulsed jet mixers (PJMs) is of particular interest. The pulsed jet mixers are designed to mix slurry of continuously sized, solid particles of waste in liquid. The WTP engineers are tasked to find the most optimal design for a mixing tank. To assist them, the engineers have turned to computers and computational fluid dynamics (CFD) to simulate numerous mixing tank designs.

The CFD models are computed on differing complex geometrical meshes that reflect the different tank designs being investigated. The models simulate transient boundary conditions within the tanks such

as the blowing and suction of the pulsed jets and the rise and fall of the free surface of the slurry. This is a multiphase problem that has required user input to properly simulate the hindered settling properties of the waste particles in the slurry. The large size of the computational meshes, in addition to the number of equations being solved on the grid, results in a computationally intensive problem capable of being solved by only a few high-end computers. These simulations are only a few examples of the computationally intensive problems that require the use of high-performance computers by Bechtel Corporation and INEEL.

This Corporate-Funded Research and Development (CFRD) grant addressed three objectives in response to Bechtel's computational needs. The first objective was to introduce and educate both INEEL and Bechtel managers and engineers to the field of HPC. This involved three separate training classes. The first class was a 4-hour class designed to introduce managers to HPC. This class was tailored to inform managers about the technical issues of HPC and how HPC may or may not translate into cost savings. The second class was a 2-day course designed for scientists and engineers. This class provided a technical overview of high-performance issues in both software and hardware, and multiprocessor computing. The third class taught the Message Passing Interface (MPI) protocol. MPI is one of the languages that controls code execution on multiprocessor machines.

The second objective was to evaluate installation, job queue, and execution issues related to multiprocessor versions of Fluent (a commercial CFD code) on two different types of parallel computer architectures. Bechtel engineers use Fluent to solve complex fluid and heat transfer problems.

The third objective was to benchmark the performance of Fluent on the SMP and cluster machines currently located at INEEL. The benchmarking process requires simulation of multiple problems of differing grid sizes and differing physics. The WTP project currently requires the solution of fluid flow, heat transfer, and multiphase problems that require large computational grids. Engineering problems for the WTP were to be benchmarked; however, because of their large size, complexity, and diversity of physics and due to complex problem setup, they were not available for this study. Nevertheless, information learned from this study was used in support of the WTP. As of the date on this report, four large multiphase simulations have been run on INEEL's multiprocessor computers, utilizing 4 to 8 processors that simulate the pulsed jet mixers.

Lessons learned from this study include the following:

- 1. High-performance hardware can achieve tremendous speeds if used efficiently.
- 2. The operating system of a computer has a large impact on the performance of a machine and the accompanying software. This is of particular importance to Bechtel. Bechtel uses the Windows operating system created by Microsoft. This operating system is developed for average consumer use. Therefore, many commands and instruction sets are not optimized to utilize a computer's processor and hardware efficiently.
- 3. One of the keys to obtaining fast execution of a software code is optimal use of a processor's pipelines and cache.
- 4. Software must be *purposefully* optimized for use on high-performance, scalar (single processor) or parallel (multiple processors) computers.
- 5. Evaluation of third-party, software vendor's claims of high performance must be done carefully. Many interdependent parameters govern efficient execution of code on computers.
- 6. Separate benchmarking of software must be done for different computer architectures and number of processors.
- 7. Separate benchmarking of software must be done for different physical models used in simulations. Different physical models will be coded differently. This translates into varying levels of execution despite the same core of the code.
- 8. Careful consideration must be made when purchasing high-performance hardware and software.

All objectives set in this CFRD were accomplished. The classes were arranged and taught by experienced HPC engineers. The classes were well attended by both Bechtel and INEEL engineers. The cluster and SMP versions of Fluent were successfully installed and benchmarked. Knowledge gained from this study supported Bechtel's WTP project. Information learned from this study assisted Bechtel's Research and Development (R&D) group in procuring and installing a five-node, dual processor cluster for use on the WTP project. The use of parallel versions of Fluent has benefited several INEEL projects and paved the way for future use of Fluent at the site. This study has prompted initiation of a direct Bechtel to INEEL Internet connection for use of INEEL computers on Bechtel projects by Bechtel engineers.

2. REQUESTED TASKS

The overall goal of this CFRD was to develop HPC applications for CFD. This is a new and emerging capability at INEEL that will be available for use on a variety of Bechtel projects. Working with the Bechtel National Research & Development group (Bechtel R&D), a HPC computing environment was developed based on general principles used in HPC and a coding protocol used for parallel computing processes, in this case MPI. The INEEL scientists and engineers, in the course of their research work, will either be required to write their own high performance computer codes or use prepackaged HPC codes.

Proposed tasks included the following:

- Development of a capability at INEEL with Bechtel R&D participation based on general principles used in HPC and the coding protocol used for parallel computing processes (i.e., MPI); Section 4
- Installation and benchmarking a parallel version of the commercial CFD software program, Fluent, on both the SMP machine and the cluster; Sections 5 and 6, respectively
- Implementation of job maintenance and job queue routines to optimize central processing unit (CPU) usage (the usage of INEEL super computers was low enough that job queues were not required during this study)
- Run computing-intensive calculations in support of the Hanford Waste Treatment Project; Section 7.

3. HIGH-PERFORMANCE COMPUTING HARDWARE AND SOFTWARE ENVIRONMENT

The INEEL HPC environment includes several hardware components. There is a Symmetric Multiprocessing machine and a 44-node Linux cluster machine. The SMP machine is a Silicon Graphics (SGI) Origin 3800. This machine contains 64 400-MHz processors with 64 GB of shared memory. The machine is connected to a RAID disk system with approximately 5 TB of disk storage. The SGI operating system (OS) is IRIX 6.5.11f.

