Beowulf cluster computing for all: The HiPerCiC project

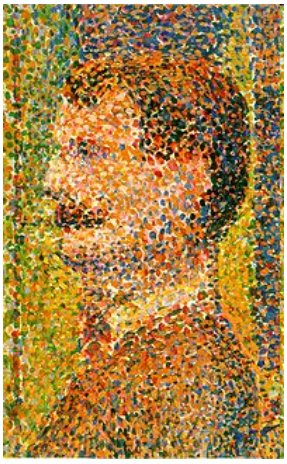
Stephanie Tanner '10, Todd Frederick '09, Jeremy Gustafson '08
Richard Brown, Project Director

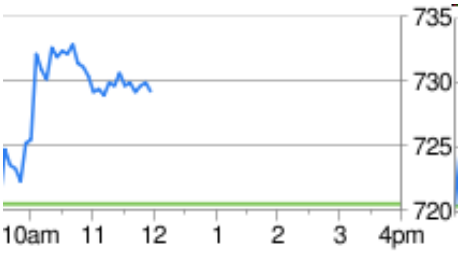
Note: This document is adapted from the contents of a poster presented in Fall, 2009. The poster may be viewed in the Regents Hall "Link" near RNS 203. -- RAB

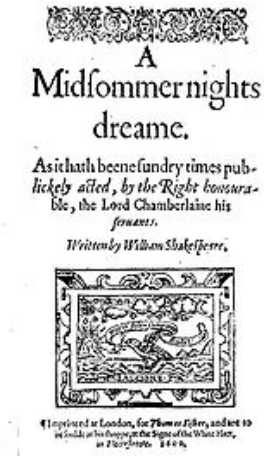
Abstract

The HiPerCiC (or ***Hi**gh **Per**formance **C**omputing **i**n the **C**lassroom*) project brings the results of Beowulf cluster computing to students and faculty who may have little or no knowledge of how to operate a computing cluster directly. In HiPerCiC, a large-scale *problem* or computing goal is identified in a target field, which may be in any academic discipline. Undergraduate research students develop programs for St. Olaf's Beowulf cluster computers that address the problem, then develop a web-based *user interface* that enables students and faculty in that target discipline to use the programs and explore the results conveniently, yielding a HiPerCiC *application*. Example HiPerCiC applications for problems in Environmental Science and in Political Science are presented.

Imagining the possibilities






A
Midſommer nights
dreame.
As it hath beene ſundry times pub-
lickely acted, by the Right honoura-
ble, the Lord Chamberlaine his
ſeruants.
Written by William Shakeſpeare.

Potential large-scale computing problems can be identified in virtually any academic discipline or combination of disciplines. The availability of powerful Beowulf cluster computing on campus at St. Olaf makes it feasible to consider ambitious investigations that might automatically have been ruled out as impractical only a few years ago. Here are some examples:

- Examine all combinations of two or three consecutive words appearing in all Shakespeare plays. Include context information such as the line or foot in which the words appear.

- Analyze the daily movement of all stock closings on the NYSE over a period of multiple years. Correlate those closing quotes with current news stories or other event streams.
- Catalog every dot in pointillist paintings by numerous artists, in order to compare the elements that contribute to interesting visual effects, or identify characteristics that distinguish the artists.

What is a Beowulf cluster?

A *computing cluster* is a system consisting of multiple networked computers that can be used together for a single computational problem. A *Beowulf cluster* is a computing cluster constructed from commodity components: ordinary computers; readily available network switches and cables; and standard open-source software. The ordinary computers of a Beowulf cluster combine to make extraordinary computations feasible.

The name "Beowulf" was chosen by the inventors of this type of cluster in honor the earliest surviving epic poem in English, with a reference to youths who become "warriors willing, should war draw near, liegemen loyal, [for] lauded deeds." (Beowulf 2007)

Riparian plants

A *riparian zone* is an area where land meets a stream, and plants that grow in a riparian zone are called *riparian plants* (Figure 1). When nitrogen-based fertilizers are used in agricultural fields near riparian zones, riparian plants may help to keep excess nitrogen from flowing into a stream and damaging its ecosystem.

Prof. John Schade and a collaborator (Schade 2006) developed a biological model describing how a riparian plant processes nitrogen (Figure 2). Tony Waldschmidt '08 programmed that model for St. Olaf's Beowulf clusters, producing millions of new data values that confirmed and extended the results in the original paper.

Figure 1: A riparian zone

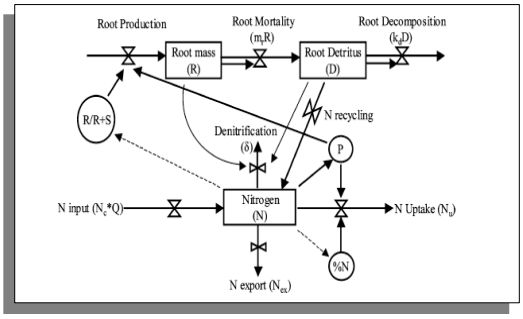


Figure 2: Biological model of nitrogen flow in a riparian plant

HiPerCiC/Riparian

The riparian-plant application described above has become the first HiPerCiC application. A prototype version of this application was created by Todd Frederick ’09 and Jeremy Gustafson ’08 in the Fall 2007 offering of CS 390, *Senior Capstone Seminar.* Stephanie Tanner ’10 rewrote the user interface and produced a complete application as a summer research project in 2009, and continues to work with Prof. Schade to refine and extend the application.

