Building an HPC Watering Hole for Boulder Area Computational Science

E. R. Jessup[1], H. M. Tufo[2], and M. S. Woitaszek[3]

[1] University of Colorado, Boulder, CO 80309, USA
{jessup, tufo}@cs.colorado.edu
http://www.cs.colorado.edu/∼jessup
[2] University of Colorado, Boulder, CO 80309, USA
http://www.cs.colorado.edu/∼tufo
[3] University of Colorado, Boulder, CO 80309, USA
matthew.woitaszek@colorado.edu
http://hemisphere.cs.colorado.edu/∼matthew

Abstract. Access to adequate computing resources is essential for any computational science initiative. Moreover, these resources can serve as a high-performance computing watering hole where ideas are exchanged, collaborations formed, and cycles drawn upon to support interaction. We describe our efforts to bolster computational science research and education in the Boulder area by providing a computational platform for the community. We provide an overview of the educational, research, and partnership activities that are supported by, or have resulted from, this effort.

1 Introduction

At the ICCS 2004 Workshop on Computing in Science and Engineering Academic Programs, we reported on our experiences with computational science and engineering (CS&E) education at the University of Colorado at Boulder (UCB) [1]. This CS&E educational activity is part of a larger effort to bolster computational science research and education in the Boulder area with the ultimate goal of building a world class computational science institute at UCB. However, given the extremely poor financial climate at UCB, the decreasing research funding available at the national level, and the fact that computational science is still not widely embraced in academia or industry we were required to focus on low-cost activities to lay the three key pillars required for this initiative: education, external partnerships, and research.

Our initiative is supported by an inexpensive cluster-based high-performance computing platform. This platform supports educational activities, provides cycles for research (and hence future funding), and provides to our external partners a low risk environment for evaluating new ideas and equipment. In essence, this platform serves as a high-performance computing watering hole where ideas are exchanged and collaborations are formed.

V.S. Sunderam et al. (Eds.): ICCS 2005, LNCS 3515, pp. 91–98, 2005.

Ten years ago acquiring such a platform would not have been feasible because of the prohibitive cost. However, the advent of clusters built using commodity off-the-shelf components (COTS) has driven down acquisition costs exponentially; it is now possible to acquire a one teraflop system with several terabytes (TB) of disk storage for approximately $250,000. Symmetrically, the advent of Linux and open source software has driven down the staffing costs to run these machines. It is our experience that such systems can be adequately administered by a handful of undergraduate and graduate students. Hence, total cost of ownership is quite small in an academic environment since grant overhead is supposed to pay for providing adequate research facilities, at least in principle.

In this paper, we describe our experiences with building a high-performance computing watering hole to support computational science in the Boulder area and discuss some of the outcomes of this effort. Specifically, we present our approaches to building the three pillars of our program. In section 2, we provide an overview of our computational facility, the means by which it was acquired, and plans to extend its capability. An overview of the educational activities supported by this facility are provided in section 3. In sections 4 and 5, we discuss the resulting partnerships and research activities, and note community involvement in section 6. Concluding remarks comprise section 7.

2 Computational Platform

Our computational platform consists of two compute clusters and a small storage cluster (see Fig. 1). The core platform was assembled over a two year period at a cost of approximately $300,000, with approximately half of that funding coming from an NSF ARI grant. It was constructed using a heterogeneous array of equipment with the goal of supporting as diverse a set of research areas as possible. For example, one cluster contains a high-performance interconnect for MPI parallel computing, another cluster uses less expensive gigabit Ethernet and is intended for batch processing of single-processor jobs, and the entire environment provides about 7 terabytes of storage for data-intensive computing.

We acquired our first cluster, Hemisphere, in January 2003. Hemisphere consists of 64 compute nodes and 2 head nodes. The compute nodes contain dual 2.4 GHz Intel P4 Xeon processors and 2 gigabytes (GB) of RAM. The compute nodes are connected using an 8 x 8 Dolphin torus for MPI communication and a gigabit Ethernet network for cluster control and I/O. We acquired our second cluster, Occam, in September 2004. It was donated by IBM via their Shared University Research (SUR) program and consists of 28 JS20 blades and a management node. Each blade contains two 1.6 GHz PowerPC 970 processors with 2.5 GB RAM and is connected by a gigabit Ethernet network for both MPI and support communication.