The Linux Stormcloud cluster system includes a total of 44 nodes purchased from Anova. Each node contains dual 1.2-MHz processors, 2 GB of memory, 20-GB hard disk, and OS version Mandrake 8.2 Linux.

Fluent licenses were purchased to run on both the SMP and Linux Cluster machines. Two serial licenses were purchased, along with 16 parallel licenses that could be used in any combination on the SMP machine or the Linux Cluster.

The MPI is a library specification for message passing, proposed as a standard by a broadly based committee of vendors, implementers, and users. MPI is designed for high performance on both massively parallel machines and on workstation clusters. MPI is available on both the SMP machine and the Linux Cluster at no cost.

4. TRAINING FOR OPTIMIZING THE HIGH-PERFORMANCE COMPUTING HARDWARE AND SOFTWARE

A detailed introduction into HPC was needed for Bechtel and INEEL managers and the scientists and engineers. The first task was to develop and teach classes on the general principles used in HPC. Three classes were formed to reach this goal.

4.1 HPC Class for Bechtel and INEEL Managers

The first class focused on giving an overview of the principles used in HPC for an audience of management personnel. Managers need to become aware and knowledgeable of the new field of HPC and its accompanying hardware, software, and user requirements. Management makes the decisions on whether or not to purchase these types of environments. This class offered a forum for them to ask questions and come away with a better understanding of why this type of technology is absolutely necessary for the success of many projects. Understanding the computing marketplace is just as important as knowing how the processors work. Managers and engineers will be faced with interpreting computer benchmarks to make a competent investment decision. Computer benchmarks measure different types of performance, some of which may apply to the task at hand and some may not.

The latest generation of processors is so fast that developers and users may not know that their code is only getting a portion of the performance that they deserve. This is a common decision point. Running the code at 40% capacity may be more cost effective than investing a few to several weeks of effort into fine tuning the code. Awareness of high-performance issues puts Bechtel and INEEL engineers and managers in a position to judge whether the tradeoff is practical and economical.

A more detailed description of what was presented in this class is given in Section 4.2.

4.2 HPC Class for Bechtel and INEEL Scientists and Engineers

The second class was tailored to satisfy the technical requirements of engineers and scientists. This class provided them with the knowledge of how to best leverage an HPC environment toward their specific work tasks. Discussions consisted of how the hardware works (i.e. processor/memory), compiler recommendations, and how to optimize source code to take advantage of the parallelism of the machine and the various interfaces that can operate within these capabilities. The class is summarized in the following subsections.

4.2.1 General Principles in High-Performance Computing

Touting faster bus times is the most popular way to characterize a computer's performance. This parameter, however, is only one of many parameters that determine the operating performance of a machine. *There is no single way to determine the performance of a machine and the software that will run on it.* This class was designed to delineate the specific parameters related to hardware performance and their subsequent impact on software development and implementation.

Twelve topics discussing computer and software were discussed. These topics are listed and briefly summarized below. HPC is not exclusive to parallel machines. There are high-performance scalar (one processor) and super scalar machines. Both parallel and scalar machines share the same basic hardware requirements. These basic requirements are discussed first before moving into parallelism.

4.2.1.1 High Performance Computing. The differences between complex instruction set computer (CISC) and reduced instruction set computer (RISC) architectures were discussed.

CISC architectures are made up of powerful command primitives. It is a bigger instruction set than that of the RISC, equating to a more powerful computer. The instruction length is variable. The pipeline handles register to memory instructions. VAX and Intel are CISC machines. CISC was the first instruction set created. At the time, computers had very little storage and a memory system that was slower than its CPU. Complex instruction sets saved time and space.

Several major advances allowed RISC architectures to become viable: more transistors per chip, caches to speed instruction fetches, increasing memory size coupled with decreasing cost, better instruction pipelining, and advanced optimizing compilers. RISC architectures possess uniform length instructions, a streamlined instruction set, simple addressing modes, and many registers. These features, and others, make RISC machines more complicated to build. This complexity makes a good optimizing compiler a must for code development. A clear advantage of RISC machines is their ability to do *two or more operations* per clock cycle of the CPU.

The efficient operation of these machines depends on the operating system making full use of the hardware and code developers making their codes amenable to streamlined pipelining and instruction and data fetching. RISC machines can attain tremendous peak performance if the pipelines are kept full.

As semiconductor density increases, it will be interesting to see what advances RISC architecture will take.

4.2.1.2 Memory. Currently, there is a large gap between CPU clock speed and memory access speed. This gap has motivated the changes in cache architecture resulting in the multilevel cache approach seen today. There are two types of memory that are semiconductor based. There is dynamic random access memory (DRAM) and static random access memory (SRAM). The 'random' in each designation refers to the ability to access memory locations in any order. DRAM is charge based. Each bit is

represented by an electric charge. DRAM must be continually refreshed and after a bit is read. DRAM offers the best price per performance. SRAM uses transistors and gates and retains memory as long as they have power. SRAM has higher access speeds but is very expensive. Small amounts of SRAM are placed on the CPU chip. Off the chip memory is made up of small amounts of SRAM called caches.

