NEH Digital Humanities Initiative
Workshop on Supercomputing & the Humanities (July 11, 2007)

Background

The National Endowment for the Humanities has been discussing with the Department of Energy's Office of Science the creation of a grant program to award hours on DOE supercomputers to humanities scholars. The Department of Energy has offered to make one million CPU hours available for the NEH to give out via a peer review process.

The purpose of this workshop, held at the NEH on July 11, 2007, was to discuss what sorts of programs already exist that might be useful to scholars in the humanities, and to help us at the NEH get a sense of the lay of the land so that we can offer program(s) that will foster productive collaborations between humanities scholars and computer scientists. Our hope is that these programs will help build bridges between humanities scholars and the supercomputing world to enable new kinds of research.

We began with a simple diagram highlighting different levels of computer power which humanities scholars might use in conducting their research. The diagram is based on a suggestion by Chris Greer at NSF, with whom we talked as we planned this workshop. The diagram appears at the end of this document. Also appended is a list of attendees.

Morning Session 1: Examples

Demonstrations/discussion of computationally-intensive humanities projects

David Bamman (Perseus Project, Tufts University)

> Digital library uses morphological analysis, morphological disambiguation, named entity recognition, citation extraction. Training syntactic parsers is very demanding. With more parallel code could use campus clusters; work would benefit from faster computers.

David Koller (Institute for Advanced Technology in the Humanities, U-Va)

> Processing raw 3-D scan data to produce models (Michelangelo's statues) and digital analysis of 3-D models -- reassembly of fragmentary artifacts. Algorithms very CPU intensive. Now using cluster computing, but rendering and reassembly would benefit from faster computers.

Anke Kamrath (San Diego Supercomputer Center)

> Testbed for the Redlining Archives of California's Exclusionary Spaces (T-RACES): analysis of maps, banking documents, city surveys, etc. from eight California cities in the 1930s and 1940s. Other projects involving historical mapping, large-scale data analysis, visualization, etc.

Vernon Burton/Kevin Franklin (Illinois Center for Computing in Humanities, Arts, and Social Science, I-CHASS)

> Important for scholars in the fields to identify the problems and pose the questions. Currently, humanists have limited access to HPC: the infrastructure emphasis is on storage and preservation, and applications are focused on presentation, visualization of data; education, outreach and training. But dealing with large datasets (particularly video) is an issue for the humanities. Examples of datasets that would benefit from (teragrid-level or higher) HPC include public census data, digital galleries. Need for resources to handle computationally-intensive data mining. The digital humanities is larger than we realize -- more than archiving and digitization.
>
> Important to have humanities scholars integrated into the work of HPC centers, not just as followers but as people setting research questions.

Discussion

The humanities need high performance computing because its problems are sometimes more complex than the sciences. Humanists and computer scientists both need to understand the big problems and challenges of the humanities. For this, we need experimentation, and documentation of failures as well as successes. Without examples of what you can and cannot do, trying to imagine grand challenges won't be helpful, it will be too theoretical.

Note that we're not about asking the grand challenges in digital humanities, but rather the grand challenges in the humanities, for which data and computational power can be useful tools. We should be working backward -- beginning with real problems and challenges of practicing humanists.

Morning Session 2: Existing HPC Programs

Presentations by Barb Helland and Bill Kramer (Department of Energy, NERSC). Lucy Nowell and Kevin Thompson (NSF Office of Cyberinfrastructure)

Scientific Discovery Through Advanced Computing (SciDAC)
http://www.scidac.gov/

Goals
Infrastructure for scientific at the petascale. Develop new tools for large data sets.

Components
Science Applications and Partnerships, Institutes, Centers for Enabling Technology

Innovation and Novel Computational Impact on Theory and Experiment (INCITE)
http://hpc.science.doe.gov/

Provides Office of Science computing resources to a small number of computationally intensive large-scale research projects that can make high-impact scientific advances. Open to national and international researchers and industry. Peer and computational readiness reviewed.. From ca. 5M CPU hours in 2004 to 250M CPU hours in 2008. All centers have support services.

