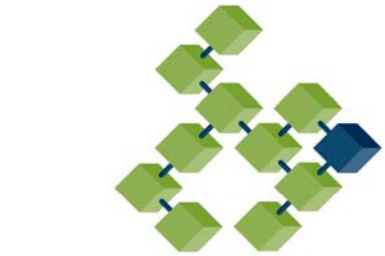

ARC CENTRE FOR
COMPLEX SYSTEMS

Technical Report

ACCS-TR-07-02

Interactively Exploring Distributed Computational Models of Biology

James Watson, Jon Kloske & Janet Wiles

November 2007

ARC Centre for Complex Systems
School of ITEE, The University of Queensland
St Lucia Qld 4072 Australia
T +61 7 3365 1003
F +61 7 3365 1533
E admin@accs.edu.au
W www.accs.edu.au

This report is also a QCIF Technical Report at
www.qcif.edu.au/research/Reports/InterModelFinalReport.pdf

Final Report

Interactively Exploring Distributed Computational Models of Biology

James Watson[1,2], Jon Kloske[3], Janet Wiles[1,3]

[1]ARC Centre for Complex Systems, [2]ARC Centre of Excellence in Bioinformatics, and
[3]School of Information Technology and Electrical Engineering

The University of Queensland

Motivation

The goal of this project was to explore the feasibility using available computing resources in student laboratories for distributed computing while retaining an interactive user interface. This work is distinct from previous efforts that have focused on distributing entire models to available resources. Funding and equipment was provided by QCIF, ITEE, ACCS and ACB.

Background

Interaction with a running computational model has proven to be a powerful method for facilitating an understanding of the model's dynamics. This is particularly true for complex systems models, where emergent behaviour is often of interest. In addition, such models form the shared language between the computational modeler and the domain expert, and allowing the domain expert to dynamically interact, visualize, and alter parameters of a simulation has proven a key factor in successful collaborations (e.g., the Neurosphere Lab project). As an added benefit, the ability to interact with a running simulation effectively allows the user to prioritize computation towards regions of interest, which may not be known *a priori*.

The main obstacle to including interactivity in biological models is their time requirements. User attention is limited to minutes at best, and large models cannot be executed within this timeframe even with the latest hardware and rigorous software optimization.

Many biological models can be broken up into smaller, independent computations that can be executed in parallel. For example, a population-based evolutionary model may include a number of computations that occur for each individual, independent of the others, before a final comparison step (such as a fitness evaluation in an evolutionary algorithm). These independent work units are small, but they are many.

Distributed computation has been chosen for a number of reasons. First, the available supercomputer facilities are unsuitable as they are shared resources, and typical wait times in the queuing systems are unsuitable for interactive modeling. Second, distributed computing, as opposed to shared-memory methods, can be scaled according to the number of available machines, and many idle machines are often available in research institutes. Finally, while scaling across many machines, distributed computing solutions can also advantage of the multi-core CPU designs that are becoming commonplace.

Existing Distributed Computing Frameworks

Three existing frameworks were identified that provide the distributed computing functionality that we require:

- Condor (http://www.cs.wisc.edu/condor/),
- Nimrod (http://www.csse.monash.edu.au/~davida/nimrod/), and
- BOINC (http://boinc.berkeley.edu/)

For the purposes of this prototype, we chose not to use Nimrod and the Grid infrastructure due to the experimental nature of the project – in particular the (initially) unknown network and CPU requirements. Condor and BOINC are ideal candidates for handling lower-level distributed functionality, such as scheduling and fault tolerance, on hardware we can control. However, a priority for the prototype was minimal impact on the existing student lab installations, since they are currently in use and are based on a disk-imaging system. A small dependency-free worker distribution was developed that did not require changes to the existing student lab system image.

Existing Infrastructure

The School of ITEE provided access to the student laboratories across three buildings. A web site showing current utilization was developed, which allowed available machines to be chosen for given simulations (see Figure 1). Approximately 400 recent desktop machines were available on a fast internal network.