When a reference can be found in cache, it is considered a 100% hit rate. In general, a hit rate of 90% in L1 cache and a 50% in L2 cache is considered acceptable. Cache works best when a program is reading sequentially through the memory. This directly translates into code development guidelines for sequential execution when possible. A new development in cache design is the Harvard Architecture approach. In this design the L1 cache is divided into two sections: one for instructions and the other for data. This separation ensures that data are ready for use immediately after the instruction is issued. The CPU does not need to search through data to get to the next instruction and vice versa.

Increasing the bandwidth transfer rate, making caches as large as possible, and increasing the width of the memory system are other ways to improve memory performance.

4.2.1.3 Floating Point Numbers. Floating point numbers provide a wide range of values while using a fixed length of storage. Limitations include the following: there is a fixed number of places of accuracy either represented in base 2, base 16, or binary coded decimal; exponents are limited to a range that can be expressed as powers of 2, 10, or 16; and the difference between two successive numbers is not uniform. Rounding of floating point numbers is due to the need to express base 10 numbers in base 2.

4.2.1.4 Compilers. Optimizing compilers attempt to translate a higher level language into the fastest possible machine language that accurately represents the high-level language source. Given an expression in a program, the compiler will look for ways to streamline it. This may mean simplifying the code, throwing out extraneous instructions, and sharing intermediate results. Further optimizations may seek to restructure the code and actually make it grow in size while keeping the number of executed instructions at a minimum.

There are many levels on which a compiler has to work. A code developer has to pick their battles. When choosing third-party software, a buyer must be informed of what areas the code has been optimized and what trade-offs result.

4.2.1.5 Profiling. In order to determine the best way to optimize a code, its general operating profile must be determined. When looking at a profile, several parameters must be determined. The user time is the time spent in user mode. The system time measures the time spent in kernel mode. The elapsed time is defined as the actual wall clock time that has passed since the program was started. The CPU time is the total of the user and system time. Percent utilization corresponds to the ratio of elapsed time to CPU time. Average real memory utilization characterizes the program's resource requirements as it executes. Shared memory space accounts for the average amount of real memory taken by the program's machine instructions. High amounts of page faults and swaps will indicate a system choked for memory. Block input/output (I/O) operations are very time consuming.

4.2.1.6 Clutter. Clutter is defined as anything that contributes to the run time without contributing to the answer. One form of clutter consists of program elements that contribute to overhead; subroutine calls, indirect memory references, tests within loops, wordy tests, type conversions, and variables that are preserved unnecessarily are a few examples. Another form of clutter consists of program elements that restrict compiler flexibility. A few examples include indirect memory references, tests within loops, and ambiguous pointers.

4.2.1.7 Loop Optimization. Operation counting is the process of surveying a loop to understand the operation mix. The information is used to direct the tuning efforts. Techniques used to optimize loops in a code include loop unrolling, nested loop optimization, loop interchange, memory referenced optimization, blocking, and out–of-core solutions.

4.2.1.8 Parallelism. Usually, codes are parallelized with a particular architecture in mind. The code may be run on a cluster of workstations, a traditional two or four CPU single workstation, or a massively parallel machine containing thousands of processors. It is fortunate that methods used to parallelize a code for a specific architecture is transferable to another. However, the field of parallel computing has not yet reached maturity. Code developers and end users still must be aware of the latest issues to determine how best to parallelize a particular code.

4.2.1.9 Shared Memory Versus Distributed Memory Multiprocessors. There is a significant distinction between shared memory and distributed memory machines. In the shared memory machine, all processors see the same memory as one global pool. The distributed machine has memory distributed between processors. This distinction is reflected in the buying price of the two architectures. Shared memory machines are much more expensive than distributed memory machines.

4.2.1.10 Distributed Memory. On the distributed memory machine, the code developer has to explicitly arrange for data transfer between CPUs if they are to work on a single problem. It takes a fair amount of intervention to get the code to run in parallel on distributed memory machines. This becomes a hindrance when migrating current running programs to distributed memory architectures. Another factor to consider is that distributed memory machines either run very quickly for certain problems or very slowly for others. Network latency is another common problem with these systems. If the code developer has balanced the simulation so each node gets done with its portion of the problem at the same time as the other processors, then they will all need to pass data to one another at the same time. For the large, thousand-node size systems, this becomes significant. The operating system must be duplicated on each machine.

4.2.1.11 Shared Memory. Shared memory machines offer many advantages over distributed memory. A few of these advantages are listed here. They run the operating system natively; they can run any application and the multiple CPUs have access to the same memory pool. However, these advantages do not come cheaply. These systems require large buses/crossbars to feed the voracious appetites of the CPUs. For example, if a machine had 24 processors, each capable of processing 64 megabytes of data per second, it would require a bus bandwidth of 600 megabytes per second. Buses and crossbars at these parameters are very expensive. Also, the number of pipelines that a machine supports and the size and type of on chip and off chip cache are reflected in the price of these machines.

4.2.1.12 Large-Scale Parallel Computing. The motivation for this study was to learn how to achieve faster compute times for large complex simulations. The majority of problems will have to be run on parallel machines. It is of interest to know what level of speedup a particular code can achieve. Gene Ahmdahl, the architect of the IBM 360 computer, stated that the performance enhancement that is possible with a given improvement in code is limited by the amount that the improved feature is used. This statement is reflected in the benchmarking of codes on various architectures. This became the motivation for the third area of investigation in this study, the benchmarking of Fluent. The specific relevance of his statement to this study and HPC in general is discussed in more detail in Section 6.

There are two main approaches to decomposing a code for parallel application: data decomposition and control decomposition. Data decomposition involves dividing up the data in a program and distributing the responsibility for each piece to each processor. With control decomposition different processors are given different jobs to do or are assigned jobs as they become available.