To make utilizing the computational platform as simple as possible, the compute clusters appear completely separate with one exception: all nodes on both clusters share the same file system space. Some users work only on Hemisphere or Occam, and some users work on both. As Hemisphere and Occam are configured

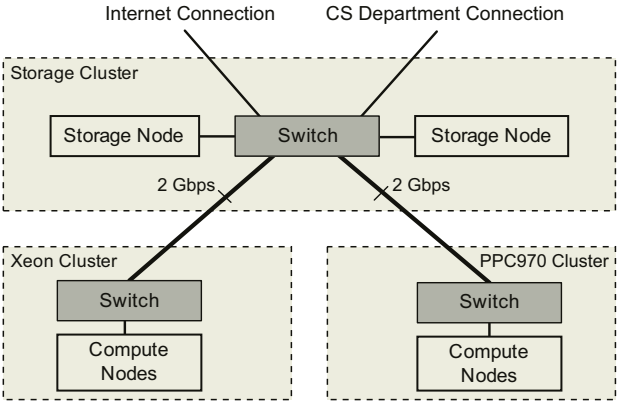

Fig. 1. UCB Computational Platform

carefully such that both systems appear similar, the shared storage allows power users to use both systems with minimal transition overhead. All users are provided with a default allocation of 1 GB in a universally accessible home directory that we backup regularly. All share an unmanaged 2.8 TB scratch workspace we call quicksand. Some users require up to 400 GB of this space, and others use none. Additional storage for large projects is available from the storage cluster provided as part of the IBM SUR grant and augmented with funds from the National Center for Atmospheric Research (NCAR). The storage cluster consists of two identical IBM x345 storage nodes, each of which provides 1.6 TB of fast Ultra320 SCSI storage. We are currently examining parallel file systems to finalize the installation with a high-performance reliable storage system.

The UCB platform is connected to NCAR via the Boulder Regional Area Network (BRAN) currently supported by OC12 fiber capable of 622 Mb/sec data transfer rates. Upgrade to gigabit Ethernet is currently underway. We use the BRAN system to connect to NCAR which will house our final computer system, CoolBlue, which is an IBM BlueGene/L (BG/L) system. This supercomputing system, to be installed in March 2005, is a one cabinet system containing 2048 compute processors arranged in an 8x16x16 3D torus, 64 I/O processors, a 4-node IBM OpenPower 720 I/O system, a p650 management node, and 8 TB of storage. At 5.7 teraflops peak (and only 12 KW) it will be our primary platform for investigating large-scale parallelism. We are currently architecting the substantial infrastructure upgrades required to seamlessly integrate CoolBlue into our computational environment.

3 HPC Educational Efforts

The Department of Computer Science at UCB was an early player in the world of CS&E education for undergraduates. Its course in high-performance scientific

computing (HPSC) was introduced in 1991. The course provided an introduction to the use of high-performance computing systems in scientific and engineering applications with undergraduate and graduate sections. Its development was supported by the National Science Foundation (NSF) under a CISE Educational Infrastructure grant awarded in 1990. The course enjoyed a successful seven year run before succumbing to technological change and a decline of student interest. The course was revived in 2002 to satisfy the increased demand for computational skills in a variety of disciplines. The source of that demand extends from academic researchers to employers in industry and the national laboratories.

To address technological developments since its first offering, we embarked on a redesign of the course. First, several structural changes were effected. To further increase the potential pool of students the numerical analysis prerequisite was removed and now appears only on the recommended course list. We moved the class project to the first semester and made HPSC a one semester course instead of two. However, as our intention was that it remain a hands-on project-based course, we kept the original four credit hour design (three hours of lecture and three hours of supervised lab per week).

In addition to structural changes, rapid changes in technology and demands from employers in industry and the national laboratories needed to be reflected in the course content and tools. Parallel programming skills are currently in high demand. As MPI is now the de facto standard for writing message-passing parallel programs we concentrate on learning MPI programming using a popular text [2]. The first half of the course is spent learning the basics of parallel programming with MPI. The remaining time is spent examining parallel architectures and algorithm development in more detail while students work on a parallel programming project of their choice. All programming and project activities revolve around using the computational platform. To encourage large-scale simulations, benchmarking, and scaling studies, dedicated access to the entire Hemisphere cluster is available through the queue. Additional HPSC course details can be found in [1].

4 Industrial and Lab Partnerships

IBM and NCAR are two Boulder area strategic partners who share an interest in promoting and strengthening computational science. IBM has been exceedingly generous in donating equipment, research support, and providing summer student internships at T. J. Watson. NCAR has supported this effort by providing money for equipment and housing equipment that is simply too costly to run at UCB. Our relationship with NCAR has also fostered the exchange of expertise and ideas in the area of computational platform management. As NCAR has over 40 years of experience running supercomputing facilities, we have learned a tremendous amount from NCAR. In return, we have used our Linux cluster expertise to help design and administer their recent cluster acquisitions.

Perhaps the most important development is the student exchange program between UCB and NCAR. Each year a select group of HPSC students are paired

with NCAR application scientists for 10-12 weeks, typically in the summer months. This offers a wonderful opportunity for cross-pollination of ideas between the scientists and newly minted high-performance computing specialists.