National Energy Research Scientific Computing Center (NERSC)
http://www.nersc.gov/

Currently the largest computational and storage resource on the Open Science Grid. The focus is on the "high end" and those moving to the "high end", and the goal is to make the science community more productive. Core strengths: expertise in compilers, libraries, programming models, MPI, I/O, performance analysis, debugging, and HPC software (optimizing parallel code). Able to provide expert advice on parallel computing, recommended methodologies, right algorithms and software, performance issues, etc. They don't pick the problems but work with problems brought by researchers.

Assumption is that applicants already have parallelized software; NERSC support staff are able to advise on optimization. NERSC staff are available for consultation during the proposal process.

TeraGrid
http://www.teragrid.org

TeraGrid ties together resources in NSF's high-performance computing programs and deliver those capabilities. Nine centers. A Grid Infrastructure Group manages and supports the distributed facility. Possible resources for humanists include:

TeraGrid Science Gateways
http://www.teragrid.org/programs/sci_gateways/programlist.php

NanoHub
http://www.nanohub.org/

Open Science Grid
http://www.opensciencegrid.org/

Hours awarded via xRAC Process. Development allocations up to 30K hours are meant to get a new user or group's feet wet. This is a complex facility, but an easy first step for new users. Short narrative -- the bar is very low for this first group (but very high for the larger awards, where you have to be a seasoned user).

Late Addition: World Community Grid*
http://www.worldcommunitygrid.org/

Harvest CPU cycles on thousands of PCs, and allow selected projects to run on this distributed network. Funded projects so far in biomedical research -- FightAIDS@Home, Human Proteome Folding, Fiocruz Genome Comparison, etc. -- but the organization welcomes projects in the humanities. In addition, they have offered to hold workshops with humanities scholars interested in using the world community grid as a computing resource.

* On July 25th -- a couple weeks after our workshop -- Robin Willner, IBM's Vice-President for Global Community Initiatives and one of the managers of the World Community Grid Project, met with us and expressed her interest in making World Community Grid resources available to humanists.

Afternoon Session 1: Ramping Up

People needed -- e.g. computer science grad students/post-docs -- to work with applicants to move up to cluster/grid, etc. Training for scholars needed -- classes, week-long workshops, summer sessions.

Approach from both directions: humanists need to know what's possible with access to high-performance computing; computer scientists need to know the opportunities and challenges of computing with humanities projects.

How to foster this mutual education/collaboration?

Interdisciplinary centers;
Graduate schools of library and information sciences;
Graduate programs in digital humanities (http://www.kcl.ac.uk/schools/humanities/cch/pg/);
Collaboration among funding agencies (e.g. DOE, NSF, IMLS, NEH, others)

Issues

Sustainability is an important issue, ensuring that when initial government funding goes away, the program can continue. How do you get from the departmental level to the higher-ups who may understand that longer term implications?

We need examples of successful collaborations -- to show both the humanists and the computer scientists what's possible, how their work and interests could benefit from collaboration, etc. And examples of failures will be useful too -- others will benefit from knowing what didn't work.

We need incentives for collaboration -- funding maybe, but others as well -- e.g. competitions with cool prizes to provide prestige, a different sort of incentive.

Maybe a set of four or five computer-enabled summer schools in different areas, with computer scientists and humanists -- e.g. analyzing census data, or running weather simulations and applying to historical moments.

Post-docs in computational humanities that are attached to institutes, labs, or centers, perhaps with an allotment of HPC hours. Emerging candidates are predisposed to and have a track-record in the digital humanities.

Afternoon Session 2: Where does the NEH fit in?

Short Term
Fund training programs, portals, web-based tools for easy access and could come with the hours. Educate humanists and computer scientists, foster collaborations, draw attention to examples of successful (and unsuccessful) collaborations.