The three classes of parallel architectures are as follows:

- SIMD (single instruction multiple data) with distributed memory (Stormcloud)
- MIMD (multiple instruction multiple data) with distributed memory
- MIMD (multiple instruction multiple data) with shared memory (Merope).

4.2.1.13 Message Passing Environments. One of the most basic parallel environments is the message passing language. This is a set of functions or functions calls that allows FORTRAN or C to split up a code for parallel execution. Data are dived up and passed between processors as messages. The message passing environment that was chosen for this study is MPI. This environment was and still is currently being developed by a consortium of computer vendors, application developers, and scientists. MPI is portable across a wide range of hardware architectures.

4.2.1.14 Benchmarking. Benchmarking is not an exact science due to the complexities and interdependencies of computer hardware and software. Code execution will vary when run on different architectures. Many websites, organizations, and books are devoted to issuing standard benchmarks. Users are always advised to benchmark for themselves whenever possible. Many standard codes used for benchmarking purposes can be obtained by a user. The most notable of these is LINPACK, a scientific linear algebra solver package. A user should always benchmark their code(s) as well.

More detail on benchmarking is provided in Section 6.

4.3 Hands-On Tutorial Message Passing Interface Class for Bechtel and INEEL Scientists and Engineers

The third class was a training class on the coding protocol MPI. MPI is used for parallel computing processes. This class focused on enhancing the INEEL engineers/scientists expertise in this protocol. Two HPC experts from the Pittsburgh Supercomputing Center (PSC) were invited to teach the class.

The PSC is a highly developed resource for HPC/MPI technology. The PSC is a joint effort between the following:

- Carnegie Mellon University
- University of Pittsburgh
- Westinghouse Electric Company.

Personnel from the PSC were chosen to teach this class for the following reasons. The PSC group had many years of experience in HPC and parallel computing. With this experience, they have recently *built* a National Science Foundation-funded, *terascale* computing system that is still the most powerful system for open research in the U.S. As a part of the Super Computing Science Consortium, they must help many users with diverse computational needs and backgrounds. They also possess skills in quantum chemistry, fluid dynamics, gravitational physics, crystallography, and other fields that enable them to help code developers ensure proper execution of a particular code on the massively parallel machines. This varied and deep expertise makes the PSC a very good source for parallel computing training.

The MPI class was taught over 2 days at INEEL. Twenty-two computers were setup as terminals

for class attendees to use for exercises. A few topics were discussed before moving into learning MPI and performing exercises. These topics included the following:

- **Faster Serial Machines versus Parallel Machines**—Current limitations on serial (one processor) machines include the speed of light, thermodynamics, and transistor switching times. These limitations can be circumvented with parallel techniques such as more processors and longer vector pipes.
- **Data Parallel—**This method of parallel computing is used in cases where the data can be worked on in parallel. Strengths and weaknesses include the following:
	- Only one executable
	- Does computations on arrays of data using array operators
	- Communicates using array shift or rearrangement operators
	- Good for problems with static load balancing that are array oriented SIMD machines
	- Scales transparently to different-sized machines
	- Wasted synchronization
	- Difficult to balance load
	- Good for the following:
		- − Finite element analysis
		- − Fluid dynamics
		- − Neural nets
		- − Weather modeling
		- − Image processing
		- − Math analysis.
- **Work Sharing**—Work sharing splits up the tasks to be done in parallel in contrast to data parallel where the data are split up. Strengths and weaknesses include the following:
	- Computations on loops are already distributed
	- Directive based and can be added to existing serial code
	- Limited flexibility
	- Efficiency is dependent on the structure of the existing serial code
	- May be poor with distributed memory
- Good for the following:
	- − Very large, complex, and old existing codes
	- − Already multitasked codes.
- **Load Balancing**—Load balancing is a method by which each CPU takes the same amount of time to compute a portion of the code before needing to communicate with other CPUs. If the computational load is unbalanced per CPU then one or more CPUs will be idle while waiting for other CPUs to finish their tasks. There are two forms of load balancing: static and dynamic.

Static load balancing requires the code developer to make each decision and assign a fixed amount of work to each CPU, a priori.

Dynamic load balancing can be accomplished in two ways:

- Task oriented—when one processing site finishes its task, it is assigned another one.
- Data oriented—when one processing site finishes its task before other sites, the site with the most work gives the idle site some of its data to process.

The rest of the class included specifics related to writing, compiling, submitting, monitoring, and executing a parallel code. Many class attendees brought their own codes to use for exercises in addition to the ones provided by PSC. The class was very interactive and many attendees were able to go beyond the training material.

5. INSTALLATION ISSUES WITH THE PARALLEL VERSIONS OF THE FLUENT COMMERCIAL COMPUTATIONAL FLUID DYNAMICS SOFTWARE ON SMP AND CLUSTER MACHINES

The second task of this CFRD was to properly install the SMP and cluster versions of Fluent.

Fluent Version 5.5 was purchased and installed on the SMP machine in September 2001. There were no major installation issues for the SMP machine. At this point, the HPC laboratory had an older cluster that consisted of 40 500-MHz nodes. These nodes were running Mandrake 7.2 Linux. Fluent 5.5 was installed on this cluster.