First, a professor or advanced student produces a *data set* to examine, created by Beowulf computing through an automated procedure controlled by that user (Figure 3). Then, that user and optionally other users can *explore* that data set. In the case of the Riparian application, a data set is generated by providing parameters for Prof. Schade’s model, and exploration includes graphing different combinations of the parameters and result values (Figure 4).

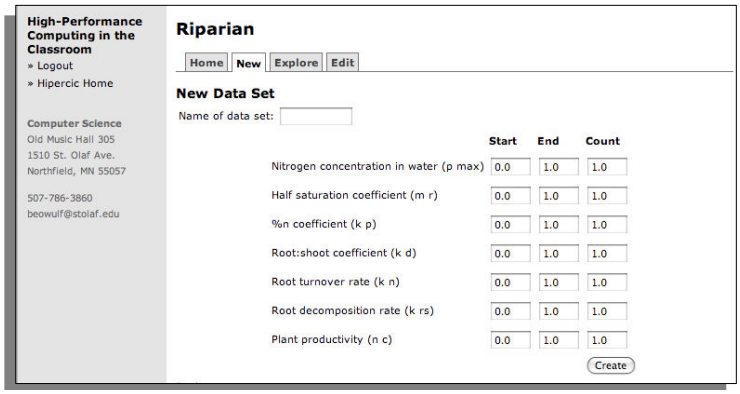


Figure 3: Creating a data set in HiPerCiC/Riparian

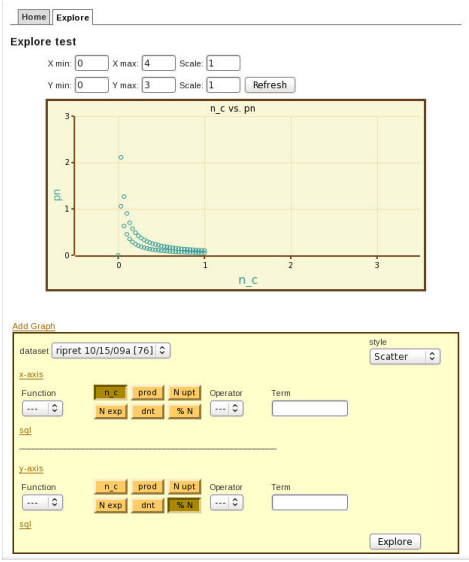


Figure 4: Exploring data in HiPerCiC/Riparian

Political Blogs

Blogs have become a formidable factor in political discourse, and have many interesting features from a political science viewpoint (e.g., anyone has access to post, no guarantee of fact checking or editorial review, rapid and wide distribution). Yet few political-science studies have been conducted to date, at least in part because it is difficult or impossible to use traditional computing methods to process the thousands of potentially significant blog pages produced each day.

Therefore, Megan Goebel '11 (co-director Christopher Chapp) is using map-reduce programming strategies (see below) on a St. Olaf's Beowulf cluster to perform political-science analysis of numerous political blogs over time, beginning with a study of approximately 400 editions of some 60 prominent liberal and conservative blogs during the 2008 election year.

HiPerCiC/Political Blogs, and beyond

Mike Holm '11 and Mary Scaramuzza '12 are developing a HiPerCiC application that will enable students and faculty in Political Science to perform their own analyses of the blog data. In this case, data sets will be generated using providing *dictionaries* or word lists that indicate some political science issue, e.g., a "horserace" dictionary indicating competitive language or a dictionary of "health care" terms. The Beowulf programming tabulates appearances of dictionary entries among the blogs, producing a data set. Students and faculty will explore those results in HiPerCiC, and will be able to download those results for further analysis with a statistics package or in a spreadsheet.

The primary technique used for our Beowulf computations on political blogs is called *map-reduce*. Google Corp. developed this strategy for processing vast quantities of data in a reasonable amount of time using computing clusters, using undergraduate-level programming. Google employs map-reduce for analyzing everything from web-page contents for its search engine to graphical images together with business information for map and GPS services. We see rich and exciting possibilities for applying powerful cluster-computing techniques creatively to other disciplines across campus, in collaboration with student researchers.

References

Beowulf Overview: Frequently Asked Questions. Downloaded Nov. 1 2009 from http://www.beowulf.org/overview/faq.html#17

Jeffrey Dean and Sanjay Ghemawat, MapReduce: Simplified Data Processing on Large Clusters. Retrieved August 22, 2008, from http://labs.google.com/papers/mapreduce.html

Schade JD, Lewis DB. 2006. Plasticity in resource allocation and nitrogen-use efficiency in riparian vegetation: Implications for nitrogen retention. *Ecosystems* 9:740-755

Acknowledgments

Funding sources Kay Winger-Blair, St. Olaf College, HHMI; faculty John Schade (Environmental Science), Chris Chapp (Political Science), Shilad Sen and Libby Shoop (Macalester Computer Science); students Mike Gesme ’10, Megan Goebel ’11, Tony Waldschmidt ’08, and Summer 2009 CS undergraduate researchers.