Interdisciplinary Cluster Computing at a Liberal Arts College

David Schaich[1*], Scott Kaplan[2], William Loinaz[1] and James Hagadorn[3]
[1]*Department of Physics, Amherst College, Amherst, MA 01002*
[2]*Department of Mathematics and Computer Science, Amherst College, Amherst, MA 01002*
[3]*Department of Geology, Amherst College, Amherst, MA 01002*
[*]*Presenter, currently at Physics Department, Boston University, Boston, MA 02215*

AAPT Topical Conference on Computational Physics for Upper Level Courses
Davidson College, Davidson, NC
27-28 July 2007

Motivation and Background

- Cutting-edge research in computational science often requires substantial computational processing and data storage.
- Computing clusters linking large numbers of “off-the-shelf” processors are an economical alternative to expensive supercomputers that permit such computationally intensive research.
- Therefore Amherst College is constructing a high-performance scientific computing cluster for interdisciplinary use in student and faculty research and training.
- The infrastructure is also accessible to the Five College Consortium of which Amherst is a part (with Hampshire, Mount Holyoke, Smith and the University of Massachusetts at Amherst).

History

- The cluster began life as a collection of PCs personally obtained by Prof. Scott Kaplan in the early 2000s.
- In 2004, Prof. Kaplan received a grant through Amherst's Faculty Research Award Program (FRAP) to purchase additional computers for the cluster, and was able to incorporate idle Macs in computer labs at Amherst College.
- In 2005, Profs. Kaplan, Loinaz and Hagadorn received an NSF grant to construct a larger high-performance cluster.
- Processors purchased through the NSF grant began to be incorporated into the cluster in early 2006.
- However, operating system upgrades made them unavailable to the cluster.

Current Status

- A full-fledged scientific computing cluster became operational in 2006, using over fifty new processors funded by the NSF. In the next year or two the cluster should reach its final size of roughly 150 such processors, possibly as many as 200.
- Since becoming operational, the cluster and has been used by researchers in six fields (computer science, physics, geology, chemistry, biology and statistics), including four undergraduate theses.
- A new computational science course is currently being developed, and will have the cluster available as a resource.

In the Curriculum

- The availability of the cluster has motivated the development of a new course in computational science, to be offered for the first time in the 2007-2008 academic year.
- A joint computer science and physics course, it will be co-taught by Prof. Lyle McGeogh (Computer Science) and Prof. David Hall (Physics).
- At least initially, the course makes only limited use of the cluster, to briefly introduce parallel processing and algorithms. The cluster's role may be expanded in the future.

Course description

Computer Science 15 / Physics 15 – Scientific Computing

This course explores how computation can be used effectively to solve problems arising in scientific disciplines. Topics include numerical integration, solving systems of equations and differential equations, root finding, the fast Fourier transform, statistical tests, random number generation, curve fitting, error analysis, and simulation of physical systems. We will emphasize ways of constructing correct, efficient algorithms and of implementing those algorithms well. No previous programming experience is required, but quantitative aptitude is essential. Students will be expected to learn the basics of programming in the first few weeks and will do substantial programming throughout the semester.

Research overview

Students and faculty in six departments have either used the cluster to perform intensive computational research, or are developing projects to run on it.

- Biology: examining evolutionary relationships between closely-related plant genera through DNA sequence analysis
- Chemistry (with Geology): modeling the dispersion of Mercury from an aging coal-fired power plant into the air and water
- Computer science: analyzing strategies for main memory management
- Geology: improving fossil visualization with X-radiographic tomography
- Physics: Monte Carlo simulations of lattice quantum field theories
- Statistics: analyzing statistical data to construct and evaluate prototype point patterns