Medium Term
Develop a program to create or use existing glue people to come with the hours offered by DOE. Make awards to humanists ready to use HPC

Long Term
Develop a program to fund computational humanities post-docs, maybe graduate students.

Next Steps

As a result of the workshop, we know more about what other agencies are doing and how the NEH might help humanities scholars learn more about, and take advantage of, high-performance computing resources. Starting in the next few months, the NEH will do the following:

- With DOE staff, develop a program to award the 1M CPU hours and NERSC support via a peer review process;

- Publish information about TeraGrid, Open Science Grid, World Community Grid, and other resources of potential interest to humanities scholars;

- Identify humanities/high-performance computing collaborations to highlight at conferences (raising awareness of possibilities for both humanists and computer scientists);

- Look into special training- and workshop funding options -- to fund technology training events that could include high-performance computing. Centers would apply to host these events.

- Encourage humanities scholars and high-performance computing centers to apply to NEH's Collaborative Research program (the most natural fit of our existing grant programs). The goal will be to put computer scientists and humanities scholars together to work collaboratively on humanities projects.

- Consider hosting prize competitions for a grand questions, à la the XPrizes;

- Investigate the idea of funding post-docs, either through the Challenge Grants or the Digital Humanities Fellowships program. Also consider graduate student funding.

Ramping Up:

From Desktop to Supercomputer

Desktop
Computing

$$$ +
Time +
Expertise

Campus
Cluster

$$$ +
Time +
Expertise

Grid
Computing

TERAGRID
Extensible Terascale Facility
ANL/UC
PSC
NCSA
PU
IU
ORNL
SDSC
TACC

$$$ +
Time +
Expertise

DOE Lab
Machine

Workshop on Humanities and High-Performance Computing
Old Post Office Building, Room 510A
Wednesday, July 11, 2007

David Bamman
Perseus Project, Tufts University
david.bamman@tufts.edu

Fran Berman
Executive Director, San Diego Supercomputing Center
fb@sdsc.edu

Vernon Burton
Director, Illinois Center for Computing in Humanities, Arts, and Social Science
vburton@ncsa.uiuc.edu

Kevin Franklin
Executive Director, Illinois Center for Computing in Humanities, Arts, and Social Science
kdf@uiuc.edu

Anke Kamrath
Division Director of User Services, SDSC
kamratha@sdsc.edu

David Koller
Institute for Advanced Technology and the Humanities, U of Virginia
dk@cs.stanford.edu

Marilyn Lombardi
Director, RENCI Center at Duke University
marilyn@renci.org

Keith Moo-Young
Dean of Engineering and Computer Science, California State University, Los Angeles
Keith.MooYoung@calstatela.edu

Department of Energy, NERSC

Barbara Helland
Program Manager, Office of Advanced Scientific Computing, Department of Energy
helland@ascr.doe.gov

Daniel Hitchcock
Senior Advisor, Office of Advanced Scientific Computing Research, Department of Energy
Daniel.Hitchcock@science.doe.gov

William T.C. Kramer
General Manager, NERSC, Lawrence Berkeley National Laboratory
kramer@nersc.gov

National Science Foundation

Lucille Nowell
Program Director, Office of Cyberinfrastructure, NSF
lnowell@nsf.gov

Kevin L. Thompson
Program Director, Office of Cyberinfrastructure, NSF
kthompso@nsf.gov

National Endowment for the Humanities

Brett Bobley
Chief Information Officer and Director, Digital Humanities Initiative
bbobley@neh.gov

Peter Losin
Technology Program Officer, NEH
plosin@neh.gov

Joel Schwartz
Senior Program Officer, Division of Research Programs, NEH
jschwartz@neh.gov

Beth Stewart
Digital Humanities Initiative
bstewart@neh.gov

Jennifer Serventi
Digital Humanities Initiative
jserventi@neh.gov

Jason Rhody
Digital Humanities Initiative
jrhody@neh.gov