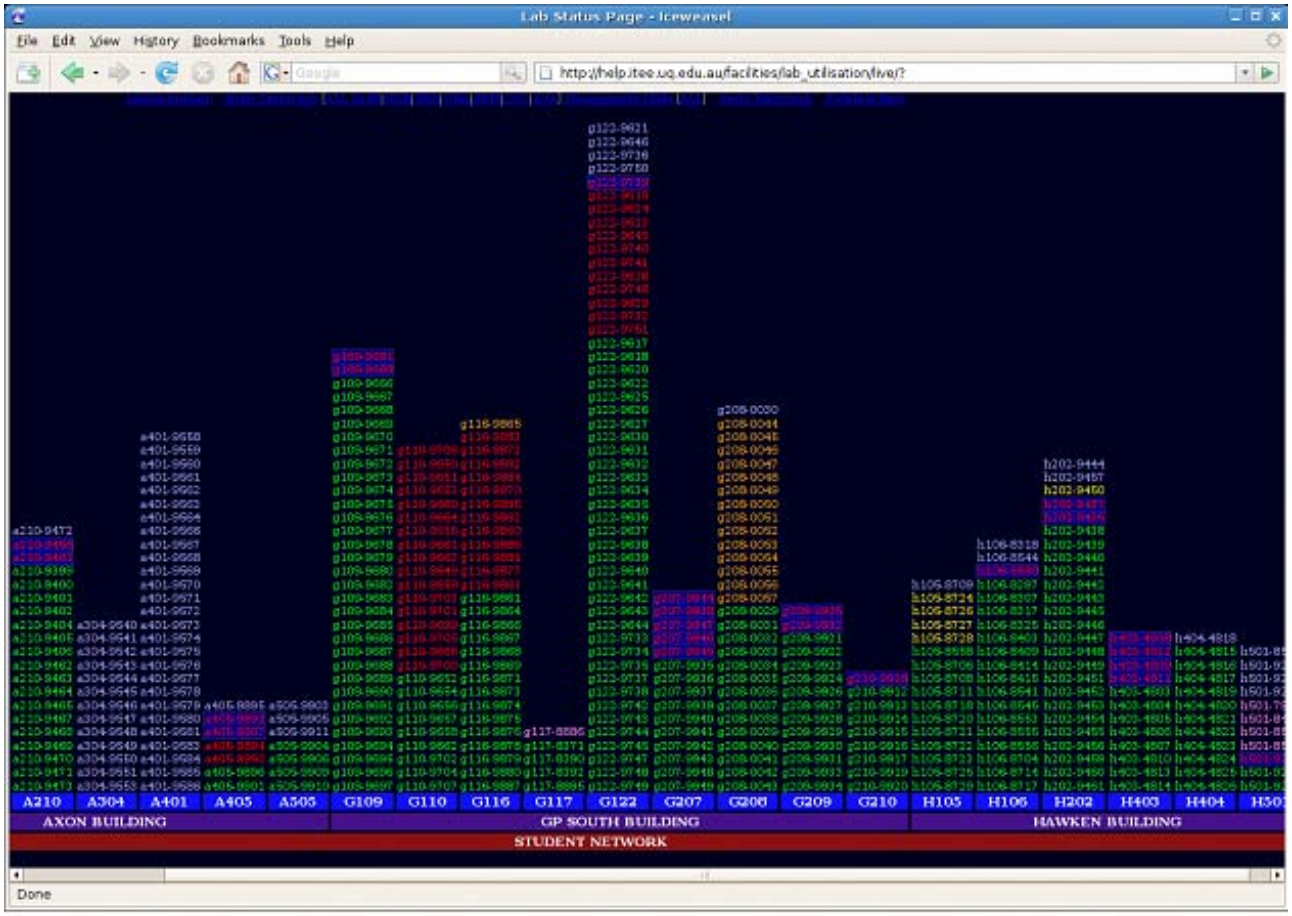

Figure 1. View of student laboratory machine and utilisation. Each column corresponds to a lab, and each label corresponds to a host name. Status (available, reimaged, logged in etc) is indicated by colours.

Distributed Genome Model

An existing monolithic computational model of evolving plant phenotypes was used as the basis for the distributed prototype. Key methods were changed into request messages sent over a TCP/IP connection, and a worker class was developed that accepted these messages, performed the computation, and returned the result to the calling process. An application controlling these workers was developed, which accepted user requests from a graphical user interface, maintained a list of requests and available workers, sent pending requests to idle workers, and updated the user display as results were returned. Figure 2 illustrates various aspects of this interface.