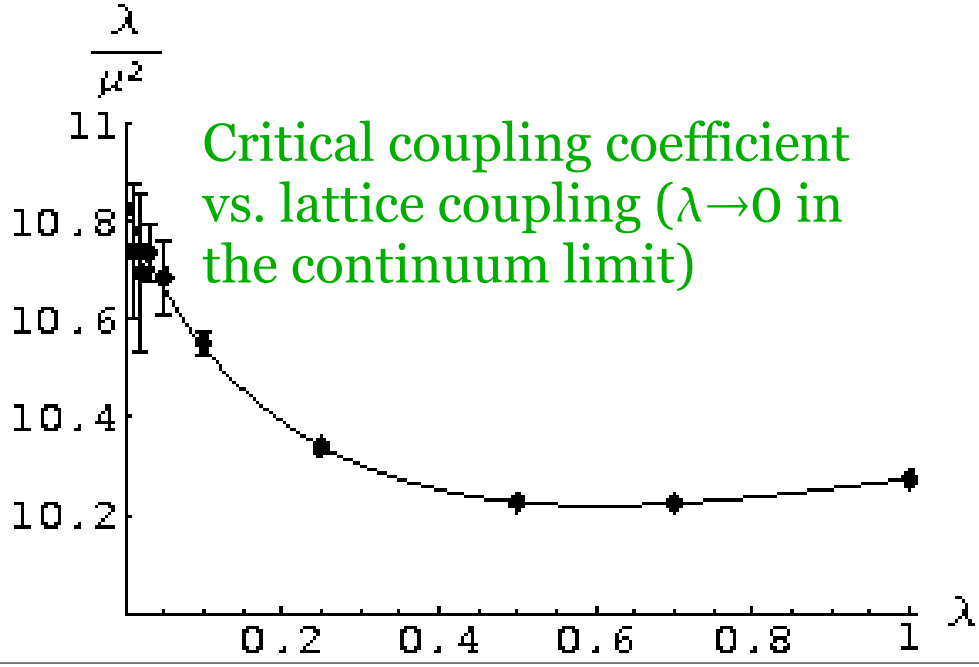
Undergraduate theses

To date, four undergraduate honors theses in three departments have been based on research performed on the cluster; students have also used the cluster for summer research.

- Owen Hoffmann (computer science), “Reference Trace Reduction via Reference Distribution Sampling”, advisor: Prof. Scott Kaplan
- Thomas Jablin (computer science), “The Effects of Hard Disk Caches on Prefetching and Clustering”, advisor: Prof. Scott Kaplan
- David Schaich (physics), “Lattice Simulations of Nonperturbative Quantum Field Theories”, physics, advisor: Prof. William Loinaz
- Joshua Shak (biology), “Phylogenetic relationships of Old World *Lycium* (Solanaceae): Reticulate evolution in the African taxa”, advisor: Prof. Jill Miller

Physics Research

- Physics research on the cluster focuses on Monte Carlo simulations of lattice quantum field theory, particularly scalar (ϕ^4) quantum field theory (Prof. William Loinaz).
- Projects focus on understanding the basic physics of these models, such as their phase structure, as well as nonperturbative features such as solitons.
- Lattice simulations are also used as a laboratory for studying improved actions and algorithms.
- Highlights include accurate measurements of the continuum critical coupling coefficient and soliton mass in ϕ^4 theory (articles in preparation).



Critical coupling coefficient vs. lattice coupling ($\lambda \to 0$ in the continuum limit)

Computer Science Research

- Computer science research on the cluster focuses on buffer cache policies and strategies for managing main memory (Prof. Scott Kaplan).

- Simulating the various ways memory management strategies can be combined, determining the effect each has on the other, allows researches to determine optimal unified main memory managers for operating system kernels, that minimize the time spent transferring data to and from the hard disk.

- The cluster has also been used to study the effects of hard disk caches on prefetching and clustering, as well as schemes to reduce reference traces without losing important information.

Geology Research

- One computational geology project (see below for another) is fossil visualization (Prof. James Hagadorn).
- This work attempts to improve the delays that occur in the process of converting attenuation projections into slice images when using X-radiographic computed tomography to visualize fossils of soft-body organisms embedded in rock.
- Such fossils of soft-body organisms are among the most important, but cannot be mechanically or chemically extracted from the surrounding rock.

Isosurface model of fossilized trilobite, from Prof. Hagadorn's Web site

Biology Research

- Biology research on the cluster examines evolutionary relationships among three closely-related genera in the tomato family (*Solanaceae*) (Prof. Jill Miller).
- The genus *Lycium* (wolfberry) is the largest of the three, with some 80 species worldwide, and has been intensely studied in recent years.
- The cluster is used to analyze DNA molecular sequence data with the goal of uncovering phylogenetic relationships among a set of *Lycium* species.