In December 2001, Fluent released the latest version of their software (Version 6.0). This was a major release and was installed on the SMP machine and everything worked correctly. In January 2002, the Stormcloud cluster was fully functional; therefore, the decision was made to move Fluent from the older cluster. The newer version was installed on Stormcloud, but it did not function correctly. The older version of Fluent was then installed on Stormcloud and it did not function correctly. Working with the Fluent technical staff to identify the issues of why Fluent would not function correctly on the Stormcloud cluster, it was determined that the version of the compiler and the compiler libraries were incompatible. Fluent had been compiled using an older version of the "gcc" compiler and its associated libraries. Fluent agreed to recompile the Linux version of Fluent 6.0 to match what the Stormcloud cluster used. Once this was complete, Fluent 6.0 installed on the Stormcloud cluster correctly.

The problem of compiler compatibility is a main issue of why some applications do not work with all Linux environments. Identifying this problem with Fluent has helped the HPC personnel correct the same incompatibility with other applications. With the input and assistance from Fluent technical personnel, the HPC personnel have been able to save time and money because the problem and solution was recognized early in the process. This learning process also applied to the Bechtel R&D group. Bechtel R&D engineers maintain their own machines, requiring expertise in workstation and cluster management.

6. BENCHMARKING RESULTS OF THE SMP ENVIRONMENT VERSES THE CLUSTER ENVIRONMENT

The third task of this CFRD was to benchmark the SMP and cluster versions of Fluent.

Various methods for measuring computer performance have been used over the years [1, 2]. Unfortunately, a universal standard for measuring computer performance does not exist. Using recommendations from literature sources, two performance metrics were chosen for this study: wall-clock time and speedup. Wall-clock time is a measure of the total time that a user would have to wait to obtain results produced by the computer code. Speedup is the time that it takes the program to execute with one processor (i.e., in our study wall-clock time) divided by the time it takes to execute with several processors (parallel processing).

A well-known law governing speedup is Amdahl's Law [1]. Amdahl's Law places an upper bound on the overall performance of a computer system executing a code by parallel processing. Amdahl's Law shows that for a given program of fixed size, using eight processors to solve the problem would at best result in an eightfold decrease in execution time and an eightfold increase in speedup (see Figure 6-1). Throughout this report, Amdahl's Law is used to describe linear speedup.

Figure 6-1. Amdahl's Law.

Figure 6-2 shows the relationship between Amdahl's Law and several different codes and machines. Parallel performance data for a Fluent code and an H code were used to compare with Amdahl's Law. Fluent code was executed on a Linux cluster distributed memory machine and a SGI 3800 SMP. The H code was executed on an SGI Power Challenge shared memory machine [3]. Appendix A contains information about Fluent code parameters. Section 3 of this report discusses the Linux Cluster and SGI Origin 3800.

Figure 6-2. Amdahl's Law performance comparison.

The results shown in Figure 6-2 suggest that other factors, besides the number of CPUs, limit speedup. These factors include algorithm complexity, network traffic, communication between CPUs, and CPU load balancing. The effects that these factors have on speedup are significant. For example, the change in speedup for each code decreases as the number of processors increases and no additional speedup is achieved using 16 processors as compared to 8 processors for the Fluent code executed on the Linux cluster.

Benchmarking of two INL machines using FLUENT 6.0 was accomplished using the performance meter included as part of FLUENT 6.0 [4]. Three case files were studied: exptn.cas, 60ft horizontal south.cas, and ovalfinvortgen.cas. A summary of the case file settings for these files is located in Appendix A. Parallel performance data results for each run are located in Appendix B. Studies were conducted from a "user's perspective." Dedicated parallel machines were not used and case results were executed from a networked MS Windows-based platform with Unix interface capabilities. Note that some studies would not be conducted on the cluster machine, unless dedicated nodes were available. This minimized skewing of results due to competition for CPU time. The dynamic load balancing capability available in Fluent 6.0 was not used because CPU speeds of the parallel machines were equal and to avoid the "time penalty" associated with load balancing. To account for statistical fluctuation, test runs were repeated three times for each case file and corresponding number of processors.

Figure 6-3, 6-4, and 6-5 show the results of the three case files when wall-clock time is plotted on a log-log plot. Results from the ovalfinvortgen and exptn cases indicate that above eight processors, the wall-clock performance degraded. The cause of this is believed to be the result of internode network communication. Case file 60ft horizontal south did not show this behavior; the reason for this is unknown.

Figure 6-3. Wall -lock performance for the ovalfinvortgen case.

Figure 6-4. Wall-clock performance for the exptn case.

Figure 6-5. Wall-clock performance for the 60-ft horizontal south case.

Figure 6-6 and 6-7 show speedup results for the three case files on each machine architecture type. Figure 6-6 shows that all three cases do not perform well on a cluster machine as compared to the shared memory machine in Figure 6-7.

Figure 6-6. Code speedup on the distributed memory machine Stormcloud.

Figure 6-7. Code speedup on the shared memory machine Merope.

Figure 6-8 is a graph of speedup showing the overall performance of the three case files and two machines relative to Amdahl's Law.

Figure 6-8. Code speedup for all cases and all architectures with Amdahl's Law.

Additional tests were conducted to evaluate the following:

- Cluster performance using the dual processor capability on each node
- Turbulence model performance
- Partition performance.

Figure 6-9 shows the wall-clock performance of the exptn.cas file using the dual processor capabilities of Stormcloud. Very little performance enhancement was observed while using the dual processor capabilities.

Figure 6-9. Wall-clock time for the exptn case, a distributed memory machine using both CPUs per node.

Appendix B contains results of the turbulence model performance. As expected, the laminar model executed with the smallest wall-clock time, while the k-ε model executed with the largest wall-clock time. Specifically, the k-ε model took approximately 60% longer to execute than the laminar model. Additionally, the k-ε model took approximately 8% longer to execute than the k-ω model. A final test to evaluate manual verses automatic partitioning was performed. The manual partitioning test case executed approximately 8% faster than the automatic partitioning.