5 Resulting Research and Funding Initiatives

Numerous research and funding initiatives have been fostered by our high-performance computing watering hole. We concentrate on those which have resulted in significant funding and are illustrative of our research interests in large-scale parallel computing, software tools to support HPC systems, and scientific application development. The following are brief overviews of four such research initiatives.

5.1 IBM BlueGene/L

In December 1999, IBM Research launched a multi-year and multi-disciplinary project called BlueGene. It is an ambitious project that currently involves more than 50 IBM researchers in laboratories worldwide. One of the stated goals of this project is to investigate biologically important phenomena such as protein folding. An equally important goal is to develop the next generation of Petascale high-performance computing architectures. In November 2001, IBM announced a partnership with Lawrence Livermore National Laboratory (LLNL) to build the BlueGene/L supercomputer, a new architecture for high performance parallel computing systems based on low cost, low power embedded PowerPC technology. The LLNL BlueGene/L system will have 65,536-nodes each capable of 5.6 Gigaflops peak performance. BlueGene/L has several promising characteristics relative to current Terascale systems. First, BlueGene/L's overall cost-performance ratio is expected to be about an order of magnitude less than the Earth Simulator's. Though it will appear three years after the Earth Simulator, its peak floating point rate is expected to be about 9 times higher, representing more than a factor of two improvement over what Moore's Law would predict. BlueGene/L has a very fast combining network that will be useful for broadcast and reduction operations, which are a weak point of all current large-scale clusters.

In collaboration with NCAR and the University of Colorado at Denver, we submitted a proposal to the NSF Major Research Instrumentation program to acquire a one cabinet BlueGene/L supercomputer. This collaboration involved twelve researchers at the three institutions. The primary goal of the project is to investigate and address the technical obstacles to achieving practical Petascale computing in geoscience, aerospace engineering, and mathematical applications using the IBM BlueGene/L system as the target compute platform. Specific topics include: scalable high-order methods for climate modeling, dynamic data driven wildfire modeling, high-fidelity multidisciplinary simulations of modern aircraft, and scalable domain-decomposition and multigrid solvers.

5.2 Grid

Another successful collaboration has been the creation of a prototype Grid between UCB and NCAR using the Globus Toolkit [3]. This grid supports work on the Grid-BGC project [4], which is funded by a three year NASA Advanced Information Systems Technology (AIST) grant. The objective of the Grid-BGC project is to create a cost effective end-to-end solution for terrestrial ecosystem modeling. Grid-BGC allows scientists to easily configure and run high-resolution terrestrial carbon cycle simulations without having to worry about the individual components of the simulation or the underlying computational and data storage systems. In the past, these simulations needed to be performed at NCAR because a direct connection to the 2 petabyte mass storage system was required. With the introduction of Grid technologies, this is no longer the case, and the simulations may be performed on less expensive commodity cluster systems instead. In addition to outsourcing compute jobs, Grid projects allow our students and facilities to contribute to application science.

In contributing to the Grid-BGC project we discovered that Grid-related projects are becoming increasingly popular in the academic and research community. As a class project for the previous year's HPSC course, a group of students associated with NCAR's Mesoscale Microscale Meteorology (MMM) division (outside the purview of our existing collaborative relationship) worked on grid-enabling a portion of a meteorology model for simultaneous execution on Hemisphere and a NCAR cluster operated by the MMM group. Our internal experiences with Grid computing have proven to be helpful in providing expertise, on top of our computational resources, to other research groups at UCB and NCAR.

5.3 Shared Parallel File Systems

We are currently involved in a collaborative research project with NCAR to construct and evaluate a centralized storage cluster using current commercially available parallel file system technology [5]. Traditional single-host file systems (e.g., those exported via Network File System (NFS)) are unable to efficiently scale to support hundreds of nodes or utilize multiple servers. We are performing a detailed analysis of IBM's General Parallel File System (GPFS), Cluster File Systems' (CFS) Lustre, TerraScale Technologies' TerraFS, and Argonne and Clemson's Parallel Virtual File System 2 (PVFS2) for use in an environment with multiple Linux clusters running with different hardware architectures and operating system variants.

A single shared file system, with sufficient capacity and performance to store data between use on different computers for processing and visualization, while still meeting reliability requirements, would substantially reduce the time, network bandwidth, and storage space consumed by routine bulk data replication. This would provide a more user-friendly computing environment, allowing scientists to focus on the science instead of data movement. While this study is of interest to NCAR, the experimentation will help us finalize the configuration of our own storage cluster. We are also proposing to construct a prototype distributed machine room to address issues to storage and computing in a geograph-

ically separated environment. Such capabilities will be required for integrating CoolBlue into our computational facility.