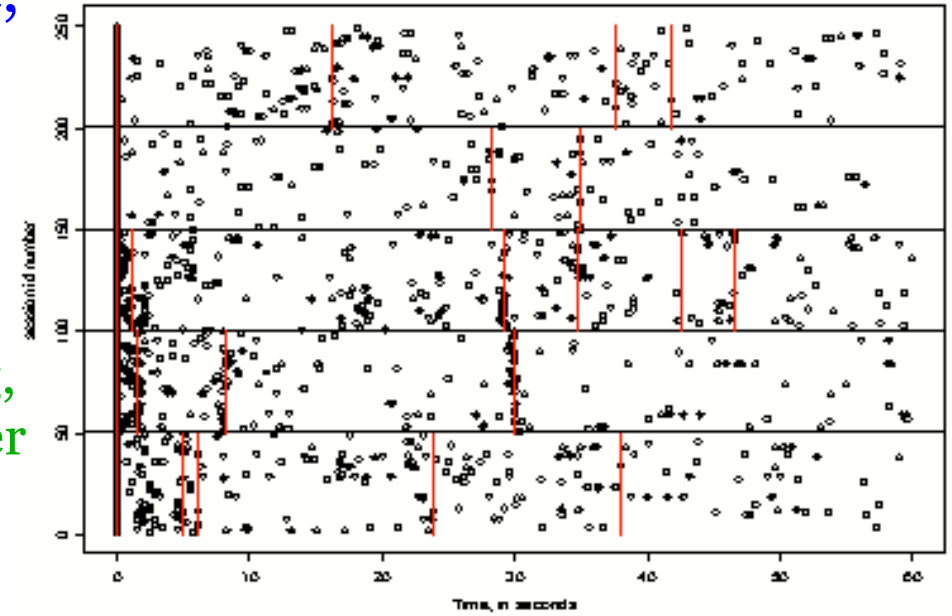
Lycium species, from Prof. Miller's Web site.

Chemistry Research

- Prof. Karena McKinney and Prof. Anna Martini (Geology) are developing a computational model for dispersion of mercury from an aging coal-fired power plant into the air and water.
- The computational capabilities of the cluster permit exploring the full 40-year historical data record, thus allowing testing and refinement of the local atmospheric chemistry and transport model.
- The power plant being studied is located at the foot of Mount Tom, roughly 15 miles from Amherst College.
- Mercury pollution, largely from coal-fired power plants, has led the Environmental Protection Agency to declare all fish in Massachusetts lakes and waterways unsafe for consumption.

Statistics Research

- Statistical research on the cluster focuses on developing and analyzing prototype point patterns.
- Such prototypes seek to identify characteristic patterns or habits through which events or individuals can be distinguished or identified.
- This can be applied, for instance, to user profiling on computer networks, as illustrated by the graph at right.