Based on results of the parallel testing, it can be concluded that Fluent's parallel scaling characteristics depend on machine type, SMP vs. cluster, and the types of models used for simulations. Specifically, Fluent's speedup characteristics are much better on the SMP machine (Merope) as compared to the cluster machine (Stormcloud). In most cases, running simulations on Stormcloud provided the best speedup performance with a maximum number of eight processors for that particular machine. In all cases studied, running simulations on Merope with 16 processors provided the best speedup performance for that particular machine. Additionally, using turbulence models results in additional computational

expense. Finally, for the partitioning study, no significant difference in wall-clock or CPU time was noted and using dual processor nodes in the cluster machine does not provide any reduction in computational expense.

7. RESULTS OF THE COMPUTING-INTENSIVE CALCULATIONS IN SUPPORT OF THE HANFORD WASTE TREATMENT PLANT

Because the Fluent case files required to run the Hanford Waste Treatment Plant simulations were not available at the time benchmarking was performed, no parallel testing was performed for the Hanford Waste Treatment Plant simulations. However, **the Hanford Waste Treatment Plant simulations have been completed, using the information learned from this study to optimize parallel performance.**

8. FUTURE DEVELOPMENTS

One of the problems identified in this study was the sharing of computer files between INEEL engineers and scientists and Bechtel. The Bechtel engineers had to transfer the files to an ftp server that INEEL personnel could retrieve. The INEEL personnel would repeat this process in the reverse direction after the results were obtained. Sometimes data could be lost due to bad network connections.

Because of this file transfer bottleneck an additional task is being prototyped. This task will be to provide a direct connection between the HPC laboratory at INEEL and the Bechtel National researchers in San Francisco. This prototype will provide the engineers/scientists at Bechtel National the opportunity to directly connect with the HPC machines and view their results on their desktop. Once this prototype is in place, it is hoped that this can provide a solution for a broader audience. Bechtel researchers across the country could access the HPC laboratory at INEEL and take advantage of the resources that are available. It also provides one location that already has the environment in place for a HPC laboratory at a cost and time savings to the other facilities.

9. SUMMARY

Bechtel and the INEEL have benefited from this study. A specific benefit to Bechtel was the availability of high –performance, ""non-mainframe" computing for Bechtel projects that allowed for a substantial increase in onsite productivity and expansion of engineering computing tasks in reference to the WTP. Training was offered that benefited both Bechtel and INEEL engineers and scientists in the emerging, state-of-the-art HPC and MPI protocol. Assistance was given in preparation of a guideline for Bechtel to improve their computing performance on engineering software, specifically, the opportunity to determine the optimum number of processors to be applied to various classes of problems.

INEEL was able to provide INEEL engineers and scientists a parallel version of Fluent that operates on both the SMP and cluster machines. This enabled INEEL personnel to observe and test the code in both environments. An unforeseen benefit was investigation with the Fluent technical personnel to determine why Fluent would not work in our existing Linux environment. This has allowed INEEL personnel to identify this problem in other applications, saving much time and effort.

An additional benefit to both Bechtel National Inc. (BNI) and INEEL is the continuing investigation in a direct connection between the HPC laboratory and other BNI facilities. Once this is proven effective, the availability of the HPC laboratory to provide more resources and services to other facilities will continue to grow.

10. ACKNOWLEDGEMENTS

The primary author would like to thank the following:

- Pam Johnson, who spent many extra hours borrowing 22 computers from INEEL staff anywhere and everywhere for use in the MPI class. She and Andrew Shewmaker spent many nights configuring these computers and ultimately delivered seamless operation during the class. Thanks, also, to those INEEL staff that loaned the computers.
- Randall Laviolette, who graciously allowed the primary author to monopolize a terminal in his laboratory to run the WTP simulations for over a 2-month period.
- The rest of the HPC CFRD investigators for professionalism and deep expertise that leaves all affected better than when we started.

11. REFERENCES

- 1. David J. Lilja, *Measuring Computer Performance,* Cambridge University Press, Cambridge, UK 2000.
- 2. Lawrence A. Crowl, "How to Measure, Present, and Compare Parallel Performance," *IEEE Parallel and Distributed Technology*, Spring 1994, pp. 9-25.
- 3. International Numerical Relativity Group, Cross Architectural Performance of the H Code, http://jean-luc.ncsa.uiuc.edu/Codes/perf/perf.html (Accessed 7.26.2002).
- 4. Fluent Inc., "Fluent 6.0 Users Guide," 2001.

APPENDIX A

CODE PARAMETERS

Appendix A Appendix A

Code Parameters **Code Parameters**

APPENDIX B

PARALLEL PERFORMANCE DATA

Appendix B

Parallel Performance Data

Run #1

Run #2

Relaxation exchanges per iteration: 156 exchanges

Time-step updates per iteration: 0.20 updates

Time-step wall-clock time per iteration: 0.278 sec (2.5%) Time-step updates per iteration: Time-step wall-clock time per iteration: 0.278 sec
Total wall-clock time: 1380.632 sec Total wall-clock time: Total CPU time: 10988.830 sec

Memory Usage

Machine: Merope No. Nodes: NA No. CPUs: 4 File Name: "exptn.cas"