5.4 Scalable Dynamical Cores

We have been working with IBM and NCAR researchers to build a scalable and efficient atmospheric dynamical core using NCAR's High Order Method Modeling Environment (HOMME) [6]. In order for this to be a useful tool for atmospheric scientists it is necessary to couple this core to physics packages employed by the community.

The physics of cloud formation is generally simulated rather crudely using phenomenological parameterizations. The dream of modelers is the direct numerical simulation of cloud processes on a global scale. Unfortunately, this requires an increase in computational power of approximately six orders of magnitude over what is currently available. A promising alternative to improve the simulation of cloud processes in climate models is a compromise technique called Cloud Resolving Convective Parameterization (CRCP, also known as Super-Parameterization). The cost of this approach is two to three orders of magnitude more expensive than traditional parameterization techniques. However, with the advent of BlueGene/L this is now tractable. NCAR researchers have built a super-parameterization package and work is underway to couple this to HOMME. The goal is to produce an atmospheric model capable of exploiting BG/L's scalability and computational power to realize practical and scientifically useful integration rates for super-parameterized climate simulation.

With help from research staff at the IBM Watson Research Center and at NCAR we have ported HOMME and CRCP to a four rack BlueGene/L system at Watson, scaled the code to approximately eight thousand processors, and achieved sustained performance of approximately 15% of peak. Through our partnership with IBM we will have access to half of the LLNL system as it is being built this spring in Rochester, Minnesota.

6 Community Research Activities

While our computational platform has provided our research group multiple opportunities for partnerships and funded research initiatives, we are actually the smallest consumer of our computational resources. We actively seek and encourage other research groups to make use of our equipment. Through exposure via the HPSC course and other channels, several groups on campus have discovered and utilize our computational platform.

The Computer Science department's numerical analysis group runs large parallel jobs for testing a new approach to Lagrange-Newton-Krylov-Schwarz class of parallel numerical algorithms, and the systems research group runs thousands of single processor jobs to simulate the behavior of wireless networks. The Electrical and Computer Engineering department similarly runs parameter studies to simulate and analyze experimental microprocessor run-time activity. The CU Center for Aerospace Structures uses the computational platform to examine

aeroelastic optimization methodologies for viscous and turbulent flows, the Department of Molecular Cellular and Developmental Biology uses the clusters to locate cellular structures present in three-dimensional digital images generated by electron microscopes, and the Solid Earth Geophysics research group uses the clusters to develop codes for cross correlation of global passive seismic data for tomography and geophysical inversion problems. In this sense, our computational platform has evolved from a single group's private cluster to a community resource. We provide managed, reliable, computing facilities at no cost, and the community response has been encouraging.

7 Conclusions

Though this effort is only two years old we have made great strides to provide a substantial set of computational resources to the Boulder Area scientific computing community to serve as a high-performance computing watering hole. This facility is at the core of our HPC educational activities and has initiated and supported a diverse set of research projects with investigators at UCB and at our strategic partners NCAR and IBM.

Acknowledgments

Support for the the University of Colorado Computational Science Center was provided by NSF ARI Grant #CDA-9601817, NSF MRI Grant #CNS-0420873, NASA AIST grant #NAG2-1646, DOE SciDAC grant #DE-FG02-04ER63870, NSF sponsorship of the National Center for Atmospheric Research, and a grant from the IBM Shared University Research (SUR) program. We would like to especially thank NCAR and IBM for their continued support, and in particular, Rich Loft (NCAR), Al Kellie (NCAR), Christine Kretz (IBM) and Kirk Jordan (IBM).

References

1. Jessup, E.R., Tufo, H.M.: Creating a Sustainable High-Performance Scientific Computing Course. In: International Conference on Computational Science. (2004) 1242
2. Pacheco, P.: Parallel Programming with MPI. Morgan Kaufmann, San Francisco, CA (1997)
3. Alliance, T.G.: The Globus Project (2003) http://www.globus.org/.
4. Cope, J., Hartsough, C., Tufo, H.M., Wilhelmi, N., Woitaszek, M.: GridBGC: A Grid-Enabled Terrestrial Carbon Cycle Modeling System. In: Submitted to CCGrid - IEEE International Symposium on Cluster Computing and the Grid. (2005)
5. Cope, J., Oberg, M., Tufo, H.M., Woitaszek, M.: Shared Parallel Filesystems in Heterogeneous Linux Multi-Cluster Environments. In: Submitted to 6th LCI International Conference on Linux Clusters: The HPC Revolution. (2005)
6. St-Cyr, A., Dennis, J.M., Tufo, H.M., Thomas, S.J.: An Efficient Non-Conforming Spectral Element Atmospheric Model using Nonlinear Operator Integration Factor Splitting. (2004)