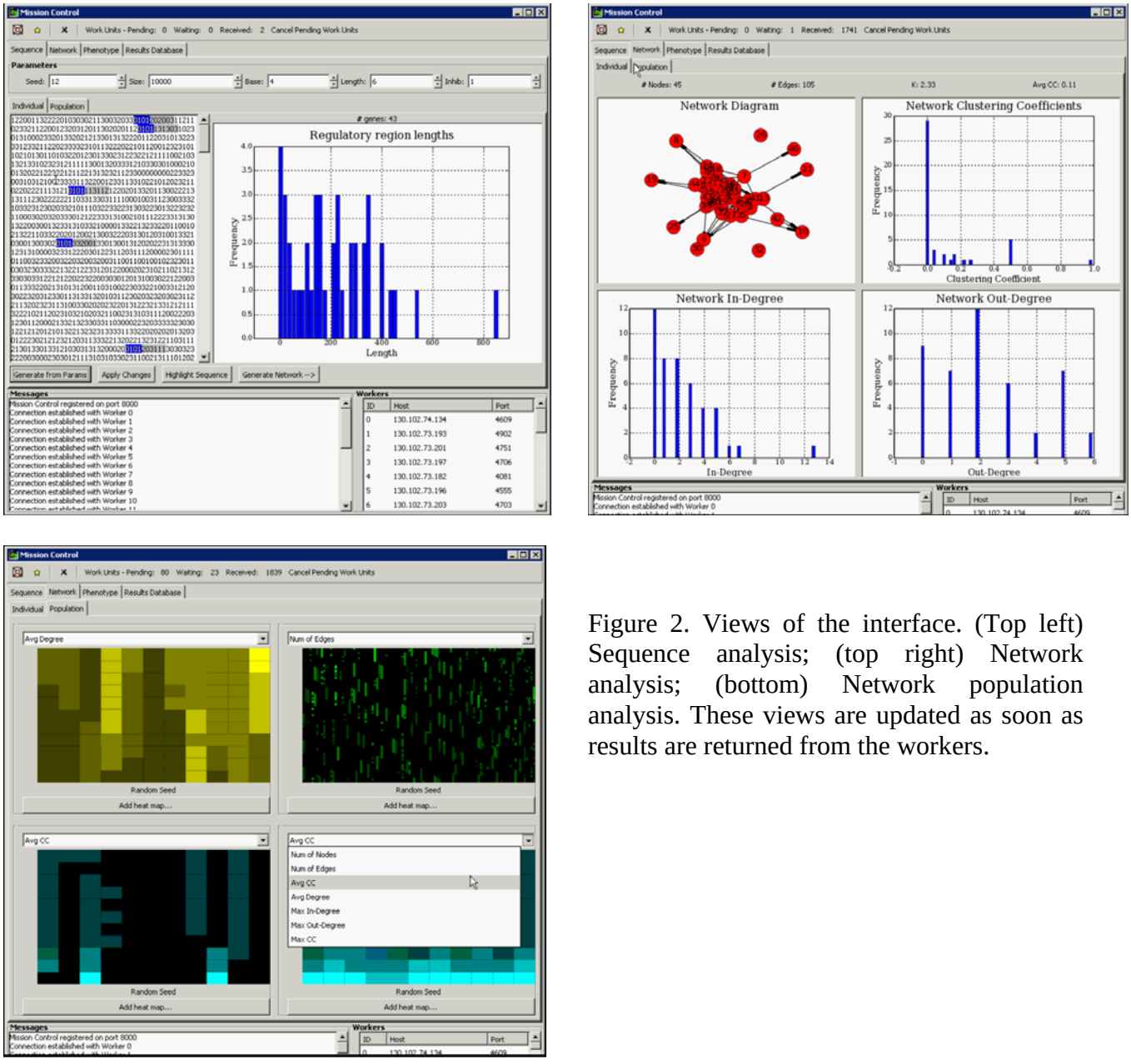

Figure 2. Views of the interface. (Top left) Sequence analysis; (top right) Network analysis; (bottom) Network population analysis. These views are updated as soon as results are returned from the workers.

Dsweep

Based on the network-event-driven system employed in the interactive prototype, a generic parameter sweep tool was developed so that arbitrary simulations could be distributed in the student labs. This has proven to be a practical tool for performing distributed parameter sweeps, with the system being used to search for regularisation parameter values in multi-purpose machine learning (a model developed by Mikael Boden), and to analyse a promoter model of *Bacillus subtilis* for transcription start site prediction (a model developed by Stefan Maetschke). Figure 3 illustrates this tool in action.