Run #1

Performance Timer for 123 iterations on 4 compute nodes Average wall-clock time per iteration: 21.731 sec Global reductions per iteration: 563 ops Global reductions per iteration: 0.000 sec (0.0%)
Global reductions time per iteration: 0.000 sec (0.0%)
Message count per iteration: 3155 messages Message count per iteration: Data transfer per iteration: 40.142 MB
LE solves per iteration: 5 solves LE solves per iteration: 5 solves LE wall-clock time per iteration: 5.861 sec (27.0%) LE global solves per iteration: 5 solves LE global wall-clock time per iteration: 0.016 sec (0.1%) AMG cycles per iteration: 8 cycles Relaxation sweeps per iteration: 826 sweeps Relaxation exchanges per iteration: 172 exchanges Time-step updates per iteration: 0.20 updates Time-step wall-clock time per iteration: 0.275 sec (1.3%) Total wall-clock time: 2672.865 sec Total CPU time: 10662.430 sec

Run #2

Run #3

Performance Timer for 123 iterations on 4 compute nodes Average wall-clock time per iteration: 21.763 sec Global reductions per iteration: 563 ops Global reductions time per iteration: 0.000 sec (0.0%) Message count per iteration: 3150 messages Data transfer per iteration: 40.142 MB

Machine: Merope No. Nodes: NA No. CPUs: 2 File Name: "exptn.cas"

Memory Usage

Machine: Merope No. Nodes: NA No. CPUs: 1 File Name: "exptn.cas"

Run #1

Memory Usage

Machine: Stormcloud No. Nodes: 16 No. CPUs: 16 File Name: "exptn.cas"

Run #1

Memory Usage

Run #3

Memory Usage

Machine: Stormcloud No. Nodes: 8 No. CPUs: 8 File Name: "exptn.cas"

Run #3

Memory Usage

Machine: Stormcloud No. Nodes: 4 No. CPUs: 8 File Name: "exptn.cas"

Run #3 (all data not available)

Process Total Memory = 267 Mbytes

Performance Timer for 123 iterations on 8 compute nodes

Memory Usage

Machine: Stormcloud No. Nodes: 4 No. CPUs: 4 File Name: "exptn.cas"

Run #3

Memory Usage

Machine: Stormcloud No. Nodes: 4 No. CPUs: 4 File Name: "exptn.cas"

Run #2

Run #3

Machine: Stormcloud No. Nodes: 1 No. CPUs: 2 File Name: "exptn.cas"

Run #1

Performance Timer for 122 iterations on 2 compute nodes Average wall-clock time per iteration: 24.250 sec
Global reductions per iteration: 575 ops Global reductions per iteration: 575 ops
Global reductions time per iteration: 0.000 sec (0.0%) Global reductions time per iteration: 0.000 sec (0.0%)
Messaqe count per iteration: 629 messages Message count per iteration: 629 message count per iteration: 629 messages and 629 mes Data transfer per iteration: 13.362 MB
LE solves per iteration: 13.362 MB LE solves per iteration: LE wall-clock time per iteration: 7.063 sec (29.1%) LE global solves per iteration: 5 solves

LE qlobal wall-clock time per iteration: 0.007 sec (0.0%) LE global wall-clock time per iteration: AMG cycles per iteration: 8 cycles
Relaxation sweeps per iteration: 618 sweeps Relaxation sweeps per iteration: 618 sweeps Relaxation exchanges per iteration: 618 sweeps 91 exchanges Relaxation exchanges per iteration: 91 exchanges per i Time-step updates per iteration: 0.20 updates
Time-step wall-clock time per iteration: 0.722 sec (3.0%) Time-step wall-clock time per iteration: 0.722 sec
Total wall-clock time: 2958.440 sec Total wall-clock time:

Total CPU time:

2958.440 sec Total CPU time:

Run #2

Run #3

Combined Usage of 2 Compute Nodes: cells faces nodes objps edges ----- ----- ----- ----- ----- Number Used: 716775 2148397 719838 404246 0

Machine: Stormcloud No. Nodes: 1 No. CPUs: 1 File Name: "exptn.cas"

Run #1

Run #2

Machine: Merope No. Nodes: 16 No. CPUs: 16 File Name: "60ft_horizontal_south.cas"

Run #1

Run #2

Machine: Merope No. Nodes: 8 No. CPUs: 8 File Name: "60ft_horizontal_south.cas"

Run #1

Run #2

Machine: Merope No. Nodes: 4 No. Processors: 4 File Name: "60ft_horizontal_south.cas"

Run #1

Run #2

Performance Timer for 200 iterations on 4 compute nodes

Machine: Merope No. Nodes: 2 No. Processors: 2 Fluent Case File: "60ft_horizontal_south.cas"

Run #1

Machine: Merope No. Nodes: 1 No. Processors: 1 Fluent Case File: "60ft_horizontal_south.cas"

Run #1

Run #2

Process Total Memory = 674 Mbytes

Machine: Stormcloud No. Nodes: 16 No. Processors: 16 Fluent Case File: "60ft_horizontal_south.cas"

Run #1

Run #2

Machine: Stormcloud No. Nodes: 8 No. Processors: 8 Fluent Case File: "60ft_horizontal_south.cas"

Run #1

Run #2

Combined Usage of 8 Compute Nodes:

Machine: Stormcloud No. Nodes: 4 No. Processors: 4 Fluent Case File: "60ft_horizontal_south.cas"

Run #1

Run #2

Process Dynamic Memory = 39 Mbytes Process Total Memory = 60 Mbytes

Machine: Stormcloud No. Nodes: 2 No. Processors: 2 Fluent Case File: "60ft_horizontal_south.cas"