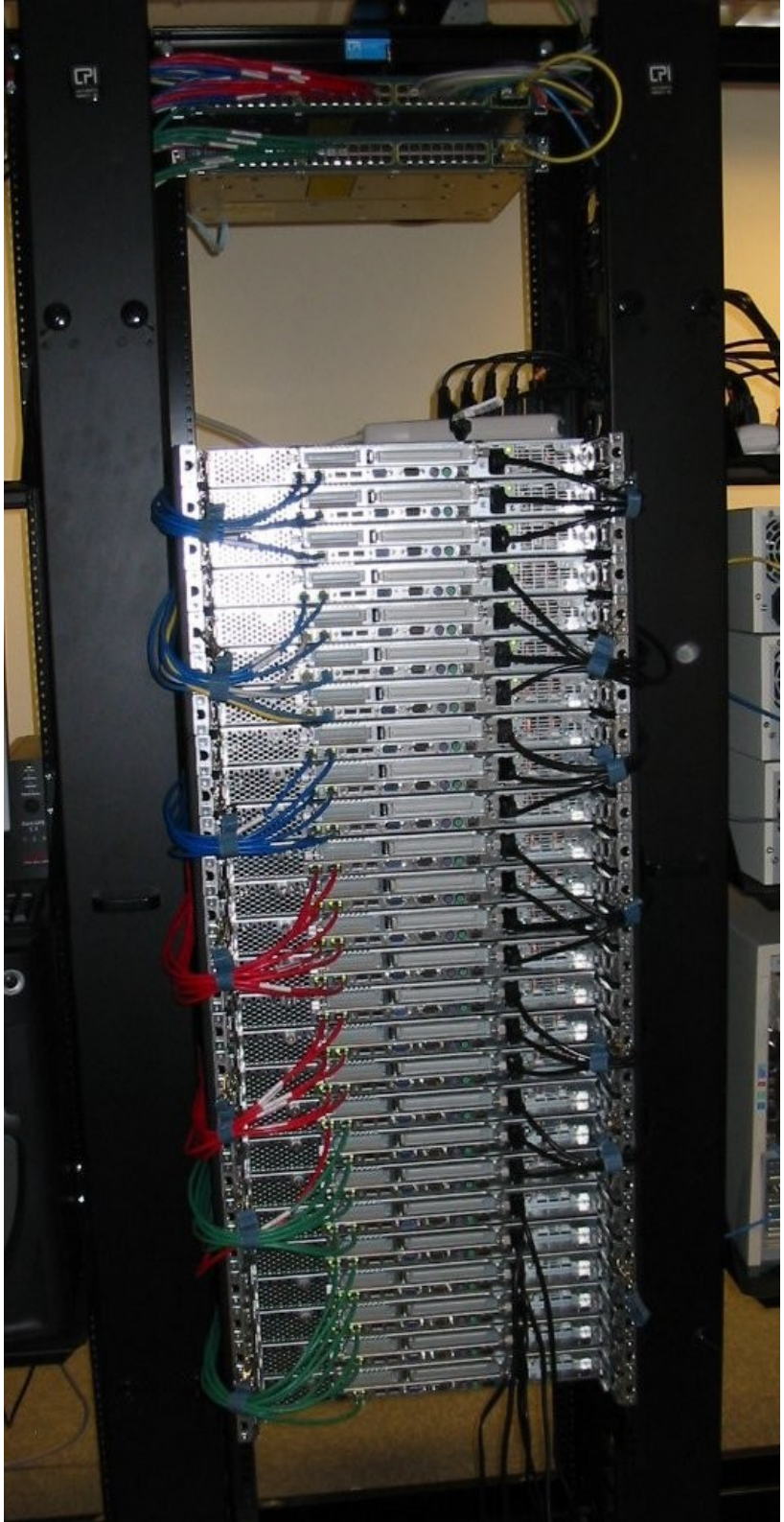
From K. E. Tranbarger and F. P. Schoenberg, "Using prototype point patterns for computer user recognition and behavior description", in review.

Hardware

- The NSF grant has been used to purchase 26 dual-processor HP DL145 G2's, each with two AMD Opteron 252 processors, 2 GB RAM, a 60 GB SATA hard disk and a 1 Gb Network Interface Card (NIC) connection. These form the core of the cluter.

- Each AMD processor is 64-bit and single-core, and clocks at 2.6 GHz using a 1 MB L2 cache.

- In addition the cluster currently includes eight single-processor 32-bit computers purchased earlier through the FRAP grant, each with 3 GHz Pentium 4 chips, 1 GB RAM, a 40 GB IDE disk and 1 Gb NIC connection.



Software

- The cluster was largely set up early in 2006, and primarily uses software from that time. The main components are:
  - Operating System: Fedora Core GNU/Linux, version 4 (www.fedoraproject.org)
  - Compiler: GCC, version 4.0.0 (gcc.gnu.org)
  - Cluster management software: Condor, version 6.6.10 (www.cs.wisc.edu/condor)
- Some of the software releases are no longer being maintained, however, and should be upgraded in the near future.

Future development

- Over the next two years the cluster should reach its full size of at least 150 processors, possibly close to 200.
- The main technical challenges that will be encountered in this expansion are supplying power to the cluster and cooling it.
- While resolving these issues may take some work, particularly setting up a powerful and reliable cooling system, both should be manageable.
- In addition, we hope to restore our ability to incorporate idle Macs in computer labs at Amherst into the cluster.
- As mentioned above, some software updates are also pending, and will likely be performed when additional processors are added to the cluster.

Acknowledgments

- Primary funding for the cluster comes from the National Science Foundation, Grant No. CNS-0521169.
- Initial funds came from a grant to Prof. Kaplan under the Faculty Research Award Program (FRAP) at Amherst College.
- The presenter was supported by a Dean's Fellowship from Boston University and a Forris Jewett Moore Fellowship from Amherst College during 2006-2007, and is currently supported by the National Science Foundation under Grant No. DGE-0221680.