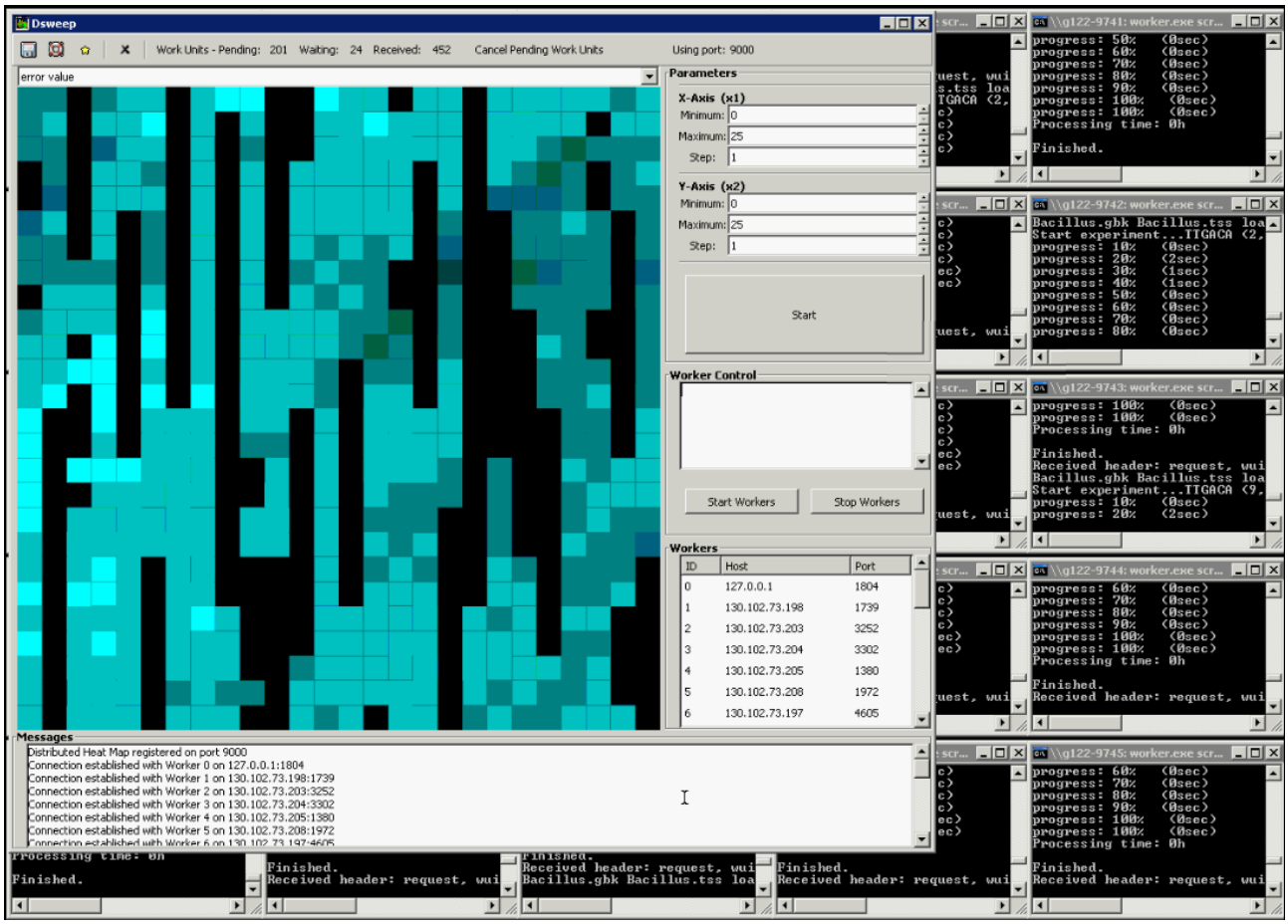

Figure 3. Distributed parameter sweep of bioinformatics models (in collaboration with IMB and ITEE researchers, Mikael Boden and Stefan Maetschke).

Conclusions and Further Work

We have successfully demonstrated dynamic interaction with a distributed computational model running in the ITEE's student labs. Significant speed-ups were gained when large simulations were performed (e.g., large populations). For single requests, the system has more overheads (costing approximately 2 seconds) than the monolithic version due to the overhead of the network communication. Distributing a biological model this way allows the researcher to be more ambitious with the size of the models they can experiment with, while retaining the benefits of dynamic interactivity.

The worker software runs on Linux and Windows, is distributed as a stand-alone binary, and has no impact on the current student image. Consequently, no administrator privileges are required to run the prototype. However, this simplicity comes at a price. Fault tolerance is only achieved through

manually re-sending requests, and there is no sophisticated scheduling algorithms that account for current machine usage. While the prototype was developed in a relatively controlled environment, security and simulation integrity are also issues that need to be addressed. The next stage is to investigate the feasibility of introducing a more mature scheduling framework.

It is expected that this work will be of interest to other research groups interested in interactive distributed computation, particularly if they have already invested in the commodity hardware required (e.g., idle machines of colleagues, student laboratories, etc.).

Publications

A description of Dsweep's application to biological modelling has been accepted for presentation at IPCAT*07:

Watson, J., Maetschke, S., & Wiles, J. Dsweep: A lightweight tool for distributed parameter sweeps. To be presented at the Seventh International Workshop on Information Processing in Cells and Tissues (28-31 August 2007, Oxford, UK).

Presentations

Presentations of the interactive distributed prototype were given on 28th June 2007 to the ARC Centre for Complex Systems, and at Complex*07.

James Watson. Interactively exploring distributed computational models of biology. Presented at the 8th Asia-Pacific Complex Systems Conference (July 2-6 2007 Gold Coast, Australia).

Contacts

James Watson
ARC Centre for Complex Systems and ARC Centre of Excellence in Bioinformatics,
The University of Queensland
jwatson@itee.uq.edu.au

Jon Kloske
School of Information Technology and Electrical Engineering,
The University of Queensland
jkloske@itee.uq.edu.au

Janet Wiles
School of Information Technology and Electrical Engineering,
The University of Queensland
j.wiles@itee.uq.edu.au