Run #1

Run #2

Machine: Stormcloud No. Nodes: 1 No. Processors: 1 **Fluent Case File: "60ft_horizontal_south.cas"**

Run #1

Run #2

Combined Usage of 1 Compute Nodes:

Process Total Memory = 20 Mbytes

Machine: Merope No. Nodes: NA No. Processors: 16 Fluent Case File: "ovalfinvortgen.cas"

Run #1

Run #2

Run #1

Run #2

Performance Timer for 85 iterations on 8 compute nodes Average wall-clock time per iteration: 0.809 sec

Global reductions per iteration: 170 ops Global reductions per iteration: 170 ops
Global reductions time per iteration: 0.000 sec (0.0%) Global reductions time per iteration: Message count per iteration: 2031 messages Data transfer per iteration: 5.164 MB LE solves per iteration: 5 solves LE wall-clock time per iteration: 0.218 sec (27.0%) LE global solves per iteration: 3 solves LE global wall-clock time per iteration: 0.004 sec (0.5%) AMG cycles per iteration: 7 cycles Relaxation sweeps per iteration: 639 sweeps Relaxation exchanges per iteration: 639 sweeps Relaxation exchanges per iteration: 109 exchanges per iteration: 109 exchanges and the second values of 68.732 second Total wall-clock time:

Total CPU time: 68.732 sec Total CPU time:

Memory Usage

Combined Usage of 8 Compute Nodes:

Machine: Merope No. Nodes: NA No. Processors: 4 Fluent Case File: "ovalfinvortgen.cas"

Run #1

Memory Usage

Combined Usage of 4 Compute Nodes:

Machine: Merope No. Nodes: NA No. Processors: 2 Fluent Case File: "ovalfinvortgen.cas"

Run #1

Memory Usage

Combined Usage of 2 Compute Nodes:

Machine: Merope No. Nodes: NA No. Processors: 1 Fluent Case File: "ovalfinvortgen.cas"

Run #1

Memory Usage

Combined Usage of 1 Compute Nodes:

Machine: Stormcloud No. Nodes: 16 No. Processors: 16 Fluent Case File: "ovalfinvortgen.cas"

Run #1

Memory Usage

Machine: Stormcloud No. Nodes: 8 No. Processors: 8 Fluent Case File: "ovalfinvortgen.cas"

Run #1

Memory Usage

Machine: Stormcloud No. Nodes: 4 No. Processors: 4 Fluent Case File: "ovalfinvortgen.cas"

Run #1

Memory Usage

Machine: Stormcloud No. Nodes: 2 No. Processors: 2 Fluent Case File: "ovalfinvortgen.cas"

Run #1

Memory Usage

Machine: Stormcloud No. Nodes: 1 No. Processors: 1 Fluent Case File: "ovalfinvortgen.cas"

Run #1

Memory Usage

Machine: Merope No. Nodes: NA No. Processors: 4 Fluent Case File: "60ft_horizontal_south.cas" Auto Partition

RUN #1

RUN #2

RUN #3

Memory Usage

RUN #1

RUN #3

RUN #4

Machine: Merope No. Nodes: NA No. Processors: 4 Fluent Case File: "60ft_horizontal_south.cas" Laminar Model

RUN #1

RUN #2

RUN #3

Machine: Merope No. Nodes: NA No. Processors: 4 Fluent Case File: "60ft_horizontal_south.cas" k-İ Model

RUN #1

RUN #2

RUN #3

Machine: Merope No. Nodes: NA No. Processors: 4 Fluent Case File: "60ft_horizontal_south.cas" k-Ȧ Model

RUN #1

RUN #2

RUN #3

Machine: Merope No. Nodes: NA No. Processors: 4 Fluent Case File: "60ft_horizontal_south.cas" Spalart-Allmaras Model

RUN #1

Performance Timer for 200 iterations on 4 compute nodes Average wall-clock time per iteration: 19.929 sec Global reductions per iteration: 184 ops

Global reductions time per iteration: 0.000 sec (0.0%) Giobal reductions time per iteration: Message count per iteration: 2607 messages Data transfer per iteration: 46.108 MB LE solves per iteration: 6 solves LE wall-clock time per iteration: 7.918 sec (39.7%)

LE qlobal solves per iteration: 4 solves LE global solves per iteration: Le global wall-clock time per iteration: 0.038 sec (0.2%)

AMG cycles per iteration: 0.038 sec (0.2%) AMG cycles per iteration: 13 cycles
Relaxation sweeps per iteration: 126 sweeps
Relaxation exchanges per iteration: 217 exchanges Relaxation sweeps per iteration: Relaxation exchanges per iteration: 217 excl
Total wall-clock time: 23985.863 sec Total wall-clock time:

Total CPU time: 3985.863 sec Total CPU time:

RUN #2

RUN #3

INEEL Contacts

Pam L. Johnson Project Investigator NSL Laboratory Idaho National Engineering & Environmental Laboratory (208) 526-9379 ppc@inel.gov

Tom D. Foust Industrial and Materials Technology Idaho National Engineering & Environmental Laboratory (208) 526-1047 foustd@inel.gov

Kurt D. Hamman Industrial and Materials Technology Idaho National Engineering & Environmental Laboratory (208) 526-4638 hammkd@inel.gov

Bechtel National Research and Development Contacts

Jon M. Berkoe Engineering and Technology Manager Bechtel National, Inc. (415) 768-2149 jberkoe@bechtel.com

Kelly J Knight Engineering Specialist Bechtel National, Inc. (208) 522-0314 kknight@bechtel.com